**NATIONAL RESEARCH UNIVERSITY**

**HIGHER SCHOOL OF ECONOMICS**

Faculty of Computer Science

Bachelor’s Programme “HSE and University of London Double Degree Programme in Data Science and Business Analytics”

**Research project report**

on the topic: "Optimization of client lists for communication by means of mathematical modeling"

**Student:**

group БПАД182 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Natalia Krauze Olegovna

Sign Name, Surname, Patronymic

18.06.2020

Date

**Supervisor**:

\_\_\_\_\_\_\_\_\_\_\_\_ Titova Natalia Nikolaevna\_\_

Name, Surname, Patronymic

Head of Online Sales Data Analysisat at Alfa-Bank and Head of SAS Department Academic Projects at the Higher School of Economics\_

Position

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Alfa-Bank and SAS company \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Company

Date \_\_27.06\_2020 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Grade Sign

**Moscow 2020**

**Content:**

**1. Introduction 3-4**

1.1 Abstract;

1.2 The role of members of the research project;

1.3 Instruments;

1.4 Main result;

**2. Review and comparative analysis 5**

2.1 Review of abstracts which you used for project and review of analogues;

**3. Selection of methods, algorithms, and models for project implementation. 6-28**

* Information about filtering of data;
* Creating pivot table;
* Analyze pivot tables;
* Creating an analytical base table;
* Normalization/Standardization;
* Table of correlations;
* K-Means Elbow Method;
* Choosing number of clusters;
* Segmentation K-Means;
* Analysis of variance;
* Analyzing segments and creating representative graphs;
* Financial and economic report;

**4. Results and conclusion 29**

**5. List of sources 30**

**Introduction:**

* 1. **Abstract:**

Nowadays one of the most popular areas of predictive analytics is the search for an optimal offer for a client or Next Best Offer. Our project is dedicated to customer analytics by unfiltered cutting data received from kid’s online store given by SAS INSTITUTE. In order to form the most suitable proposal, it is necessary to do a number of works: building analytical clustering (lifestyle - segmentation) using machine learning, identifying customer profiles based on purchases, the formation of a strategy to maintain the existing customer base through effectively built communications with customers for each segment, the choice of the best personal offer for each profile (Next Best Offer). Within the project it is necessary to allocate segments of clients which with the greatest probability will make the order in a category using results of the constructed mathematical models. Also, the project considers modern methods for detecting groups of goods characterized by mutual influence on each other in terms of demand volume (Market Basket Analysis), using Machine Learning methods and Financial and Economic analysis.

* 1. **The role of members of the research project:**

Another participant of the project besides me is Demchenko Karina. Both students from the program of Faculty of computer science, Bachelor’s Programme ‘HSE and University of London Double Degree Programme in Data Science and Business Analytics’, 2 courses. The work of Karina was dedicated to filter (‘clean’) and analyze data that she gets from our database and sort the data so that only the orders that were bought out are left in the table. While I was filtering and processing all the orders placed, cleaning up the wrong positions and analyzing the placed purchases. Then we divided the work on building pivot tables according to the obtained data. We needed to get pivot tables by purchased and placed orders. My job was to build tables by categories and subcategories of goods and time. Karina did the work on the other tables - days of the week, months, regions and others. Next, Karina combined the resulting table and made a third type of table called " Buyout". Then, my work was to make visualization (histograms, charts) on the resulting tables. The next step was to analyze the resulting summary tables and visualizations and to propose marketing hypotheses for products. We divided this work: Karina carried out the analysis for "Regions", "Categories of goods (Group 2 and Group 3)" summary tables. I analyzed data on the tables for "Days of the week", "Hours of shopping", "Months". Next, Karina began to form the analytical base table. Karina was collecting data for the purchased goods, and I was collecting data for the placed goods. Then we combined the two received tables and got the analytical base table. Then, Karina did Analytical Base table and gave it to me. With such a table I could try to implement the algorithm of k-means and all methods for preparing data for this algorithm. Then, the main goal is to analyze the obtained result and prepare to make Financial and Economic report (FER). Karina gave me the most popular combinations of the items that clients buy in the internet shop; the method called market basket analysis helps to implement it. My part was to sum up all the data and calculate the valuable results like circulation, sales, revenue from realization, gross profit, direct costs and so one. Together we invented mechanics for our companies – it is how company will promote and organize the companies for increasing the cost of check of number of items in check and select the best personal offer for each profile. Therefore, I could implement financial-economic report to prove that our mechanics for retail companies will work.

* 1. **Instruments:**

1. The unfiltered data for our project was given to us by our mentor <https://drive.google.com/open?id=1-OTO8kN5qdbMVQuNF_eRlLQ9RezcrOdc>
2. Python, Anaconda, jupyter notebook
3. Python libraries: numpy, pandas, matplotlib, sklearn, tqdm, seaborn, scipy, statsmodels
4. Microsoft Excel for Mac, version 16.33
5. Google documents
6. PowerPoint

**1.4 Main result:**

The main result of the project is an attempt to find a solution to an actual applied data analysis task. This project proposes to develop an approach using machine learning methods based on real data. We should develop a full system for segmentation of client base and definition of clients with the greatest probability of purchase in a category, create its description for business. Another purpose is selection of the best personal offer for each profile (Next Best Offer).

**2. Review and comparative analysis**

**2.1 Review of abstracts which you used for project and review of analogues:**

The project is dedicated to learning how to filter datasets and implement got data in the next purposes. There is an analogue to cope with the task. It is SAS instruments that our mentor uses at her work. Also, she works a lot of through excel tools – she created pivot tables and use sorting that is installed in excel. However, we decided to apply got knowledge of python and a powerful library ‘pandas’ . It is faster than use excel tools. On the other hand, we need to use excel and learn its implementations through online resources on YouTube that helps to understand how to operate with excel. Modern data analysis uses clustering algorithms based on separation. Our algorithm is the McKeen average method (k-means clustering; MacQueen, 1967), in which each of the k clusters are represented by a centroid. This method is spread and common use. Thus, we will try to implement it and get the result. What do companies need this for? To get to know their clients better. To find an individual approach to each client, rather than working with everyone in the same way.

Despite the fact that many companies use loyalty programs and have colossal data, their analysts first determine the person of the customer, and then analyze its behavior. As part of the task, we had to perform clustering without a teacher. One of the most popular methods is K-Means. This method allows us to do:

- Segmentation of the market (types of buyers, loyalty).

- Association of close points (shops, customer addresses, buildings) on the map.

- Analysis and markup of new data

- Image compression/

- Identify abnormal behavior detectors.

There are also Mean-Shift, DBSCAN models, but their result is difficult to interpret.

**3. Selection of methods, algorithms, and models for project implementation**

* Information about filtering of data:

Firstly, I got unfiltered data called ‘data.csv’ in the format of .csv which I opened in Microsoft Excel, and my first step was to convert the file with data to the readable file for Mac OS as our mentor works with Microsoft formats. I worked with Python3 in Jypiter Notebook as it is powerful tool to work with datasets:

The data consists of 797192 rows × 38 columns, where rows are orders and columns different ratings, including economic and user data of the order. Then, I checked that there are some bags in a table, and I need to filter it. I used commands:

df = df[(df['Количество'] > 0)]

df = df[(df['Отменено'] != 'Да')]

Rows in the 'Cancelled' column where the 'yes' is standing - we are not interested in customers who have canceled the order to compile an analytical base table.

df = df[(df['СуммаЗаказаНаСайте'] != 0) | (df['СуммаДокумента'] != 0)]

df = df[(df['КоличествоПроданоКлиенту'] > 0)]

Got the table consists of 627124 rows × 38 columns. It is table of all placed orders. This is cleaned table without null and uninterested rows.

I am creating a new table. However, to simplify my work I used excel commands to make some tasks. In the excel, I edit and remove the empty values in the "Region" column, replacing them with values from the "Store City" column. I save the result in the table via excel. The next step is removing goods without a region as it is not interesting for us to analyze such orders as they do not gibe necessary information about users. I decided to delete all the lines where it says "Delivery" in the column "Nomenclature" and respectively in the column "Group2", because delivery cannot be a complete order. The result is the filtered table which is prepared to make an analytical base table and make clustering. It is final size is 473504 rows × 42 columns called 'filtered\_dataAll.csv'

* Creating pivot table

First of all, I checked unique orders - #127 113. Further, I wrote a code for a function that will generate pivot tables with unique orders placed on the given site. There I passed on different column names each time to generate separate summary tables on different indicators. I am creating a prototype function to create a spreadsheet, so that I can then change the parameters and get the data. In the function we prescribe methods to search for one or another parameter. In this Dataset, the numbers are presented through a separator "," which prevented the function from working. For this purpose, in the table we used the function in the excel: command in excel is "=ПОДСТАВИТЬ (T2;",";".")". So, thus I generated several pivot tables for placed orders. The next step is creating some visualizations to these tables. I do it with excel tools like pie charts, graphs.

* Analyze pivot tables

1. Analysis of pivot tables by month:

In the data we got only two months: July (07) 2017 and August (08) 2017. For both months, the percentage of revenue for redeemed goods relative to placed goods is approximately 60%, which is a good indicator. A little more than half of the goods (57% on average) were redeemed from all placed items, which may be due to the fact that the number of redeemed items and unique cheques almost halved when filtering the table due to the removal of many items (orders) with erroneous data.

The percentage of unique cheques and clients is slightly higher in August (70%) than in July (68%), which is quite a high figure. This means that every month new customers come, demand and revenue increases.

*Hypothesis:* In the summer months, the buyout rate and other relative indicators are satisfactory.

Изображение выглядит как снимок экрана

Автоматически созданное описаниеИзображение выглядит как снимок экрана

Автоматически созданное описание

1. Analysis of summary tables by days of the week:

We noticed from the summary tables of purchased goods that the beginning of the week is the most profitable for our shop in terms of maximum revenue. The leaders of the days of the week with maximum revenue and margin are Monday, Tuesday, Wednesday and Thursday. The average revenue on these days is around 46,000,000 ₽. The share of redemption on these days fluctuates around 18%. However, it is interesting that the highest rate of redemption from our online store is on Sunday 66%. The number of checks aims at 18,000 ₽ these days. On the least profitable days in terms of revenue - Friday, Saturday and Sunday. Revenue is 29 669 190 ₽, 27 939 397 ₽ and 32 493 680 ₽, respectively. The number of receipts is 8,000 - 9,000 per day. The number of unique clients is half as much. And because of this, the value of customers these days is the highest. The small number of clients may be due to the fact that on weekends people rest, spend less time in the mail and websites (do not read the mailing lists), rest from computers and gadgets and spend less time at home shopping in general. While on weekdays people are more likely to check their mail and place orders (business). Also, it may be justified by the fact that many people know and believe that online stores at the weekend do not work and will deliver anyway on weekdays - so do not make orders. It may be worth making a mailing list and alerts to work at the weekend as usual and raise sales levels.

*Hypothesis:* The first half of the week (Monday, Tuesday, Wednesday, Thursday) has the highest revenue and margin.

*Hypothesis:* Customers spend less time shopping at online stores at weekends because redemption rates are lower, revenue and margin are lower, and other indicators are lower.

*Hypothesis:* It is better to do stock alerts, mailings and notifications during the first half of the working week in order to get the most out of it.

*Hypothesis:* The client's value is the same on about all days of the week.

Изображение выглядит как снимок экрана

Автоматически созданное описаниеИзображение выглядит как карта, текст

Автоматически созданное описание

1. Analysis of summary tables by hours:

From the summary tables, we can see that the shopping time on the site is definitely changing and has trends during the day. The top hours that make the largest redeemed purchases (with the highest revenue, margin, absolute margin and high average check) include: 12:00, 14:00, 15:00, 17:00, 19:00, 22:00. Worst hours of shopping on our website: from about 1:00 to 7:00. These days, most customers are obviously sleeping and so no one is shopping. However, there are exceptions such as 1:00, 2:00. It is during these hours that the revenue is high - 3-4 times more than during the hours between 1:00 and 7:00.

*Hypothesis:* 1:00, 2:00 - hours of site overload and order status updates/processing that were made the night before.

*Hypothesis:* In the evening, starting from 18-00, the revenue from the site remains at a normal level and holds until 0:00 - this is the time when customers come from work / business. They do their homework, have a rest and also check the notifications. You can make an evening newsletter too, so that the customer will pay attention exactly.

*Hypothesis:* Since the customer categories on the children's goods website are based on parents - at this time 12:00, 14:00, 15:00, 17:00 the child is in kindergarten / school and the parent can make a purchase at a time not occupied by the child for him.

*Hypothesis:* It is better to make promotions, newsletters and notifications in the hours before 'rush hour shopping' on our website, i.e. in the morning from 8-00 to 12-00 while people wake up, check their mail, smartphones, read without being distracted by our homework/work offers. To attract more unique customers.

*Hypothesis:* In the afternoon from 12:00 to 15:00, many customers have a lunch break to place an order.

Изображение выглядит как снимок экрана

Автоматически созданное описание

**Analytical Base** (**ABT**) - a flat table that is used for building analytical models and scoring (predicting) the future behavior of a subject. It was done by Karina and gave to me to interpret and make a k-means algorithm.

* Normalization/Standardization

As the data the was given in the analytical base table was ordinary for a human, but for implementing algorithms to make clusters it is needed to normalize it somehow. I used three different methods to normalize data: MinMaxScaler, Standard Scaler and RobustScaler in python libraries.

Here are little descriptions of my types of normalization:

1. StandardScaler standardizes a feature by removing the mean and dividing each value by the standard deviation. Results in a distribution with a standard deviation equal to 1 (and variance equal to 1). If you have outliers in your feature (column), normalizing your data will scale most of the data to a small interval.
2. RobustScaler standardizes a feature by removing the median and dividing each feature by the interquartile range. Outliers have less influence than with MinMaxScaler. Range is larger than MinMaxScaler or StandardScaler.
3. MinMaxScaler adds or subtract a constant. Then multiply or divide by another constant. MinMaxScaler subtracts the minimum value in the column and then divides by the difference between the original maximum and original minimum. Preserves the shape of the original distribution. Doesn't reduce the importance of outliers. Least disruptive to the information in the original data. Default range for MinMaxScaler is 0 to 1.

Scale generally means to change the **range** of the values. The shape of the distribution doesn’t change. Think about how a scale model of a building has the same proportions as the original, just smaller. Standardization generally means changing the values so that the distribution **standard** deviation from the mean equals one. It outputs something very close to a normal distribution. Scaling is often implied.

Many machine learning algorithms perform better or converge faster when features are on a relatively similar scale and/or close to normally distributed.

**StandardScaler:** it shows the best results and I decided to use it. It standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation. StandardScaler results in a distribution with a standard deviation equal to 1. The variance is equal to 1 also, because variance = standard deviation squared. And 1 squared = 1. StandardScaler makes the mean of the distribution 0. About 68% of the values will lie be between -1 and 1. I decided that deep learning algorithms often call for zero mean and unit variance. Regression-type algorithms also benefit from normally distributed data with small sample sizes.

The part of used code:

scaler = StandardScaler()

scaled\_data = scaler.fit\_transform(df)

The data had looked like:

Изображение выглядит как снимок экрана

Автоматически созданное описание

It has become after a StandardScaler:

Изображение выглядит как снимок экрана

Автоматически созданное описание

* Table of correlations

The **correlation coefficient (ρ)** is a measure that determines the degree to which two variables' movements are associated. The most common correlation coefficient, generated by the Pearson product-moment correlation, may be used to measure the linear relationship between two variables.

To calculate correlation, one must first determine the Covariance of the two variables in question. Next, one must calculate each variable's standard deviation. The correlation coefficient is determined by dividing the covariance by the product of the two variables' standard deviations.

Standard deviation is a measure of the dispersion of data from its average. Covariance is a measure of how two variables change together, but its magnitude is unbounded, so it is difficult to interpret. By dividing covariance by the product of the two standard deviations, one can calculate the normalized version of the statistic.

This is the correlation coefficient: Correlation = *ρ* =

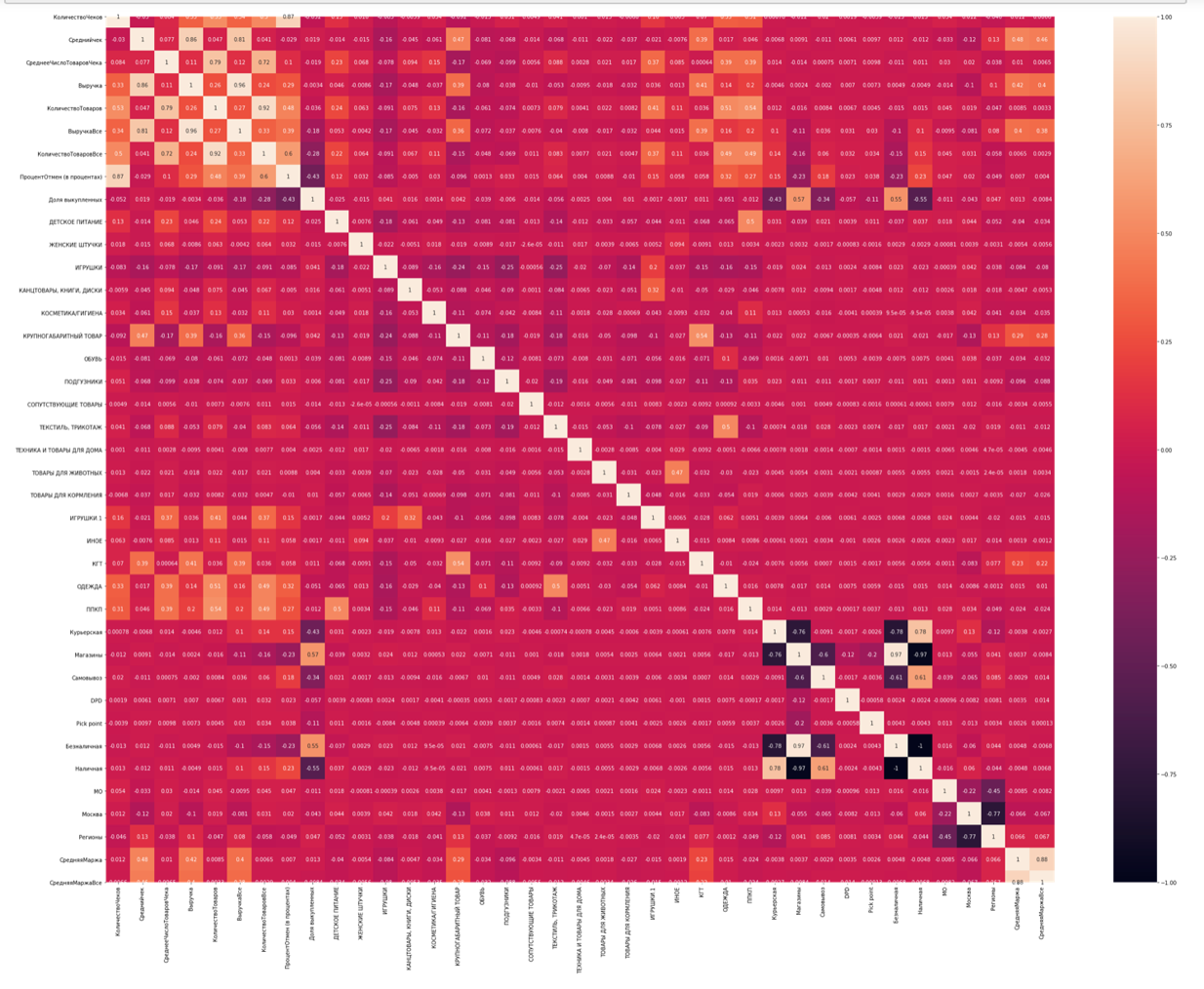
This type of statistic is useful in many ways in finance. In the case of the task it helps to determine the strong relationship between some values that would make the clustering easier and more accurate.

I used Python and its tools to make the correlation table (matrix) - a correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

The functions: creates a table. Then, I tried to visualize it with colors that will also help to

df.corr()

interpret the table. Then I construct a visualization map for this table:



and get as the result a big table with colorful indicators, The lighter the color (white) – the bigger correlation between variables. The table is correct as in the diagonal the white sells as here correlation the variable with itself. The next step is to determine and delete variables with string correlation with the help of the function:

def get\_corr\_cols(df, treshold):

corr\_matrix = df.corr().abs()

upper = corr\_matrix.where(np.triu(np.ones(corr\_matrix.shape), k =1).astype(np.bool))

to\_drop = [column for column in upper.columns if any(upper[column] > treshold)]

return to\_drop

def trimm\_correlated(df\_in, threshold):

df\_corr = df\_in.corr(method='pearson', min\_periods=1)

df\_not\_correlated = ~(df\_corr.mask(np.tril(np.ones([len(df\_corr)]\*2, dtype=bool))).abs() > threshold).any()

un\_corr\_idx = df\_not\_correlated.loc[df\_not\_correlated[df\_not\_correlated.index] == True].index

df\_out = df\_in[un\_corr\_idx]

return df\_out

So, now the table is ready for PCA and clusterising.

Thus, the number of columns was reduced from 39 to 29. Some columns were deleted.

* Choosing number of clusters;

**K-Means Elbow Method**

As k-Means is an unsupervised machine learning algorithm that groups data into k number of clusters. The number of clusters is user-defined, and the algorithm will try to group the data even if this number is not optimal for the specific case. Therefore, we have to come up with a technique that somehow will help us decide how many clusters we should use for the K-Means model. The **Elbow method** is a very popular technique and the idea is to run k-means clustering for a range of clusters k (let’s say from 1 to 10) and for each value, we are calculating the sum of squared distances from each point to its assigned center.

When the centers are plotted, and the plot looks like an arm then the “elbow” (the point of inflection on the curve) is the best value of k.

Изображение выглядит как карта, текст

Автоматически созданное описание

Here is the result of plotting, it is difficult to exactly see how many clusters we need – but 4 or 5 clusters are seemed to be optimum.

* Segmentation

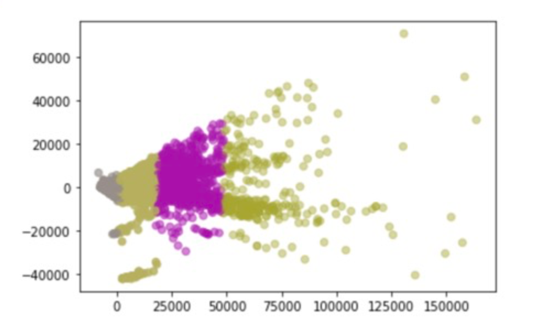
For our experiment we decided to implement k-means algorithms. In this algorithm a cluster refers to a collection of data points aggregated together because of certain similarities. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid.

However, there was a problem as firstly Karina sent me s Analytical Base table with wider range. There were a lot of variables and values. It was formed around ‘Группа3’ – it a subgroup of the ‘Группа2’ and more precisely describes the ‘Группа2’. When I implemented k-means on that table the results were not so accurate – there were 3 or 4 clusters with one segment with the greatest length (about 78% of all positions).

Изображение выглядит как еда

Автоматически созданное описание

It divides like on the graph above. With lengths of clusters equal to 2 605, 6 317 and 57 006 (out of 65 928). It was strange that one cluster is so big. I started to analyze who are people in that clyster – after some manipulations in excel it was seen that it is people with the biggest margin, bills and large items in checks. Then, I tried to make 4 segments and run k-means with it – but the results were more or less the same:



lengths of segments 2 585, 6 891, 1 299 and 55 153. (out of 65 928)

I asked Karina to remake ABT (Analytical Base Table) with not so many items in it. She left the main figures like margin, number of checks, number of items in check, revenue, number of cancels in per cent, share of repurchased goods and average margin plus she deleted variables of ‘Группа3’ and add just ‘Группа2’ and ‘Тип’. The size of new ABT became (65 928 x 39).

I made k-means with number of clusters equals to 4 first with new ABT. The next logical step is to implement PCA (Principal Component Analysis) - as our learning algorithm is too slow because the input dimension is too high, then using PCA to speed it up can be a reasonable choice. After checking the data with PCA we got that there are 2 components and tried to visualize it:

Изображение выглядит как текст

Автоматически созданное описание

The result was the same:

Изображение выглядит как карта, текст

Автоматически созданное описание

with lengths of segments 57 126, 7 483 and 1 319. (out of 65 928)

It was not good results as we'd like to keep any cluster above 50%.

Then, I used it for k-means with python library (from sklearn.cluster import KMeans) with 4 and 5 cluster on the new ABT.

For 4 clusters there is such results:

Implementing a such algorithm is easier with library and also with matplotlib it is not difficult to plot and visualize segmentation. In our case segments are formed some sort of the ‘triangle’:

Изображение выглядит как снимок экрана

Автоматически созданное описание

with length 11 633, 31 542, 9 372 and 13 542 out of 65 928.

Изображение выглядит как снимок экрана

Автоматически созданное описание

Here is the plot my segmentation. Further step was that it was needed to explore every of 4 cluster separately. I calculated length of each segment. Also, I add clusters to the not normalized data frame. As me new data frame war normalized with StandartScaler and it is impossible to analyze segments with such data. I add new column to the data frame before standardization with labels that represent every user to the appropriate cluster 0, 1, 2, 3 in the respect to 1, 2, 3, 4 cluster.

Also, as there is result of 5 clusters:

In this case with k = 5, segments are also formed some sort of the ‘triangle’:

Изображение выглядит как снимок экрана

Автоматически созданное описание

The next step also was to check the representativity of each obtained clusters:

Изображение выглядит как снимок экрана

Автоматически созданное описание

The lengths and percentage of every cluster are:

Cluster 1: ~ 33 366 clients and it is **50%** out of all customers

Cluster 2: ~ 5 696 clients and it is **8.6%** out of all customers

Cluster 3: ~ 16 652 clients and it is **25.2%** out of all customers

Cluster 4: ~ 9 377 clients and it is **14.2%** out of all customers

Cluster 5: ~ 837 clients and it is **1.2%** out of all customers

See that due to the common rule that each cluster should not be more than 10%, cluster # 2 and cluster # 5 are not appropriate to our research.

* Analysis of variance

Dispersion analysis is a method aimed at finding dependencies in experimental data by studying the significance of differences in mean values => use ONE\_WAY ANOVA

The assumptions for implementing one way ANOVA include:

- The normality criterion: each group compared should come from a population following the normal distribution.

- The variance criterion (or 'homogeneity of variances'): samples should come from populations with the same variance.

- Independent samples: performance (the dependent variable) in each sample should not be affected by the conditions in other samples.

We check the hypothesis between 4 groups:

*Null hypotheses:* groups means are equal (no variation in means of groups);

*Alternative hypotheses*: At least, one group mean is different from other groups’;

The way is computing Fisher test:

Where

–

I make the test with all important parameters like ***Average check, Average revenue, Average number of goods in check, Average margin, rare categories of goods, and some important categories;***

Изображение выглядит как рисунок, часы

Автоматически созданное описание all clusters are Gaussians, they have normal distribution.

RESULTS

~ F statistic = 2.283, Probability p = 0.077 for average check, rare categories of goods(pet’s goods and so one): As **p > a (0.05**) we state that we **do not have a main interaction effect**. This simply means that amongst group comparison identifies statistically insignificant differences.

~ F statistic = 1182.103, Probability p = 0 for average revenue: As **p < a (0.05**) we state that we **have a main interaction effect**. This simply means that amongst group comparison identifies statistically significant differences.

~ F statistic = 5206.679, Probability p = 0 for average number of goods: As **p < a (0.05**) we state that we **have a main interaction effect**. This simply means that amongst group comparison identifies statistically significant differences.

~ F statistic = 4124.682, Probability p = 0 for average margin and other rest of categories of goods: As **p < a (0.05**) we state that we **have a main interaction effect**. This simply means that amongst group comparison identifies statistically significant differences.

To sum up, We make this conclusion in most tests the average value of each of the features in each of the clusters do NOT coincide. So the interclusters are different, so we've clustered them right. We have checked that all 4 classes that are formed cannot be merged back into one common sub-sample or merged. This will be a sign that the clusters are constructed unambiguously and can be rebuilt.

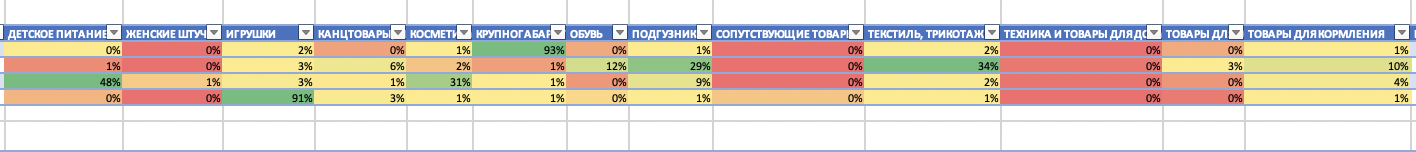
* Analyzing segments and creating representative graphs

To better understand my segments, I looked on the means of the obtained cluster table:

Изображение выглядит как снимок экрана

Автоматически созданное описание

I moved to the Excel to work with obtained data. Firstly, I reorganized data in a more representative way – added color representation of the product to see the most outstanding items. Also, I used the data representation to make reading and understanding values more comfortable and understandable: I used money representation, share representations and so one.



Furthermore, I calculated the same table but with standard deviation with Excel tool, it helped to create table with average deviations – and I represents it in per cents:

=(B3-СРЗНАЧ(B$3:B$6))/СТАНДОТКЛОН.В(B$3:B$6)

Изображение выглядит как снимок экрана

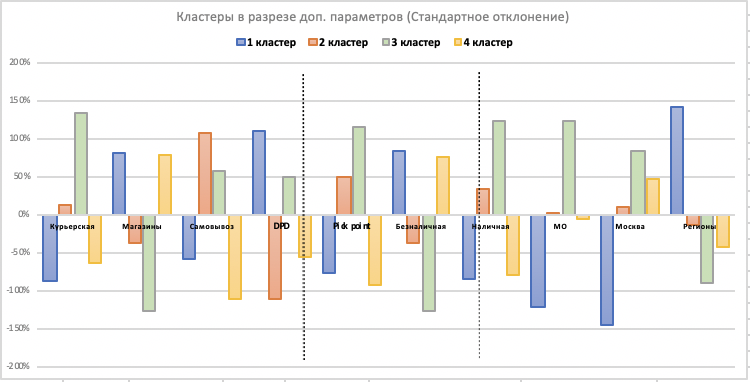
Автоматически созданное описание

Here is the little part of the obtained table above.

To better understand the given results, I used Excel manipulations to build different representatives graphs, Firstly, there graphs below show the average information about whole 4 segments:

Изображение выглядит как снимок экрана

Автоматически созданное описание



**See that we get interpreted 4 segments separately:**

Изображение выглядит как снимок экрана

Автоматически созданное описание

Cluster 1: there are ~9 372 clients (which is 14% out of all clients), from the graph above we know that they buy КГТ and (it is picks with grey color) – and mostly they buy categories in ‘Группа3’ such as АВТОКРЕСЛА(11%), ВЕЛОСИПЕДЫ/САМОКАТЫ(9%), ДЕТСКАЯ МЕБЕЛЬ, МАТРАСЫ(19%), КОЛЯСКИ(23%), ДЕТСКИЕ СТУЛЬЯ ДЛЯ КОРМЛЕНИЯ(9%). It is the most ‘mixed-haired’ group that consumes different products.

* This cluster mostly from the ‘РЕГИОНЫ’ and ‘МОСКВА’
* Clients in this segment use ‘DPD’ delivery way;
* The way of payments mostly ‘НАЛИЧНАЯ’;
* Average check is 8 814 ₽;
* Average number of items in a check is 1.5;
* Average margin is 2 760 ₽;
* Revenue or net profit is 9 558 ₽;

Изображение выглядит как кирпич

Автоматически созданное описание

Cluster 2: there are ~29 936 clients (which is 45% out of all clients), from the graph above we know that they buy ПОДГУЗНИКИ, ОБУВЬ, ТЕКСТИЛЬ/ТРИКОТАЖ, ТОВАРЫ ДЛЯ ЖИВОТНЫХ, ИНОЕ, ОДЕЖДА, КАНЦТОВАРЫ/КНИГИ/ (it is picks with grey color) – and mostly they buy categories in ‘Группа3’ such as ВЕРХНЯЯ ДЕТСКАЯ ОДЕЖДА(8%), ПОДГУЗНИКИ (27%), ОДЕЖДА ДЛЯ НОВОРОЖДЕННЫХ (0-2 лет) (7%), ОБУВЬ ДЕТСКАЯ (12%));

* This cluster mostly from ‘МОСКВА’;
* They use ‘DPD’ delivery, ‘МАГАЗИНЫ’ and ‘САМОВЫВОЗ’;
* The way of payments both equals – ‘БЕЗНАЛИЧНАЯ’ and ‘НАЛИЧНАЯ’;
* Average check is 2 354 ₽;
* Average number of items in a check is 2.66;
* Average margin is 335 ₽;
* Revenue or net profit is 2 900 ₽;

Изображение выглядит как снимок экрана, компьютер

Автоматически созданное описание

Cluster 3: there are ~10 080 clients (which is 15% out of all clients), from the graph above we know that they buy КОСМЕТИКА/ГИГИЕНА, ЖЕНСКИЕ ШТУЧКИ, ДЕТСКОЕ ПИТАНИЕ, ППКП (it is picks with grey color) – and mostly they buy categories in ‘Группа3’ such as ДЕТСКАЯ КОСМЕТИКА(9%), ЗАМЕНИТЕЛИ МОЛОКА(17%), КАШИ(9%),ПОДГУЗНИКИ(8%), ПЮРЕ(18%), СРЕДСТВА БЫТОВОЙ ХИМИИ(7%), СРЕДСТВА ГИГИЕНЫ(6%),ТОВАРЫ ДЛЯ МАМ(6%) ;

* This cluster mostly from ‘МО’, ‘РЕГИОНЫ’;
* They use ‘Pick Point’, ‘Магазины’ and ‘Курьерская’;
* The way of payments both equals – ‘БЕЗНАЛИЧНАЯ’ and ‘НАЛИЧНАЯ’;
* Average check is 2 863 ₽;
* Average number of items in a check is 6;
* Average margin is 260 ₽;
* Revenue or net profit is 4 366 ₽;

Изображение выглядит как снимок экрана, белый, компьютер

Автоматически созданное описание

Cluster 4: there are ~16 540 clients (which is 25% out of all clients), from the graph above we know that they buy ИГРУШКИ, СОПУТСТВУЮЩИЕ ТОВАРЫ (it is picks with grey color) – and mostly they buy categories in ‘Группа3’ such as ИГРУШКИ ДЛЯ ДЕВОЧЕК(17%), ИГРУШКИ ДЛЯ МАЛЬЧИКОВ(12%), ИГРУШКИ ДЛЯ РАЗВИТИЯ МАЛЫШЕЙ(28%), КОНСТРУКТОРЫ(19%), НАСТОЛЬНЫЕ ИГРЫ(6%);

* This cluster mostly from ‘Москва’, ‘РЕГИОНЫ’;
* They use ‘Pick Point’, ‘САМОВЫВОЗ and ‘Курьерская’;
* The way of payments both equals – ‘БЕЗНАЛИЧНАЯ’ and ‘НАЛИЧНАЯ’;
* Average check is 2 019 ₽;
* Average number of items in a check is 2.62;
* Average margin is 300 ₽;
* Revenue or net profit is 2 175 ₽;

Изображение выглядит как стол, лодка, сидит, грузовик

Автоматически созданное описание

The same procedure I made with k = 5 clusters and get the results:

1 segment: **baby food 19%, cosmetics hygiene %, diapers 35%** - 33 366 customers, average check **2 450**₽, number of goods **5;**

2 segment: **KGT 93%** - 5 696 customers, average check **8 794**₽, goods **2**;

3 segment: **Toys 82%** - 16 652 customers, average check **2 048**₽, number of goods **4**;

4 segment: **Textile/knitwear 84%** - 9 377 customers, average check **2 593**₽, goods **6**;

5 segment: **nappies 20%** - 837 customers, textile knitwear 22%, baby food 16%, toys 14%, cosmetics 7% - 68 customers, average check **4 282**₽, number of goods **5**;

Interesting fact about 5 segment is: (in this segment a large percentage of cancellations compared to other 2.4% - in all above only 1%).

Thus, we make a hypothesis that k – 5 is not good for our case, as maybe id the fifth segment algorithm sum some outliers or make a subgroup of group#1, as the categories of items that customers buy on segment #5 seems to be similar as in cluster#1.

**Thus, we decided to work with 4 obtained segments from k-means.**

DBSCAN

To check k-means we will implement DBSCAN as it doesn’t require the user to specify the number of clusters.

Algorithm:

1. Starts with an arbitrary point which has not been visited and its neighborhood information is retrieved from the ϵ parameter
2. If this point contains *MinPts* within ϵ neighborhood, cluster formation starts. Otherwise the point is labeled as noise.
3. If a point is found to be a core point, then the points within the ϵ neighborhood is also part of the cluster.
4. The above process continues until the density-connected cluster is completely found.

The main concept is to locate regions of high density that are separated from one another by regions of low density.

**Main points:** A point with a minimum number of min\_samples, the distance from which to the point below the threshold set by the epsilon.

**Boundary point**: A point that is not in close proximity to at least the min\_samples, but close enough to one or more base points. The border points are included in the cluster of the nearest base point.

**Noise point:** Points that are not close enough to the baselines to be considered as border points. Noise points are ignored. That is, they are not part of any cluster.

**Choose eps - size of the neighborhoods**

We can calculate the distance from each point to its closest neighbour using the NearestNeighbors. The point itself is included in n\_neighbors. The kneighbors method returns two arrays, one which contains the distance to the closest n\_neighbors points and the other which contains the index for each of those points.

Then, we sort the results and plot the visualization:

Изображение выглядит как снимок экрана

Автоматически созданное описание

Calculate the average distance by m of the closest neighbors for each point. That is, if m = 4, you need to select the three nearest neighbors, add the distances to them and divide by four. The optimal value for epsilon will be found at the point of maximum curvature.

I worked with 5 % of all dataframe. It show that optimal value of eps could be from 0.25-0.35. Look at the graph.

**DBSCAN**

I also scaled data as for k-means algorithm and implement DBSCAN with eps=0.35 and minimum number of clusters equals to 3.

Изображение выглядит как рисунок

Автоматически созданное описание

Estimated number of clusters: **4**

Estimated number of noise points: **20**

**Clusters received in DBSCAN**

Segment 1: Goods for animals I don't understand why, but it is.

Segment 2: large and baby food (average check about 8 000₽ as in k-means)

Segment 3: cosmetics/hygiene, diapers, feeding products - multicoat cluster, which buys a lot of things (in k-means it was) (average check 3 000₽)

Segment 4: toys, shoes, knitwear/textile (average check 3000₽ too)

There are visualizations to DBSCAN:

Изображение выглядит как снимок экрана

Автоматически созданное описание

Изображение выглядит как снимок экрана

Автоматически созданное описание

Изображение выглядит как снимок экрана

Автоматически созданное описание

Изображение выглядит как снимок экрана

Автоматически созданное описание

Thus, due to the representativity and on the substantial analysis of clusters, we can say that DBSCAN does not exactly approach our problem. K-means worked more accurately.

**Comparison DBSCAN and K-MEANS**

1. **Curse of Dimensionality:** DBSCAN somewhat slower than agglomerative clustering and k-means, and not quite good scales to relatively large datasets
2. **Content analysis:** Substantive validity of groups from k-means is more interpreted
3. **Consistency with expert evaluation:** We have an expert opinion from this online shop, which knows approximately target trends and customer groups
4. **Cluster Complexity:** It is desirable that each cluster be at least 10% and no more than 50%
5. **Correlation with data survey data:** The data study identified dependencies that are similar to the resulting clusters in k-means
6. Outliers can influence DBSCAN more

* Financial and economic report

The final step is to analyze the obtained results from the economic and financial side for the Internet-shop of children goods. The purpose is to understand how our analysis and research could help to increase net profit and which cases and opportunities could we get from the information that we get from the segments, market-basket analysis and next best offer.

Firstly, I learn from the open resources the average SMS cost and e-mail-cost which we could use to promote offerings to our clients from segments.

* SMS cost is 1.5 rubles per one piece;
* E-mail cost is 0,1 ruble per one piece;

(the data from open resource <https://sigmasms.ru/tarify/> and <https://sendpulse.com/ru/prices>)

**SMS marketing vs. email marketing:**

Research shows that[SMS open rates](https://www.gartner.com/smarterwithgartner/tap-into-the-marketing-power-of-sms/) are as high as 98%, compared to just 20% of all emails. And, on average, it takes 90 seconds for someone to[respond to a text](https://www.business2community.com/infographics/email-marketing-vs-sms-marketing-stats-infographic-02021390) and 90 minutes to respond to an email.

Furthermore,[75% of people](http://digitalmarketingmagazine.co.uk/mobile-digital-marketing/7-key-statistics-for-sms-marketing/558) have suggested that they’d be happy to receive an offer via SMS – this is why[65% of marketers](https://www.business2community.com/infographics/email-marketing-vs-sms-marketing-stats-infographic-02021390) say that SMS marketing is a “very effective” method for them. Thus, we decided to choose SMS marketing as the main mechanic for the shop offerings.

In the table below I calculated the main valuable measures like gross profit, direct costs, sales & marketing expenses and final net profit of whole the segments’ offerings. The expenses will be equal to 99 000 ₽, but company will obtain 6 731 000 ₽ from such a mechanic.

Изображение выглядит как снимок экрана

Автоматически созданное описание

There are two ways of doing offerings:

* Check growth mechanics, when we offer a customer to buy an amount of XXX rubles in category 1 (in category 1 and category 2) and get a discount XXX rubles;
* Mechanics for the growth of the number of goods on the check, when we offer a customer to buy an amount of XXX rubles of 2 goods in category 1 (category 1 or category 2) and get a discount XXX rubles;

The parameters in this table are calculated as follows:

* Circulation: it is the number of clients in each segment;
* Response: is an average indicator that shows what percentage of clients respond to our offer. This figure is averaged and taken from open sources and followings - that on average 3% of customers respond to our offer;
* Sales: it is circulation multiplies by response showing how much sales we will get from each segment from the mechanics;
* Revenue from realization: we set up the minimum limit on the amount of purchase like 3 000₽ for Segments #3, #3 and #4 and 9 000₽ for Segment#1 as in our example and multiplied it by the number of sales. It will demonstrate how much money in rubles we will get from the realization of such an action (For Segment#1 is 2 530 000 ₽; for Segment #2 is 2 694 000 ₽; for Segment #3 is 907 000 ₽; for Segment #4 is 1 489 000 ₽);
* Direct costs on communication: the cost of the communication that we will spend on the offerings. It is computed as sales multiplied by discount (For all 4 Segments it costs 790 000₽);
* Gross profit: Gross profit is a measure of profitability that shows the percentage of revenue that exceeds the [cost of goods sold](https://www.investopedia.com/terms/c/cogs.asp) (COGS). It is like revenue from realization minus direct costs on communication (For Segment#1 is 2 249 000 ₽; for Segment #2 is 2 245 000 ₽; for Segment #3 is 816 000 ₽; for Segment #4 is 1 340 000 ₽);
* Sales & Marketing expenses: normally includes salaries, commissions, and benefits to sales and marketing personnel, co-op advertising allowances to customers, advertising, warehouse costs, and shipping costs (For all 4 Segments it costs 99 000 ₽);
* Net Profit: represents the number of [sales](https://investinganswers.com/dictionary/s/sale) remaining after all [operating expenses](https://investinganswers.com/dictionary/o/operating-expense), interest, [taxes](https://investinganswers.com/dictionary/t/taxes) and preferred [stock dividends](https://investinganswers.com/dictionary/s/stock-dividend) (but not common stock dividends) have been deducted from a company's total [revenue](https://investinganswers.com/dictionary/r/revenue) (For Segment#1 is 2 235 000 ₽; for Segment #2 is 2 380 000 ₽; for Segment #3 is 801 000 ₽; for Segment #4 is 1 315 000 ₽);
* Minimum revenue from both 4 mechanics: it is the sum of all net profit of 4 segments – in our case it is 6 731 000 ₽;
* Discount: price or percentage that we choose and set ourselves based on our MBA and segment data;

The financial table clearly show us how much the company could get from the analysis and how much it is needed to expense to realize such a program for next best offering. I concluded that for the Segment #2, #3, #4 the minimum limit on the amount of purchase should be 3 000 ₽ because knowing the data on their average checks - up to 3000 ₽, it would be logical to increase our net profit to hang the price of one check for purchases at the online store. Thus, it will increase the margin and the average check of the client. That will make the company's profit. Based on the MBA, Karina has formulated the data for the offer to customers - so we can determine the categories of goods that can be offered in such a mechanics and the data that I obtained from the segmentation I could form different actions to offer items for clients:

* Segment #2: For the promotion it is definitely worth paying attention to the purchase of goods in the «КОСМЕТИКА/ГИГИЕНА» and «ИГРУШКИ» categories. Already to these categories it is necessary to bind «ПОДГУЗНИКИ», because almost twice as many checks with these goods are bought with «ПОДГУЗНИКИ» then without them. The limit for the check is 3 000 ₽ and the discount is 10%.
* The mechanics for Segment #2 of check growth is: ‘Purchasing a product worth 3000 ₽ from the category of hygiene and cosmetics for moms and diapers, and you get a privilege in the way of 10% on your check’ or vice versa;
* The mechanics for Segment #2 of growth of the number of goods on the check is: ‘Buy 3 products from the category of baby clothes (infant clothes (0-2 years) for the amount from 3000 ₽or products from the category of diapers and get a bonus in the form of 10% discount’ ;
* The mechanics for Segment #2 of check growth is: ‘Purchase from 3,000 ₽ for products in the categories of children clothes and children shoes and get a discount of 10%’;
* The mechanics for Segment #2 of check growth is: ‘Make a purchase from 3,000 ₽ in categories of toys and get a privilege in category top children's clothing in the form of 10%’;
* The mechanics for Segment #2 of growth of the number of goods on the check is: ‘When purchasing 5 items from 3,000 from the cosmetics and hygiene category, you will receive a 10% privilege for the diaper category’ and vice versa.
* Segment #1: For the promotion it is definitely worth paying attention to the purchase of goods in the «ТЕКСТИЛЬ, ТРИКОТАЖ, «ТОВАРЫ ДЛЯ КОРМЛЕНИЯ» and «ИГРУШКИ» categories. Already to these categories it is necessary to bind «КРУПНОГАБАРИТНЫЙ ТОВАР» as this category allocates the first segment. Promotions for this category may contain the following context: when buying "«КРУПНОГАБАРИТНЫЙ ТОВАР» for the amount of 9 000, get a discount in the way 10% on the categories of «ТЕКСТИЛЬ, ТРИКОТАЖ» and « ИГРУШКИ»”. The limit for the check is 9 000 ₽ and the discount is 10%.
* The mechanics for Segment #1 of check growth is: ‘Purchase a stroller for your baby and children's furniture for the amount of a check from 9,000 ₽ and get a 10% discount on the category of goods top children's clothing or bed linen’;
* The mechanics for Segment #1 of growth of the number of goods on the check is: ‘Buy 3 products from the category of chairs for feeding babies for the amount (chairs, food stands, study stands, bibs) from 9 000 ₽ and get 10% discount’;
* Segment #3: For the promotion it is definitely worth paying attention to the purchase of goods in the «ТОВАРЫ ДЛЯ КОРМЛЕНИЯ» and «ДЕТСКОЕ ПИТАНИЕ» categories. Already to these categories it is necessary to bind «ПОДГУЗНИКИ» as this category allocates the first segment. Promotions for this category may contain the following context: when buying "«ТОВАРЫ ДЛЯ КОРМЛЕНИЯ» for the amount of 3 000 ₽, get a discount in the way 10% on the categories of «ДЕСТКОЕ ПИТАНИЕ» or « ПОДГУЗНИКИ»”. The limit for the check is 3 000 ₽ and the discount is 10%.

- The mechanics for Segment #3 of check growth is: ‘When you buy products from 3,000 ₽ or more from the category of baby feeding products, accessories for baby feeding machines, electrical appliances, you will get the privilege of a 10% discount from the diaper category’ or vice versa;

- The mechanics for Segment #3 of growth of the number of goods on the check is: ‘Buy 5 products for the amount from 3 000 ₽ in the category of porridge and milk substitutes and puree and get 10% discount’;

- The mechanics for Segment #3 of check growth is: ‘Buy baby feeding and nutrition products on the check more than 3 000 ₽ and get 10% discount on the whole check’;

- The mechanics for Segment #3 of growth of the number of goods on the check is: ‘Buy 3 products from the category of household chemicals and baby care on the check more than 3 000 ₽ and get 10% discount from the category of baby food’;

* Segment #4: For the promotion it is definitely worth paying attention to the purchase of goods in the « ИГРУШКИ » and « КОНЦТОВАРЫ, КНИГИ, ДИСКИ» categories. Already to these categories it is necessary to bind « СОПУТСТВУЮЩИЕ ТОВАРЫ » as this category allocates the first segment. Promotions for this category may contain the following context: when buying "«ТОВАРЫ ДЛЯ КОРМЛЕНИЯ» for the amount of 3 000 ₽, get a discount in the way 10% on the categories of «ДЕСТКОЕ ПИТАНИЕ» or « ПОДГУЗНИКИ»”. The limit for the check is 3 000 ₽ and the discount is 10%.
* The mechanics for Segment #3 of growth of the number of goods on the check is: ‘Buy 3 products from the category of toys for 3 000 rubles and get 10% discount on the category of diapers’ or vice versa;
* The mechanics for Segment #3 of check growth is: ‘Purchase from 3 000 in the department of children's clothing and get 10% discount on the category of goods toys for boys or girls’.
* The mechanics for Segment #3 of growth of the number of goods on the check is: ‘Buy 4 products in the category of children's creativity and toys-constructors for the amount from 3 000 rubles and get 10% discount’
* The mechanics for Segment #3 of growth of the number of goods on the check is: ‘Buy goods in the category of active recreation with a child and board games for the amount from 3 000 rubles and get 10% discount’.
* The mechanics for Segment #3 of check growth is: ‘Buy goods in the category of educational toys for kids from 3 000 rubles and get 10% discount on diapers’ .

Thus, we have identified a large number of types of shares into categories that have been analyzed by IBA through analysis and segmentation. Each offer is lined up based on the average customer's segmentation check, his main purchases there, the number of goods on the check and the most purchased categories of goods. The next step will be to highlight some of the most successful areas of product promotion, as the implementation of all individually quite complex task in terms of sales and capacity.

1. **Results and conclusion**

In the research project, I mastered many new tools, concepts, approaches and libraries: Our project was devoted to the study of check data quality using the example of check data in retail and client segmentation and behavioral profiling. Thus, I personally learned about how client analytics is tripled in real business. I learned to select client segments taking into account preferences and profitability by building and applying results of mathematical models and sharpening of Lifestyle segmentation. Identification of client profiles based on purchases. And the most important is the formation of a strategy to retain the existing customer base through effectively structured communications with customers for each segment. Mentor taught how to work with articles, introduced machine training, gave real advice to a business analyst. Since theory and knowledge often differ from practice, she introduced and told about how it is done in real companies. We studied the very beginning of econometrics, machine learning, segmentation, data analysis, data cleansing. The project was useful because it can be used as a portfolio for an interview. The main thing is that we have a real result based on mathematical models, which can bring the customer-business to revenue.

All sources and code are stored here:

**https://github.com/natashakrauze27/Research-project-2020-DSBA.git**

1. **List of sources**

<https://habr.com/ru/company/datawiz/blog/248863/>

<https://habr.com/ru/company/sas/blog/430404/>

<https://compscicenter.ru/courses/data-mining-python/2018-spring/classes/> lecture from 1 to 8

<https://habr.com/ru/post/67078/>

<https://m.habr.com/ru/post/423597/>

<https://www.sql.ru/blogs/sasbi/1252>

<https://docs.google.com/spreadsheets/d/1woVi7wq13628HJ-tN6ApaRGVZ85OdmHsDBKLAf5ylaQ/edit#gid=0> – table with different normalizations

<https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02>

<https://sigmasms.ru/tarify/>

<https://sendpulse.com/ru/prices>

<https://www.bitdegree.org/learn/train-test-split>

<https://stackoverflow.com/questions/24147278/how-do-i-create-test-and-train-samples-from-one-dataframe-with-pandas>

<https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>

<https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/>

[**https://scikit-learn.org/stable/auto\_examples/cluster/plot\_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py**](https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py)

<https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60>

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>

<https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html>

<https://scikit-learn.org/stable/auto_examples/cluster/plot_optics.html#sphx-glr-auto-examples-cluster-plot-optics-py>

<https://www.geeksforgeeks.org/ml-optics-clustering-explanation/?ref=rp>

<https://stats.stackexchange.com/questions/88872/a-routine-to-choose-eps-and-minpts-for-dbscan>

<https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py>

https://tsput.ru/res/informat/mop/lections/lection\_7.htm

<http://contactretail.apple.com>

<http://pytolearn.csd.auth.gr/d1-hyptest/12/anova-one.html>

<https://reneshbedre.github.io/blog/anova.html>

presentations of our mentor

All sources and code are stored here:

**https://github.com/natashakrauze27/Research-project-2020-DSBA.git**