# **SPRINGBOARD ARTIFICIAL INTELLIGENCE PROJECT**

# **SALES PREDICTION**

# **FOR AN**

# **ONLINE RETAIL STORE**

**Submitted By:**

Syed Ibrahim

Varsha Ravi

Leena

**INTRODUCTION**

**PROBLEM STATEMENT**

The online retail industry has experienced significant growth, leading to an influx of data that can provide valuable business insights. However, the ability to effectively analyze this data to improve sales, customer engagement, and inventory management remains a challenge.

**OBJECTIVE**

1. Predicting Revenue:

Use regression models like Linear Regression and Random Forest to predict revenue based on transaction details such as product quantity, unit price, and country.

1. Customer Classification:

Implement K-Nearest Neighbors (KNN) to classify Revenue into meaningful categories based on features like purchasing patterns and revenue contribution.

1. Revenue Segmentation:

Apply K-Means clustering to identify top revenue-generating countries and determine peak revenue-generating hours.

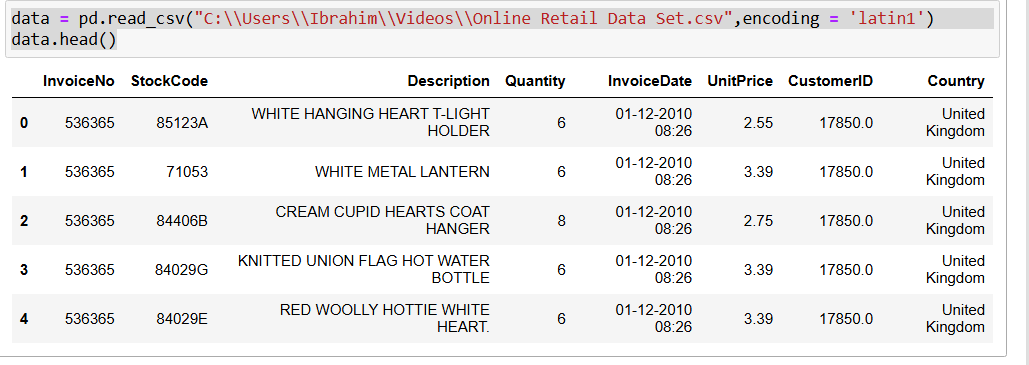
1. Insight Generation:

Analyze the results to uncover trends and actionable insights, such as high-performing products, regions, and time periods, to support data-driven decision-making.

**GOAL**:

The ultimate goal is to provide predictive models and clustering insights that improve revenue prediction accuracy, optimize marketing strategies, and identify key factors driving sales performance, thereby enhancing overall business efficiency and profitability.

**DATA LOADING**



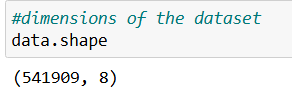
**INITIAL DATA EXPLORATION**

**Data Dictionary**

Here’s the dataset information formatted into a tabular column:

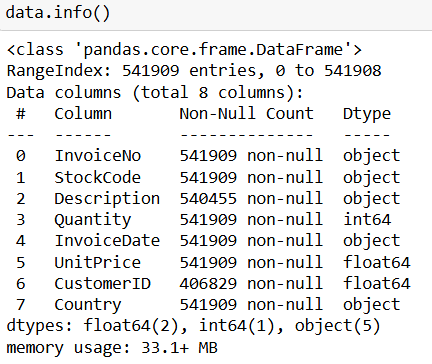
|  |  |  |
| --- | --- | --- |
| Attribute Name | Type | Description |
| InvoiceNo | Nominal | A 6-digit integral number uniquely assigned to each transaction.  If this code starts with 'C', it indicates a cancellation. |
| StockCode | Nominal | A 5-digit integral number uniquely assigned to each distinct product. |
| Description | Nominal | Product (item) name. |
| Quantity | Numeric | The quantities of each product (item) per transaction. |
| InvoiceDate | Numeric | The day and time when each transaction was generated. |
| UnitPrice | Numeric | Product price per unit in sterling. |
| CustomerID | Nominal | A 5-digit integral number uniquely assigned to each customer. |
| Country | Nominal | Name of the country where each customer resides. |

**Checking Shape Of The Data**



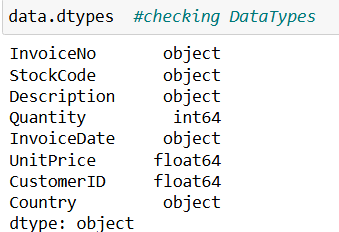
Our Data contains 541910 rows and 8 columns

**Checking Information About The Data**

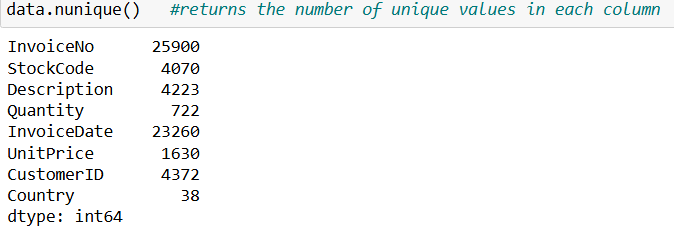


By looking at Info about Dataset it contains Null-values in description and CustomerID.

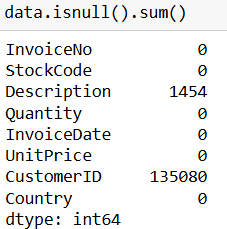
**The Dataset contains 5 Category Columns and 3 Numeric Columns**



**No Of Unique Values in Each Column:**



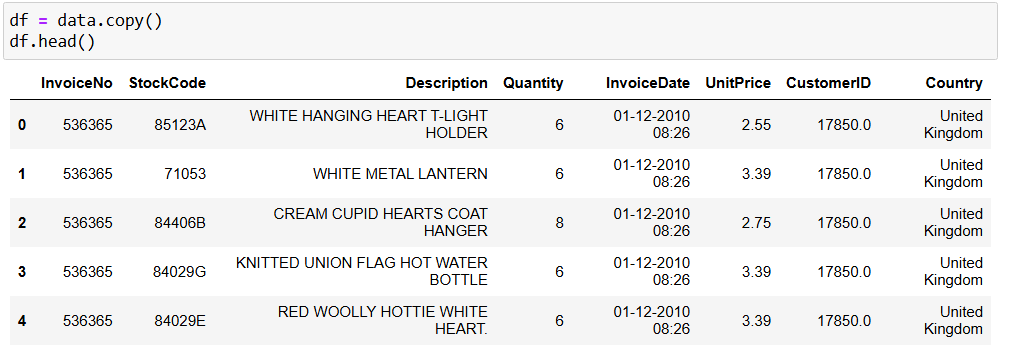
**Count the number of missing values in each column**



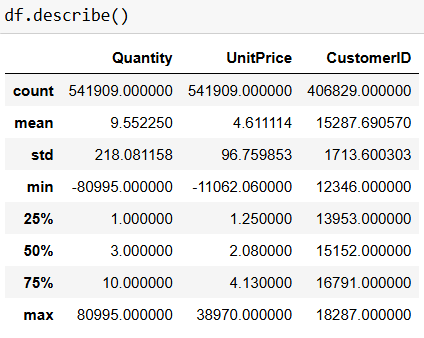
**Insights**

Customer Id contains More number of null values so we have to remove it and for Description We can remove null rows which is very low.

**Copy The Original Dataset**



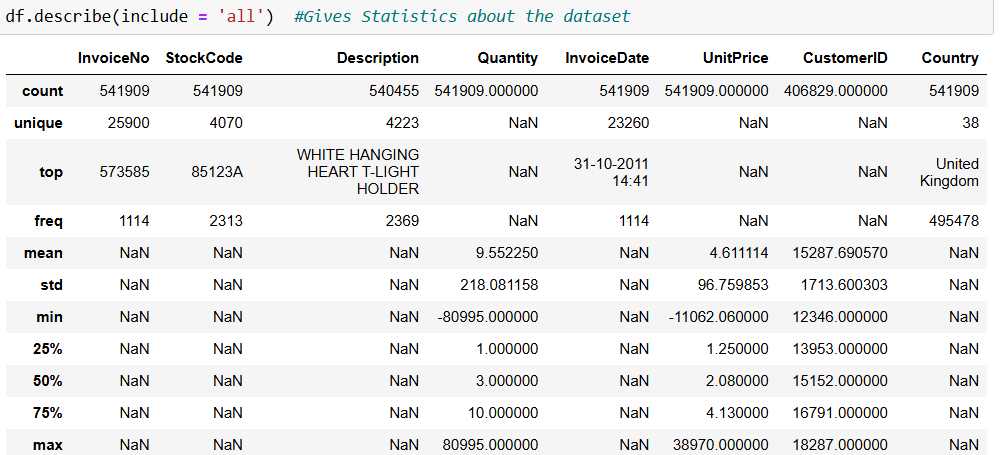
**Check Statistics about Numeric columns in the dataset**



**Insights**

1. Customer ID contains null-values.
2. The Average values of Quantity and UnitPrice are 9.55 and 4.61.
3. The Min values of Quantiy is in negative which means it is an Outlier because Quantity cannot be less than 0 and UnitPrice also.
4. The Max values of Quantity and Unitprice may be the outliers because 75 percentile of the data is less than 10 how can the remaining 25% is that much high.
5. 25% of the Data is less than 1 for Quantity and 1.25% of the Data is less than 1 for UnitPrice.
6. 50% of the Data is less than 3 for Quantity and 50% of the Data is less than 2 for UnitPrice.
7. 75% of the Data is less than 10 for Quantity and 1.25% of the Data is less than 4.13 for UnitPrice. The range between 50% and 75% of Quantity is High so maybe the distributions are long in that part.

**Check Statistics for all columns in the dataset**



Insights

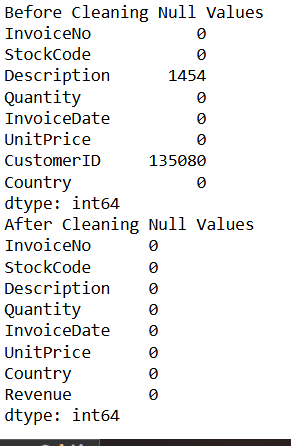
1. The Top value is given only for Categorical columns and not a numbers columns} which returns the first row in the Dataset.
2. The Nan means Not a Number which represents categorical or sentences and letters.
3. The freq row indicates the first row frequency count.
4. The Unique column gives the total number of Unique rows in each column.

**Identify what are the unique rows in country column**



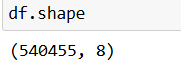
**Data Cleaning**

**Handling Null Values**

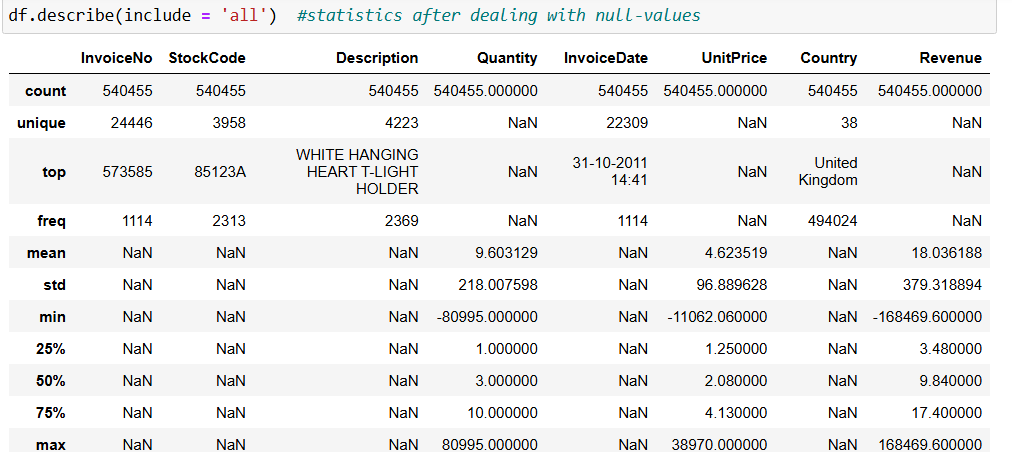


**Insights**

We removed CustomerID which has vey large no of null-values and we removed null values from description.



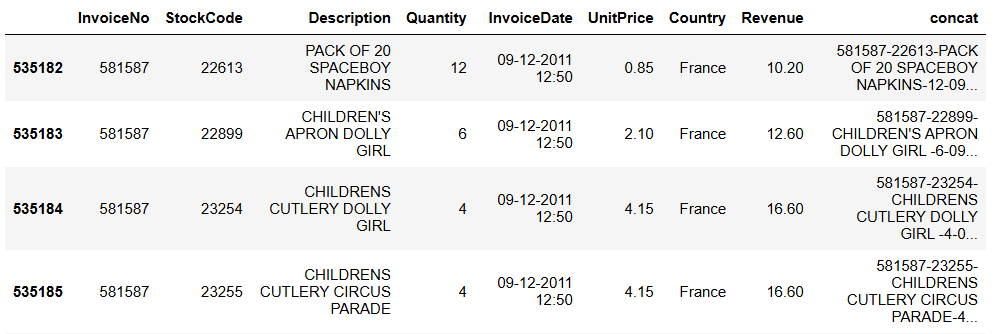
**Statistics after dealing with null-values**



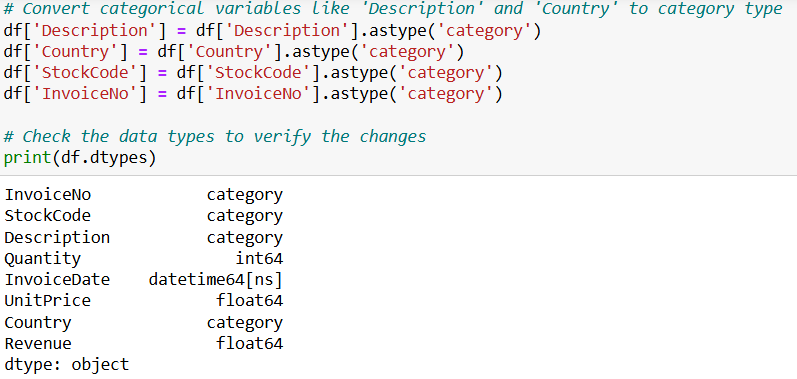
**Insights**

1. The count of all columns are same means no null values present in the data now.
2. The revenue column is newly added but the min value in negative. It can't be loss because we don't have loss column.
3. The standard deviations are relatively high, showing a large spread in the data.
4. For example, a standard deviation of 218.007598 means the data in that Quantiy column can deviate widely from its mean of 9.603129.

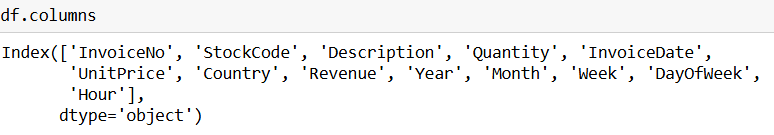
**Removing duplicates by adding temporary column and concatenating it**



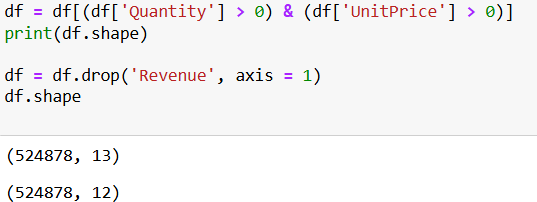
**Changing Data Types**



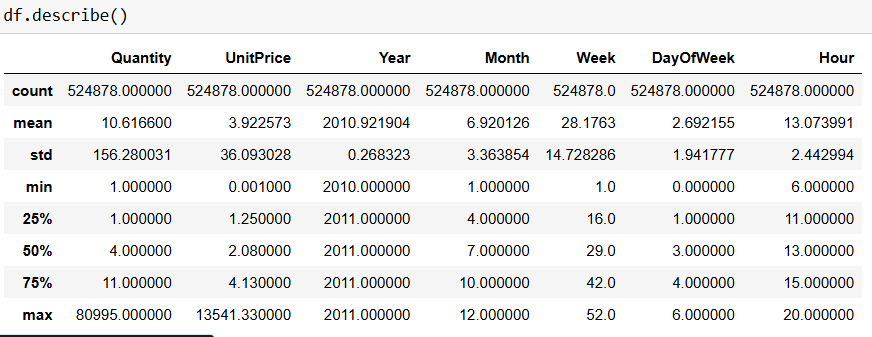
**Creating New Columns (Year, Month, Week, Day)**



### **Remove records with zero or negative amounts**



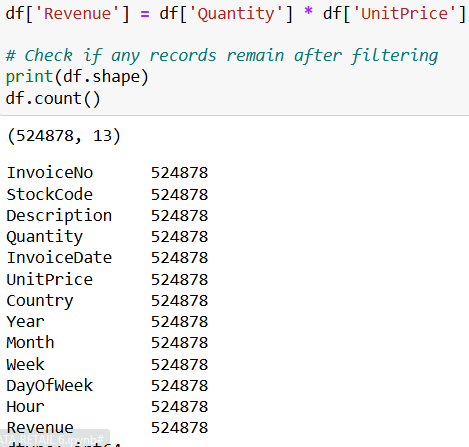
**Checking Statistics Again**



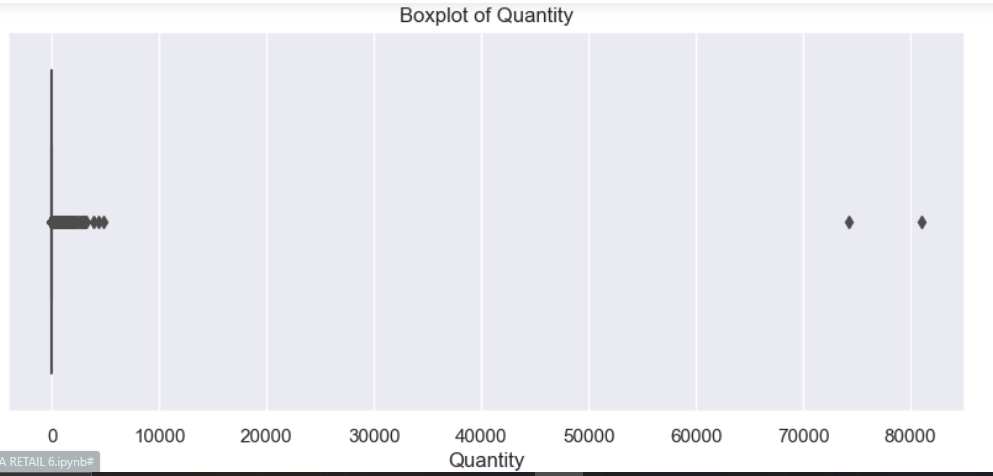
**Insights**

1. The Negative values in minimum values of Quantity and UnitPrice are removed
2. The standard deviation is still higher than mean for UnitPrice and Quantity Columns means the data is widely distributed.So we have to remove unwanted Outliers.

**We again adding revenue column to get the revenue without negative values**



**Boxplot Before Cleaning**

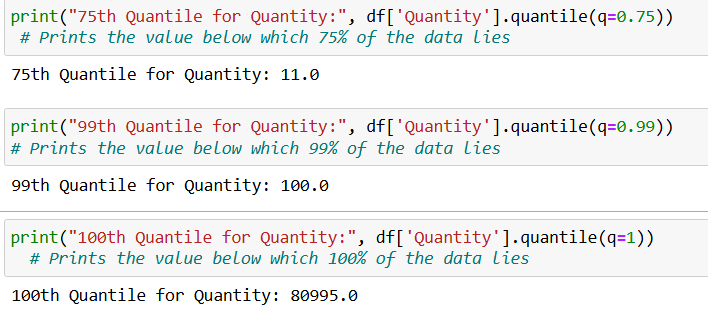


**Insights**

1. In this boxplots there are so many outliers or it didn't standardize and normalize.
2. But most of the quantity are below 75% of the dataset.
3. The 75% is 11 for Quantity means 75% of data are below 11.
4. The remaining percentile is the outliers.

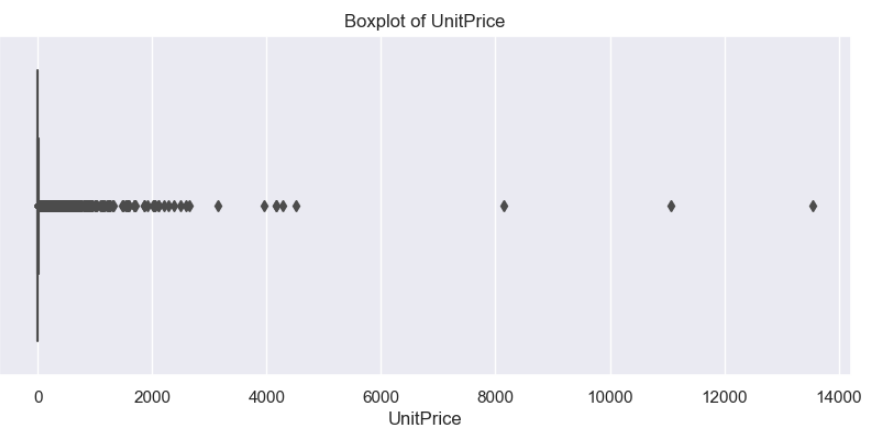
**Recommendations**

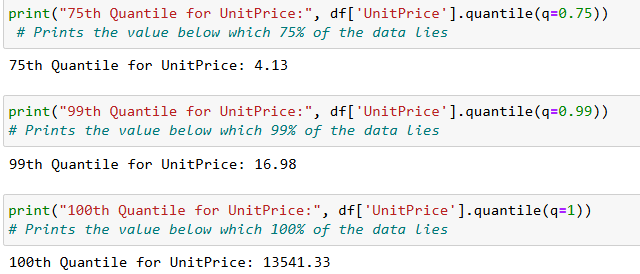
1. Cap the Outliers (Winsorization)
2. Log Transformation
3. Adjust the Plot Scale



**Insights**

1. The difference between 99 percentile and 100 percentile is very high.
2. So we have to remove the remaining 1 percentile because that is an Outlier.



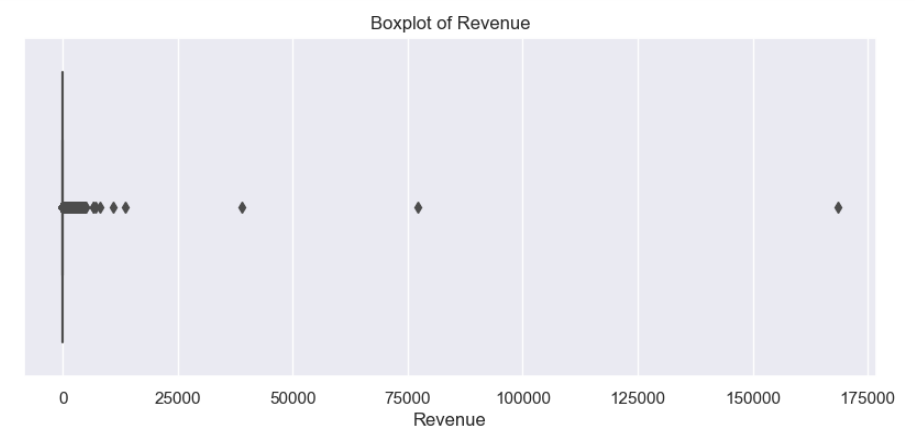


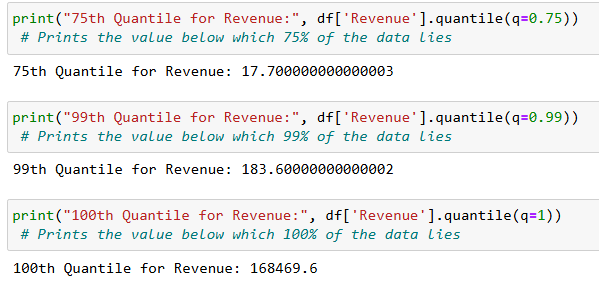
**Insights**

1. In this boxplots there are so many outliers or it didn't standardize and normalize.
2. But most of the Unitprice are below 75% of the dataset.
3. The 75% is 4 for Unitprice means 75% of data are below 4.
4. The remaining percentile is the outliers.
5. The difference between 99 percentile and 100 percentile is very high.
6. So we have to remove the remaining 1 percentile because that is an Outlier.

**Recommendations**

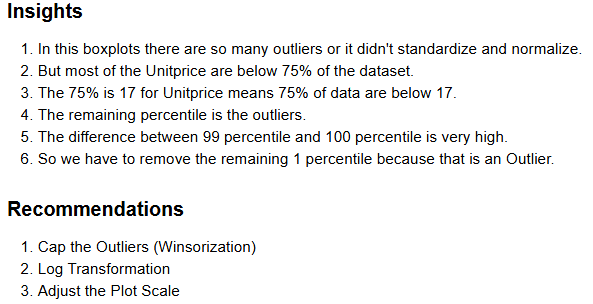
1. Cap the Outliers (Winsorization)
2. Log Transformation
3. Adjust the Plot Scale



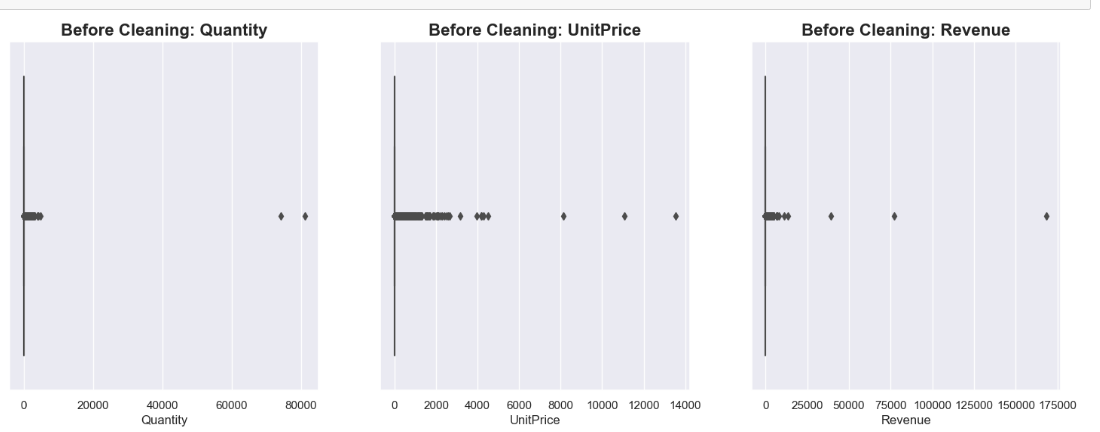


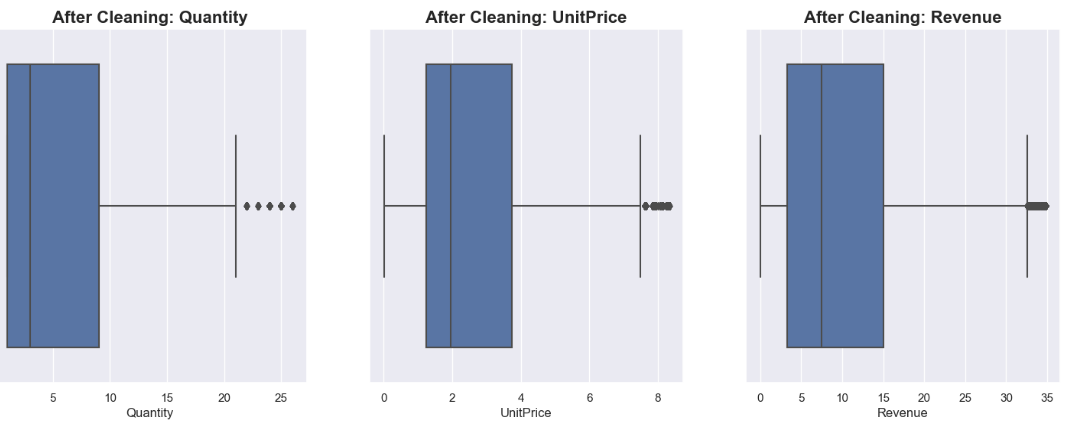
**Insights**

1. The difference between 99 percentile and 100 percentile is very high.
2. So we have to remove the remaining 1 percentile because that is an Outlier.



**Perfect Visualization of Box-Plot**





**Insights**

**Understand Boxplot**

1) The box in blue colour is the Interquantile range where 50 percentile of data are present(Lowes End = 1st Quantile, Highest End = 3rd Quantile).

2) The values below Q1(1st Quantile) to that line and above Q3(3rd Quantile) to that line are 25% and 25% of Data.

3) The values outside Min ans Max line (the two lines) are outliers.

4) The line Inside the box is Median.

**What it Indicates in Ours**

1) Quantity:

The Box Interquantile range is approx from 1 to 9.

The Median is around 3.

The Max value is 21.

The values above 21 are Outliers.

2) UnitPrice:

The Box Interquantile range is approx from 1.5 to 3.8.

The Median is around 2.

The Min and Max values are 0 and 7.5.

The values above 7.5 are Outliers

3) Revenue:

The Box Interquantile range is approx from 5 to 15.

The Median is around 8.

The Min and Max values are 0 and 33.

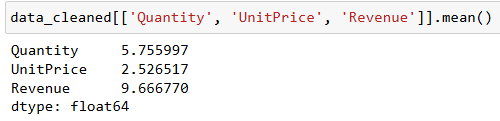
The values above 33 are Outliers.

**STATISTICS**

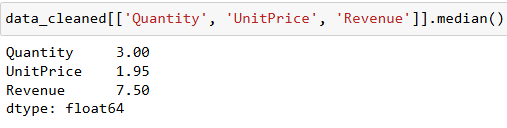
**Measures Of Central Tendency**

1. Mean
2. Median
3. Mode

1) Mean



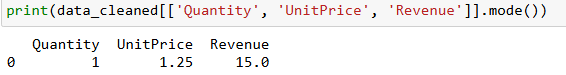
2) Median



**Insights**

The data is Right Skewed ( Mean>Median )

3) Mode



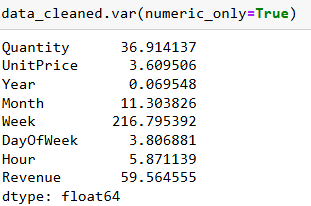
**Insights**

1. Our most sold Quantity and UnitPrice is 1 and 1.25
2. We mostly get 15 rupess as Revenue

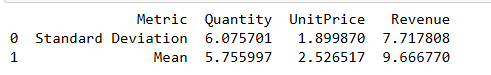
**Measures Of Dispersion**

1. Variance
2. Standard Deviation

1) Variance



2) Standard Deviation



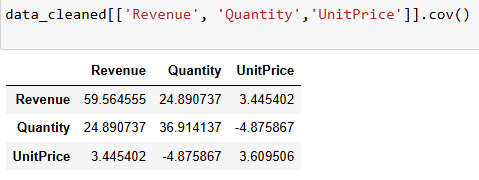
**Insights**

1. The std is very high for quantity, means it is largely distributed.
2. For Unitprice and revenue it is ok.

**Measures of Relationship between Variables**

1. Covariance
2. Correlation

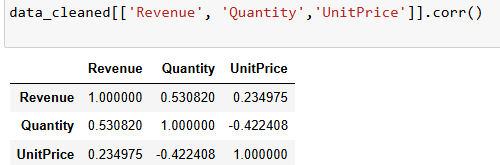
1) Covariance



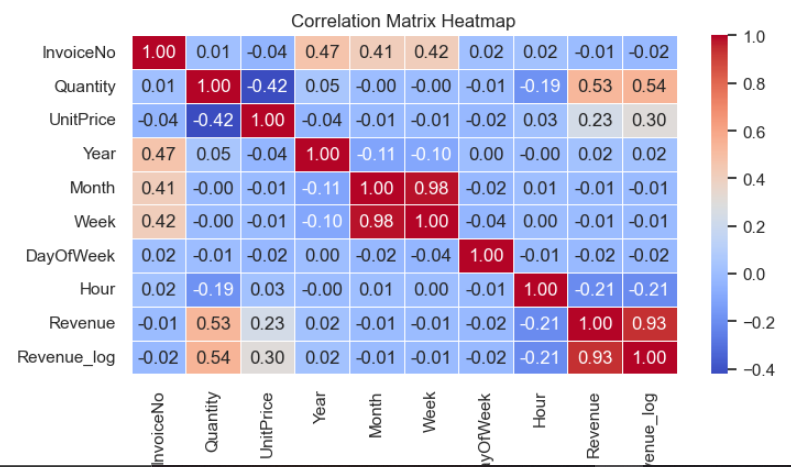
**Insights**

1. The Values for Revenue vs Quantity and Revenue vs UnitPrice are positive means the direction of that two variables is same. So it is positively Correlated. If Quantity and UnitPrice Increases, the Revenue will Automatically Increases.
2. The Values of Quantity vs UnitPrice are negayive means the direction of those variables are Opposite. So it is Negatively Correlated. If UnitPrice Increases the people who will buy the product at low price decreases.

2) Correlation



### **Correlation HeatMap Matrix**



**Insights**

1. Apart from same rows and columns Revenue and Quantity has high correlation. Means when the Quantity Increases, Revenue also Increases
2. Revenue and Revenue\_log are highly correlated bcoz it comes from Revenue
3. Revenue is slightly related with Unitprice by 0.14 mean if the Unitprice Increases Revenue also Increase
4. When the hour increases from day to night the sales decreases. People like to buy our products in daytime
5. When the Unit price Increases ,the quantity to buy the product decreases(Facts). So thats why it is in negative.

