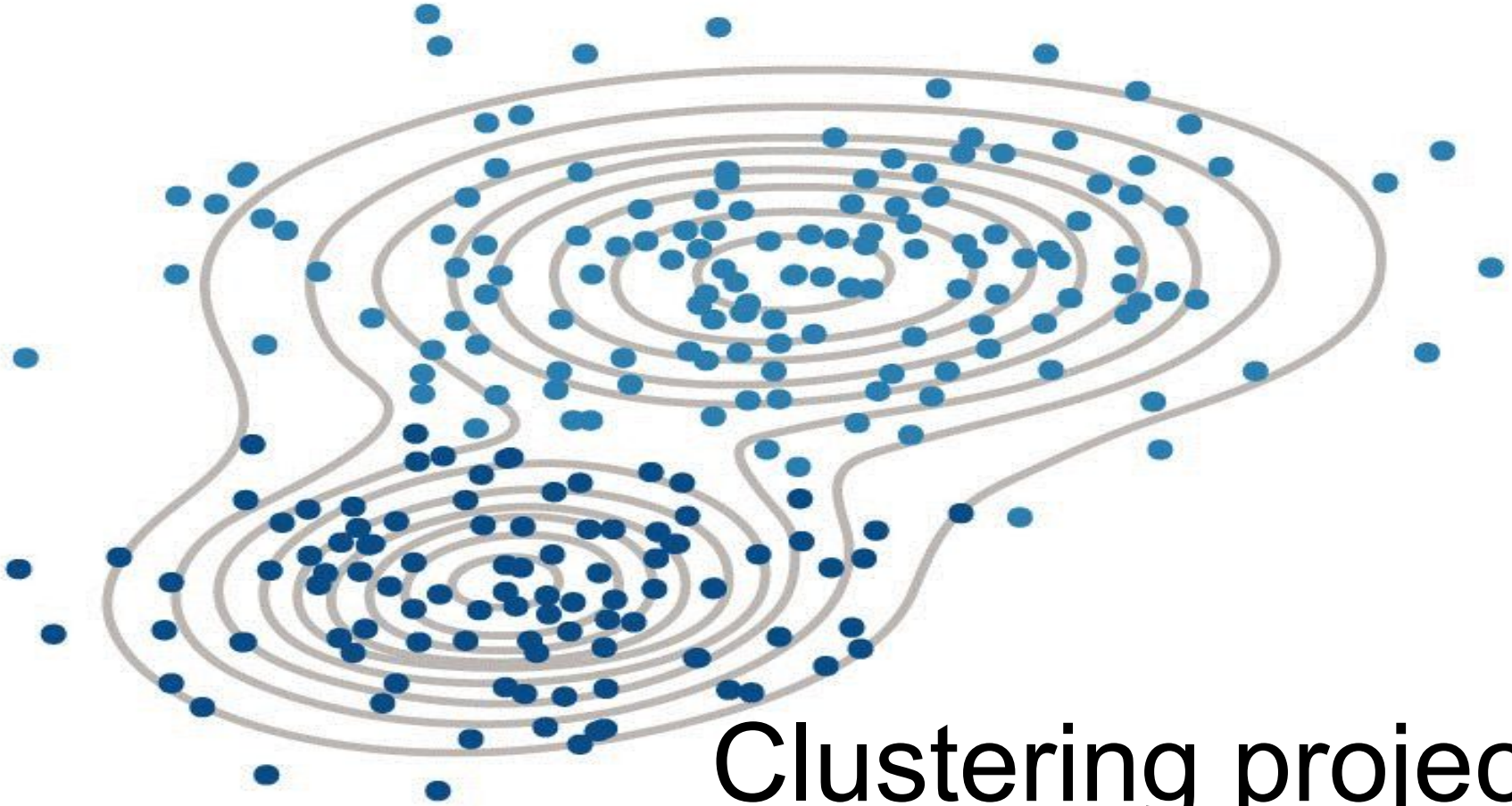


Clustering project

Table of contents

Features engineering and data preparation

Handling Null and missing values

A - Null values were clients ending the period with a balance of zero

B - Missing values were clients not having a specific account

Clients not having a specific account was encoded : 0

Clients ending the period with a balance of zero was encoded: 1

Wrangling

The dataset contains information on financial transactions completed during the year of 1995.

The data is joined on customers for clustering purposes.

Features selection

Relevant features selected to describe the customers are:

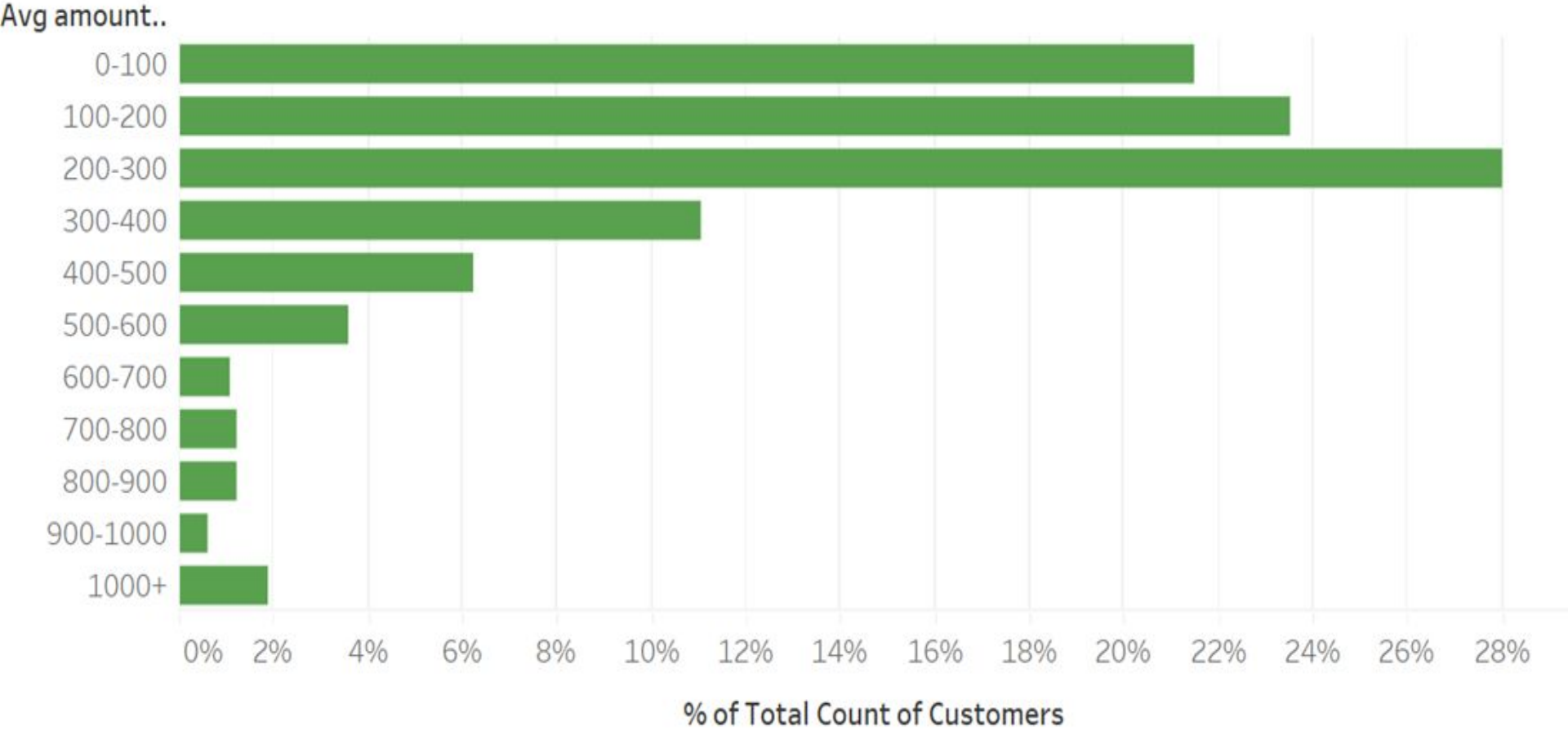
Demographic : age, income, customer loyalty and gender

Banking behaviors: average amount per transaction, checking account balance, savings account balance and credit account balance.

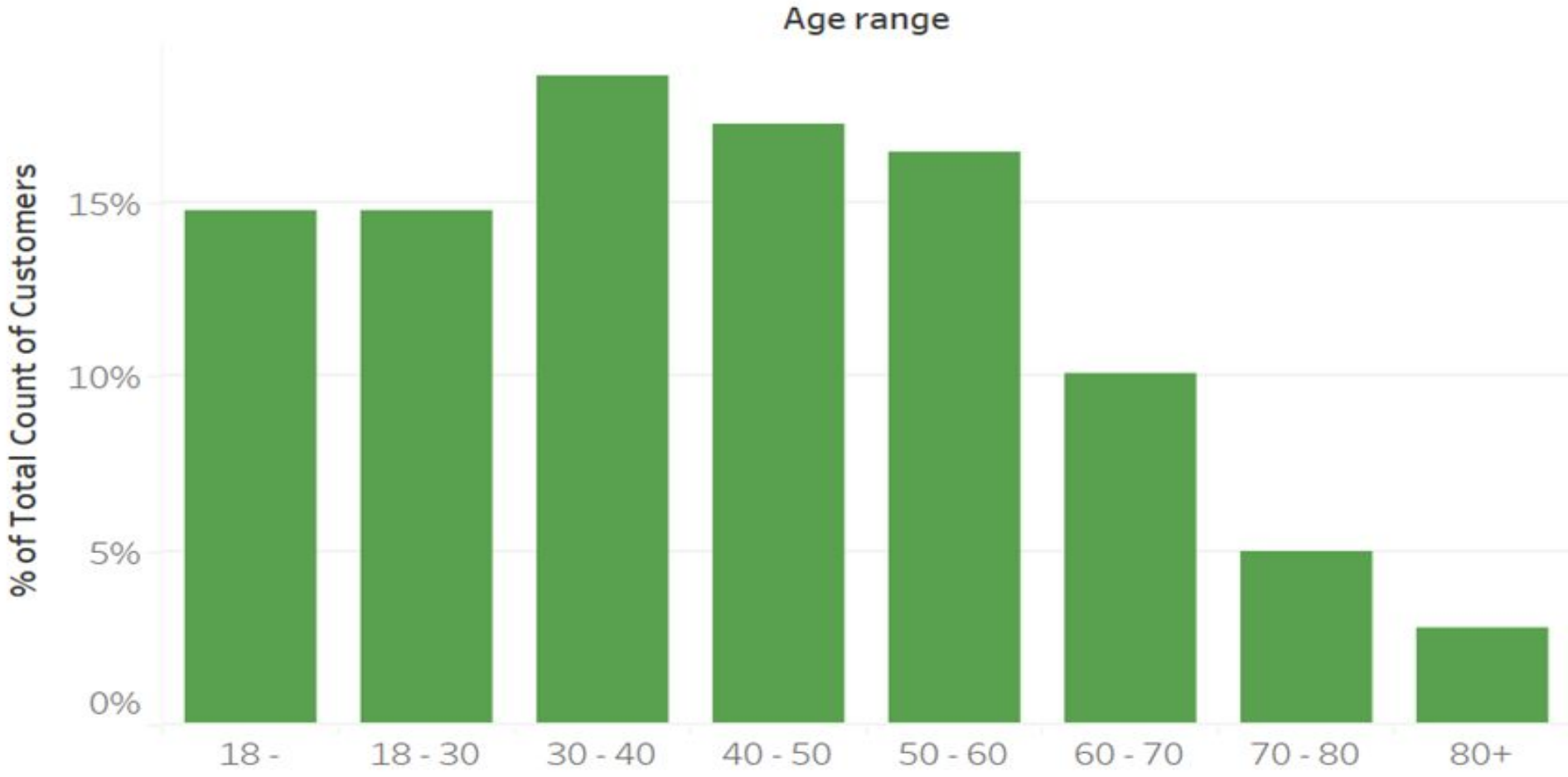
Visualization and insights

Average amount of transactions by cities

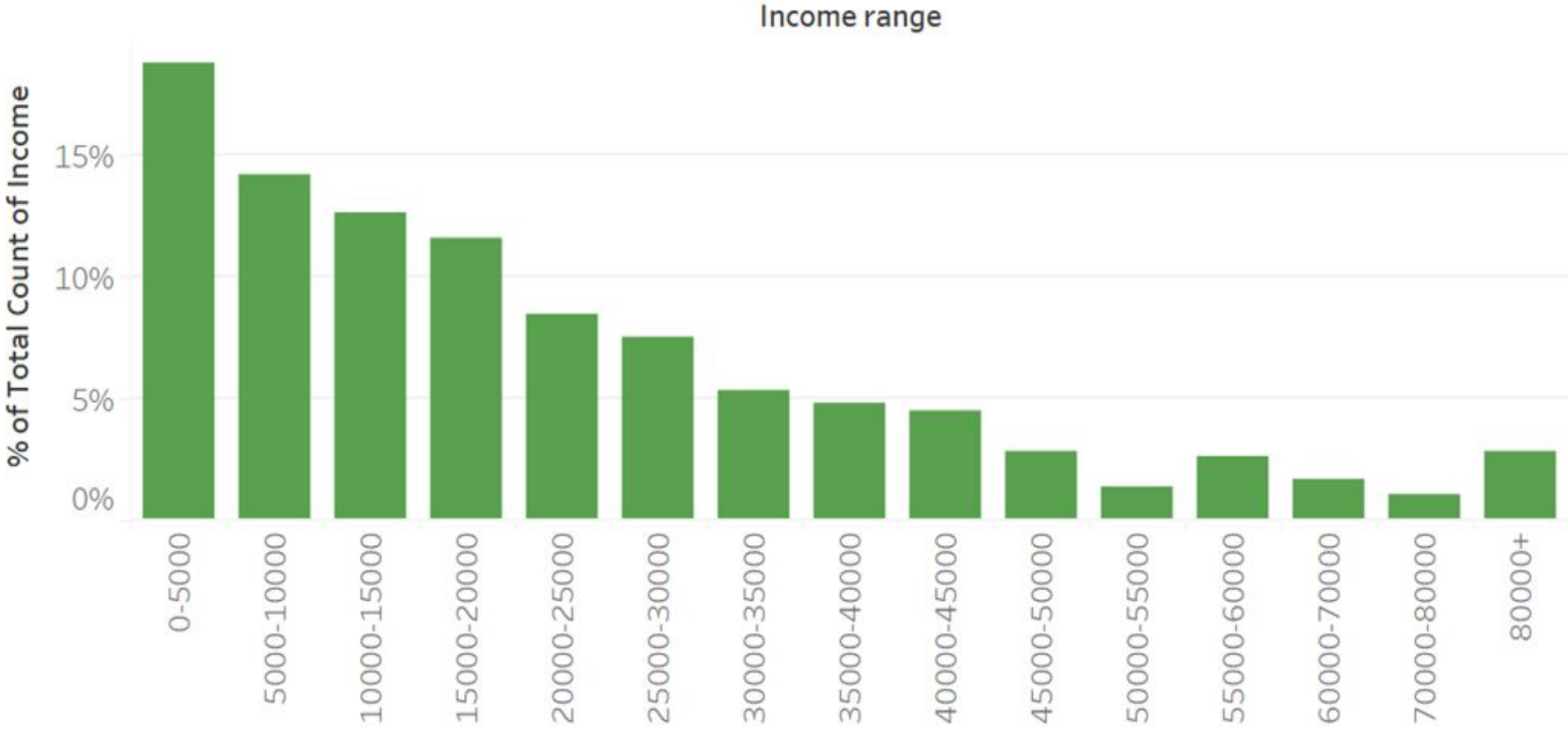
Percantage of customers by level of transaction



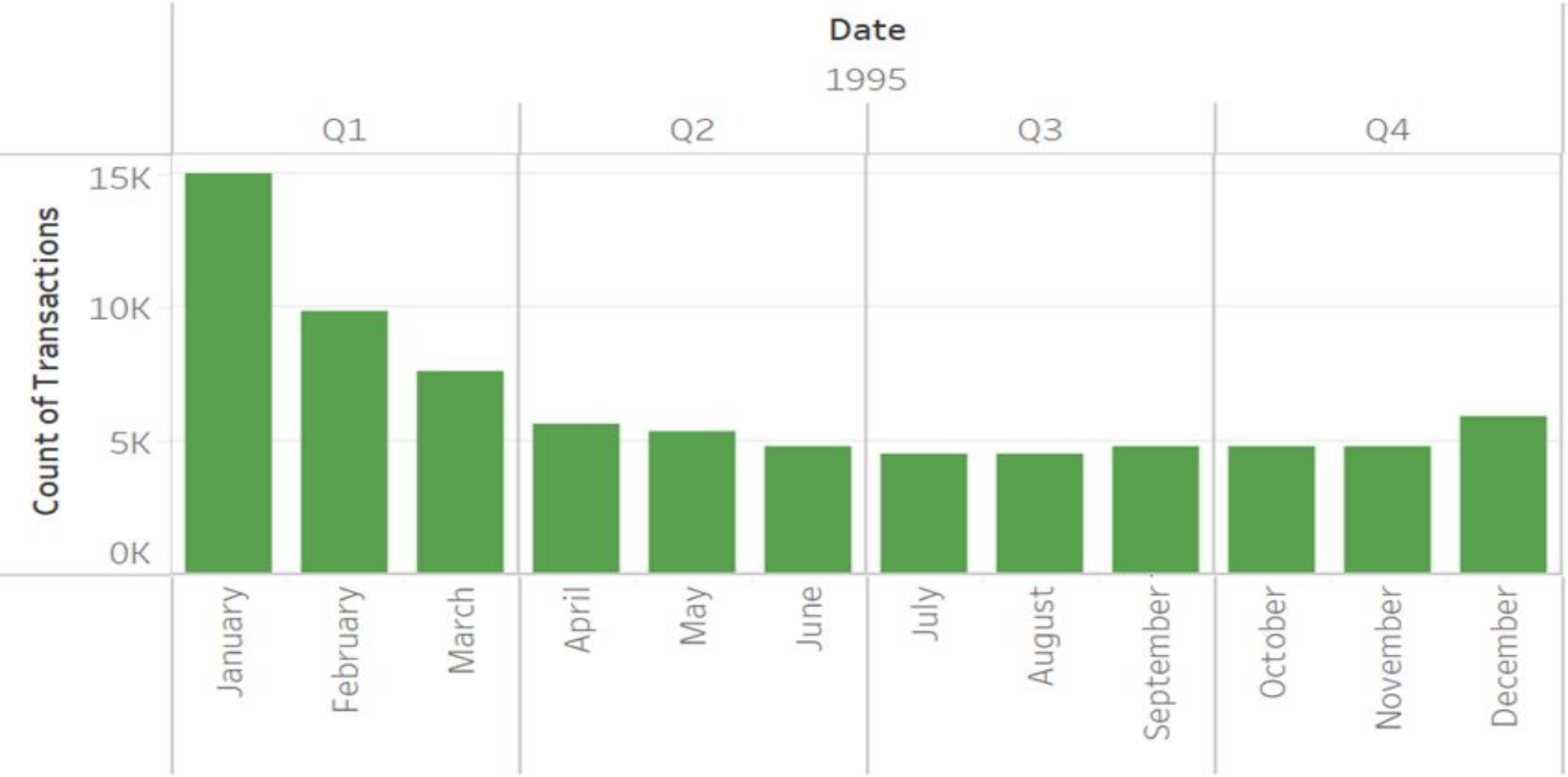
Percentage of customers by age



Percentage of customers by income range



Volume of transactions over the year



DEMOGRAPHICS

PROCESS

-Wrangled data was loaded

- Demographic information was identified and extracted

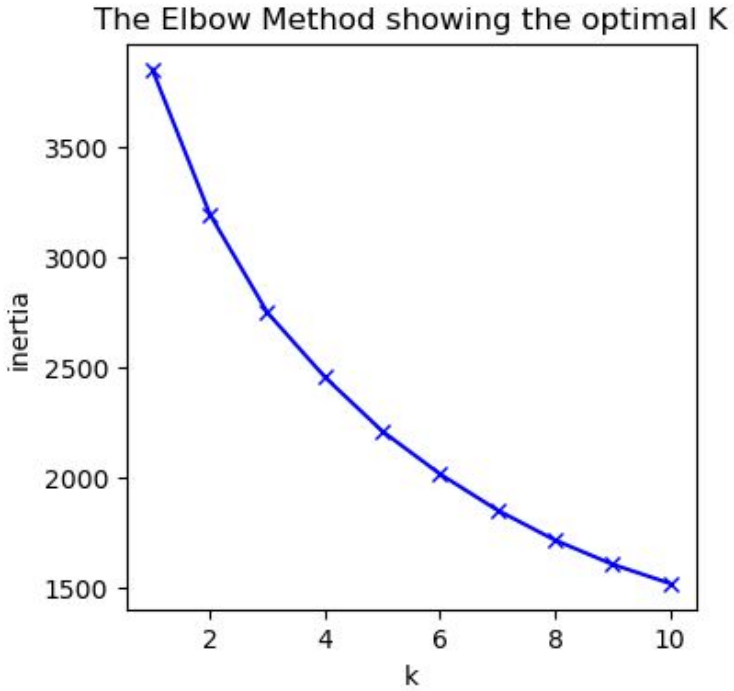
- Categorical data was transformed to numerical data

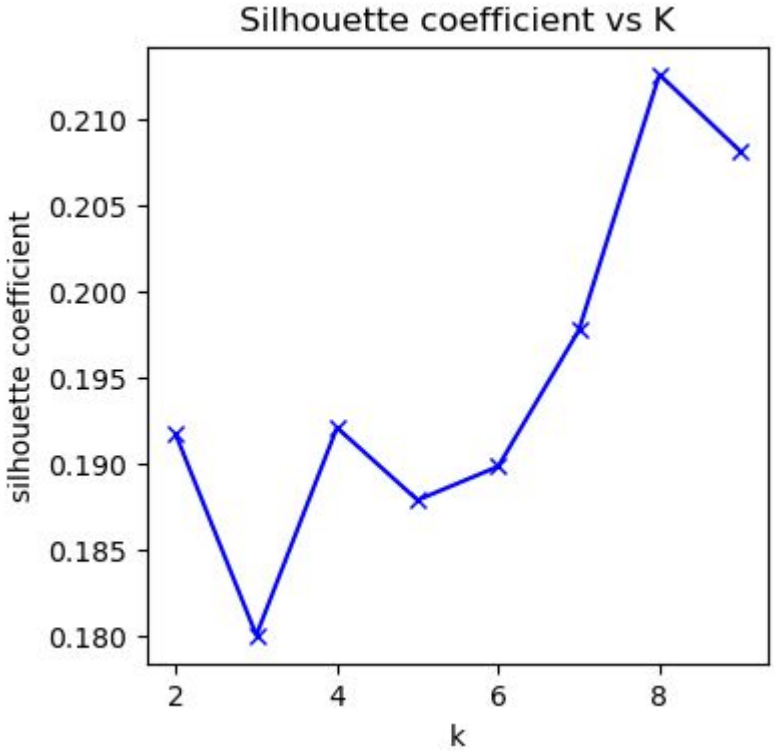
- Scaling was done to unify scales of feature values

- Reduced table features using PCA

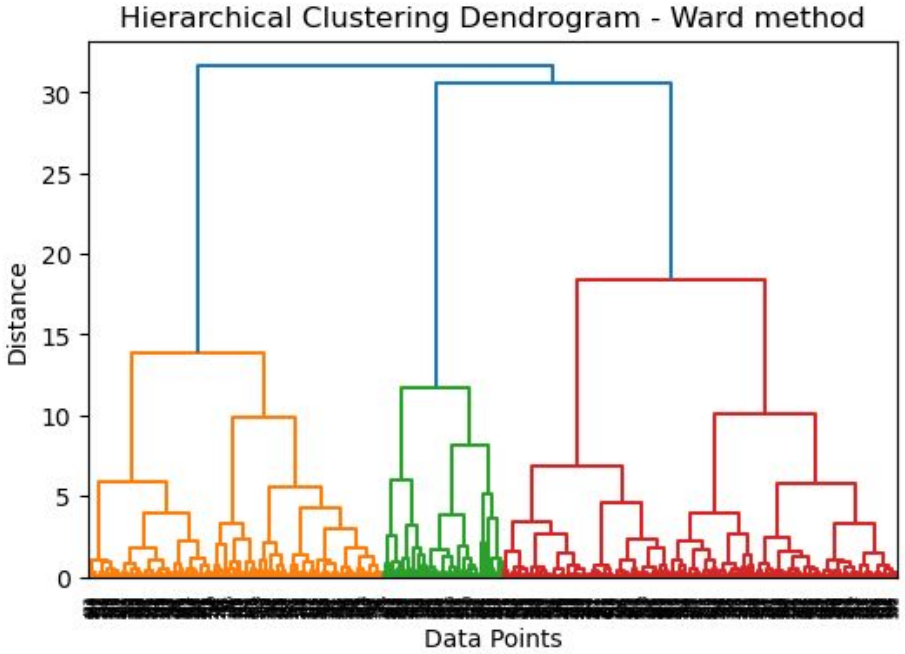
- Clusters plotted

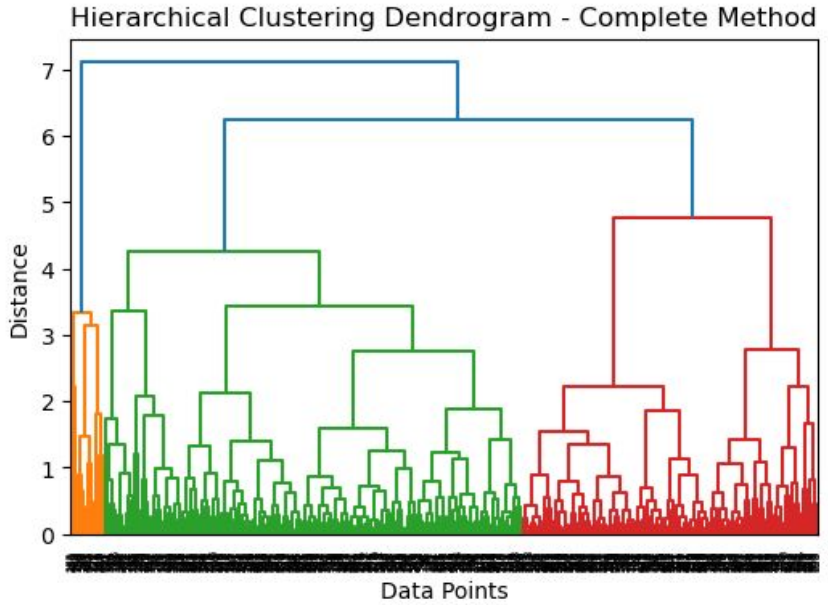
Demography - KMeans



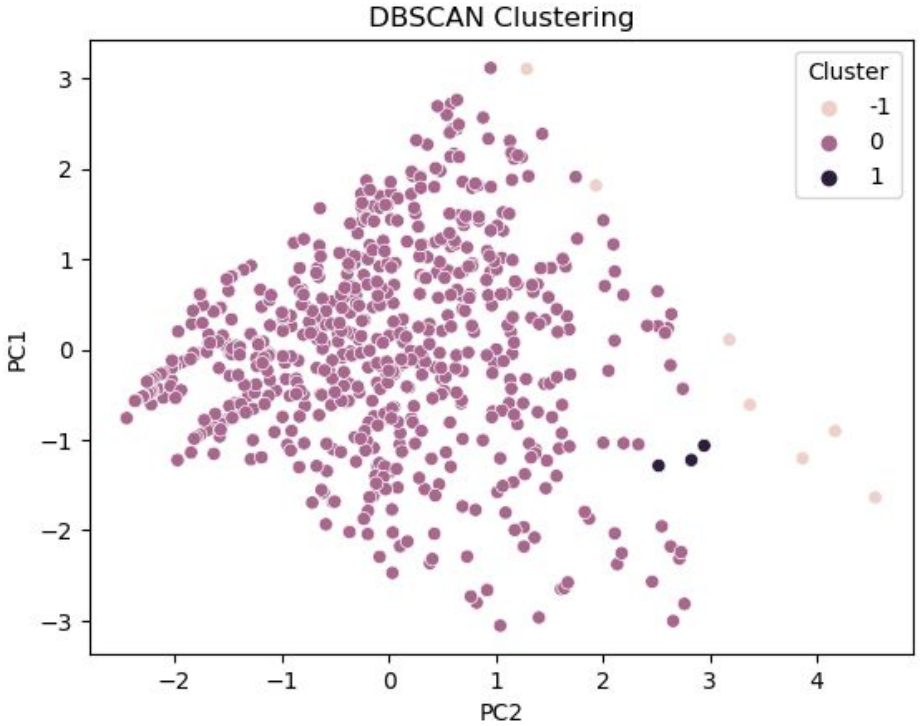


Demography - Hierarchical Clustering



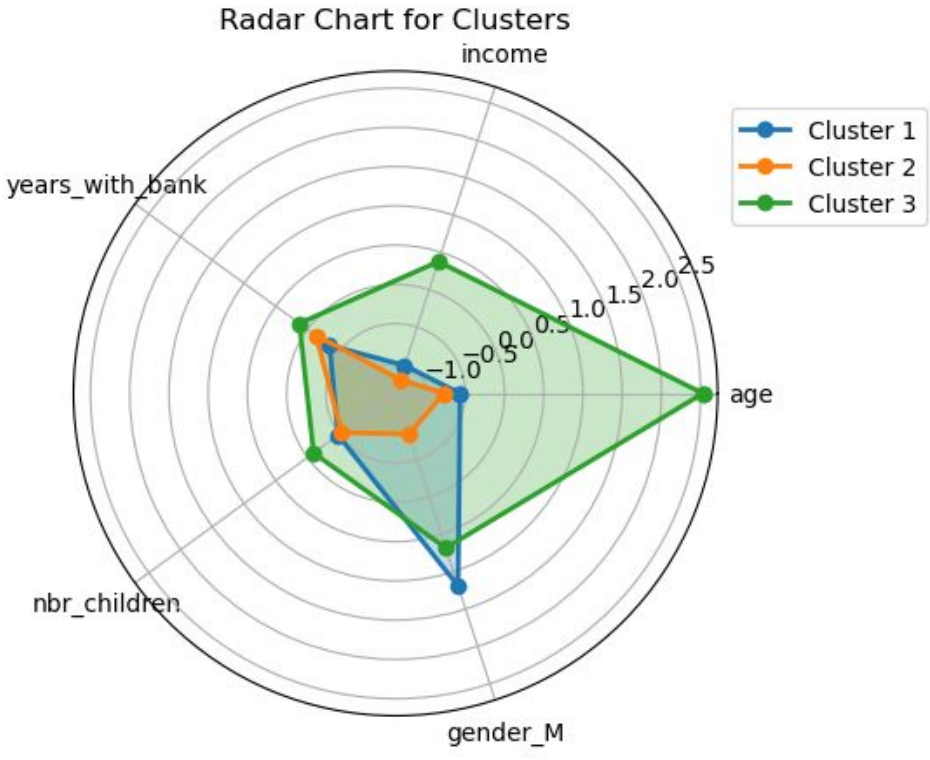


Demography - DBSCAN

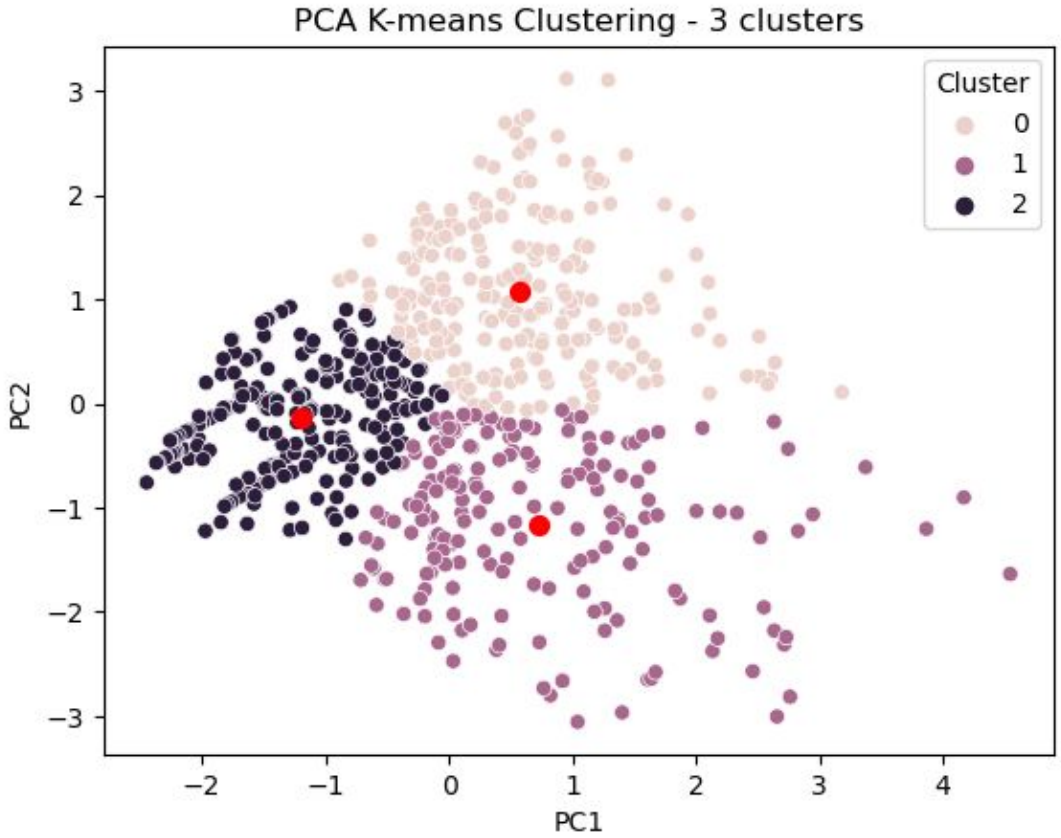


Demography - Radar Chart

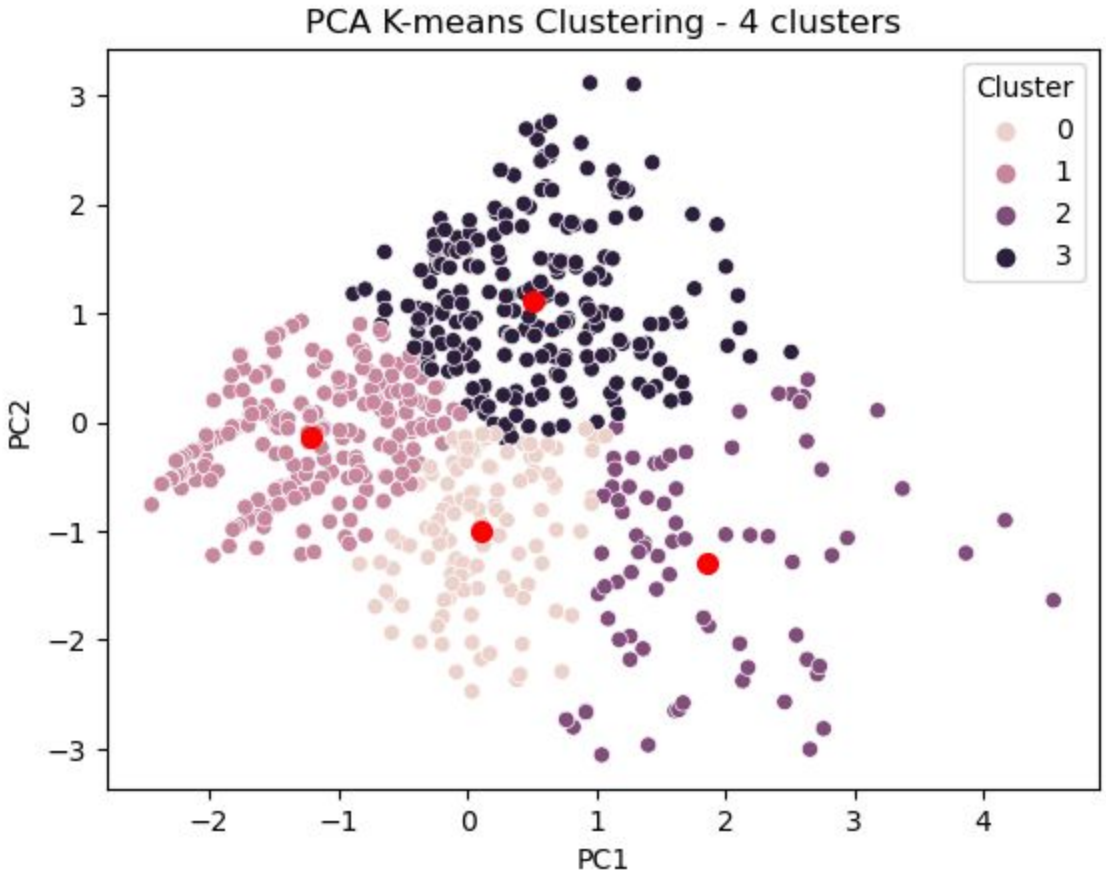
	income	age	years_with_bank	nbr_children	gender_M
0	38854.64	40.82	3.11	2.05	0.45
1	9373.45	23.55	3.03	0.14	0.41
2	21835.49	60.51	5.04	0.15	0.39



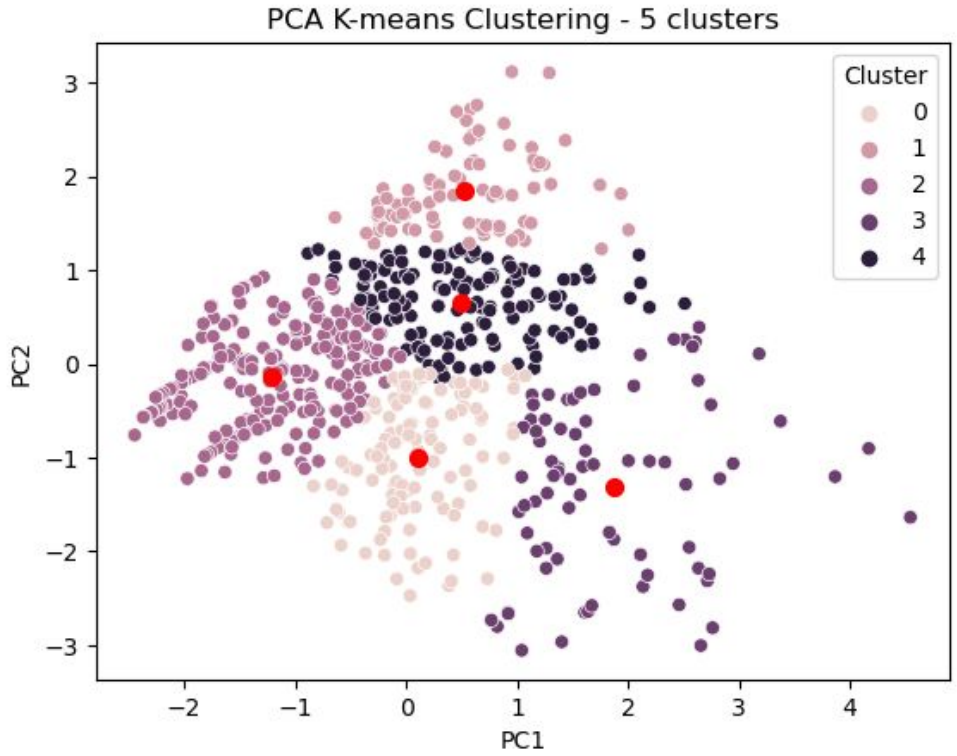
Demography - K-Means Clustering



Demography - KMeans Clustering



Demography - KMeans Clustering

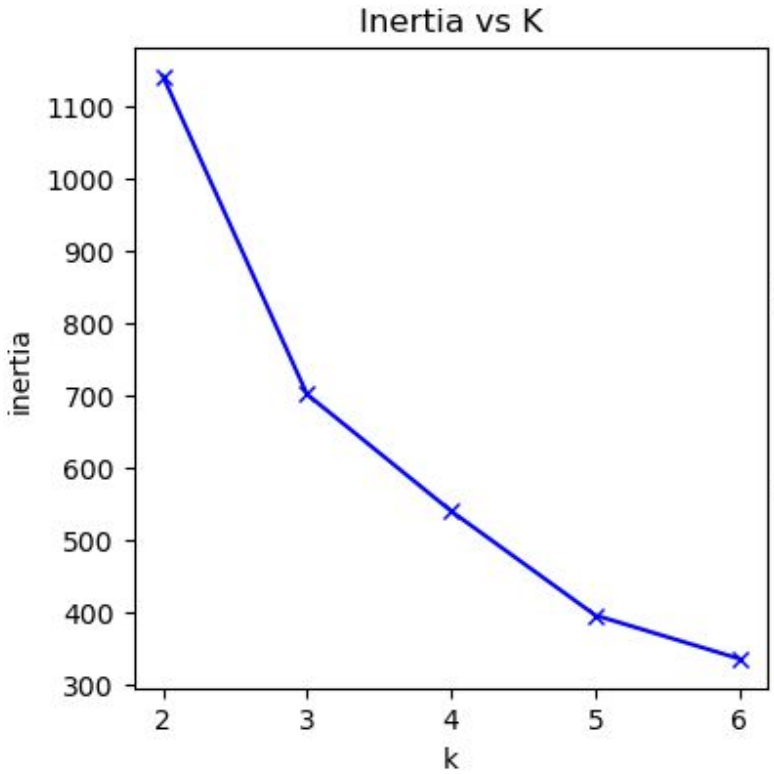


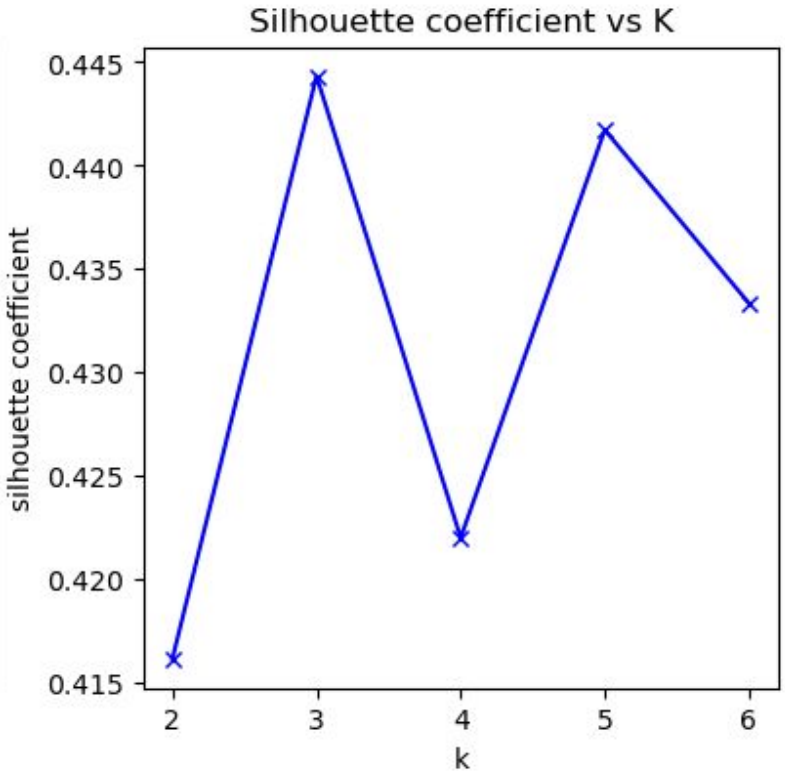
Banking Behavior Clustering

Banking Behavior - Data Preparation

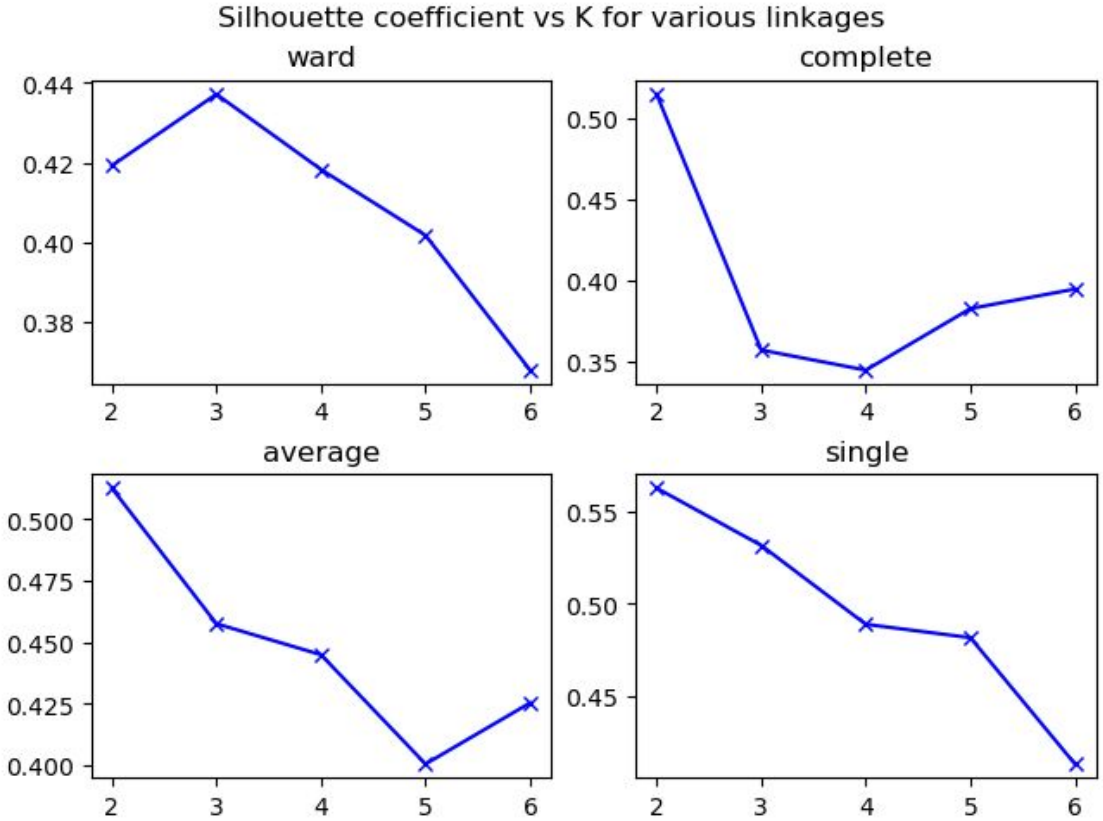
- Filtered out demographic data
- One Plus Log transformed financial data
- Standard Scaling
- PCA with 2 principal components

Banking Behavior - KMeans

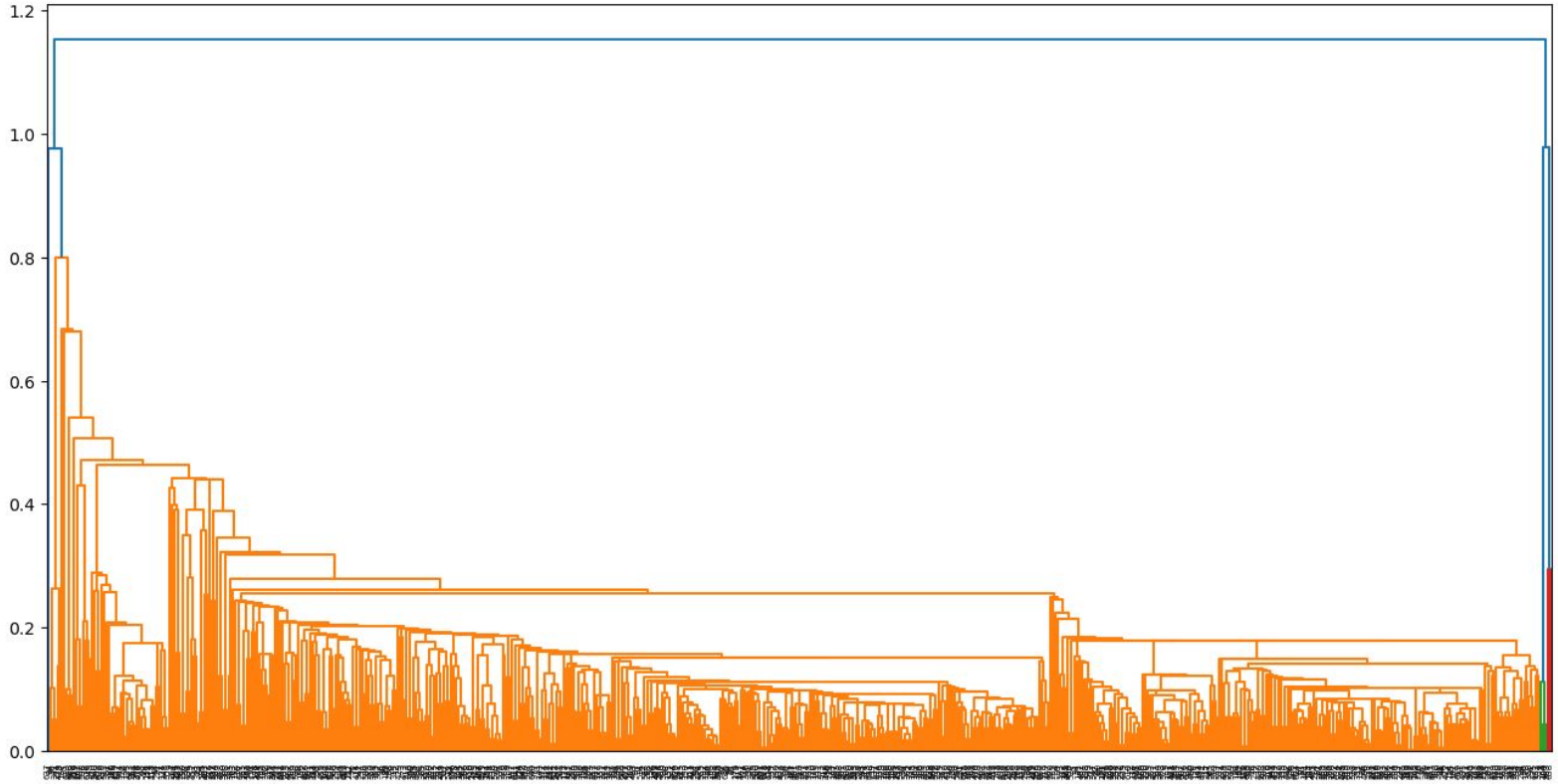




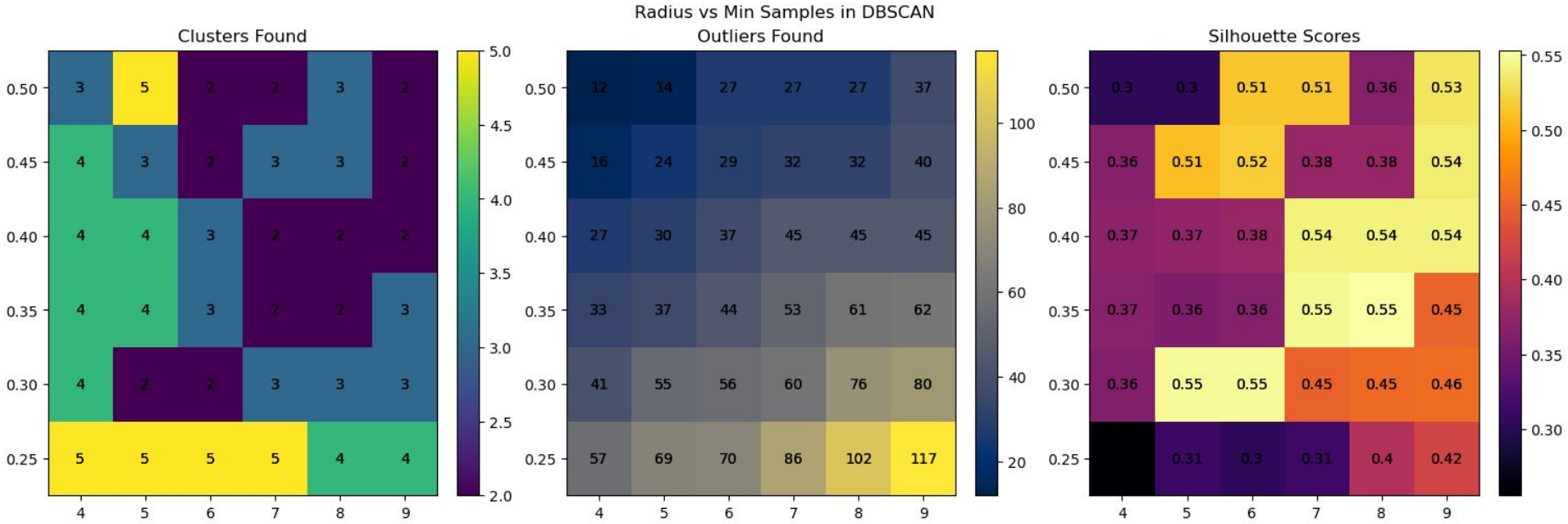
Banking Behavior - Hierarchical Clustering



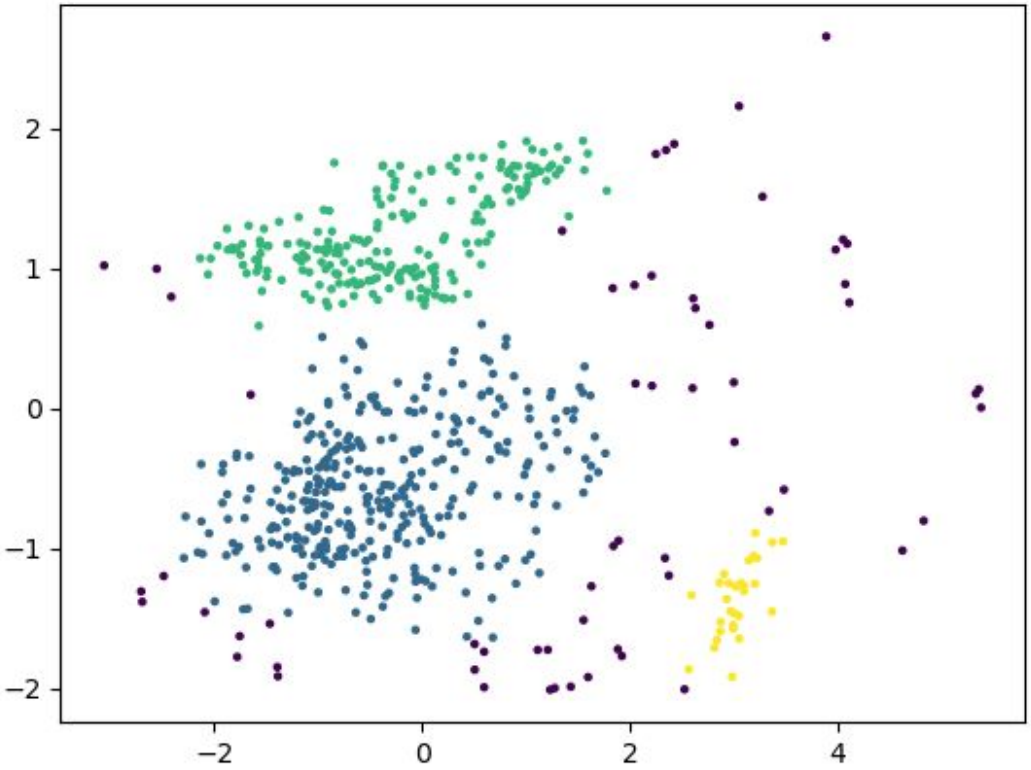
Banking Behavior - Hierarchical Clustering



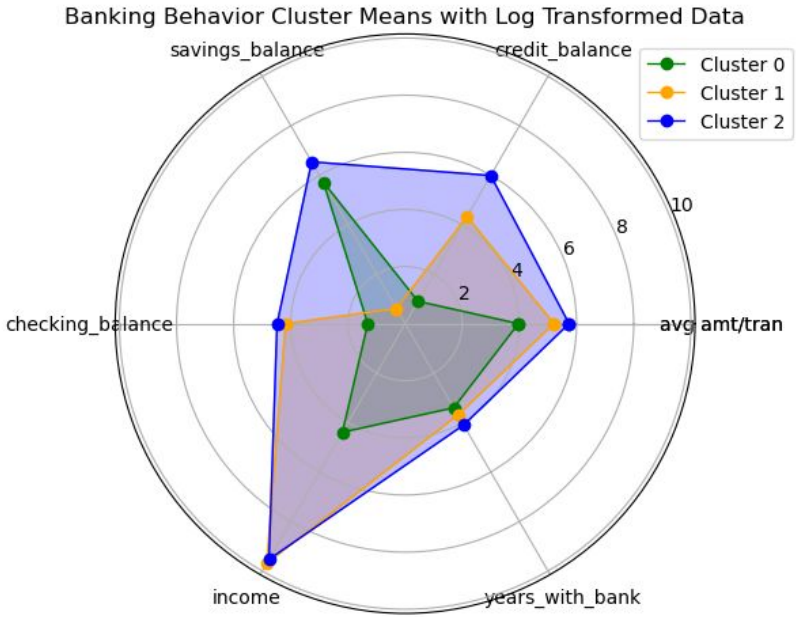
Banking Behavior - DBSCAN



Banking Behavior - DBSCAN

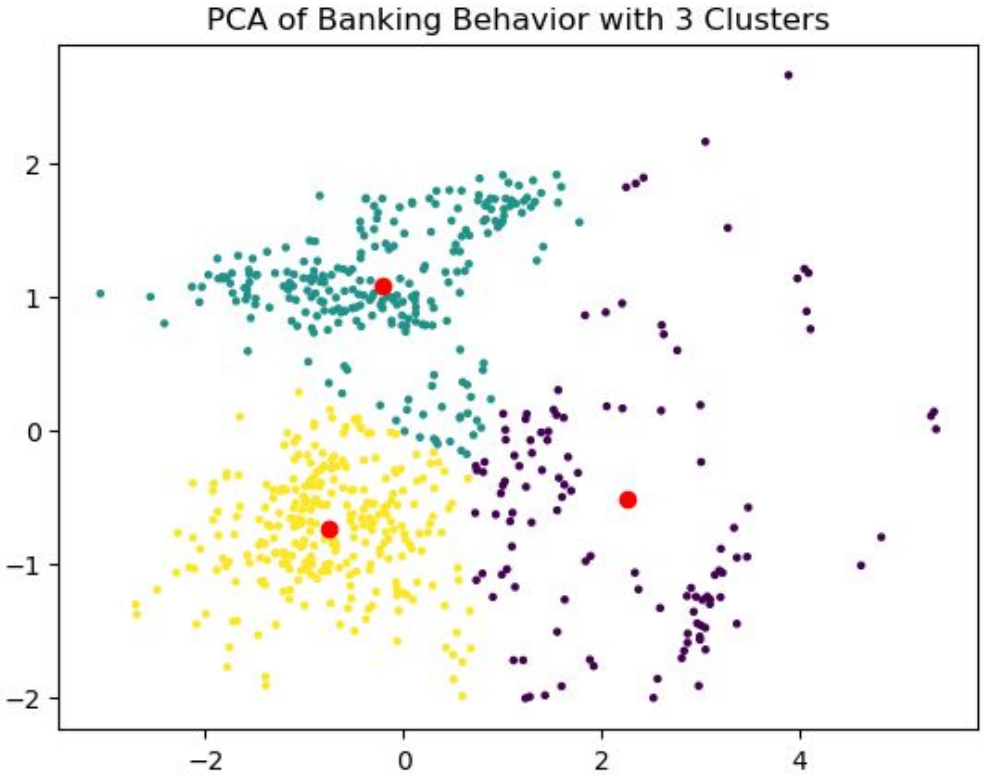


Banking Behavior - Radar Plot



	avg(amount/tran)	credit_balance	savings_balance	checking_balance	income	years_with_bank
0	88.76	77.90	825.65	89.83	8027.12	3.34
1	238.26	950.05	17.91	528.68	25514.63	3.77
2	360.33	1091.48	1471.10	614.25	25884.87	4.04

Banking Behavior - PCA Segmentation



Challenges

- Git workflow between 3 people
- Standardized procedures (same dataset, similar pipelines)
- Decision making when no clear optimum was found

Future Goals

- Try PCA with more components
- Animate Hierarchical Clustering and DBSCAN
- Compare demographic clusters with behavior clusters
- Manage outliers found with Hierarchical Clustering

Thank You