Credit Card Customer Segmentation

Team 16

Meet Our Team

Yuzhe Zheng

Leah Fowlkes

Yuchen Feng

(Bruce) Chang-Hung Hou

Problem Statement

With this dataset, we will identify multiple customer segments banks can use to personalize credit card experiences.

Stakeholders in this problem?

Bank

Identify features of customers in different clusters

Create targeted marketing strategy for each cluster

Avoid the one size fits all strategy

Remain competitive in the banking sector and increase profitability

Customers

Know which segment they belong to

Improve credit card performance for better rewards

Dataset Introduction

Dataset Features

- 8950 Rows
- 18 Columns(all numeric)
  - 1 id column
  - 14 float variables
  - 3 integer variables
- No duplicate values
- Missing Values exist
  - Credit limit(drop)
  - Minimum Payments(median)

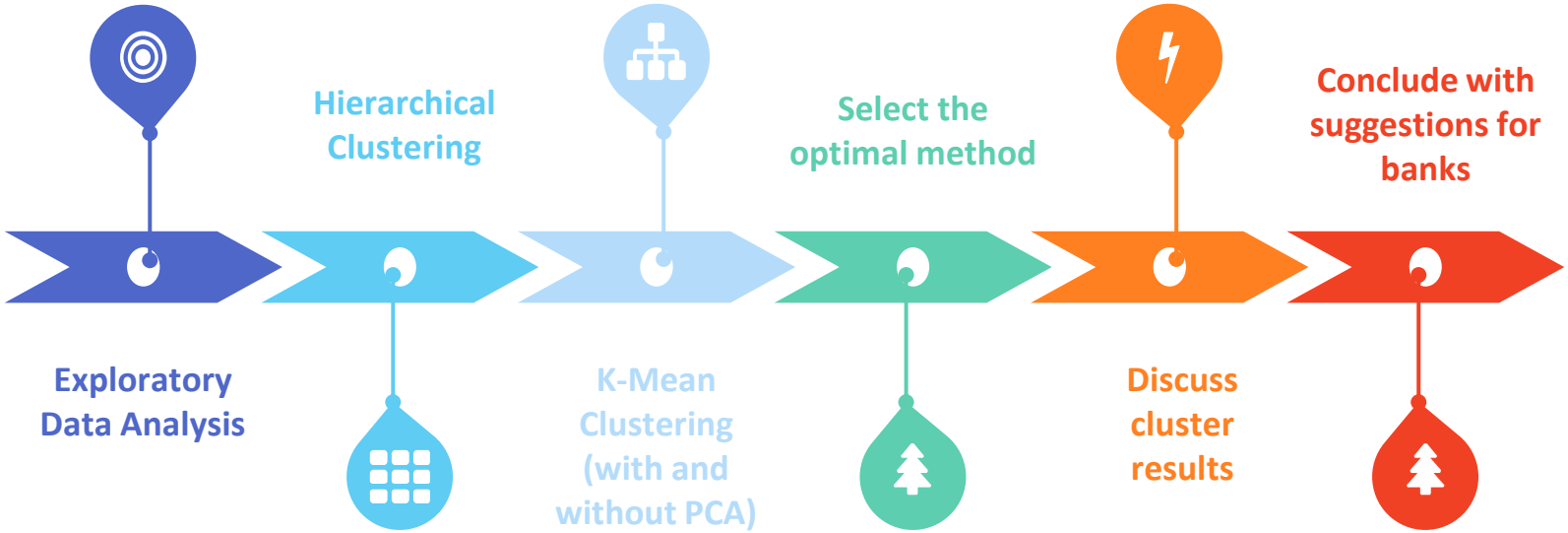
Data source

- kaggle

Column Variables

Id	Balance
Balance Frequency	Purchases
One Off Purchases	Installment Purchases
Cash Advance	Purchase Frequency
One Off Purchase Frequency	Installment Purchase Frequency
Cash Advance Frequency	Cash Advance Transactions
Purchase Transactions	Credit limit
Payments	Minimum Payments
Percent Full-payment	Tenure

Analysis Delivery

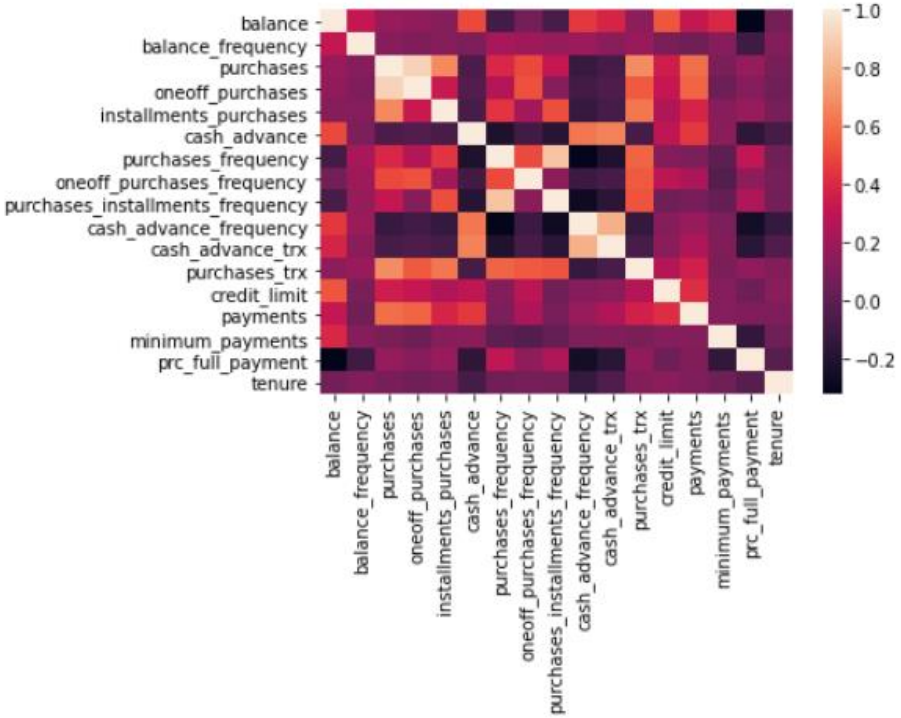


Exploratory Data Analysis

What we learned about our dataset

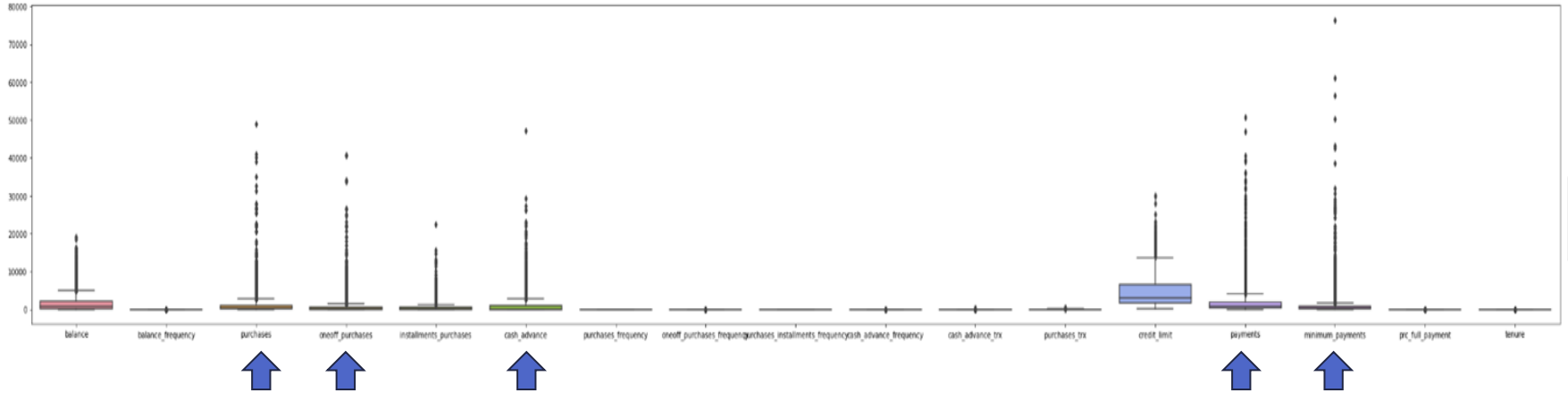
Mildly and Strongly Correlated

- Strong correlations between variables
  - Captured similar information
- Meaningful correlations like balance and credit limit
  - Mild positive correlation at 0.6
  - Seems the higher the balance, the more likely the bank will raise the credit limit



Exploratory Data Analysis

What we learned about our dataset

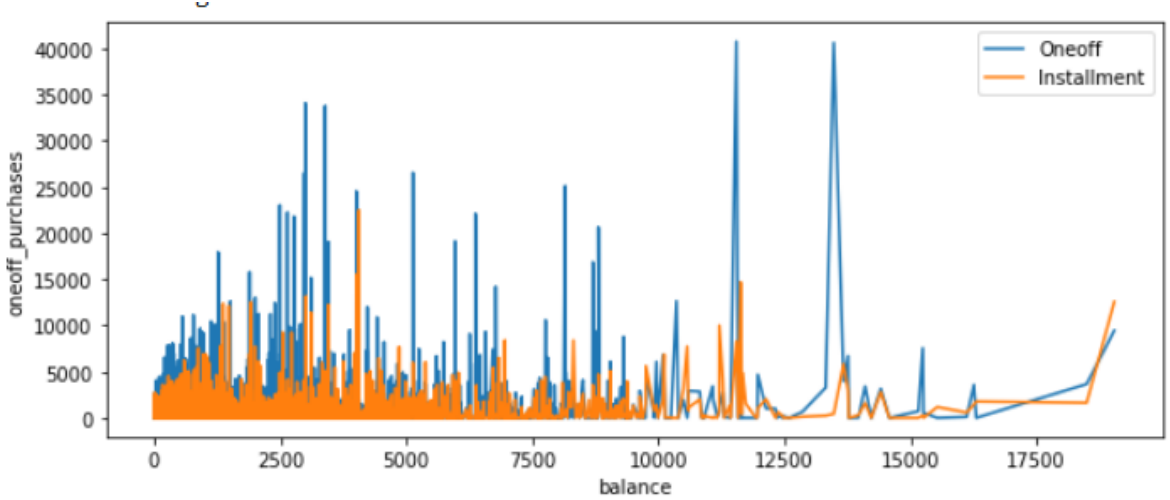


Several variables with extensive ranges of data

- Variables with wide range of data and many outliers
  - minimum_payments, payments, purchases, cash_advance, credit_purchases

Exploratory Data Analysis

What we learned about our Dataset

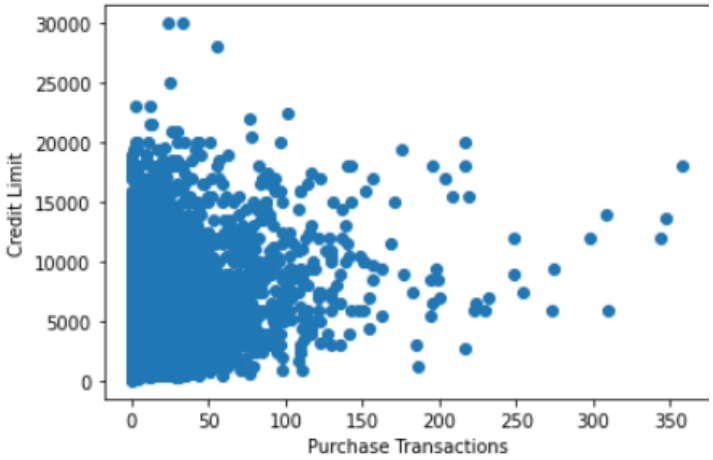


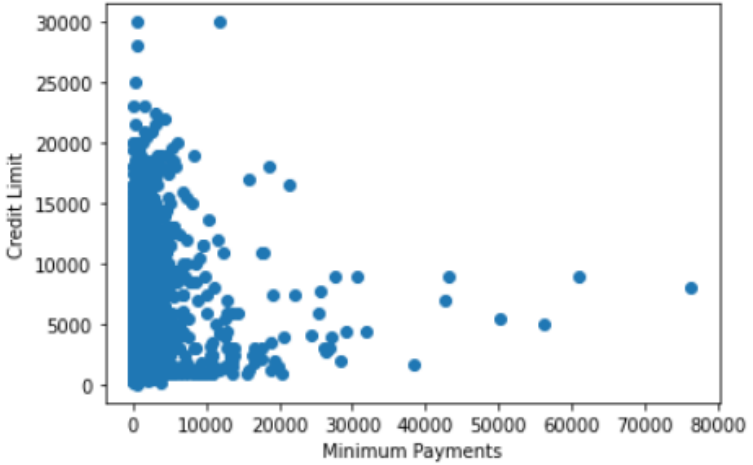
One off and Installment purchases follow similar patterns

- Follow similar patterns for card balances below $10,000
  - After $10,000, one off purchases increase randomly and strongly in credit payments
- Suggest these variables follow similar trends although their correlation score is 0.3 - 0.4

Exploratory Data Analysis

What we learned about our dataset



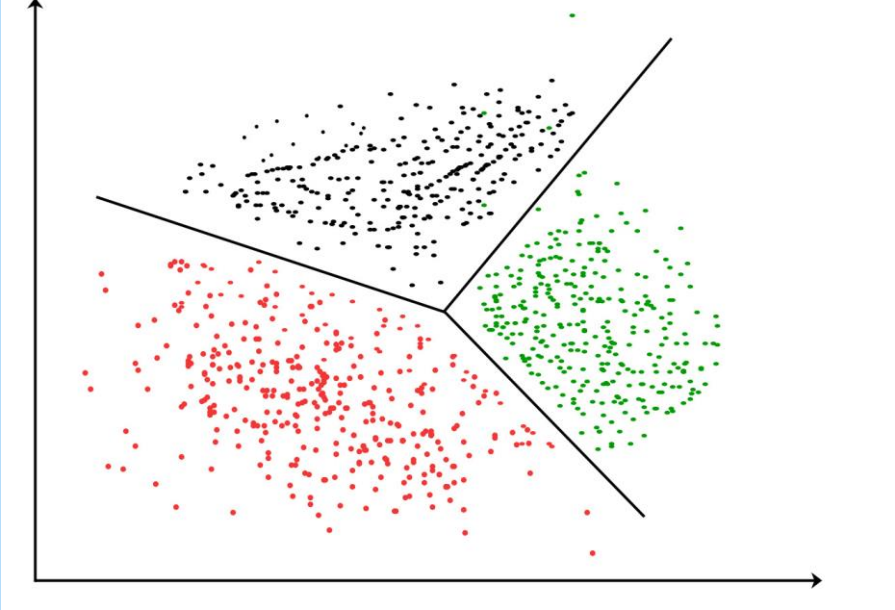


Low Purchase Transactions considered low risk

- If customers use their credit card too frequently, the bank likely considered them to be higher risk
  - Highest credit limits exist on cards where transaction count low

High Minimum Payments considered higher risk

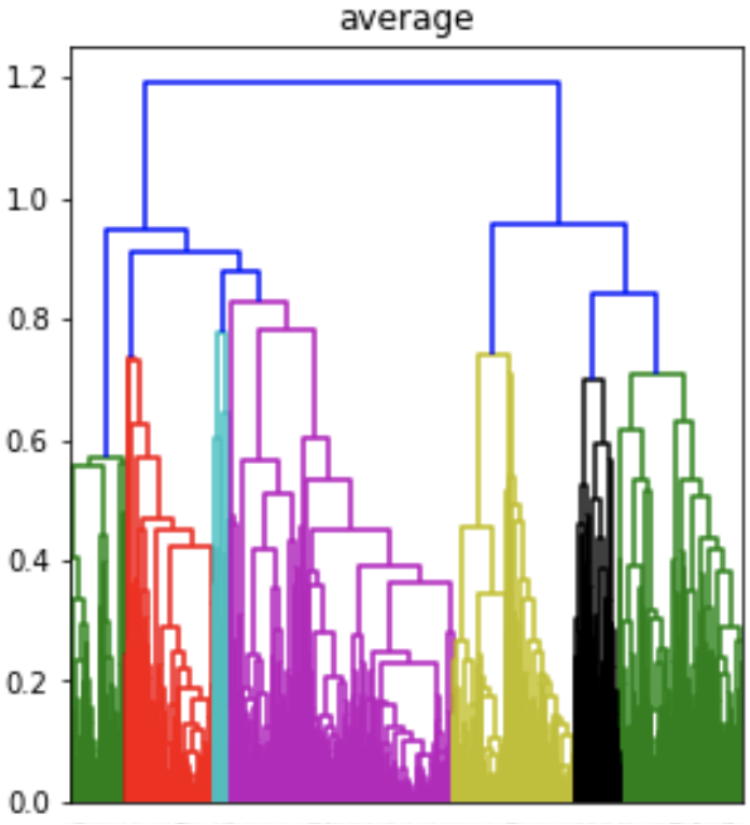
- Banks distinguish people with multiple minimum payments as high risk
  - Low credit limits when minimum payments are very high



Modeling

Unsupervised Learning

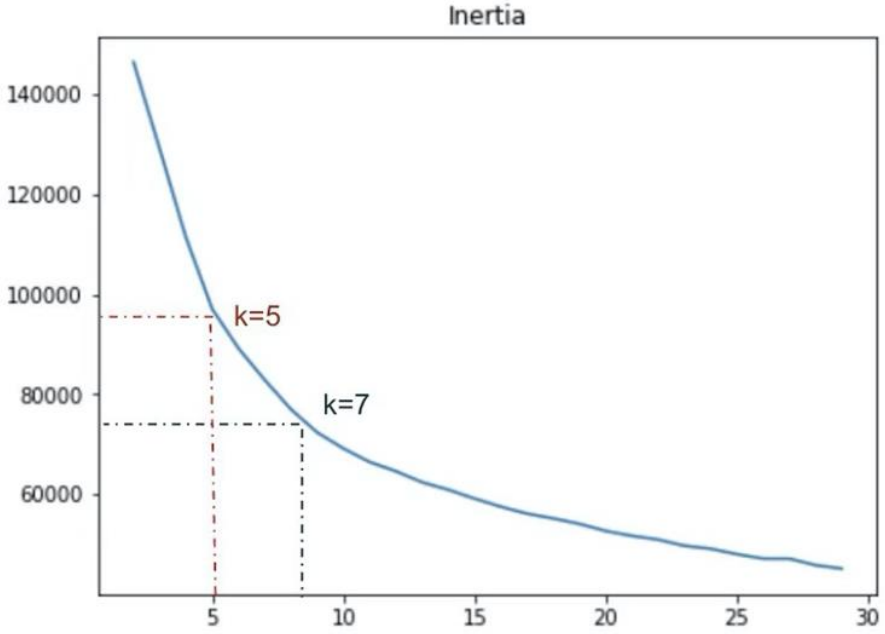
Hierarchical Clustering

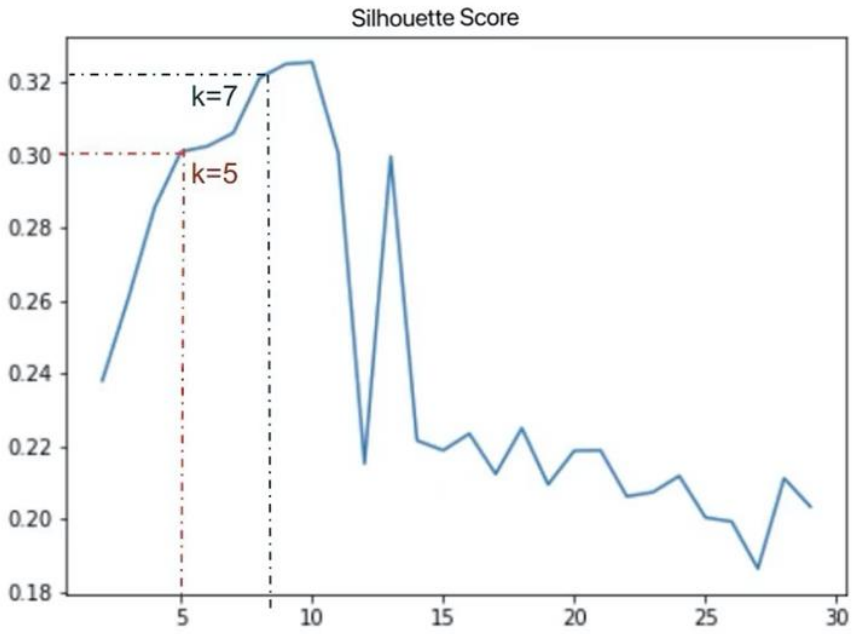


- Randomized the data for dendrogram
- Used dendrogram to visualize the relationship
- Fit number of clusters based on the dendrogram
- Choose the best model for hierarchical clustering based on the silhouette score

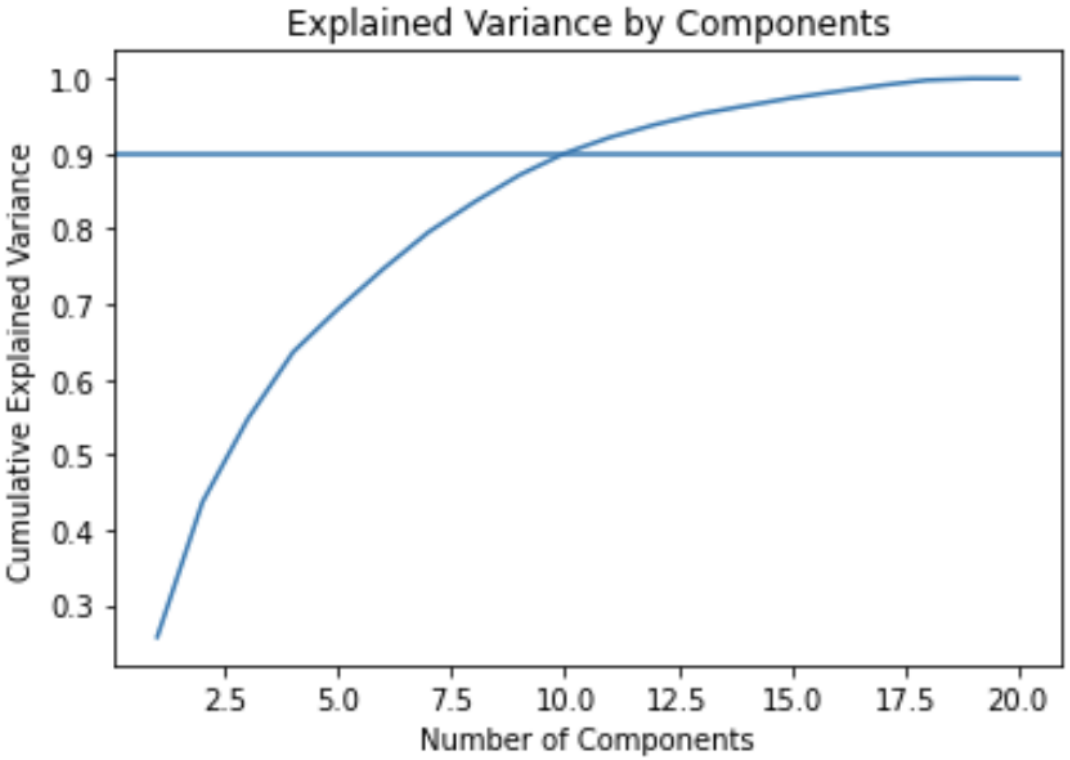
Methods	Silhouette Score
Average, 5 clusters	0.143
Average, 9 clusters	0.106

K-Means


Inertia
140000
120000
100000
80000
60000
k=5
k=7
5
10
15
20
25
30


Silhouette Score
0.32
0.30
0.28
0.26
0.24
0.22
0.20
0.18
k=7
k=5
5
10
15
20
25
30

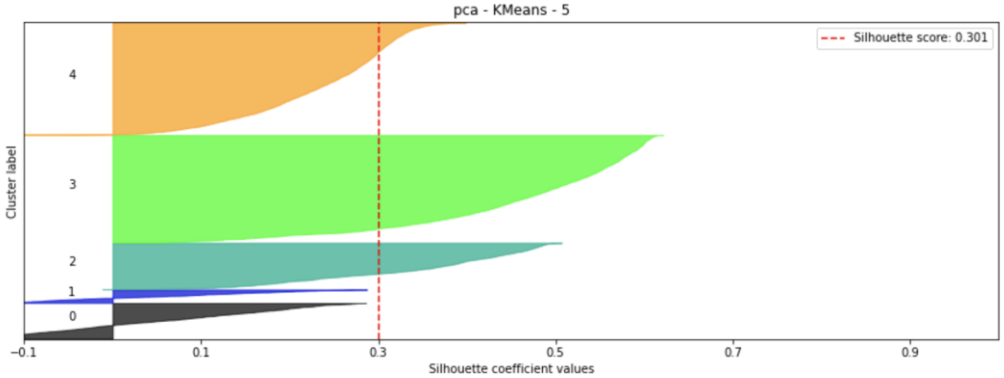
PCA

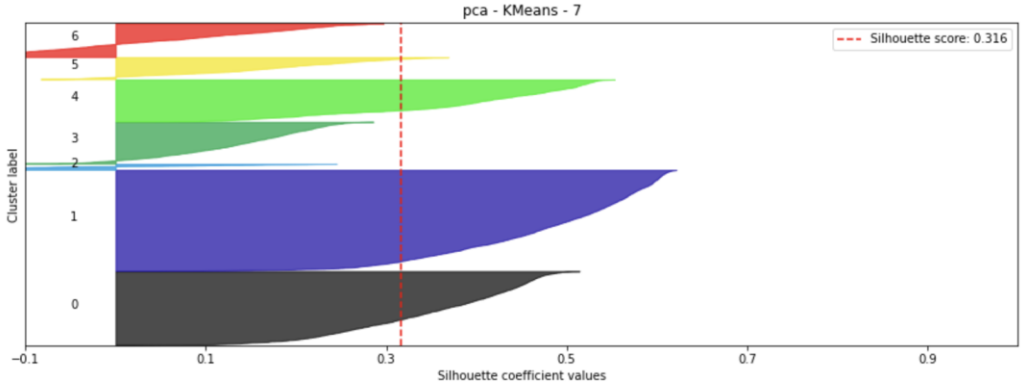


Reasons for using PCA

- Reduce high dimension
- Minimize information loss
- 90% explained variance by components
  - components = 9

K-Means





Comparing K-Means 5 & 7

- Silhouette score and number of clusters trade off
- Negative silhouette score
- Numbers observations in each cluster
- Comparison of average silhouette score

Model Overview

Final Decision

Cluster Size	Silhouette score
Hierarchical Clustering ***Without PCA***	
Average, 5 clusters	0.141
Average, 9 clusters	0.066
K-Means ***Without PCA***	
5	0.301
7	0.306
K-Means ***With PCA***	
5	0.301
7	0.316
9	0.316

Result and Discussion

Clustering Analysis

Clustering Results

What can we learn about our dataset?

Select Median for Analysis

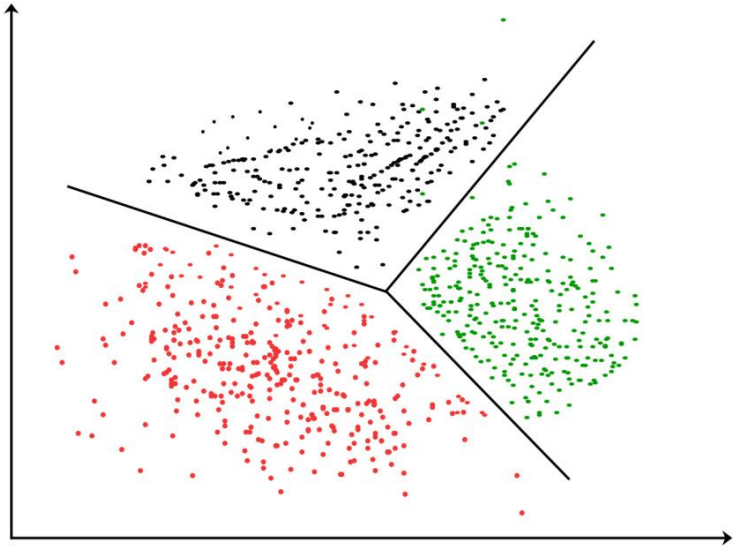
Value of mean is negatively affected by outliers. Median is not instead

Segment credit card users

Our segmentation is based on two criteria: purchasing power and customers preference of using credit cards

Conclude customized suggestions

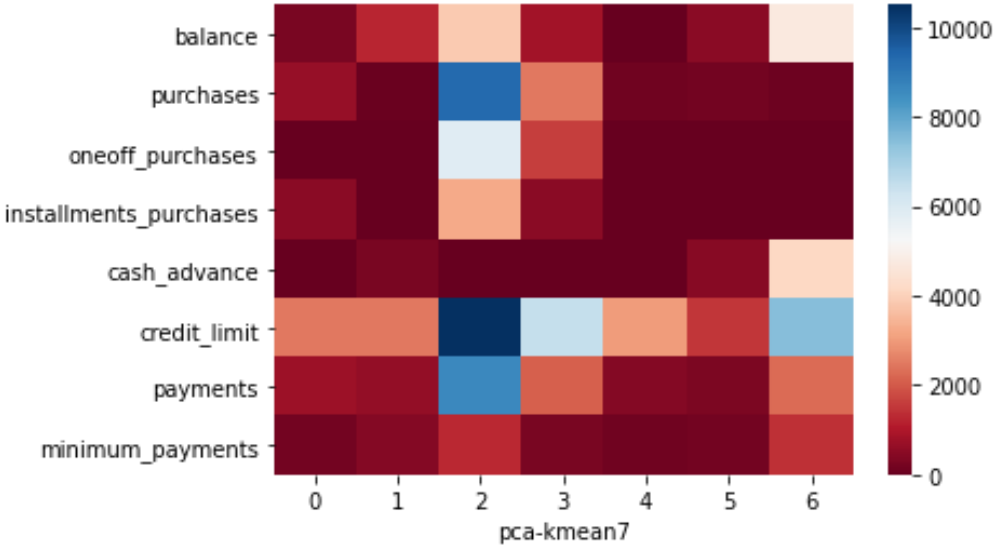
Our suggestions could help bank improve its business strategies for segmentation



Amount Analysis

Analyze the amount of each variables, all in US dollars

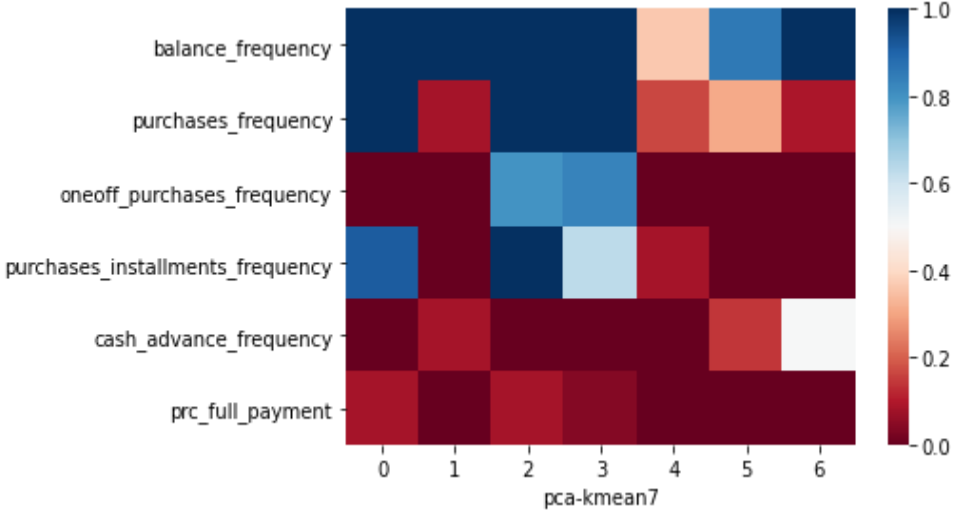
- Cluster 2 have the highest purchases and payments
- Cluster 3 have the second highest purchases and payments
- Cluster 1, 5, 6 have relatively high cash advance, ranking from the highest: 1, 5, and 6
- Cluster 0 and 4 have lower amount of balance



Frequency Analysis

Analyze the frequency of each transactions

- Cluster 2 and 3 tends to have highest purchase frequency
  - ➢ Cluster 2 makes more full payment than cluster 3
- Cluster 1, 5, 6 have relatively higher cash advance frequency but lower purchase frequency
- Cluster 0 and 4 have lower amount of balance
  - ➢ Cluster 0 has higher percentage of full payment than cluster 4
- Cluster 4 has lower balance frequency



Segmentation Strategies

Let's take a look at how bank could react to each cluster

Cluster	Characteristics	Strategies
2	Highest purchasing power	• Encourage them to use more credit cards • Offer more cash backs and discounts for credit card payment
3	Second highest purchasing power	
6	Most frequent cash advance user	• Look for what motivates them to use cash advance • If they use cash advance due to cash shortage, consider reduce credit limit • If tight budget is not the case, advertise more cash advance service
5	Cash advance user, lower repayment capacity	
1	Cash advance user	
0	Lower purchasing power	• Remain usual marketing strategy • Notify cluster 4 to pay for credit cards once limit is nearly full
4	Lower purchasing power, more reluctant to update balance	

Challenges

Limitation Records for Dendrogram

We need to randomize the data for the dendrogram. After selecting the best method from the dendrogram, we used the original dataset for our clustering process.

Lacking Practical Variables

The dataset does not contain employment, balance in checking and saving account, income, age, and FICO score. These are beneficial for credit card customers analysis, such as repayment capacity.

THANK YOU

ANY QUESTIONS?