Optimization of customer list for communication using mathematical modeling

Project of 2nd year students of FCS DSBA:

- Alexander Shirnin – DSBA181
- Maxim Shishov – DSBA182
- Gregory Antonovsky – DSBA182

Mentor of the project:

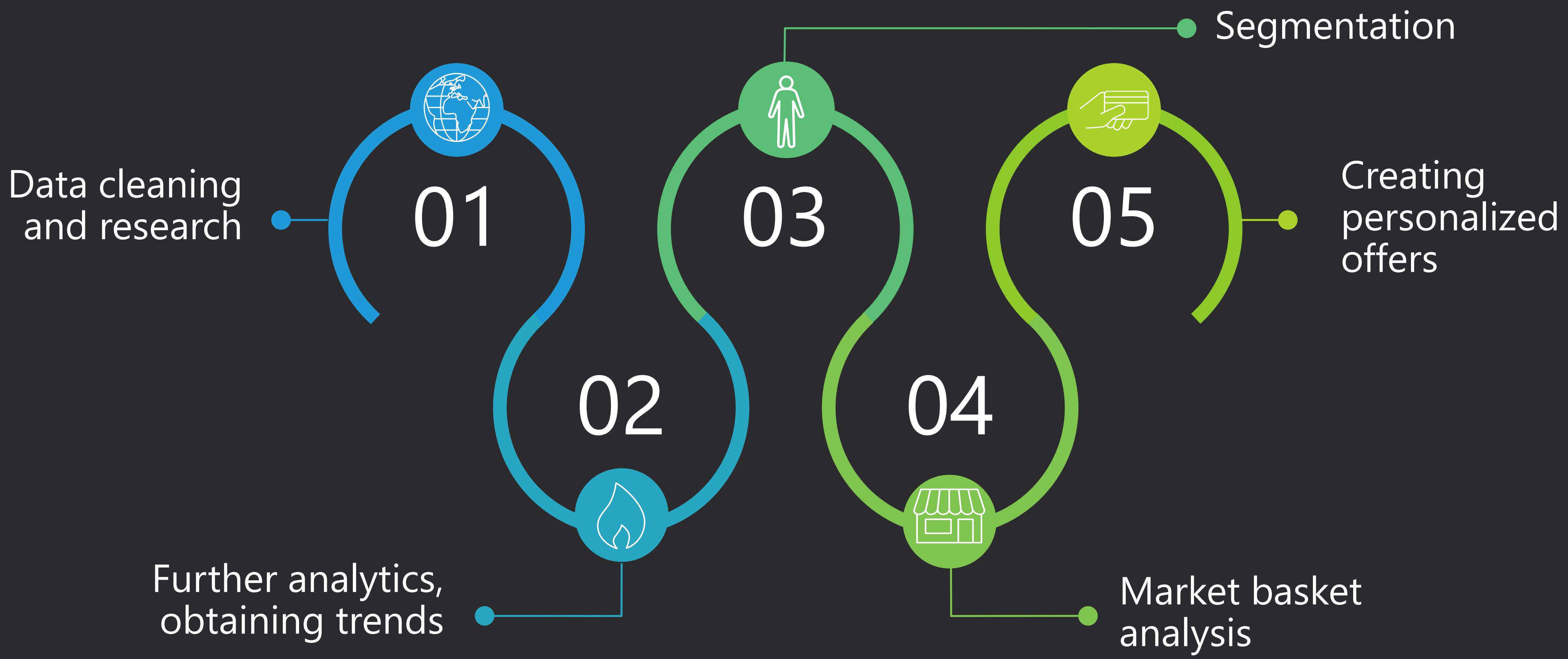
- Titova Natalya

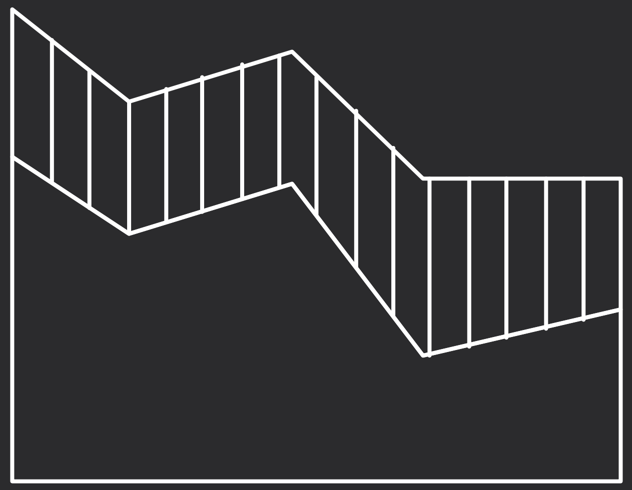
Project goals

Goal: Optimization of customer lists for communication using mathematical modeling

Tasks:
- Data filtering (Alexander)
- Data analysis (Maxim)
- Data mart and segmentation (Gregory)
- Providing Market basket analysis (Alexander)
- Creating personalized offers (Maxim)

Project plan





Data cleaning

Obtained data

Transactional data from the shop network about various goods with the following information:

- Client’s phone number, name, city
- Good price and quantity
- Delivery method
- Sales data
- Reasons why sale was cancelled

Tasks

There are issues with unprepared data:

- Missing information (NaN)
- Incorrect information
- Wrong format
- Information about customers that did not buy item
- Repeating data

Data cleaning

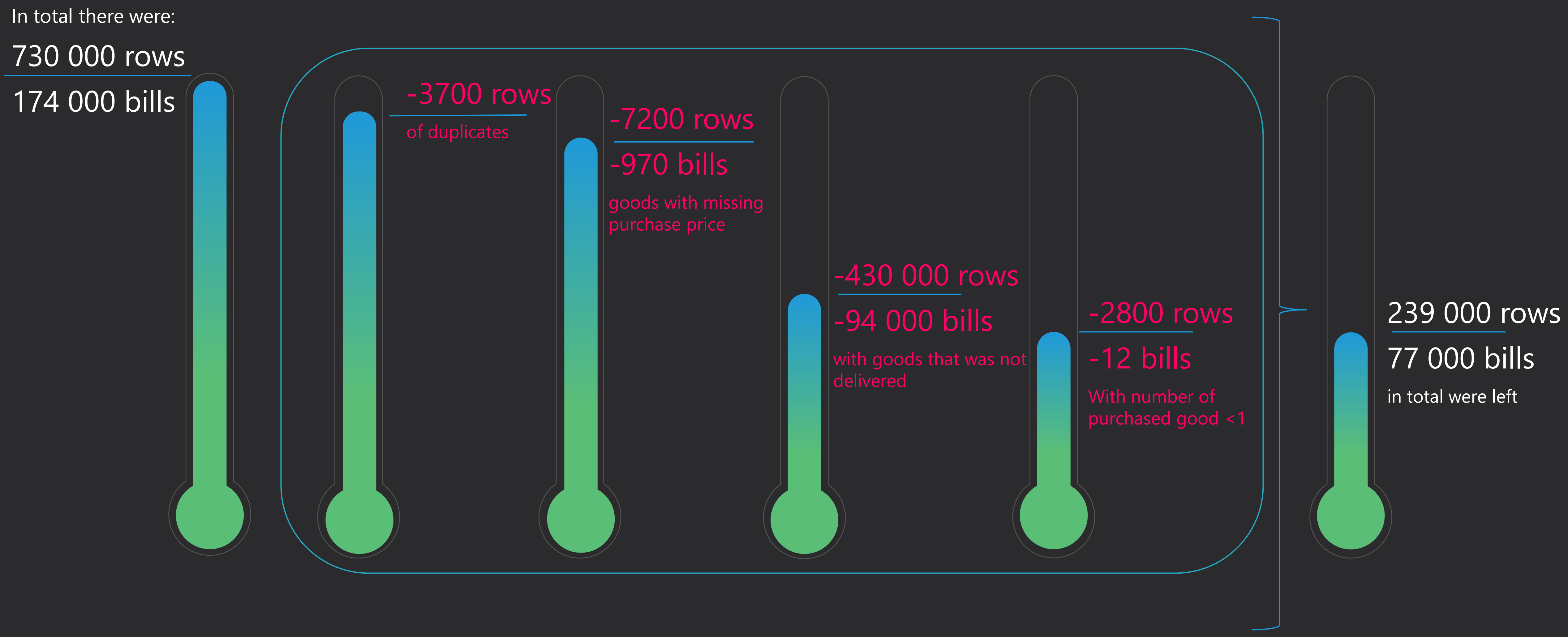
missing information

	Категория	%
0	МагазинЗаказа	99.297934
1	ГородМагазина	99.297934
2	ПричинаОтмены	90.142193
3	ПВЗ_код	28.561866
4	Группа4	17.498405
5	Маржа	14.033520
6	ЦенаЗакупки	14.033520
7	Группа3	13.623011
8	ТипТовара	13.623011
9	Группа2	13.623011

It's important to investigate reasons of missing values and decide to leave them or not:

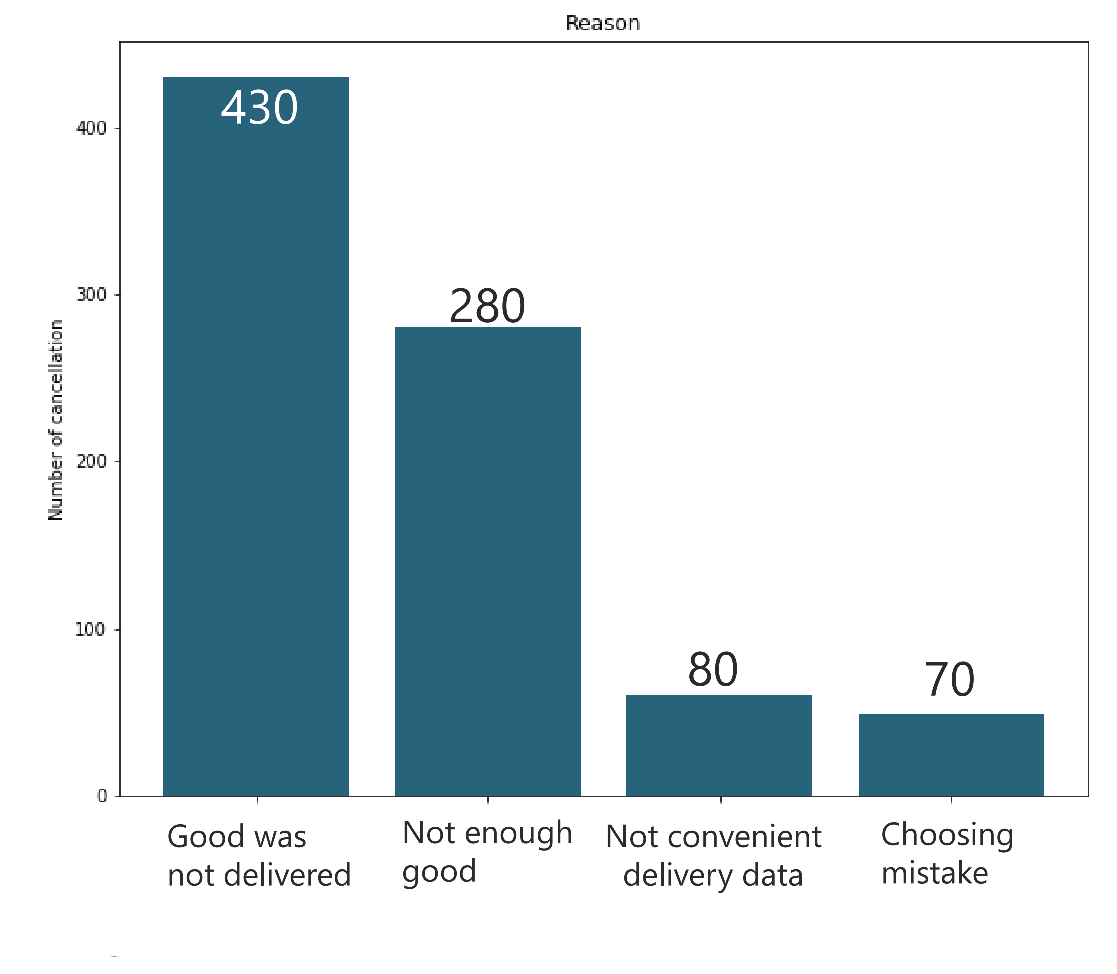
- Data is not available
- Specific data collection
- City or shop information is not shown when customer offers delivery

Data filtering
statistics



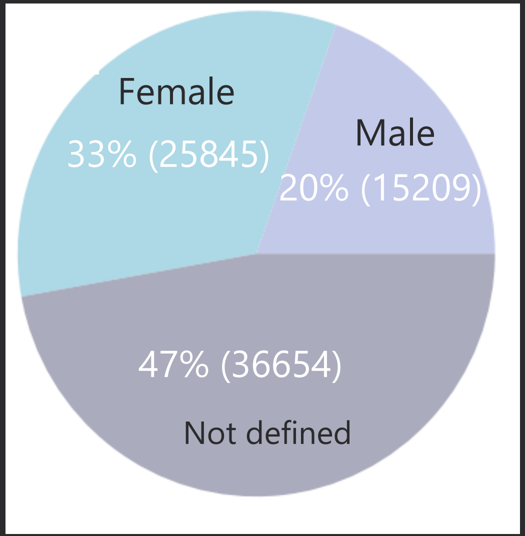
Data research and recommendations to the shop

Reasons of offer cancellations



Recommendations:

- All bought goods were purchased using cashless payment. Shop owners should check correctness of their data mining system. Otherwise, specify on customers, who doesn't use cash
- Focus on big cities in Russia, since more than 35% offers were there
- About 50% customers do not say their name, so shop might create special form of offer, that customer write it, it might be used for specialized offers



Analytics and trends

Data analytics and trends

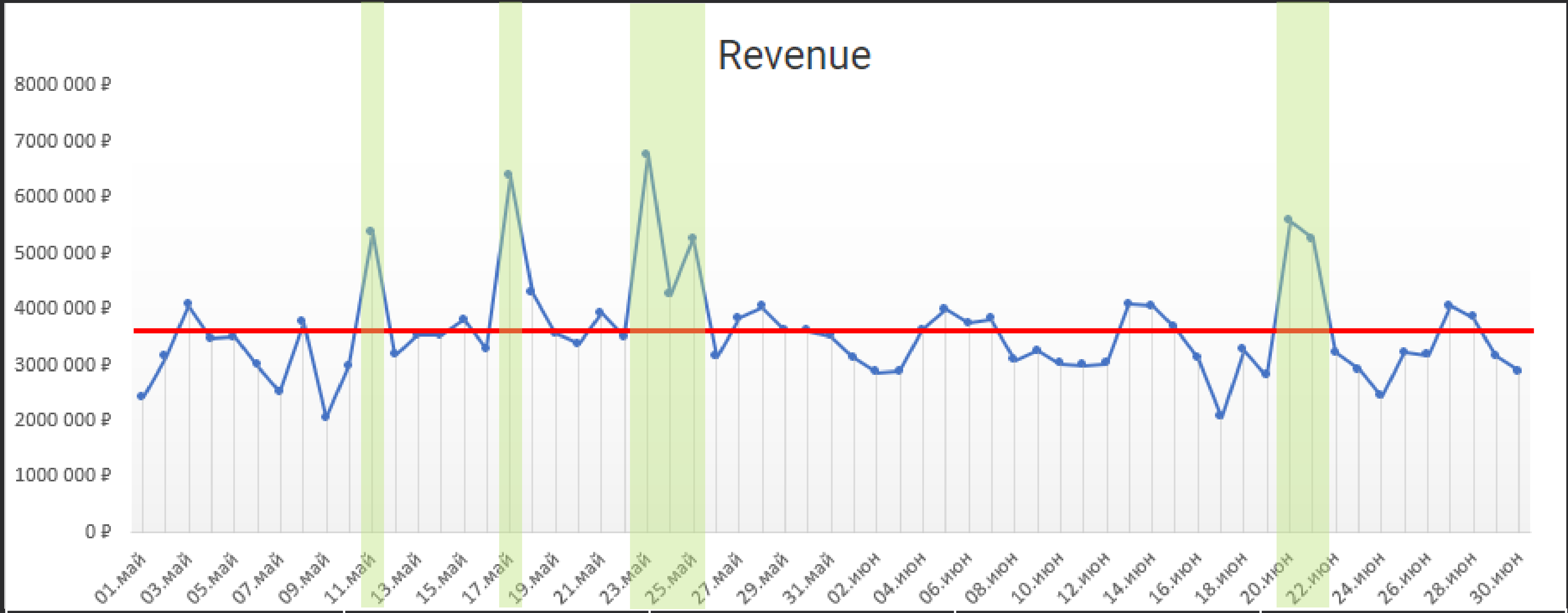
Total Revenue:
218.9 mln.

Total Profit:
49.4 mln.

% of Profit to Revenue:
22.5%

Average № of orders:
14.4 thousands

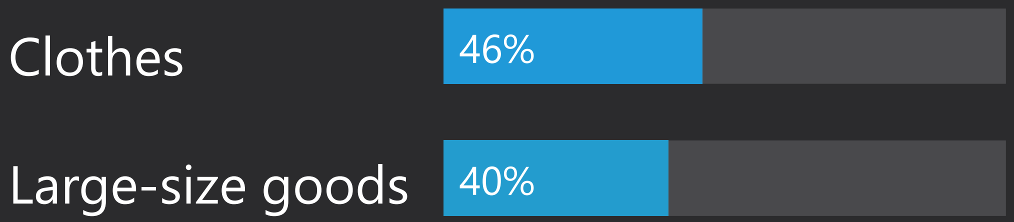
Average Revenue:
3.6 mln.



	Total Profit (Mill. of rubles)	Total Revenue (Mill. of rubles)	Unique orders (thousands)	Total quantity of good (thousands)
May	116.6	27.5	124.9	179.6
June	102.3	21.9	110.7	159.1
Changes	-14.3	-5.6	-14.2	-20.5

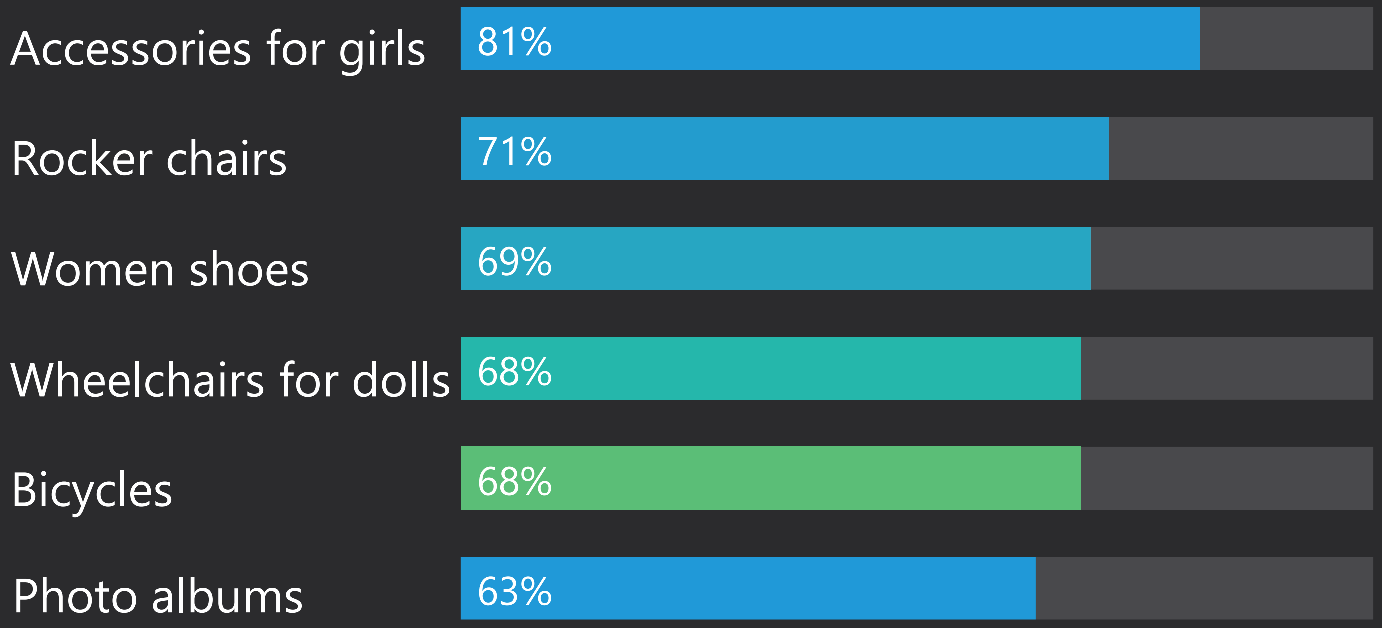
To begin with, firm we are given is highly profitable, as value for money, which is (Profit/Cost)*100%, is 28.9%, which is higher than average Russian firm in 2019.

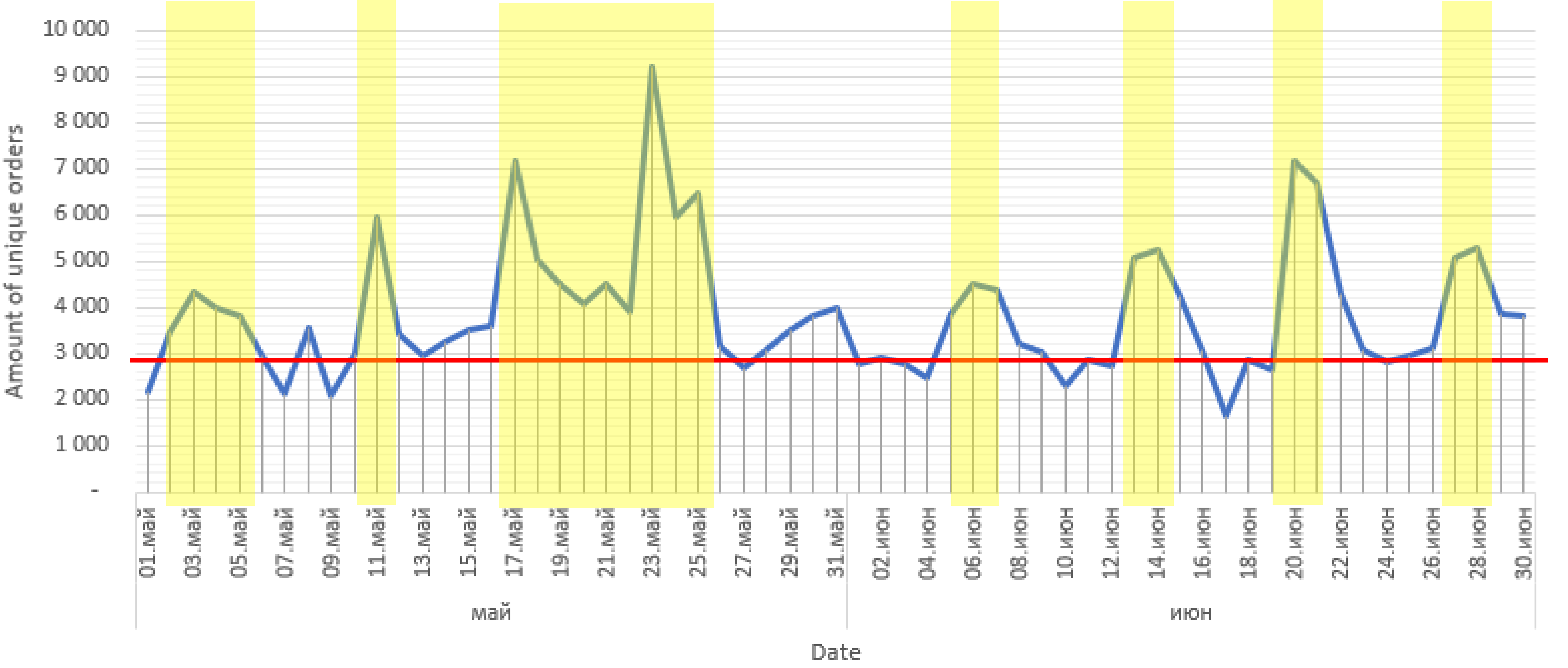
Profit/Cost is highest for categories:



Why all these goods? Because they are not so expensive to produce and must be bought not so many times.

Profit/Cost is highest for goods:

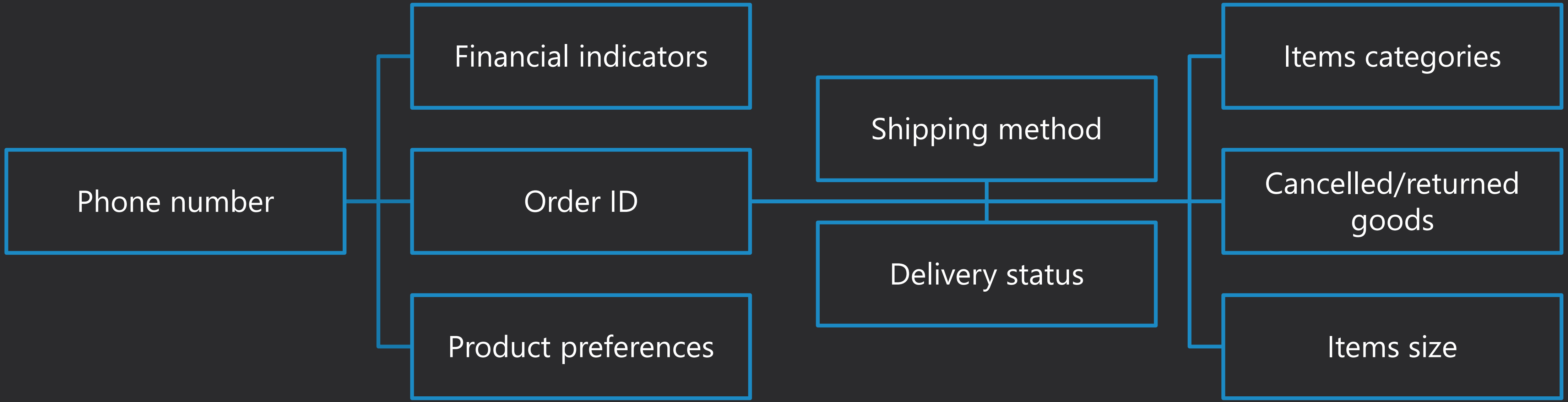




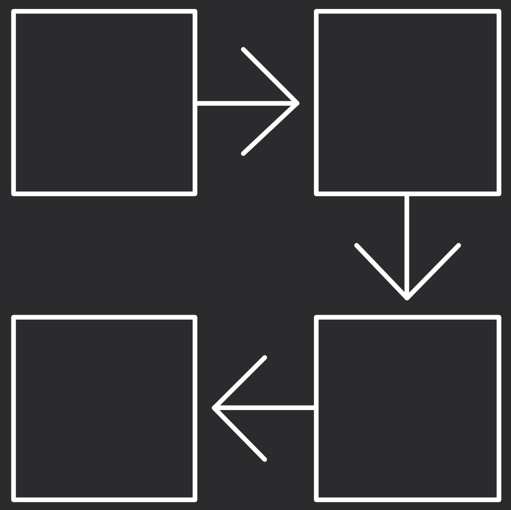
On 23 of May, we had an increase in all types of goods, but the greatest increase was in toys, with deviation of around 300% from average. Same for the 11th, 17th and 25th

Data mart

A total of 45 parameters are derived for each of ~60k client, forming a data mart used in segmentation and MBA.
For example: avg. number of items per order, total sum charged for shipping, avg. item price, etc.



To improve the quality of the data mart, strongly correlating parameters were removed, leaving 36.



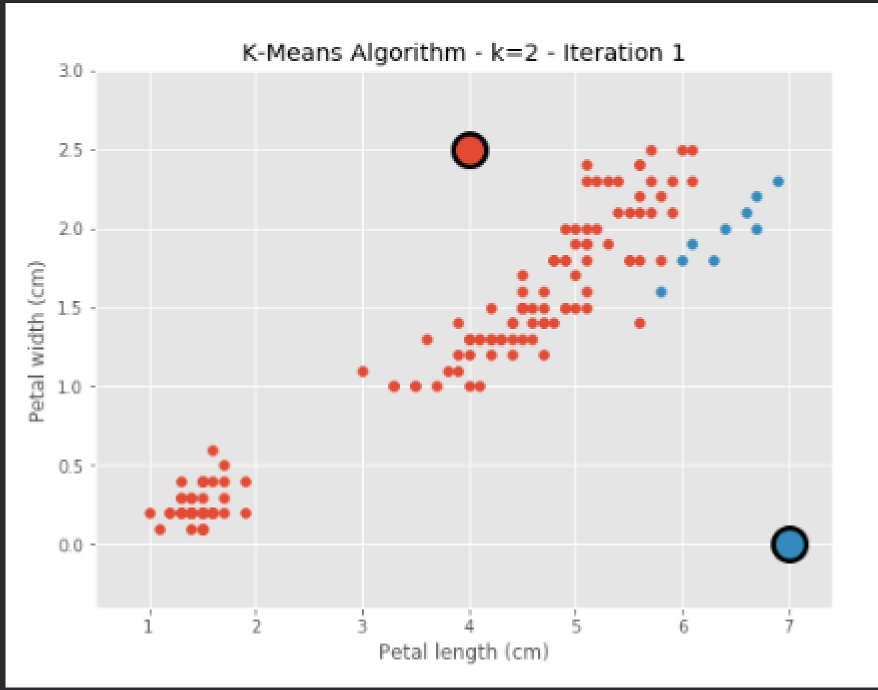
Segmentation

K-means

K-Means starts by randomly defining *k* centroids.

Loop:

- Assign each data point to the closest corresponding centroid
- The mean of values in cluster becomes the new value of the centroid



Number of clusters

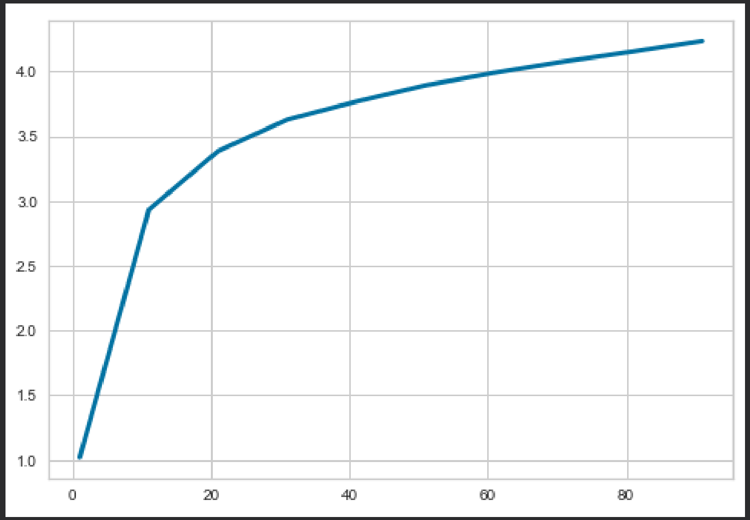
Before segmentation, the data was normalized, using min-max scaler.

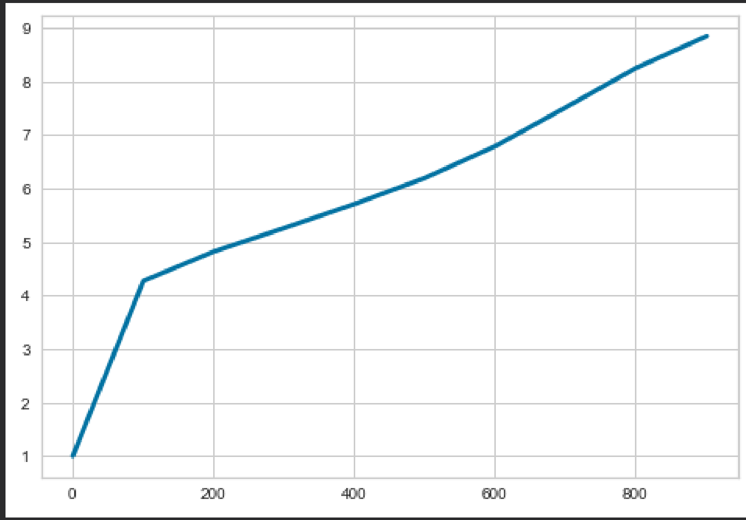
To determine the best number of clusters, the gap method was used initially.

Gap statistic compares the total within intra-**cluster** variation for different values of k with their expected values under null reference distribution of the data

$$Gap_n(k) = E^*_n\{\log W_k\} - \log W_k$$

- $E^*_n\{\log W_k\}$ – variation under reference data with a random uniform distribution
- $\log W_k$ – variation in observed data





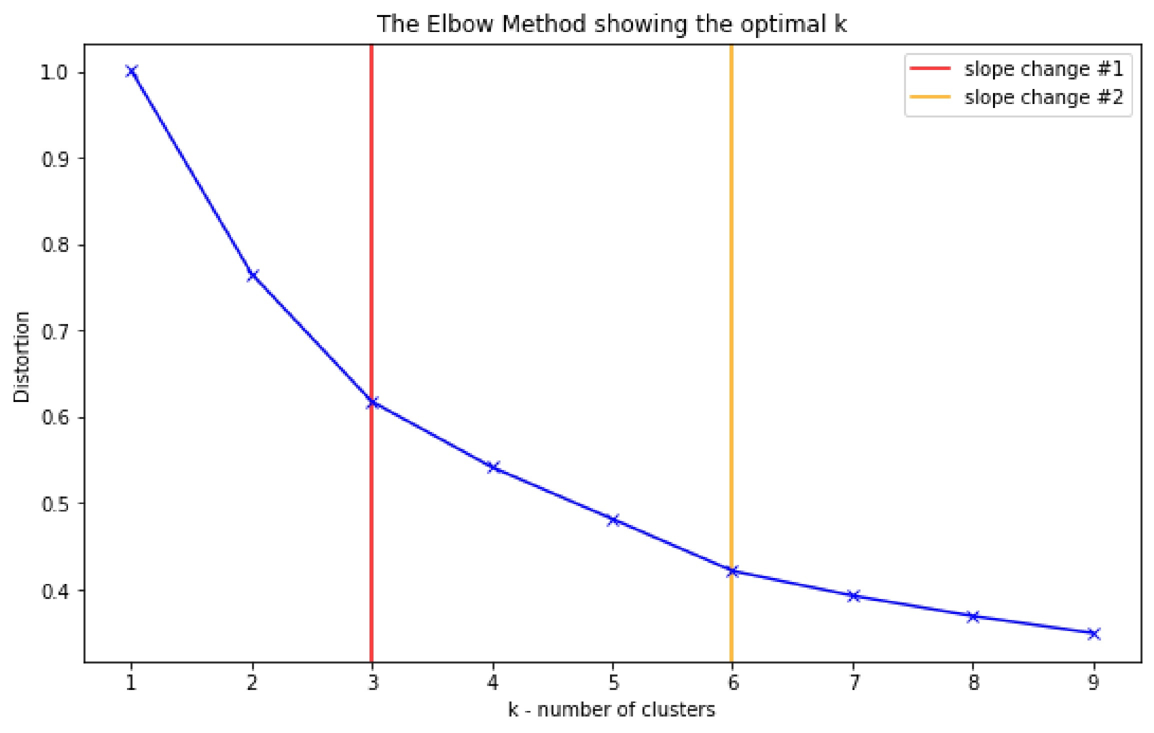
Number of clusters

Before segmentation, the data was normalized, using min-max scaler.

To determine the best number of clusters, the Elbow method was used.

- For each k, calculate the total within-cluster sum of square.
- Plot it.
- The location of "bend" is considered indicator of appropriate number of clusters

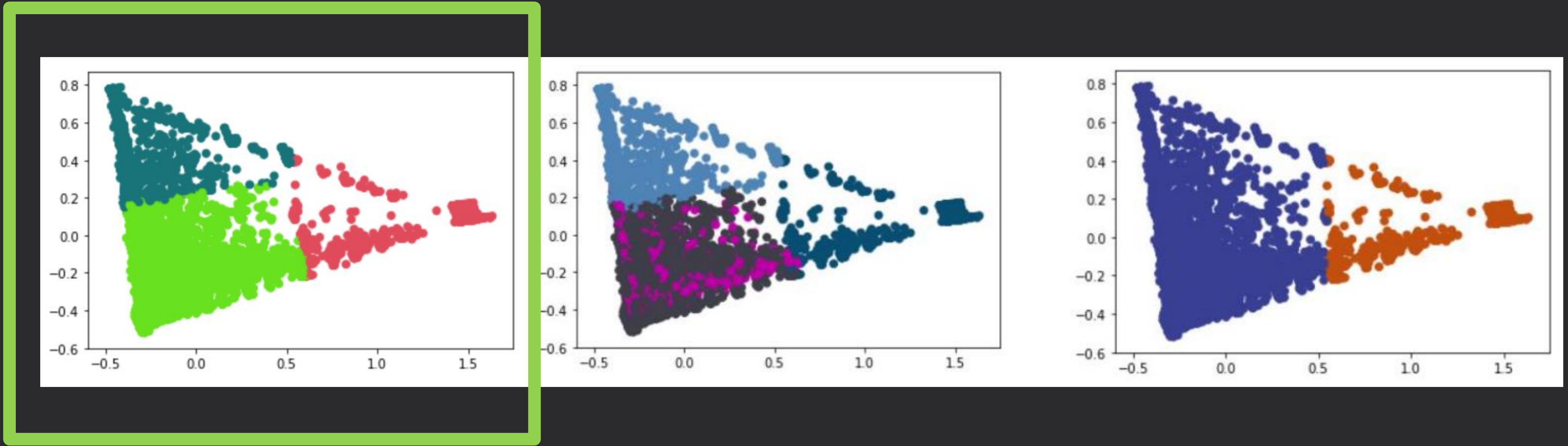
Slope changes at K=3, 6



$\Sigma_i \min(||x_i - \mu_j||^2)$, x_i – i-th item, μ_j – j-th cluster

K-means, PCA results

Principal component analysis plots after K-means clusterization for K = 3, 4 and 2



Final clusters distribution: 58%, 25%, 17%

DBSCAN

DBSCAN was tested as an alternative clustering method and performed poorly.

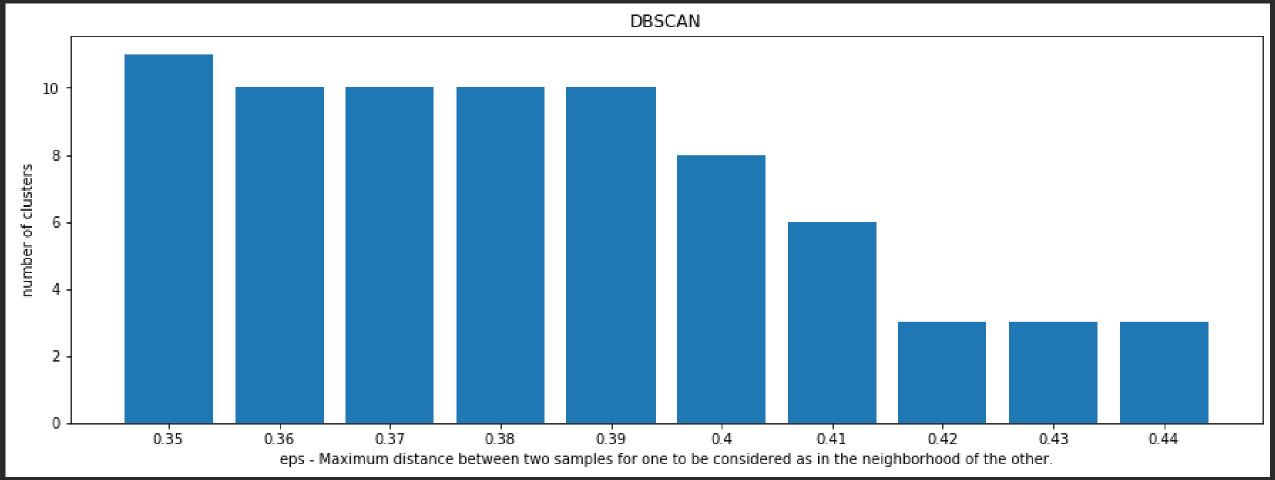
Bad cluster distribution. The model assigned 82% of data to one cluster.

Performs badly on data with different density.

Long compute time

Parameter e selection

e – maximum distance between two samples for one to be considered as in the neighborhood of the other.



Clusterization results

03

10200 (17%) clients
- **97%** oversized goods
- Highest avg. check – **8180rub**
- Avg. item price – **7000 rub**

02

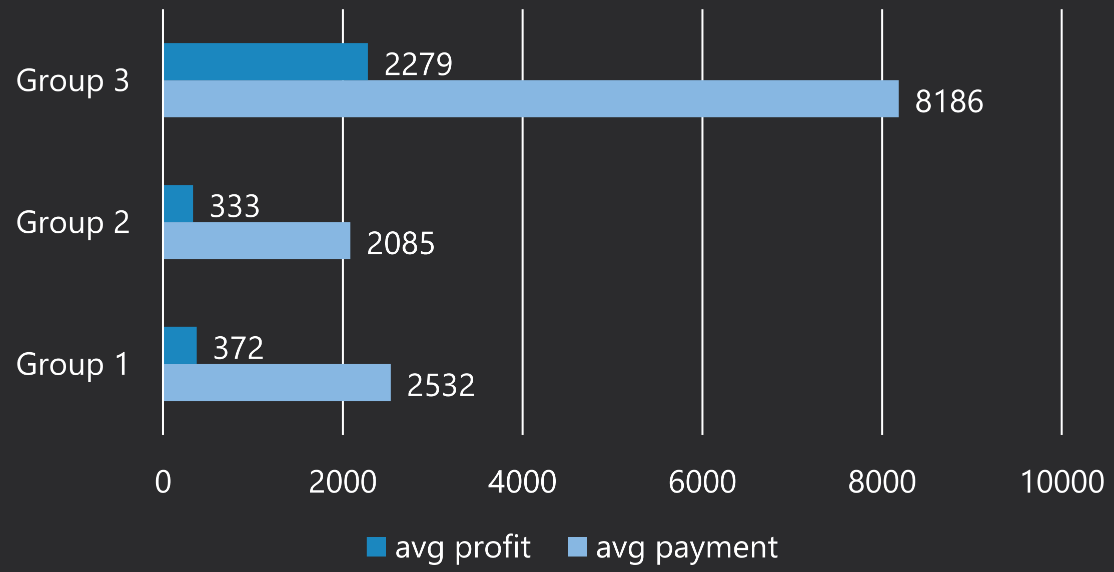
15000 (25%) clients
- Highest service charges - **40%**
- Most popular categories – **Toys 92%**

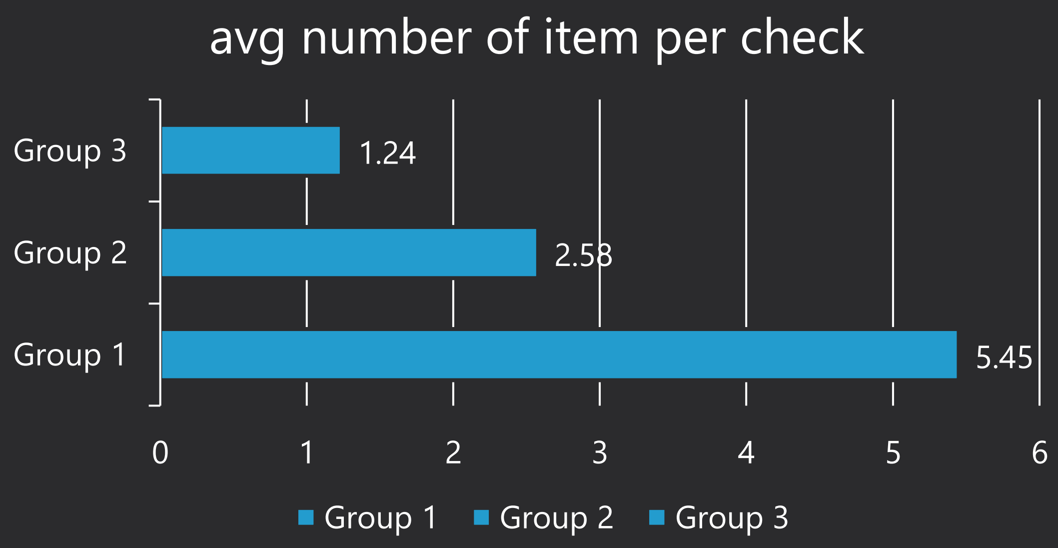
01

34800 (58%) clients
- Lowest avg. item price – **935rub**
- Lowest profit – **485rub**
- Most popular categories: diapers, textile, baby food

*Data summary for each of 3 clusters

Clusterization results





Dispersion analysis

We test the null hypothesis that the mean values of each of the parameters in each of the clusters are equal.

To test the hypothesis, we use the Fisher's test.

$F = MSE_{between} / MSE_{within}$

MSE_{within} compares N observations to the overall mean
$MSE_{between}$ compares k means to the overall mean (k – number of clusters)

$MSE_{within} = \Sigma_{k_i}\Sigma_{n_j}(Y_{ij} - mean(Y_i))^2 / (N-k)$
$MSE_{between} = \Sigma_{k_i}(n_i \times mean(Y_i) - mean(Y)^2) / (k-1)$

n_i – number of entries in i-th cluster;
N – number of entries total;
k – number of clusters;
Y_{ij} – j-th entry in i-th cluster;

As a result, we rejected the null hypothesis for all of 36 parameters, at 1% significance level.

Market Basket Analysis

Goals of MBA

What can be done with the result?

Main goal – obtain popular models of shopping, change parameters in order to increase quantity and quality of purchases

- Create personalized offers for target audience
- Optimize placement of goods on the shells in shops
- Increase or decrease prices on specific goods
- Create and manage advertising campaings

Market Basket Analysis
theory

Definition:

Analysis of market baskets (MarketBasket Analysis) - set of analytical approaches for understanding customer behavior, choosing products, determining associations and relationships between pair of products in each bill, probability of buying both goods.

Input information:

A and B that are goods, events or groups of goods
A – reason, B – consequence; In other words, If A happens, then B happens.

Main formulas for computation:

- Support = Total(A and B) in transactions / Total number of transactions
  Shows how often pair of goods appears
- Confidence = Total(A and B) in transactions / Total(A) in transactions
  Shows how often there is B in bill, if there is A
- Expected confidence = Total(B) in transactions / Total number of transactions
  How often B appears
- Lift: Confidence = Expected confidence
  Shows how many times more customer buy good B, if they buy A, then without A in a bill.

Steps of MBA

Data preprocessing:

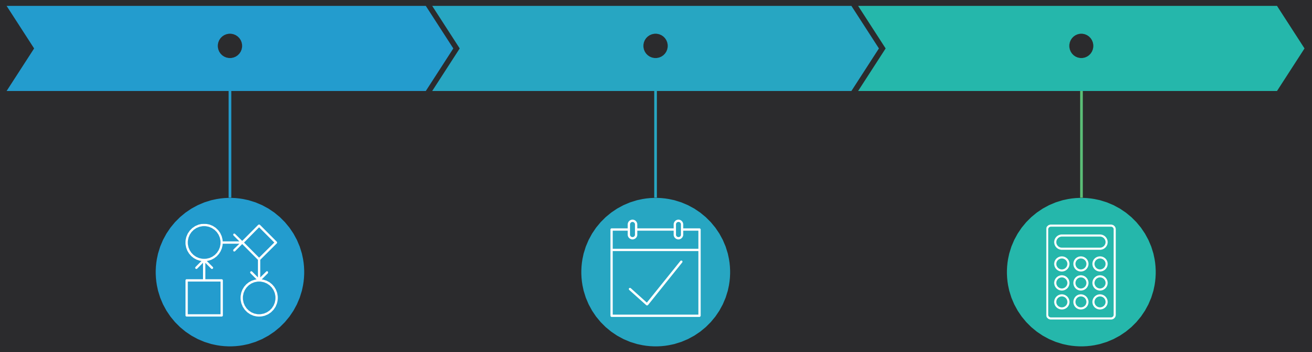
Group all transactions by clients or bills

Creating data mart:

Fill table with 0 and 1 depending on purchases

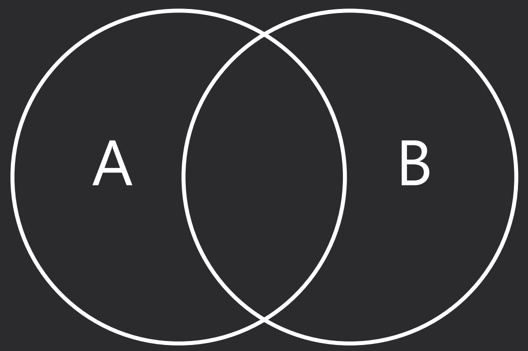
Computing MBA:

Apply appropriate algorithm



Used tools: Python + pandas + itertools

Comparison of resulting tables for one segment



MBA for clients

	Пара	support(%)	confidence(%)	lift	expected confidence(%)
54	КОСМЕТИКА/ГИГИЕНА + ПОДГУЗНИКИ	11.950857	57.332402	1.324530	43.285103
88	ПОДГУЗНИКИ + КОСМЕТИКА/ГИГИЕНА	11.950857	27.609631	1.324530	20.844857
6	ДЕТСКОЕ ПИТАНИЕ + ПОДГУЗНИКИ	9.982823	49.366542	1.140497	43.285103
84	ПОДГУЗНИКИ + ДЕТСКОЕ ПИТАНИЕ	9.982823	23.062954	1.140497	20.221841
95	ПОДГУЗНИКИ + ТОВАРЫ ДЛЯ КОРМЛЕНИЯ	8.978427	20.742534	0.992734	20.894349
151	ТОВАРЫ ДЛЯ КОРМЛЕНИЯ + ПОДГУЗНИКИ	8.978427	42.970601	0.992734	43.285103
30	ИГРУШКИ + ПОДГУЗНИКИ	8.040991	49.819625	1.150965	43.285103
86	ПОДГУЗНИКИ + ИГРУШКИ	8.040991	18.576809	1.150965	16.140208
59	КОСМЕТИКА/ГИГИЕНА + ТОВАРЫ ДЛЯ КОРМЛЕНИЯ	7.912894	37.960894	1.816802	20.894349
148	ТОВАРЫ ДЛЯ КОРМЛЕНИЯ + КОСМЕТИКА/ГИГИЕНА	7.912894	37.870977	1.816802	20.844857
32	ИГРУШКИ + ТЕКСТИЛЬ, ТРИКОТАЖ	7.645055	47.366522	1.426561	33.203296

MBA for bills

	Пара	support(%)	confidence(%)	lift	expected confidence(%)
54	КОСМЕТИКА/ГИГИЕНА + ПОДГУЗНИКИ	8.805148	51.018463	1.279196	39.883228
88	ПОДГУЗНИКИ + КОСМЕТИКА/ГИГИЕНА	8.805148	22.077320	1.279196	17.258748
6	ДЕТСКОЕ ПИТАНИЕ + ПОДГУЗНИКИ	7.557255	39.689052	0.995131	39.883228
84	ПОДГУЗНИКИ + ДЕТСКОЕ ПИТАНИЕ	7.557255	18.948454	0.995131	19.041158
95	ПОДГУЗНИКИ + ТОВАРЫ ДЛЯ КОРМЛЕНИЯ	5.764566	14.453608	0.874661	16.524814
151	ТОВАРЫ ДЛЯ КОРМЛЕНИЯ + ПОДГУЗНИКИ	5.764566	34.884300	0.874661	39.883228
59	КОСМЕТИКА/ГИГИЕНА + ТОВАРЫ ДЛЯ КОРМЛЕНИЯ	5.310226	30.768314	1.861946	16.524814
148	ТОВАРЫ ДЛЯ КОРМЛЕНИЯ + КОСМЕТИКА/ГИГИЕНА	5.310226	32.134859	1.861946	17.258748
30	ИГРУШКИ + ПОДГУЗНИКИ	4.711977	34.859316	0.874034	39.883228
86	ПОДГУЗНИКИ + ИГРУШКИ	4.711977	11.814433	0.874034	13.517125

MBA for the segment

	Пара	support(%)	confidence(%)	lift	expected confidence(%)
113	ТЕКСТИЛЬ, ТРИКОТАЖ + КРУПНОГАБАРИТНЫЙ ТОВАР	1.192735	70.967742	0.723938	98.03018
149	ТОВАРЫ ДЛЯ КОРМЛЕНИЯ + КРУПНОГАБАРИТНЫЙ ТОВАР	0.966838	66.459627	0.677951	98.03018
28	ИГРУШКИ + КРУПНОГАБАРИТНЫЙ ТОВАР	0.939731	72.222222	0.736735	98.03018
52	КОСМЕТИКА/ГИГИЕНА + КРУПНОГАБАРИТНЫЙ ТОВАР	0.560224	79.487179	0.810844	98.03018
89	ПОДГУЗНИКИ + КРУПНОГАБАРИТНЫЙ ТОВАР	0.551188	70.930233	0.723555	98.03018
77	ОБУВЬ + КРУПНОГАБАРИТНЫЙ ТОВАР	0.307220	65.384615	0.666985	98.03018

Here the most interesting pair is textile + furniture

MBA for the segment

Pairs for consideration

	Пара	support(%)	confidence(%)	lift	expected confidence(%)
54	КОСМЕТИКА/ГИГИЕНА + ПОДГУЗНИКИ	8.805148	51.018463	1.279196	39.883228
88	ПОДГУЗНИКИ + КОСМЕТИКА/ГИГИЕНА	8.805148	22.077320	1.279196	17.258748
6	ДЕТСКОЕ ПИТАНИЕ + ПОДГУЗНИКИ	7.557255	39.689052	0.995131	39.883228
84	ПОДГУЗНИКИ + ДЕТСКОЕ ПИТАНИЕ	7.557255	18.948454	0.995131	19.041158
95	ПОДГУЗНИКИ + ТОВАРЫ ДЛЯ КОРМЛЕНИЯ	5.764566	14.453608	0.874661	16.524814
151	ТОВАРЫ ДЛЯ КОРМЛЕНИЯ + ПОДГУЗНИКИ	5.764566	34.884300	0.874661	39.883228
59	КОСМЕТИКА/ГИГИЕНА + ТОВАРЫ ДЛЯ КОРМЛЕНИЯ	5.310226	30.768314	1.861946	16.524814
148	ТОВАРЫ ДЛЯ КОРМЛЕНИЯ + КОСМЕТИКА/ГИГИЕНА	5.310226	32.134859	1.861946	17.258748
30	ИГРУШКИ + ПОДГУЗНИКИ	4.711977	34.859316	0.874034	39.883228
86	ПОДГУЗНИКИ + ИГРУШКИ	4.711977	11.814433	0.874034	13.517125

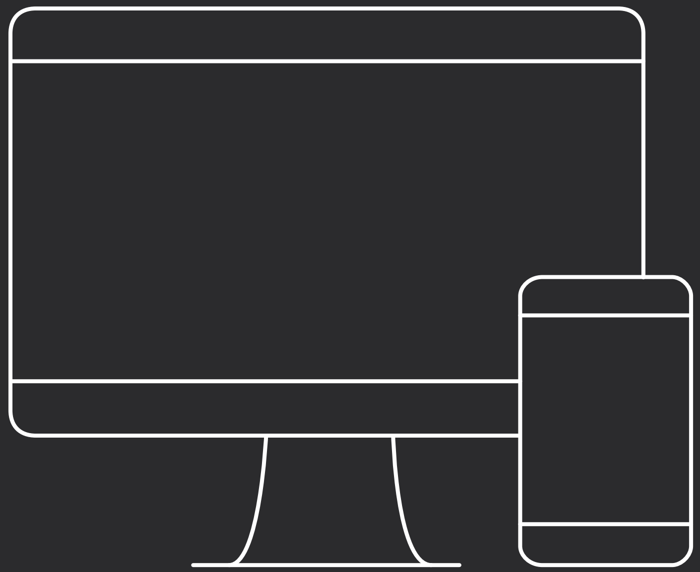
The most interesting pair is cosmetics + diapers

MBA for the segment

Pairs for consideration

	Пара	support(%)	confidence(%)	lift	expected confidence(%)
38	КАНЦТОВАРЫ, КНИГИ, ДИСКИ + ИГРУШКИ	4.134022	93.350063	0.987028	94.576874
62	КРУПНОГАБАРИТНЫЙ ТОВАР + ИГРУШКИ	3.772851	69.641026	0.736343	94.576874
110	ТЕКСТИЛЬ, ТРИКОТАЖ + ИГРУШКИ	3.206090	71.499380	0.755992	94.576874
86	ПОДГУЗНИКИ + ИГРУШКИ	2.678224	72.155689	0.762932	94.576874
146	ТОВАРЫ ДЛЯ КОРМЛЕНИЯ + ИГРУШКИ	2.411513	81.273408	0.859337	94.576874
50	КОСМЕТИКА/ГИГИЕНА + ИГРУШКИ	2.111463	86.956522	0.919427	94.576874
74	ОБУВЬ + ИГРУШКИ	1.439129	71.153846	0.752339	94.576874

The most interesting pair is stationery + toys due to high indexes



Personalized offers

Next Best Offer

Segments	Goods and Number of clients	Average price/Number of goods in cheque	Mechanic from MBA	Maximum the customers will pay (from segmentation)
Segment 1 	Cosmetics and Hygiene **3 622**	1100 rubles / 2	Diapers + Cosmetics and Hygiene; Marge of Cosmetics = 20% We suggest providing a 15% discount on cosmetics and hygiene for 1 500 ₽ spent on diapers	2 085 ₽
Segment 2 	Goods for feeding **4 197**	1800 rubles / 1.5	Diapers + Goods for feeding; Marge of Diapers = 10% We suggest providing a 5% discount on diapers for every 2 000 ₽ spent on goods for feeding	8 186 ₽
Segment 3 	Toys **13 764**	930 rubles / 1.8	Books, Disks, Stationery + Toys; marge of Toys =15% We suggest providing a 10% discount if in the same cheque you have toys + good from category books, disks, stationery.	2 532 ₽

Next Best Offer

Segments	Goal	SMS	Net profit (response percent 10%)
Segment 1 	Increase in average cheque and making wider variety of goods in cheque	If you buy diapers for more than 1 500 ₽ you get a **15% discount** on goods from category **Cometic and Hygiene**	228.5
Segment 2 	Increase in average cheque and supporting the highly-margin good	You get a **5% discount** on **goods for feeding** for every 2 000 ₽ spent on **diapers**	768.6
Segment 3 	Increase in average cheque of strong connected groups	If you have a good from category **Books, Disks Stationery**, you get a **10% discount** on one good from category **Toys**	1 148.9

Financial – economical foundation of the campaign

Cost of send SMS	Rub/one	2		
Attribute	Units of measure	Segment 1	Segment 2	Segment 3
Circulation	Units	3 622	4 197	13 764
Response	%	10%	10%	10%
Discount	Rubles	150 ₽	150 ₽	100 ₽
Minimal cheque	Rubles	800 ₽ on diapers	2000 ₽ on diapers	950 ₽
Sales	Units	362	420	1 376
Revenue	Thousands of ₽	290	840	1 307
Direct Costs	Thousands of ₽	**54.3**	**63**	**130.7**
Marketing Expenses	Thousands of ₽	**7.2**	**8.4**	**27.4**
Net Profit	Thousands of ₽	228.5	768.6	1 148.9