Faculty of Computer Science

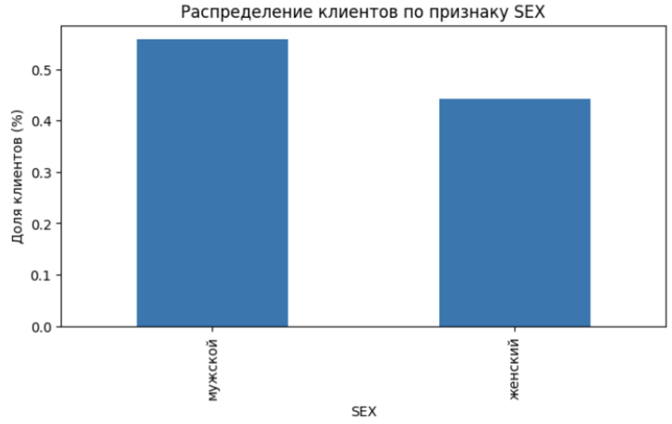
Data Science and Business Analytics

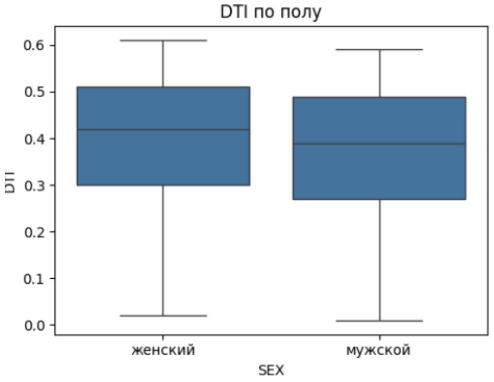
Moscow 2025

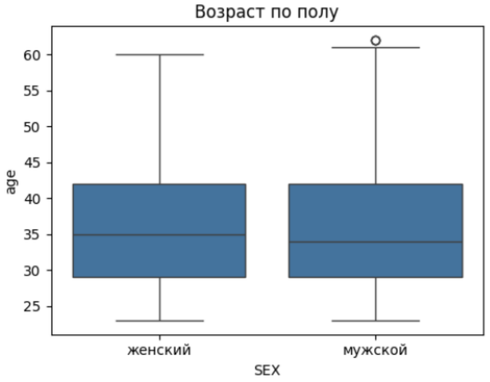
Homework #1 – Data Analysis in Business

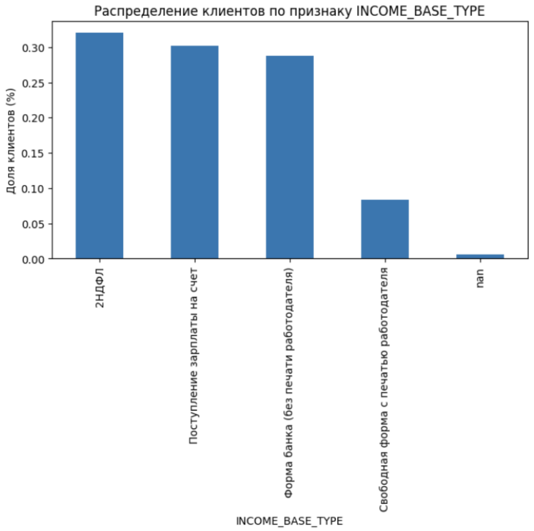
Student: Yana Antropova

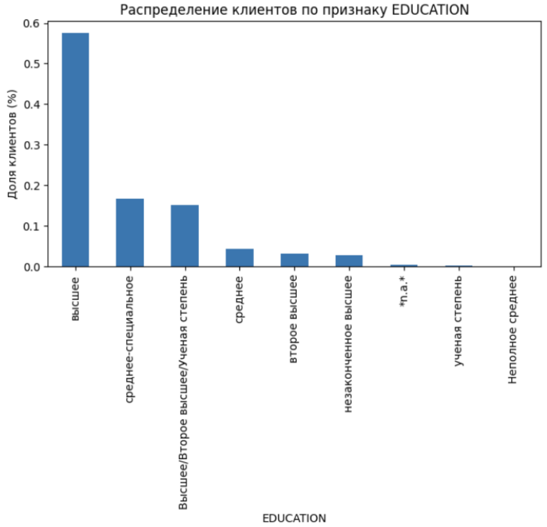
Date: November 2025

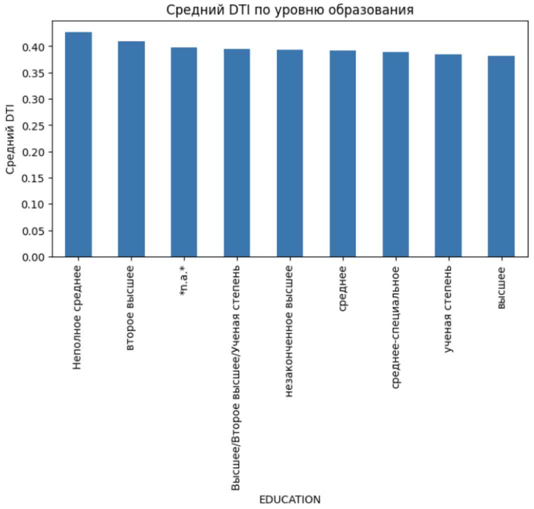


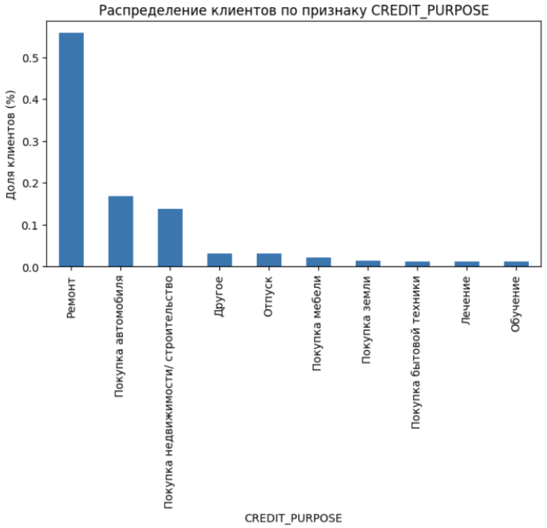












The customer base is predominantly male, educated and with official sources of income. The main purposes of loans are car repairs and purchases. A higher level of education correlates with a lower debt-to-income ratio.

Initial dataset contains **10,243 customers, 44 attributes**

	dtype	unique_values	num_zeros	num_nulls
Номер варианта	int64	1	0	0
ID	int64	10243	0	0
INCOME_BASE_TYPE	object	4	0	66
CREDIT_PURPOSE	object	10	0	0
INSURANCE_FLAG	int64	2	3964	0
DTI	float64	60	0	134
SEX	object	2	0	0
FULL_AGE_CHILD_NUMBER	float64	8	6154	1
DEPENDANT_NUMBER	int64	4	10211	0
EDUCATION	object	9	0	0
EMPL_TYPE	object	9	0	5
EMPL_SIZE	object	8	0	134
BANKACCOUNT_FLAG	float64	4	6226	2326
Period_at_work	float64	368	0	2327
age	float64	40	0	2326

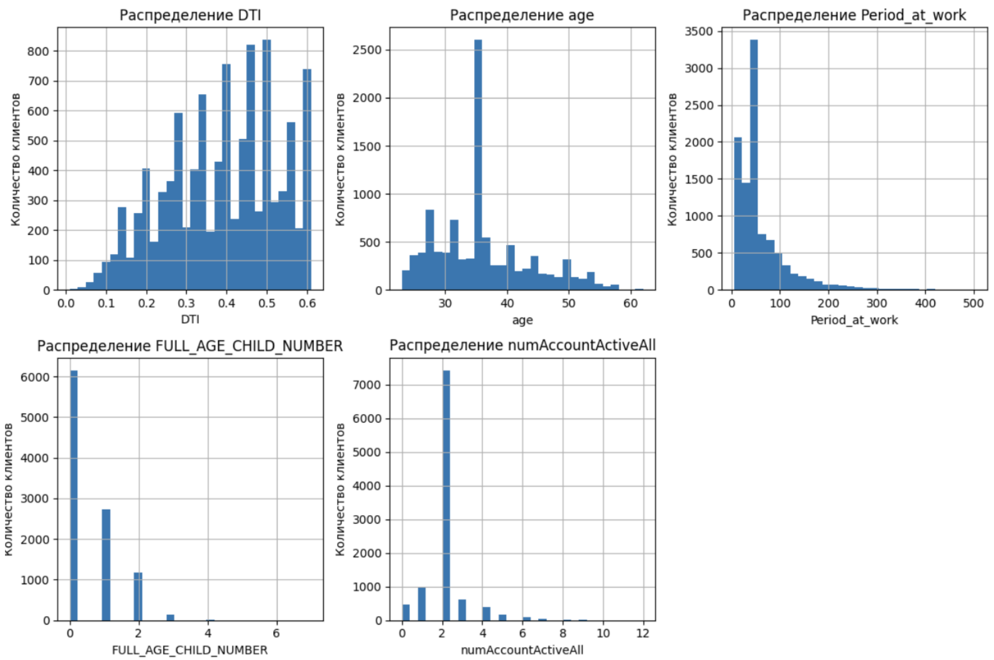
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10243 entries, 0 to 10242
Data columns (total 42 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   INCOME_BASE_TYPE       10243 non-null  int64
 1   CREDIT_PURPOSE         10243 non-null  int64
 2   INSURANCE_FLAG         10243 non-null  int64
 3   DTI                    10243 non-null  float64
 4   SEX                    10243 non-null  int64
 5   FULL_AGE_CHILD_NUMBER  10243 non-null  float64
 6   DEPENDANT_NUMBER       10243 non-null  int64
 7   EDUCATION              10243 non-null  int64
 8   EMPL_TYPE              10243 non-null  int64
 9   EMPL_SIZE              10243 non-null  int64
 10  BANKACCOUNT_FLAG       10243 non-null  float64
 11  Period_at_work         10243 non-null  float64
 12  age                    10243 non-null  float64
 13  EMPL_PROPERTY          10243 non-null  int64
 14  EMPL_FORM              10243 non-null  int64
 15  FAMILY_STATUS          10243 non-null  int64
 16  max90days              10243 non-null  float64
 17  max60days              10243 non-null  float64
 18  max30days              10243 non-null  float64
 19  max21days              10243 non-null  float64
 20  max14days              10243 non-null  float64
 21  avg_num_delay          10243 non-null  float64
 22  if_zalog               10243 non-null  float64
 23  num_AccountActive180   10243 non-null  float64
 24  num_AccountActive90    10243 non-null  float64
 25  num_AccountActive60    10243 non-null  float64
 26  Active_to_All_prc      10243 non-null  float64
 27  numAccountActiveAll    10243 non-null  float64
 28  numAccountClosed       10243 non-null  float64
 29  sum_of_paym_months     10243 non-null  float64
 30  all_credits            10243 non-null  float64
 31  Active_not_cc          10243 non-null  float64
 32  own_closed             10243 non-null  float64
 33  min_MnthAfterLoan      10243 non-null  float64
 34  max_MnthAfterLoan      10243 non-null  float64
 35  dlq_exist              10243 non-null  float64
 36  thirty_in_a_year       10243 non-null  float64
 37  sixty_in_a_year        10243 non-null  float64
 38  ninety_in_a_year       10243 non-null  float64
 39  thirty_vintage         10243 non-null  float64
 40  sixty_vintage          10243 non-null  float64
 41  ninety_vintage         10243 non-null  float64
dtypes: float64(31), int64(11)
memory usage: 3.3 MB
```

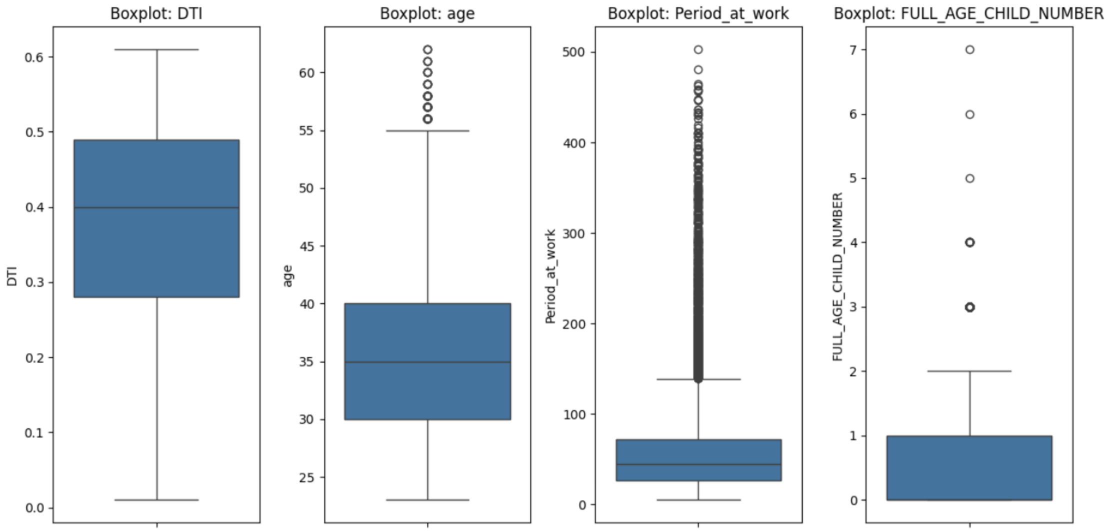
After first preprocessing:
10,243 customers, 42 attributes

There are gaps in the employment block (age, Period_at_work, BANKACCOUNT_FLAG). There are constant attributes (Номер варианта).

After cleaning the data, gaps and duplicates were removed, categorical values were unified, and uninformative features were deleted.

Missing values of numerical variables were replaced with the median, and categorical variables were replaced with the mode.





	column	count	percent of outliers
0	DTI	0	0.000000
0	age	109	1.064141
0	Period_at_work	870	8.493605
0	FULL_AGE_CHILD_NUMBER	173	1.688958
0	numAccountActiveAll	2811	27.443132

Emissions analysis showed that the Period_at_work and numAccountActiveAll attributes contain a significant number of anomalous values (8–27%) due to the long right tail of the distribution. To stabilise the data, logarithmic transformation and truncation at the 99th percentile were applied, and rare categories of the FULL_AGE_CHILD_NUMBER attribute were merged.

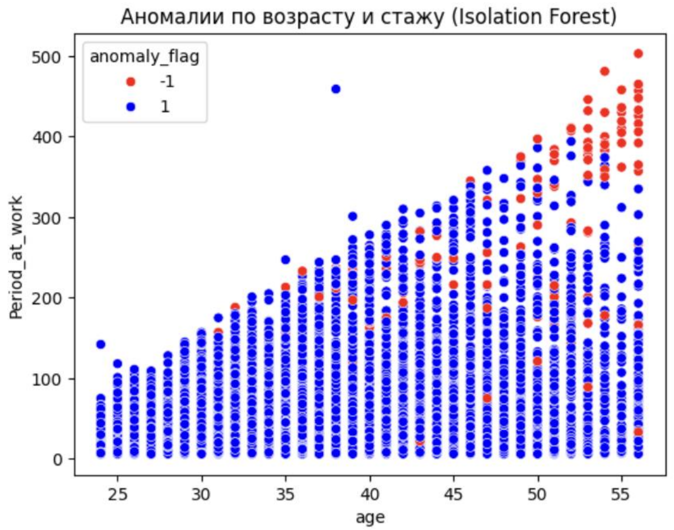
```python
from scipy import stats

z_scores = np.abs(stats.zscore(df_clean[cols_to_check], nan_policy='omit'))
outliers_z = (z_scores > 3).any(axis=1)
print(f"Аномальных строк по z-score: {outliers_z.sum()} ({outliers_z.mean() * 100:.2f}%)")
```

```
Аномальных строк по z-score: 573 (5.59%)
```

For example:

	age	Period_at_work	DTI
111	38.0	459.0	0.5

38 years = 456 months -> an almost impossible value



Isolation Forest found 205 abnormal clients (2.00%)

- Removed outliers and anomalies.
- Removed some temporary columns, like 'anomaly_flag’.
- Builded data mart.

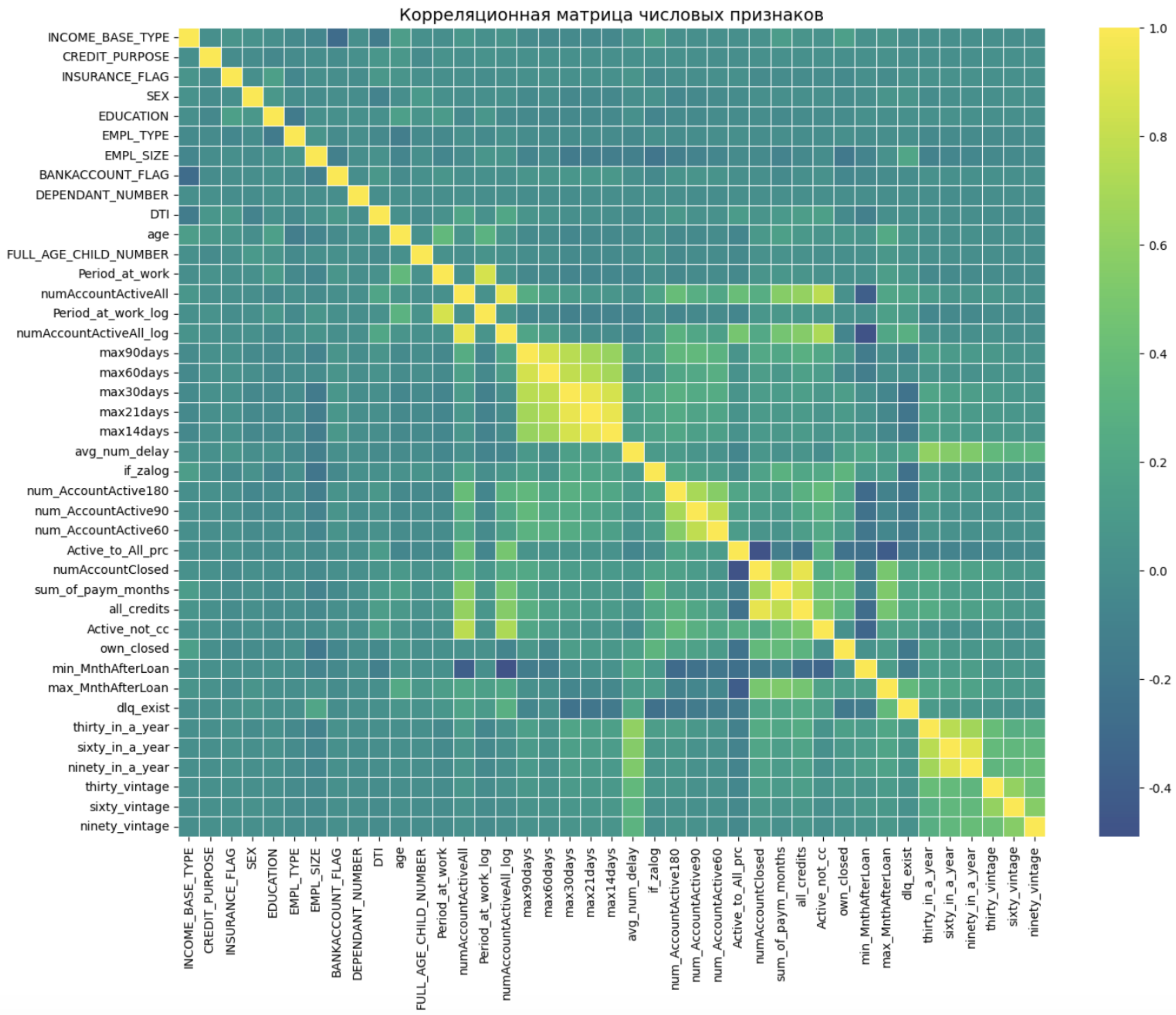
Now dataset contains
10037 rows × 41 columns

Checking non-bank customers, who do not have active accounts/loans/payment history and have no arrears:

```
rows_before = len(df_mart)

mask_nonbank_B = (
    (df_mart['BANKACCOUNT_FLAG'] == 0) &
    (df_mart['numAccountActiveAll'] == 0) &
    (df_mart['all_credits'] == 0) &
    (df_mart['sum_of_paym_months'] == 0) &
    (df_mart['dlq_exist'] == 0)
)

df_mart_B = df_mart.loc[~mask_nonbank_B].copy()
print(f"Non-bank: {mask_nonbank_B.sum()} out of {rows_before} ({mask_nonbank_B.sum()/rows_before:.2%})")
```

```
Non-bank: 0 out of 10037 (0.00%)
```

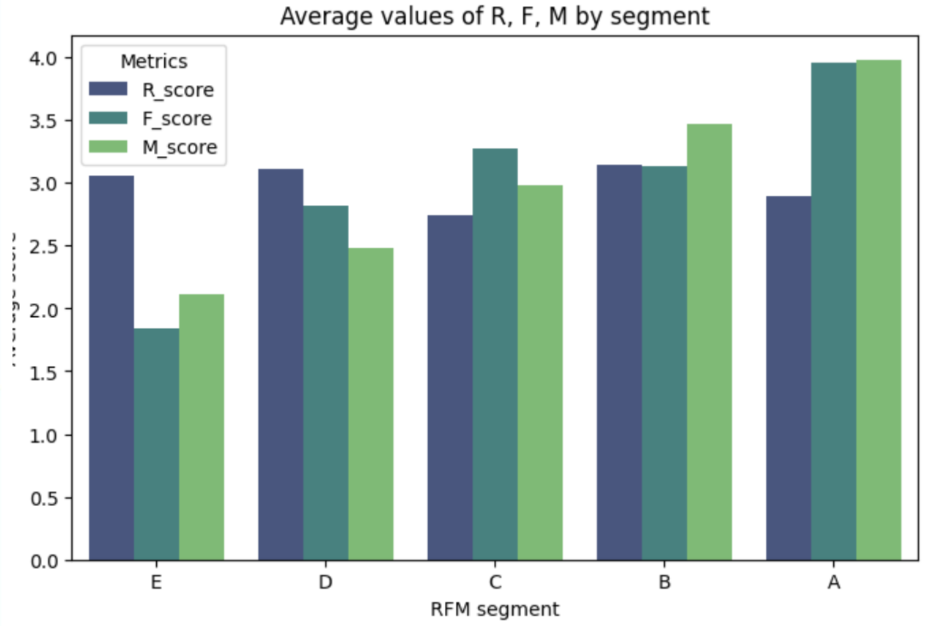
There are **no non-bank customers.**



Based on correlation analysis, highly dependent features (|r| > 0.8) were removed, including duplicate delay metrics and logarithmic values. The final set for clustering includes nine key indicators that ensure independence and diversity of factors.

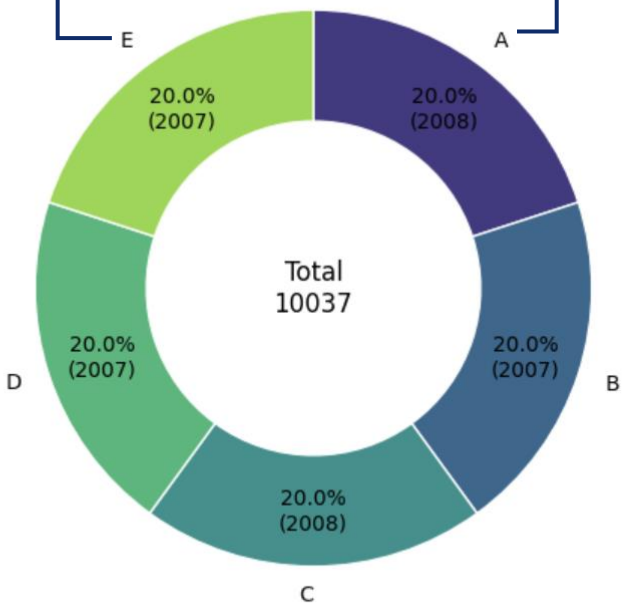
Now dataset contains
10037 rows × 9 columns

RFM segmentation



Average level of engagement: customers with moderate scores, no problems, but do not bring high profits — they can be stimulated with additional offers.

VIP and loyal customers: have the highest F_score and M_score values, low DTI and long service history.

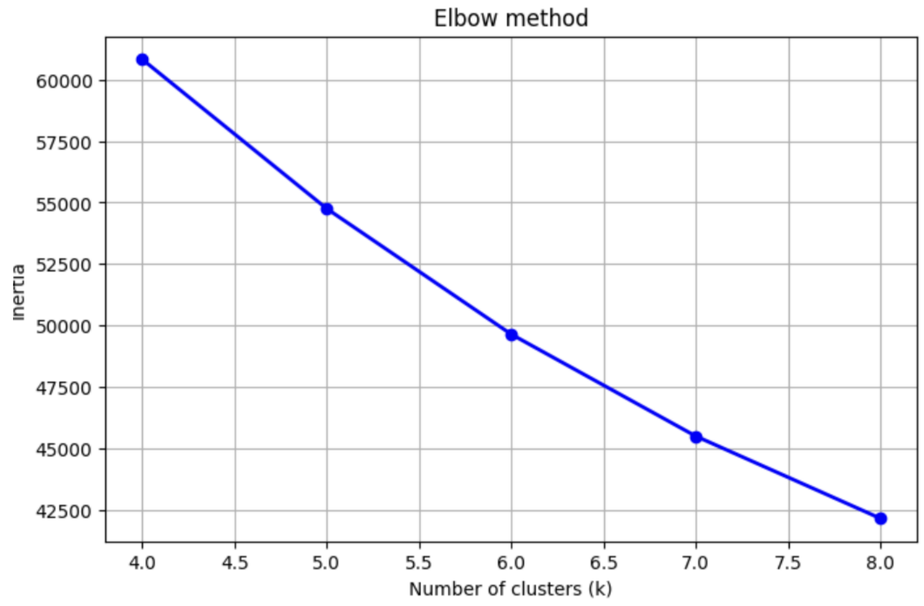


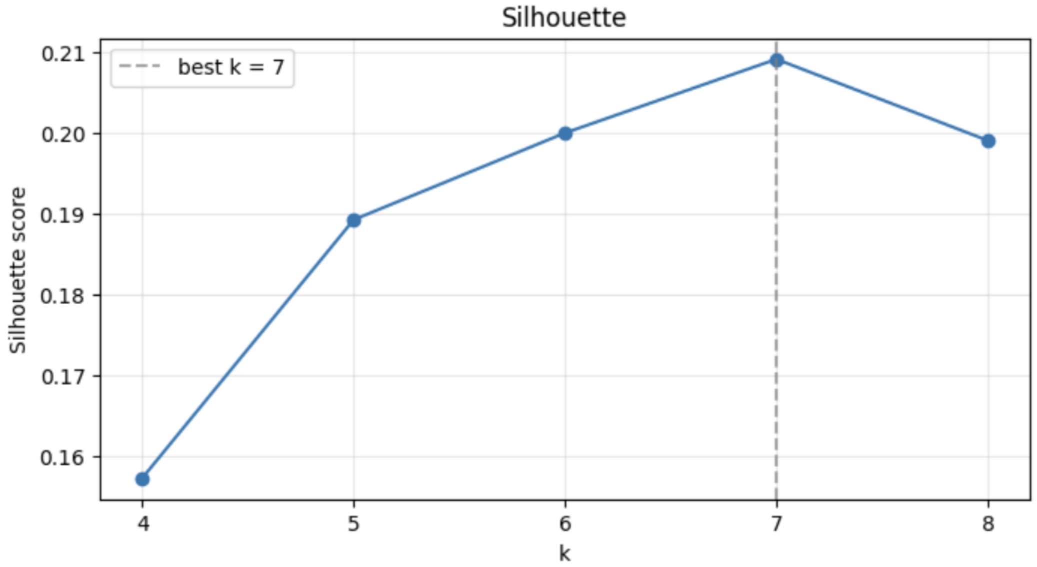
Dormant or inactive customers: average or low M_score values, but good reliability (R_score) — potentially safe, but require activation

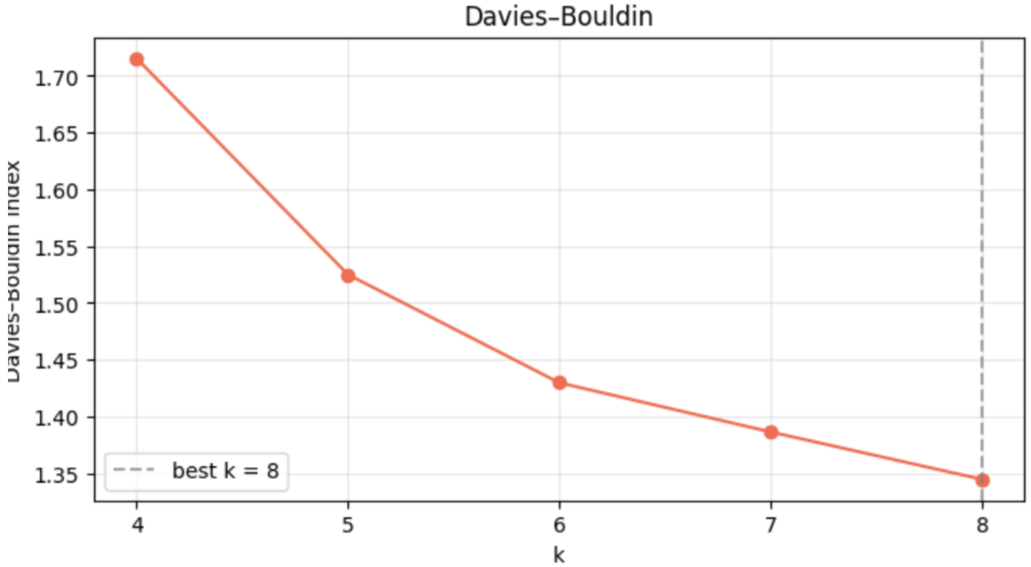
Active and stable customers: slightly lower scores than A, but also demonstrate good activity and low risk.

Least valuable or risky customers: low F_score and M_score values, high DTI and minimal experience — high risk and low profitability.

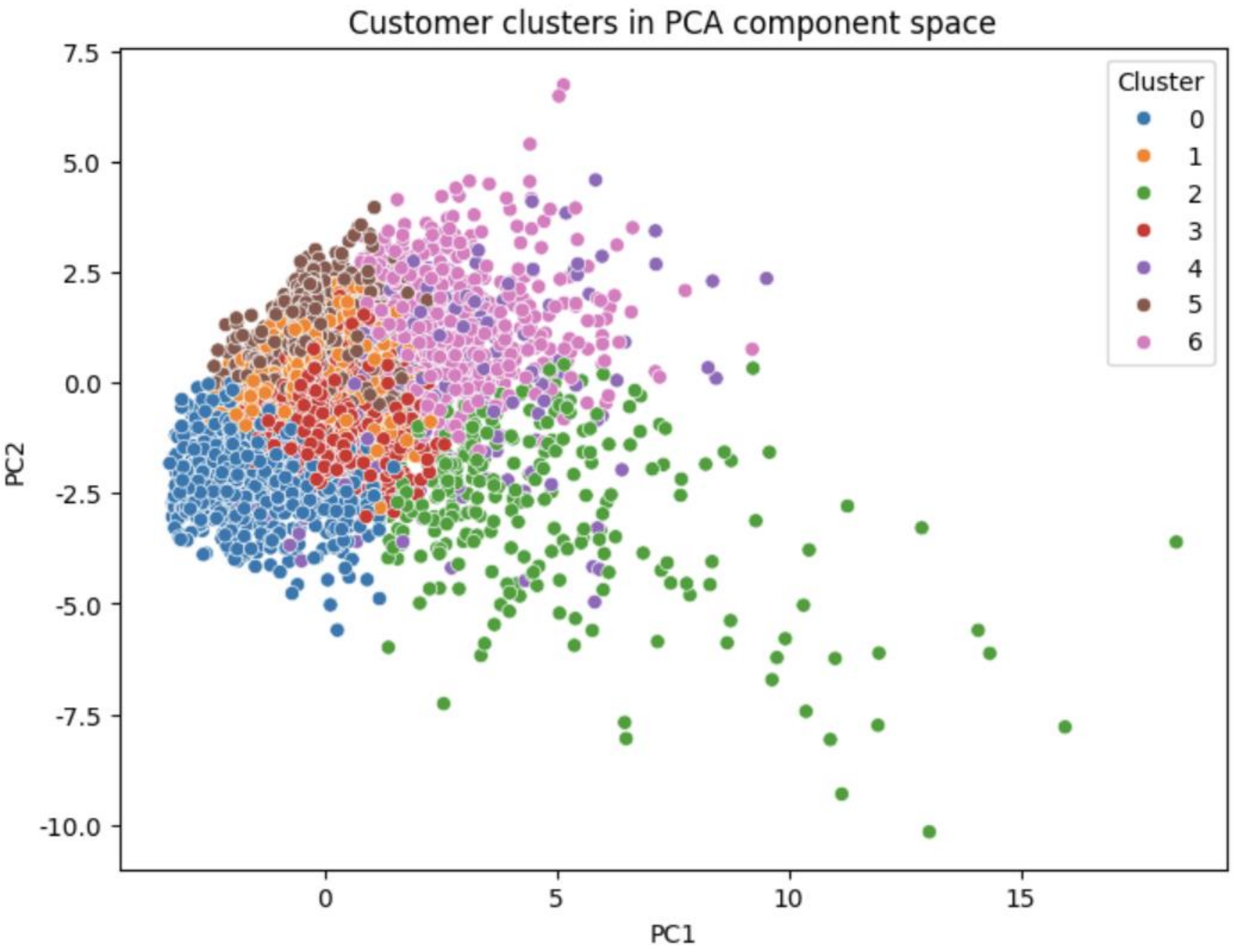
K-means





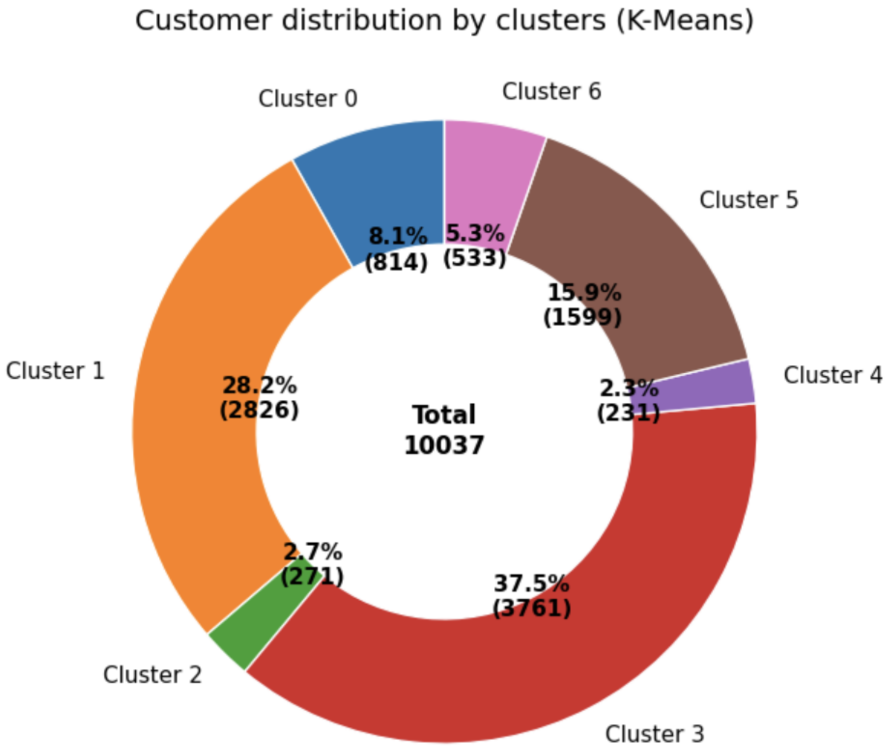


Three methods were used, and they showed that the best number of clusters is 7, so **k = 7** was selected



As we can see clusters are fairly clearly separated, especially the second green cluster, which stands out visually from the main mass of customers.

The other clusters (0, 1, 3, 5, 6) partially overlap, indicating similar customer characteristics, but the presence of separate areas shows that the K-Means algorithm has **successfully** identified distinct groups.



Most customers fall into clusters 3 (37.5%) and 1 (28.2%), indicating their predominance in the sample. The smallest clusters are clusters 2 and 4 (≈ 2–3% each), i.e. narrower or more specific customer groups

K-means clusters
Total clients: **10,037**

Cluster 0 (Young beginners) 817 (8.1%)

- Age: around 33
- DTI: low (≈0.32)
- Work experience: short (around 2.5 years)
- Activity: low (few products), about 2–3 loans
- Delinquency risk: moderate (30/90 days ≈0.6/1.0)

Cluster 1 (Stable family clients) 2,825 (28.1%)

- Age: around 35–36
- DTI: medium (≈0.38)
- Work experience: steady (around 4 years)
- Activity: moderate, about 5 loans, around 1–2 children
- Delinquency risk: low

Cluster 2 (High-risk borrowers) 273 (2.7%)

- Age: around 33
- DTI: high (≈0.41)
- Work experience: average (around 2.5 years)
- Activity: high, around 5–6 loans
- Delinquency risk: high (30/90 days ≈3.8/5.6; frequent 90-day delinquencies)

Cluster 3 (Main young group) 3,752 (37.4%)

- Age: around 32
- DTI: medium (≈0.39)
- Work experience: around 2.8 years
- Activity: moderate, around 5 loans
- Delinquency risk: minimal

Cluster 4 (Multi-loan clients with delinquencies) 235 (2.3%)

- Age: around 36–37
- DTI: medium (≈0.38)
- Work experience: around 3 years
- Activity: high, about 7 loans
- Delinquency risk: noticeable (30/90 days ≈0.8/1.6; some 90-day cases per year)

Cluster 5 (Mature reliable clients) 1,599 (15.9%)

- Age: around 47
- DTI: medium (≈0.40)
- Work experience: long (around 6.5 years)
- Activity: moderate, around 5 loans
- Delinquency risk: low

Cluster 6 (Highly active clients with high load) 536 (5.3%)

- Age: around 37
- DTI: high (≈0.44)
- Work experience: around 3.5 years
- Activity: high (multiple products, around 11 loans)
- Delinquency risk: moderate (30/90 days ≈1.0/2.4)

HSE UNIVERSITY • EST 1992