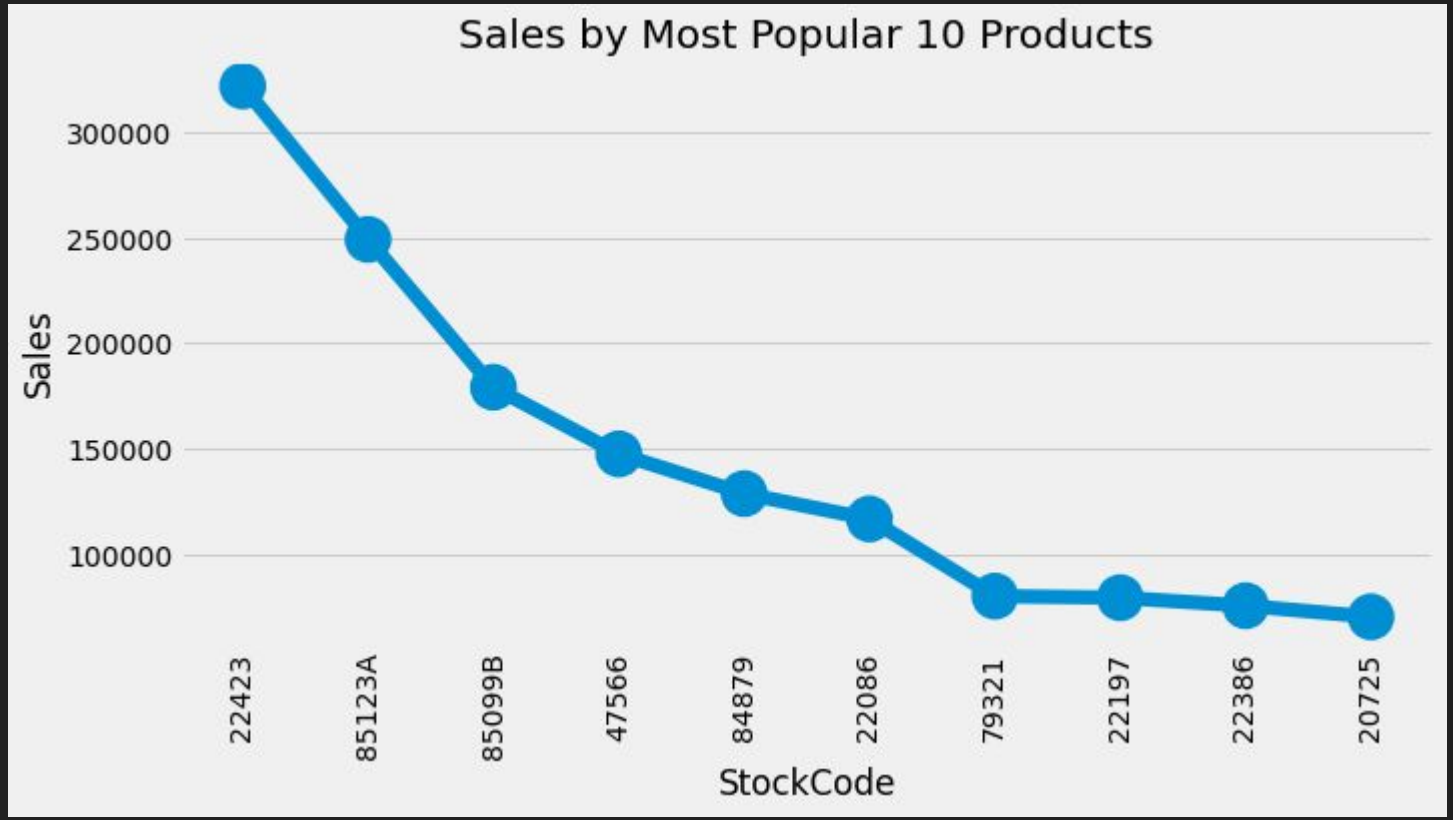
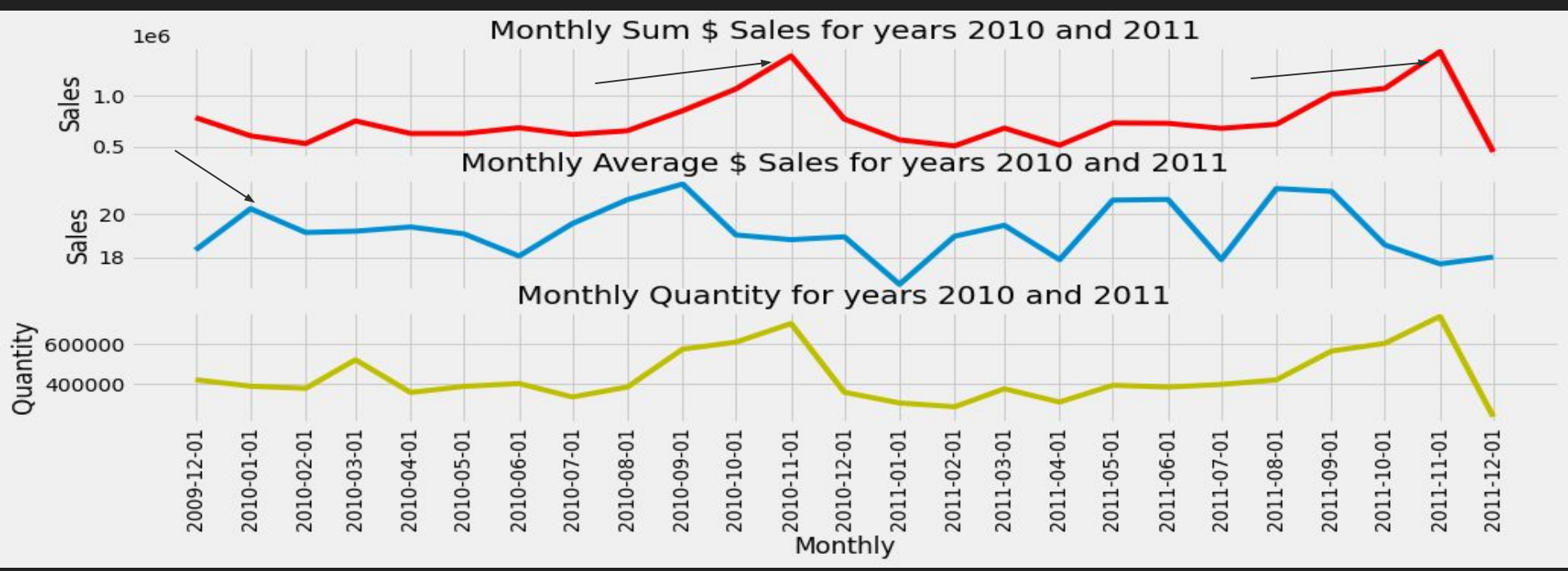
Time Series Sales Forecasting

Data Analysis

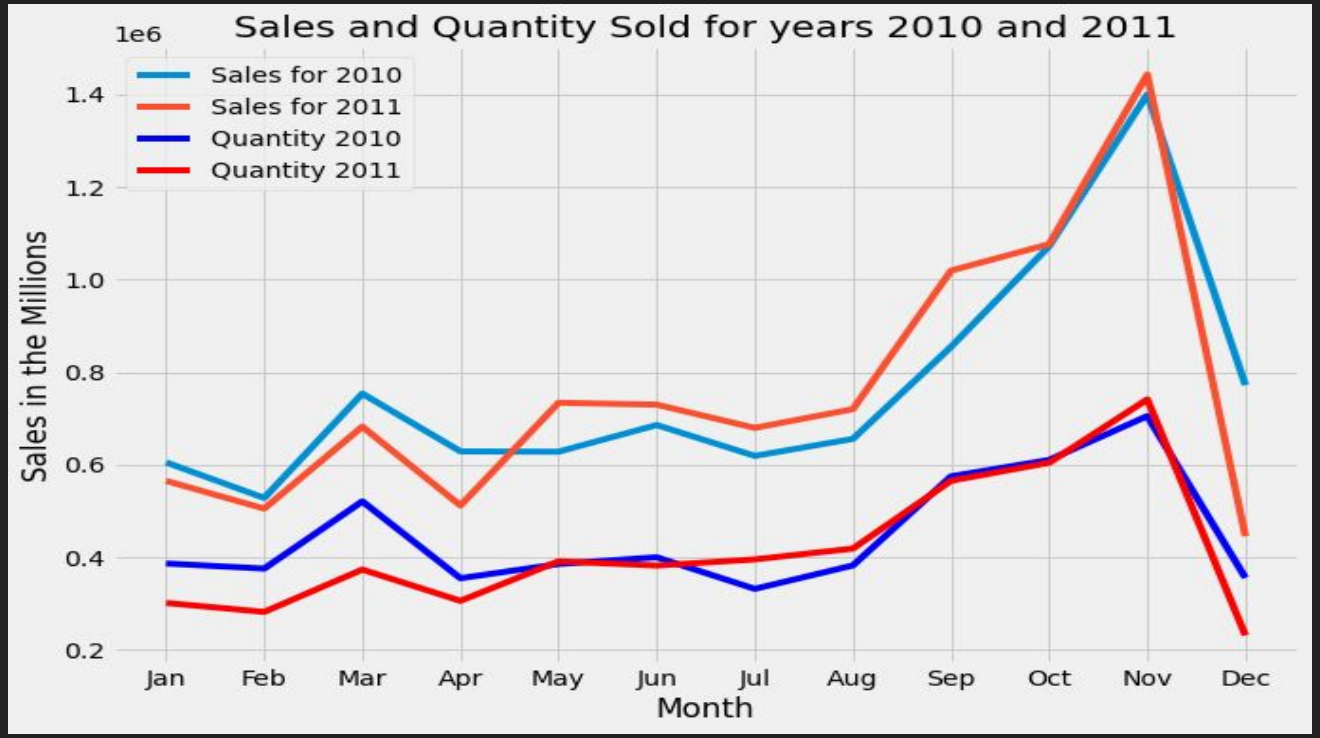
Understand the trends in the data

Link to the notebook
https://github.com/rime11/customer_segmentation_sales_forecasting/blob/master/Notebooks/arima_model.ipynb

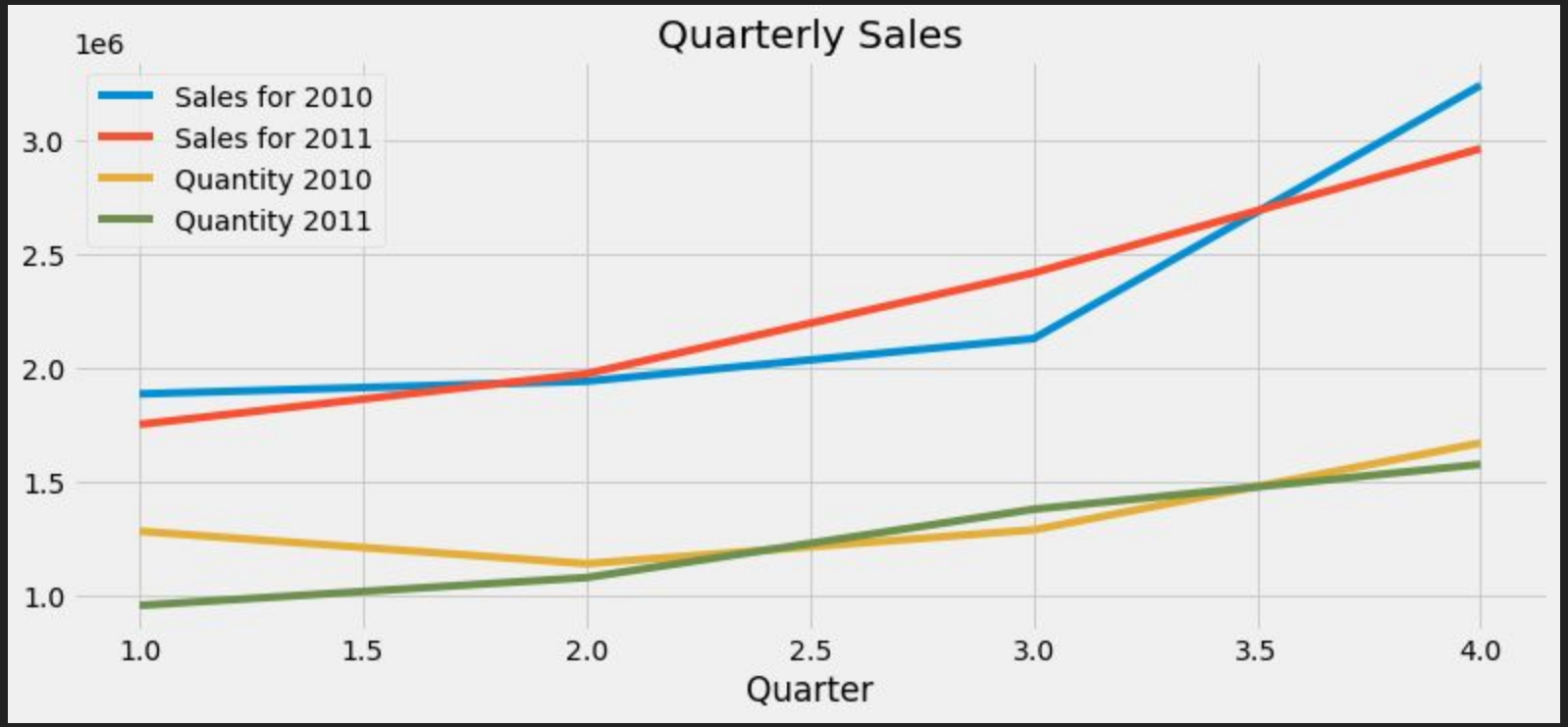

Sales by Most Popular 10 Products
300000
250000
200000
150000
100000
Sales
22423
85123A
85099B
47566
84879
22086
79321
22197
22386
20725
StockCode

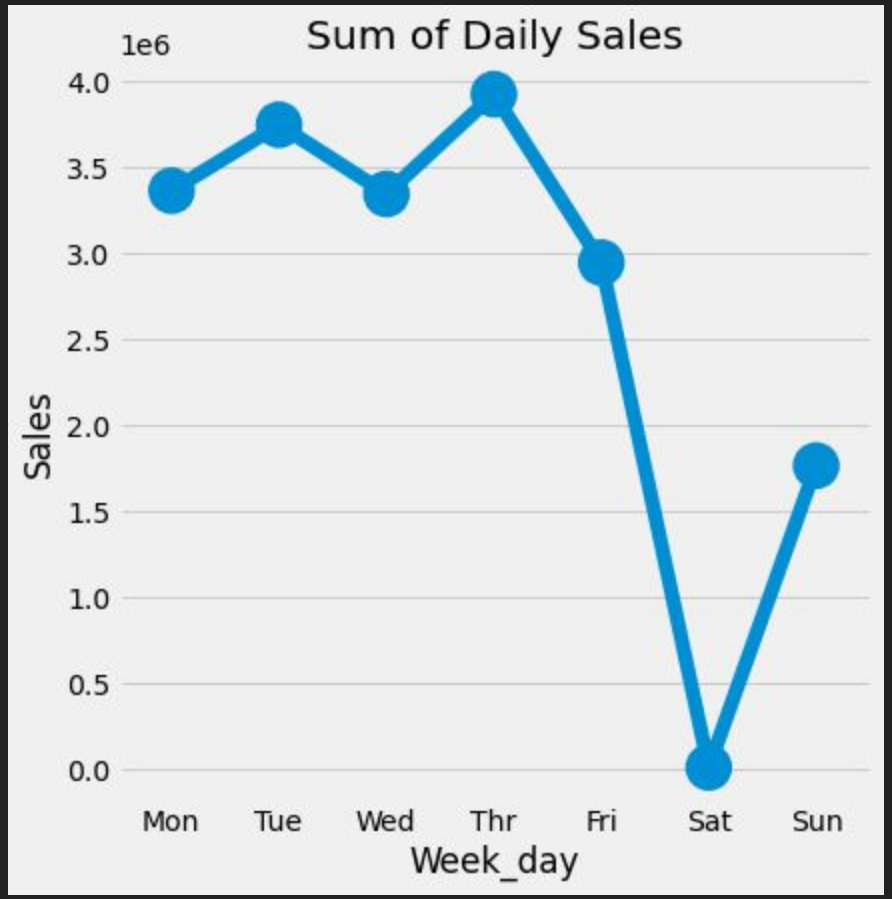


- In September 2010 there is a spike in average sales which then goes down but in sum of sales continues to rise, which means that the number of orders continues to increase of products of similar value.
- In November there is a spike in sum of sales, which is probably the Christmas rush which starts slowing down in December, meaning that most people buy their gifts in November. Since this is an online store it makes sense that people would order gifts early so that they arrive on time. It is interesting to note that in November sum of sales goes up while average sales goes down, which means that there are many small orders.

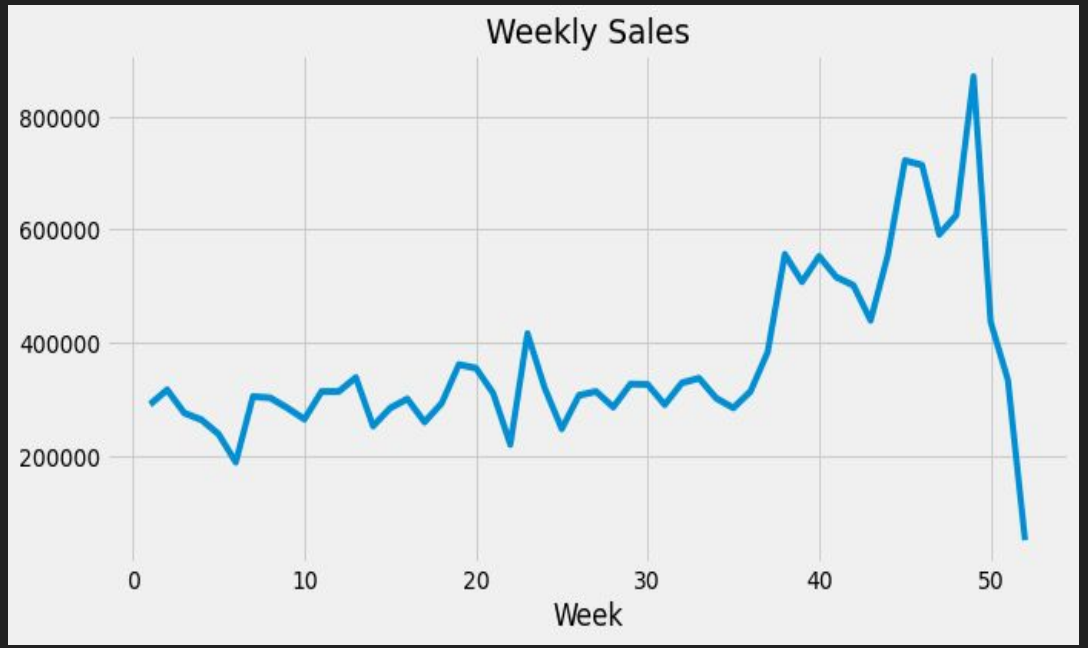


Spike in November is obvious in 2010 and 2011

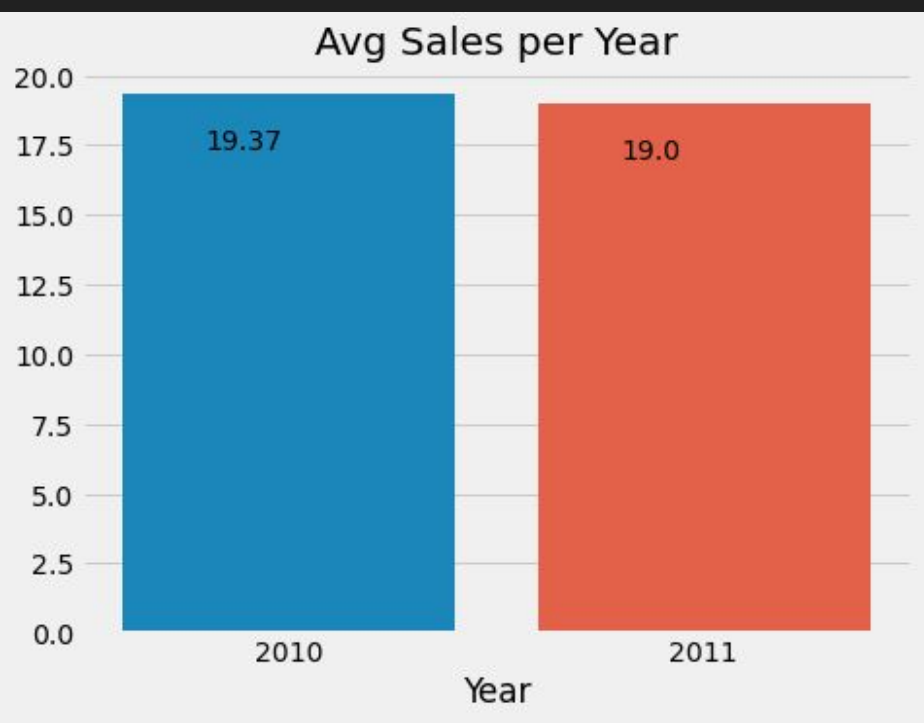

Quarterly Sales
1e6
Sales for 2010
Sales for 2011
Quantity 2010
Quantity 2011
3.0
2.5
2.0
1.5
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Quarter

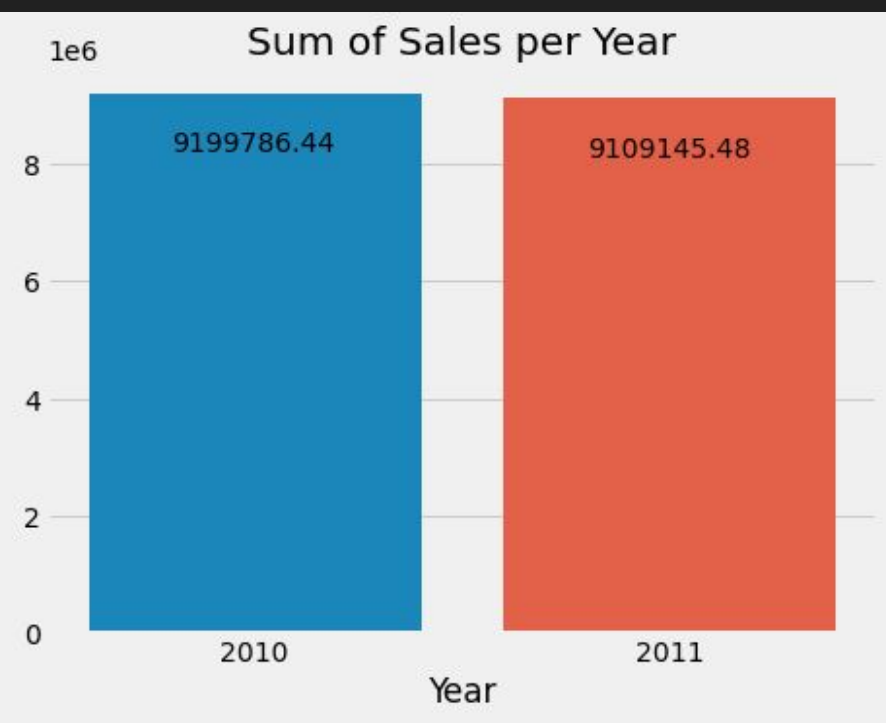


Sales spike on Thursdays and are the lowest on Saturdays



Sales increase towards the end of the year.





- There is a slight decrease in average and sum of sales from 2010 to 2011, which could mean that there is less orders being made. An investigation of the number of orders per year

The number of orders increased but the sum decreased from year 2010 to 2011. This could mean the cheaper products have become more popular, so there are more orders but the company is making less in money

Sales Forecasting

Purpose of analysis

One of the most important tasks for any retail store company is to analyze the performance of its stores. The main challenge is predicting in advance the sales and quantity of inventory required at each store to avoid over-stocking and under-stocking. This helps the business to provide the best customer experience and avoid getting into losses, thus ensuring the store is sustainable for operation. The daily sum of sales will be investigated.

What is a time series

- A time series shows how a single variable changes over time and these observations are taken over regular intervals, like weekly or monthly sales. The series would have the date at regular intervals as the index and the variable being measured, in this case total daily sales, as the dependent variable.

The features that need to be analyzed in a time series are:

- **Trend**: it is the long term change in the mean of the series, meaning it is the overall direction the data is traveling.
- **Seasonality**: is the repeating the short-term cycle in the series.
- **Noise**: is the random variation in the series.

Types of models

1. **Additive model**

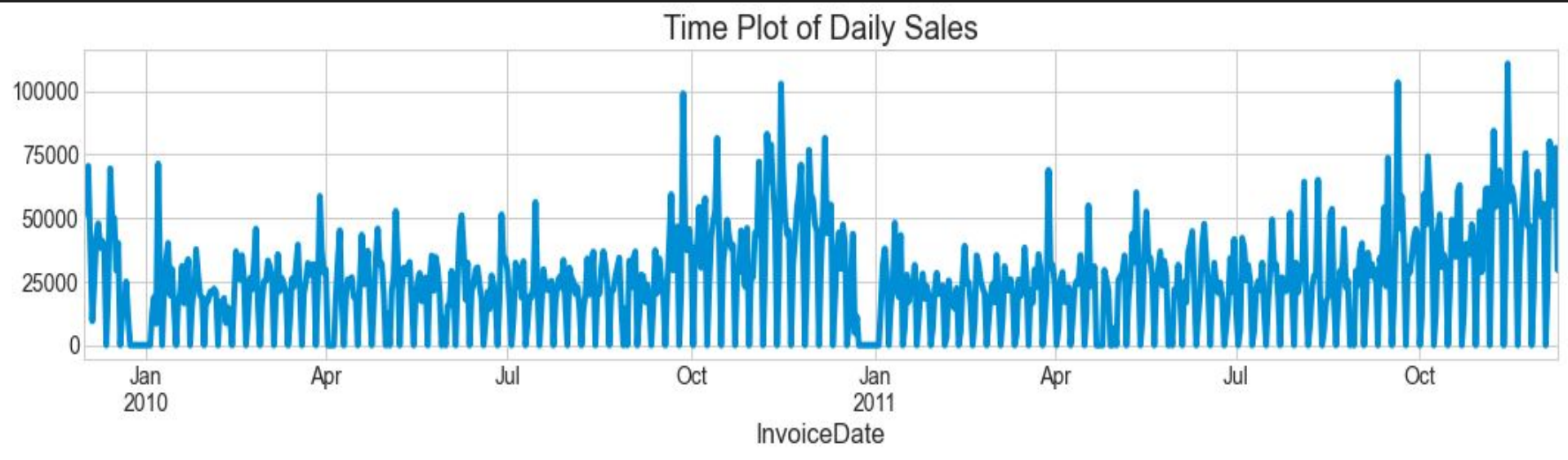
   $$y(t) = \text{trend} + \text{seasonality} + \text{noise}$$

   In the additive model, the behavior is linear where changes over time are consistently made by the same amount, meaning the linear seasonality has the same amplitude and frequency.

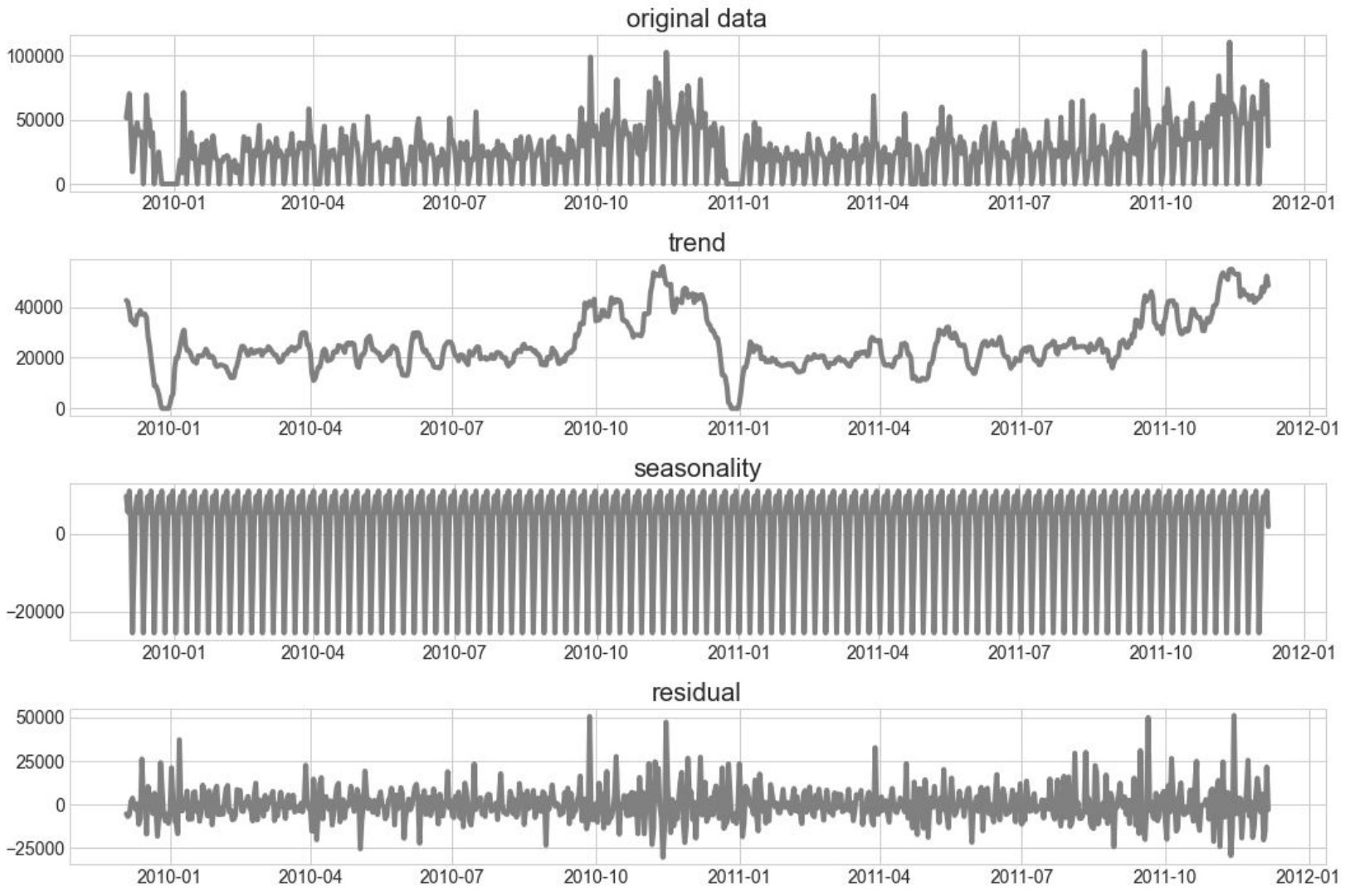
2. **Multiplicative model**

   $$y(t) = \text{trend} * \text{seasonality} * \text{noise}$$

   The multiplicative model has an increasing or decreasing amplitude and/or frequency over time. This means if the seasonal peaks change in amplitude over time then it is a multiplicative model, if they stay relatively the same then it is additive. I will plot daily sales to see which model it is.



The seems to be an additive model because the seasonal peaks do not increase in amplitude or frequency, however I can use decompose to get the trend and seasonality over time.


original data
100000
50000
0
2010-01
2010-04
2010-07
2010-10
2011-01
2011-04
2011-07
2011-10
2012-01
trend
40000
20000
0
2010-01
2010-04
2010-07
2010-10
2011-01
2011-04
2011-07
2011-10
2012-01
seasonality
0
−20000
2010-01
2010-04
2010-07
2010-10
2011-01
2011-04
2011-07
2011-10
2012-01
residual
50000
25000
0
−25000
2010-01
2010-04
2010-07
2010-10
2011-01
2011-04
2011-07
2011-10
2012-01

Observations

The time series is composed of three components: a trend-cycle component, a seasonal component, and a remainder component containing anything else in the time series. This is done to improve understanding of the series and improve forecast accuracy. So here y = trend + seasonal + remainder

In the data, the seasonal component does not change over time. It remains the same over the two years, which makes sense since I know that there is a seasonal spike before Christmas in both years. For the trend I notice that there is a downward trend right after the spike in sales before Christmas. If the three decomposed parts are added back up, the result is the original data.

Stationarity

For a data series to be forecasted, it needs to be stationary. A series is non-stationary due to a unit root, which means it shows a systematic pattern that is unpredictable. A stationary series has constant mean, constant variance and constant covariance over time.

KPSS and the Dickey Fuller tests

The most common stationary test is the KPSS test. The null hypothesis of this test is that the time series data in question is stationary; hence, if the p-value is less than the significance level,0.05 here, then I reject the null hypothesis and infer that the data is not stationary.

There the Dicky Fuller test. Here, the null hypothesis is that the data is not stationary and the alternative hypothesis is the data is stationary.

Test Results

The result of the kpss test is: p_value= 0.03⇒ this is significant, which means that the null hypothesis can be rejected and the data is not stationary

The result of the Dicky Fuller test is: p_value = p_value: 0.176⇒ this is not significant, meaning that the null hypothesis is not rejected and the data is not stationary.

So the data is not stationary and it needs to be differenced.

Linear Regression

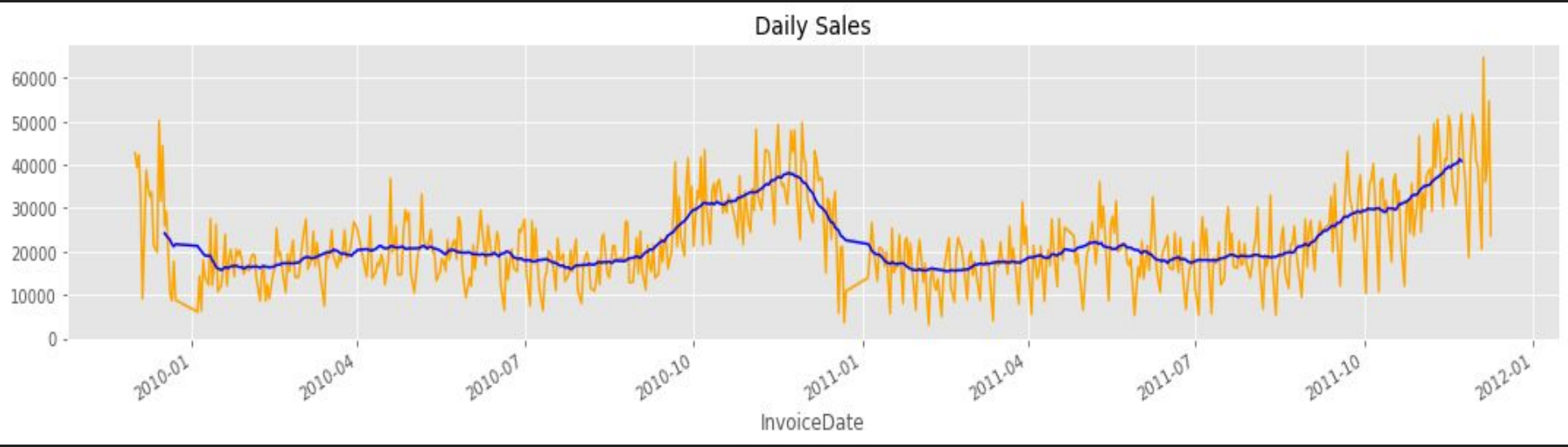
Linear regression will be used to construct a forecasting model. The linear regression model learns how to make weighted sum of its inputs to predict its target, in this case the target is sales. This is the ordinary least squared because the model tries to minimizes the squared error between predictions and targets. It looks like

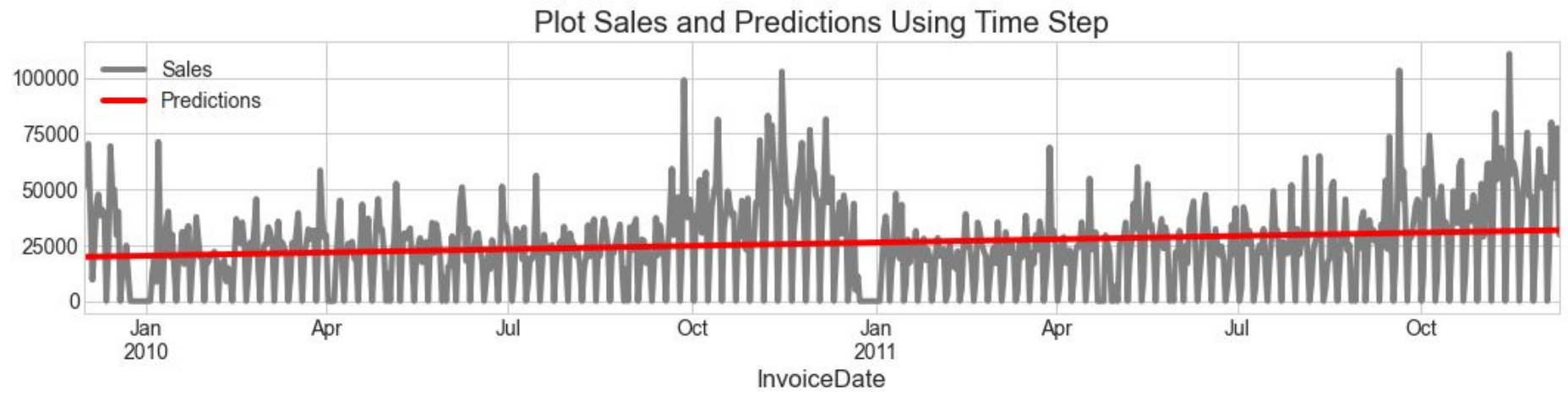
sales = weight1var1+weight2var2 + bias

For the variables: There are two kinds of features unique to time series:

- lag feature
- time step feature

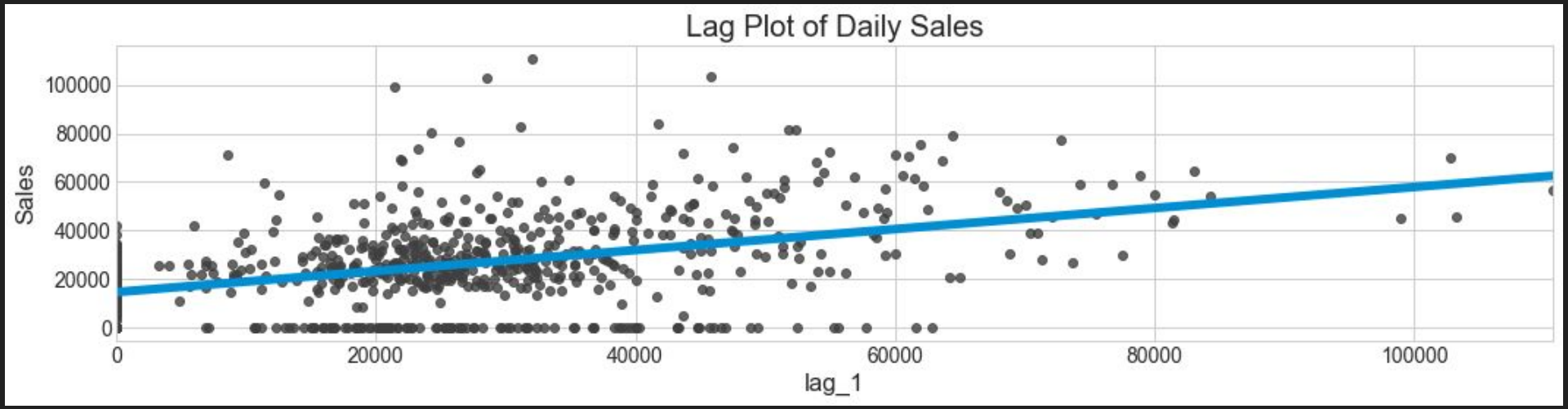
Sales with a 30 day rolling mean





There is also correlation between sales to sales from previous day. This means that high sales on one day means high sales the next day.

The square root of the mean squared error = 18822.941



There is a correlation between current sales to sales from the previous day. This means that high sales on one day means high sales the next day.

The square root of the mean squared error = 17237.29

Autoregressive Integrated Moving Average
or
ARIMA

Arima stands for Autoregressive Integrated Moving Average, which means :

- AR: stands for Autoregression, which means a model that uses the dependent relationship between an observation and some number of lagged observations.
- I: Integrated which means subtracting an observation from an observation at the previous time step in order to make the time series stationary, meaning differencing
- MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

There are three parameters that need to be tuned:

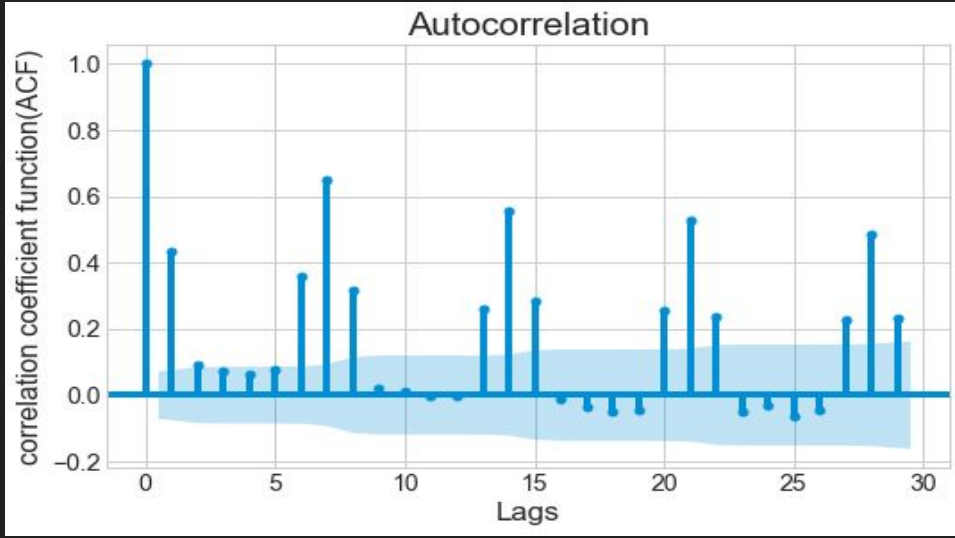
- p is the number of lags of Y to be used as predictors
- d is the minimum number of differencing needed to make the series stationary. A d =0 means the series is stationary
- q refers to the number of lagged forecast errors that should go into the Model

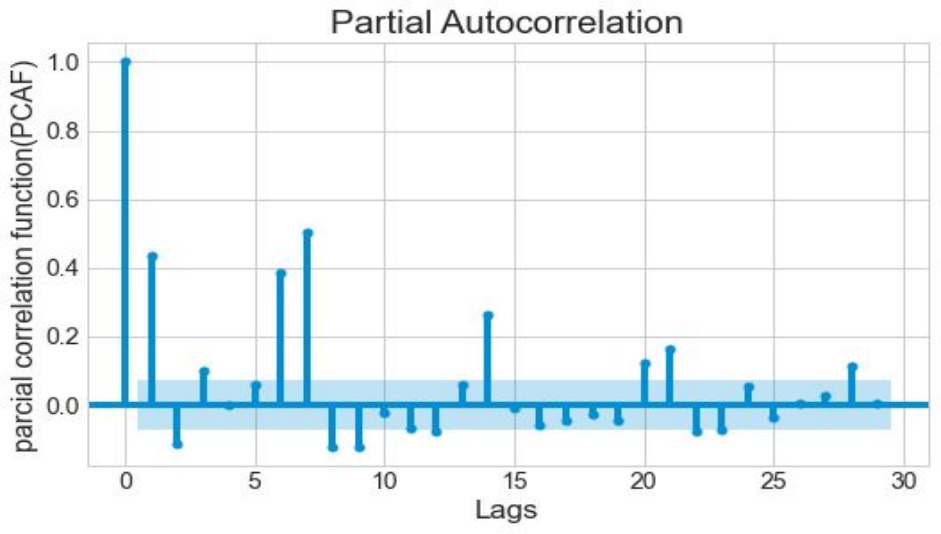
q value is determined from the acf plot

- The Autocorrelation function (ACF) plot shows how a given time series is correlated with its past values. The ACF is the plot used to see the correlation between the points, up to and including the lag unit.

P value is determined from the pacf plot

- Partial correlation, or PACF is a conditional correlation. It is a summary of the relationship between an observation with observations at prior time steps.

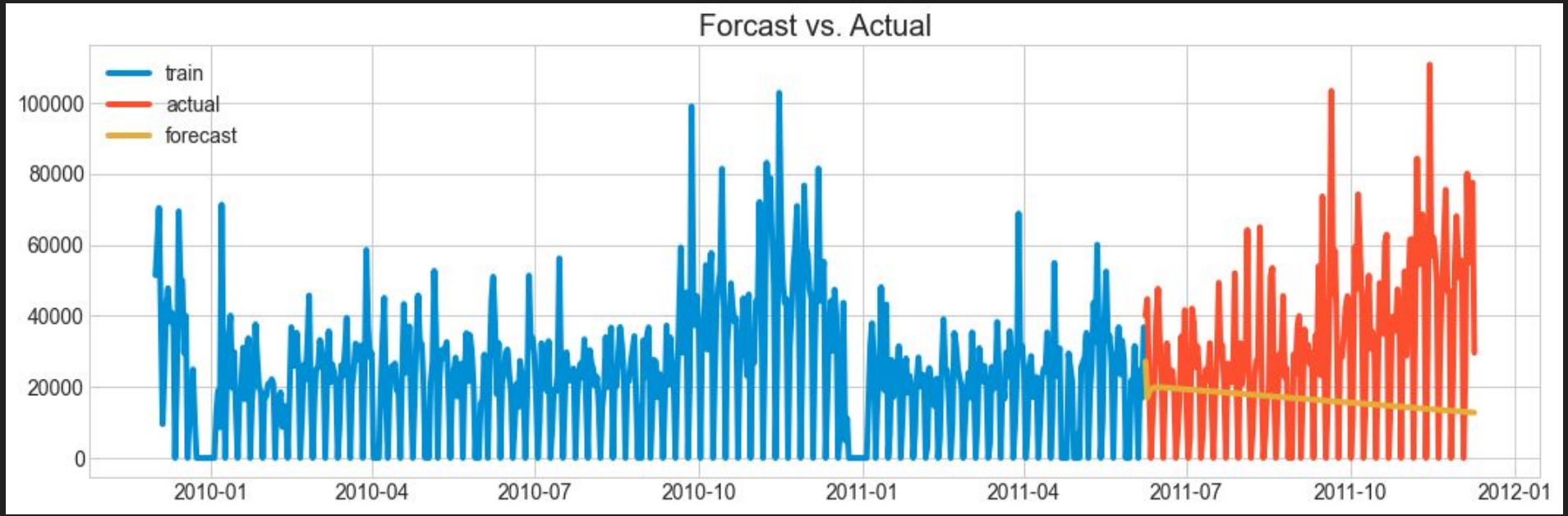




Observations

- From the ACF graph, the time series is significantly positively correlated with its past values at 1 and 2 lags.
- From the PACF lag 1, 2 and 3 are significant. Also, when auto-correlation decrease too fast it may indicate over_differencing and if it decrease too slow(stays positive for more than 10 lags) it indicates under_differencing. In the graph above, the autocorrelation decreases slowly up to 9 lags. From the PACF we get the AR or p term, from ACF we get the MA or q term. So the model will be run with p= 1, q= 3, and d=1.

p= 1, q= 3, and d=1



Root mean squared error = 27765.24

To investigate the best values for p and q, I wrote a function that takes in the data and splits it into 80% training and 20% test sets. It also takes in a tuple of p,d,q for arima model and fits and predicts then returns the root mean squared error. This function calls another function that iterates through several values for p, d and q.

The best mean squared error result =23143.73 for p=0, d=2, q=1