Rahul Singh

Walmart Sales Forecasting Model: Report

Problem Statement

Walmart is one of the largest chains of supermarkets in America housing every product that consumers require on a daily basis. We have over two years worth of sales from a total of 45 stores, required to locate key performance indicators. Many large companies have monthly goals to hit their sales target and exceed compared to last year. They use data scientists and analysts to help extract insights that can prevent any loss or forecast incoming cash flow. This process is called sales forecasting, where we determine projected sales for the coming duration of time based on previous sales data. I decided to work on creating a forecasting model to project sales for the company to help make critical decisions impacting sales. It can be used by any companies who want to forecast either ecommerce or in person sales.

Data Import/Cleaning/Wrangling

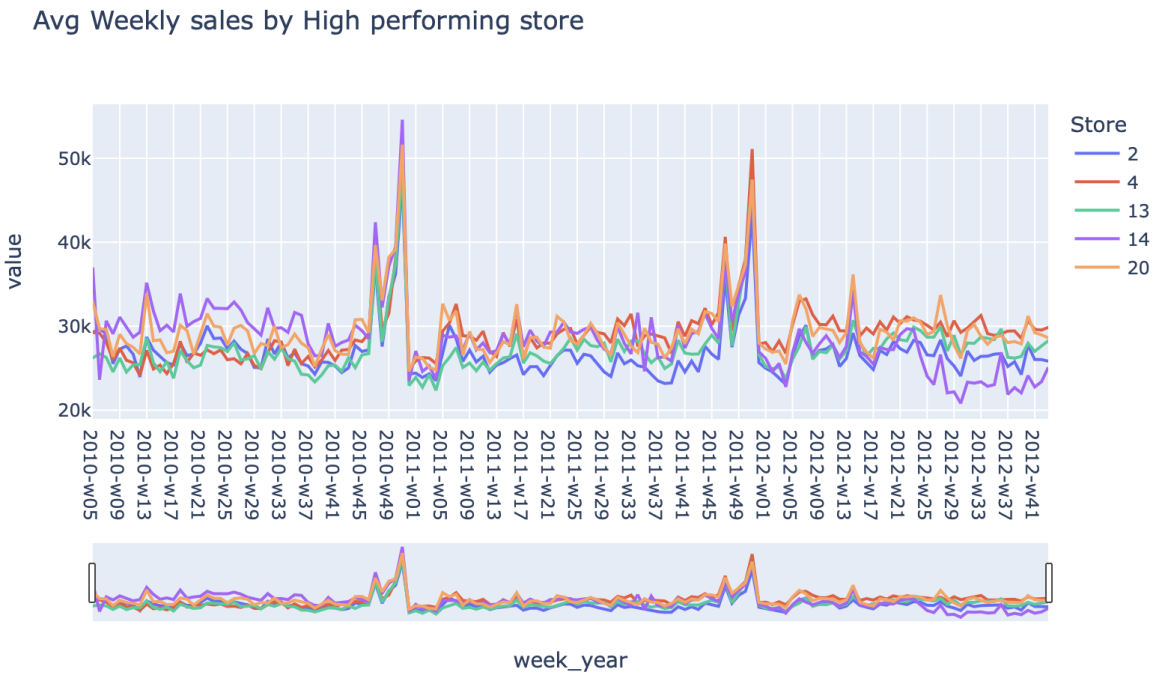
The complete dataset was found on kaggle as it was part of a case competition. The data was found at this link

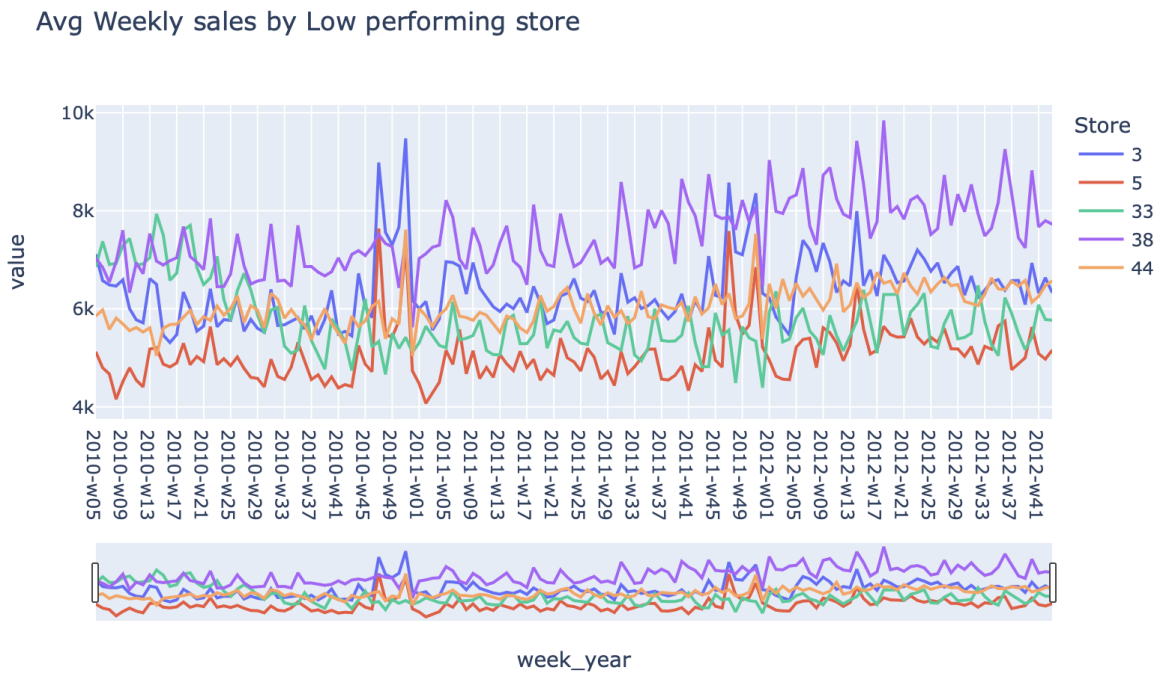
https://www.kaggle.com/competitions/walmart-recruiting-store-sales-forecasting/data

and it has three tables that need to be extracted and cleaned. Initially I reviewed the schema that was provided on kaggle platform regarding the data set. I imported all three tables labeled features,stores and trained each table into critical fields. First we evaluated the high level summary of each table and isolated each field that needs to be stored for a combined data frame. After joining all data in a high level dataframe labeled df which included stores,dept,weekly_sales,temperature,fuel_price,unemployment, size and markdown. Then I checked for null values and only the markdown field had some missing values that needed to be replaced with appropriate integer values. This concludes the initial process for this project as the data was already cleaned and wrangled beforehand.

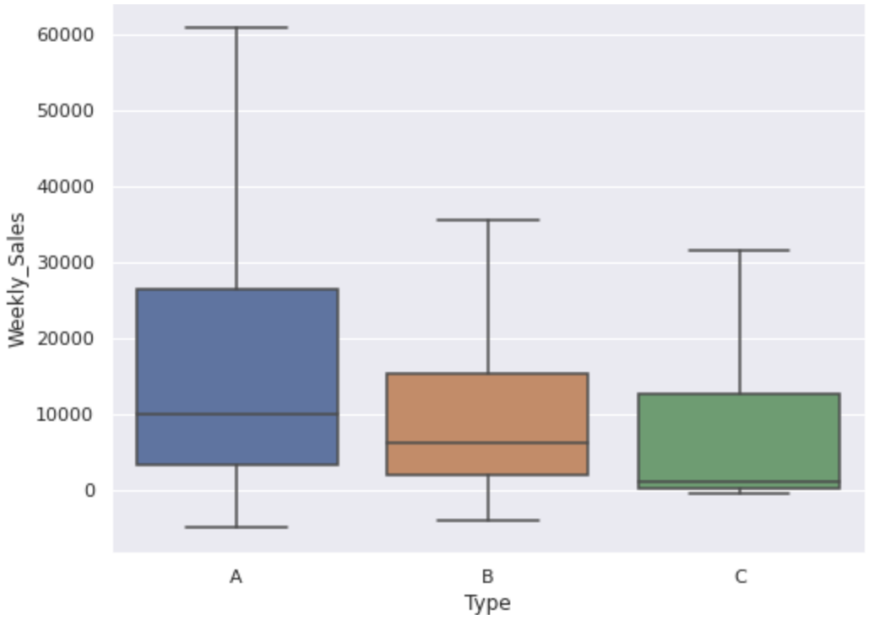
Exploratory Data Analysis

After the majority of data has been structured into respective fields, I divided the fields between categorical and numerical variables to extract any key insights before moving forward. Initially we sort the stores based on its performance and evaluate the top 5 high performing and low performing stores. I started by observing trends in the sales of all stores over the time span shown below. I noticed that we had spikes in sales during the last couples weeks of the year, this could be due to holidays or annual sales.

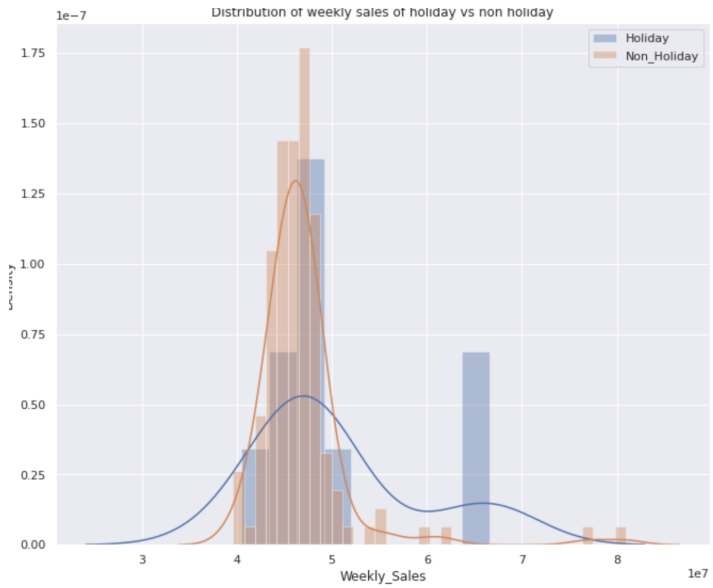




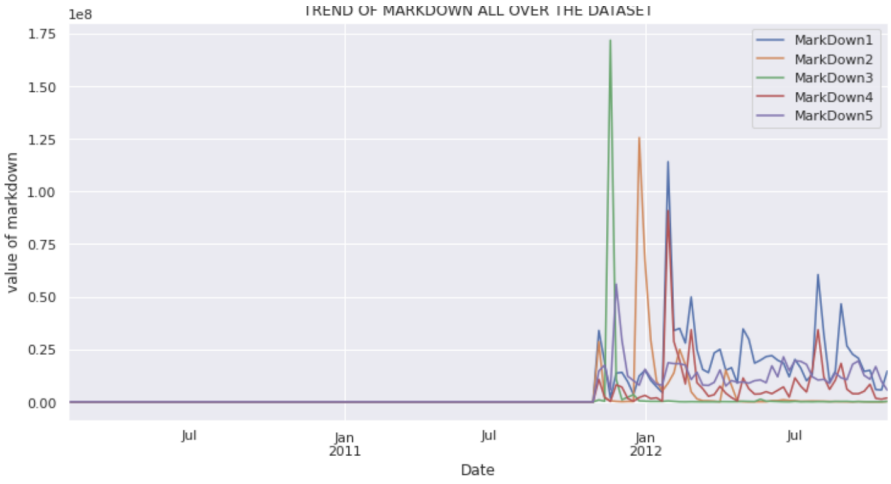
Next we started observing the categorical variable of different types of stores, the goal was to see sales output from different types of stores (A,B & C). The box plot below shows that type A stores have the highest sales with given mean and range, followed by B and C type of stores. Another key insight was store type C had the lowest sales, that needs to be further investigated.



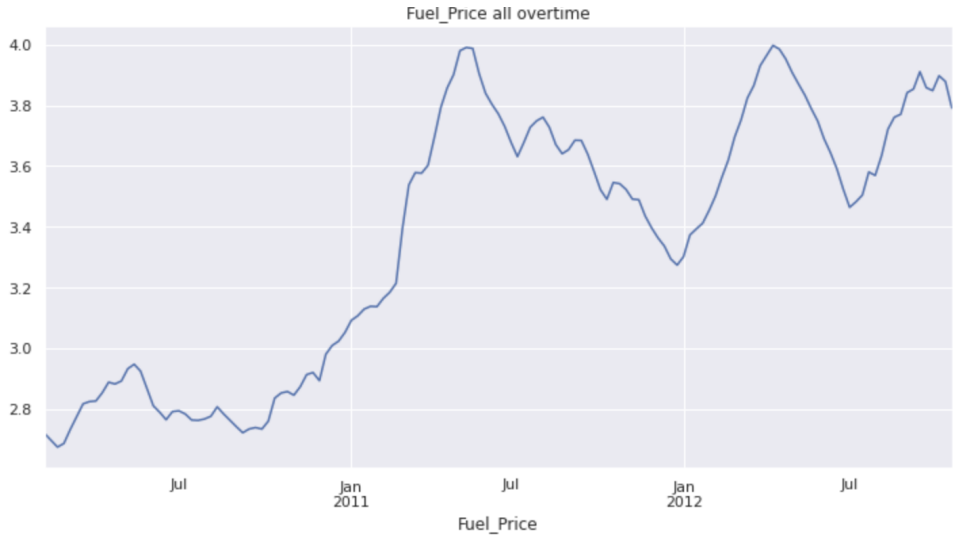
Holiday sales is another interesting field that requires exploring and once i plotted the field, to my surprise distribution of data overlapped each other. We can see it being holiday brings in a little more sales than non-holiday, which is unusual and needs to be investigated for outliers.

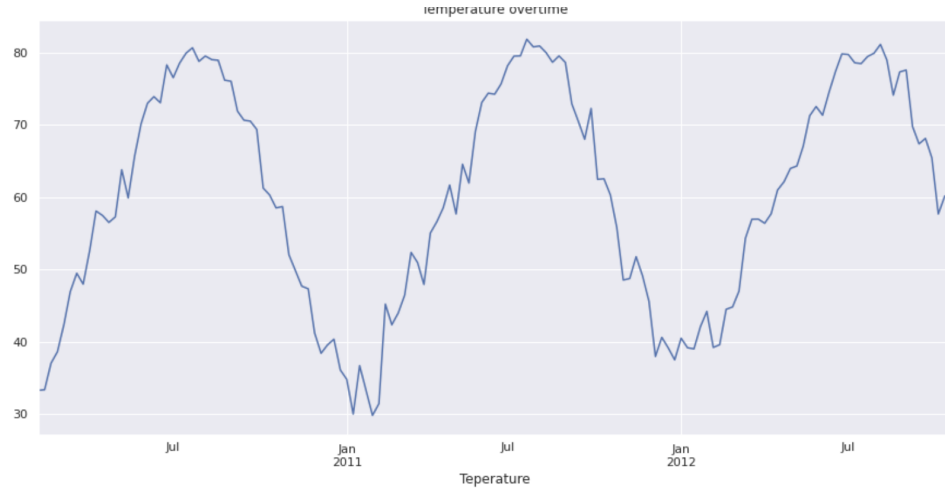


Next was the markdown field that we encoded into 1 and 0 and plotted the results against date value. As shown on the graph below we can see that the markdown on products didn't start until the beginning of 2012. This could be some economical step to help drive sales up or some global issue.

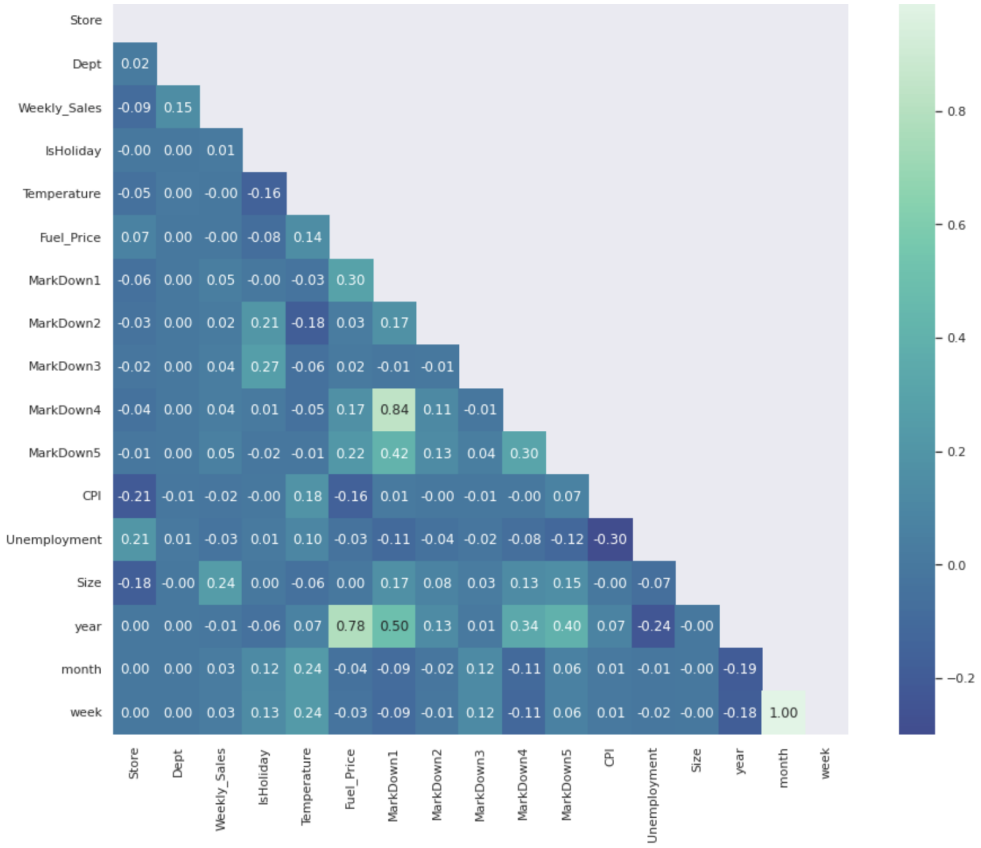


Fuel prices and temperature were the next fields that were interesting to evaluate, each variable can impact sales for a store as its driving factor for the consumer. We can see that fuel prices increase overtime and this could have an impact on sales and it constrains the consumer from visiting the store as pricing increases. Same logic can be applied towards temperature months with high temperature sales will be low as its less likely for consumers to shop at stores.





Lastly for EDA i plotted all the fields in a correlation matrix to see feature importance as shown below. The top positive correlations were size , debt and markdown compared to weekly_sales.



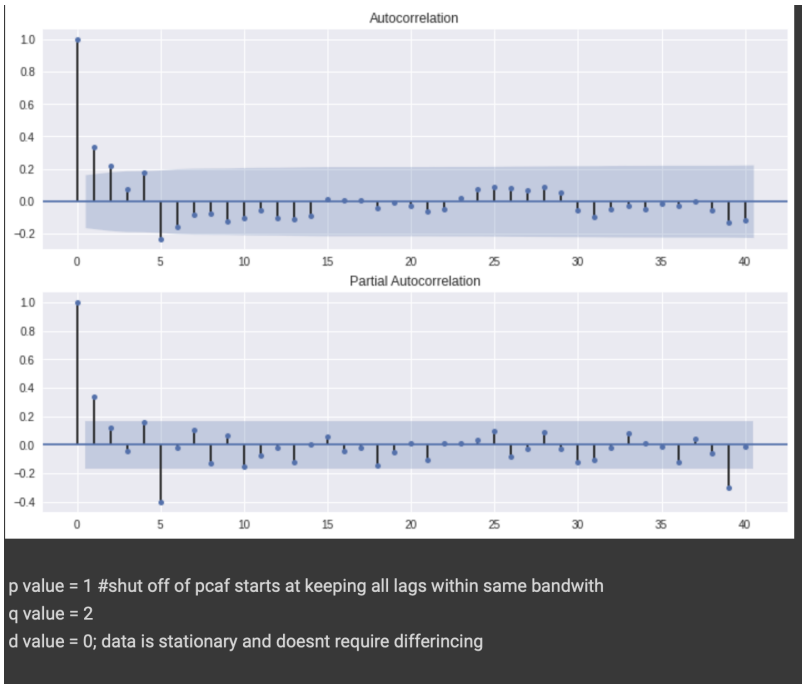
Pre-processing

After getting insights on the data, I moved to prepping the data for modeling and output results. We test the dataset for stationary to determine model needs differining before fitting. We perform stationary by performing a dickey-fuller test on weekly sales data, the hypothesis testing will be null hypothesis is non stationary if the p-value is greater than 0.05. Post results show that the dataset is stationary and we can proceed to correlation plots.

```
[ ]  adfuller_test(df2['Weekly_Sales'])

     ADF Test Statistic:-5.908297957186334
     p-value:2.675979158986027e-07
     #Lags Used:4
     Number of Observerations used:138
     Time seris : It is stationary
```

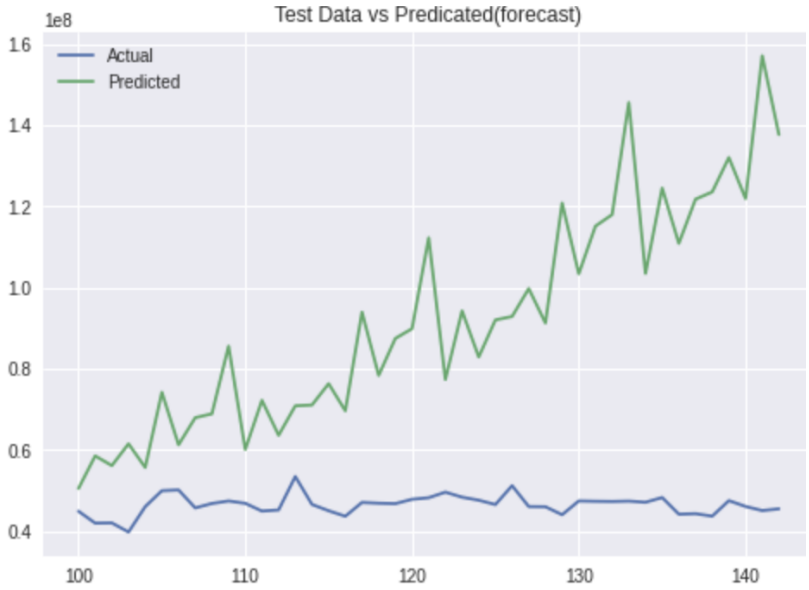
Next step is to determine the p, d, q values for our model and those are obtained through autocorrelation and partial autocorrelation graphs. Shown below are the set graphs and we can see the shut off for both graphs occur at 1 and 2, with the autocorrelation graph we get the p value and q value is obtained from the pacf graph.



Modeling

Now that we have our pdq values we can start modeling our data, first we split the data into train/test and fit the model with pdq values from the graph. Results show a large variance in the predicted and actual values for test data. Hyper Tuning the model is required to get best results output, for that we take a cross validation approach. Initially an array with pdq values ranging

from 0 to 2 and run a loop to obtain best values at pdq = 2,1,1 and seasonal pdq = 0,0,2,12 at lowest mean absolute percent error. After getting those values we are able to run the model again with the best values for the test and see our projections as shown below on the graph.



	non-seasonal_pdq	seasonal_pdq	mape
0	2,1,1	0,0,2,12	0.041689
0	2,1,2	0,0,2,12	0.044517
0	2,0,1	0,0,0,12	0.045384
0	2,0,1	0,0,2,12	0.052472
0	2,0,1	2,0,0,12	0.063504
...	...	...	...
0	1,2,0	0,1,0,12	13.320045
0	1,2,0	0,2,1,12	14.045285
0	2,2,0	0,2,0,12	14.123210
0	1,2,0	1,2,0,12	15.050972
0	1,2,0	0,2,0,12	26.850498

458 rows × 3 columns

Now that we have the best model we can forecast for the next projected 2 months, to see sales values we are expecting. Graph below shows a projection of future sales for the next 2 months and we can see that the predicted values are quite linear.

Conclusion/Future Work

The model still needs to be tuned for better output to showcase a higher percentage of confidence interval on data. We can try to use a different type of model for prediction like a random forest or linear regression.