

**CAPSTONE PROJECT**

**ONLINE RETAIL STORE ANALYSIS**

Submitted by

SIRISHA PENUMUTCHU

**Table of Contents:**

1. Problem Statement

2. Project Objective

3. Data Description

4. Data Pre-processing Steps and Inspiration

5. Choosing the Algorithm for the Project

6. Motivation and Reasons For Choosing the Algorithm

7. Assumptions

8. Model Evaluation and Techniques

9. Inferences from the Same

10. Future Possibilities of the Project

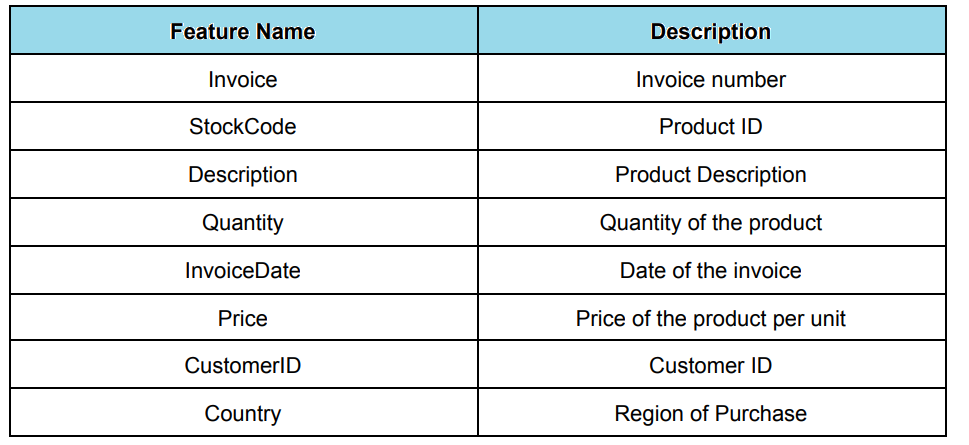
11. Conclusion

12. References

**Problem Statement 2**

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence based insights to provide the same.

Data Information:



**This dataset contains ~5.4 lakh records and 8 attributes.**

****

**Problem objective:**

1. Objective is to use the above data, find useful insights about the customer purchasing history that can be an added advantage for the online retailer.

2. Segment the customers based on their purchasing behavior using clustering algorithms.

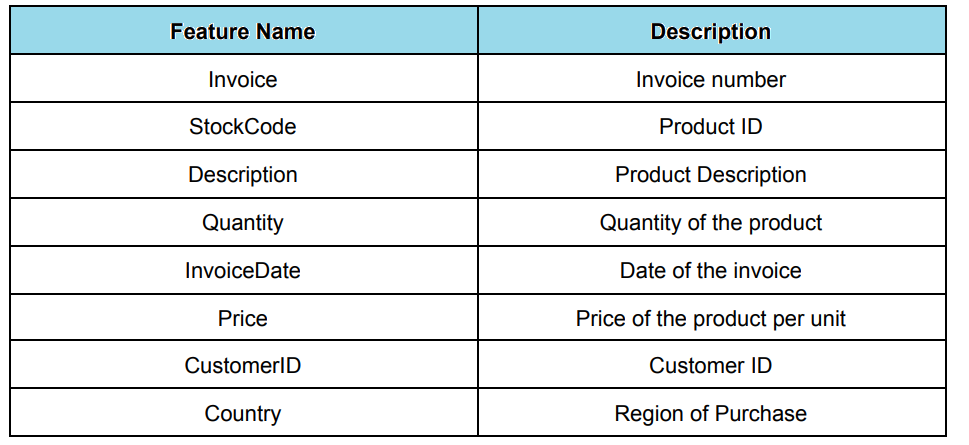
* Understand the dataset and features
* Use suitable Data Preprocessing and Feature Selection/Engineering Methods
* Fine tune the model and hyper parameters and finalize the Model
* Make the model deployment-ready by giving User-Input provision

**Approach**

* Importing all libraries which we needed.
* Perform data preprocessing technique to get balanced structured data.
* Perform statistical data analysis and derive valuable inference.
* Perform exploratory data analysis and derive valuable inference.
* Visualizing things with some plot and derive valuable inference.
* Train and test through Various different models.
* Choose the best model
* Perform hyperparameter tuning
* Make predictions based on the best model

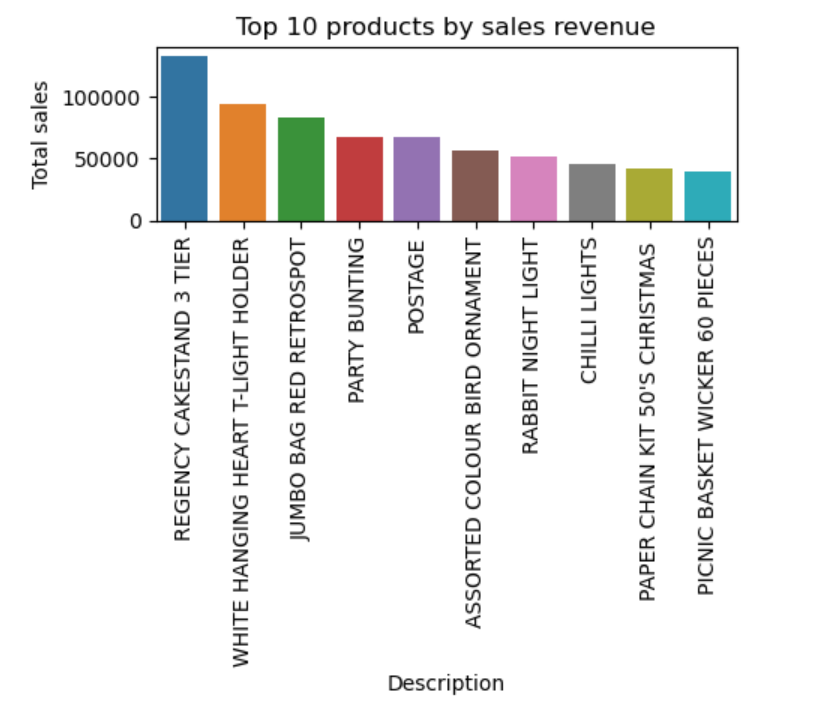
**Data Description**

This dataset contains ~5.4 lakh records and 8 attributes.

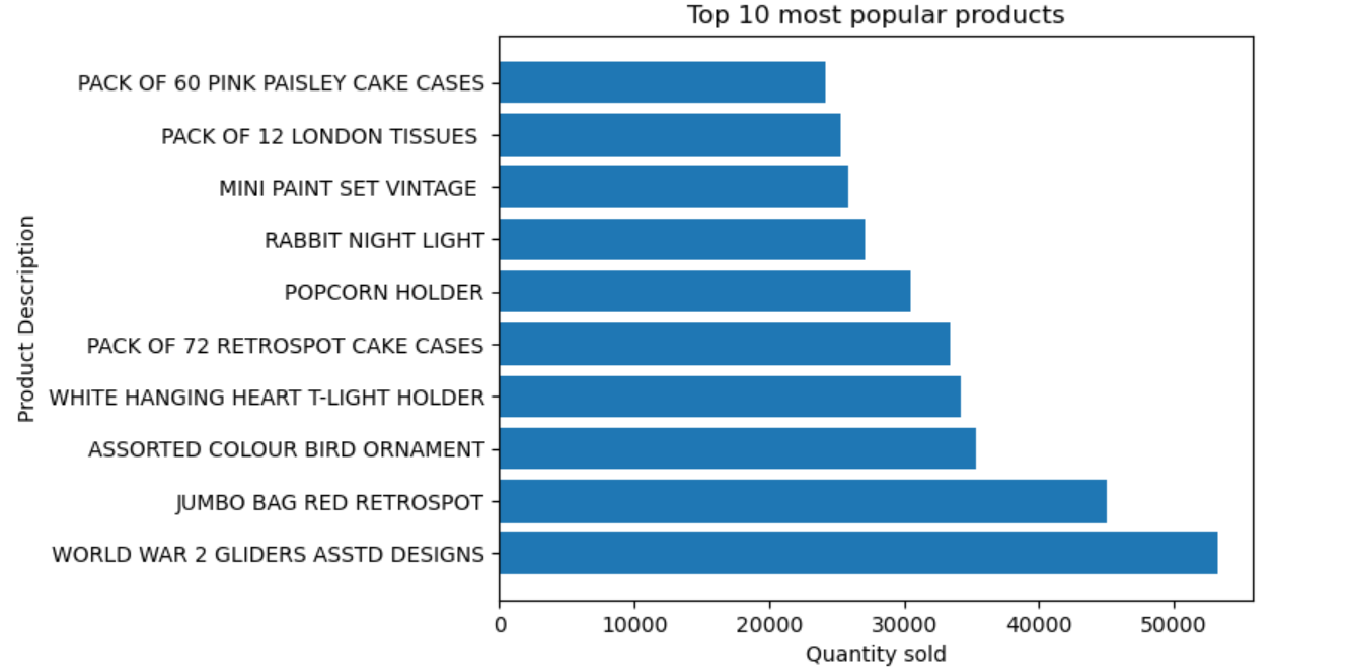


**Data Visualization:**

Bar plot showing the top 10 products with respect to sales revenue.

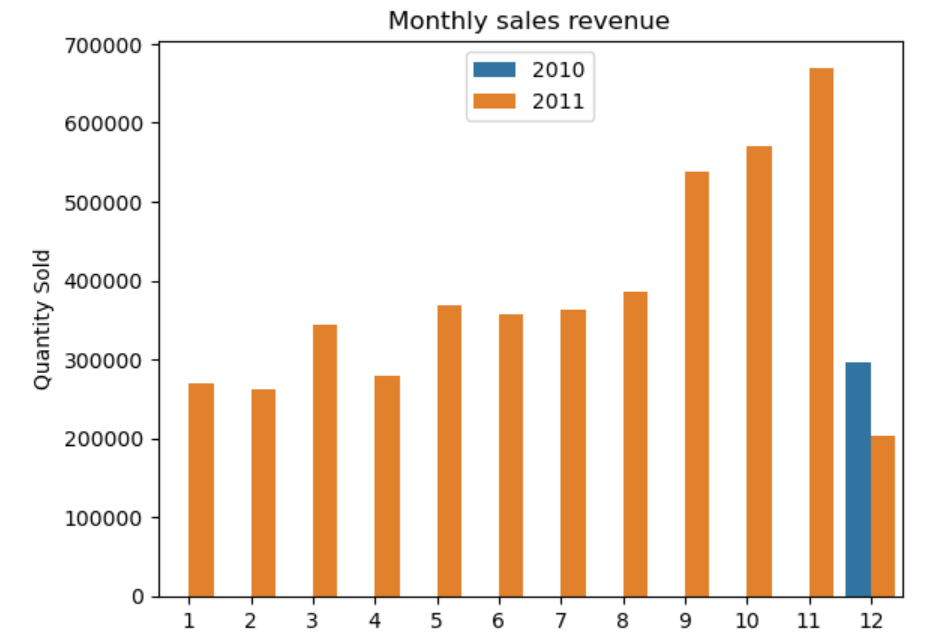


Bar plot for the Top 10 most popular products vs Quantity sold:

****

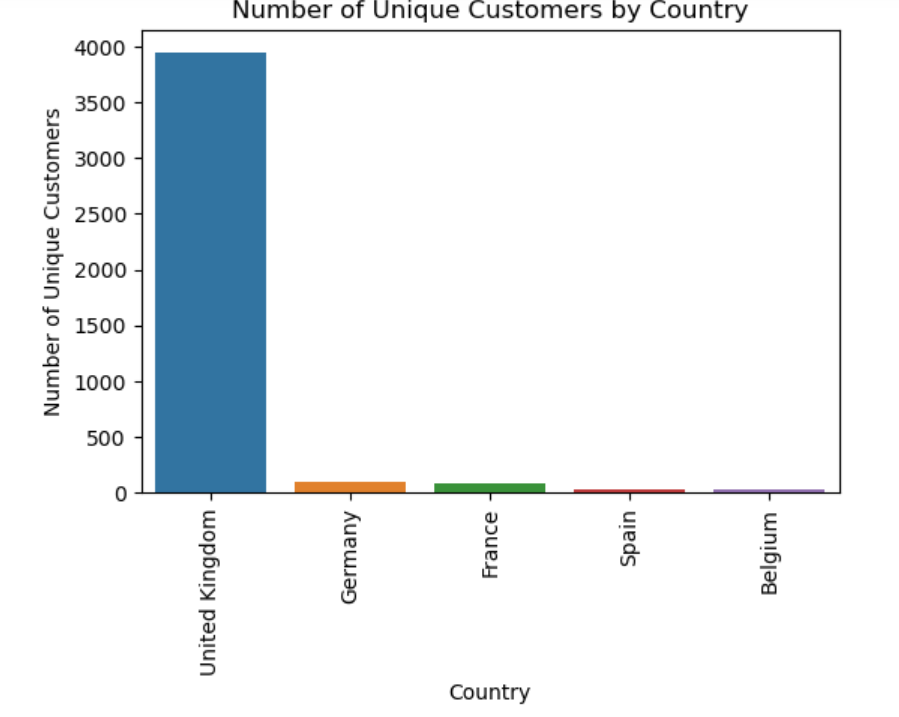
World War 2 Gliders ASSTD Designs and Jumbo Bag Red Retrospot are the top2 products with respect to highest quantity sold.

Bar plot for Monthly sales revenue vs Quantity sold



Data from December 2010 to December 2011 has been provided,with highest number of products sold in the month of Novemer 2011.

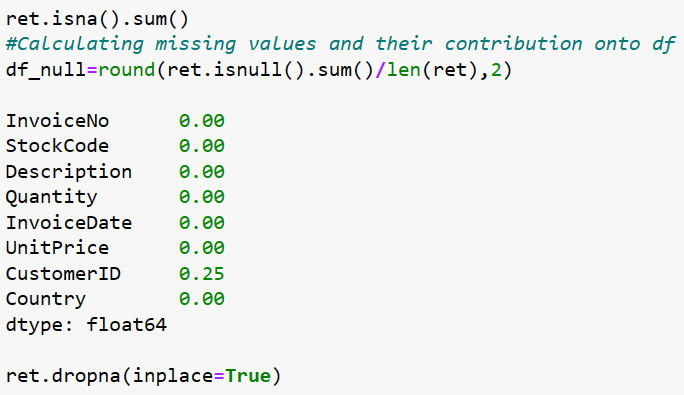
Bar plot showing the number of unique customers by country.

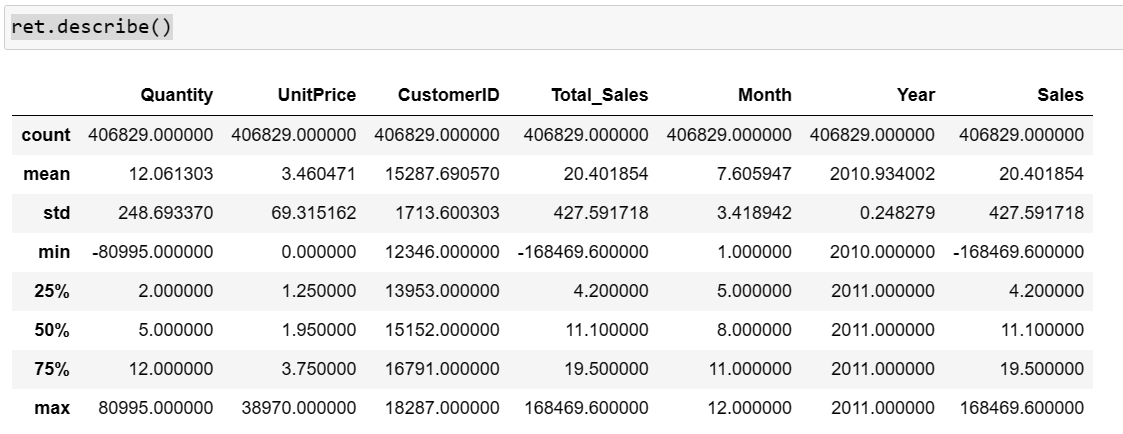


Customers are from United Kingdom.Although very few are from Germany,France,Spain and Belgium.

**Data Preprocessing steps and inspiration**

Step 1: **Checking for missing values and handling them**:

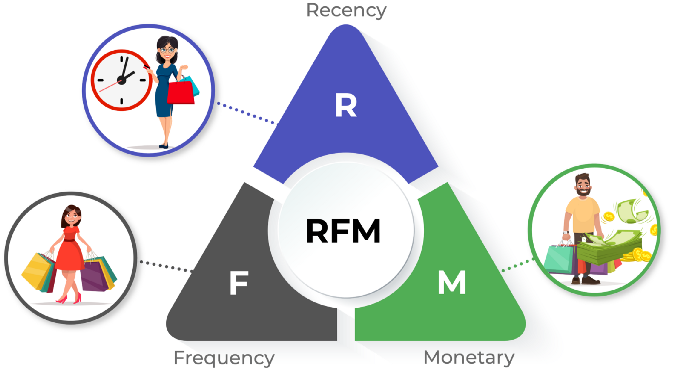


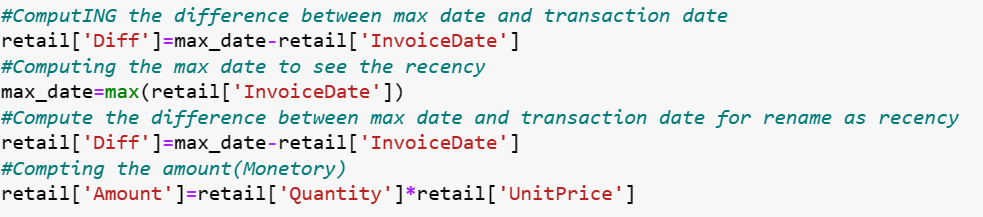


The negative values refer to the return items.Hence can be handled during outlier detection.

Step 2:**Data Preparation**:

Data Preparation We are going to analyse the Customers based on below 3 factors: **R (Recency)**: Number of days since last purchase **F (Frequency):** Number of transactions **M (Monetary)**: Total amount of transactions (revenue contributed).





Intuition behind choosing RFM analysis:

RFM Analysis is a commonly used CRM technique for years to group customers depending on their Recency, Frequency, and Monetary metrics.

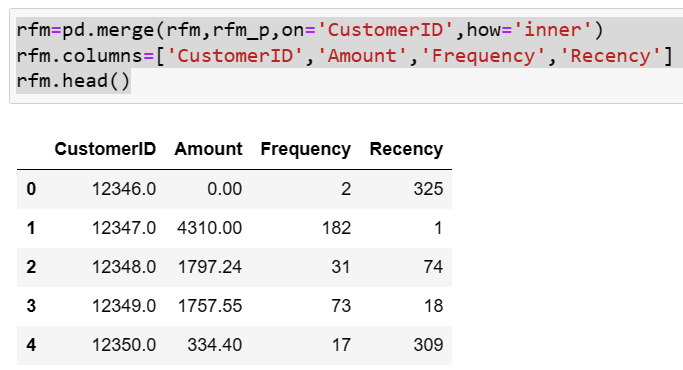
The recency metric stands for, the time elapsed since the most recent transaction the customer made. In other words, this is the freshness of a customer’s rate of giving you what you want. A good recency score for a customer is a low recency score. The best customers for us are the customers we remember their faces easier!

The frequency metric stands for, how many purchases have been made in the time range the analysis covers. As seen in its name, this is the frequency rate for the customer to give you what you want. This differs from time to time because “what you want” can sometimes be the transactions, the purchases, the visits, etc. For RFM, the more transactions the better customer is!  
Monetary metric stands for, the monetary size of the customer’s total transactions made in the time range the analysis covers. A higher spender is a better spender for RFM.

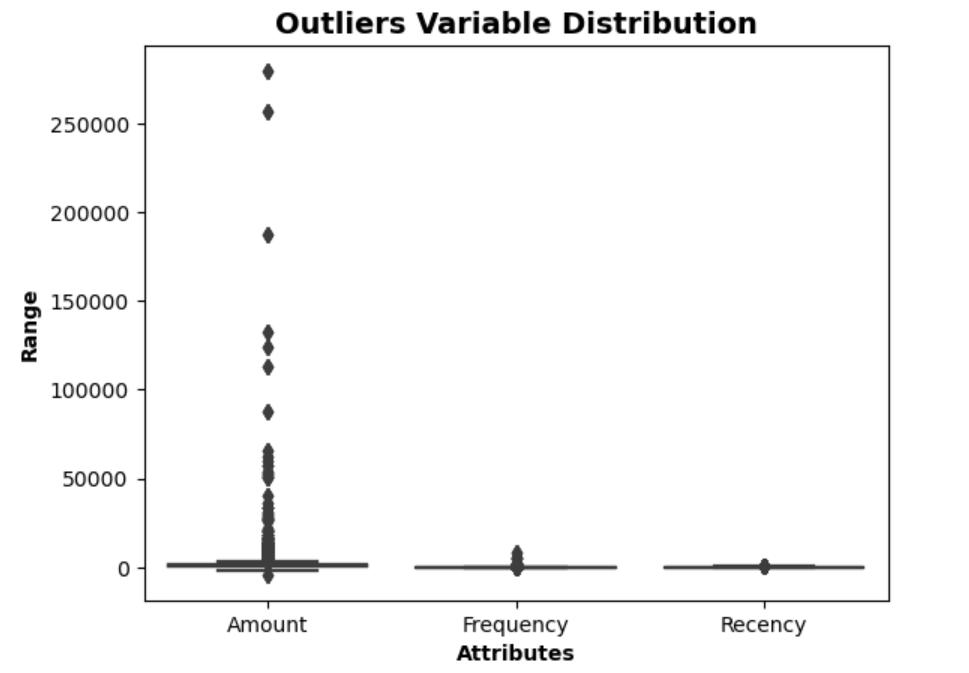
Recency: How many days ago was their last purchase?

Frequency: How many times has the customer purchased from our store?

Monetary: How much has the customer spent?

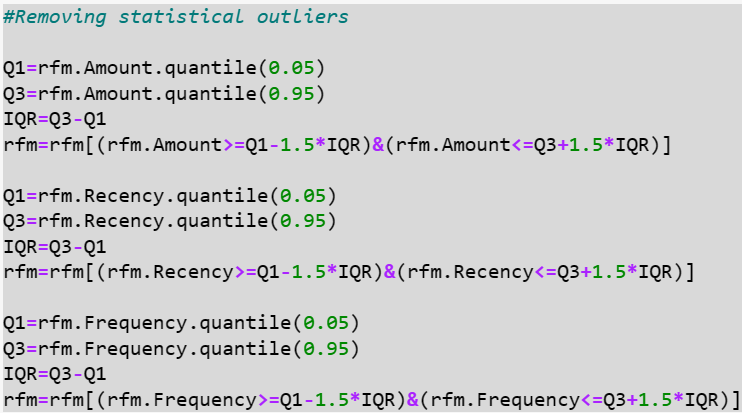


Step 3:**Outlier detection:**



**Using the Interquartile Rule to Find Outliers:** Though it's not often affected much by them, the interquartile range can be used to detect outliers. This is done using these steps:

1. Calculate the interquartile range for the data.
2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

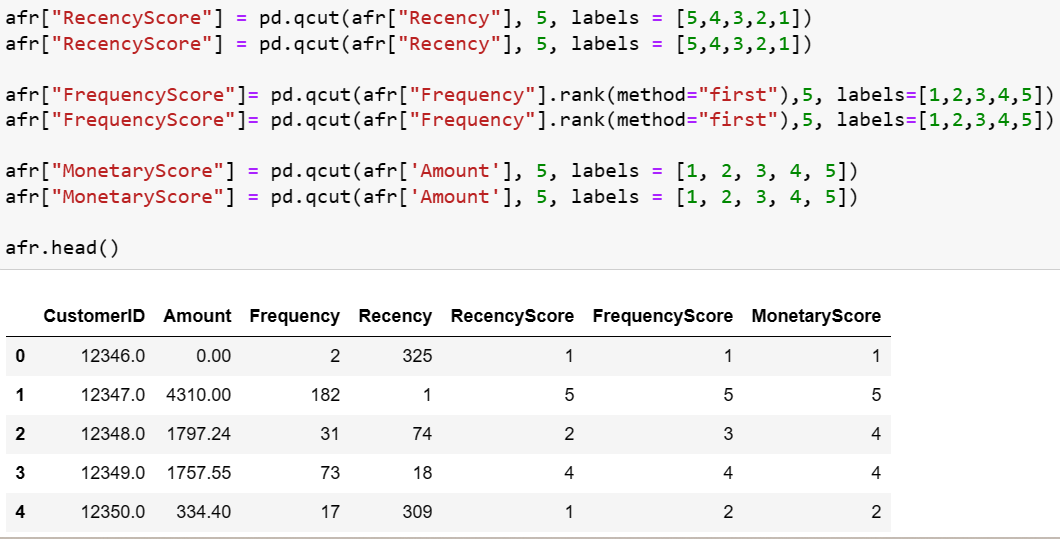


**Step:4: Assigning RMF scores for each record:**

Interpreting one of these three variables(RMF)separately does not tell US too much. For instance, a client who made a significant purchase six months ago cannot be identified as an essential customer. However, with a combination of the RFM, called RFM score, it is easier to segment it.

To conduct you RFM analysis, the first step is to rank the three variables with a score of 1 (low) to 5(important). With the variable frequency, score 1 will gather 20% of people who come less, whereas 5 means the customers who shop the most frequently. Once you assess all the rank, you concatenate them. It assigns a three-digit RFM score (from 111 to 555) to each customer. A 555 indicates that a customer has purchased a product or service very recently, most frequently, and at the highest monetary value.

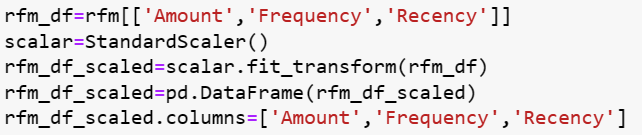
RFM is a powerful tool during your marketing campaign. It tells you which segmented group is more likely to respond to your current campaign, which client they should pay more attention to. Therefore, companies decrease the churn rate, focus on the potential and profitable customers.



Step:5: **Standardisation of data**:

Rescaling the Attributes is extremely important in rescaling the variables so that they have a comparable scale.| There are two common ways of rescaling:

Min-Max scaling Standardisation Here, we will use Standardisation Scaling.



**Choosing the Algorithm:**

Since it is an unsupervised problem,we can go for clustering techniques.

### **K-Means Clustering**

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

**Why K means clustering has been choosen:**

When to use k-means clustering in place of another type of clustering algorithm? Here are some examples of cases where we should consider using k-means clustering.

**Large datasets** : If you are working with a large dataset with many observations like this, you should go with k-means clustering than other clustering algorithms. K-means clustering is relatively **fast** compared to some other clustering algorithms.

**Popular and well studied:** The reason that k-means clustering has so many implementations across a variety of languages and libraries is that it is probably the most popular and well-studied clustering algorithm out there. This popularity confers some benefits of its own, as it will make it easier for other contributors to jump in to assist or even take over an ongoing project. If the model is going to be used to score data repeatedly, using a well studied algorithm will also reduce the burden of maintenance.

**Working of K-Means Algorithm**

The following stages will help us understand how the K-Means clustering technique works-

**Step 1**: First, we need to provide the number of clusters, K, that need to be generated by this algorithm.

**Step 2**: Next, choose K data points at random and assign each to a cluster. Briefly, categorize the data based on the number of data points.

**Step 3**: The cluster centroids will now be computed.

**Step 4**: Iterate the steps below until we find the ideal centroid, which is the assigning of data points to clusters that do not vary.

4.1 The sum of squared distances between data points and centroids would be calculated first.

4.2 At this point, we need to allocate each data point to the cluster that is closest to the others (centroid).

4.3 Finally, compute the centroids for the clusters by averaging all of the cluster’s data points.

K-means implements the Expectation-Maximization strategy to solve the problem. The Expectation-step is used to assign data points to the nearest cluster, and the Maximization-step is used to compute the centroid of each cluster.

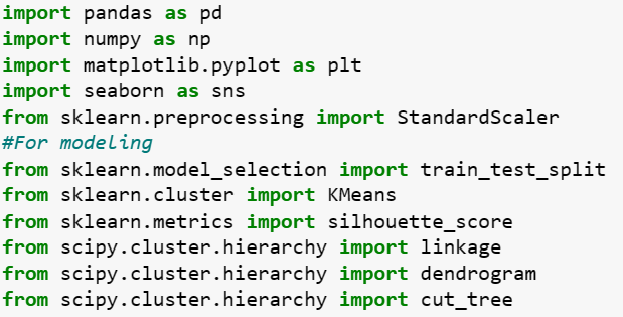
K-Means Clustering Applications: Distinct models will be created for different subgroups in a cluster-then-predict approach.

* Market segmentation
* Document Clustering
* Image segmentation
* Image compression
* Customer segmentation
* Analyzing the trend on dynamic data

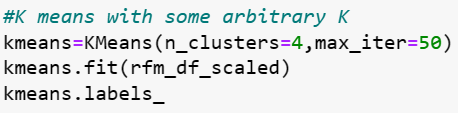
**Model Building:**

**1)Importing the libraries:**

We start off this model by importing all the necessary libraries that will be required for the process.

****

2)**Initiating the model with randomly choosing k:**



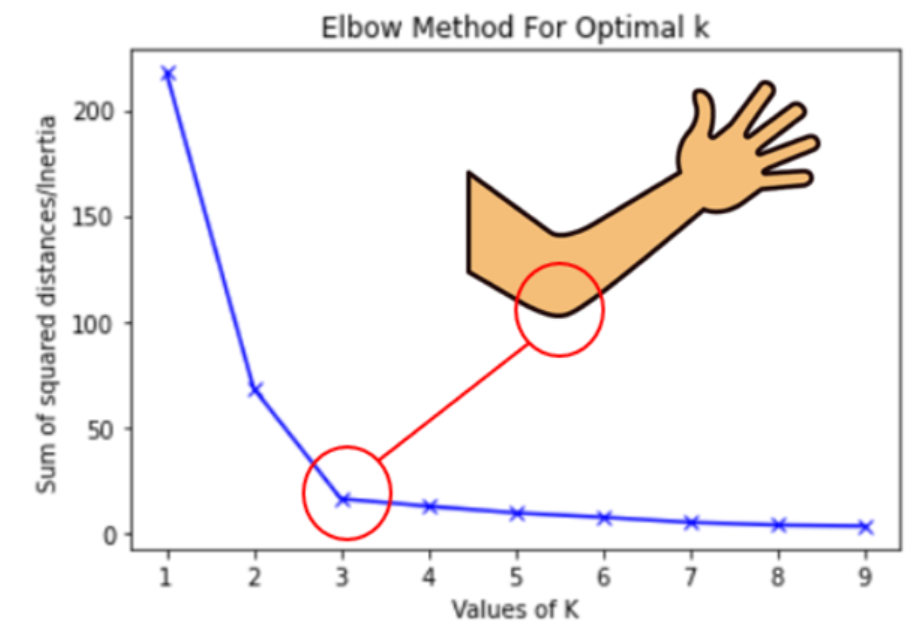
# **Model Evaluation and Techniques:**

# **K-Means: Getting the Optimal Number of Clusters.**

# **1)Elbow Curve Method**

Recall that the basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total wss measures the compactness of the clustering, and we want it to be as small as possible. The elbow method runs k-means clustering (kmeans number of clusters) on the dataset for a range of values of k (say 1 to 10) In the elbow method, we plot mean distance and look for the [elbow point](https://blogs.oracle.com/ai-and-datascience/post/introduction-to-k-means-clustering)where the rate of decrease shifts. For each k, calculate the total within-cluster sum of squares (WSS). This elbow point can be used to determine K.

* Perform K-means clustering with all these different values of K. For each of the K values, we calculate average distances to the centroid across all data points.
* Plot these points and find the point where the average distance from the centroid falls suddenly (“Elbow”).



#### 

#### 

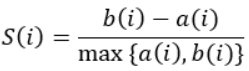
Still unsure of the elbow?Lets check the Silhouette method.

***2)Silhouette Analysis***

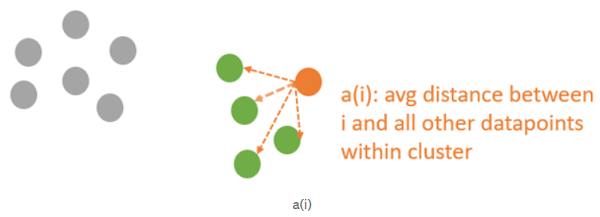
The silhouette coefficient or silhouette score kmeans is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation). The Silhouette score can be easily calculated in Python using the metrics module of the scikit-learn/sklearn library.

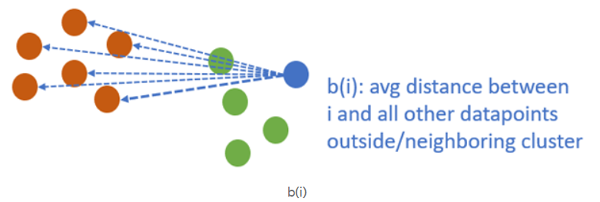
* Select a range of values of k (say 1 to 10).
* Plot Silhouette coefﬁcient for each value of K.

The equation for calculating the silhouette coefﬁcient for a particular data point:



* S(i) is the silhouette coefficient of the data point i.
* a(i) is the average distance between i and all the other data points in the cluster to which i belongs.
* b(i) is the average distance from i to all clusters to which i does not belong.



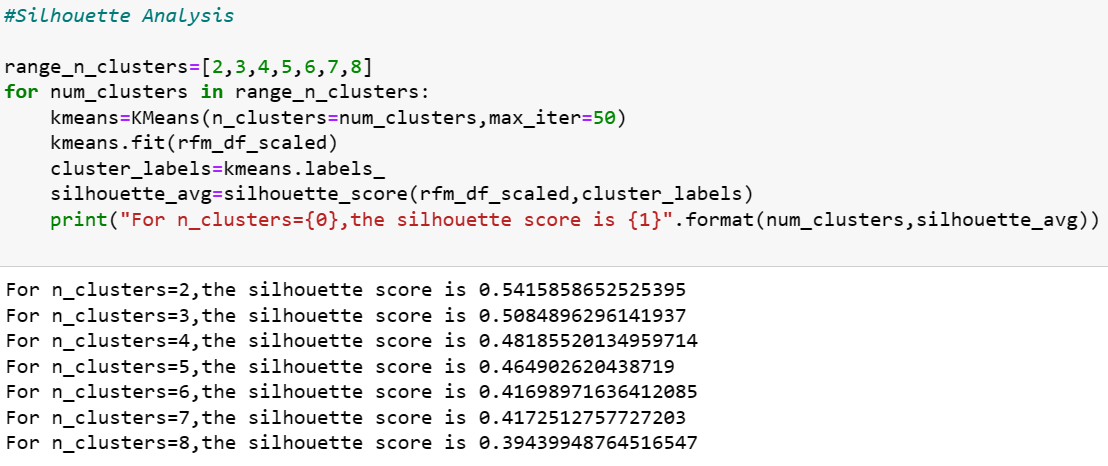


We will then calculate the average\_silhouette for every k.

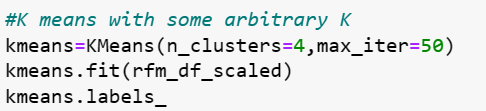
Avg silhouette

Then plot the graph between average\_silhouette and K.Points to Remember While Calculating Silhouette Coefficient:

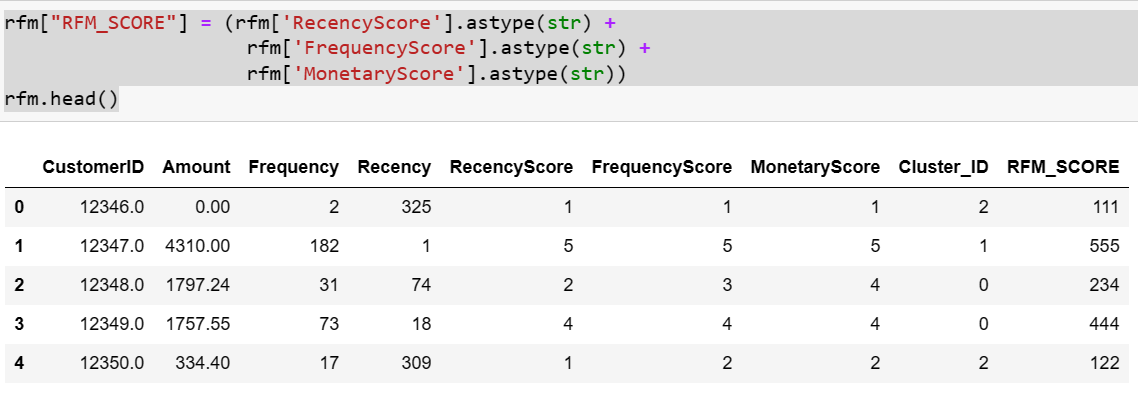
* The value of the silhouette coefﬁcient is between [-1, 1].
* A score of 1 denotes the best, meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters.
* The worst value is -1. Values near 0 denote overlapping clusters.



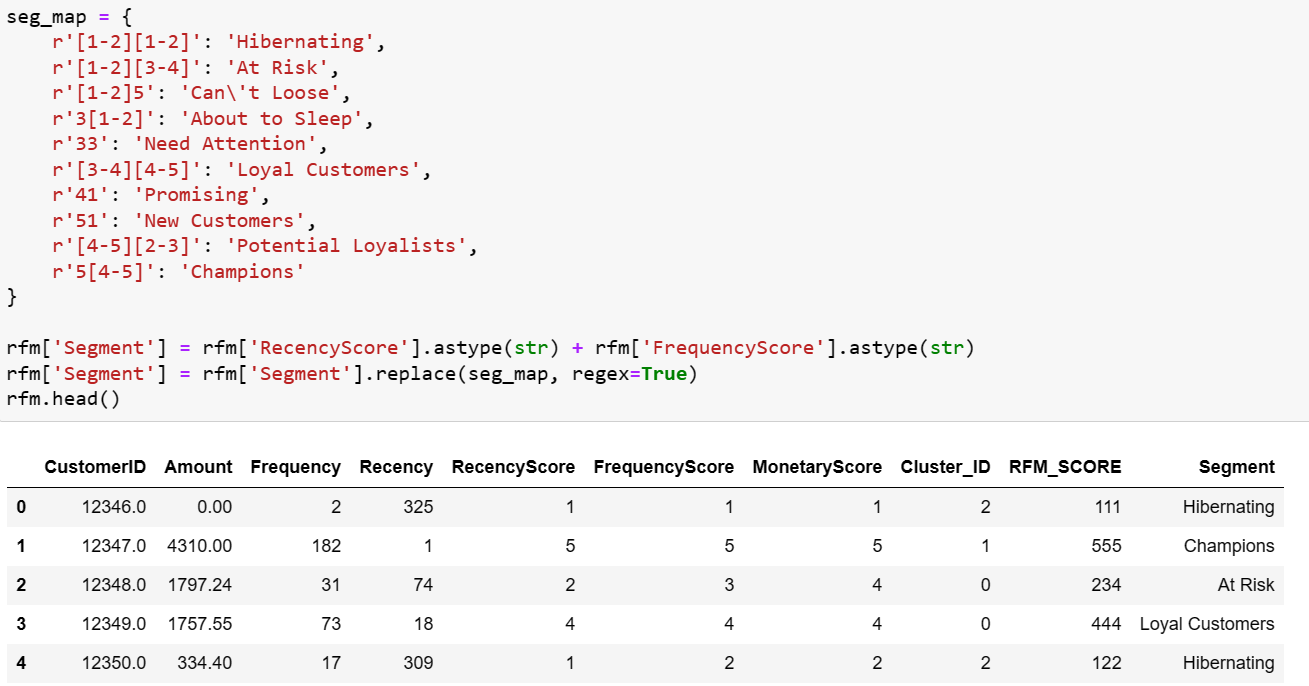
From both the above methods,K=3 seems apt.

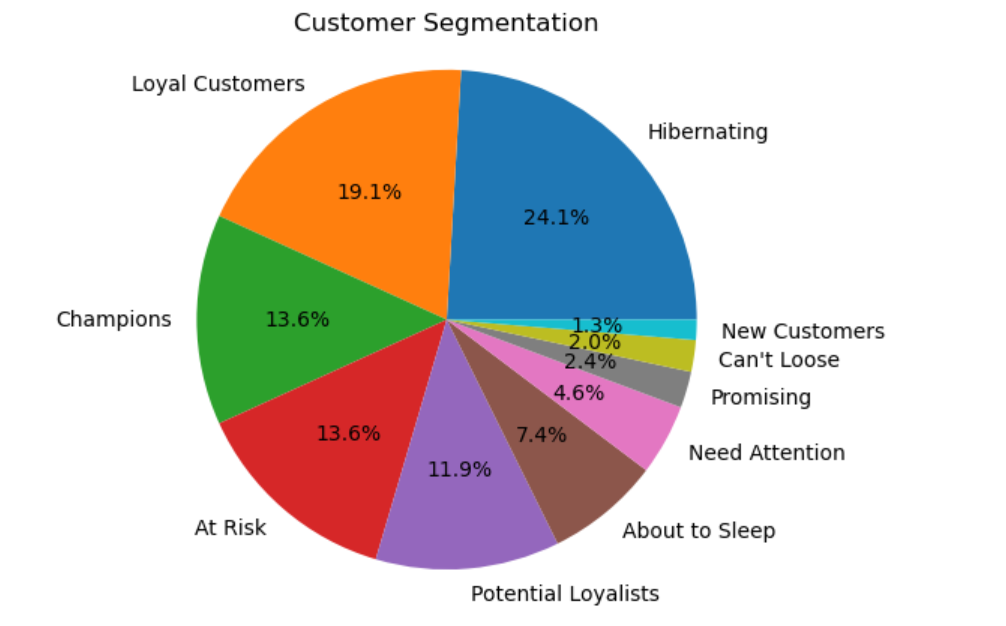


Segmenting the customers based on their purchasing behavior using clustering algorithms

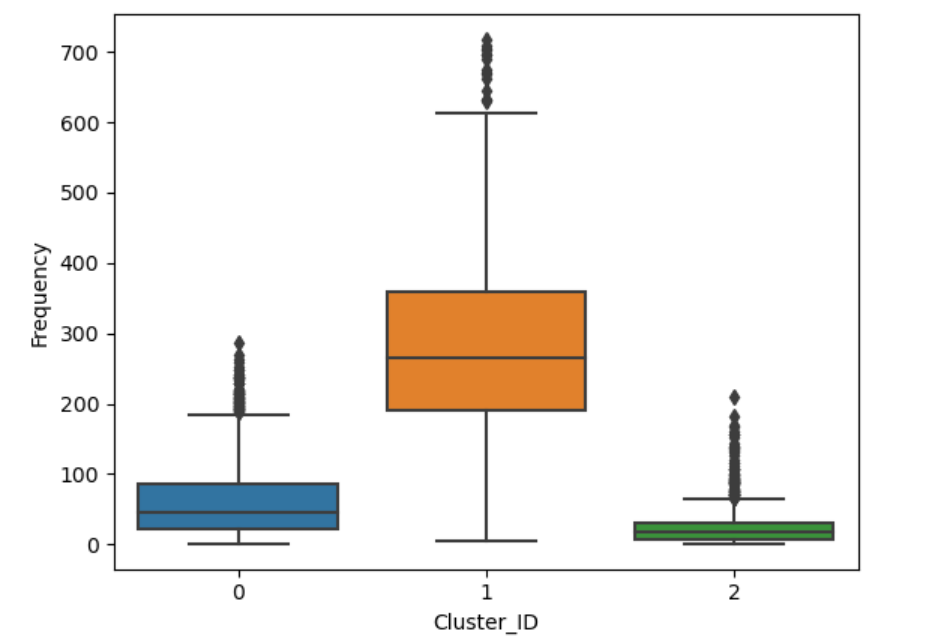


Code for bifurcating the customers into different categories.

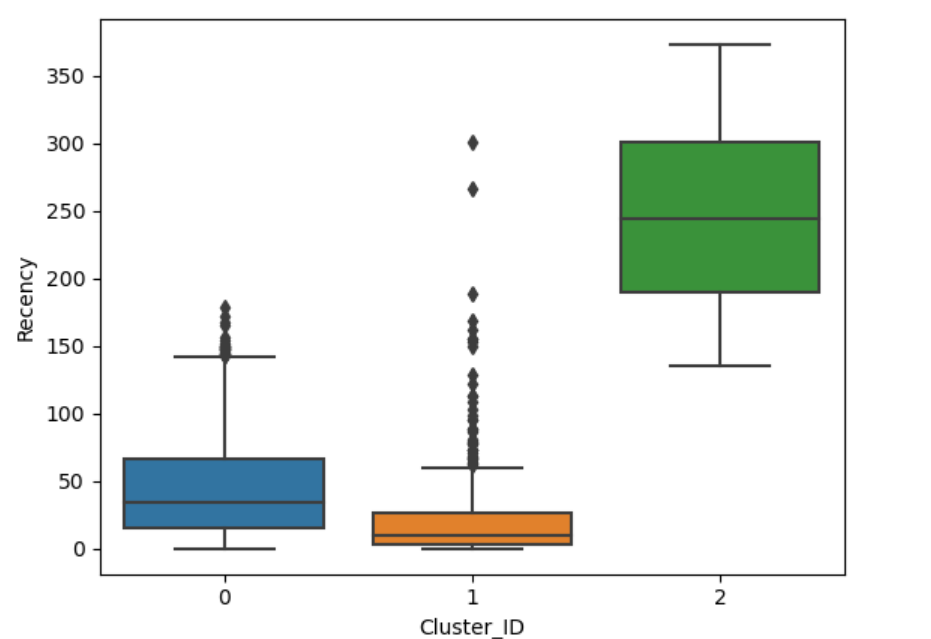
****

****

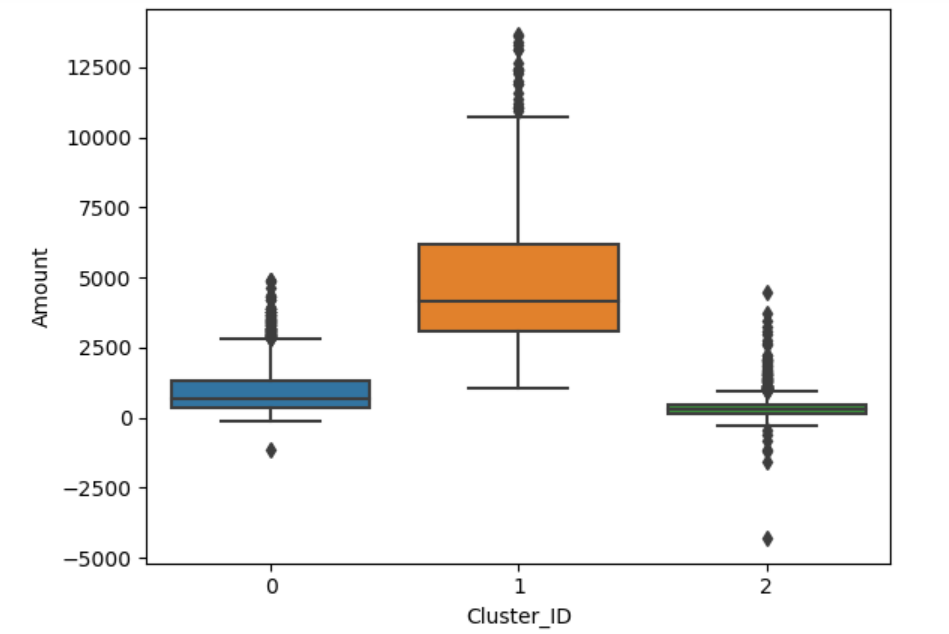
**Box plot to visualize ClusterID vs Frequency:**

****

**Box plot to visualize ClusterID vs Recency:**

****

**Box plot to visualize ClusterID vs Monetory:**

****

**Inference: K-Means Clustering with 3 Cluster Ids - Conclusion:**

* Customers with Cluster Id 1 are the customers with high amount of transactions as compared to other customers. Customers with Cluster Id 1 are frequent buyers. Customers with Cluster Id 2 are not recent buyers and hence least of importance from business point of view. Hierarchical Clustering with 3 Cluster Labels
* Customers with Cluster\_Labels 2 are the customers with high amount of transactions as compared to other customers. Customers with Cluster\_Labels 2 are frequent buyers. Customers with Cluster\_Labels 0 are not recent buyers and hence least of importance from business point of view.
* 25% Customers are Potential Customers and Champions.
* ~1% are New customers. Will have to plan some new interesting schems to attract more of them.
* ~40 % are Hibernating and At Risk Customers.Will have to plan some discounts keeping into consideration the price plasticity.

References:

<https://crunchingthedata.com/when-to-use-k-means-clustering/#:~:text=One%20of%20the%20main%20advantages,clustering%20model%20to%20be%20available>

<https://towardsdatascience.com/clustering-evaluation-strategies-98a4006fcfc>

--THE END--