CONTENTS

LIST OF ABBREVIATIONS

MCCs - Merchant Category Codes

ATM - Automated Teller Machine

MAPE - Mean Absolute Percentage Error

EDA - Exploratory Data Analysis

BDIAR - Beliefs-Desires-Intentions-Actions-Reactions

GAN - Generative Adversarial Network

t-SNE - distributed Stochastic Neighbor Embedding

MaxS - Maximal Similarity Method

AI - Artificial Intelligence

ML - Machine Learning

TERMS AND DEFINITIONS

V_1 - First Variation: stands for spent amount differences between 2020 and 2019

V_2 - Second Variation: stands for spent amount differences between 2022 and 2021

INTRODUCTION

Banking is one of the crucial industries where a tremendous amount of valuable data is generated ceaselessly. Consequently, in terms of transactional banking data, the banking industry has a substantial role in a country's economy. Because, analyzing banking transactional data, it helps to extract significant and useful knowledge from it. Moreover, due to digitalization in every sphere of the economy, data-driven techniques and artificial intelligence models make it easier to get deep insights and behavioral comprehension with much less human labor.

Hence, our goal in this exploration will be focused on identifying the presence of the agitational segments of the banking transactional data and consequently revealing the bank card holders' ways of behaving strategies. With the help of machine learning technologies, it is now within reach to challenge bulky amounts of data to examine each agent closer and consequently investigate each agent's attitude during the given timeline. In this study, for clusterization purposes, an unsupervised machine learning algorithm is used to successfully divide customers into a collection of clusters and generate underlying analyses and inferences based on them.

And, the main reason for our involvement in this study is to provide valuable insights to the business, and allowing it to manage the emerging situations effectively.

As during periods of instability, businesses face numerous challenges but also have opportunities to thrive/flourish, in case, if they adapt effectively. Here's what businesses typically require to manage instability:

- **Being flexible.** If something's not working, change it. That might mean improving your product, rethinking your strategy, or trying new ways to reach your customers;
- **Leading with clarity.** People look to leadership in tough times. Be honest, make decisions confidently, and make sure your team knows what's going on and where you're headed;
- **Watching the money.** During instability, managing costs is key. Keep an eye on spending, protect your cash flow, and make sure you have a backup plan if things get tight;

- **Thinking ahead.** What could go wrong? Plan for it. Having a few “what-if” scenarios can help you respond faster when things shift;
- **Keep improving.** Don’t just survive i.e. look for ways to grow. That might mean improving your current product, trying something totally new, or using technology in smarter ways;
- **Talking to the people.** Regular updates i.e. whether it’s to your team, customers, or investors i.e. go a long way. Be open, honest, and human;
- **Focusing on the customers.** What people need can change fast. Stay close to your customers, listen to them, and adjust what you offer based on what matters to them right now;
- **Going digital.** Whether it’s selling online, using data to make better decisions, or supporting remote work i.e. digital tools help you stay connected and responsive.
- **Strengthening the supply chain.** Relying on one supplier is risky. Build in options, go local where you can, and be ready to switch if needed;
- **Supporting the team.** Stress levels go up during uncertain times. Make sure your employees feel safe, heard, and supported. It keeps morale (and productivity) up;
- **Working together.** Partnerships can help you share risky states, try new things, or tap into resources you wouldn’t have on your own. Don’t go it alone if you don’t have to.

Thus, we can claim that businesses need to focus on flexibility, strategic thinking, and maintaining a strong connection with customers and employees during periods of instability. Adaptation and innovation often become key differentiators in navigating and managing uncertain times. Therefore, we try to offer some kind of comprehensive insights for those business needs.

In financial institutions, a massive volume of data is being generated every day, such as customer account information, transaction information, and all financially interrelated information. In this case, somehow we need to handle this huge amount of data to investigate it and attempt to take out from it comprehensive insight and understandable definite forms of information so that we could be able to discover fundamental trends and, even if they exist, relationships that are not obvious or easy to find out. Shortly, we identify them as hidden patterns, which we can also refer to as hidden knowledge.

When we try to find unusual or strange things in the transactional data, what we are really doing is looking for anomalies. These are patterns that don’t follow the normal behavior. In other words, we want to spot anything that doesn’t look like the usual spending habits of customers. This helps us find the hidden patterns, the things that might not be obvious at first glance but can tell us something important about what’s going on.

Finding anomalies is not just something small or extra. It is actually a very important task in data analysis today. In fact, in modern science and business, anomaly detection is a key step in many projects. This is especially true when we work with transactional data, like people’s purchases or money withdrawals. Anomalies can show things like changes in customer behavior, mistakes in data, or even fraud.

Because of this, anomaly detection becomes a top priority in many data-related tasks. It helps people and businesses make better decisions by showing them what needs more attention. Whether we are trying to improve customer service, find system errors, or just better understand the data, catching these strange or irregular patterns is very useful and often necessary.

Additionally, the discovery of anomalies is the essential part/segment of any real dataset, as it allows for making more accurate inferences and data-driven decisions. Essentially, finding such anomalies in datasets helps financial institutions optimize their workflow and improve overall efficiency and profitability [1].

We also need to mention that due to the rapid advancement in electronic commerce technology, the use of banking services has dramatically increased. The increasing popularity of banking services, especially as a payment mode for both online and regular purchases, has caused enormous diversification of transaction types. Based on these different types of transactions, investigations can be conducted to reveal customers' overall tactics and intentions, whether during regular or irregular periods; that is, we can uncover their behaviors and strategies during times of instability.

The financial institution is a field that performs use of statistical analysis and machine learning models for most of its operations. For instance, estimating the chances of failure or risk of investment, managing cash and other valuable assets, monitoring the cash flow, implementing new business strategies and policies, improving customer satisfaction ratio, building revenue, predicting transaction outcomes, etc.

In the same way, in this study, to achieve the above idea, we will attempt to somehow investigate the irregularity states of the given banking transactional data on debit cards by using modern scientific techniques and technologies. As mentioned, the overall prevalence of the customers' behavioral strategy is crucial for a company, especially for a financial institution, as it helps to enhance customer satisfaction, customer retention, and thus revenues and profits.

In general, business requires having an ability to adjust its customer-centric procedures in times of instability. Understanding customer behavioral strategy could be useful and efficient to reshape trade policy at all.

In this exploration, we took advantage of the K-Means clustering machine learning method to create a set of clusters. According to the obtained set of clusters, we will see how customers show their reactions in varying conditions, that is, when there is an irregular state of the dataset, especially when we will in advance take into account the pandemic period of 2020 (V_1) and the starting period of the special military operation in 2022 (V_2).

Generally, the main purpose of this exploration is to compare the (V_1) and (V_2) periods and to understand how customers behave during these periods of instability, which we refer to as the agitational segments. Here, V_1 represents the variation between 2019 and 2020, while V_2 denotes the variation between 2021 and 2022. Variation refers to the differences in spending amounts across four top categories between the corresponding years. That is, V_1 stands for spent amount differences in 2020 (in crisis) between 2019 (no crisis), and V_2 stands for spent amount differences in 2022 (in crisis) between 2021 (no crisis).

1 RELATIVE RESEARCHES

A literature review definitely provides many results and ideas on the analysis of banking transactional data, which was carried out by different ways of tactics and approaches. Many researchers have developed and implemented various analysis and forecasting models that are significant and remarkable.

Before all, we were searching for a while to collect the relevant literature for our study, so we have used multiple online literature search engines such as Google Scholar, ResearchGate, and Elsevier with the search term "Anomaly Detection." Thus, we shared some papers that we present in this study, which are dealing with anomaly identification in the observation domain. These papers gave us a huge perspicacity with reference to our study issues, and we found answers from them to our study questions.

The authors of [2] describe a possible modification of the components of the Beliefs-Desires-Intentions-Actions-Reactions agent model (BDIAR agent model), they took into account the influence of values on consumption dynamics. A method for taking into account values in the BDIAR architecture during modeling is proposed: decision-making by an agent, where the levels of architecture correspond to values, preference functions, and dynamics of the agent's state. Analyzed pseudonymized transaction data for debit cards of clients of the partner bank "St. Petersburg" separately for 2017–2019 and 2020. Demonstrated subjectivity of the influence of the environment in a crisis on the dynamics of changes in values and needs for different consumer groups, taking into account the clustering of their behavior types. Similarly, we attempt in our work also to create some basic values as top categories to be able to perform clusterization processes, and based on obtained clusters, we make our examinations and pattern recognition as well.

The authors of [3] establish a noticeable relationship between measures of realized and intrinsic predictability in both generated and real-world time series data (with the correlation coefficient being statistically significant at a 5% significance level). The discovered correlation in that research holds significant value for tasks related to evaluating time series complexity and their potential to be accurately predicted. We also try to determine some correlations in our research to make delicate predictions.

The authors of [4] describe how they try to recognize some behavioral patterns in sequences of transactional data of bank card holders in the pandemic situation of 2020 and apply them to the forecast of unsteadiness of February-March 2022 caused by customers’ anxious expectations. Their observations and definitions are almost interrelated with what we are going to perform as well.

The authors of [5] performed anomaly detection using a Wasserstein Generative Adversarial Network on a time series dataset. Generative Adversarial Networks are a popular technique to generate data from an original dataset. The main reason why they chose Wasserstein GAN is that Wasserstein GAN does not collapse, contrarily to the classical GAN, which needs to be heavily tuned to avoid the problem. They indeed investigated abnormal data in a very deep way with several techniques. This research helped me to profoundly consider how it is possible to go deeper when we talk about anomaly detection of a dataset.

The authors of [6] studied two datasets, bank customer data and German credit data, to make a classification to discover the hidden patterns. The authors identified customers' behavior and retained them because they consider customer retention an effective method for the growth of the banks. They admit that to retain customers, first it is necessary to identify which customers are active and inactive. For classification purposes, they used a supervised artificial neural network.

The author of [7] focused on developing a new model and a corresponding forecasting method related to the class of autoregressive models and eliminating the main drawback of this class of model. The new model and the corresponding method should have a high speed of calculating forecast values and be comparable with other models in the accuracy of forecasting various time series. And, they developed a new model named the forecasting method of maximal similarity (MaxS). According to the authors of [8], more versatile information concerning the MaxS method can be found. In this work, we distinguished this methodology of investigation of predictability, as it in every way answers to our requirements, which are based on discovering similarities of patterns, that is, customers' ways of behaving.

In [9], the authors sorted out an all-inclusive review of the up-to-date research on anomaly detection techniques. They sought to serve as an extensive and comprehensive review of machine and deep learning anomaly detection techniques throughout the foregoing three years, 2019-2021. Particularly, they discussed both machine learning and deep learning anomaly detection applications, performance measurements, and anomaly detection classification. Based on their research, we perceived and obtained much more wisdom regarding the differences between machine learning techniques.

The authors of [10] proposed a model of Naive Bayes and SVM (Support Vector Machine) to detect anomalies and an ensemble approach to solve the weaknesses and to remove the poor detection results. Actually, they proposed a hybrid machine learning technique that has two layers for classification.

In [11], the authors made a proposal of an intelligent data-centric evaluation framework that can identify high-quality data and improve the performance of an ML system. The proposed framework combines the curation of quality measurements and unsupervised learning to distinguish high-quality and low-quality data. We especially paid attention to how they implemented a partitional algorithm in their study.

In the [12] research paper, the authors presented an in-depth analysis of several machine-learning approaches for detecting abnormalities in Indian electricity consumption data. Specifically, the performance of three widely used machine learning algorithms, namely eXtreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), has been evaluated by them. The efficacy of each technique has been assessed using performance indicators, including mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2). The findings of this study provided us with valuable insights into the most effective machine-learning methods for identifying anomalies.

The authors of [13] provided a structured map of anomaly detection based on quantum machine learning. They have grouped existing algorithms according to the different learning methods, namely quantum supervised, quantum unsupervised, and quantum reinforcement learning, respectively. They also provided a systematic and compact understanding of the techniques belonging to each category. From their study, we got a huge new conception that emphasizes quantum machine learning.

In the [14] research paper, the authors investigated empirically the relationship between six data quality dimensions and the performance of 19 popular machine learning algorithms covering the tasks of classification, regression, and clustering to explain their performance in terms of data quality. Their experiments distinguish three scenarios based on the AI pipeline steps that were fed with polluted data: polluted training data, test data, or both. Their results highlighted how different data quality issues affect machine learning performance in distinct ways, offering valuable insights for data scientists when developing machine learning pipelines in the presence of data quality issues. Their study inspired us to implement it in the same way in the future as they did.

In the [15] research paper, the authors addressed market, operational, liquidity, and other risk types, with the objective to examine how machine learning algorithms predict, assess, and mitigate the risks and identify their advantages and challenges. Their findings offer a concise overview of current machine learning applications for multiple risk types in banking, identifying research gaps, highlighting opportunities and challenges, and providing actionable directions.

The authors of [31] research paper, examined the main theories that take into account the features of decision-making in crisis situations, on the basis of which the architecture of the model of an agent for making economic decisions is built, taking into account the values and needs of the agent. They claim also that analysis of transaction data demonstrates a change in the priorities of payment behavior as a result of the development of a crisis situation, and the classification of payment categories by groups of survival, socialization and self-development values demonstrates the increasing role of survival values.

2 MATERIALS AND METHODS

2.1 Dataset Description and Preprocessing

We need to present some information regarding data from the given transactional data. The total number of customers who hold debit cards is equal to 10,000, and the total number of processed merchant category codes is equal to 115 from January 1, 2018, until July 31, 2022. Every transaction is described by a client identifier, their debit card identifier, the date of a transaction, the amount spent, and the merchant category code. The total number of all unprocessed transactions during that period is 19,262,668 observations.

Exploratory Data Analysis (EDA) is an important embryonic phase for putting on display the deeper insights within a dataset i.e. revealing the underlying structure and key features of a dataset. Through EDA, we can identify significant variables, relationships among the variables, incompleteness, errors, and patterns in a dataset. Also, we can check for data anomalies, redundancies, dependencies, and outliers present in a dataset. Essentially, EDA is a powerful preliminary process for investigating a dataset in all directions [16]. Note that EDA is performed before data modeling. EDA, according to John Tukey, is

"Exploratory data analysis is an attitude, a state of flexibility, and a willingness to look for those things that we believe are not there, as well as those we believe to be there."

- *John Tukey*, *1977*

Additionally, we have to take into account the scalability of the machine-learning models. Thus, to improve the performance of machine learning models, transformations of the values are often applied. As we know, almost every dataset possesses inconsistent data values, which lower the machine learning model's performance. Machine learning, according to Arthur Samuel, is

"Field of study that gives computers the ability to learn [from data] without being explicitly programmed."

- Arthur Samuel, 1959

Therefore, to significantly impact the model's performance, to enhance it, it should implement data processing steps that can improve the accuracy, reliability, and quality of a dataset, making it more durable. Typically, transformations are used to make the relationships between variables more linear. In other cases, transformations are performed to make distributions closer to normal, or at least more symmetric. These transformations can include taking logarithms, exponential transformations, and power transformations.

Thus, the purpose of using any feature engineering method, especially for numerical data, is that it plays a critical role in improving the performance of machine learning models since many machine learning algorithms, such as linear regression, neural networks, and some clustering methods, assume that the data follows a normal distribution.

Hereby, as our dataset consists of numerical values, then before the dataset is forecasted, a numeric feature column must be re-scaled. Rescaling a numeric feature is extremely important for machine learning models where raw data is transformed into more meaningful features that help the model better understand the relationships in the data. There are many possible scaling methods, which are like Normalization, Polynomial Features, FunctionTransformer, KBinsDiscretizer, Logarithmic Transform, PowerTransformer (Box-Cox), QuantileTransformer, and PCA.

By the way, to be able to explore and interpret diverse characteristics and mosaics in the transactional data, we use Python libraries such as:

- Pandas, for data analysis;
- SciPy and NumPy for descriptive statistics and simple modeling;
- Seaborn for drawing attractive and informative statistical graphics;
- Sklearn for predictability and metric measuring.

Thus, after performing preliminary processing tasks such as exploratory data analysis, transformation, and normalization on our dataset, we have refined the transactional data, and now we possess 8,059,496 observations for the further tasks and operations (figure 2).

	Client	Card	Amount	MCC	Category	Top Category	DateTime	Amount_L10
0	390989	3048567	878.0000	5814	food	survival	2018-01-01	2.9435
3	475694	2884069	2564.0000	5941	fun	self_realization	2018-01-01	3.4089
5	2406107	3465449	909.3800	5411	food	survival	2018-01-01	2.9587
7	2714314	3732417	1071.0000	5411	food	survival	2018-01-01	3.0298
10	1356203	3823801	500.5000	5541	travel	socialization	2018-01-01	2.6994
11	316576	3008238	1015.3000	5411	food	survival	2018-01-01	3.0066
15	1382058	2415606	715.0000	5411	food	survival	2018-01-01	2.8543
16	2889080	2337144	779.5000	5411	food	survival	2018-01-01	2.8918
17	2454592	3463684	519.0300	5411	food	survival	2018-01-01	2.7152
18	2866255	3737744	2900.0000	6011	money	money	2018-01-01	3.4624

Figure 2 – Conversion of the raw dataset into a clean dataset before feeding it to a model, processed data

Coming to the top categories of **survival**, **socialization**, **self-realization**, and **money**, they have been formed by aggregating categories, and the categories themselves have been formed by aggregating over 300 merchant category codes (MCCs). This aggregation process of merchant category codes (MCCs) into categories and categories into top categories (as basic values) is described in detail in [2]. This approach is accepted for the purpose to be able to make assessment the customers' ways of behaving strategies. Each top category is a set of categories, as below:

- **Survival**: ‘food,’ ‘outfit,’ ‘dwelling,’ ‘health’;
- **Socialization**: ‘remote,’ ‘telecom,’ ‘travel,’ ‘nonfood’;
- **Self-realization**: ‘beauty,’ ‘fun,’ ‘kids,’ ‘charity’;
- **Money**, which means withdrawing cash from an ATM.

In comparison, we will consider their varying amounts as their ways of behaving, that is, their behavioral strategies. Our concentration will be focused on the customers who altered their strategies throughout the agitational segments [32] and those who remained consistent despite the agitational segments (figure 1).

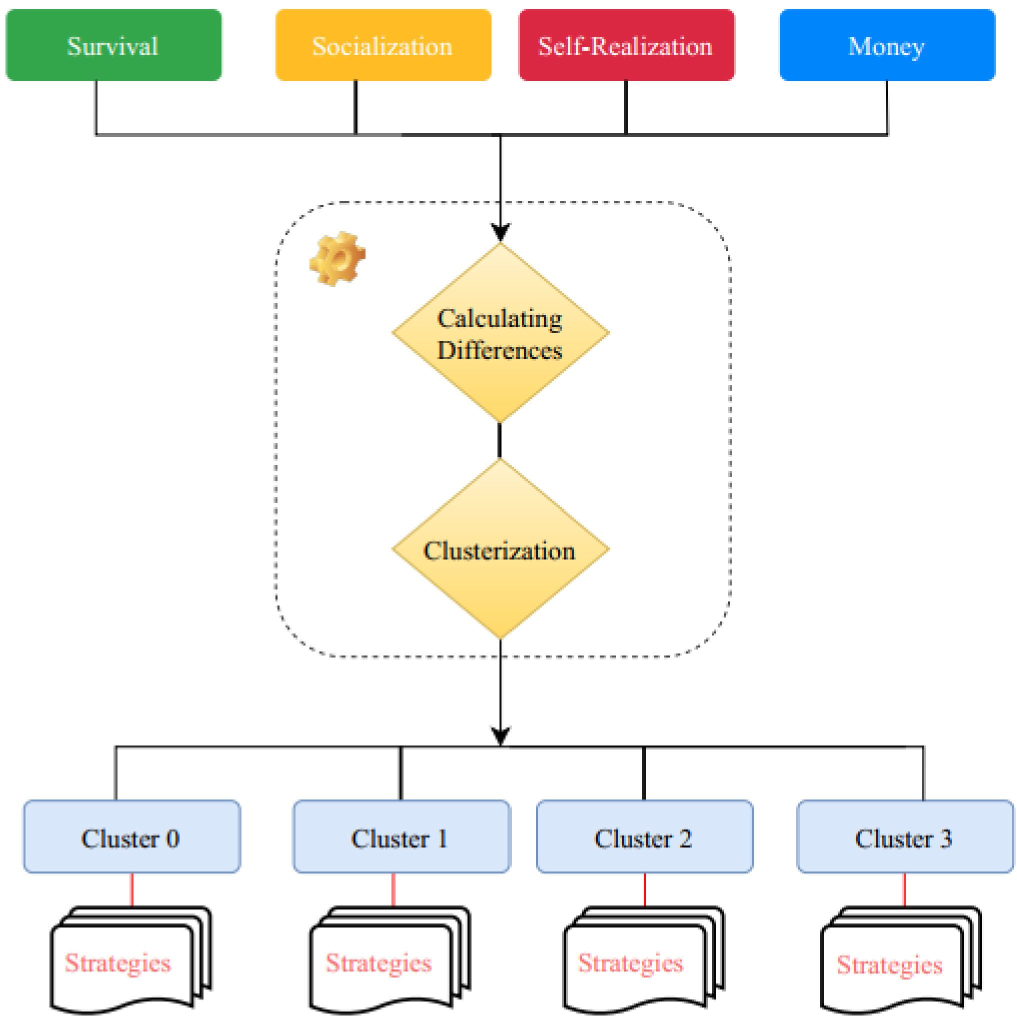

Figure 1 – Exploration scheme

Additionally, it would be preferable if we described this approach in a versatile manner. According to the authors of [31] action map of creating basic values has been developed by using BDIAR agent model and Maslow's hierarchy of needs pyramid. Below table shows action map which defines a mapping procedure from MCCs to the basic values as top categories [36].

Table 1 – Action map for three basic values as top categories

MCCs	**Consumer Interest Group**	**Value**
1	2	3
5411: 'Grocery shops, supermarkets'	Food	Survival
5814: 'Fast Food'		
5812: 'Places public food, restaurants'		
5462: 'Bakeries', 5441: 'Confectionery'		

Continuation of table 1

1	2	3
5422: 'Sale fresh and ice cream meat'	Food	Survival
5451: 'Sale dairy products retail '		
5691: 'Shops male And female clothes'	Cloth	
5651: 'Cloth For all families', 5661: 'Shoes shops'		
5949: 'Shops fabrics, threads, handicrafts, sewing'		
5211: 'Forest- And building material'	Housing	
4900: 'Housing and communal services services'		
5712: 'Equipment, furniture And household accessories (except electrical equipment)'		
5261: 'Garden accessories (including lawn care) retail'		
5714: 'Fabrics, upholstery material, urtains, blinds'		
5912: 'Pharmacies', 8043: 'Optics, optical goods And glasses'	Health	
8062: 'Hospitals', 8021: 'Dentists, orthodontists'		
8071: 'Dental And medical laboratories'		
5331: 'Universal shops'	Manufactured goods	Socialization
5999: 'Various shops And special retail stores'		
5311: 'Department Stores'		
5200: 'Goods for home'		
5399: 'Various goods general purpose'		
5931: 'Second-hand shops, shops used goods, consignment shop'		
5948: 'Leather goods, travel accessories shops '		
4111: 'Passenger transportation - suburban And local commuter services including ferries'	Transport	
5541: 'Refuelling stations (With auxiliary with/without services)'		
4121: 'Limousines And Taxi'		
4131: 'Bus lines'		
7512: 'Car rental agencies - not elsewhere classified'		
4784: 'Paid roads And bridges'		
4112: 'Passenger railway transportation'		
5533: 'Auto parts and accessories'		
4511: 'Airlines, airlines - not elsewhere classified '		
5542: 'Automatic refueling stations'		
4814: 'Telecommunications services'	Information and communication	
4816: 'Computer networks, informational services'		
5732: 'Sale electronic equipment'		
4812: 'Telecommunications equipment, including sale of tele.'		
9402: 'Postal services - only state'		
4215: 'Services courier - By air And on earth, agency for the dispatch of goods'		
4899: 'Cable And other paid television services'		

Continuation of table 1

1	2	3
6011: 'Financial institutions - automatic cash withdrawal '	Finance	Self-realization
6536: 'Money transfers from card to card - enrollment (in the country)'		
6012: 'Financial institutions – trade and services'		
6538: 'Money translations from cards on map - write-off'		
6010: 'Financial institutions - issuance cash at the box office'		
9311: 'Tax payments'		
9222: 'Fines'		
6300: 'Services insurance companies'		
6540: 'Replenishment non-banking prepaid cards, accounts'		
5943: 'Shops office, school accessories, stationery'	Children & education	
8299: 'Educational services not elsewhere classified '		
8220: 'Colleges, universities, vocational schools and technical schools'		
5921: 'Shops With sale alcoholic drinks takeaway'	Entertainment and recreation	
5813: 'Bars, cocktail bars, discos, night clubs and taverns - places where alcoholic beverages are sold'		
5993: 'Tobacco shops'		
5942: 5192: 'Books, periodic editions and newspapers'		
5992: 'Floristry'		
5735: 'Shops sound recordings'		
7832: 'Cinemas', 7995: 'Gambling games'		
5941: 'Sporting goods'		
5947: 'Shops postcards, gifts, new products And souvenirs'		
7922: 'Theatre producers (except motion pictures), ticket agencies'		
5944: 'Watch, jewelry And silver products'		
7997: 'Clubs - country clubs, membership (rest, sports), golf courses'		
5193: 'Accessories For florists, nursery And flowers'		
7941: 'Athletic fields, commercial types sports, professional sports clubs, sports promoters'		
0742: 'Veterinary services'		
7991: 'Tourist attractions And exhibitions'		
7994: 'Clubs video games', 7841: 'Video rental'		
7221: 'Photo studios'		
4722: 'Tourist agencies And organizers excursions'		

Firstly, for the training purposes, it would be better to get some insights with reference to transactional data by making some statistical exercises. The aim of this experiment is to analyze of uni-variate random variables. In this experiment we will handle the non-parametric estimation of probability density function approach. To perform this approach we need to specify and determine a density estimation technique.

As usual, as well in this experiment we will implement the Kernel Density Estimation technique, or KDE for short, as our estimated density function [33, 38].

Four random variables (categories) has been selected from the transactional dataset, below exercises has been performed just for getting some insights with reference to the transactional data.

1. **Charity**
2. **Beauty**
3. **Money**
4. **Food**

In order to make some elementary assessments according to their numerical information. Actually these empirical assessments have been performed just for getting insights and capability with reference to the statistical measurements. Furthermore, we will examine the objective in a versatile manner using all necessary statistical measurements and inferences.

1. **Charity Category**

As indicated on the graph, the charity period was taken just between one month (January, 2020), and the frequencies of the collection of facts, here as observations with the range of their amounts are given. Histogram and KDE technique [34] are combined as well. Moreover, we can get a smooth distribution estimate using the kernel density estimate that Seaborn does with kde-plot function. Lastly, we are estimating correctness of fitted distributions using two statistical tests, such as Kolmogorov-Smirnov & Chi-Square, and there is a validation of our estimated parameters using QQ biplots (figure 3).

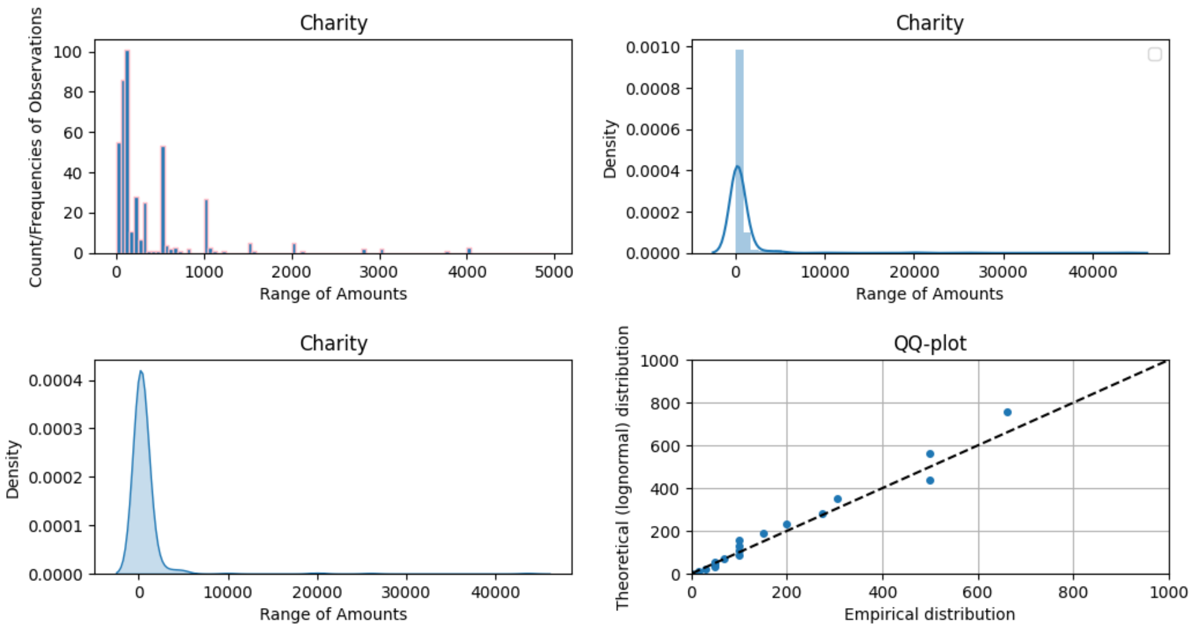

Figure 3 – Charity category's statistical measurements

Table 2 – Charity Category Analysis

Number of transactions for charity category	443
Total volume of charity	294632.52
Confidence intervals (25%, 50%, 75%)	[62.35], [93.30], [87.56]
K-S test result (statistic, p-value)	0.1272, 1.05e-06
C-S test result (statistic, p-value)	5409520.19, 0.0
Remarque	Tests show data is not normally distributed (skewed)

Average amount of spending for each corresponding month, that is from 2020-01 till up to 2021-06 of the charity category is shown below (figure 4).

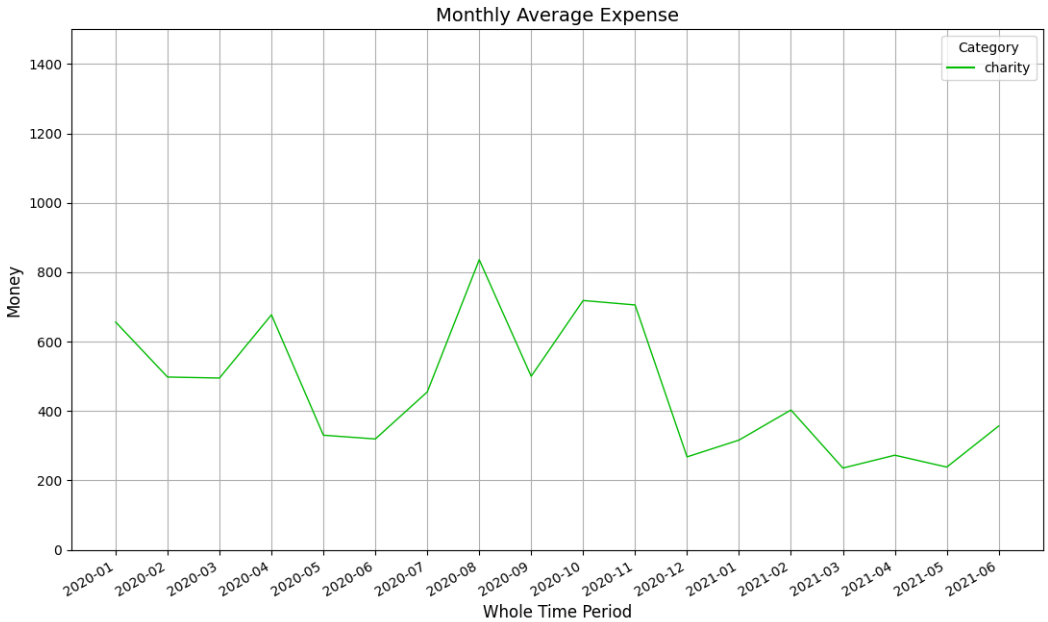

Figure 4 – Charity category's overall monthly average distribution

2. Beauty Category

As shown on the graph, the beauty period was taken just between one month, and the frequencies of the collection of facts, here as observations and the range of their amounts are given. Histogram and KDE technique is combined as well. Moreover, we can get a smooth distribution estimate using the kernel density estimate that Seaborn does with kdeplot function. Lastly, we are estimating correctness of fitted distributions using two statistical tests, such as Kolmogorov-Smirnov & Chi-Square, and there is a validation of our estimated parameters using QQ biplots (figure 5).

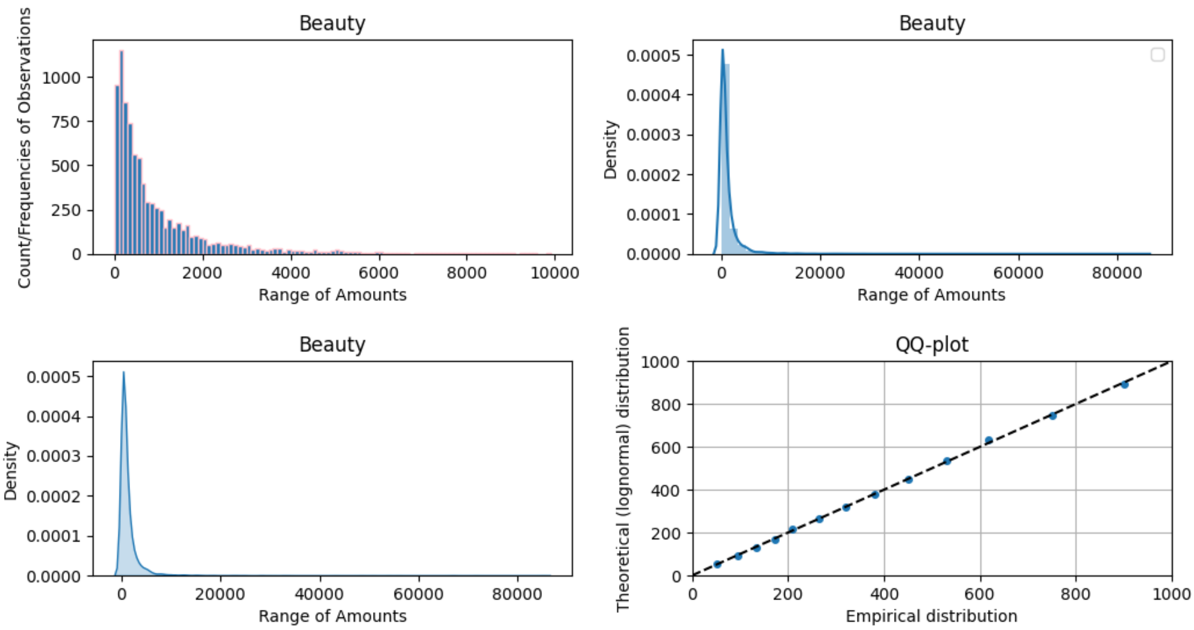

Figure 5 – Beauty category's statistical measurements

Table 3 – Beauty Category Analysis

Number of transactions for beauty category	8962
Total volume of beauty	11545726.13
Confidence intervals (25%, 50%, 75%)	[15.27], [17.39], [33.22]
K-S test result (statistic, p-value)	0.0108, 0.248
C-S test result (statistic, p-value)	60384706.09, 0.0
Remarque	K-S suggests normality; C-S suggests skewness

Average amount of spending for each corresponding month, that is from 2020-01 till up to 2021-06 of the beauty category is shown below (figure 6).

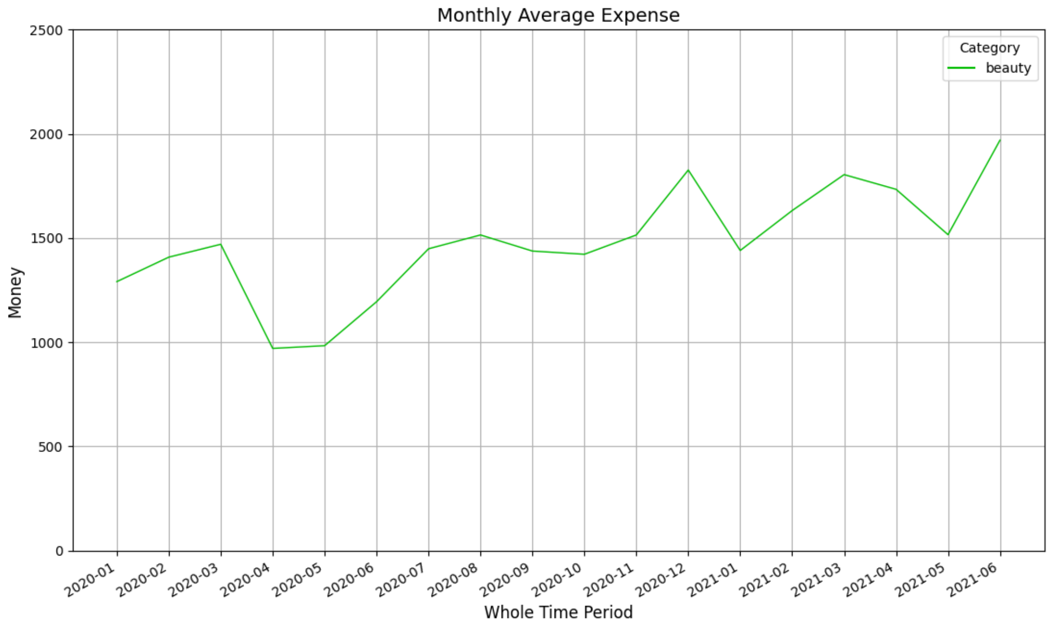

Figure 6 – Beauty category's overall monthly average distribution

3. Money Category

As shown on the graph, the money period was taken for all the time (specified), and the frequencies of the collection of facts, here as observations and the range of their amounts are given. Histogram and KDE technique are combined as well. Moreover, we can get a smooth distribution estimate using the kernel density estimate that Seaborn does with kdeplot function. Lastly, we are estimating correctness of fitted distributions using two statistical tests, such as Kolmogorov-Smirnov & Chi-Square, and there is a validation of our estimated parameters using QQ biplots (figure 7).

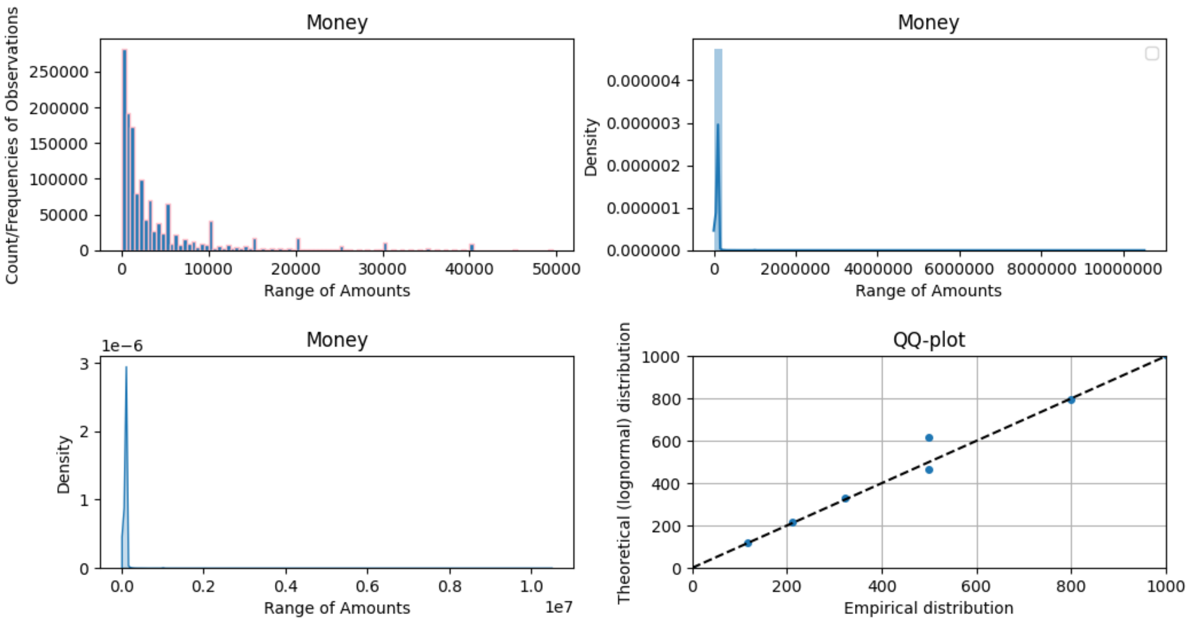

Figure 7 – Money category's statistical measurements

Table 4 – Money Category Analysis

Number of transactions for money category	1449215
Total volume of money	12810340590.02
Confidence intervals (25%, 50%, 75%)	[4.89], [6.06], [11.41]
K-S test result (statistic, p-value)	0.0589, 0.0
C-S test result (statistic, p-value)	228952372192.41, 0.0
Remarque	Both tests indicate non-normal, highly skewed data

Average amount of spending for each corresponding month, that is from 2020-01 till up to 2021-06 of the money category is shown below (figure 8).

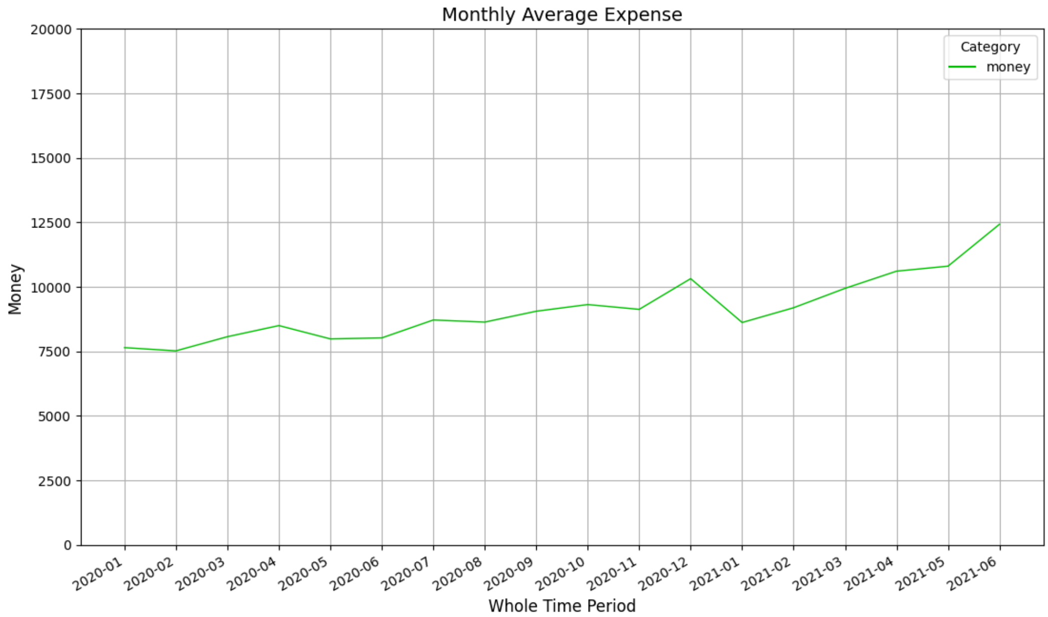

Figure 8 – Money category's overall monthly average distribution

4. **Food Category**

As shown on the graph, the food period was taken for all the time (specified), and the frequencies of the collection of facts, here as observations and the range of their amounts are given. Histogram and KDE technique is combined as well. Moreover, we can get a smooth distribution estimate using the kernel density estimate that Seaborn does with kdeplot function. Lastly, we are estimating correctness of fitted distributions using two statistical tests, such as Kolmogorov-Smirnov & Chi-Square, and there is a validation of our estimated parameters using QQ biplots (figure 9).

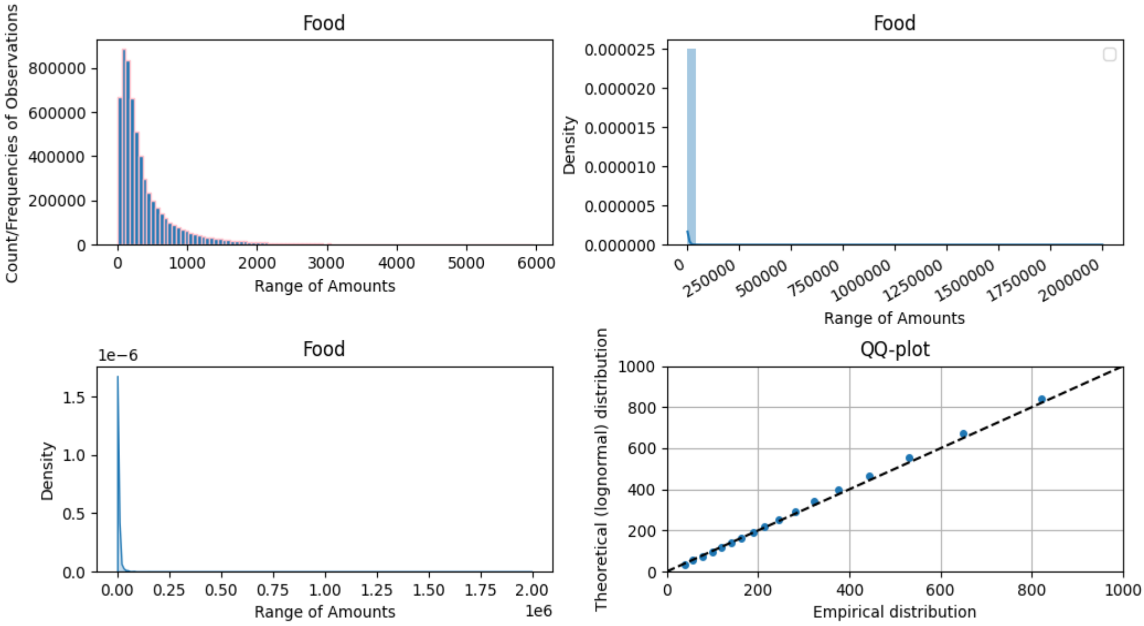

Figure 9 – Food category's statistical measurements

Table 5 – Food Category Analysis

Number of transactions for money category	6,223,113
Total volume of food	3,270,285,648.51
Confidence intervals (25%, 50%, 75%)	[0.14], [0.21], [0.56]
K-S test result (statistic, p-value)	0.0200, 0.0
C-S test result (statistic, p-value)	30,670,036,039.63, 0.0
Remarque	Data shows clear skewness; not normally distributed

Average amount of spending for each corresponding month, that is from 2020-01 till up to 2021-06 of the food category is shown below (figure 10).

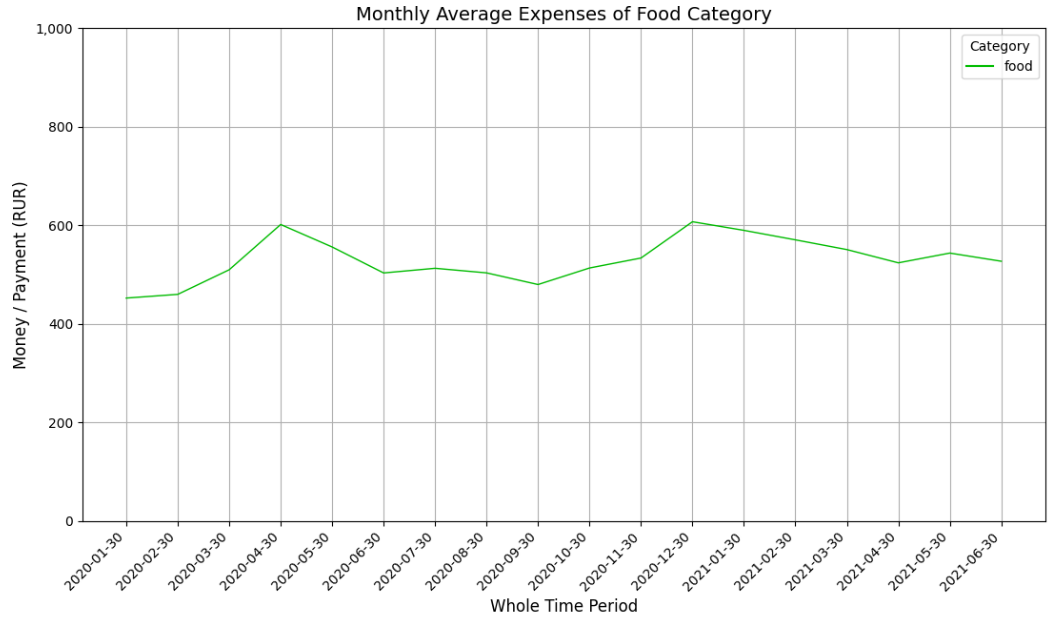

Figure 10 – Food category’s overall monthly average distribution

Initially, it will be reasonable to illustrate some graphical interpretation to get insights with reference to the given transactional data, as it helps us to be conscious of the content of the whole picture of the dataset [19]. The below graphical interpretation is monthly total expenditures of top categories (figure 11).

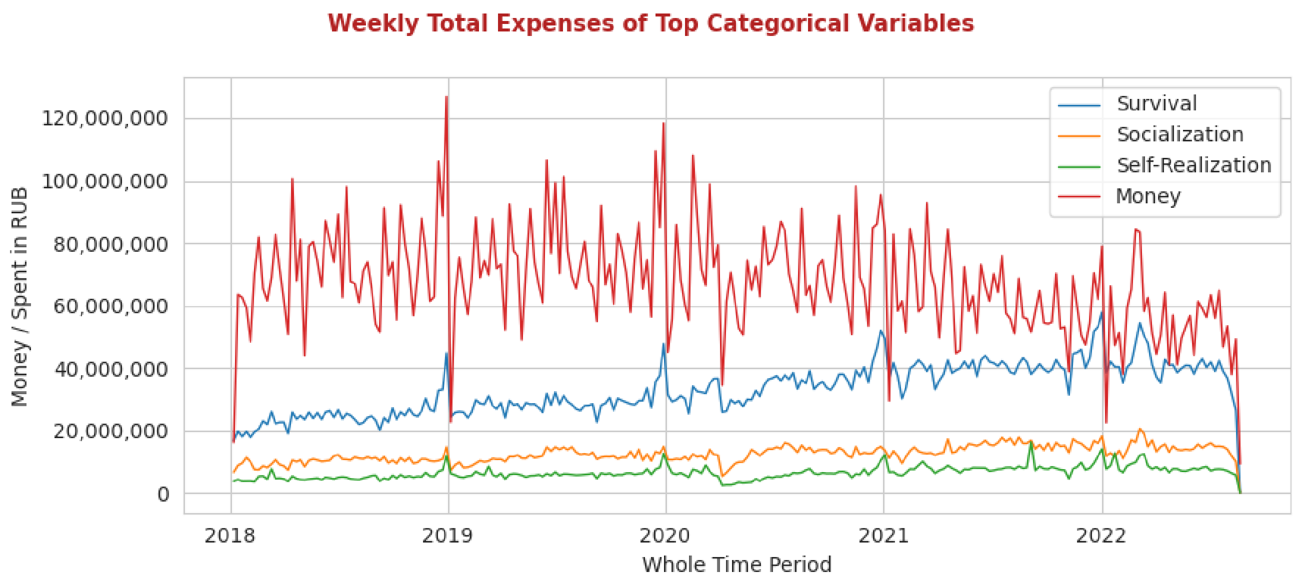

Figure 11 – Monthly total spending based on top categories

The below graphical interpretation is the total distribution of quantitative data of top categorical variables (figure 12).

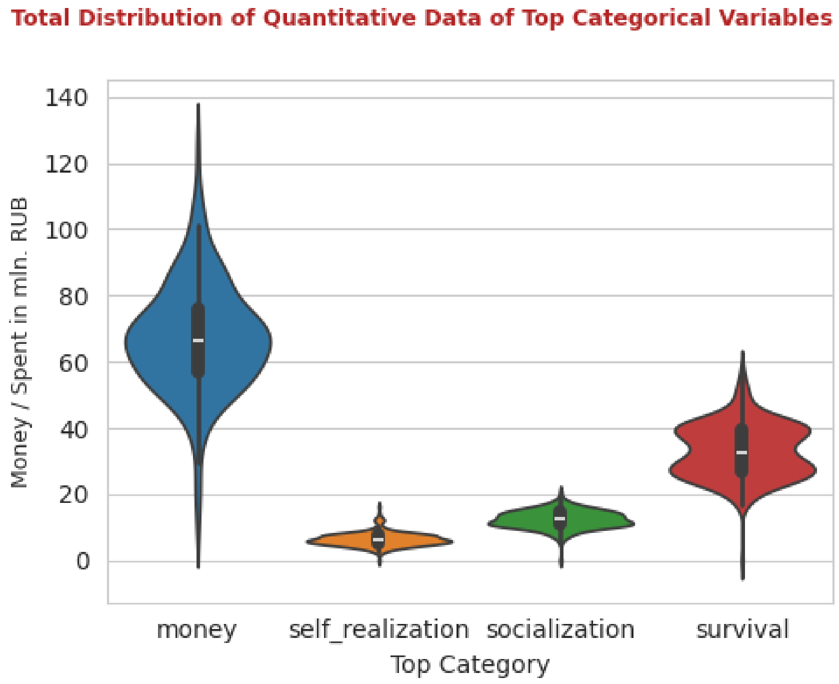

Figure 12 – Total distribution based on top categories

The below graphical interpretation is the average expenditures by customers of top categorical variables (figure 13).

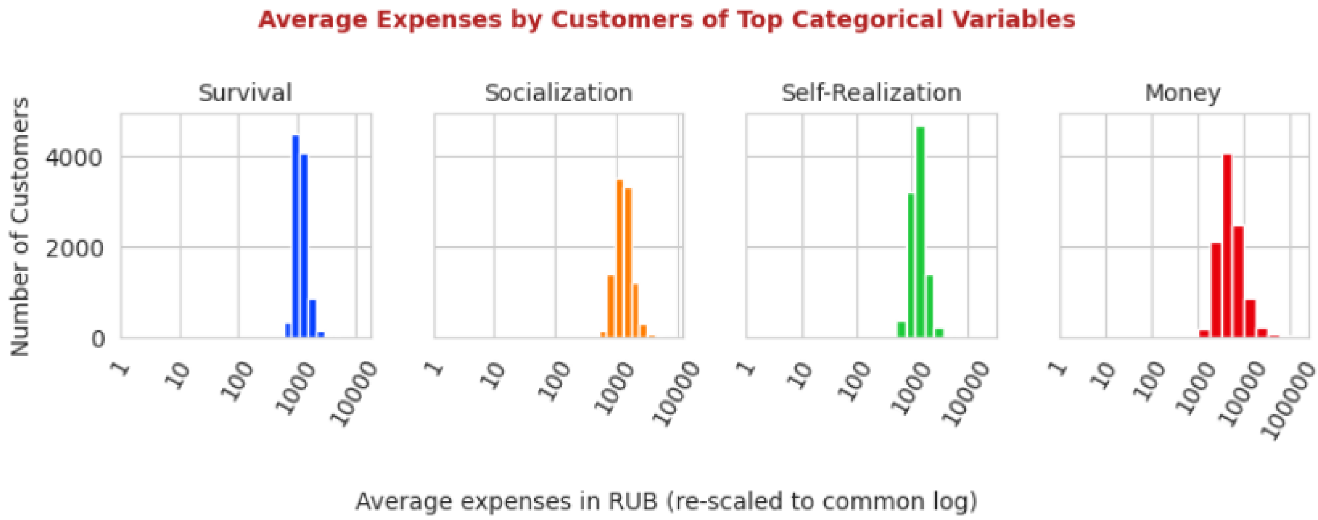

Figure 13 – Average spending by customers based on top categories

The below graphical interpretation is the portion from the transactional data that indicates weekly total expenditures of top categorical variables [17]. The taken portion segment is from March 01, 2019, till May 31, 2019 (figure 14).

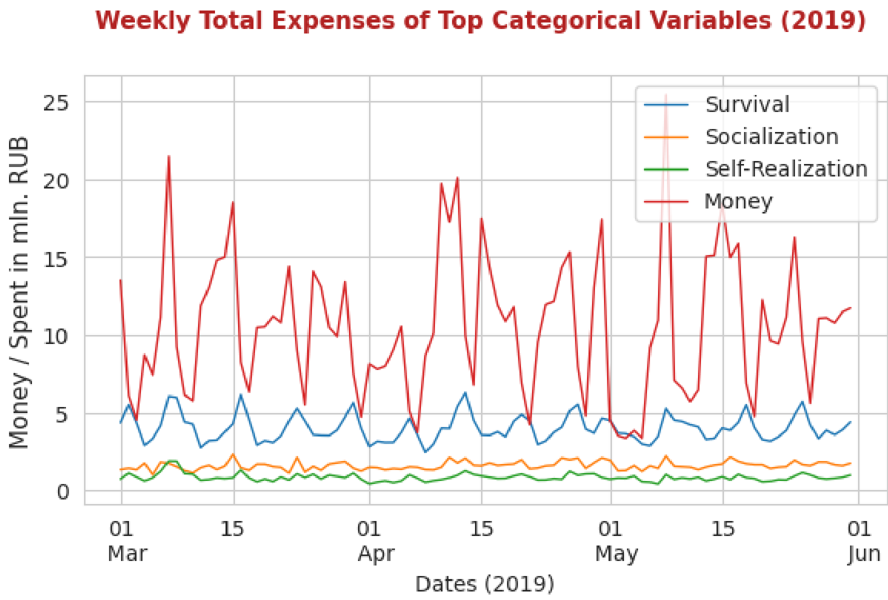

Figure 14 – Weekly spending in 2019 (no crisis)

The below graphical interpretation is the portion from the transactional data that indicates weekly total expenditures of top categorical variables and the correlation heatmap. The taken portion segment is from March 01, 2020, till May 31, 2020 (figure 15).

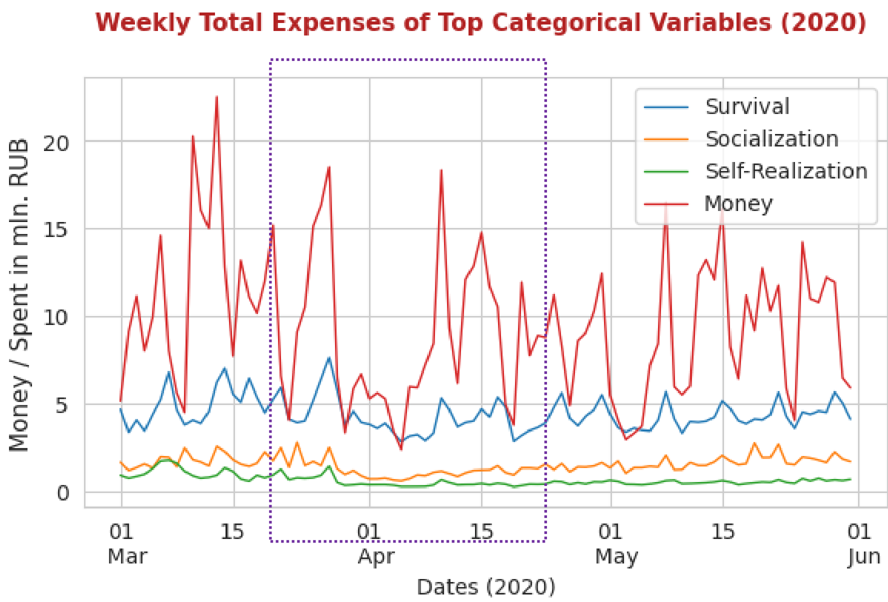

Figure 15 – Weekly spending in 2020 (crisis)

The below graphical interpretation is the portion from the transactional data that indicates weekly total expenditures of top categorical variables and the correlation heatmap. The taken portion segment is from March 01, 2021, till May 31, 2021 (figure 16).

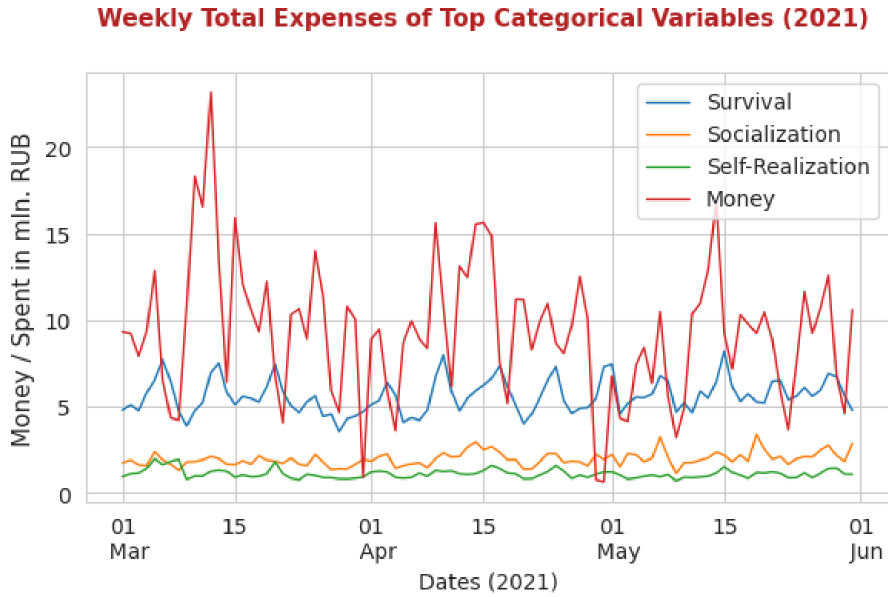

Figure 16– Weekly spending in 2021 (no crisis)

The below graphical interpretation is the portion from the transactional data that indicates weekly total expenditures of top categorical variables and the correlation heatmap. The taken portion segment is from March 01, 2022, till May 31, 2022 (figure 19).

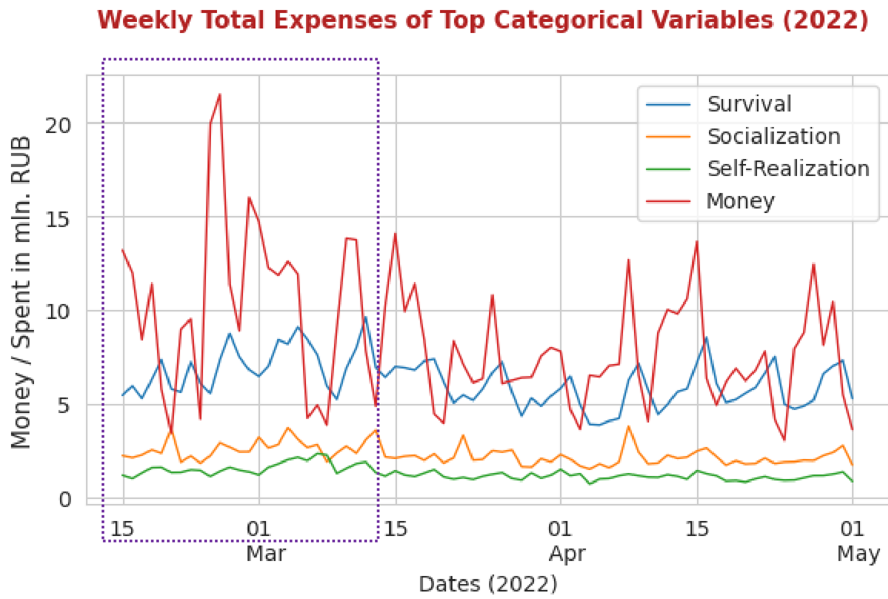

Figure 19 – Weekly spending in 2022 (crisis)

Based on the information we have gathered so far, including both the insights from our analysis and the results from the visual graphs, we can say that it is difficult to clearly identify any strong or sharp signs of abnormal behavior in the transactional data. Even when we look closely at the areas where we expected more movement or change (what we can call the agitational segments), there are no obvious irregular patterns that stand out. This means that, using the current methods and tools, we are not able to easily point to specific places in the data that show something unusual or unexpected.

This situation leads us to an important conclusion: finding anomalies in this kind of data is not a simple task. Anomalies are rare, unusual events that do not follow the usual pattern [37]. They can be signs of important behavior changes, errors, or even hidden risks. However, these kinds of hidden patterns often do not appear clearly when we only use basic graphs or general analysis. They may be buried deep in the data or masked by other regular activities. As a result, relying only on visual interpretation or simple techniques is not enough to detect them.

Because of this, it becomes necessary to develop and apply a more advanced and focused method for anomaly detection. This method should be specifically designed to search for small but meaningful differences in behavior or activity that are easy to miss with standard tools. It might involve using statistical models, machine learning algorithms, or other smart techniques that can go deeper into the data and find what the eye cannot see.

By doing this, by improving our approach to detecting unusual patterns, we will be better able to uncover hidden knowledge within the transactional data. This hidden knowledge could help us better understand customer behavior, predict future actions, or even improve decision-making in the system as a whole.

2.2 Cluster Analysis

In essence, our purpose to use cluster analysis method is to get know about customers' purchasing habits, and preferences in order to be able to group customers in a way that enhances personalization efforts. Hereby, by determining distinct segments, in fact, we are helping the business so that the business could adapt marketing strategies to meet each group's specific needs and behaviors [21].

Moreover in cluster analysis, machine learning models are being used for finding patterns within a dataset. Clustering (grouping) of observations based on their similarities is considered a fundamental searching method of learning. Therefore, clustering is being used mainly to identify the presence of instability and identify important features of the dataset.

Coming to unsupervised machine learning clustering, we can note that it is a form of clustering that is applied to unlabeled data. So, clustering methodology is a very powerful contemporary instrument ecpecially when we refer to pattern recognition. We will beneficiate from clustering methodology in our study due to its captivating capability [22].

Consequently, we can say the cluster analysis process is the very crucial and basic phase of data mining and a specific technique for statistical data analysis [35]. As mentioned, it can be used in pattern recognition, and we also take advantage of implementing it. Thus, cluster analysis helps us to recognize and identify distinct kinds of ways of behaving and things. Plus, it should be noted that clusters always have to manifest high internal homogeneity and high external heterogeneity.

Additionally, cluster analysis is in fact a powerful data mining technique which assists us to be capable to regroup the observations in terms of their characteristic indicators, that we can observe much better the whole picture of the dataset, and which helps us to make further assessments, inferences, and analyses. The primary goal of cluster analysis is to maximize the similarity of data points within a cluster while maximizing the dissimilarity between clusters [23].

We will be carrying out unsupervised machine learning [24] classification research using a clustering method named the K-Means clustering algorithm. The K-Means clustering algorithm requires the user to specify the number of centroids before implementation; that is, the k value [11, 18]. In effect, selecting the optimal k value is not straightforward and is considered one of the most challenging steps in partitional algorithms, but in our case, the k value is already predefined and it is equal to four. Hoping that this approach will support us to dive deeper to reveal more precise customers' routines to make reasonable assessments concerning behavioral strategies or hidden odd patterns of the data points during the given timeline.

Hereby, it should be noted that clustering of data points is a common form of exploratory data analysis, which is used to divide up the data points into different groups based on shared features. Data points that are similar to each other are grouped together in the same cluster, and those that are different are placed in another cluster, and so on.

To begin, we want to show some of our initial results using scatter plots that illustrate how our machine learning clustering works. Specifically, we used the K-means algorithm, which is a type of unsupervised machine learning clustering method. It helps us to group data into four clusters based on similarities without using predefined labels.

This scatter plot help us get a general sense and practical understanding of how the data is distributed and what patterns might exist. To improve the accuracy and reliability of the clustering results, we applied a standardization process beforehand. This means that all the data points were transformed to have similar scales, making it easier for the machine learning algorithm to compare them fairly and form meaningful clusters.

So, after standardization process, all the values became z-scores. This means they are on the same scale, with an average of zero and a standard deviation of one. It helps us compare the data more easily and makes the results from the clustering more reliable (figure 20).

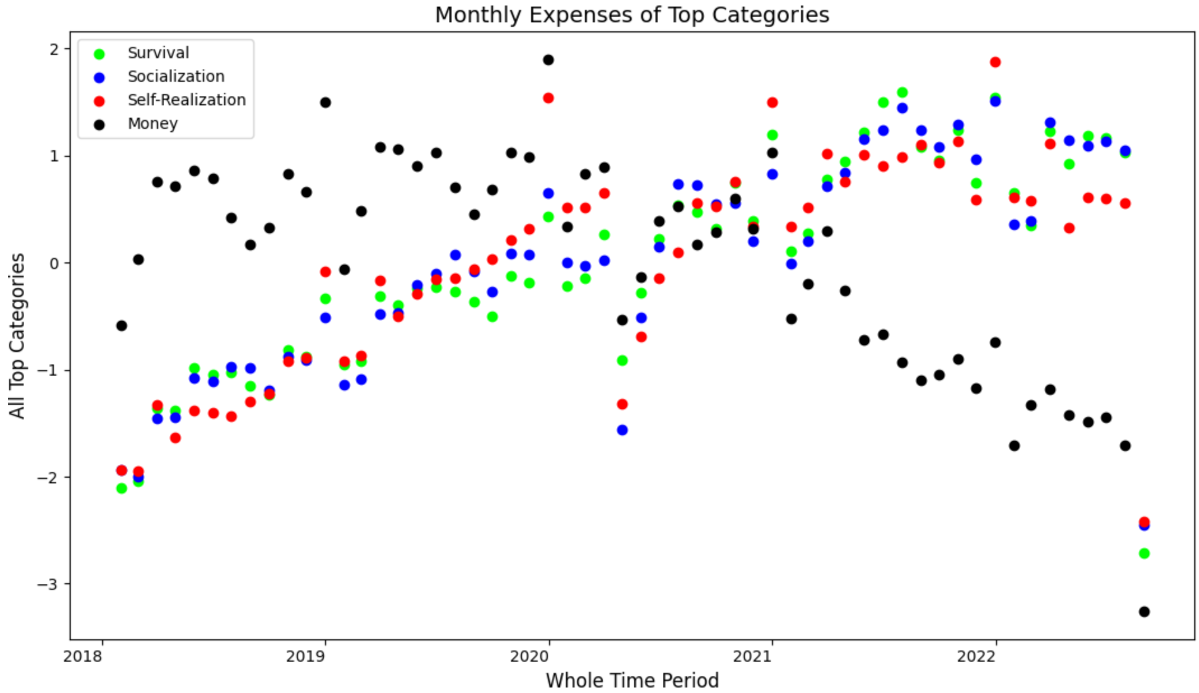

Figure 20 – Monthly total expenditures of each client and its label in terms of four basic values as top categories

When we look at the graphical interpretation of the data, we can clearly see a noticeable shift between the years 2020 and 2021, and just after the year 2022. The pattern of the data points is not consistent, which tells us that something unusual happened during this time period. Because of this, we decided to focus more closely on this specific period.

We all remember that this time was marked by the global outbreak of a serious infectious disease, i.e. the COVID-19 pandemic. It caused a lot of changes in daily life, and naturally, it also affected how people spent their money.

By studying this times of instability, we aim to understand how customer behavior changed. We want to uncover the changes in spending habits across four main categories. These shifts can be seen as the strategies people used to adapt during uncertain times.

Also, it’s worth mentioning that learning about customers’ behavior during this kind of instability can help us identify any hidden patterns or preferences. These patterns might even point to gaps in how we currently detect unusual or unexpected behavior in the data.

Identifying such issues is super important because detecting anomalies, that is, anything that doesn’t follow the usual pattern, is a key challenge in data science today. Understanding customer behavior better helps us improve these detection methods and make our models more accurate and reliable.

3 EXPERIMENTS AND FINDINGS

As mentioned earlier, the transactional data organizes all customer activity into four basic values as top categories:

- Survival;
- Socialization;
- Self-realization;
- Money.

These top categories form the basis of the machine learning clustering process, which consistently groups customers into four clusters. Keeping the number of clusters the same throughout the analysis helps maintain consistency and makes it easier to compare different time periods.

To understand how customer behavior changes, the study looks at two separate time periods, a stable time and a time of crisis. By calculating the difference in behavior for each customer between these two periods, we can see if the changes are significant.

The aim is to look not just at what people are doing at one point in time, but to see how it changes over time. This helps us better understand how customers make decisions and how outside events and circomstances can affect their spending.

In the end, this method helps us go beyond the numbers. It lets us tell the story behind the data i.e. how people's values and priorities change over time, how different groups of customers behave, and how they might react in the future.

3.1 Exploration of V1

The subtraction was performed by determining V_1, followed by the K-Means machine learning clustering algorithm

$$\boldsymbol{V_1 = Y_n^{2020} - Y_n^{2019}}, \qquad (1)$$

where (n) refers to each customer.

The first data portion was taken from '2020-03-15' and '2020-03-31'; let us name it as an irregular subset of 2020 by summing the top categories values for each customer:

$$\boldsymbol{Y_n^{2020}} = (\text{transactions } \textit{from } 2020\text{-}03\text{-}15 \textit{ till } 2020\text{-}03\text{-}31)$$

The second data portion was taken from '2019-03-15' and '2019-03-31'; let us name it as a regular subset of 2019 by summing the top categories values for each customer:

$$\boldsymbol{Y_n^{2019}} = (\text{transactions } \textit{from } 2019\text{-}03\text{-}15 \textit{ till } 2019\text{-}03\text{-}31)$$

Thus, V_1 is:

Top Category	money	self_realization	socialization	survival
Client				
224	0.0000	0.0000	-0.3669	0.3669
1108	-0.2815	-0.0803	0.0000	0.3618
1117	-0.2568	-0.1458	-0.1180	0.5207
1197	0.7031	0.0000	0.0000	-0.7031
1223	0.1555	0.0000	-0.0310	-0.1245
...	...	...	...	...
3560436	0.6446	0.2149	0.0376	0.1029
3561841	0.4255	0.0980	0.0543	-0.5779
3561945	0.0000	0.0000	-0.1501	0.1501
3562613	-0.4669	-0.0525	0.0000	0.5195
3564569	-0.0661	-0.0854	-0.0260	0.1775

Figure 21 – Difference in spending

In this part of the study, we compare two different years to understand how customer behavior changes during difficult times. The first year is 2019, it was a relatively normal year, with no major disruptions. The second year is 2020, which was heavily impacted by the COVID-19 pandemic. By comparing these two years, we aim to find out whether customers continued to spend in the same way or if their behavior changed due to the crisis. This helps us understand how people respond to external challenges and whether their financial decisions shift during the times of unstability.

Now, what do we do here? We do something very simple: we subtract the data of 2020 from the data of 2019. In other words, we take the spending values from 2019, and we subtract the 2020 values for each customer. This gives us the difference. It shows us how customer behavior has changed from one year to another.

When we look at the results, we notice that many of the values are negative. This means that in 2020, people spent less money compared to 2019. It’s a clear sign that, during difficult times, people tend to change their spending habits. They become more cautious and try to avoid unnecessary expenses.

There could be several reasons for this change. Some people might worry about losing their jobs or not having enough income in the future. Others might be concerned about rising prices or general uncertainty. Because of this, they start prioritizing basic needs and cut back on everything else. This shift in behavior is important to recognize. It shows that when people face challenges, their priorities change. They move away from comfort or luxury and focus more on safety and survival.

This information is not just useful for researchers. It’s also very helpful for businesses. When a business sees that people are spending less, they can change their own strategy. For example, instead of trying to sell expensive or luxury items, they can offer basic goods, discounts, or cheaper services. This way, the business can stay connected with what customers really need in hard times.

So, in short, this simple subtraction helps us understand big changes in customer behavior. Even if the math is easy, the message is powerful: in hard times, people spend less, and they focus more on the things that really matter.

The below graphical interpretation is the distribution of relativity of top categorical variables while performing V_1 (figure 22).

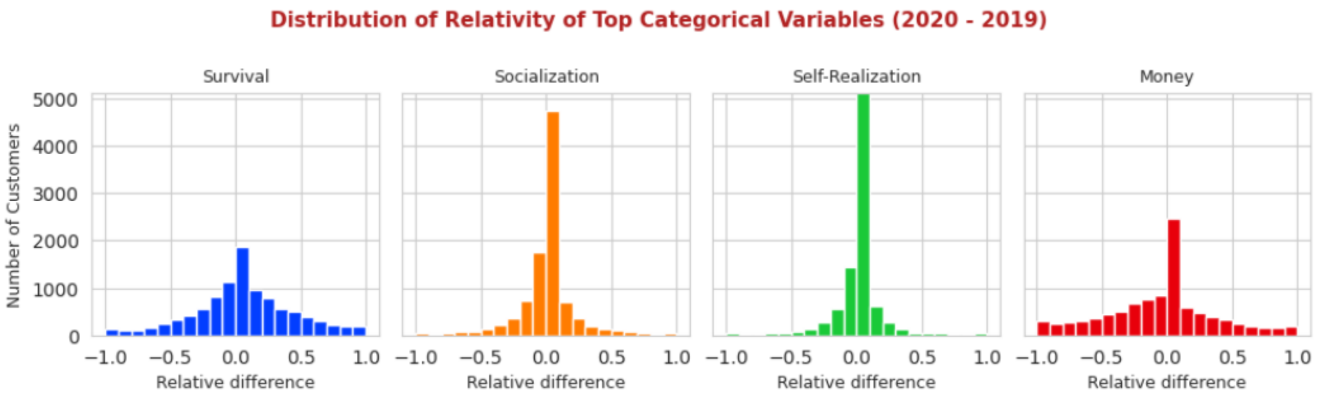

Figure 22 – Features distributions

Now, we perform the K-Means machine learning clustering algorithm to group the observations in four distinct clusters. That is, firstly, we provide the model to train it and then to get a cluster identifier that is associated with a customer. Here, the scatter plot shows how t-SNE [25] has mapped the dataset into a 2D space (figure 23). The t-Distributed Stochastic Neighbor Embedding (t-SNE/tSNE) is a dimension reduction method that is based on distances between the data points and attempts to maintain these distances in lower dimensions. This method is needed to visualize complex data structures and perform an analysis. Also, it is effective for clustering validation because it helps check if unsupervised clusters (e.g., from k-means) make sense in a reduced space.

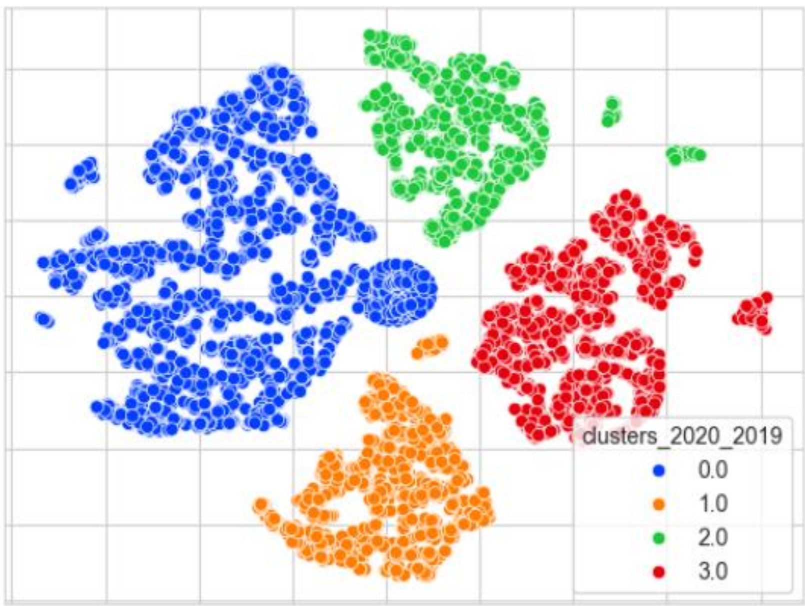

Figure 23 – t-SNE, V_1 clusters

The graph shown below in (figure 24) is a violin plot, which is used to explain how the top spending categories behave during the V_1 variation period. This kind of chart helps us look at how much money was spent in different categories and how that spending was spread out among different customers. It is a helpful way to see the differences and similarities between customers in terms of their financial actions.

To better understand this figure, it is useful to know what we mean by quantitative data interpretation. This is the process of looking closely at numbers and data to find useful information. It means not just reading the numbers but analyzing them to discover patterns, trends, or unusual behaviors. By doing this, we can learn more about how people act and make decisions based on facts. This process often includes using charts, graphs, and basic math or statistical tools to make the data easier to understand. In short, it helps turn large amounts of numbers into clear conclusions that we can use to support smart, fact-based decisions.

In this violin plot, we also look at the direction of the shift, which tells us something important about how much money is being spent:

- If the shift goes in the negative direction, it means the customer is spending less money in that top category. This could mean they are trying to save or cut back.
- If the shift goes in the positive direction, it means the customer is spending more money in that category. This might show a higher need or priority in that top category.

Overall, this violin plot is a useful visual tool that allows us to easily compare how spending changes across different categories and between different customers. It helps us understand what is happening during the V_1 variation in a simple, visual way.

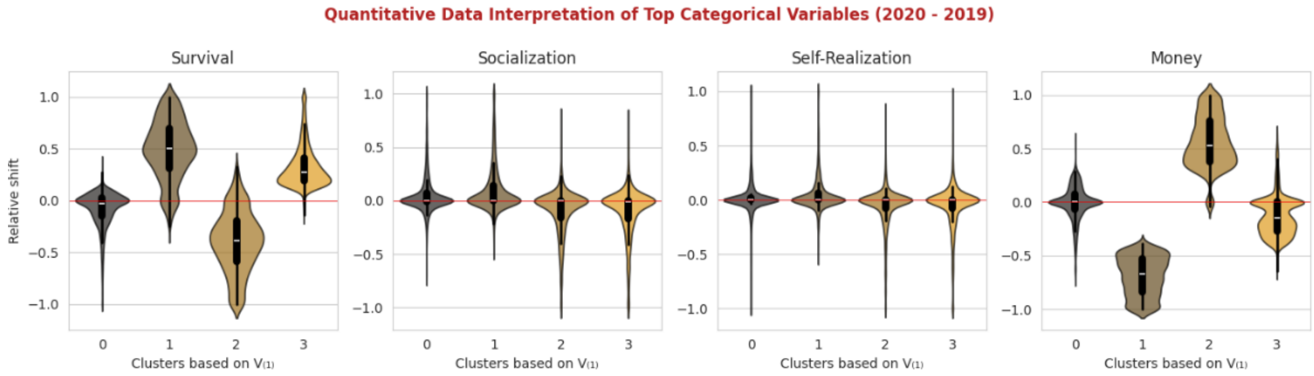

Figure 24 – Cluster population behavior shifts in V_1. Data points are primarily clustered around the median.

When we look at the figure above, we can see that Cluster 0 includes customers who showed very little change in their behavior during the V_1 variation period. These customers continued to act in the same way without making big adjustments, even when conditions might have been uncertain or stressful. This suggests that they remained calm and steady during the crisis. Because of this behavior, we can describe them as "**unflappable**" or "**stolid**" customers, that is, people who are not easily shaken and do not overreact to difficult situations. Their stable behavior may show emotional control and confidence in their chosen strategy.

Moving on to Cluster 1, we notice a different type of behavior. Customers in this group did not withdraw any money from ATMs at all. However, they spent a large amount of money in the survival category, which includes things like food, health, and other basic needs. This shows that they were careful with their finances but still made sure to cover important living expenses. Their behavior suggests a thoughtful and forward-looking mindset. These customers were preparing and planning wisely rather than reacting in panic. For this reason, we can call them "**prudent**" or "**cautious**" customers, that is, people who take care to manage risks and make careful decisions for the future.

In Cluster 2, the behavior is quite different again. Customers here withdrew a large amount of money from ATMs but spent very little on survival needs. This shows that they were possibly collecting or keeping cash without actually spending it on basic items. Such behavior may suggest a desire to hold onto money and avoid spending as much as possible, perhaps out of fear or uncertainty. Because of this, we can say these customers are showing signs of hoarding or extreme saving. They may be described as "**curmudgeon**" or "**miser**" customers, that is, individuals who are very reluctant to spend money and prefer to keep their wealth untouched.

Finally, in Cluster 3, we observe customers who withdrew only a small amount of money from ATMs but spent a noticeable amount on survival needs. This means they were not focused on gathering cash but still continued spending to support their daily lives. Their spending pattern shows a certain ease and lack of concern about the situation. They may not have adjusted much or worried too much, acting as if everything was under control. This kind of attitude can be seen as relaxed or even overconfident. Therefore, we can describe them as "**complacent**" or "**self-satisfied**" customers, that is, people who are satisfied with how things are and do not feel the need to change or react strongly.

3.2 Exploration of V_2

The subtraction was performed by determining V_2, followed by the K-Means machine learning clustering algorithm

$$\boldsymbol{V_2 = Y_n^{2022} - Y_n^{2021}}, \tag{2}$$

where (n) refers to each customer.

The first data portion was taken from '2022-03-15' and '2022-03-31'; let us name it as an irregular subset of 2022 by summing the top categories values for each customer:

$\boldsymbol{Y_n^{2022}}$ = (transactions *from* 2022-03-15 *till* 2022-03-31)

The second data portion was taken from '2021-03-15' and '2021-03-31'; let us name it as a regular subset of 2021 by summing the top categories values for each customer:

$\boldsymbol{Y_n^{2021}}$ = *(transactions from 2021-03-15 till 2021-03-31)*

Thus, V_2 is:

Top Category	money	self_realization	socialization	survival
Client				
224	0.0000	0.0000	0.1498	-0.1498
1108	-0.6039	0.1850	0.0000	0.4189
1117	-0.1964	-0.1015	-0.1100	-0.5921
1197	0.0000	-0.7738	-0.0447	0.8185
1223	0.0000	0.0000	0.0000	0.0000
...	...	...	...	...
3560436	-0.8187	0.7677	0.0000	0.0509
3561841	0.1315	-0.1917	-0.1897	0.2499
3561945	-0.9318	0.0000	0.0000	0.9318
3562613	0.3335	0.0994	-0.2316	-0.2013
3564569	0.1804	-0.1623	-0.0774	0.0593

Figure 25 – Difference in spending

Here, we carry out the subtraction operation between irregular subset 2022 and regular subset 2021; that is, differences between their transactions were calculated. The purpose of this is to attempt to see the variations in customers' existence preferences. After the subtraction operation, it is evident that we get negative signs, which imply the downward tendency of customer expenditures during the time of instability (figure 24).

The below graphical interpretation is the distribution of relativity of top categorical variables while performing V_2 (figure 26).

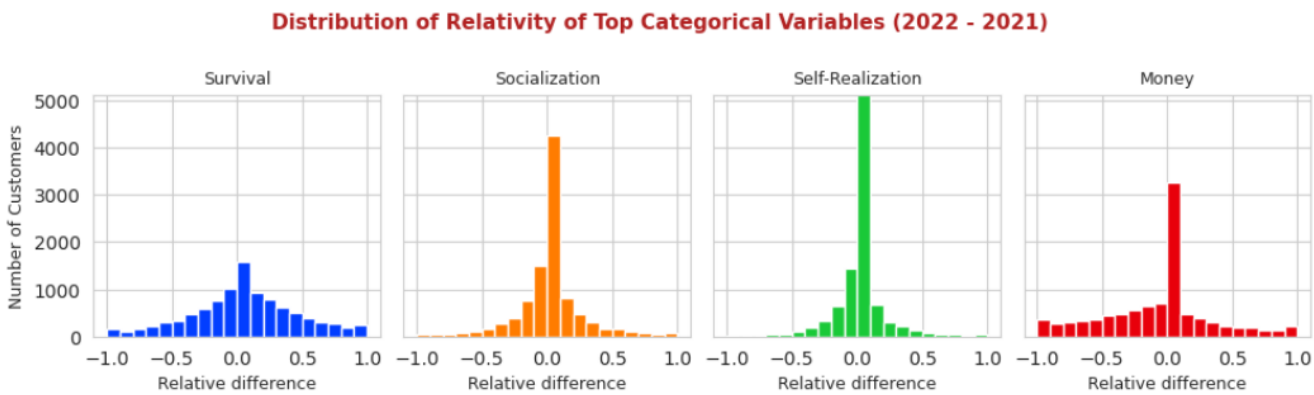

Figure 26 – Features distributions

Now, as the model already has been trained, in this case, we just get a new cluster identifier that is associated with the same customer by recalling the model. Here, the scatter plot shows how t-SNE has mapped the dataset into a 2D space (figure 27). The t-Distributed Stochastic Neighbor Embedding (t-SNE/tSNE) is a dimension reduction method that is based on distances between the data points and attempts to maintain these distances in lower dimensions. This method is needed to visualize complex data structures and perform an analysis. Also, it is effective for clustering validation because it helps check if unsupervised clusters (e.g., from k-means) make sense in a reduced space.

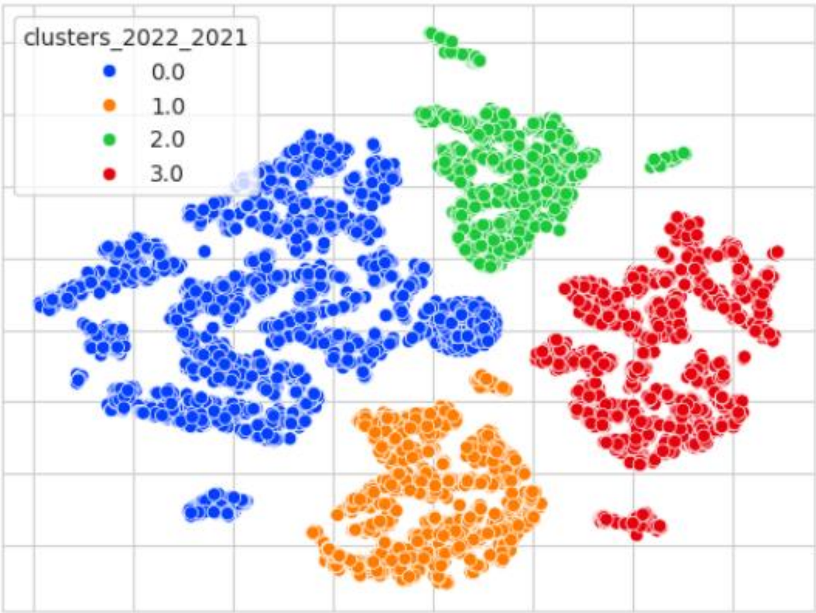

Figure 27 – t-SNE, V_2 clusters

The graph shown below in (figure 28) is a violin plot that provides a visual interpretation of quantitative data during the V_2 variation. This type of chart is used to explore the relationship and distribution of the most important categorical variables, especially in terms of how customers behaved in different spending areas. In this case, the violin plot helps us to see how spending levels are spread out among the customers during the V_2 period.

By observing the shapes and directions of these violins, we can better understand where customers are focusing their money, and how their behavior during the V_2 period reflects their priorities or strategies. This makes the violin plot a helpful tool for quickly identifying customer spending patterns in a visual and easy-to-understand way.

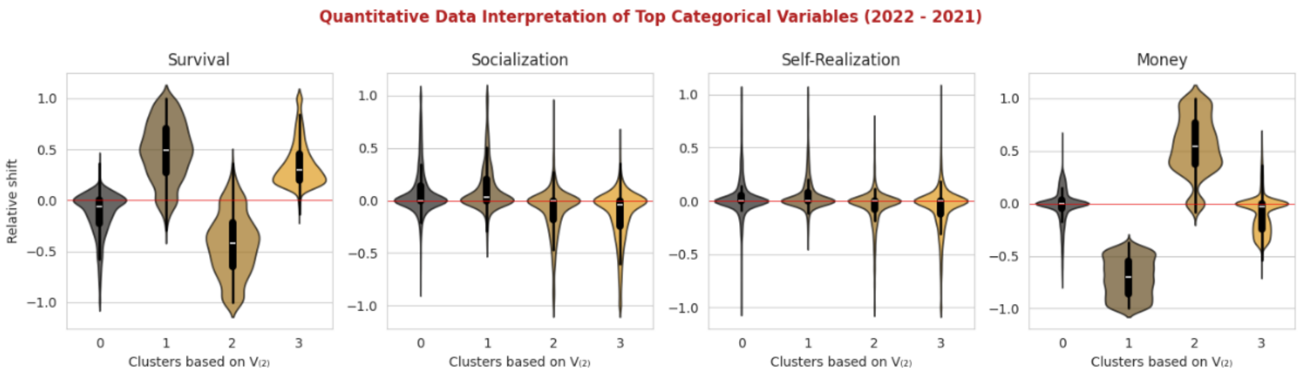

Figure 28 – Cluster population behavior shifts in V_2. Data points are primarily clustered around the median.

When we look at the figure above, we can see that Cluster 0 includes customers who showed very little change in their behavior during the V_2 variation period. These customers continued to act in the same way without making big adjustments, even when conditions might have been uncertain or stressful. This suggests that they remained calm and steady during the crisis. Because of this behavior, we can describe them as "**unflappable**" or "**stolid**" customers, that is, people who are not easily shaken and do not overreact to difficult situations. Their stable behavior may show emotional control and confidence in their chosen strategy.

Moving on to Cluster 1, we notice a different type of behavior. Customers in this group did not withdraw any money from ATMs at all. However, they spent a large amount of money in the survival category, which includes things like food, health, and other basic needs. This shows that they were careful with their finances but still made sure to cover important living expenses. Their behavior suggests a thoughtful and forward-looking mindset. These customers were preparing and planning wisely rather than reacting in panic. For this reason, we can call them "**prudent**" or "**cautious**" customers, that is, people who take care to manage risks and make careful decisions for the future.

In Cluster 2, the behavior is quite different again. Customers here withdrew a large amount of money from ATMs but spent very little on survival needs. This shows that they were possibly collecting or keeping cash without actually spending it on basic items. Such behavior may suggest a desire to hold onto money and avoid spending as much as possible, perhaps out of fear or uncertainty. Because of this, we can say these customers are showing signs of hoarding or extreme saving. They may be described as "**curmudgeon**" or "**miser**" customers, that is, individuals who are very reluctant to spend money and prefer to keep their wealth untouched.

Finally, in Cluster 3, we observe customers who withdrew only a small amount of money from ATMs but spent a noticeable amount on survival needs. This means they were not focused on gathering cash but still continued spending to support their daily lives. Their spending pattern shows a certain ease and lack of concern about the situation. They may not have adjusted much or worried too much, acting as if everything was under control. This kind of attitude can be seen as relaxed or even overconfident. Therefore, we can describe them as "**complacent**" or "**self-satisfied**" customers, that is, people who are satisfied with how things are and do not feel the need to change or react strongly.

After defining the V_1 and V_2 variations, the next step is to compare the items from V_1 and V_2. This comparison helps us understand how customer behavior may have changed over time. As we already know, each customer can either keep the same cluster identifier in both variations or have a different one. In other words, some customers may belong to the same group in both V_1 and V_2, meaning their behavior stayed consistent. However, other customers may have a different cluster identifier in V_2 compared to V_1, which means their behavior changed. This difference in cluster membership allows us to see which customers stayed loyal to their strategy and which ones changed their approach during the observed period. Hereby, we can conclude that:

- If a customer has the same cluster identifier in both the V_1 and V_2 variations, it means that this customer continued to behave in the same way during the whole time period. In other words, the customer did not change their behavior, preferences, or decision-making style, even when things around them may have changed. This kind of customer showed stability and did not shift to a new strategy. We can say that they stayed loyal to their original way of thinking and acting, without doubt or hesitation. This shows a strong level of consistency and commitment to their chosen strategy, no matter what happened during the variation period;

- If a customer has a different cluster identifier in V_1 and V_2 variations, it means that the customer changed their way of behaving over time. This change shows that the customer did not stick to their original behavior or strategy. Instead, they made a shift, possibly due to new conditions, uncertainty, or hesitation. This type of behavior suggests that the customer may have adjusted their decisions or preferences in response to outside factors, such as changes in the environment, personal situation, or available options. In this case, the customer did not follow a stable path but instead moved to a new strategy, showing flexibility or even uncertainty in their actions.

The graphical illustration below, shown in (figure 29), presents a violin plot, which is used to give a visual explanation of quantitative data. This type of chart helps us to better understand how the values of the main categorical variables are spread out and how they relate to each other. In this case, the plot focuses on showing the differences between two groups of customers: those who kept their original strategy and those who changed or shifted their strategy during the observed period. The violin plot displays not only the range of data but also the densit, that is, where most customer behavior is concentrated. This allows us to clearly compare the patterns of behavior in each group. By looking at the shape and width of each violin, we can identify whether certain actions, such as spending or withdrawing money, were more common in one group than the other. As a result, (figure 29) helps us to visually understand the impact of behavioral consistency or change among customers when it comes to their main spending and financial habits.

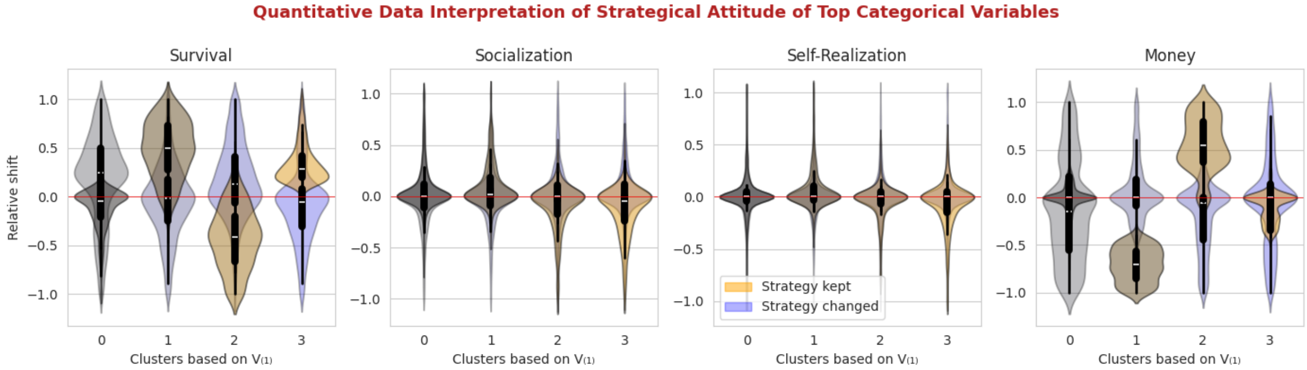

Figure 29– Cluster population behavior shifts according to V_1. Data points are primarily clustered around the median.

3.2 Cluster Migration

Life in every aspect is permanently in flux. Especially when we face life's unexpected and uncontrolled sudden undesired events or even horrible emergencies such as natural catastrophes. Then, it turns out each individual is inevitably obliged to revise his/her attitude and life existence preferences. On the other words, life is always changing, and this becomes especially clear during unexpected and difficult events, such as natural disasters or other emergencies. During such times, people often need to rethink their choices, priorities, and lifestyles. These changes can also affect how they behave as customers, including how they make decisions or stick to certain strategies.

According to this fact, we also want to explore and find the answer to the question, is there any cluster migration between the two variations in our analysis V_1 and V_2? In other words, we aimed to find out whether customers changed their group membership from V_1 to V_2, which could mean that their behavior or preferences changed during the period of observation. In effect, cluster migration occurs when customers shift from their original grouping V_1 to a new cluster V_2 due to changes in behavior, preferences, or decision-making logic.

To determine the migration process between clusters, we are free to implement the Sankey diagram [30], which is the Pythonic functionality for quantitative data interpretation. A diagram is a graphical interpretation that is used to explain the interrelated relationships and connections between the parts that it illustrates, and the Sankey diagram is one of them.

Furthermore, in addition to being visually attractive, a Sankey diagram may be used to illustrate the volume of flow between various nodes (clusters), which simplifies the understanding of complex systems. For financial transactions and environmental studies, this makes it very useful. A highly effective instrument for data storytelling, it can show both the amount and the direction of flow in a single visualization. It is a powerful data visualization tool which offers valuable insights into the flow of data. Using the Sankey diagram is crucial because it specifically represents how the data flow is subdivided at each stage [30].

To study this, we used the K-means clustering algorithm to divide customers into four groups for both V_1 and V_2. This method groups customers based on similar patterns in their behavior. After creating the clusters, we compared the group assignments for each customer in V_1 and V_2 to see if they stayed in the same group or moved to a different one.

The below graphical interpretation is the Sankey diagram, which delicately performs the migration procedure from V_1 to V_2, and as it can be seen, the clusters' customers are substantially migrated from one cluster to another (figure 30).

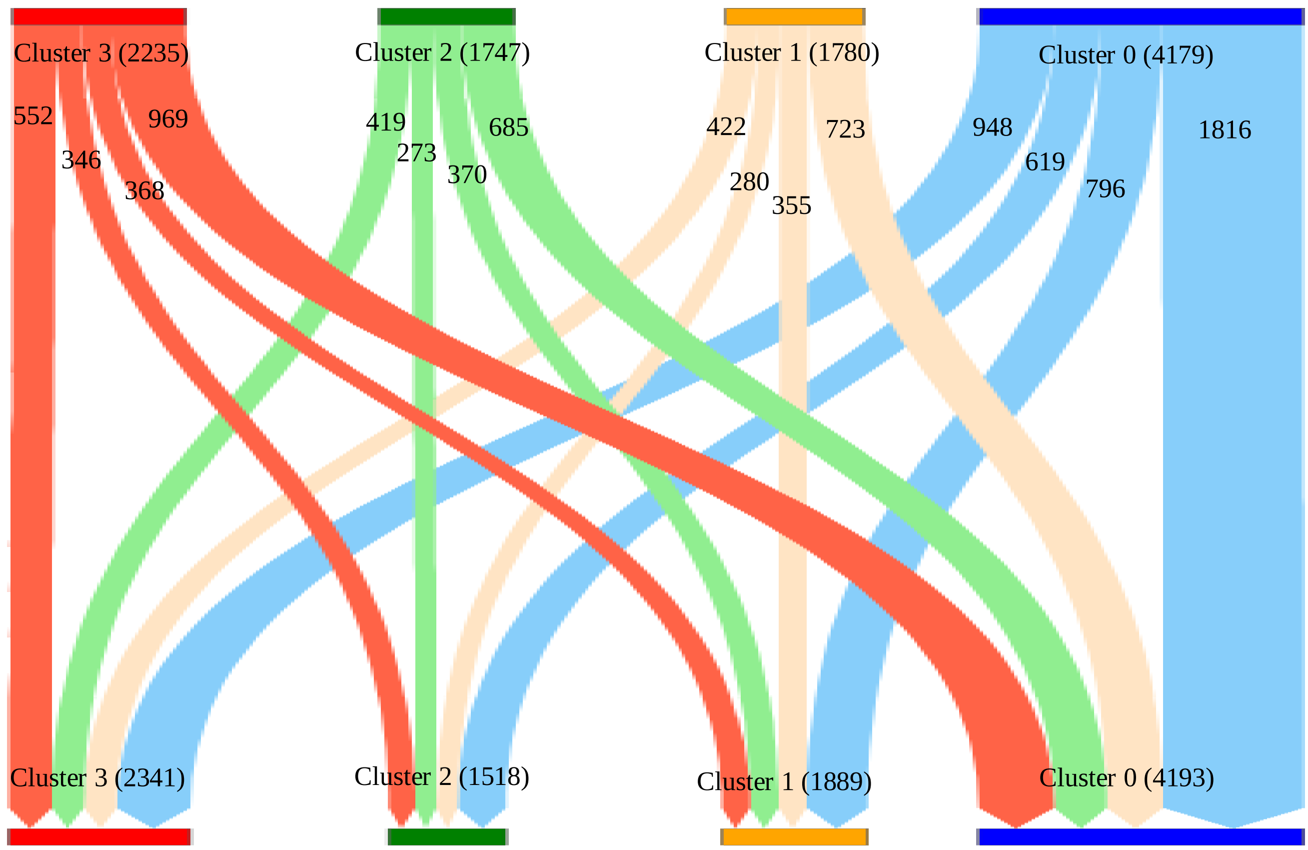

Figure 30 – Sankey diagram showing customer cluster migration from V_1 to V_2 based on K-means clustering results. Flow widths correspond to the number of customers transitioning between clusters

Customers who remained in the same group were considered non-migrants, while those who moved to a different group were considered migrants. To show these changes clearly, as it mentioned, we used a Sankey diagram, shown in (figure 30). This diagram displays how customers moved between clusters from V_1 to V_2. The thickness of the lines represents how many customers moved from one cluster to another.

The diagram also shows that many customers did, in fact, change their cluster, meaning their behavior likely changed as well. However, some customers stayed in the same group, which fits with our earlier findings about strategic consistency and loyalty.

This part of the analysis shows how customer behavior can shift in response to big life events or external pressures. It also suggests that tracking these changes over time can help improve customer models and forecasts. We can explore the meaning and possible reasons for these migrations in the further analysis.

3.3 Splitting the Expenditure by Clusters

It is also possible to understand each cluster by looking at how much they spend in the top categories: survival, socialization, self-realization, and money. Each cluster shows a different way of spending across these four top categories. For example, one cluster may spend most of its money on survival needs like food and housing, while another may spend more on things that help with social life or personal growth. By checking these spending parts, we can find a clear pattern for each group.

This helps us see what each cluster cares about the most, and it makes it easier to describe the behavior and priorities of the people in each group. By interpreting these patterns, we can better understand customer needs and preferences, which can be useful for targeted marketing, product recommendations, or strategic planning.

These spending patterns are shown in (figure 31), where we can clearly see how each cluster splits its spending among the four top categories i.e. each cluster’s portion of spending in 2020 based on top categories.

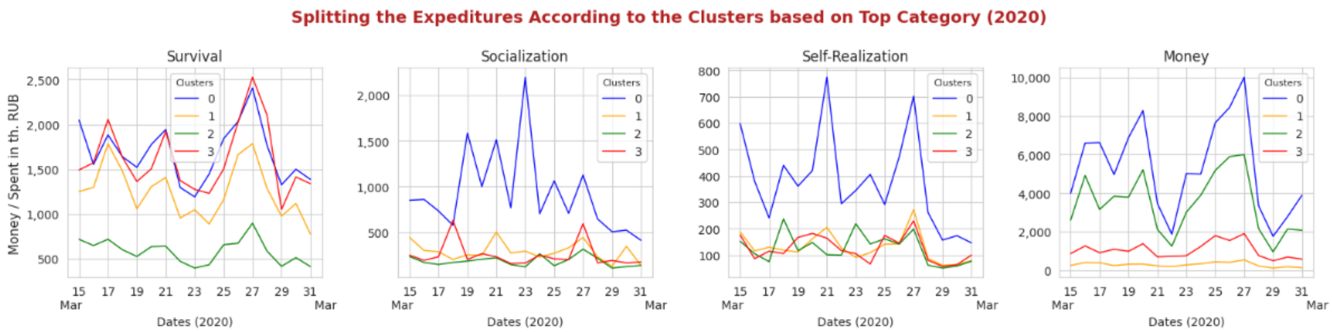

Figure 31 – Splitting the spending by each cluster in 2020

By doing this, we can see a clear pattern for each cluster. These patterns help us understand the mindset and priorities of customers in each group. This information is useful not only for describing customer behavior but also for making better business or marketing decisions.

When we look into the (figure 31), we can see the splitted expentidures by each clusters in each top categories in 2020. The customers in cluster 0 which we named them as “**unflappable**” or “**stolid**” are spending money rather more than other clusters in each top categories. The customers in cluster 1, cluster 2, and cluster 3 actually are spending much less money in socialization and self-realization top categories.

These spending patterns are shown in (figure 32), where we can clearly see how each cluster splits its spending among the four top categories i.e. each cluster's portion of spending in 2022 based on top categories.

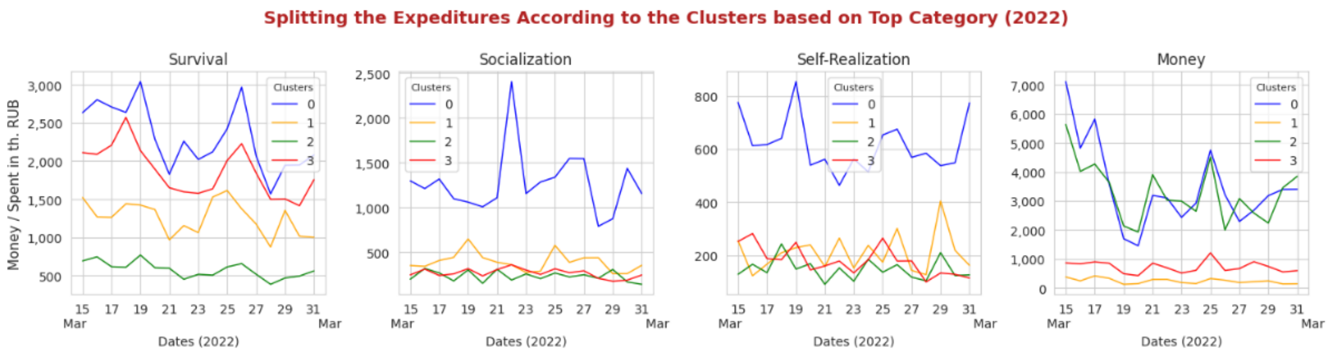

Figure 32 – Splitting the spending by each cluster in 2022

When we look into the above figure, we can see the splitted expentidures by each clusters in each top categories in 2022. The customers in cluster 0 which we named them as "**unflappable**" or "**stolid**" are as usual spending money rather more than other clusters in each top categories. The customers in cluster 1, cluster 2, and cluster 3 are as usual spending much less money in socialization and self-realization top categories.

These spending patterns are shown in (figure 33), where we can clearly see how each cluster splits its spending among the four top categories i.e. each cluster's portion of spending in 2020 based on food, health, travel, and kids categories.

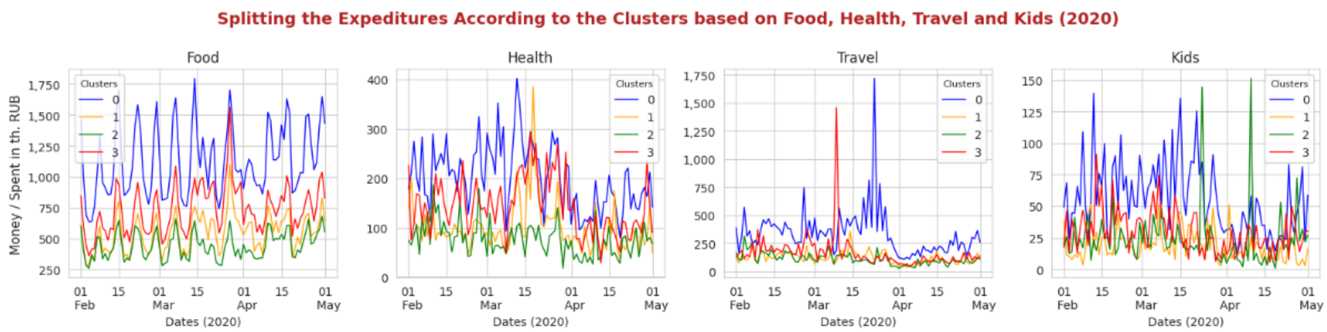

Figure 33 – Splitting the spending by each cluster in 2020

When we look at the above figure, we can notice the divided expenditures by cluster for the food, health, travel, and kids categories in 2020. It is evident that rapidly after March, there is an agitational segment in each category, which is definitely referred to as the pandemic period in 2020.

These spending patterns are shown in (figure 34), where we can clearly see how each cluster splits its spending among the four top categories i.e. each cluster's portion of spending in 2022 based on food, health, travel, and kids categories.

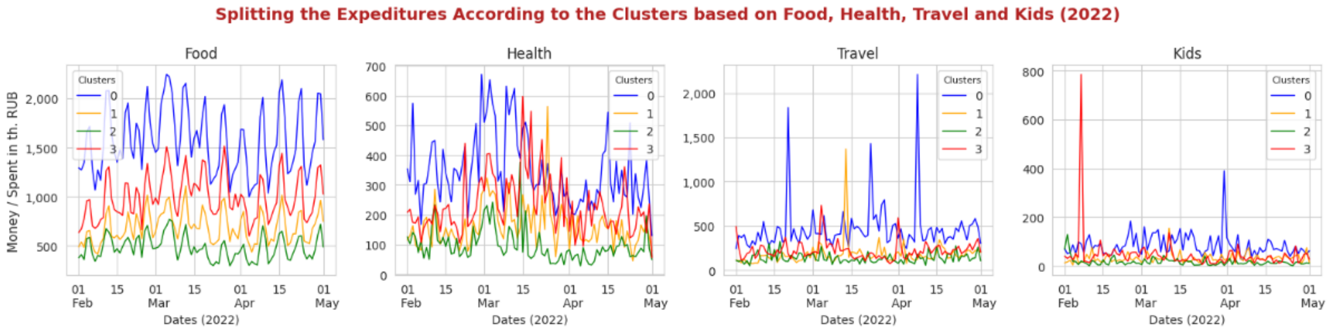

Figure 34 – Splitting the spending by each cluster in 2022

When we look at the above figure, we can realize the divided expenditures by cluster for the food, health, travel, and kids categories in 2022. It is evident that rapidly after February, there is an agitational segment in each category, which is definitely referred to the special military operation in 2022.

3.2 Forecasting

In forecasting process we will hypothesize that at the very least the customers who kept their strategy are those who, regardless of the crisis or no crisis situation, they actually keep their loyalty to ways of behaving, i.e., their strategy. Hereby [39], one of our target exploration consists of proving this assumption by showing as it implies that their forecasting loss function is smaller than who shifted their strategy.

In essence, the core idea of the hypothesis is that we actually separate costomers in two main groups:

- **Loyal Customers (loyals)**: Their consistent behavior (unchanged strategy) aligns with historical data patterns, allowing forecasting models to perform reliably. This stability results in a smaller loss function (e.g., lower prediction errors, lower loss function);

- **Strategy Shifters (switchers)**: Behavioral changes during crises disrupt historical patterns, increasing model prediction errors. The loss function grows as forecasts fail to capture new strategies.

Thus, we want to demostrate that loyal customers' forecasting loss function value have to be smaller than strategy shifters' forecasting loss function value because in case if a crisis causes others to panic and change strategies, their actions become less predictable, hence higher forecasting errors.

A forecasting loss function typically measures the difference between predicted values and actual outcomes. A smaller loss means better predictions. So the idea is that loyal customers, who don't change their strategy, have more consistent behavior, making their forecasts more accurate. Those who switch strategies might have more variability, leading to higher loss.

By the way, if we analyse a bit with reference to two mentioned groups. We may declare also that maybe loyal customers are in industries less affected by the crisis, so their behavior is stable not because of loyalty but because their environment is stable. Or perhaps loyal customers have different characteristics that make their forecasting easier, irrespective of the crisis.

Also we should pay attention to, how is "strategy" defined? It could be investment strategy, purchasing behavior, or something else. The specifics might influence how much a crisis impacts their behavior. For example, in a financial crisis, investment strategies might change more drastically than, say, consumer goods purchasing strategies.

Another point of view is that the forecasting loss function being smaller for loyal customers implies that their actual behavior deviates/diverges less from predictions. But if a crisis introduces new variables that the model doesn't account for, even loyal customers might behave unexpectedly. However, if their strategy inherently buffers against crises, their actions remain predictable.

Forecasting [27, 40] is likely the most common implementation of machine learning in the real world. Businesses forecast product demand, governments forecast economic and demographic growth, and meteorologists forecast weather. Time series forecasting is a vast topic with a lengthy history. That is, time series forecasting is a statistical/data science technique used to predict future values based on historical time-ordered data. It analyzes patterns (trends, seasonality, cycles) to make informed estimates for business, economics, weather, and more.

The basic object of forecasting is the time series, which is a set of observations recorded over time. In forecasting applications, the observations are typically recorded with a regular frequency, like daily or monthly. In this phase, we focus on applying contemporary machine learning approaches to time series data in order to provide the most accurate forecasts.

In our study, we take advantage of using the regression method to forecast a response using a set of predictors. Actually, regression analysis is one of the most important fields in statistics and machine learning. There are many regression methods available. Linear regression [28] is one of them, and it is also a type of supervised machine learning predictive algorithm.

Generally, we need a linear regression method to see the relationship of the variables. Linear regression is probably one of the most important and is a fundamental machine learning predictive algorithm that has been widely used due to its simplicity and efficiency.

By the way, a predictive algorithm is a set of mathematical equations and statistical techniques used to predict an outcome or future event based on historical data such as times-series [41]. A predictive algorithm is used to build a predictive model that can hereby forecast future trends, identify patterns in data, and make data-driven decisions. The quality of the prediction strongly depends on the quality and volume of the data.

Moreover, we also need to evaluate the performance of such a model. The purpose of performance evaluation is to provide an unbiased assessment of the out-of-sample accuracy of the selected model [29]. For this, we use the MAPE metric, which helps us to analyze the performance of the model. The lower the MAPE metric values, the better the model is. For example, a MAPE metric value of 10% means that the average absolute value of the percentage error difference between the predicted values and the actual values is 10%. A lower MAPE metric value indicates a more accurate prediction, while a higher MAPE metric value indicates a less accurate prediction. Keep in mind that MAPE works best with data without zeros and extreme values.

In addition to this, when we examine a time series pattern before the forecasting process, we have to pay attention to smoothing data techniques as well. For this, we need to use the Python window function, as it is extremely useful for smoothing or normalizing time series data, helping in trend analysis, and especially in the forecasting process. Python window functions are particularly indispensable for running totals, moving averages, and cumulative statistics [20].

For performing the forecasting process, we focus on using efficient low-level implementations of standard linear algebra algorithms such as the Frobenius norm (sometimes also called the Euclidean norm) and maximal similarity method (MaxS) which has been developed by Irina Chuchueva [7]. We aimed to use MaxS method for the purpose to perform a base line forecast for the results assessment.

MaxS method assumes that, according to the Dirichlet principle, for a pseudo random process with a finite number of internal states the values of the time series will sooner or later be repeated. The similarity function in MaxS is defined as the Pearson correlation coefficient for portions of time series X

$$\boldsymbol{L_n(i) = |corr(X_n^m, X_i^m)|,} \tag{3}$$

where i = 1,…, n-1, and n is the series portion beginning of size m for which the similarity is searched; X_i^m is a previous portion starting from i. The maximum of this function shows the portion of the series mostly alike the portion that is situated directly before the period of forecast, starting with t = T. Then the values that follow this mostly alike portion re-scaled by the ratio of those two correlated portions are the predicted values for the time series X

$$\boldsymbol{X_T^t = A * X_{j+m}^t,} \tag{4}$$

where $j = arg\ max_i(L_{T-m+1}(i))$ and $A = ((X_j^m)^T * X_j^m)^{-1} * (X_j^m)^T * X_{T-m+1}^m$.

The Frobenius norm [30] is defined as the square root of the sum of the absolute squares of its elements. It is a significant tool that enhances the efficiency and performance of neural networks and provides a way to measure the size or magnitude of a matrix

$$\|\boldsymbol{F}\| = \sqrt{\left[\sum_{i.j}\left|\boldsymbol{a}_{i.j}\right|^{\mathbf{2}}\right]}. \tag{5}$$

By using the Frobenius norm, we can create a predictive model to forecast time series values. First, we need to define a coefficient for rescaling. Let it be q, where q is equal to the square root of the sum of the absolute squares of V_2 elements over the square root of the sum of the absolute squares of V_1; that is

$$\mathbf{q} = \|\boldsymbol{V_2}\| \,/\, \|\boldsymbol{V_1}\|. \tag{6}$$

The predictive model function is simply equal to

$$\mathbf{q} * \mathbf{X}\,(\mathbf{V}_{1(i)}), \tag{7}$$

where $i \in \{V_1 \text{ values}\}$.

By the way, each data point or observation in V_1 and V_2 is already formed weekly, which means that the Frobenius norms are already in the form of the seven-day vectors. Thus, when we use this function as a predictive model, we can notice that it shows the best accuracy, and the model performance is very high. The below representation shows obtained predicted results using the Frobenius norm predictive model. Each cluster is predicted separately.

The below graphical interpretation illustrates predicted results based on each cluster. The prediction period is taken from 24 February till the end of March 2022 (figure 35).

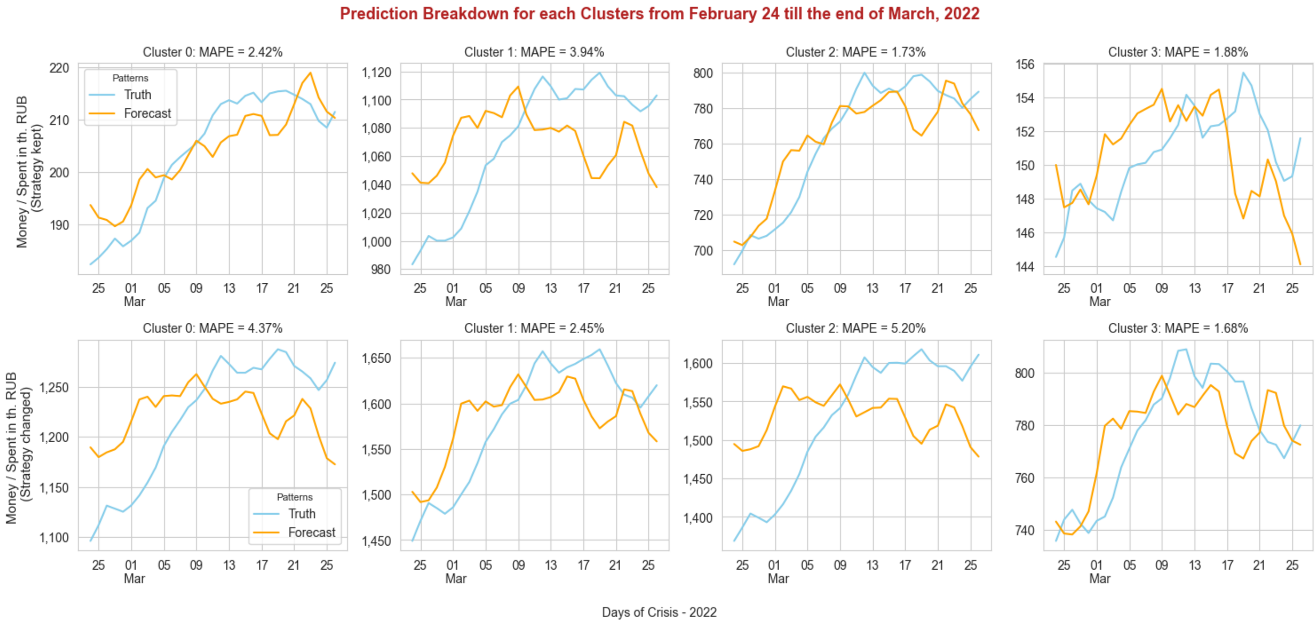

Figure 35 – Total expenditure forecasting for each cluster of customers who kept and changed their strategies

The graphical interpretation presented below illustrates the predicted outcomes for two distinct groups of customers: those who continued with their original strategy (strategy-kept customers, loyal customers) and those who made changes to their strategy (strategy-changed customers, switchers). This comparison helps to highlight the differences in behavior and outcomes between the two groups over the selected prediction period. The time frame for these predictions spans from February 24, 2022, to the end of March 2022. This period was chosen to reflect a phase of instability, allowing us to observe how strategy adjustments may have influenced customer behavior. The results are shown in (figure 36).

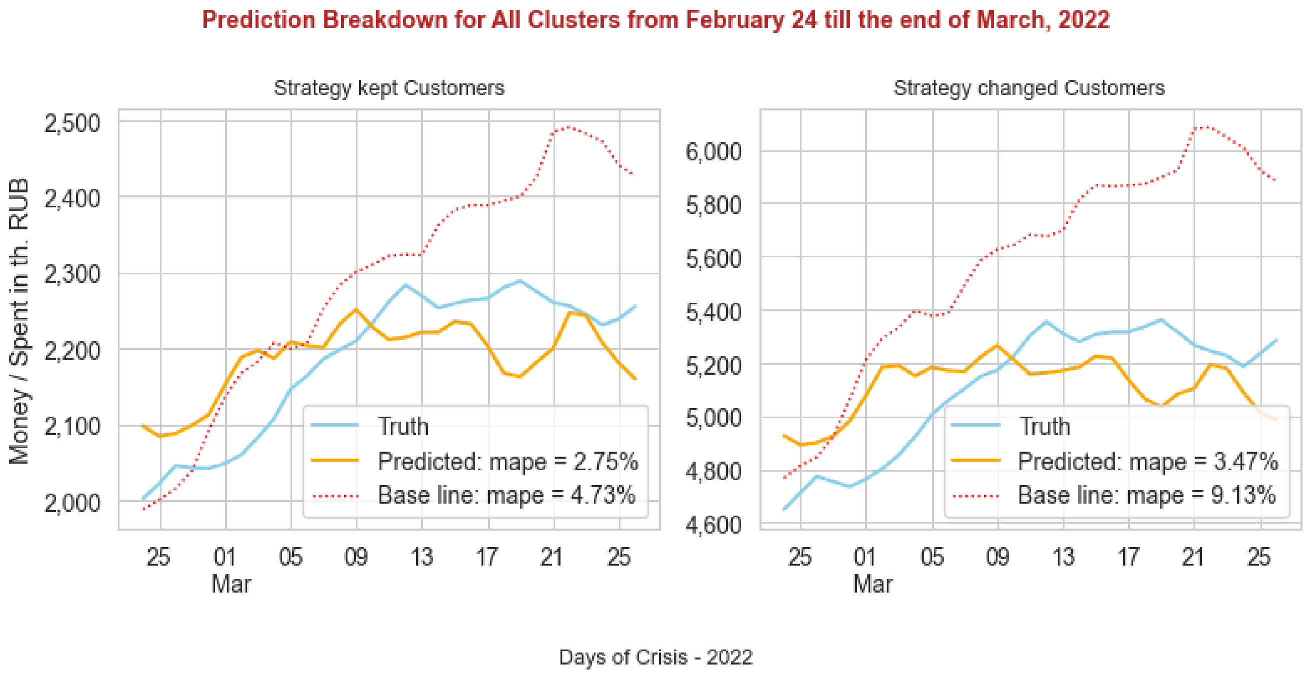

Figure 36 – Comparison of forecasting performance between customers who kept their strategy and those who shifted strategy, measured using MAPE and with one base line of the maximal similarity method

At the beginning of the forecasting process, we hypothesized that customers who kept their strategy regardless of the presence of a crisis, exhibited stronger loyalty to their strategic approach. Consequently, we expected their forecasting loss function to be lower than that of customers who changed their strategy.

Upon completion of the forecasting process, this hypothesis was evaluated by comparing the Mean Absolute Percentage Error (MAPE) across the two customer groups. As illustrated in (figure 36), the results indicate that the average forecasting loss for customers who kept their strategy was lower than for those who shifted. Specifically, the MAPE for the stable group was **2.75%**, while the MAPE for the shifting group was **3.47%**.

Table 6 – MAPE metric

	MAPE	
2022	Strategy Kept Customers	Strategy Changed Customers
All Clusters	2.75%	3.47%

These findings support our initial assumption, suggesting that customers who remained consistent in their strategic behavior were more predictable, which in turn resulted in more accurate forecasts. This confirms that strategy continuity can serve as a stabilizing factor in predictive performance, even under conditions of external uncertainty such as crises.

CONCLUSION

In this study, we learned a lot about how people change their spending habits during times of instability. By looking at different parts of the data, we were able to see how spending changed over time. We measured both average and total spending on a daily, weekly, and monthly basis across different categories. This helped us understand which areas of spending were most affected and which ones remained important during the crisis.

To dig deeper into the patterns, we used a machine learning clustering model to group customers based on similar spending behavior. This model was used throughout the research. Before applying it, we cleaned and prepared the data carefully to make sure the results were accurate. We also used quantitative data interpretation tools to show clear patterns in how people were spending. One interesting finding was that many people moved between behavior groups during the period of instability, showing that their habits were changing with the situation.

Alongside clustering, we also used forecasting methods to predict future behavior. These methods were based on efficient mathematical techniques, including the Frobenius norm and Maximal Similarity (MaxS). To check how accurate our predictions were, we used a standard error metric called MAPE (Mean Absolute Percentage Error). This helped us compare different forecasting methods and choose the ones that worked best.

Looking ahead, we believe it would be helpful to include more useful features in the model. This would make it easier to understand different types of spending and improve the quality of the clusters. Another goal is to develop a more advanced clustering approach that can give even deeper insights into how people behave during times of instability.

Improving the clustering model in future work could lead to better results and a stronger understanding of customer behavior during unstable periods. This can help both researchers and businesses better prepare for and respond to changes in the economy.

LIST OF REFERENCES

1. Daniel Schlör. Detecting Anomalies in Transaction Data // Doktorgrades der Julius-Maximilians-Universität Würzburg – 2022. P. 1–20
2. V. Guleva, A. Kovatsev, A. Surikov, P. Chunaev, G. Gornova. Value-based modeling of economic decision-making in conditions of an unsteady environment // Scientific and Technical Journal of Information Technologies, Mechanics and Optics 23 – 2023. P. 121–135.
3. A. Bezbochina, E. Stavinova, A. Kovantsev and P. Chunaev. Enhancing Predictability Assessment: An Overview and Analysis of Predictability Measures for Time Series and Network Links // MDPI – 2023. P. 2–10.
4. M. Koshkareva, A. Kovantsev. Crisis Behaviour Strategy Recognition Using Transactional Data // Procedia Computer Science 229 – 2023. P.1–217 National Center for Cognitive Research, TMO University, St. Petersburg 197101
5. Ilyass Haloui, Jayant Sen Gupta and Vincent Feuillard. Anomaly detection with Wasserstein GAN // ResearchGate – 2018.
6. Priyanka S. Patil, Nagaraj V. Dharwadkar. Analysis of Banking Data Using Machine Learning // I-SMAC – 2017.
7. I. Chuchueva, The time series extrapolation model based on maximum likeness set // Science and Education of the Bauman MSTU – 2010.
8. A. Kovantsev, P. Chunaev, and K. Bochenina. Evaluating Time Series Predictability via Transition Graph Analysis // ResearchGate – 2021.
9. S. Natha, M. Leghari, M. A. Rajput, S. S. Zia and J. Shabir. A Systematic Review of Anomaly Detection Using Machine and Deep Learning Techniques // ResearchGate – 2022.
10. Mausumi Das Nath and Tapalina Bhattasali. Anomaly Detection Using Machine Learning Approaches // ResearchGate – 2020.

11. M. Rahal, B. S. Ahmed, G. Szabados, T. Fornstedt, J. Samuelsson. Enhancing machine learning performance through intelligent data quality assessment: An unsupervised data-centric framework // ScienceDirect – 2025.
12. Ravinder M, V. Kulkarni. Performance Analysis of Machine Learning Techniques For Anomaly Detection in Indian Electricity Consumption Data // ScienceDirect – 2023.
13. S. Corli, L. Moro, D. Dragoni, M. Dispenza, E. Prati. Quantum Machine Learning Algorithms for Anomaly Detection: A review // ScienceDirect – 2024.
14. S. Mohammed, L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, H. Harmouch. The effects of data quality on machine learning performance on tabular data // ScienceDirect – 2025.
15. V. Lennart Heß, B. Damásio. Machine learning in banking risk management: Mapping a decade of evolution // ScienceDirect – 2025.
16. Uncover Insights using Exploratory Data Analysis (EDA). [Electronic resource]: – Access mode: https://businessanalyst.techcanvass.com/objective-of-exploratory-data-analysis/
17. Moments in Statistics – A Must Known Statistical Concept for Data Science [Electronic resource]: – Access mode: https://www.analyticsvidhya.com/blog /2022/01/moments-a-must-known-statistical-concept-for-data-science/
18. Discovering Patterns in Data [Electronic resource]: – Access mode: https://valdanchev.github.io/reproducible-data-science-python/notebooks /06_pattern_discovery_using_unsupervised_learning.html
19. Exploratory Data Analysis and Visualisation [Electronic resource]: – Access mode: https://valdanchev.github.io/reproducible-data-science-python/notebooks /05_data_exploration_and_visualisation.html
20. Mastering Python Window Functions [Electronic resource]: – Access mode: https://sqlpad.io/tutorial/mastering-python-window-functions-a-comprehensive-guide/
21. Mastering Cluster Analysis [Electronic resource]: – Access mode: https://dovetail.com/research/what-is-cluster-analysis/

22. Ingunn B., Bjørn-Helge M., Tormod N. New modifications and applications of fuzzy C-means methodology // ScienceDirect – 2007.

23. Cluster Analysis [Electronic resource]: – Access mode: https://julius.ai/glossary/cluster-analysis
24. Unsupervised Learning [Electronic resource]: – Access mode: https://julius.ai/glossary/unsupervised-learning
25. Mastering t-SNE: A Comprehensive Guide to Understanding and Implementation in Python [Electronic resource]: – Access mode: https://towardsdatascience.com/mastering-t-sne-a-comprehensive-guide-to-understanding-and-implementation-in-python-480929bfe6f4/
26. 5 Advanced Charting Techniques [Electronic resource]: – Access mode: https://www.dataexpertise.in/5-heatmaps-advanced-charting-techniques/
27. Linear Regression with Time Series [Electronic resource]: – Access mode: https://www.kaggle.com/code/ryanholbrook/linear-regression-with-time-series
28. 7 ML Algorithms You Can Use to Predict, Classify, & Forecast [Electronic resource]: – Access mode: https://infiniticube.com/blog/7-ml-algorithms-you-can-use-to-predict-classify-forecast/
29. Casper Solheim Bojer. Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities // ScienceDirect – 2022.
30. The Power of Frobenius Norms in Deep Learning [Electronic resource]: – Access mode: https://medium.com/@kirubasagar82/the-power-of-frobenius-norms-in-deep-learning-7afe589aaf2b
31. В. Ю. Гулева, А. Н. Кованцев, А. Г. Суриков, П. В. Чунаев, Г. В. Горнова, А. В. Бухановский. Ценностно-ориентированное моделирование принятия экономических решений в условиях нестационарности внешней среды // Журнал «Научно-технический вестник информационных технологий, механики и оптики» включен в Scopus

32. Loxton M. et al. Consumer behaviour during crises: Preliminary research on how coronavirus has manifested consumer panic buying, herd mentality, changing discretionary spending and the role of the media in influencing behaviour // J. risk Financ. Manag. MDPI, 2020.
33. A Must Known Statistical Concept for Data Science [Electronic resource]: – Access mode: https://www.analyticsvidhya.com/blog/2022/01/moments-a-must-known-statistical-concept-for-data-science/
34. Everything about Non-Parametric Probability Density [Electronic resource]: – Access mode: https://medium.com/@pritul.dave/everything-about-non-parametric-probability-density-function-kernel-density-estimator-kde-plots-a6e709d5bcf9
35. Code Basics Master [Electronic resource]: – Access mode: https://github.com/codebasics/py/tree/master
36. Negre E., Arru M., Rosenthal-Sabroux C. Toward a Modeling of Population Behaviors in Crisis Situations // Sci-Hub – 2018.
37. P. Rameshan. Leadership Behaviour in Covid-19 Crisis Response: A Personal Strategy Proposition. // IIM Kozhikode Society & Management Review – 2022.
38. Everything about Non-Parametric Probability [Electronic resource]: – Access mode: https://medium.com/@pritul.dave/everything-about-non-parametric-probability-density-function-kernel-density-estimator-kde-plots-a6e709d5bcf9
39. Differences between Correlation and Regression [Electronic resource]: – Access mode: https://byjus.com/maths/differences-between-correlation-and-regression/#
40. Tamara Louie. Applying Statistical Modeling & Machine Learning to Perform Time-Series Forecasting [Electronic resource]: – Access mode: https://www.slideshare.net/slideshow/applying-statistical-modeling-and-machine-learning-to-perform-timeseries-forecasting/124317217
41. Prediction with Times-Series [Electronic resource]: – Access mode: https://www.kaggle.com/code/kalilurrahman/store-sales-eda-prediction-with-ts/notebook