HOCHSCHULE
ESSLINGEN

Nah an Mensch und Technik.

Optimizing Sales Predictions: A Time Series Forecasting Approach with Statistical Models

In partial fulfillment of the requirements for the module

Application Project

WS 23

Name: Gee Chien Phua

Introduction

This report aims to explore the time series forecasting approach with statistical methods, such as SARIMAX. I will be utilizing a dataset from an Ecuadorian company, Corporación Favorita and it is a large grocery retailer.

There are 54 stores and 33 product families in the data. The time series starts from 2013-01-01 and finishes in 2017-08-31. I will be attempting to predict the next 16 days of grocery sales after the last date in the data.

There are 3 pieces of data that I will study on them step by step.

- Train
- Transactions
- Daily Oil Price

Train data contains time series of the stores and the product families combination. The 'sales' column gives the total sales for a product family at a particular store at a given date. The 'onpromotion' column gives the total number of items in a product family that were being promoted at a store at a given date.

Transaction data contains the number of transactions for each store, per day

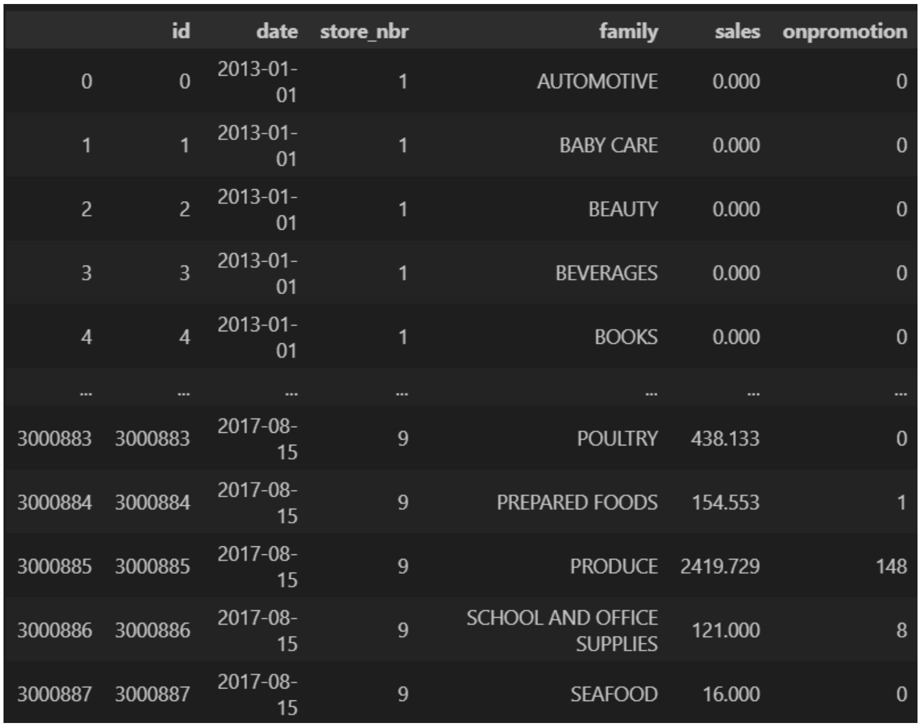
Daily Oil Price data: Ecuador is an oil-dependent country and its economical health is highly vulnerable to shocks in oil prices. This data may help us to understand which product families are affected in positive or negative ways by oil price.

Additional Notes:
Wages in the public sector are paid every two weeks on the 1st and 15th of the month. Supermarket sales could be affected by this.

Exploratory Data Analysis and Feature Engineering

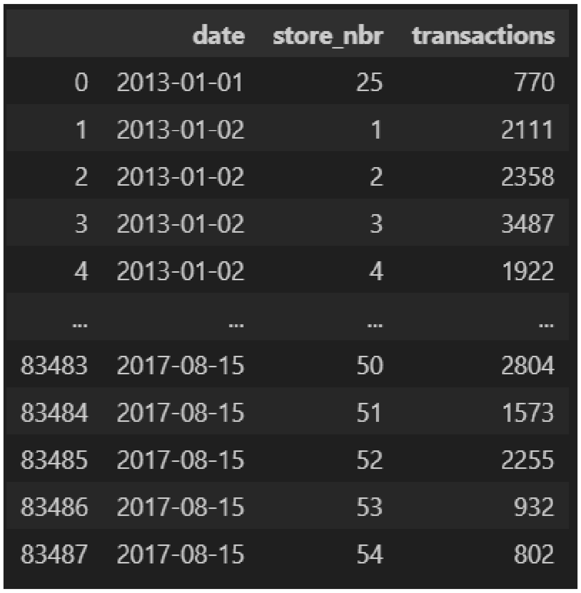
Visualizations of each datasets

Train:



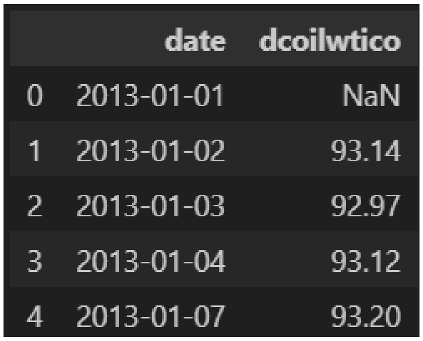
	id	date	store_nbr	family	sales	onpromotion
0	0	2013-01-01	1	AUTOMOTIVE	0.000	0
1	1	2013-01-01	1	BABY CARE	0.000	0
2	2	2013-01-01	1	BEAUTY	0.000	0
3	3	2013-01-01	1	BEVERAGES	0.000	0
4	4	2013-01-01	1	BOOKS	0.000	0
...	...	...	...	...	...	...
3000883	3000883	2017-08-15	9	POULTRY	438.133	0
3000884	3000884	2017-08-15	9	PREPARED FOODS	154.553	1
3000885	3000885	2017-08-15	9	PRODUCE	2419.729	148
3000886	3000886	2017-08-15	9	SCHOOL AND OFFICE SUPPLIES	121.000	8
3000887	3000887	2017-08-15	9	SEAFOOD	16.000	0

Transactions:



	date	store_nbr	transactions
0	2013-01-01	25	770
1	2013-01-02	1	2111
2	2013-01-02	2	2358
3	2013-01-02	3	3487
4	2013-01-02	4	1922
...	...	...	...
83483	2017-08-15	50	2804
83484	2017-08-15	51	1573
83485	2017-08-15	52	2255
83486	2017-08-15	53	932
83487	2017-08-15	54	802

Oil:

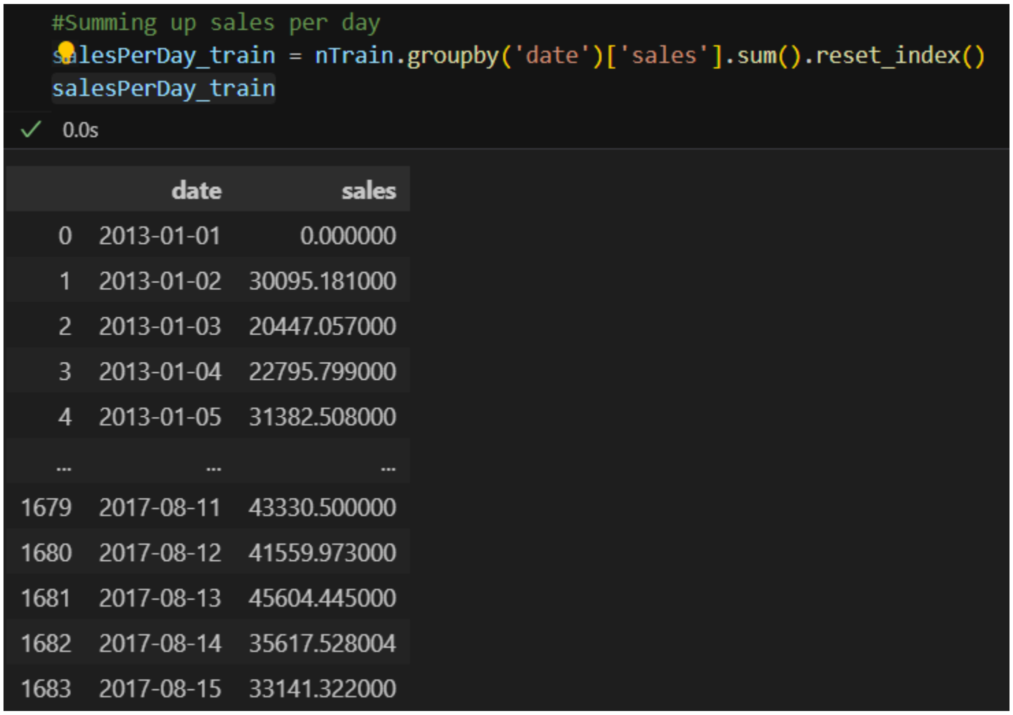


	date	dcoilwtico
0	2013-01-01	NaN
1	2013-01-02	93.14
2	2013-01-03	92.97
3	2013-01-04	93.12
4	2013-01-07	93.20

For this project, I will only focus on 1 store, store 44. This store has the highest number of sales as compared to other stores from 2013 to 2017. As such, I will filter out data that are not relevant to store 44.

Train Data Analysis

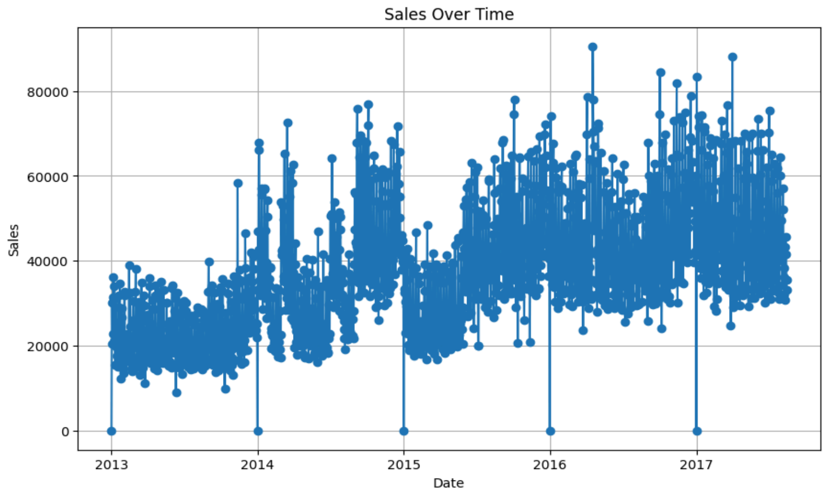
To get a better understanding of the 'sales' column, the sales of each product will be summed up, to get total sales per day, and plot it on a graph.



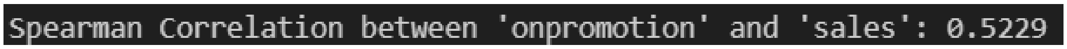
```
#Summing up sales per day
salesPerDay_train = nTrain.groupby('date')['sales'].sum().reset_index()
salesPerDay_train
```

✓ 0.0s

	date	sales
0	2013-01-01	0.000000
1	2013-01-02	30095.181000
2	2013-01-03	20447.057000
3	2013-01-04	22795.799000
4	2013-01-05	31382.508000
...	...	...
1679	2017-08-11	43330.500000
1680	2017-08-12	41559.973000
1681	2017-08-13	45604.445000
1682	2017-08-14	35617.528004
1683	2017-08-15	33141.322000



As seen above, there seems to be a gradual upward trend of total sales. All the first days of the year have 0 sales, and this may be related to the store being closed due to end of year events. There also seems to be a seasonality trend, where sales go up nearing the end of the year.

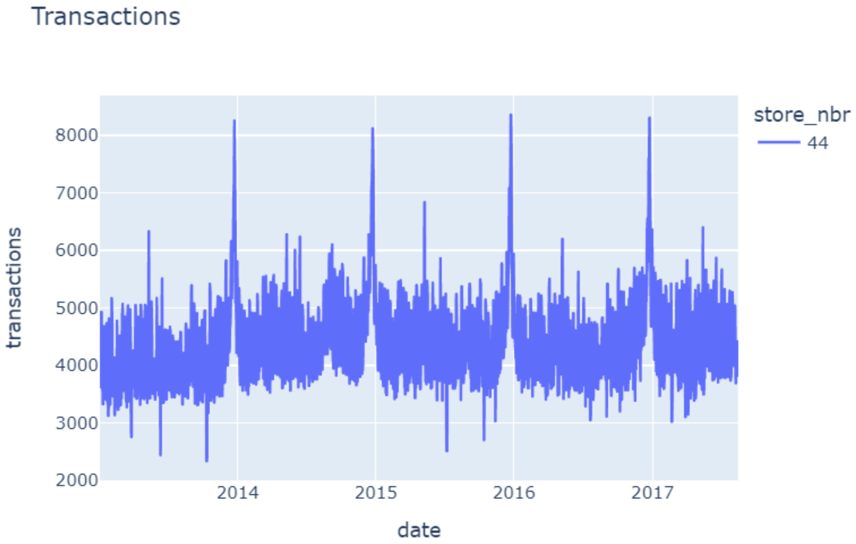


```
Spearman Correlation between 'onpromotion' and 'sales': 0.5229
```

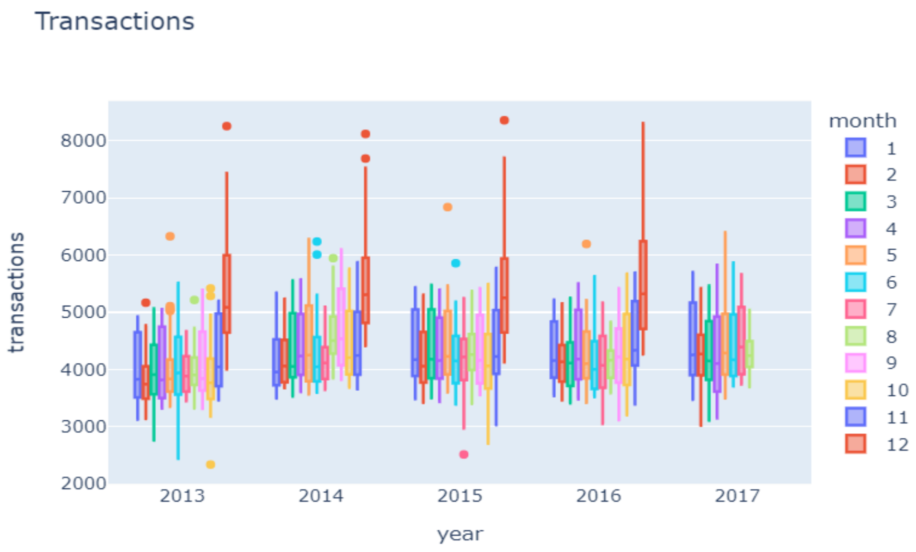
The spearman correlation between ‘onpromotion’ and ‘sales’ is also fairly high. This can also be used as an exogenous variable.

Transactions Data Analysis

I will also plot the transactions data via a line plot

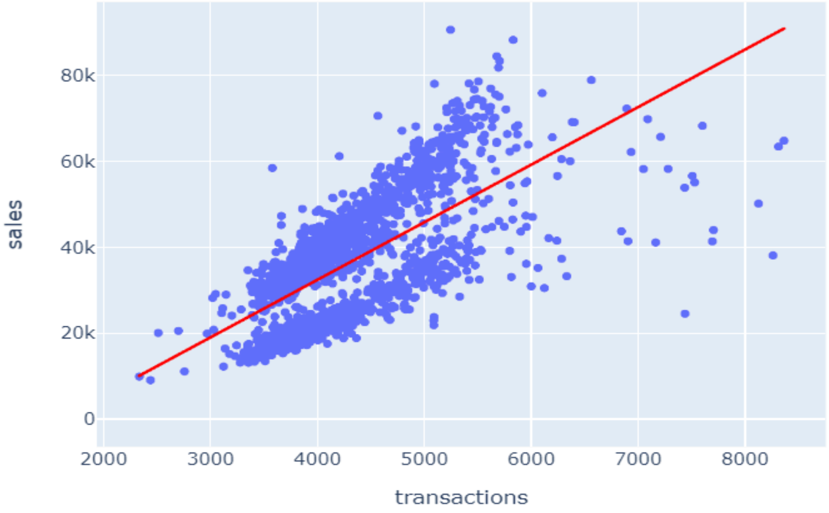


From the plotted graph above, some seasonality trends/ patterns recurring every year can be seen. A spike in transactions at the end of every year can be observed. Another graph can be plotted to better visualize the number of transactions per month, using box-plot



The box plots give us a better view of the spike in transactions. This occurs in December. The increase in transactions may be due to Christmas and New Years holidays. This seasonality trend (every 12 months) can be integrated into the model later on

When we look at the relationship between total sales and transactions, we can see that there is a high correlation via a scatter plot, and also a high value of 0.6883 using the Spearman Correlation

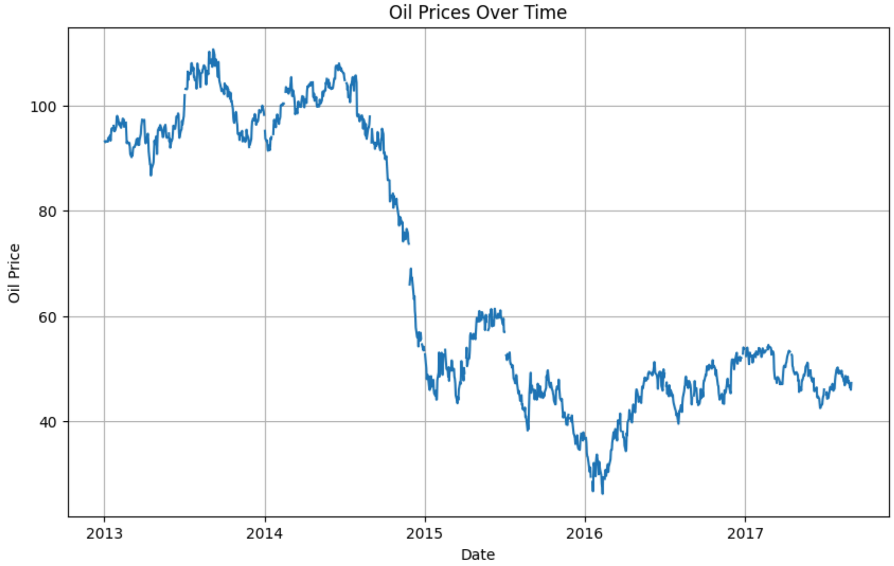


```
Spearman Correlation between Total Sales and Transactions: 0.6883
```

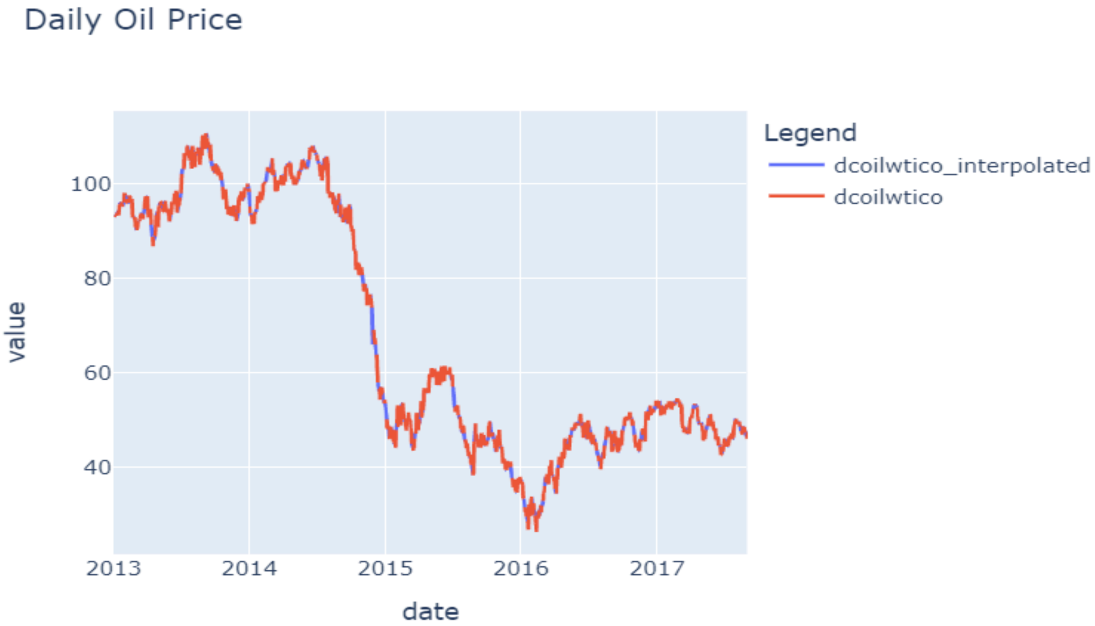
With a relatively high correlation to total sales, transactions can be used as an exogenous variable

Oil Data Analysis

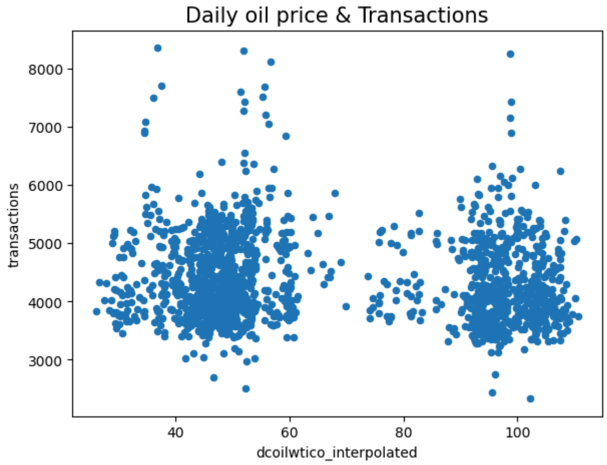
I will first create a line plot to visualize the oil data.

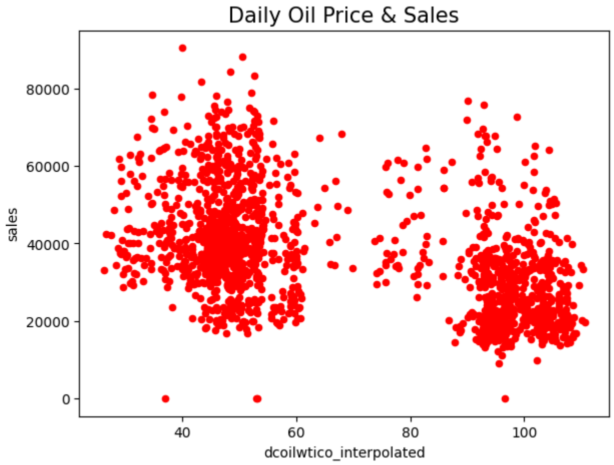


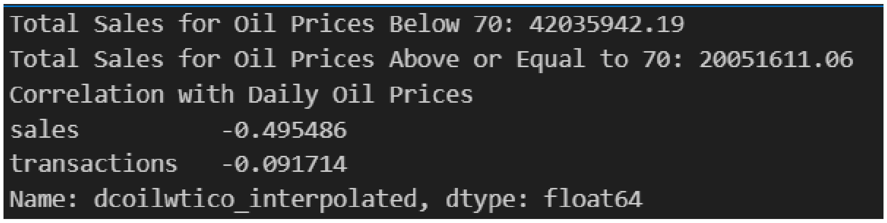
Since some data points are missing from the original oil dataset, linear interpolation can be utilized to fill in missing data (it assumes a linear relationship between the two neighboring data points)



Logically, if the daily oil price is high, it is expected that Ecuador's economy is bad and it means the price of product increases and sales decreases. There is a negative relationship here. This is further corroborated by 2 visible clusters within the scatterplot as shown below, with $70 as a 'border' between the 2 clusters. When oil prices are below $70, total sales amount to ~42 million. When oil prices are above $70, total sales amount to ~20 million. The Spearman correlation also demonstrates a fairly acceptable value of -0.495. Oil prices can also be a consideration as an exogenous variable





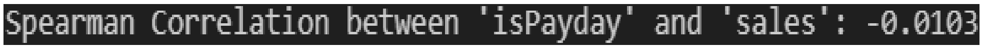


```
Total Sales for Oil Prices Below 70: 42035942.19
Total Sales for Oil Prices Above or Equal to 70: 20051611.06
Correlation with Daily Oil Prices
sales          -0.495486
transactions   -0.091714
Name: dcoilwtico_interpolated, dtype: float64
```

<u>Analysis for Additional Notes</u>

I also seek to determine the impact of paydays, occurring on the 1st and 15th of each month, on sales on those specific days. I can do that by creating a new column in the ‘train’ dataset, where it is ‘1’ on the 1st and 15th of each month, and ‘0’ for the rest of the days

However, there appears to be a weak correlation between 'isPayday' and 'sales,' as indicated by the Spearman correlation. This could be that the public sector alone does not greatly impact the number of sales on the 1st and 15th.



```
Spearman Correlation between 'isPayday' and 'sales': -0.0103
```

Given the weak correlation, incorporating 'isPayday' as an exogenous variable might not provide significant benefits to the model. Nevertheless, it's worth exploring 'isPayday' in one of the model iterations, considering that Spearman correlation might not be the most accurate measure of correlation.

SARIMAX

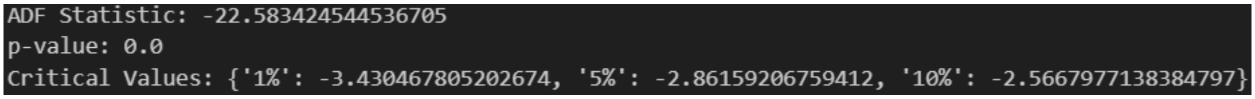
For this project, I am focusing on Autoregressive Integrated Moving Average (ARIMA) models and their extension, Seasonal AutoRegressive Integrated Moving Average with eXogenous factors (SARIMAX).

ARIMA is a widely-used model for time series forecasting that combines autoregression (AR), differencing (I), and moving average (MA) components. The AR term captures the linear dependence of the current value on its past values, the differencing term addresses non-stationarity, and the MA term deals with the dependency between a residual error and its lagged values.

Since our dataset has seasonality trends, I will use SARIMAX

Stationarity Check

The time series has to be examined for stationarity, as ARIMA models are only effective with stationary data. Stationarity means mean and variance of the time series stays relatively constant over time. The stationarity of a dataset can be assessed using the Augmented Dickey-Fuller Test. The null hypothesis (H0) assumes non-stationarity, while the alternative hypothesis (H1) suggests stationarity. If the test statistic is less than the critical values, the null hypothesis (H0) is rejected.



```
ADF Statistic: -22.583424544536705
p-value: 0.0
Critical Values: {'1%': -3.430467805202674, '5%': -2.86159206759412, '10%': -2.5667977138384797}
```

The results indicate that the p-value is very close to 0. Since it is less than the significance level (0.05), the null hypothesis of non-stationarity can be rejected. The ADF Statistic is also highly negative, which further indicates strong evidence against the null hypothesis

Grid Search

To determine the order (p,d,q) for the SARIMAX model, a grid search approach to iterate over multiple combinations of p,d,q parameters can be used. One example is the grid search loop using the ‘statsmodels’ library.

This grid search loop iterates over combinations of non-seasonal and seasonal parameters and selects the combination with the lowest Akaike Information Criterion (AIC), which is a measure of the model's goodness of fit.

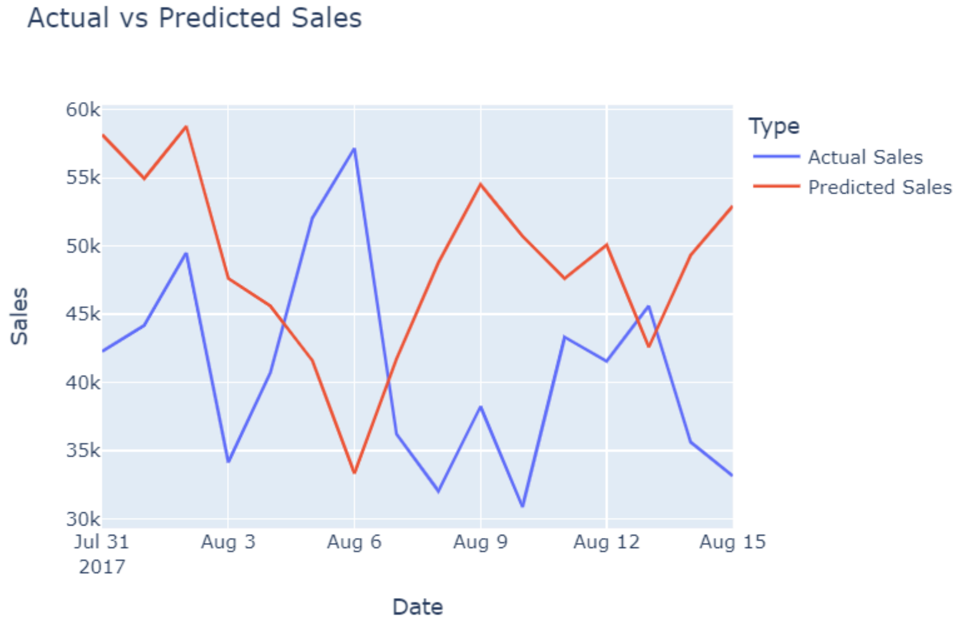
Model Fitting

I will be trying 3 different iterations of the model

1. Without any exogenous variables
2. With all exogenous variables
3. With all exogenous variables except for IsPayday

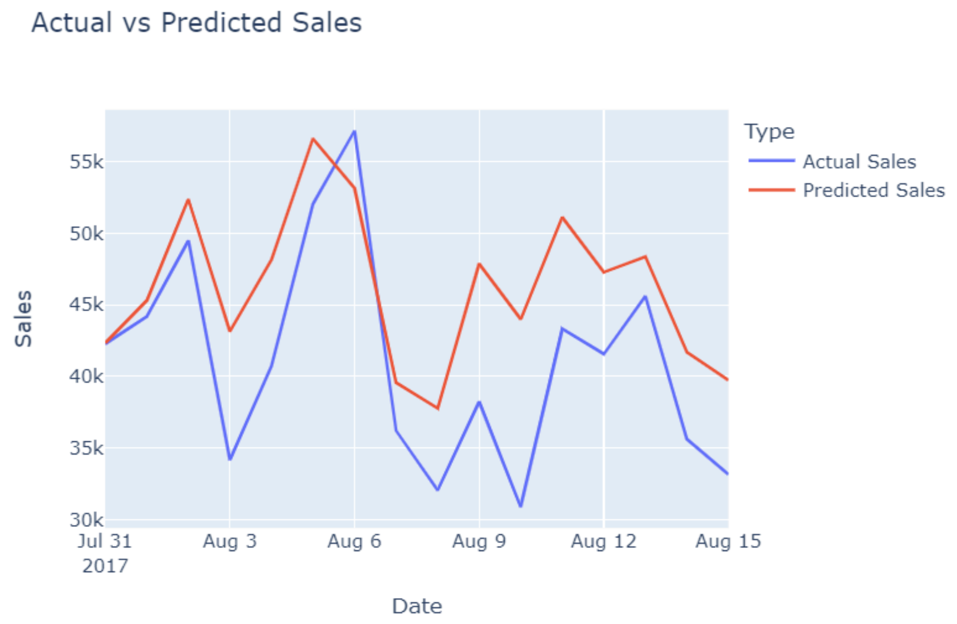
I will be forecasting the 16 days, so the data will be divided into the last 16 days and the preceding period. I will then compare the predictions of each model, and the mean squared error as compared to the test results

Model 1 - Without any exogenous variables:



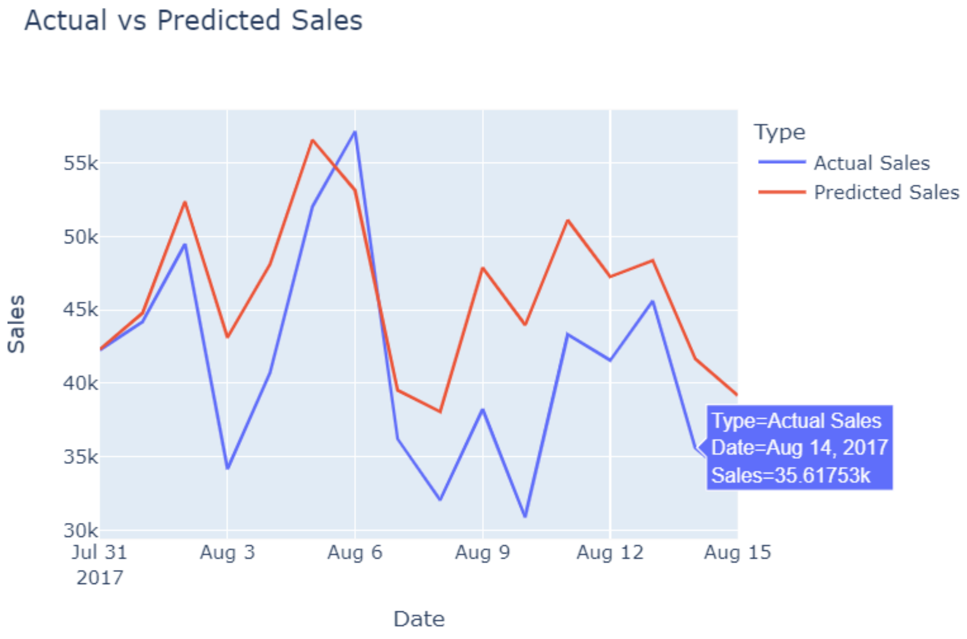
- Mean squared error: 186629630.28314942 ~ (187 million)

Model 2 - With all exogenous variables ('dcoilwtico_interpolated', 'isPayday', 'onpromotion', 'transactions')



- Mean Squared Error: 42058388.32561219 ~ (42.1 million)

Model 3 - With all exogenous variables except ‘isPayday’ ('dcoilwtico_interpolated', 'onpromotion', 'transactions')



- Mean Squared Error: 41581753.52826338 ~ (41.6 million)

As observed in Model 2, where all exogenous variables are incorporated, the model yields a Mean Squared Error (MSE) of 42.1 million, significantly improved compared to Model 1, which lacks any exogenous variables and has an MSE of 187 million. However, upon excluding exogenous variables with low correlation to 'sales,' such as 'isPayday,' from the model, the predictions become slightly more accurate. Model 3 demonstrates this improvement, showing a reduced MSE of 41.6 million.

Conclusion

SARIMAX is a powerful tool for time series forecasting, showcasing its effectiveness in relatively simple and stationary datasets and when coupled with relevant exogenous variables. However, in situations where data complexity surpasses the model's capabilities, the exploration of advanced machine learning models, such as Neural Networks, becomes imperative for achieving optimal precision and accuracy in predictions.