Homework #1

Made by DSBA212 Junior Year student,
Keseli Timur

Part I

Data Inspection and Segmentation preparation.

Feature Description

After short analysis of the description of each feature, we divided them into Categorical and Numerical.

Attributes	Описание	Description
Номер варианта	Номер варианта	Variant number
ID	Идентификатор клиента	Client ID
INCOME_BASE_	Подтверждение дохода	Income verification
CREDIT_PURPC	Цель получения кредита	Purpose of the loan
INSURANCE_FL	Страхование заемщика	Borrower's insurance when receiving a loan
DTI	debt-to-income ratio — от	debt-to-income ratio - the ratio of debt to income
SEX	Пол	Sex
FULL_AGE_CHII	Кол-во лет ребенку	Number of years of the child
DEPENDANT_NI	Кол-во иждивенцев	Number of dependents
EDUCATION	Образование	Education
EMPL_TYPE	Должность	Position
EMPL_SIZE	Зарплата	Salary
BANKACCOUNT	Кол-во аккаунтов у клиен	The number of accounts the client has. (0 - no online account, 1 - there is one online account, 2 or more - accessed the online account from another device)
Period_at_work	Время работы (кол-во дн	Working time (number of days)
age	Возраст	Age
EMPL_PROPER	Сфера бизнеса работода	Employer business area
EMPL_FORM	Организационно - правов	Organizational and legal form
FAMILY_STATUS	Семейный статус	Family status
max90days	кол-во запросов в бюро	number of requests to credit bureaus in the last 90 days
max60days	кол-во запросов в бюро	number of requests to credit bureaus in the last 60 days
max30days	кол-во запросов в бюро	number of requests to credit bureaus in the last 30 days
max21days	кол-во запросов в бюро	number of requests to credit bureaus in the last 21 days
max14days	кол-во запросов в бюро	number of requests to credit bureaus in the last 14 days
avg_num_delay	Среднее кол-во задержк	Average number of payment delays
if_zalog	Наличие залога (квартир	Presence of collateral (apartment, car)
num_AccountAc	кол-во активных счетов	number of active accounts accounts for the last 180 days
num_AccountAc	кол-во активных счетов	number of active accounts accounts in the last 90 days
num_AccountAc	кол-во активных счетов	number of active accounts accounts in the last 60 days
Active_to_All_prc	отношение активных сче	ratio of active accounts to all accounts
numAccountActi	кол-во открытых счетов	number of open accounts
numAccountClos	кол-во закрытых счетов	number of closed accounts
sum_of_paym_m	сумма платежей за посл	amount of payments for the last month (thousand)
all_credits	Кол-во кредитов	Number of credits
Active_not_cc	Активные кредитные сче	Active credit accounts but no credit card
own_closed	Кол-во закрытых кредитс	Number of closed loans
min_MnthAfterLc	минимальное кол-во мес	the minimum number of months that have passed since the last loan was taken, that is, how long ago the last loan was issued to the client
max_MnthAfterLc	кол-во месяцев прошедн	number of months past since the date of the first loan
dlq_exist	наличие просрочки на да	currently in arrears
thirty_in_a_year	просрочка больше 30 дн	overdue more than 30 days in the last year
sixty_in_a_year	просрочка больше 60 дн	overdue more than 60 days in the last year
ninety_in_a_year	просрочка больше 90 дн	overdue more than 90 days in the last year
thirty_vintage	просрочка больше 30 дн	overdue more than 30 days, ever
six ninety_vintage	просрочка больше 90 дн	overdue more than 90 days, ever

Figure 1. Feature division and description.

Feature Description

Each sample from the dataset have several categorical and numerical features:

Categorical

- INCOME_BASE_TYPE
- CREDIT_PURPOSE
- INSURANCE_FLAG
- SEX
- EDUCATION
- EMPL_TYPE
- EMPL_SIZE
- BANKACCOUNT_FLAG
- EMPL_PROPERTY
- EMPL_FORM
- FAMILY_STATUS
- If_zalog
- dlq_exist
- thirty_in_a_year
- sixty_in_a_year
- ninety_in_a_year
- thirty_vintage
- sixty_vintage
- ninety_vintage

Numerical

- Номер варианта
- ID
- DTI
- FULL_AGE_CHILD_NUMBER
- DEPENDANT_NUMBER
- Period_at_work
- age
- max90days
- max60days
- max30days
- max21days
- max14days
- avg_num_delay
- num_AccountActive180
- num_AccountActive90
- num_AccountActive60
- Active_to_All_prc
- numAccountActiveAll
- numAccountClosed
- sum_of_paym_months
- all_credits
- Active_not_cc
- own_closed
- min_MnthAfterLoan
- max_MnthAfterLoan

Total: 19 categorical, 25 Numerical

Data Analysis

We calculated the mean, standard deviation, median, minimum and maximum for each numerical feature with the given dataset and got the following results (Figure 1).

However, we see that for some attributes that have more than 50% of NaN values, and they account for around 6500 samples. Moreover, the number of such attributes is around 21, so we decided to delete all such rows. After closer inspection, we draw a conclusion that the rows that were missing a values in one of those attributes, missed all the values in other such attributes, so we are left with a dataset of around 3660 samples.

Lastly, Номер Варианта and ID are irrelevant, so they were deleted.

Attributes	# of Unique value	% of Unique valu	# of Zero values	% of Zero values	# of NaN values	% of Nan values	Mean	Median	Deviation	Minimum	Maximum
Номер варианта											
ID	10243	100.00%	0	0.00%	0	0.00%	-	-	-	-	-
INCOME_BASE_TY	4	0.04%	0	0.00%	78	0.76%	-	-	-	-	-
CREDIT_PURPOSE	10	0.10%	0	0.00%	0	0.00%	-	-	-	-	-
INSURANCE_FLAG	2	0.02%	4041	39.45%	1	0.01%	-	-	-	-	-
DTI	62	0.61%	0	0.00%	125	1.22%	0.39	0.4	0.14	0.01	0.64
SEX	2	0.02%	0	0.00%	0	0.00%	-	-	-	-	-
FULL_AGE_CHILD_	7	0.07%	6072	59.28%	0	0.00%	0.56	0	0.78	0	6
DEPENDANT_NUMI	4	0.04%	10210	99.68%	0	0.00%	0.004	0	0.08	0	3
EDUCATION	9	0.09%	0	0.00%	26	0.25%	-	-	-	-	-
EMPL_TYPE	9	0.09%	0	0.00%	12	0.12%	-	-	-	-	-
EMPL_SIZE	8	0.08%	0	0.00%	1	0.01%	-	-	-	-	-
BANKACCOUNT_FL	4	0.04%	6207	60.60%	2326	22.71%	-	-	-	-	-
Period_at_work	358	3.50%	0	0.00%	2328	22.73%	66.11	45	65.13	4	455
age	40	0.39%	0	0.00%	2327	22.72%	36.37	35	8.69	23	63
EMPL_PROPERTY	12	0.12%	0	0.00%	2327	22.72%	-	-	-	-	-
EMPL_FORM	6	0.06%	0	0.00%	6245	60.97%	-	-	-	-	-
FAMILY_STATUS	6	0.06%	0	0.00%	6245	60.97%	-	-	-	-	-
max90days	19	0.19%	1078	10.52%	6299	61.50%	1.58	1	1.83	0	25
max60days	17	0.17%	1575	15.38%	6299	61.50%	1.11	1	1.46	0	19
max30days	13	0.13%	2036	19.88%	6299	61.50%	0.92	0	1.22	0	12
max21days	12	0.12%	2376	23.20%	6299	61.50%	0.62	0	1.04	0	11
max14days	11	0.11%	2577	25.16%	6299	61.50%	0.51	0	0.92	0	10
avg_num_delay	1134	11.07%	1571	15.34%	6577	64.21%	0.06	0.01	0.12	0	0.94
if_zalog	2	0.02%	2482	24.23%	6567	64.11%	-	-	-	-	-
num_AccountActive1	7	0.07%	2609	25.47%	6567	64.11%	0.37	0	0.66	0	6
num_AccountActive9	4	0.04%	3169	30.94%	6567	64.11%	0.16	0	0.43	0	3
num_AccountActive6	4	0.04%	3359	32.79%	6567	64.11%	0.1	0	0.34	0	3
Active_to_All_prc	93	0.91%	531	5.18%	6567	64.11%	0.41	0.38	0.29	0	1
numAccountActiveAl	14	0.14%	513	5.01%	6567	64.11%	2.13	2	1.66	0	13
numAccountClosed	25	0.24%	393	3.84%	6567	64.11%	3.54	3	3.23	0	25
sum_of_paym_montl	319	3.11%	10	0.10%	6567	64.11%	79.97	61	69.54	0	548
all_credits	30	0.29%	0	0.00%	6567	64.11%	5.67	5	4.04	1	30
Active_not_cc	7	0.07%	1284	12.54%	6567	64.11%	1.06	1	1.06	0	6
own_closed	11	0.11%	2123	20.73%	6567	64.11%	0.72	0	1.11	0	11
min_MnthAfterLoan	101	0.99%	140	1.37%	6567	64.11%	14.35	10	15.5	-1	115
max_MnthAfterLoan	134	1.31%	5	0.05%	6567	64.11%	61.53	68	30.6	-1	179
dlq_exist	2	0.02%	1581	15.43%	6567	64.11%	-	-	-	-	-
thirty_in_a_year	2	0.02%	3086	30.13%	6567	64.11%	-	-	-	-	-
sixty_in_a_year	2	0.02%	3356	32.76%	6567	64.11%	-	-	-	-	-
ninety_in_a_year	2	0.02%	3419	33.38%	6567	64.11%	-	-	-	-	-
thirty_vintage	2	0.02%	3566	34.81%	6567	64.11%	-	-	-	-	-
sixty_vintage	2	0.02%	3627	35.41%	6567	64.11%	-	-	-	-	-
ninety_vintage	2	0.02%	3611	35.25%	6567	64.11%	-	-	-	-	-

Figure 2.1. Attribute analysis.

Small point: around 11 samples had no Education value, so we deleted them as well.

Data Analysis

New data distribution

	count	mean	std	min	25%	50%	75%	max
INSURANCE_FLAG	3649.0	0.620170	0.485411	0.00	0.00	1.000000	1.000000	1.000000
DTI	3649.0	0.392077	0.135591	0.01	0.29	0.410000	0.490000	0.620000
FULL_AGE_CHILD_NUMBER	3649.0	0.527268	0.757553	0.00	0.00	0.000000	1.000000	4.000000
DEPENDANT_NUMBER	3649.0	0.003837	0.066115	0.00	0.00	0.000000	0.000000	2.000000
BANKACCOUNT_FLAG	3649.0	0.309126	0.774686	0.00	0.00	0.000000	0.000000	4.000000
Period_at_work	3649.0	56.002192	53.303711	6.00	18.00	40.000000	77.000000	422.000000
age	3649.0	36.044122	8.625523	23.00	29.00	34.000000	42.000000	63.000000
max90days	3649.0	1.571389	1.852397	0.00	0.00	1.000000	2.000000	25.000000
max60days	3649.0	1.078652	1.460221	0.00	0.00	1.000000	2.000000	19.000000
max30days	3649.0	0.770622	1.211808	0.00	0.00	0.000000	1.000000	12.000000
max21days	3649.0	0.568101	1.027224	0.00	0.00	0.000000	1.000000	11.000000
max14days	3649.0	0.454097	0.897511	0.00	0.00	0.000000	1.000000	10.000000
avg_num_delay	3649.0	0.064629	0.117291	0.00	0.00	0.014706	0.075472	0.942308
if_zalog	3649.0	0.325569	0.468651	0.00	0.00	0.000000	1.000000	1.000000
num_AccountActive180	3649.0	0.373527	0.665545	0.00	0.00	0.000000	1.000000	6.000000
num_AccountActive90	3649.0	0.160592	0.431711	0.00	0.00	0.000000	0.000000	3.000000
num_AccountActive60	3649.0	0.098931	0.343041	0.00	0.00	0.000000	0.000000	3.000000
Active_to_All_prc	3649.0	0.412611	0.290466	0.00	0.20	0.375000	0.571429	1.000000
numAccountActiveAll	3649.0	2.129899	1.658513	0.00	1.00	2.000000	3.000000	13.000000
numAccountClosed	3649.0	3.544533	3.215531	0.00	1.00	3.000000	5.000000	25.000000
sum_of_paym_months	3649.0	80.027131	69.360642	1.00	30.00	61.000000	110.000000	548.000000
all_credits	3649.0	5.674431	4.026340	1.00	3.00	5.000000	8.000000	30.000000
Active_not_cc	3649.0	1.059742	1.057681	0.00	0.00	1.000000	2.000000	6.000000
own_closed	3649.0	0.724856	1.106420	0.00	0.00	0.000000	1.000000	11.000000
min_MnthAfterLoan	3649.0	14.215676	15.264026	-1.00	4.00	10.000000	19.000000	115.000000
max_MnthAfterLoan	3649.0	61.543163	30.570155	0.00	34.00	68.000000	87.000000	179.000000
dlq_exist	3649.0	0.572212	0.494826	0.00	0.00	1.000000	1.000000	1.000000
thirty_in_a_year	3649.0	0.161414	0.367963	0.00	0.00	0.000000	0.000000	1.000000
sixty_in_a_year	3649.0	0.087695	0.282890	0.00	0.00	0.000000	0.000000	1.000000
ninety_in_a_year	3649.0	0.070430	0.255906	0.00	0.00	0.000000	0.000000	1.000000
thirty_vintage	3649.0	0.030145	0.171010	0.00	0.00	0.000000	0.000000	1.000000
sixty_vintage	3649.0	0.013428	0.115116	0.00	0.00	0.000000	0.000000	1.000000
ninety_vintage	3649.0	0.017813	0.132290	0.00	0.00	0.000000	0.000000	1.000000

Old data distribution

	count	mean	std	min	25%	50%	75%	max
INSURANCE_FLAG	10242.0	0.605448	0.488778	0.00	0.00	1.000000	1.000000	1.000000
DTI	10118.0	0.388778	0.137036	0.01	0.28	0.400000	0.490000	0.640000
FULL_AGE_CHILD_NUMBER	10243.0	0.563116	0.775525	0.00	0.00	0.000000	1.000000	6.000000
DEPENDANT_NUMBER	10243.0	0.004198	0.080772	0.00	0.00	0.000000	0.000000	3.000000
BANKACCOUNT_FLAG	7916.0	0.392370	0.876974	0.00	0.00	0.000000	0.000000	4.000000
Period_at_work	7915.0	66.108528	65.132875	4.00	20.00	45.000000	87.500000	455.000000
age	7916.0	36.366978	8.690094	23.00	29.00	35.000000	42.000000	63.000000
max90days	3944.0	1.577333	1.831918	0.00	0.00	1.000000	2.000000	25.000000
max60days	3944.0	1.114097	1.462929	0.00	0.00	1.000000	2.000000	19.000000
max30days	3944.0	0.816684	1.215804	0.00	0.00	0.000000	1.000000	12.000000
max21days	3944.0	0.623732	1.041325	0.00	0.00	0.000000	1.000000	11.000000
max14days	3944.0	0.511663	0.915594	0.00	0.00	0.000000	1.000000	10.000000
avg_num_delay	3666.0	0.064530	0.117192	0.00	0.00	0.014493	0.075421	0.942308
if_zalog	3676.0	0.324810	0.468367	0.00	0.00	0.000000	1.000000	1.000000
num_AccountActive180	3676.0	0.372416	0.664178	0.00	0.00	0.000000	1.000000	6.000000
num_AccountActive90	3676.0	0.160773	0.431414	0.00	0.00	0.000000	0.000000	3.000000
num_AccountActive60	3676.0	0.099021	0.342840	0.00	0.00	0.000000	0.000000	3.000000
Active_to_All_prc	3676.0	0.412765	0.291502	0.00	0.20	0.375000	0.571429	1.000000
numAccountActiveAll	3676.0	2.126224	1.658902	0.00	1.00	2.000000	3.000000	13.000000
numAccountClosed	3676.0	3.541621	3.226280	0.00	1.00	3.000000	5.000000	25.000000
sum_of_paym_months	3676.0	79.966268	69.539844	0.00	30.00	61.000000	110.250000	548.000000
all_credits	3676.0	5.667845	4.039042	1.00	3.00	5.000000	8.000000	30.000000
Active_not_cc	3676.0	1.056583	1.056529	0.00	0.00	1.000000	2.000000	6.000000
own_closed	3676.0	0.723885	1.105428	0.00	0.00	0.000000	1.000000	11.000000
min_MnthAfterLoan	3676.0	14.347116	15.500530	-1.00	4.00	10.000000	19.000000	115.000000
max_MnthAfterLoan	3676.0	61.527748	30.597079	-1.00	34.00	68.000000	87.000000	179.000000
dlq_exist	3676.0	0.569913	0.495155	0.00	0.00	1.000000	1.000000	1.000000
thirty_in_a_year	3676.0	0.160501	0.367120	0.00	0.00	0.000000	0.000000	1.000000
sixty_in_a_year	3676.0	0.087051	0.281948	0.00	0.00	0.000000	0.000000	1.000000
ninety_in_a_year	3676.0	0.069913	0.255035	0.00	0.00	0.000000	0.000000	1.000000
thirty_vintage	3676.0	0.029924	0.170400	0.00	0.00	0.000000	0.000000	1.000000
sixty_vintage	3676.0	0.013330	0.114698	0.00	0.00	0.000000	0.000000	1.000000
ninety_vintage	3676.0	0.017682	0.131812	0.00	0.00	0.000000	0.000000	1.000000

Feature Distribution

The following distributions show the difference between Males and Females. Overall, the information for both sexes is the same, but there are some differences in distribution, such as Family Status, employment type, DTI, and number of request to the credit bureaus in the last 60 and 90 days.

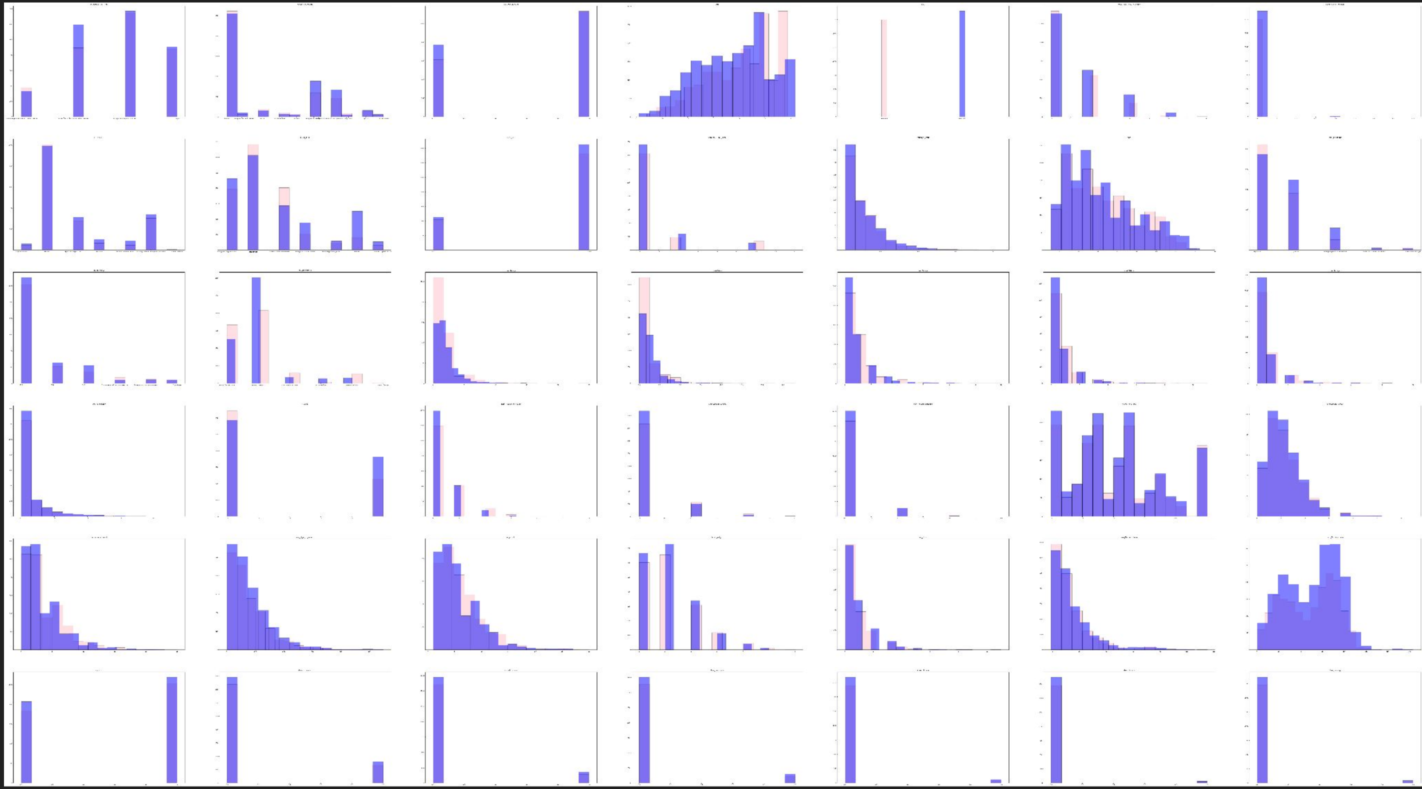

Figure 2.2. Feature distribution.

Feature Analysis

Using the following heatmap, we can see which features are redundant, and which are useful. Although many features have mild and strong correlations, we face a problem, where the feature has little correlation with any other feature, but it has a viable impact on the person's profile, such as 'DEPENDANT_NUMBER'. It shows the number of dependants in the family of a person, who requests a credit. So, we decided to leave all the features for clustering methods.

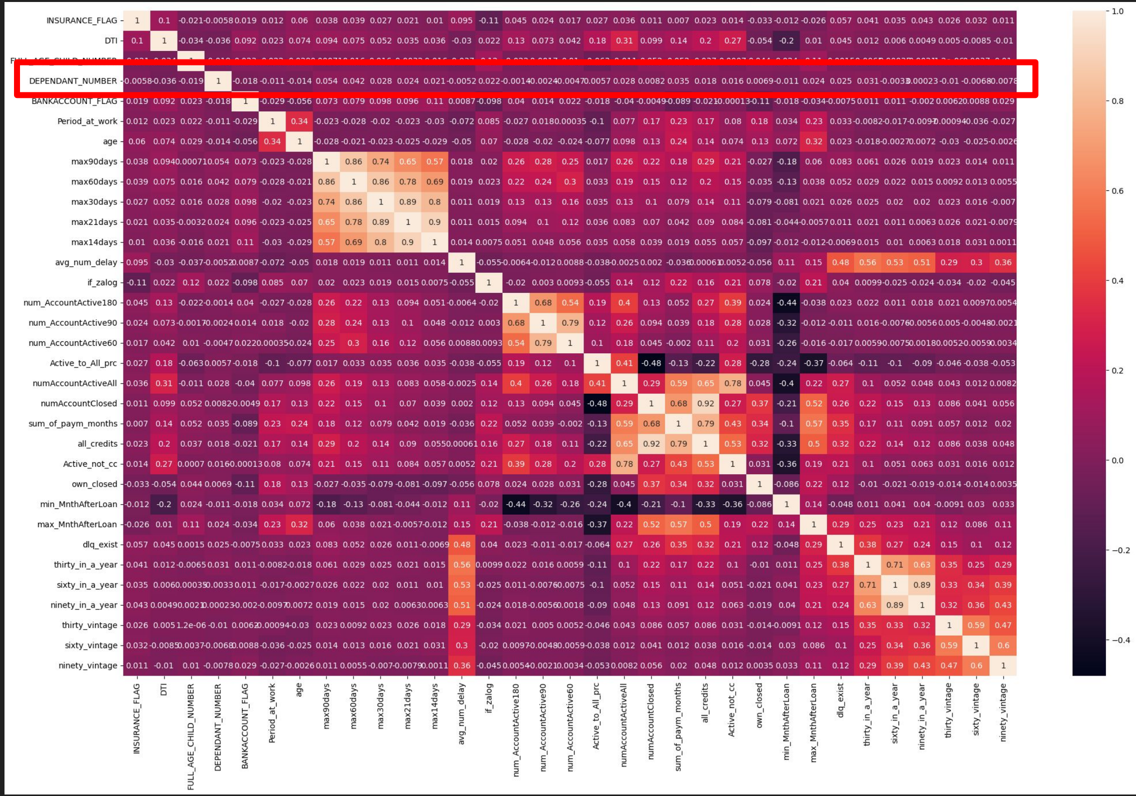

Figure 3.1 Correlation matrix.

Last Data Modifications

As we have mentioned before, there are quite a lot of categorical features, and we need to create dummy variables out of them, so we could use them in the segmentation part. Firstly, we change the ‘SEX’ feature to be 0(women) or 1(men), and reduced ‘BANKACCOUNT_FLAG’ from [0,1,2,3,4] values to [0,1,2], since it doesn’t affect the results. As for other categorical features, we used the python function, to get the dummy variables that take 0 or 1, and removed one dummy variable for each categorical features, so there would be no multicollinearity between them.

So, the resulting table has 73 features and, 3649 samples. At last, we will normalise the data.

	0	1	2	3	4	5	6	7	8	9	...	3639	3640	3641	3642	3643	3644	3645	3646	3647	3648
INSURANCE_FLAG	1.00	1.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	1.00	...	0.00	1.0	1.00	1.00	0.0	0.00	1.00	0.00	1.00	1.00
DTI	0.43	0.46	0.35	0.23	0.23	0.19	0.24	0.59	0.36	0.54	...	0.33	0.4	0.39	0.41	0.3	0.32	0.32	0.57	0.59	0.55
SEX	1.00	0.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00	...	1.00	0.0	1.00	0.00	1.0	1.00	0.00	1.00	0.00	0.00
FULL_AGE_CHILD_NUMBER	0.00	2.00	0.00	0.00	0.00	1.00	0.00	2.00	3.00	0.00	...	0.00	1.0	1.00	0.00	0.0	1.00	0.00	0.00	0.00	0.00
DEPENDANT_NUMBER	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	0.00	0.0	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
FAMILY_STATUS_гражданский брак	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	1.00	0.0	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00
FAMILY_STATUS_женат / замужем	0.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	...	0.00	0.0	1.00	0.00	1.0	0.00	0.00	1.00	1.00	1.00
FAMILY_STATUS_повторный брак	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...	0.00	0.0	0.00	0.00	0.0	1.00	0.00	0.00	0.00	0.00
FAMILY_STATUS_разведен / разведена	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	...	0.00	0.0	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00
FAMILY_STATUS_холост / не замужем	1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	...	0.00	1.0	0.00	1.00	0.0	0.00	1.00	0.00	0.00	0.00

73 rows × 3649 columns

Figure 3.2. Dummy variables in the data set.

Part II

Segmentation

Segmentation techniques

We decided to apply the following segmentation methods:

1) K-means
2) RFM

K-means is a good method for this case, since it perfectly suits for datasets with large number of features, and gives an interpretable explanation for the segmentation, however it has some drawbacks, such as changing the number of clusters can radically affect the results of all clusters, so the results might differ depending on the number of clusters.

RFM is another popular method used to cluster several groups in marketing and selling. Its benefits are that it is easy to compute and interpret, also it clearly shows the groups that buy the most, and buy the least.

K-means

We are using the elbow method to decide on the number of clusters. The elbow method uses inertia to describe the spread of points, so we plotted the number of clusters against the inertia. The graph shows the greatest fold around 3 and 5. So, we decided to stick with 5 clusters.

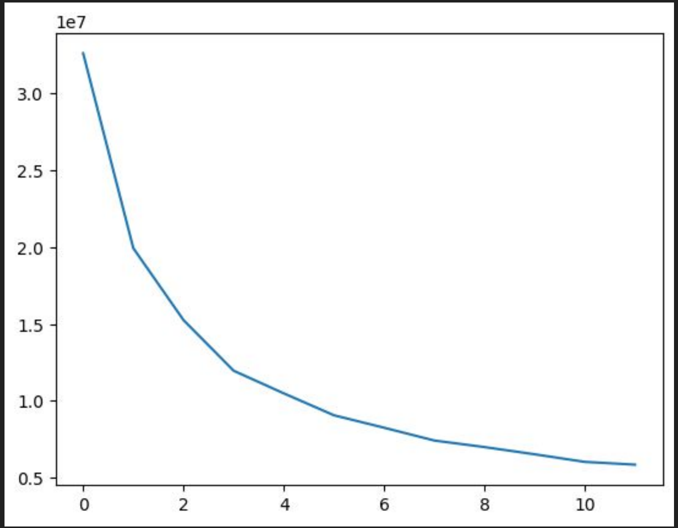

Figure 4. Inertia on y-axis, # of neighbours on x-axis.

K-means

The cluster division has given us this beautiful plot. Here we can see the spread of all 5 clusters on a two-dimensional plain. Now we are going to find the expected client from each of the 5 clusters.

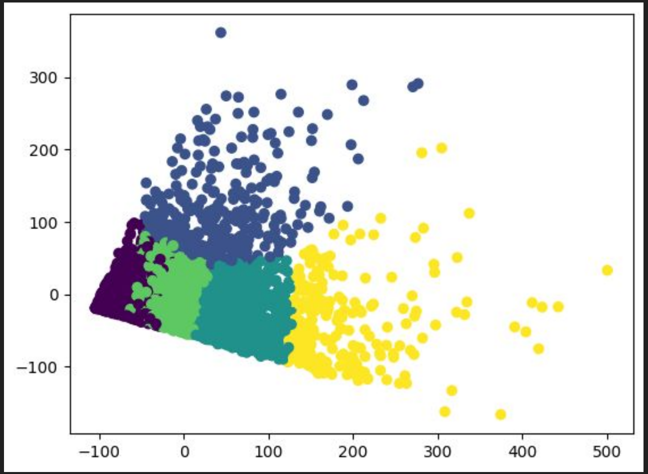

Figure 5. K-means with 5 clusters.

K-means Profiles

Cluster 1:

Gender: Woman

Age: 30

Family status: married

Education: Bachelor’s degree

Works in selling as a Specialist.

Credit history: has zero closed loans with 2 credits. Has taken a loan at least in the least 5 months and at most in the last 2 years. Has paid 25,000 roubles in the last month. Half of all accounts are active.

Cluster 2:
Gender: Man
Age: 43
Family status: married, has at least 1 child
Education: Bachelor’s degree and higher
Job: Specialist and has astounding period at work of 173 days.
Credit history: has 6 credits, and has 1 closed loan with maximum pay for the loan at 72,000 roubles.

K-means Profiles

Cluster 3:

Gender: Man

Age: 36

Family status: married

Education: Bachelor’s degree

Job: Works in a ‘OOO’ as a selling Specialist

Credit history: has 3 open and 5 closed accounts. Has paid around 178,000 roubles last month, takes loan for renovation.

Cluster 4:

Gender: Man

Age: 34

Family status: engaged, has at least 1 child

Education: Bachelor’s degree and higher

Job: Specialist in selling and has astounding period at work of 173 days.

Credit history: has 5 credits, and has 1 closed loan with average pay for the loan at 60,000 roubles. Has the highest probability to take the loan for a car.

K-means Profiles

Cluster 5:

Gender: Man

Age: 42

Family status: married, 1 child

Education: Bachelor’s degree

Job: Middle manager in the ‘OOO’, doing selling

Credit history: has 5 active accounts, has a delay in paying loans, has 2 closed loans with 14 credits. Has twice asked for a loan in the last 90 days.

RFM

For RFM method we use 3 variables: Recency, Frequency, Monetary Value.

For Recency, we will use min_MnthAfterLoan,

For Frequency: all_credits,

For Monetary Value: sum_of_paym_months.

These features describe how long ago the last loan was issued to the client, number of credits and amount of payments for the last month in thousands of roubles. So, they perfectly fit the definition of each variable in RFM.

We will divide each variable into 5 segments, such as we did in k-means, for the accuracy purposes. So, 1 means that a client has taken a loan very long time age, has taken the small number of loans and paid little money last month to cover the loans; and 5 means that the client has very recently taken the loan, has many loans and paid a substantial amount of money last month. To conclude, there are 125 segments.

RFM: Segment Division

We calculated the Total RFM Score, which is the sum of R, F and M components and got the distribution, that is close to Normal Gaussian Distribution. Now we will divide it into 5 clusters and draw a profile of a typical client from each of them.

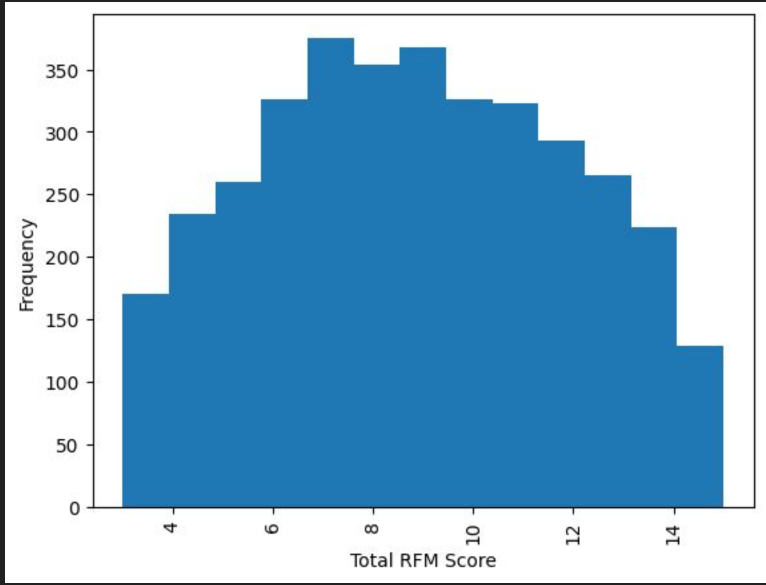

Figure 6. Total RFM Distribution.

RFM Profiles

Cluster 1:

Sex: Male

Age: 35

Family status: Married, has 1–2 children

Education: Bachelor’s Degree

Job: Earns more than 250,000 roubles, works as a selling specialist in ‘OOO’.

Credit history: very rarely takes loans, mostly for repair or renovation, but pay them back with zero days delay.

Mean RFM Score: 4

Cluster 2:

Sex: Male

Age: 33

Family status: Married, has 1 child

Education: Bachelor’s Degree

Job: Earns more than 250,000 roubles, works as a

Credit history: has 0 days in delay when paying loans, frequently takes loans, usually for a car, flat or renovation

Mean RFM Score: 7

RFM Profiles

Cluster 3:

Sex: Woman

Age: 34

Family status: Married, no children

Education: Bachelor’s Degree

Job: works as a selling specialist in ‘OOO’ or ‘OAO’.

Credit history: has around 4–5 loans from 2 different bank accounts, takes loans mostly for repair or renovation and at most has 1 day delay.

Mean RFM Score: 9

Cluster 4:

Sex: Male

Age: 20-32

Family status: Bachelor or divorced and has 1 child

Education: Bachelor’s Degree or school certificate

Job: Earns more than 250,000 roubles, works as a selling specialist

Credit history: has taken a loan in the last 8 months, has around 7 loans, usually for education, renovation or to buy furniture, last month has paid around 90,000 roubles to the banks

Mean RFM Score: 11

RFM Profiles

Cluster 5:

Sex: Male

Age: 37-40

Family status: Married, has 1 child

Education: Bachelor’s Degree

Job: Earns more than 250,000 roubles, works as a

Credit history: paid more than 150,000 roubles last month to the bank, has 10 loans, last loan took in the last 4 months, has 0 days in delay when paying loans, very frequently takes loans, usually for a flat or renovation

Mean RFM Score: 14

Conclusions

We have successfully divided the sample into five profiles, using two methods: k-means and RFM. Both of them have given us a much detailed look on the types of people that take loans. As it was observed, the profiles in both RFM and K-means have several things in common: they pay back the loans usually on time, have children and usually take loans for renovation or car repair. However, there can be drawn borders between them: they have the distinct number of loans taken and the recency of the loans. Some people very rarely take loans and others have 10 loans, where the last one was taken in the last 4 months. The other distinction is in the amount of money people pay on loans each month. People with high RFM pay more than 100,000, while low score RFM people pay approximately 50-70 thousand roubles. The last distinction is the age of profiles: most young people usually don't have enough money to take many loans and rarely take them, while at the same time middle-aged people have more than 3–5 loans, but they earn more than the youngsters.