Walmart Retail Analysis

Data Scientist Nanodegree

1- **Walmart**

Walmart is an American multinational retail corporation that operates a chain of hypermarkets, discount department stores, and grocery stores from the United States.

2- **Dataset Description**

This is the historical data that covers sales from 2010-02-05 to 2012-11-01, in the file Walmart_Store_sales. Within this file you will find the following fields:

Store - the store number

Date - the week of sales

Weekly_Sales - sales for the given store

Holiday_Flag - whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week

Temperature - Temperature on the day of sale

Fuel_Price - Cost of fuel in the region

CPI – Prevailing consumer price index

Unemployment - Prevailing unemployment rate

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
0	1	05-02-2010	1643690.90	0	42.31	2.572	211.096358	8.106
1	1	12-02-2010	1641957.44	1	38.51	2.548	211.242170	8.106
2	1	19-02-2010	1611968.17	0	39.93	2.514	211.289143	8.106
3	1	26-02-2010	1409727.59	0	46.63	2.561	211.319643	8.106
4	1	05-03-2010	1554806.68	0	46.50	2.625	211.350143	8.106

3- **Problem Statement**

My objective here is to answer some questions and Build algorithm to predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc.

4- **Exploratory Data Analysis**

We Started to explore our dataset directly as we don't have missing values, and everything looks clean. We just Splinted Date column and create new columns (Day, Month, and Year).

QUESTION 1 :- Which store has maximum sales in this dataset?

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Day	Month	Year
1905	14	2010-12-24	3818686.45	0	30.59	3.141	182.54459	8.724	24	12	2010

From the table above we see that Store 14 has the maximum weekly sales.

QUESTION 2 :- Which store has maximum standard deviation?

	Store	Weekly_Sales	
		std	mean
13	14	317569.949476	2.020978e+06

From From the table above we can see that sales in store 14 very a lot.

QUESTION 3 :- Which store/s has good quarterly growth rate in Q3'2012?

	Store	Q2_Weekly_Sales	Q3_Weekly_Sales	Growth_Rate
15	16	6626133.44	6441311.11	-0.03

	Store	Q2_Weekly_Sales	Q3_Weekly_Sales	Growth_Rate
13	14	24427769.06	20140430.4	-0.18

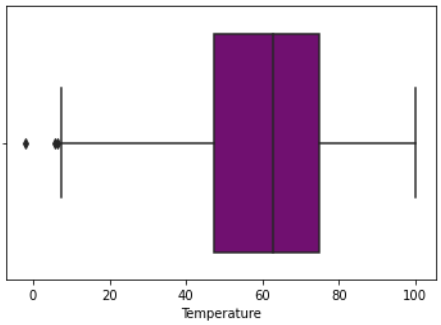
From above tables we can observe that Q3 growth rate is in losses . the Store 16 has the least loss of 3% compared the other stores and store 14 has highest loss of 18%.

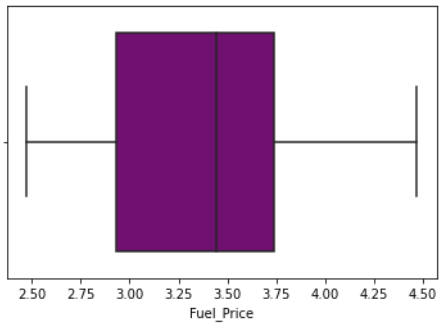
QUESTION 4:- Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together?

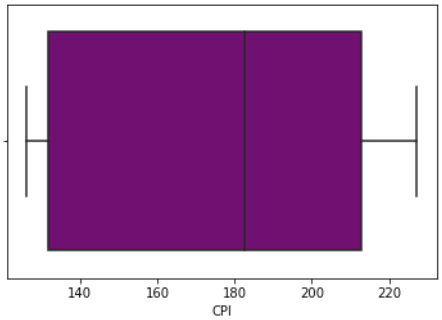
```
Holiday_Flag
0    1.041256e+06
1    1.122888e+06
Name: Weekly_Sales, dtype: float64
```

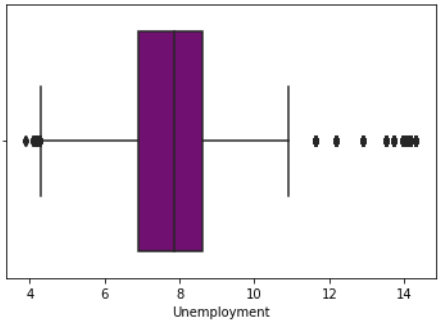
5- **Build prediction models to forecast demand**

We plotted the boxplot to look if there are any outliers

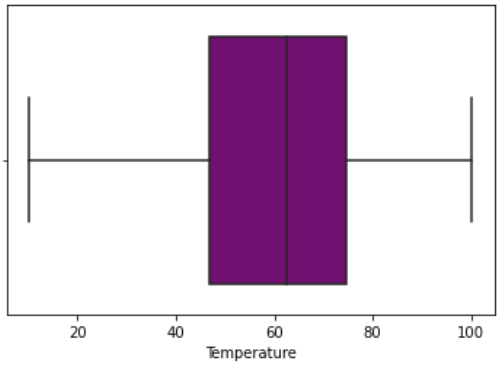


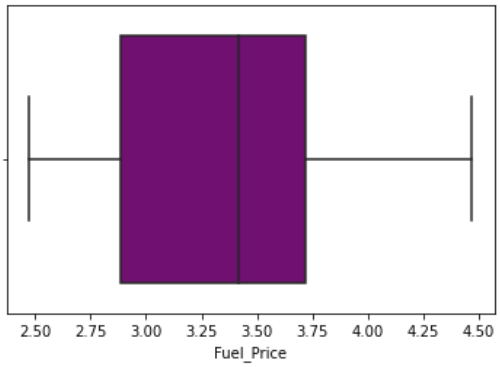


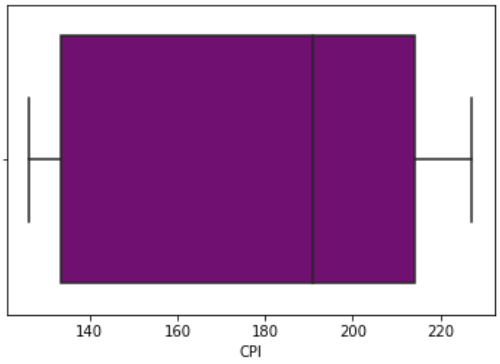


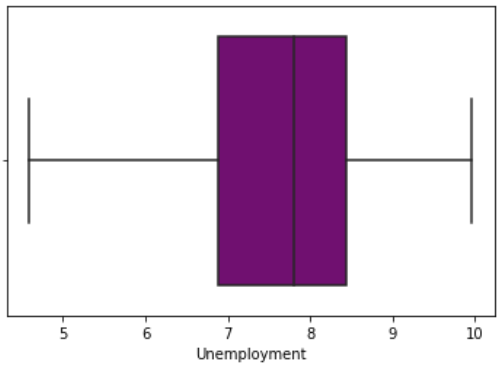


And is it shows, we have outliers, and we should deal with it before working with our model. So we dropped the outliers values.









And our boxplot looks great now!

Building the model

In this part, I tried to make a model that can predict the forecast demand.

Our feature is: Store number, Fuel Price, CPI, Unemployment.

Our target variable: Weekly Sales.

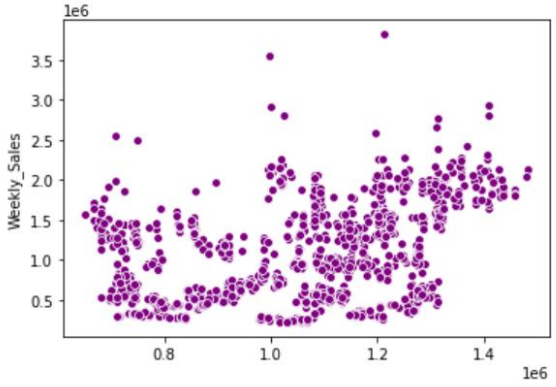
Then, we Split data to train and test (0.80:0.20) and the shape of my features and labels was:

```
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

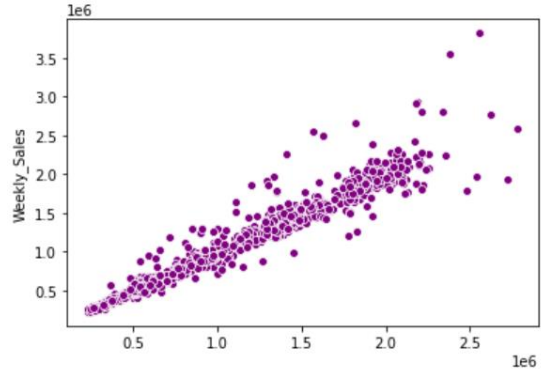
```
((4526, 4), (1132, 4), (4526,), (1132,))
```

Now for the modeling part, I tried two different models, Linear regression, and Random Forest Regressor. And here are the results:

```
Linear Regression:
Accuracy: 12.033326265084732
Mean Absolute Error: 459026.3747789028
Mean Squared Error: 300045954496.7755
Root Mean Squared Error: 547764.5064229477
```




```
Random Forest Regressor:
Accuracy: 94.49386391589769
Mean Absolute Error: 73417.7842749069
Mean Squared Error: 19098725895.97989
Root Mean Squared Error: 138198.13998741042
```



As we can see that we scored 95% accuracy in the Random Forest Regressor.

6- Conclusion:

In this project, I tried to analyze and make model to predict demand accurately and ingest their factors. First, I explored the data and see what I must change before starting the analysis. I didn't find anything to clean, and the data looks good for me. Then I did some exploratory analysis on the data. From that analysis I found out that Store 14 has the maximum weekly sales .and we infer that Q3 growth rate is in losses. And the Store 16 has the least loss of 3% compared the other stores and store 14 has highest loss of 18%. and we found that the mean sales of thanksgiving are more than the non-holiday weekly sales. And Random Forest Regressor has the highest accuracy to predict the demand retail with 94.5.

7- Improvements:

- I will ask more questions and trying more advanced predictive models.
- Trying to get newest and more data like location, offers and time. All these data can help us know how to increase our sales. Also having more data is always good think to help us improve our model results.