Java:
Automobile company

Content

1) What is at Java?
2) Dataset
3) Exploratory Data Analysis

At Java

- Enter new market with 5 products
- Current market:
    - 4 segments (A,B,C,D)

Task: Predict the new customers' segment

Dataset

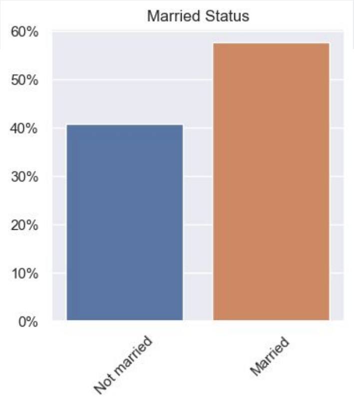
- 8068 clients records
- 11 features
- Target variable: Segmentation

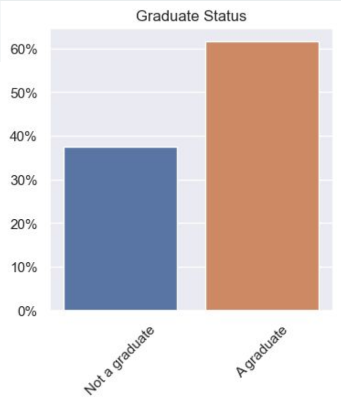
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8068 entries, 0 to 8067
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               8068 non-null   int64
 1   Gender           8068 non-null   object
 2   Ever_Married     7928 non-null   object
 3   Age              8068 non-null   int64
 4   Graduated        7990 non-null   object
 5   Profession       7944 non-null   object
 6   Work_Experience  7239 non-null   float64
 7   Spending_Score   8068 non-null   object
 8   Family_Size      7733 non-null   float64
 9   Var_1            7992 non-null   object
 10  Segmentation     8068 non-null   object
dtypes: float64(2), int64(2), object(7)
memory usage: 693.5+ KB
```

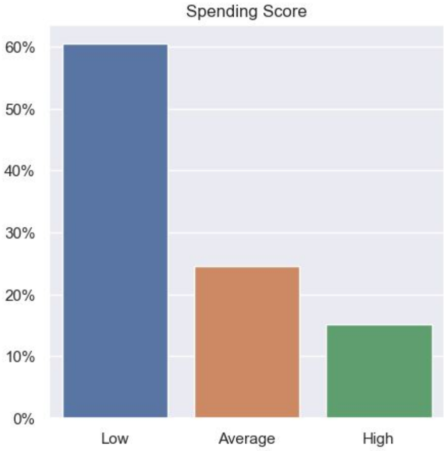
Exploratory Data Analysis

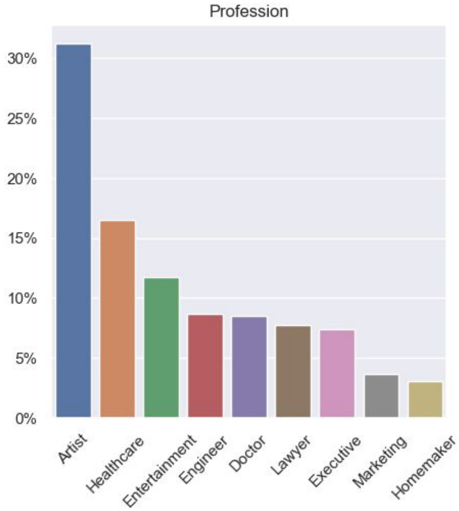
Most of the customers of the automobile company is:

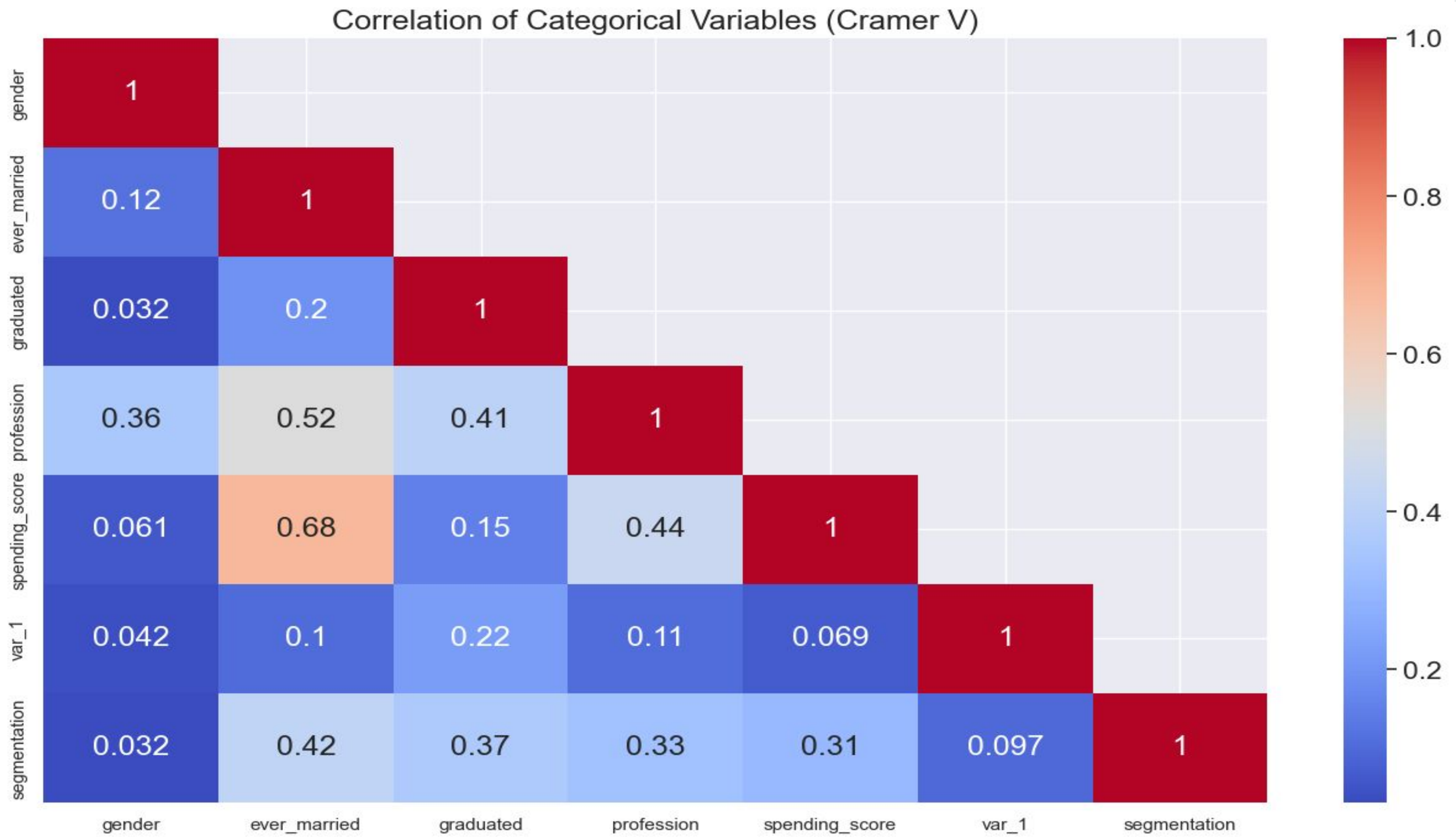
- 54.75% male
- 57.55% married
- 61.58% a graduate
- 31.18% an artist
- 60.46% of a low spender score
- 64.92% in an anonymised category 6





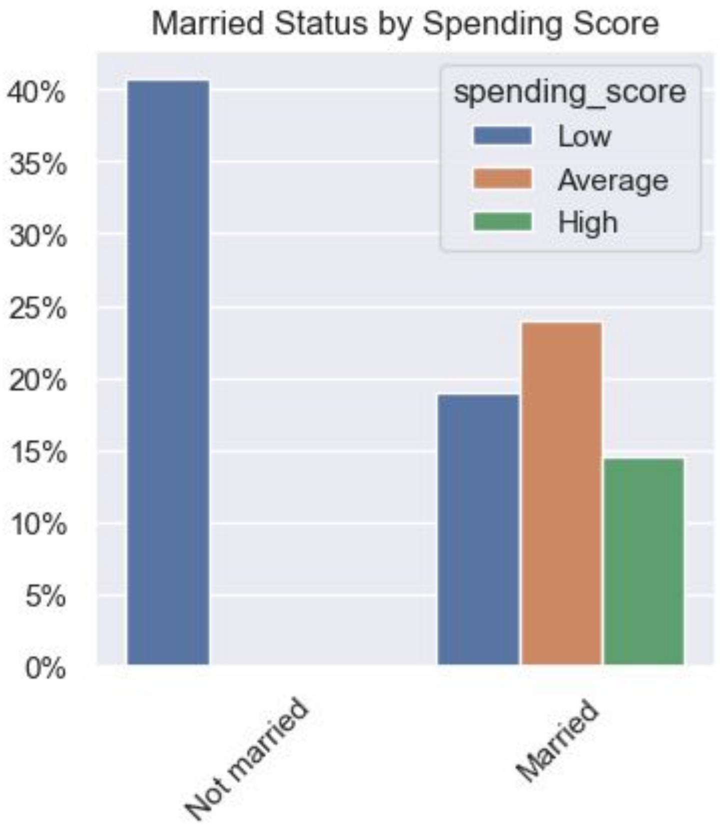


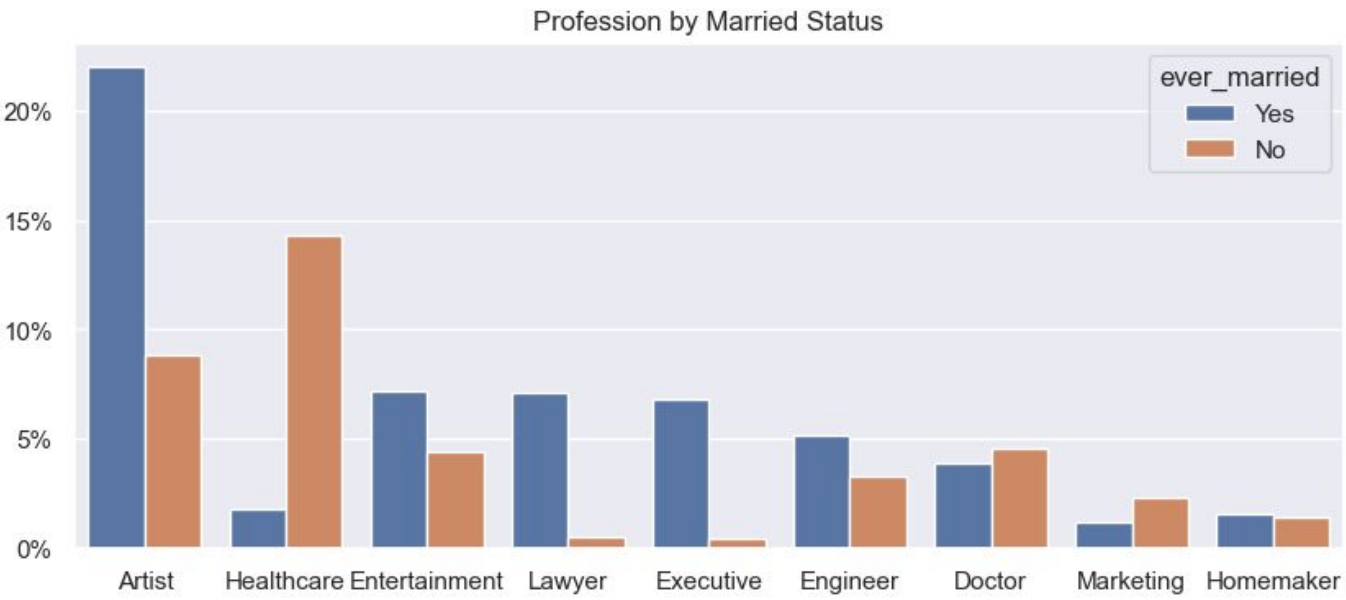



Correlation of Categorical Variables (Cramer V)
gender
ever_married
graduated
profession
spending_score
var_1
segmentation
1
0.12
1
0.032
0.2
1
0.36
0.52
0.41
1
0.061
0.68
0.15
0.44
1
0.042
0.1
0.22
0.11
0.069
1
0.032
0.42
0.37
0.33
0.31
0.097
1
1.0
0.8
0.6
0.4
0.2

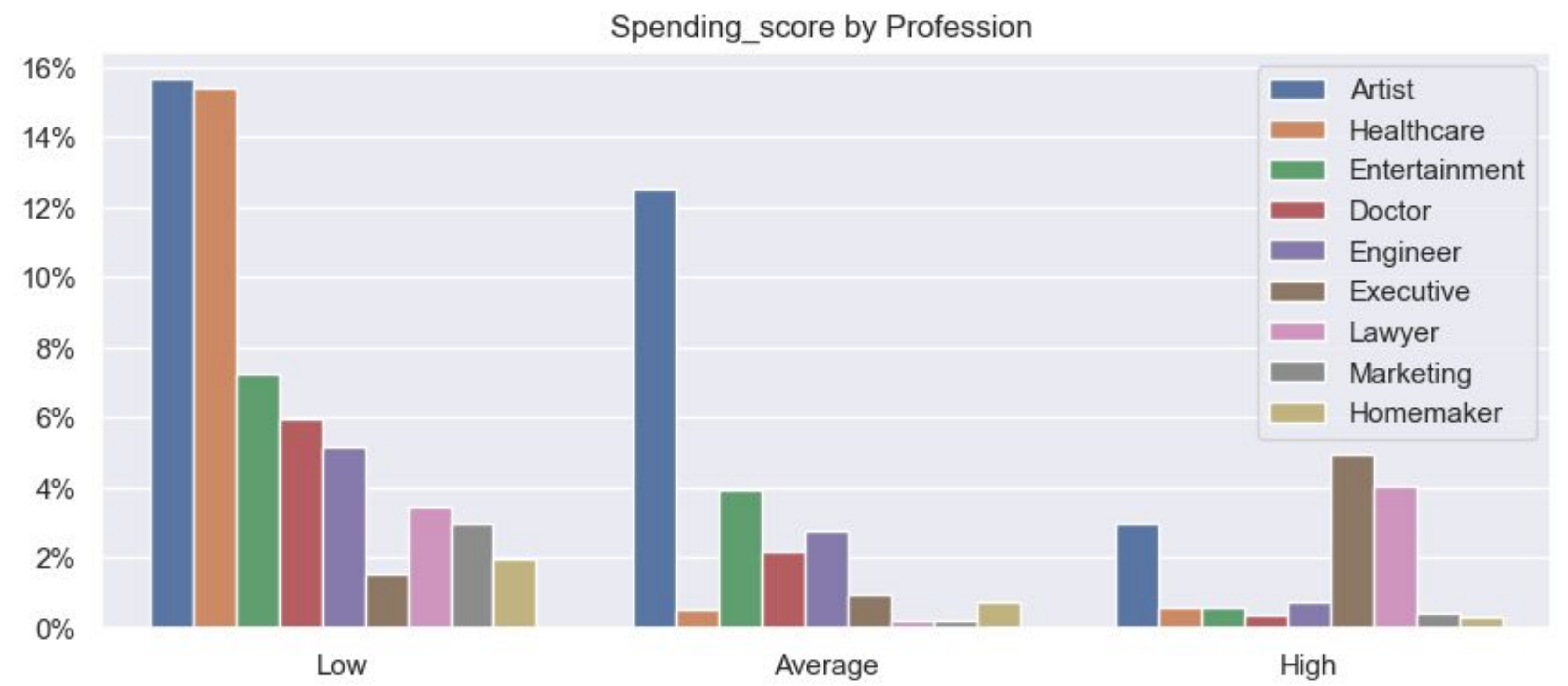
Relationships

- ever_married have the highest correlation at 0.68 with spending_score
- non-married people are all low-spenders (40%) compared to married people who fall in the 3 categories



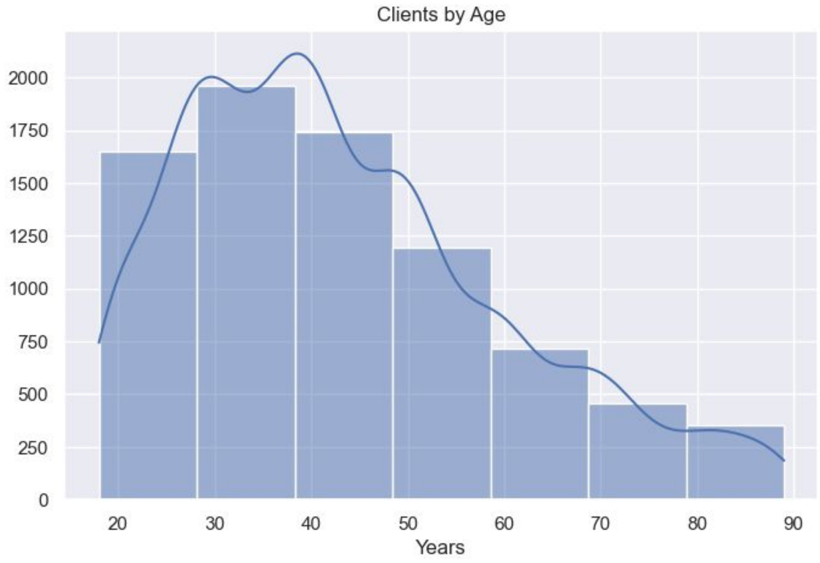


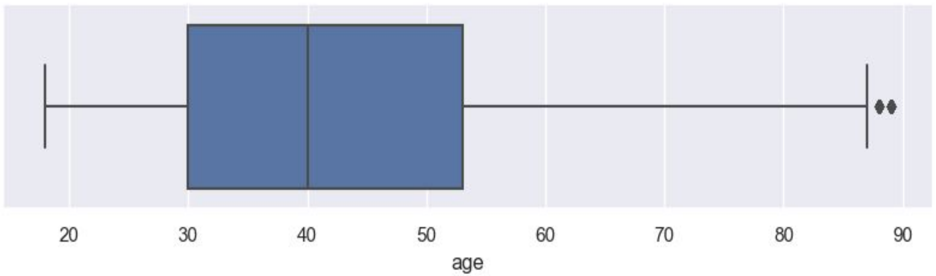
- ever_married and profession at 0.52
- single artists (22%) and married healthcarers (14%) are the biggest customers

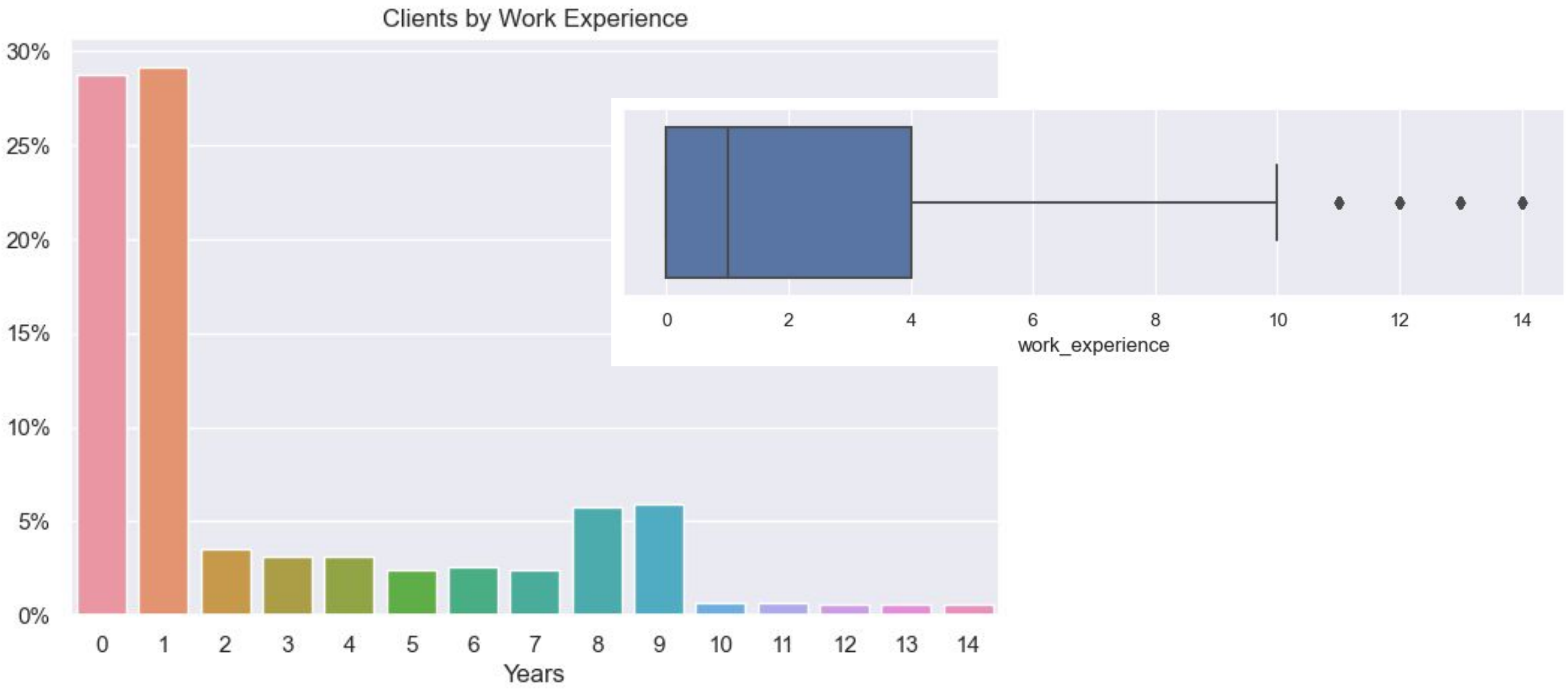


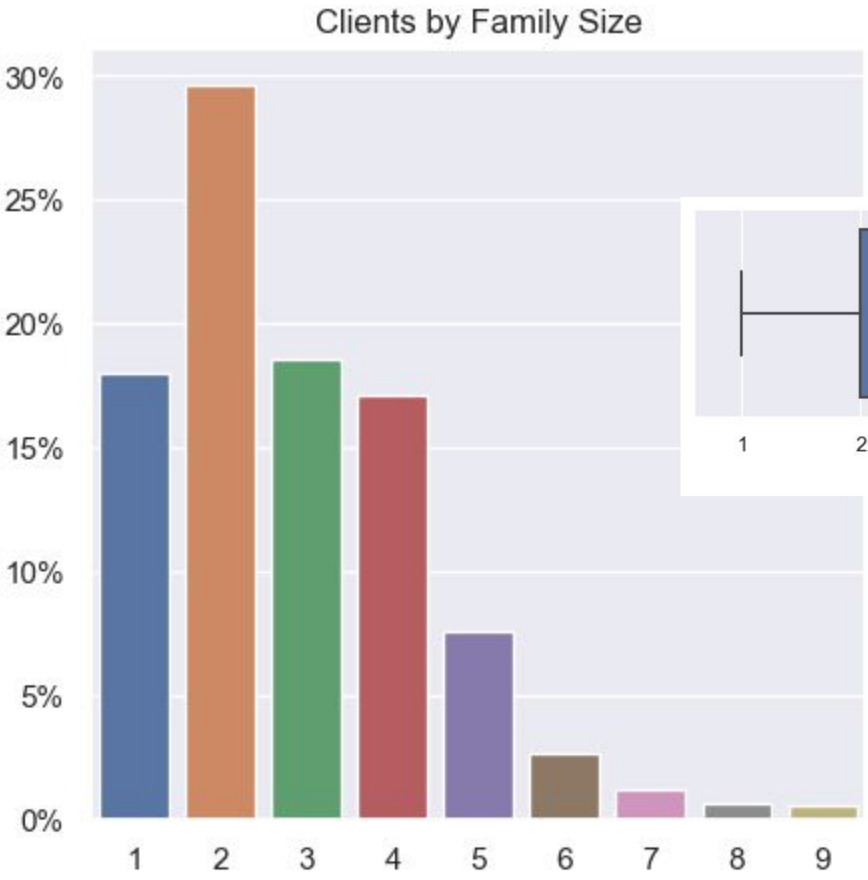
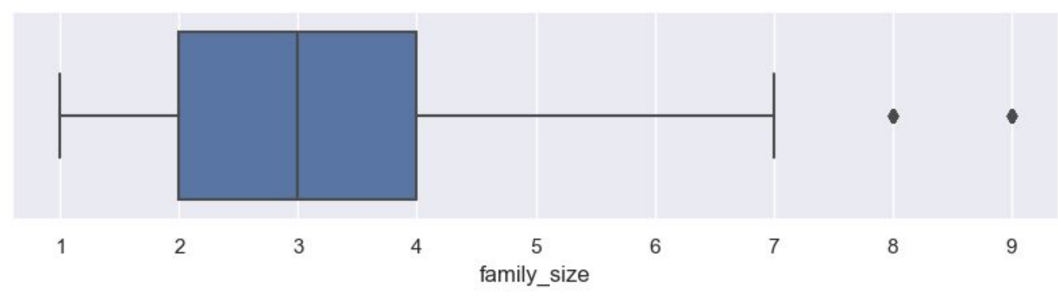
- spending_score and profession : 0.44
- artist & healthcare are the biggest customers in the low-spending-score
- artists are the biggest customers in the mid-spending-score
- executives & lawyers are the biggest customers in the high-spending-score

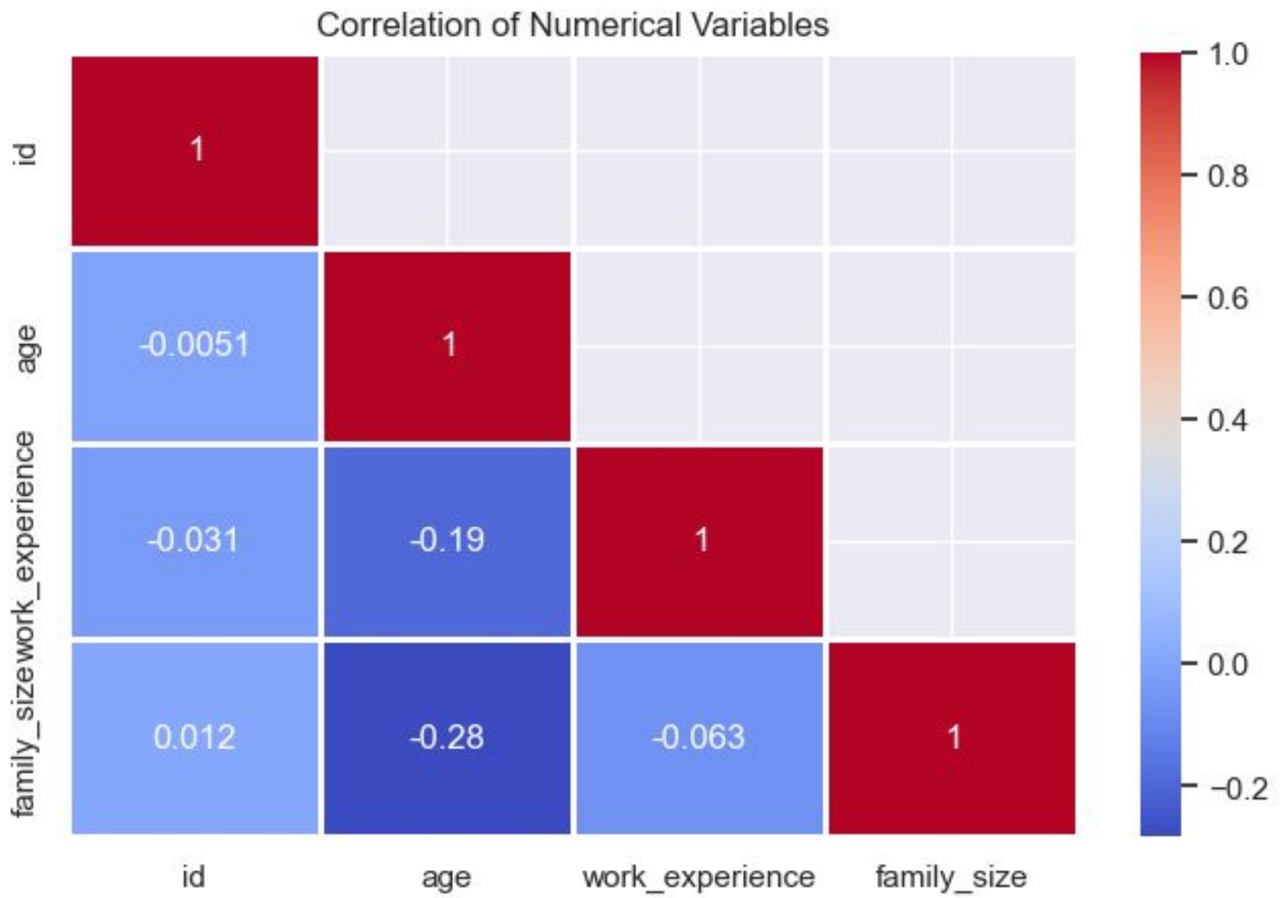
Numerical variables






Clients by Work Experience
30%
25%
20%
15%
10%
5%
0%
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Years
0
2
4
6
8
10
12
14
work_experience


Clients by Family Size
30%
25%
20%
15%
10%
5%
0%
1
2
3
4
5
6
7
8
9

1
2
3
4
5
6
7
8
9
family_size



Correlation of Numerical Variables
id
age
work_experience
family_size
1
-0.0051
1
-0.031
-0.19
1
0.012
-0.28
-0.063
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Cleaning

1) Missing values:
   a) Mode
   b) median
2) No removal of outliers
3) Cleaning columns
   a) family_size/work_experience to integer type
   b) Lower case

	percent_missing
Work_Experience	10.28
Family_Size	4.15
Ever_Married	1.74
Profession	1.54
Graduated	0.97
Var_1	0.94
ID	0.00
Gender	0.00
Age	0.00
Spending_Score	0.00
Segmentation	0.00

Encoding

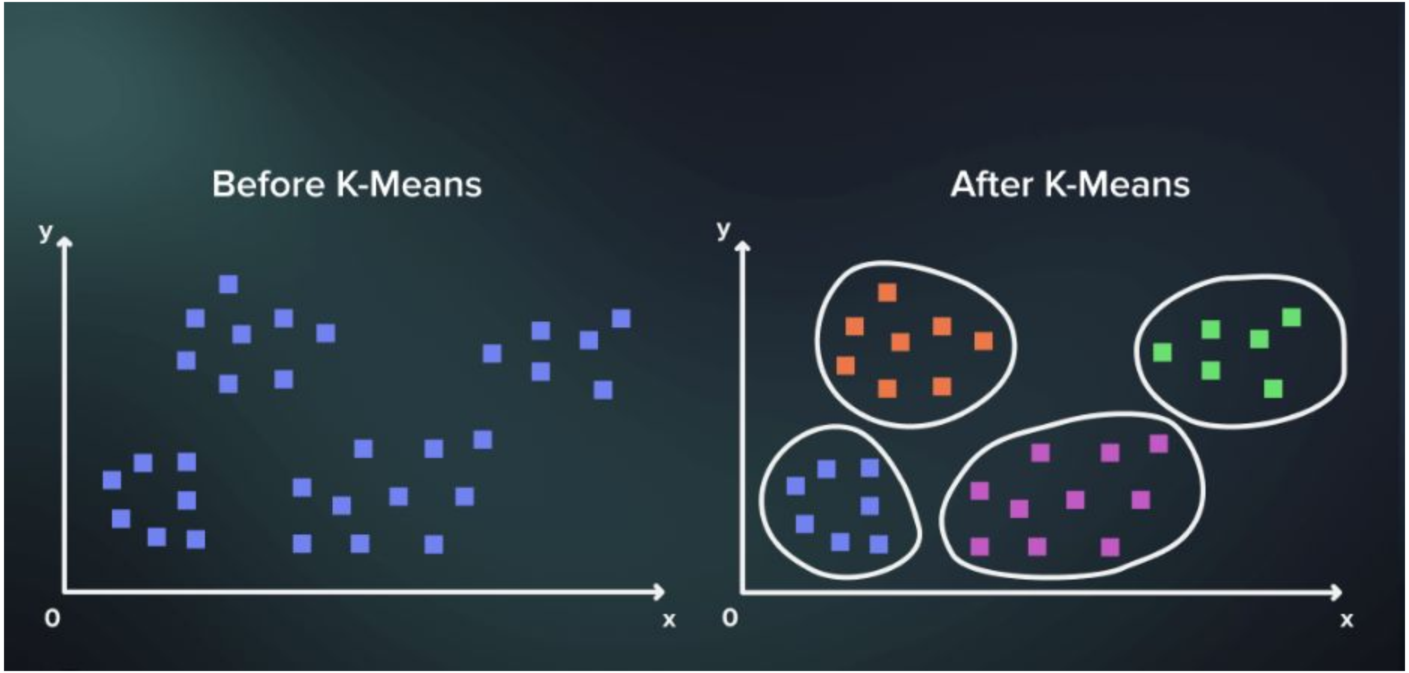
	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
0	462809	Male	No	22	No	Healthcare	1.0	Low	4.0	Cat_4	D
1	462643	Female	Yes	38	Yes	Engineer	NaN	Average	3.0	Cat_4	A
2	466315	Female	Yes	67	Yes	Engineer	1.0	Low	1.0	Cat_6	B
3	461735	Male	Yes	67	Yes	Lawyer	0.0	High	2.0	Cat_6	B
4	462669	Female	Yes	40	Yes	Entertainment	NaN	High	6.0	Cat_6	A

	id	gender	ever_married	age	graduated	work_experience	spending_score	family_size	anon_cat	Doctor	Engineer	Entertainment
0	462809	1	0	22	0	1	2.0	4	4	0	0	0
1	462643	0	1	38	1	1	0.0	3	4	0	1	0
2	466315	0	1	67	1	1	2.0	1	6	0	1	0
3	461735	1	1	67	1	0	1.0	2	6	0	0	0
4	462669	0	1	40	1	1	1.0	5	6	0	0	1

- One-hot encoding
    - Gender, ever_married, graduated
- Multicategories
    - Ordered: spending_score
    - Unordered: profession

- Standardisation: Min-Max Scale

K Means Clustering



Experimentation

#1: **32%:** just standardise-scaling all variables

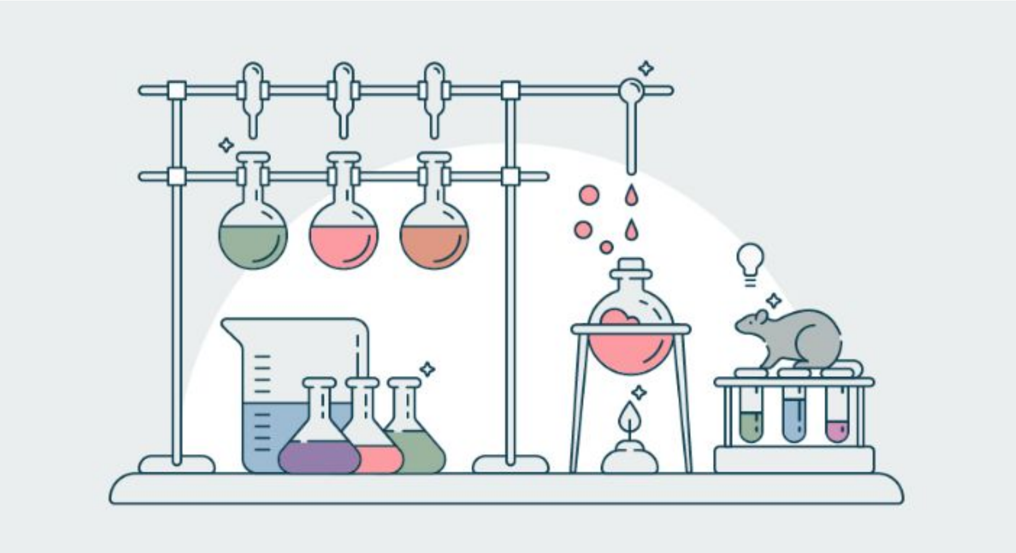
#2:**40%** 1+ deleting id, anon_cat

#3: **29%** 1+ deleted id, anon_cat, gender

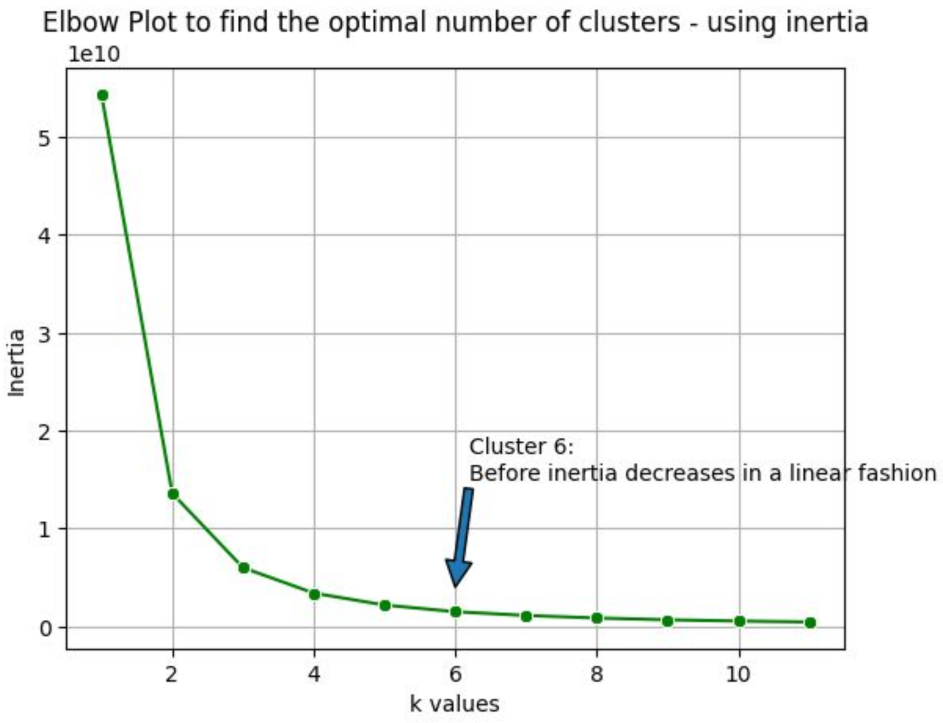
#4: 41.5% deleting id, anon_cat, min max scale all va didn't replace outliers <--- best one

#5: **17%** 1+ 4 + mean instead of median

#6: **28%** 1+ 4 + replacing outliers



Optimal number of clusters



Centroids

```
Final centroids:
[[ 4.64508230e+05  5.47856431e-01  6.05184447e-01  6.14656032e-01
   2.20887338e+00  1.29661017e+00  7.97607178e-02  1.22133599e-01
   1.08175474e-01  9.52143569e-02  1.53539382e-01  1.14656032e-02
   5.83250249e-02  4.18743769e-02 -1.37906726e-02  1.17205653e-01
  -4.27180774e-02]
 [ 4.62276731e+05  5.58662007e-01  5.95107339e-01  6.12081877e-01
   2.29505741e+00  1.31003495e+00  1.03344983e-01  8.28756865e-02
   1.03844234e-01  7.48876685e-02  1.64253620e-01  4.09385921e-02
   7.08936595e-02  3.24513230e-02 -2.61669984e-02  2.41934730e-02
   3.61637796e-02]
 [ 4.66806917e+05  5.43033761e-01  5.96766524e-01  6.34807418e-01
   2.41892534e+00  1.39752734e+00  8.36899667e-02  7.18021874e-02
   1.22206372e-01  6.60960533e-02  1.78792202e-01  3.80408940e-02
   8.41654779e-02  3.09082263e-02  1.52061186e-02 -6.97605297e-02
   3.82173672e-02]
 [ 4.60077482e+05  5.40388548e-01  5.73619632e-01  6.40081800e-01
   2.62321063e+00  1.43558282e+00  7.41308793e-02  7.00408998e-02
   1.36503067e-01  6.08384458e-02  1.63087935e-01  3.11860941e-02
   9.56032720e-02  3.98773006e-02  2.45900408e-02 -6.99732480e-02
  -3.43122241e-02]]
```

Accuracy: 30-40%

	id	gender	ever_married	age	graduated	profession	work_experience	spending_score	family_size	var_1	segmentation	labels
0	462809	Male	No	22	No	Healthcare	1.0	Low	4.0	Cat_4	D	D
2	466315	Female	Yes	67	Yes	Engineer	1.0	Low	1.0	Cat_6	B	B
7	464347	Female	No	33	Yes	Healthcare	1.0	Low	3.0	Cat_6	D	D
11	464942	Male	No	19	No	Healthcare	4.0	Low	4.0	Cat_4	D	D
13	459573	Male	Yes	70	No	Lawyer	NaN	Low	1.0	Cat_6	A	A
...	...	...	...	...	...	...	...	...	...	...	...	...
8052	467455	Female	No	37	Yes	Artist	8.0	Low	2.0	Cat_6	C	C
8053	465667	Male	No	23	No	Healthcare	1.0	Low	3.0	Cat_2	D	D
8055	461291	Male	No	18	No	Healthcare	0.0	Low	2.0	Cat_6	D	D
8059	460132	Male	No	39	Yes	Healthcare	3.0	Low	2.0	Cat_6	D	D
8065	465406	Female	No	33	Yes	Healthcare	1.0	Low	1.0	Cat_6	D	D

2485 rows × 12 columns

Next

- automate the cleaning and K Means model pipeline
- try one-hot-encoding on binary features:'gender', 'married','graduated'
- try other ways of dealing with outliers
- try to log 'age' so the distribution is closer to the gaussian distribution

Conclusion

- Possibility of 6 well targeted segments of customers
- Company’s clients are mostly young people


Thank You!
95