Customer Segmentation using K Means Clustering & RFM Analysis

Team Members:
Garveet Juneja
Vansh Chugh

Mentor :-
Mr Baan Bapat

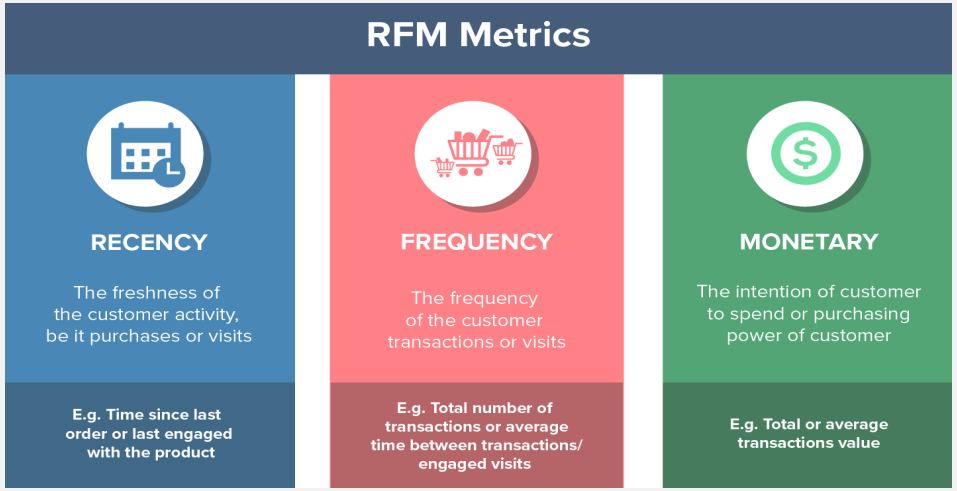
Customer Segmentation

Customer segmentation is the practice of *categorizing* customers into groups based on shared qualities so that businesses may *market* to each group *effectively and efficiently.*

Consumers of all types have *distinct* requirements and goals and respond to marketing initiatives in different ways.

As a company grows, segmenting the customers can help in enhancing marketing success by customizing the services and promotional campaigns to the customers, ultimately increasing response rates and sales.

RFM – Recency Frequency & Monetary



RFM is an effective segmentation method to enable marketers to analyze customer behavior.

It is typically used to identify groups of customers in order to develop specific campaigns to target them.

These RFM metrics are indicators of a customer's behavior because the frequency and monetary value affect a customer's lifetime value, and recency affects retention, a measure of engagement.

DATA

- The data set has been taken from a trustworthy source, consisting of the information on the products transaction and payment modes used by customers in each transaction at Big Bazaar.
- Big-Bazaar runs various loyalty programs, festive offers which provide their customer more opportunities to avail discounts. Customers can use these offers or loyalty program to either avail discount or make payment.
- So we would try to segment our customers and try to provide them with offers / discounts that attracts them the most.

OBJECTIVE

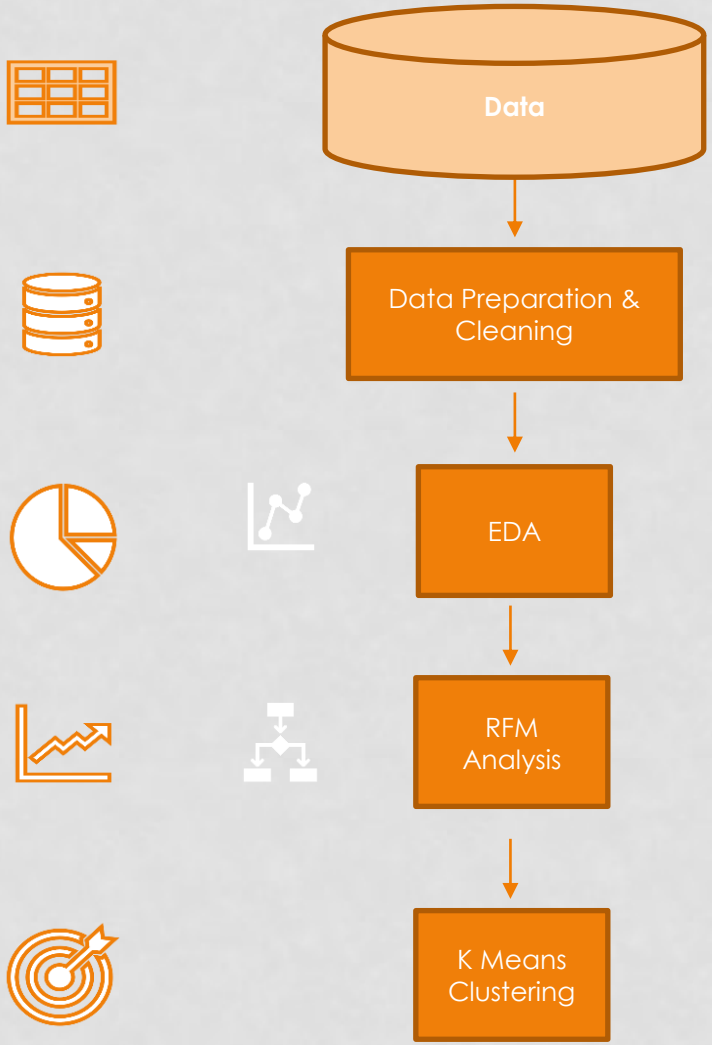
Our primary goal is to segment Big Bazaar’s customer base into homogeneous groups, understand the traits of each group, as well as, to engage them with relevant campaigns to fuel the aforementioned retail chain’s growth. Observe and analyze the strong relationship between variables of the data, if any. And,

Using K Means Clustering to create clusters of customers based on following groups:

1. **Top Customers**
2. **High Value Potential**
3. **Medium Value Attention**
4. **Low Value Customers**
5. **Lost Customers**

Based upon the above segments of customers the team shall implement the requisite campaigns to convert potential customers to star customer.

PROJECT FLOWCHART



DATA DESCRIPTION (PURCHASE DATA)

Customer ID	• Unique customer ID of the customer.
DOB	• Date of Birth of the customer.
Gender	• Gender of the customer
State	• State
Pincode	• Pincode of area where customer lives.
Transaction Date	• Date of transaction
Store Code	• Unique code of Big Bazaar store
Store Description	• Basic description of the store.
Till Number	• Counter number in the store.
Transaction Number by Till	• Unique transaction number by counter.
Promo Code	• Promotional code used in the transaction.
Promo Description	• Description of the offer.
Product Code	• Unique code of the product purchased.
Product Description	• Description of the product purchased.
Sale price after promo	• Sale price of the product after applying promotion.
Discount Used	• After promo, Customer used this discount(s) on transaction.

DATA DESCRIPTION (TENDER DATA)

Customer ID	• Unique customer ID of the customer.
DOB	• Date of Birth of the customer.
Gender	• Gender of the customer
State	• State
Pincode	• Pincode of area where customer lives.
Transaction Date	• Date of transaction
Store Code	• Unique code of Big Bazaar store
Store Description	• Basic description of the store.
Till Number	• Counter number in the store.
Transaction Number by Till	• Unique transaction number by counter.
Tender Type	• Mode of Payment
Payment amount by tender	• Amount paid using the payment mode
Payment Mode Used	• Description of mode of payment.

DATA CLEANING AND PREPARATION

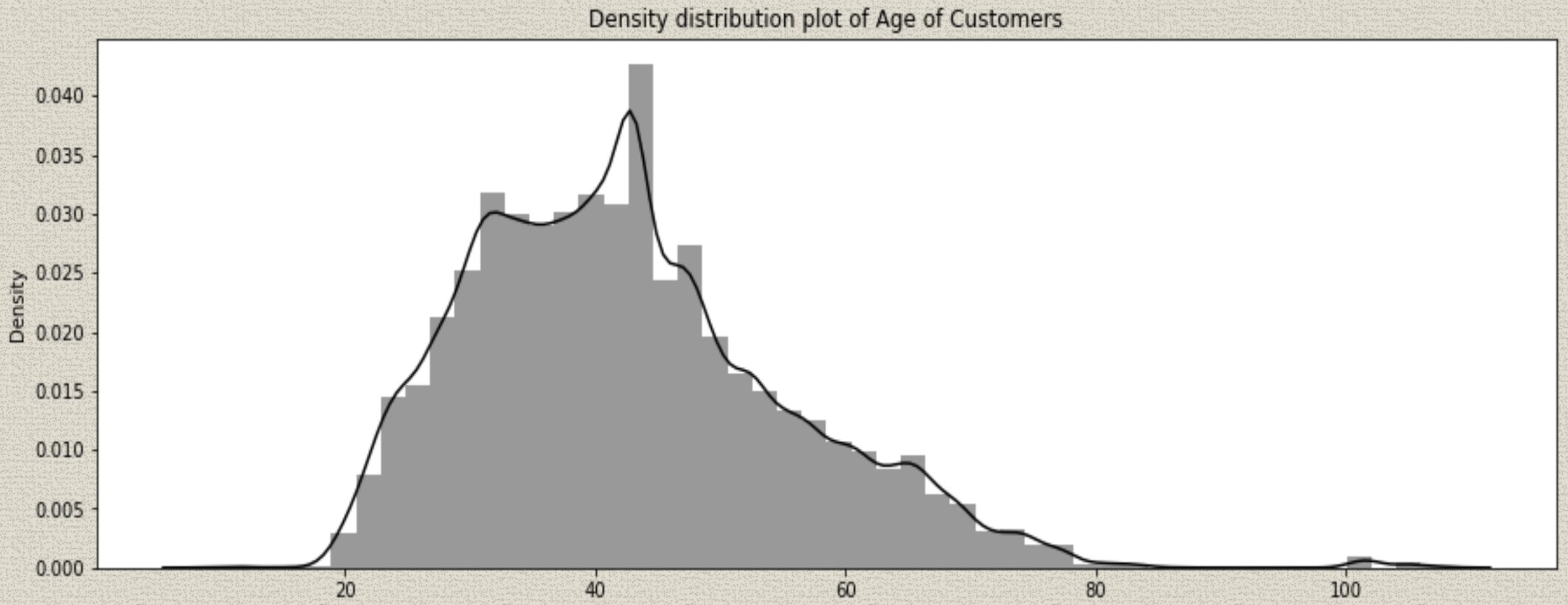
We started off by merging our two datasets (Products and Tender) by performing a left join on the two tables using a composite primary key (CustomerID + Transaction Number by till).

- Some columns were redundant for our analysis so those columns were removed from our table (On merging the two datasets the common columns in the two were repeating, they were removed before further exploration).
- There were no duplicate rows in the merged table but some columns(DOB, Gender, State, etc.) contained null values.
- A new column called “Age” was created to store the age of the customers which was calculated using the given date of birth of the customers.
- Many states were categorized into different states due to spelling mistakes so that was corrected.
- The variable promo_code was categorized into two – NONPROMO & PROMO.

Our final table consisted of 14470 and 19 columns.

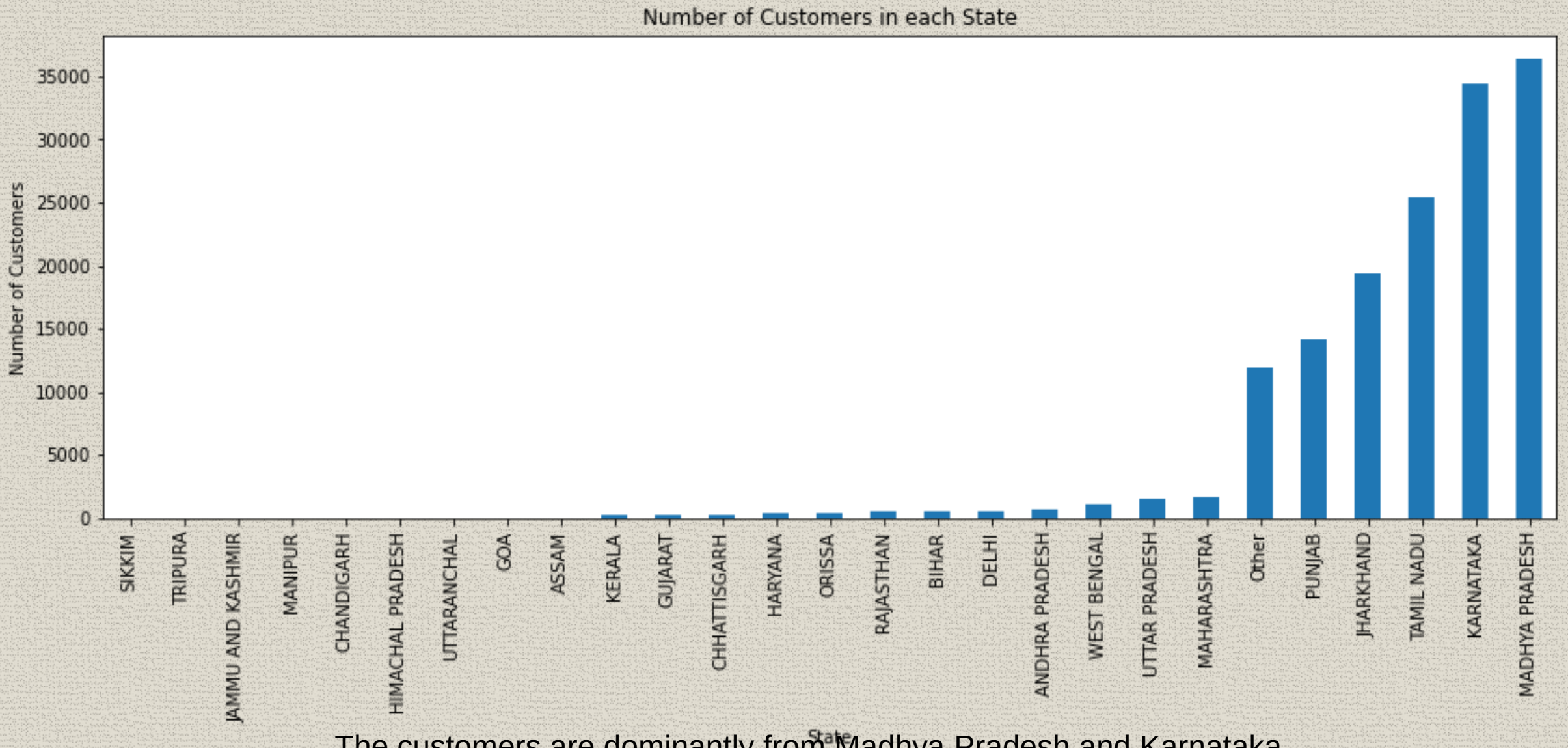
EXPLORATORY DATA ANALYSIS

Density distribution plot of Age of Customers



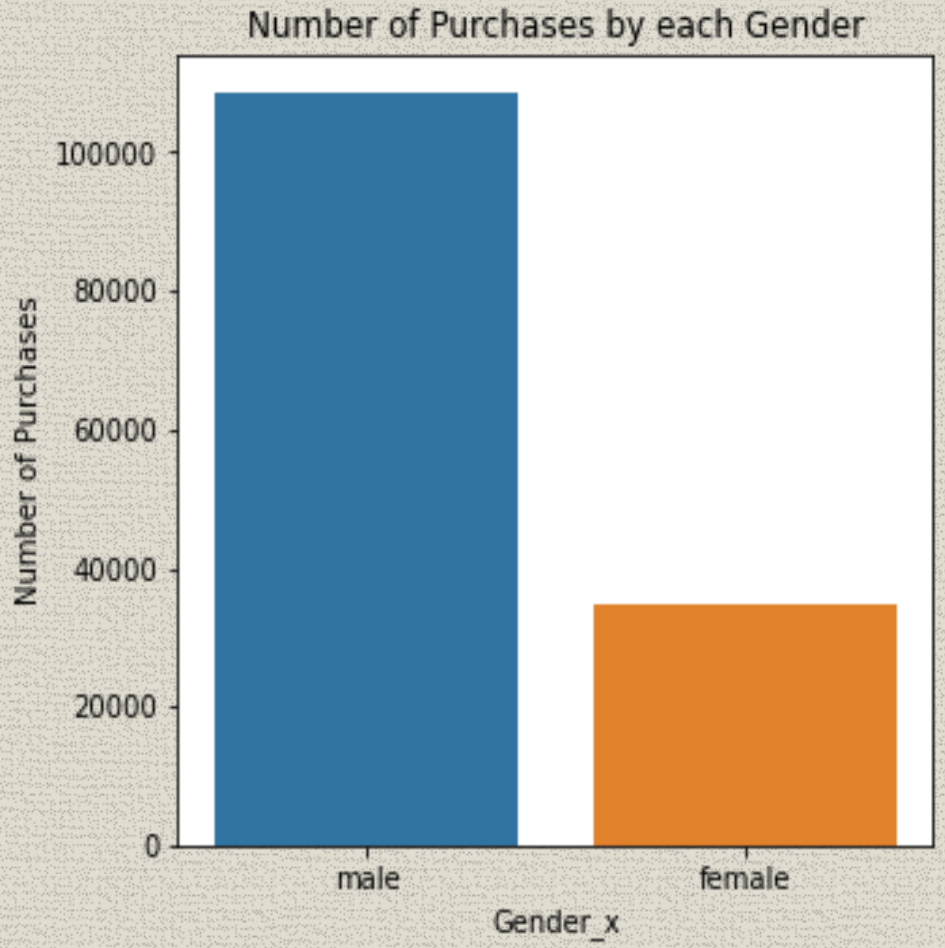
Most of the customers are young; within the age group of 25-45.

EXPLORATORY DATA ANALYSIS

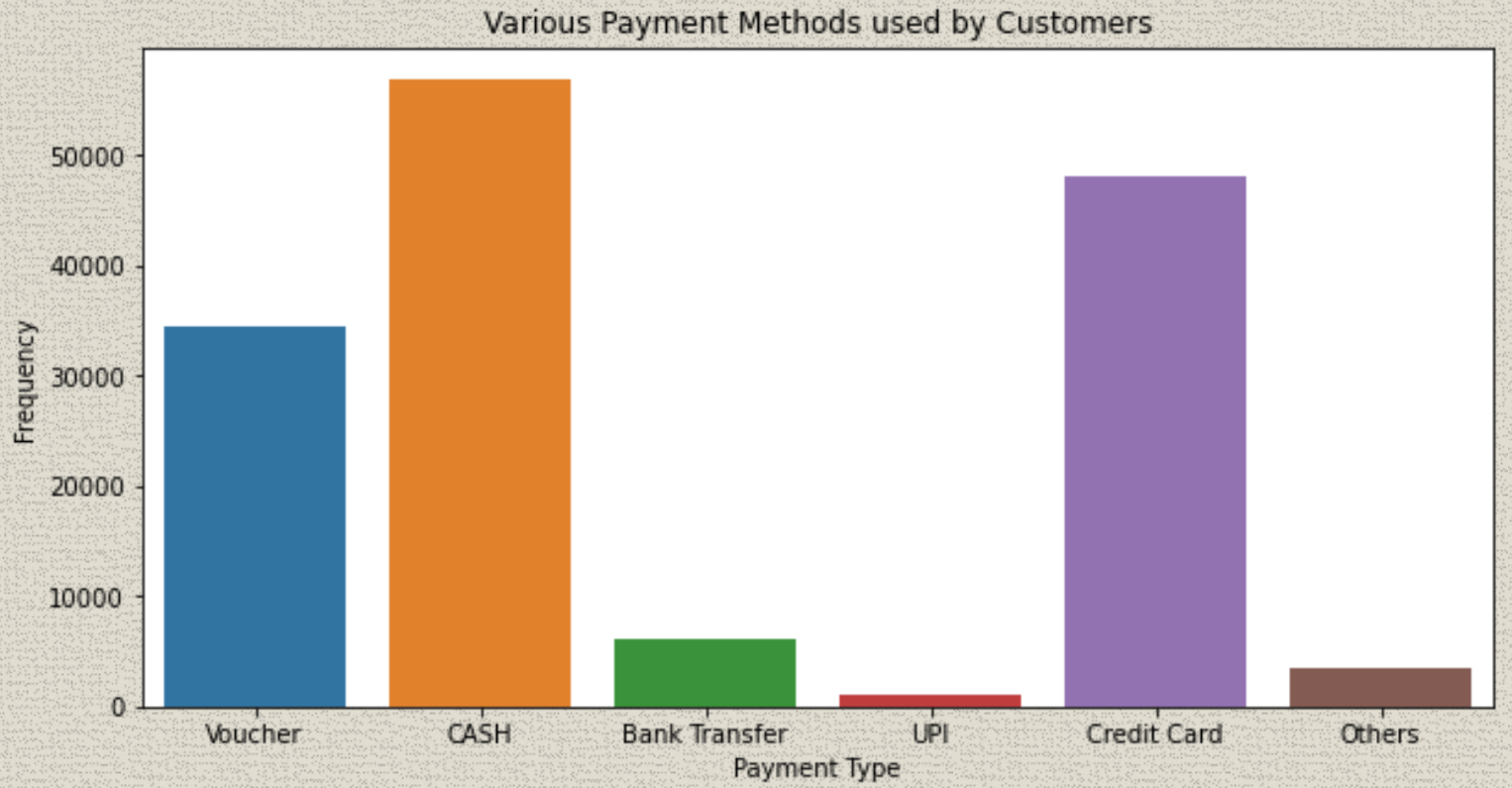


The customers are dominantly from Madhya Pradesh and Karnataka

EXPLORATORY DATA ANALYSIS

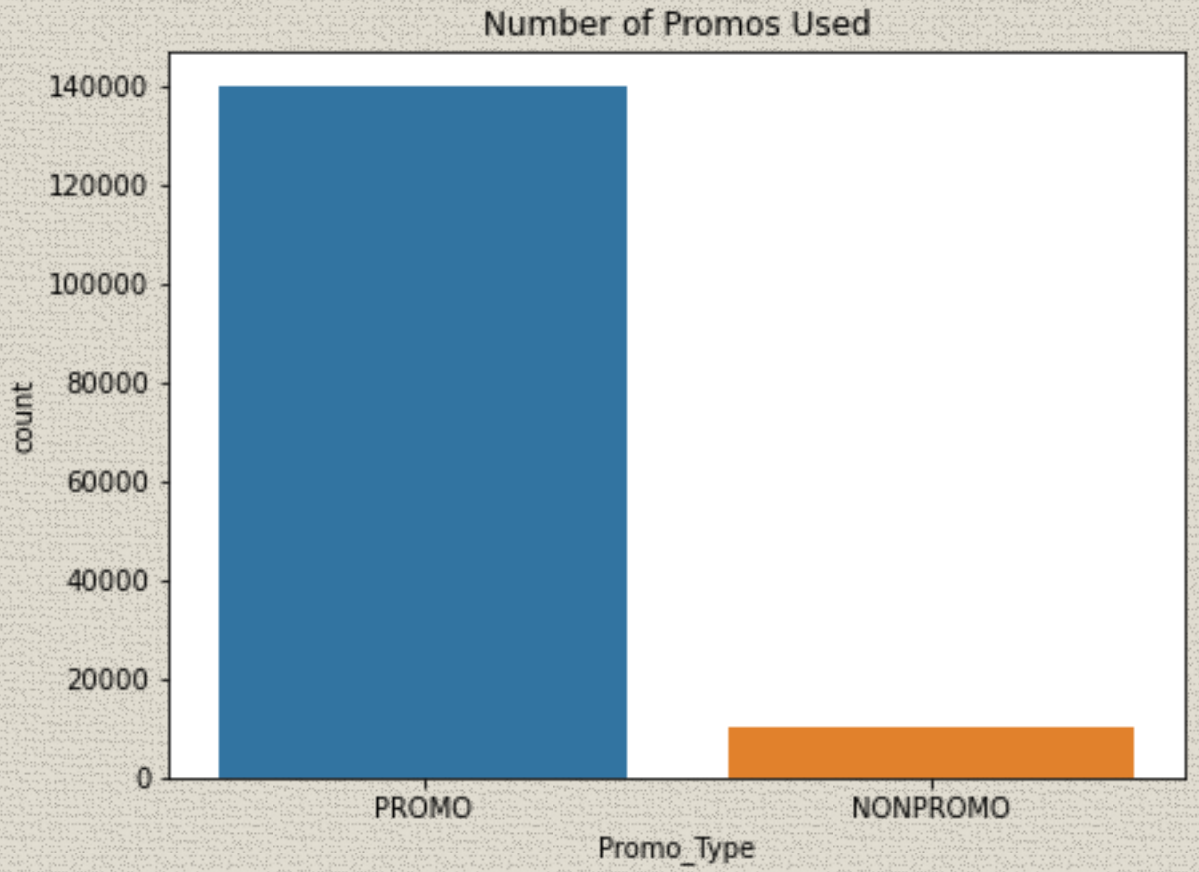


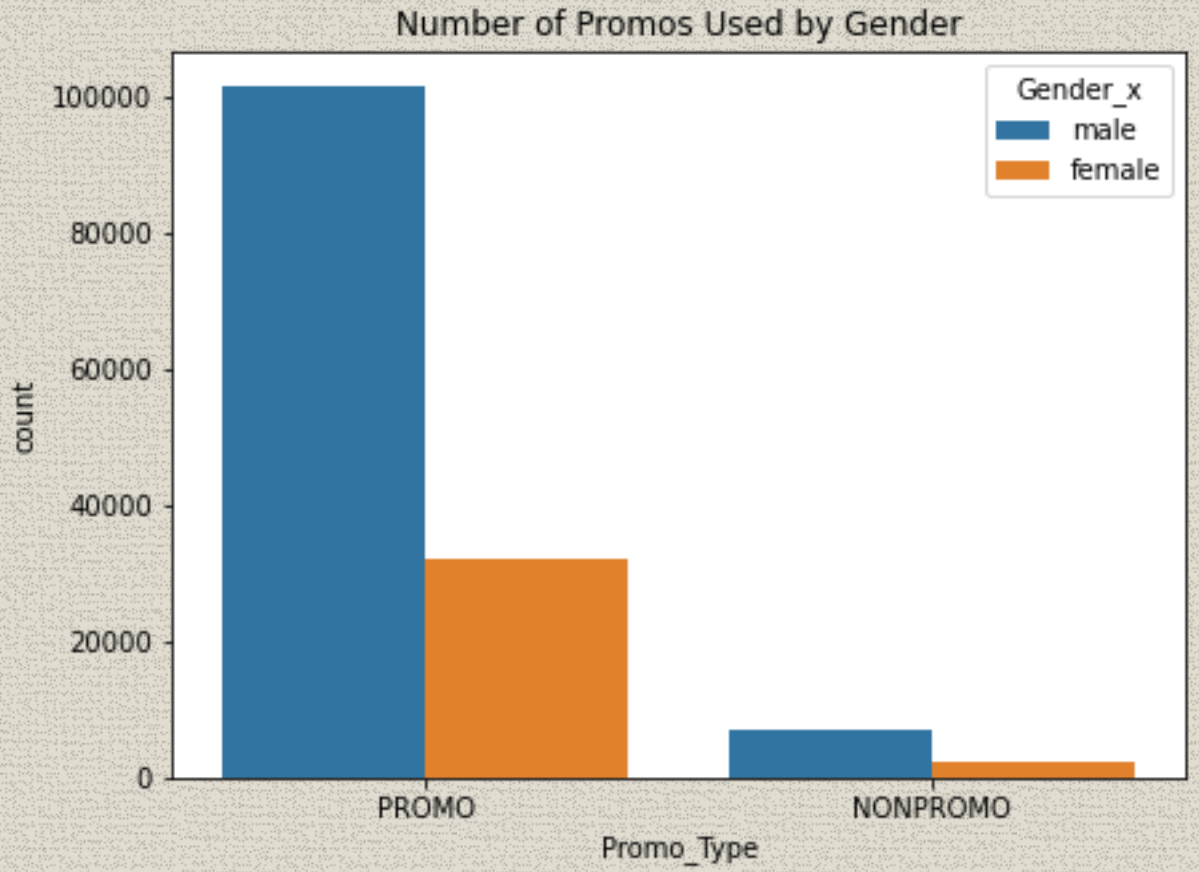
Most of the Purchases were made by Male customers (about 72%)



Most of the customers used Credit Card or Cash as a preferred mode of payment, and
Many customers also redeemed their Gift Vouchers as a mode of payment

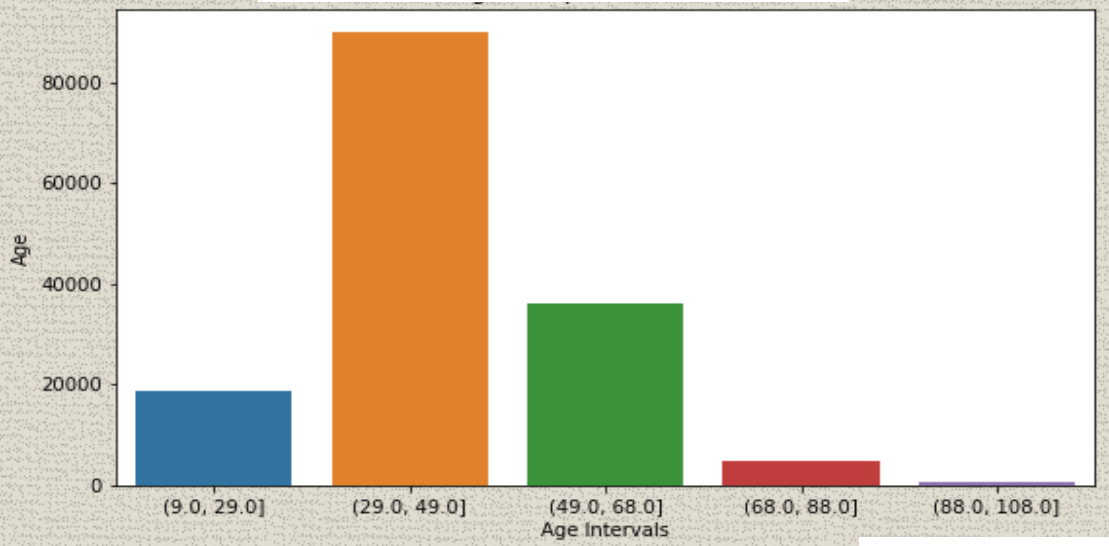
EXPLORATORY DATA ANALYSIS



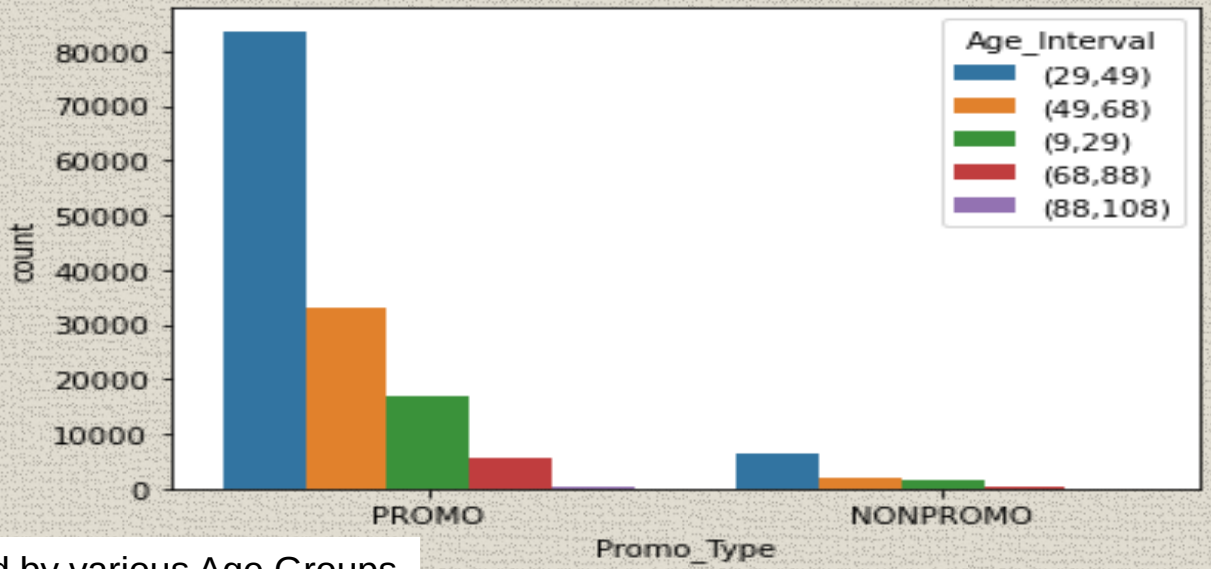


Most Purchases (about 71%) were made without the use of a promo code, and
27% of male customers used a Promo code while 25% women customers availed a discount from promo code.

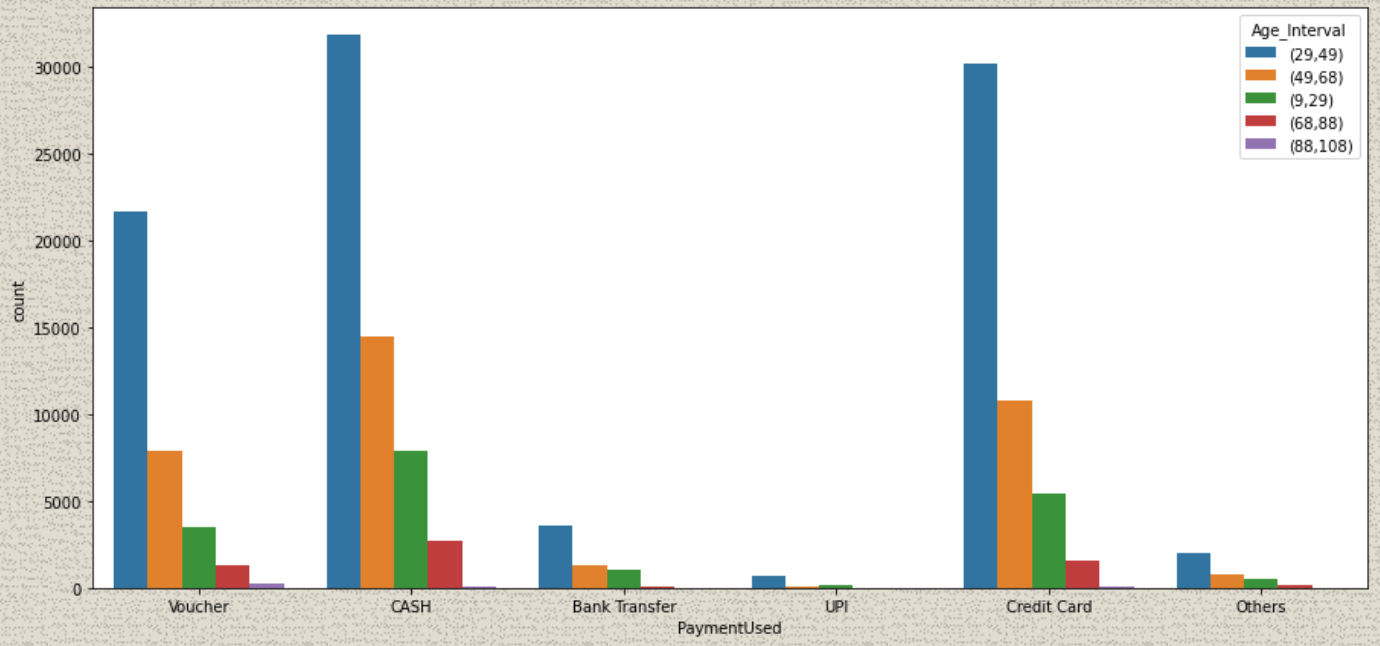
Customer by various Age Groups

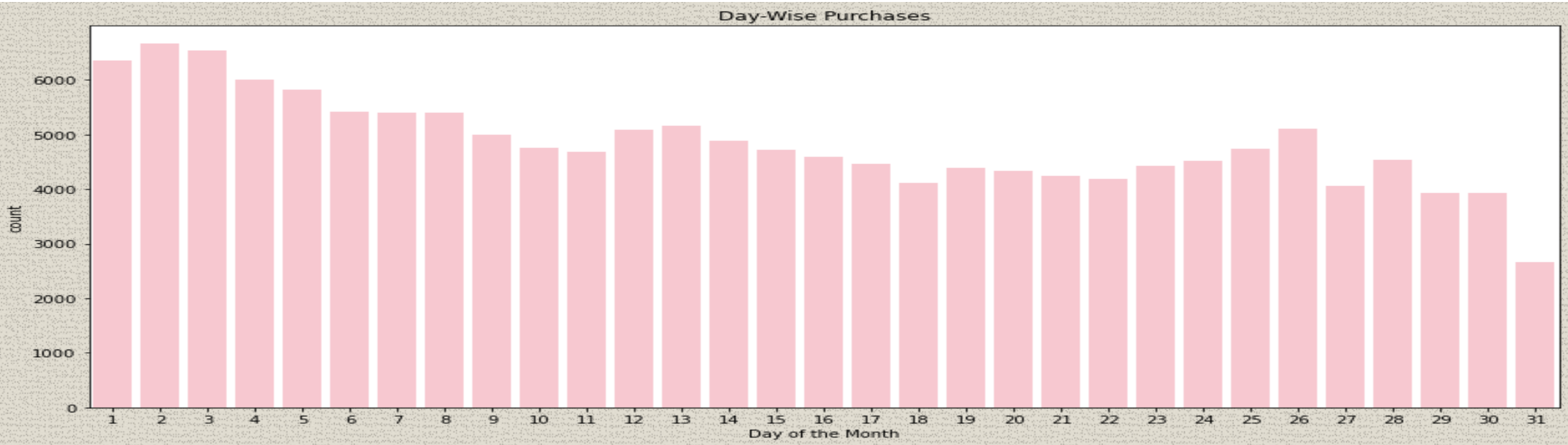


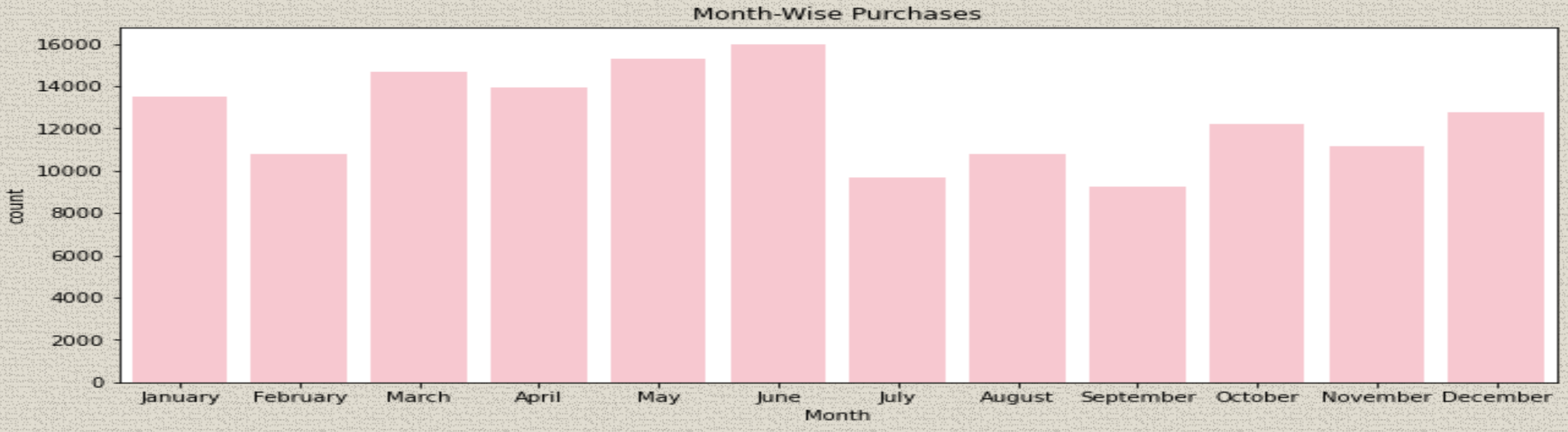
Promo Codes Used by various Age Groups



Payment methods used by various Age Groups

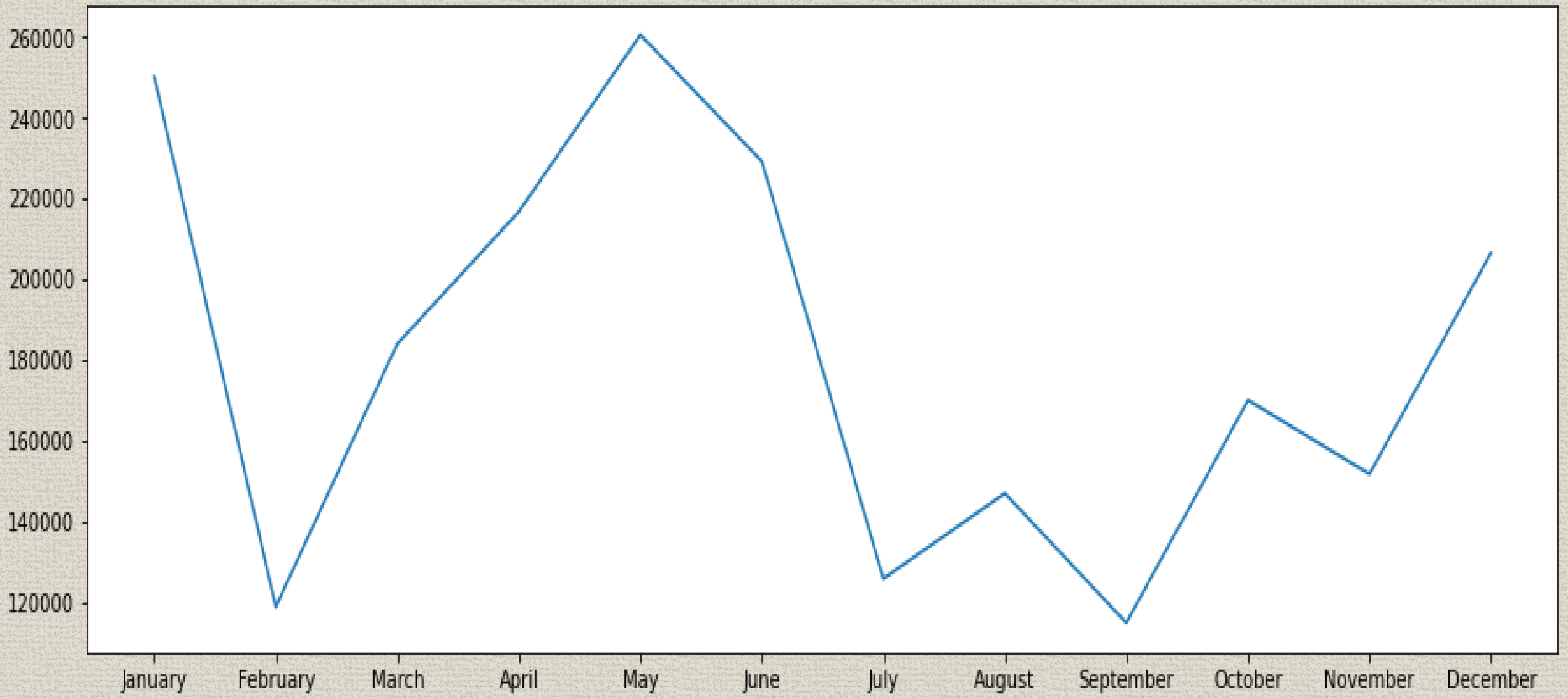



Day-Wise Purchases
count
0
1000
2000
3000
4000
5000
6000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
Day of the Month


Month-Wise Purchases
count
0
2000
4000
6000
8000
10000
12000
14000
16000
January
February
March
April
May
June
July
August
September
October
November
December
Month

EXPLORATORY DATA ANALYSIS

MONTH WISE PURCHASES IN MONETARY TERMS



RFM CLUSTERING WITH K MEANS

Let us understand how K-means clustering works:

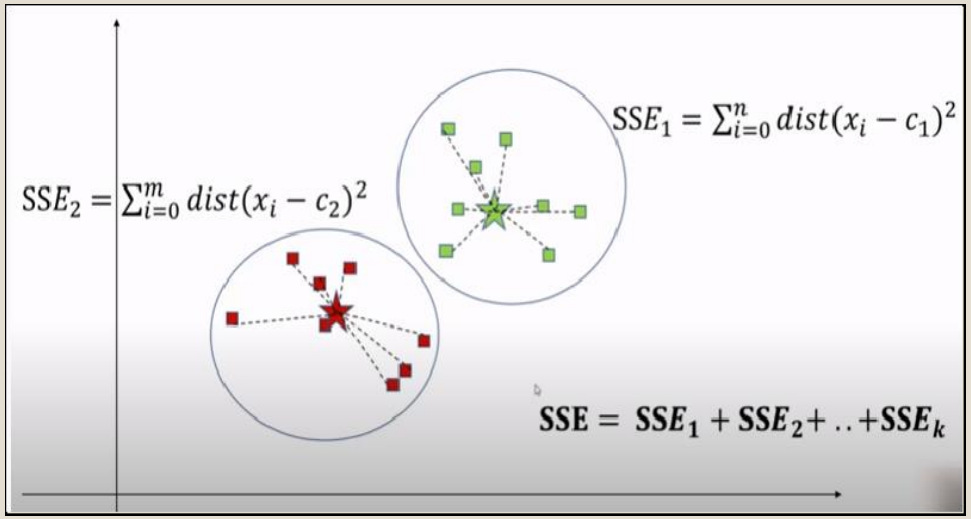
Let us consider a dataset as shown below where x and y represent the 2 different features and we are interested in finding the clusters in the dataset.

(1) Select the number of clusters for the dataset (K)

K is a free parameter that needs to be defined before the initiation of the algorithm.

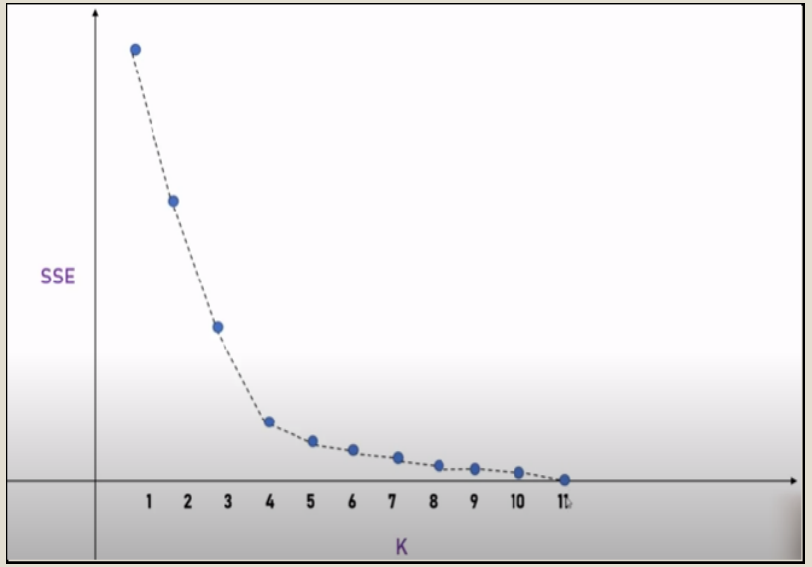
The number of K is estimated using the Elbow method. In the Elbow method, we start with some K and we try to compute the sum of squared errors.

Now, we plot these SSE = SSE1 (for K=1), SSE = SSE1 + SSE2 (for K=2),..., SSE = SSE1 + SSE2 + _ + SSEk (for K=k) as shown.



RFM CLUSTERING WITH K MEANS

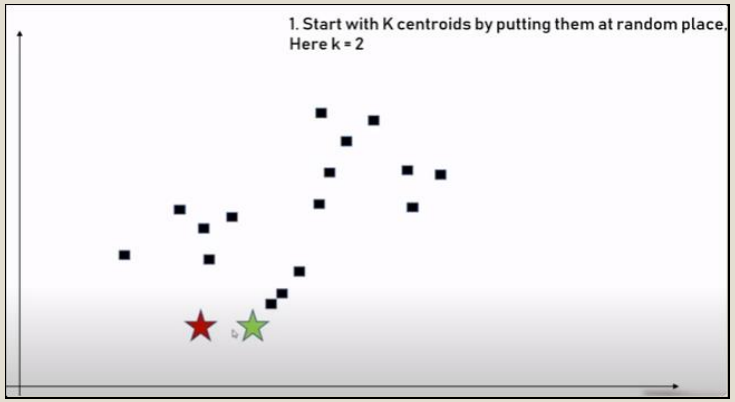
- From the plot, it can be inferred that as we increase the number of clusters, the errors get decreased. Here, the general guideline is to find out an elbow, thus the elbow on the above plot is at K = 4.
- Therefore, we can say a good number of clusters for the given dataset based on the Elbow method is 4.
- In the Elbow Method, we take that point as an elbow point which shows a sharp decrease in SSE but after that point, there is a gradual decrease.



RFM CLUSTERING WITH K MEANS

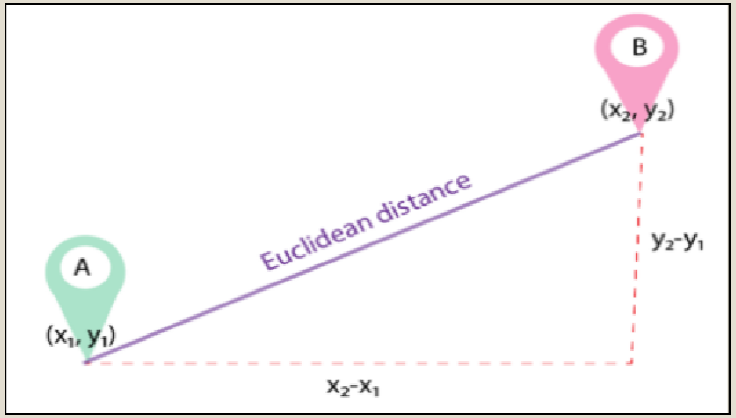
(2) Select K number of Centroids.

The next step is to identify K random points which you consider as the center of those K clusters respectively. These K random points are known as centroids.



(3) By calculating the Euclidean distance, assign the points to the nearest centroid, thus creating K clusters.

The next step is to compute the distance of every data point from the K centroids and cluster them accordingly. This distance in between the centroids and data points can be computed using Euclidean distance. A Euclidean distance is the straight line distance between 2 data points in a plane.



RFM CLUSTERING WITH K MEANS

(4) As soon as we're done associating each data point with its closest centroid, we re-calculate the means — the values of the centroids; the new value of a centroid is the sum of all the points belonging to that centroid divided by the number of points in the group. Again reassign the whole data point based on this new centroid, then repeat step 3 until the position of the centroid doesn't change.

Repeat the same process of re commuting the distance of each data point from the centroids until none of the data points change the cluster.

Finally, on performing the above stated 4 steps, we label the data points as the cluster they belong to.

Using K-means to segment customers based on RFM Variables

STEP 1: Let us first define the 3 key variables of the RFM analysis:

- Recency:

The last transaction date is 30–06–2017, we will use this data to calculate Recency.

	customerID	Recency
0	BBID_204100102	416
1	BBID_204100150	409
2	BBID_204100277	422
3	BBID_204100310	428
4	BBID_204100325	609

- Frequency

We will use Transaction No. by Till to calculate Frequency.

	customerID	Frequency
0	BBID_204100019	1
1	BBID_204100060	1
2	BBID_204100090	1
3	BBID_204100102	1
4	BBID_20410014	2

- Monetary

We will use the Transaction Amount to calculate Monetary.

	customerID	Monetary
0	BBID_204100102	20820.32
1	BBID_204100150	17390.13
2	BBID_204100277	23573.19
3	BBID_204100310	19626.70
4	BBID_204100325	1060.01

The RFM table obtained is as follows:

customerID	Recency	Frequency	Monetary
BBID_204100102	416	1	20820.32
BBID_204100150	409	1	17390.13
BBID_204100277	422	1	23573.19
BBID_204100310	428	2	19626.70
BBID_204100325	609	1	1060.01

STEP 2: We would perform scaling by the following procedure:

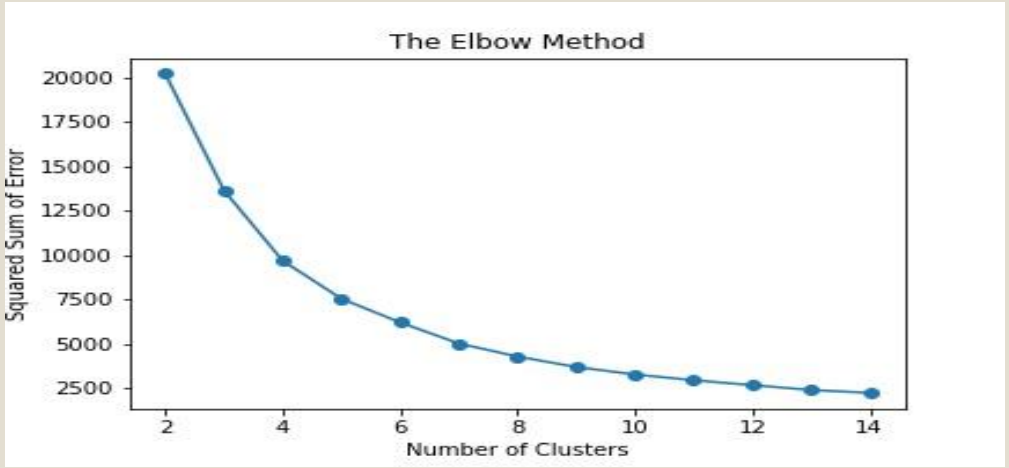
Scaling is the method used to standardize the range of features of data. Since the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms. We need to transform all the variables to one scale as we can see the range of values of each variable is very different. Since K-Means clustering is a distance-based algorithm, adjusting the range to a common range is required to avoid building biased models.

The normalized RFM variables are shown in the table below:

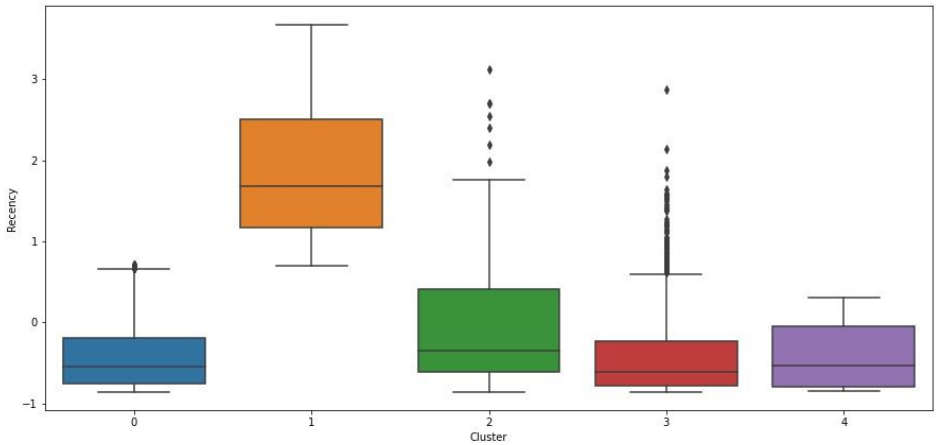
Recency	Frequency	Monetary
0.524652	0.000000	0.006407
0.515803	0.000000	0.005351
0.532238	0.000000	0.007254
0.539823	0.166667	0.006039
0.768647	0.000000	0.000326

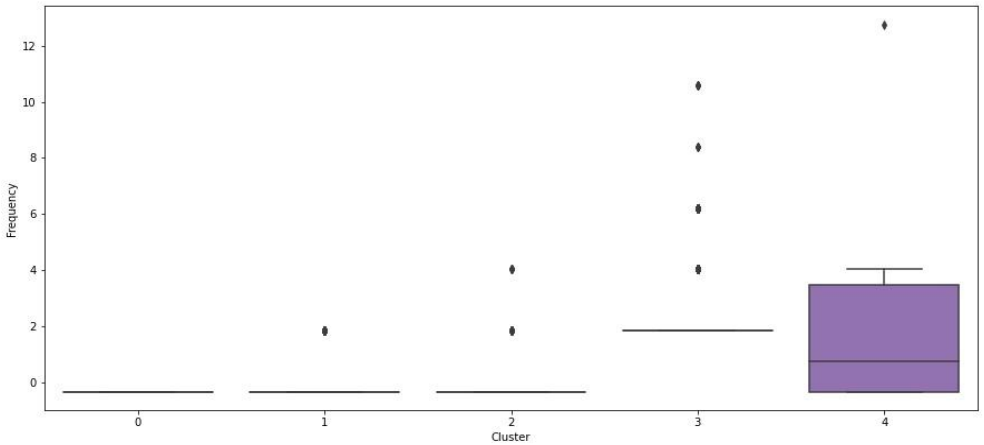
STEP 3: Application of K-means clustering:

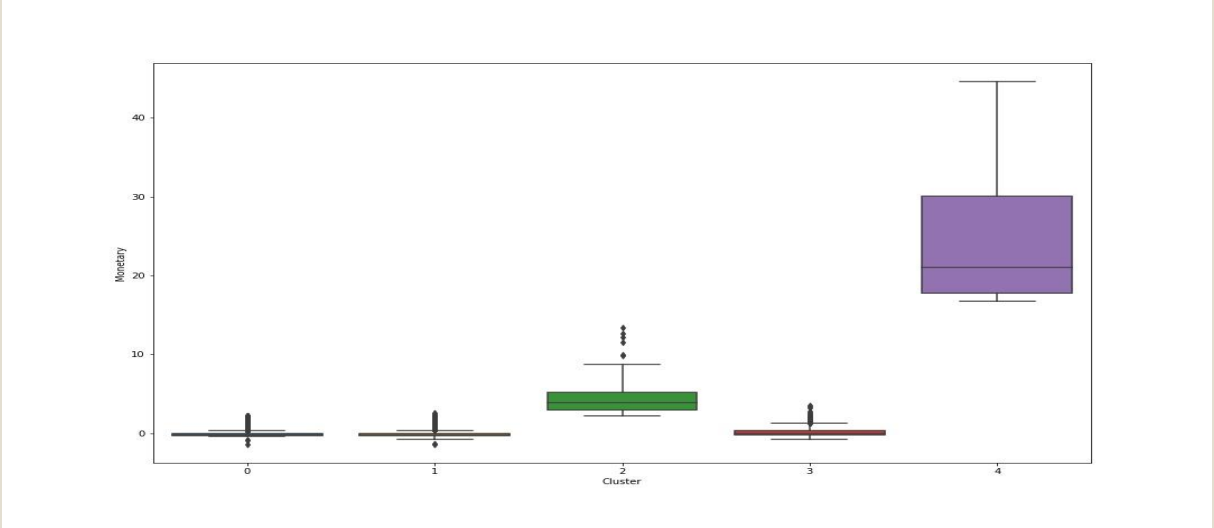
• Select the number of clusters for the dataset (K) We first initialise the number of clusters to be 12 and perform the Elbow method. Now, we plot these SSE = SSE1 (for K=1), SSE = SSE1 + SSE2 (for K=2),..., SSE = SSE1 + SSE2 + _ + SSE12 (for K=12) as shown below.



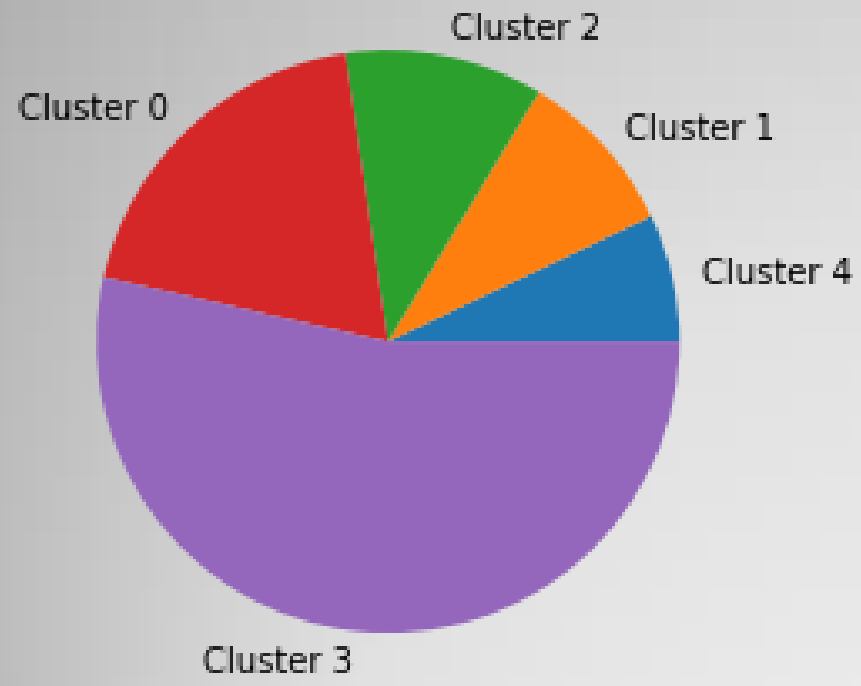
STEP 4: We will now see the individual cluster's characteristics separately on the basis of RFM variables.







Clusters and Customer Labels



- Cluster 3: Top Customers
- Cluster 4: High Value Customers
- Cluster 1: Medium Value Customers
- Cluster 0: Low Value Customers
- Cluster 2: Lost Customers

Silhouette Analysis

$$silhouette\ score = (p - q)/max(p, q)$$

Where, p is the mean distance to the points in the nearest cluster, and
q is the mean intra-cluster distance to all the points in its own cluster.

The value of the silhouette score range lies between -1 to 1 i.e A score closer to 1 indicates that the data point is very similar to other data points in the cluster, and score closer to -1 indicates that the data point is not similar to the data points in its cluster.

Silhouette score for our model = 0.9331683559417074

PROMOTIONAL STRATEGIES

To sustain and expand the business, one should realize being able to retain existing customers is as important as exploring new customers. If the rate of customers leaving is greater than the rate of new customers entering, our customers' database is actually shrinking. To a certain extent, we see customers retaining effort outweighs searching for new potential customers. Therefore, from the 5 clusters obtained using K-means to segment customers based on RFM Variables, our focus should be on the customers that lie in Cluster 3, Cluster 4, Cluster 1 and Cluster 0.

- For cluster 2 with low frequency, low monetary value and very old last transaction, the customers in this cluster(around 1095) are lost customers, rather we should focus upon other segments of customers.
- For cluster 3 with high frequency, high amount, very Recent transactions. These are the set of customers that brings recurrent revenue to our business. Hence, a plausible promotional strategy to retain them can be:
  - The introduction of "Loyalty Points" on every purchase can be redeemed after a certain amount is achieved.
- For cluster 4 with high monetary amount and recent transactions. These are the set of customers that brings good revenue to our business but are visiting the store after quite some time. Hence, a plausible promotional strategy to retain them can be:
  - The introduction of "Weekend/Monthly Discount coupons" to the customers that can be redeemed after a purchase of certain amount is achieved.

- For cluster 1 with low frequency, medium amount and not very recent transactions. These are the set of customers that can be turned into potential high value customers and bring revenue to our business.

  Hence, a plausible promotional strategy to retain them can be:

  - Providing them with a “12 - months discount coupon” that can be redeemed once per month on achieving a certain billing amount. This would ensure the association of the customers for a long period of time, as well as, would enable them to spend more in order to achieve the threshold billing amount to redeem the benefits of the coupon on a monthly basis. .

- For cluster 0 with low monetary amount and very old recent transactions. These are the set of customers that have the possibility of getting churned out if not taken care of Hence, a plausible promotional strategy to retain them can be:

  - The introduction of “Heavy Discounts and Everyday great deals coupons” for each customer, to ensure association of customers, as well as, to increase the monetary amount to redeem the benefits of the coupons .

LIMITATIONS

1. **Limited Production:** In each specific segment, customers are limited. So, it is not possible to produce products in mass scale for every segment. Therefore, company cannot take advantages of mass scale production; scale of economy is not possible. Product may be costly and affect adversely to the sales.

2. **Expensive Production:** Market segmentation is expensive in both production and marketing. In order to satisfy different groups/segments of buyers, producers have to produce products of various models, colors, sizes, etc., that result into more production costs.

3. **Expensive Marketing:** Market segmentation also results into expensive marketing. Due to different groups of buyers, the marketer has to consider all the segments in terms of needs, interests, habits, preferences and attitudes. Marketer has to formulate and implement several marketing strategies for different segments.

4. **Difficulty in Distribution:** Company needs to make the separate arrangement for each of the products demanded by different classes of customers. Salesman’s recruitments, selection, training, payments, and incentives are more difficult and costly. Company has to maintain separate channels and services for satisfying varied customer groups.

5. **Heavy Investment:** Market segmentation leads to heavy investment. In order to satisfy different needs and wants of various groups, a company has to produce variety of product lines and product items. For the purpose, the company requires to invest more on technology and other inputs that may demand heavy investment.

THANK YOU!