WELCOME TO WALMART PROJECT REPORT

Contents of the Report

1. *Introduction*
2. *Problem Definition*
3. *Exploratory Data Analysis*
4. *Data Preprocessing*
5. *Model Selection*
6. *Model Evaluation*
7. *Model Deployment*
8. *Conclusion*

INTRODUCTION

Walmart Dataset consist of many feature such as ‘store’ ,‘weekly sales’ ,’unemployment’ , ‘CPI’ , temperature ….., etc. As we have 45 stores in the dataset , and analysis revolves around these stores. By exploring the dataset it is found that this a “Time Series Data”.

And most important feature in data set is ‘weekly sales’ which is associated with Date and store relatively, therefore we are using the time series techniques and algorithms for this dataset.

Problem Definition

1.

a. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?

b. If the weekly sales show a seasonal trend, when and what could be the reason?

c. Does temperature affect the weekly sales in any manner?

d. How is the Consumer Price index affecting the weekly sales of various stores?

e. Top performing stores according to the historical data.

f. The worst performing store, and how significant is the difference between the highest and lowest performing stores.

2. Prediction for Weekly Sales for upcoming 12 weeks.

Exploratory Data Analysis

1. Walmart Dataset was present in csv format and used ‘pandas’ for importing it.

2. Df.info() is used to check the information about the dataset , and insights were there was no null values and ‘date’ column datatype was in int64.

3. df.duplicated().sum() is used to check the duplicated values and no duplicates were found.

4. Groupby function is used to check the record count for each store.

5. Random library is used for the selection of 5 random stores for further analysis.

6. Random selected stores are mapped to store 1,2,3,4,5 for ease applications.

Problems with Insights

1) a. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?

Insights – there is weak negative relationship b/w unemployment rate and weekly sales. But this relationship is more suffered by Store_5 and there is decline in unemployment rate in Store_5 as compared to any other stores.

1) b. If the weekly sales show a seasonal trend, when and what could be the reason?

Insights - All store follows a seasonal trend except Store_5==(Store 42).
The Seasonality shows that there is high Sales in store at time form mid-November to late December.

1) c. Does temperature affect the weekly sales in any manner?

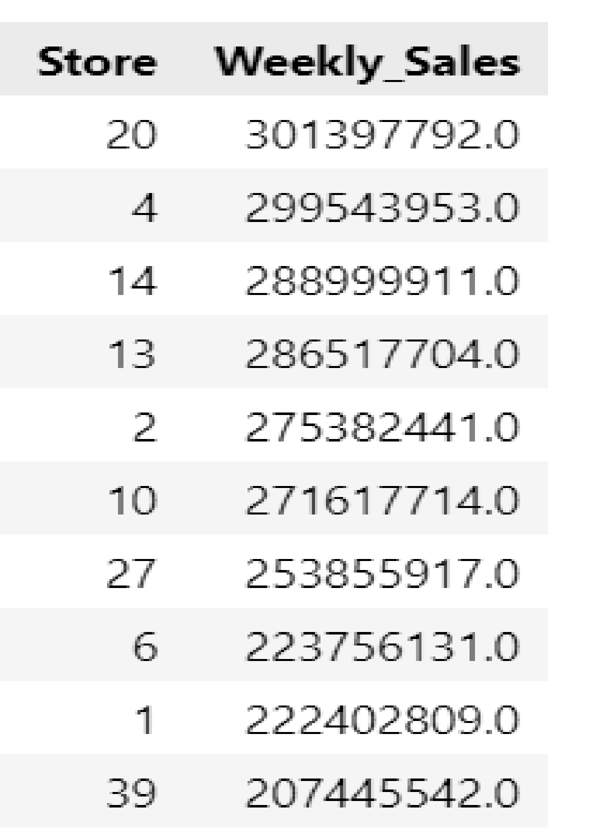
Insights – As there no strong evidences that temperature affects the weekly sales and relationship values are very weak side.

1) d. How is the Consumer Price index affecting the weekly sales of various stores?

Insights - Store 2 and Store 5 have weak positive relationship with CPI and Weekly Sales and others are near to zero. So in our dataset it is possible that some store may have slight positive relationship with CPI and Weekly_Sales but majority of stores have a zero correlation.

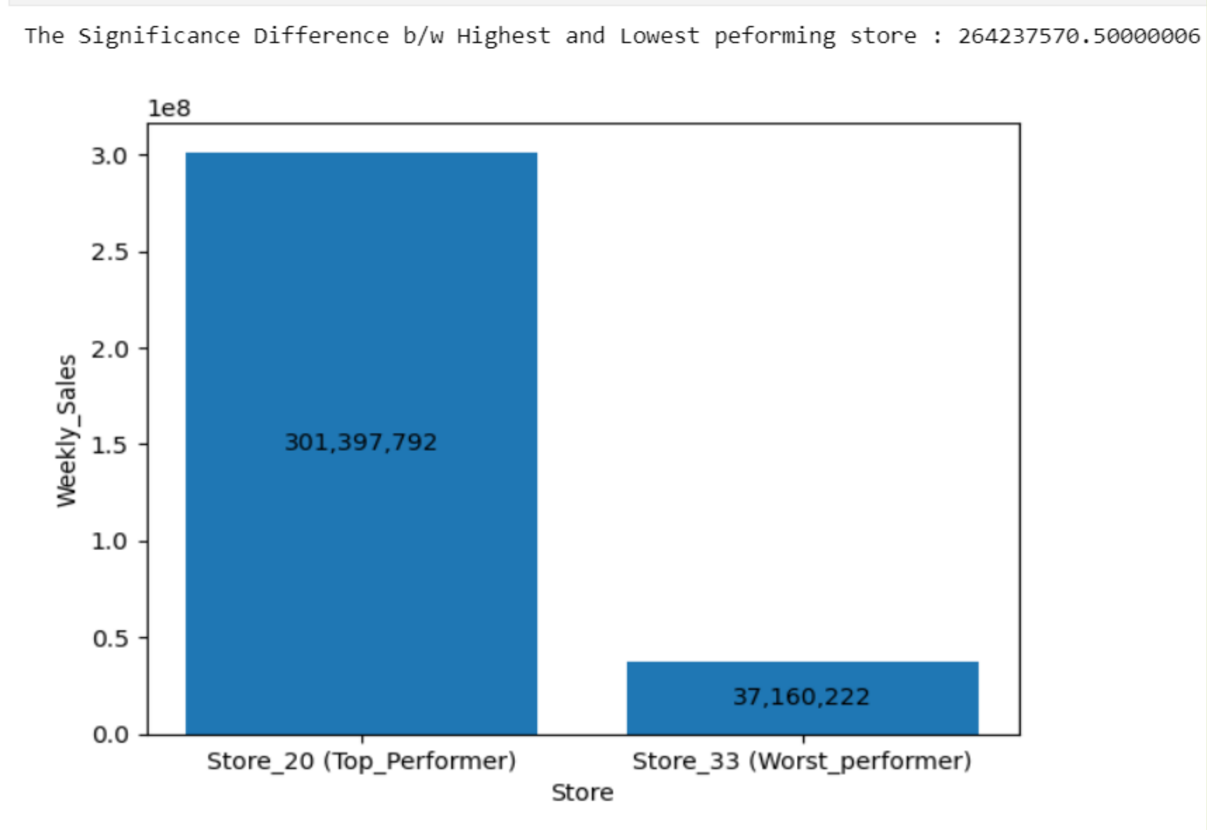
Problems with Insights

1) e. Top performing stores according to the historical data ?



Store	Weekly_Sales
20	301397792.0
4	299543953.0
14	288999911.0
13	286517704.0
2	275382441.0
10	271617714.0
27	253855917.0
6	223756131.0
1	222402809.0
39	207445542.0

1) f. The worst performing store, and how significant is the difference between the highest and lowest performing stores ?



```
The Significance Difference b/w Highest and Lowest peforming store : 264237570.50000006
```

Use predictive modelling techniques to forecast the sales for each store for the next 12 weeks.

1. Select five random stores by the help of random library and used random.seed() , Selected stores are Store 5,14,10,20,45 .

2. Created some functions for code reusability and reduce the length of codes,

a) **'decompose_visual' –** this function is used for seasonal decomposition of data and very useful for analysing trend , seasonality and residual
b) **'adf_test' -** this function is used for checking the stationarity of time series data based on a test of augmented dicky-fuller test. Where null hypothesis says that data is not stationary and vice versa for alternate hypothesis
c) **'kpss_test' –** this function is also used for checking the stationarity of time series data based on Kwiatkowski-Phillips-Schmidt-Shin test, where the null hypothesis says that data is stationary and vice versa for alternate hypothesis.
d) **'acf_pacf_test' –** this function gives the acf_plot and pacf_plot of the time series data and helps to determine and lag values of 'p' and 'q' (AR and MA respectively).
e) **'find_best_order' –** this function is used to find the best order of (p,d,q) based on ARIMA model by using 'itertools' used to create custom ranged tuples for (p,d,q) format, and based on rmse it gives 10 order of (p,d,q) and selecting order have least rmse value.

MODEL BUILDING FLOW FOR TIME SERIES DATA OF STORES

1. Using the seasonal decompose to get components of time series data such as trend , seasonality and residual.
2. Checking for the stationarity of the data which tells about the mean and standard deviation of data, if data is stationary, it is directly treated for further analysis.
3. If data found non-stationary then treating non-stationary data with ‘Log Transformation’ and ‘1st order differencing’ for making it stationary and again check.
4. Plotting ACF and PACF is used to determine the values of ‘p’ and ‘q’ for the selection of model and order (p,d,q).
5. Using itertools and ranged values of ‘p’ and ‘q’ from ACF and PACF plots is helpful finding the best order for (p,d,q) for our ARIMA model.
6. Using those selected values of p,d,q and building ARIMA model.
7. Forecasting of ARIMA model isn’t satisfied because the data consist of seasonality and it is weekly based data , therefore SARIMAX model performs robust with seasonality.
8. Building SARIMAX model by taking the same order of (p,d,q,s) , here ‘s’ stand for seasonality which we can it is ’52 weeks’.
9. Forecasting for 12 weeks for each store by SARIMAX model is more consistent then the ARIMA model.

MODEL SELECTION – *model selection is based on the ACF and PACF plot , whether it is AR , MA , ARMA or ARIMA.*

MODEL EVALUATION – *model evaluation is done by running with multiple combinations order of (p,d,q) , so that we can get best order with less RMSE.*

MODEL DEPLOYMENT – *ARIMA model didn't perform well due to unhandled seasonality in the data , here SARIMAX model comes in a picture and did far better job then ARIMA model.*

CONCLUSION – *In time series data first checkpoint is stationarity of data , second ones is the ACF and PACF plots for model selection and third is finding the best order and modelling with that order.*