Capstone Project -
Walmart Analysis

Table of contentes:

1.Problem Statement:

Walmart is an American retail corporation established in 1962 by sam walton. It operates large chains of department stores and warehouses where they sell products at discounted prices. It is the third largest publicly owned corporation in the world. It operates in several countries. Walmart has been on the receiving end for a long time mainly because of its menial treatment of employees.

- With over 2.2 million employees worldwide, Walmart has faced a torrent of lawsuits and issues with regards to its workforce.
- These issues involve low wages, poor working conditions, inadequate health care, as well as issues involving the company's strong anti-union policies.

2.Project Objectives:

Aim to build a better world - Helping people live better and renew the planet while building thriving, resilient communities, this means working to create opportunity, build a more sustainable future, advance diversity, equity and bring communities closer together.

- Prepare Data for Exploration.
- Process Data from Dirty to Clean.
- Analyze Data to Answer the Problems.
- Share Data through the art of Visualization.
- Apply Machine Learning Algorithms.
- Apply Time Series

3.Data Description:

The Walmart Data is available on kaggle or various websites where you can visit and easily download the dataset.
Also the Project completed with the Python Programming language.
Here, the dataset with 8 variables and 6435 rows.

Import the dataset

```
df = pd.read_csv('/content/Walmart (1).csv')
df.head()
```

	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

```
df.shape
```

```
(6435, 8)
```

The types of the variables.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Store         6435 non-null   int64  
 1   Date          6435 non-null   object 
 2   Weekly_Sales  6435 non-null   float64
 3   Holiday_Flag  6435 non-null   int64  
 4   Temperature   6435 non-null   float64
 5   Fuel_Price    6435 non-null   float64
 6   CPI           6435 non-null   float64
 7   Unemployment  6435 non-null   float64
dtypes: float64(5), int64(2), object(1)
memory usage: 402.3+ KB
```

Now for Online Retail Data is also available on kaggle or various websites where u can visit and download the dataset. Here, the dataset with 8 variables and 541909 rows.

Importing the dataset

```
[ ] df = pd.read_csv('/content/OnlineRetail (3).csv', encoding = 'unicode_escape', parse_dates = ['InvoiceDate'])
    df.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

```
[ ] df.shape
```

```
(541909, 8)
```

The types of the variables.

```
[ ] df.dtypes
```

```
InvoiceNo              object
StockCode              object
Description            object
Quantity                int64
InvoiceDate    datetime64[ns]
UnitPrice             float64
CustomerID            float64
Country                object
dtype: object
```

4.Data Preprocessing Steps and Inspiration:

There are no null and duplicate values in Walmart dataset but in Online Retail dataset Description and CustomerID both variables are with null and unique values, We can’t replace it so have to remove those null values with the dropna() function.

```
[5]  df.isnull().sum()

InvoiceNo            0
StockCode            0
Description       1454
Quantity             0
InvoiceDate          0
UnitPrice            0
CustomerID      135080
Country              0
dtype: int64
```

```
[6]  df = df.dropna()
```

```
df.isnull().sum()

InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

Now check duplicate values and remove them using drop_duplicates() function.

```
[8]  df.duplicated().sum()

5225
```

```
[9]  df = df.drop_duplicates()
```

```
df.duplicated().sum()

0
```

```
[11] df.shape

(401604, 8)
```

After removing the null and duplicate values our dataset is reduced with the 401604 rows.

5.Choosing the Algorithm for the Project:

- Linear Regression
- Lasso
- Ridge
- ElasticNet
- Random Forest Regressor
- Decision Tree Regressor
- KNeighborsRegressor
- Gradient Boosting Regressor

Supervised Machine Learning:- The majority of practical machine learning where you have input variables (x) and an output variable (y) and you use an algorithm to learn the mapping function from the input to the output y=f(x). The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (y) for that data.
Supervised learning problems can be further grouped into **Regression** and **Classification** problems.

The above mentioned algorithms are used for regression problems.

A regression problem is when the output variables are continuous and real values so in Walmart dataset, I used the output variable Weekly_Sales to predict the sales and for Online Retail dataset, I used the UnitPrice and Quantity as Output variable for predictions.

6.Motivation and Reasons for choosing the algorithm:

There are a lot of machine learning algorithms available but we have to try most of them and in which we get less errors and better prediction with a good accuracy score we have to choose that algorithm for our model final predictions.
With our dataset we got a better accuracy score with the Random Forest Regressor so this model is the right fit model for our analysis.

7.Assumptions:

- The chosen sample is representative of the population.
- There is a linear relationship between the independent variables and the dependent variable.
- All the variables are normally distributed; to check plot the histogram of the residuals.
- There are no outliers, if there are outliers they need to be removed using InterQuartile Range.
- The Independent variables are all linearly independent; to check plot the correlation matrix.
- The predictions from each tree must have very low correlation.
- Random forests are nonparametric and can thus handle skewed and multi-modal data

8.Model Evaluation and Techniques:

1) **Mean Absolute Error(MAE):** MAE is a very simple metric which calculates the absolute difference between actual and predicted values. Sum all the errors and divide them by a total number of observations and this is MAE and we aim to get a minimum MAE because this is a loss.
2) **Mean Squared Error(MSE):** MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted values. It represents the squared distance between actual and predicted values. We perform squared to avoid the cancellation of negative terms and it is the benefit of MSE.
3) **Root Mean Squared Error(RMSE):** As RMSE is clear by the name itself, that it is a simple square root of mean squared error.
4) **Root Mean Squared Log Error(RMSLE):** Taking the log of the RMSE metric slows down the scale of error. The metric is very helpful when you are developing a model without calling the inputs. In that case, the output will vary on a large scale. To control this situation of RMSE we take the log of calculated RMSE error and resultant we get as RMSLE.
5) **R Squared(R2):** R2 score is a metric that tells the performance of your model, not the loss in absolute sense that how many wells did your model perform. R2 squared is also known as Coefficient Determination or sometimes also known as Goodness of fit.

9.Inferences for the same:

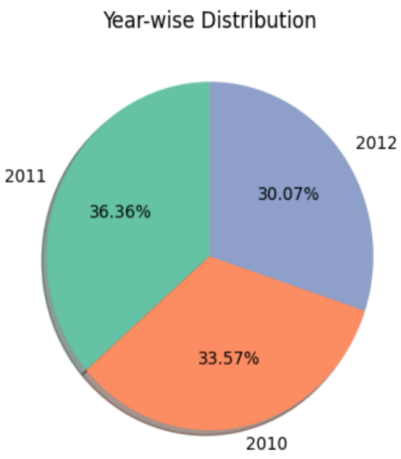
Begin to end with the walmart analysis project with respective points:

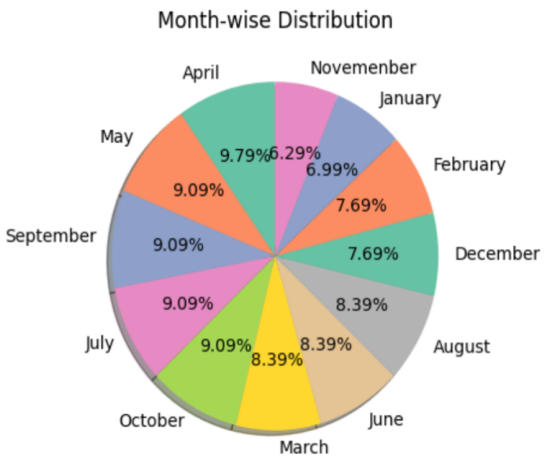
1. Import the libraries.
2. Import the dataset.
3. Check and remove null and duplicate values.
4. Fixing the outliers with InterQuartile range.
5. Split the InvoiceDate column to month, week and hour.
6. Analyzing distribution using histogram.
7. Data exploration using matplotlib visualization library.
8. Import the regressors.
9. Remove unnecessary variables.
10. Splitting the dataset into the training set and test set.
11. Training all regression models on the training set.
12. Predicting the test set results.
13. Evaluating the model's performance.
14. Apply time series.

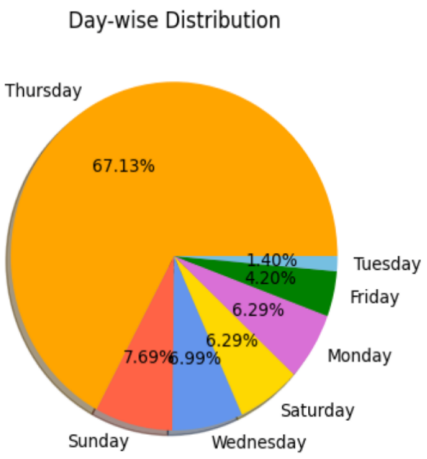
10.Future Possibilities of the Project:

Walmart adapts a low-cost and high volume strategy. The basic principle of strategy is to satisfy customers by offering low prices and exceptional customer service. The retailer offers low prices and quality customer service. Walmart should consider increasing wages, encouraging workers unions, and promoting environmental sustainability in its operations. This will effectively enhance its stature as a retail giant in the global market.

11.Conclusion:





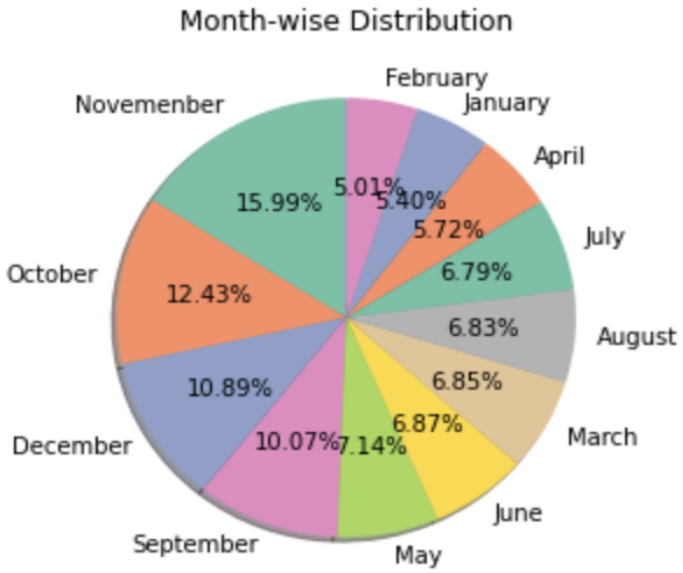


For Walmart data our analysis shows that 2011 has the most sales occurred and also sales during April month is significantly higher than other months, with 9.79%.

The middle of the week day on Thursday 67.13% compared to others day it is almost 10X times more sales occurred.

Also our analysis shows that sales during holidays weeks are significantly higher than during non-holidays weeks, with the sales doubling on average. Additionally, there is a strong seasonal component to the data. The average sales of the top performing stores are up to 500% higher than lowest performing stores.

The best model for predicting future sales is Random Forest Regressor, Which gives 95.33% accurate Walmart Sales Predictions.



For Online Retail data our analysis shows that spending during November is significantly higher than other months, with sales doubling on average.

The best model for predicting future price and quantity is the Random Forest Regressor, which achieved an accuracy score with Unit Price 47.47% and with Quantity 66.02%.

These findings have important implications for businesses as they can inform decisions about inventory, staffing, and marketing efforts. By understanding the factors that drive sales and using a reliable model to forecast future sales, businesses can better plan for the future and optimize their resources.

The Python code file link is click on here.