EDA \ Descriptive Statistic

Data Dimension: -

Columns – 10

Rows - 33045

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data-Type** |
| Date | The specific day, month, and year on which the order or supply of TMT rods was made. | Datetime |
| FY | Financial Year | Object |
| Customer Name | The party that has entered into a contract for the supply or manufacture of TMT rods. | Object |
| Dia | The thickness of the TMT rod, measured in millimetres. | Object |
| Dia Group | Assorted set of different length of rods | Object |
| Grade | The quality level of the TMT rod, indicating its strength and durability. | Object |
| Type | The category of TMT rods, such as construction or industrial use. | Object |
| Length | The size of the TMT rod, measured in meters. | Object |
| Quantity | The amount of TMT rods ordered or supplied, measured in metric tonnes. | Float |
| Rate | The cost of the TMT rods per metric tonne. | Float |

Data Description: -

Descriptive Statistic: - [Descriptive Statistic](file:///C:\Users\SONIYA\Desktop\PROJECT%202\Descriptive%20Statistic.xlsx)

|  |  |  |  |
| --- | --- | --- | --- |
|  | First Moment Business Decision | | |
|  | Mean | Median | Mode |
| Financial Year FR | - | - | FY 20 |
| Dia | - | - | 08 MM |
| Dia group | - | - | 12 MM - 32 MM |
| Grade | - | - | 500 D |
| Type | - | - | FULL LENGTH |
| Length | - | - | 12 METER |
| Quantity | 5.91 | 3.90 | 2.00 |
| Rate | 48521.11 | 45700.00 | 44000.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Second Movement Business Decision | | | | |
|  | Variance | Standard Deviation | Range | Min | Max |
| Financial Year FR | - | - | - | - | - |
| Dia | - | - | - | - | - |
| Dia group | - | - | - | - | - |
| Grade | - | - | - | - | - |
| Type | - | - | - | - | - |
| Length | - | - | - | - | - |
| Quantity | 44.46 | 6.67 | 74.02 | -32.34 | 41.68 |
| Rate | 92965999.36 | 9641.89 | 63410.00 | 19590.00 | 83000.00 |

|  |  |  |
| --- | --- | --- |
|  | Third Movement Business Decision | Fourth Movement Business Decision |
|  | Skewness | Kurtosis |
| Financial Year FR | - | - |
| Dia | - | - |
| Dia group | - | - |
| Grade | - | - |
| Type | - | - |
| Length | - | - |
| Quantity | 2.20 | 5.00 |
| Rate | 0.70 | -0.08 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Missing value | Q1(25%) | Q3(75%) | IQR | Outlier | percentage |
| Financial Year FR | 0 | - | - | - | - | - |
| Dia | 0 | - | - | - | - | - |
| Dia group | 0 | - | - | - | - | - |
| Grade | 0 | - | - | - | - | - |
| Type | 0 | - | - | - | - | - |
| Length | 0 | - | - | - | - | - |
| Quantity | 14 | 1.99 | 7.01 | 5.02 | 3170 | 9.59% |
| Rate | 16 | 41750.00 | 56000.00 | 14250.00 | 185 | 0.56% |

Insights: -

**FY – Financial Year:**

Mode – FY 20

Mode of financial year is 2020, we can say that most of the order or supply of TMT rods are made on year 2020.

This data does not contain any missing value.

**Dia**

Mode – 08 MM

Most of the rod have 08 millimeter of diameter.

This data does not contain any missing value.

**Dia Group**

Mode – 12 MM - 32 MM

Most of the group of rode is of 12 MM - 32 MM.

This data does not contain any missing value.

**Grade**

Mode – 500 D

The quality level of most of the TMT rod is 500 D.

This data does not contain any missing value.

**Type**

Mode – Full Length

The of the category of TMT rod order or supply is of Full Length.

This data does not contain any missing value.

**Length**

Mode – 12 Meter

The most ordered or supply length of TMT rod is 12 Meter.

This data does not contain any missing value.

**Quantity**

**Shape of distribution**- Left-skewed, skewness = 2.20, data are positively Skewed. It indicates that from -32.34 to 0 and 20 to 41.68 having lot of outliers.

Mean>Median

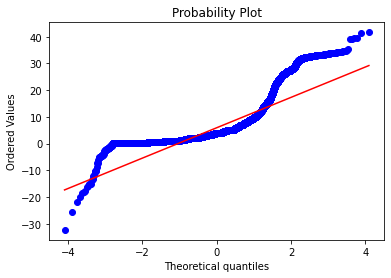
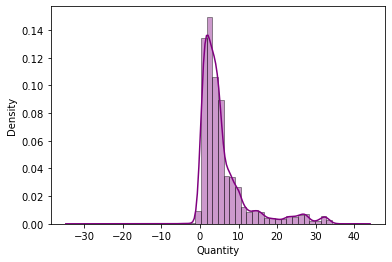
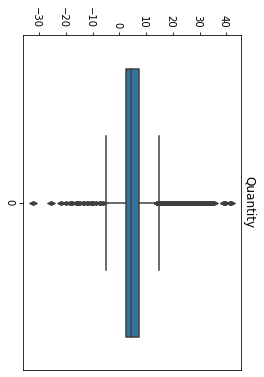
5.91 > 3.9 (Mean is influence by lower value and higher value)

**Range**- Range from -32.34 to 41.68, SD = 6.67(The SD is slightly greater than mean it indicates that there is some variability or spread in the dataset),

variance = 44.46. (The data point is quite dispersed from mean. It indicates data has highly variance; fluctuation is slightly high and there might be Outlier contribution to this larger variance.)

**Peak**- Sharp peak and long tail (Not normally distributed), Kurtosis = 5.

**Outliers**- 3170 outlier, 9.59%



**Rate**

**Shape of distribution**- The distribution is slightly distributed to right side, skewness = 0.70, data are positively Skewed. It indicates that there is extreme value on right side of the distribution.

Mean>Median

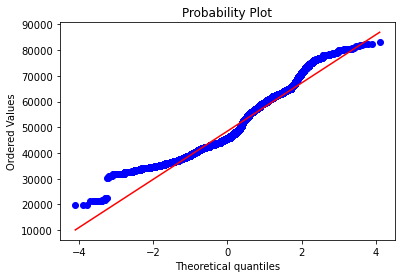
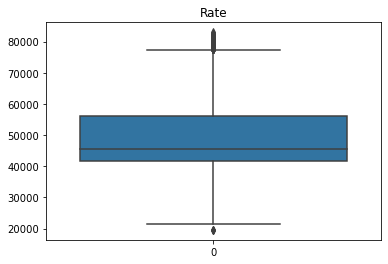
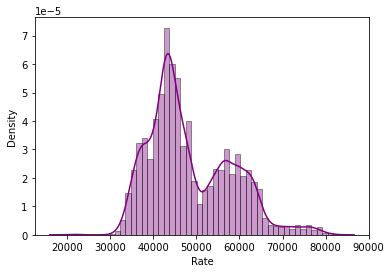
48527.11 > 45700 (Mean is influence by higher value)

**Range**- 74.02 Range from 19590 to 83000 (It indicate that the data relatively cover the narrow range of values), SD = 9641.89(There is a large amount of variability or spread in dataset. This indicate that data point is dispersed form mean),

variance = 9265999.36 (The data point is quite dispersed from mean. It indicates a wider spread of data point. High variance determine fluctuation is high.)

**Peak**- slightly flatter peak and lighter tail (Not normally distributed), Kurtosis = -0.08

**Outliers**- 185 outlier, 0.56% (Having 2 cluster called Bimodal)



**Pre-Processing –**

**Outlier Treatment –** IQR Winsorization, it has

EDA After Pre-Processing

Descriptive Statistic

Descriptive Statistic: - [Descriptive Statistic](file:///C:\Users\SONIYA\Desktop\PROJECT%202\Descriptive%20Statistic.xlsx)

|  |  |  |  |
| --- | --- | --- | --- |
|  | First Moment Business Decision | | |
|  | Mean | Median | Mode |
| Financial Year FR | - | - | FY 20 |
| Dia | - | - | 08 MM |
| Dia group | - | - | 12 MM - 32 MM |
| Grade | - | - | 500 D |
| Type | - | - | FULL LENGTH |
| Length | - | - | 12 METER |
| Quantity | 5.95 | 3.90 | 2.00 |
| Rate | 48397.42 | 45700.00 | 44000.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Second Movement Business Decision | | | | |
|  | Variance | Standard Deviation | Range | Min | Max |
| Financial Year FR | - | - | - | - | - |
| Dia | - | - | - | - | - |
| Dia group | - | - | - | - | - |
| Grade | - | - | - | - | - |
| Type | - | - | - | - | - |
| Length | - | - | - | - | - |
| Quantity | 44.01 | 6.63 | 74.02 | -32.34 | 41.68 |
| Rate | 88142651.18 | 9388.43 | 55800.00 | 21500.00 | 77300.00 |

|  |  |  |
| --- | --- | --- |
|  | Third Movement Business Decision | Fourth Movement Business Decision |
|  | Skewness | Kurtosis |
| Financial Year FR | - | - |
| Dia | - | - |
| Dia group | - | - |
| Grade | - | - |
| Type | - | - |
| Length | - | - |
| Quantity | 2.26 | 5.02 |
| Rate | 0.62 | -0.33 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Missing value | Q1(25%) | Q3(75%) | IQR | Outlier | percentage |
| Financial Year FR | 0 | - | - | - | - | - |
| Dia | 0 | - | - | - | - | - |
| Dia group | 0 | - | - | - | - | - |
| Grade | 0 | - | - | - | - | - |
| Type | 0 | - | - | - | - | - |
| Length | 0 | - | - | - | - | - |
| Quantity | 0 | 1.99 | 7.01 | 5.02 | 3145 | 9.52% |
| Rate | 0 | 41750.00 | 56000.00 | 14250.00 | 0 | 0.00% |

**Insights: -**

* **Duplicate Values –** There are 34 duplicates value were present.
* **Quantity**

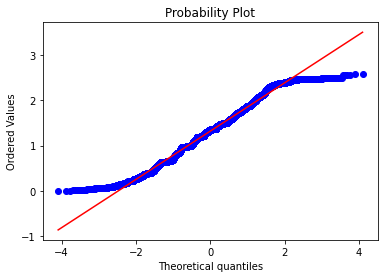
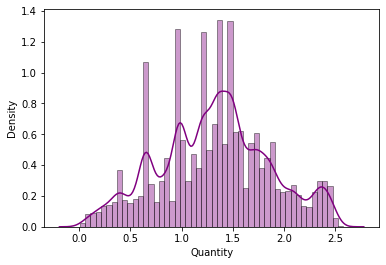
**Outlier** **Treatment** – The data has 9.59% of outlier, So I retain the outlier.

**Missing Values** – There is present of outlier so I prefer Median imputation there is extreme level of outlier median did not affect by Extreme value.

**Transformation –** Yeo Johson Transformation because Yeo Johson Transformation shows normal distribution as compare to other Transformation technique.

**Scaling –** For Quantity column I do robust scaling because Quantity column has outlier

After doing Transformation The distribution is not normally distribution but Better than before Pre-Processing. The data is slightly skewed to left side.

* **Rate**

**Outlier Treatment –** The data has 185(0.56%) outliers, So I have done Winsorization Technique using IQR for treatment of outliers.

**Missing Values –** There are 16 missing values in Rate column, Here I do Mean Imputation because the outlier is less and it will not affect the mean value.

**Transformation –** Box Cox Transformation would be the effective transformation as it shows the data normally distributed as compare to other transformation techniques.

**Scaling** – Standardization is used here as the Rate column does not contain the extreme outlier.

