LUISS

Degree Program in Marketing - Analytics & Metrics

Course of Statistics for Marketing

DIFFUSION AND PERCEPTION OF NON-ALCOHOLIC BEER: A DATA-DRIVEN ANALYSIS

Supervisor:

Prof. Francesco Salate SANTONE

Co-Supervisor:

Prof. Michele COSTABILE

Candidate:

Tommaso COSTA

Matr. 783711

Academic Year 2025/2026

Abstract

The non-alcoholic beer (NAB) category has become increasingly important in the fast-moving consumer goods sector, due to changing lifestyles, health awareness, and new consumption habits. The category encompasses complex market structures and different consumers' perceptions.
This thesis examines the consumption, perception, and valuation of non-alcoholic beer, as well as the interplay between consumer narratives and territorial market configurations in strategic decision-making.

The research employs a comprehensive analytical framework that integrates both quantitative and qualitative methodologies. Quantitatively, territorial clustering techniques are used on market data at the provincial level to find different consumption and economic profiles based on per capita consumption, pricing, income, and category share indicators. On the qualitative side, sentiment analysis and thematic clustering are performed on user-generated content from online platforms, in order to learn more about consumers' perceptions and opinions about the NAB category, and the deeper meanings of value that come with them.

The results exhibit a positive trend resulting from the sentiment about NAB, based on different value dimensions, such as sensory quality, health and sobriety, technical credibility, and social guidance. Pattern recognition results from territorial markets show different levels of category embeddedness and consumption patterns. The integrated analysis demonstrates that market structures and consumer narratives do not consistently evolve concurrently, revealing both congruences and discrepancies across territorial clusters; indicating that successful strategies in the non-alcoholic beer sector necessitate a territory-sensitive and narrative-informed methodology, merging market-level metrics with consumer interpretation processes to facilitate positioning, pricing, and market entry strategies.

Summary

Introduction

The fast-moving consumer goods (FMCG) industry has changed substantially in the last few years, due to the changing lifestyle of consumers, the growing health awareness, and evolving consumption habits. In this context, there has been a significant focus on the market of non-alcoholic beverages, which contains one of the most dynamic subcategories: non-alcoholic beer (NAB).

Non-alcoholic beer is far more than just a substitute for alcoholic beer. At first, it spread mostly because of the situational constraints, but better brewing and dealcoholization technology have improved the quality of the product, allowing it to compete on taste, style, and variety in the beer market. The category has been gradually expanding to acquire its own experiential and identity-related factors as well.

The NAB market is becoming increasingly significant, but it is still very different from other categories. This is due to the fact that consumption patterns and diffusion differ across regions. It appears that consumers' perspectives are fragmented and multifaceted, including sensory assessments, health-related motivations, technical factors, and social implications. This dual complexity suggests non-alcoholic beer as a particularly suitable topic for integrated analytical methodologies.

There is extensive research on drinking trends and moderation of alcohol consumption for health reasons, but very limited investigation into the integration of consumer narratives and territorial market structures within the NAB category. Specifically, empirical evidence is lacking regarding the alignment of consumer perceptions with observable market dynamics, or the presence of discrepancies across diverse territorial contexts.

This thesis addresses this research gap by integrating quantitative territorial market analysis with qualitative insights.

The study is guided by four research questions that examine consumer perceptions and value dimensions, the presence of distinct narratives and usage contexts, the heterogeneity of territorial markets, and the strategic implications of analyzing narratives and market profiles. The thesis enhances the comprehension of non-alcoholic beer as a hybrid category, influenced by both market dynamics and individual interpretations.

Chapter 1 reviews the relevant literature, Chapter 2 outlines the research design and methodology, Chapter 3 presents and discusses the empirical results, and the final chapter concludes with implications, limitations, and directions for future research.

1. LITERATURE REVIEW

1.1 The situation in the FMCG and food market

1.1.1 The agri-food system and the FMCG sector

Economy and society's development are deeply connected to the Fast Moving Consumer Goods sector. FMCG are very sensitive to changes in how people behave, because of their short buying cycles, low margins, and exposure to changes in demand. Most of an household's budget is employed in the food and beverage sector, therefore it embodies a very relevant category.

The global agri-food system, is going through long-term structural changes right now. This is happening because of how the consumers are behaving, the economy is growing, cities are growing, and people's lifestyles are changing. For the reasons just mentioned, the future food demand will be about 50% higher than it was in 2013, due to the expected increase of the world population to 9.7 billion by 2050. (FAO, 2017)

This rise is insightful not just for how much food is needed but also about what kinds of food are needed. As demand rises, preferences slowly change, which has a direct effect on FMCG markets. Consumers want products that are more unique and that meet not only their basic nutritional needs, but also their needs for convenience, taste, lifestyle, and above all, their own health.

1.1.2 Urbanization and shifting patterns of consumption

The process of urbanization changes the structure of demand in FMCG markets. By 2050, the number of cities inhabitants will increase significantly, from the actual 54%, it is expected to rise to more than two-thirds.

The concentration of people living in cities is not just a demographic trend; it also has a significant effect on how purchase trends and how food systems are set up.
In this context, individuals tend to want ready-to-eat or semi-ready-to-eat foods that are packaged and easy to find through modern distribution channels. The FMCG sector meets

these needs by providing a growing variety of packaged drinks and processed foods that become an important part of daily life.

As stated by the *Food and Agriculture Organization*, urbanization has made agri-food supply chains longer, and for this reason, the distance between where food is made and where it is eaten is greater. This change has made it easier for suppliers to standardize their products and made large processing and distribution companies stronger. However, it has also changed what customers expect, as they are now used to having a lot of products available and getting great service. (FAO, 2017)

1.1.3 The big picture and how price affects people in the FMCG markets

In the last few years, inflation made the cost of living increase, causing a substantial effect on how people shop, thus making FMCG markets especially sensitive to changes in price.

The *European Commission's* Consumer Conditions Scoreboard 2025 comes in handy to frame Europe's situation. In the UE, the cost of food and drinks without alcohol increased by roughly 31.9% in a three-year span (from 2021 to 2024). (European Commission, 2025) This rise directly affected the buying power of households, which made people rethink how they spend their money.

In a time when the economy is becoming less stable, the idea of perceived value is very important when making purchases. People do not just look at the price anymore; they also think about how the quality and benefits of the product work together.

1.1.4 Consumer awareness: health, well-being, and sustainability

In recent years, FMCG markets have been influenced by the rising consumer attention to health, well-being, and, consequently, food quality. The increase in the cost of living has temporarily made the environmentally friendly topics less “appealing”, but these issues are still a big part of how many people make decisions.

The *Consumer Conditions Scoreboard 2025* reports that 43% of European shoppers have thought about the environment when making at least one recent purchase. (European

Commission, 2025) Younger people, and more educated ones, are more likely to focus on this topic, which suggests that consumer priorities are changing from one generation to the next.

The *FAO* points out that modern food systems are being asked more and more to deal with the health effects of modern diets. One of the biggest threats to public health is sugar, fat, and alcohol increase in daily diets. (FAO, 2017)
In response to these worries, people are frequently choosing products that are regarded as more balanced, meaning they can obtain enjoyment while also being good for their health.

1.1.5 Effects on the beverage industry
In the context described above, the beverage industry is especially important. In fact, drinks, beyond their functional purpose, serve a social and symbolic purpose connected to social moments, identity, and lifestyle. Because of this, the industry is very sensitive to changes in customers' wants and expectations.

Along with a growing focus on health, well-being, and perceived value, the evolution of the FMCG market is allowing new products to give traditional categories a modern twist. This information sets the stage for the non-alcoholic and low-alcohol beverage market, especially for non-alcoholic beer, which represents a consistent answer to the conflicts between pleasure, socializing, and awareness. The next paragraphs will deep dive into these dynamics, focusing on the non-alcoholic beer market and what are the factors that are making it grow.

1.2 The *NAB* market
1.2.1 What is non-alcoholic beer, and how is it different from other kinds of beer?
To begin with, it is important to explain what is meant by non-alcoholic beer, because the low-alcohol and non-alcoholic product segment has a lot of different names that show how the products are made, how they are sold, and how they are regulated. When people talk about alcohol-free, non-alcoholic, low-alcohol, and no/low alcohol, they often use these words interchangeably. This makes things hard to understand for both customers and market analysts.

Beverages with low alcohol content are commonly distinguished based on their alcohol by volume (ABV). Non-alcoholic beer is usually close to 0.0% alcohol, although there could still be some of it. *Okaru* and *Lachenmeier* propose a methodical examination of these definitions. They emphasize that the non-alcoholic and low-alcohol category is not one homogeneous category but a group of products that is meant to keep the taste and smell of traditional drinks while getting rid of the alcohol. Their work underlines the existence of many production approaches to make non-alcoholic beers, such as limiting fermentation or removing alcohol later in the process. This difference is important not only from a technological point of view but also from a marketing point of view, since it directly affects the taste, body, and aroma of the final product. (Okaru & Lachenmeier, 2022)

There is not one definition that everyone agrees on at the international level. According to the *World Health Organization,* the rules for alcohol-free and low-alcohol products are very different in each country. Companies talk to customers and label their products in different ways because of this. (World Health Organization, 2023)

The non-alcoholic beer market is very different from other markets, both in terms of what is available and how people think about it. Institutional literature also emphasizes that the non-alcoholic and low-alcohol beverage sector should not be regarded exclusively in opposition to traditional alcoholic beverages. The *WHO* says that these drinks “are in between regular non-alcoholic drinks and alcoholic ones, they serve as a hybrid function that combines elements of symbolic continuity with beer and a total or partial reduction in alcohol content”. (World Health Organization, 2023)

This lack of clear definitions makes it hard for consumers to understand labels and product names. *Alcohol Change UK*, for example, says that the words "low alcohol," "alcohol-free," and "non-alcoholic" have all come from different contexts in the past and are now starting to mean the same thing, and this has caused some confusion. (Misell, 2024) The growing popularity of non-alcoholic beers in recent years has made it even more important to make their definitions clear. This is because the segment has been including products that meet not only functional needs (like driving or temporary abstinence) but also lifestyle and social related ones.

1.2.2 The processes to make non-alcoholic beer

Knowing how non-alcoholic beer is produced helps to understand the product category. The terminology for “alcohol-free” or “low-alcohol” beers refers to the different technical methods used to obtain the final product, which have a significant effect on both the ethanol content, the taste and smell of the beer.

The literature makes an important distinction between two main approaches to producing non-alcoholic beer: the first one stops alcohol from forming before fermentation, and the second one takes the alcohol out of beer that has already been fermented. This difference is important to better understand the different kinds of NAB products.

The biological approach focuses on the fermentation process, in order to produce less ethanol; the most popular way to achieve it is to change the mashing process in order to make the wort have less fermentable sugars, alternatively it Is possible to use controlled or interrupted fermentation conditions (arrested fermentation). Most of the time, beers made this way have very low alcohol content, usually less than 0.5% ABV. But they can also change the smell of the beer, often adding notes of wort or making the taste less complex.

The use of selected yeast strains limits the production of ethanol while generating relevant aromatic compounds. However, these approaches still present critical issues in terms of process standardization and replicability on an industrial scale. (Jackowski & Trusek, 2018)

An alternative approach consists of taking the alcohol out of the beer when it is already made. This method keeps some of the original smell by taking out the alcohol after fermentation. Two of the most common ways to do this are vacuum evaporation and membrane technologies, such as reverse osmosis and pervaporation.

Membrane technologies are appealing to producers because they can operate at lower temperatures, reducing the risk of losing the volatile compounds that give the beverage its taste and smell. The literature, however, suggests a compromise between the degree of alcohol elimination and the preservation of beer's sensory characteristics, often requiring an integration of biological and physical techniques to produce products of satisfactory quality. (Salanță et al., 2020)

Different ways of making the product help explain why there are so many different kinds of non-alcoholic beers on the market at the same time. These beers can have very different tastes, structures, and amounts of alcohol left over. Regulatory definitions and business names show how technology has changed over time; this means that the way the NAB segment is classified is based on the production process.

1.2.3 History of non-alcoholic beer over time

The NAB category is not a recent development; it is the result of a gradual evolution, reflecting technological, regulatory, and cultural changes in the way alcoholic beverages are produced, consumed, and perceived by consumers. During the second half of the 1900s, the first types of beer with little or no alcohol appeared. These were mostly made to meet functional and legal needs, not because there was a structured market for them.

In the beginning, the production of non-alcoholic beers was closely linked to specific contexts, such as legal restrictions, religious motivations, or the need to consume them in situations where alcohol was inappropriate, like driving or working. People often thought of the first alcohol-free products as backup options to regular beer because they were not perceived as good and as popular as other beverage offerings. (Misell, 2018) The adverse perception was mostly attributed to the production processes' technical problems. The methods used initially to cut down on or get rid of alcohol changed the smell and taste of the final product, which reinforced the idea that non-alcoholic beer is not authentic. Scientific literature substantiates that the initial technologies for dealcoholization were ineffective in preserving the sensory attributes intrinsic to beer, negatively impacting its body, aroma, and structure. (Schubert & Rettberg, 2025)

In the 2000s, when production technologies improved and consumers paid more attention to product quality, a first turning point occurred. In particular, using more advanced methods to remove alcohol and control fermentation has helped lessen the bad effects on the aromatic compounds in beer. Improvements in the dealcoholization processes have helped to close the quality gap between traditional and non-alcoholic beers, making the latter more acceptable to a wider audience.

As technology has evolved, so has the way people drink alcohol in social and cultural settings. The *World Health Organization* stated that in the last few decades, there has been a greater focus on the risks of drinking alcohol. This has led institutions and producers to seek out drinks with little or no alcohol. In this context, non-alcoholic beer was no longer just a niche product; it was regarded as a way to meet new needs for drinking. In the 2010s, drinks with little or no alcohol became more popular and accepted in the beverage markets. Part of the cause for this phenomenon was that large companies that had been making traditional beer for a long time started making them. (World Health Organization, 2023)

Non-alcoholic beer is no longer just a useful product; it is now a logical addition to an existing brand portfolio. As circumstances have changed, so have the language and the way products are marketed. Words like "alcohol-free" and "low alcohol" are starting to be associated with ideas of making choices and living in a specific manner, rather than giving up or limiting. (Alcohol Change UK, 2022)

In the last ten years, this process has helped make alcohol-free beer more common in everyday situations. NAB products are increasingly available in large retailers and in HORECA channels. This represents a substantial change in social norms around alcohol consumption, meaning that cutting or not drinking alcohol is no longer necessarily linked to social stigma. (World Health Organization, 2023)

In summary, the historical evolution of non-alcoholic beer can be seen as a journey from a marginal and functional product to a category that is increasingly integrated into the beer market. This path was made possible by the coming together of new technologies, new laws, and new ways of thinking about alcohol and consumers. This evolutionary process is the basis for understanding the current structure of the non-alcoholic beer market and the growth dynamics that will be discussed in the next paragraph.

1.2.4 The size, structure, and growth of the market for non-alcoholic beer

The size and structure of the non-alcoholic beer market should be interpreted with caution, as it is a new category where market data is often incomplete or indirect. The non-alcoholic beer market lacks established and consistent historical data on a global scale, making it necessary to rely on indicators related to consumption, availability, and purchasing behavior. An examination of the non-alcoholic beer market requires an overview of the ongoing trends that are influencing global alcohol consumption. Even though non-alcoholic beer has its own unique market, its growth is part of a larger trend of changing drinking habits that is affecting the whole alcoholic beverage market. According to data provided by *Our World in Data*, global alcohol consumption has changed over time.

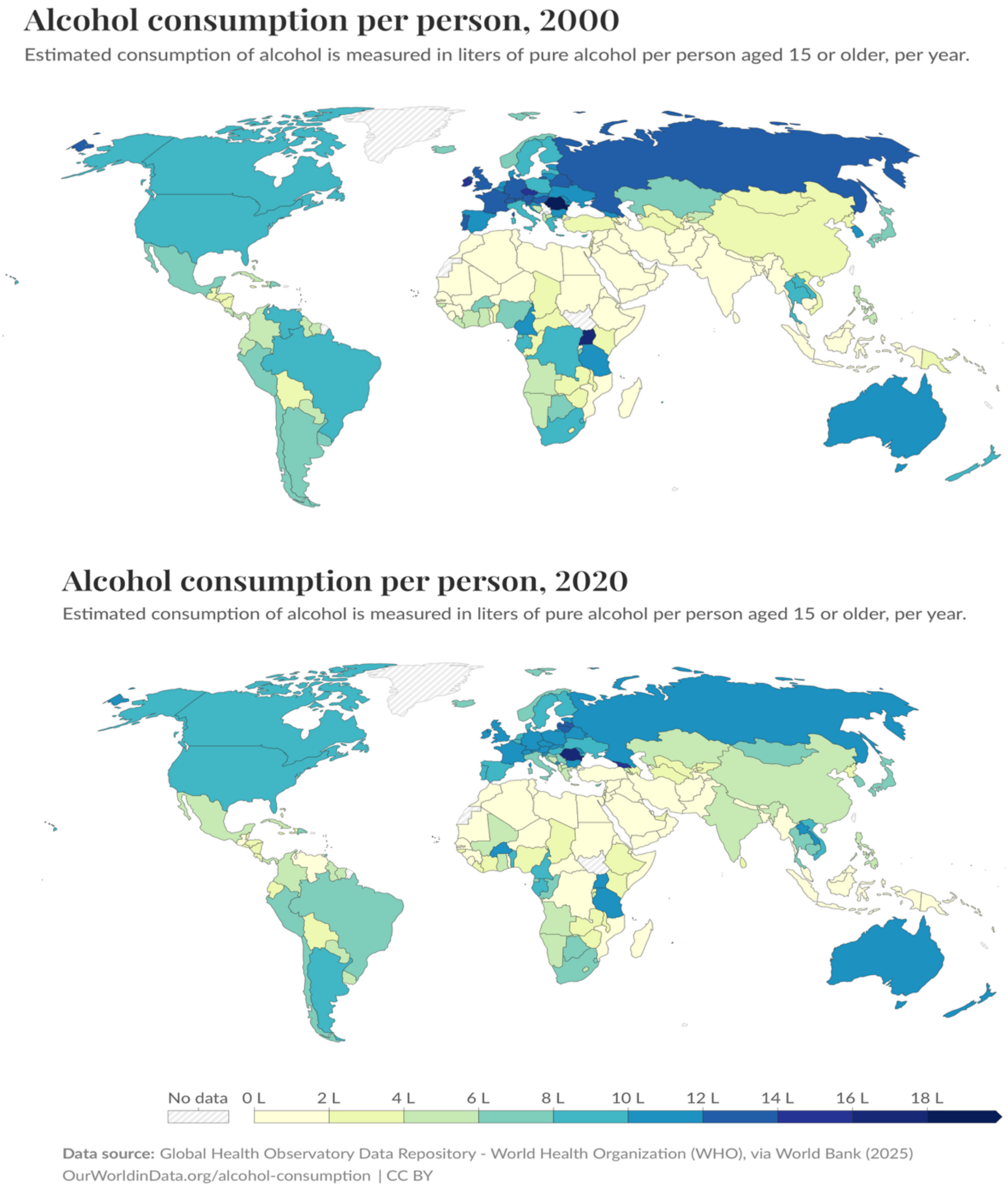

Fig. 1.1 - Alcohol consumption per person worldwide by Our World in Data

The "Alcohol consumption per person" graph shows that in many high-income countries, the amount of alcohol people drink has stayed the same or even gone down over the past decades. This means that now would be a great opportunity for low or no-alcohol drinks to become popular. (Ritchie & Roser, 2024)

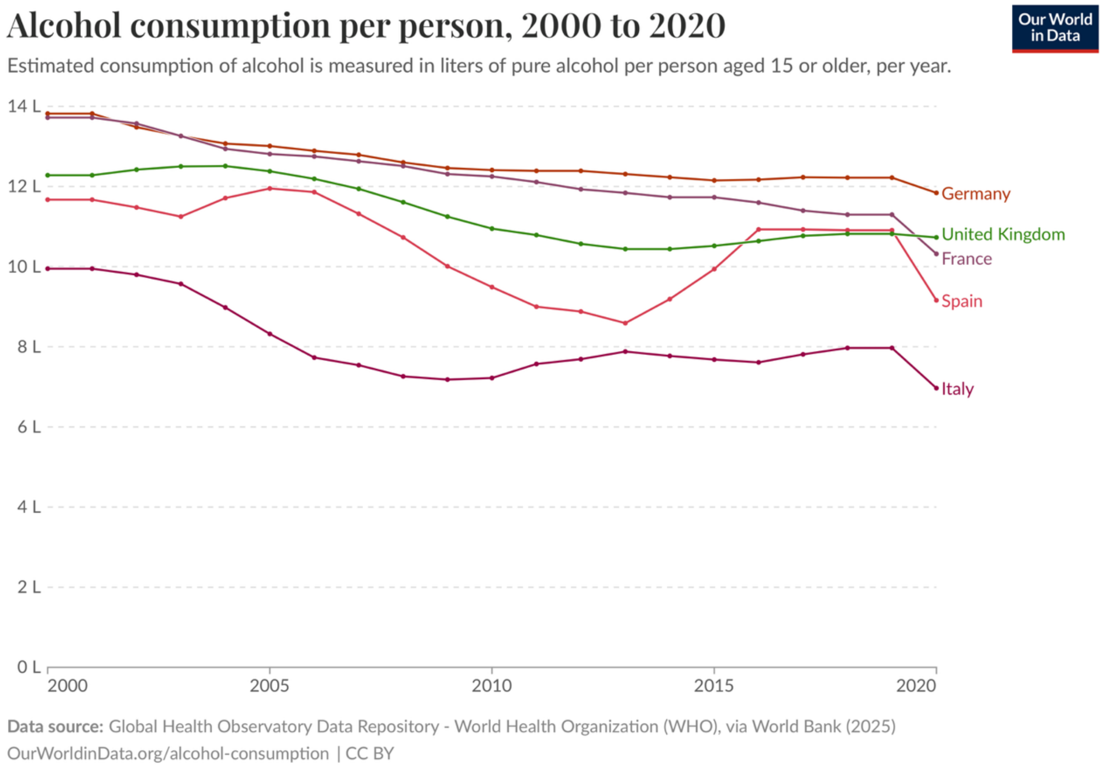

Fig. 1.2 - Alcohol consumption per person for countries by Our World in Data

In this case, non-alcoholic beer is perceived as a direct response to the evolving consumer preferences, rather than merely a byproduct of traditional beer. The Anderson et al. (2020) study clarifies some dynamics of the NAB market by showing how low- and no-alcohol beers affect the way individuals buy products. The study used real purchase data to show that the rise in non-alcoholic beer options is linked to more products being available on store shelves, and consumers slowly adopting domestic consumption. The NAB market has a very diverse availability depending on many factors. It's important to remember that NAB beers do not completely replace regular beers; instead, they tend to make them better, giving customers more options. (Anderson et al., 2020)

This information shows that the NAB segment's growth should not just be seen as taking away from other segments; it should also be seen as a way to grow the beer category. Another relevant aspect concerns the distribution of NAB products between on-trade (Horeca) channels and off-trade (large-scale retail distribution) channels. There is not a lot of specific information about non-alcoholic beers, but overall statistics on regular beers can help to understand the NAB segment better.

The Brewers of Europe report shows that most beer drinking in Italy happens in the off-trade channel, which makes up 62% of all drinking, while the on-trade channel only makes up 38%. (The Brewers of Europe, 2023)

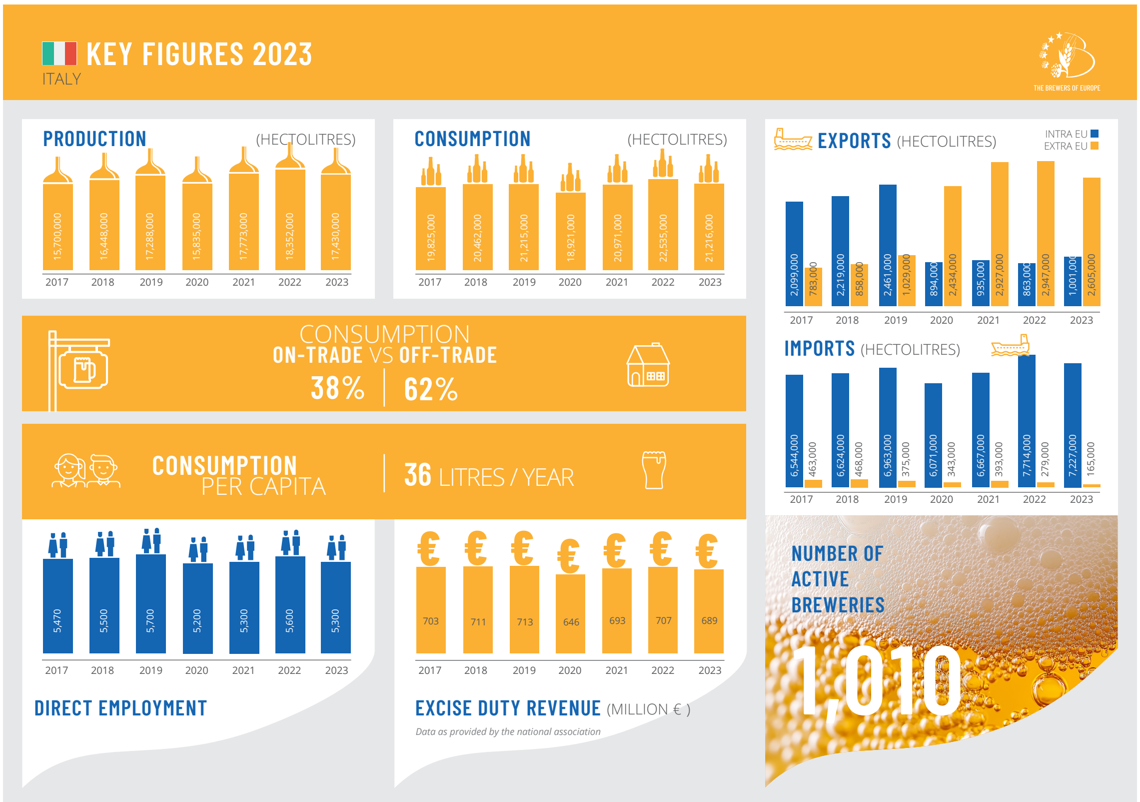

Fig. 1.3 – Report by Brewers of Europe

This distribution structure supports the non-alcoholic beer distribution, because large-scale distribution is a great way for consumers to try new options.

The traditional beer market is still clearly the biggest one when it comes to total sales. The zero and low-alcohol beverage segment is growing faster than the overall alcoholic beverage market, but it is starting from a low base. This indicates a positive trend in popularity for non-alcoholic beer, but overall data on alcohol consumption does not fully show how important it is. (World Health Organization, 2023)

Geographically, the NAB market seems to be growing the fastest in European countries. This is because traditional beer is very popular, people are becoming more health-conscious, and public policies are trying to cut down on alcohol use. Europe has historically high levels of alcohol consumption compared to the rest of the world. This makes it a great place to study how substitution and diversification work in the beverage supply. (Ritchie & Roser, 2024)

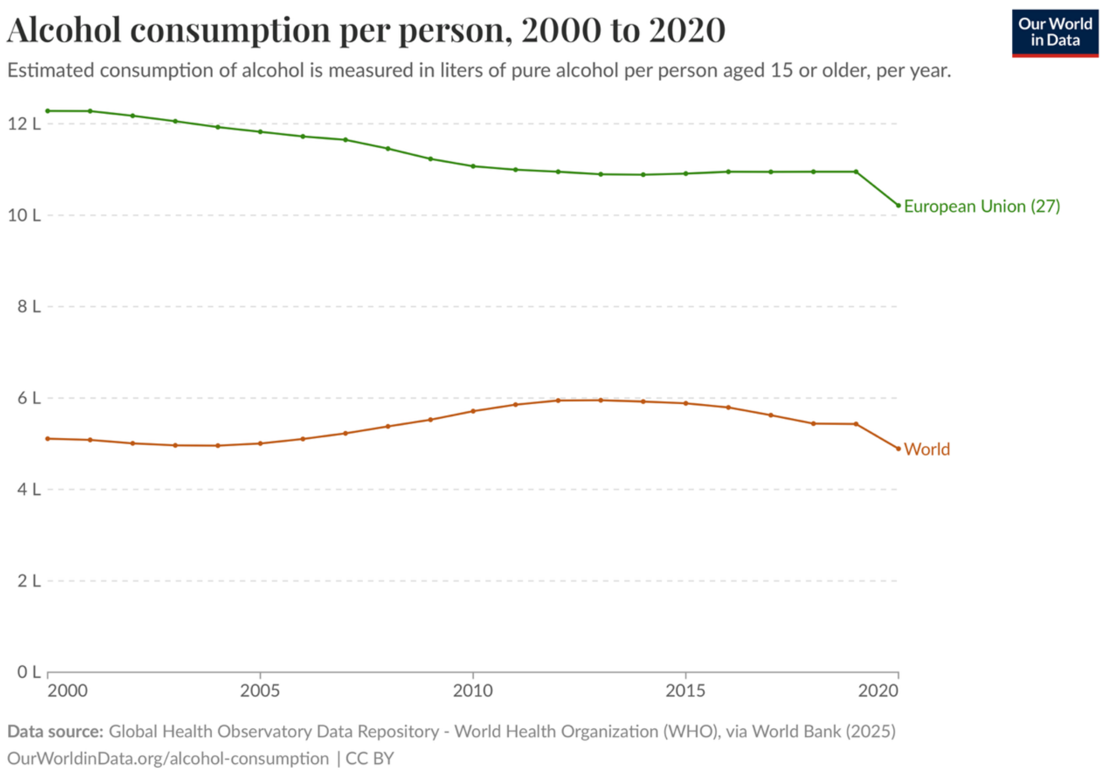

Fig. 1.4 – Comparison between alcohol consumption per person in Europe and worldwide by Our World in Data

Overall, nowadays the non-alcoholic beer market remains small in terms of total volume, but it is growing steadily and becoming more integrated into the beer market. A mix of macroeconomic factors, changes in how people shop, and the growth of the number of products that manufacturers offer is helping to shape a development path that, while not always the same, seems to be structurally different from that of traditional beer. These traits make the NAB market a very interesting place to study consumers and their point of view.

1.2.5 Health, well-being, and younger generations: a change in the relationship with alcohol

The growth of the non-alcoholic beer segment is closely linked to a broader change in the relationship between alcohol consumption, health, and lifestyles, which is particularly evident among younger generations. In recent years, the consumption of alcoholic beverages has been progressively renegotiated within a cultural framework in which physical and mental well-being play an increasingly central role in individual decisions.
A significant part of the contemporary debate on alcohol consumption revolves around the growing awareness of the risks associated with even moderate levels of consumption. Zero and low-alcohol beverages are increasingly perceived as a means of reducing exposure to alcohol-related risks without completely giving up the social practices associated with drinking. (World Health Organization, 2023)

In this sense, non-alcoholic beer fits in as a solution compatible with a more careful approach to consumption. One of the elements that emerges most clearly concerns the increasingly selective and situational nature of alcohol consumption among the younger generations.
The study by Roos et al. (2021) examines a sample of young adults between the ages of 18 and 25, and it shows that consumption is often associated with reward-type motivations, linked to the search for immediate positive effects such as socializing and mood enhancement, rather than relief-type motivations, connected to stress or discomfort management. (Roos et al., 2021)

This finding is especially important because it shows that for many young people, drinking is no longer a daily habit or a part of their routine but rather a choice they make for certain occasions. From this perspective, consumption tends to be evaluated in light of its consequences, not only in terms of health, but also in terms of performance, lucidity, and overall well-being. The growing focus on healthy lifestyles, fitness, and personal productivity contributes to making alcohol a substance to be ‘measured out’ rather than consumed systematically. The literature additionally points out that cultural representations of alcohol remain significant in influencing consumer expectations. Positive portrayals of alcohol consumption in the media, even when not directly linked to a particular brand, appear to reinforce alcohol expectancies: the belief that drinking alcohol fosters sociability, relaxation, and well-being. (Redondo et al., 2018)

However the same study states that these expectations are not necessarily linked to drinking excessively; they are more about the symbolic and social aspects of drinking. In this scenario, there is a growing tension between wanting to take part in social rituals that are usually associated with alcohol, and being more aware of the detrimental effect it has on your health. The success of non-alcoholic alternatives hinges on this very tension.

NAB beers allow people to keep the fun, taste, and social belonging aspects of drinking alcohol while lowering or getting rid of the "costs" that are associated with drinking. By looking at this evidence, it can be noticed that the non-alcoholic beer market is not just a replacement for regular alcohol. It is also a clear response to a deeper change in what younger people want and why they want it. Instead of giving up consumption, they seem to be looking for solutions that give them more freedom and control, in line with a view of well-being that includes physical health, mental balance, and social life.

1.2.6 The non-alcoholic beer consumer

To understand the non-alcoholic beer consumer, it is imperative to move beyond their classification as abstinent or exclusively health-focused. The current literature suggests that the consumption of non-alcoholic beers is part of a more complex set of behaviors and motivations, and that cutting back on alcohol doesn not mean giving up the social and symbolic aspects of drinking.

An analysis of household purchasing trends regarding the introduction of low-alcohol and non-alcoholic beers indicates that consumers of non-alcoholic beer do not entirely replace traditional beer; instead, they typically incorporate NAB products into their overall

consumption habits. (Anderson et al., 2020) These findings indicate that most people who drink non-alcoholic beer also drink alcoholic beer, not instead of it.
It is evident that situational factors have a big impact on the choice to drink non-alcoholic beer, at least when it comes to the motivation for it. People drink little or no alcohol depending on the situation, like when they are driving, working, or at a social event where drinking is not allowed. NAB customers do not stop drinking completely in this case; instead, they decide when and how much to drink. (World Health Organization, 2023)

Many consumers choose products in this category so they can drink less and still have fun with peers. This dynamic helps to understand why so many different kinds of people enjoy drinking non-alcoholic beer. Some consumers do not drink for health or personal reasons, while others switch between drinks with and without alcohol depending on the situation. There is also more demand for these products because it is getting easier to find high-quality non-alcoholic beers. People who did not like them before will probably find them interesting now. (Alcohol Change UK, 2022)

From a sociodemographic standpoint, although the literature does not offer a definitive profile, specific relevant trends are evident. The preceding paragraphs indicate that young adults tend to exhibit flexible and mindful consumption patterns. Non-alcoholic beer helps consumers to preserve the social side of drinking, while cutting down on alcohol use. This kind of behavior is part of a larger effort to change the rules about alcohol use.
In addition to why people drink, it is important to think about what stops them from drinking non-alcoholic beers. Some people still think the product tastes bad and is not real. It is not always clear what "alcohol-free" and "low alcohol" mean because they are used in different ways. The market for non-alcoholic beer is getting bigger, but a lot of people still buy it now and then because of these problems. (Misell, 2024; World Health Organization, 2023)

Consumers of non-alcoholic beer are not giving up drinking culture; they just have more choices so they can find a balance between how much they drink, the taste of it, and their social life.

1.2.7 A summary of the sub-chapter

This sub-chapter has analyzed the non-alcoholic beer market as a new part of the beverage industry; it examined what makes it different and why it is growing. The analysis has shown that non-alcoholic beer is not just a different type of regular beer; instead, it is becoming a more distinct category because of changes in technology, culture, and how people behave as consumers. The definition and analysis of production processes brought out the differences in the NAB segment, making it clear that there are different labels and alcohol thresholds, as well as differences in the quality of the products on the market. Thanks to new technologies, non-alcoholic beers' taste has improved, which has made them more popular among customers.

Non-alcoholic beer used to be hard to find, but now it is very well-distributed, because big beer companies are getting involved, and the growth rates are higher than the average for the beer market, even though the amounts are still small compared to the whole market.

There are many reasons why non-alcoholic beers are becoming more popular. First, people are becoming more aware of their health and well-being, especially young adults; in addition to that, they are seeing the bigger picture of drinking alcohol. NAB consumers appear to be part of a heterogeneous group that utilizes these products to modulate their alcohol consumption according to the circumstances, balancing social interaction, individual preference, and regulations. This subchapter has given a theoretical and market-based explanation of the NAB market and history.

1.3 Cluster Analysis

1.3.1 Basics of cluster analysis and how it has changed over time

Cluster analysis is an unsupervised learning approach that aims to uncover latent structure in unlabeled datasets. More specifically, clustering partitions observations into groups (clusters) such that within-cluster similarity is higher than between-cluster similarity. This method makes it clear that cluster analysis is different from supervised classification methods, where the labels of the observations are already known.

A brief distinction between supervised and unsupervised learning helps to better understand what cluster analysis is about. With supervised learning, data with a response variable or labels that you already know about is used, and it helps the model learn. The main goal of this method is to find a link between a group of explanatory variables and a dependent variable so that new observations can be predicted or classified. In unsupervised learning, on the other hand, there is no prior knowledge about the labels or categories that the observations belong to. Here, the analysis is exploratory, which means that it looks for patterns, structures, or regularities in the data without having a specific output in mind. Cluster analysis is a good example of this second type of analysis because its goal is not to predict a target variable but to find groups that are similar with eachother. (Jain, 2010; James et al., 2017)

Conceptually, clustering is one of the most basic ways to understand data, similar to the ways that natural sciences use classification systems. Clustering was created as a way to explore data, not to test already set hypotheses. Its goal was to find patterns and regularities in the data. (Jain, 2010)

The first uses of cluster analysis took place in the 1950s, mostly in anthropology, biology, and psychology. Later, as multivariate statistics and computers got better, clustering became a formal set of algorithmic methods. This led to a lot of writing on the subject and many different ways to do it. Some algorithms, like K-means, that were created in the early stages are still widely used today. This shows how complex the clustering problem is and how there are not solutions that work for universally. (Everitt et al., 2011)

1.3.2 Data representation, similarity, and algorithm families

At the heart of any clustering technique is a representation of observations in feature space. In this space, the similarity between observations is quantified using a similarity or distance measure. As noted in the literature, there is no universally valid measure of similarity, since the very notion of "similarity" depends on the nature of the data and the objectives of the analysis. (Jain et al., 1999)

For numerical data, one of the most commonly used measures is Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{p} \quad (x_{im} - x_{jm})^2}$$

where x_i and x_j represent two observations in the space of p variables. This measure is the basis of numerous partitioning algorithms, including K-means.

Clustering methods are commonly classified as:

- Hierarchical methods, which construct a tree structure of the data;
- Partitioning methods, which directly divide the observations into a predetermined number of clusters.

Alongside these approaches, methods based on density or probabilistic models have been developed over time. However, the choice of clustering algorithm depends on the trade-off between simplicity, scalability, and interpretability of results. (Everitt et al., 2011)

1.3.3 Data representation and similarity measures

Any clustering procedure relies on a representation of the observations in a feature space, where each object is described by a set of variables. Within this space, clustering algorithms group observations according to a measure of similarity or dissimilarity. As highlighted in the literature, the choice of this measure is a crucial step, since it directly influences the structure and shape of the resulting clusters.

For numerical data, one of the most commonly used dissimilarity measures is the Euclidean distance, defined as:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{p} \quad (x_{im} - x_{jm})^2}$$

where x_i and x_j are two observations described by p features. This distance measure underlies many partitional clustering algorithms and implicitly assumes that all variables are measured on comparable scales. Therefore, preprocessing steps like normalization or standardization are often needed to keep variables with larger ranges from taking over the distance calculation. (Jain, 2010)

In a broader sense, what constitutes a meaningful distance between observations, depends on the nature of the data and the objectives of the analysis. Consequently, clustering does not depend on a universal definition of similarity; instead, it relies on context-specific selections that require meticulous justification. (Jain et al., 1999)

1.3.4 Main families of clustering algorithms

The clustering literature traditionally distinguishes between different families of algorithms based on how clusters are constructed. A first major distinction is between hierarchical and partitional methods.

Hierarchical clustering algorithms build a nested structure of clusters, either by progressively merging individual observations into larger groups (agglomerative approach) or by recursively splitting a large cluster into smaller ones (divisive approach). A dendrogram is typically employed to show the result, which shows how clusters are hierarchically related to each other.

Conversely, partitional clustering methods split the data into a predefined number of clusters without adding a hierarchical structure, trying to find the most optimal value for a specific goal and put each observation into a single cluster. The K-means algorithm is the most popular partitional approach, because it is straightforward to understand and operates quickly. (Jain, 2010)

Beyond hierarchical and partitional methods, the literature also documents the development of clustering techniques based on density or probabilistic models. However, despite the large number of algorithms proposed over time, no single method consistently outperforms others across all applications. The choice of a clustering algorithm therefore depends on trade-offs between computational scalability, robustness to noise, and interpretability of results. (Everitt et al., 2011)

1.3.5 The K-means method

K-means is a partitional clustering algorithm designed to divide a dataset into a predefined number K of non-overlapping clusters by minimizing within-cluster variability. Given a set of observations

$$X = \{x_1, x_2, \dots, x_n\},$$

the algorithm seeks a partition

$$C = \{C_1, C_2, \dots, C_K\}$$

that minimizes the following objective function:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|^2,$$

where μ_k denotes the centroid of cluster C_k, computed as the mean of the observations assigned to that cluster. (Jain, 2010)

This formulation highlights that K-means aims to produce clusters that are internally compact, as measured by the squared Euclidean distance between observations and their corresponding centroids. From a computational perspective, there is an optimization problem with the algorithm, meaning that finding the globally optimal clustering solution becomes computationally infeasible as the size of the dataset increases. As a consequence, K-means relies on heuristic iterative procedures that converge to locally optimal solutions rather than guaranteeing a global optimum. (Jain et al., 1999)

The algorithm operates through two main iterative steps. First, each observation is assigned to the cluster whose centroid is closest according to the chosen distance metric, typically the Euclidean distance:

$$C_k = \arg \min_{j \in \{1,\dots,K\}} \| x_i - \mu_j \|^2.$$

Second, the centroids are updated by recalculating the mean of the observations within each cluster:

$$\mu_k = \frac{1}{|\, C_k \,|} \sum_{x_i \in C_k} x_i.$$

These steps are repeated until convergence, which is generally defined as the point at which cluster assignments no longer change or improvements in the objective function become negligible. (Jain, 2010)

A key implication of this iterative procedure is the sensitivity of K-means to the initial placement of centroids. Different initializations may lead to different clustering solutions, even when the same value of K is specified. To address this issue, several initialization strategies have been proposed in the literature, most notably the *k-means++* method, which aims to improve convergence behavior by selecting initial centroids that are well separated in the data space. (Arthur & Vassilvitskii, 2007; Jain et al., 1999)

Despite these limitations, K-means remains widely used due to its conceptual simplicity, ease of implementation, and favorable computational efficiency, particularly in large datasets. However, its effectiveness relies on assumptions such as approximately spherical cluster shapes and comparable variance across dimensions, which must be taken into account when interpreting the results.

1.3.6 Cluster analysis applications in beer market

Methodological literature suggests that clustering techniques are especially effective when consumer groups are not predefined and need to be directly derived from the data. The beer market serves as a pertinent application context, given that beer consumption is affected by various consumer-centric, product-centric, and situational factors.

The Calvo-Porral et al. (2018) research illustrates how cluster analysis can be used in the beer market by creating a cluster-based segmentation of beer drinkers to find out if beer demand is the same across all groups, or if there are different types of beer consumers.

The authors used survey data from 592 beer drinkers and a multivariate method that combines factor analysis and clustering techniques to find important groups based on hidden dimensions of consumption. (Calvo-Porral et al., 2018)

The first step in the study is to use factor analysis to group a lot of observed variables into a small number of hidden factors. Brand loyalty, perceived quality, familiarity, willingness to pay a higher price, purchase intention, and value for money are some of them. The clustering process uses these scores for the factors as input variables. The authors firstly perform a hierarchical cluster analysis using Ward's method to find out how many clusters there are. Then, to improve the final solution, a K-means clustering using Euclidean distance is executed.

The analysis produces a five-cluster solution, classifying distinct beer consumer segments identified as "beer lovers," "circumspect seniors," "social drinkers," "homelike women," and "beer to fuddle consumers". These groups have very different drinking habits, such as how much beer they drink, how much they care about price, how good they think the beer is, how loyal they are, and when they drink. This evidence backs up the idea that people who drink beer should not be thought of as one big group, but as a bunch of smaller groups, each with its own reasons and traits for drinking beer.

The following graph shows the average scores for the main factors that affect consumption in each of the five groups. This is a visual way to show the groups that were found. The figure shows that not all beer drinkers are the same, which makes it easy to compare the profiles of different groups. This shows that each group has different levels of loyalty, perceived quality, willingness to buy, and willingness to pay more. (Calvo-Porral et al., 2018)

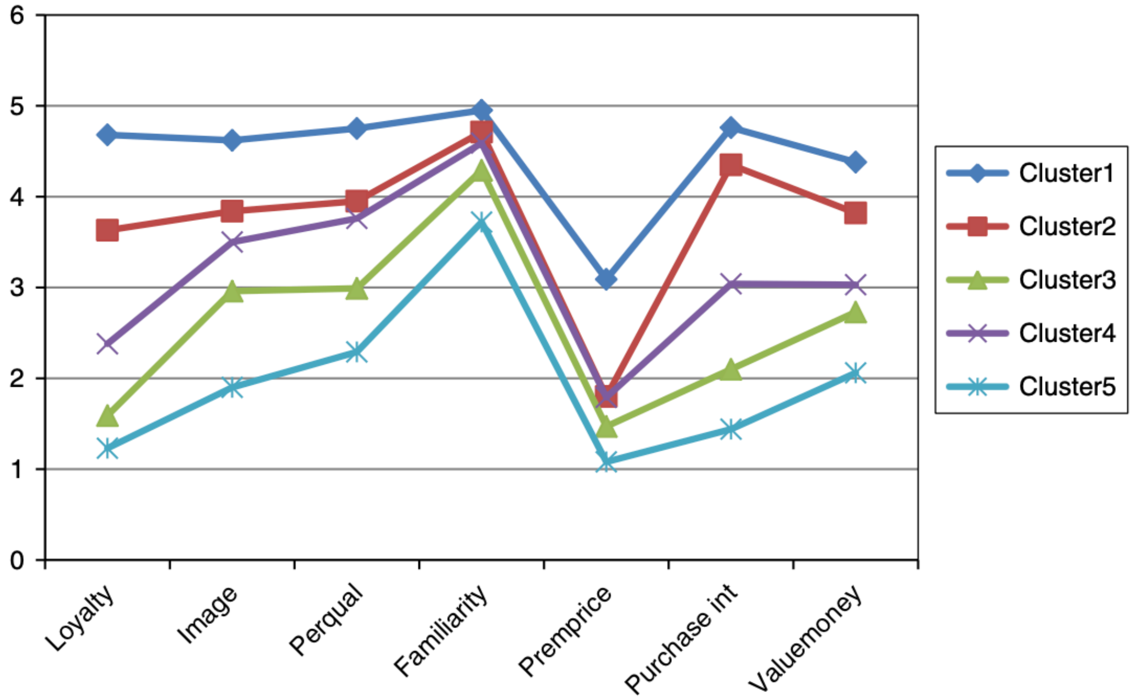

Fig.1.5 - adapted from (Calvo-Porral et al., 2018)

This application successfully finds hidden types of beer drinkers using K-means clustering. The study backs up the use of cluster analysis as a method for market segmentation in the beer industry, by showing that it is important to combine statistical criteria with interpretability when defining and describing consumer segments. (Jain, 2010)

1.5 Sentiment Analysis

1.5.1 Fundamentals of sentiment analysis

Sentiment analysis, also called opinion mining, refers to a group of methods used to find and study subjective information in text data. The main goal is to figure out what kind of evaluative orientation a text has, which is usually done by figuring out if the sentiment is positive, negative, or neutral (Pang & Lee, 2008). Text is not organized like numbers, which makes it harder to analyze but also very useful for learning about how people think and feel about subjects.

The growing significance of sentiment analysis in marketing and consumer research is linked to the strong increase of user-generated content on digital platforms. The researcher did not ask for online reviews, comments, and discussions on social media, so they are less likely to be biased. As Liu (2008) pointed out, sentiment analysis makes it possible to turn these qualitative expressions into structured information in a systematic way.

There are different levels at which sentiment analysis can be done, such as the document level, the sentence level, or the aspect level. Even though these levels are different, the main job is still the same: figuring out the polarity and, in some cases, the strength of the sentiment expressed. It is crucial to acknowledge that sentiment analysis does not aim to comprehensively explain the semantic meaning of language. Instead, it concentrates on approaching the evaluative aspect in written expressions (Liu, 2008). This characteristic makes it particularly suitable for research that examines and articulates how individuals discuss topics.

1.5.2 Main methodological approaches

The literature on sentiment analysis discusses various methodological approaches, each characterized by distinct assumptions, data requirements, and analytical objectives. One way to group methods is to make a distinction between machine learning-based methods, and lexicon-based methods. (Pang & Lee, 2008)

Lexicon-based methods use existing dictionaries that connect words and phrases to polarity labels or sentiment scores. You can figure out how a text makes you feel by adding up the meanings of its words and phrases. Liu (2008) says that one of the best aspects about these methods is that they do not need labeled training data; this makes them great for exploratory analyses. (Liu, 2008)

In contrast, methods based on machine learning see sentiment analysis as a supervised classification problem. In this case, models are trained on labeled datasets where texts are clearly linked to sentiment categories. The patterns learned are then used to make predictions about new observations (Pang & Lee, 2008). These methods might be better at making predictions, but they only work well if there is a substantial quantity of good labeled data, and if the training and application domains are the same. Hybrid methods combine lexical resources with data-driven methods in addition to these two main groups. The choice of methodology depends on the research goals, the nature of the textual data, and the desired balance between interpretability and analytical accuracy.

1.5.3 Rule-based and lexicon-based sentiment analysis

Rule-based and lexicon-based sentiment analysis are two of the oldest and most well-known methods for opinion mining. This method is based on the idea that you can tell how someone feels by the words they use that have a positive or negative meaning. There is a set polarity or valence for each word in sentiment lexicons. To find the overall sentiment of a text, you add up the individual values. (Liu, 2008)

Lexicon-based methods often use language rules to deal with modifiers that change how strong or weak a feeling is. Negations, intensifiers, and mitigating expressions can have a big effect on the meaning of a sentence. To partially capture these effects, rule-based changes are used (Pang & Lee, 2008). These kinds of approaches may achieve more than just frequency-based representations because they use both lexical knowledge and heuristic rules.

The fact that rule-based sentiment analysis is straightforward to understand is one of its best features. Sentiment scores are based on clear lexical resources, so you can determine which words or phrases in the text caused certain classification results. Liu (2008) says that this openness makes lexicon-based methods especially useful in both exploratory and applied research settings, where figuring out what triggers sentiment is just as important as measuring its spread.

The literature also points out some of its own flaws. Lexicon-based methods have difficulties catching sarcasm, irony, and specialized language that changes depending on the situation. Static dictionaries do not always express that the meaning of a word can change depending on the situation (Pang & Lee, 2008). So, even if rule-based sentiment analysis is a good starting point for analyzing great quantity of text, the results need to be carefully thought about, especially when it comes to complicated consumer conversations.

1.5.4 Performance and evaluation considerations

There are plenty of practical and theoretical issues that render it hard to know how well sentiment analysis systems work. Sentiment analysis is different from many other types of classification tasks because it has to deal with subjective labels, language that can be understood in different ways, and meanings that change depending on the situation. Performance evaluations are unable to utilize simple measures of accuracy because how language is understood can affect how well an element is classified. (Pang & Lee, 2008)

Sentiment classification is particularly susceptible to class imbalance due to the prevalence of neutral or mildly positive expressions in real-world text data compared to negative ones. In these situations, a classifier that only predicts the majority class may be very accurate but not very useful for analysis. This is why it is so important to look at other ways to measure how well a classification works that show different parts of it. (Liu, 2008)

Two common measures are precision, which shows the percentage of correctly identified sentiment instances out of all instances assigned to a certain class, and recall, which shows the percentage of relevant instances that are correctly identified. These metrics show different parts of performance, and there are often trade-offs between them. The F1-score is often used as a single summary measure when the class distributions are not even. It is the harmonic mean of precision and recall. (Pang & Lee, 2008)

In addition to numbers, it can be helpful to look at sentiment analysis systems in a more qualitative way. You can find systematic errors by checking some of the outputs by hand. Something might be misclassified if sarcasm, negation scope, or language that is only used in a certain field is present. Liu (2008) says that these mistakes do not always show up in overall performance metrics, but they can make it hard to understand results in applied research settings. This means that performance evaluation in sentiment analysis should be thought of as a process that combines numbers with an understanding of the situation.

1.5.5 Strengths and limitations of sentiment analysis

Sentiment analysis has a number of benefits that make it especially useful for marketing and consumer research. One of its main strengths is that it can analyze a lot of text data that was created on its own. Sentiment analysis works with content that the researcher does not ask for, which lowers the risk of social desirability bias and gives researchers access to more natural ways that people express their opinions. (Liu, 2008)

It can also grow, which is another important benefit. Automated methods for sentiment analysis can look at thousands or even millions of pieces of text. This makes it easier to find general trends in how people feel about aspects that would be hard to find by hand. Sentiment analysis also works well with quantitative methods because it lets you add feelings and perceptions to decision-making frameworks that are based on numbers.
Even though these are good aspects, the literature always talks about important issues. One of the biggest problems is that language depends on its context. Words can have different meanings depending on they are used. Automated methods struggle to detect sarcasm and irony. Pang and Lee (2008) contend that human annotators may possess divergent perspectives on sentiment labels, underscoring the inherent subjectivity of the task.

Another important aspect to think about is how specific a domain is. Words can mean different concepts in different areas, which can change how they are judged. This means that sentiment lexicons and classification models made for one area may not work well in another one. This means that you should be careful when looking at sentiment analysis results, especially in real life, where small differences in how people talk can make a big difference. (Pang & Lee, 2008)

For these reasons, sentiment analysis should not be regarded as a precise method for measuring something; instead, it should be thought of as a way to analyze written data. With the right methodological knowledge, it can tell a lot about how people perceive various issues. But its flaws show how important it is to be careful when interpreting the results, and to use other methods to analyze them as well.

1.5.6 Applications of sentiment analysis in the beverage industry

An example of the application of sentiment analysis to consumer perceptions in the beverage sector is provided by the study of Mostafa (2013), which investigates brand-related sentiment expressed on social media platforms. The author analyzes user-generated content collected from Twitter, with the objective of assessing consumers' emotional orientation toward well-known brands through automated text mining techniques. (Mostafa, 2013)

The study is based on a dataset of 3,516 tweets referring to sixteen global brands and applies a lexicon-based sentiment analysis approach to classify textual content according to sentiment polarity. Tweets are assigned sentiment scores by matching words against a predefined sentiment lexicon, allowing the author to compute aggregate indicators of positive, negative, and neutral sentiment.

This methodology enables the transformation of unstructured textual data into quantitative measures that can be systematically compared across brands. The results show that consumer sentiment expressed on social media is not evenly distributed across brands, with some exhibiting predominantly positive sentiment and others displaying more polarized or negative evaluations.

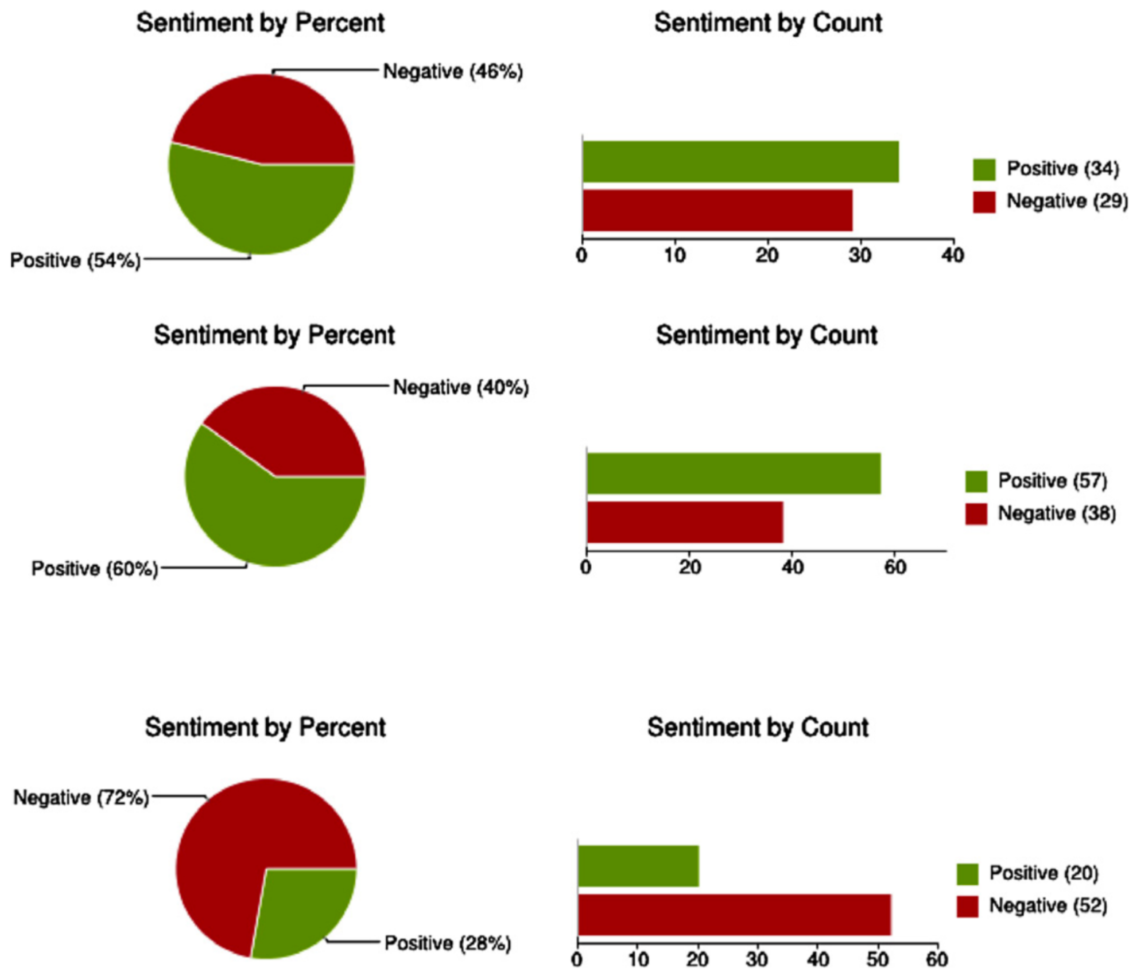

Fig.1.6 - adapted from (Mostafa, 2013)
"Sentiment analysis for a random tweets sample-after eliminating neutral tweets-for Lufthansa (top), DHL (middle) and T-Mobile (bottom) brands"

In particular, the distribution of sentiment scores highlights that a substantial share of online consumer discourse is characterized by neutral or mildly affective language, while extreme positive or negative expressions occur less frequently. These findings underline the importance of considering the full distribution of sentiment scores rather than relying solely on average indicators.

Beyond polarity classification, the study also illustrates how sentiment analysis can be combined with exploratory text mining techniques, in order to uncover patterns in consumer discourse. The author employs word frequency analysis to identify significant themes pertaining to consumer perceptions of a brand, including price, quality, and service. This integrative approach demonstrates that sentiment analysis can provide both quantitative insights into public sentiment and qualitative data regarding the underlying reasons for such sentiments. (Mostafa, 2013)

Overall, the study shows how useful sentiment analysis can be for figuring out how people feel about brands and products in the beverage and fast-moving consumer goods industries, where brand image and consumer attitudes are very important. Researchers and professionals use sentiment analysis to keep track of how people's feelings about a product or service change over time, using data from social media. It also helps them obtain more information from random comments made online, that would be hard to get from regular market research methods.

2. Methodology

2.1 Research framework and research questions

The market for non-alcoholic beer has become a separate segment within the broader beer market. As mentioned in the literature review, people who drink non-alcoholic beer do so for different reasons; they often choose these alternatives when there are ethical or health "barriers", for instance, when working, driving, or playing sports are involved. (Roos et al., 2021)

This situational aspect of consumption leads to significant variability in consumer perceptions and evaluations of non-alcoholic beer. Individuals may have different expectations and levels of acceptance for the product. They often have to choose between functional benefits, like not drinking alcohol, and sensory or symbolic aspects that people usually think of when they think of drinking beer. This is why looking at only overall consumption data may not be enough to understand how customers act. (Roos et al., 2021)

In this case, user-generated content on the web could represent an efficient way to learn how people express their feelings and perceptions. Sentiment analysis and text-based techniques have been widely utilized in marketing research to evaluate individual sentiments and to derive insights from large quantities of unstructured text data. Moreover, exploratory methodologies, including clustering, have been utilized to reveal latent patterns in human behavior and cognition, especially in scenarios characterized by varied motivations and innovative consumption practices.

Behavioral data needs to be added to perceptual evidence to find out how important new consumption trends are in the market. Territorial and market-level analyses allow to investigate the transformation of consumer perceptions into distinct consumption and value profiles across geographic regions, thereby enhancing the understanding of market structure and opportunities. (Calvo-Porral et al., 2018; Mostafa, 2013)

This study utilizes a comprehensive empirical methodology, analyzing both online consumer data and market data from diverse contexts, based on this framework. The objective is to examine both consumer perceptions and market distinctions to obtain insights that facilitate informed decision-making in the non-alcoholic beer sector.

The study is based on the following research questions:

- RQ1: How do consumers perceive non-alcoholic beer, and which dimensions of value emerge from online consumer discourse?
- RQ2: Are there distinct consumer narratives surrounding non-alcoholic beer that imply different value propositions and usage contexts?
- RQ3 Do territorial markets exhibit heterogeneous consumption and value patterns for non-alcoholic beer, indicating differentiated market opportunities?
- RQ4: How can consumer narratives and territorial market profiles be jointly interpreted to inform strategic decisions in positioning, pricing, and go-to-market strategies for non-alcoholic beer?

Together, these research questions frame the empirical analysis as a unified investigation of consumer meaning and market structure, setting the basis for an integrated interpretation of perceptual and behavioral evidence.

2.2 Research design and methodological approach

The empirical analysis uses an exploratory and data-driven research design to look at the non-alcoholic beer market from both a perceptual and behavioral point of view. Because the category is both hybrid and situational, the study looks for patterns and structures that come directly from the data.

The analysis is based on two parts that work together to answer the research questions. The first one analyzes how people think about non-alcoholic beer by looking at online content generated by users. The second one looks at market heterogeneity through data analysis on consumption and value in different areas. This two-part approach captures how people talk about the product and how these opinions affect real-world market results.

The first step in the analysis is to frame what people say about non-alcoholic beer online to find out their thoughts, opinions, and what they value. This step gives an idea of what consumers think the product means and how they use it. After that, a territorial clustering analysis is done to find different market profiles based on indicators of consumption and value. The study's goal is to find out if perceptual patterns are the same as or different from territorial market structures by comparing and combining the results of these two analyses.

The study uses unsupervised learning techniques, which are suitable for situations where there are heterogeneous motivations and not many assumptions made ahead of time. Sentiment analysis and thematic clustering are used to organize unstructured text data, while clustering techniques are used on quantitative market data to find areas that exhibit similarities. It is important to note that these methods are not used to make predictions but rather to help with strategic analysis and understanding.
Overall, this research design reads the non-alcoholic beer market in a way that makes sense and fits together, seeing consumer behavior and territorial performance as two sides of the same coin.

2.3 Software and analytical tools

The empirical analysis utilized a suite of analytical instruments for applied data analysis and marketing research.
Python was the main programming language used in the study due to its flexibility and libraries, which are suitable for working with, analyzing, and modeling data. It lets analysts compute all the steps of the analytical process in one place, for instance, actions like cleaning, changing, and modeling data. This study utilized Python for data preprocessing, pattern analysis, and both qualitative and quantitative evaluations. (Booth, 2019)

For text analysis, the same Python environment was used to group online comments and figure out how people felt about them. The literature review indicates that sentiment analysis was employed to verify public opinion regarding user-generated content in its entirety. After that, clustering methods were used to find common themes and stories in what consumers were saying. (Mostafa, 2013)

The scripts were all written in Visual Studio Code (VS Code), which is an integrated development environment (IDE). There are special add-ons for Python development in

VS Code that let you run, debug, and organize your code all in one place. This made it easier to use scripts and analytical workflows when doing research. (Microsoft, n.d.)
Circana's Unify+ platform offered market and territorial data. Circana provides syndicated retail data that are widely used in market analysis and offer detailed information on sales, volume, value, and distribution across geographic areas. These data were used to construct territorial indicators and to support the identification of different market profiles relevant for the analysis. (Circana., n.d.)

Apify, a web scraping and data extraction platform that automates the acquisition of structured data from online sources, was used to collect the online comments employed in the text-based analysis. Apify has pre-made scraping tools called actors that can be used to extract various types of content from websites without having to write your own scraping scripts from scratch. This study employed Apify's actor named "YouTube Comments Scraper" to extract user comments from YouTube videos related to non-alcoholic beer. The platform allows collecting comments and related metadata in a structured way. This method made it possible to obtain a complete dataset of user-generated content that could be used for text analysis. The dataset could be reproduced or expanded in future analyses using the same extraction parameters. (Apify, n.d.)
The combination of Python-based tools, a flexible development environment, and organized market data for a full analysis allowed the study to reach its goals.

2.4 Datasets and data preprocessing

2.4.1 Territorial market dataset

The territorial cluster analysis uses a structured dataset that combines market performance indicators for non-alcoholic beer with certain socio-demographic variables at the geographic level. The dataset was made by combining retail data from Circana (full year 2024) with official demographic statistics from ISTAT. The goal was to get a picture of both how people use the data and how local markets are structured.

Circana's standard territorial aggregation defines the geographic area as the unit of analysis. Each row in the dataset represents a different Italian province with data about the NAB market available, for a total of 88 rows. The columns show market performance indicators and contextual variables that are used to describe how areas differ from each other. (Circana., n.d.)

The first group of variables captures the economic and volumetric performance of non-alcoholic beer within each geographic area:

- Value sales: total monetary sales of non-alcoholic beer in the area;
- Volume sales: total sales volume, expressed in liters;
- Volume sales share to category: share of non-alcoholic beer volume relative to the total beer category in the same area. This variable provides a measure of category penetration, and it is expressed in hectoliters;
- Incremental volume sales: additional volume generated over the reference period, used as an indicator of category growth dynamics;
- Unit price index on category: relative price index comparing the average price of non-alcoholic beer to the overall beer category, capturing local pricing positioning;
- Volume sales change vs year ago: percentage change in volume sales compared to the previous year, used to assess recent performance trends.

These variables were selected to jointly represent size, penetration, growth, and price positioning of the category across territories.

In order to account for structural differences between geographic areas, the dataset was enriched with socio-demographic indicators obtained from official national statistics. The variables were manually integrated using publicly available data from ISTAT.

- Average income: average disposable income of residents in the geographic area, used as a proxy for purchasing power;
- Population density: number of residents per square kilometer, used to distinguish between more urbanized and more rural areas, used to compute other per capita variables.

These indicators allow the analysis to control for differences in economic conditions and urban structure that may influence consumption behavior.

To ensure comparability across geographic areas of different sizes, two per capita variables were constructed:

- Value sales per capita: calculated as total value sales divided by the resident population of the area;
- Volume sales per capita: calculated as total volume sales divided by the resident population of the area.

The use of per capita indicators lets the analysis look at how much people buy instead of how big the market is overall. This gets rid of the bias that comes from areas with a heavy number of residents.

Before developing the clustering analysis, all the variables were checked to see whether they were consistent and complete. Market data and socio-demographic data were aligned at the same geographic level, and variables expressed in different units were standardized in later stages of the analysis to ensure balanced contribution to the clustering process.
The dataset from this project shows territorial markets in a multidimensional way by putting together performance, pricing, penetration, and contextual traits.

2.4.2 Online consumer comments dataset

The sentiment analysis is based on a text-based dataset composed of comments from non-alcoholic beer videos on YouTube. The choice of the social network was made relying on the fact that it has open and spontaneous discussions that are often about product reviews, brand content, and conversations that users start about their experiences with the products.

The choice of YouTube videos was based on a category-driven approach. The goal was to obtain a sample consisting of a wide range of consumer conversations about non-alcoholic beer. Instead of using just one type of content, the videos were chosen to show different ways of talking about and discussing non-alcoholic beer on the platform.

22 videos from the following three categories were chosen:

- General content about non-alcoholic beer: such as videos that explain or talk about the category as a whole. These videos usually talk about broader concepts, like what the product consists of, why people buy it, and how it compares to regular beer.

- Reviews of NAB brands: where creators rate and compare different non-alcoholic beers. People who comment on this kind of video usually give more detailed and experience-based feedback, often focusing on taste, expectations, and perceived trade-offs.
- Branded NAB contents: comments on these videos allow to grasp how people feel about brands and how they react to branded communication.

This sampling strategy was used to make sure that the discussions had a range of tones and topics, which lowered the chance of overrepresenting one single point of view. The dataset includes a wide range of consumer conversations, from general thoughts about a category to specific opinions about a product or brand.

Before filtering and cleaning, there were 2084 comments scraped from the chosen videos. Each observation (row) in the dataset gathers one user comment. In the original dataset, there were comments in many languages, and they were all somewhat related to the research topic. For this reason, many steps were taken to prepare the final dataset for analysis.

The first selection criterion concerned language. Only comments written in English were retained in order to ensure linguistic consistency across the dataset. This choice was made to avoid distortions arising from multilingual text processing and to maintain comparability between observations. A second criterion concerned comment length. Extremely short comments, typically consisting of fewer than three words, were removed. These comments were excluded because they generally lack sufficient semantic content to meaningfully describe consumer perceptions or evaluations. In addition, comments that were clearly unrelated to non-alcoholic beer were excluded. This category includes, for example, generic reactions to video content or comments referring exclusively to production quality, background music, or unrelated topics.

The resulting dataset therefore consists only of comments that directly address the product category, specific brands, or consumption-related aspects. The entire process concerning data cleaning and preparation will be further discussed in the next paragraphs.

The dataset includes the following variables:

- Comment text: a free-text field containing the content of the user's comment. This field is populated only for rows classified as comments, while it is empty for rows representing the video entry.
- Video Title: the title of the YouTube video. This variable allows each comment to be linked to the specific content that generated the discussion.
- Record Type: a categorical variable indicating whether the row refers to a video record, to a comment record, or to a comment's reply.
- Comments Count: the total number of comments associated with the video. This value provides contextual information on the overall size of the discussion.
- Vote Count: the number of likes/upvotes received by each comment. This variable can be used as an indicator of visibility or engagement at the comment level.

Overall, the dataset combines a textual component, represented by user comments, and basic contextual metadata at the video and comment level. This structure allows comments to be analyzed while preserving the link to the originating video and to simple engagement indicators.

2.5 Territorial clustering methodology

The territorial clustering analysis aims to find groups of similar geographic NAB markets by looking at consumption levels, price positioning, and socio-demographic traits all at once. Given the fact that the research was exploratory, and there were no predefined territorial segments, an unsupervised and data-driven approach was used. The structured analytical pipeline that was used for the clustering process will be explained step by step in the next paragraphs, from preparing the data to validating the model and generating the output.

2.5.1 Dataset preparation and feature selection

The analysis is based on the territorial dataset described in Section 2.4.1, where each observation corresponds to a geographic area. The final set of variables used as input for the clustering includes:

- value sales per capita
- volume sales per capita
- average income
- unit price index relative to the beer category
- volume sales share to category

This feature set was selected to capture complementary dimensions of market performance: normalized consumption intensity, purchasing power, relative pricing, and category penetration. Absolute sales measures were deliberately excluded to avoid size effects driven by population density.

Data handling and preliminary inspection were performed using the *pandas* library. At this stage, variables were checked for missing values and internal consistency. No observations were removed, as the dataset did not present incomplete records.

2.5.2 Scaling and normalization

Since the chosen variables adopt different numerical scales, they had to be normalized before executing the clustering. Scaling was used to make sure that all of the features had the same effect on the distance calculation that the clustering algorithm used.

A RobustScaler was used, implemented into the *scikit-learn* library. This scaler uses the median and the interquartile range to standardize variables, which makes it less sensitive to extreme values than normalization techniques that use the mean. (Pedregosa et al., 2011)

```
from sklearn.preprocessing import RobustScaler

scaler = RobustScaler()
X_scaled = scaler.fit_transform(X)
```

The use of robust scaling was motivated by the presence of moderate outliers in some per capita indicators and price-related variables.

2.5.3 Clustering algorithm and exploratory configurations

The clustering analysis was implemented using the K-Means algorithm provided by the *scikit-learn* library. (Pedregosa et al., 2011) K-means partitions observations into a predefined number of clusters by minimizing the within-cluster variance.

```
from sklearn.cluster import Kmeans

kmeans = KMeans(n_clusters=3, n_init=20, random_state=42)
labels = kmeans.fit_predict(X_scaled)
```

In the exploratory phase, different configurations were tried out by changing both the number of clusters and the set of variables. In the analytical workflow, these different pipelines are called E1, E2, and E3.

Before scaling, the first configuration (E1) applies a *log1p* transformation to the population density variable. This transformation compresses the long right tail of the distribution, making it less likely that very dense urban areas will affect distance calculations. However, this compression also makes the important differences between urban and rural settings less clear; as a result, it is harder to understand the cluster centroids. Furthermore, the internal validation metrics were not as strong as those in the final configuration.

The second configuration (E2) uses a *winsorization* method, which cuts off the population density variable at certain percentile thresholds (for example, p1–p99) while keeping its original scale. This choice strikes a balance between robustness and interpretability by limiting the effect of outliers without completely changing the distribution. This method keeps more of the original information than E1 does. The resulting segmentation was better than E1; however, the absolute values of some variables, for instance the number of residents, were still too dominant, resulting in clusters biased by those variables, resulting in a lack of interpretability.

The last configuration (E3) keeps population density in its original form and only uses strong scaling before clustering. This specification found the best balance between statistical separation and substantive clarity. E3, in particular, made centroid profiles easier to understand and gave stronger internal validation metrics, such as the highest *Silhouette Score* and *Calinski–Harabasz* index among the tested configurations. Using raw density suggests that there is still some sensitivity to extreme values, but the strong scaling method and the overall feature selection strategy lessen this effect.

2.5.4 Cluster validation and model selection

Internal validation metrics are used to check how the clustering algorithm performed on the dataset. Internal validation examines the data's internal structure and assesses the efficacy of clustering observations in relation to their separation from other clusters.

The *silhouette score*, from the *scikit-learn* library, as the main way to choose a model. (Pedregosa et al., 2011) For each observation, the silhouette score shows how different the average distance to points in the same cluster is from the distance to points in the nearest alternative cluster. The metric ranges from −1 to 1, with higher values indicating better-defined clusters. (Sci-kit Learn Documentation, 2025)

The average silhouette score for the last clustering configuration (E3) is **0.398**, which means that the separation and internal cohesion levels are satisfactory for an exploratory territorial segmentation.

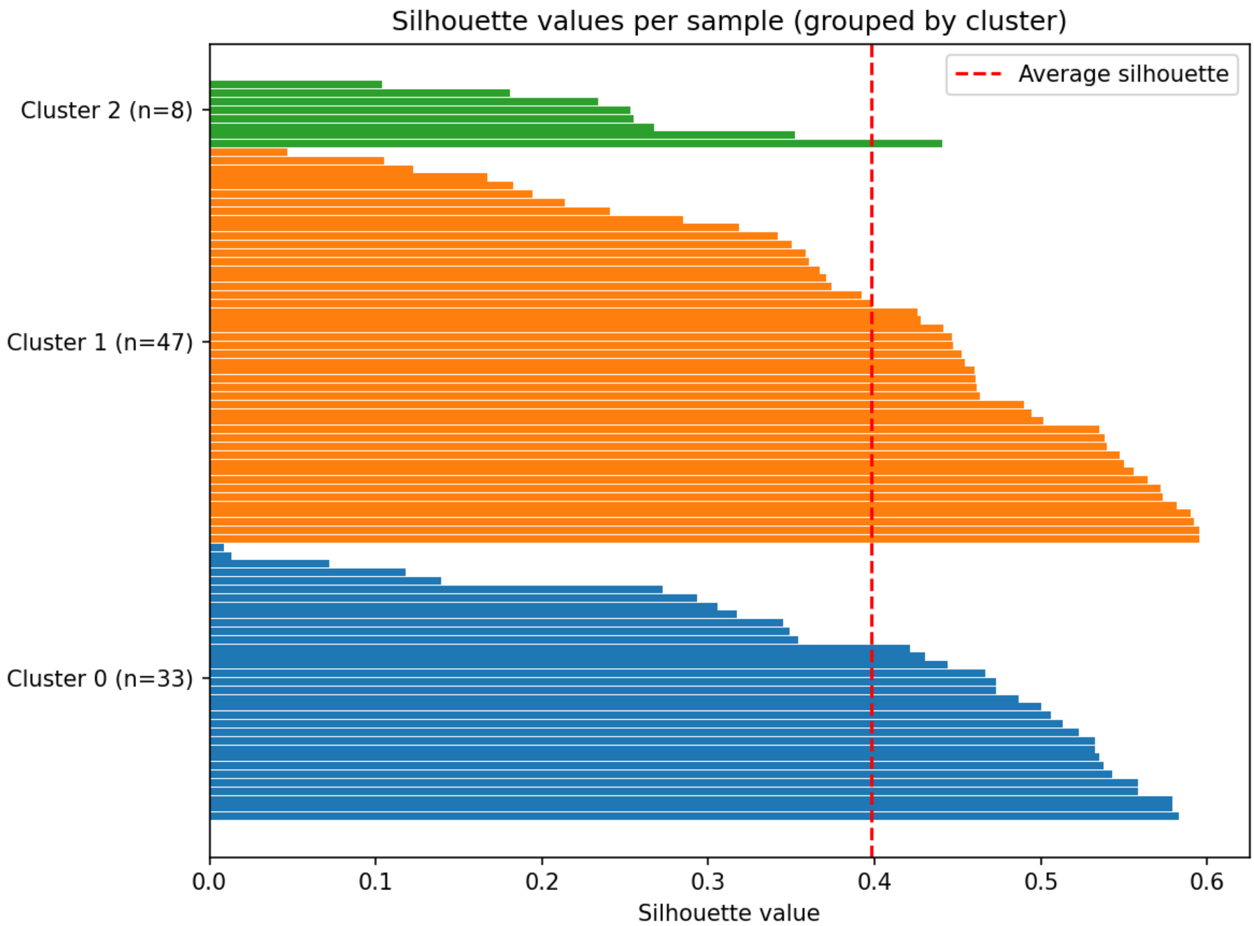

Figure 2.1 -. Number of geographic areas by cluster in the final k-means solution.

The Calinski–Harabasz index was also calculated as an additional diagnostic, resulting in a good value of **62.47**. This index ranges from 0 to +∞, it looks at the ratio of between-cluster dispersion to within-cluster dispersion. Higher values mean that the cluster structures are better defined. (Sci-kit Learn documentation, 2025)

```
from sklearn.metrics import silhouette_score

silhouette_avg = silhouette_score(X_scaled, labels)


from sklearn.metrics import calinski_harabasz_score

ch_score = calinski_harabasz_score(X_scaled, labels)
```

2.5.5 Cluster size and feature distribution

After choosing a model, the distribution of observations across clusters was checked to make sure that the segmentation did not create useless or unrepresentative groups.

```
import pandas as pd

cluster_sizes = pd.Series(labels).value_counts().sort_index()
```

The resulting clusters vary in size, indicating heterogeneity within territorial markets. Importantly, no cluster contains an insignificant number of observations, and the smallest cluster is not influenced by isolated or extreme cases.

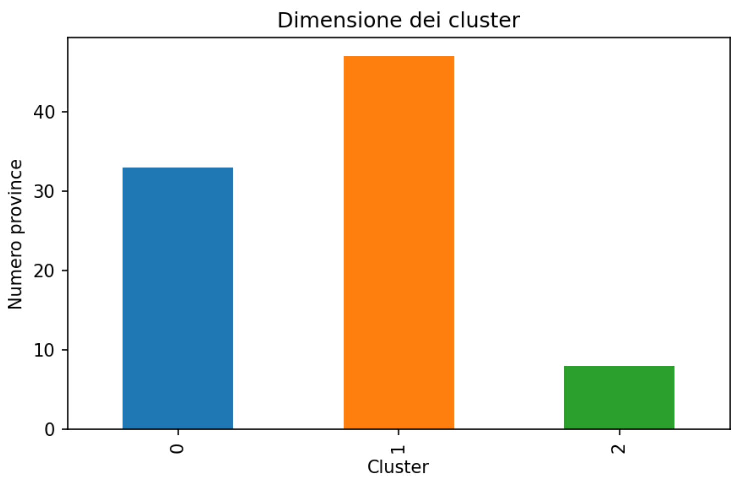

Figure 2.2 - Number of geographic areas by cluster in the final k-means solution.

Descriptive statistics and boxplots become essential to assess the distributions of the input variables by cluster, and to see how well clusters fit together. This step is a diagnostic check to make sure that clusters are different in more than one way and are not just based on one main feature. Standard Python visualization libraries are used to generate boxplots (*matplotlib*) that grouped observations by cluster membership and compared the distributions of the main input variables. The interpretation of the plot will be discussed in the third chapter.

```python
import matplotlib.pyplot as plt

df_plot = df.copy()
df_plot["cluster"] = labels

df_plot.boxplot(column=feature_names, by="cluster",
figsize=(12, 6))
plt.suptitle("")
plt.title("Feature distributions by cluster")
```

2.5.6 Cluster centroids and dimensionality reduction

To summarize cluster characteristics in a compact and comparable form, cluster centroids were computed on the scaled feature space and, subsequently, inverse-transformed to the original scale. This procedure allows centroid values to be expressed in meaningful market units.

In addition, centroid values were expressed relative to the overall mean of each feature. This representation highlights whether a given cluster is characterized by systematically above- or below-average values along each dimension.

```python
centroids_scaled = kmeans.cluster_centers_
centroids = scaler.inverse_transform(centroids_scaled)

centroids_df = pd.DataFrame(centroids, columns=feature_names)

centroids_relative = centroids_df / df[feature_names].mean()
```

Principal Component Analysis (PCA) was used as a last diagnostic step on the dataset to get a low-dimensional view of the observations. PCA is a statistical method that aims to lower the number of dimensions in a dataset while keeping as much of its variability as possible. To do this, new variables called principal components are created. These are linear combinations of the original variables that are not correlated with each other. Each new component maximizes the remaining variance in the data. (Jollife & Cadima, 2016)

From a methodological point of view, PCA is more of a descriptive and exploratory tool than an inferential one. It does not depend on distributional assumptions and is great for exploring multivariate numerical data, where the goal is to make it easier to understand while losing as little information as possible. The method is flexible because the components it creates depend on the dataset's structure rather than on a set of predetermined basis functions. (Jollife & Cadima, 2016)

This study only used PCA for visual and diagnostic purposes, to qualitatively compare how close different areas were to each other across clusters. The clustering solution was found in the full feature space, and the PCA projection was not used to change or add to the clustering process.

```
from sklearn.decomposition import PCA

pca = PCA(n_components=2, random_state=42)
X_pca = pca.fit_transform(X_scaled)
```

Figure 2.3 shows the two-dimensional PCA projection of the observations, with colors showing which cluster they belong to. This is an extra diagnostic tool that supports the internal validation metrics talked about in the previous sections.

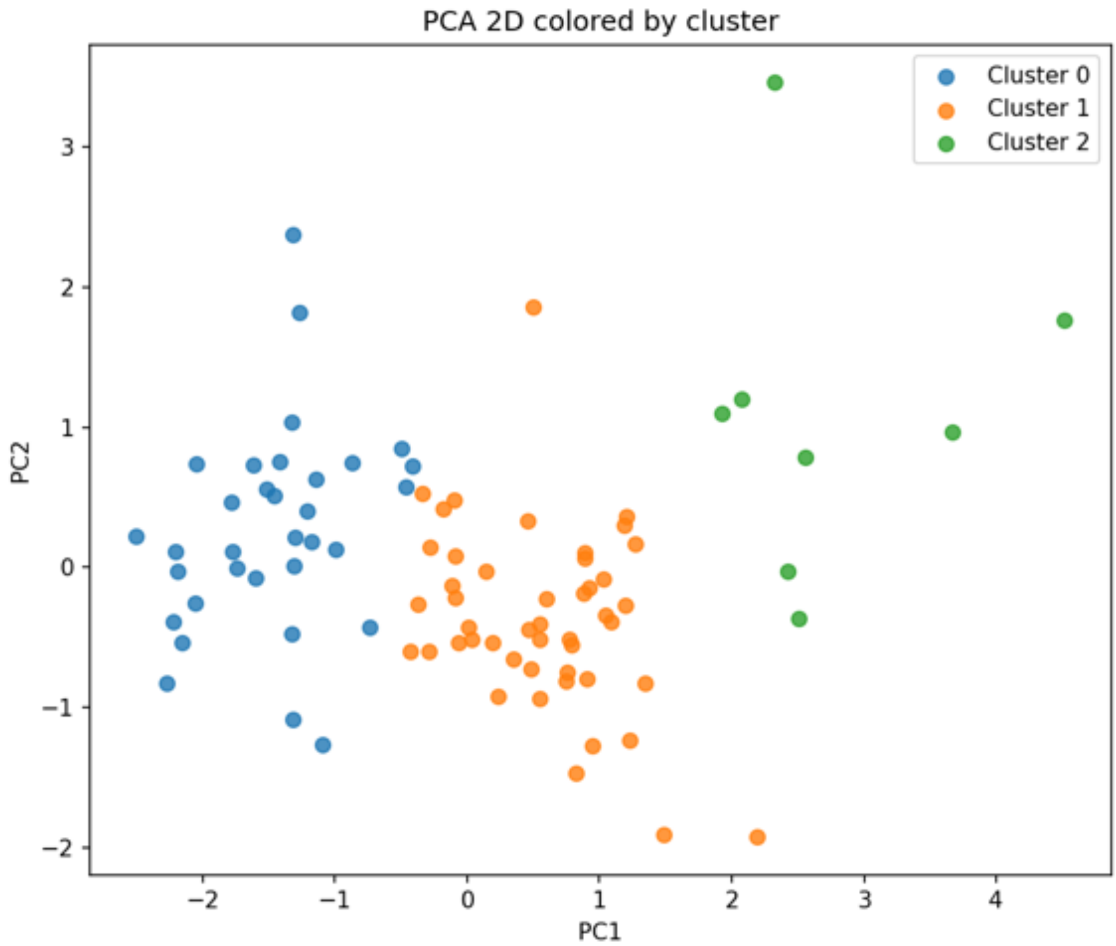

Figure 2.3. Two-dimensional PCA projection of geographic areas, colored by cluster membership.

2.5.7 Output of the clustering model

The final output of the territorial clustering analysis consists of a cluster label assigned to each geographic area, together with cluster-level summaries of the underlying market and socio-demographic variables. These outputs constitute the empirical basis for the territorial results presented in Chapter 3, where clusters are interpreted in substantive terms and linked to strategic implications for the non-alcoholic beer market.

2.6 Online consumer sentiment and thematic analysis

The second part of the empirical analysis analyzes consumers' opinions through YouTube-collected data about non-alcoholic beer. Online comments represent a great resource for people to say what they think without being influenced by the research setting. This part is a mix of functional features (no alcohol) and symbolic and experiential aspects that are usually associated with drinking beer. A two-step analytical method is used to deep dive into these aspects. A sentiment analysis is performed to find out how people feel about the products and to find out the general polarity and strength of the opinions that people share online. The second step consists of using thematic clustering to explore common and hidden topics in the analyzed comments. The following sections explain in detail how the sentiment analysis was structured, and how the thematic clustering was performed.

2.6.1 Dataset preparation and filtering

The sentiment and thematic analysis is based on a dataset composed of user comments collected from YouTube videos related to non-alcoholic beer, as described in Section 2.4.2. Each observation corresponds to an individual user comment and includes the textual content of the comment, together with basic contextual metadata, starding with a dataset composed by .

Before performing any textual analysis, a structured data preparation and filtering process was applied in order to ensure the relevance, coherence, and interpretability of the final dataset. Given the exploratory nature of the study and the heterogeneity of user-generated content, several filtering criteria were introduced to reduce noise and retain only comments that were meaningful for the research objectives.

The first filtering step concerned language consistency. Since the sentiment analysis tools employed in this study are optimized for English-language text, only comments written in English were retained, using *pandas* library. A strict language-detection procedure was applied in order to minimize the inclusion of comments written in other languages that could bias the sentiment scores and distort word-level analyses.

```
from langdetect import detect_langs

MIN_PROB_EN = 0.90

def is_english_strict(text: str, min_prob: float = MIN_PROB_EN)
-> bool:
    if not isinstance(text, str) or not text.strip():
        return False
```

```python
    try:
        langs = detect_langs(text)
        probs = {str(x.lang): x.prob for x in langs}
        return probs.get("en", 0.0) >= min_prob
    except Exception:
        return False

df_clean["is_en"] = df_clean["comment"].apply(is_english_strict)
df_en = df_clean[df_clean["is_en"]].copy()

df_en.to_csv("outputs/comments_en_strict.csv", index=False)
```

A second filtering criterion focused on comment length. Extremely short comments, typically consisting of only one or two words, were excluded from the analysis using the *langdetect* library. Although such comments may express a generic reaction, they usually lack sufficient semantic content to support reliable sentiment scoring or thematic interpretation. These observations were therefore separated from the main analytical dataset.

```python
import pandas as pd

MIN_WORDS = 3

df = pd.read_csv(INPUT_PATH)
df["word_count"] = df["comment"].fillna("").str.split().str.len()

short_df = df[df["word_count"] < MIN_WORDS]
df_clean = df[df["word_count"] >= MIN_WORDS].copy()
```

In addition, comments that were clearly off-topic were removed. This category includes messages unrelated to non-alcoholic beer, such as remarks exclusively referring to video production quality, background music, or unrelated discussions not connected to the product category. The removal of off-topic content was necessary to ensure that the analysis focused on consumer perceptions of the product rather than on peripheral aspects of the platform or the video itself.

After applying these filtering steps, the resulting dataset is composed by **1503** observations, consisting exclusively of English-language comments that directly address non-alcoholic beer, specific brands, consumption experiences, or category-related considerations, and that contain sufficient textual substance for analysis.

2.6.2 Sentiment analysis approach

This study employs a lexicon-based sentiment analysis approach to assess the emotional orientation of online consumer discourse. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis model designed for short, informal texts that are common on social media. VADER uses a list of words that have been given valence scores to show how strong or weak their emotional effect is. There is a lexical part to the model and a set of heuristic rules that are supposed to find language features that are common in user-generated content, such as negations, intensifiers, capitalization, punctuation emphasis, and emotive symbols. This combination allows VADER to get close to the contextual modulation of sentiment without needing supervised learning or labeled training data. (Hutto & Gilbert, 2014)

For every piece of text, the model makes four sentiment indicators: the ratios of negative, neutral, and positive sentiment, as well as a combined score called the compound score. The compound score is a weighted sum of the lexical valence scores that has been normalized. It can be anywhere from −1 (very negative) to +1 (very positive). This scalar measure shows the overall polarity and intensity of sentiment very clearly. The authors propose a uniform methodology for classifying sentiment labels based on threshold values assigned to the compound score. Comments with a compound score of +0.05 or higher are classified as *positive*, and comments with a compound score of −0.05 or lower are classified as *negative*. Comments that fall between these two scores are considered *neutral*. (Hutto & Gilbert, 2014; Liu, 2008)

```python
from vaderSentiment.vaderSentiment
import SentimentIntensityAnalyzer

analyzer = SentimentIntensityAnalyzer()

def assign_sentiment_label(text):
    scores = analyzer.polarity_scores(text)
    compound = scores["compound"]
    if compound >= 0.05:
        return compound, "positive"
    elif compound <= -0.05:
        return compound, "negative"
    else:
        return compound, "neutral"
```

Descriptive statistics and graphical diagnostics are being used to see how sentiment is spread across the dataset after giving it sentiment labels and compound scores. This step allows to make an initial guess about the overall emotional balance of online discussions.

Out of the 1,503 comments included in the final dataset, 64.4% are classified as positive, 21.8% as neutral, and 13.8% as negative. The distribution of compound scores shows a mean value of 0.35 and a median of 0.43, indicating that sentiment scores are predominantly located in the positive range. At the same time, the standard deviation of 0.45 highlights a non-negligible dispersion around the mean, suggesting heterogeneity in emotional intensity across comments.
Extreme sentiment values are present but limited in number. The compound score ranges from −0.99 to +0.99, while the interquartile range spans from 0.00 (25th percentile) to 0.74 (75th percentile). This pattern indicates that most observations express mild to moderately positive sentiment, whereas strongly polarized evaluations represent a smaller subset of the dataset.

After the comment-level sentiment analysis, lexical research is conducted to find the words that are most often used to describe good and bad feelings. The only goal of this step is to show how different emotional orientations in online consumer discourse are linked to different linguistic elements. Lowercasing, removing punctuation, and getting rid of common stop words are all basic preprocessing steps that break comments down into separate tokens. Then, each token is connected to the comment's sentiment score, which allows to figure out simple lexical indicators like word frequency and average sentiment intensity. These are measured by the mean compound score for each word. (Liu, 2008)

These lexical patterns are summed up using word clouds that are specific to sentiment. Word clouds are a simple way to see important words because they show words in sizes that are proportional to how important they are. People often use them as exploratory tools in text analysis to combine large amounts of unstructured text data in a way that is easy to understand, even though they do not give exact numbers. (Heimerl et al., 2014)At this stage, the analysis remains intentionally exploratory, providing a transparent overview of the lexical elements associated with sentiment polarity without introducing additional modeling assumptions. The following chapter will discuss the results of this lexical research, showing the output of it.

```
from collections import defaultdict
import re

WORD_PATTERN = re.compile(r"\b[a-z]+\b")
word_scores = defaultdict(list)

for _, row in df.iterrows():
    tokens = WORD_PATTERN.findall(row["comment"].lower())
    for token in tokens:
        word_scores[token].append(row["compound"])

word_mean_sentiment = {
    w: sum(scores) / len(scores)
    for w, scores in word_scores.items()
    if len(scores) >= 5 }
```

2.6.3 Semantic representation on consumer comments

Sentiment analysis and lexical exploration can provide an initial examination of the emotional tone and significant terminology in online consumer dialogues; however, they analyze the superficial aspects of the text. Frequency-based and bag-of-words models do not take into account semantic similarity and have problems with sparsity.

To overcome these limits, text can be shown in a multidimensional vector space, where each comment is linked to a dense numerical vector. (Aggarwal, 2018) In this representation, commonly referred to as *embedding*s, textual units that convey similar meanings are located closer to each other in the vector space, while semantically dissimilar texts are positioned farther apart.

Before generating embeddings, the dataset undergoes an additional preparation and filtering phase, built on the filtering steps already described in the previous paragraphs, aimed at improving the semantic coherence of the textual representations.

First, comments previously identified as off-topic or excessively short are excluded from the embedding pipeline. These observations typically lack meaningful narrative content and could introduce noise into the vector space. Second, basic text normalization procedures are applied, including lowercasing and the removal of non-informative symbols, in order to reduce superficial variation that does not contribute to semantic meaning.

A crucial preprocessing step concerns the normalization of brand and product references. Online discussions about non-alcoholic beer frequently include explicit mentions of brand names and product variants. While such references are relevant at the descriptive level, they might dominate similarity computations if left as they are in the dataset, leading to groupings driven primarily by brand naming rather than by underlying consumer narratives. To mitigate this effect, different surface forms referring to the same brand or product are replaced with standardized placeholder tokens. The following code snippet is provided for illustrative purposes only. The complete implementation includes a more extensive set of patterns covering additional brand names, product variants, and spelling variations.

The following list summarizes the results of the last data filtering, in order to have a more defined picture of the clean dataset that will be used in the analysis:

- Total comments before filtering: 1503
- Off-topic comments removed: 538
- Short comments removed: 149
- Total comments after filtering: 678
- Normalization coverage for "BRAND_" tokens found in the comments: 343
- Normalization coverage for "PRODUCT_" tokens found in the comments: 48

```
import re

BRAND_MAP = {
    r"\bguinness\s*0\.0\b|\bguinness\s*zero\b|\bguinness0\b":
"BRAND_GUINNESS",
    r"\bheineken\s*0\.0\b|\bheineken0\b": "BRAND_HEINEKEN",
    r"\bbud\s*zero\b|\bbudzero\b": "BRAND_BUD",
}

def normalize_brands(text):
    t = text.lower()
    for pattern, token in BRAND_MAP.items():
        t = re.sub(pattern, token, t)
    return t

df["text_for_embedding"] =
df["comment"].apply(normalize_brands)
```

Once preprocessing is completed, each comment is transformed into a dense numerical vector through the embedding process. The embedding model encodes the contextual and distributional properties of the text, producing a fixed-length vector for each comment. These vectors provide a compact and information-rich representation of consumer narratives, suitable for quantitative comparison. (Aggarwal, 2018)

The following code snippet illustrates the core step of embedding generation:

```
from sentence_transformers import SentenceTransformer

model = SentenceTransformer("all-MiniLM-L6-v2")
embeddings = model.encode(df["text_for_embedding"].tolist())
```

The output of this step is a matrix of numerical vectors, where each row corresponds to a single comment. At this stage, the embeddings are not analyzed directly; rather, they serve as an intermediate representation that enables the identification of latent structure in the data. (Aggarwal, 2018) Such representations are particularly effective when combined with unsupervised learning techniques, which can exploit semantic proximity in the vector space to uncover recurring patterns in textual data, as will be seen in the next paragraph.

2.6.4 Thematic clustering of consumer comments

When embeddings put consumer comments into a continuous semantic space, it's possible to find patterns in online conversations by grouping comments that are semantically similar.

To achieve this, an unsupervised clustering technique is utilized on the embedding vectors, aiming to reveal underlying thematic structures within consumer narratives. Clustering is a key method in text mining when there is not any labeled data and the goal is to explore. Clustering algorithms use semantic proximity to group text units that have similar meaning, even if they are very different at the lexical level. This is possible by working directly on vector-based representations. In this case, each cluster can be thought of as a group of comments that all have the same main idea or subject matter. The embedding matrix from the previous step is used in the clustering process. A partition-based clustering algorithm puts each comment into one cluster based on where it is in the semantic space. This choice is made considering that the goal of the analysis is to find a small number of recurring narrative patterns, rather than improving predictive performance (Aggarwal, 2018).

```
from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=K, random_state=42)
cluster_labels = kmeans.fit_predict(embeddings)
```

Validation focuses on the internal coherence and robustness of the clustering solution. From a descriptive perspective, the resulting clusters show a balanced distribution of observations. In particular, the four clusters contain 212, 209, 158, and 99 comments respectively, for a total of 678 textual units included in the analysis.

Importantly, no cluster is composed of a marginal or negligible number of observations. Even the smallest cluster includes nearly one hundred comments, indicating that all clusters represent substantive and recurring patterns in consumer data rather than isolated or noisy cases. This size distribution supports the robustness of the clustering solution and provides a solid basis for subsequent thematic interpretation, which will be explained in the third chapter.

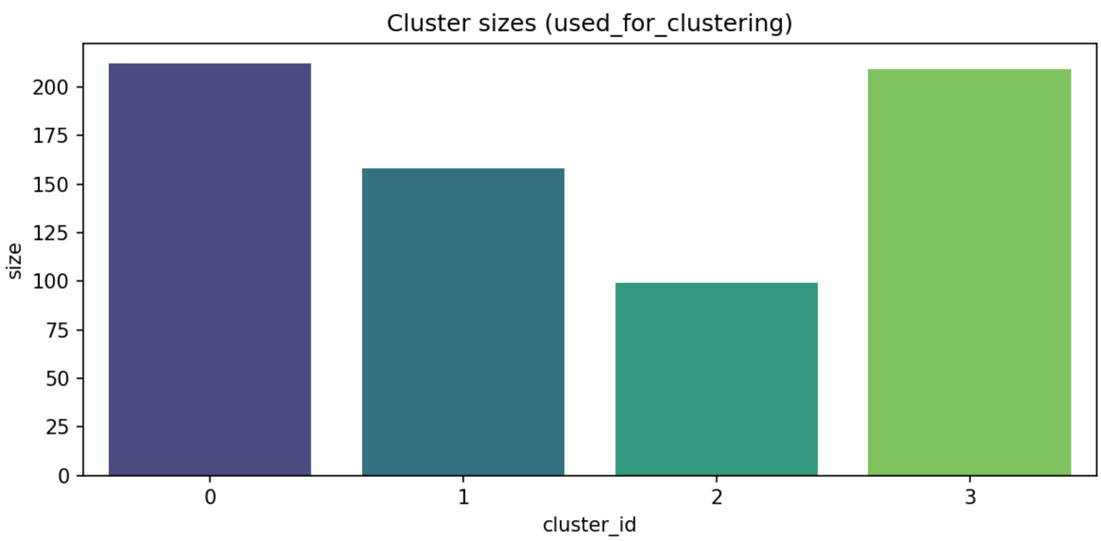

Figure 2.4. distribution of thematic clusters.

2.6.5 Analytical outputs and transition to results

The analytical pipeline described in this section produces a structured dataset in which each consumer comment is enriched with multiple analytical attributes. In particular, each textual unit is associated with a sentiment score and polarity label, word-level lexical indicators derived from the sentiment-based exploration, and a cluster label obtained from the embedding-based thematic clustering procedure.

This integrated structure enables the joint analysis of emotional orientation and semantic content. Sentiment-related information captures the affective dimension of online discourse, while cluster assignments reflect the underlying thematic organization of consumer narratives. Importantly, these outputs are generated independently of any substantive interpretation at this stage and are treated as analytical descriptors rather than final results.

Thanks to the methodological explanation of the models in chapter 2, the technical basis to understand the results and implications discussed in chapter 3 are set.

3. Results

3.1 Introduction to empirical results

This chapter discusses and analyzes the empirical results of using the analytical models from Chapter 2. The analysis is structured around the research questions that guided the study, aiming to link quantitative data to both economic and managerial insights. In particular, the chapter first addresses RQ3, by identifying homogeneous territorial market profiles through a cluster analysis based on per-capita consumption indicators, income levels, and price- and share-related metrics. It then focuses on RQ1 and RQ2, by examining consumer sentiment and the main thematic narratives emerging from online comments. Finally, the results are jointly interpreted in order to address RQ4, exploring the consistency between structural market characteristics and consumer perceptions, and deriving actionable strategic implications. This chapter does not revisit the technical parts of building or testing the models, as was done in the previous chapter. Instead, it focuses on the interpretation of empirical results, especially on how they can help managers make concrete choices about the NAB market.

3.2 Results of the territorial clustering analysis

Before presenting the results in detail, it is important to remind how to read the quantitative indicators that were used in the clustering, already explained in paragraph 2.4.1. The variables related to consumption are expressed as per-capita measures, deriving them by dividing the total value and volume of the province by the number of its residents. This method makes it possible to make meaningful comparisons between provinces with different population sizes and makes sure that the clusters found show differences in consumption behavior and market structure, not just differences in population size. The Unit Price Index on Category is also a relative price measure. Higher values mean that the average unit price is higher than the category benchmark. The Volume Sales Share to Category shows how well the NAB category performs compared to the beer market. In the following paragraphs numerical values are interpreted in relative and economic terms, focusing on whether provinces and clusters are above or below the overall sample average and what this means for market positioning.

3.2.1 Distribution and size of territorial clusters

The final clustering solution (E3) identifies three distinct territorial clusters based on per-capita consumption indicators, average income, and relative price and share metrics. From a model quality perspective, the selected configuration achieves an average silhouette score of approximately 0.40 and a Calinski–Harabasz index of 62.47, representing the best balance between cohesion and separation among the alternative specifications tested.

In terms of size, the clusters are distributed as follows:

Cluster 0: 33 provinces

Cluster 1: 47 provinces

Cluster 2: 8 provinces

This distribution makes it clear that the sizes of the clusters are not equal. Cluster 1 is the biggest group, with the most provinces and what can be called the "core" territorial profile of the market.

A small cluster just means that there are only a few areas with very different per-capita consumption patterns. In fact, as shown by the cluster profiles discussed in the following section, the provinces belonging to Cluster 2 exhibit consumption intensities that are significantly higher than the sample average, which naturally leads to their separation into a smaller and more specialized group.

The distribution of cluster sizes suggests that the market is made up of a lot of provinces with moderate and intermediate consumption patterns, as well as a few high-intensity markets that are easy to spot when looking at per-capita data. This diversity creates a significant foundation for distinct strategic interpretations, which are elaborated in the following analysis of cluster profiles.

Figure 3.1 provides a joint view of purchase power and consumption intensity of the provinces, by plotting income and per-capita volume simultaneously. The visualisation highlights a clear separation of clusters in the income–consumption space: Cluster 0 concentrates in the low-consumption region, Cluster 1 occupies the central area of the plot, and Cluster 2 groups provinces with substantially higher per-capita volumes. By adding the marker size (value per capita), it is possible to distinguish the high-value per capita provinces from the low-value per capita ones at a glance.

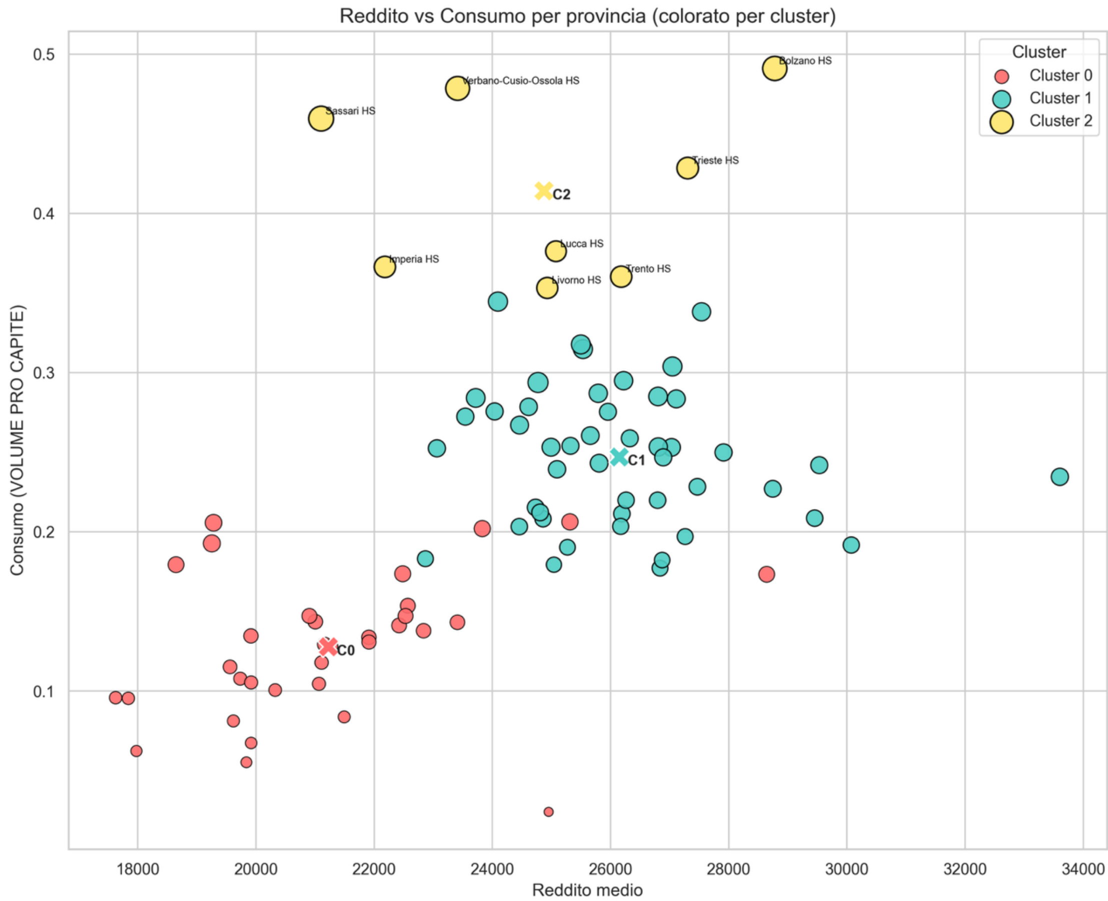

Figure 3.1 Income-Volume-Value scatter plot from territorial clustering analysis

Figure 3.2 shows how the main economic and consumption indicators are spread out across clusters. The boxplots show clear differences in both central tendency and dispersion. There is not much overlap between Cluster 2 and the other groups for per-capita value and volume. Price-related indicators, on the other hand, show more variation and some overlap, which suggests that price levels alone cannot explain why people consume differently.

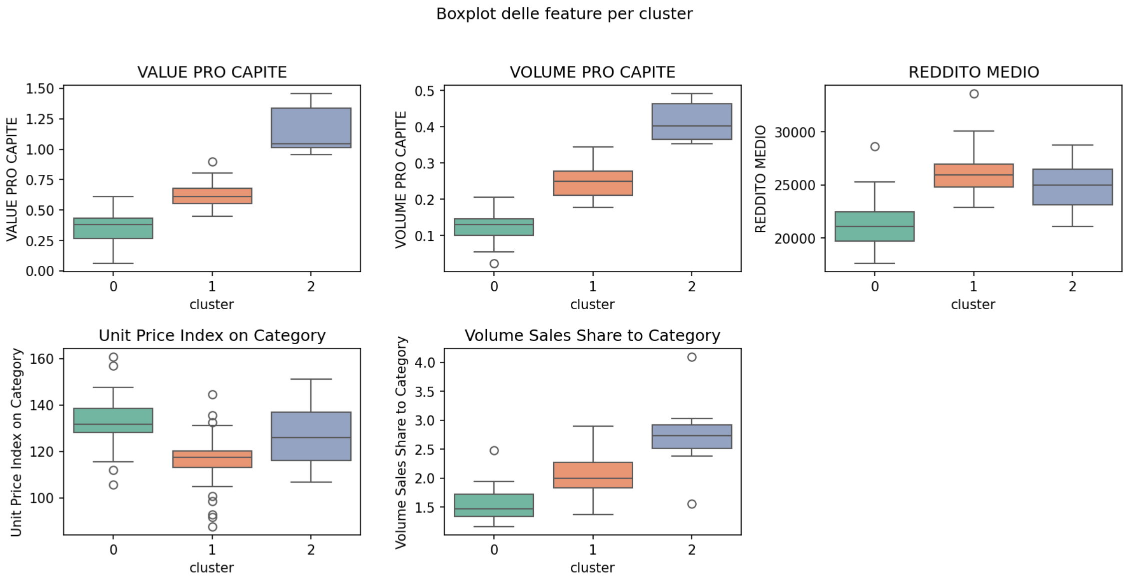

Figure 3.2 Distribution of per-capita consumption, income, unit price index, and category volume share across territorial clusters

3.2.2 Economic and consumption profiles of the clusters

The final clustering solution found three territorial clusters that have very different economic and consumption profiles. These differences appear consistently across all the key indicators examined.

Cluster 0 – Markets with High Prices and Low Intensity (n = 33)

Cluster 0 has the lowest consumption intensity per resident of the three groups. The average value consumed per person is approximately 0.36, and the average volume per person is about 0.13. Both of these numbers are much lower than the global sample averages. These numbers show that, on average, people in these provinces drink and spend a lot less non-alcoholic beer than people in other areas.

Although people in Cluster 0 do not purchase significantly, the Unit Price Index is still quite high, with an average value of roughly 132.7. In Cluster 1, it is about 116.5. This means that consumers in these markets tend to buy more expensive items or higher-end formats, but they do not do it as often. The average income level in this group is also the lowest of the three, around €21,220. This supports the idea that this is a market where the category is present but not deeply embedded in everyday life.

From a geographical point of view, Cluster 0 includes provinces like Ravenna, Rimini, Rome, Naples, and Palermo, which are made up of big cities and areas that are favorable to tourism along the coast. The fact that people are not purchasing much (value ≈ 0.36; volume ≈ 0.13) and prices are high (price index ≈ 133) suggests that some of the value may come from consumers who do not reside there or only visit occasionally, rather than from stable demand from people who live there.

Cluster 1: Balanced and mature consumption markets (n = 47)

Cluster 1 has the most complete and structurally balanced market profile. Provinces in this group exhibit a per-capita value consumption of approximately 0.62 and a per-capita volume of about 0.25. Both of these values are higher than the global average, and significantly greater than those in Cluster 0. This shows that residents are consuming more regularly and in a more consistent way.

Additionally, Cluster 1 has the highest average income level, about €26,148, and a Unit Price Index of about 116.5, which is lower than the other clusters. This combination suggests a balanced price-volume mix, where higher levels of consumption are attained without relying on higher unit prices. The Volume Sales Share to Category, which is about 2.05, backs up the idea that non-alcoholic beer is well-established in local consumption baskets.

Milan, Turin, Bologna, Florence, Venice, and Genoa are all examples of provinces in this group. Overall, the findings for Cluster 1—moderately high per-capita value and volume, high income, and controlled price levels—support the idea that it is the core and mature part of the market, where the category is well established and consumption remains stable over time.

Cluster 2: Markets with High-Intensity Consumption (n = 8)

Cluster 2 is clearly a group of markets with an extensive amount of consumption. The average value per person in this group is about 1.15, and the average volume per person is approximately 0.41. These numbers correspond to approximately twice the global sample mean, which indicates that Cluster 2 is the most unique group when it comes to consumption. Cluster 2 has an average income level of €24,872, which is similar to the level of Cluster 1, yet lower than its peak. However, it has the highest Volume Sales Share to Category (about 2.75) and a Unit Price Index of around 127.2, indicating that it has high purchasing intensity. This shows that the category has a lot of penetration and can keep prices high, meaning that the market is very favorable toward this category.

Bolzano, Trento, Trieste, Imperia, and Sassari are examples of provinces in this group that always rank among the best in terms of per-capita value consumption. The data indicate that factors besides income, such as how people in the area spend their money, how tourism changes, or very effective distribution channels, are important for driving demand. For this reason, Cluster 2 can be considered a group of key strategic markets where targeted investments are likely to pay off in significant amounts.

To further synthesize the differences observed across territorial clusters, Figure 3.3 reports the normalized cluster centroids expressed as ratios with respect to the global sample mean. This representation highlights whether cluster-level averages lie above or below the overall benchmark.

Values greater than one indicate that the cluster average exceeds the global mean for a given indicator, while values below one signal lower-than-average levels. This visualization complements the distributional analysis by providing a compact and comparative summary of the cluster profiles.

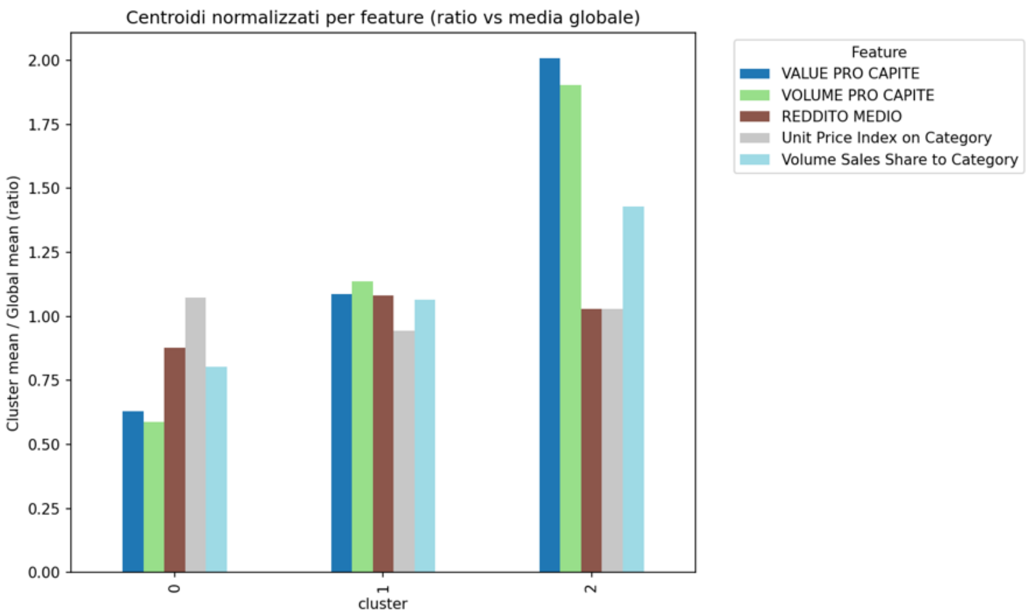

Figure 3.3 - Normalized cluster centroids expressed as ratios relative to the global sample mean

The clusters demonstrate a clear gradient of market intensity, from low-intensity, high-price but occasional markets (Cluster 0) to balanced and structurally mature markets (Cluster 1) to a small group of highly intensive and penetrated markets (Cluster 2). The significant variations between clusters offer strong evidence for different strategic approaches, which will be addressed in further detail in the following sections.

3.2.3 Geographical patterns of territorial clusters

Along with the economic and consumption profiles discussed in the previous sections, the territorial clusters also have clear patterns in where they are located. Figure 3.4 shows how provinces are distributed in clusters.

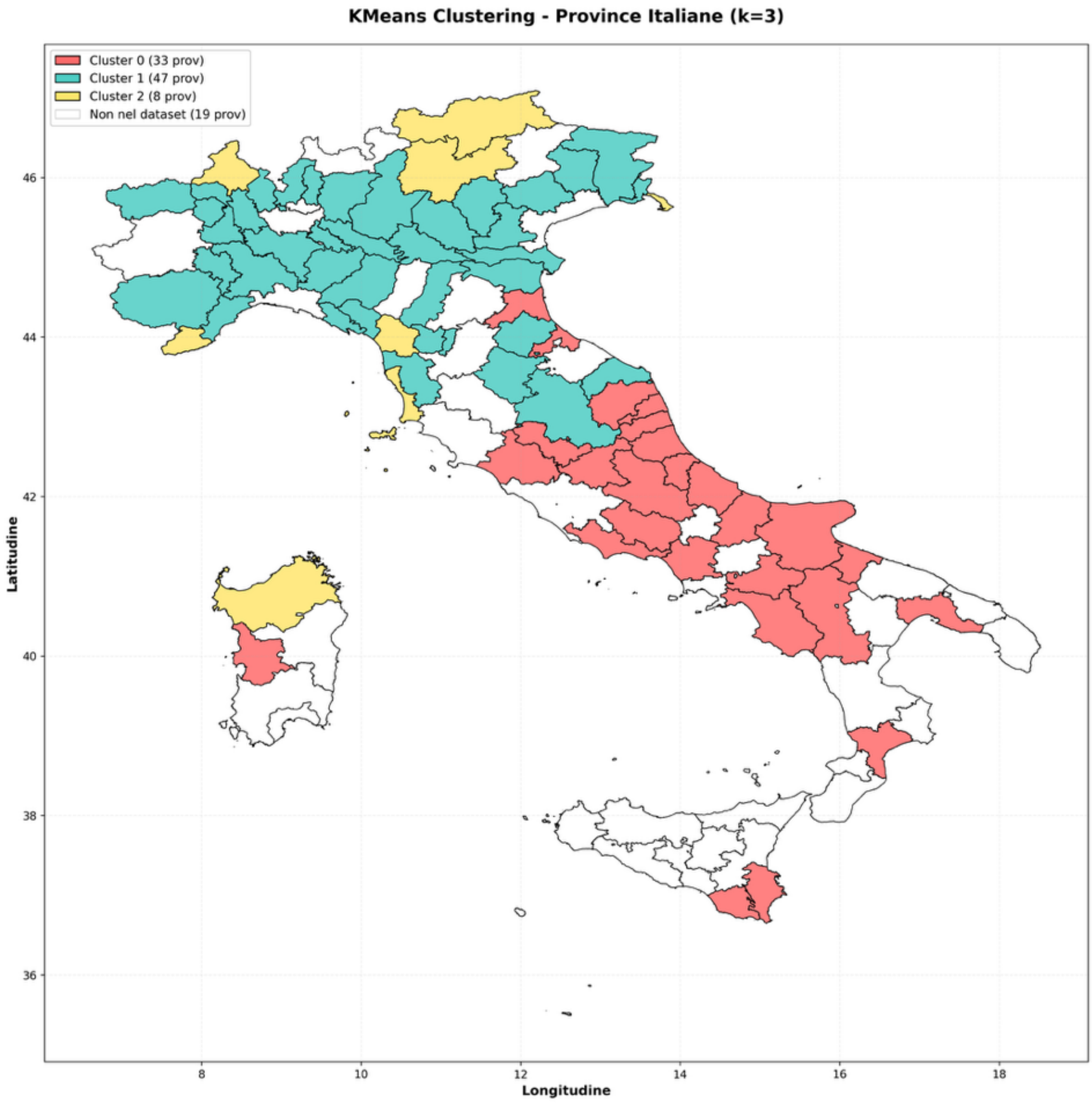

Figure 3.4 - Geographical distribution of territorial clusters across Italian provinces

Cluster 1 is mostly concentrated in Northern and Central Italy, particularly in economically developed areas characterized by dense retail networks and higher purchasing power. This geographical concentration is in line with its role as the market's core, which is stable and well-structured, where consumption patterns are more stable and consistent across provinces.

Cluster 0 is more frequent in the Southern and Central parts of the country, and in coastal provinces. This spatial pattern suggests that there are markets where consumption may be less steady and more affected by phenomena like tourism or demand from people who do not live there constantly, rather than by regular consumption from residents. The large area covered by this cluster supports the idea that it is composed of a set of heterogeneous and less structurally embedded markets.

On the other hand, Cluster 2 covers provinces in different macro-areas. There is not spatial continuity; this suggests that there are cultural or structural factors in certain areas that make the category more relevant. There is not a clear regional concentration, which means that high consumption intensity is not just due to geography, but also to certain local conditions that go beyond traditional territorial boundaries. The way clusters are distributed across different areas shows that the segmentation is strong and that strategies should be different in each area.

3.3 Results of the sentiment analysis

3.3.1 Sentiment distribution in online comments

This section examines the overall distribution of sentiments conveyed in online reviews of non-alcoholic beer. The research utilizes **1,503** YouTube comments put through to a strict English language detection procedure and subsequently analyzed with the VADER sentiment classifier. Before going into more detailed lexical and thematic analyses, the goal is to give an overall sense of how people feel about the category. Overall, the sentiment distribution is clearly skewed toward positive reviews. There are 64.4% positive comments, 21.8% neutral comments, and 13.8% negative comments. This initial evidence suggests that the online discussion about non-alcoholic beer is mostly positive, with negative comments accounting for a small part of what users have posted.

Figure 3.5 shows how the compound scores are divided into three groups: positive, neutral, and negative. Neutral comments have a value very close to zero, which shows that they do not contain sufficient emotional intensity. Positive comments can be anywhere from moderately favorable to scores that are very close to the top of the scale. This suggests that the levels of enthusiasm are not the same. Negative comments are less common than neutral ones, but they are more varied and lower on the scale. This comparison shows that VADER's sentiment labels are different not only in terms of polarity but also in terms of how much they internal variability. Comments that are both positive and negative may represent a wide range of emotional intensity.

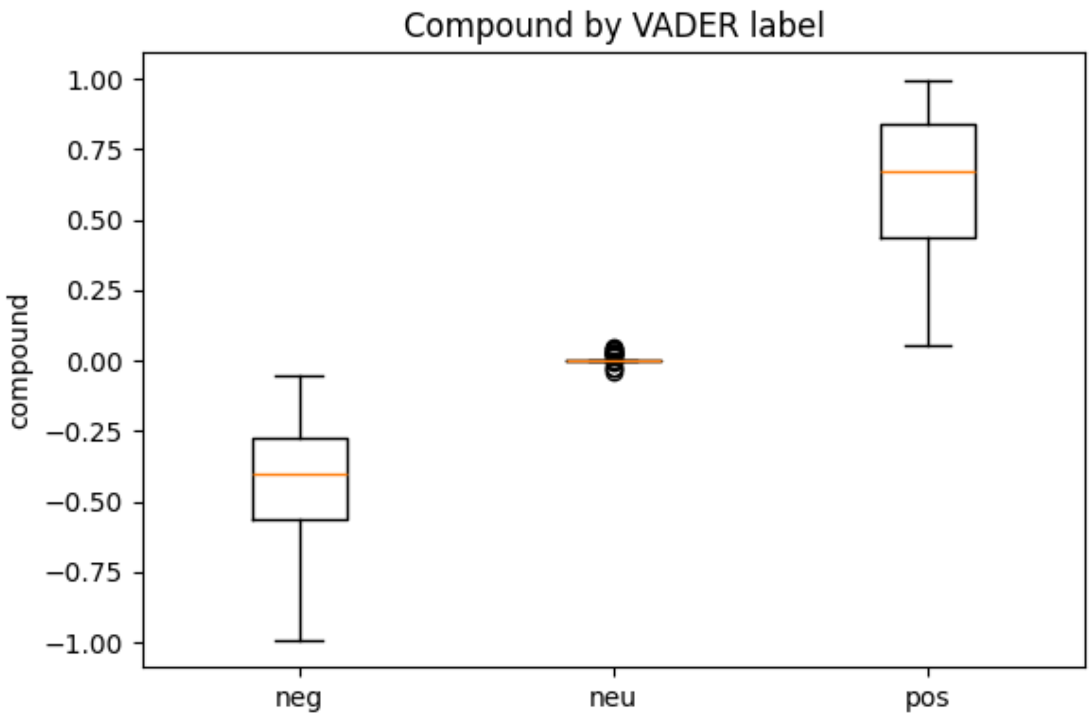

Figure 3.5 - Boxplot of VADER compound score distribution

The large amount of positive comments shows that consumers are not just accepting the product, but also showing their appreciation. At the same time, the large number of neutral comments suggests that there is a lot of informational or descriptive discourse, where individuals talk about products, experiences, or contexts without making strong emotional judgments.

The small amount of negative comments indicates that being openly unhappy is not a common view in the conversations that were looked at. It is important to note that the main point of the analysis does not appear to be the negative reviews. They are not presented alone; instead, they are mixed with positive and neutral contributions, resulting in a mostly positive mood. This balance shows that there are critical points of view without affecting the overall positive tone of the conversation. These results, when looked at as a whole, give back a rough idea of how people feel when they talk about non-alcoholic beer online. The next goal is to find the words and phrases that are most closely connected to the polarization of the sentiment.

3.3.2 Lexical and polarity-related evidences

The last part concerned the distribution and intensity of sentiment. This section addresses the word patterns that generate positive and negative comments. Two word clouds from comments that were either positive or negative were generated. The size of a word reflects the combined effect of term frequency and average sentiment polarity, allowing the most salient sentiment-related terms to emerge.

The negative word cloud (Figure 3.5) reveals a lexicon characterized by strong emotional intensity and explicit rejection. Highly frequent terms in this word cloud, such as "sucks", "shit", "awful", "sad", and "ruined" indicate that individuals frequently express negative feelings in harsh and direct ways, especially on the internet.

Along with general complaints, the negative word cloud encounters words related to health and serious medical problems, like "liver" and "cancer". The VADER model classifies these words as negative because they have a negative meaning in the sentiment lexicon. But their presence in negative sentiment outputs necessitates careful interpretation within the context of non-alcoholic beer. In this particular field, health-related terms may not always indicate disapproval or unfavorable assessments of the product itself. Conversely, discussions that present non-alcoholic beer as a safer alternative to conventional alcoholic drinks may reference conditions such as liver disease or cancer, highlighting risk mitigation and long-term health advantages. Consequently, a portion of the sentiment polarity associated with these terms is indicative of a lexicon-driven constraint of the model, rather than a definitive negative consumer perception of the category. This ambiguity indicates that negative sentiment within the health-related lexical domain may partially reflect broader perceptions of alcohol-related harm, rather than direct product dissatisfaction. So, negative evaluations seem to come from more than just rejection or disappointment.

Figure 3.6 – Negative sentiment WordCloud

The structure of the positive word cloud (Figure 3.7) is significantly distinct to the negative one. Several words pertain to having fun, being grateful, and helping others. "Beer", "great", "love", "best", and "enjoy" are the most important words, indicating that positive sentiment is strongly tied to perceived product quality and overall consumption experience. Words like "athletic," "Heineken," and "Guinness" show how specific brands can change how people think about the category, in addition to general evaluative words. This suggests that positive emotions are often linked to product experiences rather than abstract attitudes toward the category.

Words like "thank," "cheers," and "video" also point to a social and community aspect, in which individuals show their appreciation not only for the product but also for the people who made the content and the experiences they shared. The positive lexical field appears to be bigger and more complex than the negative word cloud. It is based on brand names, emotional language, and real life experiences.

Figure 3.7 – Positive sentiment WordCloud

By looking at all the lexical evidence, it can be seen that there is a substantial distinction between positive and negative sentiment expression. Several words in positive comments are about fun, quality, and social experience; the negative ones, on the other hand, use fewer but stronger words.

3.4 Thematic clustering

The sentiment analysis examined the way individuals felt about non-alcoholic beer in online conversations, revealing that the vast majority had a favorable opinion. But sentiment polarity alone does not explain why people rate the way they do, or what themes shape consumer discourse. To mitigate this problem, the current section uses a thematic clustering method to find primary narratives through which consumers express their perceptions of non-alcoholic beer.

3.4.1 Thematic clustering

The clustering analysis was conducted on a filtered subset of the original dataset, retaining only comments that were topically relevant and sufficiently informative, as described in chapter 2. Following the exclusion of off-topic and very short comments, a total of **678** comments were included in the clustering procedure.

In line with the approach detailed in the previous chapter, brand and product mentions were normalized prior to clustering in order to reduce lexical noise and prevent brand-specific references from driving the formation of clusters. This step ensured that comments were grouped based on shared semantic content rather than on the presence of specific brand names.

The clustering procedure identified four distinct thematic clusters, each oen of them representing a recurring narrative pattern within the online discourse.

Cluster sizes are relatively balanced: Cluster 0 is the largest group, with 212 comments, closely followed by Cluster 3 with 209 comments. Cluster 1 contains 158 comments, while Cluster 2 is the smallest cluster, including 99 comments. This distribution suggests that online conversations about NAB are structured around multiple coexisting narratives rather than being dominated by a single theme.

The following section offers a detailed interpretation of each cluster, examining their defining themes and illustrating them through representative comments.

3.4.2 Interpretation of consumer narratives

The thematic clustering analysis reveals four distinct narratives that influence online discussions regarding non-alcoholic beer. Each cluster shows a different way in which consumers think about their experiences, motivations, and evaluations, highlighting the multidimensional nature of consumer discourse within this category.

Cluster 0: Product experience and sensory evaluation.

The largest cluster consists of direct reviews of non-alcoholic beer as a product, with a strong focus on taste, style, and quality. People in this group often assess how much different types of beer (like IPAs, lagers, and stouts) are close to non-alcoholic alternatives. Brand names are common, but they are mostly used to talk about how the drink tastes, and how it makes the consumers feel overall. This narrative reflects a consumer perspective focused on product performance, where NAB judgment is based on the same standards as regular beer, not as a separate or secondary category.

Cluster 3: Health, sobriety, and alcohol-related reflections.

The conversations contained in this cluster are also part of bigger discussions about health, sobriety, and drinking. Customers often talk about their own experiences with cutting back on or stopping drinking, medical reasons, or lifestyle choices. Non-alcoholic beer is marketed as an opportunity to drink in social occasions without the risks associated with drinking alcohol. This narrative goes beyond the product's features; it links non-alcoholic beer to moral, medical, and identity-related issues that have to do with health and self-control.

Cluster 1: Production process and technical curiosity.

People who wrote comments pertaining in this cluster are more interested in the technical and process aspects of the category. They talk about how to brew, what ingredients to use or how to carbonate. These comments do not show strong emotional judgments; they show interest and a desire to learn more about non-alcoholic beer. This narrative shows that there are people who want to know how NAB is made and how different technical choices change the taste and quality of the beer.

Cluster 2: Recommendations and category discovery.

The smallest cluster is primarily oriented toward recommendation, discovery, and peer advice. People talk about which non-alcoholic beers are "worth trying," sharing their opinions on the best choices, and offering their thoughts to help others decide. Compared to cluster 0, this group does not spend as much time on detailed sensory evaluation as they do on giving a general overview and helping others find their way through the category. This narrative positions non-alcoholic beer as an emerging market where consumers actively contribute to collective learning and orientation.

Taken together, these narratives show that consumer discourse on non-alcoholic beer is organized around multiple complementary perspectives. While some users focus on sensory quality and product performance, others frame non-alcoholic beer in terms of health, technical interest, or social recommendation. This plurality of aspects provides a richer understanding of consumer perceptions and sets the stage for synthesizing the underlying value dimensions emerging from the clustering analysis.

To provide a visual representation of the semantic structure underlying online discussions, the embedding space of consumer comments was projected into two dimensions using dimensionality reduction techniques (PCA followed by UMAP). Figure 3.8 illustrates the resulting semantic map, where each point represents a comment, and spatial proximity reflects semantic similarity between textual perceptions.

The visualization reveals the presence of distinct yet partially overlapping semantic regions. While some clusters occupy relatively compact areas of the embedding space, Cluster 1 appears more dispersed across dimensions. This pattern reflects the heterogeneous nature of production- and process-related comments, which tend to intersect multiple perceptual domains rather than forming a narrowly defined one. As a result, Cluster 1 functions as a "semantic bridge" connecting experiential evaluations, health-related concerns, and recommendation-driven discussions.

The axes of the plot represent abstract semantic dimensions derived from the embedding space, they do not correspond to directly interpretable variables; interpretation therefore relies on relative distances and cluster structure rather than on axis values.

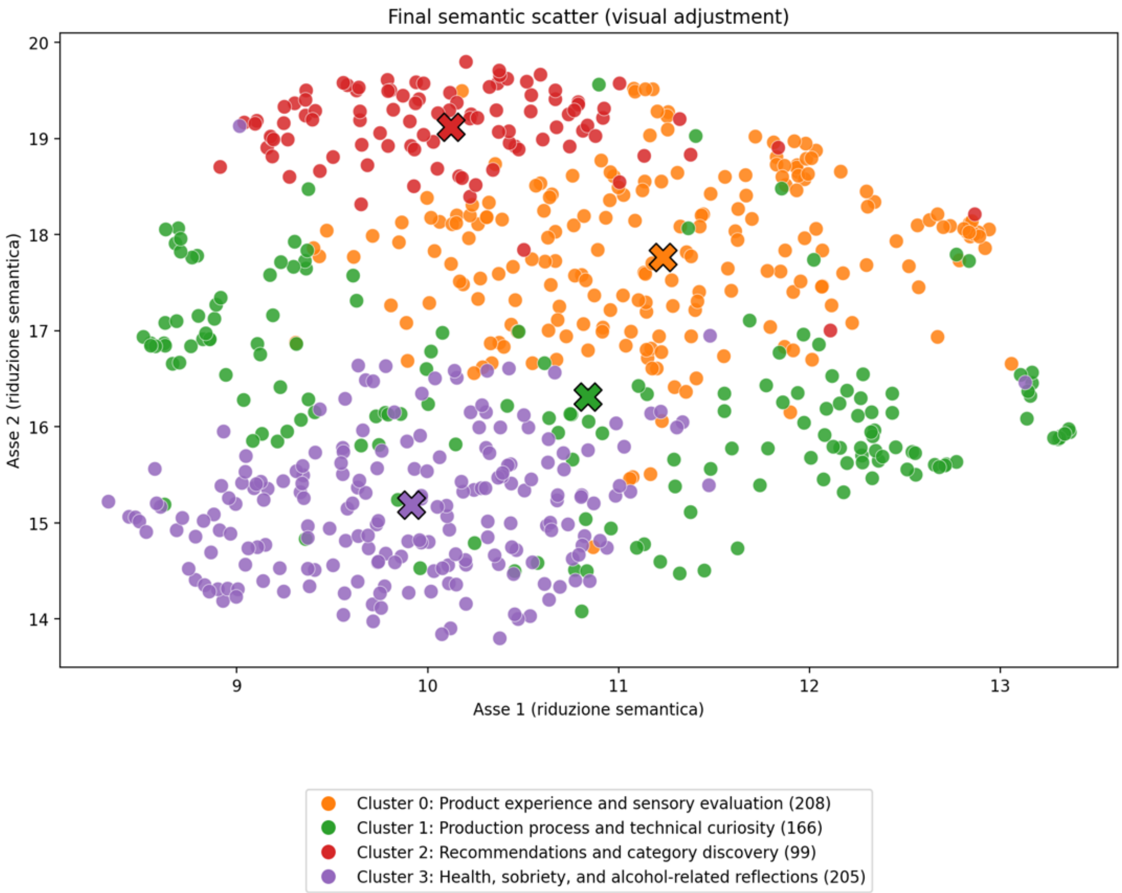

Figure 3.8 – Semantic map of online consumer comments based on embedding space

3.4.3 Synthesis of emerging value dimensions

The analysis of the four thematic clusters enables the recognition of a more extensive range of value dimensions that influence consumer discussions about non-alcoholic beer. The clusters do not provide separate narratives; instead, they can be understood as complementary expressions of a limited number of underlying dimensions through which consumers evaluate and make sense of the category.

The primary dimension concerns the quality of the product and sensory performance. This dimension includes aspects related to taste, style, and similarity to traditional alcoholic beer. People who drink non-alcoholic beer judge it based on the same quality standards that they use for regular beer, such as the intensity of the flavor, and consistency across styles. In this context, non-alcoholic beer is not regarded as a less desirable option; instead, it appears as a product that is expected to satisfy the same experiential standards as its alcoholic counterparts.

The second dimension is about health, sobriety, and drinking in a responsible manner. In this case, non-alcoholic beer is shown to be an opportunity to have enjoyable times while also taking care of your health, medical issues, and lifestyle choices. This demonstrates that consumers also buy items to help them control themselves effectively, lower risk factors, and stay healthy in the long run. Not drinking alcohol is valuable, but so is being able to enjoy the social and ritual aspects of drinking beer without any negative effects.

When it comes to technical credibility and production transparency, a third dimension emerges. This kind of consumers want to learn more about the brewing process, the ingredients, and the technological solutions that can reduce or eliminate alcohol while preserving the taste. Consumers think that craftsmanship, technical skill, and innovation are all essential elements of value; making the category more authentic and credible.

At last, the fourth dimension concerns social guidance and category navigation. Consumers actively share tips, comparisons, and shortcuts to help each other learn, given that the non-alcoholic beer market is still new and diverse. In this dimension, value is created through peer endorsement and shared discovery, positioning consumers as pivotal intermediaries in the development of category understanding.

By looking at these dimensions together, they show that non-alcoholic beer is a complex category that can be looked at from many points of view, such as health, technical, and social. This synthesis provides a unified framework for examining consumer perceptions, and serves as a conceptual bridge between the thematic clustering results and the subsequent integrated, market-oriented analysis.

3.5 Integrated interpretation and strategic implications

3.5.1 Alignments and mismatches between markets and perceptions

The analyses in the previous sections provided two different views on the NAB category. The territorial analysis identified three separate market clusters, each with its own level of consumption, price structure, and economic profile. At the same time, the thematic clustering brought out a set of value dimensions that show how consumers feel about

NAB. This section brings these points of view together through an examination at how consumer narratives align, or fail to align, with the structural features of the territorial clusters that were found in Section 3.2.

Cluster 1, which is the core and structurally balanced market, shows a strong alignment. This group has relatively high and stable per-person consumption, higher income levels, controlled unit prices, and a large share of local consumption baskets. In these provinces, non-alcoholic beer is assessed based on flavor, stylistic variation, and consistency with conventional alcoholic beers. This alignment indicates that where the category is well-established and frequently consumed, consumer perceptions transition from solely functional interpretations to more product-focused and experiential evaluation criteria.

Cluster 0 acts differently, that is because it includes provinces with low consumption intensity, higher unit prices, and lower average income levels. These markets do not seem to be as deeply rooted in everyday consumption, and demand is often driven by tourism or special occasions. In this setting, consumer narratives are more focused on health, sobriety, and responsible use. NAB is mainly seen as a way to cut down on drinking while keeping the social or ritual parts of drinking, not as a product to be judged on taste or technical grounds. This setup shows a partial misalignment. Even though the category exists and is priced relatively high, consumer discussion is still focused on functional and instrumental value dimensions. This suggests that experiential credibility has not yet fully developed.

Cluster 2, which stands for a small number of high-intensity consumption markets, has a more complex structure. These provinces have consumption levels that are much higher than the sample average, a lot of people in the high category, and prices that stay the same. Even though there are not many of them, and they are not all in the same place, they provide valuable structural contexts for the category. Narratives associated with these markets usually include evaluations of products and technical curiosity. Consumers appreciate NAB not only for its sensory qualities, but also because it is a part of stable and conscious consumption habits. This alignment indicates that significant engagement with the category cultivates more nuanced and complex narratives, in which experiential, technical, and lifestyle-related aspects coexist.

The social guidance and category navigation dimension is a cross-cutting aspect of all clusters, but its intensity varies. Peer recommendations and shared evaluations exist across all markets, but they are especially important in Cluster 0, where consumption is less regular and evaluative cues are weaker. In such circumstances, informal advice helps people deal with uncertainty and makes them feel safer in a category that is less popular and more fragmented. In contrast, in Cluster 1 and Cluster 2, where the category is more established or heavily consumed, social guidance enhances rather than replaces product-based evaluation.

In general, the integrated analysis shows that territorial market structures and consumers' perception do not evolve in the same way in all situations. There is a strong connection between structural conditions and experiential narratives in Cluster 1 and Cluster 2, yet not in Cluster 0. In this case, the category is present on the market, but consumer discourse is mostly functional. These results show how important it is to read consumer narratives in the context of specific territorial arrangements. They additionally provide a structured way to figure out how they can affect business implications.

3.5.2 Business Implications

The combined analysis of market structures and consumer narratives in different areas gives companies that sell NAB a set of useful concepts. The results do not suggest a single strategy; instead, they show that effective business decisions should depend on the specific territorial configuration in which the category operates, since consumer perceptions and value drivers vary widely between clusters.

In Cluster 1, which refers to the core and structurally balanced markets, the strong link between stable consumption patterns and product-centered narratives suggests that companies should focus on strategies for consolidation and differentiation. In these markets, consumers already think of NAB as a separate type of beer, judging it based on taste, style, and how well it fits with traditional beer standards. So, businesses should focus on increasing the value of their products through a wide range of styles, quality signaling, and a wide range of products. Pricing strategies can stay competitive without relying on premium positioning to keep people purchasing, while communication strategies can credibly emphasize sensory performance and craftsmanship.

Cluster 2, which has high-intensity consumption and strong category engagement, stands out as a group of strategic lead markets. Here, consumers share complex narratives that include their experiences, their technical curiosity, and their motivations based on their identity. These situations are great for targeted investments, testing new ideas, and leading changes for a brand. Companies can use these markets for testing new styles, formats, or methods for making NAB products because customers are already interested in them and the category is already seen as trustworthy. Cluster 2 can be used to test advanced ideas that could later be used in other markets.

For Cluster 0, where consumption is still low and consumer narratives are mostly about health and functional aspects, a different strategic logic applies. Although the unit prices are relatively high, the category is not fully integrated into daily life yet, and the experiential value dimensions are still not very clear. In these markets, business implications point to the need for reframing and educating individuals about the category. Instead of focusing on how advanced their products are, companies might do better to use communication strategies that slowly change consumers' perceptions about their products from just their practical benefits to their further social and experiential benefits. Lower-risk entry points, clearer usage occasions, and reassurance about taste and quality may help bring consumers closer to the category.

The importance of social guidance and peer endorsement across all clusters shows how informal information channels can affect what people purchase. User-generated content, recommendations, and shared reviews are all helping to make a category that continues to evolve and has an array of products, less uncertain. Companies can help this process by making it easier for consumers to leave honest reviews and by encouraging consumer communities. They should also avoid sending messages that are too strict, as these could make the reviews less honest.

Overall, the results show that strategic decisions in the NAB category should take into account both the territory and the narrative at once. Aligning product portfolios, communication strategies, and investment priorities with the most important value dimensions in each territorial cluster can make the market work better and help the category grow over time. These implications give the final discussion of the results and the answers to the research questions that guided this study a clear structure.

3.6 Final discussion of the results and research questions

This chapter analyzed the NAB segment by combining territorial market analysis with sentiment- and narrative-driven insights from online consumer discussions. The results provide a coherent response to the four research questions that directed the study, while also underscoring broader interpretative implications.

RQ1: How do consumers perceive non-alcoholic beer, and which dimensions of value emerge from online consumer discourse?

In relation to RQ1, the results indicate that consumer perceptions of NAB are predominantly positive, yet cannot be comprehensively understood solely through sentiment polarity. In addition to overall emotional orientation, online discussions show that consumers evaluate at and judge the category based on different value dimensions, such as sensory quality, health and sobriety, technical credibility, and social guidance. This suggests that NAB cannot be understood from a single perspective, but rather through various, partially overlapping viewpoints.

RQ2: Are there distinct consumer narratives surrounding non-alcoholic beer that imply different value propositions and usage contexts?

The thematic clustering in response to RQ2 confirms that there are distinct consumer narratives. These narratives suggest different value propositions and approaches to use them, from product-centered evaluations based on taste and quality to functional frameworks related to health, moderation, and responsible consumption. The findings indicate that narratives do not merely convey opinions; they actively shape the understanding and application of NAB in everyday contexts.

RQ3 Do territorial markets exhibit heterogeneous consumption and value patterns for non-alcoholic beer, indicating differentiated market opportunities?
In relation to RQ3, the territorial analysis reveals significant variability among markets regarding consumption intensity, price structure, and category embeddedness. The identification of distinct territorial clusters demonstrates that NAB does not adhere to a uniform market pattern; rather, it displays differentiated configurations that indicate varying levels of structural integration and consumption relevance.

RQ4: How can consumer narratives and territorial market profiles be jointly interpreted to inform strategic decisions in positioning, pricing, and go-to-market strategies for non-alcoholic beer?

Finally, the integrated interpretation of consumer narratives and territorial market profiles answers to RQ4. The findings indicate that market structures and consumer perceptions do not consistently progress at the same pace: some territorial clusters exhibit a significant correlation between consumption patterns and experiential narratives, whereas others reveal distinct mismatches. This joint reading underlines how important it is to take into account both market-level indicators and narrative analysis to make strategic decisions about positioning, pricing, and go-to-market approaches.

Overall, the analysis demonstrates that it is necessary to go beyond the general consumption metrics or sentiment scores to understand the NAB category, by looking at both the territorial market configurations and the narratives consumers construct.

Conclusions

This thesis aimed to investigate the consumption, perception, and evaluation of non-alcoholic beer (NAB) by consumers, and how these perceptions interact with territorial market structures to inform strategic decision-making. The study sought to deliver a comprehensive understanding of an evolving category, both economically and culturally, by integrating quantitative territorial analysis with sentiment and narrative insights from online consumer discourse.

The findings theoretically expand the literature on low- and no-alcohol products by considering non-alcoholic beer as a hybrid category. NAB is not only regarded as a functional replacement for alcoholic beer; it is also seen as a product with value that derives from numerous factors, such as taste, health and sobriety, technical credibility, and social interaction. This multidimensionality challenges reductive interpretations of the category, and underscores the significance of meaning-making processes in influencing consumption behavior.

The thesis methodologically illustrates how important it is to combine territorial market segmentation with narrative analysis of user-generated content. Market data indicates where and how much non-alcoholic beer is consumed, but narrative analysis reveals how consumers perceive and justify the category in different situations. These approaches work optimally when used together, especially in new or rapidly evolving categories.

From a managerial standpoint, the results show that successful strategies in the NAB category cannot depend on a "one size fits all" logic. Territorial markets differ not only in how much consumers purchase but also in how they perceive the category.

The study is subject to several limitations. First, the analysis of consumer perceptions is based on YouTube comments, which originate with a self-selected audience who may not be representative of the general population. Second, sentiment analysis relies on a lexicon-driven model, which might not fully capture contextual or domain-specific nuances. Finally, the empirical analysis is limited to a single national context, making it difficult to generalize the results to other countries or contexts outside of Italy. These constraints are not intended to invalidate the findings, but rather delineate the parameters for their interpretation.

Future research may enhance this study by expanding the analysis to additional countries, comparing various low- and no-alcohol categories, or employing a longitudinal approach to investigate the evolution of consumer narratives as the category develops. Subsequent research may incorporate experimental or survey methodologies to confirm and improve the insights obtained from online discourse.

In conclusion, this thesis demonstrates that comprehending the evolution of non-alcoholic beer necessitates transcending aggregate consumption metrics, and focusing on the narratives that consumers employ to explain drinking without alcohol. The study provides a holistic framework to assess the present condition and future prospects of the non-alcoholic beer category, by connecting territorial market structures with consumer meaning-making processes.

Bibliography

Aggarwal, C. C. (2018). Machine learning for text. In *Machine Learning for Text*. Springer International Publishing. https://doi.org/10.1007/978-3-319-73531-3

Alcohol Change UK. (2022, April). *Alcohol-free and low alcohol drinks*. Alcohol Change UK.

Anderson, P., Jané Llopis, E., O'donnell, A., Manthey, J., & Rehm, J. (2020). Impact of low and no alcohol beers on purchases of alcohol: interrupted time series analysis of British household shopping data, 2015-2018. *BMJ Open*, *10*(10). https://doi.org/10.1136/bmjopen-2019-036371

Apify. (n.d.). *Apify*.

Arthur, D., & Vassilvitskii, S. (2007). *k-means++: The Advantages of Careful Seeding*.

Booth, T. (2019). *Python Data Analytics*.

Calvo-Porral, C., Orosa-González, J., & Blazquez-Lozano, F. (2018). A clustered-based segmentation of beer consumers: from "beer lovers" to "beer to fuddle." *British Food Journal*, *120*(6), 1280–1294. https://doi.org/10.1108/BFJ-11-2017-0628

Circana. (n.d.). *Unify+*.

European Commission. (2025). *Consumer Conditions Scoreboard*. https://doi.org/10.2838/5794128

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis 5th Edition Cluster Analysis 5th Edition WILEY SERIES IN PROBABILITY AND STATISTICS Cluster Analysis 5th Edition*.

FAO. (2017). *The future of food and agriculture - Trends and challenges*.

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 1833–1842. https://doi.org/10.1109/HICSS.2014.231

Hutto, C. J., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. http://sentic.net/

Jackowski, M., & Trusek, A. (2018). Non-Alcoholic beer production-an overview. *Polish Journal of Chemical Technology*, *20*(4), 32–38. https://doi.org/10.2478/pjct-2018-0051

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

James, Gareth., Witten, Daniela., Hastie, Trevor., & Tibshirani, Robert. (2017). *An introduction to statistical learning : with applications in R*. Springer : Springer Science+Business Media.

Jollife, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065). https://doi.org/10.1098/rsta.2015.0202

Liu, B. (2008). *Sentiment Analysis and Opinion Mining*. University of Illinois at Chicago.

Microsoft. (n.d.). *Visual Studio Code*.

Misell, A. (2018, March). *Review: Alcohol-free and low alcohol beer*. Alcohol Change UK.

Misell, A. (2024, April). *Low-alcohol, alcohol-free, non-alcoholic: What does it all mean?* Alcohol Change UK.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, *40*(10), 4241–4251. https://doi.org/10.1016/j.eswa.2013.01.019

Okaru, A. O., & Lachenmeier, D. W. (2022). Defining No and Low (NoLo) Alcohol Products. *Nutrients*, *14*(18). https://doi.org/10.3390/nu14183873

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. In *Foundations and Trends in Information Retrieval* (Vol. 2, Issue 2).

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, andÉdouard, & Duchesnay, Fré. (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* (Vol. 12). http://scikit-learn.sourceforge.net.

Redondo, I., Russell, C. A., & Bernal, J. (2018). To brand or not to brand a product placement? Evidence from a field study of two influence mechanisms of positive portrayals of alcohol in film. *Drug and Alcohol Review*, *37*, S366–S374. https://doi.org/10.1111/dar.12671

Ritchie, H., & Roser, M. (2024, January). *Alcohol Consumption*. Our World in Data .

Roos, C. R., Bold, K. W., Witkiewitz, K., Leeman, R. F., DeMartini, K. S., Fucito, L. M., Corbin, W. R., Mann, K., Kranzler, H. R., & O'Malley, S. S. (2021). Reward drinking and naltrexone treatment response among young adult heavy drinkers. *Addiction*, *116*(9), 2360–2371. https://doi.org/10.1111/add.15453

Salanță, L. C., Coldea, T. E., Ignat, M. V., Pop, C. R., Tofană, M., Mudura, E., Borșa, A., Pasqualone, A., & Zhao, H. (2020). Non-alcoholic and craft beer production and challenges. *Processes*, *8*, 1–22. https://doi.org/10.3390/pr8111382

Schubert, C., & Rettberg, N. (2025). Unveiling the sensorial, chemical, and flavor stability in Pilsner style non-alcoholic beer produced by fermentation with maltose intolerant yeast, thermal dealcoholization, and combined production techniques. *European Food Research and Technology*, *251*(8), 2357–2371. https://doi.org/10.1007/s00217-025-04767-1

Sci-kit Learn documentation. (2025). *calinski_harabasz_score*.

Sci-kit Learn Documentation. (2025). *Silhouette_Score*.

The Brewers of Europe. (2023). *KEY FIGURES in Italy* (Vol. 700).

World Health Organization. (2023). *A Public Health Perspective on Zero- and Low-Alcohol Beverages. Brief 10, 2023*. World Health Organization.