**D213 Task 1:**

***Time Series Analysis***

**I. Research Question**

Research Question:

Based on our continued efforts to reduce customer churn over the past two years, how well does our telecom company’s last two months of revenue compare to the potential forecasted amounts produced from a Time Series Analysis on the first one year and 10 months of revenue data?

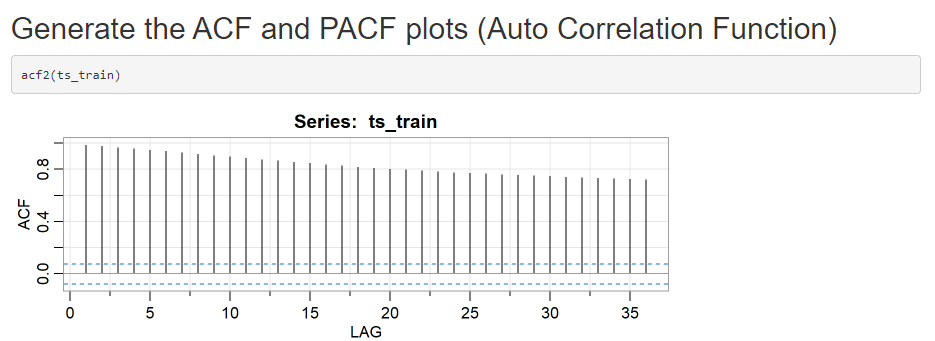
Goals/Objectives:

1. Identify the best ARIMA model to fit our training data
2. Based on the Lyung-Box Statistic, identify the ARIMA model that is above the dotted line. This ARIMA model would be considered a good fit.
3. Find the ARIMA model with an AIC and BIC under 910.

**II.  Method Justification**

Assumptions of a Time Series Model:

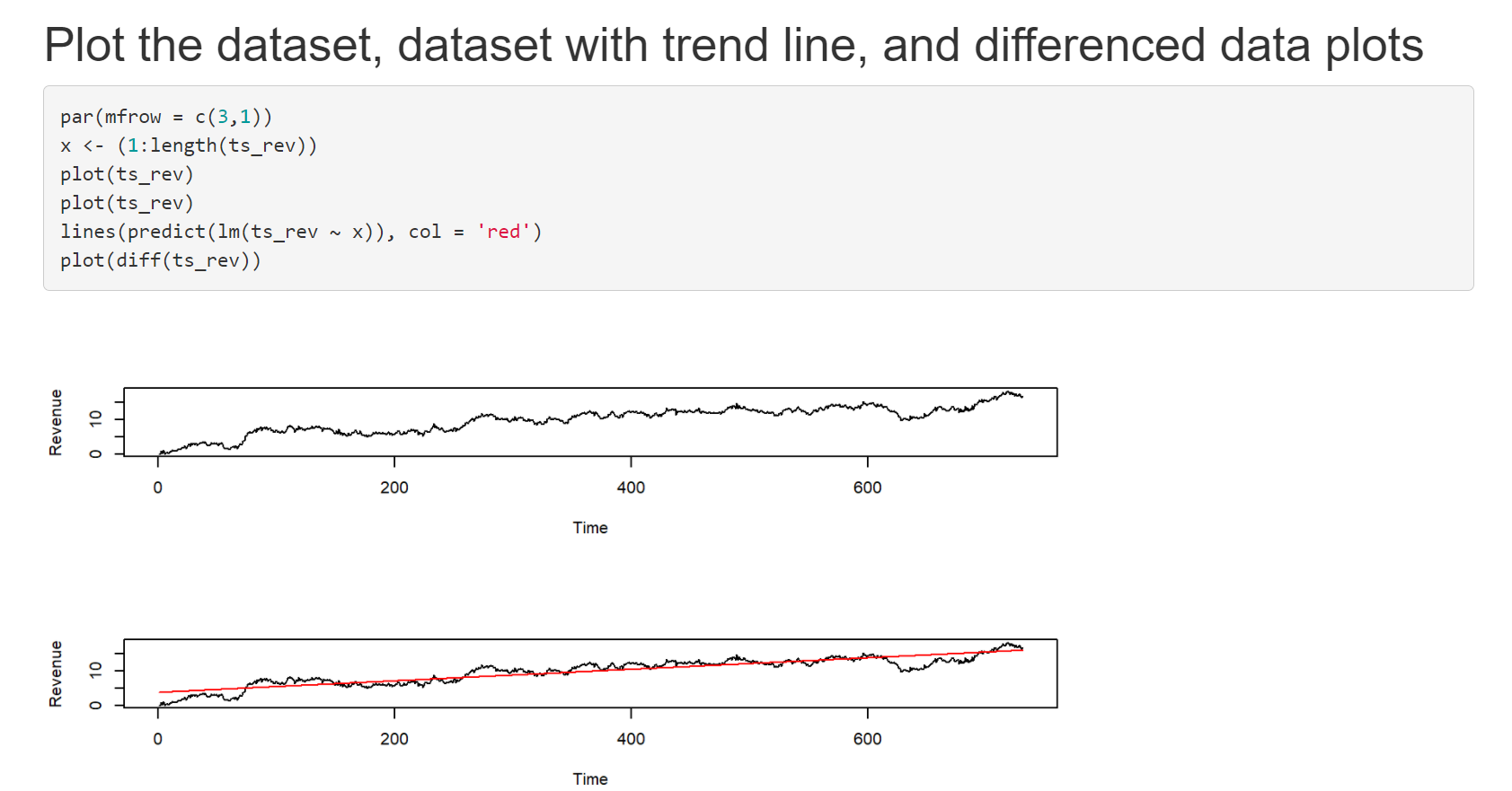
1. The data must be stationary. This means that the mean, variance, and autocorrelation are constant over time. This was tested by using the adfuller test (adf.test())
2. The data must be autocorrelated. I plotted and used the autocorrelation function to view the autocorrelation of our data. There is a high, constant amount of autocorrelation seen over each lag in our data. This can be viewed below:



1. The Time Series Model I am using in this analysis is ARIMA. ARIMA is only based on the first assumption of stationarity.

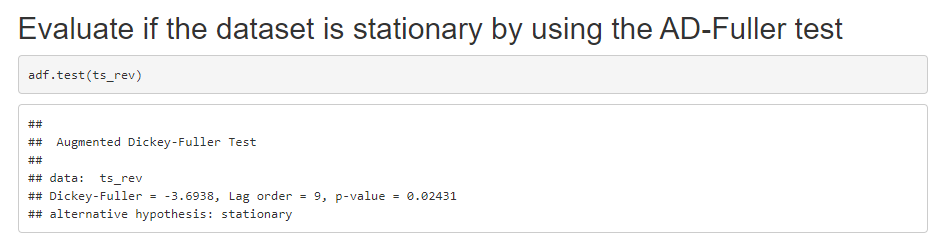
**III:  Data Preparation**

Please see below for a line graph visualization the realization of the time series:



Since there were no gaps in the two years of our data, I did not need to account for gaps. Each day over the two year period is accounted for in our data. I converted the raw data into a time series object by using as.ts() and setting it to a new variable (ts\_rev).

To evaluate the stationarity of our data. I used the Ad-Fuller test with the function adf.test(). The test favored the alternative hypothesis that the data was stationary with a statistically significant p-value of under 0.05. Please see below for the output:



Steps to prepare my datasets:

1. Load the raw data into a Dataframe in R with read.csv()
2. Convert the dataframe to a time series object by using as.ts()
3. Split the data by indexing. I had to index the first 671 days for the training set and used the entire dataset for the test dataset. This is because you cannot use a typical train/test split like you would with regression. It does not make sense in the context of a time series analysis to do the typical train/test split. It is better to build the model using less data, forecast the model for your desired time amount, and compare/test your model’s forecast to the actual test data.
4. I exported each of the datasets once I finished indexing by using write.csv()

Please see “ts\_train.csv” and “ts\_test.csv” for the split train/test copies of my cleaned dataset.

**Please note that train/test set splits the popular way with randomizing your training and testing sets is NOT beneficial nor necessary for a time series analysis.**

**IV:  Model Identification & Analysis**

Lack of Seasonality:

Please see “Task 1 – Lack of Seasonality.PNG” and “Task 1 – Decomposed Time Series.PNG” to see the lack of seasonality. The decompose() function throws an error when there is no seasonality present in our data. The Decomposed Time Series can show how there is no seasonal pattern to our data, but only that it trends upwards.

Trends:

Please see “Task 1 – Trends.PNG” to see the plotted time series data and the trend line added showing the direction of the data over time.

Auto Correlation Function:

Please see “Task 1 – Auto Correlation Function Viz.PNG” for the ACF and PACF outputs of our training data. There is continuous auto correlation seen. In the PACF, there is a cut-off at lags 1 & 2.

Spectral Density:

Please see “Task 1 – Spectral Density.PNG” for a periodogram of the spectral density. Spectral Density is used to show data stability. In the second graph, the data is unstable within the first 1%, but stabilizes.

Decomposed Time Series:

I used a Simple Moving Average to decompose and smooth out our time series. There is no seasonality present in the data and this can be seen in both “Task 1 – Lack of Seasonality.PNG” and “Task 1 – Decomposed Time Series.PNG”.

Residuals of Decomposed Series:

The Residuals can be seen in “Task 1 – Residuals 2-1-0 ARIMA.PNG”. These are the standardized residuals from our time series. It is part of the output from SARIMA() with our identified best parameters.

Identify ARIMA:

Please see “Task 1 – Identify ARIMA Model (2-1-0).PNG” for the code used and output given to find the best ARIMA model for our data. There is no seasonality with this data.

Forecast:

Please see “Task 1 – 60 Day Forecast (Forecast Lib).PNG” for a screenshot of the forecast and its graphed output. Predictions can be seen in “Task 1 – Prediction Intervals of Forecast.PNG”.

Outputs and Calculations:

Please see “Task-1-R-Markdown.HTML” for the outputs/calculations in my code. Also, please see the attached for various screenshots of graphs and outputs.

Code:

Please see “Task 1 – TSA.R” for a script of my entire code and “Task-1-R-Markdown.HTML” for the R Markdown version of my code.

**V:  Data Summary & Implications**

Selection of ARIMA Model:

Our final ARIMA model found by auto.arima() was (2,1,0). This means that p, or the lag of auto regression, is 2, d , or the differencing needed to stationarize the series, is 1, and q, or the number of moving average terms, is 0.

Prediction Interval:

The prediction intervals can be viewed at 80% and 95% confidence by using forecast() on our best ARIMA model found by auto.arima(). The 80% confidence forecasting interval is seen by the dark grey in “Task 1 – SARIMA Forecast 60 Days.PNG” and blue in the “Task 1 – 60 Day Forecast (Forecast Lib).PNG” screenshots I provided. The light gray areas on both .PNGs represent the 95% confidence interval. The 95% confidence interval is about (9.7, 20) at the end of the forecasted 60 days.

Justification of Forecast Length:

I chose to set the forecast for the next 60 days based on our research question and stakeholders’ needs. Our stakeholders want to evaluate the current prior two months of revenue progress with the focus of minimizing customer churn against the forecasted expectations of our all-time data. From there, stakeholders can decide whether they would like to continue their current methods to minimize customer churn or try other methods. 60 days is a great length of time to implement changes and see the effects of those changes.

Model Evaluation & Error Metric:

* AIC & BIC:
  + The auto.arima() function found the best ARIMA model with the lowest AIC and BIC values. AIC is a measure of goodness of fit of our model and higher values when compared to other estimated models are bad. BIC is the Bayesian Information Criteria which penalizes free parameters more strongly than AIC does. The lowest pairing of AIC and BIC can indicate the best model fit for our dataset. I also added the goal to keep both AIC and BIC under 910. Our model satisfies this goal as it produces an AIC of 886.17 and a BIC of 904.2.
* Ljung-Box Statistic:
  + In the Ljung-Box Statistic of the SARIMA output, all our p-values are above the blue dotted line (our threshold). Any p-values under our threshold would indicate a bad fit of our model. However, all p-values shown are above our threshold and indicate a good fit.
* RMSE:
  + The Residuals Root Mean Squared Error was 0.465563. This is in millions and is a low error rate compared to the range of our data.

60-Day Forecast against test dataset:

Please see “Task 1 – SARIMA Forecast 60 Days.PNG” and “Task 1 – 60 Day Forecast (Forecast Lib).PNG” for the 60-day forecast against the actual next 60 days from the test dataset. Also, please see “Task-1-R-Markdown.HTML” for the annotated code and output of my forecasts.

Recommendation:

Based on the forecasted 60 days presented by our model versus our actual 60 days of revenue data, I recommend that the stakeholders continue their current efforts to minimize customer churn. In most of the forecasted 60 days, our revenue levels were above the forecasted numbers. However, causality does not equal causation. A deeper dive will be needed to verify that the reduction in customer churn caused the increase in revenue.