# PROGRAM DESCRIPTION

In this project, I was required to do **RFM (recency, frequency and monetary)** analysis on the online retail dataset.

- **Recency** factor describes how recently customer makes the purchase.
- **Frequency** factor describes how much a customer purchase.
- **Monetary** factor describes what amount a customer spends in total.

Based on the above information, we try to segment our customers into various groups and try to design business strategies accordingly.

The whole process that I followed can be described as follows:

- The program begins by asking the user to input minimum and maximum values of the number of clusters to try. From these values, we create a range which is then used to instantiate the class. I decided to use between **2** and **8** clusters.
- I created a class called **RFM_analyzer** which performs all the task. This class is initiated with **"input_path"** which describes the location of the dataset and **"num_clusters"** that define a range of cluster values to try.
- After initializing the above class object, **"initialize_analysis ()"** method is called upon the class object. This class function is responsible for starting the analysis.
- The **"initialize_analysis ()"** method then calls another method called **"perform_analysis ()"** where begins the main logic of this entire project.
- The first step in **"perform_analysis ()"** method is to preprocess data which is executed by the function called **"read_and_preprocess_data ()"** which loads data from the file and then removes the missing values.
- After preprocessing the data, I call another method, **"extract_information ()'.** This function is core of the program where I extract information based on **3** factors discussed above. First, I evaluate the **recency** factor by calculating the last purchase date for any customer. Then, I subtract that date from the latest invoice date of all the customers. This value is stored in a column called **"Diff".** After this, I calculated the **frequency** of purchase of the customers by calculating the total of **"InvoiceNo"** for each of the customers identified by the **"CustomerID"** column of the dataset. The **monetary** analysis is done using two columns identified as **"Quantity"** and **"UnitPrice"** which tells us about the quantity purchased by each customer and price of unit product respectively. All this information is then merged into a data frame with column names **"CustomerID", "Recency", "Frequency"** and **"Amount'.**
- After extracting thse required information, it is imperative that we remove outliers from the data. For this, I created a function called **"remove_outliers ()"** to which I pass **3** columns- **"Recency", "Frequency'** and **"Amount"** as the other column called **"CustomerID'** does not qualify for outlier detection.

- After outlier removal from the dataset, the last step left to perform before I could use it to perform analysis, was to scale the data. I used **"MinMaxScaler ()'** This scaler transforms each of the **3 (recency, frequency, and amount)** values individually so that each of these features are in their respective ranges.
- The next step is to determine the optimal number of clusters into which we are going to segment our customers. To do this, I created a function called **"calculate_optimal_cluster_value ()"** to which I pass the final data frame and a range of cluster values from which it chooses one as the optimal. This function performs **"KMeans"** clustering algorithm for cluster values ranging from **2** to **8.** For each of these cluster values, the **inertia** values are calculated. Using these **inertia** values, I plot a curve called **"elbow'** curve where the purpose is to identify the cluster value where the maximum dip in the value of **inertia** is noticed. This value leads to the most compact cluster.
- After identifying the optimal cluster value, the last but final step was to perform **"KMeans'** using this cluster value. The algorithm returns the **cluster ids** of for each of the data points.
- The last step was to create **box plots** which helps us in visually understand the different groups of customers.

## CONCLUSION

Performing above analysis, I was able to segment the customers into **3** groups. Out of these **3** groups, the **3rd** group seems to be the most profitable for the business as customers in this group spend more money **(**high **monetary** value**)**, purchase more products (high **frequency** value) and also more often (less **recency** value). This group is followed by group number **2** and then number **1.**

**PLEASE NOTE THAT WE TREAT LOW RECENCY VALUE AS A GOOD SIGN FOR THE BUSINESS BECAUSE RECENCY VALUE HERE IS CALCULATED AS THE DIFFERENCE BETWEEN THE END DATE IN THE DATASET AND THE LAST PURCHASE DATE OFA CUSTOMER. SO, LOW RECENCY VALUE MEANS THAT CUSTOMER PURCHASES MORE OFTEN.**