Department of Economics and Management

Cybercrime and fraud detection

Data analysis of Darknet market activities: preliminary data investigation of Alphabay

SUPERVISOR Gianluigi Me

CANDIDATE Francesco Tramontano

ID number 246591

ACADEMIC YEAR

2021/2022

ABSTRACT

Criminal activities have drastically risen during the past two decades. Despite the presence of law enforcement and security organizations, it's becoming more difficult to face this fast expansion, especially the marketplaces for the illicit trading and consumption of goods. Manual assessments are not sufficient anymore to analyse fastly and efficiently every case on the Dark Web, including Dark Net Markets (DNM).

In this work, the main aim is to extract meaningful patterns from the contents of a dark web market using data mining techniques. The whole work is conducted on the RapidMiner data mining tool, which provides easy-to-use operators, and on Gephi to analyse graphs.

The database, of a drug marketplace on Tor, we are going to analyse requires some data cleaning and data mining because of the way it is built, the anonymity and many missing attributes, which make this database hard to understand without a thorough cleansing of all the values that are not useful and reliable.

Methods for data mining include the generation of Association Rules, Text mining and Data extraction which help us in finding anomalies and relationships in data.

In this study, we offer a methodology for manipulating data to analyse the illicit drug traffic occurring on such platforms. This study focuses on AlphaBay drug selling and on how to conduct, through Exploratory Data Analysis, a preliminary data investigation.

INDEX

INTRODUCTION

The environment of the dark web

Since the emergence of the Dark Web at the beginning of the 21st century, researchers have investigated Dark Web activity using a variety of approaches and viewpoints. The majority of these studies focused on terrorists and extremist groups, but few of them investigated studying dark web markets, where the majority of illegal operations and malevolent transactions occur. Dark Web is regularly witnessed in the development of new marketplaces in a variety of industries. Products and services in these marketplaces are quite diverse, ranging from narcotics, guns, pornography, and child abuse to malware, software vulnerabilities, and hacking lessons, as well as the trade of documents, false IDs, stolen credit cards, and the hire of hitmen.

In general, researchers that evaluated web resources contemplated employing data mining and machine learning approaches to uncover hidden patterns in large web data. The identified hidden patterns result in the discovery of new information that aids in comprehending various elements of web resources. In criminology, which is our area of study, and specifically e-markets on the dark web, data mining plays a significant role in discovering new advantageous information about malicious products promoted in such markets, discovering relationships between products and vendors, and determining the trading volume of those vendors. These statistics help deduce if the vendor is a person or a business, the latter of which may indicate the presence of organised crime. In addition, it aids in comprehending the social and psychological structure of criminal and harmful societies. The data provided by Data Mining techniques on dark web markets aids security agencies of many specialities in their investigations and studies of crime and its perpetrators.

Accessing the dark web, extracting and organising data, cleaning and transforming the data, and lastly mining the data are the main steps required to obtain meaningful information from a dark web resource. In the following sections, I provide the mining process upon which my proposed strategy is based. The technique contains a concise overview of data mining, data pre-processing, and the data mining methodologies applied.

The rise of cryptomarkets

From February 2011 to November 2013, Silk Road held a monopoly-like status. Other cryptomarkets emerged, including Black Market Reloaded, Sheep Marketplace, and Atlantis, although their earnings were very minuscule. In November 2013, the operator of Silk Road was

discovered, arrested, and the site was shut down. This resulted in significant changes to the ecosystem: erstwhile also-rans (Black Market Reloaded and Sheep Marketplace) experienced a sudden surge of Silk Road clients; several short-lived markets (such as Black Flag) failed to fill the void left by Silk Road's death. This chaotic situation stabilised roughly a month later with the emergence of Silk Road 2.0, which was operated by former Silk Road employees and utilised an interface visually strikingly similar to Silk Road's; several other marketplaces (Pandora, Hydra, Evolution, Agora, Cloud 9, etc.) also emerged within a few months, resulting in a diverse ecosystem that flourished throughout 2014. Indeed, the total income of all marketplaces rapidly surpassed that of the Silk Road during its heyday. In November 2014, government authorities again intervened with Operation Onymous, which resulted in the closure of Silk Road 2.0 and many lesser-known sites (e.g., Cloud 9). Since November 2014, traffic has been predominantly centred on the Agora and Evolution markets.

The marketplace AlphaBay

Supposedly, the AlphaBay marketplace was created in the middle of 2014. It became live on December 26, 2014, shortly after the occurrence of Operation Onymous. Similar to the Evolution marketplace, AlphaBay was allegedly founded by "carders," or those who traded stolen credit card information and other financial credentials. However, AlphaBay rapidly added entries for illegal substances. Initially, AlphaBay was a rather little marketplace, dwarfed by Evolution and Agora. As Evolution closed its doors in the middle of 2015, its exposure began to expand significantly, and it became one of the leading marketplaces by the end of the year. Supposedly, by 2016, it was the unchallenged leader of the cryptocurrency business. We will be able to corroborate these historical narratives through our investigation.

The Dark Web is the hidden portion of the Internet that employs anonymizing software. The Dark Web's dangerous environment creates significant national security concerns. The National Security Agency reported in 2013 that for more than a decade, coordination and communication between al-global Qaeda's leadership predominantly occurred on the Dark Web. In addition, these covert Web layers enabled terrorists to gather funds, disseminate recruiting material and acquire illegal explosives and weapons. In reality, the weapons used in the 2015 terrorist assault in Paris were purchased through a German Darknet merchant on Dark Web. The spread of digital vulnerabilities through Dark Web channels is also worrying. The May 2017 WannaCry

ransomware outbreak, which infected over 200,000 machines in 150 countries and caused billions in losses, started on the Dark Web. In 2017, approximately 6,300 Dark Web markets offered more than 45,000 ransomware packages for sale, boosting income from $250,000 in 2016 to $6.2 million in 2017. In addition to botnets for distributed denial-of-service attacks, confidential material and hacker services are now available for purchase on the Dark Web. This portion of the Internet is frequented by cybercriminals, terrorists, hacktivists, and non-state actors who utilise the Dark Web's anonymity to circumvent regulations.

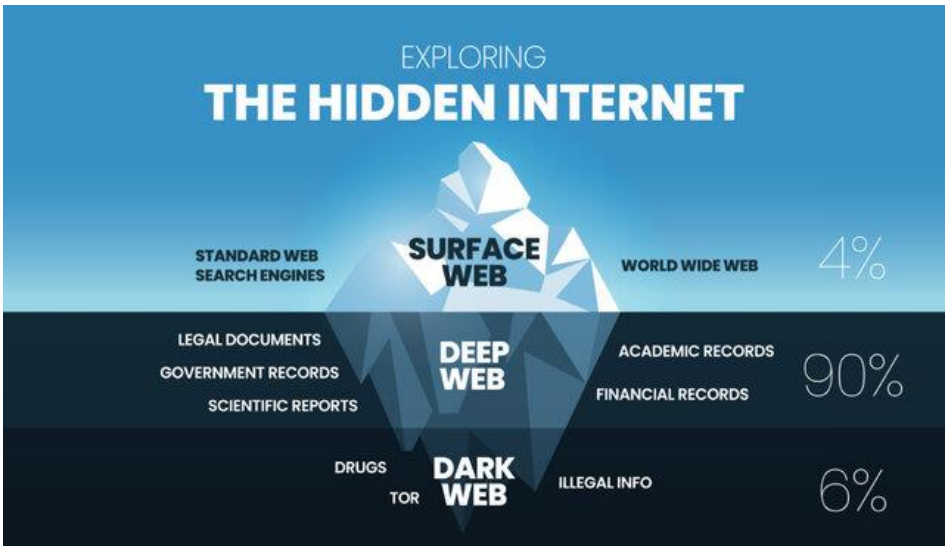


Figure 1. Internet iceberg *(Source:hhtps://stock.adobe.com/)*

The majority of online activity is conducted on the "surface web" portion of the Internet. It is the portion of the World Wide Web indexed by popular search engines such as Google and Yahoo. The majority of the Internet resides on the Deep Web, where material may only be accessed with an authentication method such as a username and password. Email databases, legal papers, and financial information are common types of information on the Deep Web. Dark Web is a subset of the Deep Web that uses anonymizing technologies to conceal user names, locations, and activities. Thus, it is home to political activists in authoritarian states, whistleblowers, and journalists, in addition to tens of thousands of black market sites.

Figure 1 depicts the various Internet tiers and their usual applications. The Dark Web is a haven where criminals may disguise their communications and organise illegal activities, such as the acquisition of malware, guns, child pornography, classified information, and drugs. In 2019, it is anticipated that transactions on the Dark Web will surpass $1 billion (Kumar & Rosenbach, 2019). Tor (The Onion Router) is a well-known anonymizing programme for accessing the Dark Web. The online underground economy is fueled by cryptocurrencies such as Bitcoin, which are decentralised digital currencies that employ peer-to-peer technology to conduct instantaneous transactions. Cryptocurrencies employ public keys or electronic addresses to safeguard all transactions, which are recorded in a blockchain, a public, unchangeable ledger. Users are allowed pseudonymity since their electronic addresses are not directly associated with personally identifying information. Nonetheless, when combined with Tor's anonymity technology, bitcoin addresses and node Internet Protocol (IP) addresses are decoupled, rendering transactions untraceable. Thus, Tor and three cryptocurrencies have facilitated the growth of Darknet markets. From $250 million in 2012 to over $872 million in 2018, the value of Bitcoin transactions on the Dark Web increased. When both buyer and seller profit from a trade, a market is thriving. To this purpose, formality, trust, and evaluation can facilitate the conditions under which parties can organise the transfer of money, products, and services. These terms also apply to the sale of illegal products and services on the Dark Web. Silk Road, the first marketplace for illicit goods on the Dark Web, utilised a feedback system to rank buyers, sellers, and escrow services in order to increase the confidence of its users in their transactions. This featured a service to hedge against the volatility of the bitcoin price. Indeed, facilitating confidence in the acquisition of unlawful products and services is the Darknet markets' principal role. Because it reduces the perceived danger of fraud and law enforcement action, the feedback loop is even more vital for protecting individual reputations than the open Internet. Therefore, the language and etiquette utilised in Darknet markets to facilitate unlawful trade can provide important insight into this anonymous environment. Understanding the inner workings of Darknet marketplaces and quantifying the traits and relationships specific to sellers and customers engaging in the trading of illicit products and services is facilitated by data analytics Classifying Dark Web dangers accurately is crucial for cyber reconnaissance, surveillance, and protection.

CHAPTER 1: DATA PREPARATION.

1.1 INTRODUCTION OF THE DATABASE

The database that required analysis concerned the sale of illicit substances in a Tor (the onion router) marketplace. Tor is currently one of the primary internet avenues utilised by both drug sellers and buyers. Tor's overlay, which lets users conceal their IP address and remain entirely anonymous when dealing online, is the primary reason why the majority of people who wish to sell illegal things use Tor. The Alphabay marketplace is not the only marketplace on the dark web; many others have been shut down throughout the years, and many others are still being used by criminals to trade illicit items, including firearms, narcotics, phoney identities, and more. Due to their anonymity, which makes every stage of the retail process considerably safer for drug dealers and their customers, these marketplaces are wildly popular with drug sellers. Moreover, thanks to the usage of cryptocurrencies, transactions are considerably more secret and expedient than with traditional methods. In comparison to the experience a customer may have with traditional e-commerce businesses, the popularity of such marketplaces stems from the simplicity with which such platforms may be used and the user-friendly interface that makes them seem extremely familiar. As noted previously, the database pertains to drug sales in Tor Marketplace. It relates specifically to the Alphabay marketplace and the transactions conducted in February 2016 by anonymous sellers. *Name*, *Vendor_name*, *Quantity_g*, *Price_bc*, *Marketplace*, *Escrow*, *Payment*, *Product_Type*, *Description*, *Shipping_options*, *Ships_from*, *Ships_to*, *Feedback* and *DateOCrawling* are the 14 attributes present in the database, which are recorded in the columns of the data set.

In addition, there are 68,652 observations (as shown by the rows of the dataset). In order to provide a much clearer picture of the dataset at hand, each characteristic will be described in the following part.

I. *Name*.

The Name parameter indicates the listing's name in the Alphabay marketplace. It is the name that the seller picked for the drug's listing when he placed it up for sale. This attribute often displays the product name, quantity and weight.

II. *Vendor_ Name*.

The Vendor Name is the name linked with the Vendor by the Nucleus marketplace. They are not anonymised, but for privacy reasons, this is one of the attributes that will be anonymized.

III. *Quantity_g*

It should indicate the quantity (in grammes) of drugs being sold by the vendor in their listings. However, this attribute is fully composed of missing values.

IV. *Price_bc*

The price bc property was anticipated to display the bitcoin price that the vendor asked for his medicine listing. Unfortunately, missing values are also present in this column.

V. *Marketplace*

This property reveals which market was utilised to list a particular deal. It is Alphabay, and each observation yields identical results. This feature might be beneficial for merging or comparing datasets from various marketplaces.

VI. *Escrow*

It is a binary variable, made up of 1 and 0, respectively present and not present.

On the other hand, it is linked to a third party collecting and keeping payments until the buyer reports receiving the product.

VII. *Payment*.

The payment method used to complete transactions. This feature has revealed that every transaction was conducted using USD.

VIII. *Product_Type*

The category of the item being offered in the listing. It separates the many sorts of sold drugs, including benzos, opioids, cannabis, and others.

IX. *Description*

This property contains far more useful information regarding the descriptions of the listings. This information pertains to the sold drug's consumption, the vendor, and the sale's amount and cost.

X. *Shipping Options*.

This property discloses the shipping method utilised by the vendor to send the purchased item. It displays the estimated delivery time and the courier service utilised to send the merchandise to the customer.

XI. *Ships_from*

This attribute reveals the location from which the vendor sends certain drugs. This can relate to a nation or global shipment (Worldwide), resulting in a significantly more anonymous transaction.

XII. *Ships_to*

The syntax is identical to that of *Ships_from*, except the nation to which the seller is exporting his product is revealed.

XIII. *Feedback*

This attribute displays the feedback provided by the customer to the vendor. Typically, this feedback is based on the buyer's happiness with how the sale went.

This may be a significant feature for a fraud study, since it may reveal whether or not the sale was fraudulent.

XIV. *DateOCrawling*

An attribute in date format covers just two days, the 23rd of march 2016 and the day after. Useless for our purposes.

1.2: DATA CLEANSING

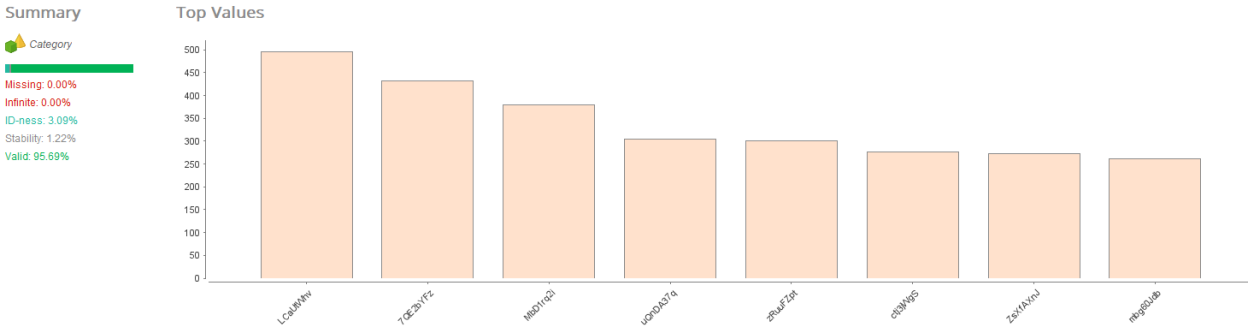
We begin by inserting a node that retrieves the connection to our database *Offer_alphabay_feb_RED.sql*, then we use the Read Database node to read it. These two nodes are then followed by a General Statistics operator that permits the collection of statistical information regarding the dataset's composition. Before beginning our studies, we must do data cleansing to remove superfluous information from the dataset. We are concerned with both the amount and quality of data. To produce a legitimate database, we must therefore identify and eradicate errors and meaningless duplicates to discover key insights and make a more accurate analysis. We examined the dataset and performed the following data cleansing procedures:

- We cleansed our data by deleting, via the *Select Attribute* operator, the attributes *DateOCrawling, Marketplace, Payment, Price_bc* and *Quantity_g* as we noticed that they did not substantially contribute to the whole analysis. In particular, *Marketplace* and *Payment* have just a single value (respectively ‘alphabay’ and ‘USD’) and, for what concerns the other two attributes *Price_bc* and *Quantity_g*, they are composed solely of missing values.

- Then, almost 30,000 duplicates were removed, using the *Remove Duplicates* node, and all the missing values, using the *Filter Examples* operator. In the original database, many observations of the variable *Feedback* were missing values. However, I decided not to remove this attribute. Instead, I removed any rows having missing values. At the end of this process, the example set was composed of approximately 38,000 observations (compared to the approximately 68,000 observations after the removal of duplicates).
- I noticed from the statistics on the raw (uncleansed) data – via the *General Statistics* node - that in the attribute *Vendor_name* there were 80 nominal values 'Non definito'. By using the *Filter Examples* operator, we removed these values by applying the filter on the *Vendor_name* attribute. The *Feedback* attribute also has some irregularities. Despite the presence of many missing values, it also has a multitude of useless 'Listing Feedback Buyer Date Time Comment' values that I have examined and have chosen to merge with the already present missing values. This operation was carried out using the *Declare missing value* operator in which the expression *'length (Feedback) <46'* was inserted among the parameters.

1.3: EDA

From the results obtained through the preprocessing stage, deeper analyses were made using different data exploration techniques. This was done in order to acquire some relevant data from the dataset at hand. Through the RapidMiner result page's visualisation area, it was possible to generate several charts to analyze the critical aspects to be examined. The first variable of relevance was *Vendor_name*, however, its values have been anonymized for security reasons. As you can see in Figure 2, we have a total amount of different vendors which is equal to 1,346 and the vendor who appears more times has the pseudonym *LCaUtWhv*, with 496 occurrences representing the 1,29% of the total listings.



1,346 Distinct Values:

Value	Count	Percentage
LCaUtWhv	496	1.29%
7QE2bYFz	432	1.12%
MbD1rq2i	380	0.99%
uQnDA37q	305	0.79%
zRuuFZpt	301	0.78%
ctj3jWgS	277	0.72%
ZsXfAXnJ	272	0.71%
mbg60Jdb	262	0.68%
6TsmOyFf	260	0.68%
gV97IAJP	259	0.67%
Q5NfnVry	244	0.63%

Figure 2. *Vendor_name* statistics

Another crucial attribute to examine is *Product_Type* to evaluate how many types of drugs are and what is the most common drug sold. The analysis shows that there is a total of 13 categories of drugs in the data set, with cannabis being the most prevalent, accounting for 26,94% of the total number of listings. This number can be represented graphically in the form of a pie chart, as seen in Figure 3 under the heading "Product_Type Pie chart." This is a graphic representation of how the drug market has been segmented.

The *Ships_from* attribute deserves further analysis as it could be important to verify which drugs are sold by a particular country and which type of drug is exported most from each country. With the use of some operators, it has been possible to create a very efficient visualization to obtain the answers to the questions just exposed. As we can see from the following Heatmap in Figure 4, there is much interesting news. For example, the types of cannabisANDhashish drugs are blatantly sold in large quantities by the united states while India provides multiple prescription and benzos drugs. Netherlands and UK are the main European suppliers. The UK sells almost all types of drugs in discrete quantities, while the Netherlands mainly supplies ecstasy, dissociative and cannabisANDhashish.

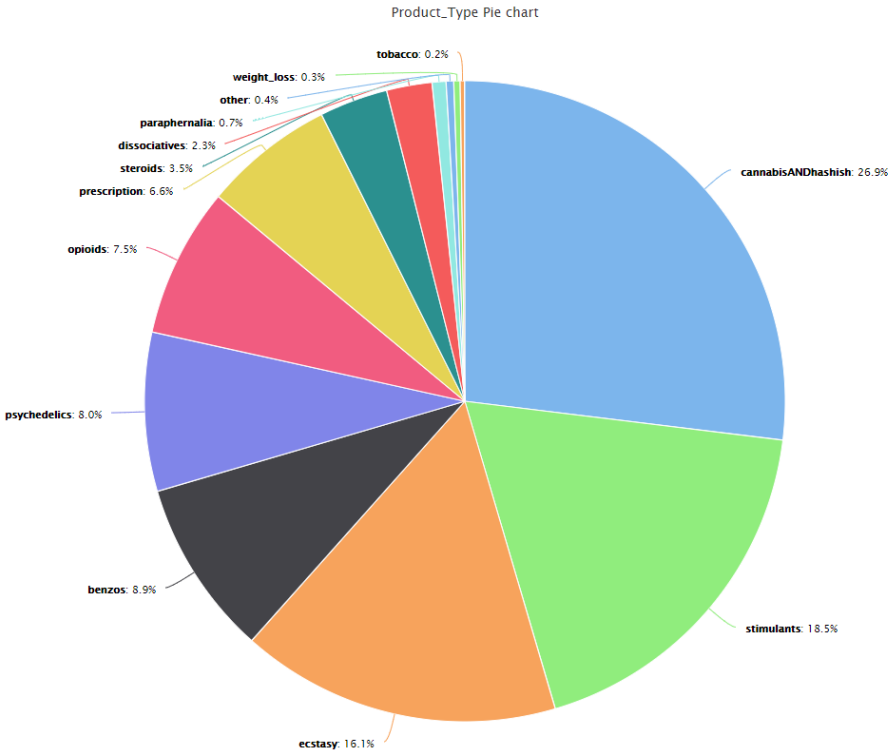


Figure 3. *Product_type* Pie chart

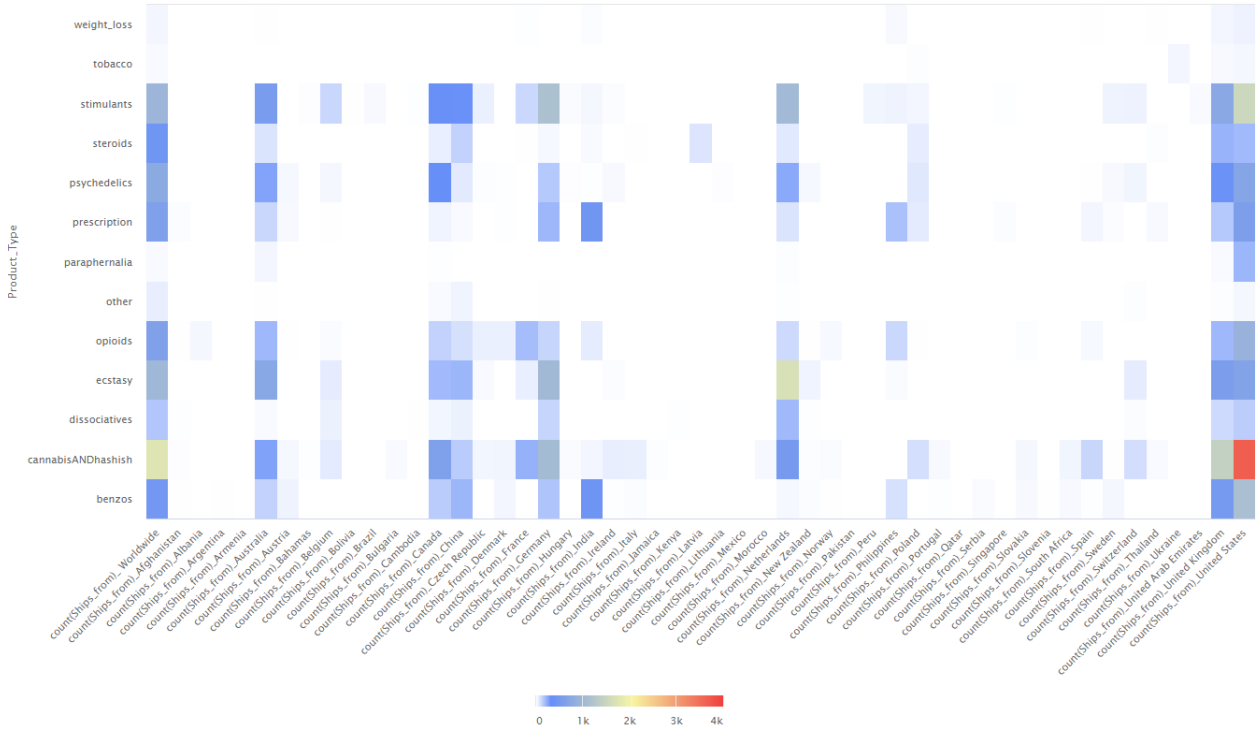


Figure 4. *Ships_from* and *Product_Type* heatmap

CHAPTER 2: DATA MINING.

Our database has several attributes with nominal values that make our dataset difficult to interpret, but at the same time, these values make multiple text extraction operations possible to derive useful information. The *Name* attribute has a very high id-ness, but by extracting the text we were able to derive the precise names of many drugs within the dataset. These operations are possible through the *Generate Extract* operator combined with the use of regular expressions, also known as regex.

2.1 REGULAR EXPRESSION

A regular expression, regex (sometimes called a rational expression) is, in theoretical computer science and formal language theory, a sequence of characters that define a search pattern. Usually, this pattern is then used by string searching algorithms for "find" or "find and replace" operations on strings. A regex pattern matches a target string. The pattern is composed of a sequence of atoms. An atom is a single point within the regex pattern that it tries to match to the target string. The simplest atom is literal, but grouping parts of the pattern to match an atom will require using ( ) as metacharacters. Metacharacters help form: atoms; quantifiers tell how many atoms (and whether it is a greedy quantifier or not); a logical OR character, which offers a set of alternatives, and a logical NOT character, which negates an atom's existence; and backreferences to refer to previous atoms of a completing pattern of atoms. A match is made, not when all the atoms of the string are matched, but rather when all the pattern atoms in the regex have matched. The idea is to make a small pattern of characters stand for a large number of possible strings, rather than compiling a large list of all the literal possibilities.

Depending on the regex processor there are many metacharacters, characters that may or may not have their literal character meaning. A regular expression often called a pattern, specifies a set of strings required for a particular purpose. A simple way to specify a finite set of strings is to list its elements or members.

2.1.2: Regex syntax

The syntax of regular expressions is very complex and long, however among the main metacharacters and those used in the study we mention:

- ^ - Matches the starting position within the string. In line-based tools, it matches the starting position of any line.
- **.** - Matches any single character except a newline character.
- ***** - Matches the preceding character or subset zero or more times.
- **+** - Matches the preceding character or subset one or more times.
- **?** - Matches the preceding character or subset zero or one time.
- **[xyz]** - A bracket expression. Matches a single character that is contained within the brackets.
- **{n,m}** - The m and n variables are non-negative integers. Matches the preceding character at least n and at most m times.
- **\d** – Matches a digit character.
- **\w** – Matches any word character including underscore.

2.2. DATA EXTRACTION.

Through the combination of the *Generate Extract* operator and the regex functions, some new attributes have been generated. These new variables will be crucial in conducting further preliminary investigations of our case.

2.2.1: Drugs and quantity extraction.

The process to derive the exact name of the drug and its quantity is quite complex, as it provides a concatenation of 4 identical *Generate Extract* operators, but which have different parameters and regex functions.

The first extraction in *Name* attribute mining is rather complex due to the very articulated and long regular expression:

\w{4,10}[\s]+[0-9]*[.,]?[0-9]+[\s]?[mgMGOZozmlMLmrLmG]+[.]?[\s]+

In this way, the operator is looking for any word linked to digits which in turn are linked to various case-sensitive units of measure, thus obtaining a new variable that I called *drug* which contains the specified name of the drug with its quantity in any unit of measure.

Subsequently, another regex helps us to create a new variable *new-drug* which contains only the name of the drug in question, simply using the regex '^[a-zA-z]{4,10}'.

At the moment, we just have to find the quantity of the drug. Thus, the regex '\w{4,10}[\s]+[0-9]*[.,]?[0-9]+[\s]?[mgMG]{2,2}[\s]+' is crucial in deriving the drugs measured in milligrams, and finally, using the latest *Generate extract*, the extraction of the quantity takes place. The last regex of this extraction is pretty simple: '[\d]+'. It just scans digits and decimals taken one or more times and stores them in the variable quantity_mg. The *Filter Example* operator then allows us to remove missing values and useless words.

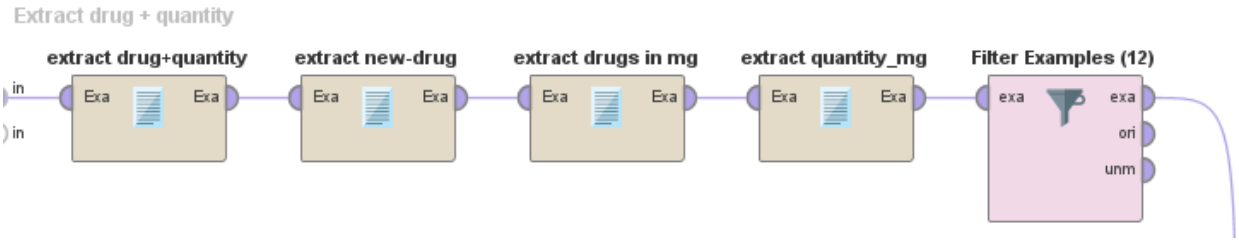


Figure 5. RapidMiner extraction subprocess

However, after obtaining the new variables new-drug and quantity_mg, I need to make some adjustments. *New-drug* attribute has plenty of specific terms but they have duplicates as there are the same values but with uppercase or lowercase letters. To correct this, we make use of the *Generate Attribute* operator, in which we select our variable new-drug and apply the function 'lower ([new-drug])'. Considering instead the other attribute obtained *Quantity_mg*, it has the digits but they are evaluated as nominal values. The *Parse Number* operator is the best choice to solve this issue since it applies the conversion. In this subprocess of drug and quantity extraction, there are two example outputs which give us some useful insights. Using the *Aggregate* operator with the function count, we can observe the best-selling drugs, indeed it is clear from Figure 6 that Xanax is definitely the most requested, followed by Adderall, modafinil and oxycodone.

The second output is linked to the *Quantity_mg* attribute that we have obtained. With a total of approximately 4000 observations, it can be seen from the following barplot in Figure 7 that nearly half of these observations cover the range from 0 to 13 mg. This data is interesting as it points out that the quantities of drugs purchased are really small and one of the reasons for this data may be linked to the lack of reliability of consumers towards these channels.

Row No.	Product_Type	new-drug	count(new-drug) ↓
68	benzos	xanax	343
500	stimulants	adderall	146
352	prescription	modafinil	99
240	opioids	oxycodone	97
395	prescription	tramadol	83
3	benzos	alprazolam	74
66	benzos	valium	74
21	benzos	diazepam	64
258	opioids	tramadol	57
293	prescription	ambien	53
310	prescription	cialis	53
400	prescription	viagra	53
17	benzos	clonazepam	52
40	benzos	lorazepam	50

Figure 6. Most frequent drugs

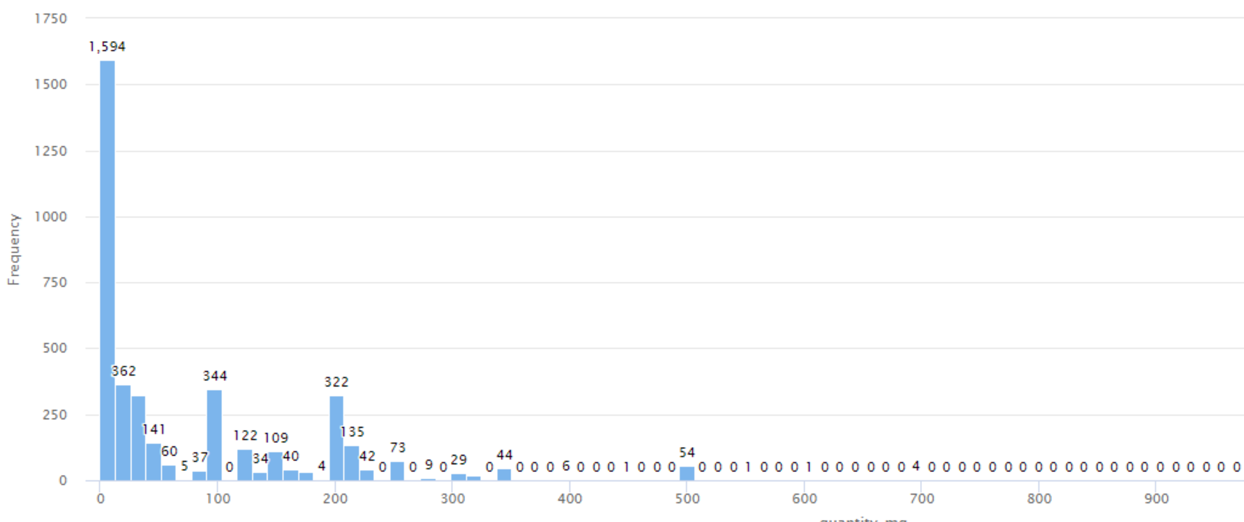


Figure 7. *Quantity_mg* distribution

2.2.2: Email extraction

The *Description* attribute has a large amount of information, indeed from its nominal values, it can be seen that some vendors put their email to have a contact channel with buyers. Our goal is to derive those email addresses to verify the presence or absence of anomalies. The use of the *Generate Extract* operator is also necessary here, as well as a new regular expression aimed at detecting text fields that include a leading word with any numbers, the '@' character and another word followed by the '. ' and from the domain. We did this using the regex: '[a-zA-Z0-9 .-] + @ [a-zA-Z -] + [.] [\ w] +'. Furthermore, we use the operator *Obfuscate* to avoid privacy issues and anonymize all the values of the new attribute.

We found 49 different emails and, looking at the following figures, with a basic graph analysis using the Gephi software, it is easy to sample just the nodes with an in-degree of two or higher. This indicates that it was feasible to include only e-mails or users that were related to two or more e-mails or users. This approach revealed that only three people are linked with two distinct e-mail addresses: "d3RROOso," "gEX6nZSn," and "Lwpfpbmi." This indicates that simply by employing this approach it is impossible to discover criminal groups as there are no users connected to the same email. However, it is easier to spot frauds or transactions from the same individual if they have many email addresses, making them less anonymous.

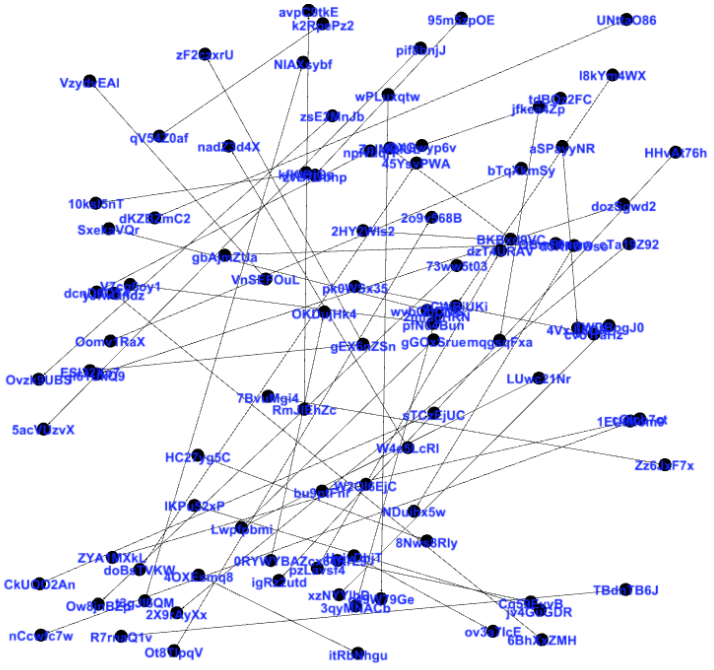


Figure 6. Complete *Vendor_name*/*new-email* graph

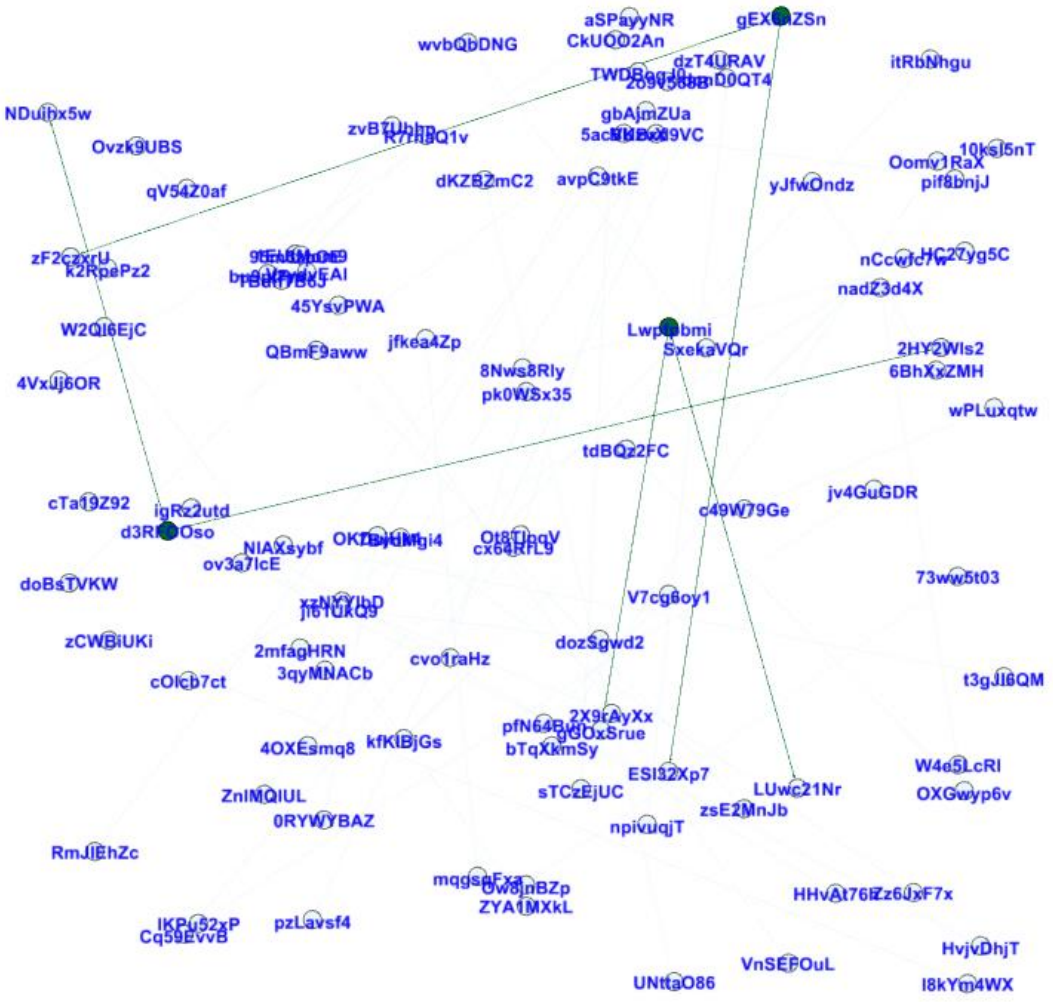


Figure 7. Vendors with more than one email

2.2.3: Shipping information extraction

Another variable of interest is *Shipping_options* as it has a lot of useful information such as the duration of the shipment and its price. Using two *Generate Extract* operators, here we get the item/order price through the regex '[UuSsDd]{3,3}[\s+]{2,2}[\d.]+\s+[/]+\s[itemorder]{4,5}' and the duration of the shipment with '[\d\sdays]{6,8}'. To get only the digits of the amount of days and the price, we use two more regular expressions: respectively '[0-9]*' and '\d{1,3}(?:[.,]\d{3})*(?:[.,]\d{2})'. We include these results in the new attributes *Ship_duration_days* and *Ship_cost_USD* and then transform the nominal values into numbers through the *Parse numbers* operator. The following horizontal bar plot in Figure 8 gives some insights.

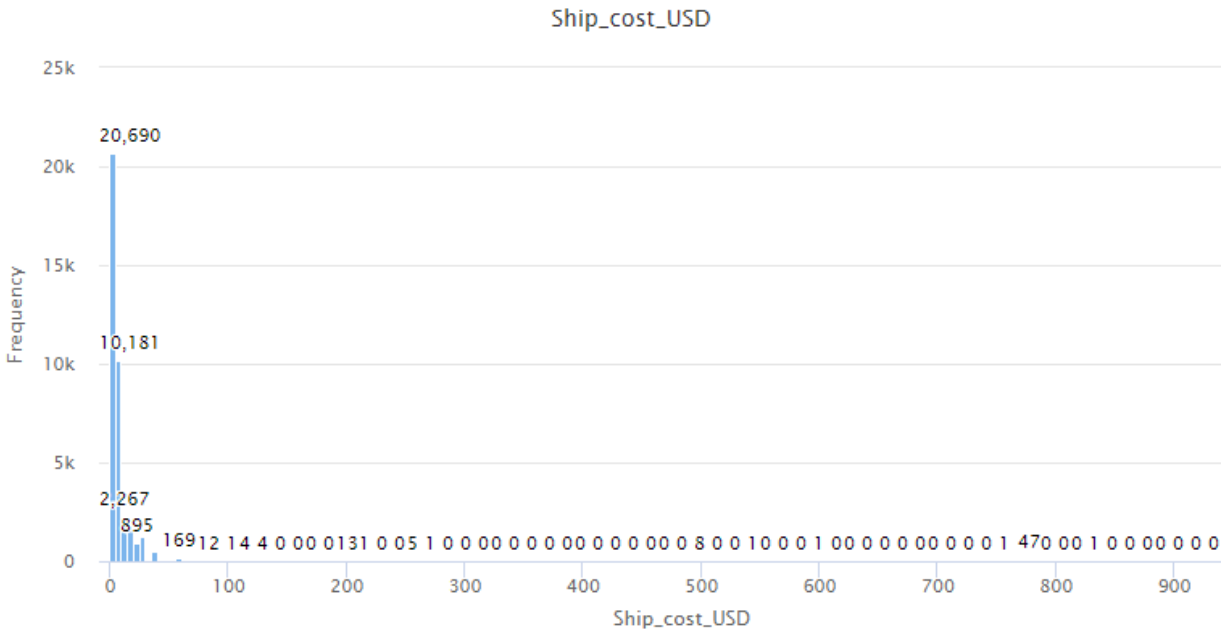


Figure 8. *Ship_cost_USD* bar plot

Regarding the duration of the shipments, there are some outliers there too, as eight observations have a value equal to 1000. However, more than 50% of the values are between one and five days.

2.3 SENTIMENT ANALYSIS

This portion of the article will discuss one of the most complex, yet extremely valuable, analyses that can be performed on an unstructured document. In the context of exploratory data analysis, sentiment analysis and text mining are crucial if the data that scientists need to obtain conceal meaningful information. Therefore, Sentiment Analysis converts the language of a document into structured data, making it suitable for quantitative analysis. The *Feedback* column of this database was subject to Sentiment Analysis. However, it was required to eliminate all missing values from the column before the model could be applied to the database with greater precision and efficacy. This reduced the number of variables in the dataset to 4132, which were then examined using an *Extract sentiment* node. This node may extract sentiment from each word in a document, which in this example is represented by the words of the column *Feedback*, by assigning a positive or negative value to each token based on the sentiment of each word. Additionally, the node allows users to contribute additional phrases they believe should be examined with a different attitude. The decision was made to add 26 words to this list. An emotion was assigned to them based on a

prior analysis performed by the node. This was done in order to make the overall model more particular and efficient for the study of a drug market dataset, such as by assigning a positive value to the word "stealth," which the algorithm had previously assigned a negative meaning. In addition, the node assigns a final score to each piece of feedback based on the average sentiment of each word in the text. It was then decided to take the *Generate Attributes* procedure a step further by passing it a function that assigns an emotion to each piece of input. Each piece of feedback was classified as negative, positive, or neutral. Thus, if the detected score was more than zero, the feedback would be deemed "Positive." If the score was below 0, it would be designated as "Negative." If the value was 0, it would be called "Neutral."

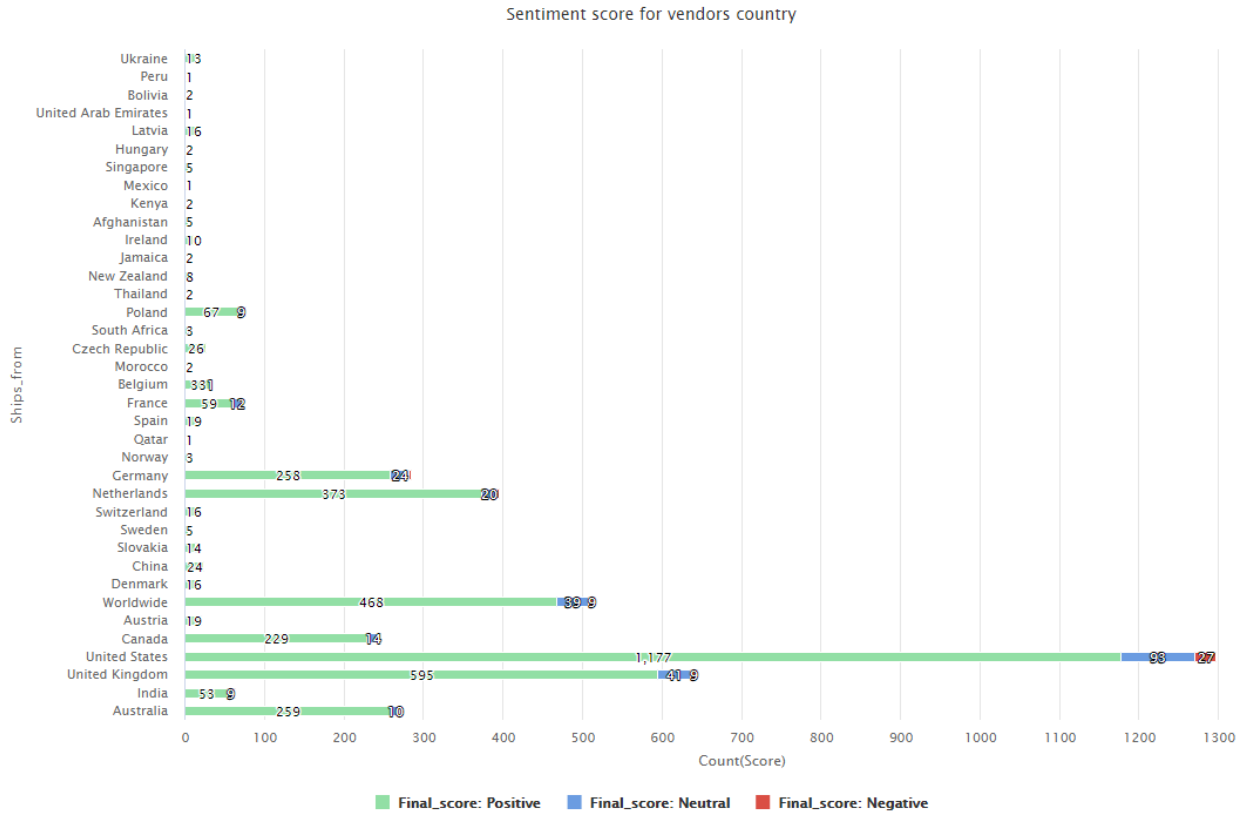


Figure 9. Sentiment score for shipping country

It is interesting to relate these feedbacks with other attributes to understand if there may be anomalies regarding positive or negative feedback.

In the representation in Figure 9, we can see how some countries from which the shipments start have a higher positive feedback percentage than others. First of all, it is obvious that most of the feedback comes from drugs shipped from the USA, in fact, they also have a greater number of

negative feedbacks. Overall, the feedbacks are almost all positive, which confirms that most of the sales satisfy customers and apparently there seem to be few scams if any.

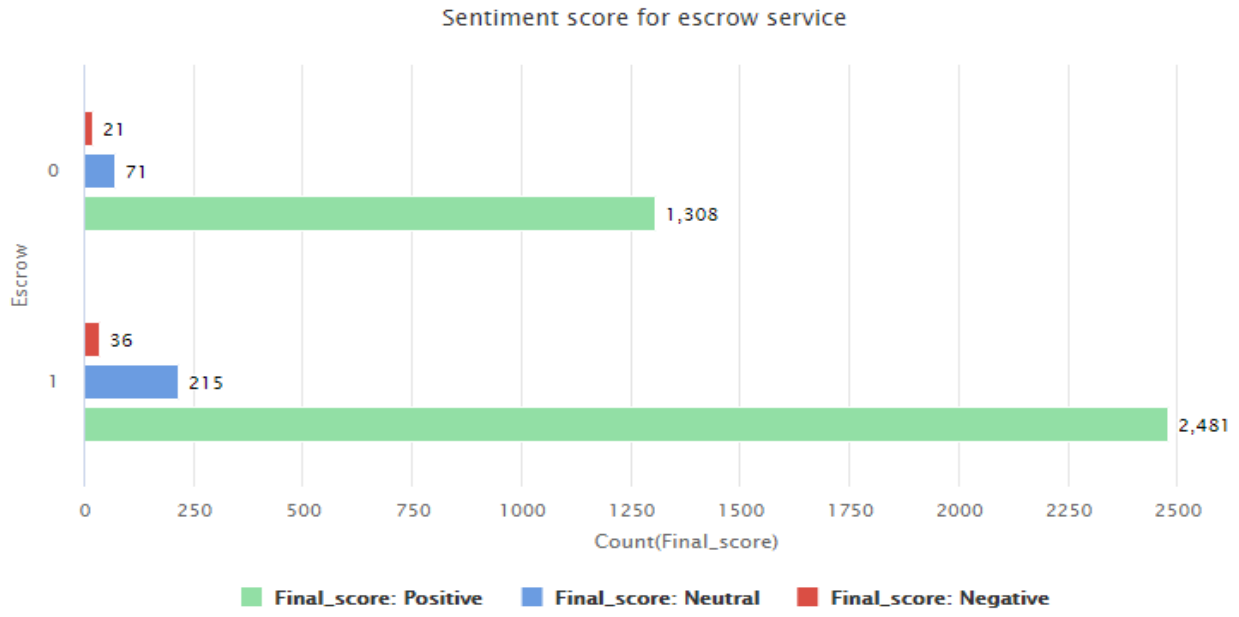


Figure 10. Sentiment score for *escrow* service

The second representation in Figure 10, on the other hand, describes the relationship with the escrow service. As we can see, the ratio of negative and positive feedback is similar in both cases, while the percentage of neutral feedback rises slightly in cases of sales with escrow service.

2.4 ASSOCIATION RULES

We applied association rules to our cleaned dataset. To get a formal definition of such rules we can say that "Association rules detect frequently occurring relationships between items" (Agrawal, Imielinski et al. 1993). Originally, they were introduced in a market basket analysis context to detect which items are frequently purchased together. However, they can be used also for fraud detection, as they might bring our attention to some relationships that may imply fraud. The goal of our analysis is to find frequently occurring and hidden relationships between the various features (variables and values) of the database.

In the following association rules, any relationships between *ships_from*, *product_type* and *final_score* attributes are analyzed. Let's define a three-step procedure:

- *Step 1.* Identify the frequent item sets. The frequency of an item set is measured using its support which is the percentage of total rows in the database that contains the item set. The support of an item set x is given by the frequency of that item set in all the rows in the dataset, or equivalently:

$$Support(\{x\}) = \frac{number\ of\ rows\ with\ x}{total\ number\ of\ rows}$$

The most frequent itemsets are the ones with higher support values. As shown in Figure 11, the most frequent itemsets are x_3,x_4 and x_5= {Final_score=Positive, Product_Type= cannabisANDhashish, Ships_from = United States} given in rows no. 3,4, and 5. The support is equal to 0.126, meaning that 37.8% of rows contain itemsets x_3,x_4 and x_5.

- *Step 2*. Derive the most frequent association rules of the type [x][y] (antecedent) => [z] (consequent). The strength of an association rule can be quantified by means of its confidence. The latter is defined as the conditional probability of the rule consequent, given the rule antecedent, or equivalently:

$$Confidence(\{x\} \Rightarrow \{y\}) = \frac{freq(X \cup Y)}{freq(X)}$$

Itemset x_2 and x_4 deserve attention: [Final_score=Positive][Ships_from = United States]=> [Product_Type=cannabisANDhashish] and [Final_score=Positive][Ships_from= United Kingdom] => [Product_Type=cannabisANDhashish] have a confidence of approximately 43% and 36% respectively. These results suggest that the cannabis and hashish products with the best sentiment scores come from the US and the UK.

- *Step 3.* Calculate the measure of the importance of a rule, also known as lift. The lift of a rule is the ratio of the observed support to that expected if X and Y were independent, or equivalently:

$$Lift(\{x\} \Rightarrow \{y\}) = \frac{Confidence(x \Rightarrow y)}{freq(x) \cdot freq(y)}$$

The lift assumes values between 0 and infinity. A lift value greater than 1 indicates that the rule body (antecedent) and the rule head (consequent) appear more often together than

expected, this means that the occurrence of the rule body has a positive effect on the occurrence of the rule head. A lift smaller than 1 indicates that the rule body and the rule head appear less often together than expected, this means that the occurrence of the rule body has a negative effect on the occurrence of the rule head. If the lift values are equal to 1, then the rule head and the rule body are independent. A value of lift greater than 1 voucher for the high association between {Y} and {X}. The greater the lift value, the greater the chances of finding {X} and {Y} together.

Once all association rules have been found, they can be closely inspected and validated. An association does not necessarily imply fraud, but it's a least worth the effort to further inspect the relationship between these variables. In the workflow, I created an "association rules" subprocess consisting of the following nodes:

- The *Select Attribute* operator to single out the variables of interest.
- The *Nominal to Binomial* operator which converts the nominal value to binomial ones. This is an essential step for applying the *Create Association Rules* node.
- The *FP-Growth* operator is based on an efficient algorithm for calculating frequently co-occurring items in a transaction database. This node creates filters the itemset according to their support
- Finally, the *Create Association Rules* operator outputs the antecedents (premises) and consequents (conclusions). This node generates a set of association rules from the given set of frequent itemset given by the *FP-Growth* operator. The resulting table also displays the support, confidence, and lift of each AR.

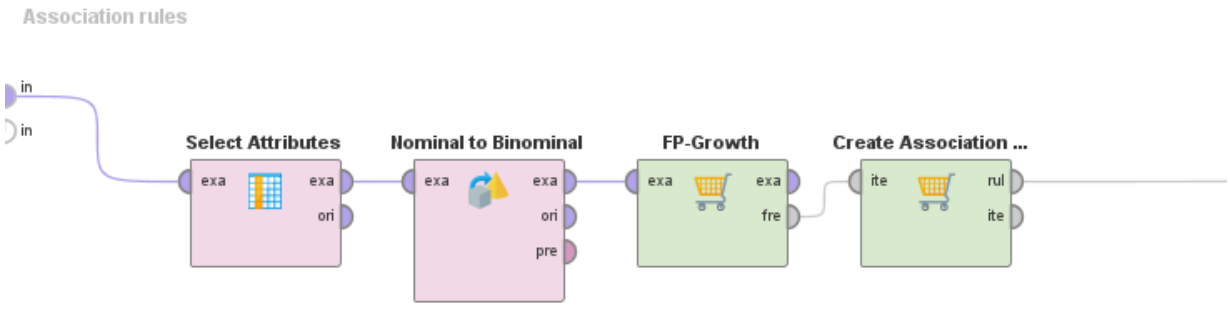


Figure 11. RapidMiner *Association rules* subprocess

By looking at the results outputted by the association rules, we obtained that the same three elements say, A, B, and C have been associated multiple times in different fashions: for instance, elements A and B are associated with *Product_Type* C; vice versa, compression type C appears to be associated with elements A and B.

- In terms of support, i.e., the number of transactions (in percentage) supporting the association rule, the highest was obtained with the ARs in rows no.3,4 and 5 with support of 0.126.
- The highest confidence which explains the strength or reliability of the association rule was obtained by the AR in row no.6, with confidence of 0.894.
- Finally, in terms of lift, that is the factor by which the confidence exceeds the expected confidence, it was not possible to calculate the lift in this case of AR, but in the next case dealing with the selection of *new-drug* and *ships_from* attributes for further creation of association rules, the lift value is equal to 3.054.

These results suggest that the cannabis and hashish products with the best sentiment scores come from the US and the UK.

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
1	Final_score = Positive, Product_Type = cannabisA...	Ships_from = United Kingdom	0.051	0.176	0.815	-0.529	0.051	∞	1.214
2	Final_score = Positive, Ships_from = United Kingd...	Product_Type = cannabisANDhashish	0.051	0.355	0.919	-0.237	0.051	∞	1.549
3	Final_score = Positive, Product_Type = cannabisA...	Ships_from = United States	0.126	0.435	0.873	-0.454	0.126	∞	1.770
4	Final_score = Positive, Ships_from = United States	Product_Type = cannabisANDhashish	0.126	0.443	0.876	-0.444	0.126	∞	1.794
5	Product_Type = cannabisANDhashish, Ships_fro...	Final_score = Positive	0.126	0.886	0.986	-0.159	0.126	∞	8.776
6	Product_Type = cannabisANDhashish, Ships_fro...	Final_score = Positive	0.051	0.894	0.994	-0.063	0.051	∞	9.440

Figure 12. *Association rules* results – 1

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
1	new-drug = adderall	Ships_from = United States	0.030	0.789	0.992	-0.046	0.020	3.054	3.518

Figure 13. *Association rules* results - 2

However, in the second case involving the *new-drug* attribute, only one AR was released which exhibits high confidence and lift values. As can be seen from the following table, the specific drug 'adderall' is mainly shipped from the United States.

2.5 BENFORD'S LAW

Benford's Law is a popular tool for fraud detection since it identifies deviations from the expected distribution that need further review. It is even legally admissible as evidence in the US in criminal cases at the federal, state, and local levels. I applied Benford's Law (also known as the Newcomb-Benford law or Law of First Digits) exclusively to numerical variables: in this case *Ship_cost_USD*, *Quantity_mg* and *Ship_duration_days* variables. Firstly, the law relies on the idea that the distribution of digits in multi-digit natural numbers is not random; instead, it follows a predictable pattern. Secondly, it is applicable exclusively to "natural numbers" (in the fraud examination sense).

Unfortunately, there is no standard "Benford's Law" operator that could allow us to compare the pattern of data according to Benford's law and the pattern hiding in our data. Ergo, we manually 'built' this functionality through a RapidMiner subprocess (consisting of many different operators). The main steps are:

- Removal of the zeroes as, generally, Benford's Law only accepts 1-9 values.
- Transformation of all numerical values into polynomial values.
- Then, the creation of two new attributes *digit*, and *digit_complex*, by using two distinct function expressions.
- Lastly, the results are aggregated into a new variable, namely, *Benford's*.

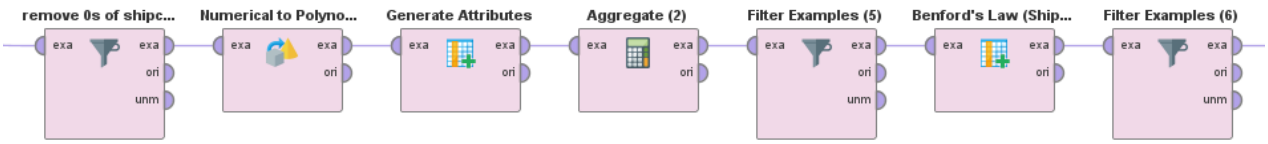


Figure 14. RapidMiner *Benford's law* subprocess

Now, we can discern if any anomaly is present by looking at any deviation of the frequencies from Benford's values. As seen in the following figures, the green bars in the bar plot represent

Benford's one, while the blue ones represent our data. There is no significant difference between Benford's distribution (in green) and the distribution of the first digits of the variable *Ship_cost_USD* (in blue). The same can't be said for *Quantity_mg* and *Ship_duration_days* attributes. Particularly, the former has a lack of high numbers while, in the latter, digit 9 is far off from its expected distribution, implying that there may be some anomalies in the dataset.

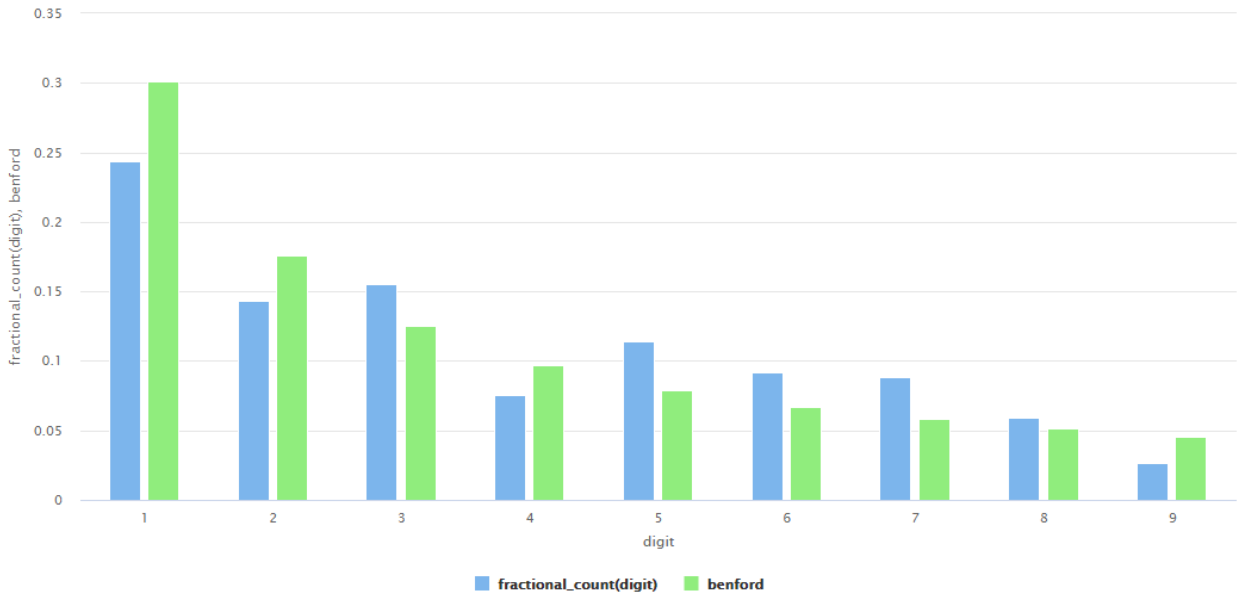


Figure 15. *Ship_cost_USD* distribution

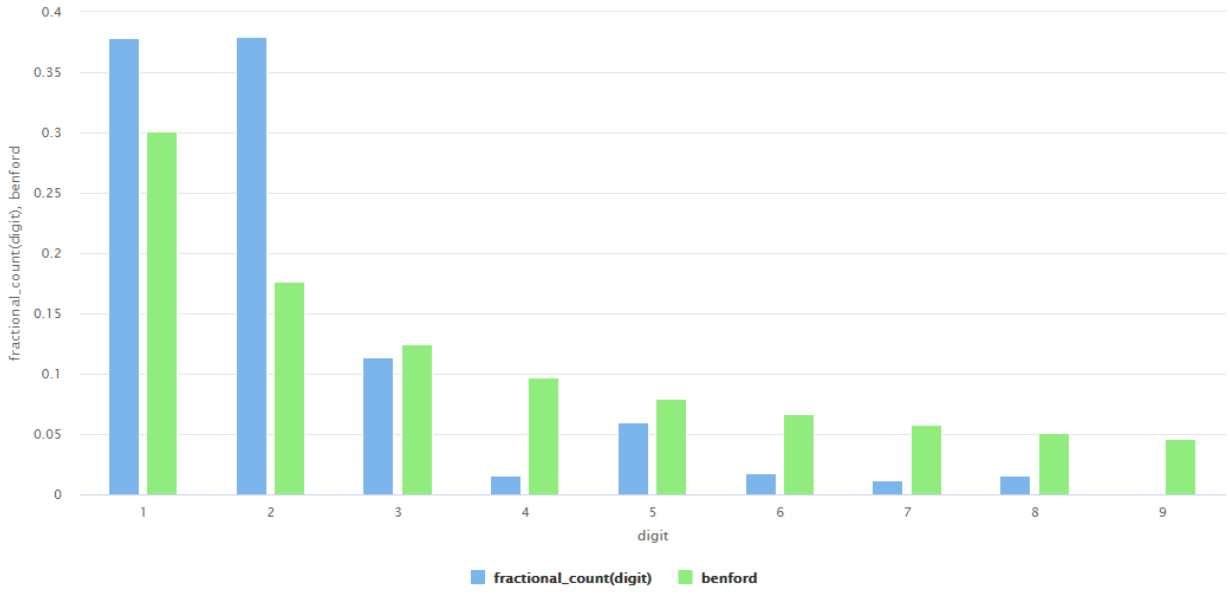


Figure 16. *quantity_mg* distribution

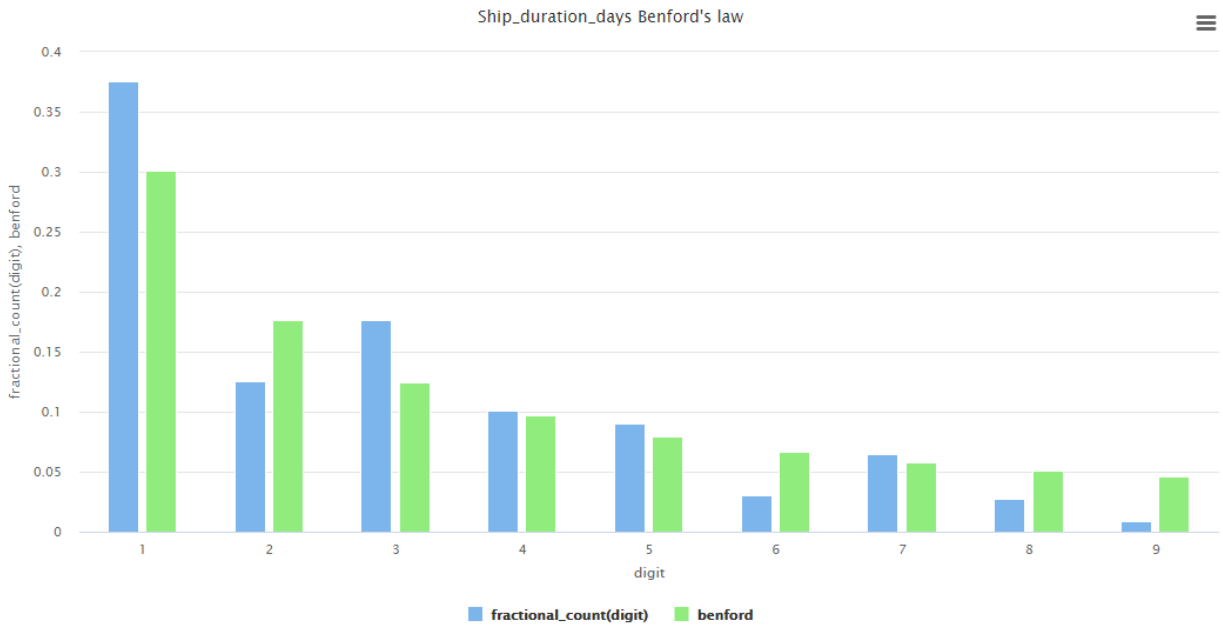


Figure 17. *Ship_duration_days* distribution

2.6 TEXT ANALYSIS

Description attribute is a delicate variable that deserves a separate analysis. it is composed of nominal values which, if analyzed, can provide a large amount of information. text analysis is an excellent practice for verifying the presence and frequency of words that may be useful for our investigations. On RapidMiner, the process is quite complex. Initially, nominal values must be transformed into text with the *Nominal to text* operator. The output is then linked with a very important operator, which is *Process Document From Data*. The main tokenizing operations take place within it. This process, in Figure 18, begins with transforming the complete text into many different tokens, and then the text is transformed into all lowercase.

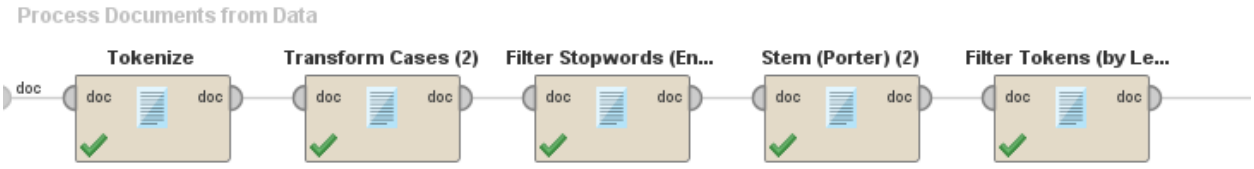


Figure 18. RapidMiner *Tokenization* subprocess

Then there is the removal of stopwords that are not useful for our purposes, followed by stemming and token filtering by length which serve to elide the suffixes of some words to enclose different variants of words in a single token.

Following this process, it is then necessary to transform the numerical data into binomial to connect them to the *FP-Growth* attribute that will illustrate the results, expressed in terms of frequency and support.

Size	Support	Item 1	Item 2
1	0.249	time	
1	0.234	list	
1	0.229	us	
1	0.215	get	
1	0.198	price	
1	0.191	free	
1	0.183	pill	
1	0.182	good	
1	0.178	weed	
1	0.177	mdma	
1	0.170	profil	

Figure 19. *Tokens* frequency (size 1)

Size ↓	Support	Item 1	Item 2
2	0.087	time	list
2	0.121	time	us
2	0.100	time	get
2	0.078	time	price
2	0.079	time	free
2	0.098	time	make
2	0.084	time	custom
2	0.093	time	packag
2	0.080	time	take
2	0.084	time	address
2	0.079	time	deliveri
2	0.088	list	us
2	0.090	us	get
2	0.089	us	make
2	0.081	us	custom
2	0.082	us	name
2	0.080	us	packag
2	0.095	us	address
2	0.079	us	pgp
2	0.081	get	price
2	0.087	pill	mdma

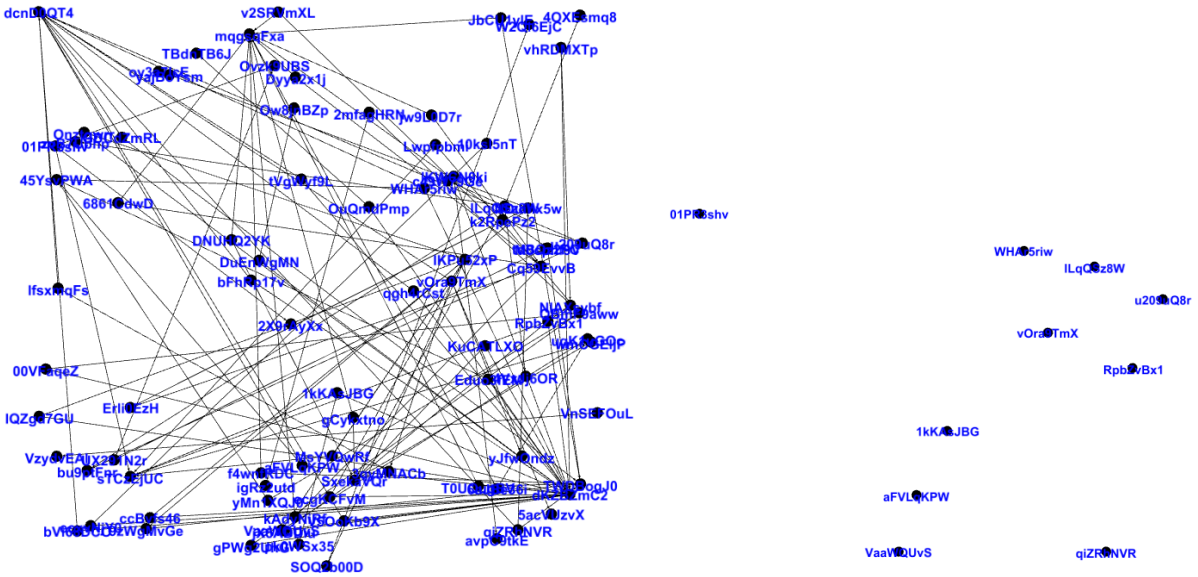
Figure 20. *Tokens* frequency (size 2)

As you can see from Figure 19, the most common words are *time*, *list*, *us*, *get*, *price and free*. All these tokens have the support of about 2, so these words are found in about 20% of the descriptions. Among the most frequent tokens found, *MDMA* (3,4-methylenedioxymethamphetamine drug) and refund deserve attention, which points out how sellers want to promote their products by seeking consumer trust.

Figure 20 shows the frequencies of sets with a size greater than 1. From this table, we highlight the line that we have marked in blue. The combination of 'us' and 'pgp' tokens is interesting as it manifests how anyone on the dark web pays attention to privacy and security. In fact, pgp turns out to be the acronym of 'Pretty Good Privacy' which it's actually an incredibly powerful asymmetric encryption technology that has been protecting all sorts of sensitive communication.

2.7 ANALYSIS OF VENDOR SHIPPING LOCATIONS

In this section, I want to deepen the aspect of shipments by drug sellers. The main objective is to check if a seller ships from more than one country. The RapidMiner subprocess has been named *vendor shipping locations*. In this subprocess a selection of the affected attributes is initially carried out, all those values containing 'worldwide' are removed from the Ships_from attribute to obtain more precise results and, through some join and grouping operations, a CSV file is generated which is then imported to the Gephi software. The CSV is represented in the form of a directed graph as can be seen in Figures 21/22.



Figures 21/22. Vendors shipping locations graphs

On the Gephi workspace, it is possible to perform many different types of queries, in this case, the query concerns the out-degree range filter. By setting a value for example 3, we can see which and how many vertices, ie vendors, have 3 or more different shipping locations. The node which has the highest value of out-degree is IlqQSz8W, with six different shipping locations.

How to interpret these results? There is no exact solution, for instance, the seller who ships from six different countries, ships from Argentina, India, Serbia, Slovenia, Pakistan and Thailand. The causes may be different, but in this situation, we can only suppose. For example, the seller might have multiple locations around the world and ship from countries where transportation is most convenient or efficient. On the other hand, it could also be more than one reseller working under one name. It must also be considered that not all countries can ship products to some countries and that shipping from one country can have higher prices than another due to the taxes that are applied. However, having more shipping locations makes your market much more anonymous.

CONCLUSIONS

At the end of the analysis of this database, we can briefly sum up the main results and draw some final considerations.

Darknet markets have invaded the particular world of the internet with crime, and have established themselves as the most efficient black markets globally for many illegal products such as firearms, or in this case drugs. However, this illegal trading hides several pitfalls and the sellers, as we have seen, do not guarantee absolute efficiency or reliability. There are many techniques used by sellers to make money in this market and to organize scams.

Starting from a database containing raw data, we were able to draw results that help to warn of some anomalies within the activities of the darknet market.

Considering the whole process, general cleaning of the data was carried out, collecting the generally most influential information and discarding the less important ones for investigations. Subsequently, each attribute was analyzed in more detail, examining the values and their distribution. The most common drugs found are cannabis and hashish, while the most frequent shipping country is the USA.

The process of extracting data using regular expressions was crucial in obtaining new variables with specific observations. The email extraction finds out to be relevant in the process of being warned of anomalies. The Gephi graph enabled the evaluation of potential cybercrime networks

inside the industry. This was made feasible by the ability to visually represent the relationships between certain emails and vendors. However, it must be noted that 3 distinct sellers used more than a single email.

In contrast, sentiment analysis has been the most effective method for identifying potential instances of fraud within the dataset. It was feasible to discern which merchants had earned positive reviews and which had received negative ones based on customer feedback. It can be noticed that negative or positive evaluations are correlated with the seller and the sort of drug he has sold. In instances when the products never reached the customer or did not offer the desired results, negative feedback was released. This has provided evidence of the occurrence of fraud instances. However, the majority of user reviews have been positive, indicating that this database does not include a large number of vendors who have likely participated in defrauding the purchasers.

In section 2.7, it would be plausible to assume that they had made sales in various regions of the globe (thus creating a network with other vendors). The graph illustrating this evaluation demonstrates that there have been instances in which sellers form a network by selling their products from many locations. This technique should be expanded to identify transnational drug traffickers with many bases of operations worldwide.

Many hidden relationships were discovered through the creation of association rules, which highlighted in particular the provenance of different drugs. Shipping information is of high importance when dealing with darknet transactions, identifying the origin of the seller's country. Through the analysis of the database of the Alphabay darknet market, it has been made an attempt to depict a more exhaustive and efficient picture of the environment that surrounds this anonymous world. This work has the main role of conducting a preliminary data investigation to assess whether there are anomalies in these markets of illicit products. Gephi and especially RapidMiner are optimal software which suits perfectly this work, making it easier and more efficient to use than other coding software.

This work and in particular this data analysis cannot be sufficient at all to conduct a complete investigation, but with this study, it has been shown how the new data science techniques prove to be excellent, fast and efficient to carry out preliminary investigations and to discover many insights.

BIBLIOGRAPHY

- [1] B. Hawkins, "Under The Ocean of the Internet - The Deep Web," 15 5 2016. [Online]. Available: https://www.sans.org/readingroom/whitepapers/covert/ocean-internet-deep-web-37012.
- J. T. P. Surbhi K. Solanki, "A Survey on Association Rule Mining," in 2015 Fifth International Conference on Advanced Computing & Communication Technologies, 2015.
- J. I. N. S. Vijayarani, "Preprocessing Techniques for Text Mining - An Overview," International Journal of Computer Science & Communication Networks, vol. 5, no. 1, pp. 7-16, 2015.
- L. H. P. P. C. Pritam C. Gaigole, "Preprocessing Techniques in Text Categorization," in National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2013), 2013.
- D. R. M. M. L. S. Q. R. Julian Broséus, "A Geographical Analysis of Trafficking on a Popular Darknet Market," Forensic science international, vol. 277, pp. 88-102, 2017.
- I. Ladegaard, "Crime Displacement in Digital Drug Markets," International Journal of Drug Policy, vol. 63, p. 113–121, 2019.
- F. C. B. B. Paolo Spagnoletti, "An Investigation on the Generative Mechanisms of Dark Net Markets," in Proceedings of the 13th Pre-ICIS Workshop on Information Security and Privacy, 2018.
- D. L. N. A. S. A. M. P. M. S. Ben R. Lane, "The Dark Side Of The Net: Event Analysis Of Systemic Teamwork (East) Applied To Illicit Trading On A Darknet Market," in Proceedings of the Human Factors and Ergonomics Society 2018 Annual Meeting, 2018.
- M. S. L. S. W. L. Andres Baravalle, "Mining the Dark Web: Drugs and Fake Ids," in Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference, Barcelona, Spain, 2016.
- Martin Erwig and Rahul Gopinath. “Explanations for Regular Expressions”. In: Fundamental Approaches to Software Engineering. Ed. by Juan de Lara and Andrea Zisman. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 394–408. ISBN: 978-3-642- 28872-2.
- Zsolt Nagy. Regex Quick Syntax Reference: Understanding and Using Regular Expressions. Apress, 2018.
- Xuan Zhang, ” A Framework for Dark Web Threat Intelligence Analysis”. International Journal of Digital Crime and Forensics, 2018, 10 (4)
- W. Park, "A Study on Analytical Visualization of Deep Web," 2020 22nd International Conference on Advanced Communication Technology (ICACT), Phoenix Park, PyeongChang, Korea (South), 2020, pp. 81-83, doi: 10.23919/ICACT48636.2020.9061283.
-] H. Alnabulsi and R. Islam, "Identification of Illegal Forum Activities Inside the Dark Net," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, Australia, 2018, pp. 22-29, doi: 10.1109/iCMLDE.2018.00015.