Credit Card Customer Clustering Optimized with Multiclass Machine Learning Modelling

Numpy Team Members:

- Fajar Setiawan (jarbling@gmail.com)
- Fransiskus Alvin (fransiskusalvin@yahoo.com)
- Panji Agdiwijaya (agdiwijaya@gmail.com)

Purwadhika JCDS 1204 – Final Project

Outline

01 Introduction & Workflow Process

02 Background, Business Problems, and Goals

03 Data Understanding & Cleaning

04 Clustering

05 Exploratory Data Analysis, Multiclass Machine Learning Modelling & Evaluation

06 Project Recommendation, Business Impact, & Future Research

07 Model Deployment (Flask Dashboard)

1. Introduction & Workflow Process

Introduction

Hello! We Are ***NumPy*** Team

Fajar Setiawan

Management Int Business Graduate

Project Administrator – Technology Industry

Fransiskus Alvin

Mechanical Engineering Graduate

Service Marketing – Automotive Industry

Panji Agdiwijaya

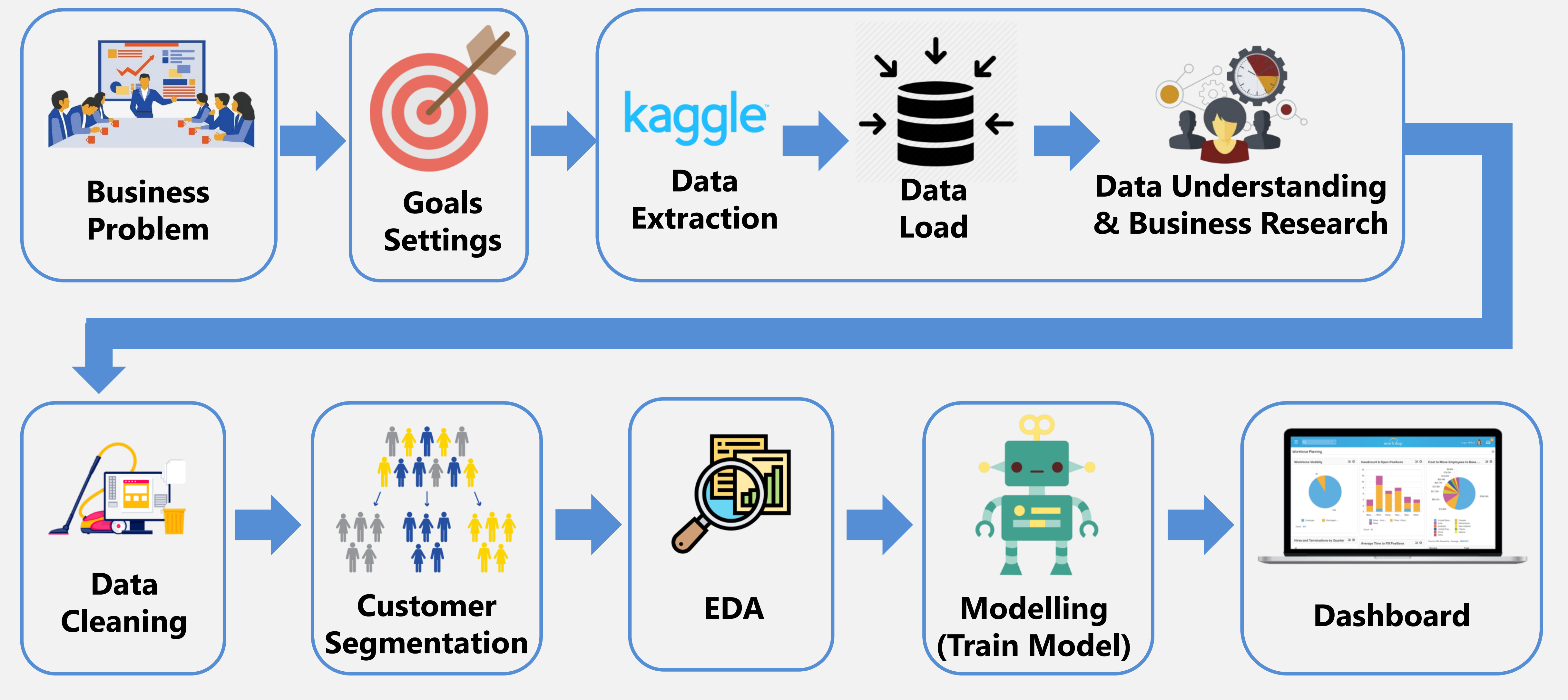
Public Administration Graduate

Consultant – Recruitment Industry

Introduction

In this project, we **position ourselves** as a part of the **Data Scientist** Team in one of the **Bank in USA**. We are assigned to **help the marketing team** to create a **segmentation of credit card customers** based on their **behavior**.

Workflow Process



2. Background, Business Problem, and Goals

Background

Almost every American, it seems, gets a new credit card offer in the mail almost every week. Credit cards are highly profitable, but only if the customers stays around for a while. It **costs about 80 dollars** to **acquire a new credit card customer** who **returns about 120 dollars per year** in profit, but **only if the customers keeps the card**. If customers **drops the card after a few weeks, or doesn't use the card**, the issuer will **lose that 80 dollars, plus some more money spent trying to reactivate them**.

Source: https://www.dbmarketing.com/articles/Art175.htm

Business Problem

- **Customers loyalty** is one of the key to survive in this credit card business competition
- The **cost of acquiring new customers** is estimated at **five times** the rate of **retaining existing ones**
- In order to retain **Customers**, we must first understanding our **Customers Type and Customers Behaviour**
- Previously, our bank **only has 1 product of credit card**, resulting **low customer loyalty** because **inaccurate marketing program**
- After do long research, our management decides to make 3 different products: **Business Unlimited (High), Business Cash (Medium), and Performance Business (Low).**
- In other hand, the company **doesn't know which customers belongs to which products effectively**

Source: https://www.dbmarketing.com/articles/Art175.htm

Goals

Our end goal is to retain customer

Understanding **Customers Type** and **Customer Behaviour** through **Customer Data Clustering**

Define product details based on **Clustering Results** to ensure that customers get the proper product

Help Marketing Team to define new Customers Type through **Multiclass Classification Machine Learning Technique**

3. Data Understanding & Data Cleaning

Data Collection

Credit Card data shape is **8950** rows and **18** features

- **CUST_ID** : Identification of Credit Card Holder
- **BALANCE** : Balance amount left in their account to make purchases
- **BALANCE_FREQUENCY** : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
- **PURCHASES** : Amount of purchases made from account
- **ONEOFF_PURCHASES** : Maximum purchase amount done in one-go
- **INSTALLMENTS_PURCHASES** : Amount of purchase done in installment
- **CASH_ADVANCE** : Amount of Cash Money user take from credit card
- **PURCHASES_FREQUENCY** : How frequent the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
- **ONEOFF_PURCHASES_FREQUENCY** : How frequent Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
- **PURCHASES_INSTALLMENTS_FREQUENCY** : How frequent purchases in installments are being done (1 = frequently done, 0 = not frequently done)
- **CASH_ADVANCE_FREQUENCY** : How frequent user take money from credit card
- **CASH_ADVANCE_TRX** : Number of Transactions made with "Cash in Advanced"
- **PURCHASES_TRX** : Number of purchase transactions made
- **CREDIT_LIMIT** : Limit of Credit Card for user
- **PAYMENTS** : Amount of Payment done by user
- **MINIMUM_PAYMENTS** : Minimum amount of payments made by user
- **PRC_FULL_PAYMENT** : Percent of full payment paid by user
- **TENURE** : Tenure of credit card service for user

Our dataset is taken from kaggle https://www.kaggle.com/arjunbhasin2013/ccdata

Data Cleaning (Pre-Processing)

01

Drop Unnecessary Columns

Since **CUST_ID** has object and has no relation for analysis, we will drop CUST_ID

02

Fill Missing Values

- **MINIMUM_PAYMENTS** is filled with 0 assuming the customers haven't made any PAYMENTS (PAYMENTS = 0)
- **MINIMUM_PAYMENTS** is filled with same value of PAYMENTS because the customers have PAYMENTS data recorded

03

Drop Missing Value

Missing Value on **CREDIT_LIMIT** is dropped because there is only 1 CREDIT_LIMIT data that has null value

4. Clustering

Data Clustering

Based on problems and added by research results, we utilize 3 features that might be the factors for customer segmentation:

BALANCE 

PURCHASES 

CREDIT_LIMIT 

Source:
- https://creditcards.chase.com/
- https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/Financial%20Services/Latest%20thinking/Payments/MoP19_New%20frontiers%20in%20credit%20card%20segmentation.ashx

For clustering we use 3 algorithms:

KMeans

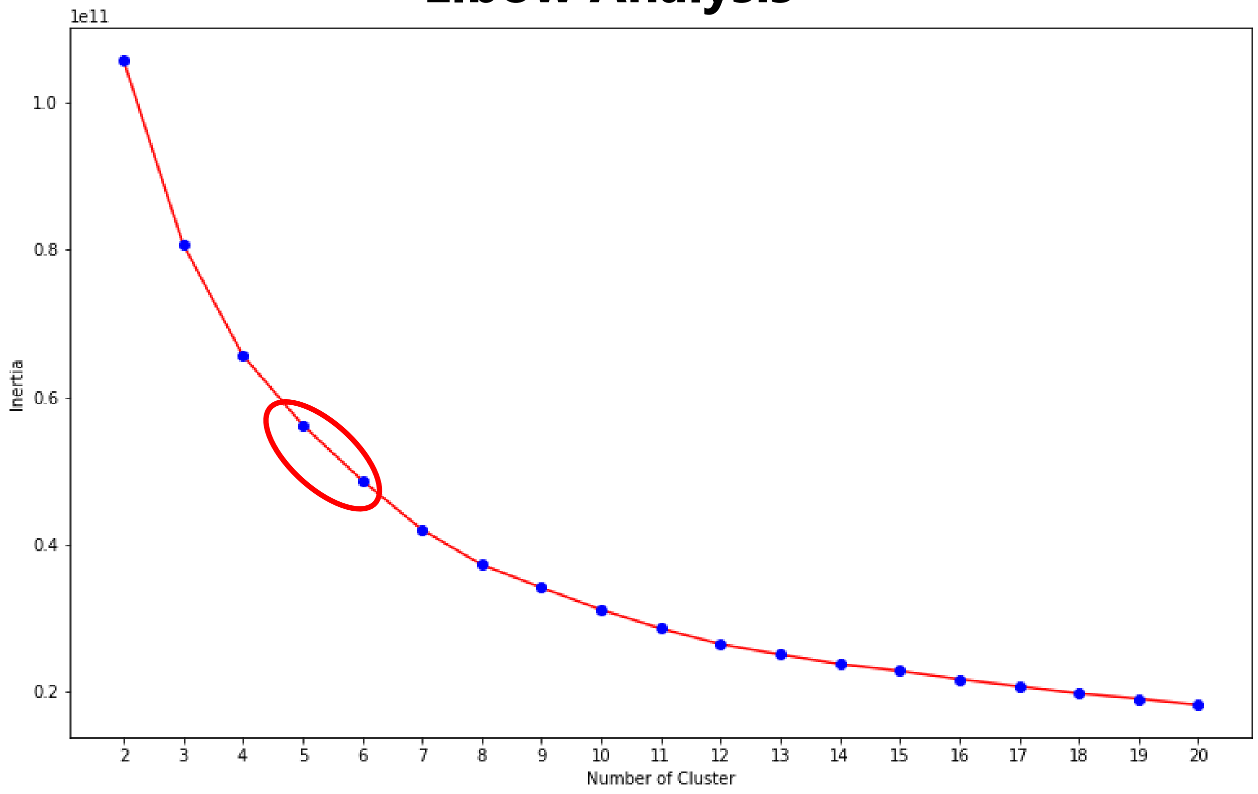
AHC

Gaussian Mixture

Data Clustering - KMeans

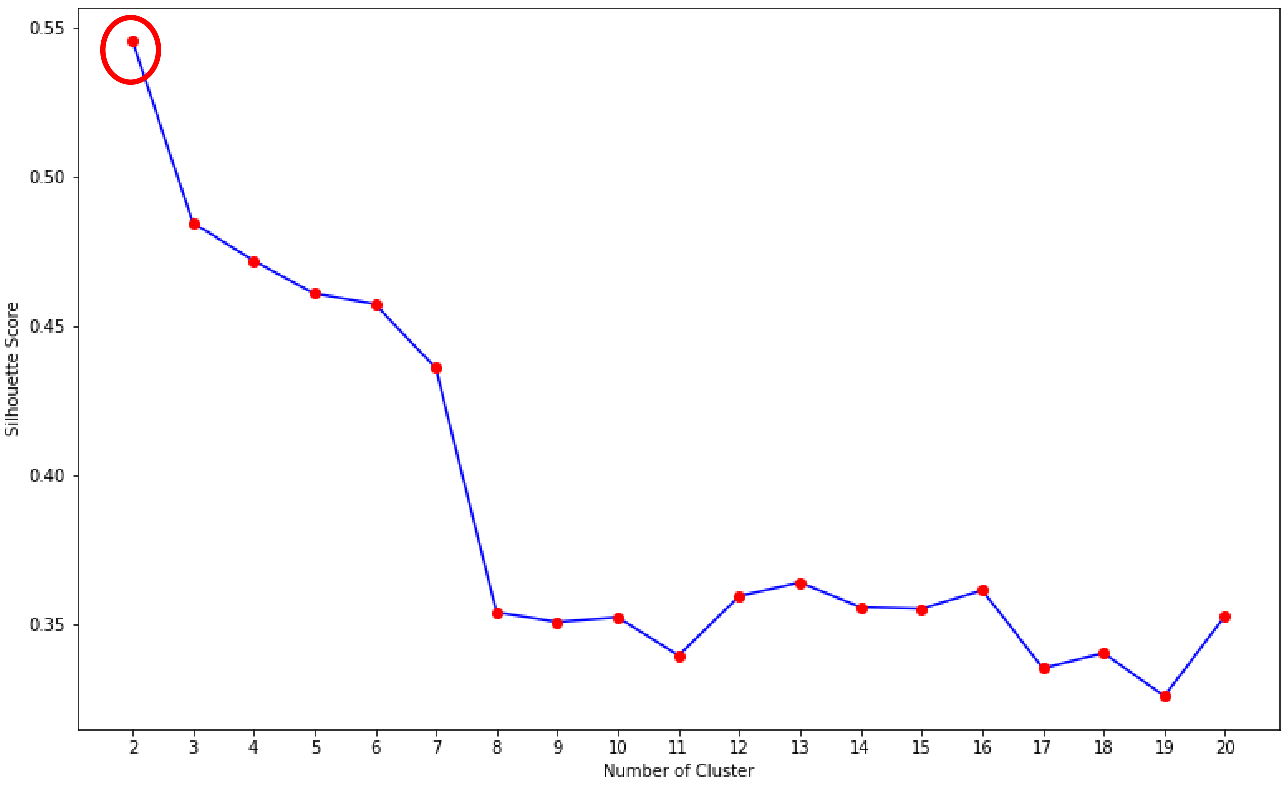
Ideal Number Of Clusters Analysis

Elbow Analysis



From Elbow Analysis the best number of cluster is 5

Silhouette Score

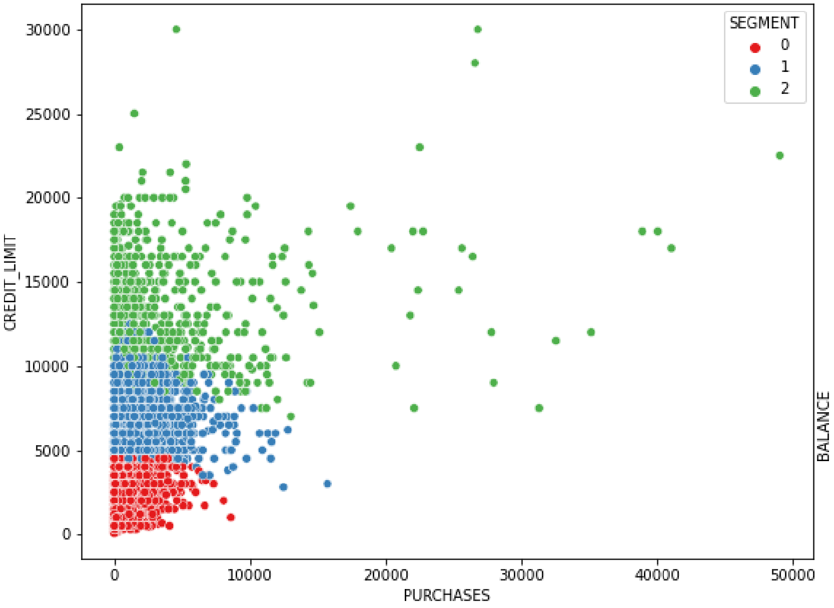


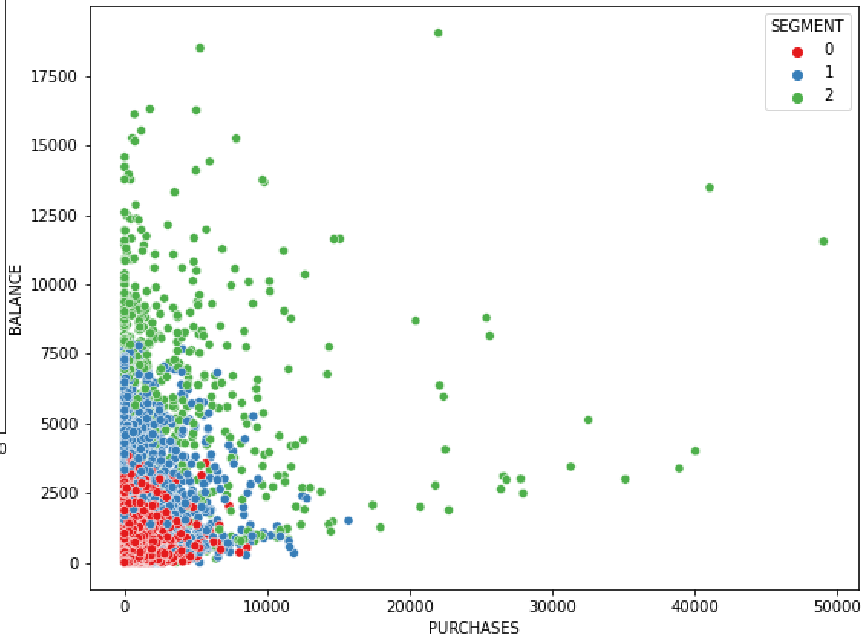
From Silhouette score, the best number of clusters is 2 with 0.55 score

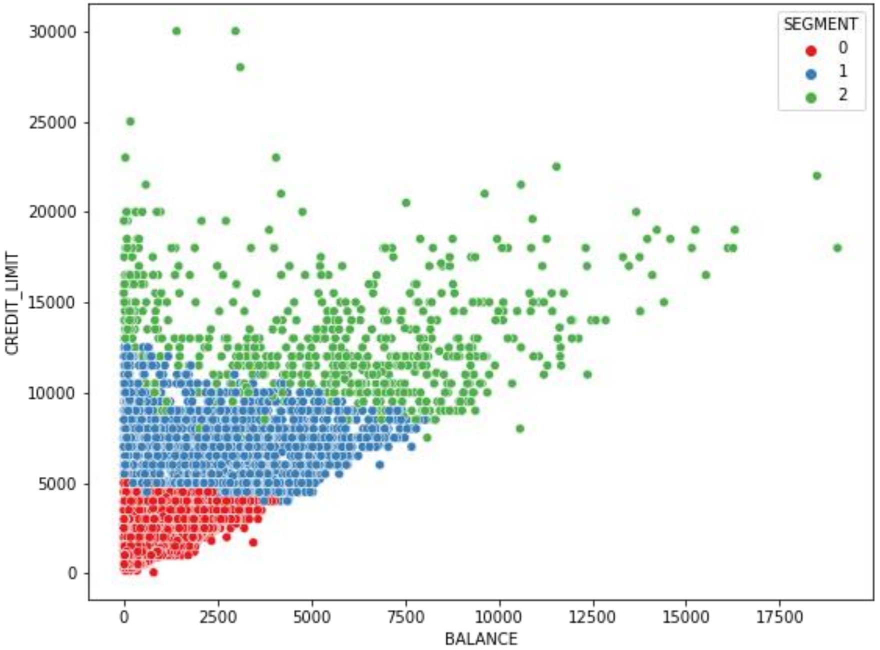
Yet, due to the product availability (3 products) we choose 3 clusters

Data Clustering - KMeans

KMeans Clustering Results

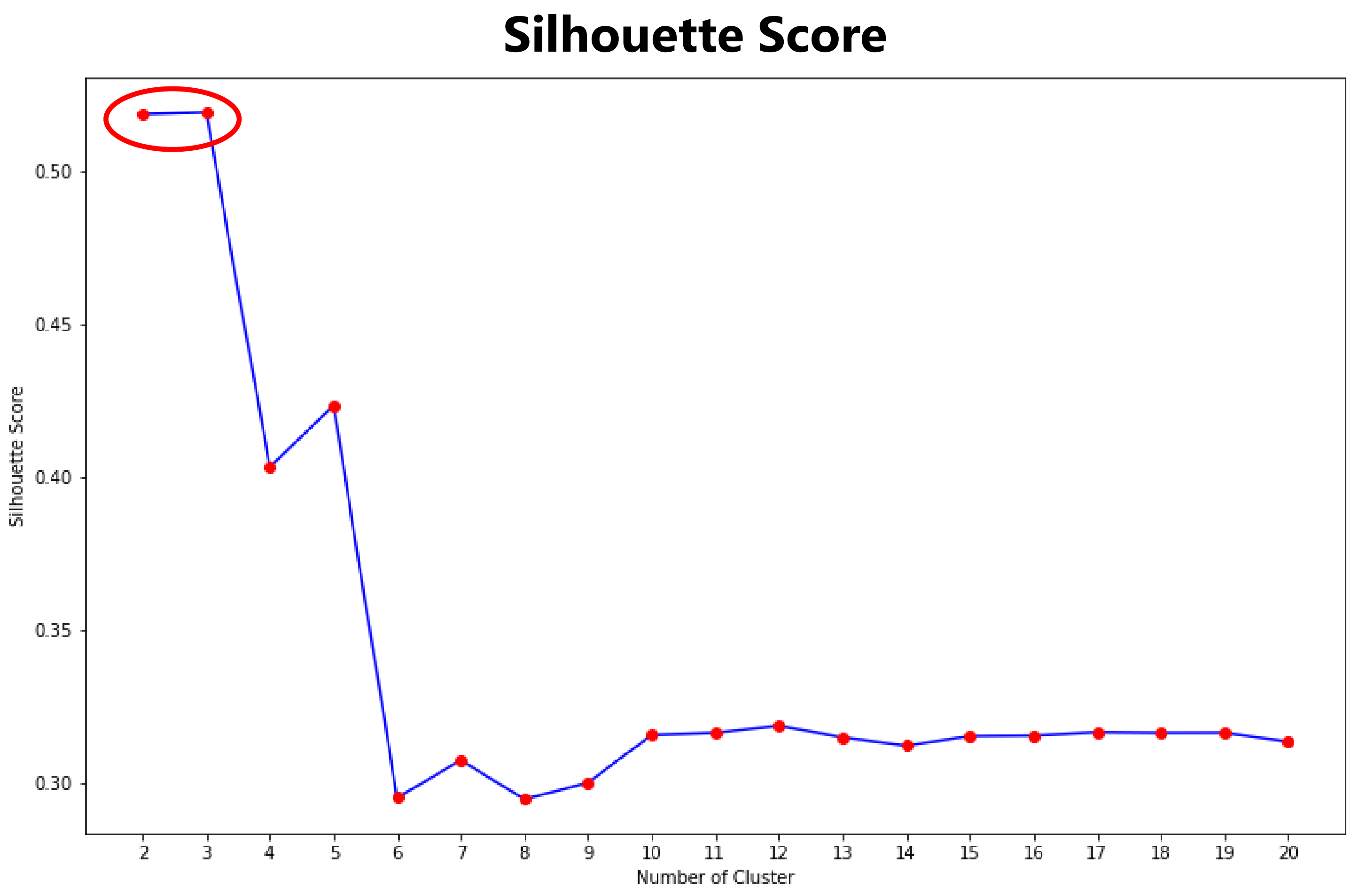






Data Clustering - AHC

Ideal Number Of Clusters Analysis

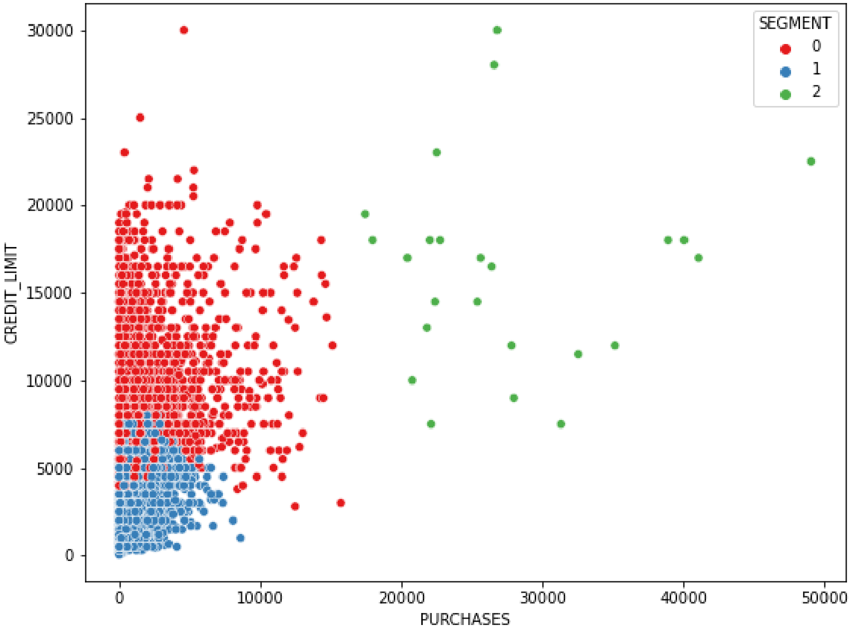


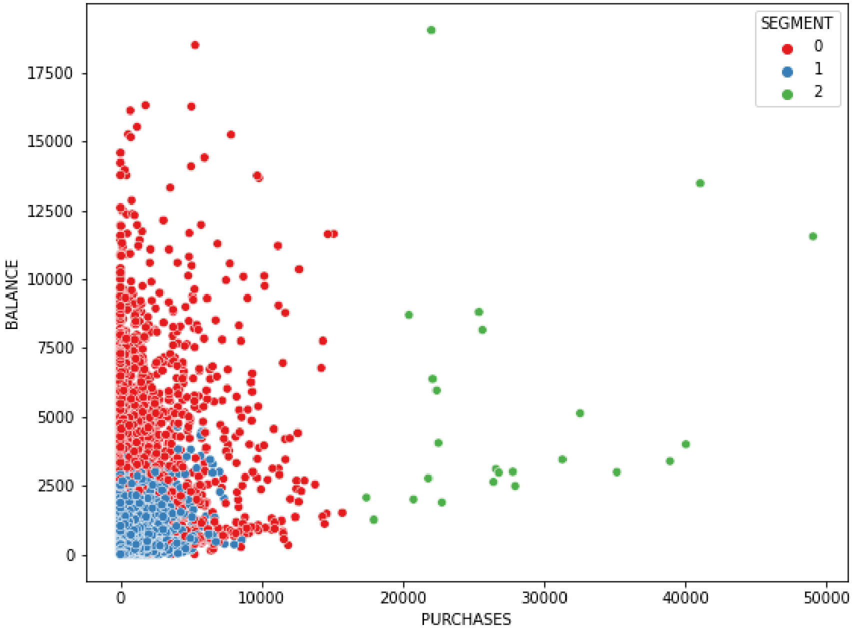
From Silhouette score, the best number of components (using AHC) is 2/3 with 0.52 Silhouette Score

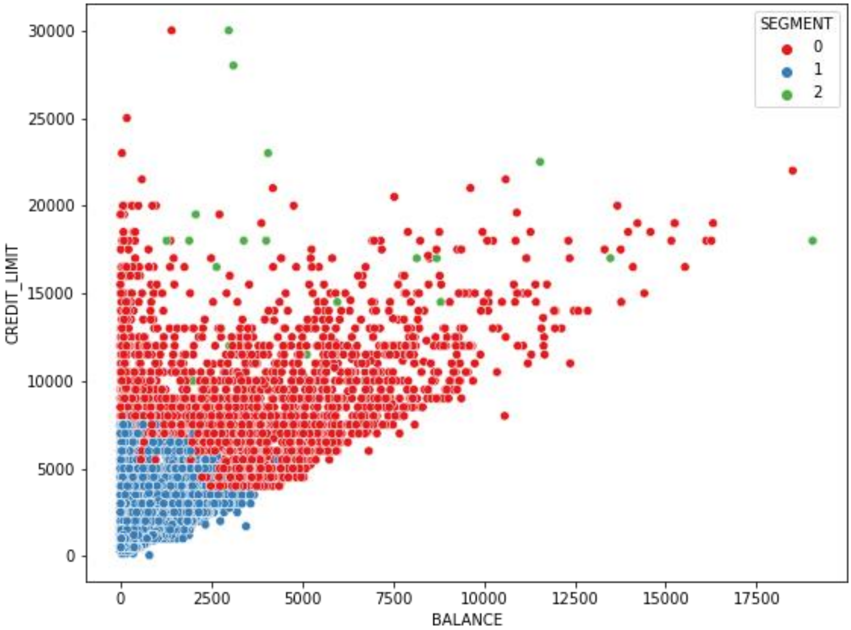
Yet, due to the product availability (3 products) we choose 3 clusters

Data Clustering - AHC

AHC Clustering Results



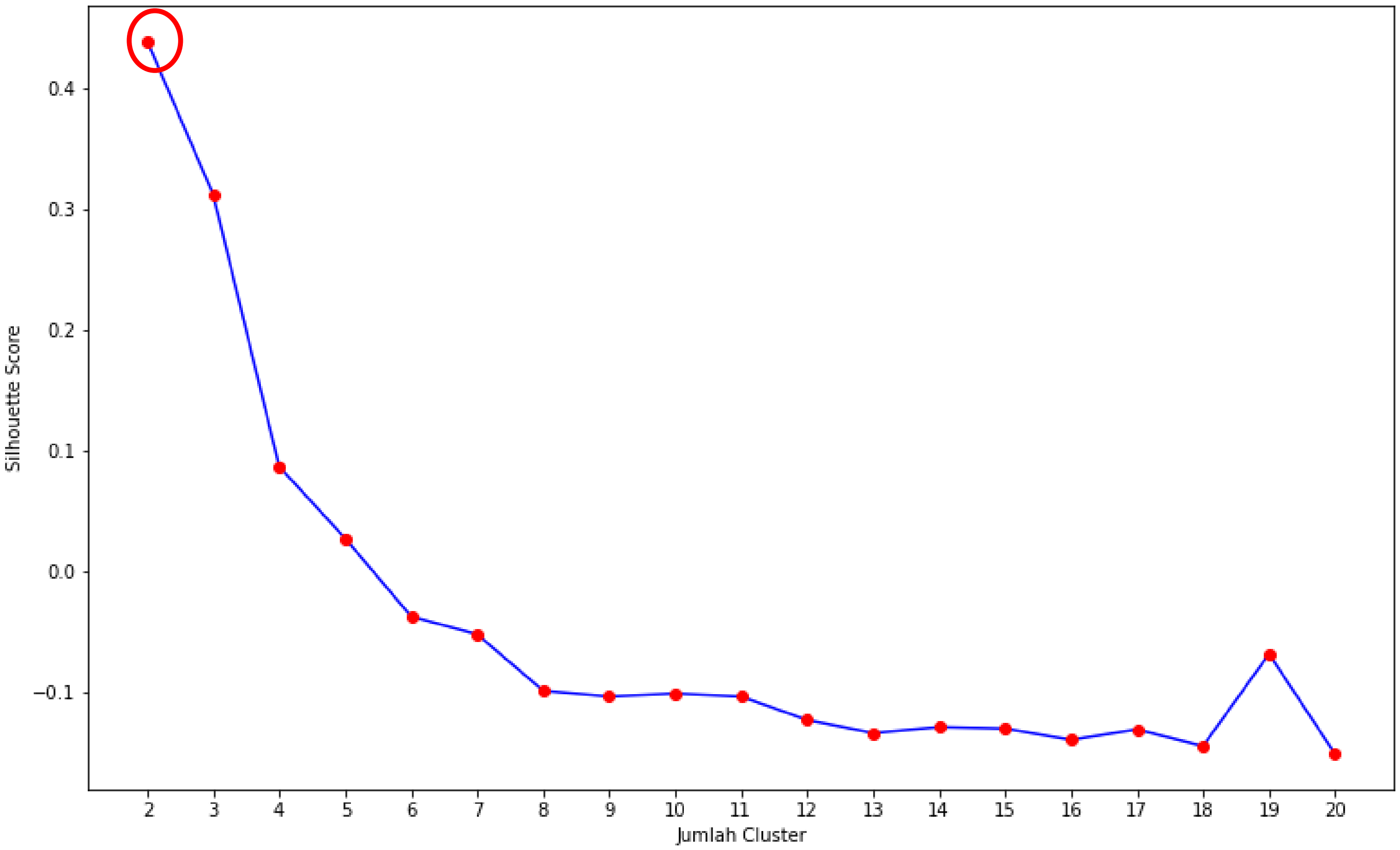




Data Clustering – Gaussian Mixture

Ideal Number Of Clusters Analysis

Silhouette Score

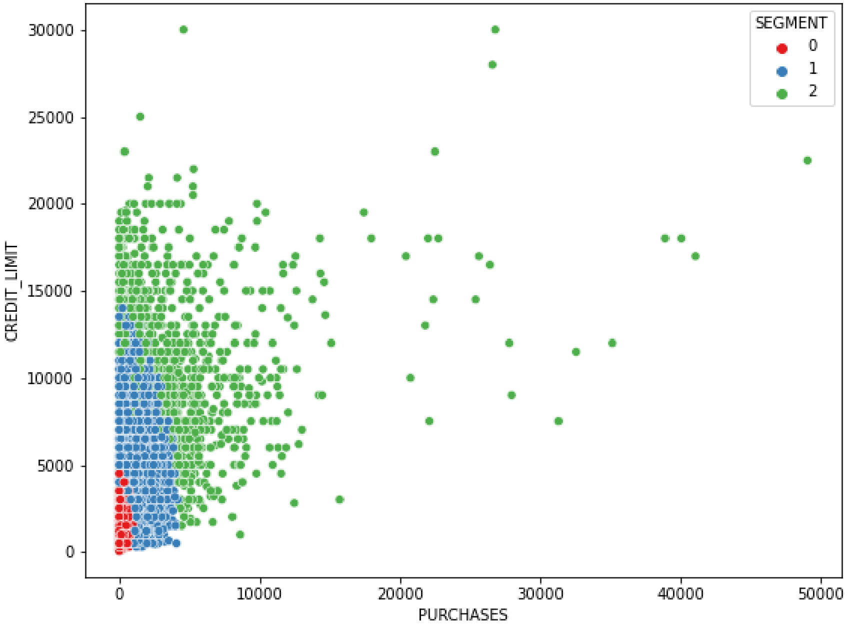


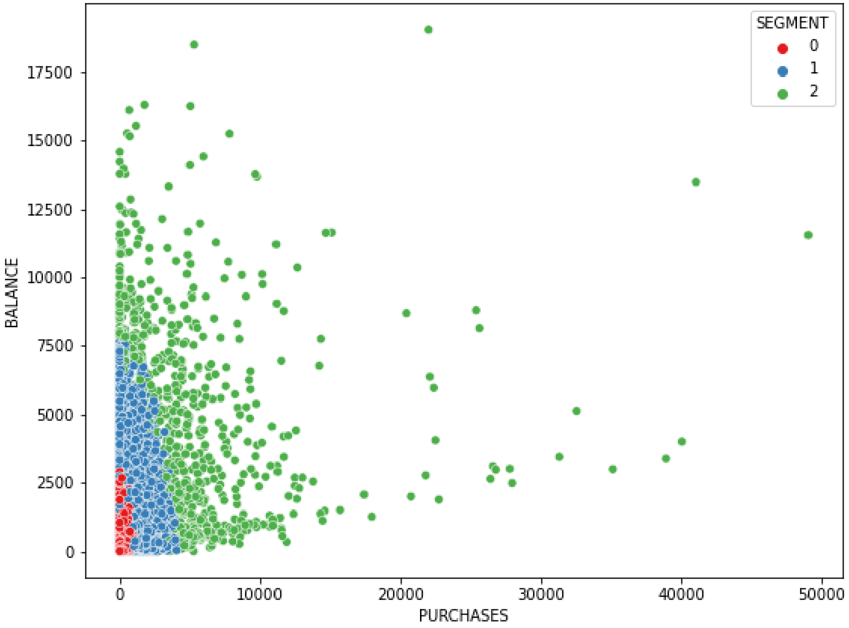
From Silhouette score, the best number of components (using Gaussian Mixture) is 2 with 0.44 Silhouette Score

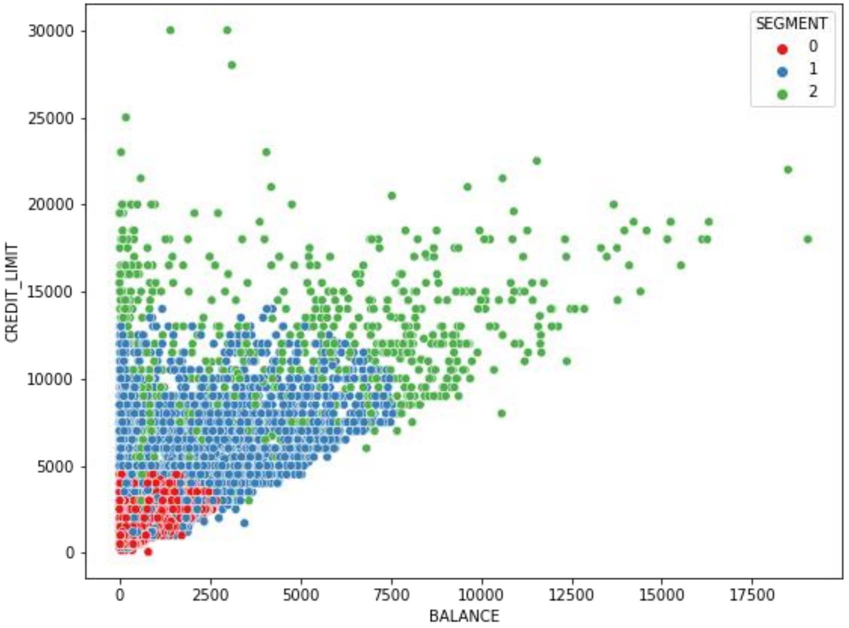
Yet, due to the product availability (3 products) we choose 3 clusters

Data Clustering – Gaussian Mixture

Gaussian Mixture Clustering Results







Data Clustering – Summary

Silhouette Score Summary

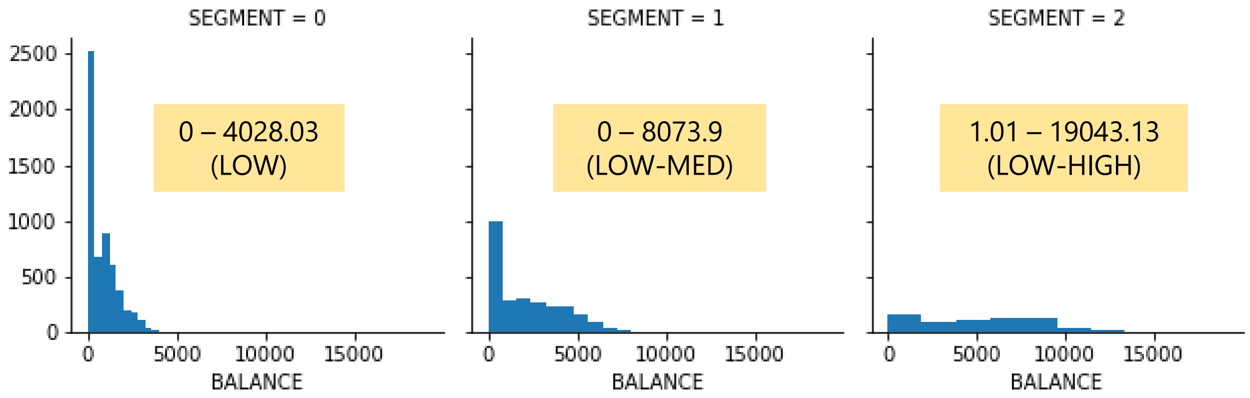
		Number of Clusters																		
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Clustering Method	**KMeans**	0.55	0.48	0.47	0.46	0.46	0.44	0.35	0.35	0.35	0.34	0.36	0.36	0.36	0.36	0.36	0.34	0.34	0.33	0.35
	AHC	0.52	0.52	0.40	0.42	0.30	0.31	0.29	0.30	0.32	0.32	0.32	0.32	0.31	0.32	0.32	0.32	0.32	0.32	0.31
	Gaussian	0.44	0.31	0.09	0.03	-0.04	-0.05	-0.10	-0.10	-0.10	-0.10	-0.12	-0.13	-0.13	-0.13	-0.14	-0.13	-0.14	-0.07	-0.15

- From the Silhouette Score using three different methods (KMeans, AHC, Gaussian Mixsture), the best number of clusters obtained is 2.
- Nevertheless, we choose to use **3 clustering** due to **Business Demand and Simulation**.
- Within 3 clustering, AHC method has better Silhouette Score (0.52) compared to KMeans (0.48). However, we choose **KMeans** method because has **better seperation of grouping**

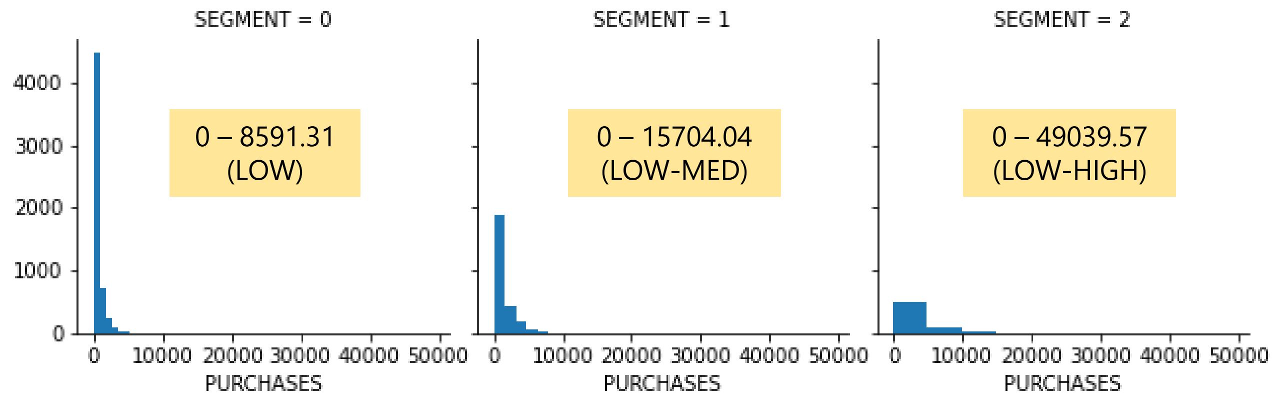
Data Clustering – Segment Interpretation

SEGMENT VS FEATURES DATA DISTRIBUTION

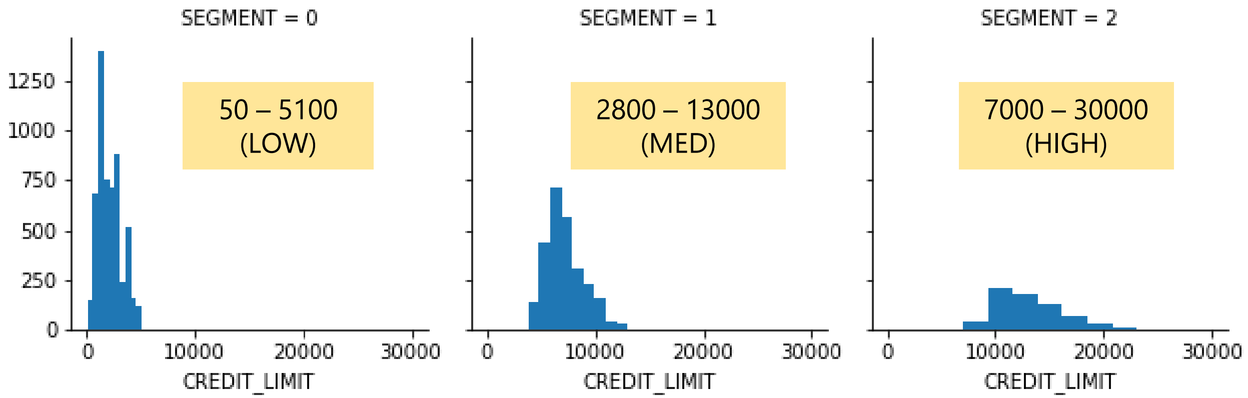
BALANCE VS SEGMENT



PURCHASES VS SEGMENT

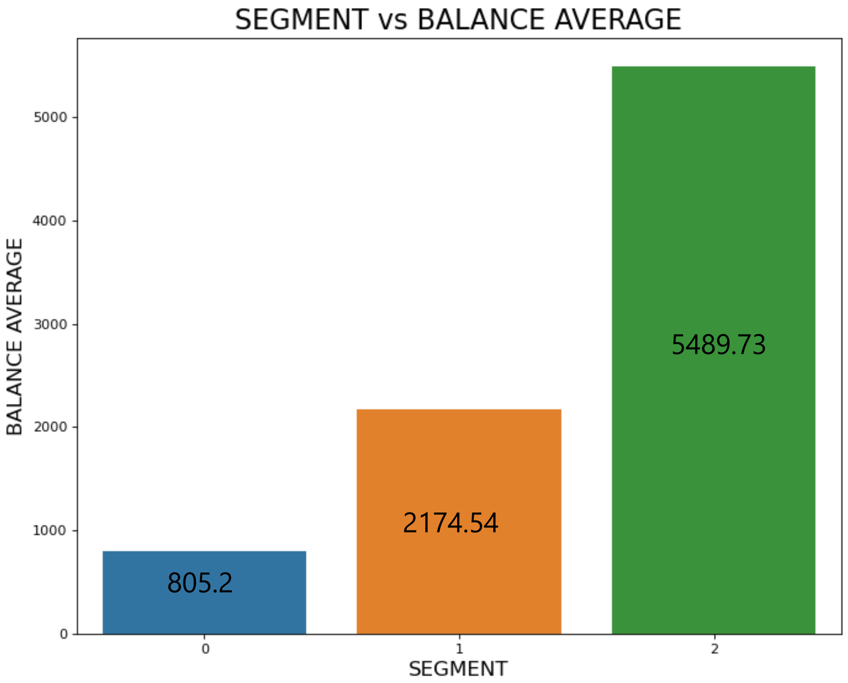


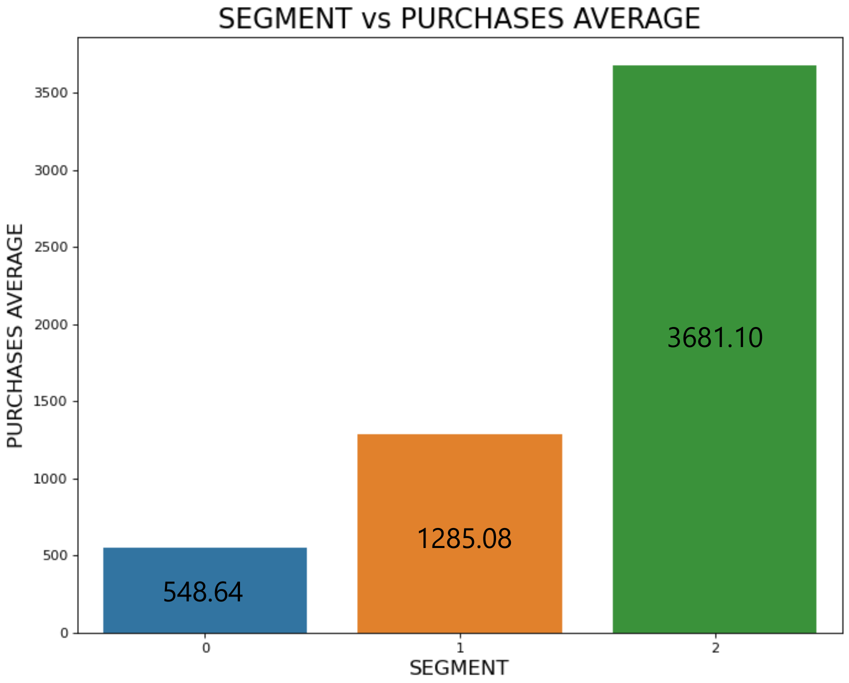
CREDIT_LIMIT VS SEGMENT

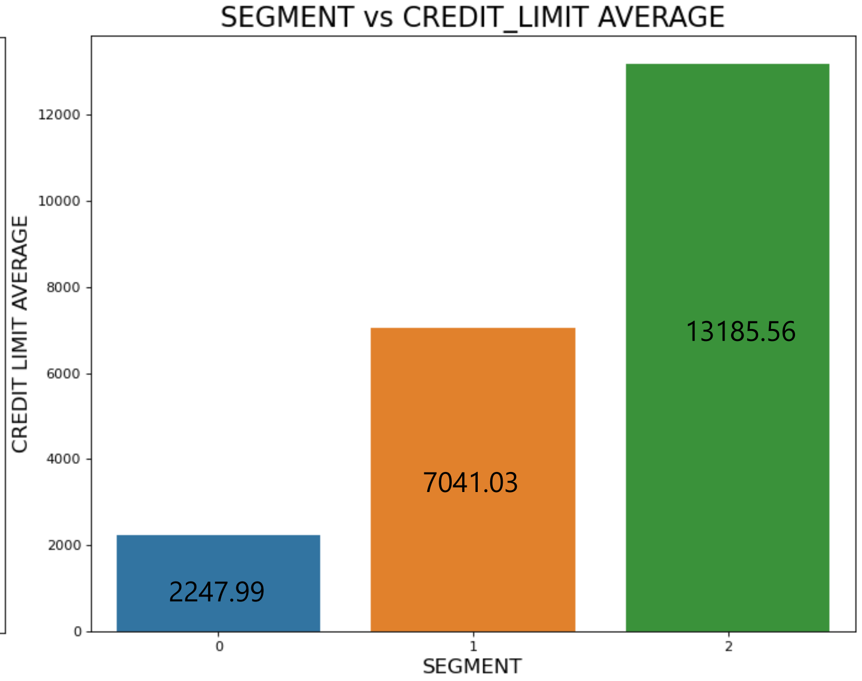


Data Clustering – Segment Interpretation

SEGMENT VS AVERAGE FEATURES

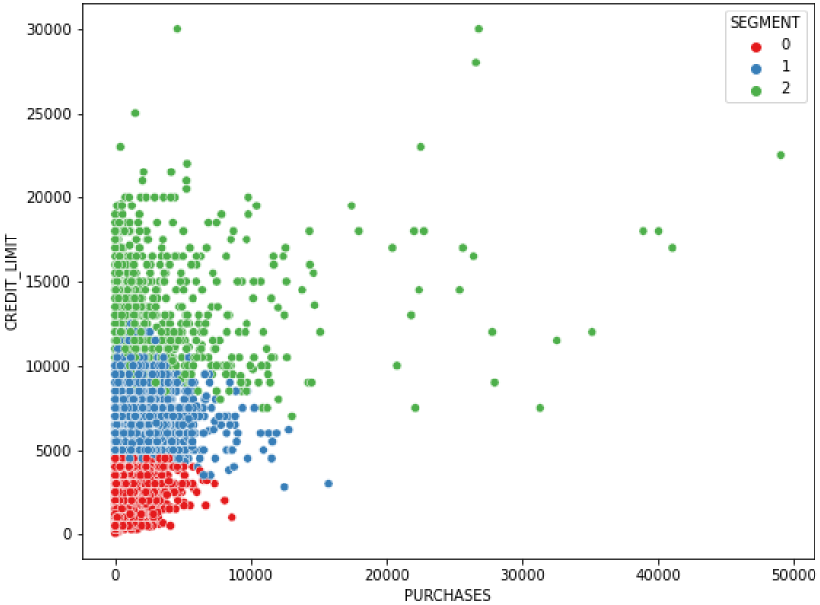


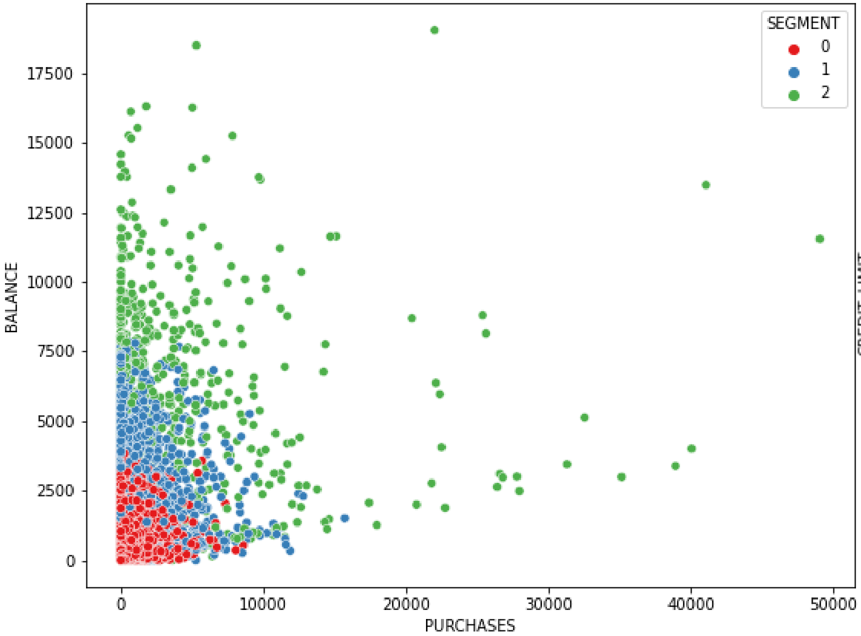


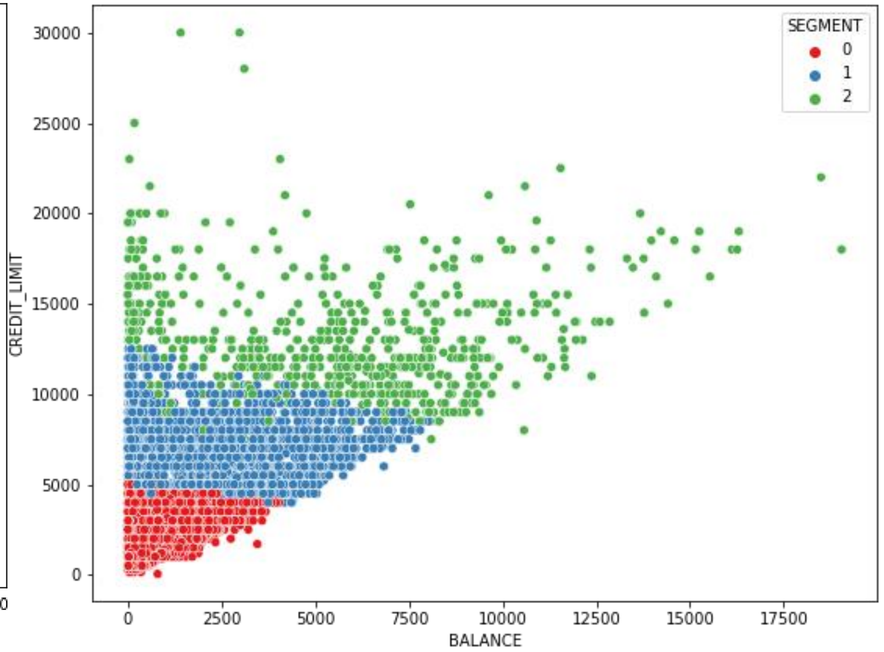


Data Clustering – Segment Interpretation

KMeans Clustering Results







SEGMENT 0 : LOW CUSTOMERS This customer group indicates a large group of customers who have **LOW BALANCES**, **small spenders (LOW PURCHASES**) with the **LOWEST CREDIT LIMIT**.

SEGMENT 1 : MEDIUM CUSTOMERS This customer group indicates a small group of customers who have **LOW-MEDIUM BALANCES**, **intermediate spenders (LOW-MEDIUM PURCHASES)** with **intermediate CREDIT LIMIT**.

SEGMENT 2 : HIGH CUSTOMERS This customer group indicates a small group of customers who have **LOW-HIGH BALANCES**, **high spenders (LOW-HIGH PURCHASES)** with **HIGH CREDIT LIMIT**.

5. EDA, Multiclass Machine Learning Modelling & Evaluation

Exploratory Data Analysis

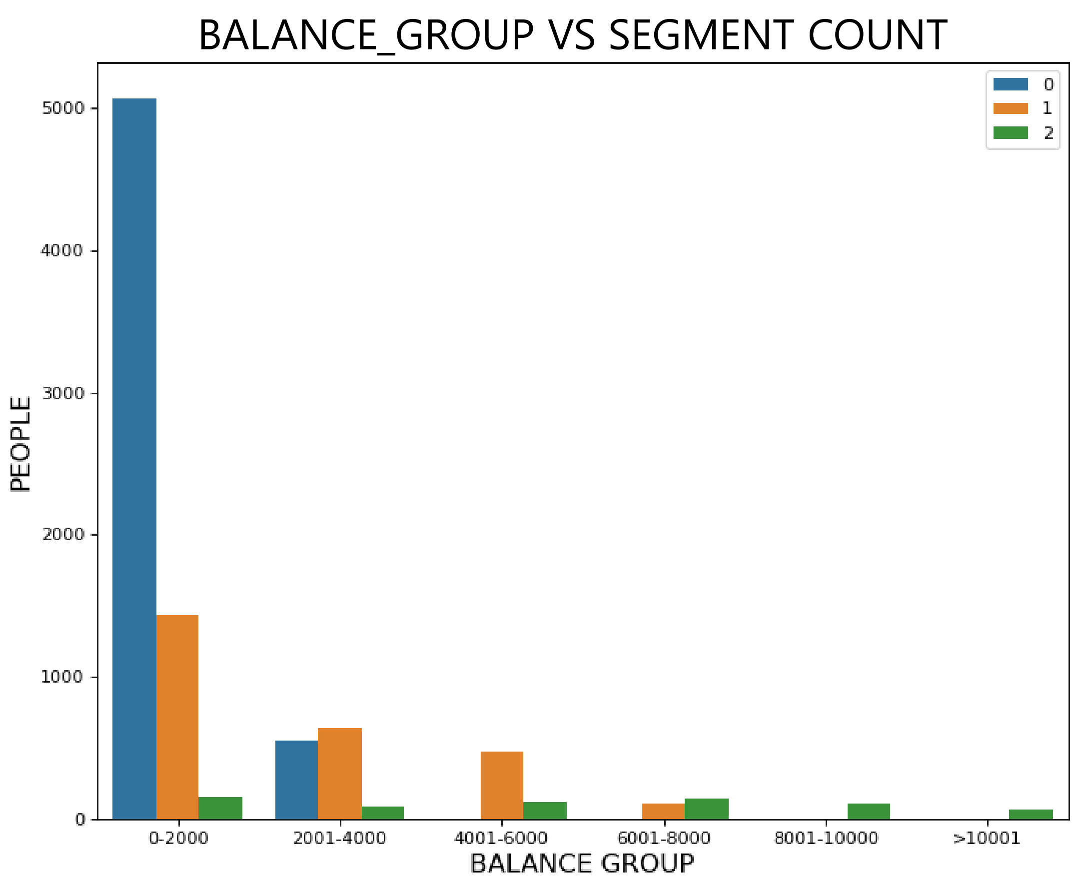
For EDA we do the following steps below:

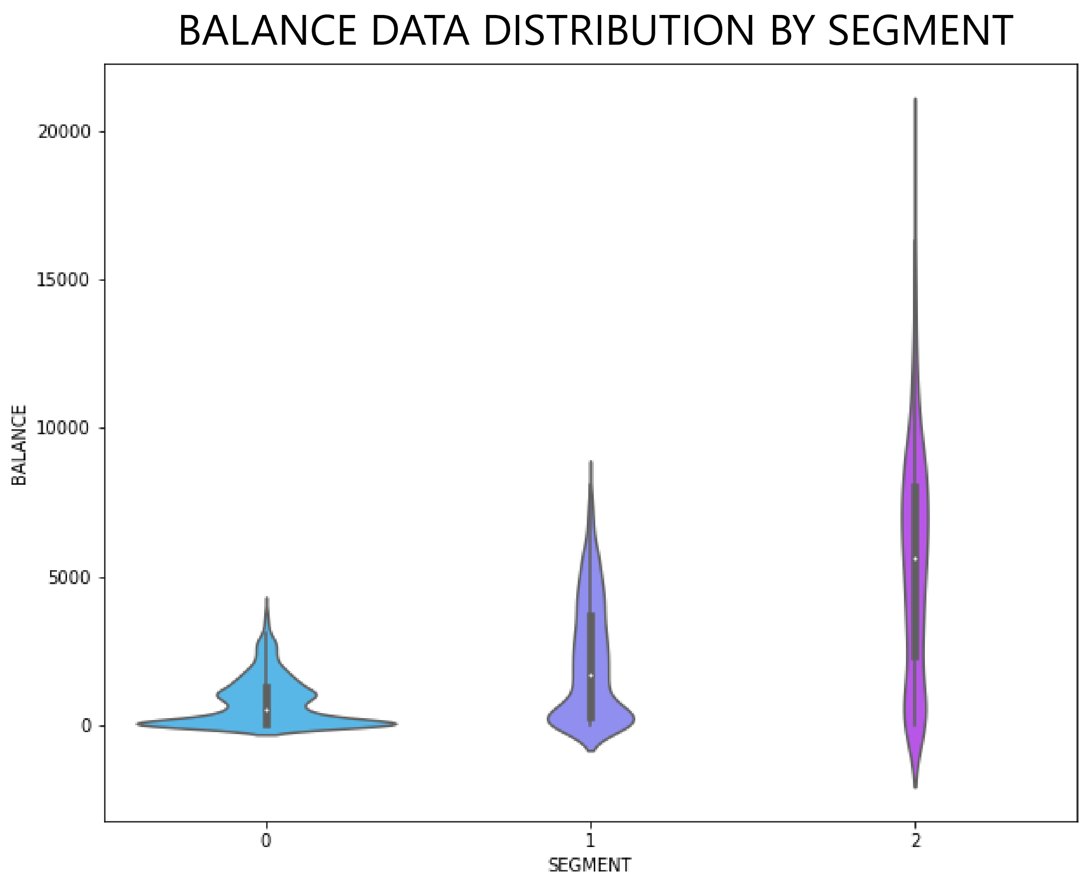
- **Binning**
- **Aggregating Columns**
- **Visualization**
- **Insight & Conclusion**

BUSINESS QUESTIONS

- What features which have impact to SEGMENT?

Exploratory Data Analysis

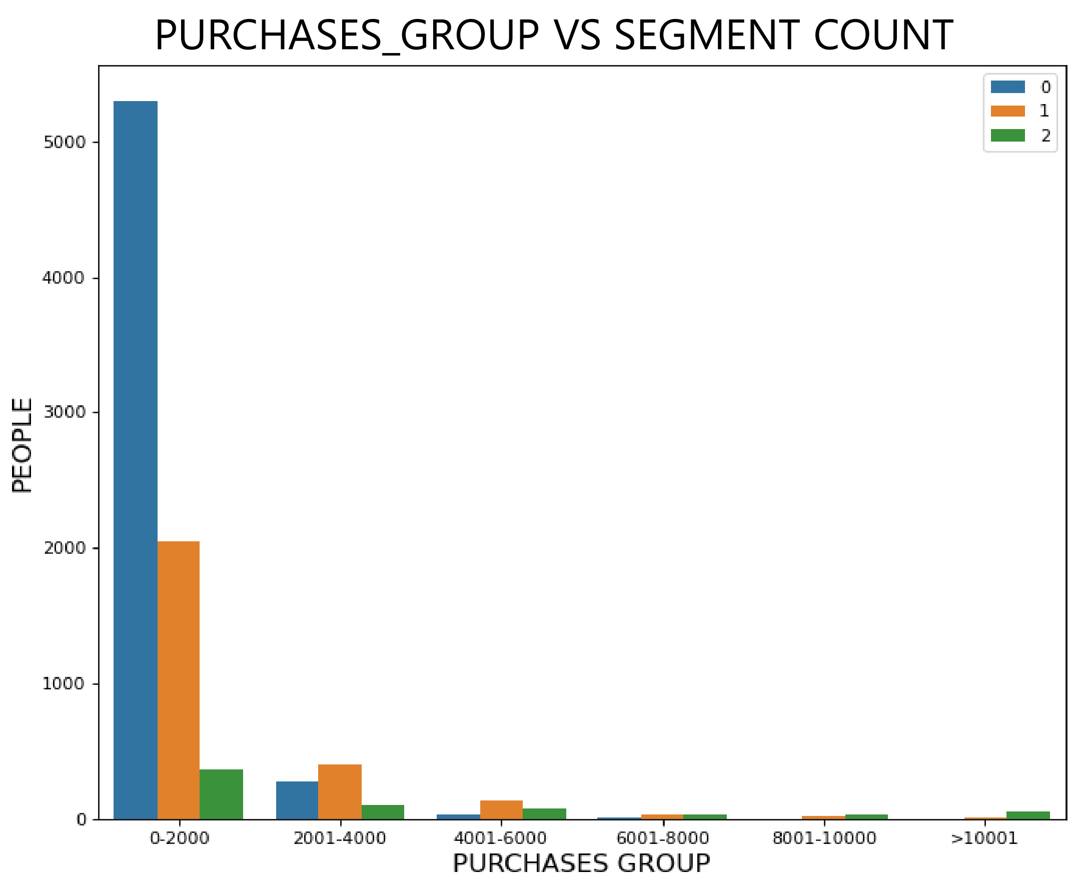


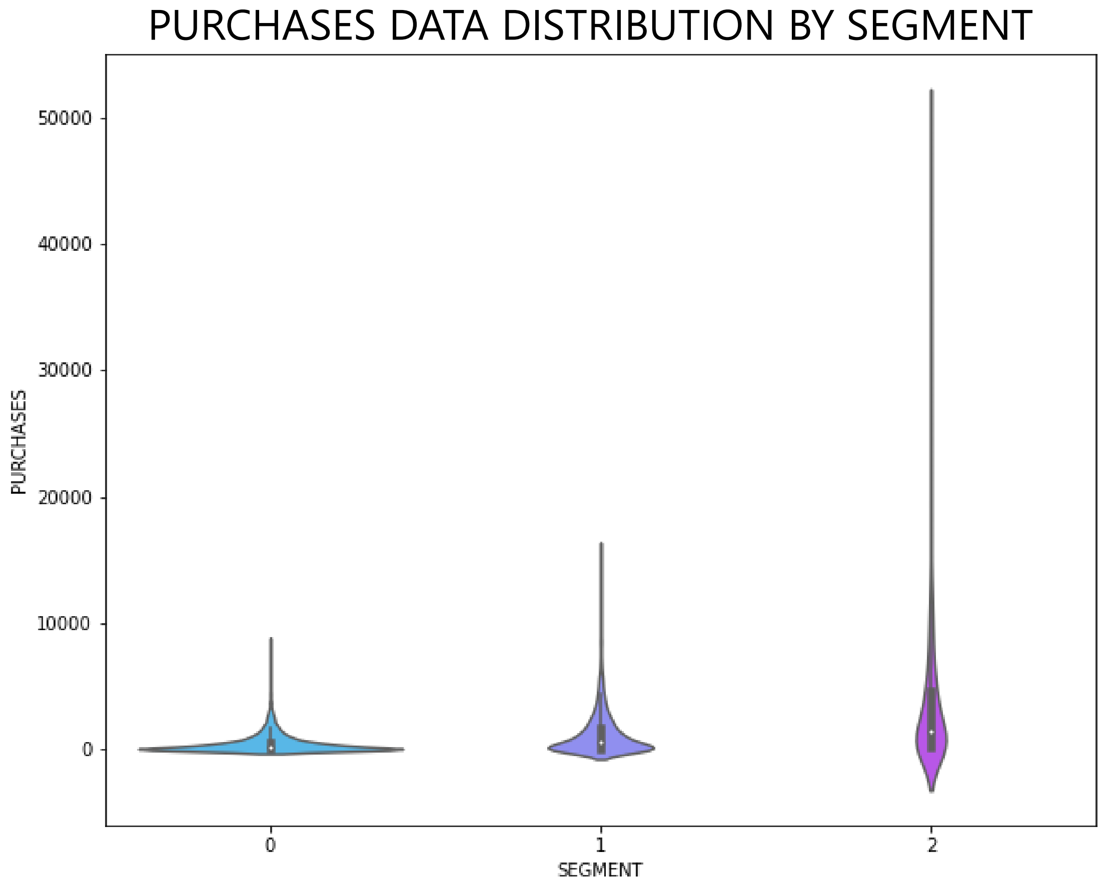


- SEGMENT 2 customer's AVERAGE BALANCE are 5489.73 dollars
- SEGMENT 1 customer's AVERAGE BALANCE are 2174.54 dollars
- SEGMENT 0 customer's AVERAGE BALANCE are 805.2 dollars

- SEGMENT 2 customer's BALANCE RANGE is 0 - >10001 dollars
- SEGMENT 1 customer's BALANCE RANGE is 0 - 8000 dollars
- SEGMENT 0 customer's BALANCE RANGE is 0 - 4000 dollars

Exploratory Data Analysis



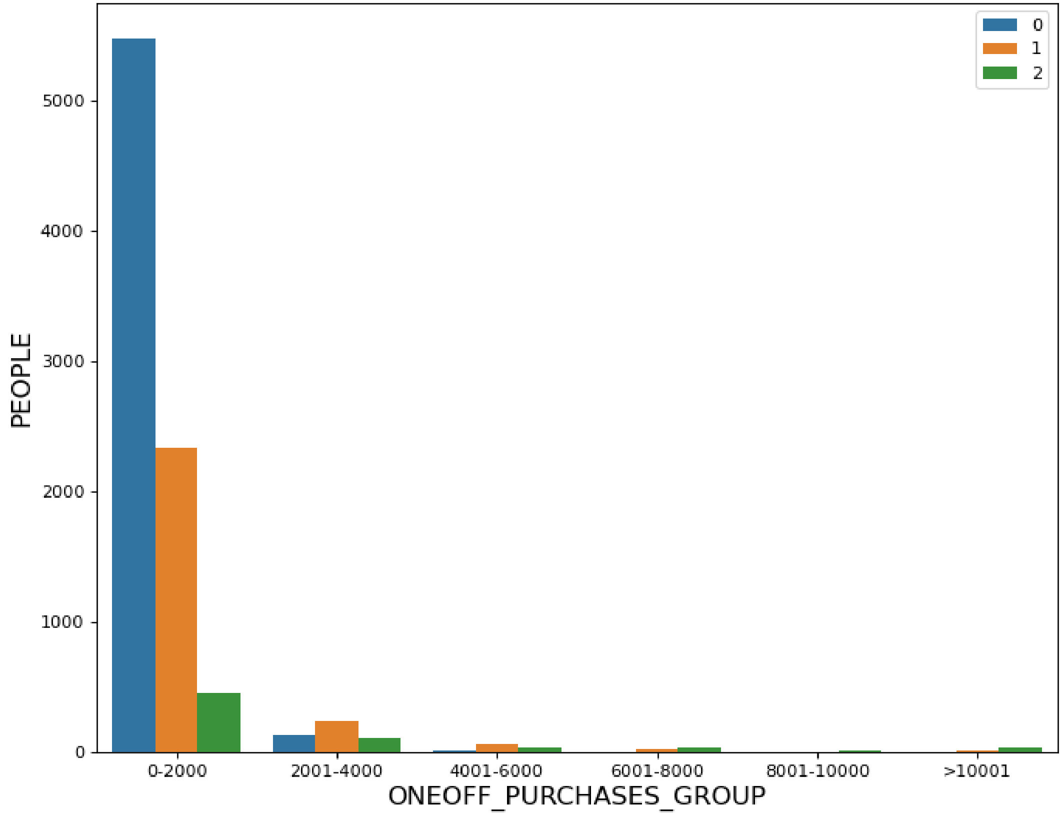


- SEGMENT 2 customer's AVERAGE PURCHASES are 14927.14
- SEGMENT 1 customer's AVERAGE PURCHASES are 1320.44
- SEGMENT 0 customer's AVERAGE PURCHASES are 791.53

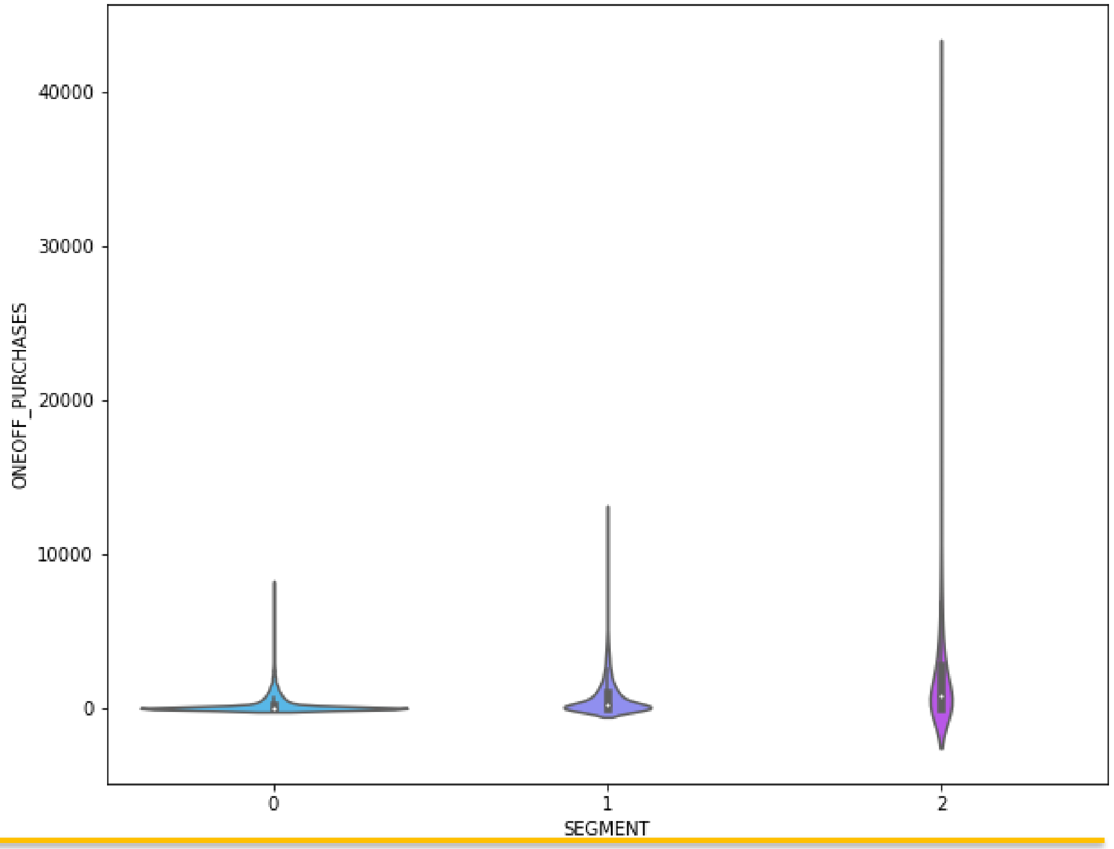
- SEGMENT 2 customer's PURCHASES RANGE is 0 - >10001 dollars
- SEGMENT 1 customer's PURCHASES RANGE is 0 - >10001 dollars
- SEGMENT 0 customer's PURCHASES RANGE is 0 - 8000 dollars

Exploratory Data Analysis

ONEOFF_PURCHASES_GROUP VS SEGMENT COUNT



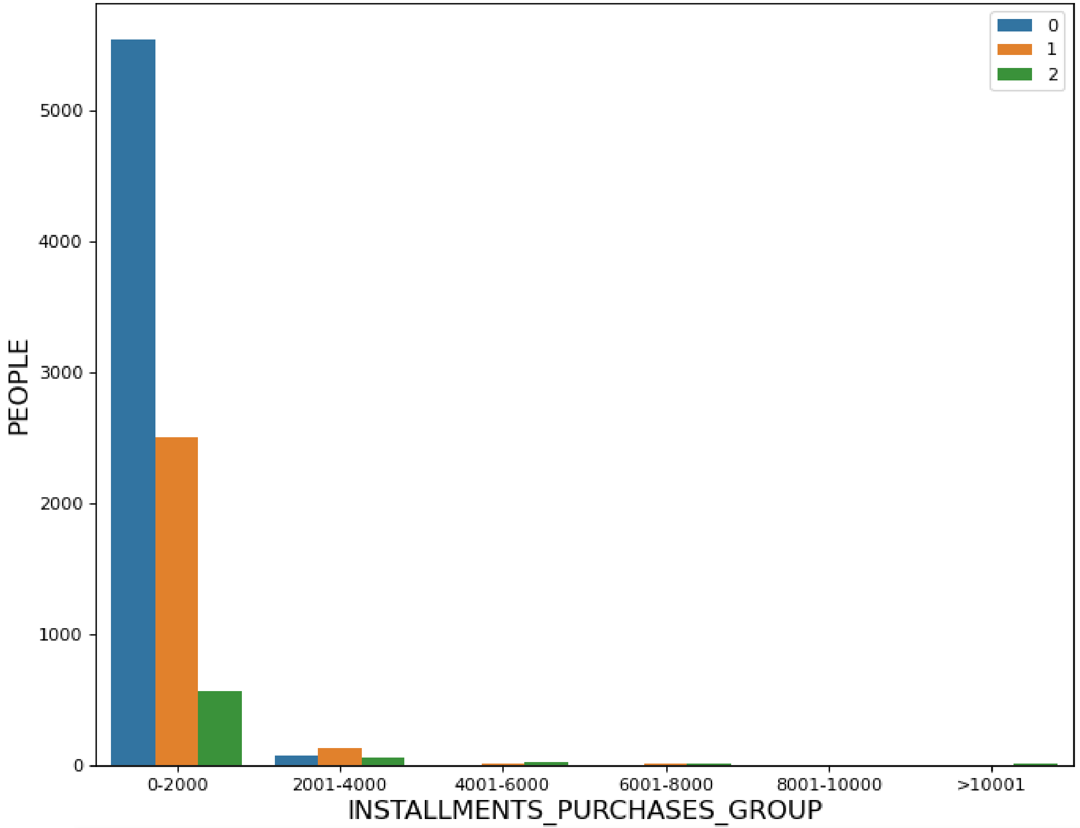
ONEOFF_PURCHASES DATA DISTRIBUTION BY SEGMENT



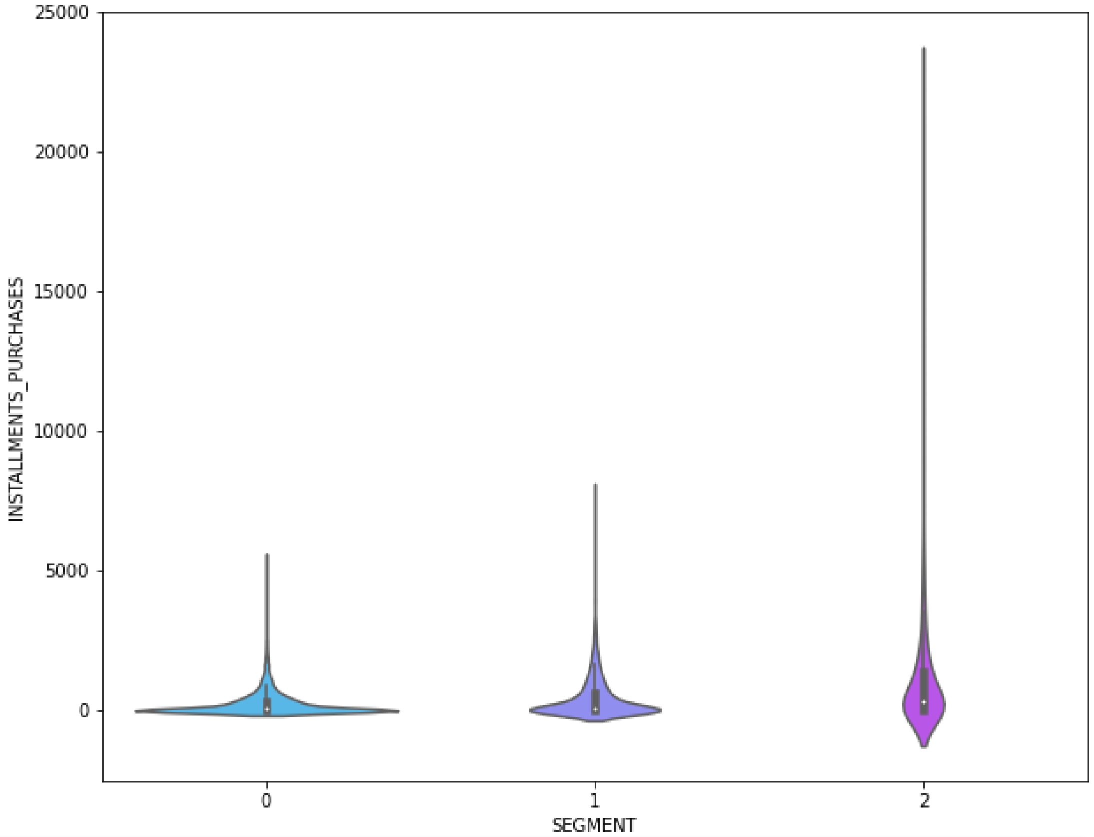
- SEGMENT 2 customer's average ONEOFF_PURCHASES are 12487.68
- SEGMENT 1 customer's average ONEOFF_PURCHASES are 734.17
- SEGMENT 0 customer's average ONEOFF_PURCHASES are 433.49
- SEGMENT 2 customer's ONEOFF_PURCHASES RANGE is 0 - >10001 dollars
- SEGMENT 1 customer's ONEOFF_PURCHASES RANGE is 0 - >10001 dollars
- SEGMENT 0 customer's ONEOFF_PURCHASES RANGE is 0 - 10000 dollars

Exploratory Data Analysis

INSTALLMENTS_PURCHASES_GROUP VS SEGMENT COUNT



INSTALLMENTS_PURCHASES DATA DISTRIBUTION BY SEGMENT

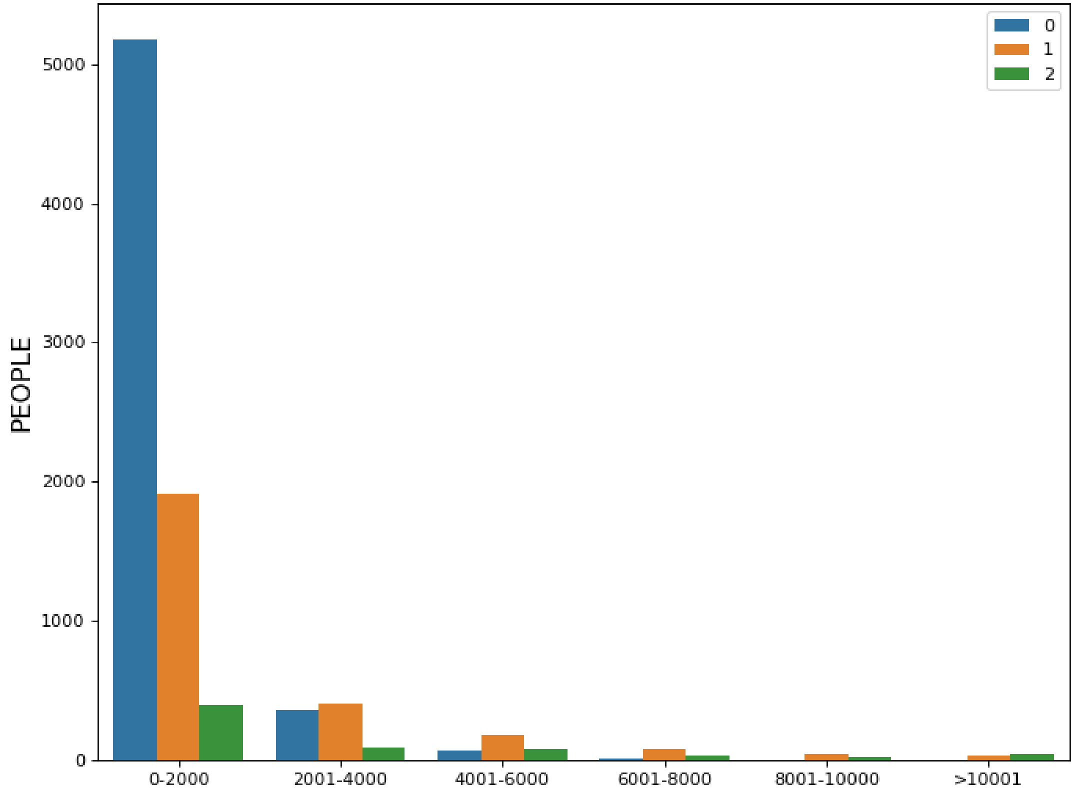


- SEGMENT 2 customer's average INSTALLMENT_PURCHASES are 2439.46
- SEGMENT 1 customer's average INSTALLMENT_PURCHASES are 586.33
- SEGMENT 0 customer's average INSTALLMENT_PURCHASES are 358.38

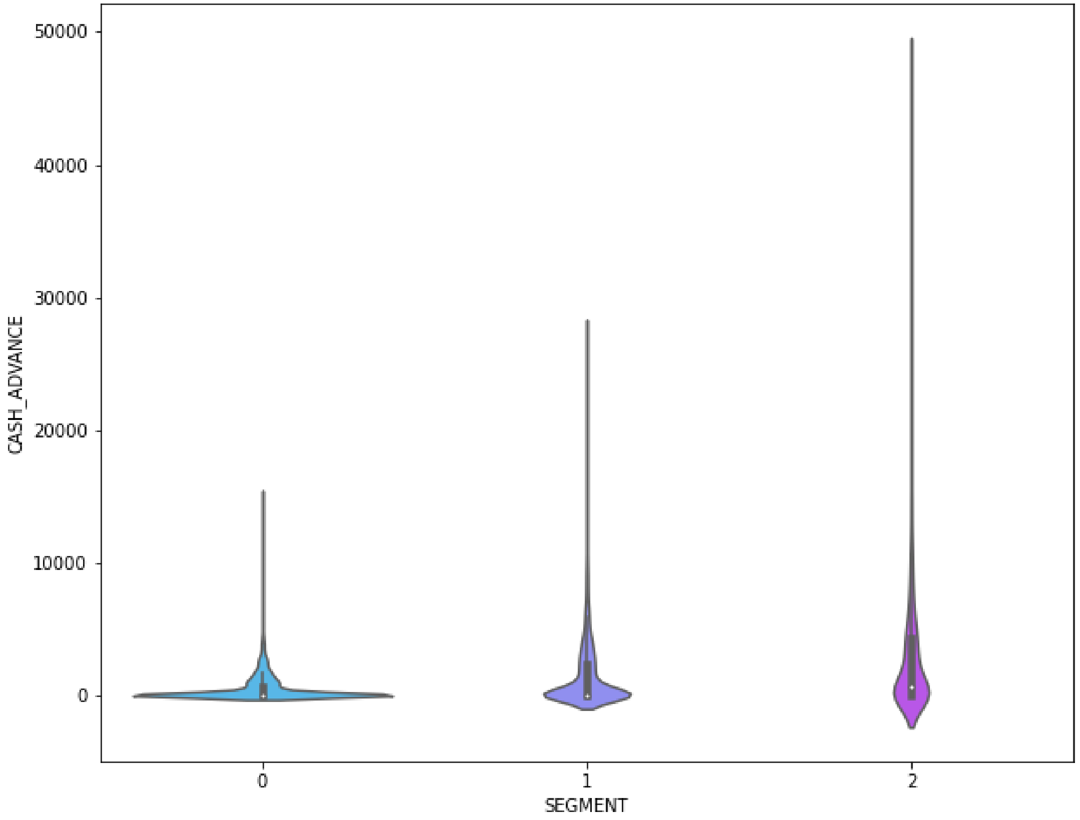
- SEGMENT 2 customer's INSTALLMENT_PURCHASES RANGE is 0 - >10001 dollars
- SEGMENT 1 customer's INSTALLMENT_PURCHASES RANGE is 0 - 8000 dollars
- SEGMENT 0 customer's INSTALLMENT_PURCHASES RANGE is 0 - 6000 dollars

Exploratory Data Analysis

CASH_ADVANCE_GROUP VS SEGMENT COUNT



CASH_ADVANCE DATA DISTRIBUTION BY SEGMENT

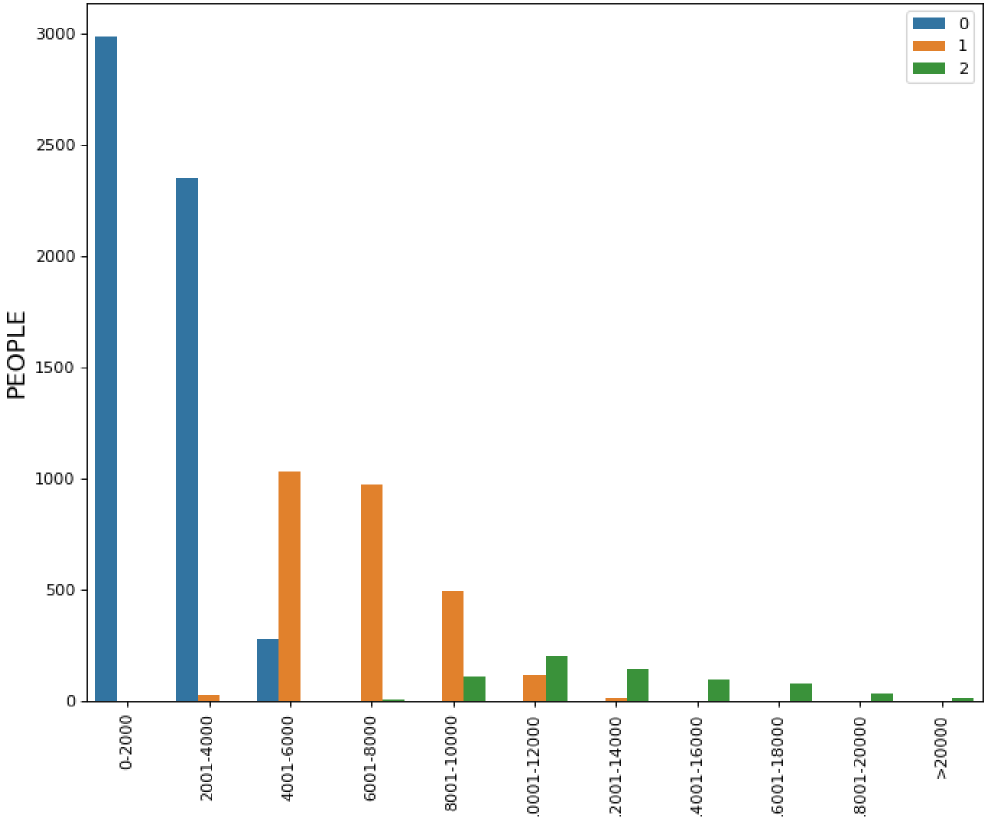


- SEGMENT 2 customer's AVERAGE CASH_ADVANCE are 2439.46
- SEGMENT 1 customer's AVERAGE CASH_ADVANCE are 586.33
- SEGMENT 0 customer's AVERAGE CASH_ADVANCE are 358.38

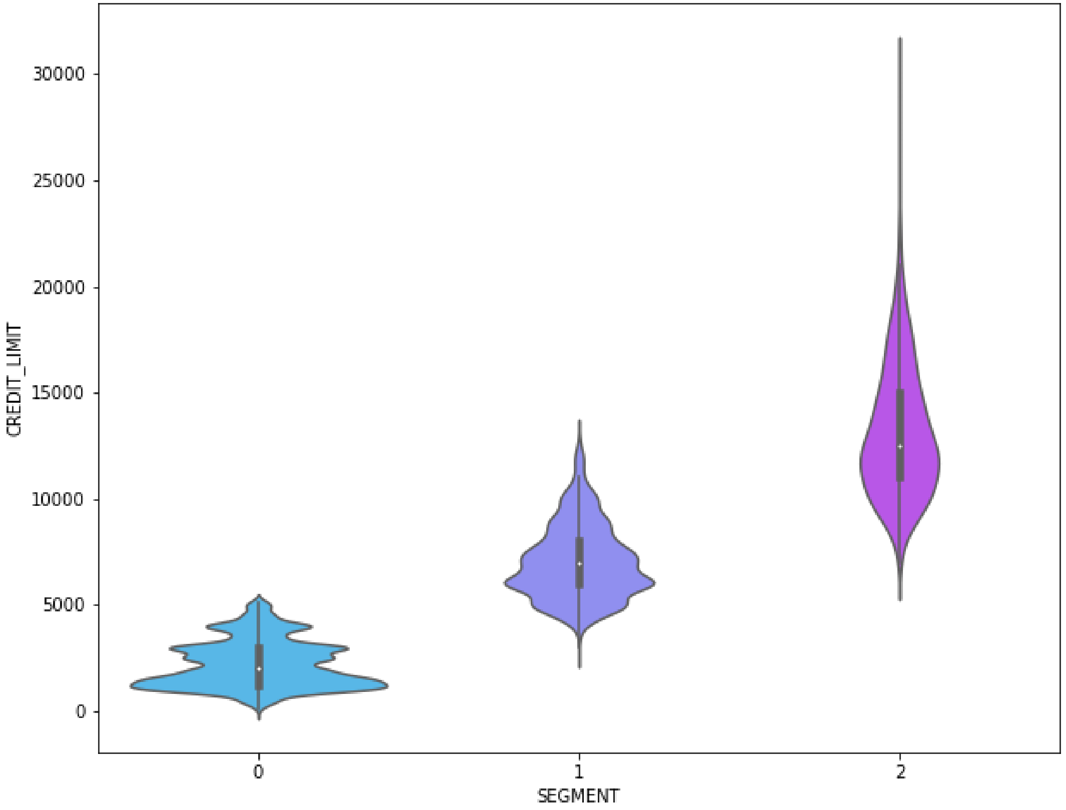
- SEGMENT 2 customer's CASH_ADVANCE RANGE is 0 - >10001 dollars
- SEGMENT 1 customer's CASH_ADVANCE RANGE is 0 - >10001 dollars
- SEGMENT 0 customer's CASH_ADVANCE RANGE is 0 - >10001 dollars

Exploratory Data Analysis

CREDIT_LIMIT_GROUP VS SEGMENT COUNT



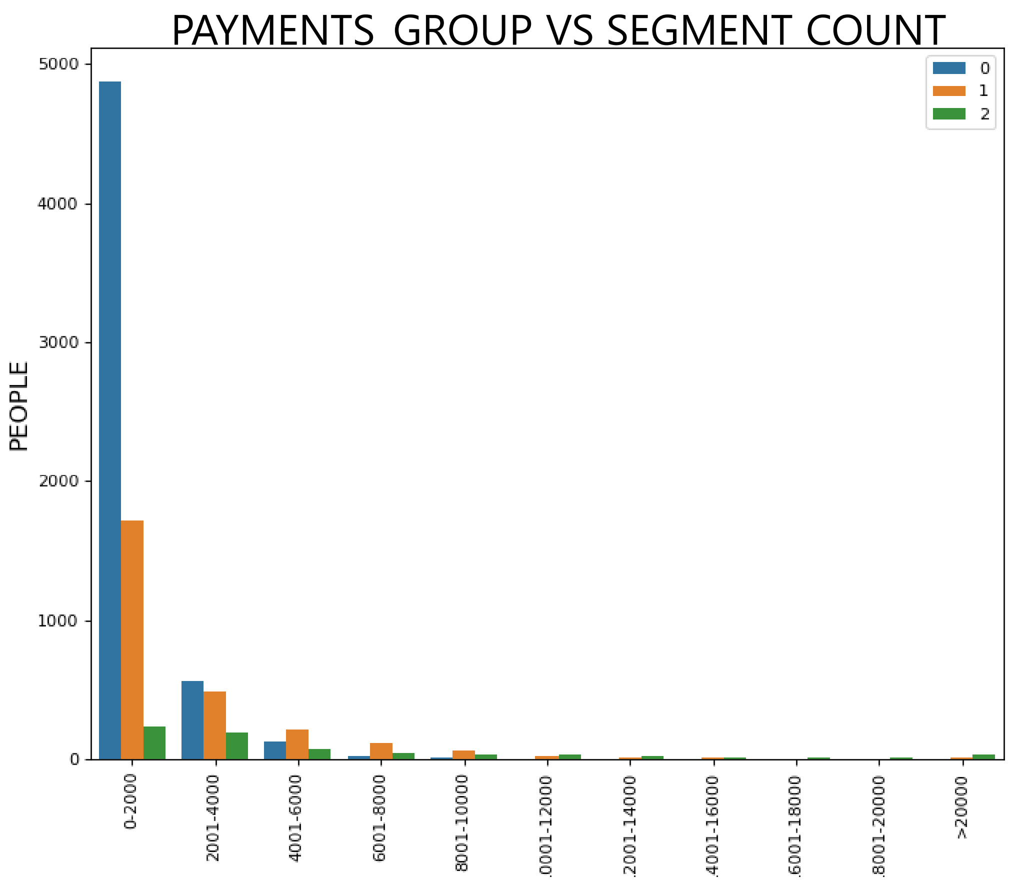
CREDIT_LIMIT DATA DISTRIBUTION BY SEGMENT

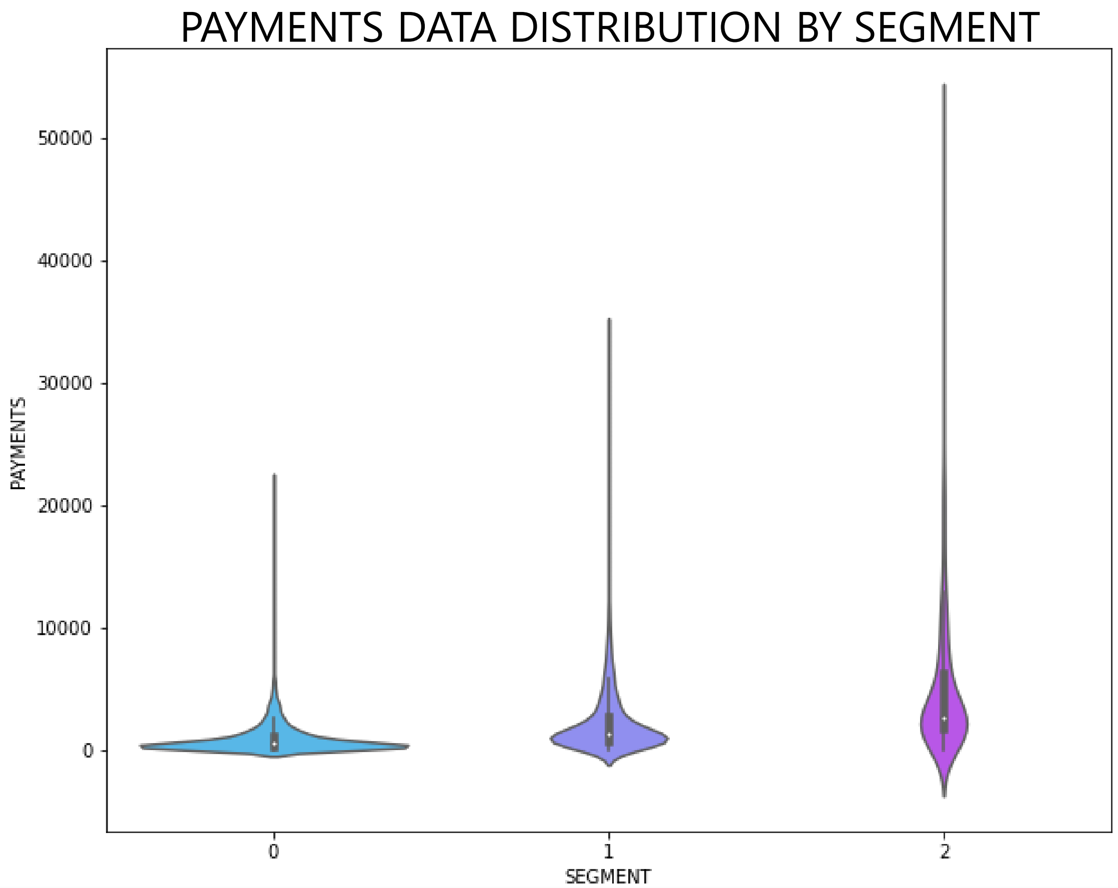


- SEGMENT 2 customer's AVERAGE CREDIT_LIMIT are 11041.28
- SEGMENT 1 customer's AVERAGE CREDIT_LIMIT are 8661.60
- SEGMENT 0 customer's AVERAGE CREDIT_LIMIT are 3713.53

- SEGMENT 2 customer's CREDIT_LIMIT RANGE is 6000 - >20001 dollars
- SEGMENT 1 customer's CREDIT_LIMIT RANGE is 2000 - 14000 dollars
- SEGMENT 0 customer's CREDIT_LIMIT RANGE is 0 - 6000 dollars

Exploratory Data Analysis



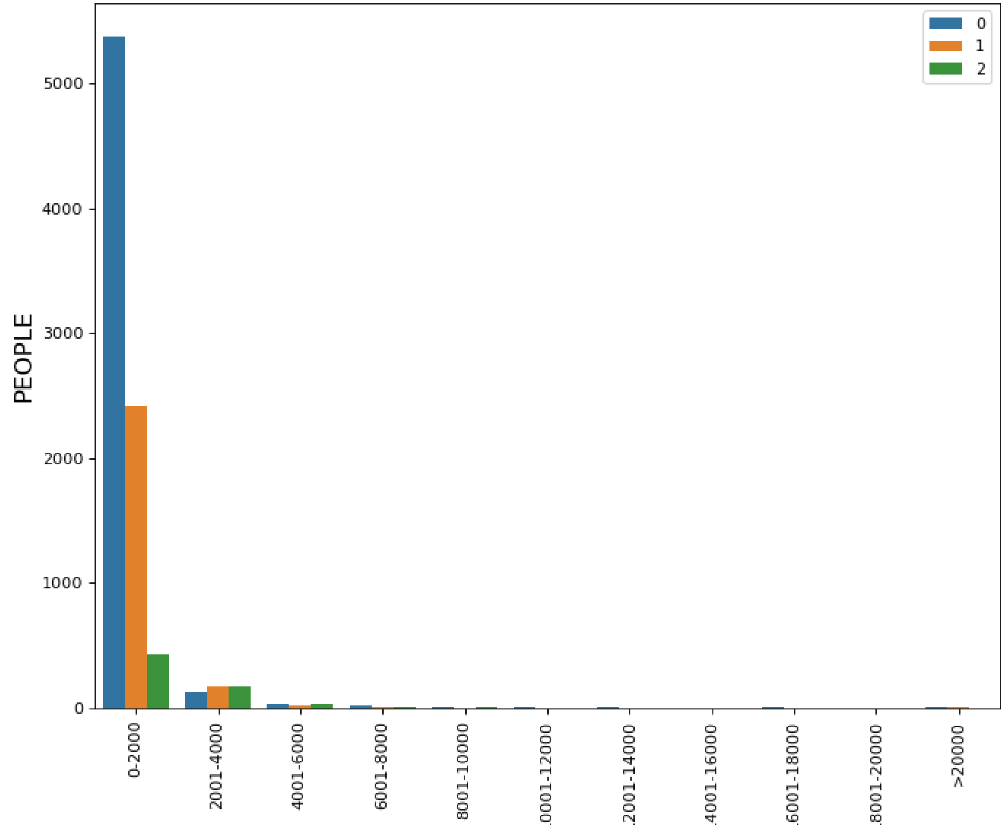


- SEGMENT 2 customer's AVERAGE PAYMENTS are 13622.30
- SEGMENT 1 customer's AVERAGE PAYMENTS are 3216.83
- SEGMENT 0 customer's AVERAGE PAYMENTS are 1346.86

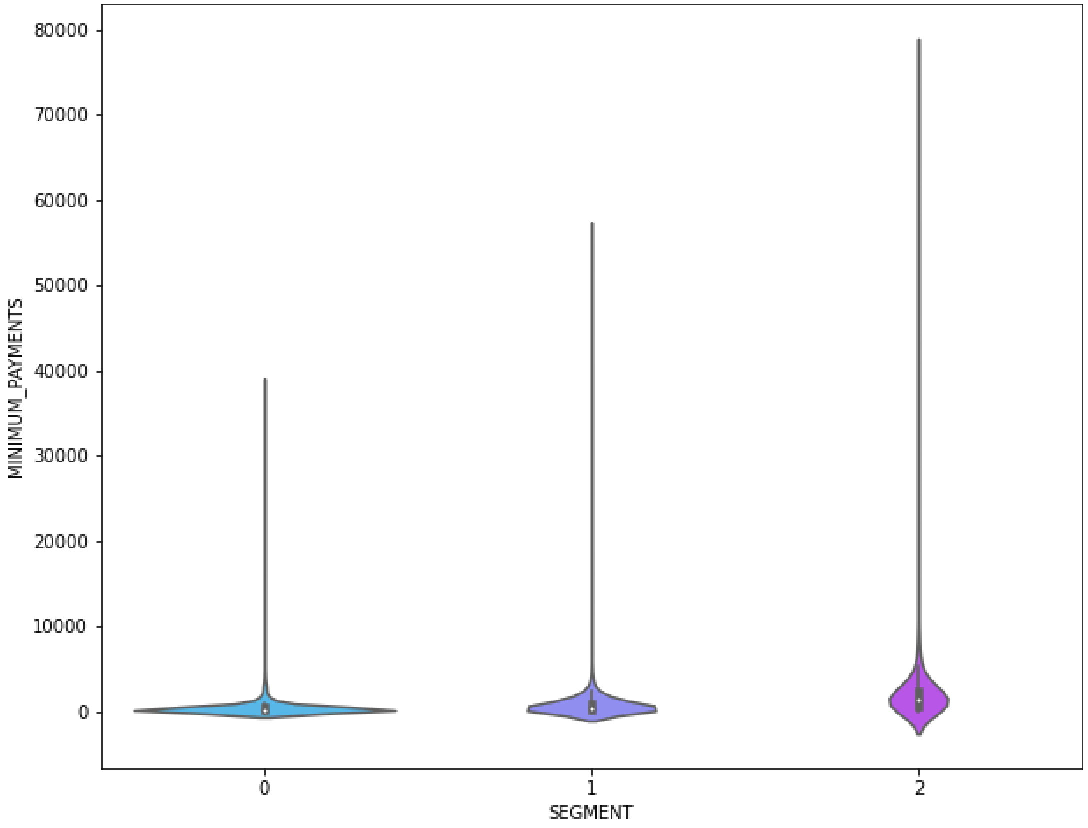
- SEGMENT 2 customer's PAYMENTS RANGE is 0 - >20001 dollars
- SEGMENT 1 customer's PAYMENTS RANGE is 0 - >20001 dollars
- SEGMENT 0 customer's PAYMENTS RANGE is 0 - >20001 dollars

Exploratory Data Analysis

MINIMUM_PAYMENTS_GROUP VS SEGMENT COUNT



MINIMUM_PAYMENTS DATA DISTRIBUTION BY SEGMENT



- Segment 2 customer's AVERAGE MINIMUM PAYMENTS are 2631.86
- Segment 1 customer's AVERAGE MINIMUM PAYMENTS are 1718.09
- Segment 0 customer's AVERAGE MINIMUM PAYMENTS are 532.50

- SEGMENT 2 customer's MINIMUM_PAYMENTS RANGE is 0 - >20001 dollars
- SEGMENT 1 customer's MINIMUM_PAYMENTS RANGE is 0 - >20001 dollars
- SEGMENT 0 customer's MINIMUM_PAYMENTS RANGE is 0 - >20001 dollars

Exploratory Data Analysis

EDA Summary

- **BALANCE** has low impact to SEGMENT
- **PURCHASES** has low impact to SEGMENT
- **ONEOFF_PURCHASES** has low impact to SEGMENT
- **INSTALLMENT_PURCHASES** has low impact to SEGMENT
- **CASH_ADVANCE_PURCHASES** has low impact to SEGMENT
- **CREDIT_LIMIT** has significant impact to SEGMENT
- **PAYMENTS** has low impact to SEGMENT
- **MINIMUM_PAYMENTS** has low impact to SEGMENT
- Customer SEGMENTATION influenced by many Features

EDA Recommendation

- Based on our analysis, we recommend to use **all features for Machine Learning**

Machine Learning

Since our dataset have 3 SEGMENT/**Multiclass** where:

- **SEGMENT 0**: LOW CUSTOMERS
- **SEGMENT 1**: MEDIUM CUSTOMERS
- **SEGMENT 2**: HIGH CUSTOMERS

With SEGMENT Composition below:

SEGMENT	Composition (%)
0	62.8
1	29.66
2	7.54

We will focus to obtain Machine Learning Model with the best **Accuracy Score**

Machine Learning Algorithms

- For this model we will **use all features**, because from EDA results, the customer segmentation is affected by all features from dataset

Random Forest

Logistic Regression

KNN Classifier

	Accuracy Score (Test)	Accuracy Score (Train)
Random Forest Base Model	0.99	1
Random Forest Tuned	0.93	0.94
Logistic Regression Base Model	0.92	0.91
KNN Base Model	0.98	0.99

Random Forest result accuracy is already **good**, in other hand, this model is categorized as **Strong Learner** model which causing the model might be only memorizing the data, and not learning the pattern. So we want to decrease accuracy score to get a **Good Learner** and get more suitable confusion matrix through Hyper Parameter Tuning.

Machine Learning Summary

Random Forest Tuned Classification Report

	precision	recall	f1-score	support
0	1.00	0.94	0.97	1124
1	0.82	0.99	0.90	531
2	1.00	0.63	0.77	135
accuracy			0.93	1790
macro avg	0.94	0.85	0.88	1790
weighted avg	0.94	0.93	0.93	1790

Random Forest Tuned Confusion Matrix

	Pred 0	Pred 1	Pred 2
Akt 0	1059	65	0
Akt 1	4	527	0
Akt 2	0	50	85

Machine Learning Summary

- From the initial machine learning modelling, there are **no overfit result** on all over model algorithm
- We suggest to use **Random Forest Tuned**, because after analysis it has the best **accuracy score 93%** (not so high) with the **most suitable confusion matrix**
- How this model will help bank company?
    - This model will allow bank marketing team to **take actions** on **identified** as "customer segment", furthermore the development of these model **should contribute** to **bank revenue management**.
    - These prediction models **enable marketing teams to mitigate profit loss** derived from **customer churn** caused **by unsuitable marketing program**

MACHINE LEARNING RECOMMENDATION:

- This Machine Learning could be used for **customer segmentation** based on their **credit card usage behaviour**
- The result from this project could be used by marketing team to **offer suitable product** for **customers** based on **their segmentation** which is predict through **Machine Learning Model**

6. Project Recommendation, Business Impact, & Future Research

Project Recommendation

Detail Product Suggestion:

- **PERFORMANCE CREDIT CARD** DETAIL PRODUCT:
  - GET REWARDS with Monthly **Minimum Purchases 500 dollars**
  - **CREDIT_LIMIT: 5000 dollars**

- **BUSINESS CASH CREDIT CARD** DETAIL PRODUCT:
  - GET REWARDS with Monthly **Minimum Purchases 1200 dollars**
  - **CREDIT_LIMIT: 13000 dollars**

- **BUSINESS UNLIMITED CREDIT CARD** DETAIL PRODUCT:
  - GET REWARDS with Monthly **Minimum Purchases 3500 dollars**
  - **CREDIT_LIMIT: 30000 dollars**

We Suggest:

- Offer **LOW CUSTOMERS SEGMENT** with **PERFORMANCE CREDIT CARD**
- Offer **MEDIUM CUSTOMERS SEGMENT** with **BUSINESS CASH CREDIT CARD**
- Offer **HIGH CUSTOMERS SEGMENT** with **BUSINESS UNLIMITED CREDIT CARD**

Business Impact

From our research (https://www.statista.com/statistics/816735/customer-churn-rate-by-industry-us/), current **credit card churn is about 25%.** With the help of our Clustering and Multiclass Machine Learning Modeling, we simulate that credit card churn **will drop into 7%.**

Assuming number of customers and lost per customers as down below:

	VALUE
Number of Credit Card Customer Simulation	1,000,000
Cost per Customer Lost Simulation (In Dollar)	80

Attached below is **rough calculation** Lost Customer Cost without Machine Learning vs Lost Customer Cost with Machine Learning

	Percentage Churn	Number of People Churn	Lost Customer Cost Calculation
Current Credit Churn Situation (Without Machine Learning)	25%	250,000.00	20,000,000.00
Credit Churn Simulation Using Machine Learning (Accuracy 93%)	7.00%	70,000.00	5,600,000.00

14,400,000 dollars

Using our Multiclass Machine Learning Modelling, our company could save money around **14,400,000 dollars!!!**

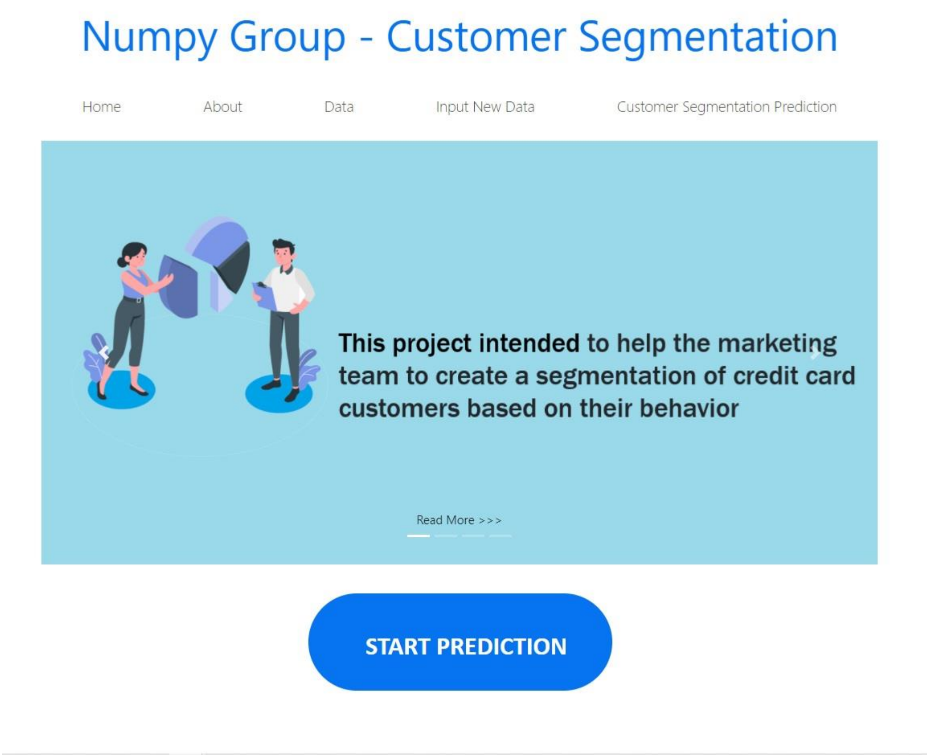
Future Works

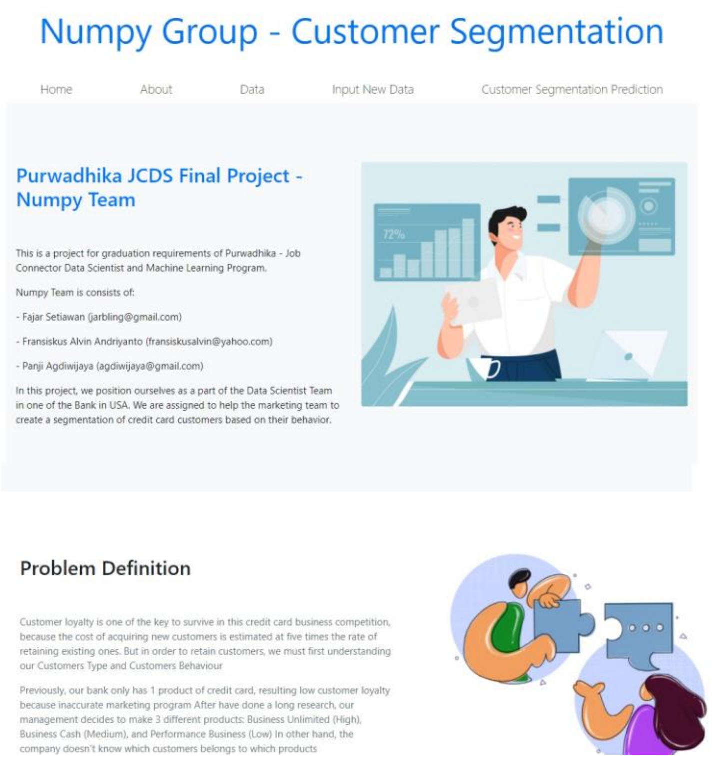
For further research, Customer Behaviour (**Payment history**, **Length of credit history**, **new credits**, **Variety of credit products** (installment loans, finance company accounts, mortgage loans and so on)), could be included into the dataset in hope to improve the models and measure the importance of these features

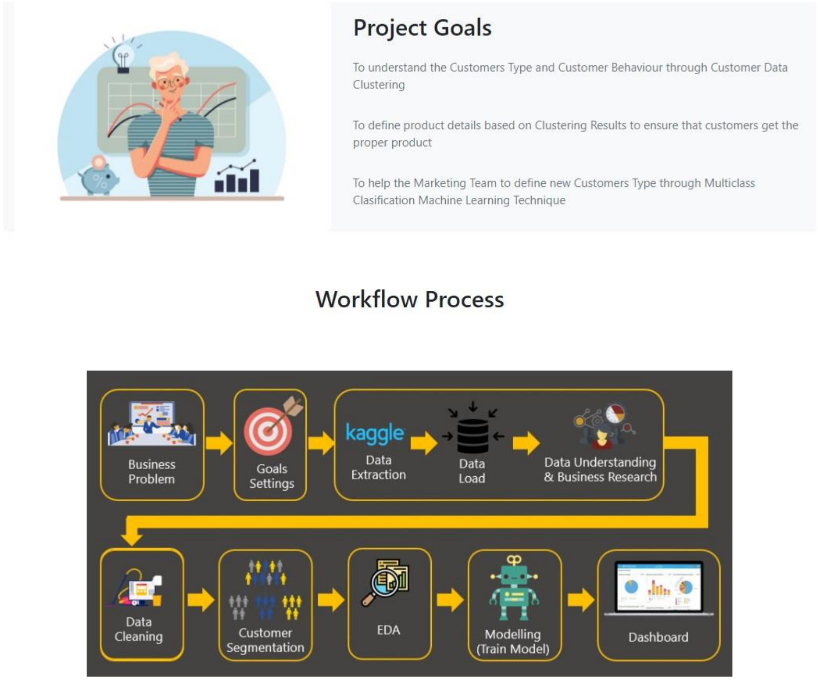
7. Model Deployment (Flask Dashboard)

Model Deployment (Flask Dashboard)

Preview

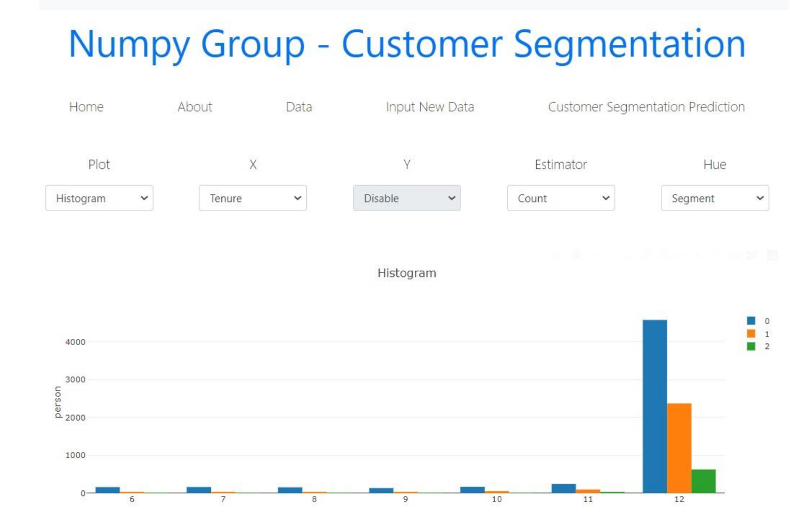


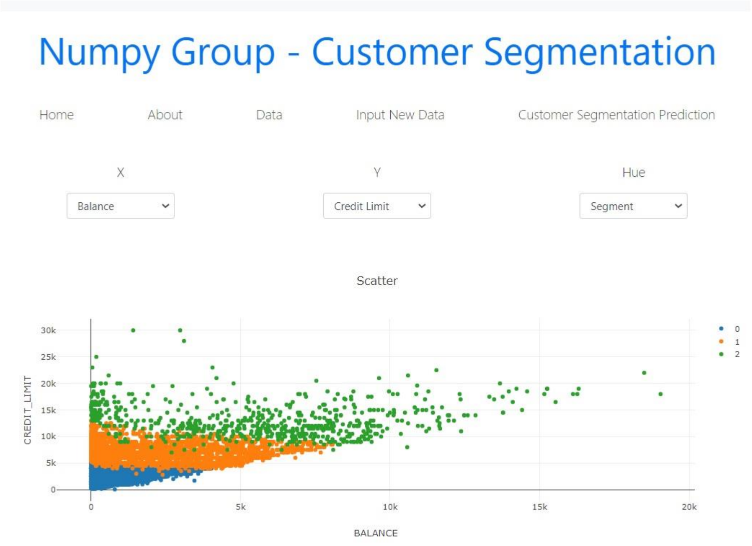


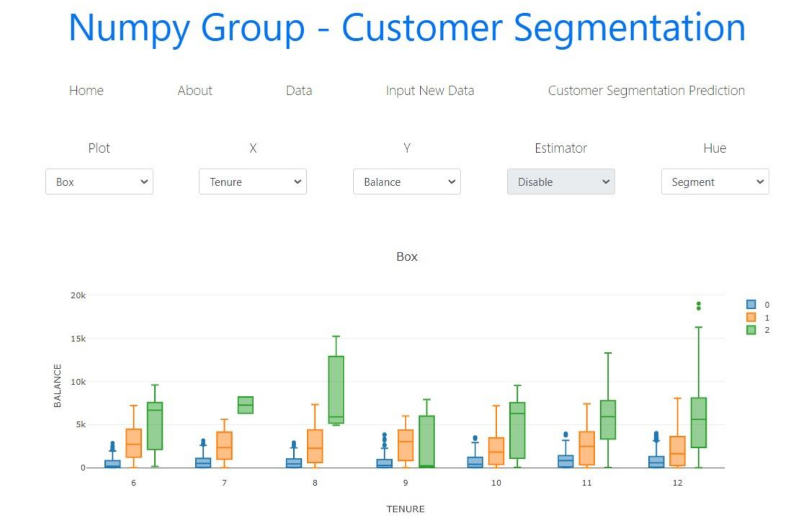


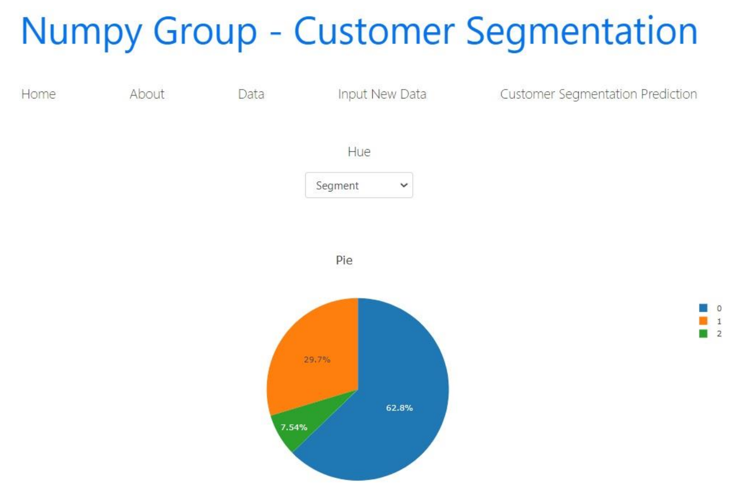
Model Deployment (Flask Dashboard)

Preview









Model Deployment (Flask Dashboard)

Preview

THANK YOU