CREDIT CARD CUSTOMER SEGMENTATION

Kardelen E.

Contents

1. Introduction
2. Basic Data Exploration
3. Choosing Cluster Number
4. Analyzing Results
   4.1 Numerical Columns
   4.1 Categorical Columns
5. Summary of the Clusters
6. Conclusion - Potential Next Steps

1. Introduction

In this project, we’ll play the role of a data scientist working for a credit card company. Our dataset contains information about the company’s clients and we're asked to help **segment the clients into different groups** in order to apply different business strategies for each type of customer.

For instance, the company could provide higher credit limits for customers that use the card a lot, but spend little money, or even create incentives for those with high income who don't use the card as much as the company expects. In order to apply different strategies, the company needs different groups of customers.

We’ll use the **K-means algorithm** to segment the data. The company expects to receive a group for each client and an explanation of the characteristics of each group and the main points that make them different.

2. Basic Data Exploration - Data Dictionary

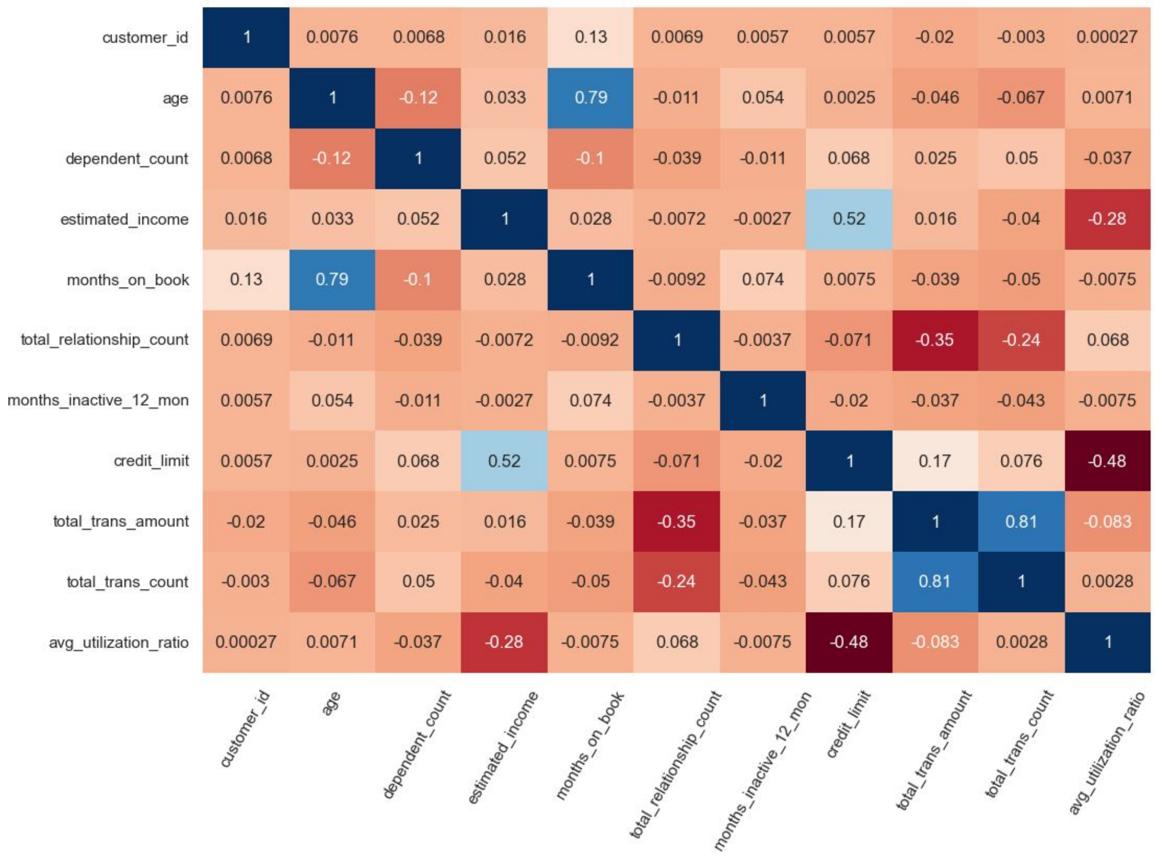
Here's the data dictionary:

- customer_id: unique identifier for each customer.
- age: customer age in years.
- gender: customer gender (M or F).
- dependent_count: number of dependents of each customer.
- education_level: level of education ("High School", "Graduate", etc.).
- marital_status: marital status ("Single", "Married", etc.).
- estimated_income: the estimated income for the customer projected by the data science team.
- months_on_book: time as a customer in months.
- total_relationship_count: number of times the customer contacted the company.
- months_inactive_12_mon: number of months the customer did not use the credit card in the last 12 months.
- credit_limit: customer's credit limit.
- total_trans_amount: the overall amount of money spent on the card by the customer.
- total_trans_count: the overall number of times the customer used the card.
- avg_utilization_ratio: daily average utilization ratio.

2. Basic Data Exploration - Correlations

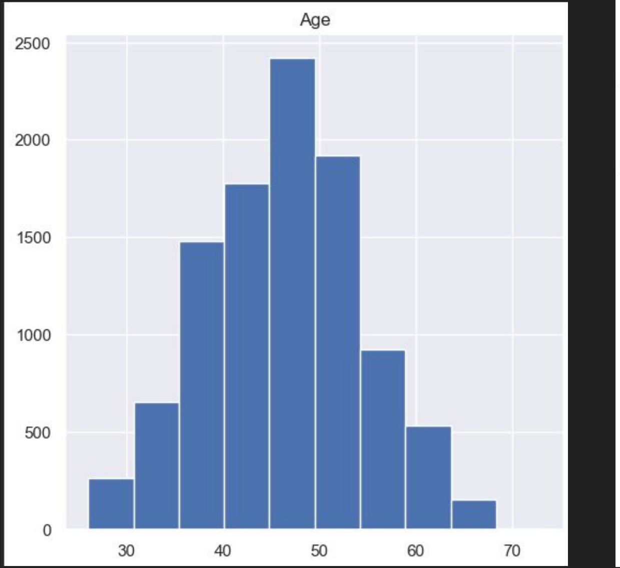
The columns age and months_on_book are correlated, as one would expect.

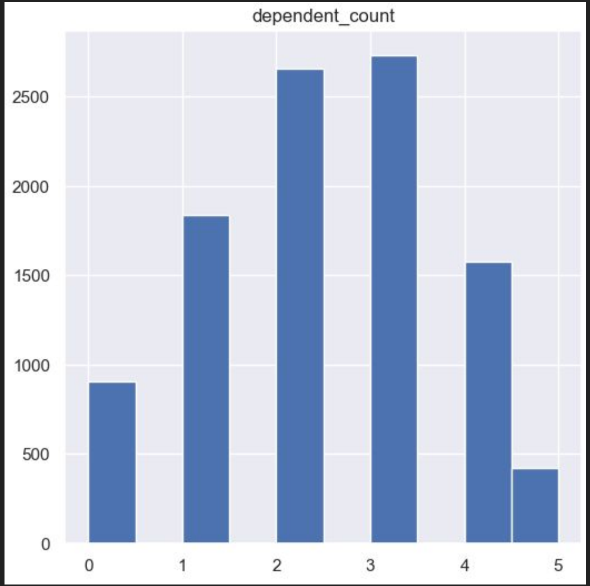
The columns estimated_income and credit_limit are also highly correlated, as well as total_trans_amount and total_trans_count.

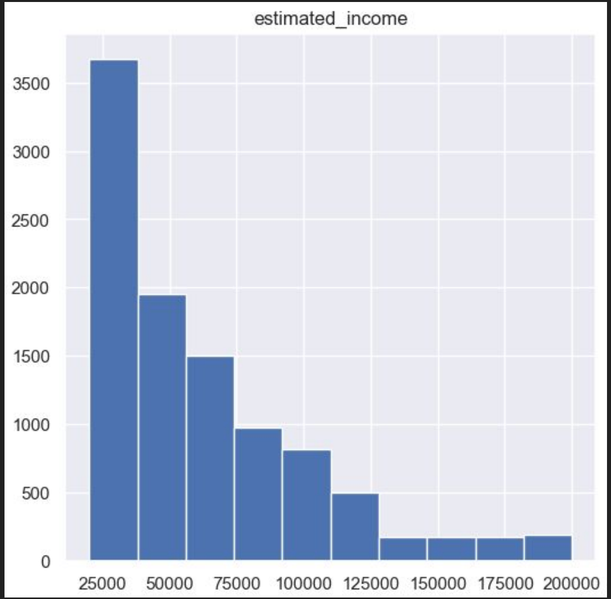


	customer_id	age	dependent_count	estimated_income	months_on_book	total_relationship_count	months_inactive_12_mon	credit_limit	total_trans_amount	total_trans_count	avg_utilization_ratio
customer_id	1	0.0076	0.0068	0.016	0.13	0.0069	0.0057	0.0057	-0.02	-0.003	0.00027
age	0.0076	1	-0.12	0.033	0.79	-0.011	0.054	0.0025	-0.046	-0.067	0.0071
dependent_count	0.0068	-0.12	1	0.052	-0.1	-0.039	-0.011	0.068	0.025	0.05	-0.037
estimated_income	0.016	0.033	0.052	1	0.028	-0.0072	-0.0027	0.52	0.016	-0.04	-0.28
months_on_book	0.13	0.79	-0.1	0.028	1	-0.0092	0.074	0.0075	-0.039	-0.05	-0.0075
total_relationship_count	0.0069	-0.011	-0.039	-0.0072	-0.0092	1	-0.0037	-0.071	-0.35	-0.24	0.068
months_inactive_12_mon	0.0057	0.054	-0.011	-0.0027	0.074	-0.0037	1	-0.02	-0.037	-0.043	-0.0075
credit_limit	0.0057	0.0025	0.068	0.52	0.0075	-0.071	-0.02	1	0.17	0.076	-0.48
total_trans_amount	-0.02	-0.046	0.025	0.016	-0.039	-0.35	-0.037	0.17	1	0.81	-0.083
total_trans_count	-0.003	-0.067	0.05	-0.04	-0.05	-0.24	-0.043	0.076	0.81	1	0.0028
avg_utilization_ratio	0.00027	0.0071	-0.037	-0.28	-0.0075	0.068	-0.0075	-0.48	-0.083	0.0028	1

2. Basic Data Exploration - Charts





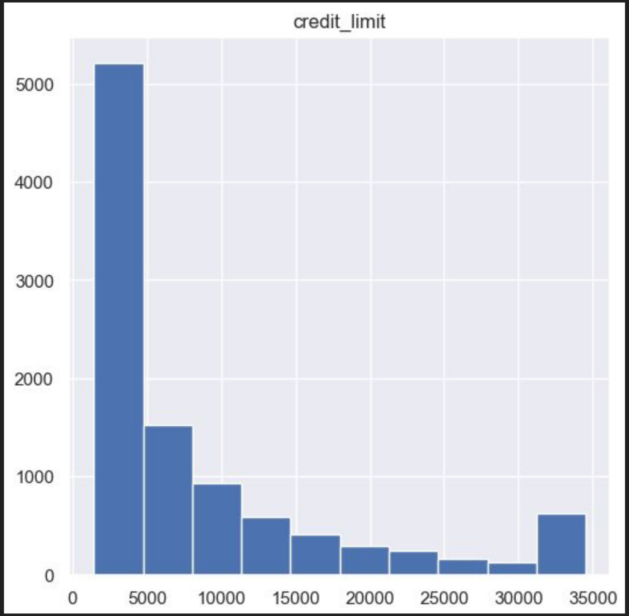


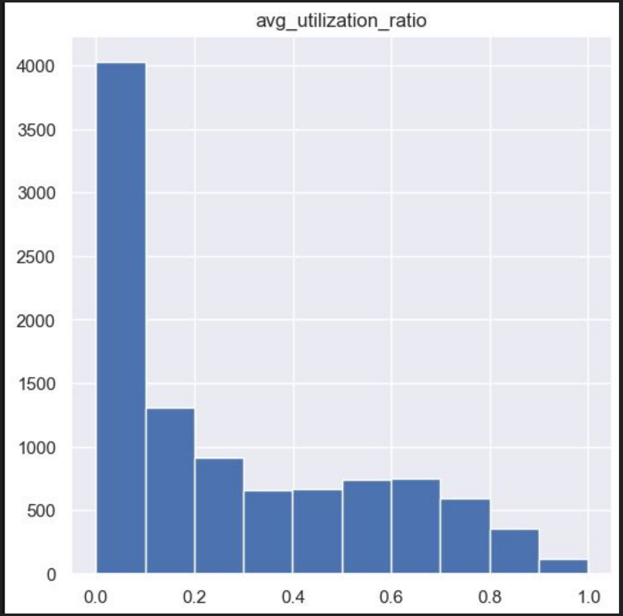
These charts show the distribution within the columns.

2. Basic Data Exploration - Charts

Several observations:

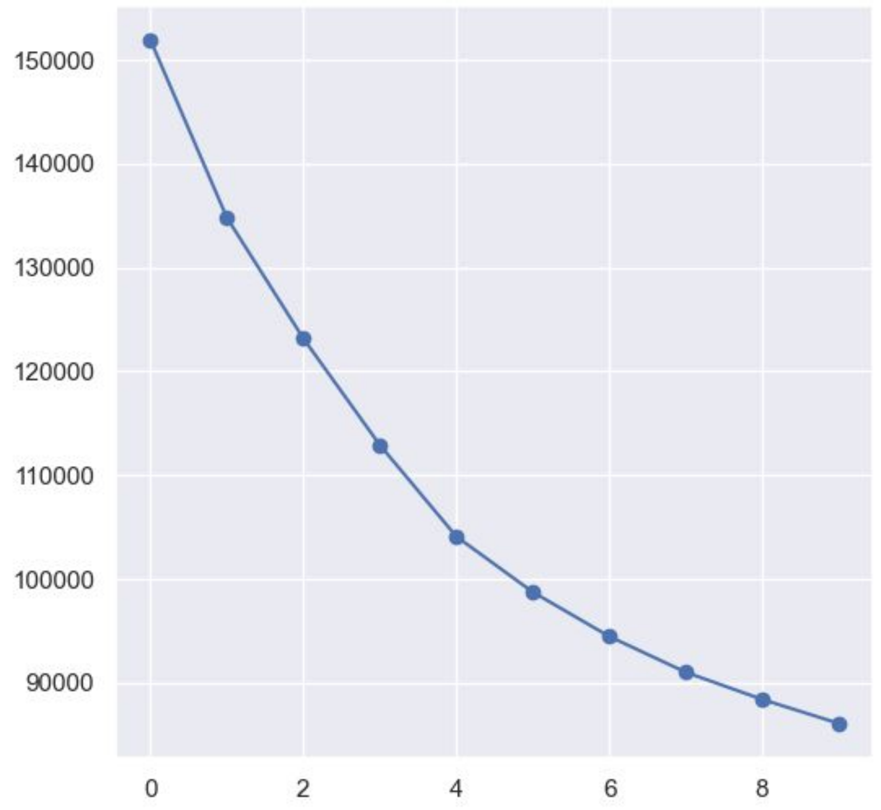
- age and dependent_count columns have normal distribution.
- estimated_income, credit_limit and avg_utilization_ratio columns are right-skewed. We may see sharp differences between clusters in terms of these columns.





3. Choosing Cluster Number

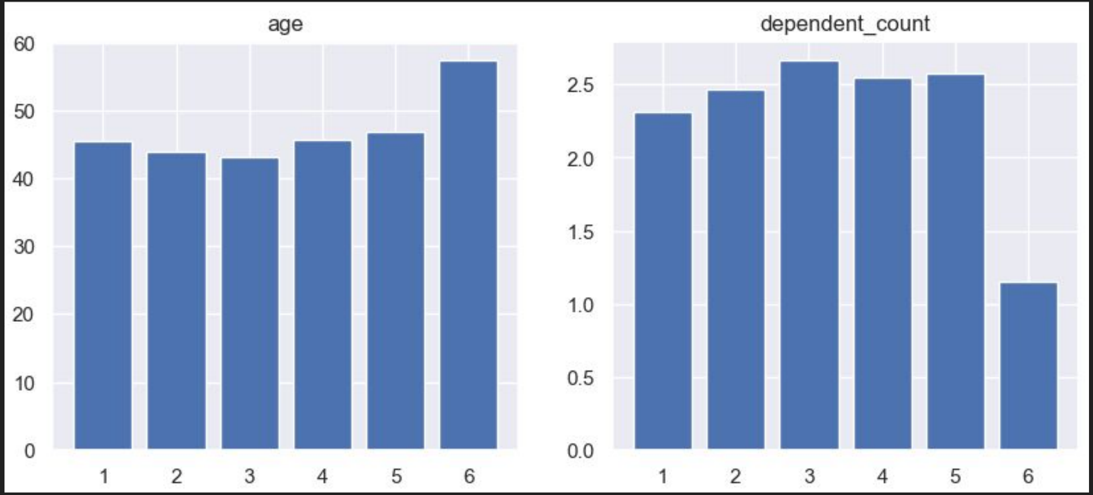
In order to decide on how many clusters we want to create, we will utilize the elbow curve. We wrote a code that used the numbers 1-10 as number of clusters and calculated their inertias. You can see their visualization on the right. Although we don't see a sharp elbow, we can see that the decrease in the inertia slows down after number 6. For this reason, we will choose to segment our data into 6 clusters.

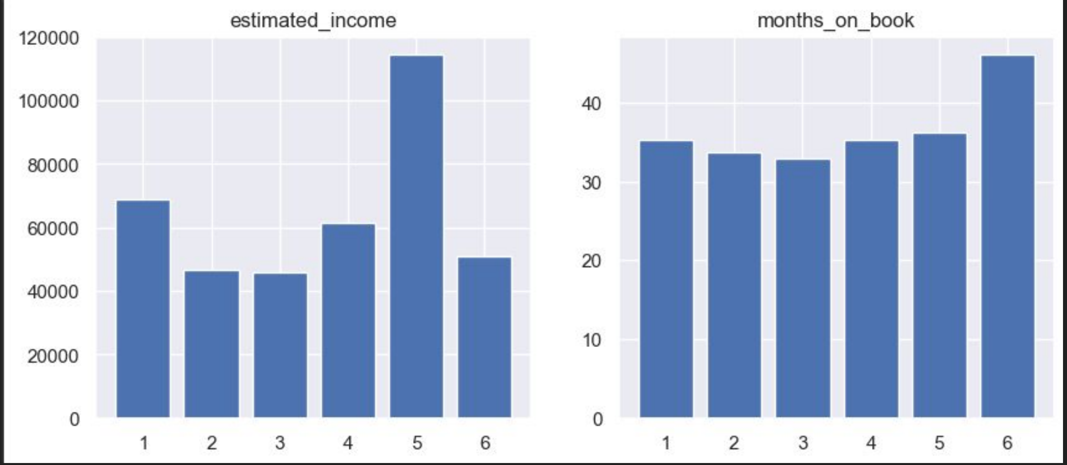


1. Analyzing Results

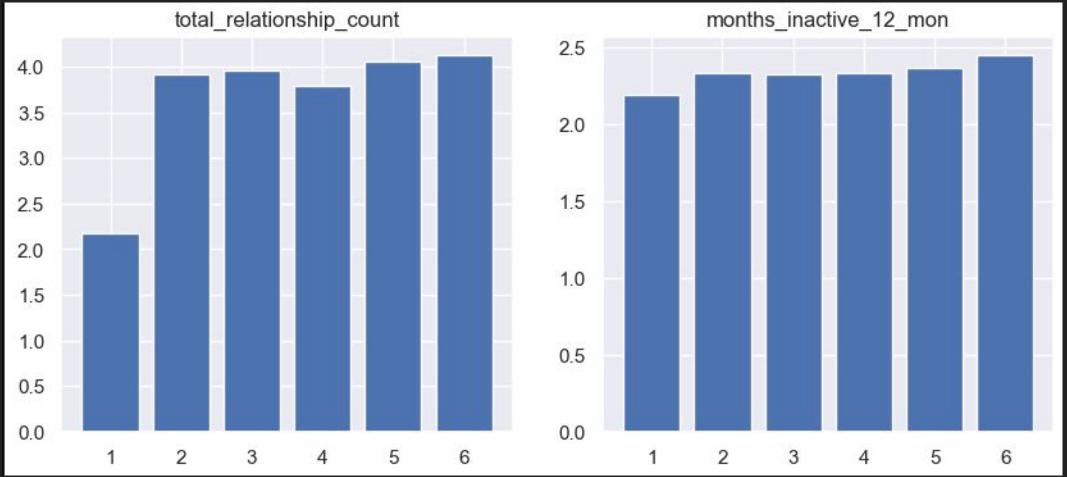
   4.1 Numerical Columns

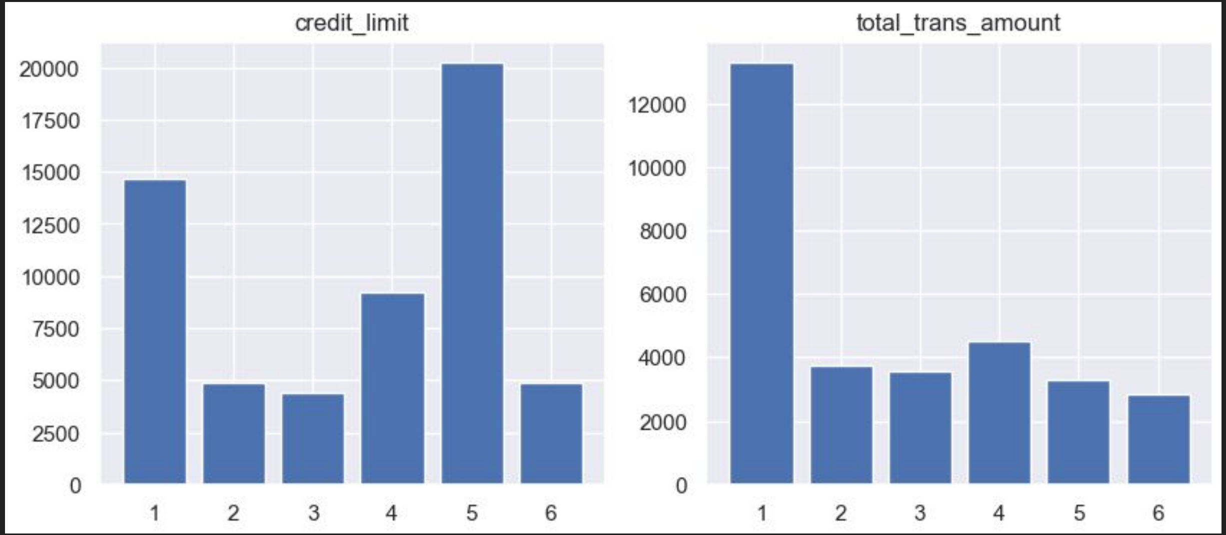
Numbers 1-6 on the x axis represent the cluster numbers. The charts show the mean of each cluster (for example, the mean age of cluster 6 is above 50)





1. Analyzing Results
   4.1 Numerical Columns

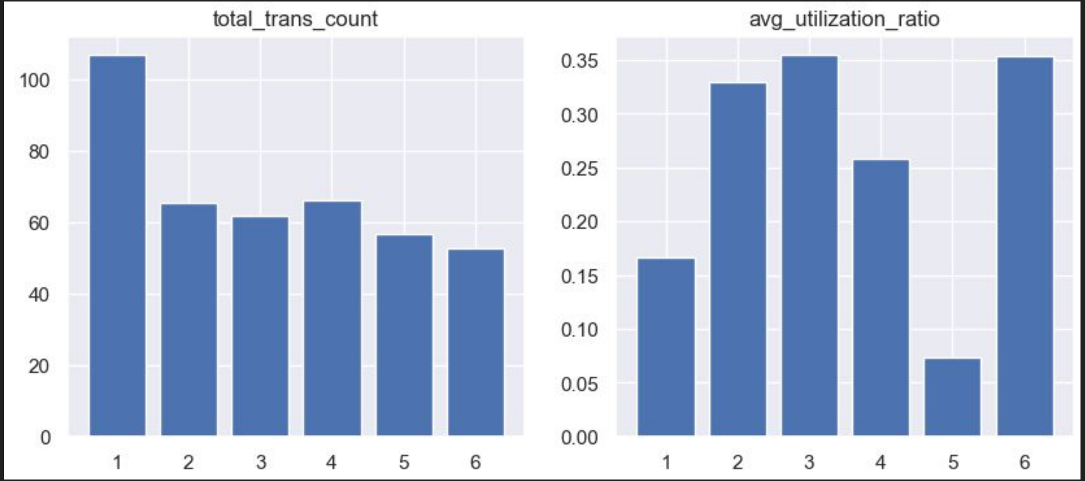




4. Analyzing Results
4.1 Numerical Columns

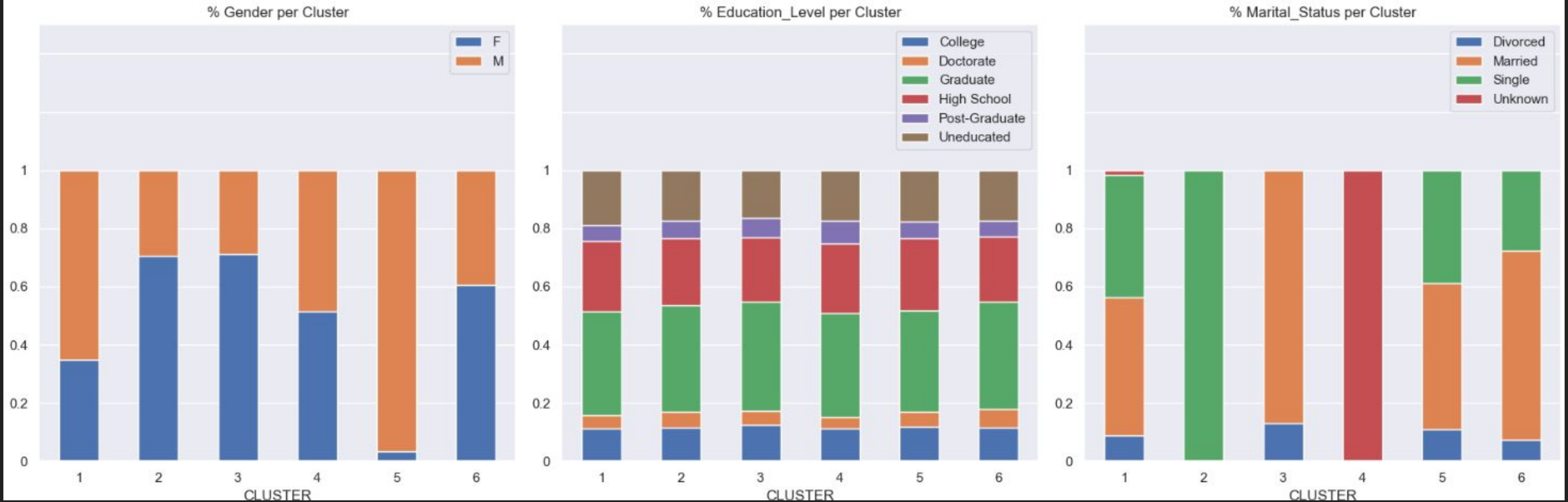
Some observations:

- Cluster 6 has the highest mean age. The rest of the clusters are similar in terms of mean age, with Cluster 3 being the youngest albeit not by much.
- Cluster 6 also has the least amount of dependents, with 3 having the most.
- Cluster 5 has the highest estimated income, with Clusters 2 and 3 being at the bottom with very similar numbers.
- The mean values for months_on_book are similar, with Cluster 6 on top.
- Cluster 1 has the lowest number in terms of total_relationship_count.
- Cluster 5, which has the highest estimated income also has the highest credit_limit.
- Despite having only the second largest credit_limit, total_trans_amount of Cluster 1 dwarfs that of Cluster 5. This information could be valuable for a campaign targeting Cluster 5.
- Cluster 6 has one of the lowest credit_limit mean values. However, they have the highest avg_utilization_ratio. Cluster 5 has the lowest.



1. Analyzing Results

4.2 Categorical Columns



4. Analyzing Results

4.2 Categorical Columns

Some remarks about categorical columns:

- Clusters 2 and 3 consist primarily of women, whereas Cluster 5 is almost entirely male. The rest are more uniform.
- There aren't big differences in terms of education level between clusters. The largest group is always "Graduate".
- Cluster 2 entirely consists of singles. Most customers on Cluster 3 are married.
- Marriage status of customers in Cluster 4 are unknown.

5. Summary of the Clusters

Cluster 1:

- 2/3 males
- Lowest total_relationship_count
- Second largest credit_limit, highest total_trans_amount by far
- Similar number of single and married customers

Cluster 2:

- 2/3 females
- One of the lowest estimated income
- Mostly singles

Cluster 3:

- 2/3 females
- One of the lowest estimated income
- Most amount of dependents
- Mostly married

Cluster 4:

- Almost the same number of males and females
- Marriage status is mostly unknown

Cluster 5:

- Mostly males
- Highest estimated income
- total_trans_amount unusually low despite the income
- Similar number of single and married customers

Cluster 6:

- Around 60% females
- Highest mean age
- Least amount of dependents
- One of the lowest credit_limit, highest avg_utilization_ratio
- More than half are marrieds

6. Conclusion - Potential Next Steps

We used K-Means to segment the customers into six clusters and explained the characteristics of each cluster. Stakeholders can use this data for various applications such as targeted marketing and campaigns.

Potential next steps include:

- trying out different numbers of clusters,
- researching other ways to find the best number of clusters and comparing the results with the results from the elbow method,
- using fewer variables in the clusterization.