# D213 Advanced Data Analytics

Professor William Sewell

Mackenzie Simon

# **Part I:  Research Question**

**A1: Our goal is to use the ARIMA model in order to forecast future revenue. We want to answer if our current business model will lead towards higher future revenue.**

**A2: Our objective for the data analysis is to predict future revenue using the ARIMA model. Once we build our ARIMA model, we can forecast for future revenue. Based on our ability to forecast future revenue, we can budget for minimizing customer churn.**

# **Part II:  Method Justification**

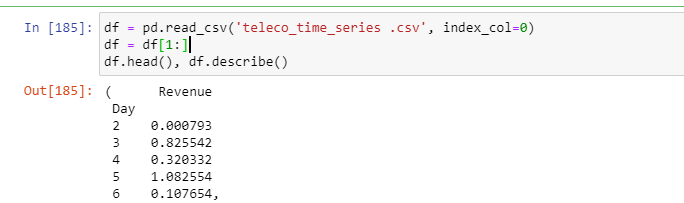
**B: There are multiple assumptions when it comes to time series analysis.** For our time series there is t**he assumption that mean, and variance are consistent** over time. We call this stationarity. In addition to using stationarity, we use something called auto correlation. Auto correlation measures the relationship between a variable’s current and past values. We use lagged values of the target variable as our X variables for auto regression. Prabhakaran (2020) describes ARIMA as a class of models that explains a given time series based on its own past values.

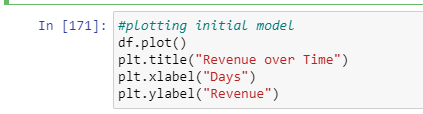
# **Part III:  Data Preparation**

**C1/2:**

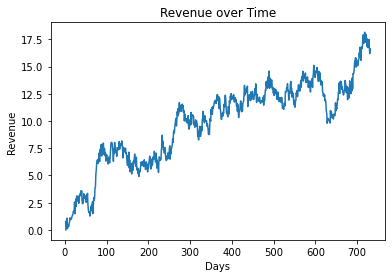
**For our model, I created a data frame with day as our row index and revenue as our column values. I decided to drop the first row because the revenue value was zero. I then graphed the time series using the matplotlib library. For the steps in our model, I just kept the time series to individual day.**

**Formatting Code:**





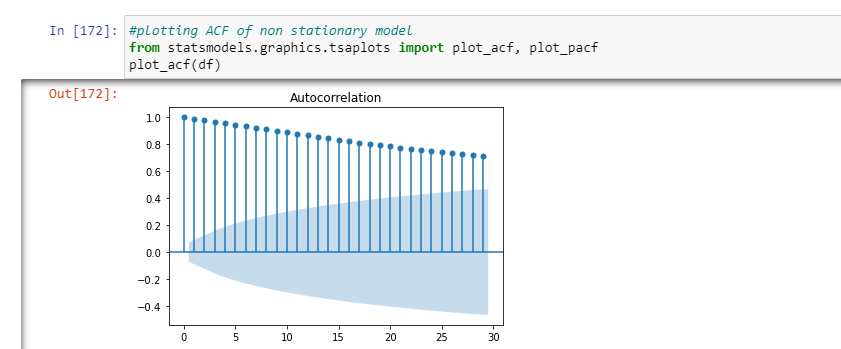
**Time Series Plot:**



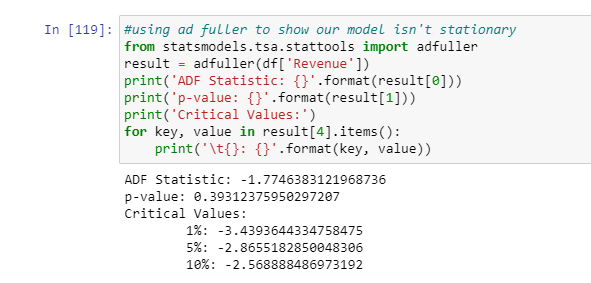
C3:

In order to evaluate the stationarity of the series, I looked at both the auto correlation plot, and the augmented dickey fuller test. The spikes for our auto correlation plot are statistically significant for lags up to 30. This means that the revenue prices are highly correlated with each other proving that the model is not stationary. Our Dickey Fuller test also gives us an output p value of .393, meaning that we do not reject the null hypothesis. By not rejecting the null hypothesis, we are showing that the model is not stationary.

**Auto Correlation Plot:**

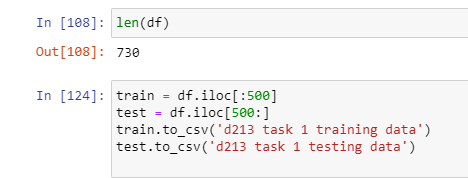


**Dickey Fuller Test:**



C4:

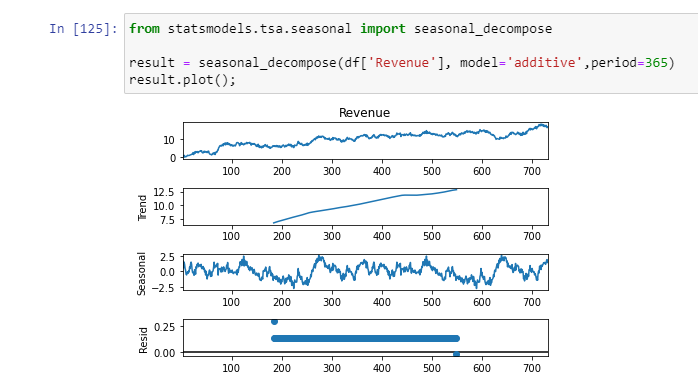
In order to prepare our data for an ARIMA model, I split the data into training and test sections. For our training section, I set the data to the first 500 rows focusing on revenue. For our testing set, I set the data from row 500 to the end for revenue. I then wrote train and test to csv for our project submission.



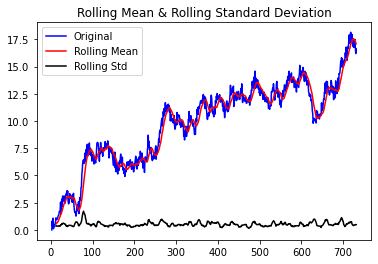
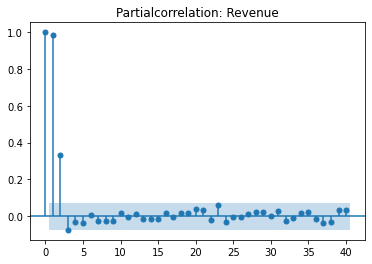
# **Part IV:  Model Identification and Analysis**

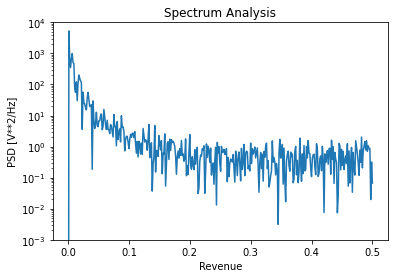
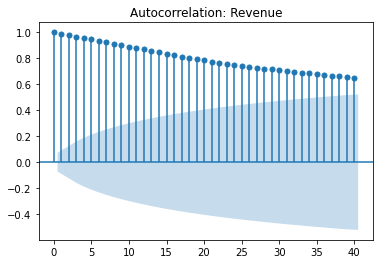
D1:

Through the statsmodels library, we can use the seasonal decompose function to deconstruct our time series. For our seasonal\_decompose function, I set the period to one year or 365 days because we had two years of data.

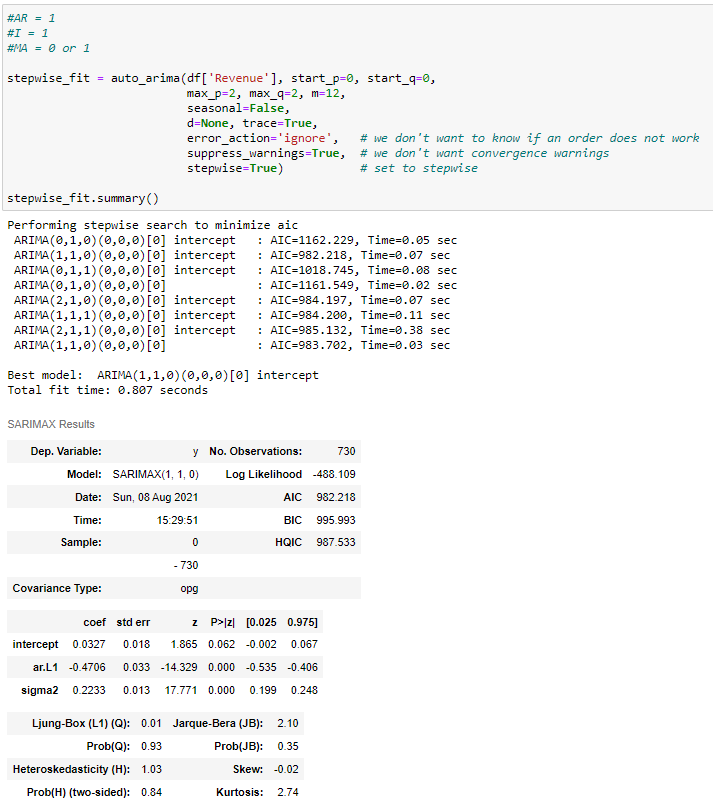


Looking at the decomposition model, I see a lack of seasonality. The seasonal trend series doesn’t show a strong correlation to our revenue. In addition, the max and min of our seasonal series are -2.5 and 2.5. The seasonal max and min values are insignificant compared to our revenue and trend lines. Overall, there is a positive trend line, as we saw earlier with our non-stationarity results. There is a lack of trend for our residuals. We see the highest amount of noise located in our early revenue earnings.



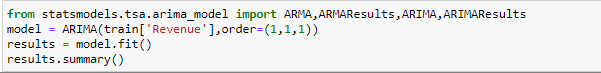
D2: In order to setup our ARIMA model, we must get a value for p,d, and q. P is the order of the AR term. D is the number of differencing required to make the time series stationary. Q is the order of the MA term. In order to obtain these values, we used a step wise auto\_arima function from the pmdarima library.



After running our stepwise auto\_arima function, we look for the combination with the lowest AIC. We see that we can use 1 for AR, 1 for I, and 0 or 1 for MA. For our ARIMA model I decided to use (1,1,1).

D3/D4/D5:

After obtaining our ARIMA parameters we can start building our model. We fit the ARIMA model to our parameters (1,1,1).



.

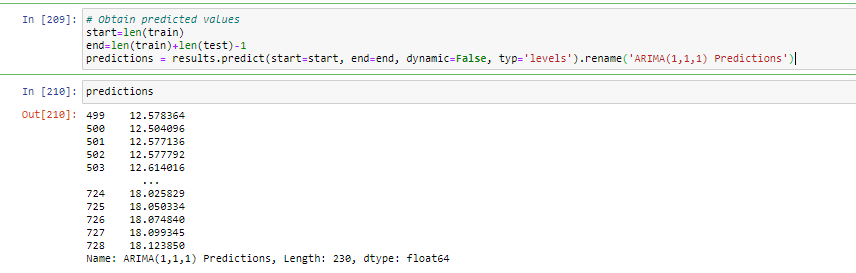
ARIMA output:

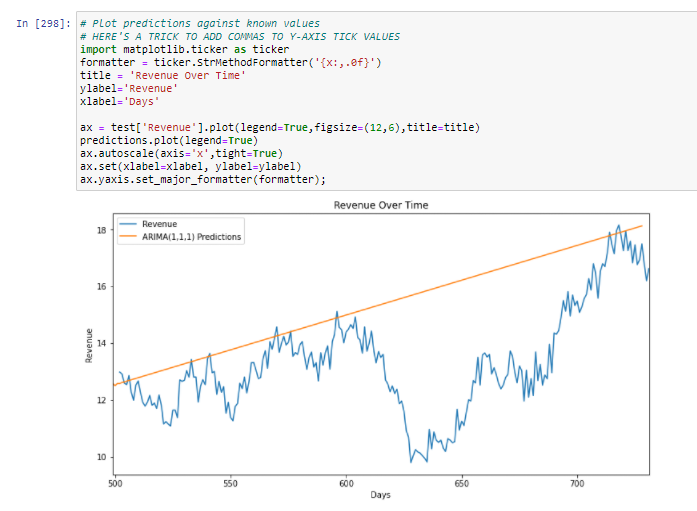
|  |  |  |  |
| --- | --- | --- | --- |
| ARIMA Model Results | | | |
| **Dep. Variable:** | D.Revenue | **No. Observations:** | 499 |
| **Model:** | ARIMA(1, 1, 1) | **Log Likelihood** | -332.334 |
| **Method:** | css-mle | **S.D. of innovations** | 0.471 |
| **Date:** | Sun, 08 Aug 2021 | **AIC** | 672.667 |
| **Time:** | 15:29:53 | **BIC** | 689.517 |
| **Sample:** | 1 | **HQIC** | 679.280 |
|  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **const** | 0.0245 | 0.014 | 1.709 | 0.087 | -0.004 | 0.053 |
| **ar.L1.D.Revenue** | -0.4914 | 0.074 | -6.651 | 0.000 | -0.636 | -0.347 |
| **ma.L1.D.Revenue** | 0.0137 | 0.082 | 0.167 | 0.867 | -0.147 | 0.174 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Roots | | | | |
|  | **Real** | **Imaginary** | **Modulus** | **Frequency** |
| **AR.1** | -2.0351 | +0.0000j | 2.0351 | 0.5000 |
| **MA.1** | -72.9955 | +0.0000j | 72.9955 | 0.5000 |

We can then use our testing and training data to predict future values. We will graph our test data section and plot it on our current graph.

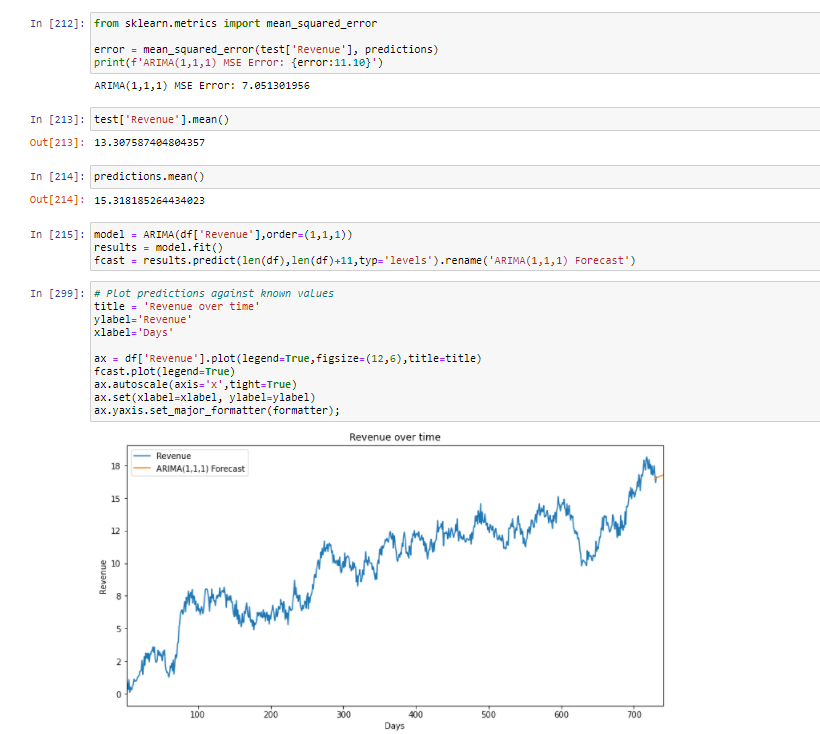




Looking at our current fitted predictions, the ARIMA model looks good. There is a little bit of overfitting for our ARIMA model, but over all I would move forward.

We can then use our ARIMA model to predict for the future. I decided to plot for the next 11 days because we are looking at revenue per day. Our ARIMA model predicts a positive trend in revenue.

ARIMA MSE, FORECAST:



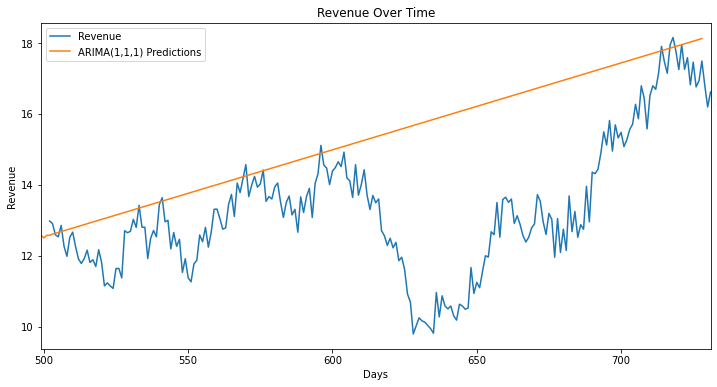
# **Part V:  Data Summary and Implications**

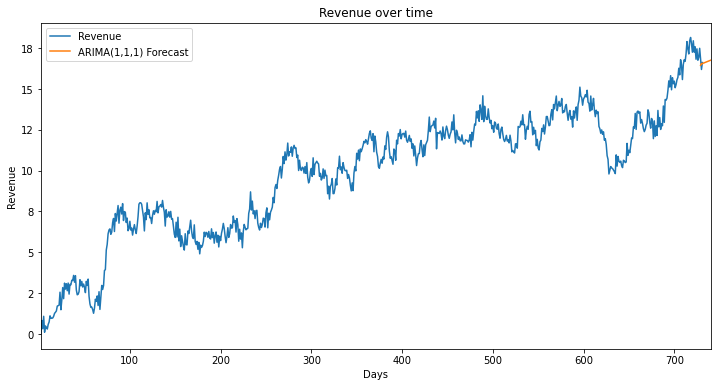
E1:

In order to select our ARIMA model, we decided to use (1,1,1). We gained these parameters from our stepwise auto\_arima function. Our ARIMA results had a low AIC at 672.67 with a mean squared error of 7.05. Our mean squared error falls within a realistic range for our revenue. We set our prediction interval for the next 11 days. I justified the forecast length of 11 days because our time series is revenue per day.

E2:

Test:



Forecast: 

E3:

Based on our results, I would move forward with our model despite there being a small amount of overfitting. Yui (2021) describes overfitting as allowing additional features — in this case, it would be lagged errors. I would suggest continuing our current business model because our revenue will continue to increase. We could also try running the model with the parameters of (1,1,0) to see if there is less overfitting.

# References:

Prabhakaran, S. (2021, June 14). *ARIMA model - Complete guide to time SERIES forecasting in Python: ML+*. Machine Learning Plus. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/.

Yiu, T. (2020, April 26). *Understanding ARIMA (time Series Modeling)*. Medium. https://towardsdatascience.com/understanding-arima-time-series-modeling-d99cd11be3f8.