LUISS

Corso di laurea in Analisi e Misure di Marketing

Cattedra: Customer Intelligence e Logiche di Analisi dei Big Data

RELATORE Prof. Emanuele Frontoni

CORRELATORE Prof. Luca Romeo

CANDIDATO Lorenzo Nanni

# "An analytical trajectories method for understanding shoppers' buying patterns in intelligent retail environment for a Business Analysis purpose."

Anno Accademico 2022/2023

# SUMMARY

**"An analytical trajectories method for understanding shoppers' buying patterns in intelligent retail environment for a Business Analysis purpose."**

# 1. INTRODUCTION

Artificial intelligence is a branch of computer science that has been booming in recent years. It deals with the development of algorithms and systems that are capable of performing tasks that would normally require the intervention of human intelligence. Being an ever-evolving science, it is not possible to identify a founder, but at the same time it is possible to identify and give credit to the people who, in the mid-twentieth century, were already theorising about the existence of a new branch of computer science that aimed to replicate the functioning of the human mind in performing tasks.

They are John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon, who coined the term[1] for the first time in a lecture at Darmouth College; Alan Turing, who proposed the Turing test with his studies in this field, which is still in vogue today despite the rapid change in technology; Herbert Simon and Alan Newell, who first developed a programme that worked by machine learning.

With the exponential increase in the computational capacities of computers, artificial intelligence has been able to make enormous leaps forward in a very short time, so much so that nowadays it has so many facets that it is present in the everyday life of every individual, when searching on browsers, the results provided are the work of artificial intelligence algorithms that filter out the most relevant results, facial unlocking on devices exploits learning algorithms centred on the analysis of physiognomic data of faces and recognition of people.

There are countless examples that can be given in this regard, but, at this point, given the very high level of specialisation and integration of artificial intelligence in the everyday world, the question arises as to how the available tools can best be exploited.[2]

Specifically, in this thesis, the focus of the research will be on artificial intelligence devices that can be used in retail environments in such a way as to achieve intelligent retail spaces that can provide databases that allow for data analysis leading to benefits on the user side by improving their shopping experience and on the retailer side whose improvements can lead to an increase in the average consumer's receipt and thus to a higher overall turnover.

[1] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, Claude E. Shannon (1955) 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955' from https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904

[2] Emanuele Frontoni & Marina Paolanti ‘Introduction, Definition and taxonomy: AI, Machine Learning and Deep Learning' Customer Intelligence e Logiche di Analisi dei Big Data (LUISS 2023)

Going into more detail, this thesis will address the issue of analysing customer trajectories in retail environments to obtain datasets of observations that can be analysed by special data analysis programmes to deduce important information for use in business analysis.

The aforementioned topic will be addressed in its various related aspects. The focus will be on the technical research methodologies, i.e. the scripts with which to manage and interpret large datasets, from the point of view of the artificial intelligence technologies to be used, and finally from the point of view of the outputs researched through the collection of datasets and their analysis.

The important objective of this thesis is to be able to find a methodology to extrapolate results from datasets collected through guided analysis of hybrid retail environments to findings concerning consumer purchasing habits and how to set up one's own retail environment in order to increase one's business.

Artificial intelligence is increasingly present within everyday life, the amount of tasks it is able to reproduce and the intelligence with which it learns and improves over time is set to increase exponentially in the coming years. Therefore, being able to utilise the power of neural networks and sensors for a business purpose will be a fundamental skill to be able to produce an up-to-date and attractive offer to the market.

The presence of hybrid retail environments where cameras, proximity sensors, eye movement trackers, facial, emotion and behavioural recognition are added to the classical store fittings could become an everyday occurrence in the next few years, being able to have an accurate report and survey methodology already in these years could be a great source of competitive advantage.

The possibility of being able to dispose of a dataset with such a number of observations that it is possible to extend the observations obtained in hybrid environments to all retail solutions is the objective that this thesis sets itself. Being able to provide through business analysis and data analysis of the aforementioned datasets allows for a complete integration between the different information and therefore a solution that is as effective as possible and extremely anchored to analytical deductions, thus succeeding as far as possible in reducing human error or the margin left to qualitative deductions, which having no scientific reference would lead any decision towards an initially formal and when applied substantial inaccuracy.

## Main contributions

The main objective of this thesis is to create a methodology that combines tools and analysis techniques belonging to the world of data science to obtain answers that can be read from a business analysis perspective.

The retail sector in recent years has seen an increasing need to acquire information on customer behaviour and how they move around the store. The information gathered makes it possible to optimise shop layouts, improve offerings, innovate products and maximise profits by monitoring customer behaviour and choices in real time.3

[3] Contigiani Marco (2017) 'Machine vision and IoT applications in intelligent retail environments' from https://iris.univpm.it/handle/11566/245516

There is a need to research and test strategies that transform the retail space into a more advanced environment under the data customer analytics side, in fact, the point of sale is no longer the only place where shopping takes place and for this reason it must be able to quickly align itself with customer needs, transforming spaces and the shopping experience.

Through the use of integrated devices in retail environments based on artificial vision, indoor tracking systems and distributed sensors for environmental monitoring, this can be achieved.

The main focus of the research activity was therefore the creation of a survey methodology based on the use of indoor tracking sensors associated with the baskets and trolleys used by customers inside the store, which communicated in real time with an anchor in the ceiling of the store itself, in order to collect the trajectories travelled by customers and monitor their movements.

Understanding and translating the output from programmes analysing datasets from retail environments into a business perspective is a source of competitive advantage for retailers, as it enables them to adapt marketing strategies to improve the customer experience and make informed decisions on product placement and space management.

Clustering procedures are a set of algorithms that have the function of grouping individuals and objects according to their similarities, their application is of particular importance in the present study as they allow the analysis of datasets by providing a grouping of observations depending on the variable of interest studied that contains homogeneous points within it and different points between the different clusters.

The output from programmes using analysis by means of unsupervised learning algorithms can be read in a different key than a basic statistical interpretation, in fact, depending on the study to be conducted, it is possible to modify the search parameters to obtain output that allows investigations aimed at the efficiency of a company section or a particular process.

The managerial implications that will be discussed within this thesis will concern the use of the information collected and read from a business perspective for the formulation of strategies aimed at increasing the sales efficiency of the store.

The proper creation of a setup consisting of artificial intelligence devices that allow for data collection guided by the objectives sought, and a proper process through code steps on software that from the datasets obtained in the previous phase extrapolate insights and output consistent with the objective to be pursued, would offer a key to understanding observations that have never been analysed and understood until then.

# 2. STATE OF THE ART

In the following, a state-of-the-art analysis of previous research on the phenomenon will be carried out in order to obtain valuable information on previous research conducted on the subject and to be able to highlight possible research gaps that can be filled by this thesis.

Initially, the state of the art that has previously examined the role of artificial intelligence in assisting in obtaining data in the retail environment will be presented. Subsequently, previous research that has focused on topics similar to those proposed by this thesis will be analysed in order to obtain important information regarding the methodology used in conducting the analyses, the technologies used in the previous studies and, finally, the conclusions obtained from their research.

The third and final step of the state-of-the-art analysis will concern the section of studies conducted and the theoretical background of the machine learning methodologies that were used for the analysis of the two datasets.

## 2.1 LITERATURE REVIEW PART ONE

The first step will be crucial for the correct circumscription of the problem, the literature that will be included in this section will provide useful input to understand the implementation of integrated artificial intelligence solutions in retail environments: their history, their use and the problems associated with such hybrid environment solutions.

Customers must be able to do their business without being influenced by cameras or sensors, and the latter must be placed in special spaces that cannot be easily detected by customers. In order to work, the hybrid environment must have a proven balance between the part used for data collection and the classic sales part; papers that have studied the possibilities in the integration of artificial intelligence in the retail environment, the way data is collected and the ability to integrate these solutions in a hybrid environment are reported as a first step in analysing the state of the art.

Artificial intelligence will have a decisive impact on the future of retail, it is already influencing the retail sector by providing a better shopping experience and personalisation for customers.

According to Abhijit Guha, artificial intelligence will have an increasingly significant impact on retail[4]; managers should think about the implementation of hybrid retail environments by optimising factors such as the amount of value created for the customer, the study concludes by offering an interesting insight into the

[4] Abhijit Guha et al. (2021) 'How artificial intelligence will affect the future of retailing' from https://www.researchgate.net/publication/349050063_How_artificial_intelligence_will_affect_the_future_of_retailing

implication of ethical concerns involved in the use of artificial intelligence devices that could invade consumer privacy.

The use of artificial intelligence is changing the way retail shops operate; in fact, automated shops powered by artificial intelligence will be the next revolution in physical retail. The study proposed by Rajasshrie Pillaia[5] explores predictors of consumer intention to shop in AI-powered automated shops, incorporating AI-specific factors such as perceived satisfaction, adaptability to the individual user, and the ability to interact into the model of technology acceptance and usage. The results show that innovation and consumer optimism affect perceived ease and perceived usefulness, while insecurity negatively affects perceived usefulness. Perceived ease of use, perceived usefulness, perceived pleasure, personalisation and interactivity are significant predictors of consumer purchase intention in the cases analysed.

More specifically, this thesis focuses on the impact of artificial intelligence technologies on consumer behaviour in retail environments.

Another important contribution that will be mentioned in this section is the one offered by V.Kumar, in fact, the use of physical artificial intelligence devices within the store is considered to have all the implications mentioned throughout the thesis, such as data collection, creation of prediction models etc., but at the same time the interaction of customers with this new presence needs to be studied in order to exclude possible negative behavioural implications and closures[6].

The study was conducted on 231 participants who had visited a hybrid retail environment at least once, their experience with such a new configuration was evaluated through the factors of perceived ease of use, security, overall user experience, privacy, social influence and propensity to shop in shops with retail environments integrated with cameras, sensors etc.

The study reveals that factors such as privacy and ease of use were critical.

The paper just proposed was included within the literature review analysis as it offers insight into the evolution of metrics and analysis being performed in the retail world.

The retail sector is becoming more and more data-driven, the data available for quantitative analysis on customers, products and operations is becoming more and more, which leads to a strong need for the adoption of increasingly sophisticated metrics and analysis tools that are better trained to be able to exploit the large amount of data involved.

Finally, the author argues that as the retail industry is in a time of great change driven by technology, consumer tastes, economic pressures, competition and government regulations, only analysis can help retailers excel in this dynamic and ever-changing environment.

The only way to be able to embrace the big change that is taking place and to be able to turn it into a competitive advantage is to use appropriate metrics and effective analytics, to be able to make the customer

[5] Rajasshrie Pillaia et al. (2020) 'Shopping intention at AI-powered automated retail stores (AIPARS)'
[6] V.Kumar et al. (2021) ‘Transformation of Metrics and Analytics in Retailing: The Way Forward’ from https://www.emerald.com/insight/content/doi/10.1108/IJRDM-09-2020-0350/full/html

experience, customer engagement and the creation of social connections between consumers and data-driven better and more interactive.

The application of artificial intelligence in the retail sector brings with it a number of challenges and opportunities that will arise as the available technologies improve.

L.Cao with the paper 'Artificial intelligence in retail: applications and value creation logics'[7] describes how artificial intelligence can be used in the retail world, the implications that will have a greater practical development and will be more and more widespread are the interactions between salespeople and customers, warehouse management and data analysis to improve operational efficiency.

The implications of the use of artificial intelligence in the retail world will evolve hand in hand with the evolution of machine learning, natural language processing and predictive analysis techniques.

The correct, intelligent and strategic use of artificial intelligence would consist of gaining an important competitive advantage due to the ductility of the technology and the almost infinite data and information gathering possibilities.

The paper cited above attempts to answer three questions on how retailers can benefit from AI. What are the main strategies of retailers to improve AI-related data management; how do retailers use AI to provide solutions in business processes; and what are the value creation logics of AI applications in retail.

The adoption of such technologies to support retail environments will lead to a change in sales channels, allowing integration where previously there was division and sectoriality.

Robert Zimmerman in his scientific article 'Enhancing brick-and-mortar store shopping experience with an augmented reality shopping assistant application using personalised recommendations and explainable artificial intelligence'[8] looks at the future of retail and the adoption of omnichannel, which leverages digital technologies, such as augmented reality virtual assistants, to enhance the customer shopping experience.

It describes the introduction of an augmented reality application to enhance the shopping experience in physical shops; in fact, the application provides personalised recommendations to customers using artificial intelligence.

Augmented reality within the application is used to provide additional information on products, their features and prices.

The introduction of gamification (i.e. the application of game elements and mechanics in the context of retail environments to engage customers, stimulate their interest and encourage them to interact with the

[7] 'Artificial intelligence in retail: applications and value creation logics' by L.Cao (2021) from https://www.researchgate.net/publication/350133675_Artificial_intelligence_in_retail_applications_and_value_creation_logics

[8] Robert Zimmermann et al. (2022) ‘Enhancing brick-and-mortar store shopping experience with an augmented reality shopping assistant application using personalized recommendations and explainable artificial intelligence’ from https://www.emerald.com/insight/content/doi/10.1108/JRIM-09-2021-0237/full/html

brand or shop through playful dynamics) makes it an incentive for customers to visit new areas of the shop and leads to interaction with the products displayed in the shops.

## 2.2 LITERATURE REVIEW PART TWO

In the second literature review section, scientific papers will be analysed that report past cases in which artificial intelligence devices have been integrated into retail spaces to obtain data and information for analysing shopping behaviour within physical stores.

1) 'Customers' Activity Recognition in Intelligent Retail Environments' by Emanuele Frontoni, Paolo Raspa, Adriano Mancini, Primo Zingaretti & Valerio Placidi[9]

The paper under review describes the execution of a project proposing an artificial intelligence system consisting of low-cost vision systems embedded in special anchors located in the ceiling of retail shops for the purpose of analysing human behaviour and collecting statistical data for customer analysis.

The aim of the project is to be able to engage consumers in a more direct way in order to improve their shopping experience, their satisfaction and consequently the quantities they then buy.

The implemented system uses an RGBD sensor, pointed vertically at the individual, to count the number of people and their interaction with the products on the shelves.

An important achievement of the study, which represents a novelty compared to previous studies on the subject, is achieved through the use of hand movement data and their x, y, z labelled with product/non-product characteristics and the id of the person performing the interaction.

This allows the sequence of interactions to be used to represent a heatmap on the image of the shelf, which provides information on the sequence of interactions: the red areas represent products taken by consumers and subsequently put back on the shelf, while the areas marked with the colour green represent positive interactions, i.e. areas in which customers took and subsequently purchased the product.

[9] Emanuele Frontoni, Paolo Raspa, Adriano Mancini, Primo Zingaretti & Valerio Placidi (2013), 'Customers' Activity Recognition in Intelligent Retail Environments' pages 532-539 from https://link.springer.com/chapter/10.1007/978-3-642-41190-8_55#auth-Emanuele-Frontoni

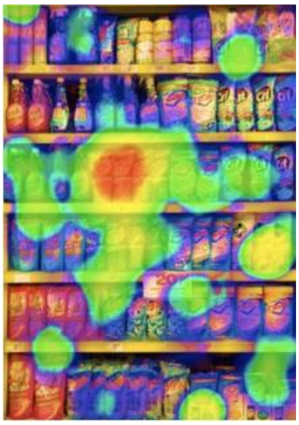

*Figure 1 - Example of a heatmap of a shelf view*

The final step of the above-mentioned project is to provide the shop with a fully automated user interaction model that can also provide information on shelf replenishment, stock-outs, crowded areas and the customer walk map within the store.

The aim is also to have a quick system to evaluate the effects of changes in shop layout, shelf layout and communication on selected observation features.

The great merit of the study is also the practical demonstration of the cost-effectiveness of the proposed solutions, to make an intelligent retail space from which to extrapolate data for business analysis purposes does not require a large economic intervention, the technology used to obtain data and measurements in the study is defined as cheap and readily available.

2) Modelling and Forecasting Customer Navigation in Intelligent Retail Environments'.
Marina Paolanti, Emanuele Frontoni, Rocco Pietrini, Daniele Liciotti and Adriano Mancini[10]

The paper was included within the state of the art of the thesis as it shows an interesting study on the consumer shopping experience within retail shops. An intelligent navigation system called sCREEN (Consumer REtail ExperieNce) is presented that uses Ultra-wideband technology to track consumer movements within retail spaces without the intervention of tags or other devices deemed invasive. The paper presents the design of an intelligent mechatronic system using Hidden Markov Models (HMM) to predict consumers' shopping choices, the representation of shoppers' attraction to shelves/categories, and usual retail scenarios such as product out-of-stocks or changes in shop layout.

On the other side of the user, i.e. from the consumer's point of view, the system can present the consumer with information on the location of the searched item, a map of the walking distance to the product location in the retail space, and the number of an aisle where the product is located

[10] Marina Paolanti, Emanuele Frontoni, Rocco Pietrini, Daniele Liciotti and Adriano Mancini (2018) ‘Modelling and Forecasting Customer Navigation in Intelligent Retail Environments' from https://link.springer.com/article/10.1007/s10846-017-0674-7

Results from a German supermarket show effectiveness of over 76%.

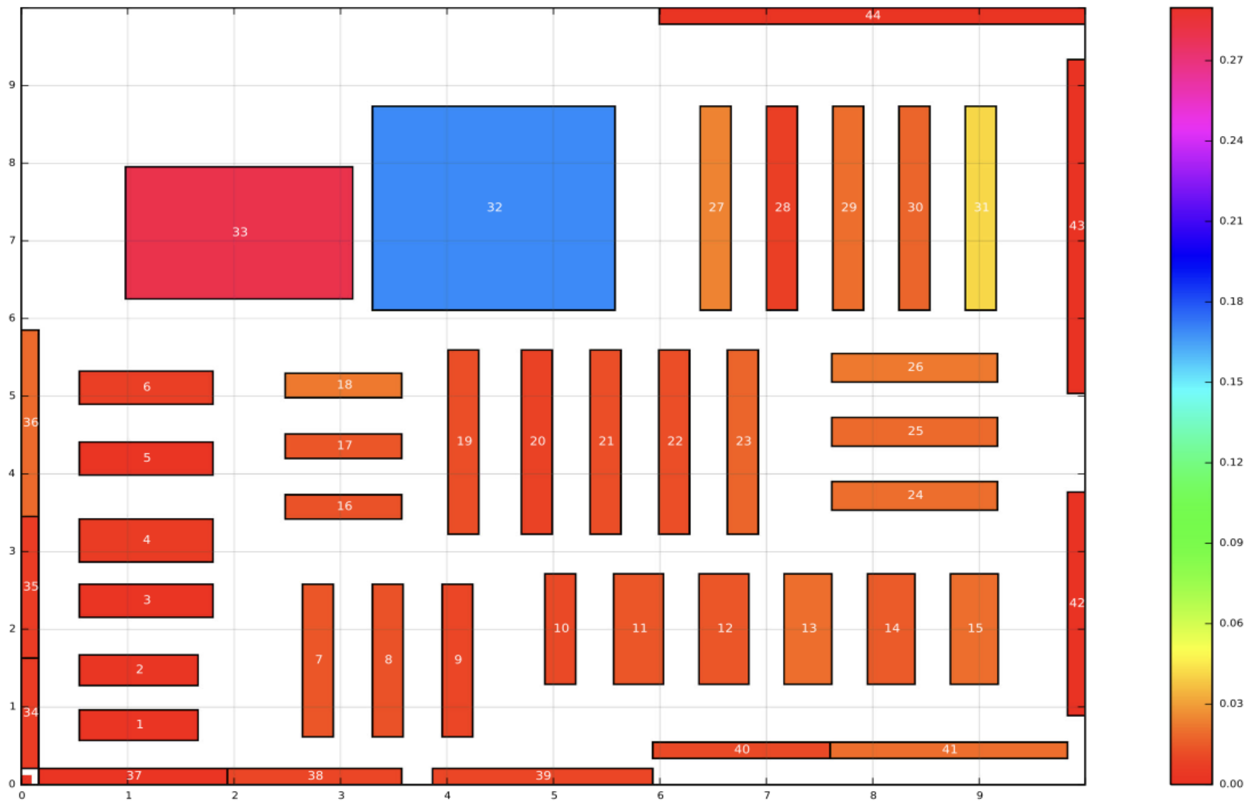

*Figure 2 Example of a store shelf heatmap obtainable via HMM*

In summary, the sCREEN system presented in the paper has a relevant impact for the purposes of this thesis, as it does not require any active labelling and/or other devices that could make the consumer's shopping experience less pleasant. The implication it has for the user is at the same time very satisfying, in fact, the dual purpose of the system allows both a data collection for internal analysis uses and an interface that can be positioned, for example, on shopping trolleys and baskets with which to offer a personalised shopping experience by suggesting paths to take, navigation for the search of a specific article and prediction of the probability of attraction to a particular shelf and/or product.

In the future, the integration of an artificial intelligence model such as sCREEN could be integrated with intelligent robotics systems to assist customers who are blind, elderly or have other impairments that affect their ability to carry out their activities in the retail environment with complete autonomy.

3) "People Detection and Tracking from an RGB-D Camera in Top-View Configuration: Review of Challenges and Applications" by Daniele Liciotti, Marina Paolanti, Emanuele Frontoni & Primo Zingaretti[11]

[11] Daniele Liciotti, Marina Paolanti, Emanuele Frontoni & Primo Zingaretti (2017) ‘People Detection and Tracking from an RGB-D Camera in Top-View Configuration: Review of Challenges and Applications’ from https://www.researchgate.net/publication/320267380_People_Detection_and_Tracking_from_an_RGB-D_Camera_in_Top-View_Configuration_Review_of_Challenges_and_Applications.

This paper is an analysis of the use of RGB-D cameras for customer tracking purposes. The investigation shows the successful use of vertically positioned RGB-D devices for counting and behaviour analysis purposes.

Within the analysis presented in this thesis, of particular interest for its intended purpose and to gain an insight into the present state of the art is section 3.2 entitled 'Intelligent Retail Environment'.

Indeed, one area of research on this topic is the detection of interactions between people and the environment, and its many applications in the fields of smart retail environments and smart shelves.

The objective of the paper, like that of this thesis, is to present a low-cost integrated system consisting of an RGB-D camera and connected software capable of monitoring buyers.

RGB-D cameras are installed above the shelves and detect the presence of people who are uniquely identified; through depth frames, the system detects the interactions of shoppers with the products on the shelf and determines whether a product is picked up or if the product is picked up and then put back down, and finally, whether or not there is contact with the products.

The operation of retail monitoring of consumer behaviour is technically explained in this way: the autonomous and low-cost system employed is based on a software infrastructure linked to a network of video sensors, with a series of computer vision algorithms embedded in RGB-D cameras distributed within the store.

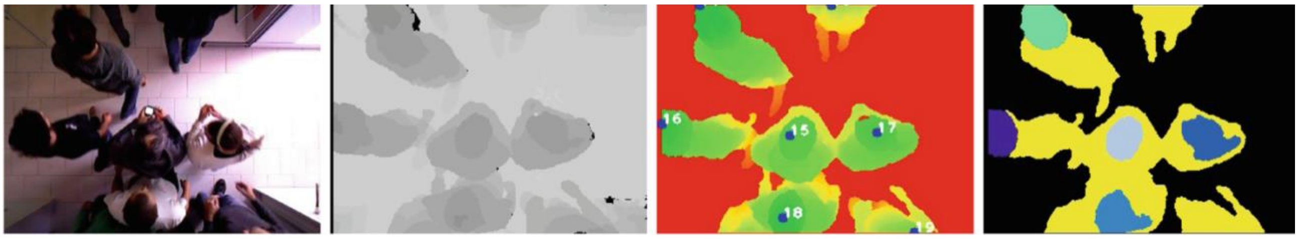

*Figure 3RGB-D camera operation*

4) 'A Deep Learning-Based System for Product Recognition in Intelligent Retail Environment' by Rocco Pietrini, Luca Rossi, Adriano Mancini, Primo Zingaretti, Emanuele Frontoni & Marina Paolanti[12]

This paper was included in the literature review section because it shows an interesting research on the construction of a pipeline that aims to recognise the products on a shelf from a photo of the shelf itself.

The pipeline consists of a first neural network that identifies individual products on the shelf and a second network, which associates the image created by the first network with a vector called embedding that describes the distinctive features of the image.

[12] Rocco Pietrini, Luca Rossi, Adriano Mancini, Primo Zingaretti, Emanuele Frontoni & Marina Paolanti (2022) 'A Deep Learning-Based System for Product Recognition in Intelligent Retail Environment' from https://www.researchgate.net/publication/360646101_A_Deep_Learning-Based_System_for_Product_Recognition_in_Intelligent_Retail_Environment

Using the embedding vector, the above-mentioned pipeline is able to measure the degree of similarity between the image it analysed and the embedding vectors in the dataset used for comparison by means of 'cosine similarity'.

The vector that is deemed to have the highest cosine similarity is labelled with an EAN code.

This paper aims to present a new pipeline for EAN code recognition of products on the shelves. A comprehensive dataset suitable for such research could lead to a correct reading of all product types within retail spaces in the future.

5) 'The objective way to detect the path to purchase by clustering shoppers' trajectories' by Marina Paolanti, Emanuele Frontoni et al[13]

The last contribution belonging to the literature review section on the testimony and study of previous literature concerning data analysis experiments using artificial intelligence devices is the most significant paper for the research to be conducted in this thesis as it addresses the same research topic as my work with the use of slightly different devices and clustering analysis.

In more detail, this paper addresses the use of technology called Ultra-WideBand (UWB) to enhance the consumer shopping experience within smart retail spaces.

This study is of great value for the sought-after objective of the thesis as it introduces the analysis of customer trajectories within the store by dividing them, as with the datasets studied in my thesis, into carts and baskets.

The methodology of conducting the research involved the clustering of the dataset to identify common subsets and trajectories to understand customer shopping behaviour within the store. The technology presented in the paper is considered non-invasive and the study conducted produces an output that if interpreted correctly can provide valuable insights to improve the shopping experience by making it personalised and tailored to the real needs of consumers.

The combination of UWB's Ultra-WideBand technology and the analysis of trolley and basket trajectories can enable a better arrangement of products on the shelves and shop layout to provide an easier and more rewarding shopping experience for consumers.

The state of the art listed above has a very specific purpose, in fact, the demonstrations that with a low budget one can make an intelligent retail space are numerous and detailed, at the same time the previous literature has provided an opportunity to explore the possibilities of new research and new ways of data collection.

[13] Marina Paolanti, Emanuele Frontoni, Daniele Liciotti, Rocco Pietrini , Adriano Mancini (2023) "The objective way to detect the path to purchase by clustering shoppers' trajectories" from https://www.researchgate.net/publication/320257107_Modelling_and_Forecasting_Customer_Navigation_in_Intelligent_Retail_Environments

The datasets used in the analysis section of the thesis collected by Grottini Lab were created through the presence of anchors embedded in the ceiling that communicated with tags embedded in shopping trolleys and baskets.

The data was sampled at a time interval of one second; in fact, every second the sensors send a signal to the anchors on the ceiling, which then allows the reconstruction of the position in which the consumer is located, the time of detection and the use of the shopping object, i.e. trolley or basket.

This data allows the calculation of further features, in fact, the second-by-second measurement of the customer's position also allows the calculation of the time the consumer stayed in a sector, the route he travelled, the chronological order in which he chose to visit the different sectors, the attendance in a sector on a particular day of the week, the preference to use the trolley or basket to visit a particular sector, the sectors that led the consumer to spend more time in them and many other possible outputs.

The technology used cannot be defined as invasive since it is completely integrated into the environment and almost not visible; all subsequent deductions in the paper will therefore be taken as actual and not influenced by different consumer behaviour since in the presence of external devices that could change the purchase intention or the journey and time within the store.

The scientific contributions mentioned above are of great relevance for the purpose of this thesis, as it was possible to see that all the analyses conducted previously used a very specific asset of technologies and tools for analysing the collected observations.

Also in this thesis, a sensor system associated with trolleys and baskets used by customers within retail environments will be used, proposing an investigation methodology that utilises inexpensive, invisible customer detection systems and data analysis software that follows and searches for a particular clustering procedure to obtain features that behave similarly to one another.

To be able to expand the analysis from a purely sampling observation to one with a view on the entire population requires a sufficient amount of data to allow us to have as little error as possible in asserting that the results obtained in circumscribed environments are as faithful and applicable to all retail spaces; to do this, the analysis I conducted had as its dataset the data obtained in two stores with an 'intelligent retail environnement' at the locations of Rome and L'Aquila.

The company that conferred and collected the data for me personally is 'Grottini Lab - Big Data science for Marketing Retail Strategy'. Through the use of sensors associated with the trolleys and baskets inside two stores, it was possible to obtain data on the trajectories of consumers within the retail environment and was able to break down the predictions obtained by aoi (area of interest)[14] .

[14] Artificial intelligence to support marketing strategies at the point of sale - Grottini Lab from https://www.grottinilab.com/it/tecnologie-servizi

## 2.3 LITERATURE REVIEW PART THREE

In the last section on state-of-the-art analysis and research, scientific papers will be proposed concerning the methodology of codes and algorithms that will be used for clustering operations on store datasets.

This section is also fundamental for the purposes of this thesis, in fact, being able to obtain a theoretical foundation of unsupervised learning methodologies is a fundamental step in being able to understand the correct functioning of data analysis and to allow a correct interpretation of the output provided by data analysis software.

In turn, the research will consist of a section on the proposal of previous literature on the topic of unsupervised learning algorithms, followed by the second section where scientific articles on k-Means and Spectral Clustering algorithms will be documented.

### K-Means Clustering

In the following, the state of the art on the use of the unsupervised clustering learning method 'K-Means' will be presented in order to theoretically introduce the clustering of store data.

Clustering operations use iterative techniques to group different observations in a dataset into clusters that possess similar characteristics.

These groupings are made because they are useful for exploring the data, identifying anomalies in the observations and finally for making estimates on the analyses to be carried out later.

Another plus that is achieved through a clustering process is the ability of clustering models to identify relationships in a dataset that may not be logically observable by simple exploration or observation.

When configuring a clustering model using the K-means method, it is necessary to specify a target number k indicating the number of centroids desired in the model. The centroid is a representative point of each cluster. The K-means algorithm assigns each input data point to one of the clusters by minimising the sum within the cluster of squares.

A more detailed explanation of how this algorithm works can be found in the chapter devoted to the explanation of data analysis methods and techniques, but in this section the focus will be shifted to the previous literature and research on unsupervised learning methodologies and more specifically on the k-Means clustering mode.

The mathematical operation of the k-Means algorithm is explained in detail within the scientific paper 'A review on k-means data clustering approach' by Shraddha Shukla and Naganna S., this paper[15] , in fact describes the operation of the algorithm in every detail, for the analysis I will conduct I will emphasise the section of the paper relating to application in data mining.

[15] Shraddha Shukla and Naganna S. (2014) 'A review on k-Means data clustering approach' from http://www. irphouse.com

Going into more detail, the article discusses the K-means clustering algorithm, a technique widely used in data mining to cluster similar data points. The article explains the basic approach of K-means clustering, in which data points are assigned to the closest cluster based on the centroid. The article also mentions the limitations and applications of the K-means clustering algorithm. Furthermore, the article provides an introduction to data mining and its importance for extracting useful information from large amounts of data.

Spectral clustering

Spectral clustering was the second clustering algorithm used for clustering the sectors belonging to the two stores.

A correct definition of this algorithm is provided by Ulrike von Luxembrug in her scientific paper 'A tutorial on spectral clustering' (2007)

The article[16] introduces the concept of spectral clustering as a modern and popular clustering algorithm that outperforms traditional algorithms such as k-Means. Spectral clustering is simple to implement and can be solved efficiently using linear algebra methods. The article provides a step-by-step introduction to the mathematical objects used by spectral clustering, including similarity graphs and Laplacians of graphs. The article then presents different approaches to the derivation of spectral clustering algorithms and explains why they work. Practical problems related to Spectral clustering are also discussed, as well as the various extensions and literature related to spectral clustering.

A further important contribution from the literature to better interpret and contextualise the Spectral clustering algorithm is the scientific paper 'On spectral clustering: Analysis and an algorithm'[17] by Ng, Andrew Y, Jordan, Michael I and Weiss, Yair.

The cited paper discusses the use of spectral clustering algorithms; these algorithms use the upper eigenvectors of a matrix derived from the distance between observations in the dataset.

However, these algorithms present unsolved problems, in fact, several authors disagree on exactly which eigenvectors to use and how to derive clusters from them, and many of these algorithms have no proof that they can actually calculate reasonable clustering. The article presents a simple spectral clustering algorithm that can be implemented via the Matlab library and uses the tools of matrix perturbation theory to analyse the algorithm. The article compares Spectral clustering algorithms with generative models and k-Means and suggests that spectral methods are a very promising alternative for conducting clustering analyses.

[16] Ulrike von Luxburg (2007) 'A tutorial on spectral clustering' from https://link.springer.com/article/10.1007/s11222-007-9033-z

[17] 'On spectral clustering: Analysis and an algorithm' Advances in neural information processing systems: pp. 849-856 by Ng, Andrew Y, Jordan, Michael I and Weiss, Yair. (2002)

# 3. DATASET DESCRIPTION

The datasets are composed of the observations collected within the two stores, the one relating to the L'Aquila store is called the Store 1 dataset while the one relating to the Rome store is called the Store 2 dataset, the number of observations relating to each dataset is listed below:

- The dataset for Store 1 consists of 732114 observations and 11 columns.
- The Store 2 dataset consists of 971653 observations and 11 columns.

The names of the 2 datasets are as follows: Store1full.csv and Store2full.csv, they are two files in.csv format, the columns are labelled with the following indices:

*id_people_rtls; description; time_start_aoi; time_end_aoi; distance_aoi; time_start_full; time_end_full; distance; sector; first_sector; last_sector.*

Once the features have been listed, we will proceed to a brief description of them.

The first feature in the dataset is based on the differentiation of the customers, in order to make the measurement carried out on a person unique, a code is generated in the feature *id_people_rtls*; in fact, each person who enters the store is associated with a random string of letters and numbers that identifies the individual by associating them with the measurements that will then be described by the subsequent features.

*Description: this* indicates the type of object used by the consumer during his or her purchase path; in particular, it determines whether the user within the store used a shopping cart or a basket.

*Time_start_aoi*: this feature is in hour format and indicates the exact time when the individual identified by the value *id_people_rtls* entered the physical study area within the retail space.

*Time_end_aoi*: this feature is in hourly format and like the feature analysed above indicates the exact time when the individual identified by the value *id_people_rtls* exited the physical study area within the retail space.

*Distance_aoi*: this feature is in numerical format; it represents the measure in metric terms of the space travelled by the individual within the area of interest.

*Time_start_full*: this feature is in time format and indicates the time of entry into the smart retail space.

*Time_end_full*: this feature, like the previous one, is in time format and in turn indicates the time of exit from the intelligent retail space.

*Distance*: This feature is in numerical format and indicates the total distance travelled by the individual within the retail space.

*Sector*: this feature indicates the area of interest (i.e. section of the store) in which the surveys are carried out.

*First_sector*: this indicates the first area of interest crossed by the individual during his or her journey within the retail space.

*Last_sector:* as for the feature mentioned above, this indicates the last area of interest traversed by the individual under investigation, within the retail space where the surveys are carried out.

From the two datasets described above, a clustering analysis will be carried out on Google Colaboratory with the aim of obtaining a survey methodology, by means of appropriate cluster analysis algorithms, aimed at defining the technical and methodological steps for the clustering of the observations obtained from the surveys within the smart retail spaces.

Both datasets are files in .csv format, which is a format used for storing tabular data; the two datasets are organised in rows and columns, each row representing a specific observation while the columns contain the various information related to that specific observation.

Before studying and analyzing the dataset, it was necessary to conduct a correlation analysis of the variables included in the model to explore the relationships between them.

Using Pearson's correlation coefficients, it is indeed possible to quantify whether the correlation present between the analyzed variables is positive (i.e., the variables tend to increase and decrease simultaneously) or is negative (i.e., the opposite of the previous situation) and also whether this correlation is to be considered strong (if it has values close to +1 or -1) or weak (with values close to 0).

Therefore, in the preliminary phase of studying the dataset, a correlation study was also carried out, using a correlation_matrix, between the distance and distance_aoi features.

This analysis is an important procedure prior to carrying out clustering processes as it allows the identification of any strongly correlated variables, in fact, should a strong correlation between two variables be discovered, they may provide redundant information for the clustering procedure.

The strong correlation between the variables could also influence the clustering procedure due to a high weighting of certain characteristics in the operation of the clustering algorithm and change the formation of the clusters.

In summary, this procedure is to be implemented both for a better management and interpretation of the observations constituting the datasets and to improve the quality of the clustering processes that will be carried out later.

Below are the correlation matrices for dataset ‘Store 1’ and dataset ‘Store 2’.

| | distance_aoi | distance |
|---|---|---|
| distance_aoi | 1.000000 | -0.007949 |
| distance | -0.007949 | 1.000000 |

*Dataset ‘Store 1’ correlation matrix*

| | distance_aoi | distance |
|---|---|---|
| distance_aoi | 1.00000 | 0.00411 |
| distance | 0.00411 | 1.00000 |

*Dataset ‘Store 2’ correlation matrix*

Correlation values range from -1 to +1; the first value, i.e., the negative one, indicates maximum negative correlation, while +1 in turn indicates perfect positive correlation: Values close to 0 indicate no correlation.

As can be deduced from these correlation matrices, there is no correlation between the variables studied, so it is possible to begin processing the dataset.

# 4. RESEARCH METHODOLOGY

This section will show in a very concise manner the dataset analysis procedures that were used to achieve the objectives discussed later in this thesis. The analysis consists of loading two .csv files containing two datasets of the Store in the city of L'Aquila and the Store in Rome respectively.

Preliminary processing enabled the reduction of outliers, correction of data types, removal of columns and normalisation of the data using the MinMax Scaler technique.

Once the preliminary phase had been carried out, the clustering algorithm 'k-Means' was initially applied for the column sector, followed by the unsupervised learning algorithm 'Spectral Clustering', again for the feature sector; finally, the two different algorithms were evaluated using the silhouette score to ascertain their accuracy and fit to the model.

Going into more detail, we can divide the clustering analysis into several sections in order to have as detailed a technical reconstruction of the entire process as possible.

## 4.1 PRACTICAL DEVELOPMENT OF THE ANALYSIS

The initial phase is the phase relating to the exploratory analysis of the dataset; it has the important purpose of enabling a deeper understanding of the type of data and variables that make up the two datasets under examination.

The exploratory analysis allows us to obtain information on the quality and mode of the collected data, this fundamental step has the great merit of showing us that there are no null values within the dataset.

The commands that were executed were *data.head()*, *data.info()* and *data.describe()*. Another very useful code step in this phase was certainly the data["id_people_rtls"].value_counts(), which made it possible to count the code strings that were associated by the artificial intelligence to the individual within the store; in fact, in order to measure the individual movements within the intelligent retail environment, sensors and cameras associated a unique code to the individual customer inside the shop, so that the person inside the shop could be tracked and a history of his movements and trajectories could be kept.

This made it possible to have a set of data that had its own intrinsic meaning; in fact, each row of the dataset does not correspond to a single individual, but the dataset was structured in such a way as to report for each row the reference sector in which the observation of the relative movement trajectory was taken, therefore the person's id was not given as a differentiating criterion. This was a step that made it possible to disassociate the person from the entire trajectory by being able to extrapolate only the observation of a particular 'sector'.

The next part corresponds to data cleaning, in fact, before being able to proceed with the standardisation of values, it is usually necessary to clean up the dataset by removing observations that contain missing values. In this precise case, as there were no missing values, the data cleaning activity focused mainly on the removal of values that did not have a useful purpose for research within the data analysis. In fact, distance values that had values below a personally established threshold were removed as they were not useful for the purposes of processing.

I then proceeded to standardise the values by creating values with a mean of 0 and variance between them equal to 1. This served to avoid some columns having values that were too far apart and prevented proper cluster construction.

In the next step, I proceeded with a 'data drop' operation to avoid overfitting the model and simplify the dataset to improve clustering.

Finally, the cleaned dataset, standardised and enriched with new columns expressing useful features for the analyses sought, was subjected to two clustering processes with different unsupervised learning algorithms.

Clustering was not carried out on all variables within the dataset but only on the feature sector, as an output analysing the clustering of the data for this variable was of interest.

The clustering algorithm used was the K-means algorithm, the number of clusters of which is not chosen automatically by this algorithm but needs to receive the relevant numerical input manually.

To overcome this problem, the elbow method was used, which allowed a precise number of clusters to be used for the clustering procedure.

## 4.2 CLUSTERING

The cluster analysis that was conducted on the two datasets is a data analysis technique that explores naturally occurring groups within a dataset, known as clusters.

Clustering is a powerful machine learning method that involves the grouping of data points. With a set of various data points, the dataset under consideration can be clustered using a clustering algorithm to classify each point within a particular group.

Indeed, data points in the same cluster contain similar characteristics or properties. On the other hand, points in separate clusters contain highly unique characteristics or properties; different from those of observations belonging to different clusters.

Clustering is an unsupervised learning method as it is a technique that does not require the use of data to train the model with associated labels or targets. In fact, clustering algorithms aim to identify the intrinsic structure within datasets by subdividing them into groups or clusters based on the similarities between observations.

## 4.3 UNSUPERVISED LEARNING

Unsupervised learning is so called because it uses machine learning algorithms for the purpose of analysing and grouping unlabeled data sets.

The above-mentioned algorithms have the function of finding hidden patterns or groupings of data without the need for manual classification and labelling.

Its ability to discover similarities and differences in datasets makes it ideal for exploratory data analysis, cross-selling strategies, customer segmentation, image recognition and trajectory analysis.[18]

The most common situations in which unsupervised learning is used can be grouped into three main tasks: grouping of observations based on similarities between data, association of observations and reduction of dataset dimensionality.[19]

There are several methods of learning this algorithm, the main ones being:

1. Grouping

Clustering, the method used for data processing within this thesis, is a data mining technique that groups observations belonging to a dataset consisting of a set of data that have not been labelled according to their similarities or differences.

Clustering algorithms are then used to process information from raw, unclassified data and reorder it into a series of internally homogeneous and heterogeneous groups representing structures or patterns in the information.

Clustering algorithms can be classified into exclusive, overlapping, hierarchical and probabilistic.

2. Exclusive and overlapping clustering

Exclusive clustering is a form of unsupervised clustering in which an observation can only belong to one cluster.

The k-Means clustering algorithm used with the store datasets is an example of exclusive clustering.

K-Means clustering is a widely used example of an exclusive clustering method, its operation consists of assigning data points to K groups, where K represents the number of clusters.

Each K-value determines the creation of a cluster and a related centroid; thus, given the dataset and the set of points consisting of the individual observations, clustering will work by grouping the data points closest to a given centroid in the same category. A high K value indicates smaller clusters and higher granularity, while a lower K value indicates larger clusters and relatively lower granularity.

[18] ‘Unsupervised Learning’ IBM from https://www.ibm.com/it-it/topics/unsupervised-learning

[19] Machine Learning and Deep Learning Tasks Supervised vs Unsupervised, Regression vs Classification' Emanuele Frontoni and Marina Paolanti

3. Hierarchical clustering

Hierarchical clustering, also known as hierarchical cluster analysis (HCA), is an unsupervised clustering algorithm that in turn can be classified into two types: agglomerative or divisive.

Agglomerative clustering works like this: data points are initially isolated as separate clusters and then iteratively joined on the basis of similarity until a single cluster is obtained.

Four different methods are used to measure data similarity:

- Ward's bond: This method states that the distance between two clusters is defined by the increase in the sum of squares after the clusters have merged.
- Average tie: The method is defined by the average distance between two points in each cluster.
- Complete binding: The method is defined by the maximum distance between two points in each cluster.
- Single bond: The method is defined by the minimum distance between two points in each cluster.

Divisive clustering can be defined as the opposite of agglomerative clustering, as it adopts a 'top-down' approach, i.e., from the top down. In this mode, a single cluster of data is divided according to the differences between the points within the dataset. These clustering processes end with a graphical visualisation of cluster creation by means of a dendrogram, a tree diagram documenting the joining or splitting of data points at each iteration.

## 4.4 FOCUS ON CLUSTERING METHOD: K-MEANS AND SPECTRAL CLUSTERING

1. k-Means algorithm

The K-means algorithm groups the data trying to separate the samples into n groups of equal variance.

This algorithm requires a manual step, i.e., that the number of clusters be specified, in fact, within my data analysis, the number of clusters obtained through the elbow method analysis was specified.

This method is widely used for its ability to adapt well to a large number of samples and has been used in a wide range of applications in numerous fields.

The k-means algorithm divides a set of N samples into K disjointed clusters, each described by the average of the samples in the cluster. The averages of the observations belonging to a single cluster are commonly called the 'centroids' of the cluster.[20]

The centroid is defined as the point belonging to the feature space that averages the distances between all the data belonging to the cluster associated with it. It thus represents the midpoint of the cluster and in general,

[20] https://scikit-learn.org/stable/modules/clustering.html

and precisely because it is calculated as the average of the observations belonging to the individual cluster, it is not one of the points in the dataset.[21]

The operation of the unsupervised K-means clustering algorithm is as follows: as mentioned above, it does not recognise the classes present in the input dataset, therefore, the first step following the algorithm is the manual decision of the number of clusters into which the input dataset is to be divided.

The number of clusters chosen to carry out the algorithm is designated by the letter K, hence the name of the K-means method. K indicates the number of clusters chosen while means refers to the use of centroids or midpoints.

Subsequently, the two datasets were subjected to the elbow method test to determine the number of K centroids belonging to the feature space into which the two datasets were to be subdivided, followed by the calculation of the distance of each point in the dataset with respect to each centroid, and consequently each point in the dataset was automatically associated with the cluster connected to the closest centroid.

The elbow method mentioned above is the most objective method for deciding the optimal number of clusters. The use of the "elbow" or "knee of a curve" as a cutoff point is a common methodology in mathematical optimisation to choose a point at which the diminishing returns in adding numbers of clusters are no longer worth the additional cost. In clustering, this means that one should choose a number of clusters such that adding another cluster does not provide better modelling of the data.

The rule behind this method is that increasing the number of clusters will naturally improve the fit (explaining greater variation), as there are more parameters (more clusters) to use, but continuing to add clusters to the algorithm will result in over-fitting, which will be visible from the elbow graph by flattening the curve.

The first clusters found will add a lot of information to the model by having to explain a lot of variation, as the data within the dataset can indeed be attributed to numerous groups, but if the number of clusters is increased too high, the number of clusters will exceed the actual number of groups that can be created with the observations within the dataset and this will lead to the greater subdivisions added being useless.

### 2. Spectral Clustering

Spectral clustering is an exploratory data analysis technique that reduces complex multidimensional data sets into clusters of similar data in a smaller number of dimensions. The aim is to group the entire spectrum of

[21] Emanuele Frontoni and Marina Paolanti Lab: machine learning in python ScikitLearn' Customer Intelligence e Logiche di Analisi dei Big Data (LUISS 2023)

unorganised data points (the eigenvalues) into different groups based on their similarity.[22] In this way, similar data, regardless of characteristics, are grouped around common points.

Spectral clustering operates as follows:

First, the similarity matrix between each pair of observations within the two datasets is created, then the k eigenvalues of the Laplacian matrix (matrix named after Pierre-Simon Laplace) are found.

The Laplacian matrix is used in spectral clustering because it relates the properties of a graph to the spectrum, i.e. the eigenvalues and eigenvectors of the matrices associated with the graph, such as the adjacency matrix or the Laplacian matrix. In practice, its task is to define the feature vector of each object.

In the last step, K-means clustering is performed on the new features found in the previous steps to group the observations of the dataset into k classes.

Spectral clustering is an unsupervised clustering technique derived from graph theory, where graphical data were used. Fortunately, the spectral clustering method is flexible and allows non-graphical data to be clustered as well.

[22] 'On Spectral Clustering: Analysis and an algorithm' Andrew Y. Ng, Michael I. Jordan, Yair Weiss

# 5. SCRIPT MODE

The datasets used contain respectively: 732,114 observations for the Rome store dataset and 971,653 observations for the L'Aquila store 2 dataset.

The datasets were analysed using the data analysis software Google Colaboratory, or 'Colab' for short, is a product of Google Research. Colab allows anyone to write and execute arbitrary Python code through the browser and is particularly suitable for developing machine learning models, data analysis and algorithm training[23]. More technically, Colab is a hosted Jupyter notebook service that does not require any configuration to be used, while providing free access to computing resources, including GPUs.

Both datasets were subjected to the same data analysis activity, i.e. they were run by the same script, since the aim of my analysis was to find interpretable differences between the different outputs, and this could only be done by obtaining the same information in order to be able to subsequently compare and interpret it.

In the following, the salient steps of the script procedure will be reported and will be interpreted in such a way as to provide a technical rationale for the presentation of results in the following chapter.

## 5.1 ANALYSIS OF CODE PASSAGES

```
import pandas as pd ## for data manipulation
import numpy as np ## for mathematical operations
import matplotlib.pyplot as plt ## for plotting graphs
from sklearn.preprocessing import MinMaxScaler ## for scaling the data
from sklearn.cluster import KMeans ## for clustering analysis
from sklearn.cluster import SpectralClustering ## for Spectral clustering
from sklearn.metrics import silhouette_score ## for silhouette score
```

This code step is inserted prior to the analysis of both databases and has the task of importing a number of libraries into Google Colaboratory.

Python libraries are collections of predefined and pre-compiled code and serve to provide additional functionality to the basic functions of the Python language. They consist of a series of modules and packages that in turn contain a set of functions and classes that can be imported into one's own programmes to extend their functionality.

[23] https://research.google.com/colaboratory/faq.html

For the creation of the output of the trajectory analysis within the stores, the libraries present in the previous code passage were used, in particular, the import sections of pandas, numpy, matplotlib were inserted in the upper passage.

Pandas was imported as it is the library used for data manipulation in Python useful for cleaning, manipulating and analysing data and especially for reading .csv files, the format in which Grottini Lab provided the store databases; the Numpy library performed the valuable task of support for numerical data management and statistical operations, more specifically also in the processing of arrays and more generally multidimensional data and matrices; the matplotlib library is a data visualisation library in Python that offers tools for the creation of graphs, histograms and scatter plots.

In addition to the libraries, classes were imported from the ScikitLearn library. This is an open source Python library created for machine learning that offers within it a wide range of tools and algorithms for building machine learning models. The main functions for which this library was used extensively in my work, especially for the section on clustering, are the use of clustering algorithms, and normalisation of data; in addition, other key features of this library were the possibility of feature selection, evaluation of model performance and visualisation of results.

Classes were imported from the Scikit Learn library:

- MinMaxScaler used to scale data into a specific range.
- KMeans, a class used for the creation of cluster analysis using the K-Means algorithm, an unsupervised algorithm.
- SpectralClustering, a class used for the creation of cluster analysis using the Spectral Clustering algorithm, an unsupervised algorithm.
- Silhouette score, in order to be able to calculate the 'silhouette score', i.e. a metric for assessing the quality of clustering based on the distance between points within individual clusters and between clusters themselves.

In summary, this script imports a number of Python libraries useful for cluster analysis, including tools for data manipulation, data visualisation, data scaling and cluster analysis using algorithms such as K-Means and spectral clustering.

```
data = data[data['distance_aoi'] > 0.4]
data = data[data['distance'] > 100]
print(f'Number of rows: {data.shape[0]} and number of columns: {data.shape[1]} in the data after removing outliers')
```

The upper code step is another salient aspect of the script that should be mentioned in the section on the commentary on the survey methodologies, as it represents one of the key points through which the two datasets were modified in such a way as to have a more satisfactory output that was not polluted by trajectory values that did not lead to any improvement in model performance.

In fact, so-called threshold levels were introduced to establish a selection criterion for the observations collected by the sensors within the store. In this case, trajectories were 'discarded' from the total number of observations within the dataset if they were recorded with a distance within a particular sector of the store of less than 40 centimetres and a total distance travelled within the store of less than 100 metres.

These precautionary measures were taken for the filtering of the dataset because the shelf where the survey is carried out within the two stores has measurements between 75 cm and 130 cm in width, so assuming that a person would walk into a particular category up to the middle of the first shelf and then return back, they would at worst walk at least 75 cm.

Through these selection criteria the observations were reduced for both datasets from 732,114 to 709,545 for Store 1 and from 971,653 to 941,538 for Store 2. Since each row of the dataset belongs to a particular observation of the trajectory of a single individual within a specific sector, as per dataset description see Chapter x paragraph X, 20/30 thousand data of respective customers were not eliminated, but rather either single observations of a particular sector if this was the only problematic one within the dataset, or entire trips within the store if the total distance travelled was below the predefined threshold value of 100 metres in total.

```
data['time_start_aoi'] = pd.to_datetime(data['time_start_aoi'])
data['time_end_aoi'] = pd.to_datetime(data['time_end_aoi'])
data['time_start_full'] = pd.to_datetime(data['time_start_full'])
data['time_end_full'] = pd.to_datetime(data['time_end_full'])
```

Since the file of collected observations had values for features that could not be used for clustering analyses, a conversion via the command pd.to_datetime() was necessary to have the values in the numpy-defined data type datetime64[ns], which then allowed a timestamp object that in turn allowed dates and times to be handled in an efficient manner by the algorithm.

```
data['duration_full'] = (data['time_end_full'] - data['time_start_full']).dt.total_seconds()
data['duration_aoi'] = (data['time_end_aoi'] - data['time_start_aoi']).dt.total_seconds()
data['day_of_week'] = data['time_start_aoi'].dt.dayofweek
```

Three columns were added in order to better visualise the dataset, the duration_full column indicating the total time spent within the store by each individual consumer analysed, the duration_aoi column indicating the total time spent within the area of interest by the individual and finally the day_of_week column indicating the day of the week on which the measurements were taken.

```
sector_df = pd.DataFrame(data['sector'].value_counts())
sector_df = sector_df.reset_index()
sector_df.columns = ['sector', 'total_trips']
```

Subsequently, in order to carry out the clustering analyses with the two datasets, it was decided to create a new dataframe called sector_df that would split each value of the sector column into a single value in order to have a dataframe that could be used more efficiently to pursue the clustering objectives on the selected features.

```
sector_df['avg_duration'] = data.groupby('sector')['duration_aoi'].mean().values ## average duration per sector
sector_df['avg_distance'] = data.groupby('sector')['distance_aoi'].mean().values ## average distance per sector
sector_df['basket_count'] = data[data['description'] == 'Basket']['sector'].value_counts().sort_index().values ## number of baskets per sector
sector_df['cart_count'] = data[data['description'] == 'Cart']['sector'].value_counts().sort_index().values ## number of carts per sector
```

The previous step served to prepare the dataset for the creation of columns that would correctly express the features to be analysed for the cluster. The first column introduced in the sector_df dataframe was the one called avg_duration, which made it possible to obtain through a passage with the groupby method a column expressing the average duration of the path in the area of interest in a particular sector.

The second column performs the same procedure as above, but goes on to form the avg_distance column, which similarly to the top column expresses the average distance travelled by a profiled individual within the area of interest in a particular sector. Finally, the number of baskets and trolleys respectively present in each sector are calculated.

```
days = ['monday', 'tuesday', 'wednesday', 'thursday', 'friday', 'saturday', 'sunday']
dfs = {}
for day in days:
df = data[data['day_of_week'] == days.index(day)]['sector'].value_counts().sort_index()
dfs[day] = df
```

Next, in order to better visualise the relationship of the previously created variables with the day of the week on which the particular trajectory analysed by the sensors and anchors occurred in the store, a Pandas

DataFrame dictionary containing the sector counts for each day of the week (represented by the string days) was added.

```
days = ['monday', 'tuesday', 'wednesday', 'thursday', 'friday', 'saturday', 'sunday']
dfs = {}
for day in days:
df = data[data['day_of_week'] == days.index(day)]['sector'].value_counts().sort_index()
dfs[f'{day}_trips'] = df
```

Script used within the data analysis process to have a numerical and counting representation of individual trajectories identified and categorised by day of the week.

After concatenating the datasets[24] above and creating an organised database for the research purposes of the thesis, the script then moved on to the plotting section.

Plotting in Python consists of creating graphs for data visualisation using the previously imported plotting libraries; it is a fundamental part of data analysis as the code steps explained in the next step will then be justified in the section on presenting the output.

```
ax.bar(sector_df['sector'], sector_df['monday_trips'], label='monday')
ax.bar(sector_df['sector'], sector_df['tuesday_trips'], bottom=sector_df['monday_trips'], label='tuesday')
ax.bar(sector_df['sector'], sector_df['wednesday_trips'], bottom=sector_df['monday_trips'] + sector_df['tuesday_trips'], label='wednesday')
ax.bar(sector_df['sector'], sector_df['thursday_trips'], bottom=sector_df['monday_trips'] + sector_df['tuesday_trips'] + sector_df['wednesday_trips'], label='thursday')
ax.bar(sector_df['sector'], sector_df['friday_trips'], bottom=sector_df['monday_trips'] + sector_df['tuesday_trips'] + sector_df['wednesday_trips'] + sector_df['thursday_trips'], label='friday')
ax.bar(sector_df['sector'], sector_df['saturday_trips'], bottom=sector_df['monday_trips'] + sector_df['tuesday_trips'] + sector_df['wednesday_trips'] + sector_df['thursday_trips'] + sector_df['friday_trips'], label='saturday')
ax.bar(sector_df['sector'], sector_df['sunday_trips'], bottom=sector_df['monday_trips'] + sector_df['tuesday_trips'] + sector_df['wednesday_trips'] + sector_df['thursday_trips'] + sector_df['friday_trips'] + sector_df['saturday_trips'], label='sunday')
ax.set_ylabel('Number of visits')
ax.set_xlabel('Sector')
```

[24] Missing script passages in the full script in the appendix

```
ax.set_title('Number of visits per day of the week for each sector')
```

This script represents the first of the bar graphs inserted for the graphic visualisation of feature interactions in the dataset. The function used for the following steps is plt.subplots(). The axis names are inserted after the axis name of interest x label or y label, in the example of the script step taken above the crossing of features for graphical visualisation was the variables 'Number of visits' and 'Sector'.

In addition to this plotting, graphs were analysed for the visualisation of the interaction between the features 'number of baskets' or 'carts' per different 'sector'; the following plot depicts the variable 'average duration' in relation to the feature 'sector' in order to graph the variation of the average time spent by consumers in the different sectors traversed during their purchase path within the intelligent retail environment of the two stores.

Next, in order to obtain an estimate of the average distance travelled every single day of the week in a particular sector, advanced feature plotting was carried out, whereby in addition to the previous plotting that would have combined the distance travelled in a particular sector, the breakdown by day of the week was also added to the graph returned as output by the Google Colaboratory programme

```
scaler = MinMaxScaler()
data_scaled = scaler.fit_transform(sector_df[['total_trips', 'avg_duration', 'avg_distance', 'basket_count', 'cart_count']])
data_scaled = pd.DataFrame(data_scaled, columns=['total_trips', 'avg_duration', 'avg_distance', 'basket_count', 'cart_count'])
data_scaled
```

The following step was reported in the data analysis description because it performs a fundamental step in the creation of the dataframe for clustering analysis, the final objective of the script process. The function used is called 'MinMaxScaler' and had the precise objective of normalising the columns relating to the features of the total trajectories travelled, average duration, average distance and number of baskets and trolleys; this resulted in the new dataframe having the values in the columns mentioned above, values within a given range.

Once the preprocessing phase and the dataframe backup for the use of several clustering methods had been completed, we then moved on to the central phase of the data analysis work, i.e. the implementation of the clustering procedures and the search for the optimal number of clusters.

```
wcss = []
for i in range(1, 11):
kmeans = KMeans(n_clusters=i, init='k-means++', random_state=42)
kmeans.fit(kmeans_scaled)
```

```
wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss) ## plotting elbow curve
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
```

As previously mentioned, a preparatory step for clustering by means of K-Means is the search for the optimal number of clusters, as a clustering method requires the manual input of the number of centroids into which the dataframe is to be divided. The most commonly used methodology, and also the one employed in this script, is the 'Elbow Method'.

The 'WCSS' is also displayed, which stands for 'Within-Cluster Sum of Squares' and is a measure of the sum of squares of the distances of each point in a cluster from the centroid representing the cluster itself. The aim of the 'K-Means' clustering method is to minimise the value of the WCSS, thus finding centroids that have a sum of the squares of the distances of each point belonging to the given cluster that is as small as possible. The elbow method is, in fact, used to observe when the curve begins to level off, the levelling off the curve being the point at which the addition of further clusters will not provide any significant improvement in the reduction of the WCSS.

In this case, by means of the plt.plot() command, it was possible to observe graphically, through the Elbow Method curve, how the WCSS score value changes as k (expression of the number of clusters) changes.

```
kmeans = KMeans(n_clusters=4, init='k-means++', random_state=42)
kmeans.fit(kmeans_scaled)
kmeans_df['cluster'] = kmeans.labels_
kmeans_df
```

This script reports the use of the KMeans clustering algorithm to divide the observations of the dataframe into different clusters. Using the previous elbow method, the optimal number of clusters identified was 4

```
ax[0, 0].scatter(kmeans_df[kmeans_df['cluster'] == 0]['avg_duration'], kmeans_df[kmeans_df['cluster'] == 0]['avg_distance'], c='red', label='Cluster 1')
ax[0, 0].scatter(kmeans_df[kmeans_df['cluster'] == 1]['avg_duration'], kmeans_df[kmeans_df['cluster'] == 1]['avg_distance'], c='blue', label='Cluster 2')
ax[0, 0].scatter(kmeans_df[kmeans_df['cluster'] == 2]['avg_duration'], kmeans_df[kmeans_df['cluster'] == 2]['avg_distance'], c='green', label='Cluster 3')
```

```
  ax[0, 0].scatter(kmeans_df[kmeans_df['cluster'] == 3]['avg_duration'], kmeans_df[kmeans_df['cluster'] == 3]['avg_distance'], c='black', label='Cluster 4')
  ax[0, 0].set_xlabel('Average Duration')
  ax[0, 0].set_ylabel('Average Distance')
  ax[0, 0].set_title('Duration vs. Distance')
  ax[0, 0].legend()

  ax[0, 1].scatter(kmeans_df[kmeans_df['cluster'] == 0]['cart_count'], kmeans_df[kmeans_df['cluster'] == 0]['basket_count'], c='red', label='Cluster 1')
  ax[0, 1].scatter(kmeans_df[kmeans_df['cluster'] == 1]['cart_count'], kmeans_df[kmeans_df['cluster'] == 1]['basket_count'], c='blue', label='Cluster 2')
  ax[0, 1].scatter(kmeans_df[kmeans_df['cluster'] == 2]['cart_count'], kmeans_df[kmeans_df['cluster'] == 2]['basket_count'], c='green', label='Cluster 3')
  ax[0, 1].scatter(kmeans_df[kmeans_df['cluster'] == 3]['cart_count'], kmeans_df[kmeans_df['cluster'] == 3]['basket_count'], c='black', label='Cluster 4')
  ax[0, 1].set_xlabel('Cart Count')
  ax[0, 1].set_ylabel('Basket Count')
  ax[0, 1].set_title('Carts vs Baskets')
  ax[0, 1].legend()
  #weekday vs weekend graphic
  kmeans_df['weekday'] = kmeans_df['monday_trips'] + kmeans_df['tuesday_trips'] + kmeans_df['wednesday_trips'] + kmeans_df['thursday_trips'] + kmeans_df['friday_trips']
  kmeans_df['weekend'] = kmeans_df['saturday_trips'] + kmeans_df['sunday_trips']
  ax[1, 1].scatter(kmeans_df[kmeans_df['cluster'] == 0]['weekday'], kmeans_df[kmeans_df['cluster'] == 0]['weekend'], c='red', label='Cluster 1')
  ax[1, 1].scatter(kmeans_df[kmeans_df['cluster'] == 1]['weekday'], kmeans_df[kmeans_df['cluster'] == 1]['weekend'], c='blue', label='Cluster 2')
  ax[1, 1].scatter(kmeans_df[kmeans_df['cluster'] == 2]['weekday'], kmeans_df[kmeans_df['cluster'] == 2]['weekend'], c='green', label='Cluster 3')
  ax[1, 1].scatter(kmeans_df[kmeans_df['cluster'] == 3]['weekday'], kmeans_df[kmeans_df['cluster'] == 3]['weekend'], c='black', label='Cluster 4')
  ax[1, 1].set_xlabel('Weekday Trips')
  ax[1, 1].set_ylabel('Weekend Trips')
  ax[1, 1].set_title('Weekday vs Weekend Trips')
  ax[1, 1].legend()
```

```
kmeans_df['weekday'] = kmeans_df['monday_avg_duration'] + kmeans_df['tuesday_avg_duration'] + kmeans_df['wednesday_avg_duration'] + kmeans_df['thursday_avg_duration'] + kmeans_df['friday_avg_duration']
kmeans_df['weekend'] = kmeans_df['saturday_avg_duration'] + kmeans_df['sunday_avg_duration']
ax[1, 0].scatter(kmeans_df[kmeans_df['cluster'] == 0]['weekday'], kmeans_df[kmeans_df['cluster'] == 0]['weekend'], c='red', label='Cluster 1')
ax[1, 0].scatter(kmeans_df[kmeans_df['cluster'] == 1]['weekday'], kmeans_df[kmeans_df['cluster'] == 1]['weekend'], c='blue', label='Cluster 2')
ax[1, 0].scatter(kmeans_df[kmeans_df['cluster'] == 2]['weekday'], kmeans_df[kmeans_df['cluster'] == 2]['weekend'], c='green', label='Cluster 3')
ax[1, 0].scatter(kmeans_df[kmeans_df['cluster'] == 3]['weekday'], kmeans_df[kmeans_df['cluster'] == 3]['weekend'], c='black', label='Cluster 4')
ax[1, 0].set_xlabel('Weekday Duration')
ax[1, 0].set_ylabel('Weekend Duration')
ax[1, 0].set_title('Weekday vs Weekend Duration')
ax[1, 0].legend()
```

This script has been included in the section commenting on the salient steps of the dataset analysis, since it describes in a technical manner the code steps performed for the creation of scatter plots or scatter graphs, which are present in the section presenting the results. The first scatter plot is used to graphically display the positioning of each cluster in relation to the variables average duration and average distance; the second scatter plot shows the graphical position and dispersion of the clusters relative to the different sectors in a graph with the basket or cart on its Cartesian axes; the third scatter plot created for both stores represents the positioning of the clusters in a Cartesian graph with 'Weekday duration' on its x-axis, i.e. the duration of transit for the different sectors by customers on weekdays, and 'Weekend duration' on its y-axis, which measures the average duration of transit in the different sectors on weekend days, i.e. Saturday and Sunday. Ending the phase of representation of the clusters, we describe the fourth and last scatter plot created for the study of the clusters, it represents the graphical arrangement of the clusters on a Cartesian plane having as variables on the x-axis 'Weekday trips' that is the number of times consumers have transited through a given sector and on the y-axis instead 'Weekend Trips' which measures the same value as the previous feature but for weekend days (Saturday and Sunday)

```
kmeans_df['cluster'].value_counts()
kmeans_df[kmeans_df['cluster'] == 0]
kmeans_df[kmeans_df['cluster'] == 1]
kmeans_df[kmeans_df['cluster'] == 2]
kmeans_df[kmeans_df['cluster'] == 3]
```

In order to finish the clustering procedure using K-Means and to be able to better understand the scatter plots through the arrangement of the sectors within the clusters, the total count of the number of sectors within the individual clusters was carried out and then the list of individual sectors per cluster.

```
spectral = SpectralClustering(n_clusters=3, affinity='nearest_neighbours')
spectral.fit(spectral_scaled)
spectral_df['cluster'] = spectral.labels_
spectral_df
```

Next, with the code step above, we proceeded to conduct the cluster analysis with the second algorithm, i.e., by means of the spectral clustering algorithm.

The analysis was conducted with this specific algorithm as it is particularly useful in clustering large datasets such as those analysed in this thesis, data which have a complex structure and a distribution that can be modelled more easily than by a partitional algorithm such as the K-Means mentioned above.

Unlike the previous clustering method, a set of parameters must be chosen to implement Spectral Clustering; the parameters used to implement spectral clustering in this dataset analysis were:

- Choice of the number of clusters (K): the dataset was chosen to be divided into 3 clusters.
- Choice of affinity: affinity was set to 'nearest neighbours' which means that the closest possible distance between the points was used for the clustering process and for the creation of the affinity matrix.

```
ax[0, 0].scatter(spectral_df[spectral_df['cluster'] == 0]['avg_duration'], spectral_df[spectral_df['cluster'] == 0]['avg_distance'], c='red', label='Cluster 1')
ax[0, 0].scatter(spectral_df[spectral_df['cluster'] == 1]['avg_duration'], spectral_df[spectral_df['cluster'] == 1]['avg_distance'], c='blue', label='Cluster 2')
ax[0, 0].scatter(spectral_df[spectral_df['cluster'] == 2]['avg_duration'], spectral_df[spectral_df['cluster'] == 2]['avg_distance'], c='green', label='Cluster 3')
ax[0, 0].scatter(spectral_df[spectral_df['cluster'] == 3]['avg_duration'], spectral_df[spectral_df['cluster'] == 3]['avg_distance'], c='black', label='Cluster 4')
ax[0, 0].set_xlabel('Average Duration')
ax[0, 0].set_ylabel('Average Distance')
ax[0, 0].set_title('Duration vs. Distance')
ax[0, 0].legend()
ax[0, 1].scatter(spectral_df[spectral_df['cluster'] == 0]['cart_count'], spectral_df[spectral_df['cluster'] == 0]['basket_count'], c='red', label='Cluster 1')
```

```
  ax[0, 1].scatter(spectral_df[spectral_df['cluster'] == 1]['cart_count'], spectral_df[spectral_df['cluster'] == 1]['basket_count'], c='blue', label='Cluster 2')
  ax[0, 1].scatter(spectral_df[spectral_df['cluster'] == 2]['cart_count'], spectral_df[spectral_df['cluster'] == 2]['basket_count'], c='green', label='Cluster 3')
  ax[0, 1].scatter(spectral_df[spectral_df['cluster'] == 3]['cart_count'], spectral_df[spectral_df['cluster'] == 3]['basket_count'], c='black', label='Cluster 4')
  ax[0, 1].set_xlabel('Cart Count')
  ax[0, 1].set_ylabel('Basket Count')
  ax[0, 1].set_title('Carts vs Baskets')
  ax[0, 1].legend()
  #weekday vs weekend graphic
  spectral_df['weekday'] = spectral_df['monday_trips'] + spectral_df['tuesday_trips'] + spectral_df['wednesday_trips'] + spectral_df['thursday_trips'] + spectral_df['friday_trips']
  spectral_df['weekend'] = spectral_df['saturday_trips'] + spectral_df['sunday_trips']
  ax[1, 1].scatter(spectral_df[spectral_df['cluster'] == 0]['weekday'], spectral_df[spectral_df['cluster'] == 0]['weekend'], c='red', label='Cluster 1')
  ax[1, 1].scatter(spectral_df[spectral_df['cluster'] == 1]['weekday'], spectral_df[spectral_df['cluster'] == 1]['weekend'], c='blue', label='Cluster 2')
  ax[1, 1].scatter(spectral_df[spectral_df['cluster'] == 2]['weekday'], spectral_df[spectral_df['cluster'] == 2]['weekend'], c='green', label='Cluster 3')
  ax[1, 1].scatter(spectral_df[spectral_df['cluster'] == 3]['weekday'], spectral_df[spectral_df['cluster'] == 3]['weekend'], c='black', label='Cluster 4')
  ax[1, 1].set_xlabel('Weekday Trips')
  ax[1, 1].set_ylabel('Weekend Trips')
  ax[1, 1].set_title('Weekday vs Weekend Trips')
  ax[1, 1].legend()
  spectral_df['weekday'] = spectral_df['monday_avg_duration'] + spectral_df['tuesday_avg_duration'] + spectral_df['wednesday_avg_duration'] + spectral_df['thursday_avg_duration'] + spectral_df['friday_avg_duration']
  spectral_df['weekend'] = spectral_df['saturday_avg_duration'] + spectral_df['sunday_avg_duration']
  ax[1, 0].scatter(spectral_df[spectral_df['cluster'] == 0]['weekday'], spectral_df[spectral_df['cluster'] == 0]['weekend'], c='red', label='Cluster 1')
  ax[1, 0].scatter(spectral_df[spectral_df['cluster'] == 1]['weekday'], spectral_df[spectral_df['cluster'] == 1]['weekend'], c='blue', label='Cluster 2')
  ax[1, 0].scatter(spectral_df[spectral_df['cluster'] == 2]['weekday'], spectral_df[spectral_df['cluster'] == 2]['weekend'], c='green', label='Cluster 3')
  ax[1, 0].scatter(spectral_df[spectral_df['cluster'] == 3]['weekday'], spectral_df[spectral_df['cluster'] == 3]['weekend'], c='black', label='Cluster 4')
  ax[1, 0].set_xlabel('Weekday Duration')
  ax[1, 0].set_ylabel('Weekend Duration')
  ax[1, 0].set_title('Weekday vs Weekend Duration')
  ax[1, 0].legend()
```

The script shows the visualisation of the results of the spectral clustering; as in the case of the scatter plots of the clusters found using the K-Means algorithm, the graphs show the relationships between two variables positioned on the Cartesian X and Y axes, respectively.

The combinations of variables used to show the dispersion of the clusters obtained are the same as those used previously for 'k-Means' clustering in order to be able to propose a graphical comparison of the effectiveness of the two methods used for the survey.

# 6. OUTPUT

After having described the salient steps of scripts aimed at giving a technical background to the outputs, in this section, we will go on to analyse the graphical results of the analyses carried out on the two datasets and, subsequently, the outputs originating from the implementation of the clustering algorithms, visualised by means of scatter plot type graphs, used to succeed in representing the clusters of the observations relating to the Rome and L'Aquila shop.

In order to be able to better observe the results produced, it is first necessary to make a brief survey of the retail spaces in which the surveys were carried out; the topographies of the individual shops will be given below.

The areas marked in red are the aoi (area of interest) sections, they are named after the reference sector and indicate the detection area in which the sensors detect consumers and mark the observation as a trajectory relative to the particular transit sector.

The retail environments, as can easily be seen from the inserted topographies, have a layout of the sectors and their scans that do not coincide with each other, however, before we begin to show the outputs of the data analysis obtained from the two shops, it is necessary to emphasise some common points between the layouts of both shops:

- The entrance is on the side adjacent to the cash desks.
- The frozen food, water and pet food sectors are located at the same time as the checkouts, being sectors that offer heavy products such as water crates, products that customers tend to buy on impulse before making payment, or products that require special refrigerated counters for storage.

Observing common elements between the different retail environments in which the surveys that will subsequently make up the datasets are carried out is of paramount importance when analysing the outputs; in fact, a similarity in a particular configuration between environments will lead the results to assume similar configurations.

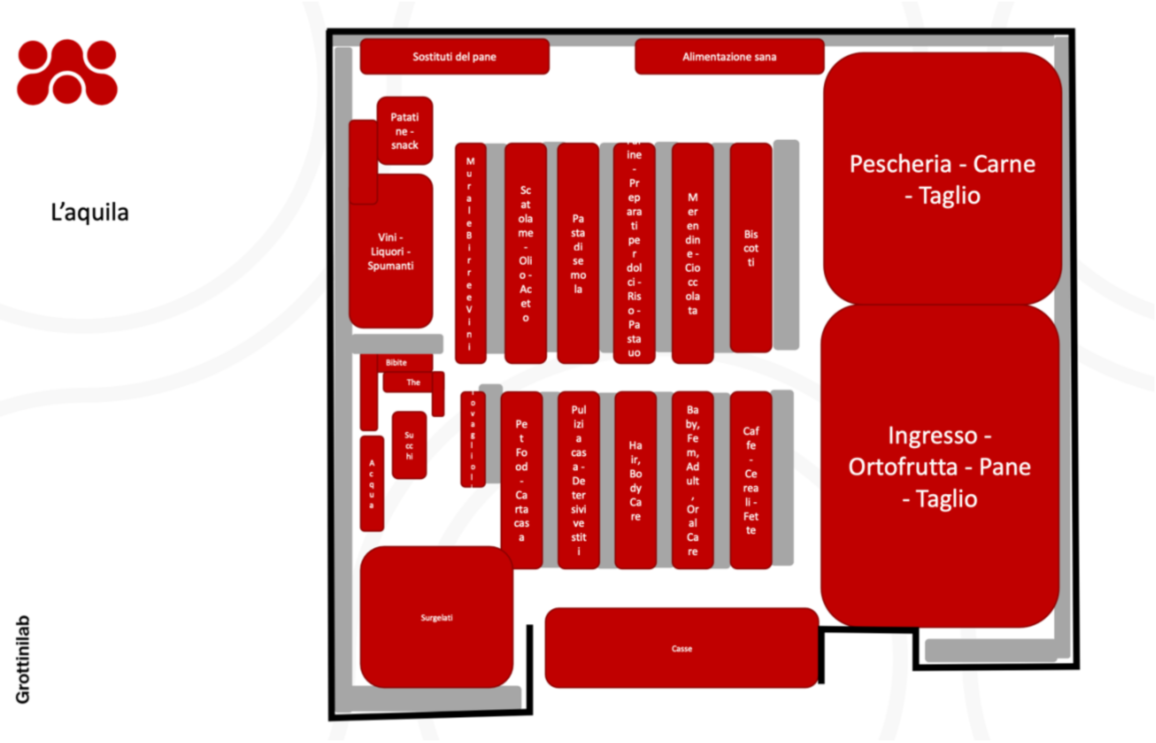

*Figure 4: Store 1 map L'Aquila*

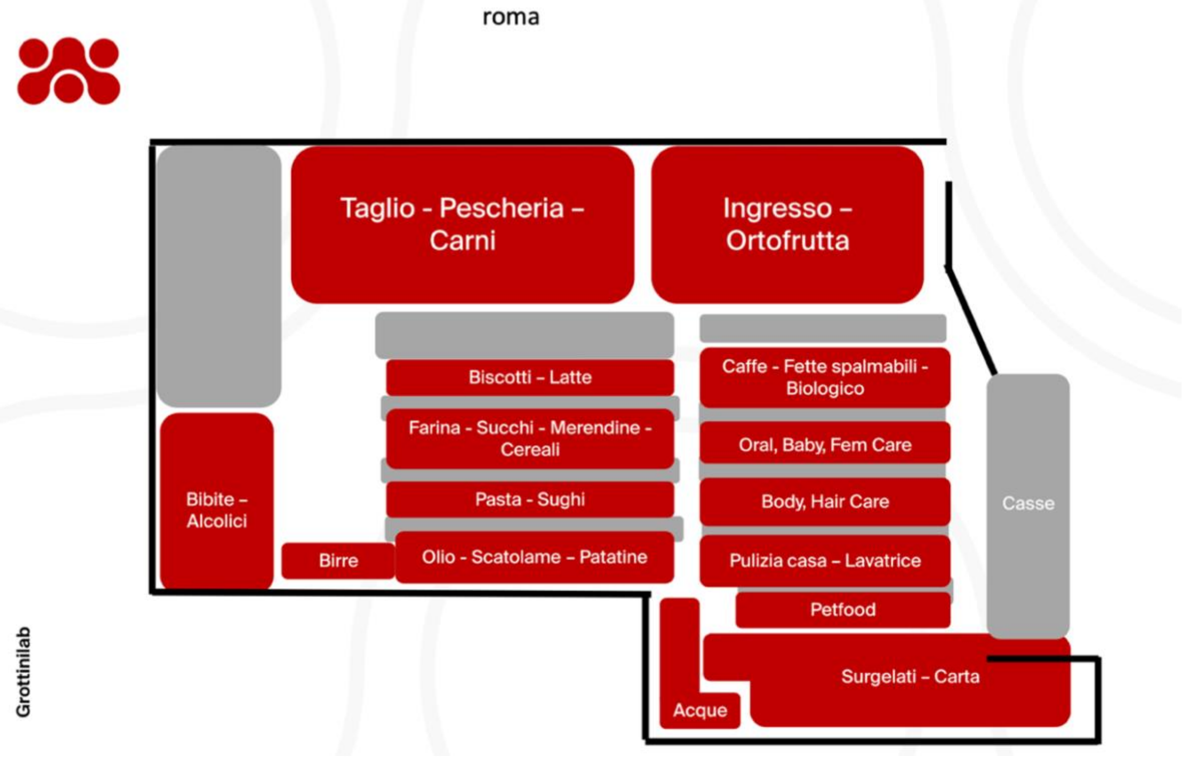

*Figure 5: Store 2 map Rome*

## 7.1 OUTPUT SURVEY METHODOLOGY

The first part on output analysis examines the detections that can be made on the databases obtainable through artificial intelligence devices listed above.

Before carrying out the clustering analysis, which is the real aim of this work, a series of information can also be obtained by means of preliminary analyses, which include optimising the columns of the dataset, clustering and making certain key features explicit for a reading of this dataframe.

The precise coding steps performed and how they are output are explained more technically and precisely in the previous chapter, but in order to introduce the two scatter plots relating to the two inserted stores, it is worthwhile to resume, albeit briefly, what they refer to.

The two present scatter plots refer to the number of visits in each different sector per day of the week; the Cartesian x-axis of this graph shows the different sectors belonging to the supermarkets, while the y-axis shows the numerical value of total weekly visits.

The day of the week on which these measurements are taken is indicated by the colour of the column, i.e., blue for Monday, orange for Tuesday, green for Wednesday, red for Thursday, purple for Friday, brown for Saturday and pink for Sunday.

This graph offers the possibility of intercepting which sectors are most preferred by customers when they enter the retail space, furthermore, the composition of the column with sections referring to the different days of store attendance allows the understanding of the correlation between weekly and working time with the passage in a relative sector.

In the following, observations that can be made on the graphical 'scatterplot' outputs are shown, these reflections are examples of the focus points on which the analysis of the graphs should be more focused. Since it is not the purpose of this thesis to observe and analyse the output for a particular purpose, empirical deductions that can be made on these graphical representations are shown below.

From the pictures it is easy to see that the sectors, for both stores, that are most visited regardless of the day of the week are: Fruit and vegetables, Cold meats and dairy products, Cutlery and other sectors that normally have products consumed by the majority of customers; in fact, in turn, the least visited sectors for both stores are those that offer products and services that are more specific and targeted to a particular clientele, among these we find the sectors related to products for consumers with food intolerances or preferences such as gluten-free, vegan, nutraceutical, or those that offer products to a targeted clientele that responds to characteristics of gender or precise social status such as those related to sanitary towels and nappies

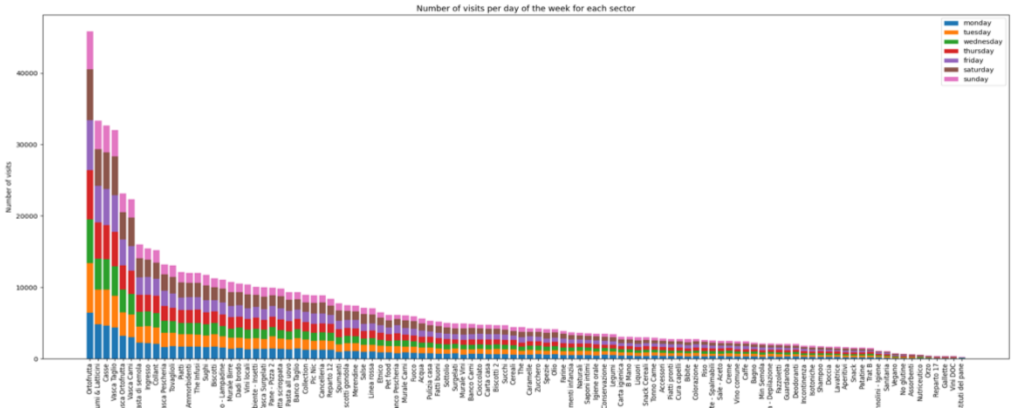

*Figure 6 graph 'Number of visits per day of the week for each sector' for Store 1*

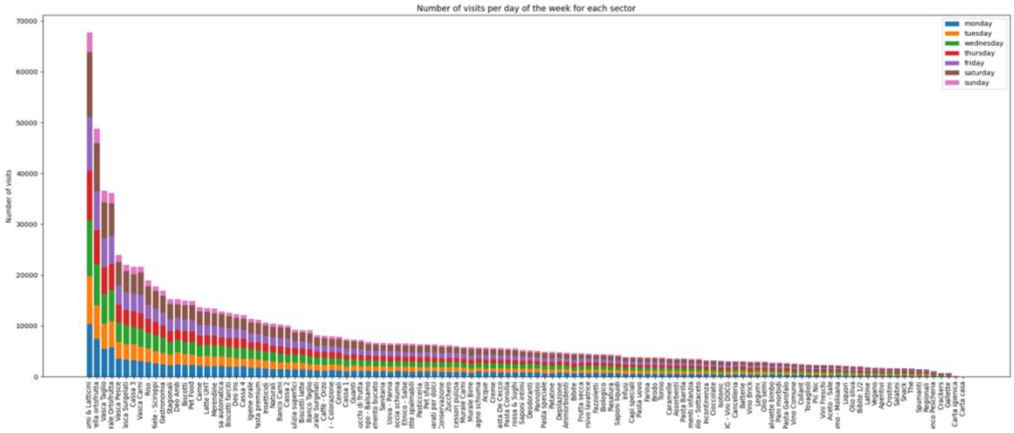

*Figure 7 graph 'Number of visits per day of the week for each sector' for Store 2*

The scatter plot shown in Figure 8 and Figure 9 observes the passage through a particular area by customers who chose the basket or trolley at the entrance. The relevance of this observation is not only for statistical purposes, in fact, it allows further reflections, for example, it is possible to study how within a store the association between the factor of use of the basket of with less time spent inside the store, or a shopping carried out with less time available and less quantity of products purchased and, the use of the trolley, instead, as an indicator of more prolonged spending and with relative greater purchase of products; Furthermore, it is possible to study how the layout of the shelves within the store and the perception of speed in reaching and choosing products influences the purchasing behaviour of these two categories of consumers.

There are no proven strategies for the areas where there is an imbalance between the use of the two means for the collection of the sought after products, some areas are seen as time consuming, such as the cutting counter which in both stores has non-proportional concentrations of users with the baskets and trolleys, this thought may depend on numerous factors such as the fact that to get the product in these types of sections of the store you need to be served by staff, this additional step increases the waiting and retrieval time for the product itself, therefore, shoppers who initially chose to use the basket for quicker shopping may be reluctant to transit and purchase products in these departments, opting more for products already pre-wrapped and available on the shelves.

To increase the flow of customers in these sectors, it is necessary to act on the perception of the timing of service delivery or to attack complementary strategies such as the presence of a self-service section to reduce waiting times or to provide time-conscious consumers with a valid motivation to transit and eventually purchase in a particular sector within the retail environment.

Another motivation could be the difficulty in transporting products to the checkout and then home to products that are deemed too heavy or bulky, this often happens in the case of shelf space for products containing large quantities of product such as crates of soft drinks, detergent bottles or bulky products such as the area of interest of nappies. In this case, more should be done on the availability of products with reduced weight and size in order to improve customer perception.

The factors relating to the perception of a particular sector by individuals passing through a retail space are diverse, and it is not possible to categorise them through an observation of customer behaviour in two stores, but perception tests should be carried out if an imbalance between the two components is revealed and an attempt is made to induce users of either medium to frequent more sectors they would tend to ignore.

The column graphs obtained by analysing the dataframes for the two stores are shown below:

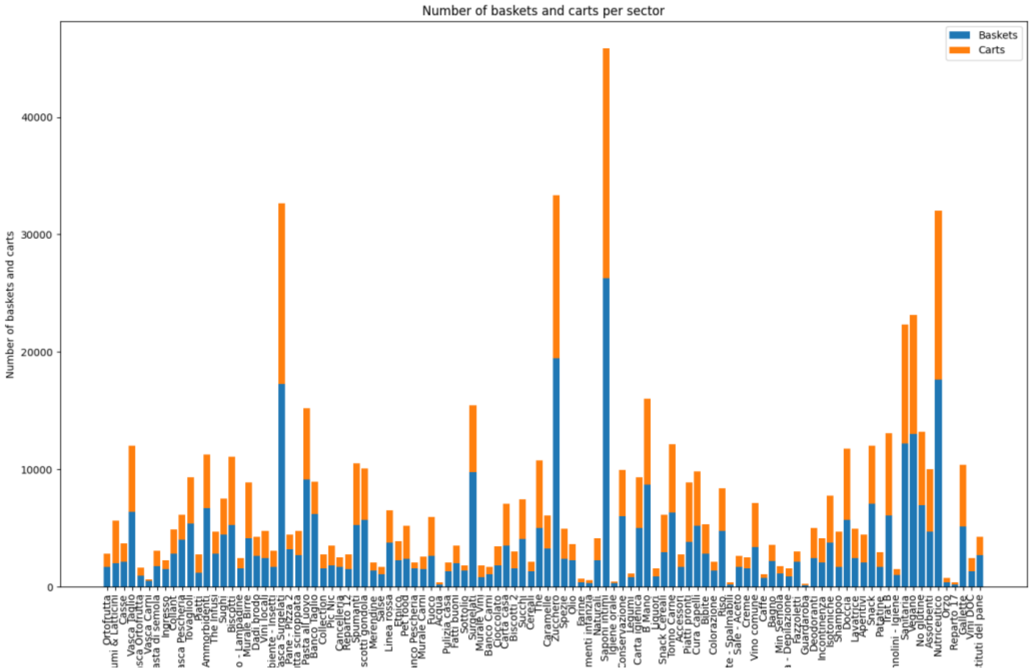

*Figure 8 graph 'Number of baskets and carts per sector' for Store 1*

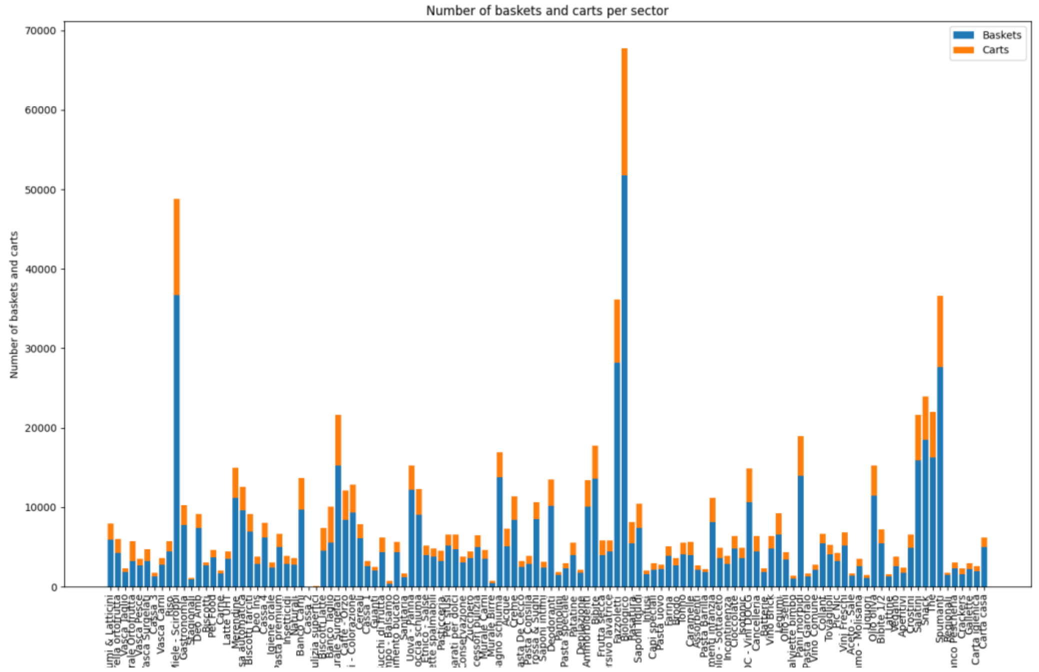

*Figure 9 graph 'Number of baskets and carts per sector' for Store 2*

Also, in the section on the analyses and depictions that can be conducted in order to obtain interesting statistics on the store are the graphical outputs relating to the plotting of feature combinations with the relevant sectors of interest. As for the present graph, the plotting methods and the organisation of the related dataframes has been explored in detail in the previous chapter.

In this case Figure 10 and Figure 11 show the column graphs relating to the length of time spent in the different store sectors, there is no correlation between the sectors with the longest length of time spent in Store 1 and those relating to Store 2, in fact, for the store located in the capital city of Abruzzo, the sectors that receive the most attention from customers are those relating to frozen foods, intimate soaps and tampons, while for the store located in the capital city they are the sector for syrups, the cutting counter and the sector relating to paints and colours.

Below are the bar graphs for the two stores:

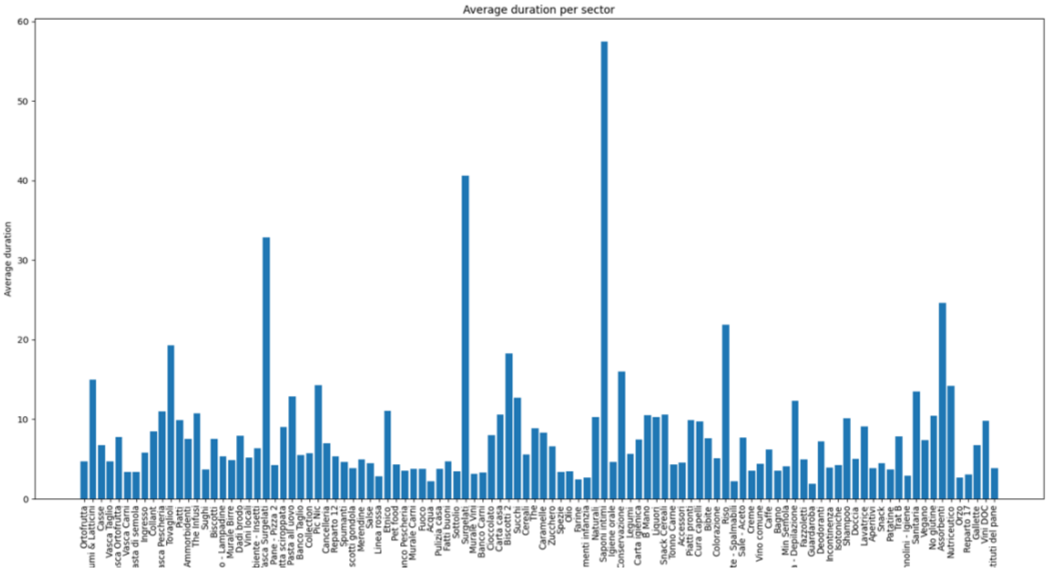

*Figure 10: 'Average duration per Sector' graph Store 1*

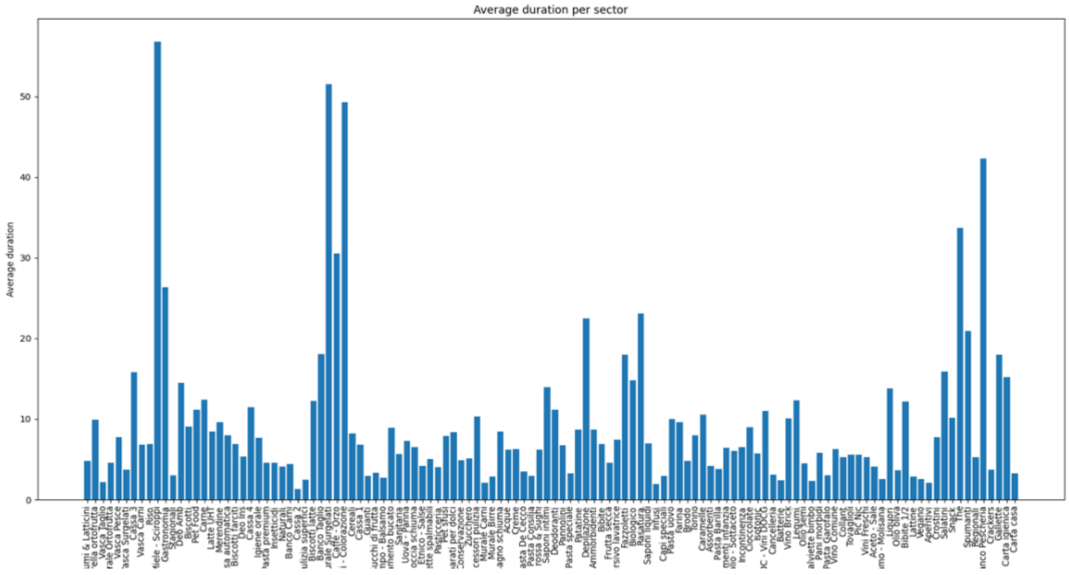

*Figure 11: 'Average duration per Sector' graph Store 2*

As reported for the previous graphs now in Figure 12 and Figure 13 the column graphs are inserted by means of the representation of the two datasets; the columns belonging to the graphs are positioned within a Cartesian axis having on the X axis the denomination of the analysed sectors and on the Y axis the metres travelled within the sector. There is a correlation between the sectors with greater time and greater distance (intimate soaps for Store 1 and syrups for Store 2), this confirms the probable presence of targeted activities carried out in these sectors and eliminates the possibility of any detection errors by the data collection procedure using the artificial intelligence infrastructure.

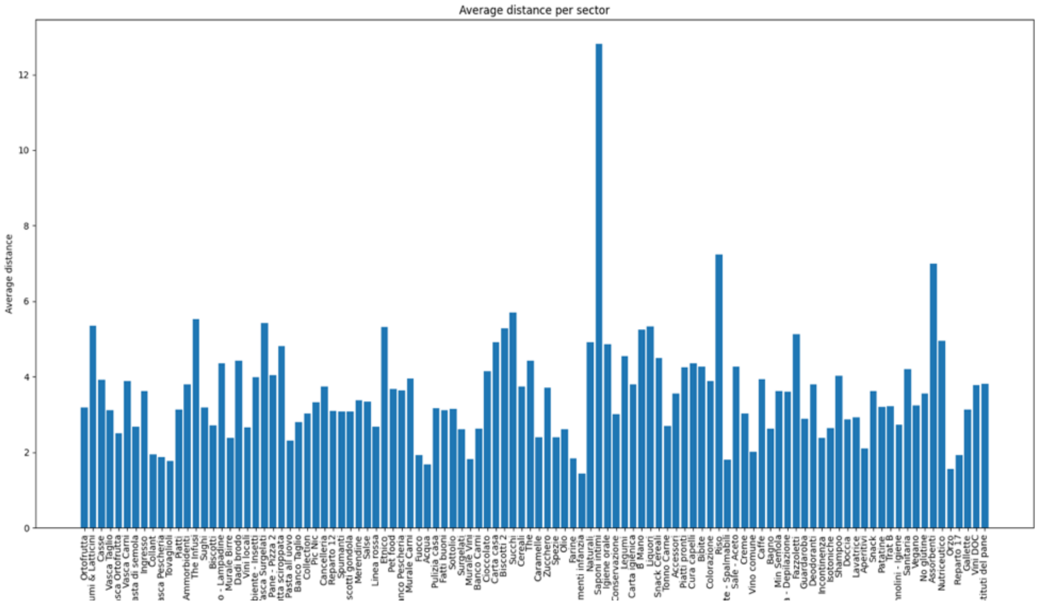

*Figure 12: 'Average distance per Sector' graph Store 1*

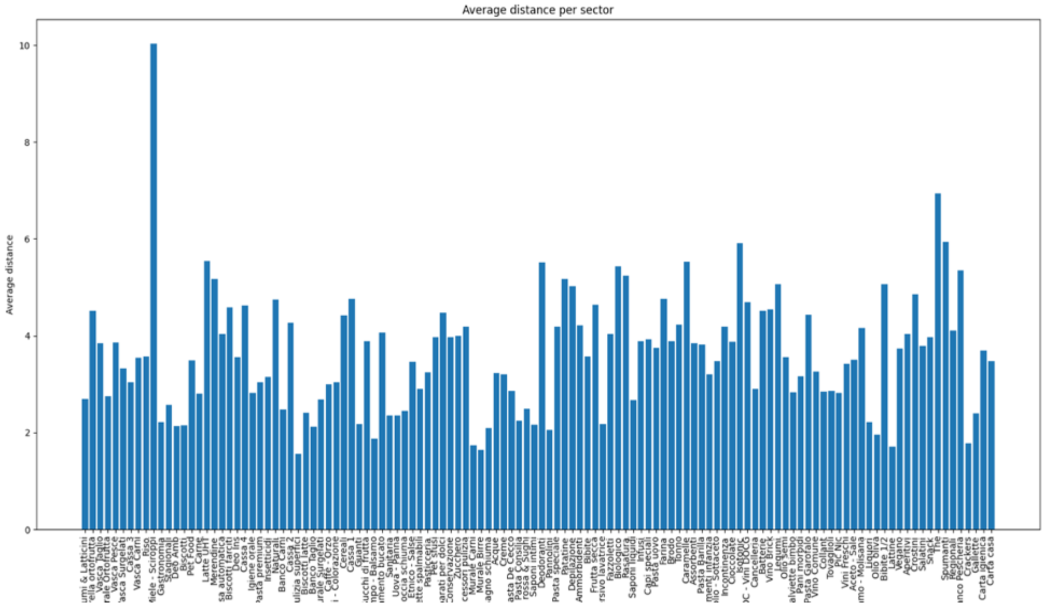

*Figure 13: 'Average distance per Sector' graph Store 2*

## 7.2 BUSINESS ANALYST: IMPROVING BUSINESS PROCESSES THROUGH CLUSTERING

As described in the methodology chapter, the elbow method was tested prior to the clustering operation, followed by the clustering of the dataframe.

The clustering methods have already been discussed from a technical point of view, but in this chapter, in addition to showing the results obtained from the datasets of the two stores under analysis, the main objective is to show the investigation methodology that needs to be carried out in order to gain important insights, irrespective of the type of dataset available.

Clustering analyses of facts are a method that can be implemented by means of dataset analysis software, which allows in datasets of high numerosity such as the present one, the aim of including them within the survey to be conducted on datasets containing information on consumer trajectories within retail spaces is to create clusters, i.e. groups, containing observations, which have the characteristic of being internally homogeneous and externally heterogeneous.

This makes it possible to find a similarity in the behaviour of the clustered features within the dataset and to carry the discovery made on the data into the actual operation of the store.

A discipline that deals with identifying business problems, finding solutions to improve the efficiency of business processes and the expression of business value to customers is Business Analysis.

The implementation of such a strategy involves the steps followed for the analysis carried out in this thesis, starting with the processing of data and arriving at the definition of a research model to obtain the necessary information for the business purpose.

It is a fundamental activity for any company that wants to identify and solve problems and adapt to market changes in order to maintain an important competitive advantage.

In fact, retail environments are a prime example of activities that require periodic analysis and monitoring of the observations collected in order to identify as quickly and effectively as possible problems that lead to low system efficiency or undermine its full sales and turnover capacity.

Below we will show the output obtained from the cluster survey carried out in the two stores in Rome and L'Aquila and then go on to visualise the output of the programme from a business analysis perspective in order to better understand the key aspects in which to focus in order to correctly read the functioning of the different sectors within the retail environments.

Cluster analysis can be carried out with all types of unsupervised learning, in this thesis the K-Means method and the method called Spectral Clustering were chosen as they are particularly effective in correctly handling datasets containing a large number of observations. Obtaining datasets by means of artificial intelligence devices that record a series of complex features for every single interaction of the consumer within the store, the datasets that will be created will most likely contain a complex number of observations that need to be managed and clustered with unsupervised learning algorithms that correctly analyse this type of dataset.

Figures 14 and 15 show the representation of the clustering activity by means of the K-Means methodology, in the pictures the positioning of the clusters within different Cartesian axes, the Cartesian axes in each individual ridge represent a feature, their combination allows us to understand the behaviour of the clusters as a result of the interaction with other features.

The interactions sought in these two images concern the crossing of features:

- Average distance travelled within the area of interest and duration of the trip within the survey area
- Sector cluster based on use of baskets or trolleys by customers observed
- Graphical representation cluster positioning according to features travel time in the survey area on a weekday and travel time on a weekend day
- Graphical representation of cluster positioning according to the features of the number of passages in the survey area during a weekday and those made on weekend days

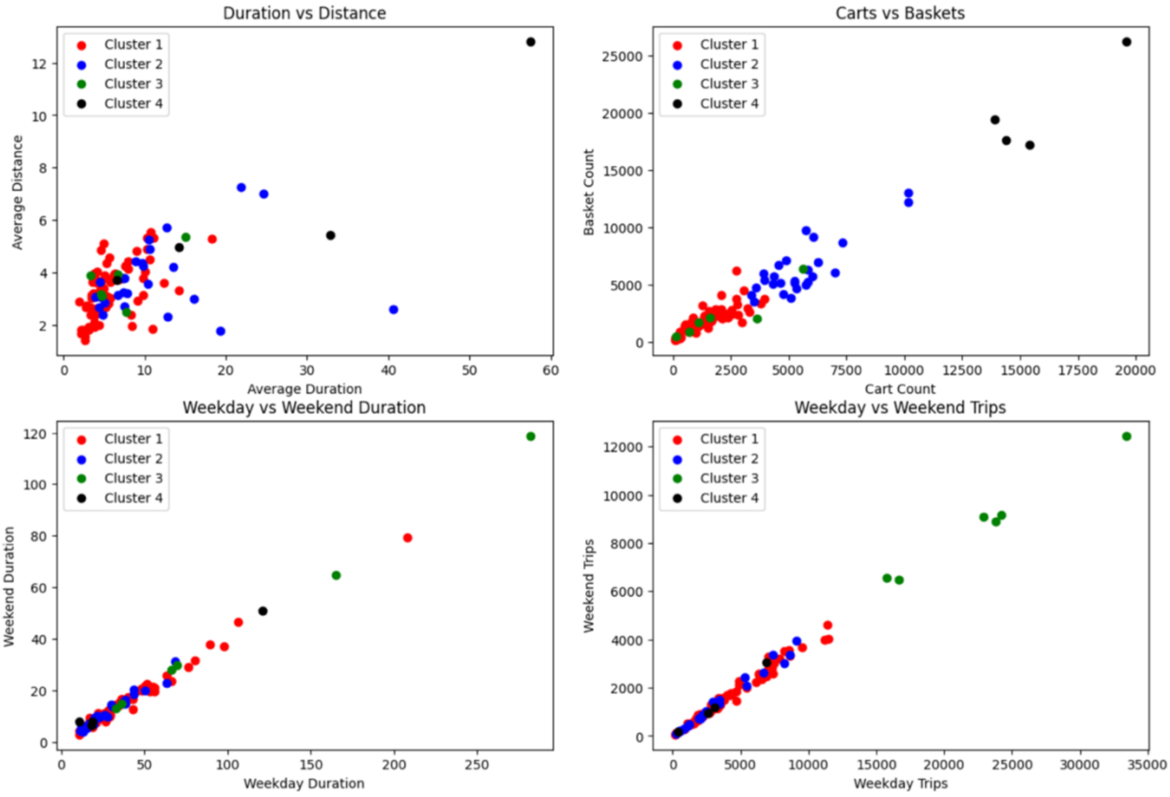

*Figure 14: Graph 'Plotting the clusters in subplots (k-Means)' Store 1*

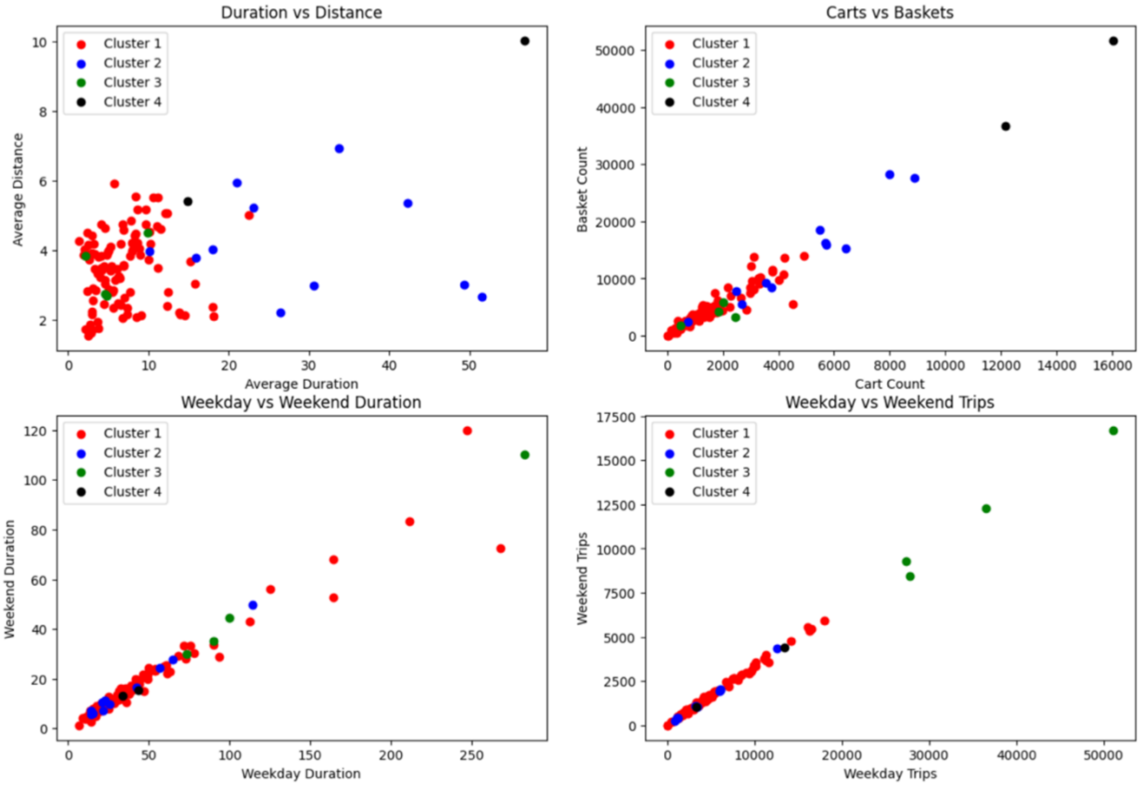

*Figure 15: Graph 'Plotting the clusters in subplots (k-Means)' Store 2*

The dots in the graphs above depict the position of the observations in the dataset and their colour indicates their belonging to the 4 clusters created.

Therefore, at this point in the search, the sectors belonging to the various clusters must be made explicit so that we can observe the sectors that assume similar behaviour for both stores.

Executing the command for the desrciption of elements within clusters initially reports that the following sectors are present within the clusters:

- Store 1: Cluster 0 = 70 sector; Cluster 1 = 26 sector; Cluster 2 = 6 sector ; Cluster 3 = 4 sector

- Store 2: Cluster 0 = 103 sector; Cluster 1 = 11 sector; Cluster 2 = 4 sector; Cluster 3 = 2 sector

In relation to Store 1 the 70 sectors belonging to Cluster 0 are: Entrance, Pantyhose, Fish Tank, Dishes, Infusions, Sauces, Brico - Bulbs, Stock cubes, Local wines, Room Deo - Insects, Bread - Pizza 2, Cake mixes - Fruit in syrup, Cutlery counter, Collection, Pic Nic, Stationery, Snacks, Sauces, Ethnic, Pet food, Fish counter etc.

For each cluster identified, it is possible to identify the sectors within it, this process serves to gain a better understanding of which sectors exhibit similar behaviour, being able to compare stores in which consumers exhibit similar behaviour such as dwell time, average distance travelled and use of carts or baskets can be a valuable indicator to understand the common responses of customers to a particular product configuration and location within the store.

Representation by Spectral Clustering:

Subsequently, Figures 16 and 17 also show the cluster representations obtained by means of the unsupervised learning method called Spectral Clustering; the same feature crossings as in the previous clustering activity were also used for this type of cluster representation.

In this particular case, 4 clusters were used with Spectral Clustering

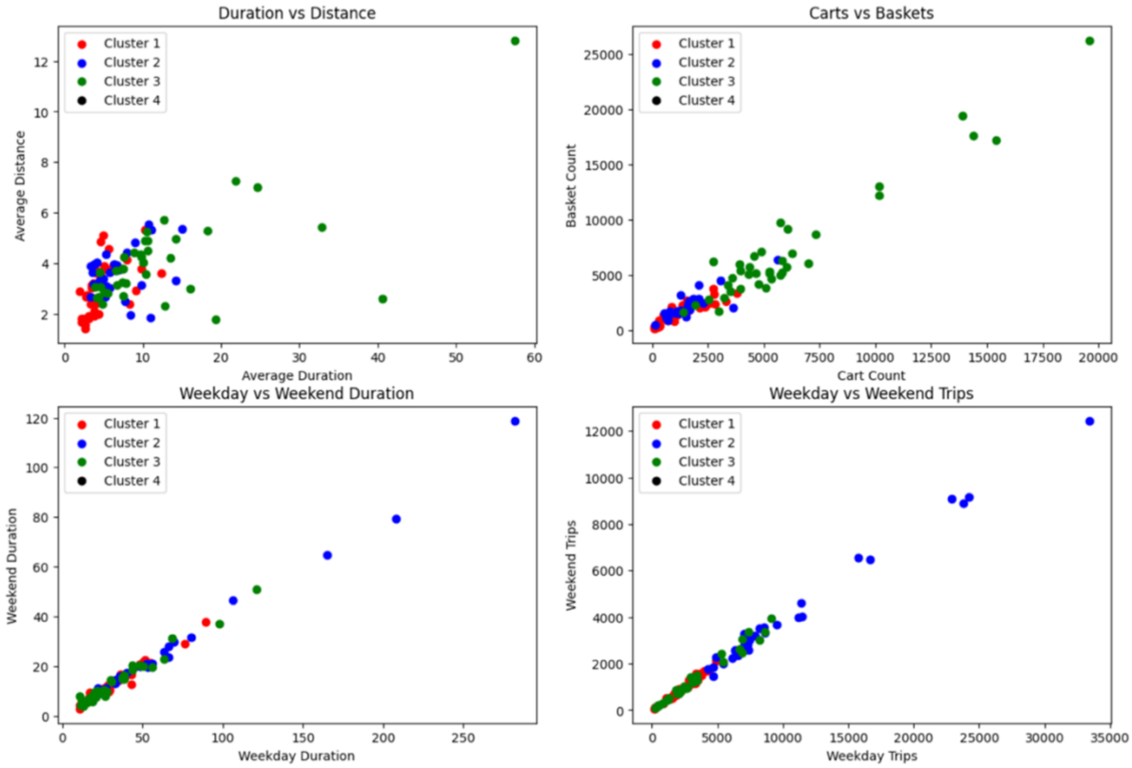

*Figure 16: Graph 'Plotting the clusters in subplots (Spectral Clustering)' Store 1*

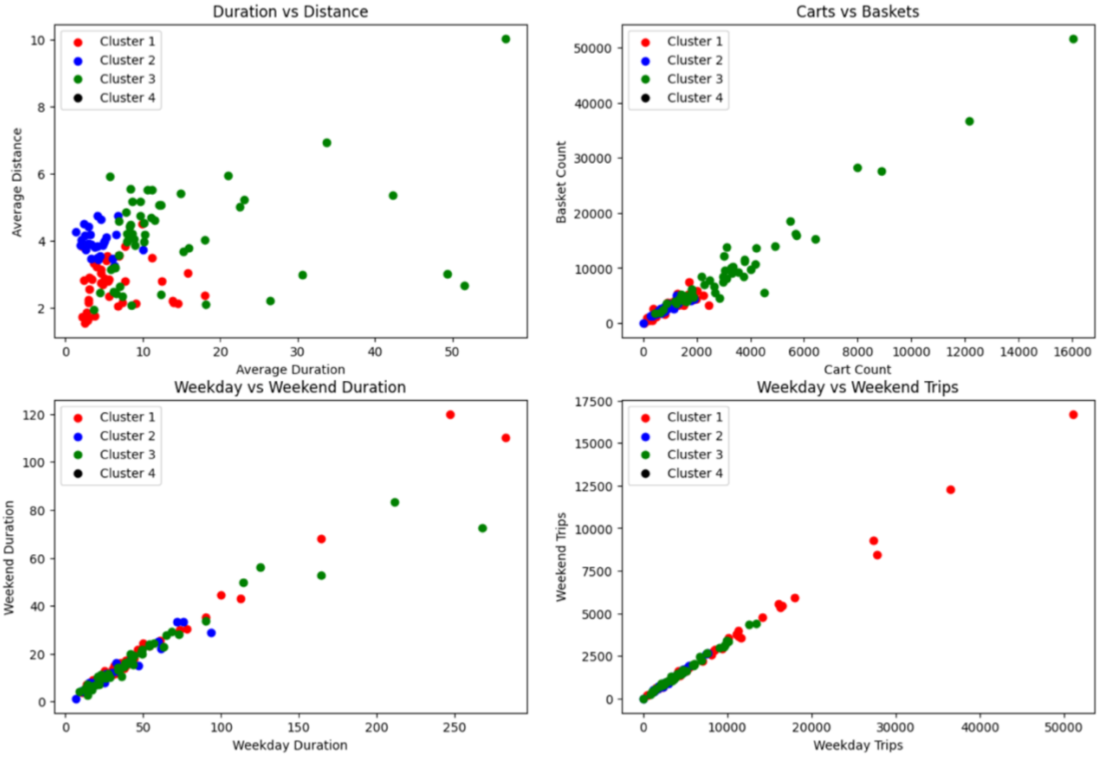

*Figure 17: Graph 'Plotting the clusters in subplots (Spectral Clustering)' Store 2*

## 7.3 SILHOUETTE SCORE: COMPARE CLUSTERING ALGORITHMS

The silhouette score is a metric used to study the proximity of points belonging to a cluster with points belonging to neighbouring clusters; it takes a value within the range of -1 to +1.

A value close to 1 indicates that the sample is far away from neighbouring clusters, thus confirming the effectiveness and validity of the clustering procedure just performed; a value of 0 indicates that the sample is on a decision boundary between two neighbouring clusters, in which case it would be appropriate to change the number of clusters in order to perform a different clustering; finally, a value equal to or close to negative -1 means a wrong assignment of values to the clusters.[25]

Below (Figure 18 and Figure 19) are the silhouette score representations of store 1 and store 2.

[25] Selecting the number of clusters with silhouette analysis on KMeans clustering' Scikit-learn.org

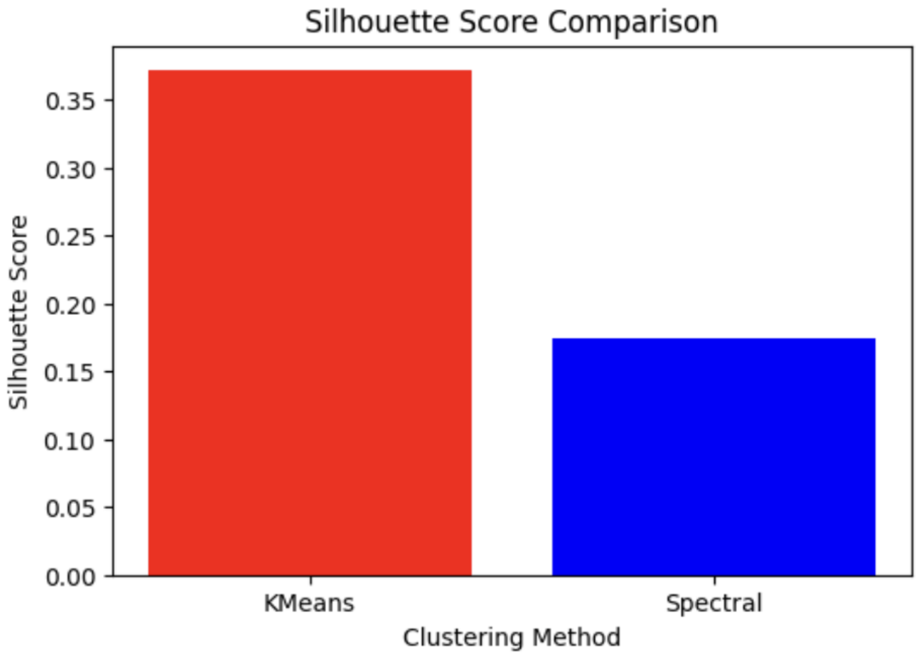

*Figure 18: Graph 'Silhouette Score Comparison (KMeans and Spectral Clustering)' Store 1*

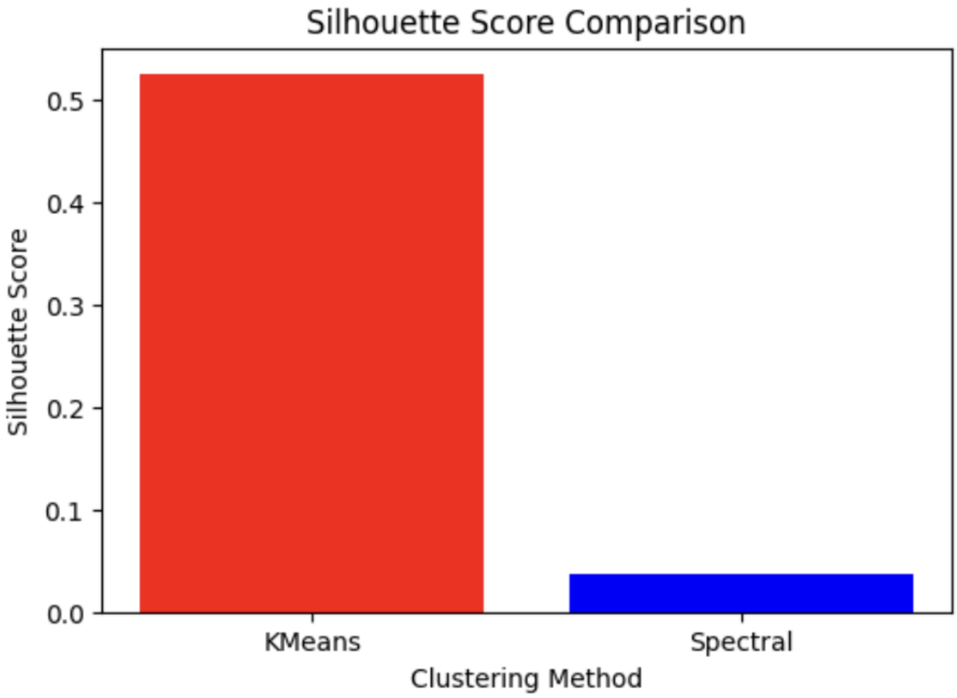

*Figure 19: Graph 'Silhouette Score Comparison (KMeans and Spectral Clustering)' Store 2*

As can be seen in Store 2, there is a Silhouette Score for the procedure by Spectral Clustering close to 0, in which case a change in the cluster numbering can be opted for in order to achieve a more effective division of the observations by Spectral Clustering.

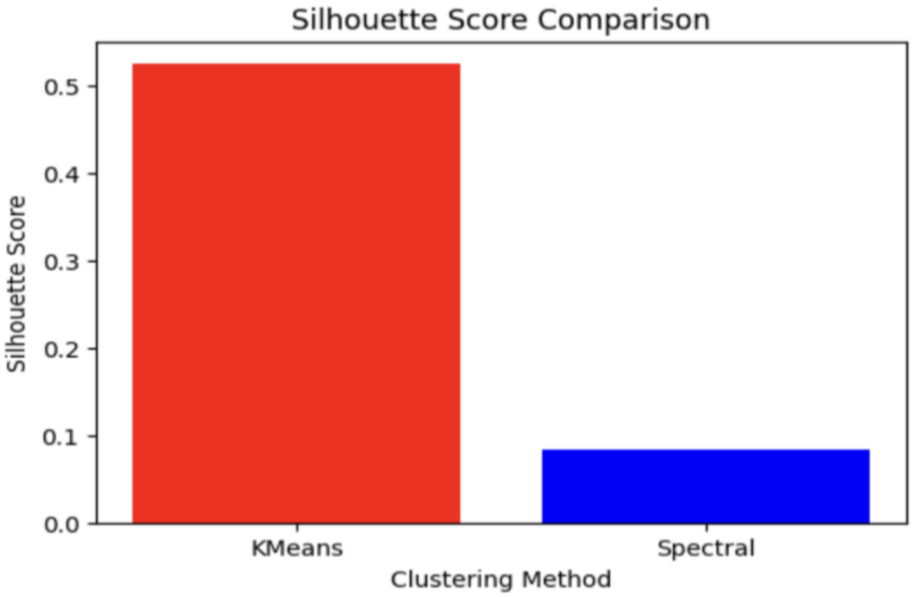

*Figure 20: 'Silhouette Score Comparison (KMeans and Spectral Clustering (n_clusters=4))' graph Store 2*

Increasing the number of clusters to 4 increases the silhouette score for the clustering procedure to a value close to 0.1 (0.08340433003492764), a value that is not equally satisfactory for defining optimal clustering by spectral clustering.

To observe whether the spectral clustering procedure was not effective or the number of clusters selected was incorrect, the number of clusters searched for was subsequently increased again to 5.

Figure 16 shows the silhouette score comparison of the two clustering methods used.

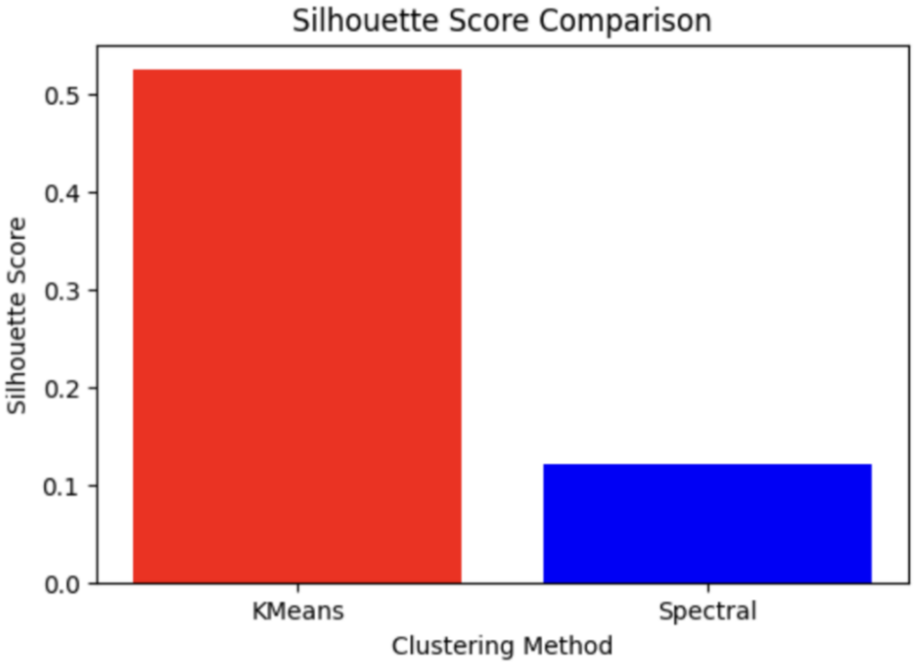

*Figure 21: 'Silhouette Score Comparison (KMeans and Spectral Clustering (n_clusters=5))' graph Store 2*

The value assumed by the Silhouette Score for the Spectral Clustering procedure after the modification made is greater than 0.1 (0.12066745165317455), from these changes it can be deduced that in Store 2 the observations in order to be clustered more accurately need to use a higher number of clusters, but at the same time the more selective and precise clustering does not significantly improve the efficiency of the clustering model.

# 7. INTELLIGENT RETAIL ENVIRONMENTS AND GDPR

*Written section with the participation of Privacy Officers lawyers experienced in EU Regulation No. 679/2016 (GDPR) and data protection and privacy:*

*Lawyer Telese Sofia, Lawyer Inzerilli Lucia.*

The use of artificial intelligence in retail allows the collection of databases useful not only for a reliable history of customer behaviour, but also for all forecasting activities that can be implemented by processing this information with learning algorithms that, as in this thesis, manage to produce output useful for defining strategies and implementations aimed, for the most part, at increasing turnover.

At the same time, however, when customer profiling tools are implemented, such as cameras for counting customers or vertically on the shelves for detecting interactions with products, trajectory sensors for studying and predicting customers' purchasing behaviour, visual scanner sensors for obtaining data on visual points of interest, etc., it is important to consider the subtle boundary that separates the use of artificial intelligence for behavioural study purposes declined under the different researches that can be conducted with these databases and the personal privacy of the individual whose personal space is invaded as a result of the surveys conducted.

It is useful to approach this topic from two different perspectives, on the one hand from a more ideal, ethical and theoretical point of view, and on the other hand to look at the problem from a more practical and concrete point of view and, above all, to be able to define a broad strategy on how to make it legally possible for companies to profile their customers without incurring fines from the privacy guarantor.[26]

To address the issue of ethics in the collection of personal information for a purpose. Of profiling, it is necessary to define a code to adhere to in order to be ethical and respect the consumer's privacy as much as possible.

The main actions to take if you are collecting or processing.

Responsible data management: customer data must be used responsibly and stored securely, their deletion and anonymisation must not only be an obligation or constraint on the processing of personal data but a basic ethical principle when dealing with sensitive data.

If the data collected is to be shared with third parties, it must be ensured that it is handled with parties that will have data storage methods and techniques that guarantee data integrity.

Ensure transparency on data collection: it must be made clear how any personally identifiable information acquired within the smart retail space is used and for what purposes. The retention of customer data must be

[26] The ethics of customer behavioural tracking' Jascha Kaykas Wolff 2021

clear and defined, they must remain in the company's datasets for no longer than a period of time considered necessary, the length of retention must coincide with the company's own privacy policy.

Ensure transparency on how customer data are shared: Should the sharing of customer identification information with third parties be required or necessary, standards of data security and integrity required both by the guarantor and established internally for the ethical handling of such data must be adhered to. Customer data, if shared, will require the explicit prior consent of the data subjects unless otherwise stated.

Be transparent about customers' rights: Data collected through artificial intelligence technologies for customer monitoring must be used for purposes indicated prior to the collection and acceptance of consent, customer data must be used in a way that meets their expectations, provides them with access to their data and allows them to delete or correct the information collected. All stakeholders involved must also abide by these rules.

Use trusted third-party tools: customer data should only be collected through trusted third-party tools that make the integrity of the collected data one of the core values underpinning customer research and sampling and have a proven track record of compliance with industry standards and best practices. They in turn need to meet the specific requirements for safeguarding customer privacy and at the same time provide detailed and available information on how the data collected from customers is used.

Ethical monitoring entails obligations on the part of the company itself and probably additional costs and/or lower revenues. At the same time, being recognised as a company that makes ethical data collection and processing one of the founding values on which the entire activity of analysis, research and customer profiling is based, would allow the positioning of the company itself through a fundamental value of distinction from competitor activities and a relative increase in customer loyalty.

To sum up, we can state that ethicality in the monitoring, collection, processing, deletion and use of data must be a fundamental concept, if the artificial intelligence tools are to be implemented, must invade the privacy of the individual in order to function properly.

The argument concerning the ethicality of processing is a concept that must be internal to the company promoted to the subjects whose data is processed and various stakeholders; at the same time, one must be compliant in order to be able to carry out one's data collection activity without incurring fines or suspension of activity due to the intervention of bodies such as the privacy guarantor, who is in charge of overseeing the correctness with which companies process personal data.

Of particular interest for the topic addressed in this thesis is Article 35 of the GDPR[27] which mentions in its first paragraph:

[27] https://gdpr-text.com/it/read/article-35/

- Article 1: "*Where a type of processing operation is likely to present a high risk to the rights and freedoms of natural persons when in particular it involves the use of new technologies, taking into account the nature, subject-matter, context and purposes of the processing, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data. A single assessment may examine a set of similar processing operations presenting similar high risks.*"

In the third sub-section, in fact, cases are listed where the implementation of a DPIA (Data Protection Impact Assessment) is necessary and mandatory; the cases envisaged by the Privacy Guarantor in the article are as follows:

(a) '*where a systematic and comprehensive assessment of personal aspects relating to natural persons takes place, which is based on automated processing, including profiling, and on which decisions are based that have legal effects or affect such persons in a similar significant way*'; an activity that will certainly be carried out where data are collected by stores or third parties for the implementation of databases, such as the two datasets analysed and clustered within this thesis

(c) '*large-scale systematic surveillance of an area accessible to the public*', since the activity is carried out in retail environments that are open and accessible to the public, DPIA is also necessary in this case in order to be compliant with the GDPR.[28]

The GDPR in Article 25 stipulates that the data controller must guarantee for the processing of the data mentioned above:

- by design (privacy by design); and
- by default (privacy by default);

only those personal data are processed that are necessary in relation to each specific purpose of the processing, the amount of data collected and the duration of their storage do not exceed the minimum period necessary to achieve the purposes pursued.

It is therefore important that the collection of data be limited to the pursuit of the stated purposes and that the data be used only for the pursuit of the stated purposes.

In any case, it is important to emphasise that the choice of the most appropriate measures to comply with the principles of privacy by design and by default are not and cannot be defined a priori by the legislator, but are left to the individual data controller, who has the task of identifying them on a case-by-case basis, on the basis of his risk assessment, with reference to the specific processing carried out and the type of data used.

Therefore, the profiling and collection of customer data is not only possible from a technical point of view, i.e. by looking at artificial intelligence devices integrated in retail environments, the focus of study of the entire

[28] Stefan Brandes, 'Compliance, ESG, Data Protection and GDPR, Whistleblowing' 24 Ore Group, 2022, p.163-178

project, but it is also possible from a legal point of view, since it is a topic already covered and provided for by the articles of the GDPR.

The document presents guidelines on the burdens the data controller must assume and all the precautions he or she must take with regard to the specific quality of the data he or she holds.

In fact, if some information is not needed for research purposes, it should not be contained in the datasets and features of the databases; the two datasets belonging respectively to the Stores of Rome and L'Aquila do not have a set of information that would have no purpose from the point of view of my research, they are in fact completely anonymised data, there are no features that indicate the sex of the person or allow any precise identification of the individual, in this precise case, in fact, an automatic anonymisation of the data took place, associating an id represented within the dataset under the heading id_people_rtls composed of a string of numbers and letters aimed solely at the recognition of the total trajectories travelled within the store.

## 7.1 METHODOLOGIES FOR COMPLIANCE IN SURVEYING

This just described is an example of how one can effectively protect oneself and at the same time exercise customer profiling, which is not portrayed as an illicit activity but as a regulated data and information collection process that can therefore be carried out in compliance with the data protection and processing rules of the country in which the surveys are carried out.

Having therefore set out the theoretical methodologies to be compliant with the obligations imposed by the GDPR, these will now be transferred to the model of intelligent retail that has been analysed in this thesis, it is necessary to find a proven methodology to carry over the theoretical notions described through the previously mentioned articles of the GDPR to enable the widespread use of artificial intelligence devices in the field of retail for the purpose of creating intelligent retail spaces.

Whatever artificial intelligence additions are made to support the personal data collection activity, a series of procedures must be put in place in order to be able to make the whole system compliant in order to carry out the survey activities aimed at creating datasets containing personal customer data and also to sell the configurations to other realities.

Whenever sensors and devices are placed in spaces open to the public that have the precise purpose of customer profiling, two fundamental principles must be put in place to be able to comply with the obligations imposed by the GDPR on the protection of personal data and thus be able to carry out this activity. They are called Privacy by default and Privacy by design. The first principle concerns all the precautionary measures to be put in place prior to the use of such customer tracking technologies, such as the definition of data anonymisation strategies and the reduction of the information collected to the minimum required for the declared surveys; the second principle, Privacy by design, instead, concerns all the strategies that must be put in place to limit access to the data collected by only those personnel interested in the use and analysis of the data collected, thus excluding persons who would have no purpose in viewing and using the datasets.

Having, therefore, observed the principles of Privacy by Design and Privacy by Default, it is possible to obtain legitimate processing of personal data, which, in conjunction with the DPIA[29], i.e. the impact assessment that the data controller must carry out in order to mitigate the risks of data processing, makes it possible to carry out customer analysis activities in complete transparency, thus transforming physical stores into intelligent retail environments that, through dataset analysis processes, are able to improve the store's sales efficiency; Furthermore, the implementation of such policies would allow third party companies to sell the sensor and configuration assets to make retail environments intelligent if there is a collaborative relationship between the company that owns the store and the management figure of the infrastructure related to the components intended for data collection.

[29] Data Protection Impact Assessment (DPIA) – Garante della Privacy from https://www.garanteprivacy.it/valutazione-d-impatto-della-protezione-dei-dati-dpia-

# 8. THE NEW ROLE OF THE PHYSICAL RETAIL SPACE

It is certainly not news that the purchase of products takes place on many channels, the physical store being only one of the channels available for customers to make purchasing choices. Despite the fact that the physical store is the oldest and most proven retail channel present, other types of channels are often preferred for purchasing products because they allow a quick comparison of product types, greater assortment and more distributors selling products at more competitive prices not having to bear the high fixed costs incurred by those with a physical sales channel.

The Covid 19 pandemic was another strong negative for physical retail, leading customers to become increasingly familiar with online shopping. Despite this, a survey conducted by Shopkick (an American company that has created a shopping app for smartphones and tablets that offers users rewards for purchases made on online and offline platforms) found that the role of physical retail outlets is still central; the study found that 93% of Americans still prefer to shop for essentials in physical stores.[30]

The world of physical retail has undergone dramatic changes over the past decade. Sales through online channels have increased by an average of 15 per cent every year since 2010 and are expected to reach 25 per cent of the US retail market by 2024[31] .

The omnichannel marketplace allows consumers to interface with the product via multiple physical and online touchpoints, omnichannel customers increasingly expect to be able to use a combination of channels in their interactions with retailers, and retailers at the same time are seeking an offering that includes a seamless and integrated shopping experience.

Physical stores, therefore, are not destined to disappear, but must take on a new role and be rethought.

There are many tried and tested methods designed to make the retail space more intelligent and rewarding on the consumer shopping experience side, so-called BOPIS (Buy Online Pick up In Store) areas have been created in recent years to try to reduce the time spent shopping and make the product evaluation and purchasing system more fluid.

At the same time, however, with this thesis we want to propose a new role that retail spaces must assume in order to survive, in fact, through the creation of databases by means of detection devices it is possible to gain access to a series of information that could never be obtained through a different sales channel.

[30] Praveen Adhi ‘Reimagining the role of physical stores in an omnichannel distribution network' McKinsey (2021)

[31] 'Omnichannel Operations: Challenges, Opportunities, and Models' Stefanus Jasin, Amitabh Sinha & Joline Uichanco (2019) pages 8-10

The core of the concept on the future of retail spaces is indeed the uniqueness unlike other channels of having physical persons buying, these within retail spaces take on unfiltered attitudes as is the case with online channels but the latter are profiled for their actual attitude.

The search possibilities with the observations collected in the physical stores are almost infinite, by initially deciding on the features to be analysed, the parameters with which to carry out these searches and the technology used to do so, it is possible to structure databases capable of responding to all the data analyses one wishes to conduct.

In the present thesis, a wide-ranging investigation was shown to search for sectors that responded equally to the selected features by means of clustering measures, but the malleability in the construction of datasets makes it possible to search to an extent that can be considered exhaustive for the answers to be obtained.

Thus, the physical shop will increasingly take on an omnichannel role by being able to integrate into the classic physical store model the best features of online channels, such as: optimisation of the store for the improvement of the user's shopping experience, more competitive products and positioning of these in more strategic locations.

In this way, the physical store will not only have the advantages of the physicality of the store, i.e. the ability to view, test and evaluate products in person, but also the advantages of an analysis of data created on the human behaviour of individuals within the store.

# 9. CONCLUSIONS AND FUTURE DEVELOPMENTS

The aim of the thesis was to provide an investigation methodology that would enable a new research approach within physical retail spaces; that is, the creation of an artificial intelligence sensor system capable of recognising consumers, following them within the store and analysing aspects of interest for the purposes sought.

The research conducted through the use of such sensors proved to be a powerful business analysis tool, the data that was collected and subsequently analysed provided valuable information on the trajectories travelled and the behaviour of consumers within the retail space, thus enabling the identification of trends, navigation patterns and areas of interest within the store.

In this thesis, an investigation mode was identified and explained under the different sections of which it is composed. The proposed hardware set-up is based on the use of artificial intelligence devices that communicate with a special anchor located in the ceiling of the store. The sensors detect individuals within the retail space and record their movements and duration of transit; the data collected by the sensors allow the study of trajectories, duration times within the sectors and the detection of all the other features analysed in this thesis.

Subsequently, a cluster analysis was conducted, the ultimate goal of which is the construction of clusters by means of unsupervised learning algorithms, in order to observe in greater detail how the sectors, on which the clustering procedures were carried out, assumed similar behaviour according to the selected features.

An important future development of this research would consist in the integration of the data collection component with the activities carried out within the store, in fact if the datasets provided had adequate data collection integrated with promotional activities carried out and the relative time frame in which these were in place, or with the presence of non-routine products or products not available in other seasons on the shelves, it would be possible to arrive at conclusions that cannot be deduced with the available information.

By means of a data collection and dataset creation that includes the presence of the moment of detection and the other perspectives listed above, it is possible to carry out analyses that can study the effectiveness of a promotional activity, implement strategies to modify the purchase behaviour in areas that are not optimal under the various observation parameters, optimise the purchase and sales strategy of seasonal products, and modify the assortment of products in the sector.

These are just some of the objectives that can be implemented if the situation of the physical retail environment is studied using the survey methodology proposed in this thesis.

Furthermore, an interesting future development for this investigation would be to be able to develop the following analysis on different datasets from different stores, being able to manipulate elements of difference and uniqueness between the different stores analysed would allow the study of the influence of various factors in the shopping experience of individual stores and also allow the creation of models that study the optimal combination of features to make the retail environment as optimal as possible.

At the same time, the presence of more advanced systems or integrated with other devices that are able to combine the analysis and study of trajectories with tools that are able to perform a more invasive tracking such as eye tracking to obtain which products or services attract the most attention, would be able to complete customer profiling by integrating numerous observation parameters and guaranteeing more integrated datasets that allow clustering research to be conducted on further features.

This new methodology would also be applicable through an integration between online and offline retail channels. Being able to integrate the datasets with the detections obtainable through artificial intelligence devices in physical retail spaces with the profiling that can be implemented through online channels would allow for greater integration and differentiation of the provenance of the data, making the mechanism more effective and truer to reality.

The future developments listed above are profiling methodologies that follow several steps. In fact, the consumer's purchasing behaviour generates data that is then analysed, and the analyses conducted are carried out from a business analysis perspective, which provides important insights into the efficiency of the sales process. They certainly represent a major step forward compared to the research methods conducted so far, but the ultimate goal of the analysis and profiling process is to eliminate the steps in the analysis process through the creation and implementation of artificial intelligence algorithms that enable automatic and real-time analysis, providing immediate insights that ensure timely and informed decisions.

Automated profiling, carried out at the same time as the consumer is carrying out his or her shopping process, would allow the creation of a personalised shopping experience even in physical retail spaces dependent on the individual's cluster of reference. Automated and real-time customer clustering, as carried out for the sectors in this thesis, would allow the creation of preset shopping experiences for clusters of consumers created on the study of analysed feature intersections.

A further useful future development for the research conducted would be the creation of an association mode matching the individual cluster to a specific consumer profile. In fact, the topic just presented is included within the chapter of future research developments as the latter could be improved and rendered more comprehensive through the addition of further AI devices within retail environments, such as RGBD sensor, sCREEN navigation system, Top-View Configuration, i.e. systems seen in previous research, they are listed within the scientific papers in the second section of the literature review.

The trajectory clustering algorithm will be used, in addition to the further procedures created through the collection of information with the various devices, to deduce the most common trajectories of visitors, exploiting the traces of the global positioning system and the sensors[32].

Through the creation of an artificial intelligence device system composed of an optimal combination of the above-mentioned sensors, a more in-depth profiling of customers would be feasible, which does not occur through the sole use of sensors aimed at trajectory tracking.

In fact, this would allow the creation of a profile of the Buyer Personas that interface with the Store and a relative association with the Sector belonging to a single cluster.

Through the infrastructure of artificial intelligence devices, it is possible to create datasets enriched with significant features for the research to be conducted.

In this way, the clustering procedures will be more complex and exhaustive at the same time, as the clustering of the feature sector will depend on more features than the clustering performed on the datasets used in this thesis.

[32] Pierdicca Roberto; Paolanti Marina; Vaira Raffaele; Marcheggiani Ernesto; Malinverni Eva Savina; and Frontoni Emanuele (2019 ‘Identifying the use of a park based on clusters of visitors' movements from mobile phone data’ *Journal of Spatial Information Science*: No. 19, 29-52.

# 10. APPENDIX

The repository, where is possible to view the entire script used for the clustering procedures of the thesis, is accessible via the appropriate link: https://github.com/lona-collab/Master-Thesis-Lorenzo-Nanni.

# 11. EXTENDED ABSTRACT

EXTENDED ABSTRACT OF THE ELABORATE: "An analytical trajectories method for understanding shoppers' buying patterns in intelligent retail environment for a Business Analysis purpose." by Lorenzo Nanni.

Artificial intelligence is a branch of computer science that has been booming in recent years. It deals with the development of algorithms and systems that are capable of performing tasks that would normally require the intervention of human intelligence. Being an ever-evolving science, it is not possible to identify a founder, but at the same time it is possible to identify and give credit to the people who, in the mid-twentieth century, already theorised the existence of a new branch of computer science that aimed to replicate the functioning of the human mind in performing tasks.

They are John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon who coined the term for the first time in their lecture at Darmouth College; Alan Turing who, with his studies in this field, proposed the Turing test which, despite the very rapid change in technology, is still in vogue today; Herbert Simon and Alan Newell who first developed a programme that worked by machine learning.

With the exponential increase in the computational capacities of computers, artificial intelligence has been able to make enormous leaps forward in a very short time, so much so that nowadays it has so many facets that it is present in the everyday life of every individual, when searching on browsers, the results provided are the work of artificial intelligence algorithms that filter out the most relevant results, facial unlocking on devices exploits learning algorithms centered on the analysis of physiognomic data of faces and recognition of people.

There are countless examples that can be given in this regard, but, at this point, given the very high level of specialisation and integration of artificial intelligence in the everyday world, the question arises as to how the available tools can best be exploited.

In particular, in this thesis, the focus of the research will be on artificial intelligence devices that can be used in retail environments in such a way as to achieve intelligent retail spaces that can provide databases that allow for data analysis leading to benefits on the user side by improving their shopping experience and on the retailer side whose improvements can lead to an increase in the average consumer's receipt and thus to a higher overall turnover.

Going into more detail, this thesis will deal with the subject of analysing the trajectories made by customers in retail environments in order to obtain datasets of observations that can be analysed by special data analysis programmes to obtain important information to be used in the field of business analysis, these datasets will be analysed by means of a series of steps, the focus of the thesis is not to obtain an output sought to justify deductions on customer behaviour within the store, but rather to develop a methodology indicating the steps to be followed for the profiling and study of trajectories in hybrid retail environments.

The main objective of this thesis is to create a methodology that combines tools and analysis techniques from the world of data science to obtain answers that can be read from a business analysis perspective.

Understanding and translating the output from programmes that analyse datasets composed of surveys carried out in retail environments into a business perspective is a source of competitive advantage for retailers, as it enables them to adapt marketing strategies to improve the customer experience and make informed decisions on product placement and space management.

Clustering procedures are a set of algorithms that have the function of grouping individuals and objects based on their similarities, their application is of particular importance in the present study as they allow for the analysis of datasets by providing a clustering of observations depending on the variable of interest studied that contains homogeneous points within and different points between the different clusters.

The output from programmes using analysis by means of unsupervised learning algorithms can be read according to a different key than a basic statistical interpretation, in fact, depending on the study to be conducted, it is possible to modify the search parameters in order to obtain output that allows investigations aimed at the efficiency of a company section or a particular process.

This thesis is subdivided according to the different steps of which the proposed investigation methodology is composed. In fact, after an initial chapter concerning the literature review composed of both scientific papers analysing unsupervised learning algorithms and scientific articles concerning previous studies carried out on the subject, the sections where the most technical procedure will be shown, i.e. the processing and application of clustering algorithms to datasets, will be dealt with first, followed by the chapters concerning the keys to reading and interpreting the output produced by software.

The main objective of this thesis is to create a methodology that combines tools and analysis techniques from the world of data science to obtain answers that can be read from a business analysis perspective.

Understanding and translating the output from programmes that analyse datasets composed of surveys carried out in retail environments into a business perspective is a source of competitive advantage for retailers, as it enables them to adapt marketing strategies to improve the customer experience and make informed decisions on product placement and space management.

Clustering procedures are a set of algorithms that have the function of grouping individuals and objects based on their similarities, their application is of particular importance in the present study as they allow for the analysis of datasets by providing a clustering of observations depending on the variable of interest studied that contains homogeneous points within and different points between the different clusters.

The output from programmes using analysis by means of unsupervised learning algorithms can be read according to a different key than a basic statistical interpretation, in fact, depending on the study to be conducted, it is possible to modify the search parameters in order to obtain output that allows investigations aimed at the efficiency of a company section or a particular process.

The main objective of this thesis is to create a methodology that combines tools and analysis techniques belonging to the world of data science to obtain answers that can be read from a business analysis perspective.

The retail sector in recent years has seen an increasing need to acquire information on customer behaviour and how they move around the store. The information gathered makes it possible to optimise shop layouts, improve offerings, innovate products and maximise profits by monitoring customer behaviour and choices in real time.

There is a need to research and test strategies that transform the retail space into a more advanced environment under the data customer analytics side, in fact, the point of sale is no longer the only place where shopping takes place and for this reason it must be able to quickly align itself with customer needs, transforming spaces and the shopping experience.

Through the use of integrated devices in retail environments based on artificial vision, indoor tracking systems and distributed sensors for environmental monitoring, this can be achieved.

The main focus of the research activity was therefore the creation of a survey methodology based on the use of indoor tracking sensors associated with the baskets and trolleys used by customers inside the store, which communicated in real time with an anchor in the ceiling of the store itself, in order to collect the trajectories travelled by customers and monitor their movements.

Understanding and translating the output from programmes analysing datasets from retail environments into a business perspective is a source of competitive advantage for retailers, as it enables them to adapt marketing strategies to improve the customer experience and make informed decisions on product placement and space management.

Clustering procedures are a set of algorithms that have the function of grouping individuals and objects according to their similarities, their application is of particular importance in the present study as they allow the analysis of datasets by providing a grouping of observations depending on the variable of interest studied that contains homogeneous points within it and different points between the different clusters.

The output from programmes using analysis by means of unsupervised learning algorithms can be read in a different key than a basic statistical interpretation, in fact, depending on the study to be conducted, it is possible to modify the search parameters to obtain output that allows investigations aimed at the efficiency of a company section or a particular process.

The managerial implications that will be discussed within this thesis will concern the use of the information collected and read from a business perspective for the formulation of strategies aimed at increasing the sales efficiency of the store.

The proper creation of a setup consisting of artifical intelligence devices that allow for data collection guided by the objectives sought, and a proper process through code steps on software that from the datasets obtained in the previous phase extrapolate insights and output consistent with the objective to be pursued, would offer a key to understanding observations that have never been analysed and understood until then.

The aforementioned topic will be addressed in its various related aspects. The focus will be on the technical research methodologies, i.e., the scripts with which to manage and interpret large datasets, from the point of view of the artificial intelligence technologies to be used and their cost of integration with retail environments, and finally from the point of view of the outputs sought through the collection of datasets and their analysis.

The important objective of the thesis is therefore to find a methodology to extrapolate results from datasets, the observations of which are collected through guided analyses of hybrid retail environments, to identify important insights into consumer purchasing habits and how to set up one's retail environment in order to increase one's business through the aforementioned targeted analyses.

Artificial intelligence is increasingly present within our lives, the amount of tasks it can reproduce and the intelligence with which it learns and improves over time is set to increase exponentially in the coming years, so being able to utilise the power of neural networks and sensors for a business purpose will be a key skill to be able to produce an up-to-date and market-attractive offering.

The presence of hybrid retail environments where cameras, proximity sensors, eye movement trackers, facial, emotion and behavioural recognition are flanked by classic store fixtures could become an everyday occurrence in the coming years, and having an accurate report and survey methodology could be a great source of competitive advantage.

The possibility of being able to dispose of a dataset with such a number of observations that it would be possible to extend the observations obtained in hybrid environments to all retail solutions is the objective that this thesis sets itself. Being able to provide through business analysis and data analysis of the aforementioned datasets allows a complete integration between the different information and therefore a solution that is as effective as possible and extremely anchored to analytical deductions, thus succeeding in reducing human error or the margin left to qualitative deductions which, having no scientific reference, would lead any decision towards an initially formal and when applied substantial inaccuracy.

In this thesis, an analysis of the literature review concerning the subject is initially carried out with research that will focus on the previous research concerning the use of artificial intelligence for business functions. Subsequently, scientific papers will be analysed by researchers who have already implemented artificial intelligence solutions in the retail world in different ways and finally, this chapter will conclude with a part containing more theoretical papers to be able to better understand the functioning of the clustering algorithms and, more generally, the data analysis methodologies used to illustrate the investigation methodology proposed in the thesis.

Going into greater detail, without listing the supporting scientific papers, initially in this section of the thesis we will present the state of the art that has previously examined the role of artificial intelligence in the context of assisting in obtaining data in the retail environment, then we will move on to the second step where we will analyse previous research that has focused on topics similar to those proposed by this thesis in order to obtain important information regarding the methodology used to conduct the analysis, the technologies used in the previous studies and finally the conclusions obtained from their research; finally, the third and final step of the literature review analysis will concern the section of studies conducted and the theoretical background of the machine learning methodologies that were used to analyse the two datasets.

The datasets, used to show the survey methodology in practice, are composed of the observations collected inside two physical shops, the first one, i.e., the one relating to the store located in L'Aquila, is called the Store 1 dataset while the one relating to the store in Rome is called the Store 2 dataset, the number of observations relating to each dataset is listed below:

- The dataset for Store 1 consists of 732114 observations and 11 columns.
- The Store 2 dataset consists of 971653 observations and 11 columns.

The names of the 2 datasets are as follows: Store1full.csv and Store2full.csv, they are two files in.csv format, the columns are labelled with the following indices:

id_people_rtls; description; time_start_aoi; time_end_aoi; distance_aoi; time_start_full; time_end_full; distance; sector; first_sector; last_sector.

The thesis will then proceed to illustrate the dataset analysis procedures that were used to achieve the objectives shown in the following chapters. The analysis starts from the loading of two .csv files containing two datasets respectively of the Store in the city of Rome and the Store in L'Aquila into the Google Colaboratory analysis software.

Preliminary processing enabled the reduction of total observations through the application of outlier removal, correction of data types of different columns, removal of columns, and normalisation of data through the MinMax Scaler technique.

Once this preliminary phase had been carried out, the K-means clustering algorithm was initially applied for the column sector and then the unsupervised learning algorithm 'Spectral Clustering', again for the feature sector; finally, the two different algorithms were evaluated using the silhouette score to assess their accuracy and fit to the model of the algorithms just listed.

Going into more detail, we can divide the clustering analysis into several parts in order to have a technical reconstruction of the entire process as detailed as possible.

The cluster analysis that was conducted on the two datasets is a data analysis technique that explores naturally occurring groups within a dataset, known as clusters.

Clustering is a powerful machine learning method that involves the grouping of data points. With a set of various data points, the dataset under consideration can be clustered by using a clustering algorithm to classify each point within a particular group.

Indeed, data points in the same cluster contain similar characteristics or properties, on the other hand, data points in separate clusters contain highly unique characteristics or properties; different from those of observations belonging to different clusters.

Clustering is an unsupervised learning method as it is a technique that does not require the use of data to train the model with associated labels or targets. In fact, clustering algorithms aim to identify the intrinsic structure within datasets by subdividing them into groups or clusters based on the similarities between observations.

The datasets were analysed, as reported above, using the data analysis software Google Colaboratory, or 'Colab' for short, is a product of Google Research. Colab allows anyone to write and execute arbitrary Python code through the browser and is particularly suitable for developing machine learning models, data analysis and algorithm training. More technically, Colab is a hosted Jupyter notebook service that does not require any configuration to be used, while providing free access to computing resources, including GPUs.

Both datasets were subjected to the same data analysis activity, i.e. they were subjected to the same script run, since the objective of the analysis conducted is to illustrate the survey methodology for gaining insight also through the interpretable differences between the different outputs, and this could only be done by obtaining the same information in order to subsequently compare and interpret it.

In the thesis, the salient passages of this analysis and the discussion regarding the interpretation of the output sought will be shown.

The first part on output analysis examines the detections that can be made on the databases obtainable through artificial intelligence devices listed above.

Before carrying out the clustering analysis, which is the real aim of this work, a series of information can also be obtained by means of preliminary analyses, which include optimising the columns of the dataset, clustering and making certain key features explicit for a reading of this dataframe.

The precise code steps performed and the output methods are explained in greater technicality within the entire work; in the chapter devoted to reporting on script highlights, scatter plots are presented for the two inserted stores that refer to the various representations of the interpretations of the data sought.

The first two scatter plots refer to the number of visits in each different sector per day of the week; the Cartesian x-axis of this graph shows the different sectors belonging to the supermarkets, while the y-axis shows the numerical value of total weekly visits.

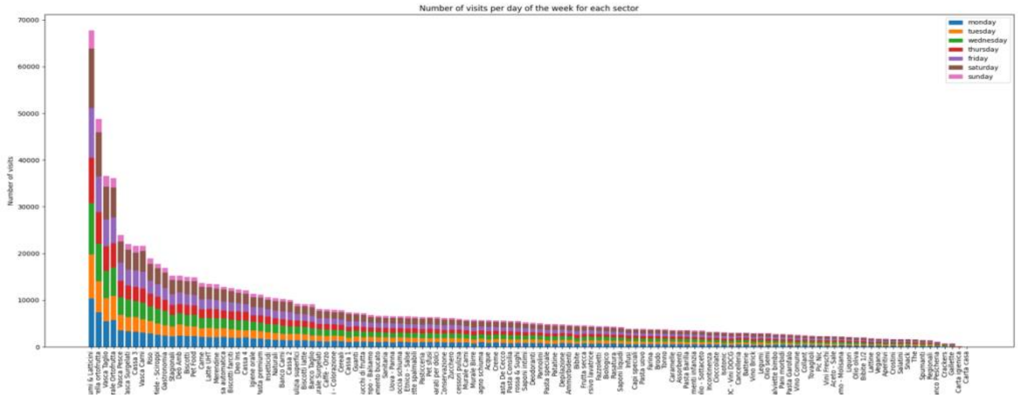

*Figure 1: example of scatterplot 'Number of visits per day of the week for each sector' of Store 2*

Further scatterplots were then graphed concerning the number of trolleys and baskets used in the different sectors, column graphs of the average distance travelled and average time spent by customers in the different sectors. Below are some examples of the graphs shown in the thesis to support the technical part in describing the survey methodology.

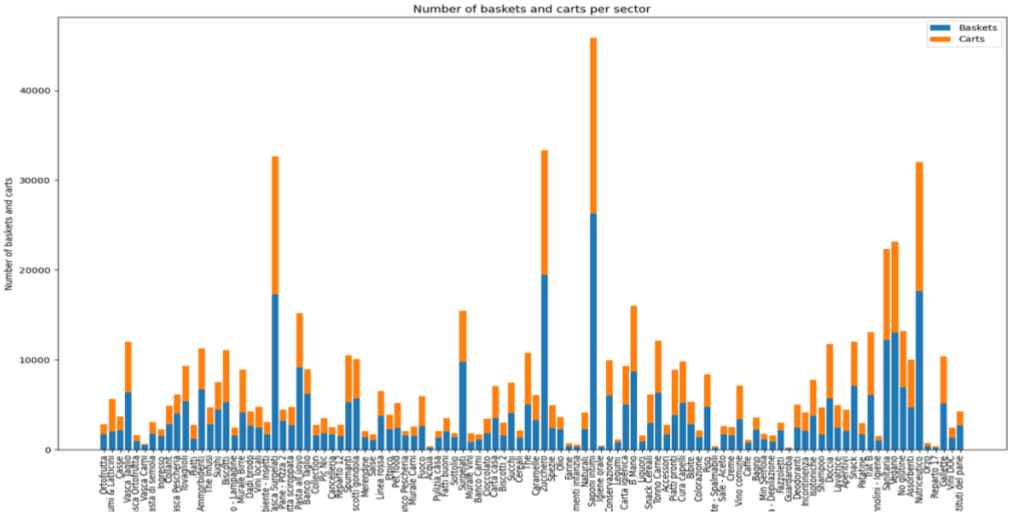

*Figure 2 example of graphic 'Number of baskets and carts per sector' for Store 1*

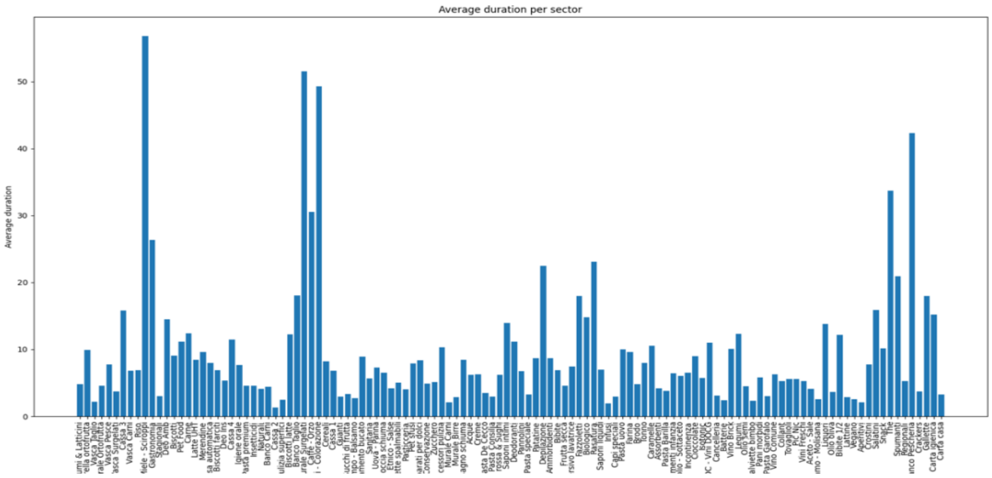

*Figure 3: Example of graphic 'Average duration per Sector' Store 2*

These graphs were then analysed from the point of view of business analysis interpretation, i.e. an investigation methodology was sought which, through the graphing of the observations obtained, would allow a reading of the problems and insight to be evaluated of the store.

There are no proven strategies for the analyses conducted in the two stores but only implementations that vary according to the findings made, for example, areas where there is an imbalance between the use of the two means for the collection of the sought after products, some areas are probably seen as time consuming, such as the cutting counter which in both stores has non-proportional concentrations of users with baskets and trolleys, this thinking may depend on a number of factors such as the fact that in order to obtain the product in these types of sections of the store it is necessary to be served by the staff, this additional step increases the waiting and picking time for the product itself, therefore, shoppers who initially chose to use the basket for quicker shopping may be reluctant to pass through and purchase products in these departments, opting more for products already pre-wrapped and available on the shelves.

Subsequently, practical indications were given to support the the analysis conducted, e.g., for the surveys carried out in the two available datasets, and with reference to the example given above, it was hypothesised that in order to increase the flow of customers in these sectors, it is necessary to act on the perception of the timing of the service provision or to attack complementary strategies such as the presence of a self-service section to reduce waiting times or to provide time-conscious consumers with a valid motivation to pass through and eventually purchase in a particular sector within the retail environment.

The factors relating to the perception of a particular sector by individuals passing through a retail space are diverse, and it is not possible to categorise them through an observation of customer behaviour in two stores, but it is necessary to carry out perception tests if an imbalance between the two components is revealed and attempt to induce users of either medium to frequent more sectors they would tend to ignore.

Then further scatterplots were shown for a strategic reading of the results obtained from the preliminary process of analysing and restructuring the features of the dataset.

Subsequently, the thesis dealt with the section on how to cluster the datasets formed from the observations collected in the store, going on to present in practice the clustering mode and algorithms conducted in the present study.

The clustering methods were discussed from a technical point of view, and in the following chapter, the results obtained from the datasets of the two stores under analysis were shown, the main objective is the investigation methodology that needs to be carried out in order to gain important insights, regardless of the type of dataset available.

Clustering analyses are in fact a method that can be implemented by means of dataset analysis software, which allows, in datasets of such high numerosity as the present, the creation of groups containing homogeneous points within the cluster and heterogeneous points between the different groupings.

This quality of unsupervised learning methods makes it possible to find within the dataset a similarity in the behaviour of clustered features and to carry the discovery made on the data into the actual operation of the store.

A discipline that deals with the identification of business problems, the search for solutions to improve the efficiency of business processes and the expression of business value to customers is Business Analysis, so within this thesis, an attempt was made to transport the evidence found through data analysis into the field of business decisions.

The implementation of such a strategy involves the steps followed for the analysis performed in this thesis, starting with the processing of data and arriving at the definition of a research model to obtain the necessary information for the business purpose.

The business analysis component is a fundamental activity for any company that wants to identify and solve problems and adapt to market changes in order to maintain an important competitive advantage.

In fact, retail environments are a prime example of activities that require periodic analysis and monitoring of the observations collected in order to identify as quickly and effectively as possible problems that lead to low system efficiency or undermine its full sales and turnover capacity.

Subsequently, the output of the cluster survey in the two stores in Rome and L'Aquila was shown, and then the output of the programme was visualised from a business analysis perspective in order to gain a better understanding of the key aspects in which to focus in order to correctly read the functioning of the different sectors within the retail environments.

Cluster analysis can be performed with all types of unsupervised learning, but in this thesis, the k-Means clustering method and the Spectral Clustering method were chosen as they are particularly effective in correctly handling datasets containing a large number of observations. Obtaining datasets by means of artifical intelligence devices that record a series of complex features for each individual consumer interaction within the store, the datasets that will be created will most likely contain a complex number of observations that need to be managed and clustered with unsupervised learning algorithms that correctly analyse this type of dataset.

An example of the representation of the clustering activity by means of KMeans methodology used in the thesis is the following, the Cartesian axes in each individual row represent a feature, their combination allows us to understand the behaviour of the clusters as a result of the interaction with the other features.

The interactions sought in these two images concern the crossing of features:

- Average distance travelled within the area of interest and duration of the trip within the survey area
- Sector cluster based on use of baskets or trolleys by customers observed
- Graphical representation cluster positioning according to features travel time in the survey area on a weekday and travel time on a weekend day
- Graphical representation of cluster positioning according to the features of the number of passages in the survey area during a weekday and those made on weekend days

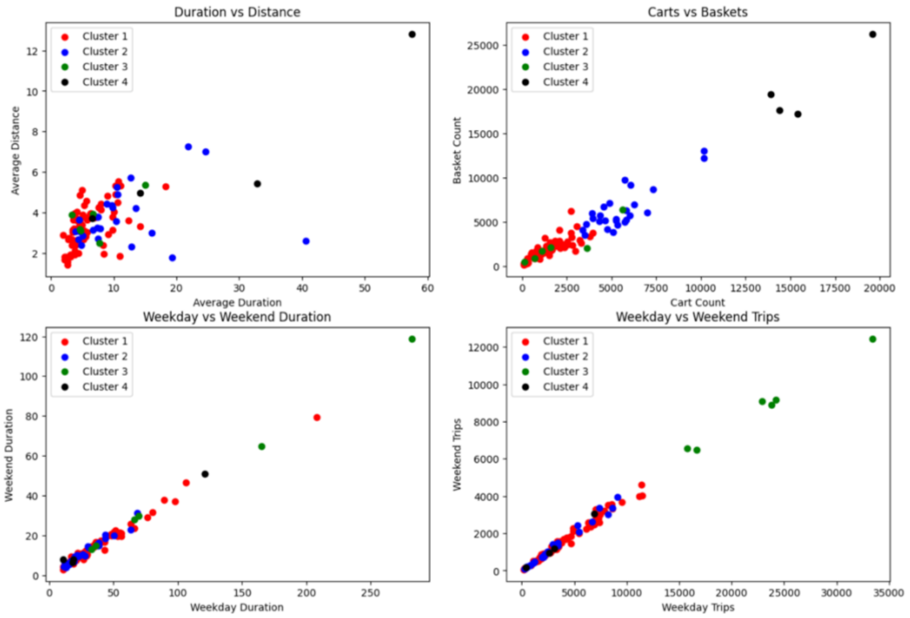

*Figure 4: Example of graphic 'Plotting the clusters in subplots (k-Means)' Store 1*

Subsequently, the thesis degressed the legal feasibility of the implementation of artificial intelligence devices and the utilisation and analysis of perosnal information, this section was included as the purpose of the thesis was to illustrate the process of analysing customer observations, it was of relevant importance to also investigate the actual feasibility from a legal perspective.

Having initially enunciated the theoretical methodologies to be compliant with the obligations imposed by the GDPR when carrying out such activities, now, these have been subsequently carried over to the smart retail model that has been analysed in this thesis, a proven methodology needs to be found to carry over the theoretical notions described through the articles of the GDPR previously to enable the widespread use of artificial intelligence devices in the field of retail for the purpose of creating smart retail spaces.

Whatever artificial intelligence additions are made to support the activity of collecting personal data, a series of procedures must be put in place to be able to bring everything up to standard in order to be able to carry out the survey activities aimed at creating datasets containing personal customer data, and also to be able to sell the configurations to other realities.

Whenever sensors and devices are placed in spaces open to the public that have the precise purpose of customer profiling, two fundamental principles must be put in place to be able to comply with the obligations imposed by the GDPR on the protection of personal data and thus be able to carry out this activity. They are called Privacy by default and Privacy by design. The first principle concerns all the precautionary measures to be put in place prior to the use of such customer tracking technologies, such as the definition of data anonymisation strategies and the reduction of the information collected to the minimum required for the declared surveys; the second principle, Privacy by design, instead, concerns all the strategies that must be put in place to limit access to the data collected by only those personnel interested in the use and analysis of the data collected, thus excluding persons who would have no purpose in viewing and using the datasets.

Having, therefore, observed the principles of Privacy by Design and Privacy by Default, it is possible to obtain the DPIA that would allow the customer analysis activity to be exercised with complete transparency, thus transforming physical stores into intelligent retail environments that through processes of dataset analysis are able to improve the sales efficiency of the store; Furthermore, the implementation of these policies would allow third party companies to sell the sensor and configuration assets to make retail environments intelligent if there is a partnership between the company that owns the store and the management figure of the infrastructure linked to the components intended for data collection.

Subsequently within the thesis, the topic of the new role to be played by the physical retail space was addressed. Although the physical store is the oldest and most tried and tested retail channel present, other types of channels are often preferred for the purchase of products as they allow a quicker comparison of product types, greater assortment and more competitive prices without having to incur high fixed costs as those incurred by those with a physical sales channel.

The Covid 19 pandemic was another strong negative for physical retail, leading customers to become increasingly familiar with online shopping. Despite this, a survey conducted by Shopkick (an American company that has created a shopping app for smartphones and tablets that offers users rewards for purchases made on online and offline platforms) found that the role of physical retail outlets is still central; the study found that 93% of Americans still prefer to shop for essentials in physical stores.

Physical stores, therefore, are not destined to disappear, but must take on a new role and be rethought.
There are many tried and tested methods designed to make the retail space more intelligent and rewarding on the consumer shopping experience side, so-called BOPIS (Buy Online Pick up In Store) areas have been created in recent years to try to reduce the time spent shopping and make the product evaluation and purchasing system more fluid.
At the same time, however, the aim of this thesis is to propose a new role that retail spaces must assume in order to survive; in fact, through the creation of databases by means of surveying devices, it is possible to gain access to a range of information that could never be obtained through a different sales channel.
The core of the concept on the future of retail spaces is indeed the uniqueness unlike other channels of having physical persons buying, these within retail spaces take on unfiltered attitudes as is the case with online channels but the latter are profiled for their actual attitude.

The search possibilities with the observations collected in the physical stores are almost infinite, by initially deciding on the features to be analysed, the parameters with which to carry out these searches and the technology used to do so, it is possible to structure databases capable of responding to all the data analyses one wishes to conduct.
In the present thesis, a wide-ranging investigation was shown to search for sectors that responded equally to the selected features by means of clustering measures, but the malleability in the construction of datasets makes it possible to search to an extent that can be considered exhaustive for the answers to be obtained.

Thus, the physical shop will increasingly take on an omnichannel role by being able to integrate the best features of online channels into the classic model of the physical store, such as: optimisation of the store for the improvement of the user's shopping experience, more competitive products, and positioning of these in more strategic locations. In this way, the physical store will not only have the advantages of the physicality of the store, i.e. the possibility of viewing, testing and evaluating products in person, but also the advantages of analysing data created on the human behaviour of individuals within the store.

Ending with the conclusion, I stated that the objective of the thesis was to provide an investigation methodology that would allow for a new research approach within physical retail spaces; the creation of a detection system using artificial intelligence sensors capable of being able to recognise consumers, follow them within the store and analyse aspects of interest for the purposes sought. The sensor-based research proved to be a powerful business analysis tool, the data that was collected and subsequently analysed provided valuable information on the trajectories travelled and the behaviour of consumers within the retail space, thus enabling the identification of trends, navigation patterns and areas of interest within the store.

Finally, to summarise, a survey mode has been identified in this thesis and explained under the different sections of which it is composed. The proposed set-up is based on the use of sensors communicating with a special detection sensor that allows the study of trajectories, duration times within sectors and all the other features analysed previously.

The analysis conducted thereafter was a cluster analysis, the goal of which was the construction of clusters by means of various unsupervised learning algorithms in order to observe in greater detail how the sectors, on which the clustering procedures were carried out, behaved in a similar manner according to the selected features. At the same time, if the datasets provided had adequate detection indications, if the store map plans were updated with the promotional activities carried out and the time frame in which these were in place, and if the presence of non-routine products or products not available in other seasons on the shelves were indicated, it would be possible to arrive at conclusions that cannot be deduced with the available information.

By means of a data collection and dataset creation that includes the presence of the moment of detection correlated with the presence of in-store promotional activities, presence of non-ordinary product displays (seasonal products, bundles, cross-selling incentives), it is possible to carry out analyses that can study the effectiveness of a promotional activity, implement strategies to modify the purchase behaviour in areas that are not optimal under the various observation parameters, optimise the purchase of seasonal products, and modify the assortment of products in the sector.
These are just some of the objectives that can be implemented if the situation in the physical retail environment is studied using the survey methodology proposed in this thesis.

Furthermore, an interesting future development for the thesis would be to be able to develop the proposed analysis on different datasets from different stores, being able to manipulate elements of difference and uniqueness between the different stores analysed, this would allow the study of the influence of various factors in the shopping experience of individual stores and also allow the creation of models that study the optimal combination of features to make the retail environment as optimal as possible.

At the same time, the presence of more advanced systems or integrated with other devices that are able to combine the analysis and study of trajectories with tools that are able to perform a more invasive tracking, such as eye tracking to obtain which products or services attract the most attention, would be able to complete customer profiling by integrating numerous observation parameters and guaranteeing more integrated datasets that allow clustering research to be conducted on further features.
This new methodology would also be applicable through an integration between online and offline retail channels; being able to integrate the datasets with the detections obtainable through artificial intelligence devices in physical retail spaces with the profiling achievable through online channels would allow for greater integration and differentiation of the data's provenance, making the mechanism more effective and truer to reality.

The future developments listed above are profiling methodologies that follow several steps. In fact, the consumer, through his purchasing behaviour, generates data that is then analysed, and the analyses conducted are carried out from a business analysis perspective, whereby important insights are deduced for the efficiency of the sales process. They certainly represent a major step forward compared to the research methods conducted so far, but the goal of the analysis and profiling process is to eliminate the steps in carrying out the analysis through the creation and implementation of artificial intelligence algorithms that enable automatic analysis carried out in real time, providing immediate insights that ensure timely and informed decisions.

Automated profiling, carried out at the same time as the consumer is carrying out his or her shopping process, would allow the creation of a personalised shopping experience even in physical retail spaces depending on the individual's cluster of reference. Automated and real-time customer clustering, as carried out for the sectors in this thesis, would allow the creation of preset shopping experiences for clusters of consumers created on the study of analysed feature intersections.

A further useful future development for the research conducted would be the creation of an association mode matching the individual cluster to a specific consumer profile. In fact, the topic just presented is included within the chapter of future research developments as the latter could be improved and rendered more comprehensive through the addition of further AI devices within retail environments, such as RGBD sensor, sCREEN navigation system, Top-View Configuration, i.e. systems seen in previous research, they are listed within the scientific papers in the second section of the literature review.

Through the creation of an artificial intelligence device system composed of an optimal combination of the above-mentioned sensors, a more in-depth profiling of customers would be feasible, which does not occur through the sole use of sensors aimed at trajectory tracking.

# 12. BIBLIOGRAPHY AND SITOGRAPHY

- John McCarthy, Marvin L. Minsky, Nathaniel Rochester, Claude E. Shannon (1955) ‘A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955’ from https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904

- Emanuele Frontoni & Marina Paolanti ‘Introduction, Definition and taxonomy: AI, Machine Learning and Deep Learning’ - Customer Intelligence e Logiche di Analisi dei Big Data (LUISS 2023)

- Contigiani Marco (2017) ‘Machine vision and IoT applications in intelligent retail environments’ from https://iris.univpm.it/handle/11566/245516

- Abhijit Guha et al. (2021) ‘How artificial intelligence will affect the future of retailing’ from https://www.sciencedirect.com/science/article/pii/S0022435921000051

- Rajasshrie Pillaia et al. (2020) ‘Shopping intention at AI-powered automated retail stores (AIPARS)’ from https://www.sciencedirect.com/science/article/pii/S0969698919302887

- V.Kumar et al. (2021) ‘Transformation of Metrics and Analytics in Retailing: The Way Forward’ from https://www.emerald.com/insight/content/doi/10.1108/IJRDM-09-2020-0350/full/html

- L.Cao (2021) ‘Artificial intelligence in retail: applications and value creation logics' from https://www.researchgate.net/publication/350133675_Artificial_intelligence_in_retail_applications_and_value_creation_logics

- Robert Zimmermann et al. (2022) ‘Enhancing brick-and-mortar store shopping experience with an augmented reality shopping assistant application using personalized recommendations and explainable artificial intelligence’ from https://www.emerald.com/insight/content/doi/10.1108/JRIM-09-2021-0237/full/html

- Robert Zimmermann et al. (2022) ‘Customers’ Activity Recognition in Intelligent Retail Environments’ from https://link.springer.com/chapter/10.1007/978-3-642-41190-8_55#auth-Emanuele-Frontoni

- Marina Paolanti, Emanuele Frontoni, Rocco Pietrini, Daniele Liciotti e Adriano Mancini (2018) ‘Modelling and Forecasting Customer Navigation in Intelligent Retail Environments’ from https://link.springer.com/article/10.1007/s10846-017-0674-7

- Daniele Liciotti, Marina Paolanti, Emanuele Frontoni & Primo Zingaretti (2017) ‘People Detection and Tracking from an RGB-D Camera in Top-View Configuration: Review of Challenges and Applications’ from https://link.springer.com/chapter/10.1007/978-3-319-70742-6_20

- Rocco Pietrini, Luca Rossi, Adriano Mancini, Primo Zingaretti, Emanuele Frontoni & Marina Paolanti (2022) ‘A Deep Learning-Based System for Product Recognition in Intelligent Retail Environment’ from https://www.researchgate.net/publication/360646101_A_Deep_Learning-Based_System_for_Product_Recognition_in_Intelligent_Retail_Environment

- Marina Paolanti, Emanuele Frontoni et al. (2023) ‘The objective way to detect the path to purchase by clustering shoppers’ trajectories”

- ‘Intelligenza artificiale a supporto delle strategie di marketing nel punto vendita’ – Grottini Lab from https://www.grottinilab.com/it/tecnologie-servizi

- Shraddha Shukla and Naganna S. (2014) ‘A review on k-means data clustering approach’, International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 17 pp. 1847-1860

- Ulrike von Luxburg (2007) ‘A tutorial on spectral clustering’, Cornell Univeristy- Computer Science, Data Structures and Algorithms

- ‘On spectral clustering: Analysis and an algorithm’ Advances in neural information processing systems: pp. 849–856 di Ng, Andrew Y, Jordan, Michael I and Weiss, Yair. (2002)

- ‘What is unsupervised learning?’ IBM

  https://www.ibm.com/it-it/topics/unsupervised-learning

- Emanuele Frontoni and Marina Paolanti ‘Machine Learning and Deep Learning Tasks Supervised vs Unsupervised, Regression vs Classification’ Customer Intelligence e Logiche di Analisi dei Big Data (LUISS 2023)

- ‘2.3. Clustering’ scikit-learn.org

- ‘Lab: machine learning in python ScikitLearn' Emanuele Frontoni and Marina Paolanti, Customer Intelligence e Logiche di Analisi dei Big Data (LUISS 2023)

- ‘Google Colaboratory’ research.google.com/colaboratory

- ‘Selecting the number of clusters with silhouette analysis on KMeans clustering’ Scikit-learn.org

- ‘The ethics of customer behavioural tracking' Jascha Kaykas Wolff (2021)

- ‘Article 35 GDPR. Data protection impact assessment’

  https://gdpr-text.com/it/read/article-35/

- Stefan Brandes ‘Compliance, ESG, Data Protection and GDPR, Whistleblowing’, Gruppo 24 Ore, 2022, p.163-178

- Praveen Adhi ‘Reimagining the role of physical stores in an omnichannel distribution network' McKinsey (2021)

- ‘Data Protection Impact Assessment (DPIA)’ – Garante della privacy from https://www.garanteprivacy.it/valutazione-d-impatto-della-protezione-dei-dati-dpia-

- Stefanus Jasin, Amitabh Sinha & Joline Uichanco (2019) ‘Omnichannel Operations: Challenges, Opportunities, and Models’ pages 8-10 from https://link.springer.com/chapter/10.1007/978-3-030-20119-7_2

- Pierdicca Roberto; Paolanti Marina; Vaira Raffaele; Marcheggiani Ernesto; Malinverni Eva Savina; and Frontoni Emanuele (2019 ‘Identifying the use of a park based on clusters of visitors' movements from mobile phone data’ *Journal of Spatial Information Science*: No. 19, 29-52.