PROJECT REPORT

Black Friday Customer Purchase Behavior Against Different Products

Submitted towards partial fulfillment of the criteria for the award of
PGPDSE by **Great Lakes Institute of Management**

Submitted by:

Group No.2 of **Batch: June 2019**

Group Members:
Ritesh Chaudhury
Madasu Leena
Sowmya Madabushi
Abdul Khaleel Shaik
Karthikeyan Thallam

Mentor:
Dipanjan Goswami

CERTIFICATE OF COMPLETION

I hereby certify that the project titled **Black Friday Customer Purchase Behavior Against Different Products** was undertaken and completed under my supervision by Ritesh Chaudhury, Madasu Leena, Sowmya Madabushi, Abdul Khaleel Shaik, Karthikeyan Thallam of Post Graduate Program in Data Science and Engineering (PGPDSE).

Dipanjan Goswami

Date:10/11/2019

Place: Hyderabad

Declaration

I declare that the project entitled **Black Friday Customer Purchase Behavior Against Different Products** is a project work carried out by us under the supervision and guidance of Dipanjan Goswami for the award of degree PGPDSE, and this has not been previously submitted for the award of any Degree, Diploma or other similar title of any other University/ Institute.

Date: 10/11/2019
Place: Hyderabad

Group No.2:
Ritesh Chaudhury
Madasu Leena
Sowmya Madabushi
Abdul Khaleel Shaik
Karthikeyan Thallam

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our mentor Dipanjan Goswami for providing his invaluable guidance, comments and suggestions throughout the course of this project. We value the assistance of Great Learning, Hyderabad campus. Learning from their knowledge helped me to become passionate about my research topic.

We will be failing in our duty if each one of us don't express our gratitude for other team members, for the valuable contributions during course of this project.

Abstract

This documentation shows the exploratory analysis of the customer purchase behavior against different products, clustering and classification analysis to predict customer purchase behavior using the features available. The purpose of this study was to observe and analyze the costumer behaviors of Black Friday sales.

The data-set contains 550000 rows and 12 columns to explore. Exploratory analysis is done after the data cleansing by changing the variables to correct format. Uni-variate analysis for every feature followed by bi-variate analysis to see how each feature is related, affecting the purchase to find insights and understand the dependency of purchase with each is covered in the exploratory analysis. The insights that are expected and found are also discussed. Categorical variables are label encoded. The results show that data is perfectly divided into 3 clusters and labels as 0,1,2.After performing PCA Decision Tree and KNeighbors Classifier gives more accuracy than Logistic Regression and GaussianNB Classifiers. The results also shows that when the dataset is considered to predict purchases then RMSE value is 2687 and R square accuracy score is 0.71.

- Techniques used: Exploratory analysis, Clustering and Predictive Modelling
- Tools: Python and Tableau
- Domain: Retail

Table of Contents

I. INTRODUCTION

1.Problem statement:

The aim here is the store wants to know get an idea about customer purchase behavior against different products.

2.Data-set:

The dataset contains demographics (age, gender, marital status, city_type, stay_in_current_city), product details (product_id and product category) and Total purchase_amount from last month.

3.Shape:

55000 rows and 12 columns

II. LITERATURE

Columns:

Variable	Definition
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins
Occupation	Occupation (Masked)
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belongs to other category also (Masked)
Product_Category_3	Product may belongs to other category also (Masked)
Purchase	Purchase Amount (Target Variable)

III. DATA CLEANING

As a business grows and matures, the size, number, formats, and types of its data assets change along with it. Evolutions in payroll systems, new network hardware and software, emerging supply-chain technologies, and the like can all create the need to migrate, merge, and combine data from multiple sources. "Dirty" data — data that contains redundancies, includes duplicate records, is missing information, or has been otherwise corrupted in the process of being imported or merged — is one inevitable result. Data transformation, which involves “massaging” data to make its fields and formats confirm to those of its destination, can also be the source of hair pulling and sleepless nights. The art and science of handling these odious tasks is called "data cleansing."

The following changes have been done for better analysis, visualization and model building. The changes done for the required columns are as below:

1.Null-Value Imputation:

- We have various imputation Techniques by using inbuild Libraries in the python or we can also use basic imputation method which is mean, mode, median.
- But here in the dataset we are imputing the null values with 0 because Product_Category_2 , Product_Category_3 is dependent Product_Category_1 and so if we impute by mode it will be bias for other categories .

2.Dealing with Categorical Data:

Generally, there are many ways to encode the categorical columns like:

- Label encoding
- One Hot encoding
- But for this data we have lot of odd columns which have already labelled but there are some columns which have to be encoded.
- columns: - Age, Gender, City_Category
- For age column we can create different groups.
- 0-17----Teenage.
- 18-25---young.
- 26-35 and 36-45---median aged people.
- 46-50 and 51-55 and 55+---senior aged peoples.
- There is some cleaning which have to done on the "Stay_In_Current_City_Years" as there is a category which is "4+" we have to remove + from the columns as we can considering 4 and 4+ into one category.

3.Outlier Treatment: -

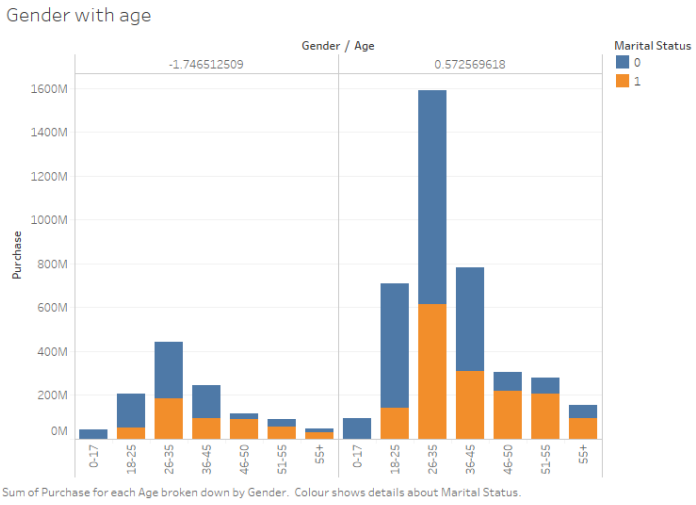
Outliers in the data are present in purchase column. The outliers from the data are not removed as their removal might affect the sales data i.e., the purchase which has made the highest monetary on removal might lead to a loss of information. Instead such values are being considered as extreme values.

IV. EXPLORATORY DATA ANALYSIS

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed insights about the Purchase Behavior:

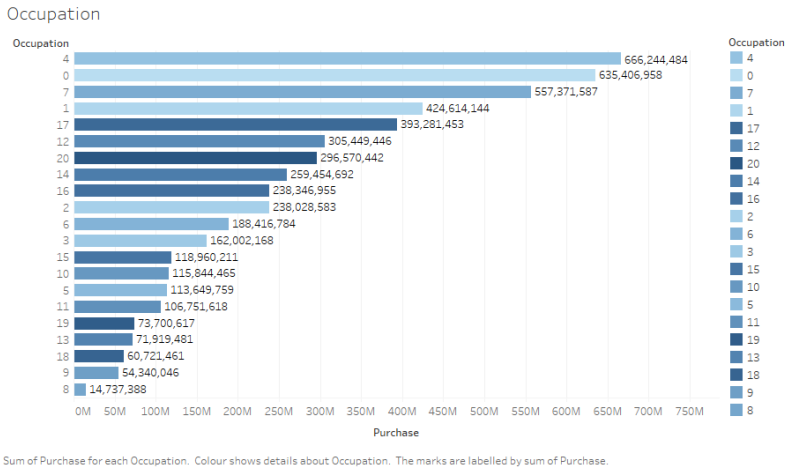
1.Gender vs Age with Marital status



AGE	GENDER	PURCHASE
0-17	F	42385978
	M	92527205
18-25	F	205475842
	M	708372833
26-35	F	442976233
	M	1588794345
36-45	F	243438963
	M	783130921
46-50	F	116706864
	M	304136539
51-55	F	89465997
	M	277633647
55+	F	45782765
	M	154984610

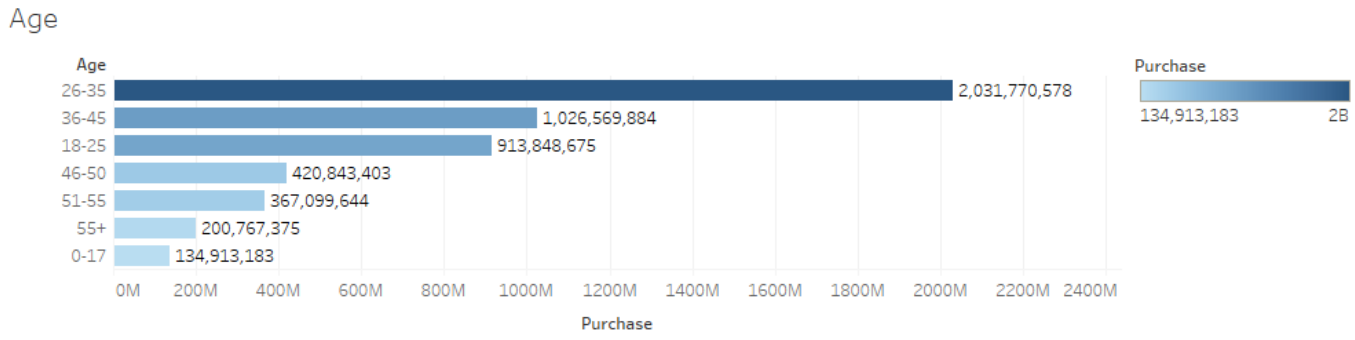
Here we can clearly see that age between 26-35 and mostly men are going with higher purchase. Most of the revenue contribution is done by customers between age 18 and 45.

2.Occupation & Purchase:



Occupation 4 is making the highest purchase in terms of monetary but based on this it cannot be concluded that the income of that particular occupation is high as nothing about the income is mentioned in the data.

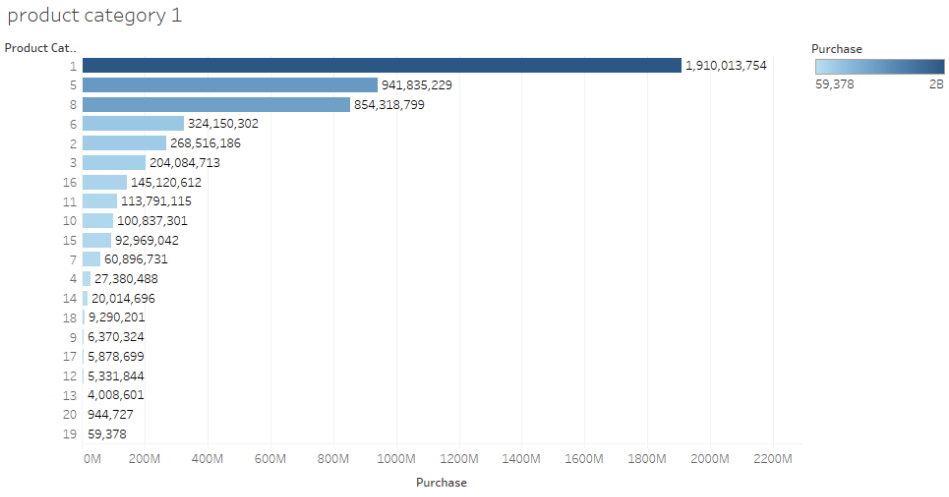
3.Age & Purchase



Sum of Purchase for each Age. Colour shows sum of Purchase. The marks are labelled by sum of Purchase.

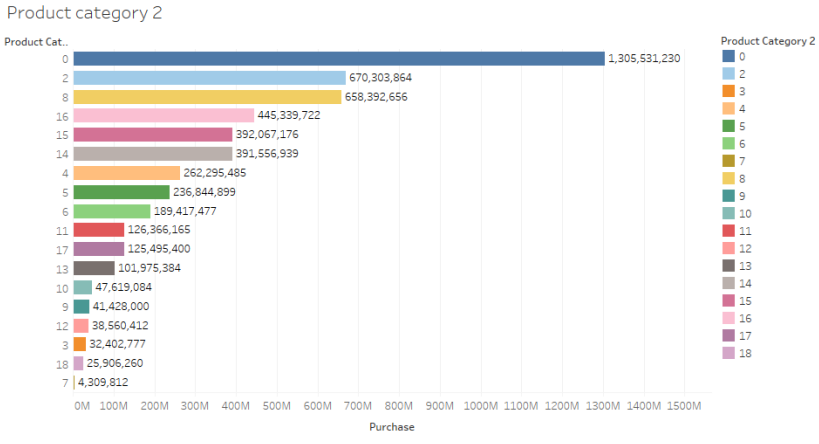
Here when purchase is compared with Age we can infer that age between 26-35 have higher purchase count.

4.Product Category 1 with Purchase



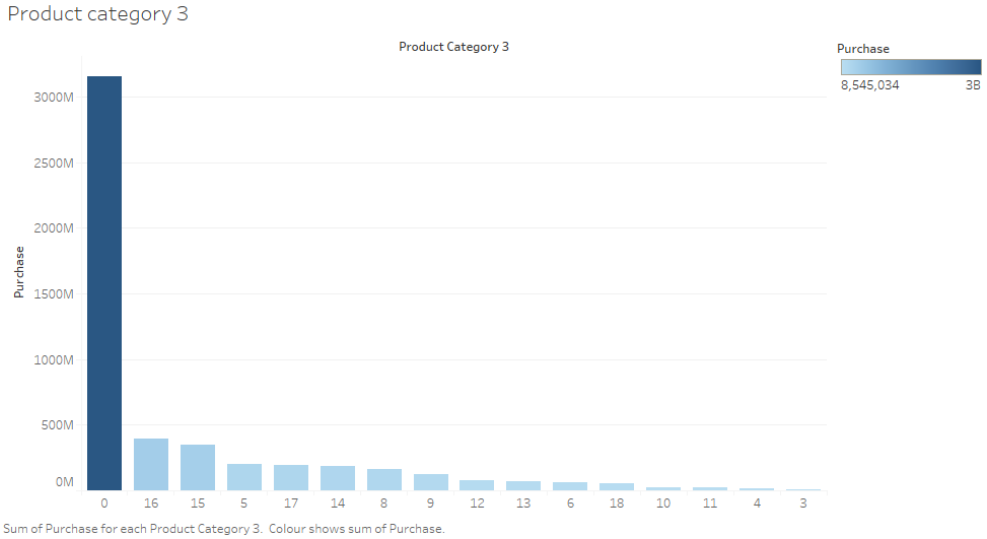
Sum of Purchase for each Product Category 1. Colour shows sum of Purchase. The marks are labelled by sum of Purchase.

5.Product Category with Purchase



Sum of Purchase for each Product Category 2. Colour shows details about Product Category 2. The marks are labelled by sum of Purchase.

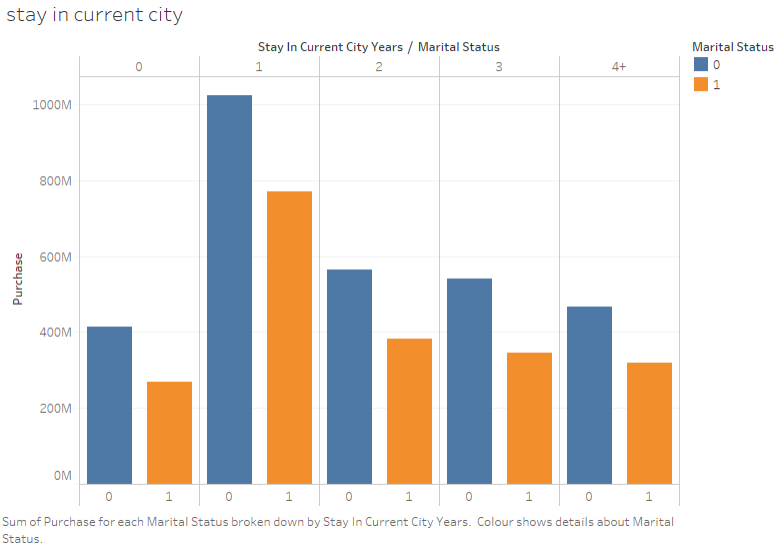
6.Product Category 3 with Purchase



The above 3 graphs show that:

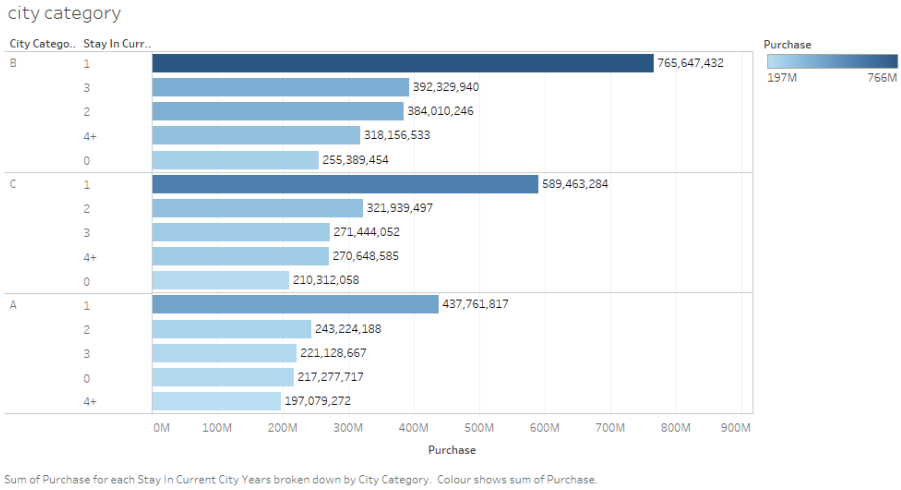
- Product_Category_1 the highest purchase is for category-1.
- Product_Category_2 the highest purchase is for category 2 as 0 is ignorable as we imputed it.
- Product_Category_3 the maximum purchase is for 0 which is to be ignored as it's a null value imputed with 0 so lets take next highest purchase for category 16 which is very less compared to Product_category_1 and Product_Category_2.

7.Stay in current city years & Marital status



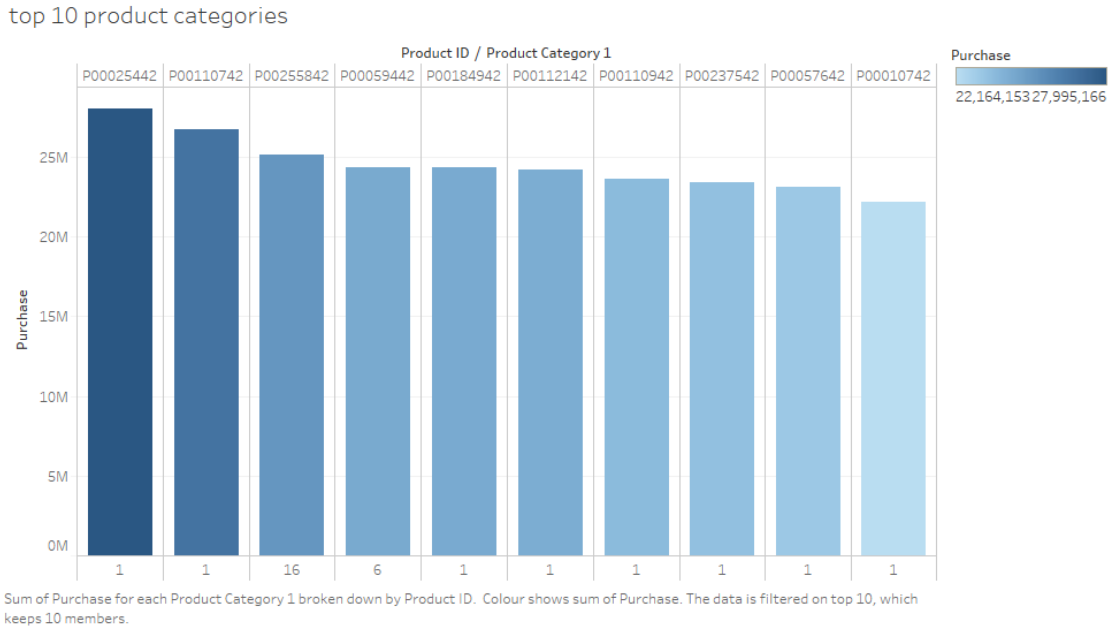
Stay in current city Years vs Marital Status Graph shows that customers who stayed only for 1 year and who are un married do higher purchases than compared to people who are married and number of years stayed in a city more than a year.

8.Purchase with stay in current city year broken down with city category



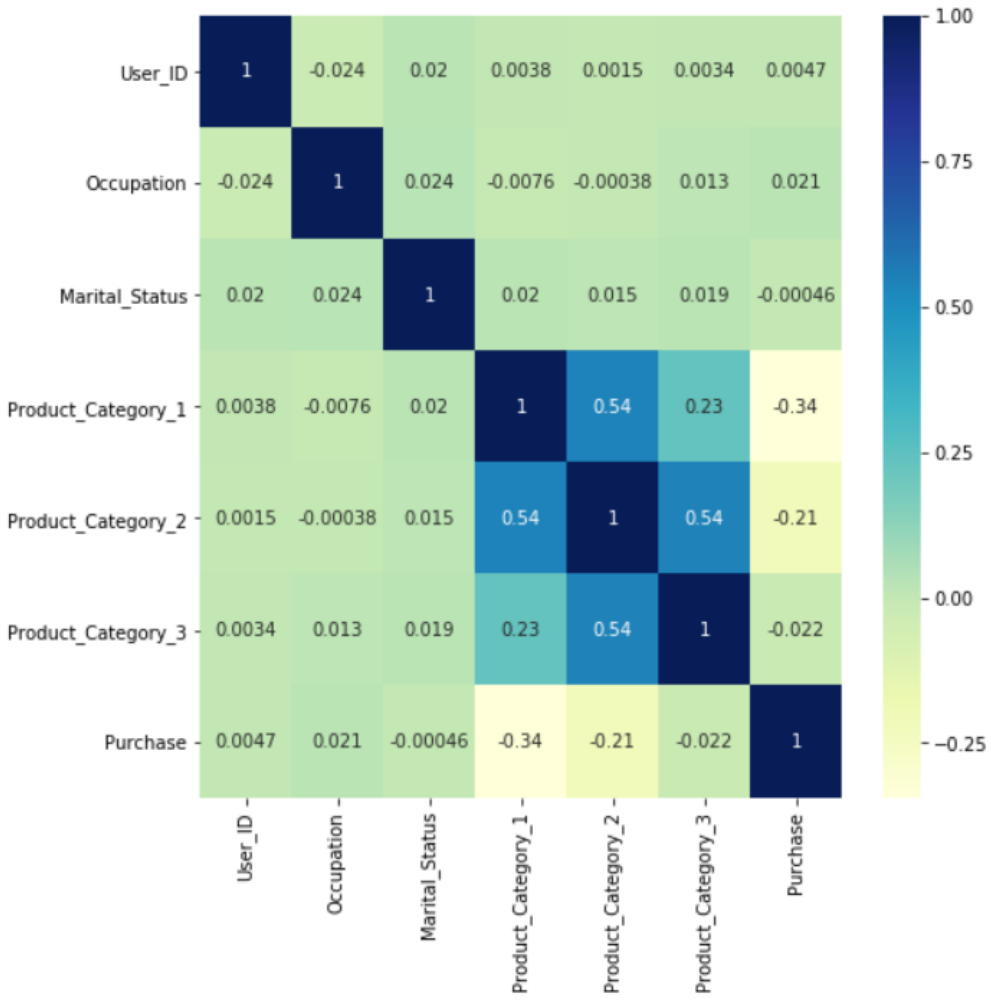
Here we can see the sum of purchases for each stay in current city years which is broken down by city category where City Category B and staying for 1 year in a city purchase rate is more. On the other hand Customers who stayed in city for more than 4 years are less willing to shop more do high purchases. People in city B contribute more to the revenue which may be due to city's population or higher buying capacity of the customers.

9.Top 10 product categories



This shows top 10 categories which shows top 10 customers product IDs.

10.Correlation Matrix



- As we can see correlation between the features is very less.
- Max co-relation is between Product_category_1 and Product_category_2.
- And also for Product_category_2 and Product_category_3.

Feature Extraction

1.Frequency Monetary Analysis:

This method used for analysing customer value. It is commonly used in database marketing and direct marketing and has received particular attention in retail and professional services industries evaluates which customers are of highest and lowest value to an organization based on purchase recency, frequency, and monetary value, in order to reasonably predict which customers are more likely to make purchases again in the future.

RFM stands for the three dimensions:

• Frequency – How often do they purchase? • Monetary Value – How much do they spend?

Based on the EDA on the features a frequency and monetary analysis is done in order to find out the products with high monetary and high frequency.

	Product_ID	Product_Score	Frequency
0	P00000142	0.612766	1152
1	P00000242	0.200000	376
2	P00000342	0.129787	244
3	P00000442	0.048936	92
4	P00000542	0.079255	149

	User_ID	User_Score	Frequency
0	1000001	0.034113	35
1	1000002	0.075049	77
2	1000003	0.028265	29
3	1000004	0.013645	14
4	1000005	0.103314	106

As we can see that product ID P00000142 have higher frequency and other product IDs next which are more important for revenue as most of the customers tend to purchase those products more. When we look at User Ids 1000005 are more valuable customers as that id frequency is higher where store can offer discounts to such kind of customers.

2.Scaling:

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

The data frame formed through frequency monetary analysis is subject to scaling in order to cluster the data, as it helps to eliminate the redundant data and ensure that the clusters are of good quality.

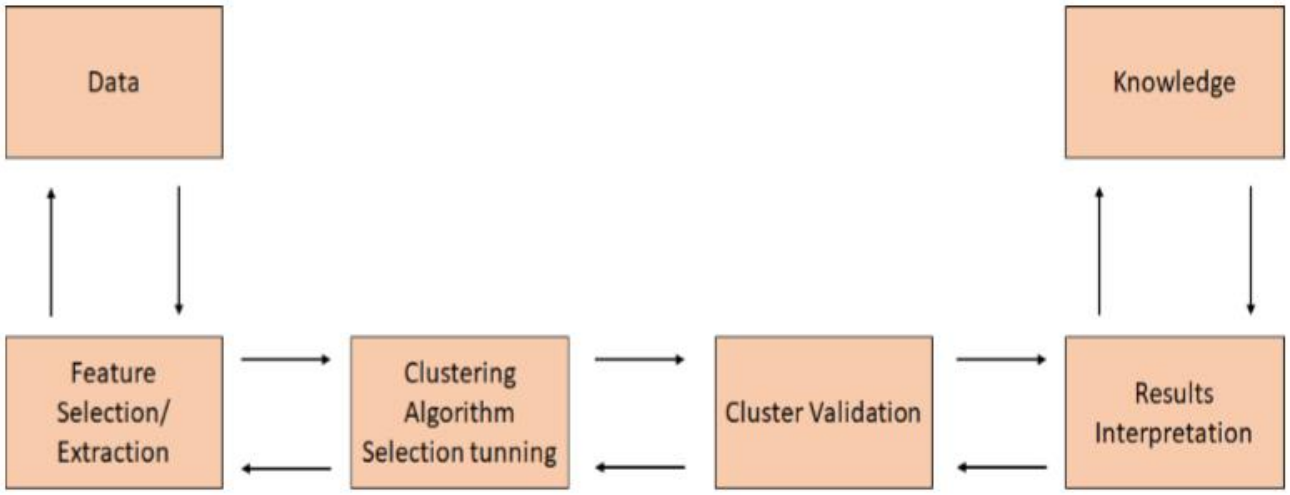
Scaling becomes essential as the Euclidean distance is very sensitive to the changes in difference

V. CLUSTERING:

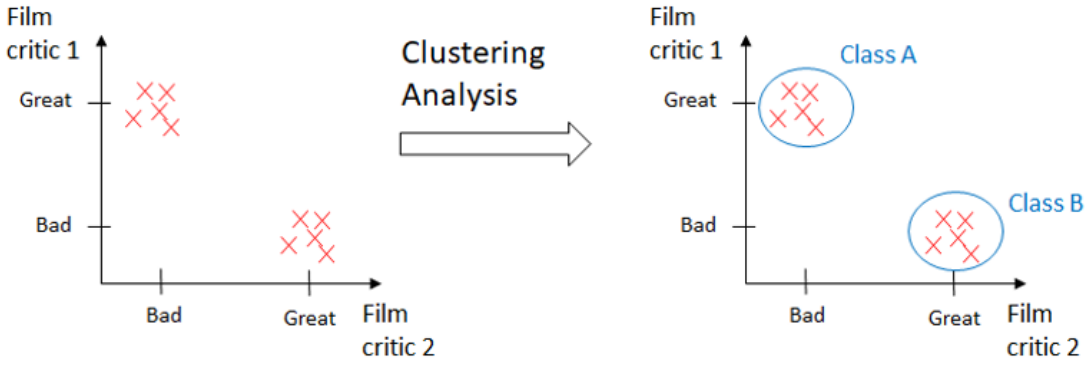
Clustering algorithms can utilize distance metrics to compute the data points to form groups. But since the data can be of different scales, it might impact on the cluster creation, since the clusters are grouped by calculating the distances. Different scales in the columns can be an issue.

Distances between data points and their cluster centres and the points attached to a cluster were used to measure the clustering quality among the three different standardization methods, the smaller the value of the sum of squares error the higher the accuracy, the better the result.

The overall process that we will follow when developing an unsupervised learning model can be summarized in the following chart:



Clustering Analysis: In basic terms, the objective of clustering is to find different groups within the elements in the data. To do so, clustering algorithms find the structure in the data so that elements of the same cluster (or group) are more similar to each other than to those from different clusters. In a visual way: Imagine that we have a dataset of movies and want to classify them. We have the following reviews of films:



The machine learning model will be able to infer that there are two different classes without knowing anything else from the data.

These unsupervised learning algorithms have an incredible wide range of applications and are quite useful to solve real world problems such as anomaly detection, recommending systems, documents grouping, or finding customers with common interests based on their purchases.

Some of the most common clustering algorithms, and the ones that will be explored through out the article, are:

• K-Means

• Hierarchical Clustering

• Density Based Scan Clustering (DBSCAN)

• Gaussian Clustering Model

K Means Clustering:

K-Means algorithms are extremely easy to implement and very efficient computationally speaking. Those are the main reasons that explain why they are so popular. But they are not very good to identify classes when dealing with in groups that do not have a spherical distribution shape.

The K-Means algorithms aims to find and group in classes the data points that have high similarity between them. In the terms of the algorithm, this similarity is understood as the opposite of the distance between data points. The closer the data points are, the more similar and more likely to belong to the same cluster they will be.

Key Concepts

• Squared Euclidean Distance

The most commonly used distance in K-Means is the squared Euclidean distance. An example of this distance between two points x and y in m-dimensional space is:

$$d(\boldsymbol{x}, \boldsymbol{y})^2 = \sum_{j=1}^{m} \left(x_j - y_j \right)^2 = \left\| \boldsymbol{x} - \boldsymbol{y} \right\|_2^2$$

• Cluster Inertia

Cluster inertia is the name given to the Sum of Squared Errors within the clustering context, and is represented as follows:

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} w^{(i,j)} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{\mu}^{(j)} \right\|_2^2$$

Where μ(j) is the centroid for cluster j, and w(i,j) is 1 if the sample x(i) is in cluster j and 0 otherwise.

K-Means can be understood as an algorithm that will try to minimize the cluster inertia factor.

Algorithm Steps

1. First, we need to choose k, the number of clusters that we want to be finded.

2. Then, the algorithm will select randomly the the centroids of each cluster.

3. It will be assigned each datapoint to the closest centroid (using euclidean distance).

4. It will be computed the cluster inertia.

5. The new centroids will be calculated as the mean of the points that belong to the centroid of the previous step. In other words, by calculating the minimum quadratic error of the datapoints to the center of each cluster, moving the center towards that point

6. Back to step 3.

Challenges of K-Means

• The output for any fixed training set won't be always the same, because the initial centroids are set randomly and that will influence the whole algorithm process.

• As stated before, due to the nature of Euclidean distance, it is not a suitable algorithm when dealing with clusters that adopt non-spherical shapes.

Points to be Considered When Applying K-Means

• Features must be measured on the same scale, so it may be necessary to perform z-score standardization or max-min scaling.

• When dealing with categorical data, we will use the get dummies function.

• Exploratory Data Analysis (EDA) is very helpful to have an overview of the data and determine if K-Means is the most appropriate algorithm.

• The mini batch method is very useful when there is a large number of columns, however, it is less accurate.

How to Choose the Right K Number ?

Choosing the right number of clusters is one of the key points of the K-Means algorithm. To find this number there are some methods:

• Field knowledge

• Business decision

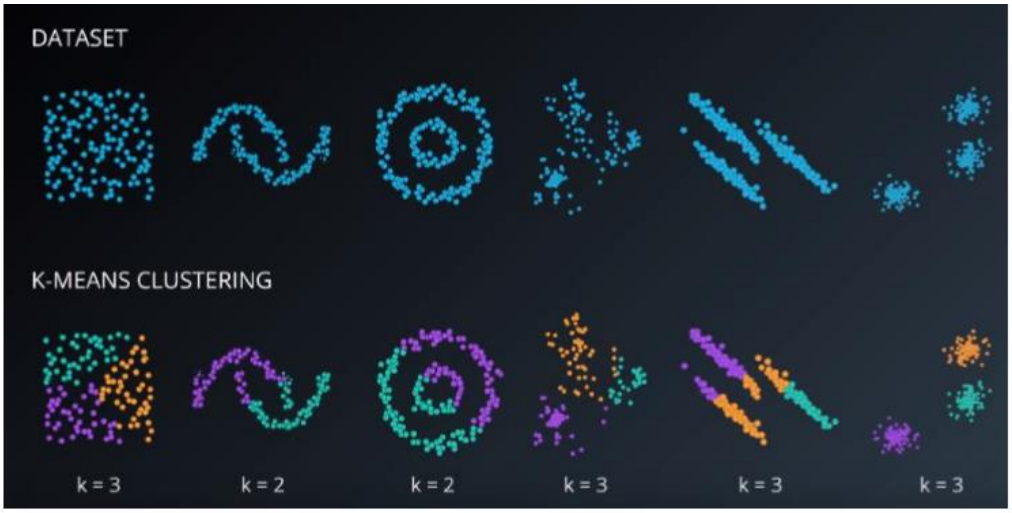
• Elbow Method

As being aligned with the motivation and nature of Data Science, the elbow method is the preferred option as it relies on an analytical method backed with data, to make a decision.

K-Means Limitations

Although K-Means is a great clustering algorithm, it is most useful when we know beforehand the exact number of clusters and when we are dealing with spherical-shaped distributions.

The following picture show what we would obtain if we use K-means clustering in each dataset even if we knew the exact number of clusters beforehand:



It is quite common to take the K-Means algorithm as a benchmark to evaluate the performance of other clustering methods.

Hierarchical Clustering

Hierarchical clustering is an alternative to prototype-based clustering algorithms. The main advantage of Hierarchical clustering is that we do not need to specify the number of clusters, it will find it by itself. In addition, it enables the plotting of dendrograms. Dendrograms are visualizations of a binary hierarchical clustering.

Hierarchical clustering will not be implemented since the data set has over 5,00,000 rows and the only disadvantage in hierarchical clustering is that it is computationally expensive. Hence it will not be implemented.

Hierarchical clustering is one of the popular and easy to understand clustering technique. This clustering technique is divided into two types:

1. Agglomerative

2. Divisive

• Agglomerative Hierarchical clustering Technique: In this technique, initially each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed.

The basic algorithm of Agglomerative is straight forward.

• Compute the proximity matrix

• Let each data point be a cluster

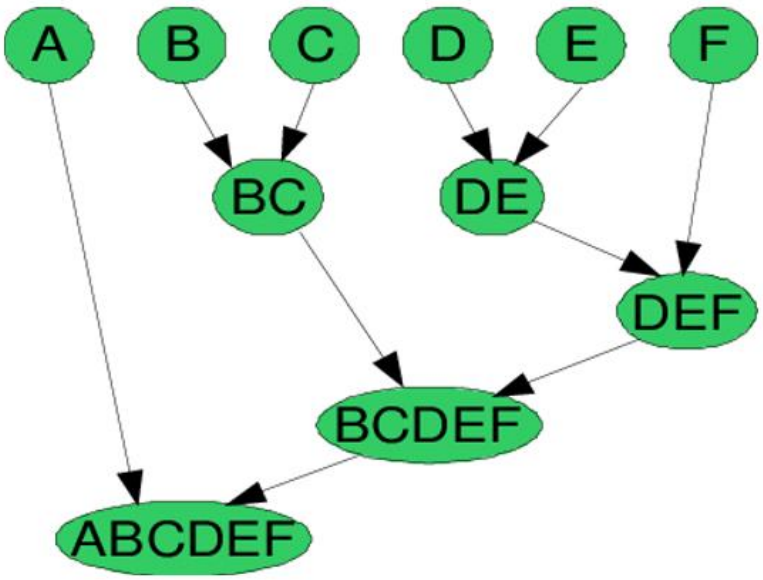
• Repeat: Merge the two closest clusters and update the proximity matrix

• Until only a single cluster remains

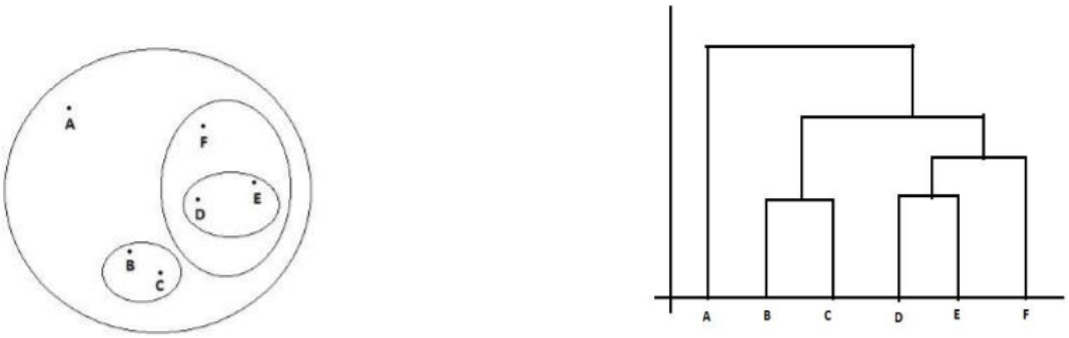
Key operation is the computation of the proximity of two clusters

To understand better let's see a pictorial representation of the Agglomerative Hierarchical clustering Technique. Let's say we have six data points {A, B, C, D, E, F}.

• Step- 1: In the initial step, we calculate the proximity of individual points and consider all the six data points as individual clusters as shown in the image below.



The Hierarchical clustering Technique can be visualized using a Dendrogram. A Dendrogram is a tree-like diagram that records the sequences of merges or splits.



Dendrogram representation

• Divisive Hierarchical clustering Technique: Since the Divisive Hierarchical clustering Technique is not much used in the real world, I'll give a brief of the Divisive Hierarchical clustering Technique.

In simple words, we can say that the Divisive Hierarchical clustering is exactly the opposite of the Agglomerative Hierarchical clustering. In Divisive Hierarchical clustering, we consider all the data points as a single cluster and in each iteration, we separate the data points from the cluster which are not similar. Each data point which is separated is considered as an individual cluster. In the end, we'll be left with n clusters.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Density-Based Spatial Clustering of Applications with Noise, or DBSCAN, is another clustering algorithm especially useful to correctly identify noise in data.

DBSCAN Assigning Criteria

It is based on a number of points with a specified radius ε and there is a special label assigned to each datapoint. The process of assigning this label is the following:

• It is a specified number (MinPts) of neighbour points. A core point will be assigned if there is this MinPts number of points that fall in the ε radius.

• A border point will fall in the ε radius of a core point, but will have less neighbors than the MinPts number.

• Every other point will be noise points.

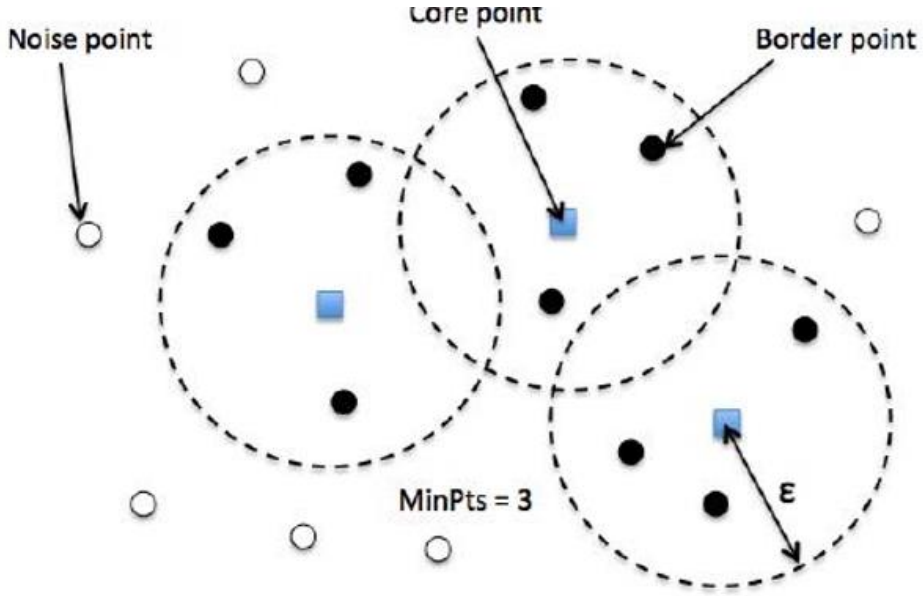
DBSCAN Algorithm

The algorithm follows the logic:

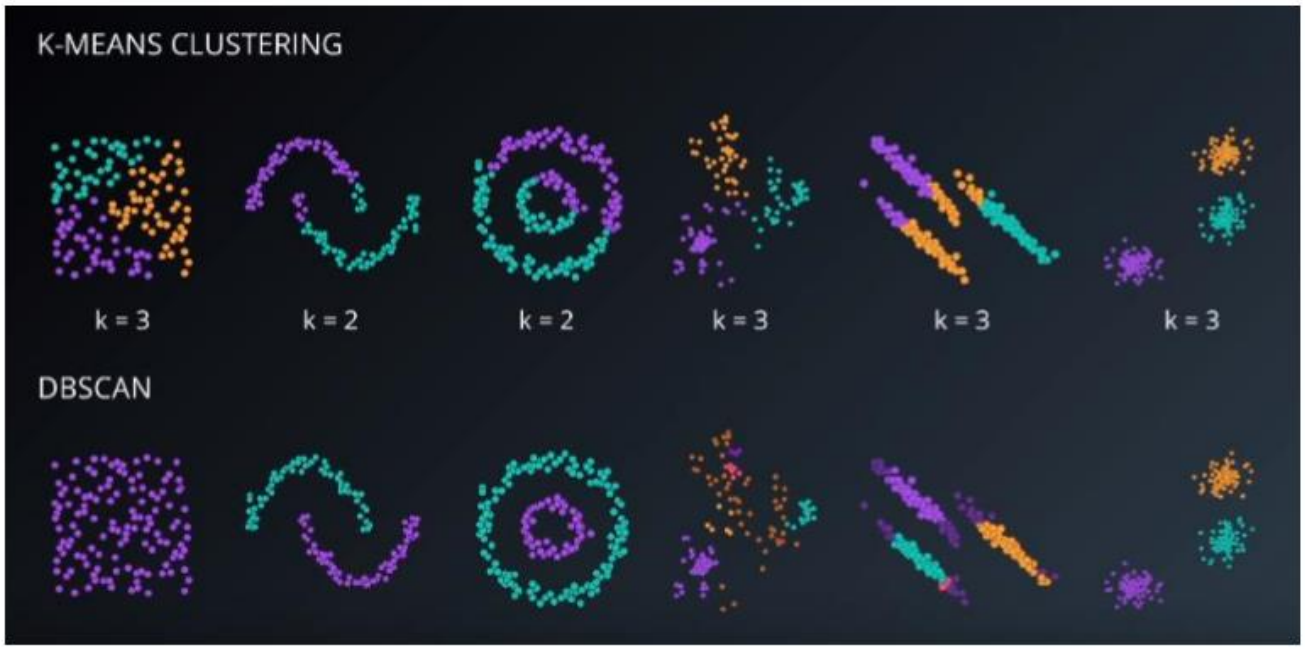
1. Identify a core point and make a group for each one, or for each connected group of core points (if they satisfy the criteria to be core point).

2. Identify and assign border points to their respective core points.

The following figure summarize very well this process and the commented notation.



DBSCAN vs K-Means Clustering



DBSCAN Advantages

• We do not need to specify the number of clusters.

• There is high flexibility in the shapes and sizes that the clusters may adopt.

• It is very useful to identify and deal with noise data and outliers.

DBSCAN Disadvantages

• It faces difficulties when dealing with border points that are reachable by two clusters.

• It doesn't find well clusters of varying densities.

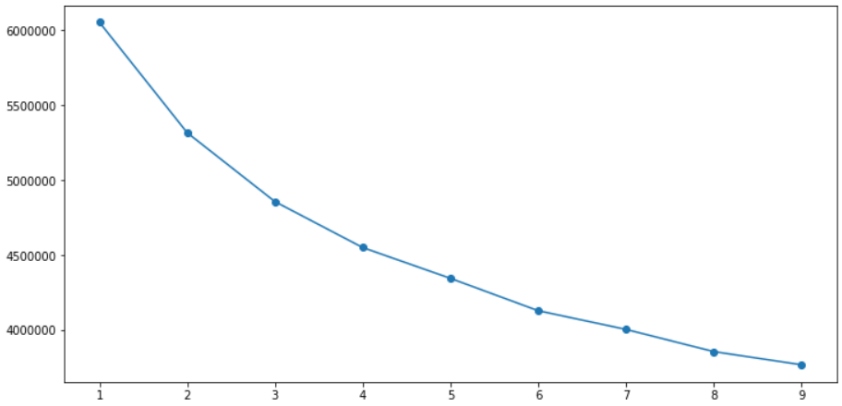
Among all these clustering methods K means is the suitable method for clustering as it shows clear clustering with different color format.

Principal component analysis (**PCA**): It is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components. Principal Component Analysis (**PCA**) is **used to** explain the variance-covariance structure of a set of variables through linear combinations. It is often **used as** a dimensionality-reduction technique.

K – means implementation on the dataset:

K-means clustering is applied to the scaled data to perform product segmentation and obtain the clusters based on various frequency and monetary metrics such as product cluster with low monetary and low frequency, low frequency and high monetary, high frequency and high monetary. In this as the variables are not correlated and its not possible to do regression clustering is done check the accuracy after applying k means clustering and PCA.In this dataset when K means is applied without PCA then the elbow plot we got is

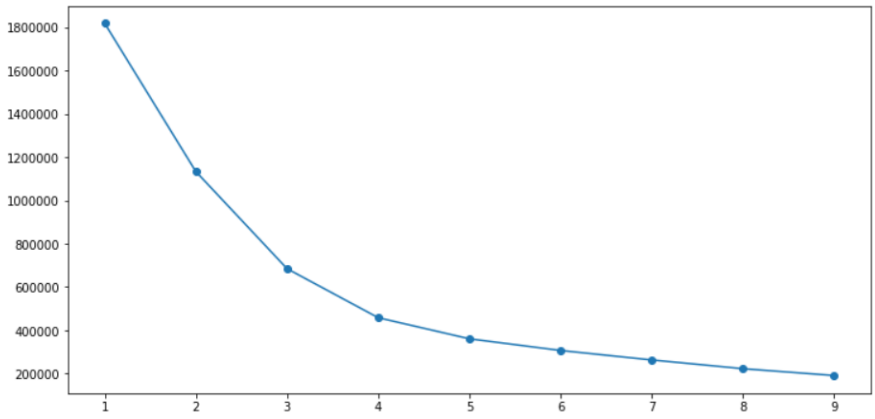
	custers_num	cluster_errors
0	1	6.050748e+06
1	2	5.314795e+06
2	3	4.855745e+06
3	4	4.549778e+06
4	5	4.344351e+06
5	6	4.128801e+06
6	7	4.004235e+06
7	8	3.856358e+06
8	9	3.768640e+06



- we are not able to see any elbow but we can see that the drop-in error is constant after 3 clusters.
- From this we cannot interpret that how many clusters we can use to take predicted labels.

When K means is applied with PCA then the elbow plot we got is

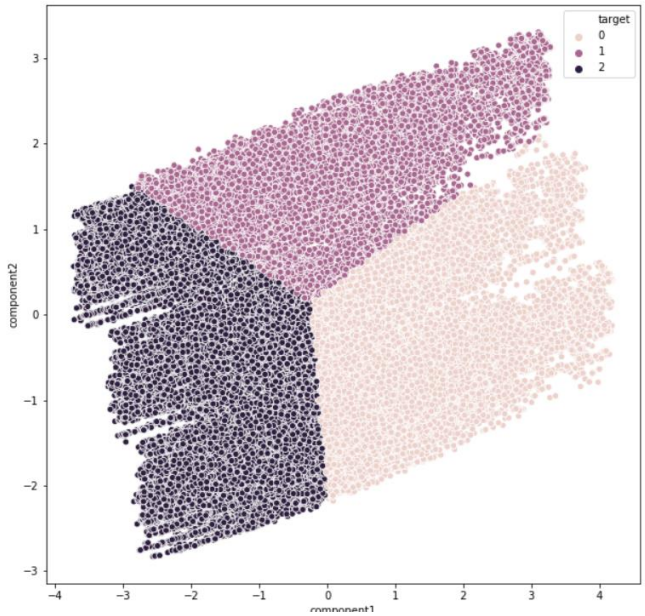
	custers_num	cluster_errors
0	1	1.817899e+06
1	2	1.133282e+06
2	3	6.848156e+05
3	4	4.583770e+05
4	5	3.609536e+05
5	6	3.070289e+05
6	7	2.632322e+05
7	8	2.228667e+05
8	9	1.913940e+05



- After PCA we can see that our elbow plot is changing and we can see that the elbow is forming at 3 or 4.
- From this we can see that at after 3 and 4 the error drop is almost constant.
- We can consider both number of clusters and make a model around it and check by performing supervised-learning or by looking clusters visually.

VI. MODEL BUILDING:

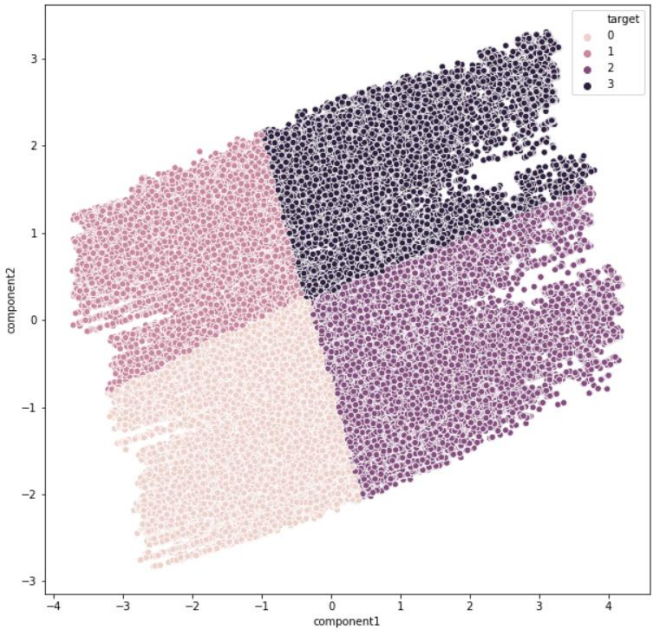
Considering 3 clusters:



Model	Accuracy
Logistic Regression	**0.97**
DecisionTreeClassifier	**0.99**
KneighborsClassifier	**0.99**
GaussianNB	**0.97**

- From this we can see data is perfectly divided into 3 clusters and labels as 0,1,2.
- As performed PCA on the dataset so cannot explain the components.

Considering 4 clusters:

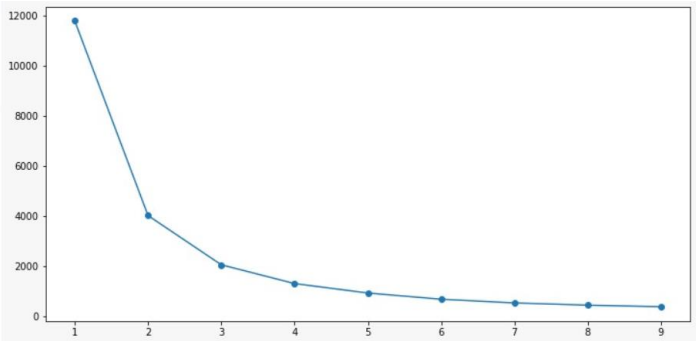


Model	Accuracy
Logistic Regression	**0.94**
DecisionTreeClassifier	**0.99**
KneighborsClassifier	**0.99**
GaussianNB	**0.95**

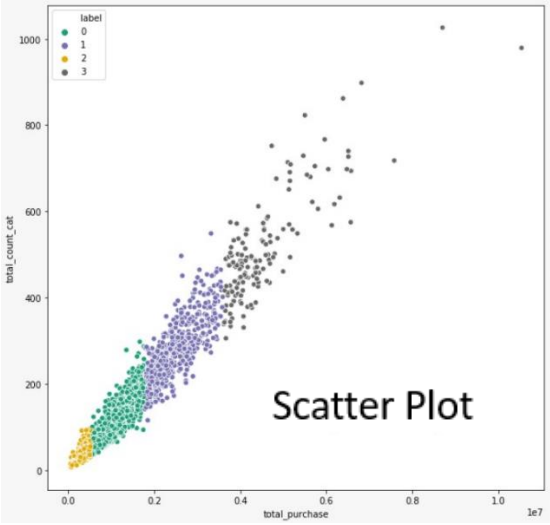
- From this we can see that whole data is clustered into 4 clusters which is labled as 0,1,2,3.
- From performing kmeans clustering and taking making 3 and 4 clusters we can see that we are able to achieve good accuracy on 3 clusters so we can consider 3 clusters.

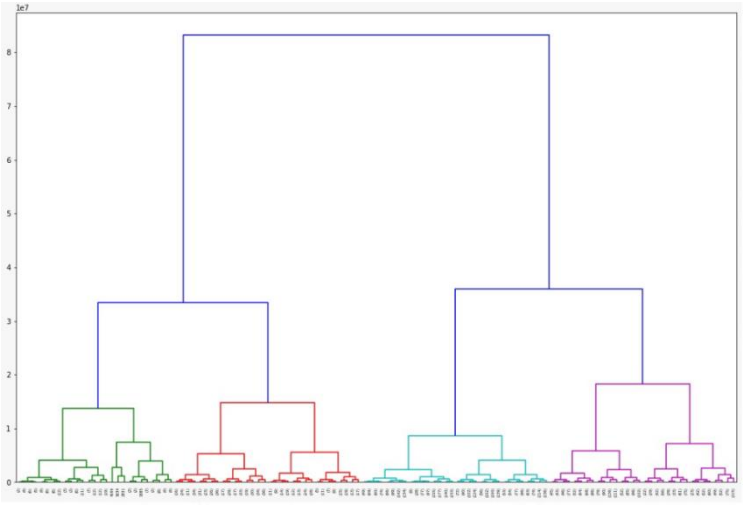
Final model of clusters based on purchase

As we cannot derive patterns from base model user IDs are divided according to their purchases as the user IDs were repeating for various purchases and clustering is done.



Elbow Plot



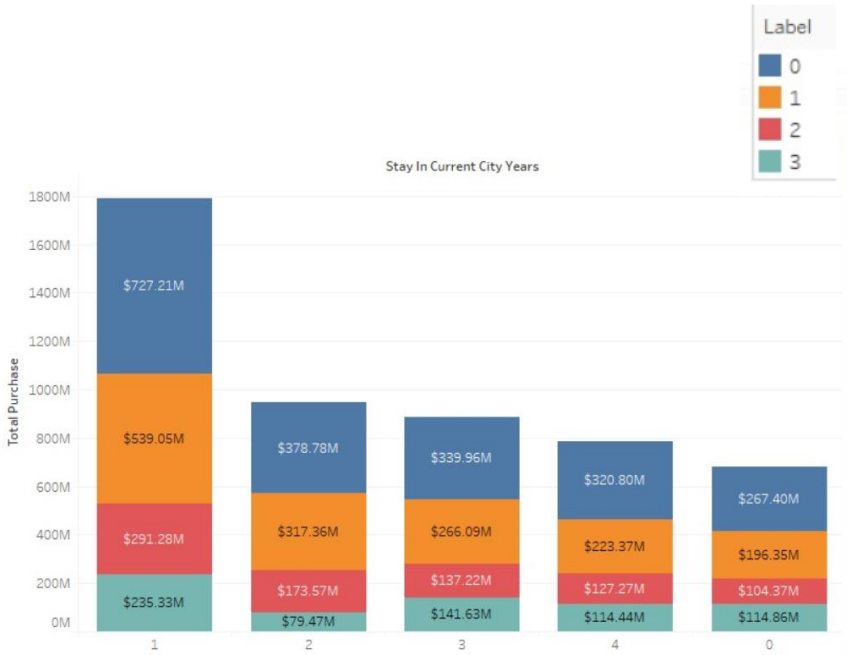


Agglomerative Clustering

```
LogisticRegression accuracy 0.9445701357466063
LogisticRegression f1_score 0.8946428571428572
RandomForestClassifier accuracy 0.9949095022624435
RandomForestClassifier f1_score 0.9857866076238381
GaussianNB accuracy 0.9621040723981901
GaussianNB f1_score 0.9458001198912626
```

Data is reframed by using pivot for user ID and gathered the purchase for user IDs and got 4 distinct clusters as shown in Agglomerative clustering and scatter plot.

Insights from clusters



- The most revenue has been generated from Cluster 0
- In every cluster we can see that people who have stayed for one year in the current city have spent more during the sale.
- Top 5 occupations who made more purchases
- Purchase patterns in different cities with respect to clusters.

VII. INFERENCES

- Cluster 3 users may be wholesale or premium buyers as the purchase quantity and total purchase are high.
- Cluster 1 users may be compulsive shoppers who shop at any time irrespective of sales
- Cluster 0 users may be sales addicts who shop moderately.
- Cluster 2 users may be hunters who particularly buy specific products in a single quantity as their quantity and total purchase are low.

Conclusion

With the machine learning algorithms like clustering algorithms and similar product recommendation used, the solution was built with ease providing benefits to the loyal customers. These strategies were identified with ease using the machine learning methods in these applications. This approach can be used in any retail market sales data where understanding customer requirement and building strategies that would increase sales to the company and also satisfy the customers that ought to buy the products by providing offers, discounts and complementary or combos to the customers.