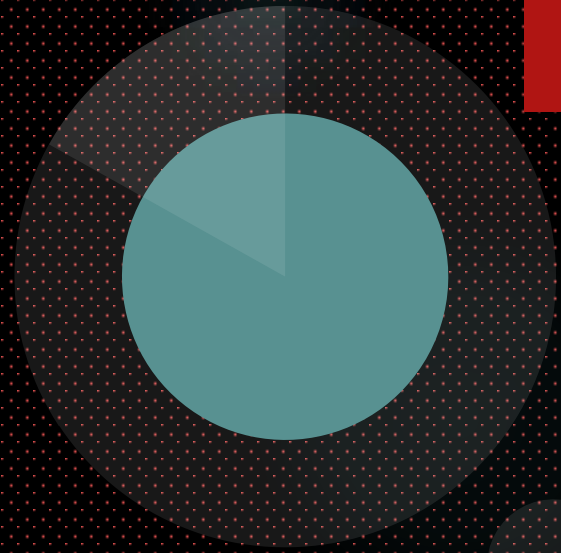


# Amazon Sales Data Analysis

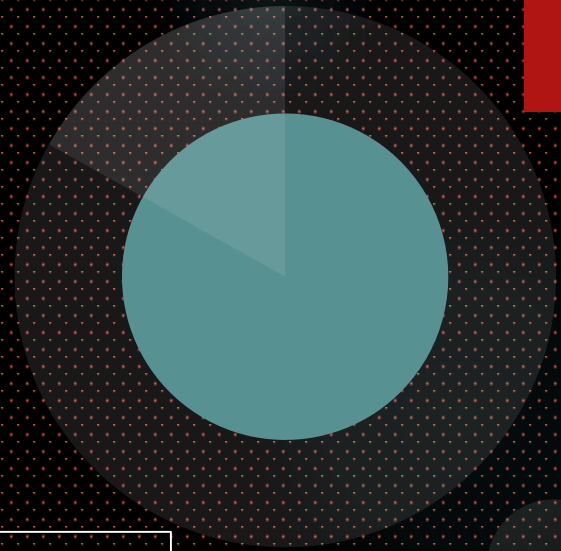


# Objective

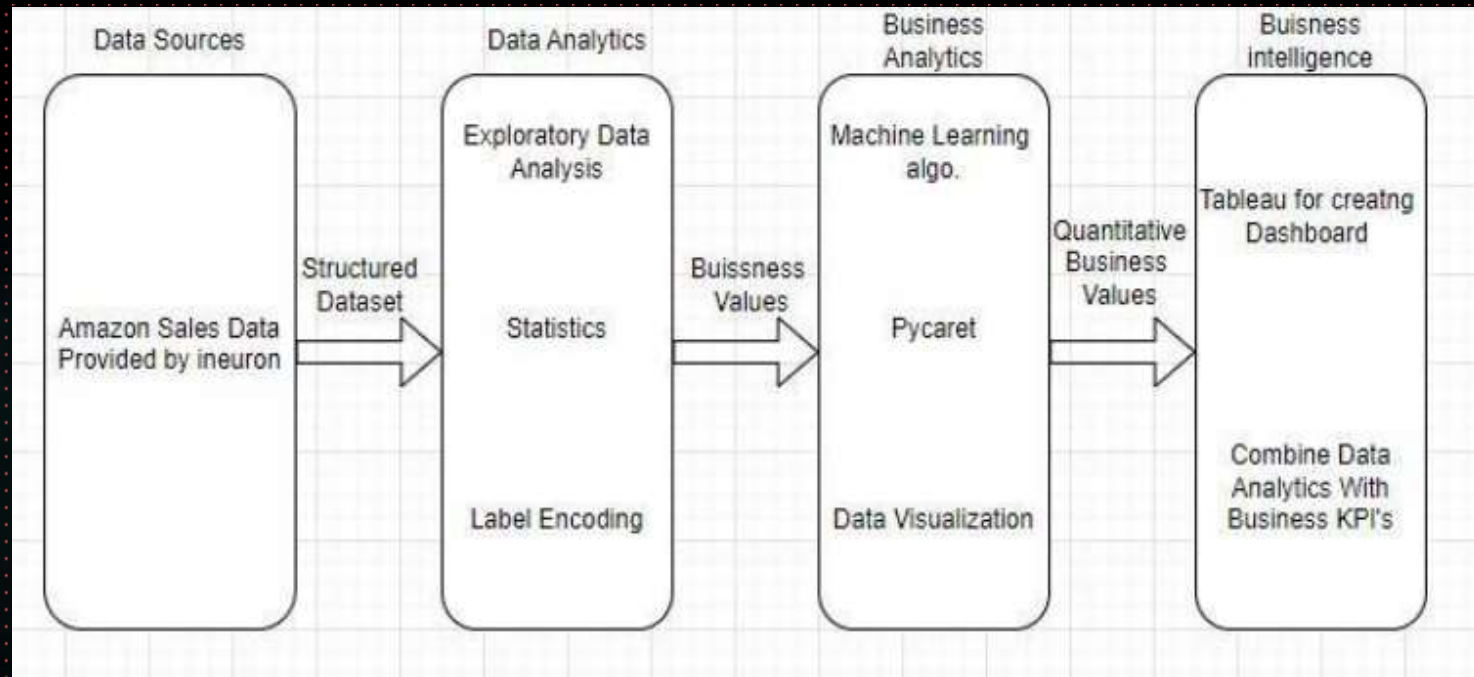
- Development of a **predictive** model for predicting sales.
- Perform ETL (Extract-Transform-Load) on dataset.
- Develop dashboard by using tableau.

# Benefits

- Better understand and optimise revenue generation in future
- Maximize forecasting accuracy
- Make current sales experience our top priority



# Architecture



# Data Preprocessing:

- Importing necessary libraries for data analysis such as : Pandas, Numpy, Matplotlib & Seaborn etc.
- Using `pd.read_csv()` function stores the data in pandas dataframe named data.
- Using `data.columns` showing columns present in dataframe.
- `info()` function show basic information of dataframe like null value count of each column and their data type
- Changing the data type of different column for model training and analysis
- Using `describe` function on dataframe for getting basic stats of numerical dataset
- Adding extra column to dataframe which contain only month, year and month with year
- Using `isnull().sum()` checking out total null value in all the column of dataframe

# Exploratory data

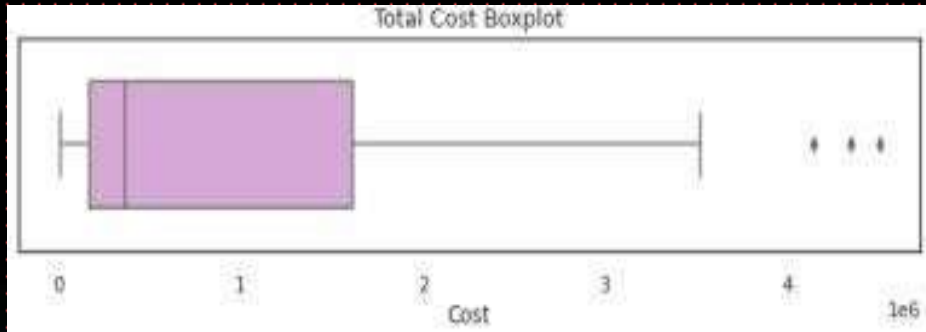
## Analysis

Checking Outliers in the dataframe by using Box Plot

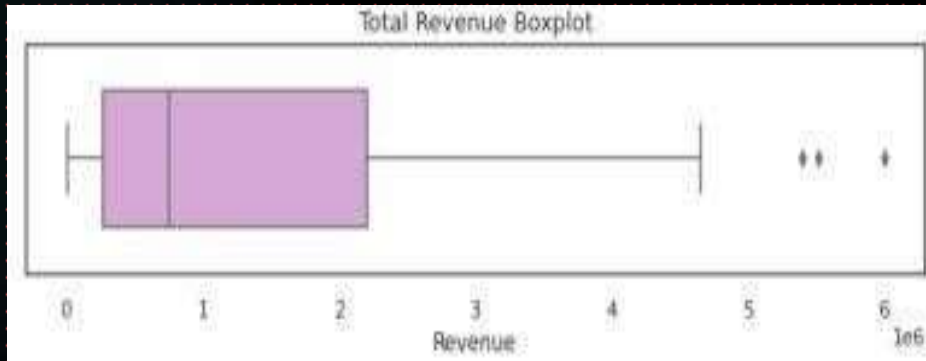
- Box Plot for Total Profit : Here we detect outliers in the specified column using the Z-score method and found 7 outliers.



- Box Plot of Total Cost : found 5 outliers in Total Cost column



- Box Plot of Total Revenue : Found 6 outliers in Total Revenue column



- Creating a bar chart for Total Revenue and Order Month : where it showcases the number of order purchased in particular month.



- Calculating the total revenue for each group with respect to Item Type and then sorting them in descending order.
- Calculating the total profit for each group with respect to Item Type and then sorting them in descending order.

- Calculating correlation of 'Total Revenue', 'Total Cost' and 'Total Profit' columns present in dataframe.

```
print(df[['Total Revenue', 'Total Cost', 'Total Profit']].corr())
```

	Total Revenue	Total Cost	Total Profit
Total Revenue	1.000000	0.983928	0.897327
Total Cost	0.983928	1.000000	0.804091
Total Profit	0.897327	0.804091	1.000000



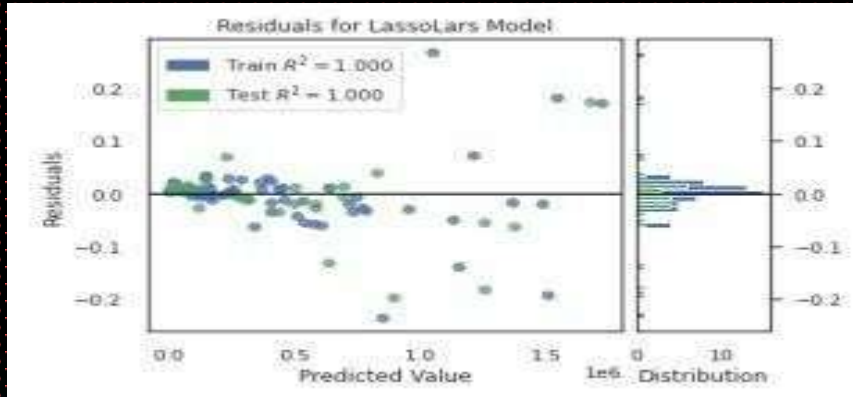
# Predictive Analytics :

- Label Encoding of Item Type, Sales Channel and Order Priority for model training.
- Dropping columns Region, Country, Order Date MonthYear, Order ID and Ship Date.

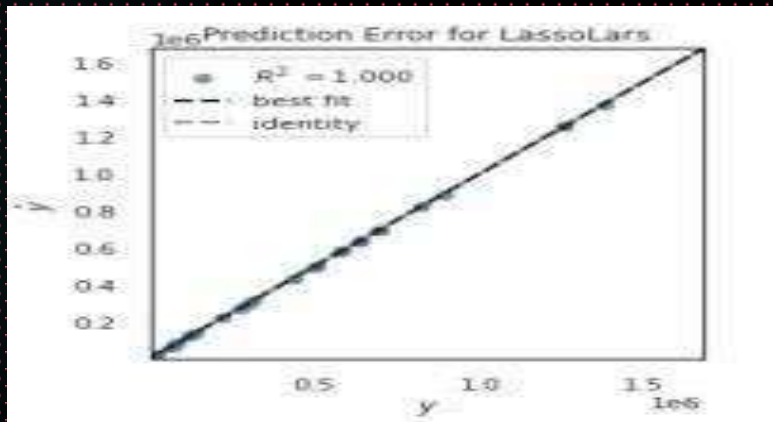
## ► **Pycaaret library :**

- PyCaret is an open-source, low-code machine learning library in Python.
- Allows users to quickly and easily build, compare, and deploy machine learning models on structured and tabular data.
- Reduce the amount of code needed to build a model.
- It provides preprocessing and feature engineering functions.
- Automatic model selection and hyperparameter tuning.

- Plotting residuals for Lasso Least Angle Regression based trained model

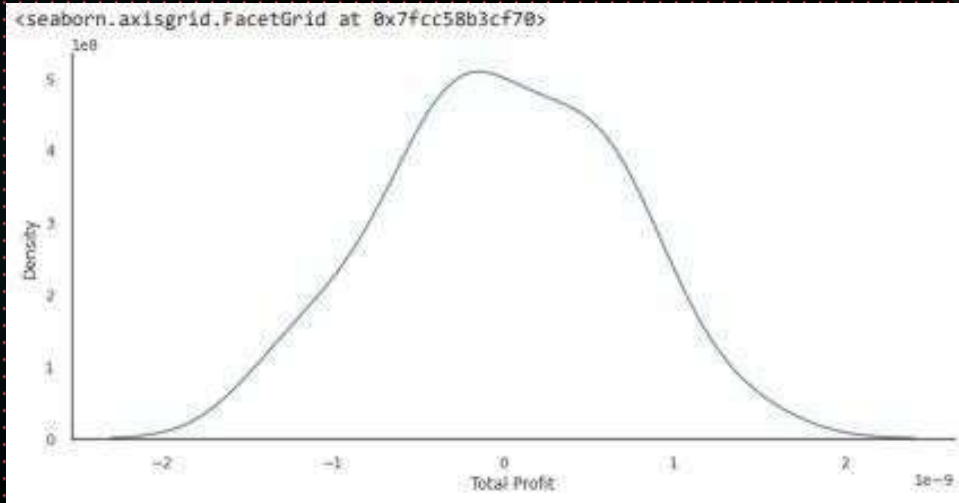


- Plotting prediction error plot for Lasso Least Angle Regression based trained model



# Implementation of Linear Regression

- Selecting the independent variable and the target variable.
- Splitting the data into training and testing datasets.
- Standardizing the dataset.
- Performing fit transform on X\_train dataframe.
- Performing fit transform on X\_test dataframe.
- Applying Linear Regression on X\_train and y\_train.
- Calculating mean squared error.
- Creating kernel density estimate plot



- Plotting the predicted values against the actual values to visualize how well the model is fitting the data.

