Data Wrangling

Steps involved:

1. ***Loading the Dataset***: **Pandas** dataframe library was used to load the online retail dataset to Jupyter notebook in CSV format. The data consist of the following columns: Invoice No, Stock Code, Descriptions, Quantity, Invoice Date, Unit Price, CustomerID and Country with 541909 rows.

2. ***Identifying missing values in retail data***: Missing values have a major effect on the dataset, it causes biased estimates and reduces statical power and accuracy of the analysis. The only columns with missing values are CustomerID and Description columns. The **Description** column has an approximate 0% of missing values in the data, hence the missing values can be removed without it affecting the data-this strategic approach is a **listwise deletion**.

   The **CustomerID** has approximately 25% missing values in the retail dataset. It may not be so wise or feasible to follow the same strategic approach as that of the Description column, because this might affect the quality of the data. Many approaches can be taken but the approach I took was filling the null values with the next CustomerID in the column. After this is done, you can check the number of missing values in the dataset. CustomerID and Description now have no missing values.

3. ***Duplicate values***: Distribution of data is **skewed** or distorted affecting patterns and relationships in data. All duplicates in the dataset were removed by using the **drop_duplicates()** method.

4. ***Mismatched Data types***: No mismatched data type was found in any column of this dataset. All data in their respective columns match the same data type.

5. ***Invalid values***: The column **Quantity** has invalid values; it has some negative values, and the quantity should not have negative values. These negative values are replaced with nan (not a number).

6. ***Inconsistent Formatting***: The data format in the data is not inconsistent. They all follow the same pattern “YYYY-MM-DD". Hence the data format is consistent.

Inconsistent formatting causes misleading insights that hinders statical performance.

7. Outliers: A scatter plot visual is used to show outliers between Unit Price and Quantity. From the visualization there are 3 points that deviate away from the data.

Aderounmu Adeyemi