WALMART DATA ANALYSIS

Sainath Vineeth Raju Putta
June 21st,2022

Summary

Historical sales data for 45 Walmart stores located in different regions are available. There are certain events and holidays which impact sales on each day. The business is facing a challenge due to unforeseen demands and runs out of stock some times, due to inappropriate machine learning algorithm. Walmart would like to predict the sales and demand accurately. An ideal ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc. The objective is to determine the factors affecting the sales and to analyze the impact of markdowns around holidays on the sales.

Outline

- Business Problem
- Data
- Checking multicollinearity
- Regression modelling
- Results

Business Problem

- Store which makes the highest amount of Sales.
- Which season makes the highest amount of sales.

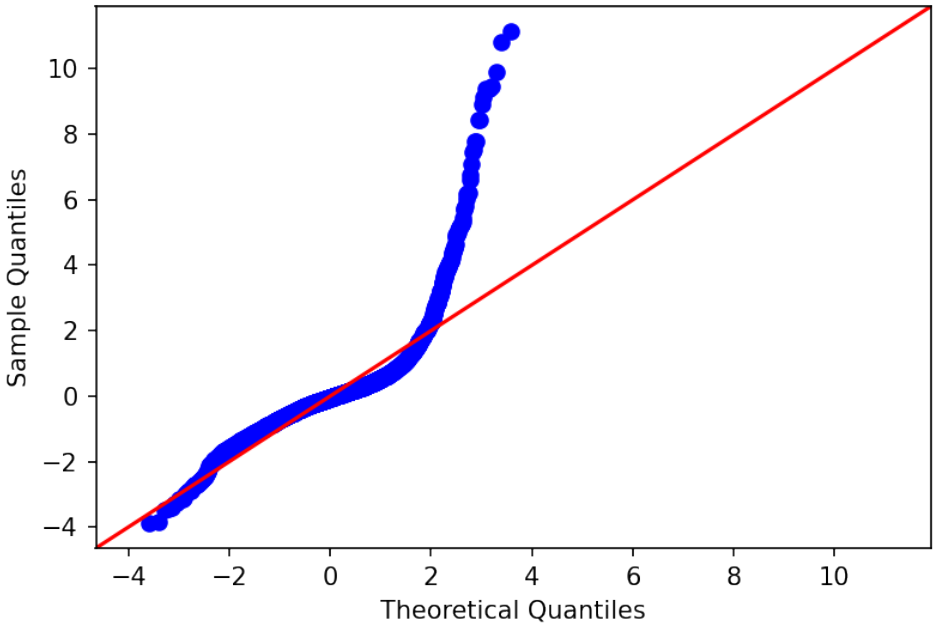
Data Exploration.

- The data set is a single data set and it contains of columns, Store, Date, Weekly_Sales, Holiday_Flag, Temperature, Fuel_price, CPI and Unemployment.
- It has 6435 entries of data.
- As I checked for any clean or missing data, nothing is found and the data types are assigned as per needed other than date column.
- The Date column needs to be sliced into year, weekday and month column.
- The data we have is between the year 2010 to 2012
- They are 4 holidays as per the data explanation given. Super Bowl, Labour Day, Thanks giving, Christmas.

Checking multicollinearity

- Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. We use this method to check the multicollinearity between variables.
- Firstly found few outlier and deleted them.
- Checking for multicollinarity variables using variance inflation factor (Vif).
- Found temperature, fuel price, unemployment and CPI are highly correlated and hence dropped them.

QQ Plot for OLS model.

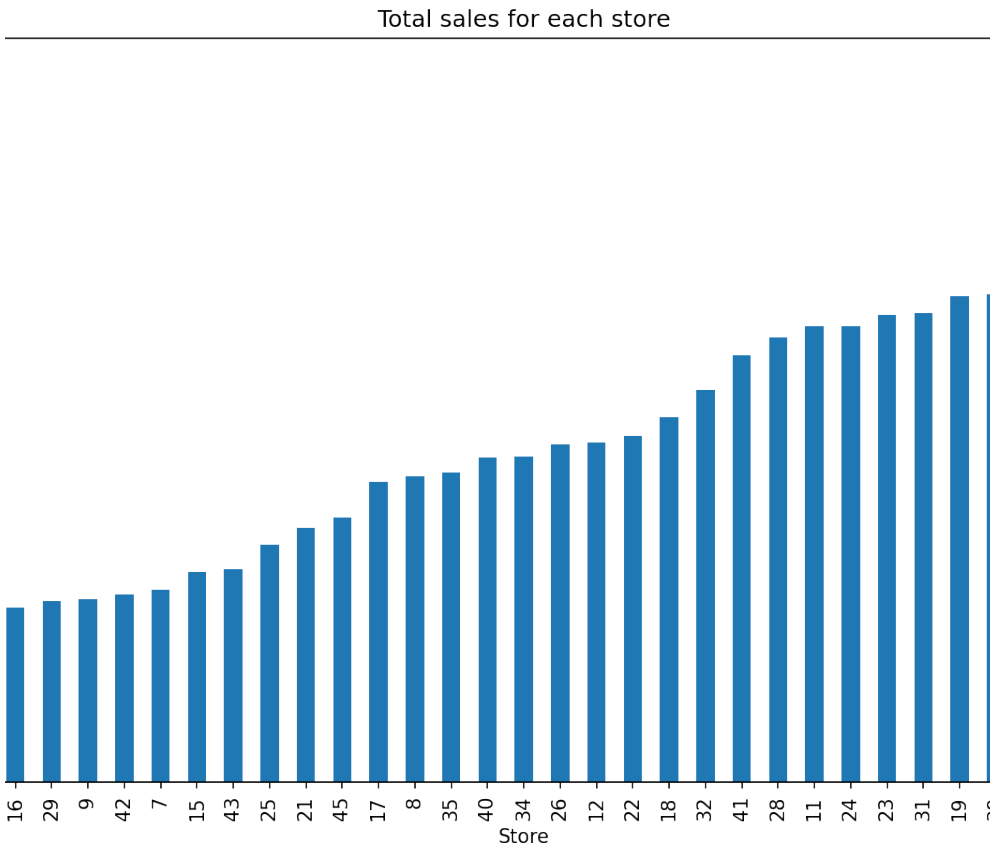


As the QQ plot says we needs to normalise and scale the data.

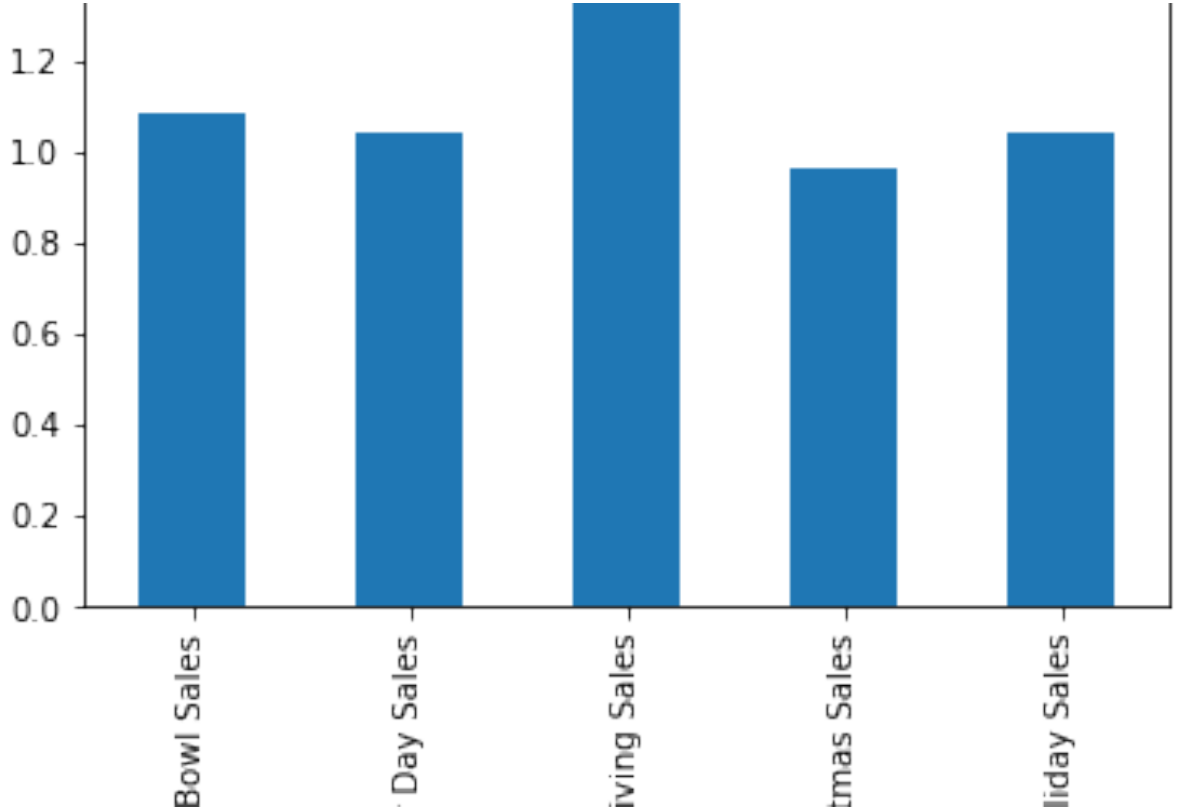
Changing data as per needed for modelling.

- We have created a 80/20 training data set and testing data.
- Standardisation of data is done which means bringing all the variables into a similar scale.
- Linear regression model is built which has a R square value of 90.2% for training data and 92.2% of testing data.

Results



Store 20 makes the the highest amount of sales of all time and store 33 makes the minimum sales.



Thanks Giving makes the highest amount of sales.

Thank You!

Email: vineeth810@gmail.com
GitHub: @vineeth810
LinkedIn: linkedin.com/in/vineeth810