UNSUPERVISED LEARNING

Credit Card Transaction Data for Clustering, Customer Segmentation and Analysis

Sri Nithya K (14423326) **Rithwik G (14423558)**

Abstract

The data used for this study is the credit card transaction details of 9000 unique customers which includes details like balance, purchases, purchase frequency, installment purchases, installments purchase frequency, payments, one-off purchases, one-off purchase frequency, tenure etc. The analysis done on this dataset after performing unsupervised learning techniques like principal component analysis and clustering algorithms like K-means are presented with effective visualizations and tabulation of results.

Introduction

Unsupervised learning is a branch of machine learning which deals with unlabeled data i.e., each individual input does not have a labelled output. In this project, we used an unlabeled data set of credit card transaction details of nearly 9000 customers and performed clustering algorithm called K-means.

To perform K-means clustering, a dimension reduction algorithm called principal component analysis (PCA) is performed and compared with another algorithm called Random Projection. For K-means algorithms, we need to specify the number of clusters that we want to divide the data into. Obtaining optimal number of clusters can be done using methods like Elbow method or Silhouette score method.

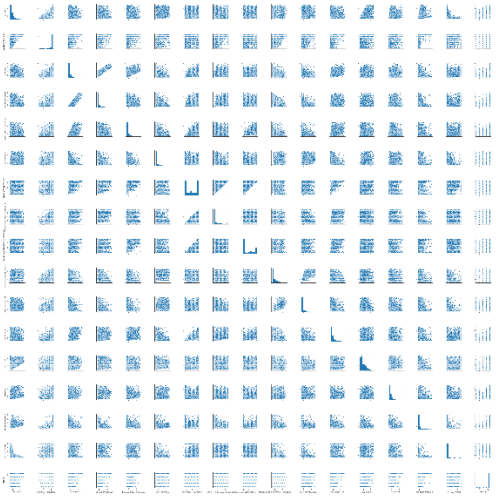
Based on the clusters obtained, segmentation of customers and their analysis based on various combination of features is done to gain useful insights that can be used by the credit card service provider for marketing purposes. Effective visualizations can be used by the credit card users to make conscious decisions regarding the expenditure and payments to keep their credit score in check.

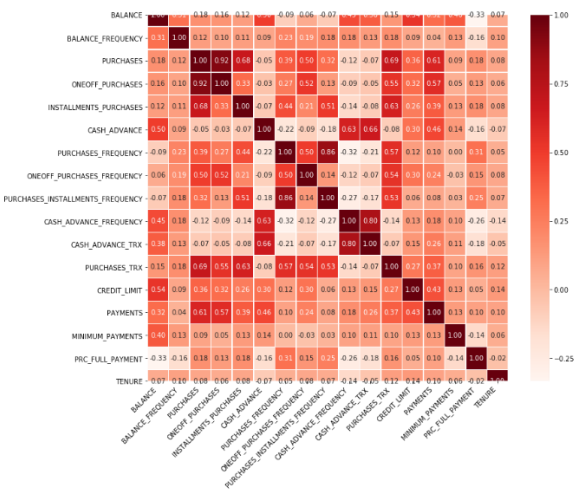
Data Pre-Processing

The first step in this project is to nullify the inconsistencies that are present in the dataset. The dataset used in this project is collected from Kaggle. Kaggle is a platform for data science which is used by many data scientists to design models for solving specific problems. The datasets available on this platform can be raw and must be pre-processed before using it for further computations. As a part of it, we identified the entrees that have N/A or null values. The result was significantly less and thus we deleted these rows to have consistent data. Then we performed normalization using Standard Scaler to gain computational efficiency.

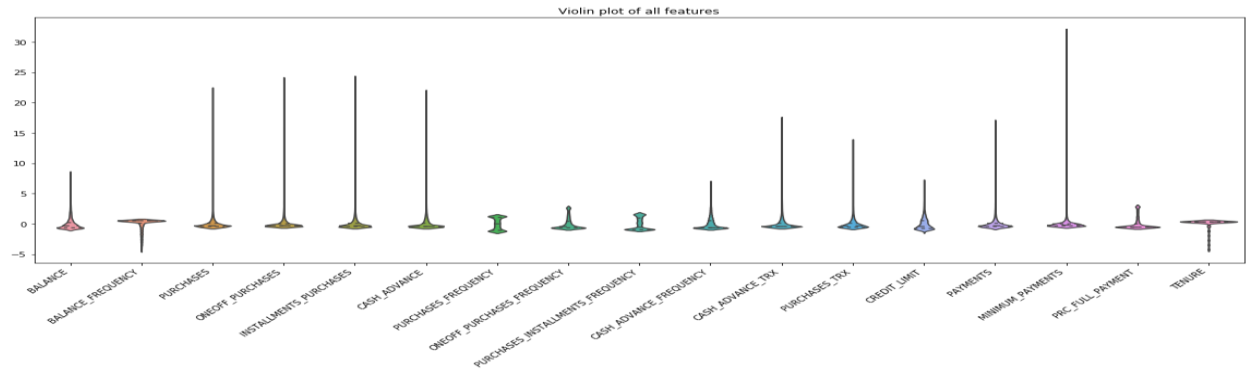
Exploratory Data Analysis

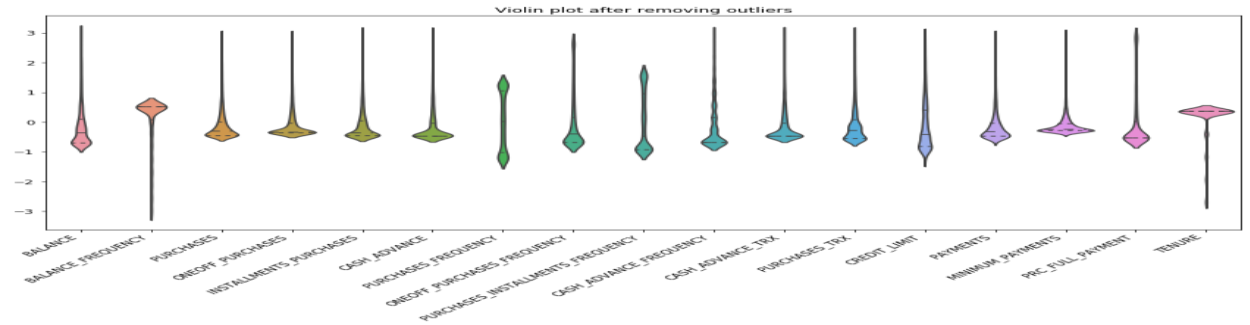
To obtain any visible patterns in the data, we created pair plots for all the features. For easy recognition of patterns or relations between any two features, we created a heatmap of the correlation matrix. We observed that the features “PURCHASES” and “ONEOFF_PURCHASES” are highly correlated whereas the features “PURCHASES_FREQUENCY” and “CASH_ADVANCE_FREQUENCY” have the least correlation.





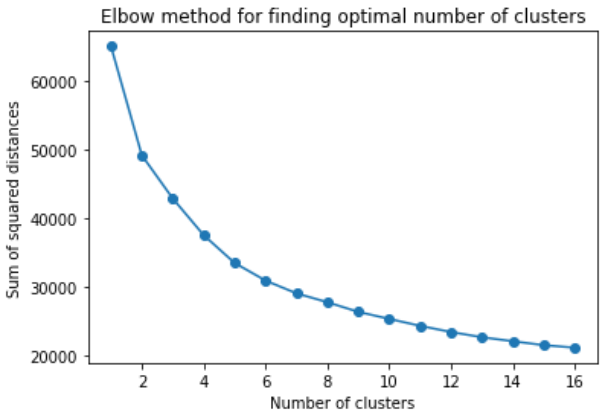
Outliers in a dataset may be responsible for unnecessary noise during computations. Removing outliers can be used to avoid any human error while collecting data. We used violin plots to visualize the outliers and used z-score to remove outliers. Z-score is a measure that describes a value’s relation with the means of its group values. It is measured as a distance in terms of standard deviations. The violin plots before and after removing outliers are depicted as follows.

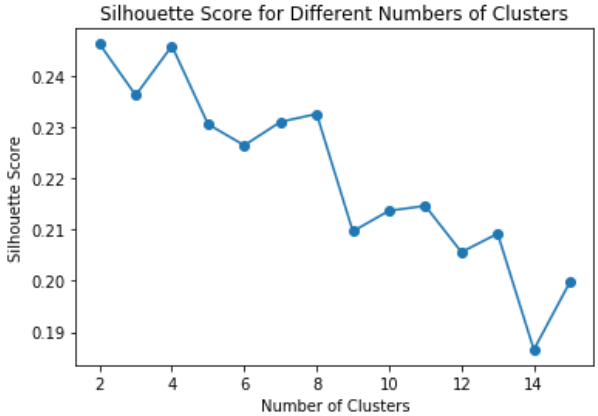




Elbow method and Silhouette score

To perform K-means clustering, obtaining the optimal number of clusters to perform the algorithm is a requisite. We used two methods, Elbow method and Silhouette score method to obtain it.



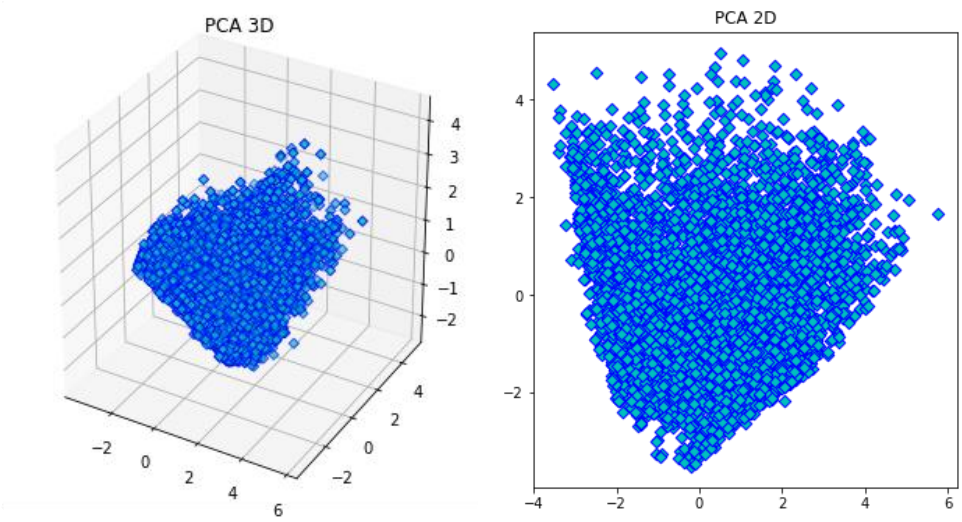


Elbow method calculates the within cluster sum of squares i.e., sum of the squares of distance from one point to its center for all values in the specified range of clusters and visualizes a curve. The optimal number of clusters can be obtained at the point where the decrease in the curve becomes significantly less.

Silhouette score method is another statistical measure that tells how well the point is matched with its assigned cluster. The higher the silhouette score for the cluster, higher is the match to the cluster of the point.

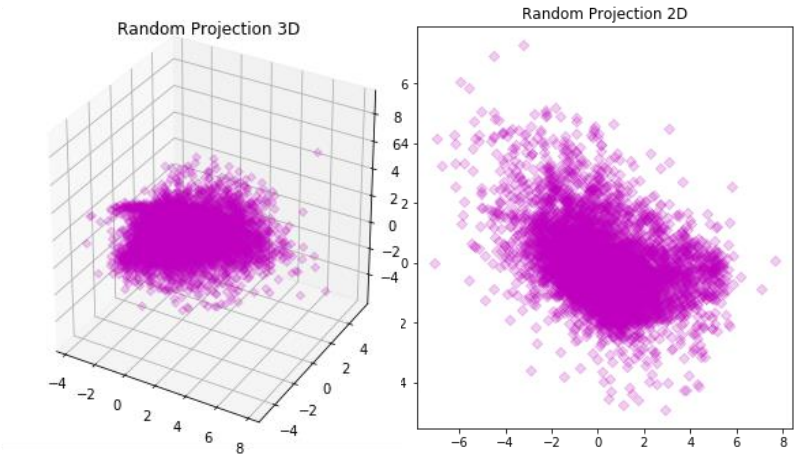
Both these methods have displayed that the optimal number of clusters for this dataset is 4.

Principal Component Analysis



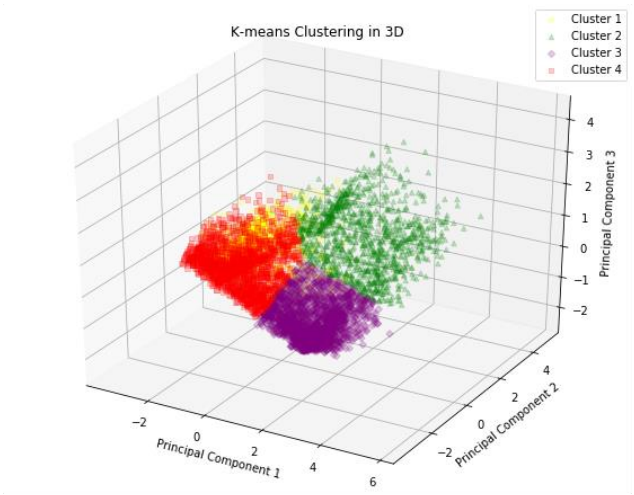
Principal Component Analysis (PCA) is a dimensionality reduction method used to transform the original features of the data set into a new set of uncorrelated features by retaining most of the variance among the features. The visualizations in 3D and 2D after dimensionality reduction is as shown above.

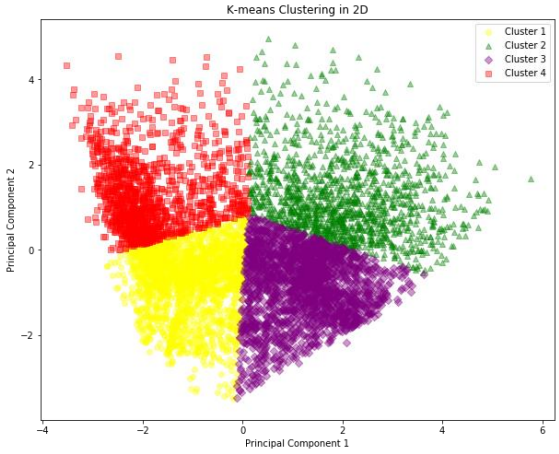
Random Projection is another dimensionality reduction method. It is a technique used to project high dimensional data into a low dimensional subspace using a random matrix. It provides good approximation with reduced computational costs. One projection in 3D and 2D using random projections is as follows.



K-means Clustering

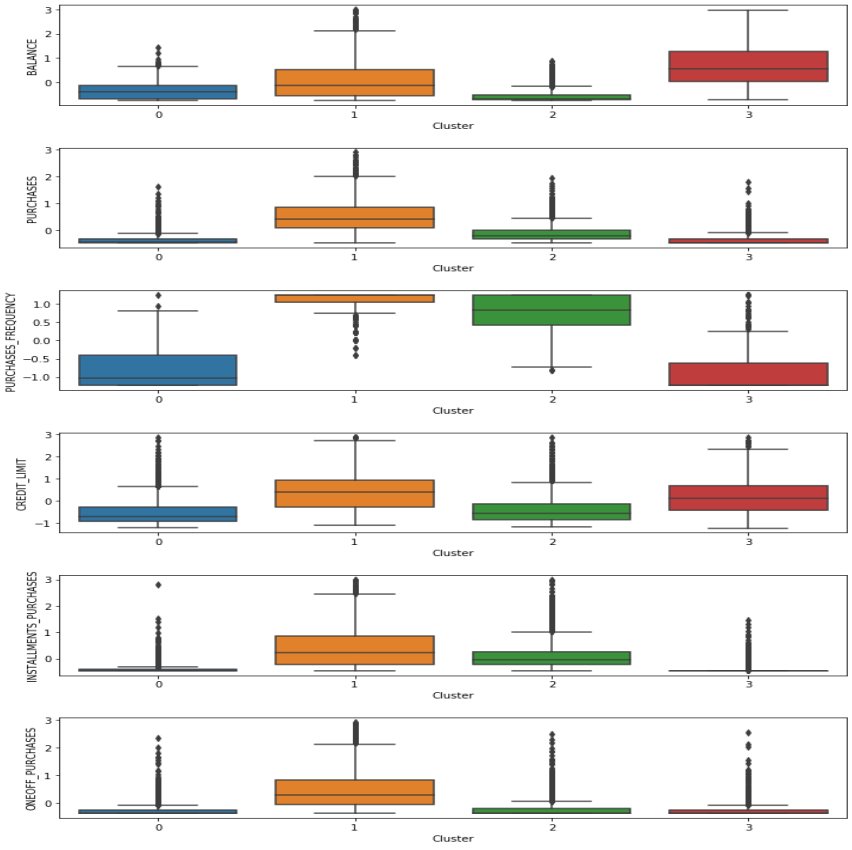
K-means clustering partitions the data into non overlapping subsets or groups or clusters using a defined number of clusters. Mostly, K-means clustering is used in customer segmentation tasks, anomaly detection tasks etc. After getting the optimal number of clusters as 4 from the elbow method and silhouette score method, K-means clustering is performed on the data. The following representations are the clusters in both 3D and 2D format for the purpose of effective visualization of separation between clusters.





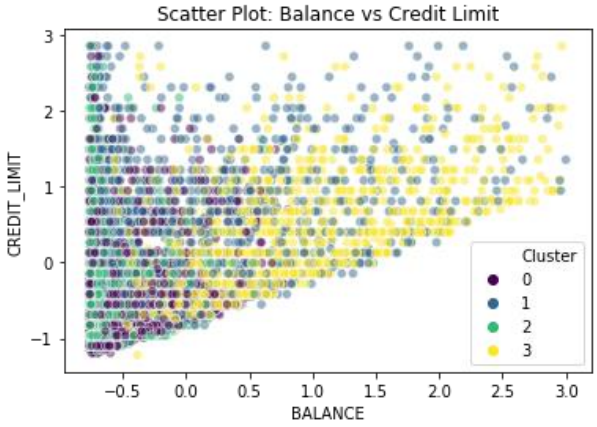
Segmentation and Analysis

For the next step of the project i.e., customer segmentation, we have created box plots for a few highly correlated features (balance, purchases, purchase frequency, credit limit, installment purchases, one-off purchases) of the data set to observe for any visible patterns and tabulated the observations and analysis.



	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Balance	Low	Low/Medium	Low	Medium/High
Purchases	Low	Low/Medium	Low/Medium	Low
Purchase frequency	Low/Medium	High	High	Low/Medium
Credit limit	Low	High	Medium	High
Installment purchase	Low	High	Medium	Low
Oneoff purchases	Low	Medium	Low	Low

Another segmentation based on the features balance and credit limit are as follows. The customer segmentation based on this profile has the following observations. Cluster 0, 1, 2 consists of customers with low to medium balance whereas Cluster 3 consists of customers with medium to high balance.



Conclusion

The aim of this project is to provide the credit card service providers with useful insights of the usage of the specific company's credit card which will be helpful in market research purposes to further their business requirements. Also, the effective visualizations used in this project lets the public gain information on how their respective transactions are affecting their chances of a good credit score and other related aspects. It helps the customers make conscious decisions regarding their credit card usage.

The use of K-means clustering is an effective method in projects involving customer segmentation, anomaly detection tasks, healthcare services etc. Principal component analysis (PCA) is the most effective method in dimensionality reduction tasks with utmost accuracy in preserving the original variance among features of the data set.

Data Set: Credit Card Transaction Data Set