Tidyquant Assignment

KRISHNA PURI

Task Overview

· Identify potential customer segmentation using RFM Model and provide some meaningful insights from each segment.

·What decision company should take based on the data insights?

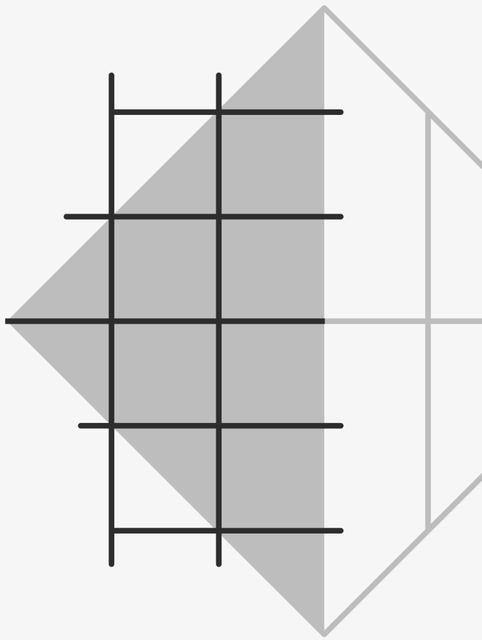
·Prepare a PPT slide for the demonstration of your data modelling work.

· Demonstrate your understanding of the data.

· Do EDA (Exploratory Data Analysis) and provide a summary
· Explain your data modelling approach.

· What sort of modelling & segmentation is best fit for this data?

· Please find out the customers who are 'champions', 'Potential customers' and 'need attention'



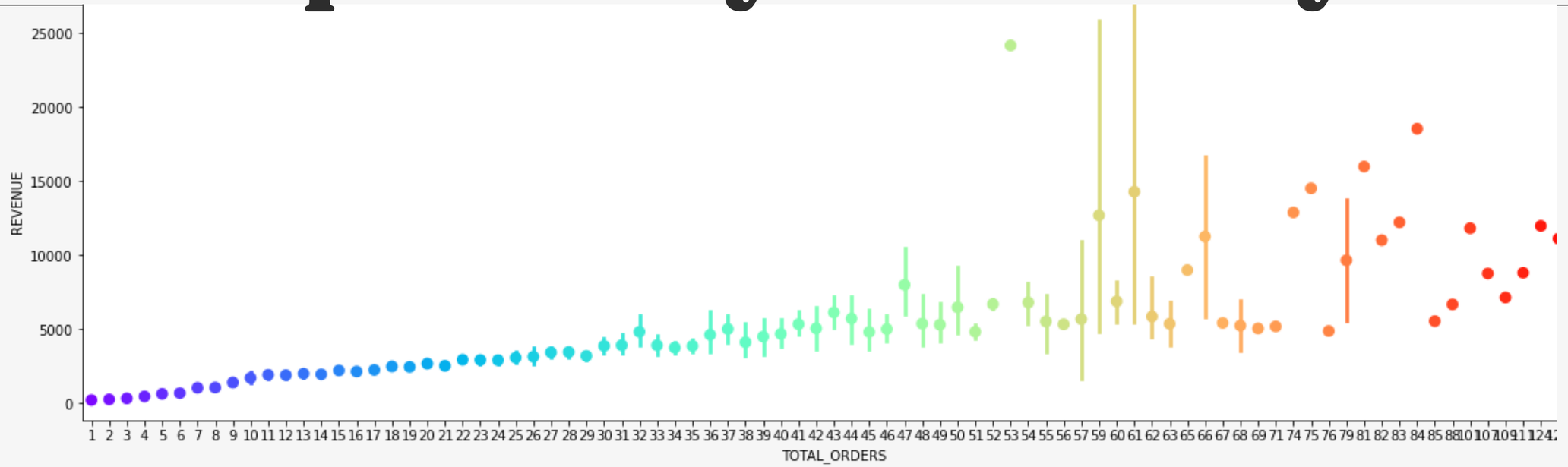
Agenda

- Data set Insights
- Exploratory Data Analysis
- Data Pre-Processing and Cleaning
- Model Building / Segmentation

Data Set Insights

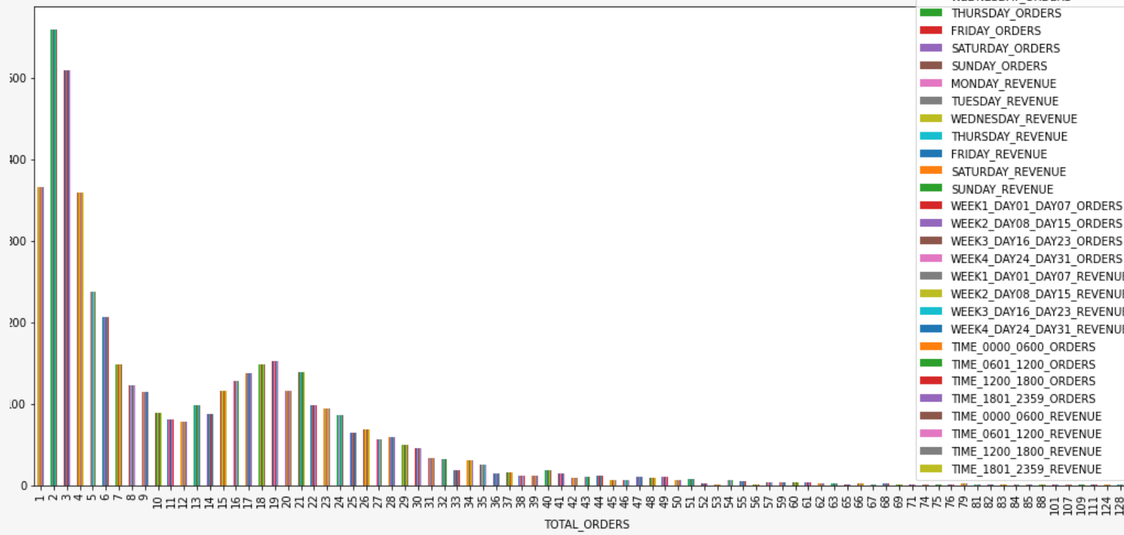
- The data set has a total of **5000 rows and 42 columns**
- There are **2 columns with missing values** in the data set
- df.columns gave us columns/features some of them are- **'CustomerID', 'TOTAL_ORDERS', 'REVENUE', 'AVERAGE_ORDER_VALUE'**
- **df.info()** gave us their data types like **CustomerID - int64, TOTAL_ORDERS - int64, AVERAGESHIPPING - float64, FIRST_ORDER_DATE - datetime64**, and more
- The **describe()** method returns a description of the feature and their interrelations- **count, mean, std, min, 25%, 50%, 75%, max**
- There are a **few outliers**
- The **correlation matrix** shows the interrelation between features- **Total Orders and Revenue are highly correlated, while average order value and Revenue are less correlated**

Exploratory Data Analysis



- the **highest, lowest and average amount of orders** placed- 34847.4, 38.5, 1681.52384
- **customers who generate revenue > avg revenue** (above average category)= 1845
- **customers who generate revenue < avg revenue** (below average category)= 3155
- **Maximum, minimum, and avg amount of orders placed**- 156, 1, 12.8704
- **Day in weak has maximum order is Thursday = 32**
- Saturday orders = 31 second-highest
- **maximum order and revenue in day are-**

Maximum order is 47 , customer id=10 , revenue = 14309



Segmentation and RFM Model

The **RFM model** is based on three quantitative factors:

*__Recency__: How recently a customer has made a purchase (DAYSSINCELASTORDER)
*__Frequency__: How often a customer makes a purchase (AVGDAYSBETWEENORDERS)
*__Monetary__: How much money a customer spends on purchases (REVENUE)

Segmentation:
-**Champions Customers:** Bought recently bought, buy often and spend the most => recency-latest | frequency-high | momentary-high
-**Potential Customers**: Recent customers with average frequency=>
recency-latest | frequency-avg
-**Need attention**: Below average recency and frequency, Some time since they've purchased=>
recency-low | frequency-low

Analysis Process

I have summarized the whole process into four steps:

1. **Import and clean the data**
2. **Create the RFM table and calculate the RFM quantiles**
3. **Perform clustering and generate cluster label**
4. **Analyze the box plots to get the customer segment to find out the customers who are 'champions', 'Potential customers', and 'need attention**

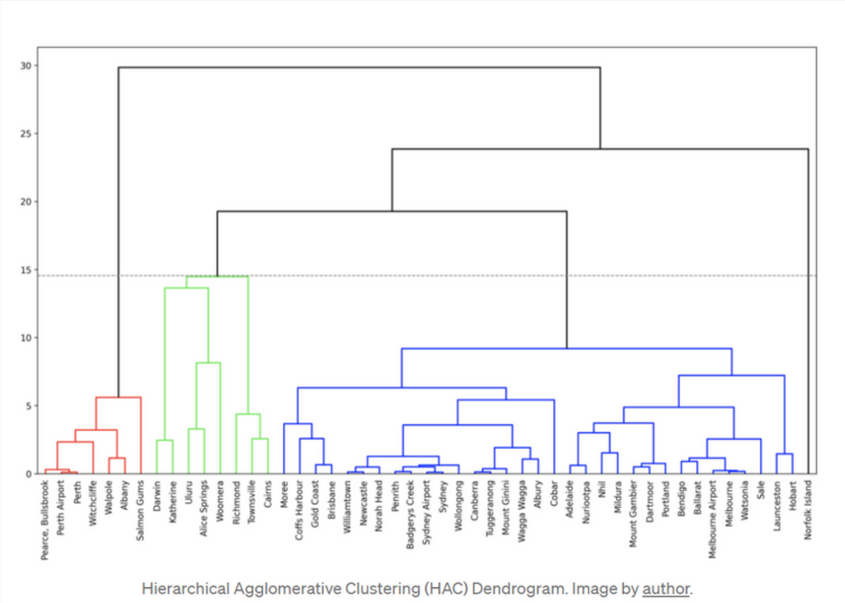
Looking at the problem statement, it's clear that this is a **classification problem** as we need **3 categories of customers: ' champions', 'Potential customers', and 'need attention.** So we can use **K-means or Hierarchical clustering.**

Why Hierarchical clustering is preferably used here over K-means is because:

K means rely on the distance to centroid to create clusters but here the **boundaries of clusters may not be well defined as there are outliers present.**

Hierarchical Cluster Working

- Initially, it assumes every point is a cluster in itself, lets's suppose there are 6 points
- The closest 2 points are clustered and stored in dendron
- Similarly, another dendron is made of two other closest points
- Now the 3rd point is merged/clustered to one of the above clusters based on its distance from them
- The above steps are repeated and stored order-wise in a dendrogram
- Based on the no. of clustures we intend to get, the dendogram is cut accordingly



Hierarchical Agglomerative Clustering (HAC) Dendrogram. Image by author.

Linkage types

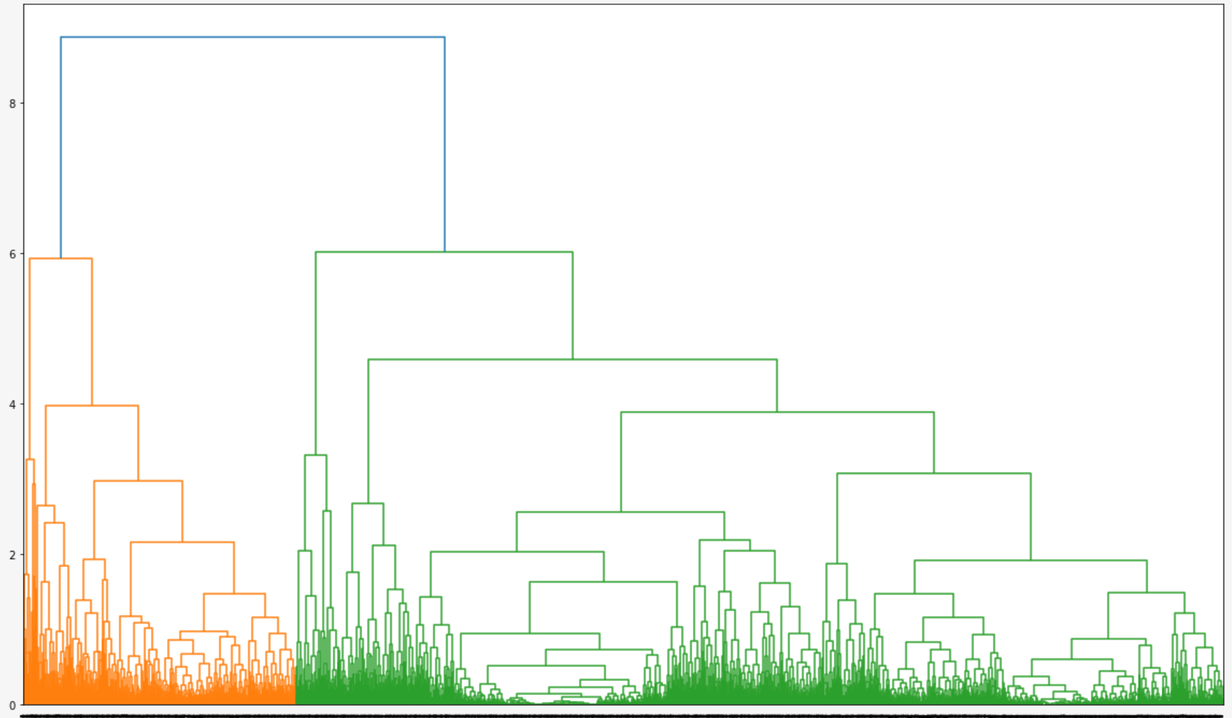
There are multiple ways to link the points together. I **used 'complete' linkage** other linkage types are-

- '**Average**': uses the average of the distances of each observation of the two sets, i.e., finds the mid-point between observations.
- '**Single**': uses the minimum of the distances between all observations of the two sets, i.e., looks for the closest point within the cluster of points at that stage (instead of cluster mid-point used in 'average').
- '**Complete**' or 'Maximum': uses the maximum distance between all observations of the two sets. E.g., if the point is closer to the farthest point of Cluster A than the farthest point of Cluster B, then such a point would be added to Cluster A.
- '**Ward**': minimizes the variance of the clusters being merged. This is very similar to minimizing Within Cluster Sum of Squares (WCSS) used by K-Means.

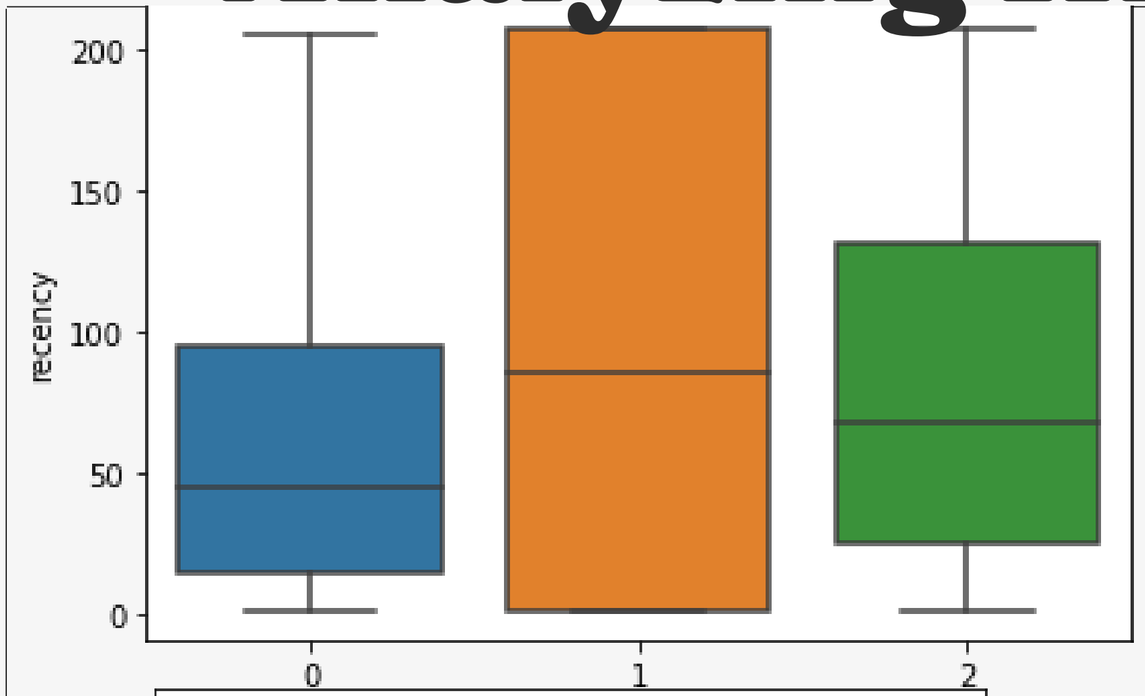
'Ward' linkage is typically the one that is used most often. It is also the default option for sklearn's implementation of HAC, which we will explore in the Python section below.

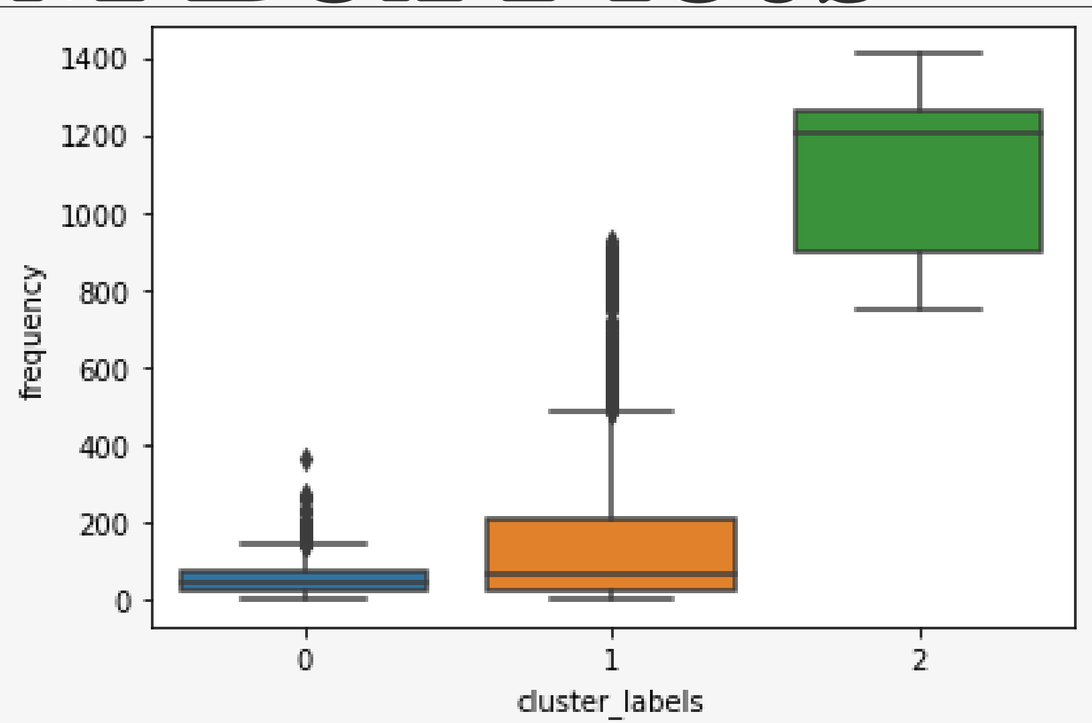
Hierarchical Cluster Steps

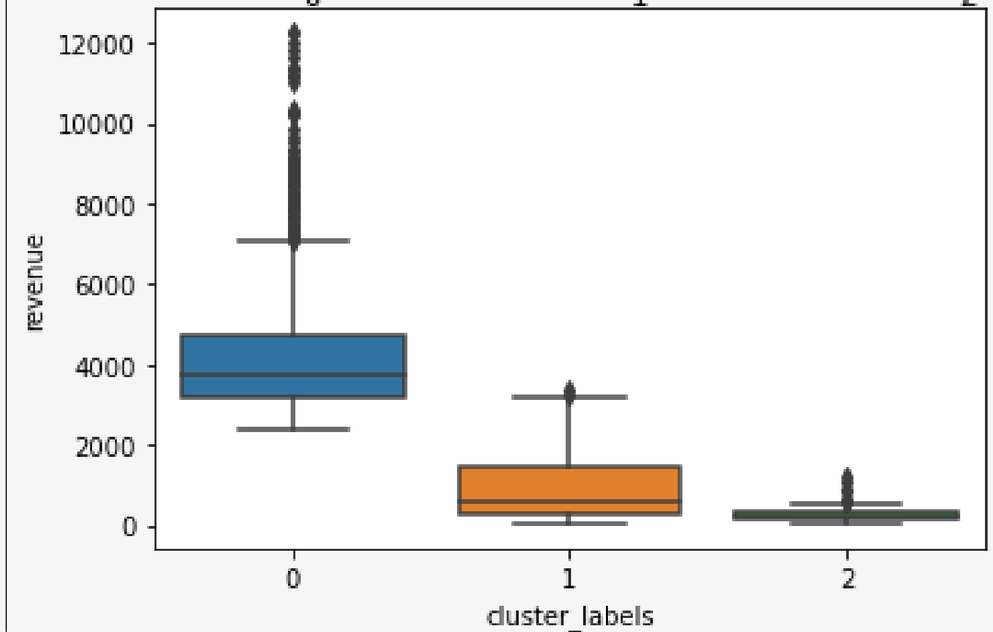
- Created RFM table
- Handled outliers
- Performed Scaling
- Performed Hopkins Statistic- checks if points are well spaced for clustering
- Built Model
- Plotted RFM parameters of clusters
- Drew conclusions and did customer segmentation accordingly



Analyzing RFM Box Plots







Analyzing using box plots of 3 clusters (blue, orange, and red) :

-The box in the box plot indicates the range in which the middle 50% of all the data lies

-the upper horizontal edge of the box is q3 and the similarly lower one is q1 25% of data lies above and below q3 and q1 respectively (q3 to q1 is IQR)

-The solid line in the box is median

-outside of wiskers are outliers

Conclusions

cluster 1 > cluster 0> cluster 2 (recency or R)
cluster 2 > cluster 0 > cluster 1 (Frequency or F)
cluster 2> cluster 1 > cluster 0 (revenue or M)

-**Champions Customers**:=> recency-high | frequency-high | momentary-high; So, **cluster 2 are Champions Customers since they have the highest Frequency and highest monetary (Revenue)**

-**Potential Customers**:=> recency-high | frequency-avg; So, **cluster 1 are Potential Customers since they have the highest recency and average frequency**

-**Need attention**:=> recency-low | frequency-low; So, **cluster 0 are Need attention Customers since they have low recency and a low frequency**