# CREDIT CARD CUSTOMER SEGMENTATION ANALYSIS

UNSUPERVISED LEARNING

CAPSTONE

# OVERVIEW OF THE STUDY

- Unsupervised Learning is the most interesting branches of Machine Learning.
- The dataset chosen for this Unsupervised Learning project is Credit Card Dataset from Kaggle
- The main goal of this case is to develop a customer segmentation to define marketing strategy.
- The sample Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months.
- The dataset has18 behavioral variables
- This dataset is also available from Kaggle and the link is given below

https://www.kaggle.com/arjunbhasin2013/ccdata

# OVERVIEW OF THE DATASET

- The dataset of credit card contains **8950 rows** and **18 columns.** The following are the variables in the dataset
- **Attributes list :** They are based on customer behavior namely customer id, balance, balance frequency, purchases, purchases frequency, one-off purchases, one-off purchases frequency, cash advance, cash advance frequency, installments purchases, installments purchases frequency, cash advance transactions, purchases transactions, credit limit, payments, minimum payments, percent full payment, tenure
- A basic exploratory data analysis reveals that there were two columns with missing values **minimum payments and credit limit** and they were replaced with mean values
- **Customer ID** values were unique and the column was **removed** from further analysis as there was no useful information

# EXPLORATORY DATA ANALYSIS

- Most customers have a Tenure of 12 years

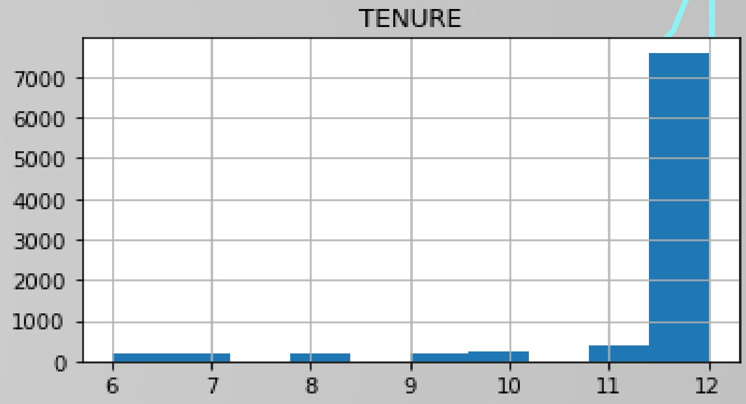
- The average **Balance** is about 1600 but a balance of 0 is more common

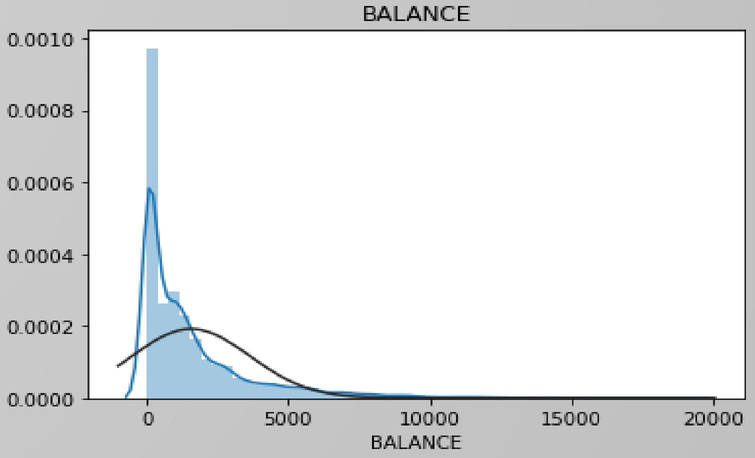
- **Balance Frequency** for most customers is updated frequently ~ 1

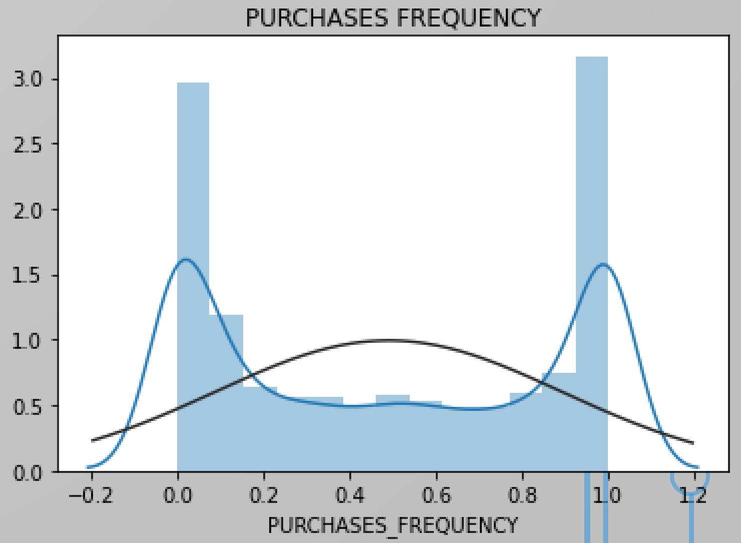
- Very small number of customers pay their balance in full **prc full payment** ~ 0

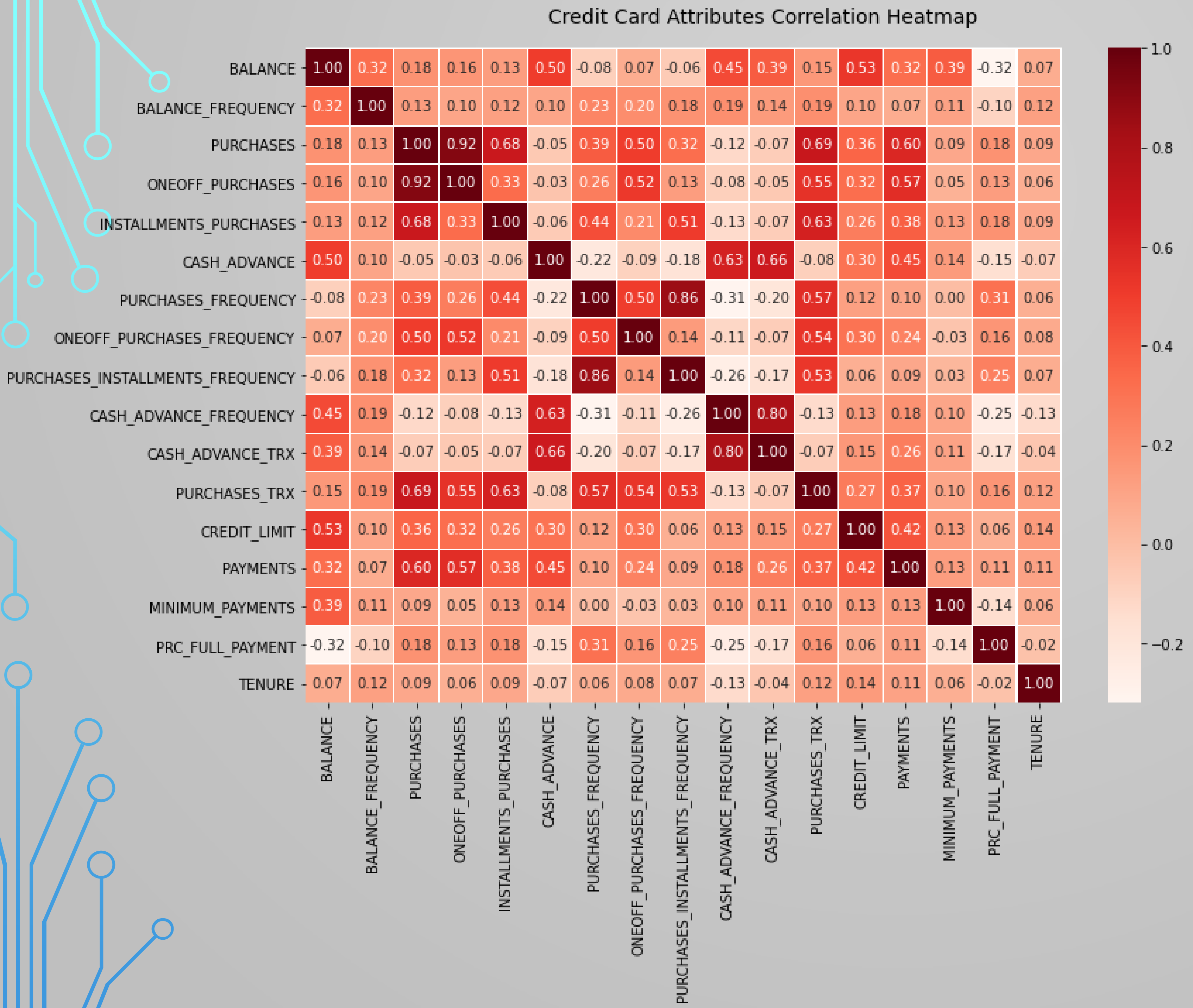
- For **Purchases frequency**, there are two distinct group of customers

- Also, from **Oneoff purchases frequency** and **Purchases installment frequency** most users don't do one off purchases or installment purchases frequently









Credit Card Attributes Correlation Heatmap

|  | BALANCE | BALANCE_FREQUENCY | PURCHASES | ONEOFF_PURCHASES | INSTALLMENTS_PURCHASES | CASH_ADVANCE | PURCHASES_FREQUENCY | ONEOFF_PURCHASES_FREQUENCY | PURCHASES_INSTALLMENTS_FREQUENCY | CASH_ADVANCE_FREQUENCY | CASH_ADVANCE_TRX | PURCHASES_TRX | CREDIT_LIMIT | PAYMENTS | MINIMUM_PAYMENTS | PRC_FULL_PAYMENT | TENURE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BALANCE | 1.00 | 0.32 | 0.18 | 0.16 | 0.13 | 0.50 | -0.08 | 0.07 | -0.06 | 0.45 | 0.39 | 0.15 | 0.53 | 0.32 | 0.39 | -0.32 | 0.07 |
| BALANCE_FREQUENCY | 0.32 | 1.00 | 0.13 | 0.10 | 0.12 | 0.10 | 0.23 | 0.20 | 0.18 | 0.19 | 0.14 | 0.19 | 0.10 | 0.07 | 0.11 | -0.10 | 0.12 |
| PURCHASES | 0.18 | 0.13 | 1.00 | 0.92 | 0.68 | -0.05 | 0.39 | 0.50 | 0.32 | -0.12 | -0.07 | 0.69 | 0.36 | 0.60 | 0.09 | 0.18 | 0.09 |
| ONEOFF_PURCHASES | 0.16 | 0.10 | 0.92 | 1.00 | 0.33 | -0.03 | 0.26 | 0.52 | 0.13 | -0.08 | -0.05 | 0.55 | 0.32 | 0.57 | 0.05 | 0.13 | 0.06 |
| INSTALLMENTS_PURCHASES | 0.13 | 0.12 | 0.68 | 0.33 | 1.00 | -0.06 | 0.44 | 0.21 | 0.51 | -0.13 | -0.07 | 0.63 | 0.26 | 0.38 | 0.13 | 0.18 | 0.09 |
| CASH_ADVANCE | 0.50 | 0.10 | -0.05 | -0.03 | -0.06 | 1.00 | -0.22 | -0.09 | -0.18 | 0.63 | 0.66 | -0.08 | 0.30 | 0.45 | 0.14 | -0.15 | -0.07 |
| PURCHASES_FREQUENCY | -0.08 | 0.23 | 0.39 | 0.26 | 0.44 | -0.22 | 1.00 | 0.50 | 0.86 | -0.31 | -0.20 | 0.57 | 0.12 | 0.10 | 0.00 | 0.31 | 0.06 |
| ONEOFF_PURCHASES_FREQUENCY | 0.07 | 0.20 | 0.50 | 0.52 | 0.21 | -0.09 | 0.50 | 1.00 | 0.14 | -0.11 | -0.07 | 0.54 | 0.30 | 0.24 | -0.03 | 0.16 | 0.08 |
| PURCHASES_INSTALLMENTS_FREQUENCY | -0.06 | 0.18 | 0.32 | 0.13 | 0.51 | -0.18 | 0.86 | 0.14 | 1.00 | -0.26 | -0.17 | 0.53 | 0.06 | 0.09 | 0.03 | 0.25 | 0.07 |
| CASH_ADVANCE_FREQUENCY | 0.45 | 0.19 | -0.12 | -0.08 | -0.13 | 0.63 | -0.31 | -0.11 | -0.26 | 1.00 | 0.80 | -0.13 | 0.13 | 0.18 | 0.10 | -0.25 | -0.13 |
| CASH_ADVANCE_TRX | 0.39 | 0.14 | -0.07 | -0.05 | -0.07 | 0.66 | -0.20 | -0.07 | -0.17 | 0.80 | 1.00 | -0.07 | 0.15 | 0.26 | 0.11 | -0.17 | -0.04 |
| PURCHASES_TRX | 0.15 | 0.19 | 0.69 | 0.55 | 0.63 | -0.08 | 0.57 | 0.54 | 0.53 | -0.13 | -0.07 | 1.00 | 0.27 | 0.37 | 0.10 | 0.16 | 0.12 |
| CREDIT_LIMIT | 0.53 | 0.10 | 0.36 | 0.32 | 0.26 | 0.30 | 0.12 | 0.30 | 0.06 | 0.13 | 0.15 | 0.27 | 1.00 | 0.42 | 0.13 | 0.06 | 0.14 |
| PAYMENTS | 0.32 | 0.07 | 0.60 | 0.57 | 0.38 | 0.45 | 0.10 | 0.24 | 0.09 | 0.18 | 0.26 | 0.37 | 0.42 | 1.00 | 0.13 | 0.11 | 0.11 |
| MINIMUM_PAYMENTS | 0.39 | 0.11 | 0.09 | 0.05 | 0.13 | 0.14 | 0.00 | -0.03 | 0.03 | 0.10 | 0.11 | 0.10 | 0.13 | 0.13 | 1.00 | -0.14 | 0.06 |
| PRC_FULL_PAYMENT | -0.32 | -0.10 | 0.18 | 0.13 | 0.18 | -0.15 | 0.31 | 0.16 | 0.25 | -0.25 | -0.17 | 0.16 | 0.06 | 0.11 | -0.14 | 1.00 | -0.02 |
| TENURE | 0.07 | 0.12 | 0.09 | 0.06 | 0.09 | -0.07 | 0.06 | 0.08 | 0.07 | -0.13 | -0.04 | 0.12 | 0.14 | 0.11 | 0.06 | -0.02 | 1.00 |

## Correlation Heat Map

1. Balance has a higher level of correlation with Cash Advance, Cash Advance Frequency and Credit Limit

2. Purchases has a higher level of correlation with One off Purchases, Purchases TRX, Installment Purchases and One Off purchases frequency

3. Payments variable has a high correletion with Purchases and one off Purchases and to some extent with Credit limit

4. Tenure has a negative correlation with Cash Advance and Cash Advance Frequency variables

# AGENDA

- Clustering Methods for Analysis

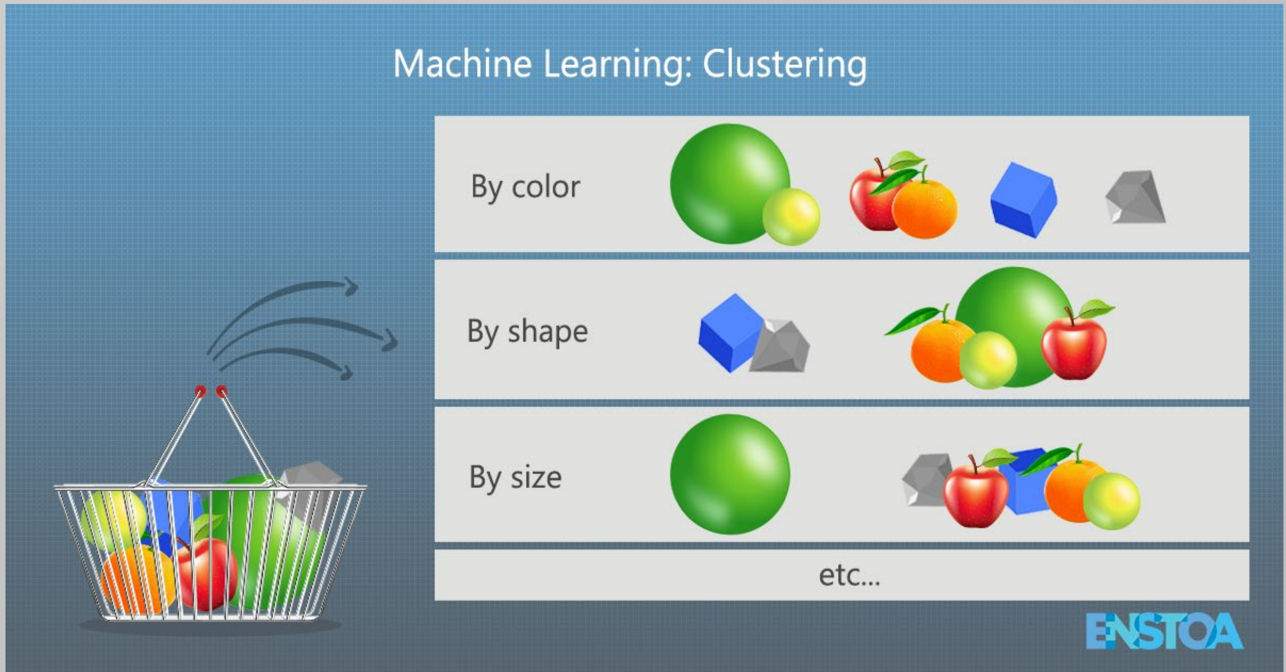
  *(A). KMeans*

  *(B). Agglomerative Hierarchical*

  *(C). DBSCAN*

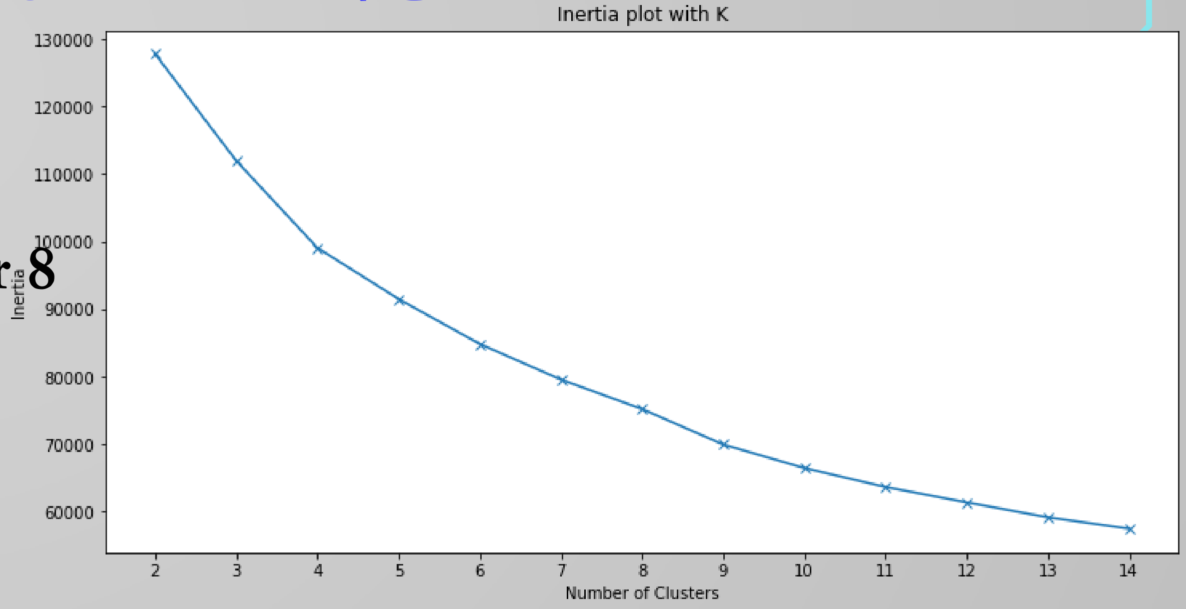
  *(D). Gaussian Mixture*

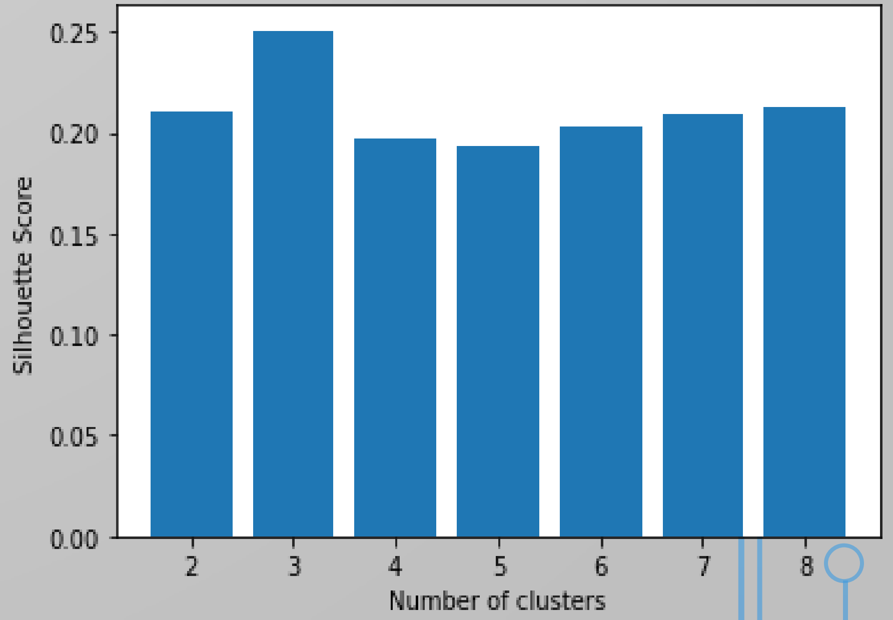
- Evaluation and Model Selection
- Visualization of Clusters
- Interpretation of Clusters
- Conclusions and Recommendations



# KMEANS CLUSTERING

- First the data was standardized for analysis
- We got the **Inertia plot** with **elbow at 4 or 8**
- From the **Silhouette scores**, the **3-cluster solution** is more suitable for the data
- The tuned **hyperparameters** for KMeans are **n_clusters = 3, n_init=1000, max_iter=400, init='k-means++'**
- The **silhouette score** Kmeans is **0.2509**



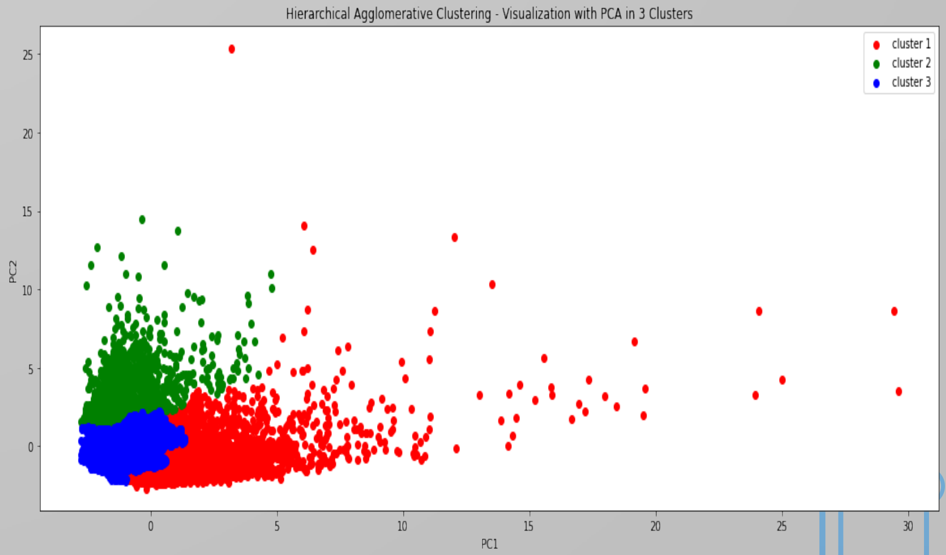


# AGGLOMERATIVE CLUSTERING

- Data was standardized for analysis
- Agglomerative Clustering with various linkage methods "ward", "average", "complete"
- The tuned hyperparameters for Agglomerative Clustering are

**n_clusters = 3, affinity='euclidean’, linkage="ward"**

- The **silhouette score** Agglomerative Clustering is **0.1731**
- PCA was used to visualize the results of clustering

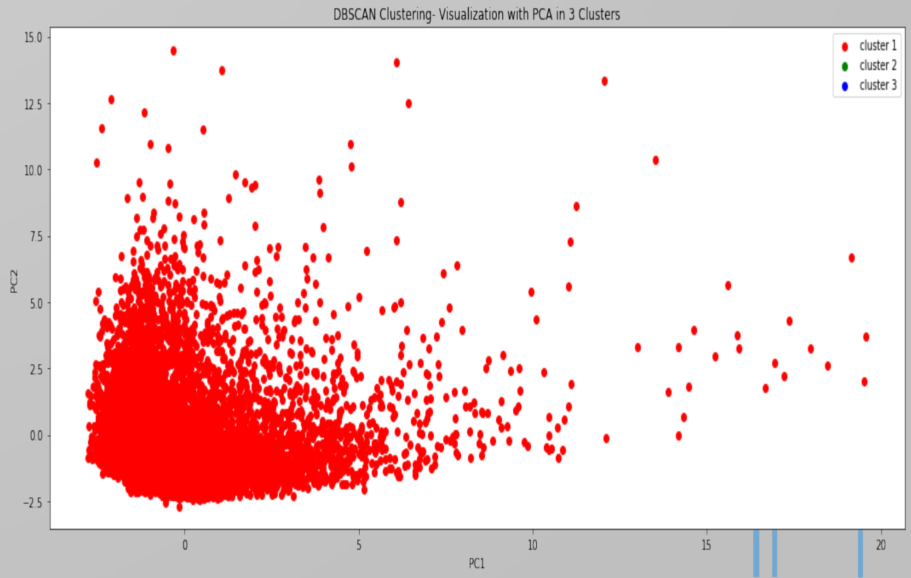


# DBSCAN

- Data was standardized for analysis
- DBSCAN was conducted with various values of eps [0.01,0.1,1,2,3,4,5,6,7,8,9,10], min_samples and metric
- The tuned hyperparameters for DBSCAN are eps=4, metric="euclidean"

and min_samples=3

- The silhouette score **DBSCAN is 0.6239**
- PCA was used to visualize the results of clustering
- DBSCAN results for clusters have noise



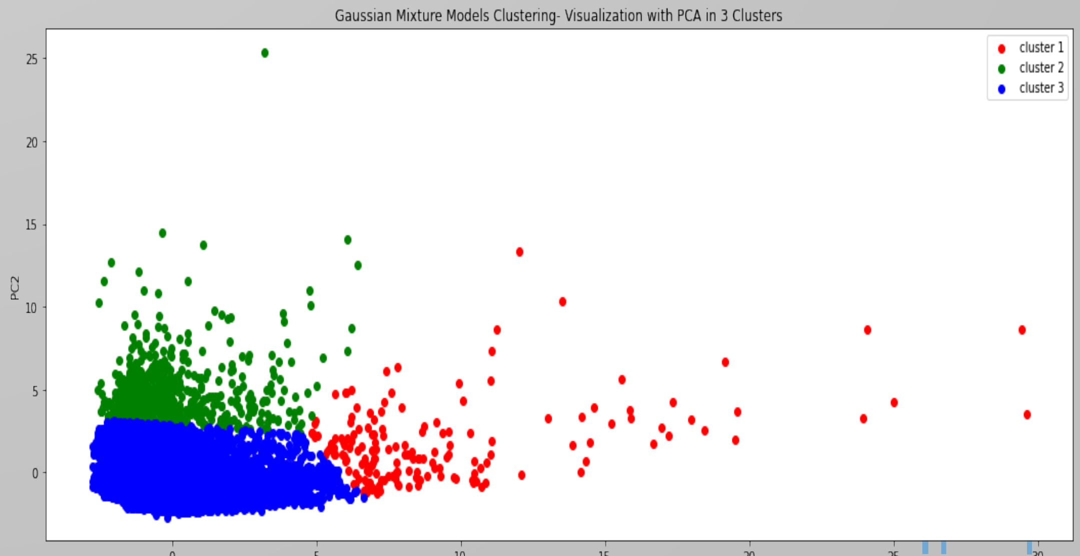
# GAUSSIAN MIXTURE MODELS

- Data was standardized for analysis
- Gaussian Mixture Models analysis
- was conducted with various covariance types
- The tuned hyperparameters for Gaussian Mixture Models are n_clusters = 3

affinity='euclidean’, linkage="ward"

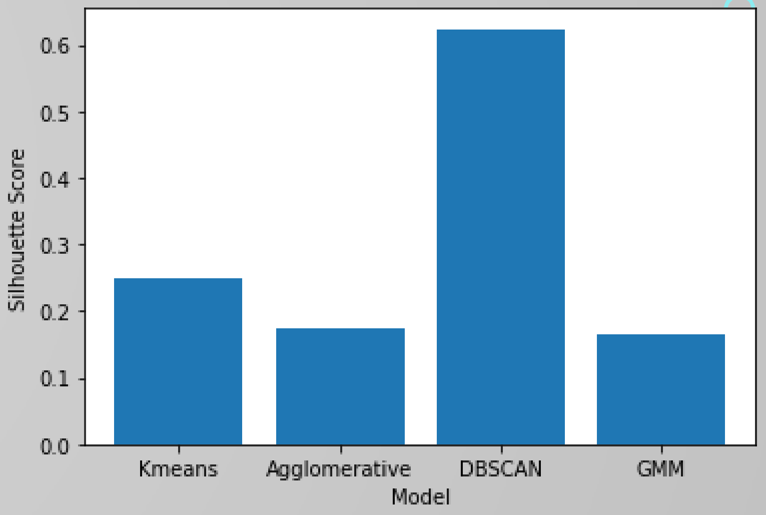
- The silhouette score **Gaussian Mixture Models is 0.1657**
- PCA was used to visualize the results of

clustering

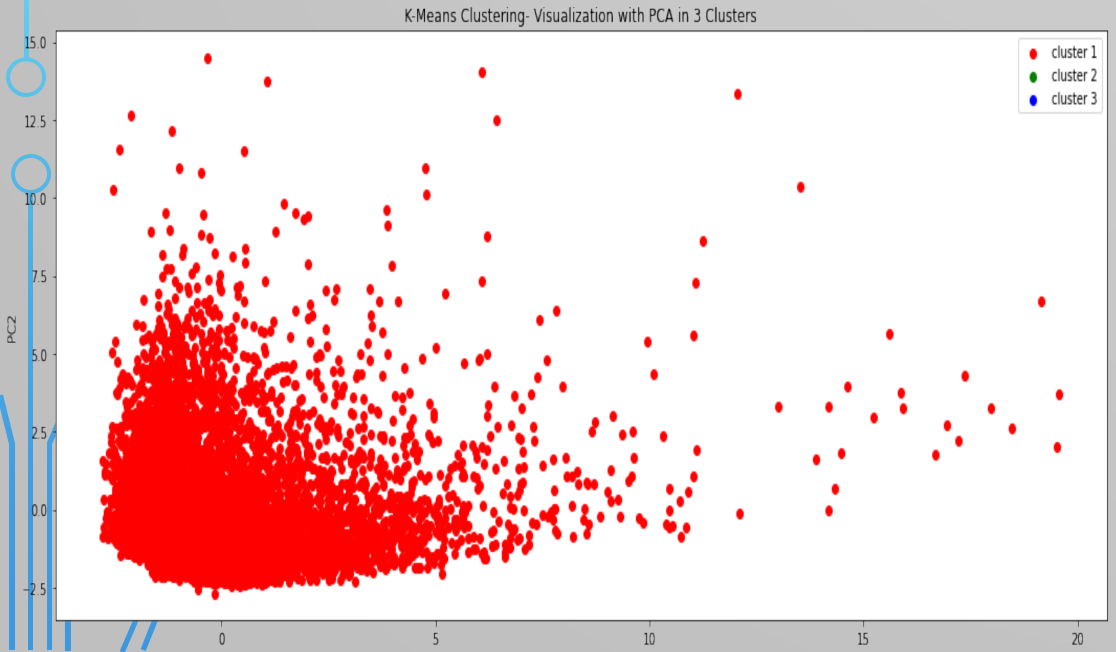


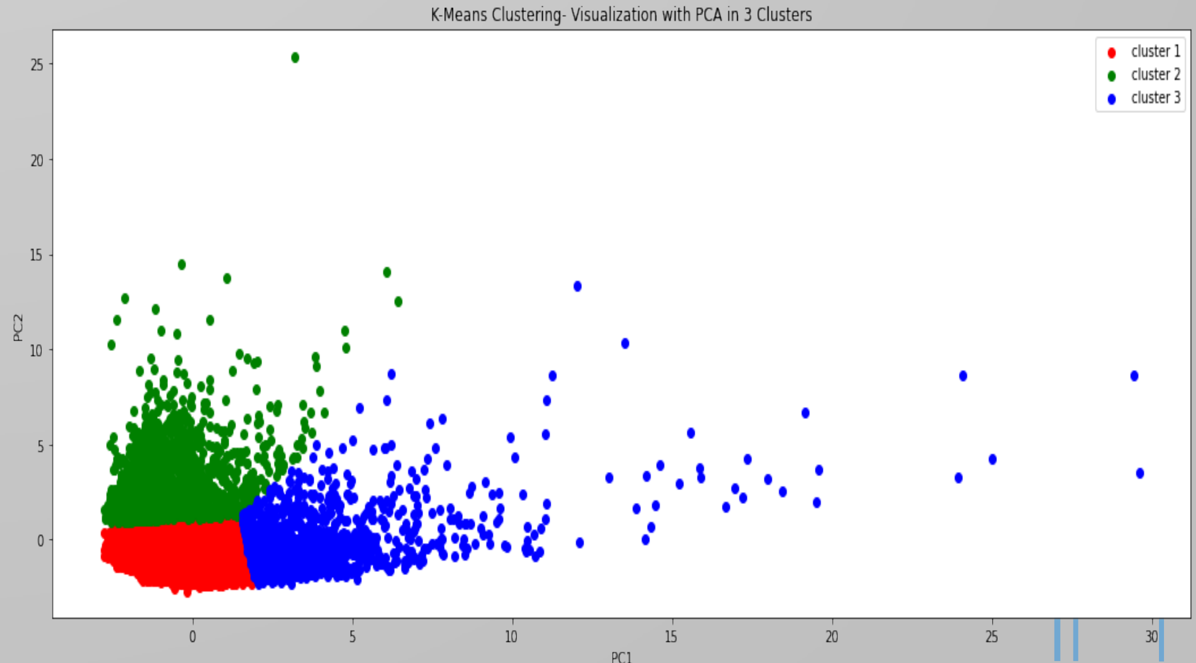
# FINAL MODEL SELECTION

- For the final selection of models, we compared the silhouette scores of all the models
- From the silhouette scores, DBSCAN is the best model
- But PCA reveals a different story. DBSCAN clusters are not distinct and most values are clustered as noise.



## Kmeans Clustering





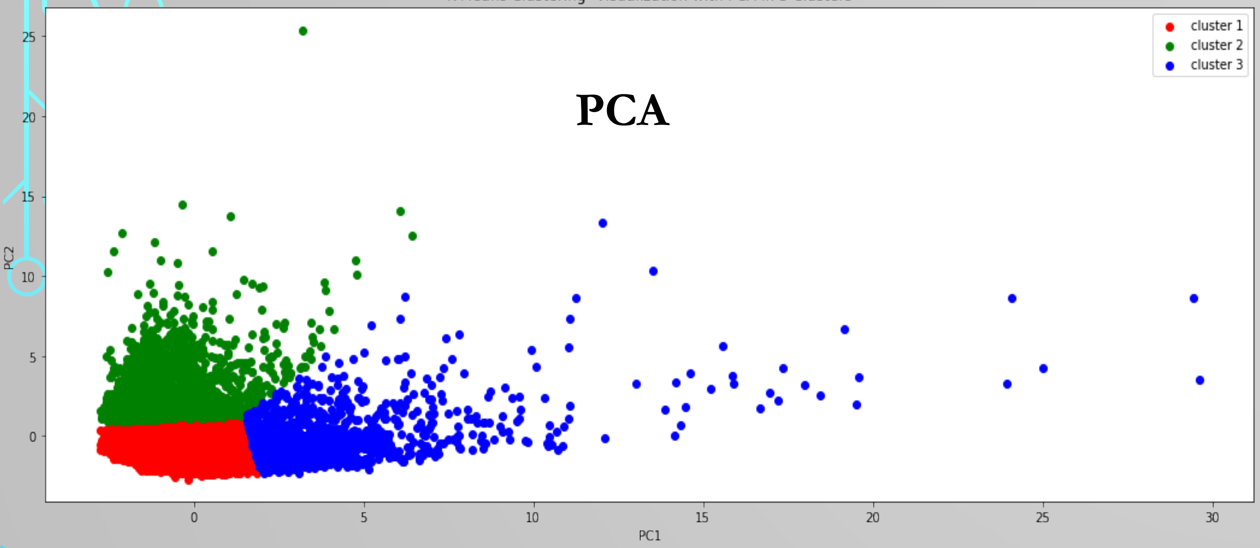
# INTERPRETATION OF CLUSTERS

- Boxplots were obtained for various input attributes and they were **grouped by clusters**
- The results of cluster analysis is given in the table

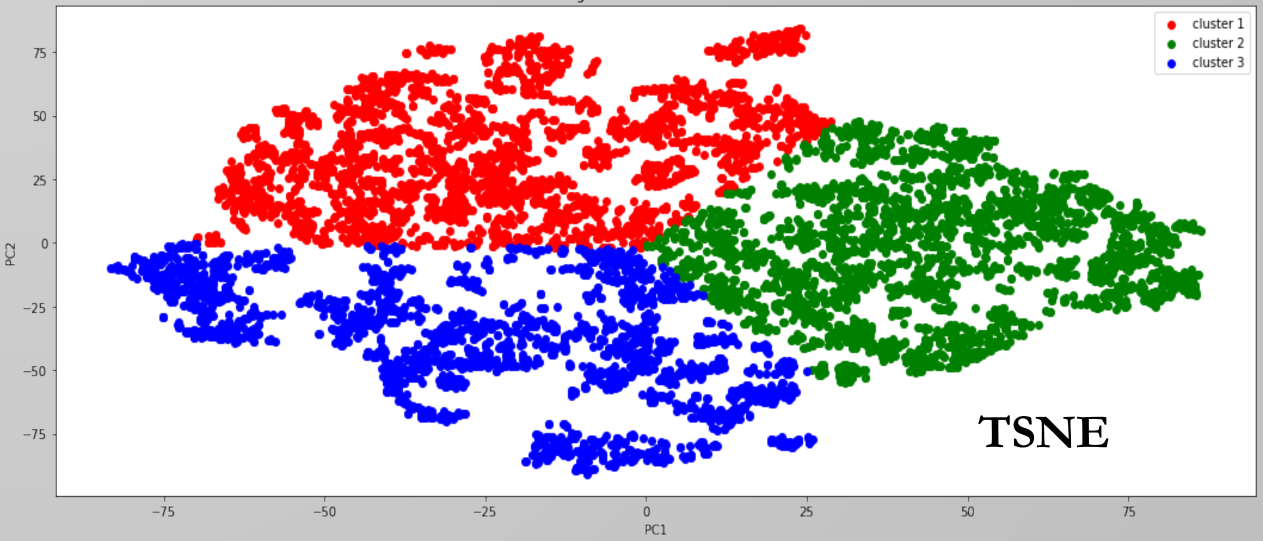
| | **cluster 0** | **cluster 1** | **cluster 2** |
|---|---|---|---|
| **BALANCE** | low | **med** | **high** |
| **BALANCE_FREQUENCY** | med | med | med |
| **PURCHASES** | **high** | **low** | **low** |
| **PURCHASES_FREQUENCY** | **high** | **med** | low- |
| **CASH_ADVANCE** | low | low | high |
| **MINIMUM_PAYMENTS** | **low** | low | med |
| **CREDIT_LIMIT** | **high** | **low** | **high-** |
| **PAYMENTS** | **high** | low | **med** |
| LABELS | Active User | Cautious Spender | Average Joe |

# VISUALIZATION OF CLUSTERS

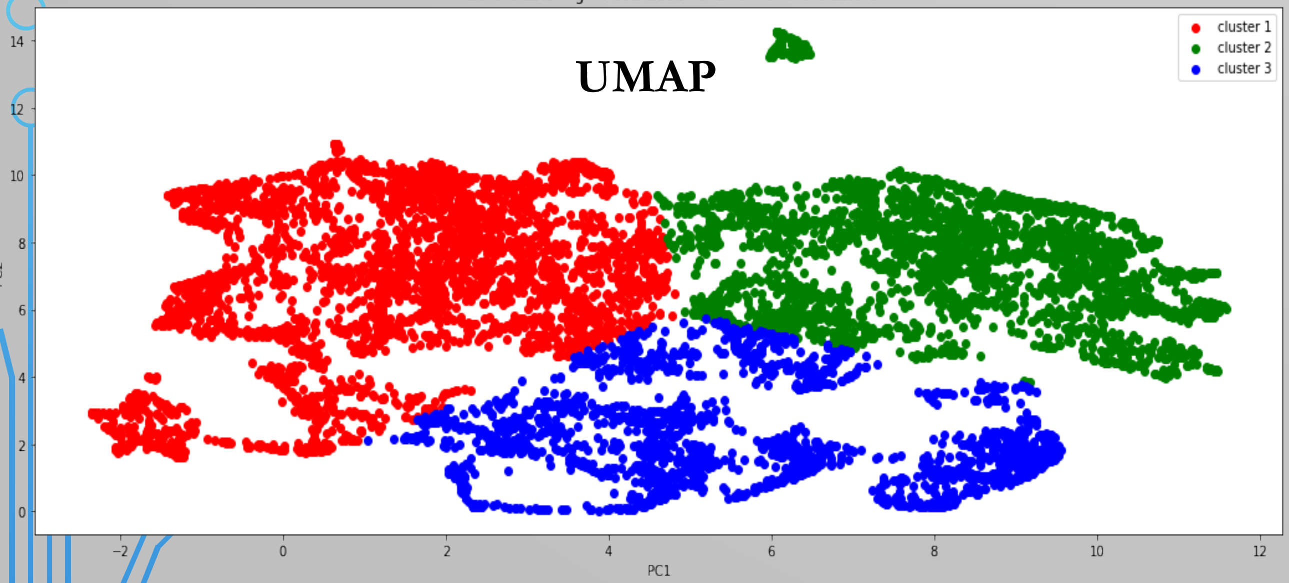
K-Means Clustering- Visualization with PCA in 3 Clusters



K-Means Clustering - Visualization with t-SNE in 3 Clusters



K-Means Clustering - Visualization with UMAP in 3 Clusters



# CONCLUSIONS AND RECOMMENDATIONS

- In this project, we have performed data preprocessing, looked at various clustering metrics (inertias, silhouette scores), experimented with various Clustering algorithms (KMeans Clustering, Agglomerative Hierarchical Clustering, DBSCAN Clustering, and Gaussian Mixture Clustering), data visualizations
- Finally, we have segmented the customers into three smaller groups: **the Active Users**, **the Cautious Spenders**, and **the Average Joe**
- The **Active User** is the group, marketing team should focus on as they are ideal group
- The **Cautious Spenders** is the difficult to strategize but should not be neglected
- The **Average Joe** has healthy finances and low debts. While encouraging these people to use credit cards more is necessary for the company's profit, business ethics and social responsibility should also be considered.
- **Advanced data preparation:** Building an enriched customer profile by deriving "intelligent" **KPIs** like limit usage(balance to credit limit ratio), average amount per purchase can be used for to improve accuracies
- Using **dimensionality reduction** not just for visualization but actual reduction of features could potentially improve the cluster analysis and hence the accuracy associated

# THANK YOU!