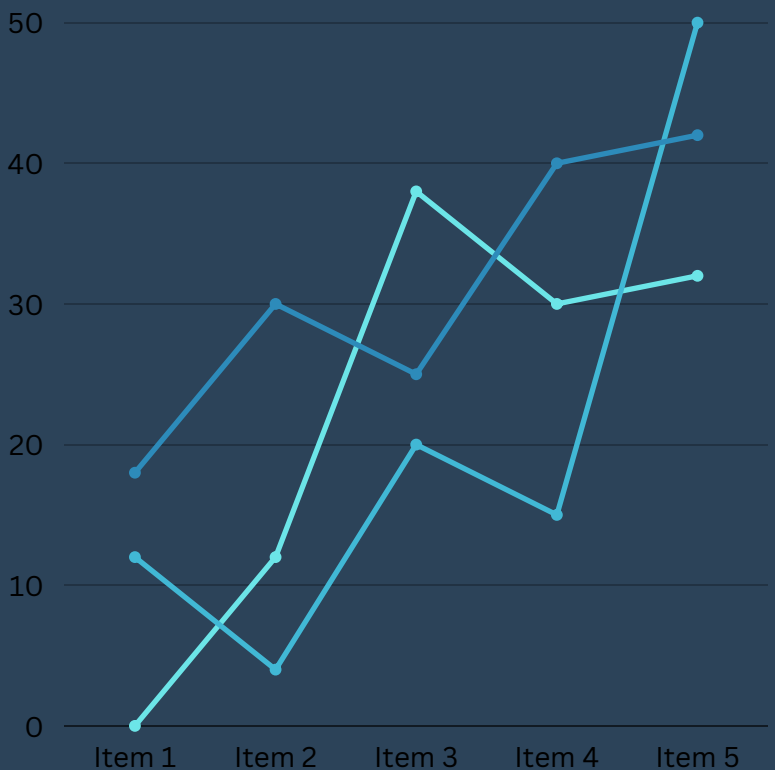
Time series forecasting project report

Walmart Sales analysis

Inna Krmoyan
Irena Torosyan
Gor Mkrtchyan



Walmart
Save money. Live better.

1. Abstract

Many different American multinational retail corporations that operate a chain of hypermarkets, discount department stores, and grocery stores that sell household goods which are geographically located at various locations. It is not always possible for the retailers to understand the condition of the market at different geographical locations. The retail store corporations need to understand the conditions of their markets to intensify the sales of their products for a large number of customers to get attracted in that particular direction. Time Series Forecasting helps the retailers to understand and visualize the big picture of the sales depending on various time periods. By forecasting the sales, we can get a general idea of the upcoming years and predict the future possible outcome of the bought and sold. If the retailer sees that they need some change in their business plans and strategies, then those are done in the retail store's objective so that it brings more profit to the company. Forecasting also helps the organizations understand how to make their customer happy so that they always go to their stores to get their desired products before a certain time that they want (e.g. The customer might want a Christmas tree early in November and if a store has it then they will be satisfied and go to that store most of the times). When customers are happy, they prefer the store that provides all the resources they need to be satisfied. As a result, sales in the store where the customers buy more items increase, resulting in more profit. Forecasting also helps the retailer to understand when a certain product is demanded so that they increase their sales at that given period of time. In this research report, we will make an attempt to understand the driving factors that lead to huge sales, try to find trends, and seasonality, and understand how time series forecasting can benefit the chosen company that is Walmart.

Keywords: Time series forecasting, sales predictions, trend, seasonality

2. Introduction

As being curious young enthusiasts, we are very interested in the future and its all possible developments. We strive to know whether it is possible to predict what can be awaiting for a certain company in the future, with the help of time series forecasting and its tools it surely is possible to predict patterns to a certain extent.

The aim of our project is to understand the reasons, relationships between the variables, the main driving factors that effect the sales so that we can predict the possible future pattern of the sales of ‘‘Walmart’’. This huge chain of retail stores sell mostly household products and gains profit by that. There are many other businesses like “Walmart” that sell goods for people’s houses but many of them fail to succeed. There are many reasons why that happens such as the company fails to make a good evaluation of the location of their stores as they do not have a good understanding of their customer needs. The rate of sales or shopping may increase on special occasions, which periodically results in less efficient items. To address this, the relationship between customers and retailers is examined, and any necessary adjustments are made in order to increase the profit.

All in all we want to understand our data make any necessary changes to make good visualizations that will show us the fluctuations of the sales (which is our main variable) and the relationship between it and the remaining variables.

3. Analysis Question

There are many organizations that find it difficult to forecast sales because of some reasons such as constant introduction of new products, seasonal/weather changes, etc. Retailers have moved to large-scale demand forecasting that can handle a lot of transaction data in an effort to solve these problems. Retailers can mine this data and predict future consumer behavior by gathering them. Retailers have the chance to optimize their revenue system by using the capacity to forecast at such a big scale, which enables them to make better decisions on

promotions and price. For our project we will face the challenge of making valid sales predictions that will actually benefit the company itself and other stakeholders who want to make reasonable changes to their business plans.

4. Data Description

4.1. About the dataset

The file has information about the Weekly Sales of 45 stores for the years 2010-2012 including the factors affecting Sales such as Holidays, Temperature, Fuel Price, CPI, and Unemployment.

4.2. Columns, number of observations, the covered period of time

Column names:
- Store - Store Numbers ranging from 1 to 45
- Date - The Week of Sales. It is in the format of dd-mm-yyyy. The date starts from 05-02-2010
- Weekly_Sales - The sales of the given store in the given week
- Holiday_Flag - If the week has a special Holiday or not. 1-The week has a Holiday 0-Fully working week Holiday events are given in the description.
- Temperature - Average Temperature of the week of sales
- Fuel_Price - Price of the Fuel in the region of the given store
- CPI - Customer Price Index
- Unemployment - Unemployment of the given store region.

Number of observations: 6435
The covered period of time / frequency: 2010 - 2012 / annual

4.3. Getting to know the dataset

For our project, we will be using Python (notebook: Jupyter Notebook)

Firstly we load the dataset and get some information about the variables.

	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
...	...	...	...	...	...	...	...	...
6430	45	28-09-2012	713173.95	0	64.88	3.997	192.013558	8.684
6431	45	05-10-2012	733455.07	0	64.89	3.985	192.170412	8.667
6432	45	12-10-2012	734464.36	0	54.47	4.000	192.327265	8.667
6433	45	19-10-2012	718125.53	0	56.47	3.969	192.330854	8.667
6434	45	26-10-2012	760281.43	0	58.85	3.882	192.308899	8.667

6435 rows × 8 columns

Table 1: Walmart dataset

```
<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
```

Table 2: Information about the columns

4.4. Finding mean, median, max, min etc.

	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
count	6435.000000	6.435000e+03	6435.000000	6435.000000	6435.000000	6435.000000	6435.000000
mean	23.000000	1.046965e+06	0.069930	60.663782	3.358607	171.578394	7.999151
std	12.988182	5.643666e+05	0.255049	18.444933	0.459020	39.356712	1.875885
min	1.000000	2.099862e+05	0.000000	-2.060000	2.472000	126.064000	3.879000
25%	12.000000	5.533501e+05	0.000000	47.460000	2.933000	131.735000	6.891000
50%	23.000000	9.607460e+05	0.000000	62.670000	3.445000	182.616521	7.874000
75%	34.000000	1.420159e+06	0.000000	74.940000	3.735000	212.743293	8.622000
max	45.000000	3.818686e+06	1.000000	100.140000	4.468000	227.232807	14.313000

Table 3: Getting descriptive variables of each column

As we can see in the table we get count, mean, standard deviation, minimum and maximum values, and quantiles for each column.

5. Exploratory Analysis

5.1. Data Preparation

In order to have an easier dataset to work with we decided to divide the column 'Date' to its every component (Day, Month, Year, also Year-Month). As it would be much easier to access certain data from a time period that we desire to make some predictions about it based on the given month, day, year or month-year. Then we created a new data frame based on the monthly data that we extracted.

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	YM	Year	Month	Day
0	1	05-02-2010	1643690.90	0	42.31	2.572	211.096358	8.106	02-2010	2010	02	05
1	1	12-02-2010	1641957.44	1	38.51	2.548	211.242170	8.106	02-2010	2010	02	12
2	1	19-02-2010	1611968.17	0	39.93	2.514	211.289143	8.106	02-2010	2010	02	19
3	1	26-02-2010	1409727.59	0	46.63	2.561	211.319643	8.106	02-2010	2010	02	26
4	1	05-03-2010	1554806.68	0	46.50	2.625	211.350143	8.106	03-2010	2010	03	05

Table 4: Modifying the column "Date"

	Store	Weekly_Sales	YM	Year	Month	Day	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
0	1	1316899.31	01-2011	2011	01	28	0	43.83	3.010	212.197058	7.742
1	1	1319325.59	01-2012	2012	01	27	0	54.26	3.290	220.078852	7.348
2	1	1327405.42	01-2011	2011	01	21	0	44.04	3.016	211.827234	7.742
3	1	1345454.00	10-2010	2010	10	22	0	69.86	2.725	211.861294	7.838
4	1	1351791.03	09-2010	2010	09	24	0	80.94	2.624	211.597225	7.787

Table 5: Creating "Walmart_monthly" data frame

5.2 Top 10 stores

In order to explore the data, one of the best ways to do so is to understand which are the stores that make the most profit.

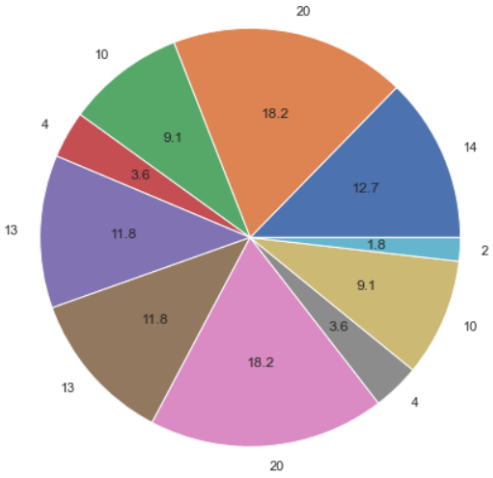


Figure 1: Top 10 stores

As we can identify from the pie chart above, the store that makes the most profit is the 20th one.

5.3. Weekly Sales

Now we will focus on our main variable of the dataset that is “Weekly_Sales”.

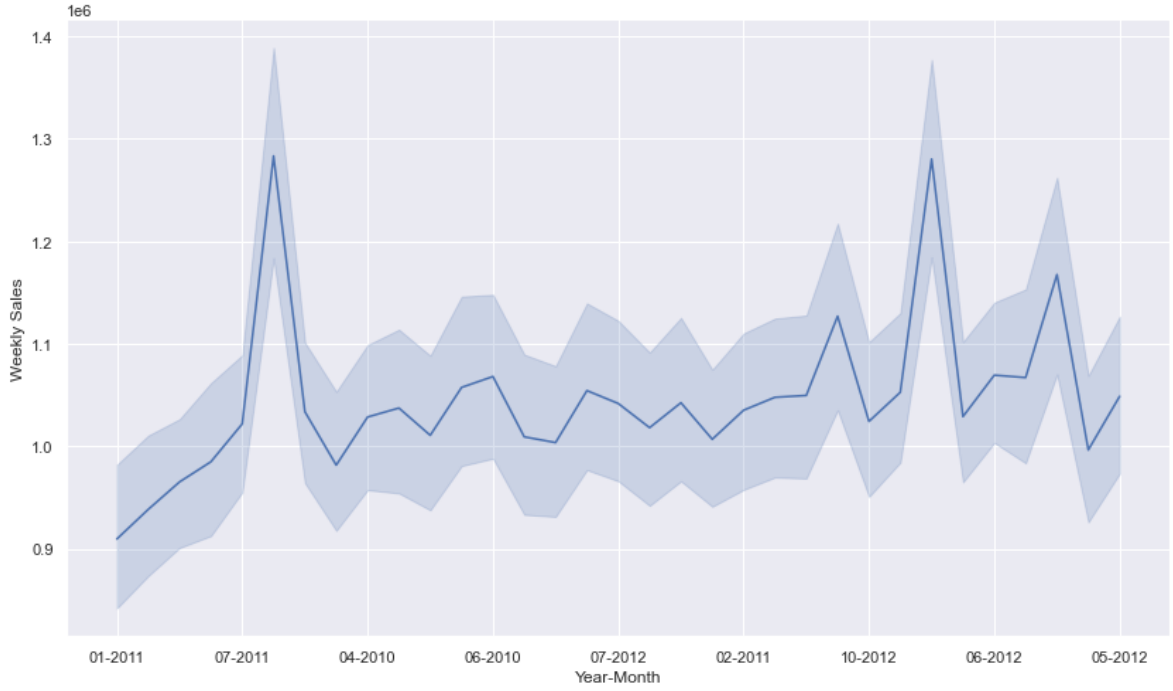


Figure 2: Plot of Weekly_Sales

We can definitely see trend and some kind of seasonality in our plot. In order to capture those clearer we will extract them from the plot.

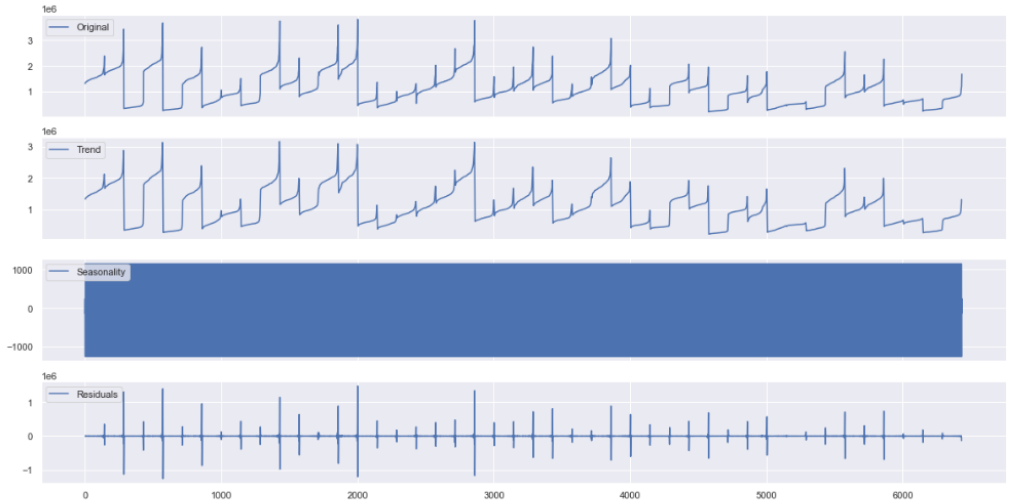


Figure 3: Plot of the original data, trend, seasonality and residuals

Weekly Sales in each store can be identified easier with the following plots.

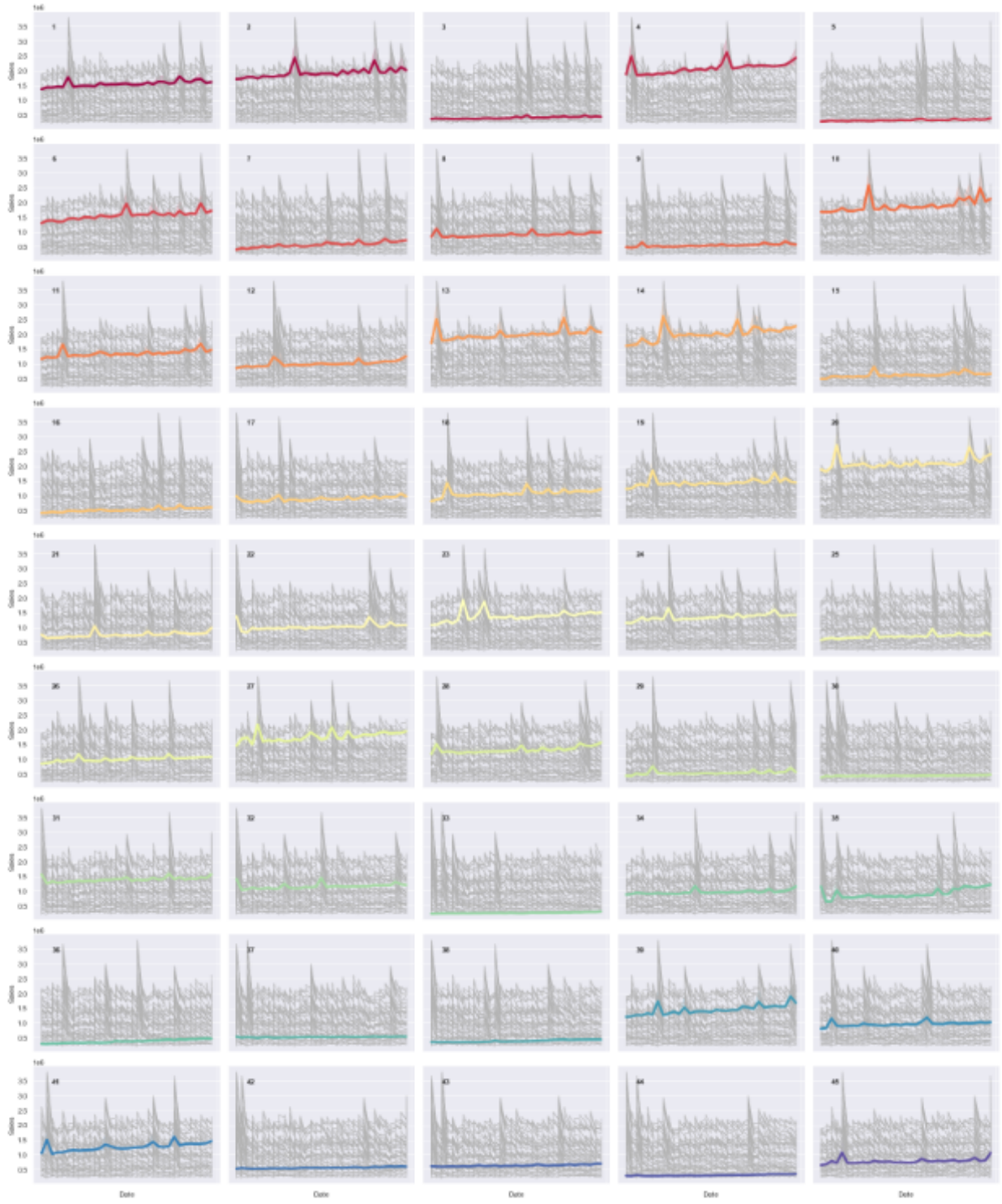


Figure 4: Each store's weekly sales

For understanding whether the holidays have an impact on the sales we modified our dataset to display the November 25th plot for us to identify the trends in the plot. We chose November 25th as it is Thanksgiving.

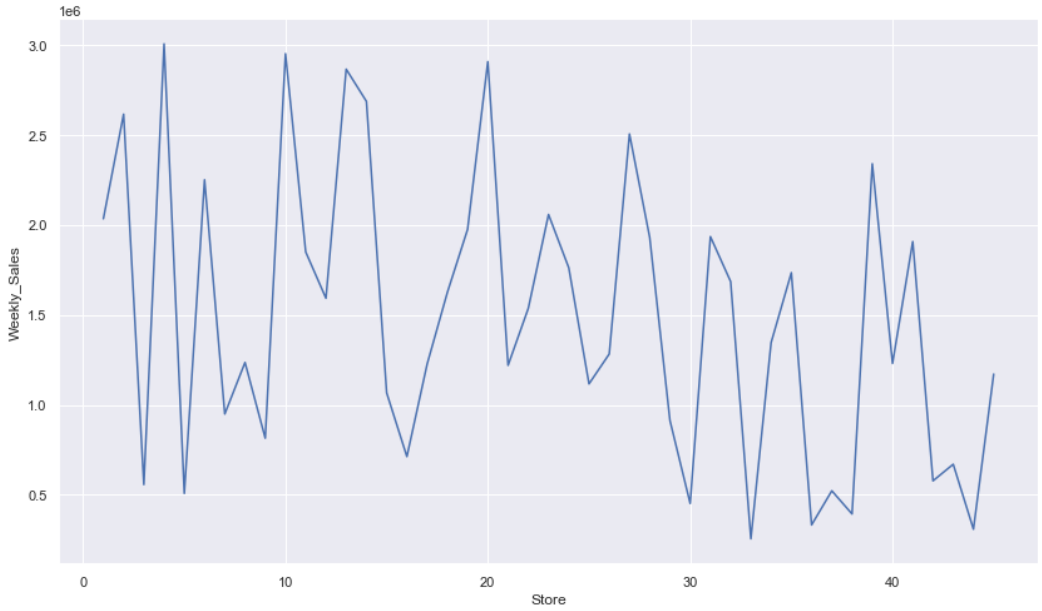


Figure 5: Thanksgiving plot

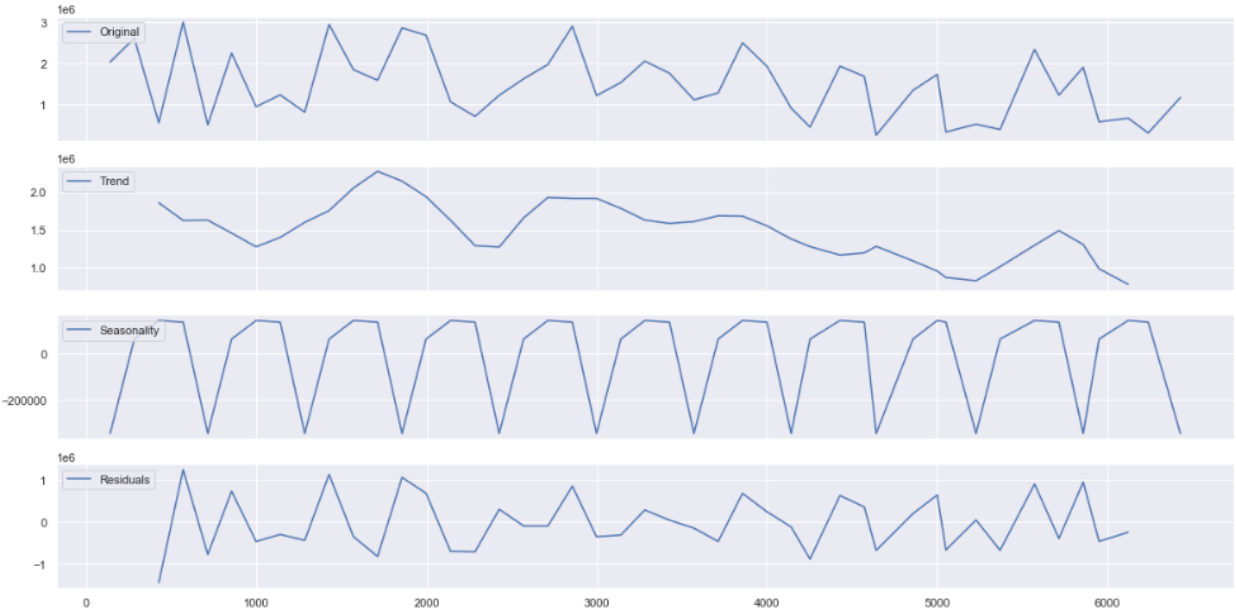


Figure 6: plot for Thanksgiving, trend, seasonality, residuals

We can identify that there is seasonality and trend as people tend to go to supermarkets a lot during those days.

5.4. Correlations

To understand the relationship between each variable it is necessary to take a look at their correlations.

	Store	Weekly_Sales	Holiday_Flag	Temperature
Store	1.000000e+00	-0.335332	-5.188199e-18	-0.022659
Weekly_Sales	-3.353320e-01	1.000000	3.689097e-02	-0.063810
Holiday_Flag	-5.188199e-18	0.036891	1.000000e+00	-0.155091
Temperature	-2.265908e-02	-0.063810	-1.550913e-01	1.000000
Fuel_Price	6.002295e-02	0.009464	-7.834652e-02	0.144982
CPI	-2.094919e-01	-0.072634	-2.162091e-03	0.176888
Unemployment	2.235313e-01	-0.106176	1.096028e-02	0.101158

	Fuel_Price	CPI	Unemployment
Store	0.060023	-0.209492	0.223531
Weekly_Sales	0.009464	-0.072634	-0.106176
Holiday_Flag	-0.078347	-0.002162	0.010960
Temperature	0.144982	0.176888	0.101158
Fuel_Price	1.000000	-0.170642	-0.034684
CPI	-0.170642	1.000000	-0.302020
Unemployment	-0.034684	-0.302020	1.000000

Table 6: Correlation between the variables

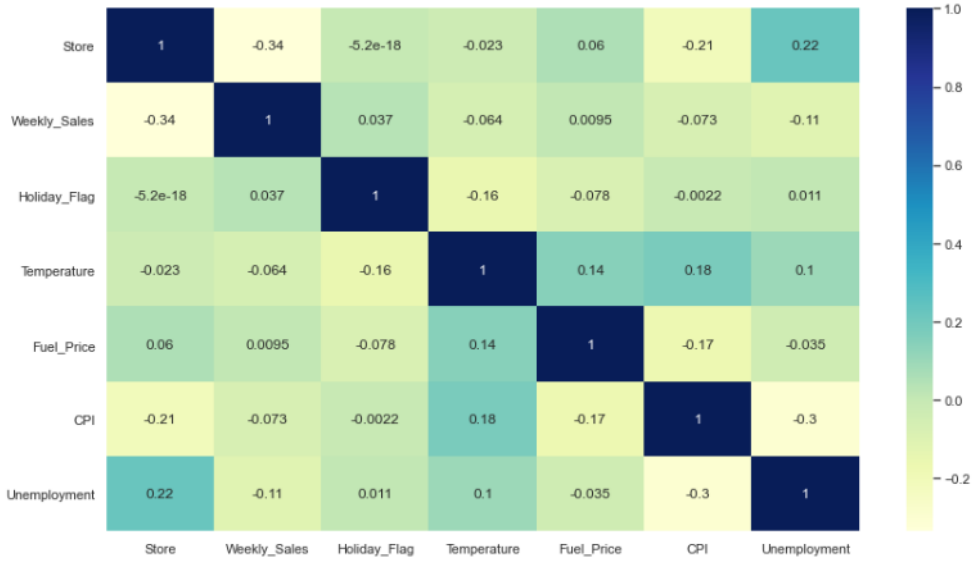


Figure 7: The correlation heatmap

6. Literature Review

6.1. Wine dataset

The first literature that we found is the wine sales analysis for time series forecasting. The main goal for that project was to perform forecasting analysis on the Rose and Sparkling dataset. They tried to analyse the wine dataset by using Linear Regression, Naïve Model, Simpleand Moving Average models, Simple, Double and Triple Exponential Smoothing.

We chose this dataset's report as a literature review as it was very nicely done and everything written was very clear and informative.

6.2. Amazon dataset

This paper attempts to forecast future sales at Amazon.com, Inc. based on historical sales data. Firstly, it proposes three possible forecasting approaches according to the historical data pattern, that is Holt-Winters exponential smoothing, neural network autoregressive model and ARIMA(Autoregressive integrated moving average). Secondly, it specifies certain accuracy measures which well determine the suitability of the forecast methods on the available sales data. Finally the three methods will be implemented to forecast Amazon's quarterly sales in 2019. The results can help Amazon well manage its future operations.

We chose this dataset's report as it is quite related to our theme and analysis question, therefore it was very helpful to look through it.

6.3. Supermarket dataset

This research work has proposed a FB Prophet tool for the sales prediction of the supermarket data. The proposed research work has examined few forecasting models such as- The additive model, the Autoregressive integrated moving average (ARIMA) model, FB Prophet model. From the proposed research work, it is concluded that FB Prophet is a better prediction model in terms of low error, better prediction, and better fitting.

Again as the Amazon's dataset report this one coincides with our theme and analysis question, so it was very useful to see how others interpret the topic we chose.

7. Estimation I

7.1 ADF & KPSS

```
ADF Statistic: -4.624149
p-value: 0.000117
Critical Values:
        1%: -3.431
        5%: -2.862
        10%: -2.567
```

Figure 8: The ADF Statistic

In figure 8 we calculated the ADF statistic for our model.

```
KPSS Statistic: 1.8459
p-value: 0.010000
Critical Values:
        10%: 0.347
        5%: 0.463
        2.5%: 0.574
        1%: 0.739
```

Figure 9: The KPSS Statistic

In figure 9 we calculated the KPSS statistic for our model.

ADF was done in order to test the null hypothesis of whether a unit root is present in the time series sample. The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity.

The KPSS test figures out if a time series is stationary around a mean or linear trend, or is non-stationary due to a unit root. A stationary time series is one where statistical properties — like the mean and variance — are constant over time.

P-value is greater than, let's say α=0.05, therefore we cannot reject the null hypothesis and conclude that time series possess unit root. Therefore the series is not stationary.

7.2 Dependence order of the model

First of all the ACF and PACF of the regular model were checked.

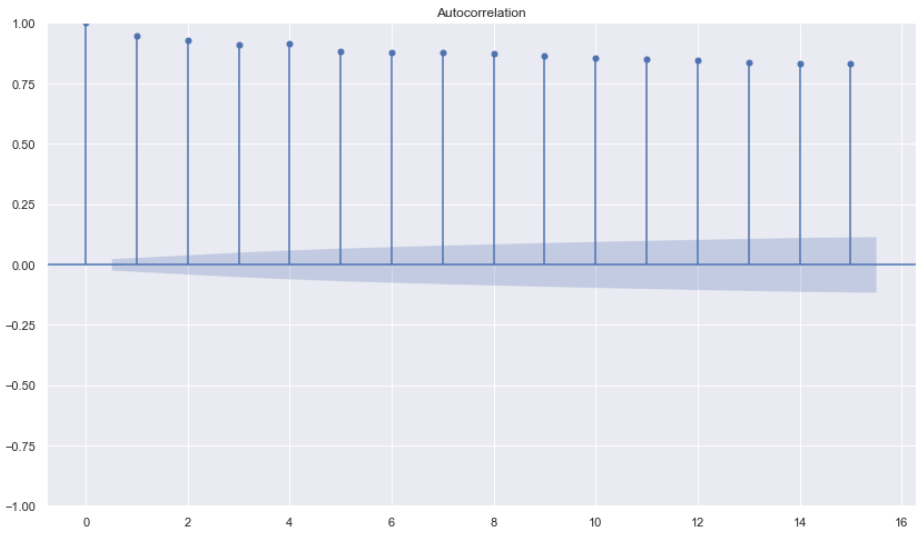


Figure 10: The ACF of the original model

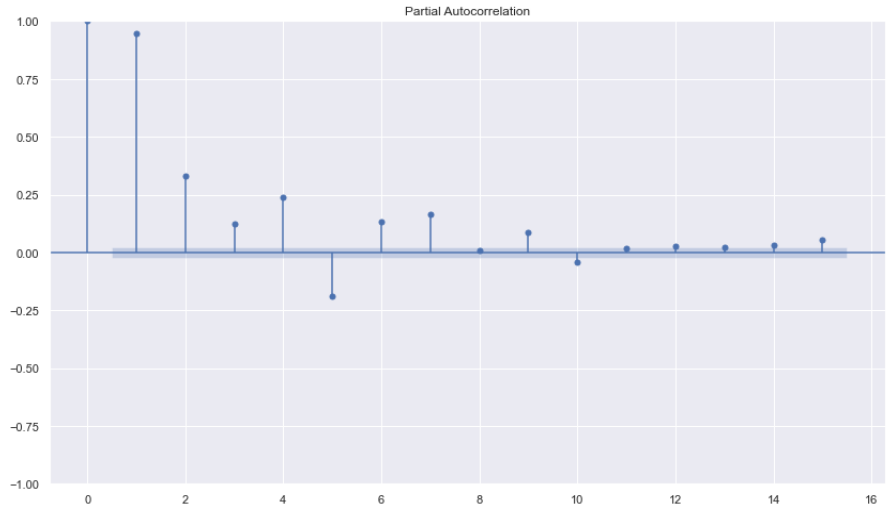


Figure 11: The PACF of the original model

ACF is the correlation between a time series with a lagged version of itself. Whereas the PACF gives the partial correlation of a stationary time series with its own lagged values.

In both ACF and Pacf we have a lot of significant lags

We will split our data into a train-test and take a look at ACF and PACF again. The main purpose of the train-test split procedure is to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

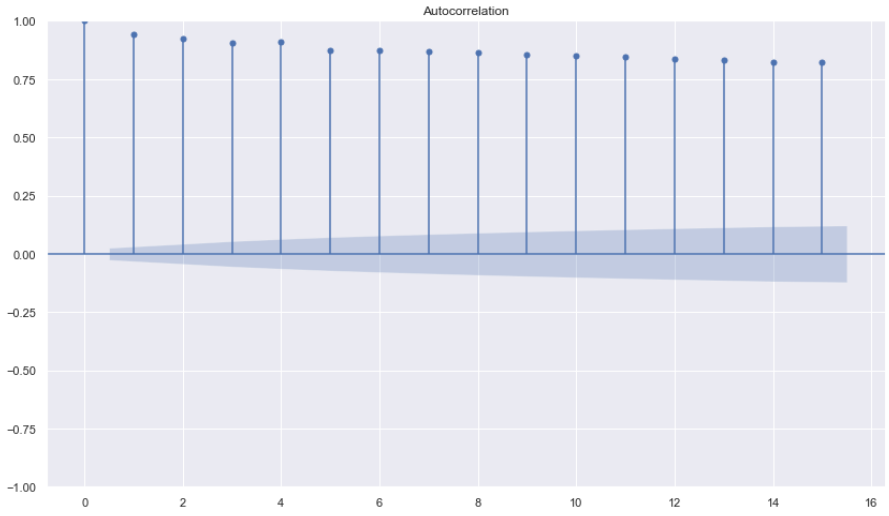


Figure 12: The ACF of the train model

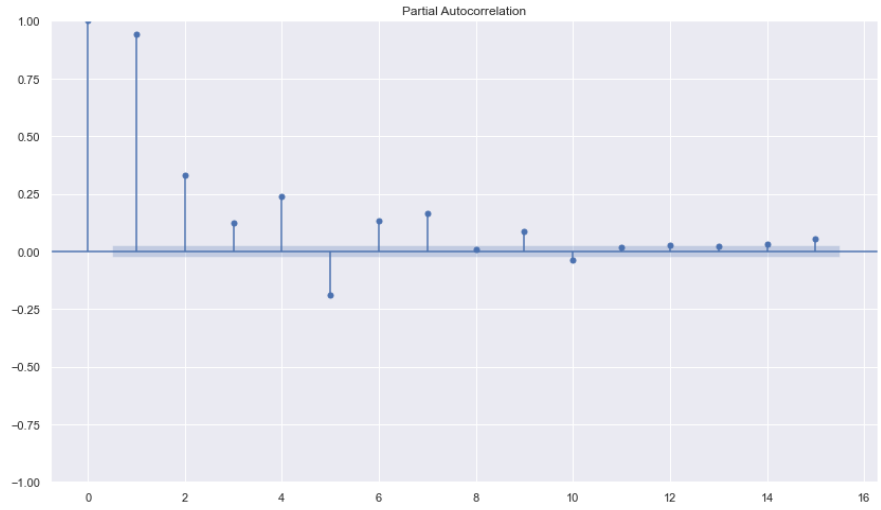


Figure 13: The PACF of the train model

7.3 Residual Diagnosis

SARIMAX Results

Dep. Variable:	Weekly_Sales	No. Observations:	5791
Model:	SARIMAX(1, 0, 1)x(1, 1, [1, 2], 7)	Log Likelihood	-78544.734
Date:	Sat, 03 Dec 2022	AIC	157101.468
Time:	22:22:01	BIC	157141.445
Sample:	0	HQIC	157115.376
	- 5791		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9838	0.004	254.320	0.000	0.976	0.991
ma.L1	-0.4549	0.010	-44.977	0.000	-0.475	-0.435
ar.S.L7	-0.9563	0.044	-21.941	0.000	-1.042	-0.871
ma.S.L7	-0.0273	0.038	-0.715	0.475	-0.102	0.048
ma.S.L14	-0.9647	0.038	-25.572	0.000	-1.039	-0.891
sigma2	5.646e+10	2.44e-13	2.32e+23	0.000	5.65e+10	5.65e+10

Ljung-Box (L1) (Q):	5.16	Jarque-Bera (JB):	91502.99
Prob(Q):	0.02	Prob(JB):	0.00
Heteroskedasticity (H):	0.40	Skew:	0.02
Prob(H) (two-sided):	0.00	Kurtosis:	22.49

Figure 14: The SARIMAX Results for train model

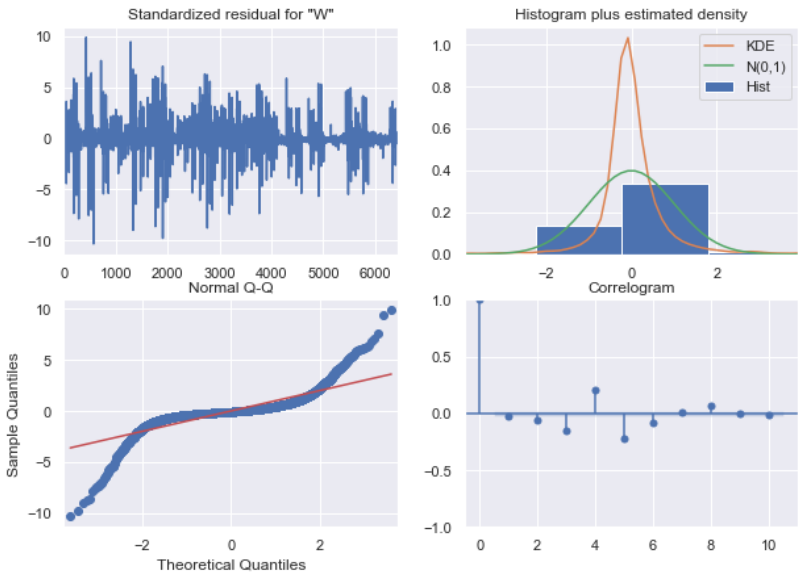


Figure 15: The model diagnosis

H_0: The residuals are independently distributed.

H_1: The residuals are not independently distributed; they exhibit serial correlation.

If the p-value is less than some threshold (e.g. $\alpha = .05$), we reject the null hypothesis and conclude that the residuals are not independently distributed

7.4 Ljung-Box test

	lb_pvalue
1	3.030982e-03
2	8.286938e-03
3	1.146842e-14
4	1.884345e-52
5	2.084839e-110
6	5.129220e-120
7	4.682792e-119
8	3.099964e-121
9	1.597722e-120
10	7.142265e-120

Figure 16: The Ljung-Box test p-value results

The Ljung-Box test is applied to the residuals of our time series after fitting an ARMA(p,q) model to the data. The purpose3 of the test is to examine the autocorrelations of the residuals. If the autocorrelations are very small, we conclude that the model does not exhibit significant lack of fit.

As we can identify our p-values are small therefore we reject null hypothesis so the residuals are not independently distributed.

7.5 Holt Winter's test

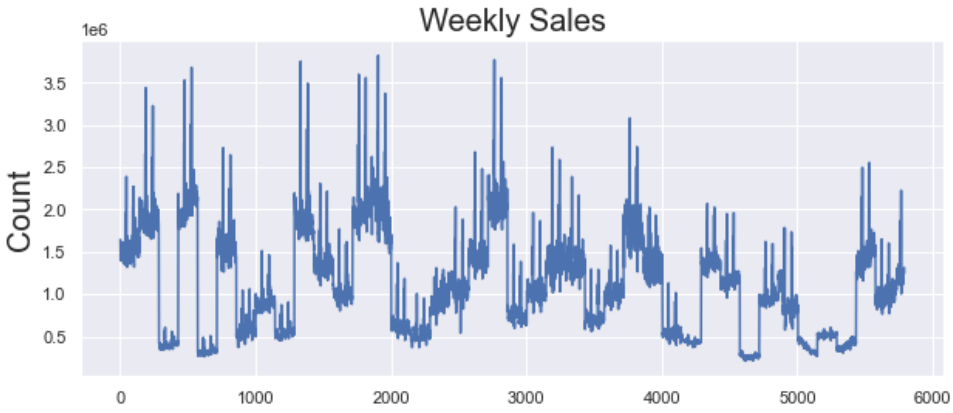


Figure 17: The Weekly Sales plot

We have to estimate the model on the train test and check whether the series is additive or multiplicative.

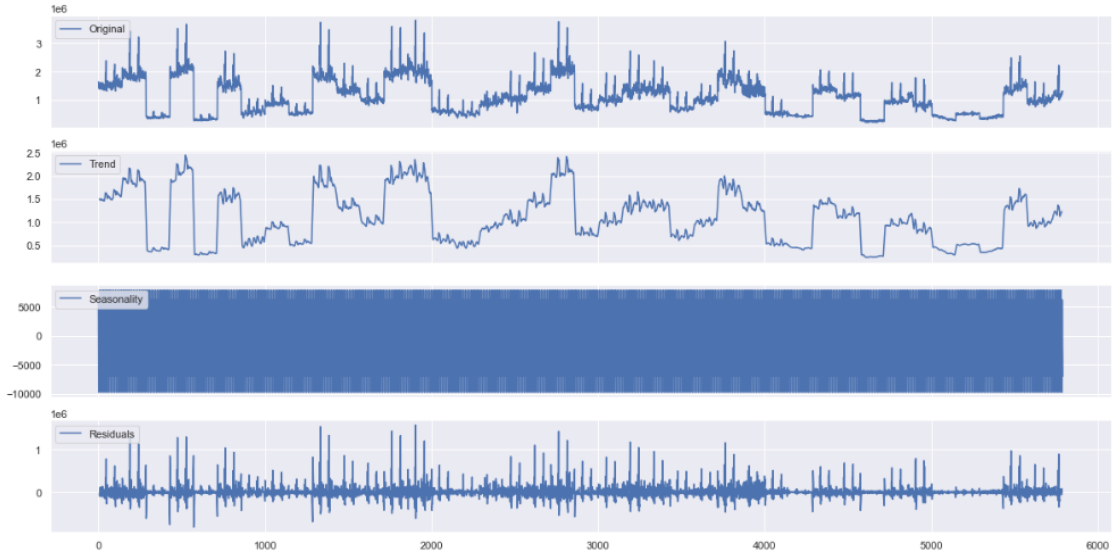


Figure 18: The Additive Decomposition

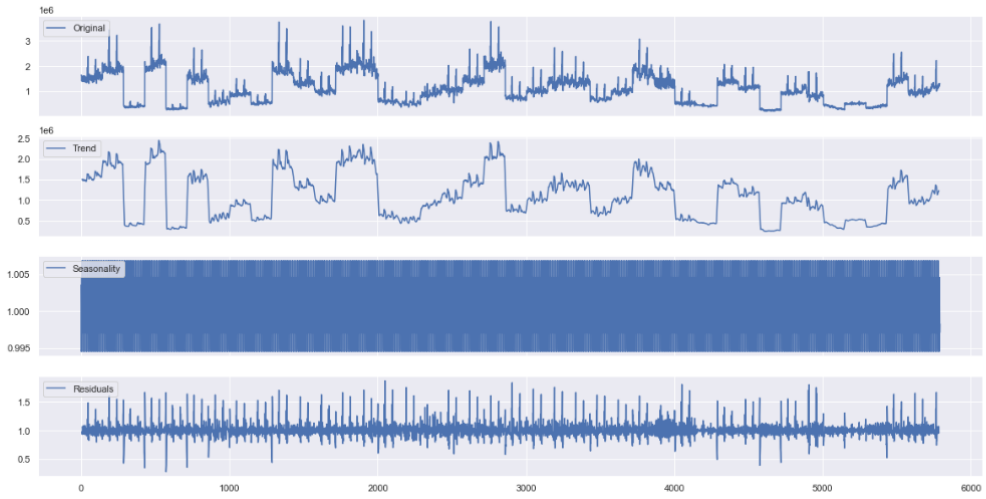


Figure 19: The Multiplicative Decomposition

As we can see our data has trends and seasonality which can be seen from the graphs. The time series are additive as when we decompose it we can see that the residuals are white noise (the mean is concentrated around 0) whereas when we plot for multiplicative we can see that it is not the case there and the mean is around 1.

7.6 Calculating MSE and Comparing with the best SARIMA

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	YM	Year	Month	Day	first_diff
0	1	05-02-2010	1643690.90	0	42.31	2.572	211.096358	8.106	02-2010	2010	02	05	NaN
1	1	12-02-2010	1641957.44	1	38.51	2.548	211.242170	8.106	02-2010	2010	02	12	-1733.46
2	1	19-02-2010	1611968.17	0	39.93	2.514	211.289143	8.106	02-2010	2010	02	19	-29989.27
3	1	26-02-2010	1409727.59	0	46.63	2.561	211.319643	8.106	02-2010	2010	02	26	-202240.58
4	1	05-03-2010	1554806.68	0	46.50	2.625	211.350143	8.106	03-2010	2010	03	05	145079.09
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5786	41	13-05-2011	1270025.74	0	50.29	3.767	192.826069	6.934	05-2011	2011	05	13	25068.83
5787	41	20-05-2011	1244542.33	0	41.11	3.828	192.831317	6.934	05-2011	2011	05	20	-25483.41
5788	41	27-05-2011	1278304.33	0	50.56	3.795	192.836565	6.934	05-2011	2011	05	27	33762.00
5789	41	03-06-2011	1297584.95	0	54.81	3.763	192.841813	6.934	06-2011	2011	06	03	19280.62
5790	41	10-06-2011	1311690.11	0	61.10	3.735	192.847061	6.934	06-2011	2011	06	10	14105.16

Figure 20: The Train-test data

SARIMAX Results

Dep. Variable:	Weekly_Sales	No. Observations:	6435
Model:	ARIMA(1, 1, 0)	Log Likelihood	-86810.112
Date:	Sat, 03 Dec 2022	AIC	173624.224
Time:	22:42:40	BIC	173637.763
Sample:	0	HQIC	173628.909
	- 6435		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3464	0.004	-88.434	0.000	-0.354	-0.339
sigma2	3.064e+10	3.95e-14	7.76e+23	0.000	3.06e+10	3.06e+10

Ljung-Box (L1) (Q):	16.94	Jarque-Bera (JB):	158555.45
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.36	Skew:	-0.90
Prob(H) (two-sided):	0.00	Kurtosis:	27.25

Figure 21: ARIMA(1, 1, 0) Results

```
the error of the AR(1) is 102418443778.6022
the error of the Exponential Smoothing is 491071548866.8271
the error of the SARIMA is 463509463096.31055
```

```
6435    745676.927755
6436    750736.515756
6437    748983.670619
6438    749590.926798
6439    749380.548791
           ...
7074    749434.679240
7075    749434.679240
7076    749434.679240
7077    749434.679240
7078    749434.679240
Name: predicted_mean, Length: 644, dtype: float64
```

Figure 22: Model forecast

Based on MSE we can infer that exponential smoothing does the best forecast.

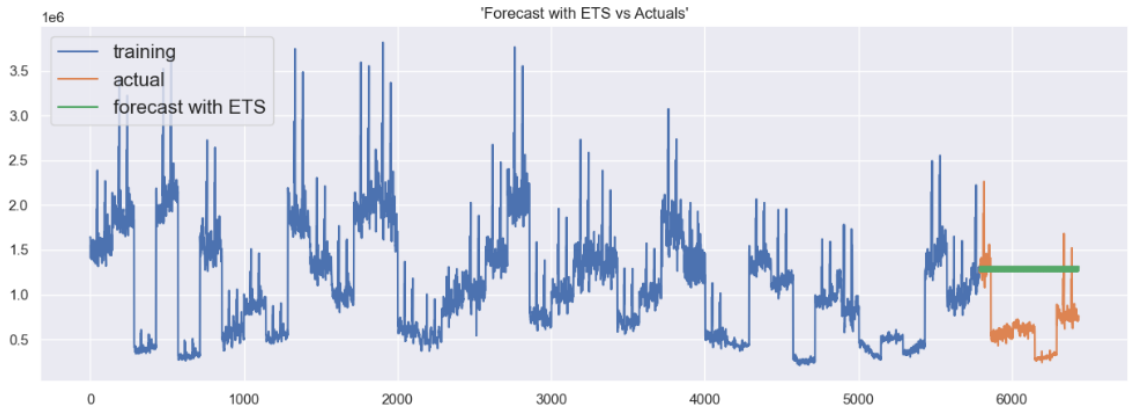


Figure 23: Forecast with ETS vs Actual

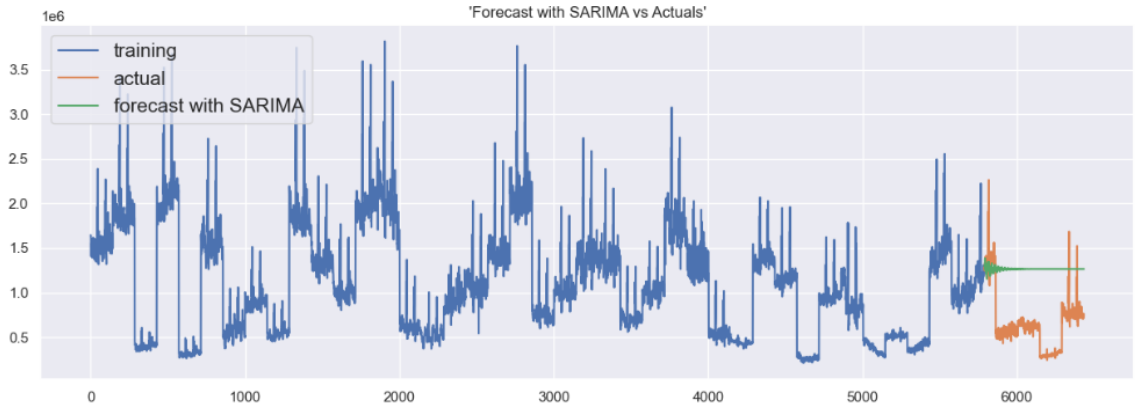


Figure 24: Forecast with SARIMA vs Actual

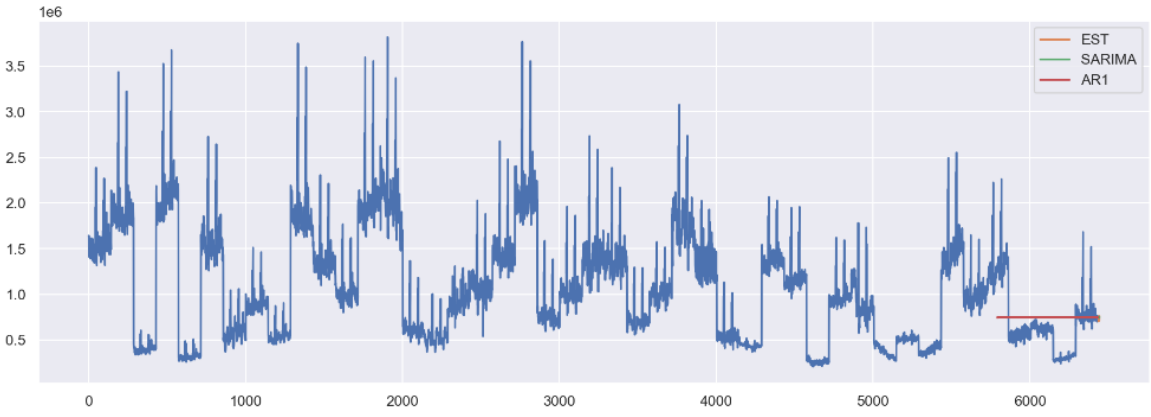


Figure 25: Forecast with AR1

We can identify fluctuating seasonality that happens mostly because of holidays.

8. Estimation II

Estimating a VAR model. A VAR model is a generalization of the univariate autoregressive model for forecasting a vector of time series.It consists of one equation for each variable in the system. The right side of each equation contains a constant as well as the lags of all the variables in the system.

	AIC	BIC	FPE	HQIC
0	32.31	32.32	1.082e+14	32.32
1	27.91	27.92	1.326e+12	27.92
2	27.77	27.78	1.145e+12	27.77
3	27.74	27.76	1.118e+12	27.75
4	27.68	27.70	1.050e+12	27.69
5	27.58	27.61	9.524e+11	27.59
6	27.55	27.57	9.184e+11	27.56
7	27.51	27.54	8.833e+11	27.52
8	27.47	27.51	8.552e+11	27.49
9	27.45	27.49	8.369e+11	27.47
10	27.45	27.50	8.365e+11	27.47
11	27.43	27.48	8.204e+11	27.45
12	27.43	27.48	8.176e+11	27.45
13	27.42	27.48*	8.104e+11	27.44
14	27.42	27.48	8.083e+11	27.44
15	27.41*	27.48	8.054e+11*	27.44*

Figure 26: VAR order selection

Selecting an appropriate VAR model by using information criteria. So as we can see the 15th lag has the lowest AIC. Therefore we take that model.

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Wed, 07, Dec, 2022
Time:                     20:14:01
--------------------------------------------------------------------
No. of Equations:         2.00000    BIC:                    27.9174
Nobs:                     6434.00    HQIC:                   27.9133
Log likelihood:          -108043.    FPE:                 1.32323e+12
AIC:                      27.9111    Det(Omega_mle):      1.32199e+12
--------------------------------------------------------------------
Results for equation Weekly_Sales
================================================================================
                    coefficient       std. error           t-stat            prob
--------------------------------------------------------------------------------
const              60179.604325      9220.772667            6.527           0.000
L1.Weekly_Sales        0.945013         0.004080          231.601           0.000
L1.Temperature       -45.254726       124.845770           -0.362           0.717
================================================================================

Results for equation Temperature
================================================================================
                    coefficient       std. error           t-stat            prob
--------------------------------------------------------------------------------
const                3.848718         0.312028           12.335           0.000
L1.Weekly_Sales     -0.000000         0.000000           -1.659           0.097
L1.Temperature       0.940554         0.004225          222.630           0.000
================================================================================

Correlation matrix of residuals
               Weekly_Sales  Temperature
Weekly_Sales       1.000000     0.009903
Temperature        0.009903     1.000000
```

Figure 27: Model Fit Summary Table

For the Weekly Sales equation the constant coefficient and L1.Weekly_Sales are significant for p=0.05.

For the Holiday_Flag equation the constant coefficient and L1.Temperature are significant for p=0.05.

	Weekly_Sales	Temperature
1	30382.767517	-4.756946
2	1859.672350	0.236761
3	-171976.386312	5.594303
4	164525.952435	-0.883742
5	-87846.693283	10.561770
...	...	...
6430	-27375.706654	-0.240008
6431	2252.709554	0.181565
6432	-15903.476457	-10.243194
6433	-33667.653193	1.557606
6434	24019.168253	2.052755

Figure 28: Residual Results for the Weekly Sales and Temperature

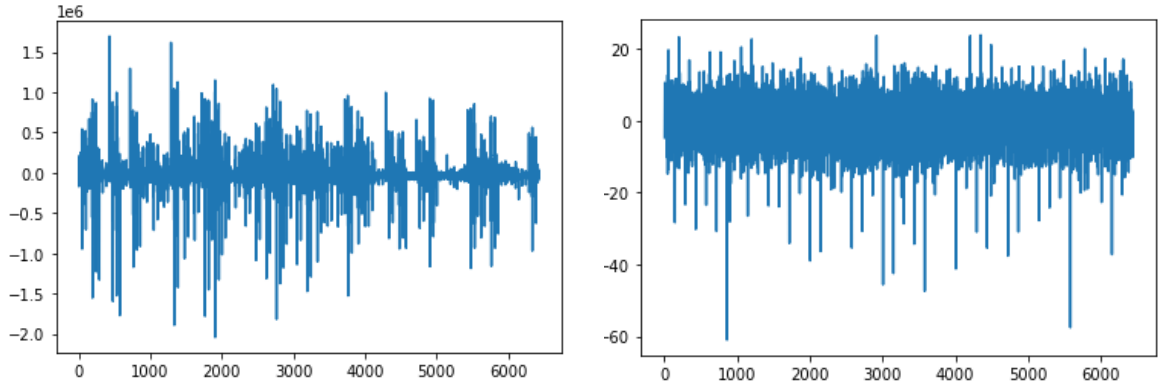


Figure 29: Plots for residual results (Right one for Temperature left one for Sales)

	lb_stat	lb_pvalue
1	634.530514	5.169894e-140
2	640.240906	9.407054e-140
3	759.589074	2.512628e-164
4	1430.773393	1.467752e-308
5	1767.890095	0.000000e+00
6	1768.244917	0.000000e+00
7	1780.907748	0.000000e+00
8	1817.643210	0.000000e+00
9	1817.656568	0.000000e+00
10	1817.662210	0.000000e+00

	lb_stat	lb_pvalue
1	173.393264	1.343030e-39
2	180.502645	6.373070e-40
3	243.970813	1.317618e-52
4	528.215307	5.283550e-113
5	531.025242	1.600577e-112
6	536.954709	9.157710e-113
7	618.161574	2.985653e-129
8	654.277346	4.960354e-136
9	678.921385	2.348016e-140
10	786.088864	2.018173e-162

Figure 30: p-values for residual results (Right one for Temperature left one for Sales)

We can see from the outputs that all the p-values are less than 0.05, so the residuals are correlated and we can conclude that they are not independent.

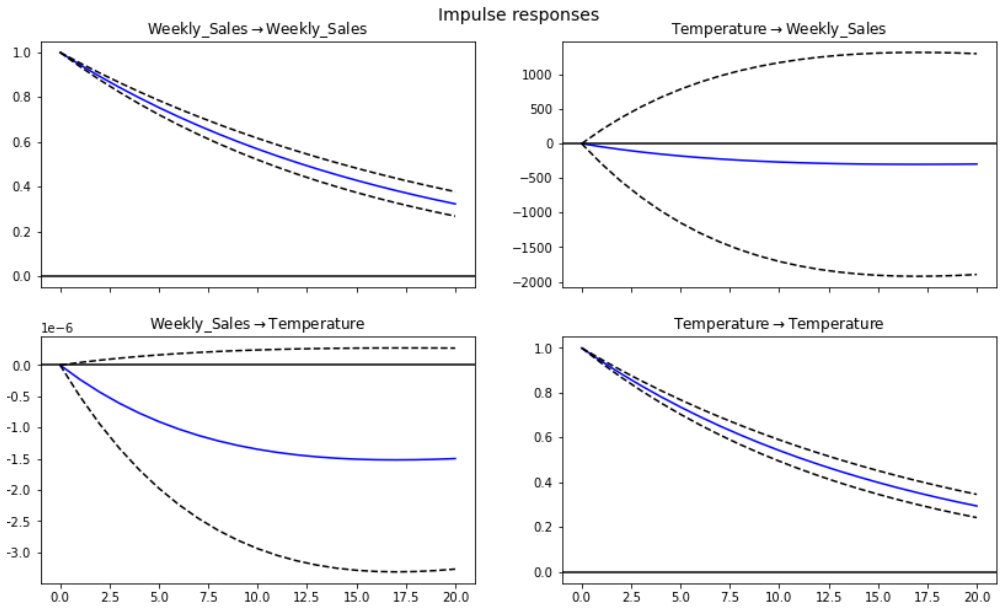


Figure 31: Impulse responses

```
array([[7.75992443e+05, 5.90261039e+01],
       [7.90831589e+05, 5.91881395e+01],
       [8.04847447e+05, 5.93371428e+01],
       [8.18085877e+05, 5.94740771e+01],
       [8.30590173e+05, 5.95998380e+01],
       [8.42401207e+05, 5.97152580e+01],
       [8.53557569e+05, 5.98211106e+01],
       [8.64095689e+05, 5.99181145e+01],
       [8.74049963e+05, 6.00069375e+01],
       [8.83452865e+05, 6.00881996e+01],
       [8.92335055e+05, 6.01624766e+01],
       [9.00725481e+05, 6.02303031e+01],
       [9.08651477e+05, 6.02921752e+01],
       [9.16138848e+05, 6.03485532e+01],
       [9.23211962e+05, 6.03998643e+01],
       [9.29893827e+05, 6.04465046e+01],
       [9.36206167e+05, 6.04888413e+01],
       [9.42169498e+05, 6.05272151e+01],
       [9.47803187e+05, 6.05619413e+01],
       [9.53125528e+05, 6.05933125e+01],
       [9.58153791e+05, 6.06215993e+01],
       [9.62904286e+05, 6.06470526e+01],
       [9.67392415e+05, 6.06699043e+01],
       [9.71632723e+05, 6.06903693e+01]])
```

Figure 32: 6-month predictions

9. Conclusion

By using various methods to construct different kinds of models to explore our data, we came to the conclusion that holidays are actually the cause of the trends and seasonality in our time series as during those times the sales have big shocks. That is because people shop more during the holidays such as Christmas, Thanksgiving and so on. By forecasting the series we understood that Walmart company uses a great technique of keeping their holiday sales on peak as we noticed seasonality that indicates some repetitions of data over the years. There are also some peaks that are trends that may be caused because of some in-store sales and many other factors. All in all to summarize we answered our analysis question by investigating our target variable which was Weekly_Sales and with the help of it and other useful columns came to the conclusion that the most responsible component of the fluctuations of the data are the holidays.

10. References

[1] Jha, B. K., & Pande, S. (2021, May 6). *Time Series Forecasting Model for Supermarket Sales using FB-Prophet*. https://ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/9418033

[2] Khan, K. (2022, January 13). *Project - Time Series Forecasting - Wine Sales Analysis*. https://www.scribd.com. https://www.scribd.com/document/552600283/Time-Series-Forecasting-Project-Report

[3] Singh, B., Kumar, P., Sharma, N., & Sharma, K. P. (2020, April 20). *Sales Forecast for Amazon Sales with Time Series Modeling*. https://ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/9071463