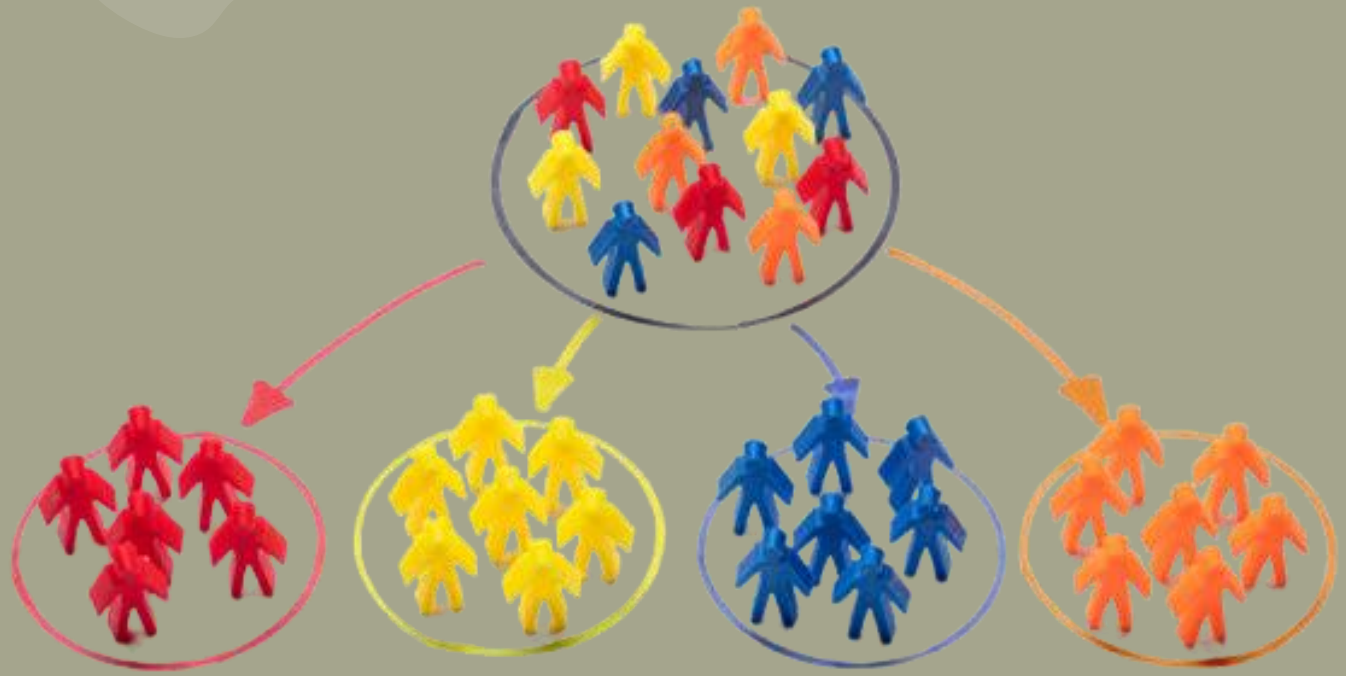
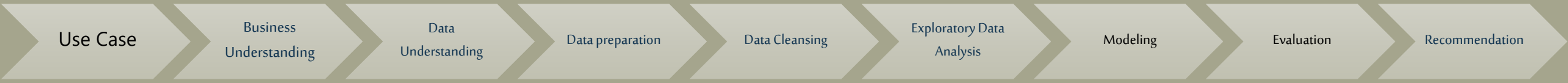


Capstone Project
Customer Segmentation

Project by Auti Rupali

Mind Map Customer Segmentation



Use Case: Customer Segmentation

Use Case Summary

Objective Statement:

- Get business insight about how many product sold every month.
- Get business insight about how much customer spend their money every month.
- To reduce risk in deciding where, when, how, and to whom a product, service, or brand will be marketed.
- To increase marketing efficiency by directing effort specifically toward the designated segment in a manner consistent with that segment's characteristics.

Source Data:

Online retail dataset
by UCI Machine Learning Library.
Customer Segmentation

Challenges:

- Large size of data, can not maintain by excel spreadsheet.
- Need several coordination from each department.
- Demography data have a lot missing values and typo.

Methodology / Analytic Technique:

- Descriptive analysis
- Graph analysis
- Segment Analysis

Business Benefit:

- Helping Business Development Team to create product differentiation based on the characteristic for each customer.
- Know how to treat customer with specific criteria.

Expected Outcome:

- Know how many product sold every month.
- Know how much customer spend their money every month.
- Customer segmentation analysis.
- Recommendation based on customer segmentation.

Business Understanding

- Retail is the process of selling consumer goods or services to customers through multiple channels of distribution to earn a profit.
- This case has some business question using the data:
- How many product sold every month?
- How much customer spend their money every month?
- How about Customer segmentation analysis?
- How about recommendation based on customer segmentation?

Data Understanding

- Data of Retail Transaction from 01 December 2010 to 09 December 2011
- Source Data: Online retail dataset by UCI Machine Learning Library. Customer Segmentation
- The dataset has 8 columns and 541909 rows.
- Data Dictionary:

➢ InvoiceNo: Invoice number uniquely assigned to each transaction.

➢ StockCode: Product (item) code.

➢ Description: Product (item) name.

➢ Quantity: The quantities of each product (item) per transaction.

➢ InvoiceDate: The day and time when each transaction was generated.

➢ UnitPrice: Product price per unit in sterling.

➢ CustomerID: Customer number uniquely assigned to each customer.

➢ Country: The name of the country where each customer resides.

Data preparation

Code Used:

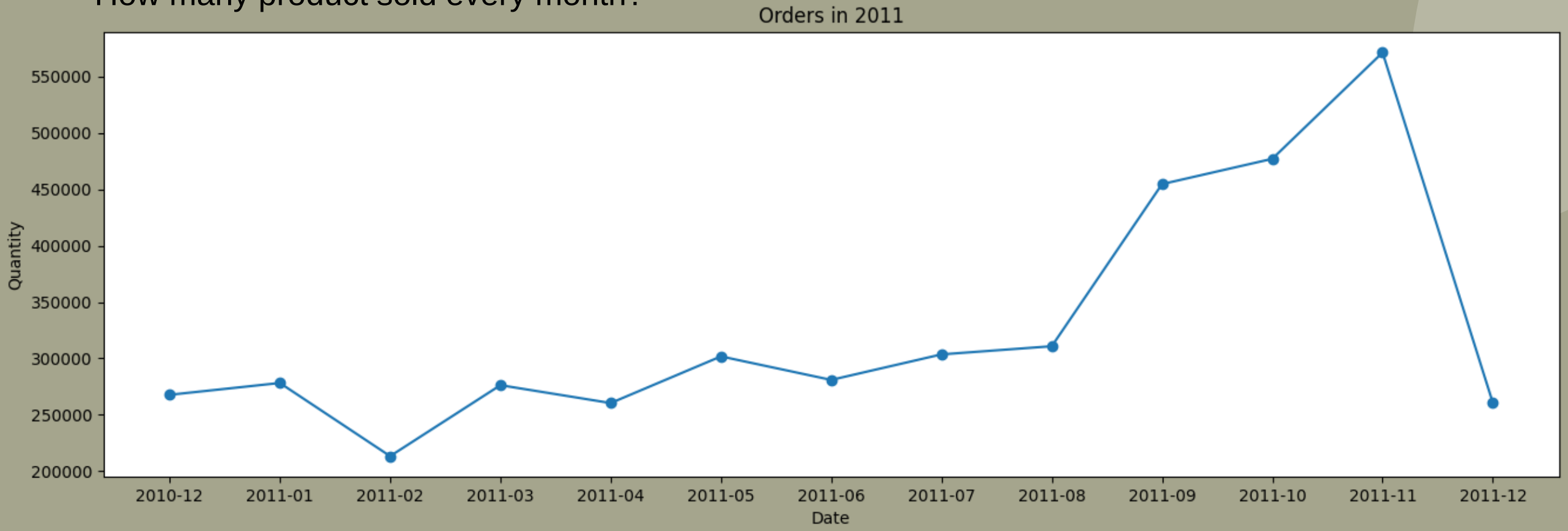
- Python Version: 3.7.6
- Packages: Pandas, Numpy, Matplotlib, Seaborn, Sklearn, and Feature Engine

Data Cleansing

- There are about 25% of Null CustomerID in the data. We need to remove them as there is no way we can get the number of CustomerID.
- There are few records with UnitPrice<0 and Quantity<0. We need to remove them from the analysis. This could represent cancelled or returned orders.
- There is more than 90% of 'United Kingdom' customers, therefore we will restrict the data to only United Kingdom customers.

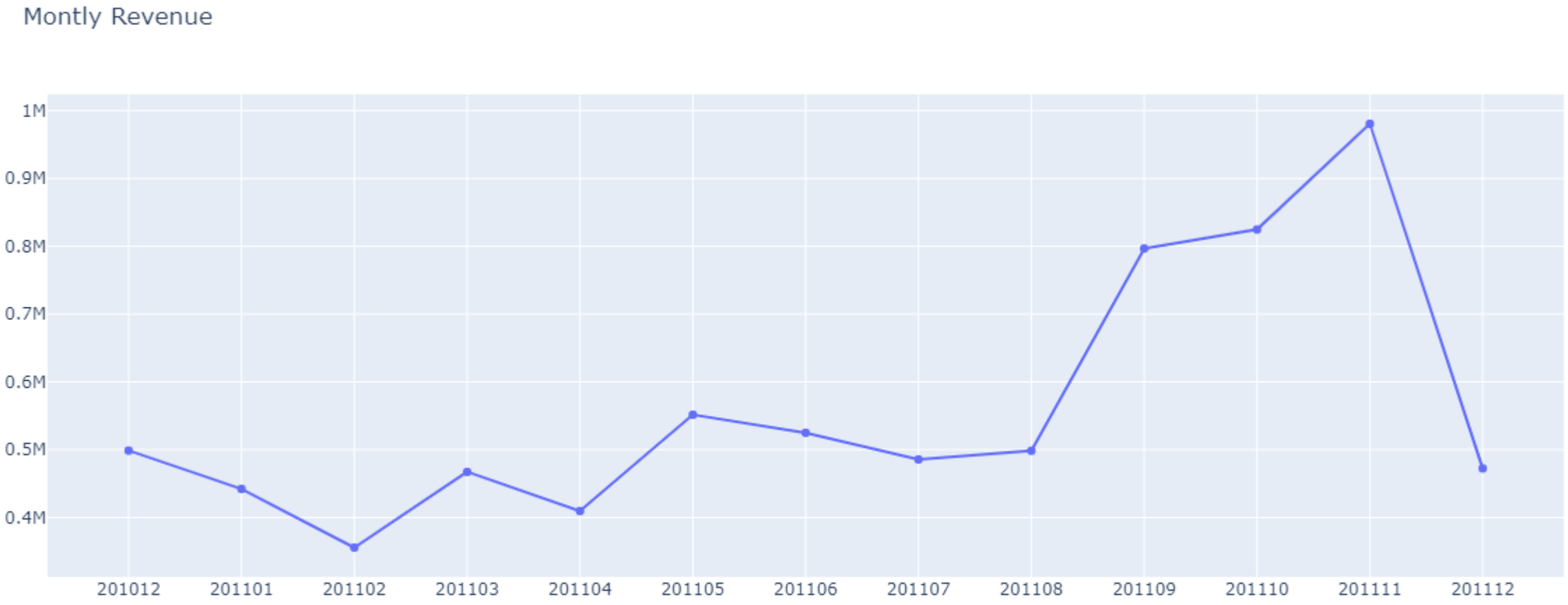
Exploratory Data Analysis

- How many product sold every month?



Product sold in November has highest quantity that has around 13,41% product sold from all transaction along 1 year. Therefore the business team can increase sales in this month such as promoting new products to customers in this month.

- How much customer spend their money every month ?

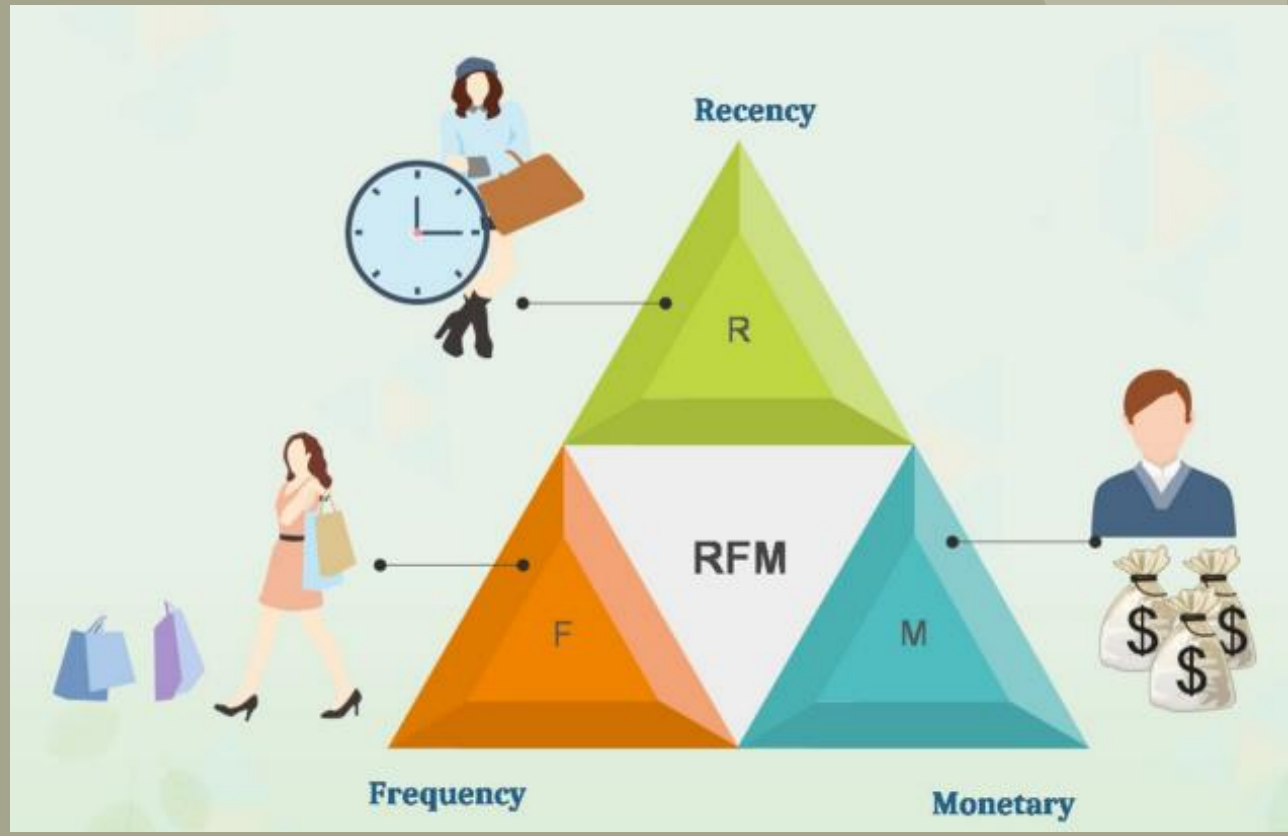


Revenue in November has highest amount that has 13,41% revenue from total revenue along 1 year. Therefore the business team can replicate the success of sales strategies in November to be implemented in other months

Modeling Data: RFM Quantiles

Recency Frequency Monetary (RFM)

RFM analysis allows you to segment customers by the frequency and value of purchases and identify those customers who spend the most money.



- Recency — how long it's been since a customer bought something from us.
- Frequency — how often a customer buys from us.
- Monetary value — the total value of purchases a customer has made.

Modeling Data: RFM Quantiles

- RFM Quantiles
- Split the metrics into segments using quantiles.
- We will assign a score from 1 to 4 to each Recency, Frequency and Monetary respectively.
- 1 is the highest value, and 4 is the lowest value.
- A final RFM score (Overall Value) is calculated simply by combining individual RFM score numbers.

Segment	RFM Score
Best Customers	111
Loyal Customers	F=1
Big Spenders	M=1
Almost Lost	134
Lost Customers	344
Lost Cheap Customers	444

Modeling Data: RFM Quantiles

Modeling Data: K-Means Clustering

- K-Means clustering algorithm is an unsupervised machine learning algorithm that uses multiple iterations to segment the unlabeled data points into K different clusters in a way such that each data point belongs to only a single group that has similar properties.
- K-means gives the best result under the following conditions:

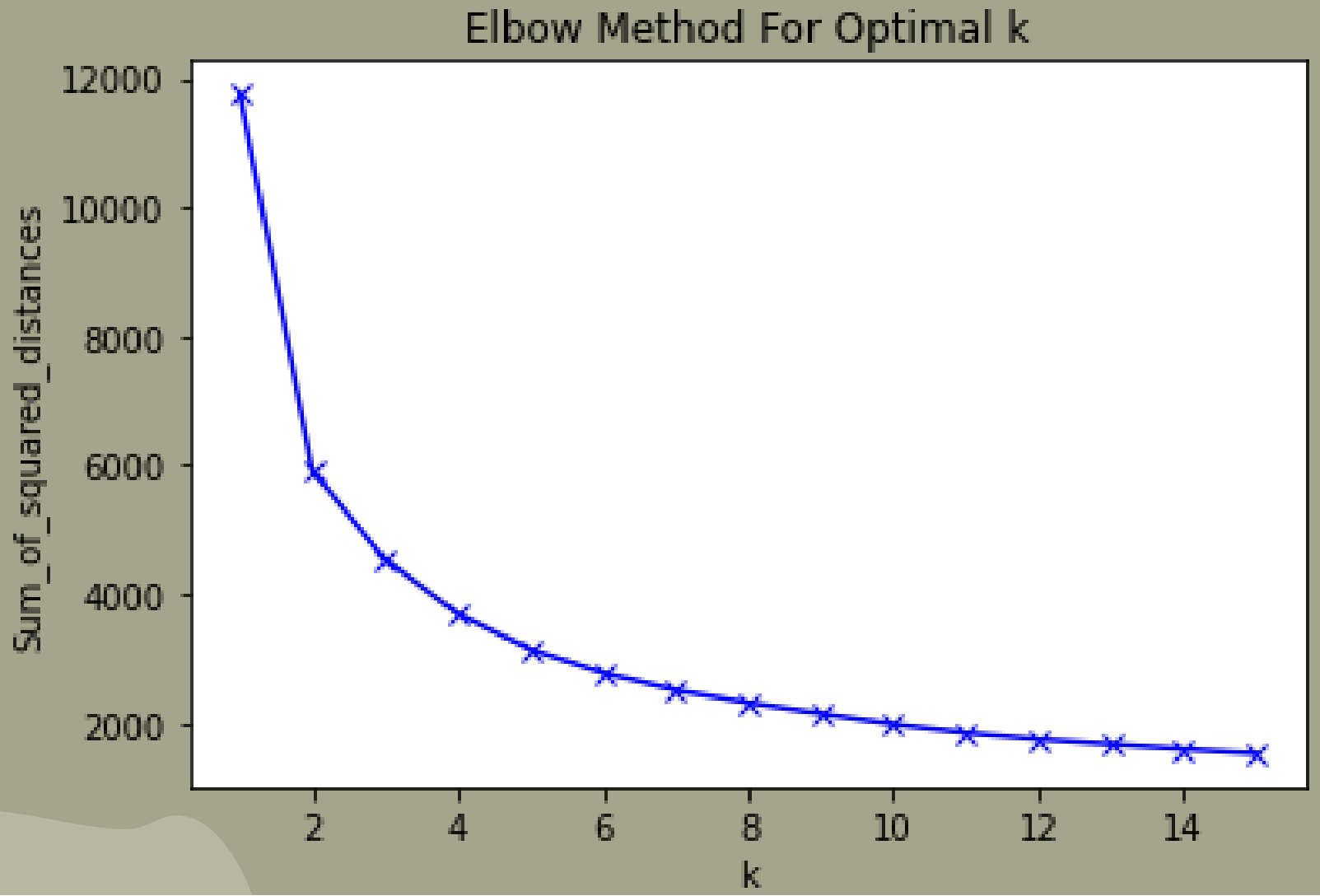
➢ Data’s distribution is not skewed

➢ Data is standardised

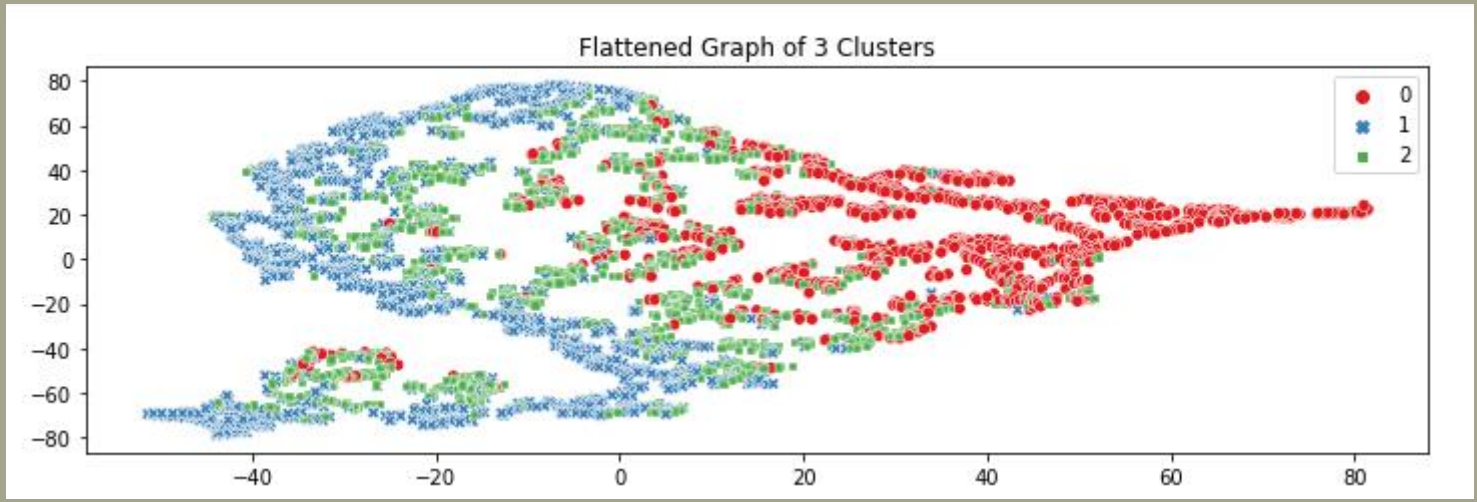
- The data is highly skewed, therefore I will perform log transformations to reduce the skewness of each variable and I standardised the data.

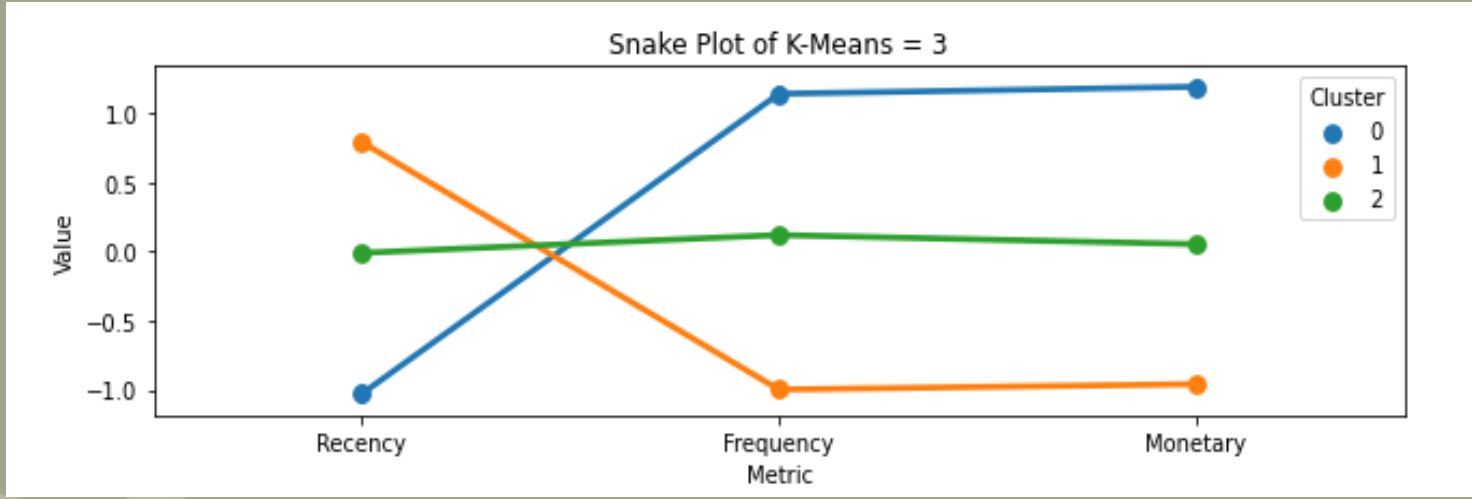
Modeling Data: K-Means Clustering

- Finding the optimal number of clusters (Finding K) using Elbow Method.

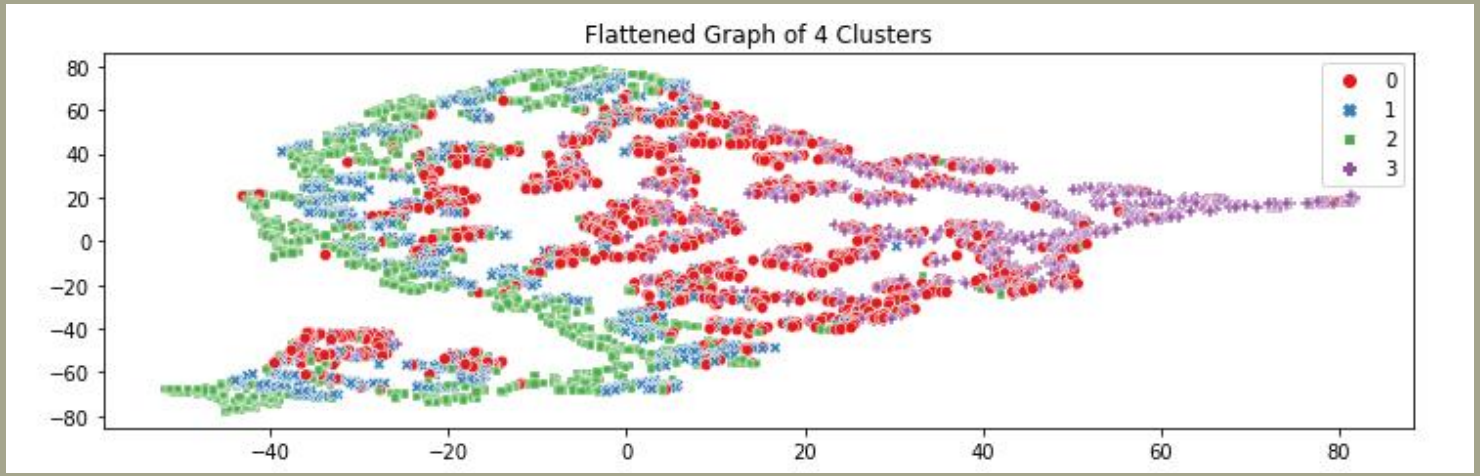


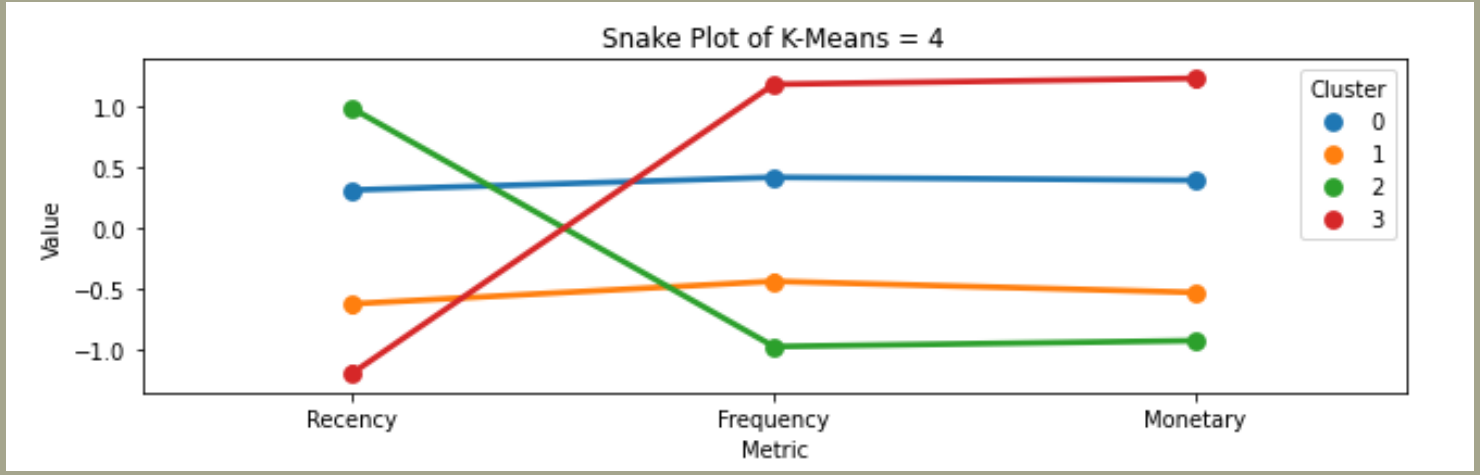
Modeling Data: K-Means Clustering



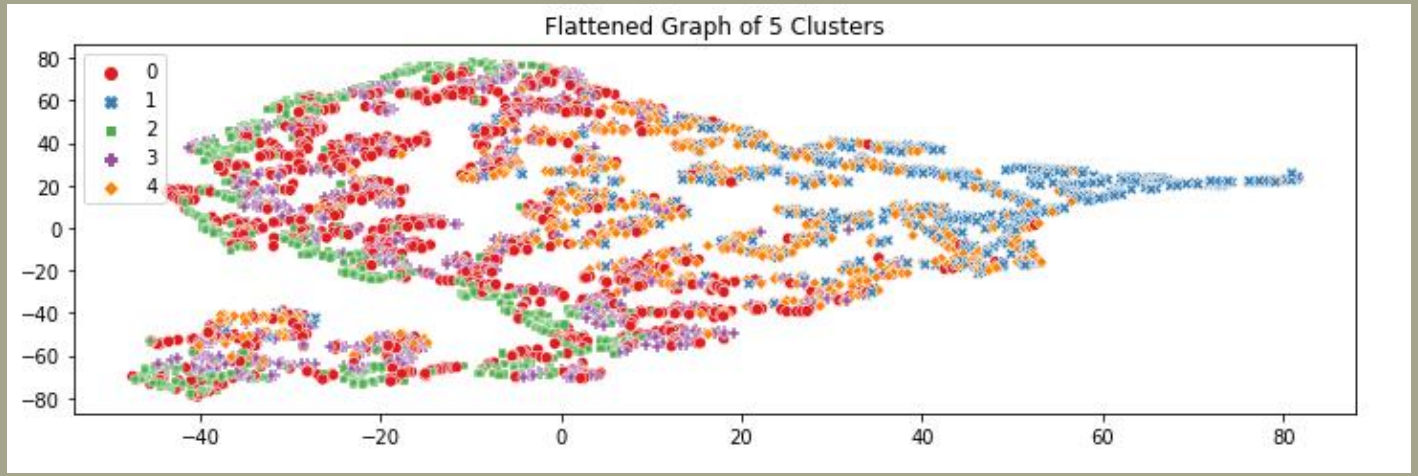


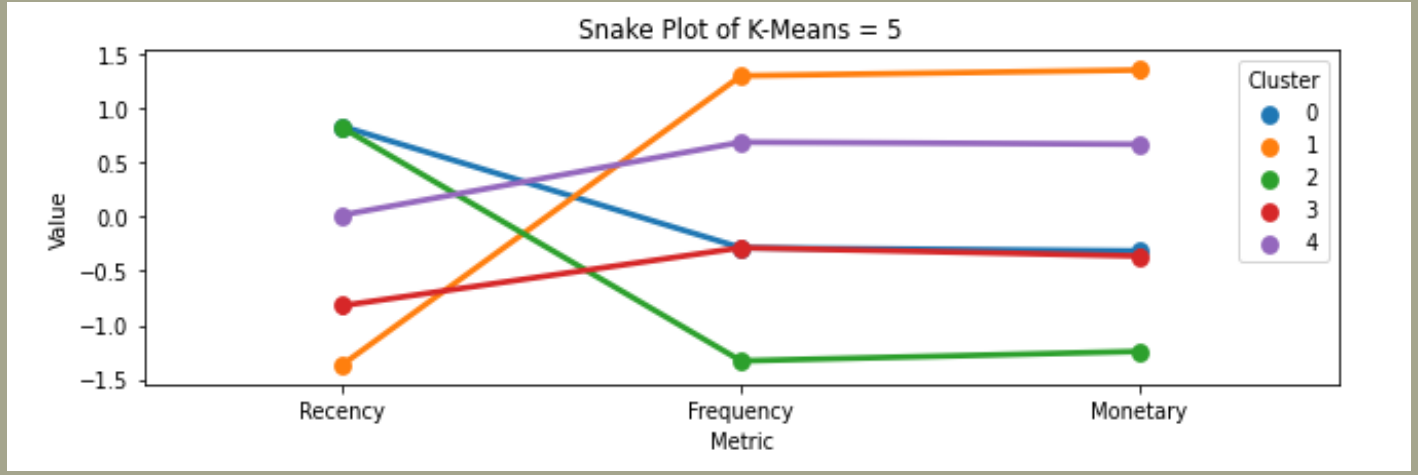
Modeling Data: K-Means Clustering





Modeling Data: K-Means Clustering





Evaluating Model: K-Means Clustering

Davies Bouldin Score is a metric for evaluating clustering algorithms.

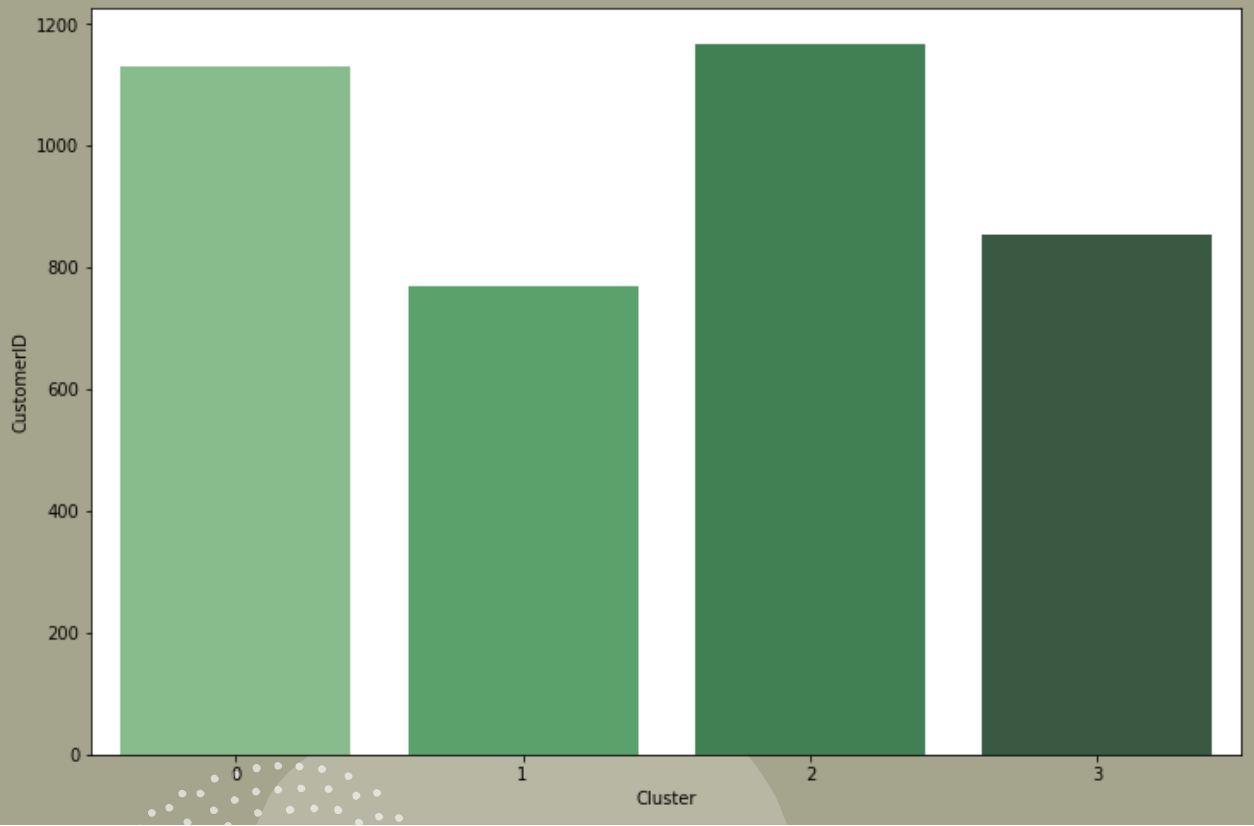
The smaller Davies Bouldin Score is The more optimal the cluster.

K-Means Cluster	Davies Bouldin Score
3	1.119
4	1.065
5	1.067

K-Means 4 clusters has lowest Davies Bouldin Score than other cluster.
Therefore the optimum cluster is 4.

K-Means 4 Clusters

Loyal Customers	: 29%
Almost Lost	: 20%
Lost Cheap Customers	: 30%
Best Customers	: 21%



Interpretation of the clusters formed using k-means:

- "Cluster 0" has 29% customers. It belongs to the “Loyal Customer" segment as they Haven’t purchased for some time, but used to purchase frequently (F=2) and spent a lot.
- "Cluster 1" has 20% customers. It can be interpreted as “Almost Lost". They purchase recently (R=2). However they do not purchase frequently and do not spent a lot.
- "Cluster 2“ has 30% customers. It can be interpreted as "Lost Cheap Customers". Their last purchase is long ago (R=4), purchased very few (F=4) and spent little (M=4).
- "Cluster 3“has 21% customers. It belongs to the "Best Customers" segment which we saw earlier as they purchase recently (R=1), frequent buyers (F=1), and spent the most (M=1).

Recommendation

- Recommendation for “Best Customers" segment:

Focus on increasing customer purchases therefore it is necessary to form a cross/Up Selling Strategy.

- Recommendation for “Loyal Customers" segment:

The business team must optimize the budget campaign and the time campaign for this customer segment in order to maintain their loyalty and increase their value.

- Recommendation for “Almost Lost" segment:

This customer segment is very at risk for churn, so focus on activating customers and making repurchases by forming a Reactivation Strategy, Retention Strategy.

- Recommendation for “Lost Cheap Customers" segment:

This customer segment has churned, so the focus of the campaign is to reactivate the customer by forming a Reactivation strategy.

Thank You

autirupalivikas@gmail.com

Rupali Auti

github.com/RupaliAuti09