# Part 1 : Cleaning the Data

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.

This data is not useful when we perform data analysis because it may provide the inaccurate results. In our dataset there are four columns with either missing or incorrect data. These columns are:

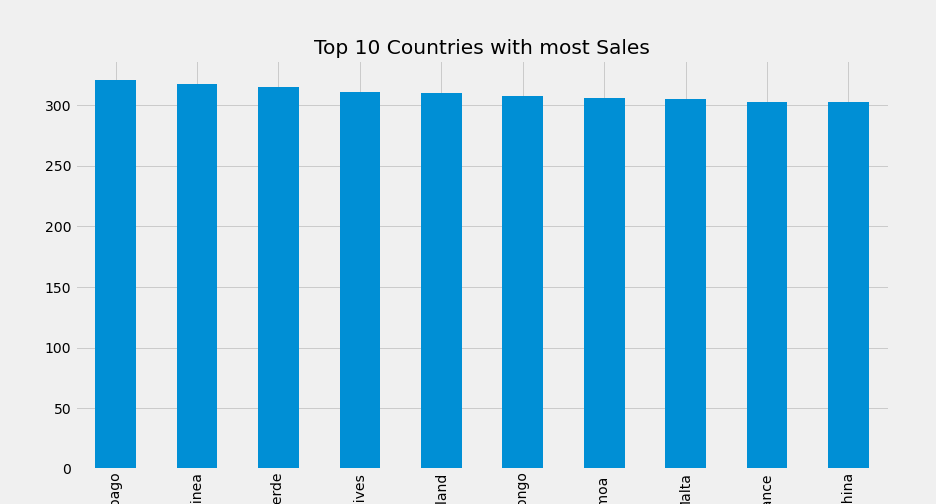
* Country   (either missing OR will be a number as a string)
* Item Type   (either missing OR won't be a valid Item Type from the other ones listed)
* Order Priority   (either missing OR won't be a valid priority code of 'C', 'H', 'M', 'L', or 'NULL')
* Order ID   (either missing OR won't be a number)

We first find all the entries that are either missing or invalid and then replace those entries with NULL if the column is of string type and with 0 if the column is of numeric type. And finally we remove all these rows and then save the cleaned data into csv file called **MM\_Sales\_clean.csv**.

# Part 2: Exploratory Data Analysis with Reports & Visualizations

## Country Rankings

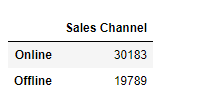
In this part we will find which countries we sell most so we can pick a new location to build shipping center. We first do some analysis to find the top 10 countries where we sell the most and then visualize it into bar chart for better understanding. The bar chart of top 10 countries where our company sells the most is given below. From this graph we observed that our company sells most to Guinea but in this country, we have already built a shipping center. The second country where we sell most is Democratic Republic of Congo so now, we already have a shipping center in Guinea so we must build shipping center in Democratic Republic of Congo.



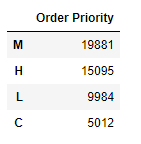
## Count of Sales Channel and Order Priority

In this part we found the count of the online and offline orders that our company takes and the count of orders by each order priority types for better understanding of the distribution of orders by Sales Channel and Order Priority.

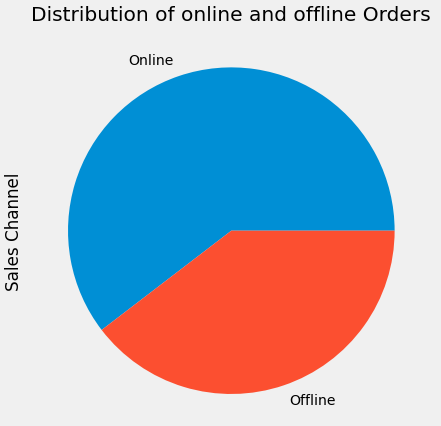
The count of the orders by Sales Channel is given below:



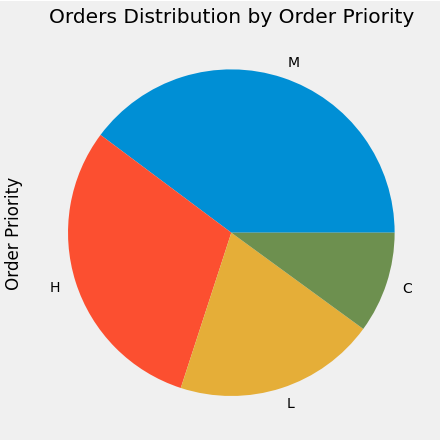
The count of the Orders by Order Priority type is given below:



The pie Chart for orders by Sales Channel is shown below:



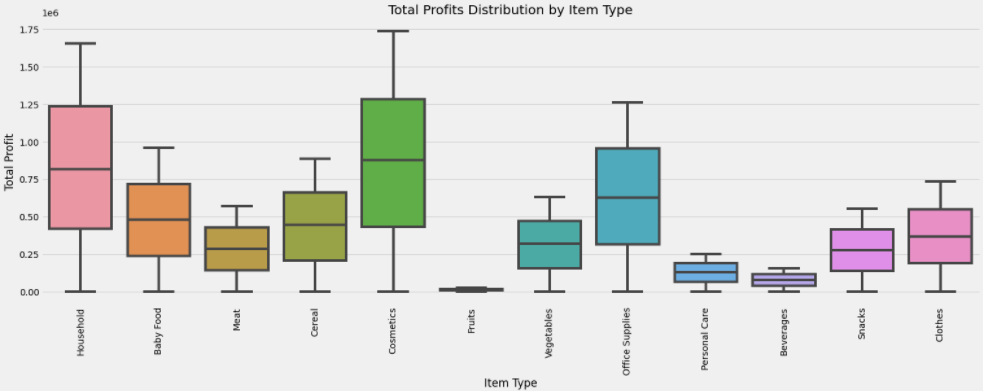
The pie chart for orders by Order Priority type is given below:



From above tables we observed that our company has more online orders as compare to offline orders and the Order Priority sequence from most to least are M/H/L/C.

## Profit by Item Type

In this part we visualize the Profit Produced by each item type in a box plot so that we can get insights about which item producing the most amount of the profit and which producing the least. The box plot of the Profit distribution by Item Type is shown below.

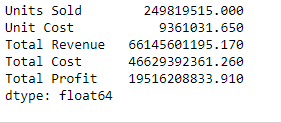


This plot clearly shows that most of the profit is generated by the Cosmetics than Household and Office Supplies type of items. And the least amount of profit generated by Fruits. With this kind of analysis, we can make decision about for which item type we should take most orders so our profits can be increased.

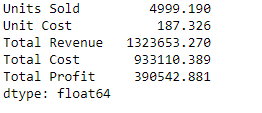
## Descriptive Statistics

In this part we calculate the sum, mean and maximum values for columns Units Sold, Unit Cost, Total Revenue, Total Cost and Total Profit.

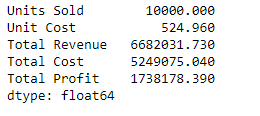
The sum of the specified columns is shown below:



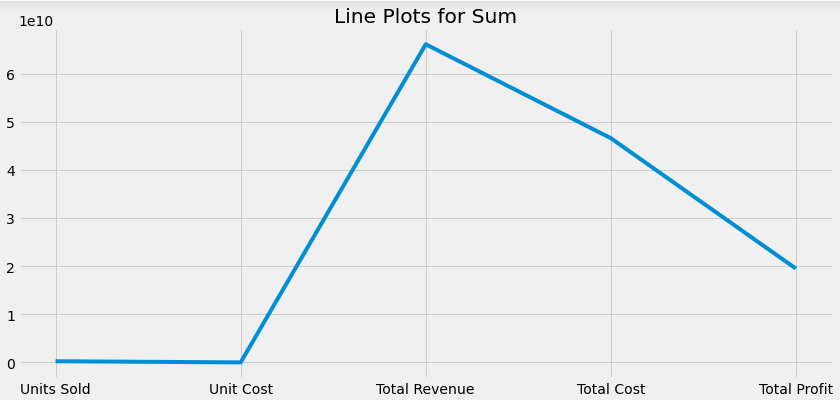
The mean of the specified columns is shown below:



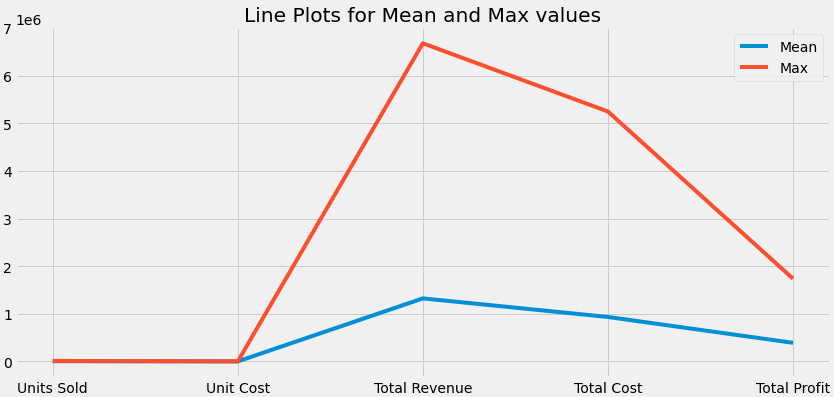
The max of the specified columns is shown below:



The line plot for the sum values of the specified columns is shown below:



The line plot for the mean and maximum values of the specified columns is shown below:



# 

# 

# Part 3: Cross-referencing the Data

In the last part we grouped the data by region and create a dictionary with Region as keys and Countries as list values belonging to each region and the stores the data into csv file called Countries\_by\_Region.csv which is shown below:

