Time Series Project Report

Pratyusha Bala(221372)
Anirban Ghosh(221271)
Rajdeep Adhya(221385)
Abhraraj Halder(221255)
Souvik Roy(221433)

Instructor: Dr. Amit Mitra

November 14, 2023

Contents

1 Introduction 2

1.1 Overview . . . 2

2 About the Dataset 2

3 Exploratory Data Analysis : 2

3.1 Testing for trend existence . . . 4

3.1.1 Implementing Relative Ordering Test for trend existence . . . 4

3.1.2 Estimation of Trend . . . 4

3.2 Detrending the data : . . . 6

3.2.1 Seasonality Check using Friedman' test . . . 6

3.3 Testing for randomness of a time series data: . . . 7

4 Modelling 7

4.1 Train/Val split . . . 7

4.2 ACF and PACF . . . 8

4.3 Fitting . . . 8

5 Conclusion 9

1 Introduction

1.1 Overview

Forecasting sales, revenue, and stock prices is a classic application of machine learning in economics, and it is important because it allows investors to make guided decisions based on forecasts made by algorithms.In this project we have calculated the fitted the future sales at Walmart based on heirarchical sales in the states of California, Texas, and Wisconsin.

2 About the Dataset

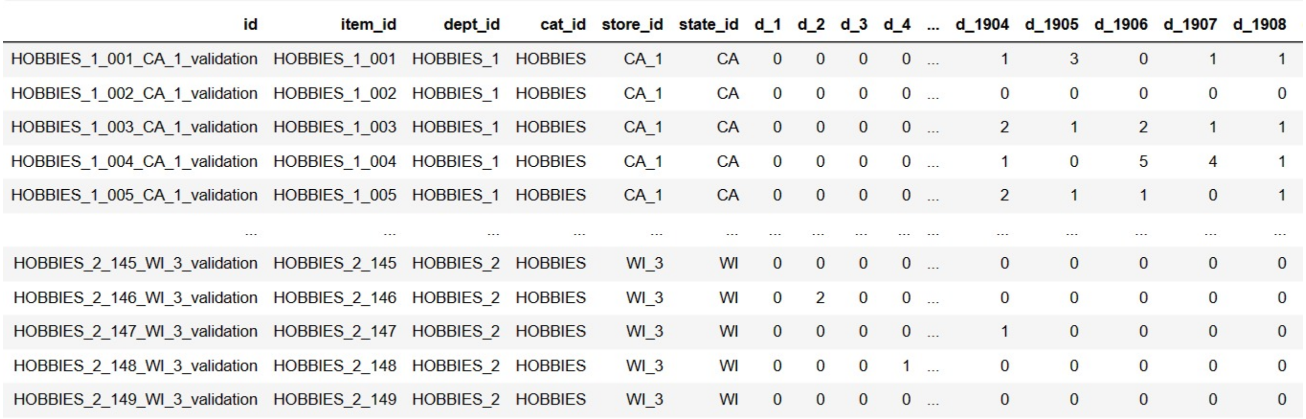
The dataset consists of three .csv files

- **calendar.csv :** This Dataset contains the dates on which the products are sold. The dates are in yyyy/dd/mm format.
- **sales_train_validation.csv :** Contains the historical daily unit sales data per product and store for the time period of [d_1 to d_1913]
- **sales_train_evaluation.csv :** It will include sales for [d_1 - d_1941]

In this Project we need to forecast sales d_1914 - d_1941. These rows are from evaluation set. We train our model on the data on d_1-d_1913 and test our model on the data on d_1914 - d_1941. Since we have the data on the sales of different products of different categories form different stores of three states of USA. Now we have to visualize the dataset.

3 Exploratory Data Analysis :

Now we will try to visualize the sales data and gain some insights from it. Now let's have a look on the sales data we have.
Training Data:



id	item_id	dept_id	cat_id	store_id	state_id	d_1	d_2	d_3	d_4	...	d_1904	d_1905	d_1906	d_1907	d_1908
HOBBIES_1_001_CA_1_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	1	3	0	1	1
HOBBIES_1_002_CA_1_validation	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	0	0	0	0	0
HOBBIES_1_003_CA_1_validation	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	2	1	2	1	1
HOBBIES_1_004_CA_1_validation	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	1	0	5	4	1
HOBBIES_1_005_CA_1_validation	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...	2	1	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
HOBBIES_2_145_WI_3_validation	HOBBIES_2_145	HOBBIES_2	HOBBIES	WI_3	WI	0	0	0	0	...	0	0	0	0	0
HOBBIES_2_146_WI_3_validation	HOBBIES_2_146	HOBBIES_2	HOBBIES	WI_3	WI	0	2	0	0	...	0	0	0	0	0
HOBBIES_2_147_WI_3_validation	HOBBIES_2_147	HOBBIES_2	HOBBIES	WI_3	WI	0	0	0	0	...	1	0	0	0	0
HOBBIES_2_148_WI_3_validation	HOBBIES_2_148	HOBBIES_2	HOBBIES	WI_3	WI	0	0	0	1	...	0	0	0	0	0
HOBBIES_2_149_WI_3_validation	HOBBIES_2_149	HOBBIES_2	HOBBIES	WI_3	WI	0	0	0	0	...	0	0	0	0	0

Figure 1: Sales_Train_Validation Data

Description of the Dataset : This data represents the daily sales of different products which belong to the different categories in different stores of three states in USA.

As this is a multivariate time series time series data so we have extracted the daily sales of a particular product of a particular store.

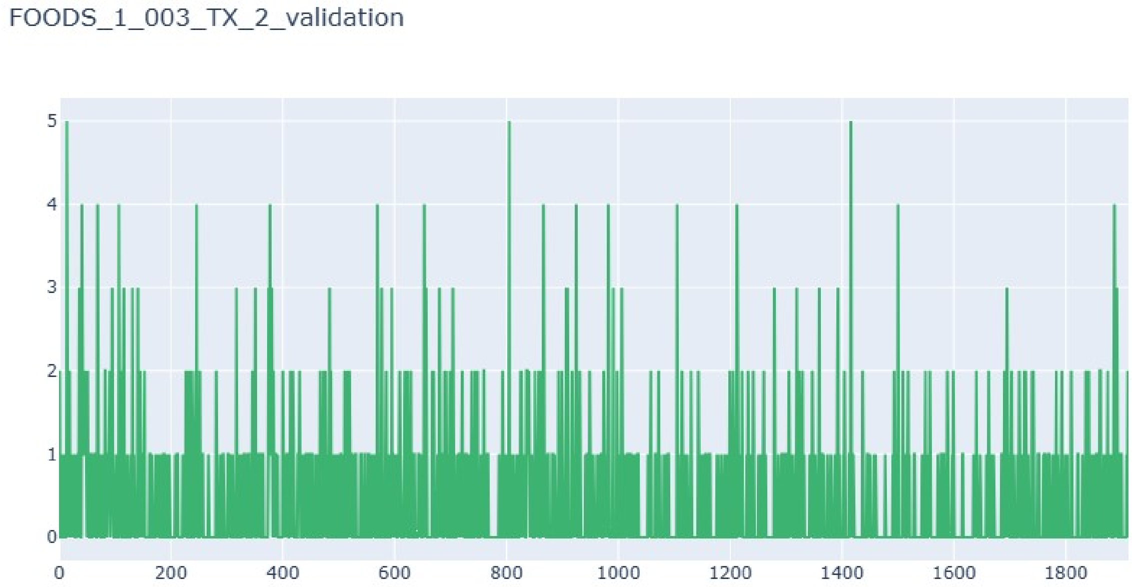
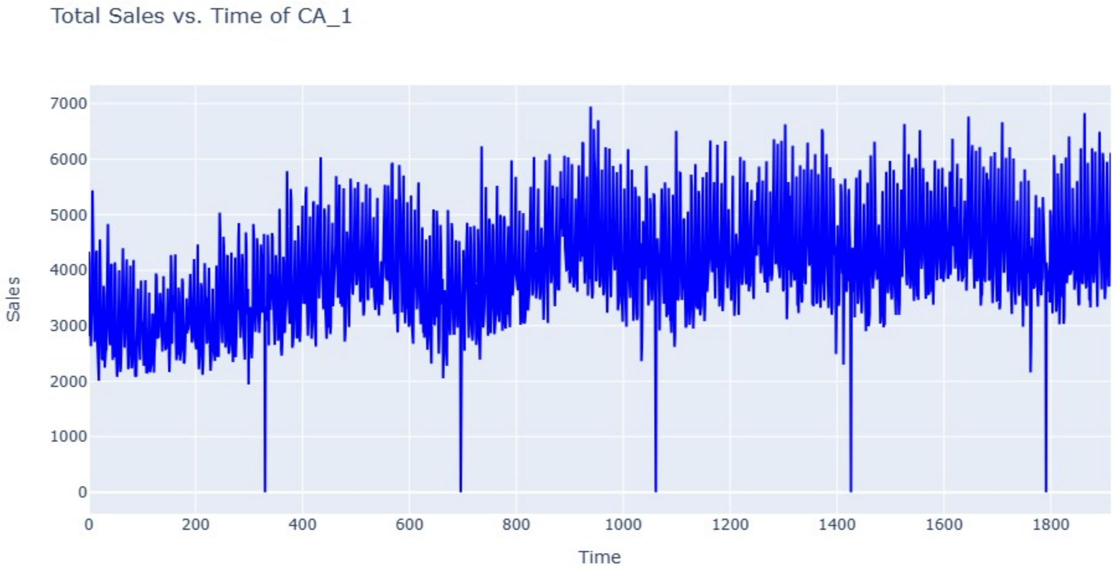


Figure 2: Time series of sales of the product FOODS-1-003-TX-2-validation

For Example : Let us consider a product **FOODS_1_003_TX_2_validation**. We have plotted the daily time series sales of this product. These are sales data of a randomly selected product from randomly selected stores in Texas in the dataset. As expected, the sales data is very erratic, owing to the fact that so many factors affect the sales on a given day. On certain days, the sales quantity is zero, which indicates that a certain product may not be available on that day.

As there is significant noise in this data and there is no significant pattern, so we can not handle it directly. Instead, we can consider the total sales of all products of a particular store in **California** to visualize the daily sales in that particular store.

Daily Sales of a store 1 in California :



3.1 Testing for trend existence

Now we have to test for the existence of the trend in the underlying time series data.

3.1.1 Implementing Relative Ordering Test for trend existence

This is a non-parametric test procedure used for testing existence of trend component. Here we are testing H_0: No trend against H_A : Trend is present.

Here We observe Kendell's τ where

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

where Q counts the no of decreasing points in the time series and is also the number of discordances.
Under the null hypothesis of no trend

$$E(\tau) = 0, Var(\tau) = \frac{2(2n+5)}{9n(n-1)}$$

Asymptotic test for H_0: No trend is based on the statistic

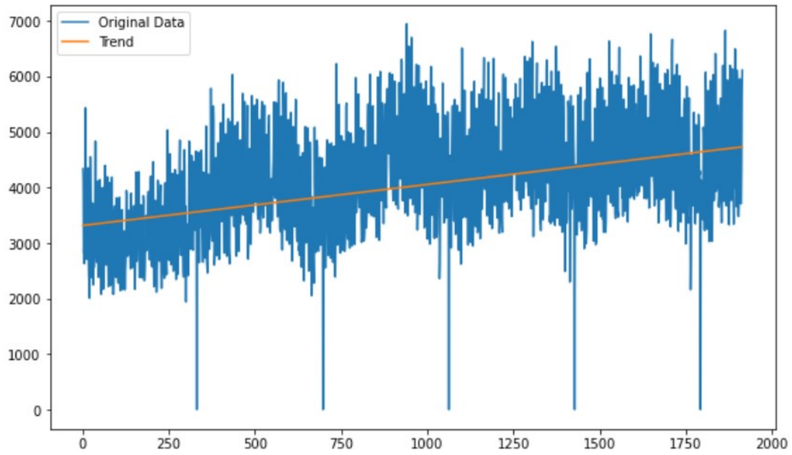
$$Z = \frac{\tau - E(\tau)}{\sqrt{V(\tau)}} \sim N(0,1)$$

We would reject the null hypothesis of no trend at level of significance α if observed $|Z| > \tau_{\frac{\alpha}{2}}$
Here in this data, Kendall's $\tau = 0.27$ and p value is less than 0.05 so We will reject the Null hypothesis of no trend.
Hence there is a significant trend in the data.

3.1.2 Estimation of Trend

Now we have to estimate the trend of the time series data. To estimate the trend we use several approaches

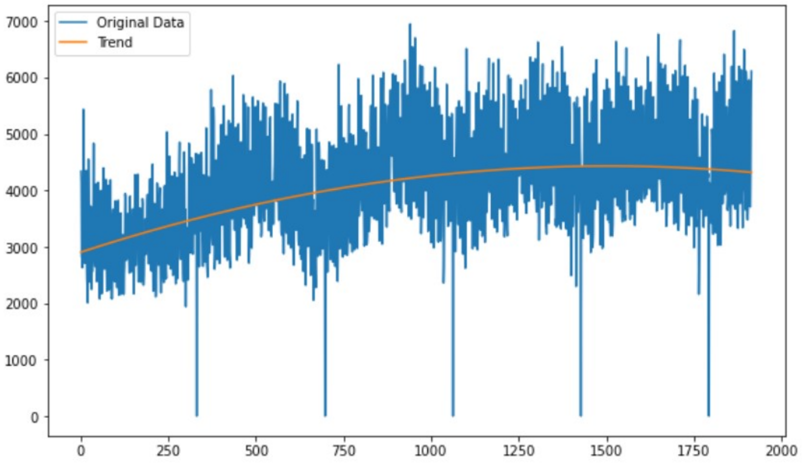
- **Trend estimation by fitting linear trend line using least square method:** Here We fit a trend line using linear regression approach by assuming normality.
  The fitted line is
  $$Y_t = \alpha + \beta t + \epsilon_t$$
  where t = 1,2.....n and $\epsilon_t \sim N(0, \sigma^2)$ independently.



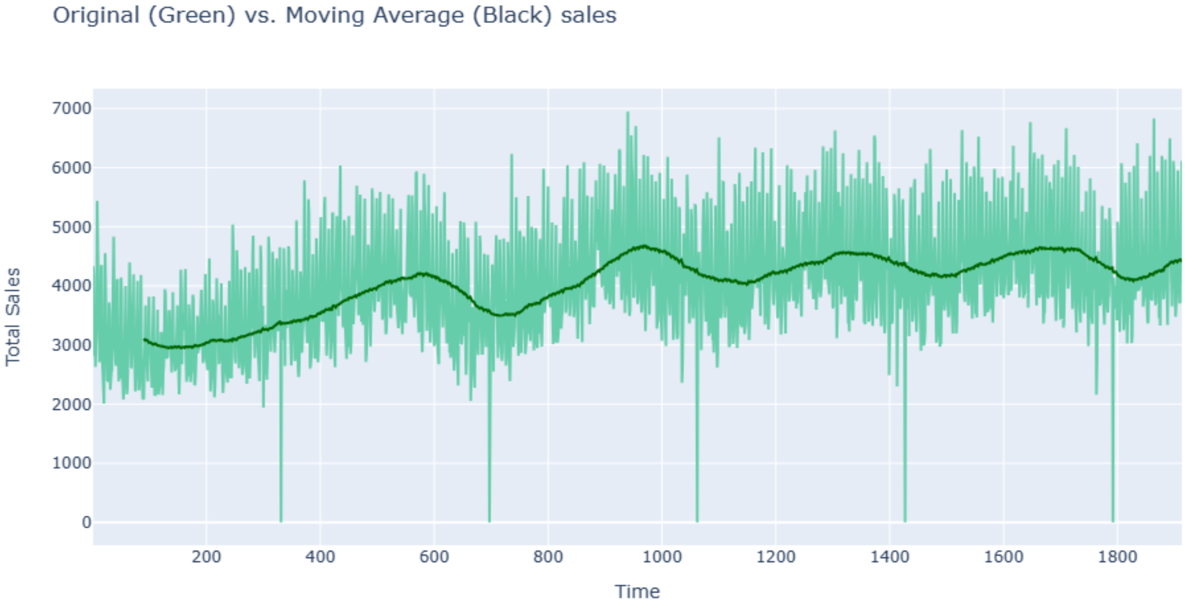
- **Fitting Polynomial trend of degree 2:**
  Here We fit a trend line using a polynomial of order 2.The fitted quadratic trend line is

$$Y_t = \alpha + \beta t + \gamma t^2 + \epsilon_t$$

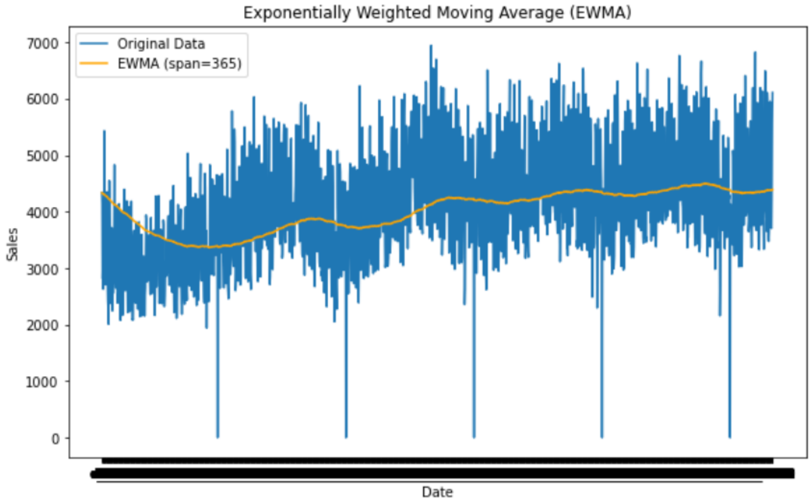
  where t = 1,2.....n and $\epsilon_t \sim N(0, \sigma^2)$ independently.



- **Using Moving Average:**
  Here we will fit equal weighted moving average and the weight equals to the window length of the moving average. The choice of window length depends upon the pattern of the data.For example, if the data has a large amount of noise, a larger window size may be needed to achieve the desired level of smoothing. On the other hand, if the signal is already relatively smooth and free of noise, a smaller window size may be sufficient.
  Here we have chosen a window length of 90 for convenience by visually inspecting the data.

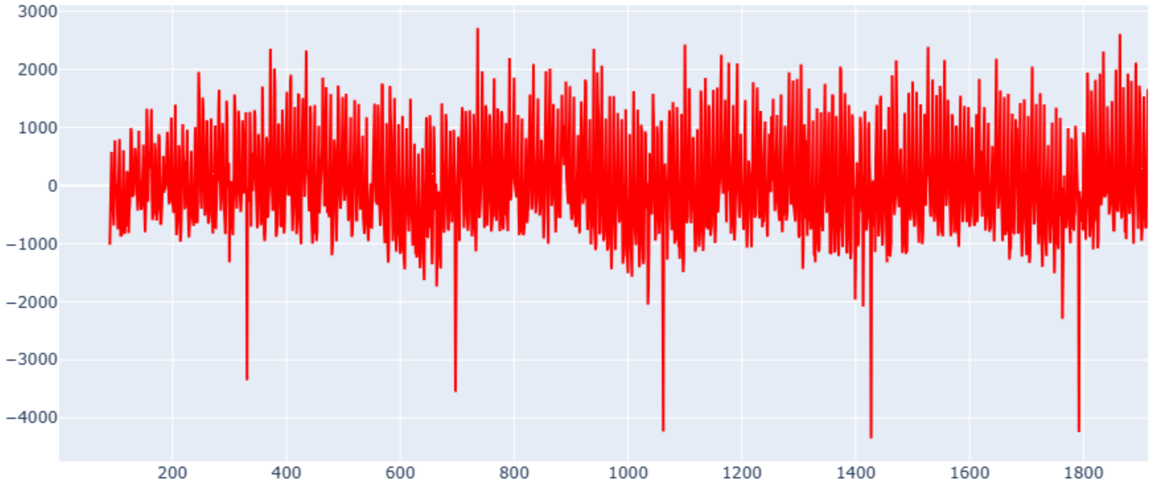


- **Exponentially weighted Moving Average:**
  EWMA is an example of one sided moving average filtering with weights decreasing exponentially inside moving average window as one moves further and further away from the time point at which trend is estimated.



3.2 Detrending the data :

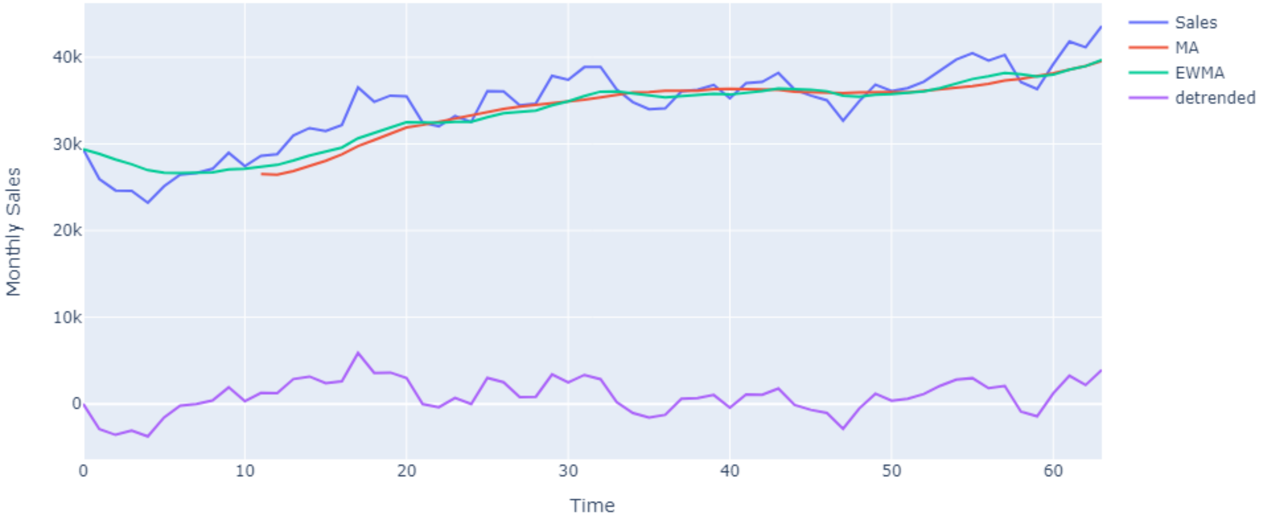
Now we will detrend the data to further visualize the data and to check for seasonality in the detrended data.



From the above figure it is clearly seen that the data is free from time trend and here we may also test for existence of trend in the detrended data.

3.2.1 Seasonality Check using Friedman' test

Here we use **Friedman's test** for testing seasonality which involves estimation and elimination of trend, if required. We work with the detrended values here . The observed value of the test-statistic is $X = 25.861$ which is greater than $\chi^2_{12-1} = 19.675$, hence, we reject the null hypothesis and thus seasonality exists in the data.



3.3 Testing for randomness of a time series data:

- **Turning point test:** This is a non parametric test procedure for testing randomness of a time series data.
  H_0 : series is purely random (does not contain any deterministic component) against alternative hypothesis
  Here We apply turning point test on the original time series data and conclude that the series is not fully random
  The test statistic is
  $$z = \frac{P - E(P)}{\sqrt{var(P)}}$$
  where P is the total number of turning points
  Under null hypothesis,
  $$E(P) = \frac{2}{3}(n-2), V(P) = \frac{16n-29}{90}$$
  Here the value of the test statistic is -19.48 and hence the null hypothesis is rejected.

4 Modelling

Now, we need to model the data. We need to first check which model will fit this data.

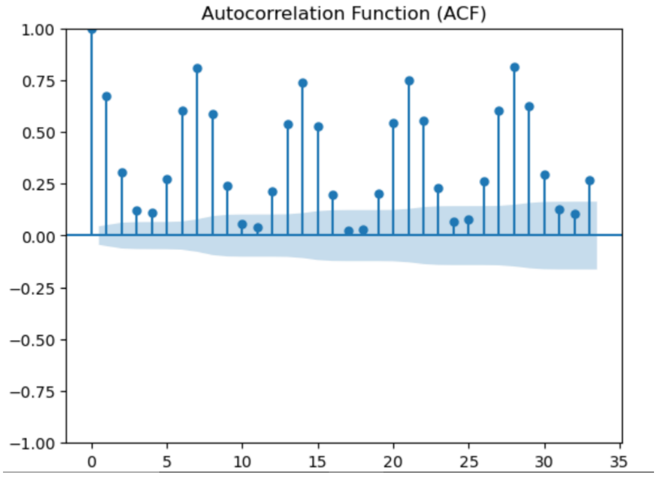
4.1 Train/Val split

First, we need to divide the dataset into training and validation sets to train and validate our models. We will use the last 30 days' sales as the validation data and the sales of the 70 days before that as the training data. We need to predict the sales in the validation data using the sales in the training data.

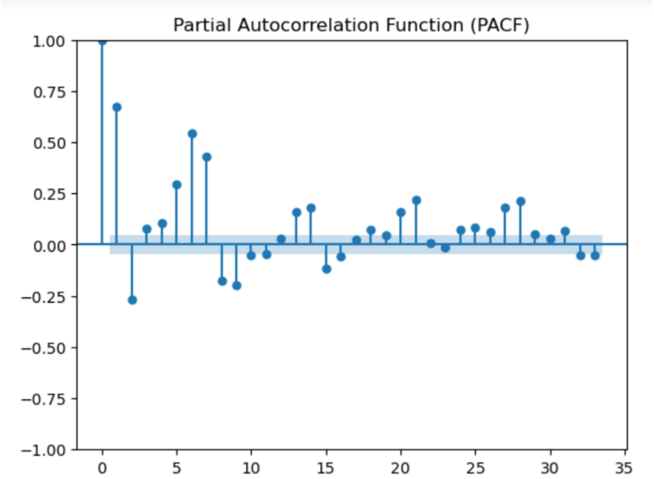
4.2 ACF and PACF

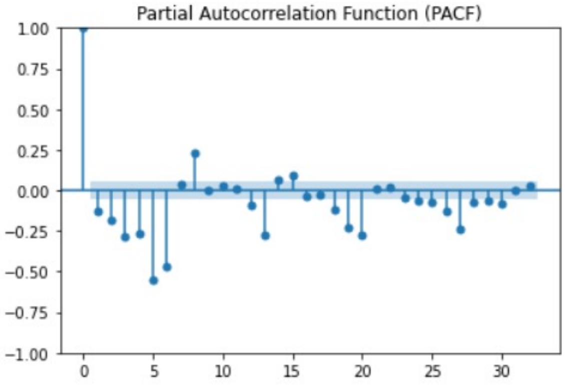
First let us check the ACF and PACF plot and identify the model. Since, this is real life data, obviously there will be some ambiguity. From both ACF and PACF plots, it's not clear whether they are tailing off or cutting off, so let us discard MA and AR processes. First let us get the differenced(d=1) time series, and test for stationarity by ADF test. The ADF test rejects the null hypothesis i.e the differenced time series is stationary. Hence, the we can fit ARIMA model, with d=1. To get p value for AR for this, we will look at PACF plots of the differenced time series. The no of large spikes looks to be 1, Thus it's AR(1).

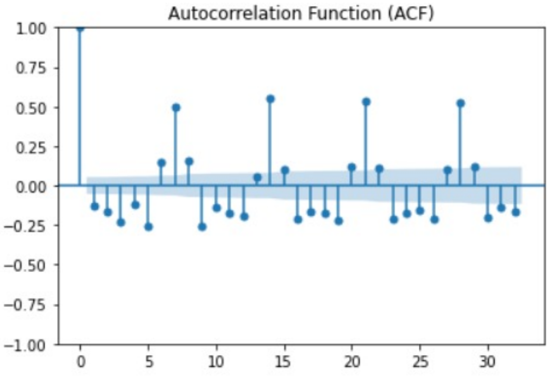
0.5



0.5





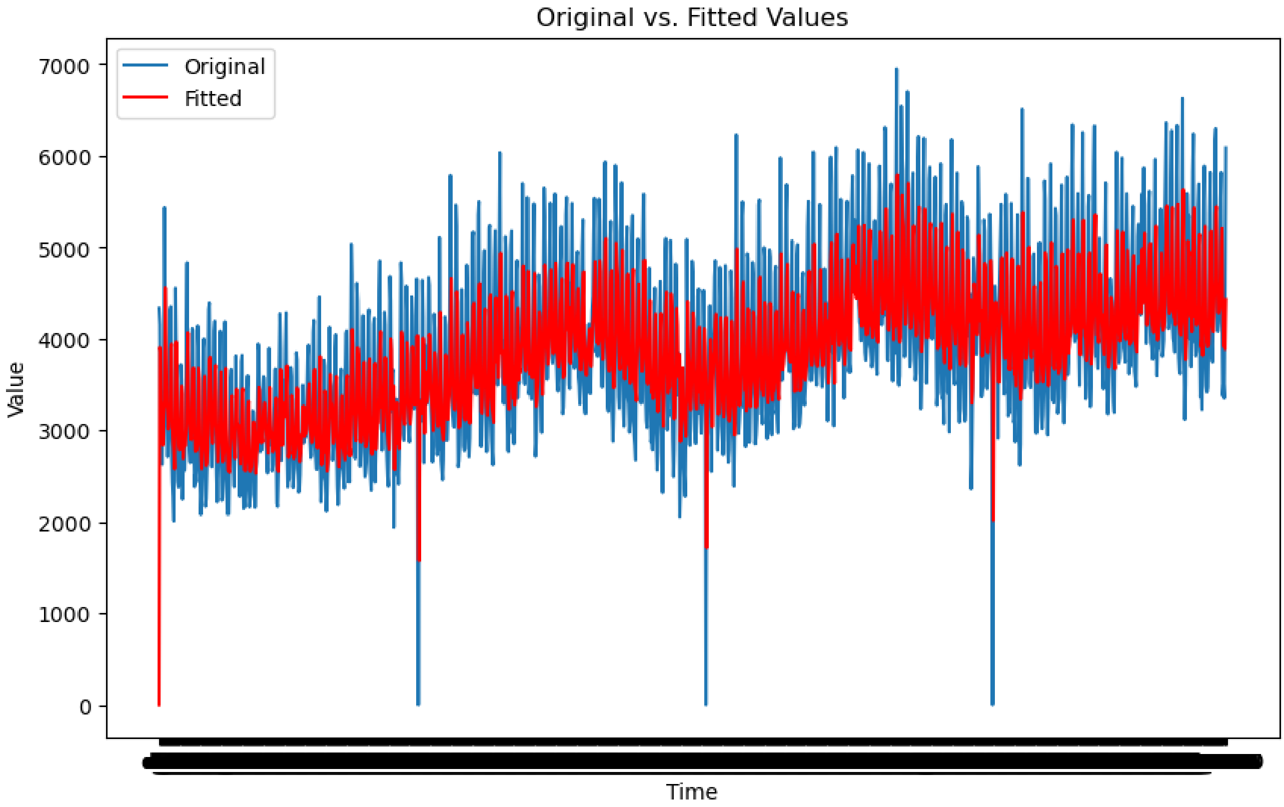


To get q value, we will look at ACF plot of the differenced time series. The no of large spikes looks to be 1. Thus it might be MA(1).

So, let us fit ARIMA(1,1,1) model.

4.3 Fitting

Lastly, let us fit the ARIMA(1,1,1) model. We can see that the fit is quite well. The RMSE for this model is 739.59.



5 Conclusion

We have calculated the total Walmart sales value of the store 1 of California, tested for the existence of different components of time series with different parametric and non-parametric tests and lastly, identified and fitted ARIMA(1,1,1) model to this time series data.