DAI-101 Assignment 1

Nikhil Yadav, 23116069, Electronics and Communication

# Data Cleaning and Univariate Analysis

After loading the dataset with Pandas library as Data Frame ‘df’, we try to get the general view of the data and determine what kind of data we are dealing with, by using the command ‘df.head()’ to read the first few lines or records of the data.

Now, we visit each column of data one-by-one to fix: Missing Values, Duplicate Values, and Incorrect Values.

By using the command ‘df.info’ we get the information about the datatypes of each column, and we can deal with it accordingly. The basic commands we use are: ‘*df[“column name”].isnull()*’, ‘*df[“column name”**].duplicated(*)’, ‘*df[“column name”**].describe()*’, ‘*df[“column name”].**info()*’

First, we remove the entries from the data frame containing same Sales\_Order #. Then for the Date column, we first convert its type to Date\_time, so that we can use it more efficiently to fill the missing values in the Day and Month columns.

By plotting the density of Day, we can see when most of the orders were made.

After fixing the values of Day, Month and Year, we plot the Density curve of Customer\_Age to see what age people buy the product more. We can also see outliers using Boxplot of this column to determine if we have orders from extreme ends of the age range.

Plotting the value\_counts of gender column as pie chart gives us the information about our more prominent customers.

In the Countries column we find a lot of errors in the filled data. Like the number of Spaces given in typing the name of a same country differs which, the pandas data frame considers unique as this is an Object data\_type. We fix it by removing all the spaces from the said column.

By plotting the Bar Plots of the countries, we can see which country places the most orders and the distribution of orders among the countries.

From the State column, we extract the value\_counts for each state to specifically see if we have a certain state ordering more.

We can see here, only one order lacks the product description about which we cannot do much so we can either drop that particular entry or leave it empty.

Now from the Profit and Cost columns we can get their totals by the command ‘df[‘ Profit ’].sum()’ and ‘df[‘ Cost ’].sum()’.

# Bivariate & Multivariate Analysis

Plotting a scatterplot between Profit and Cost using seaborn library, we can see that Profit increases as the Cost of the product increases, which means that we can still increase the prices of these products.

A boxplot between Age\_Group and Profit gives the idea of profit generated by each of the age groups and we can see here, Young Adults group sometimes generated higher profit.

A bar plot between Profit and States gives us the information about the profit generated by each individual state.

By using ‘groupby’ command with profits and countries we can see which country generated more profit. Similarly, we plot a pie chart showing profit generated by both the genders.

A heatmap generated between Country, Customer\_Gender, and Order\_Quantity depicts which gender, and from which country orders the most. As a result, we can see that Australian women ordered the most and women from United Kingdom ordered the least.

Another boxplot using the columns Gender, Age\_group, Profit shows the characteristics of profit generated by each of the genders, of each age group.