Walmart Forecasting

Problem Statement

As the largest retail corporation of the world, Walmart must have the most customers and experience the most changes in a year. With things like seasonality and different seasonal sales going on, it is hard to predict Walmart price. The goal of the project is to create a model that will more accurately predict the price of Walmart's sales with time-series analysis and other regression models.

1. Data Collection

The Walmart data is collected from Kaggle (Here)

2. Data Wrangling

Missing Values:

The only missing values we had in the dataset was some of our stock and sp500 prices. We simply used forward filling for those missing values, since the missing prices occured on Sundays, which we could use the price of that Friday.

The first thing I did was to merge the S&P 500 data and the stock price of Walmart with our original dataframe.

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	stock_price	sp500
0	1	2010-02-05	1643690.90	0	42.31	2.572	211.096358	8.106	53.45000	1066.19
1	1	2010-02-12	1641957.44	1	38.51	2.548	211.242170	8.106	52.89999	1075.51
2	1	2010-02-19	1611968.17	0	39.93	2.514	211.289143	8.106	53.49001	1109.17
3	1	2010-02-26	1409727.59	0	46.63	2.561	211.319643	8.106	54.07001	1104.49
4	1	2010-03-05	1554806.68	0	46.50	2.625	211.350143	8.106	54.14000	1138.69

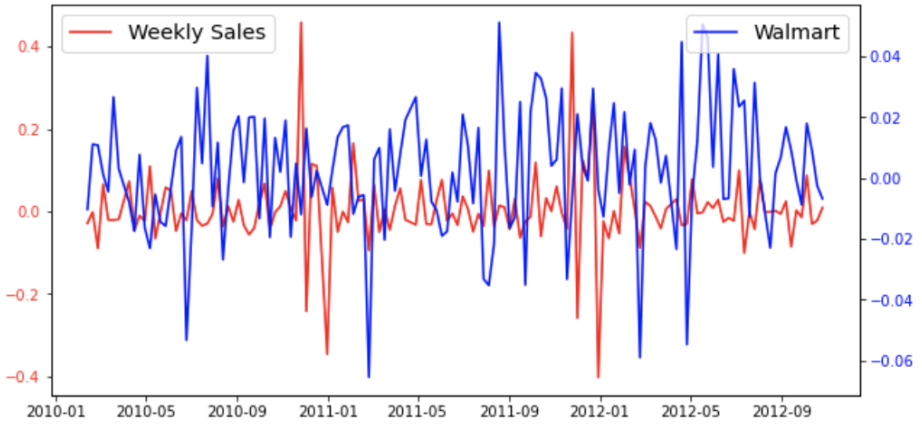
There are 45 stores in this dataset, and each of them has 131 dates associated with them. The target variable of our study is the weekly sales category in which we are trying to conduct supervised learning to make our own regressor to forecast the weekly sales price.

Alternative Dataset:

Although we have 45 stores' data over a span of roughly two and a half years, we still do not have enough data to generate a model for each of the stores we have. What I did instead is to group by all of our datasets together to get the total weekly sales of all of our stores. For data like temperature, fuel price, and unemployment of our group by data, we got the average number of all 45 stores to get the number.

3. EDA

Weekly Sales Vs. Walmart Stock Prices (Percentage Change)

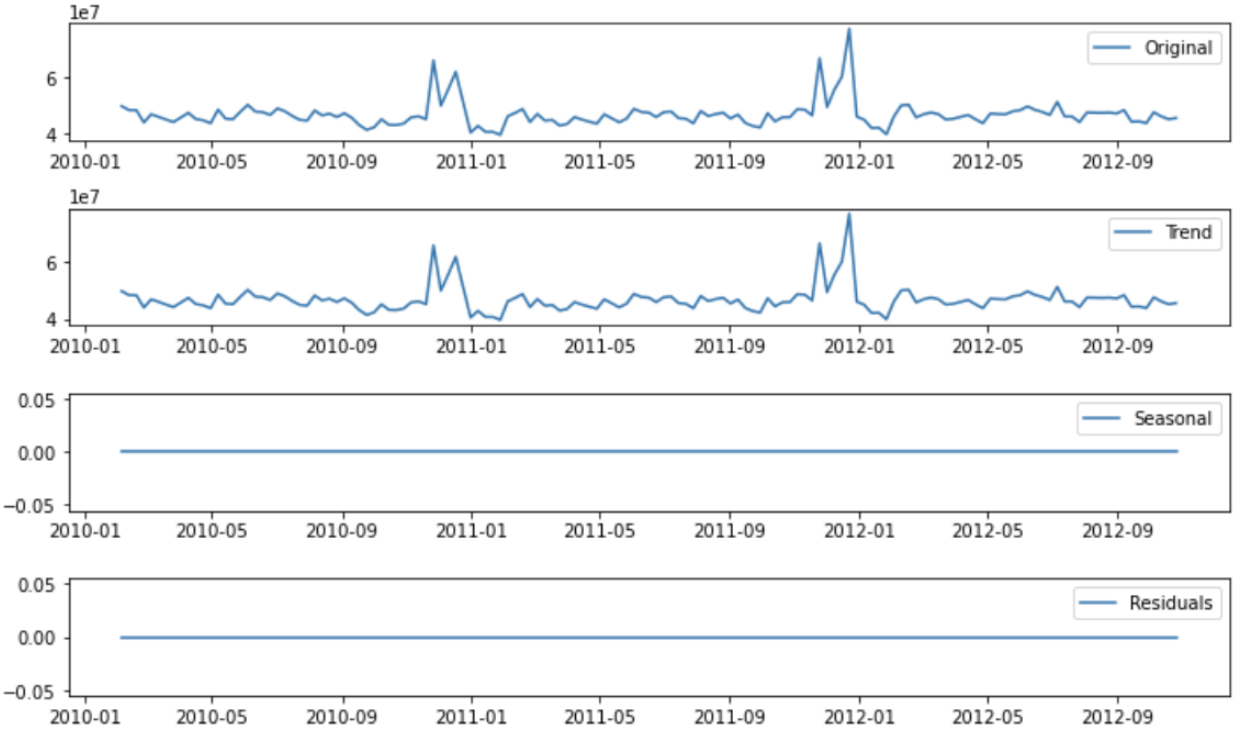


Weekly Stocks Vs. S&P 500

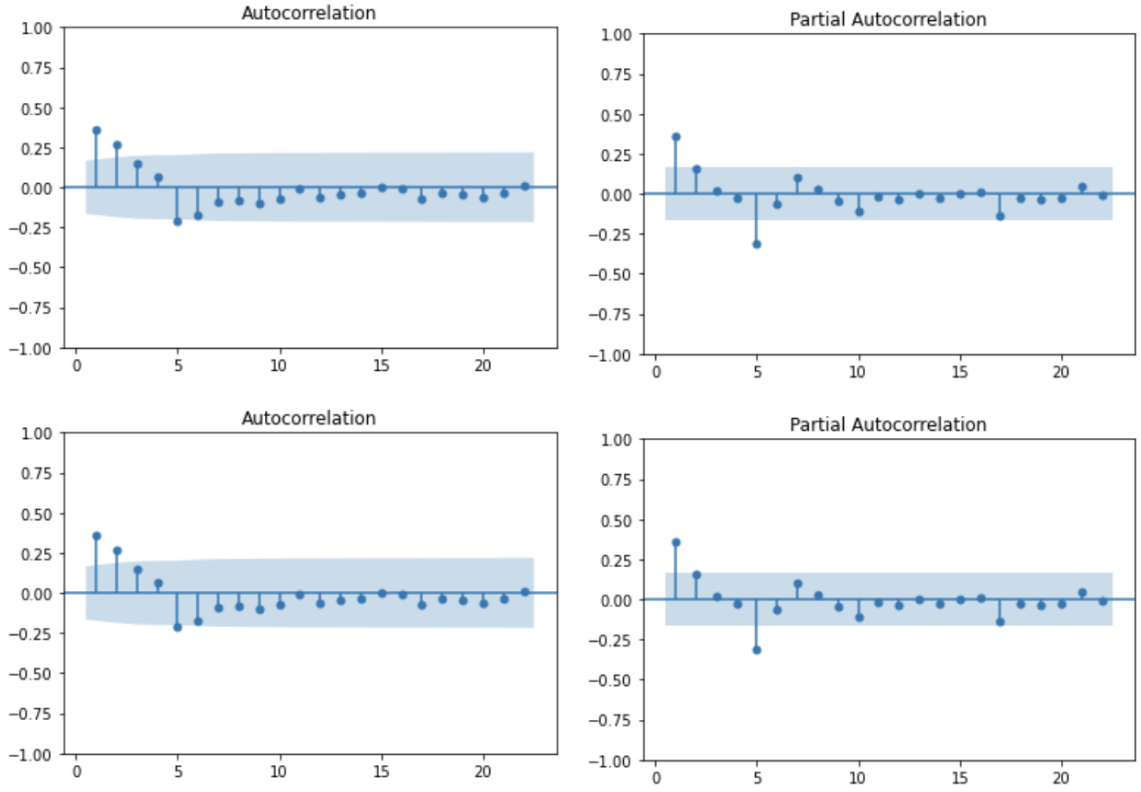
We can see some correlation between the two stock prices.

After our AD Fuller test, we have also concluded that the Walmart Weekly Sales prices is indeed a random walk.

Seasonality



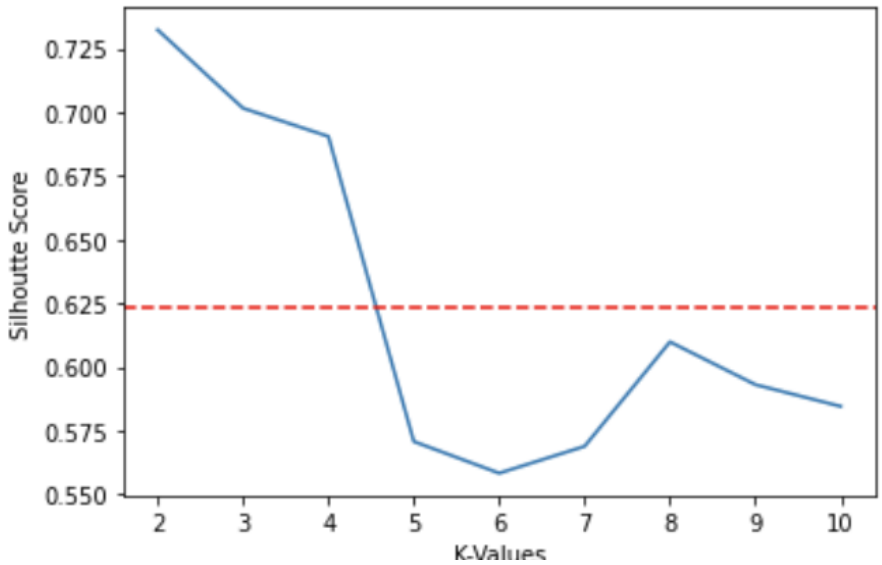
Autocorrelation & Partial Autocorrelation

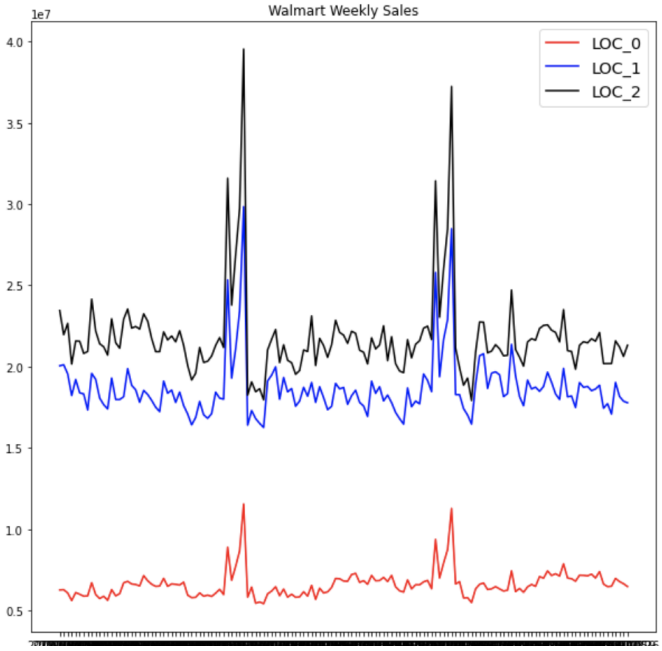


From the ACF and PACF, we can see that we need to build an ARIMA model for our data, since both the ACF and PACF have tails cut off.

4. Pre-Processing

We did two things for pre-processing. First of all, since we have 45 stores, we could try to cluster them into different groups for easier analysis. The reason for that is we could split the data into several groups based on their temperature and gas price, since in a similar geological location, the stores might share similar temperature and gas prices. From our silhouette score, we split the data into three groups.





Moreover, we can see that location seems like it has some impact on the weekly sales, as the average weekly sales of stores grouped by locations deviate.

Alternative Data with locations and various similar dates

	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	stock_price	sp500	loc_0	loc_1	loc_2
0	2010-02-05	1643690.90	0	42.31	2.572	211.096358	8.106	53.45000	1066.19	0	1	0
1	2010-02-12	1641957.44	1	38.51	2.548	211.242170	8.106	52.89999	1075.51	0	1	0
2	2010-02-19	1611968.17	0	39.93	2.514	211.289143	8.106	53.49001	1109.17	0	1	0
3	2010-02-26	1409727.59	0	46.63	2.561	211.319643	8.106	54.07001	1104.49	0	1	0
4	2010-03-05	1554806.68	0	46.50	2.625	211.350143	8.106	54.14000	1138.69	0	1	0

Data with unique dates and total weekly sales

	time	Weekly_Sales	Walmart	sp500	Temperature	Fuel_Price	CPI	Unemployment	Holiday_Flag
0	2010-02-05	49750740.50	53.45000	1066.19	34.037333	2.717844	167.730885	8.619311	0
1	2010-02-12	48336677.63	52.89999	1075.51	34.151333	2.694022	167.825608	8.619311	1
2	2010-02-19	48276993.78	53.49001	1109.17	37.719778	2.672067	167.871686	8.619311	0
3	2010-02-26	43968571.13	54.07001	1104.49	39.243556	2.683933	167.909657	8.619311	0
4	2010-03-05	46871470.30	54.14000	1138.69	42.917333	2.731200	167.947628	8.619311	0

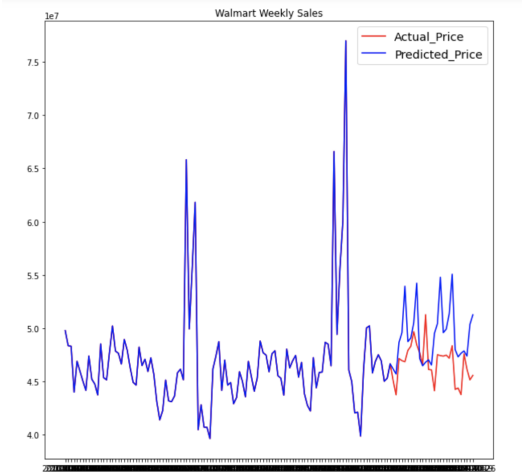
The other thing we did was one hot encoding for our categorical variables.

	time	Weekly_Sales	Walmart	sp500	Temperature	Fuel_Price	CPI	Unemployment	Holiday_Flag	Target
0	2010-02-05	49750740.50	53.45000	1066.19	34.037333	2.717844	167.730885	8.619311	0	-0.028423
1	2010-02-12	48336677.63	52.89999	1075.51	34.151333	2.694022	167.825608	8.619311	1	-0.001235
2	2010-02-19	48276993.78	53.49001	1109.17	37.719778	2.672067	167.871686	8.619311	0	-0.089244
3	2010-02-26	43968571.13	54.07001	1104.49	39.243556	2.683933	167.909657	8.619311	0	0.066022
4	2010-03-05	46871470.30	54.14000	1138.69	42.917333	2.731200	167.947628	8.619311	0	-0.020184
...	...	...	...	...	...	...	...	...	...	...
133	2012-09-21	44354547.11	74.45000	1460.15	67.924889	3.907911	176.242124	7.237333	0	-0.013970
134	2012-09-28	43734899.40	73.80000	1440.67	68.754444	3.854578	176.373588	7.237333	0	0.087613
135	2012-10-05	47566639.31	75.13000	1460.93	65.973111	3.845222	176.505052	6.953711	0	-0.030234
136	2012-10-12	46128514.25	75.81000	1428.59	58.342667	3.896733	176.636515	6.953711	0	-0.021811
137	2012-10-19	45122410.57	75.62000	1433.19	60.705333	3.880000	176.652613	6.953711	0	0.009346

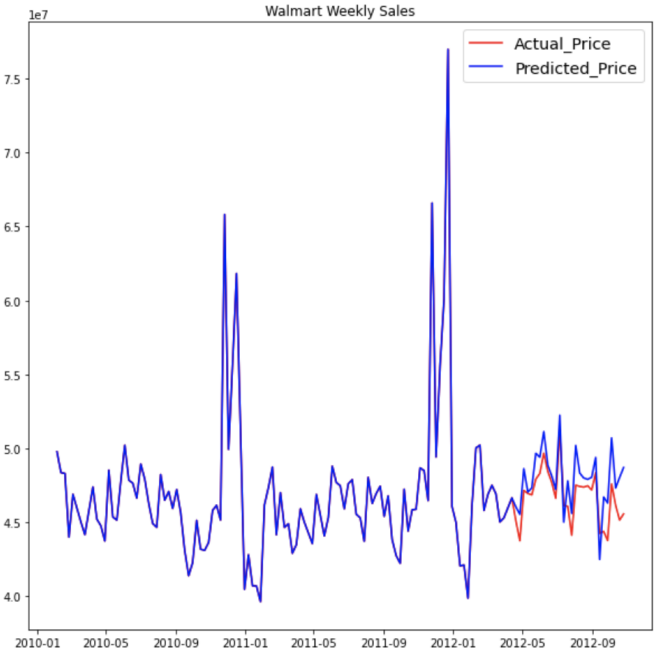
Also, for our Sarimax model, we have a target variable, which is the percentage change of weekly sales compared to the weekly sales after that specific week. Essentially, we exclude the last record, and our goal is to use everything we have in one day, including the weekly sales, to predict the weekly sales for next week.

5. Modeling

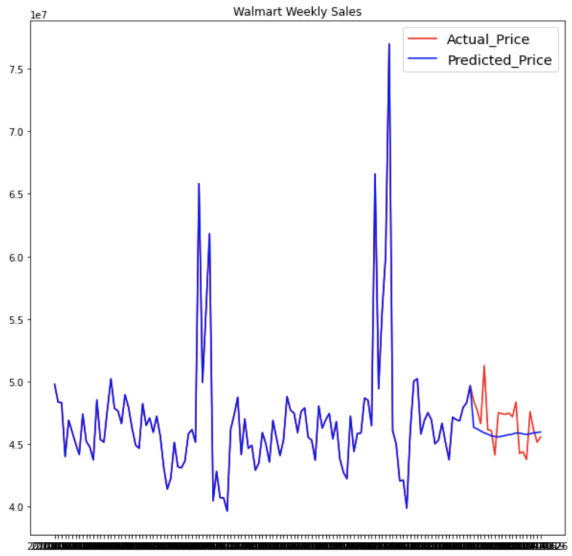
I. ARIMA Model for all stores (One variable)



II. SARIMAX with Exogenous Variables

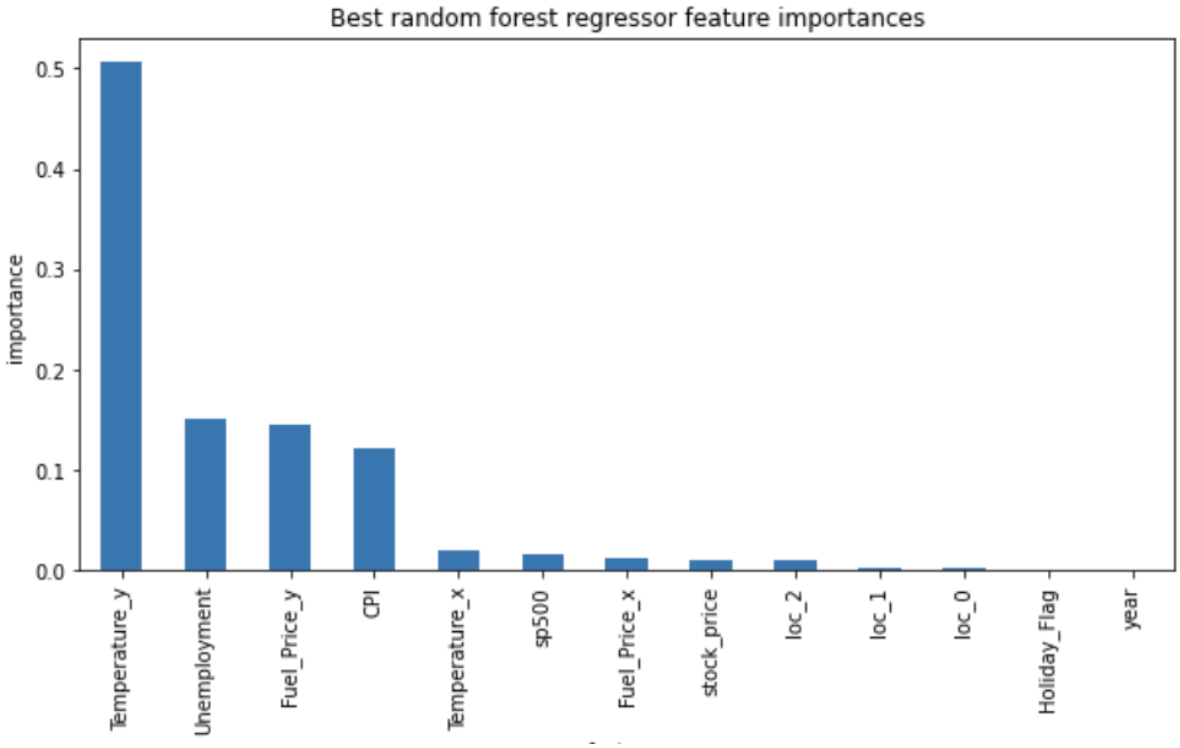


III. LSTM with two layers (Parallel Time Series). It uses five variable to predict weekly sales in the following date



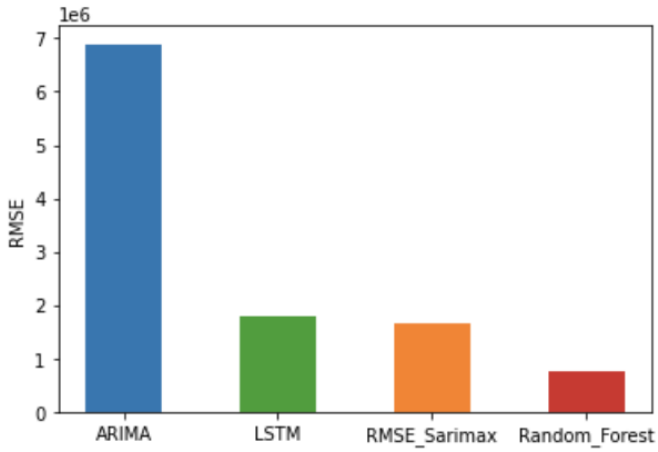
IV. Random Forest

For our random forest model, we dropped the date and instead used all the other variables we have to see whether we could predict an accurate weekly sales price from other variables of our study.



6. Model Selection

To choose our final model, we use the RMSE of the models as our metric to measure the model's performance.



The Sarimax model with exogenous variables gives fairly accurate predictions, and it is a time series model.

7. Future Improvements

One of the main problems that caused the high RMSE and bad performance of our LSTM models is that we do not have sufficient data for training. If we were to have more data with unique dates, we would have produced a better LSTM model. Moreover, we tried to train our data by grouping the stores in terms of the temperature and fuel price to see whether the store location would play a big factor; however, we do not have enough data. Moreover, a good model for this problem would be to use multiple inputs with LSTM model, but we do not have enough data.