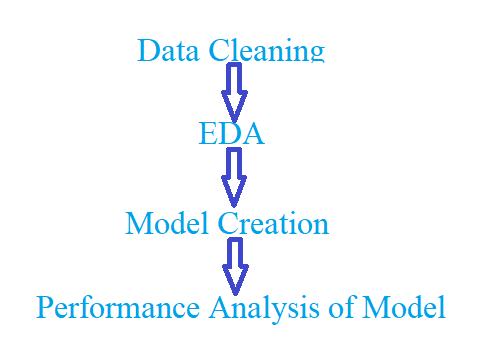
Summary of Daily Offers Analysis

Data Description

* We have a data set that contain the attributes related to sales of the steel item during a certain period. Some of the attributes are **id**, **item\_date**, **quantity tons**, **customer**, **country**, **status**, **selling\_price** etc.

Workflow



Data Cleaning

* Removing **Null** values from the data set.
* **Imputing** Null values
  + For categorical attribute we replace Null values with the **mode** of the attribute.
  + For numerical attribute we replace Null values with the **mean** if the data is **Normally** distributed about the mean and has few **outliers**.
  + For numerical attribute we replace Null values with the **median** if the data is **skewed** towards left or right. i.e data is not normally distributed. Such data contains outliers on either side of the data.
* Detecting **outliers**
  + Dectecting outliers with the help of **Boxplot**, **Scatter plot** and **Histogram**.
* Outlier transformation
  + Applied **log transformation** to mitigate the effect of outliers.
  + Imputed the value of the outliers with mean or median according to the distribution of the dataset.

Exploratory Data Analysis

Chart, bar chart

Description automatically generated

Majority of the sales of the steel was on date 1970-08-23.

Chart, histogram

Description automatically generated

Maximum sales revenue is generated from the country with code 27.0

Chart, bar chart, histogram

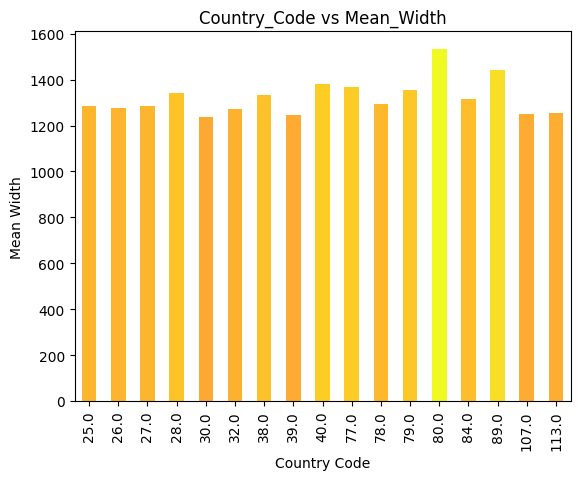
Description automatically generated

Maximum number of unique item are sold to country with code 26.0

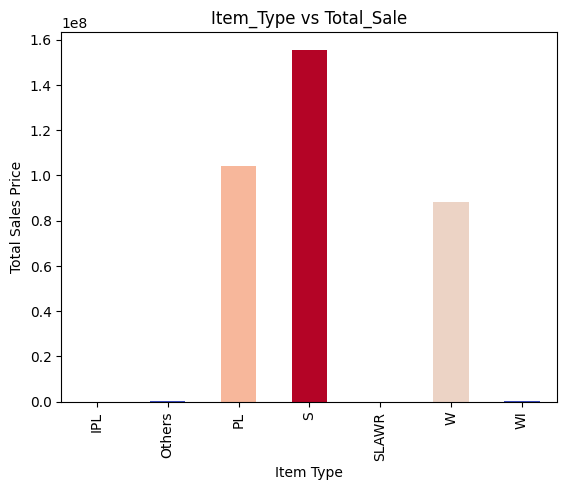
Chart, bar chart, histogram

Description automatically generated

The steel item sold to country with code 80.0 has the higher thickness on average



The average width of the steel item sold to each country is similar.



Maximum revenue is generated by selling of steel item of type “S”

Chart, histogram

Description automatically generated

Steel item of type “W” is sold maximum times.

Chart, bar chart, histogram

Description automatically generated

Mean thickness of steel item “IPL” is maximum.

Chart, bar chart, histogram

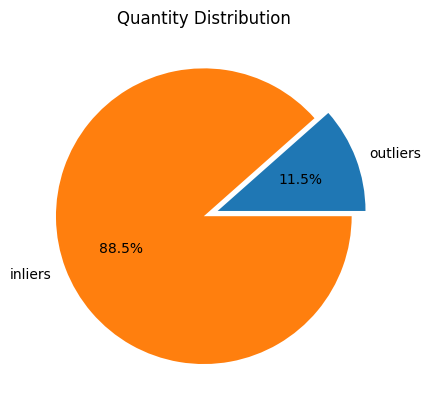
Description automatically generated

Mean width of the steel item “IPL” is maximum.

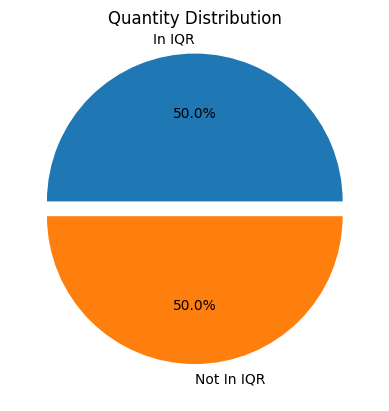
Chart, histogram

Description automatically generated

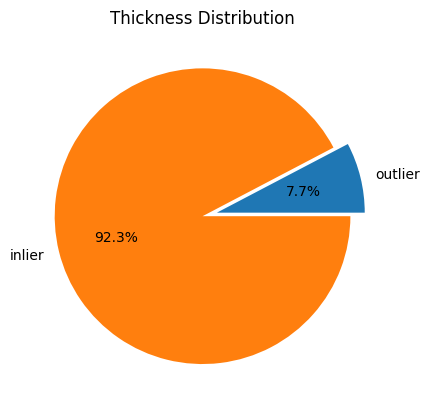
Steel item of type “S” is sold in maximum quantity.



11.5% of the data points in the “quantity tons” attribute are outliers. i.e they are extreme values which largely affect the mean quantity of the steel sold.



50% of the data points lies in the Inter Quartile Range of the attribute “quantity tons”.

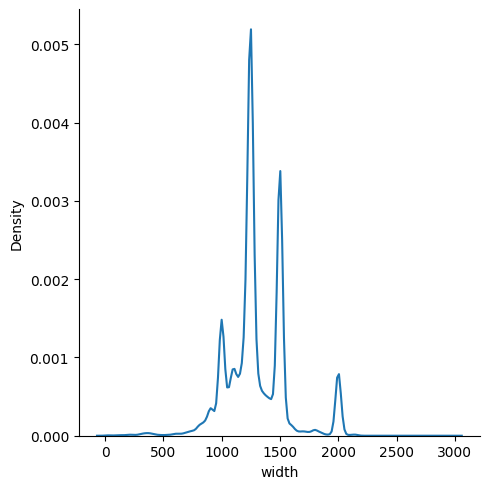


7.7% of the data points in the attribute “Thickness” are extreme. i.e far away from the measure of central tendency.

Chart, pie chart

Description automatically generated

45.6% of the data points lies outside the Inter Quartile Range of the attribute “Thickness”.



Density plot of the “width” shows that most of the data points are concentrated at mean but there a number of outliers. The smaller spikes are of different size and shape suggest that dataset have a more complex distribution that can not be easily characterized by a single distribution.

Chart, bar chart

Description automatically generated

Correlation map between different features of the dataset. Higher the correlation between independent features , we can replace those features with single entity.

Model Creation

We have drawn three types of plot for the analysis of performance of the model. Below are the mentioned details that can be used for the interpretation of the plots:-

* **Probability Distribution Plot of Residuals**

When examining the probability distribution plot of the residuals, there are a few inferences than can be drawn:

**Normality**:-If the residuals follow a normal distribution, this suggests that the model is a good fit for the data.

**Skewness**:-If the residuals are skewed, then this suggest that the model may not fit the data well, and there may be some nonlinearity or other issue with the model.

**Outliers**:-If there are outliers in the residuals, then this suggest that the model may not fit the data well, and there may be some influential observations that are affecting the model.

**Heteroscedasticity**:-If the spread of the is not constant, then this suggest that the model may not fit the data well, and there are some problems with the variance of the model.

* **Plot between Residual and Original value**

**Positive Relationship**:-If there is a positive relationship between residual and original value then it suggests that our model is under-predicting the high value and over-predicting the lower value. This can happen when the model is not able to capture the true functional form of relationship between the variables, or when there are more important variables that are not included in the model.

**Negative Relationship**:- If there is a negative relationship between residual and original value then it suggests that our model is under-predicting the low value and over-predicting the higher value. This can happen when the model is able to capture the underlying patterns and relationships in the data, but may be overfitting or overestimating the true relationship.

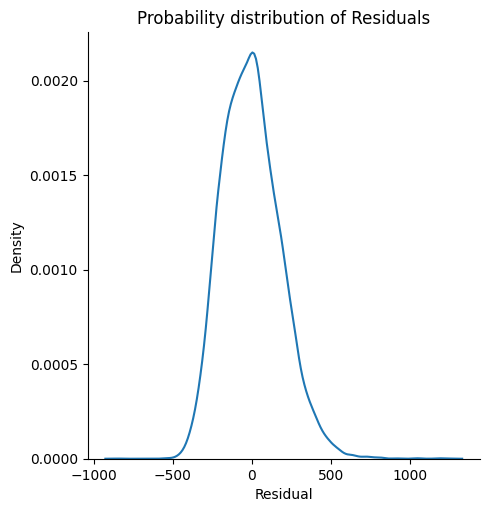
In general the scatter plot between original value and residual should show no systematic pattern or trend, with the residual evenly distributed around zero.

* **Plot between Predicted value and Original value**

In general, a good model will produce a scatter plot in which the points are clustered around the diagonal line, and the degree of variation around the line is minimal.

A poor model will produce a scatter plot in which the points are scattered randomly around the plot, indicating the mode is not able to make accurate prediction.

* Ridge Regression Model



Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

* DecisionTree Regressor Model

Chart, line chart, histogram

Description automatically generated

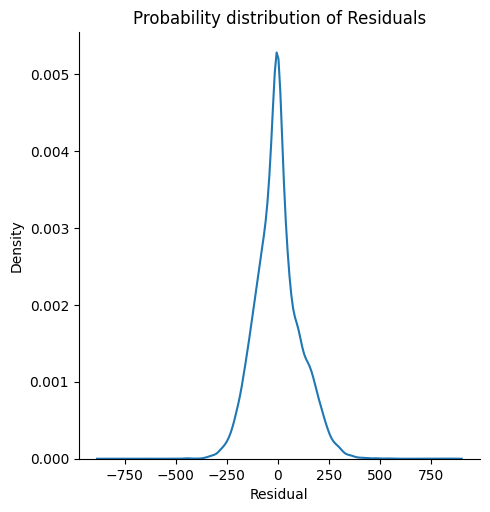
Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

* RandomForest Regressor Model



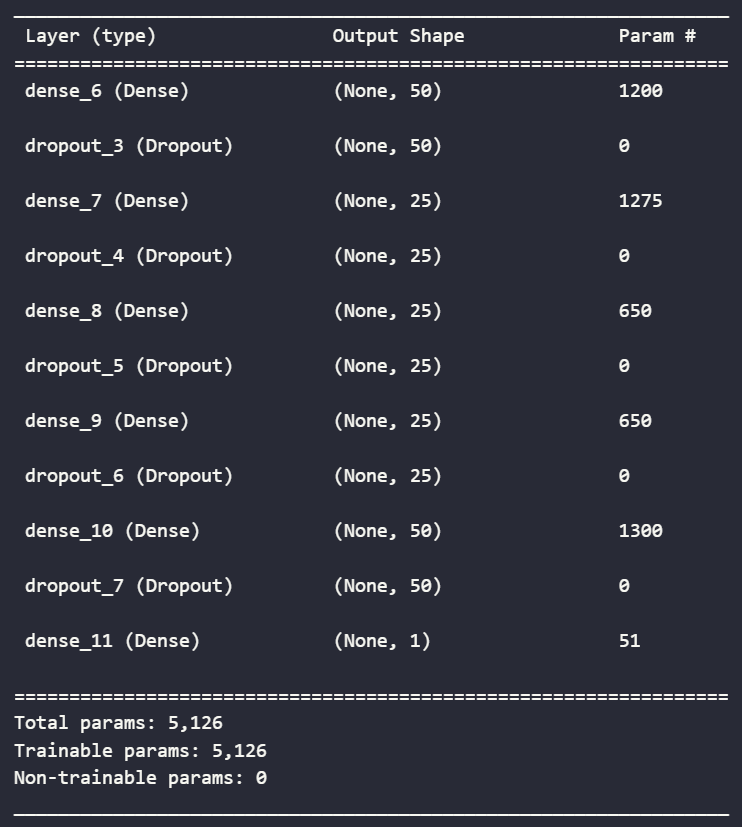
Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

* Artificial Neural Network

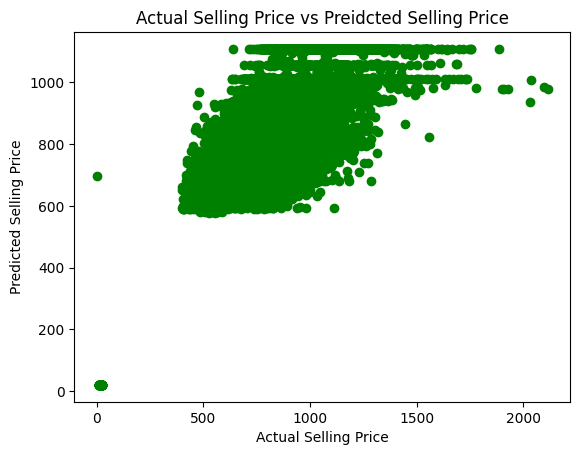


Architecture of ANN

Chart, histogram

Description automatically generated

As the number of iteration increased the error in the output decreased.



Chart, scatter chart

Description automatically generated

Model Comparison

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated