Lora Milam

Western Governor’s University

D213 Advanced Data Analytics

11 August 2023

D213 Performance Assessment Task 1

**1 Introduction**

Lora Milam Masters Data Analytics (8/11/2023) Program Mentor:d213@wgu.edu

* 1. **Research Question**

This analysis will utilize time series modeling techniques to examine a revenue dataset obtained from a telecommunications company. The primary objective is to address the research question:

Is it possible to make reliable forecasts for a period of 120 days ahead to aid decision-makers in strategizing for that time period?

* 1. **Objective and Goals**

As mentioned previously, the primary objective is to create a forecast of 120 days of revenue.

Additional goals would include:

1. Determine whether the data is stationary; if not, make stationary.
2. Determine if there are any seasonal components.
3. Determine the best order for the ARIMA model.
4. Create ARIMA model utilizing training data.
5. Verify ARIMA model with test data.

**2 Method Justification**

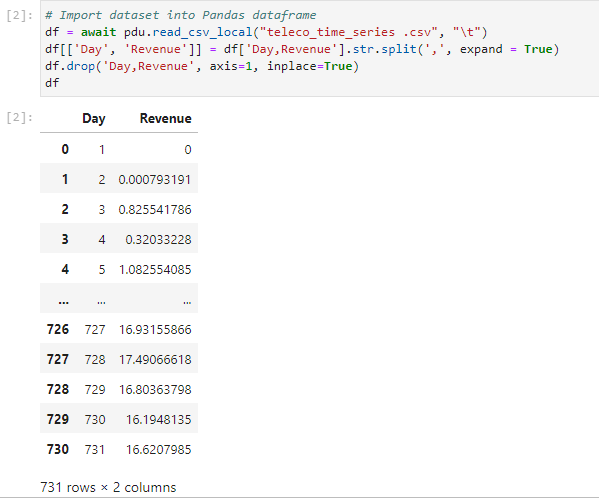
**2.1 Summary of Assumptions**

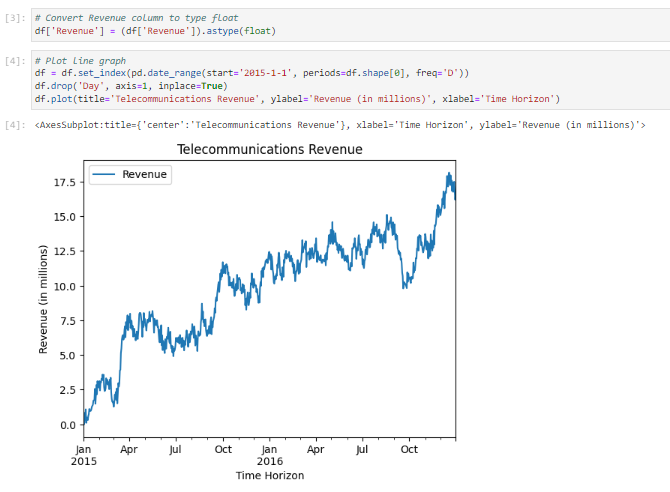
The time series model in this analysis is the ARIMA model. The ARIMA model assumes that the given data is stationary, having no trends or irregularities. The distribution of stationary data doesn’t change over time, and it’s mean, variance and autocorrelation are all constant. The ARIMA model also assumes the given data is autocorrelated. Autocorrelation states that past values can be prognostic of current and future values In this analysis, the time series model examines one set of variables over time and assumes the relationship to be univariate.

**3 Data Preparation**

**3.1 Line Graph Visualization**





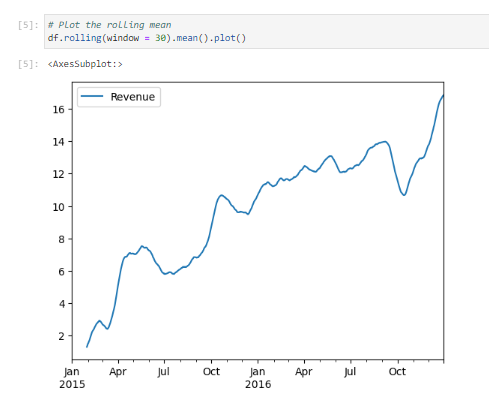


**3.2 Time Step Formatting**

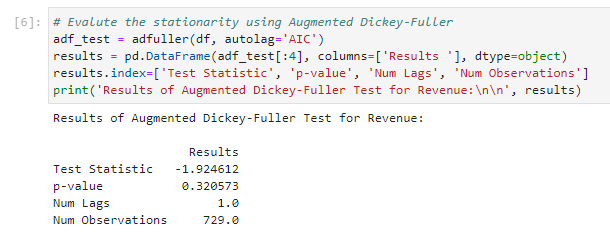
The time series of this dataset is revenue over 731 days. After examination of summaries, the data shows no null values. There are also no duplicates within the “Days” column. This indicates that there are no gaps or missing values for the days or revenue values.

**3.3 Stationarity**

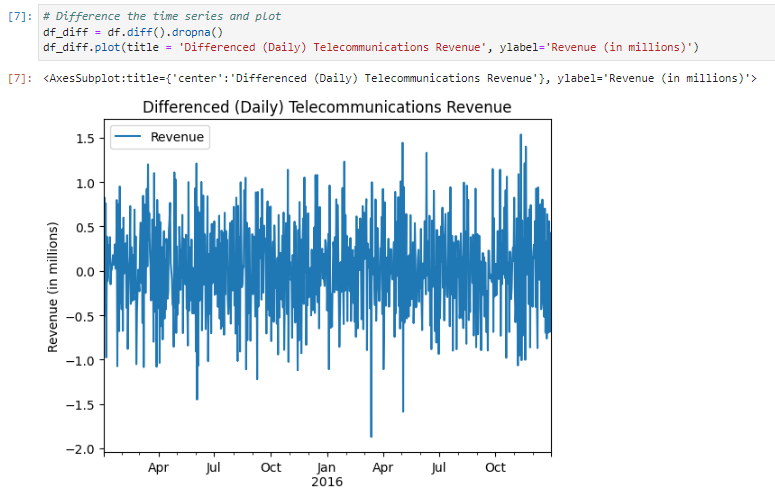
In order to assess stationarity, a plot of the rolling mean of the data was generated, as depicted below.



The graph shows that the data is not stationary. It can be further analyzed by utilizing the Augmented Dickey-Fuller test.



Generally if the p-value is 0.05 or lower, the result is trumpeted as significant, but if it is higher than 0.05, the result is non-significant and tends to be passed over in silence (Why the P-value is significant). The p-value for this data is greater than 0.05, at 0.320573. This further instantiates that the data is not stationary. To achieve stationarity, next the data will be differenced.



As scene in the previous plot, the differenced data is stationary. To analyze the data further, plot the rolling mean of the differenced data.

A screen shot of a graph

Description automatically generated

The differenced rolling mean plot shows stationarity. It can be further analyzed by utilizing the Augmented Dickey-Fuller test.

A screenshot of a computer program

Description automatically generated

The p-value for this data is less than 0.05, at 0. This means stationarity has been achieved.

**3.4 Steps to Prepare the Data**

The data preparation steps include:

1. Load libraries
2. Import dataset
3. Plot dataset
4. Plot dataset rolling mean
5. Evaluate dataset stationarity using Augmented Dickey-Fuller
6. Plot differenced time-series
7. Plot differenced rolling mean
8. Evaluate differenced dataset stationarity using Augmented Dickey-Fuller
9. Split differenced dataset in 80-20 train-test split

**3.5 Prepared Dataset**

The prepared dataset will be provided as ‘D213\_task1\_train\_clean.csv’ and ‘D213\_task1\_test\_clean.csv’.

A screenshot of a computer

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A white card with red and blue text

Description automatically generated

**4 Model Identification and Analysis**

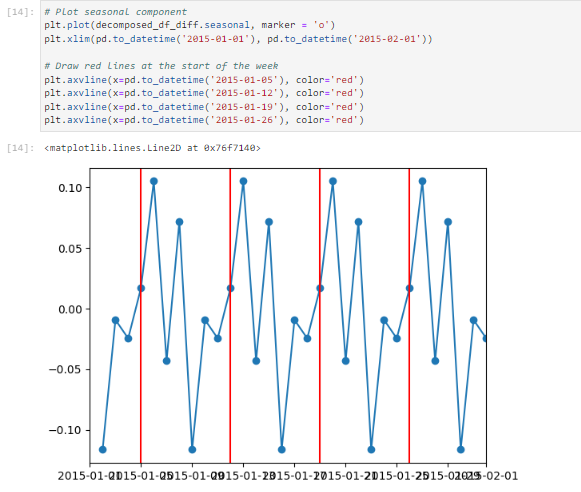
**4.1 Report Findings and Visualizations**

**4.1.1 Seasonality**

The below plots suggest no seasonality.

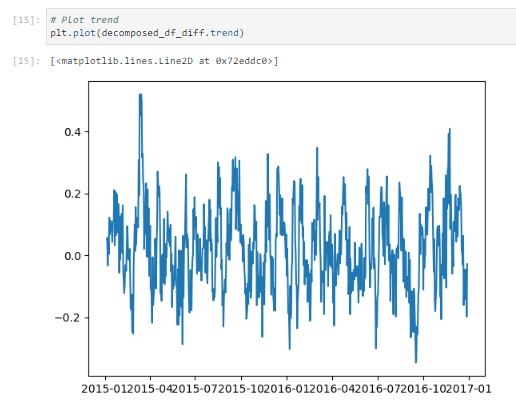
A screenshot of a bar code

Description automatically generated



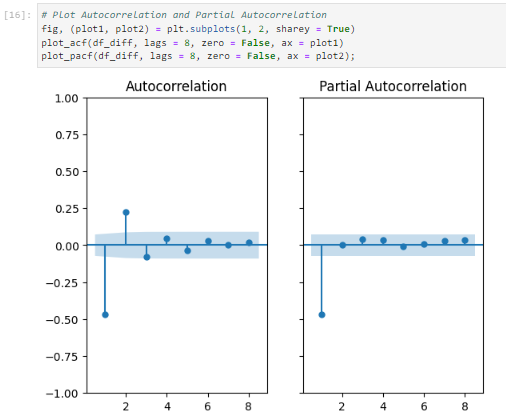
**4.1.2 Trends**

The plot below does not have any apparent trends.



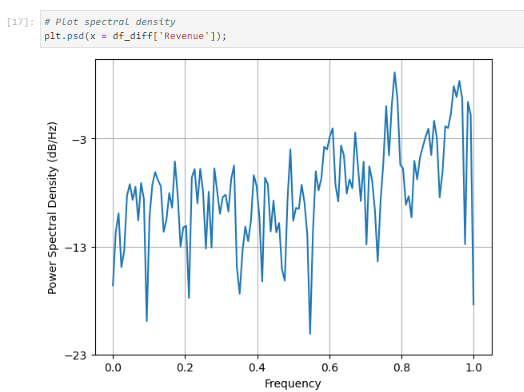
**4.1.3 Autocorrelation and Partial Autocorrelation**

ACF/PACF does not discern any obvious seasonality or trends.



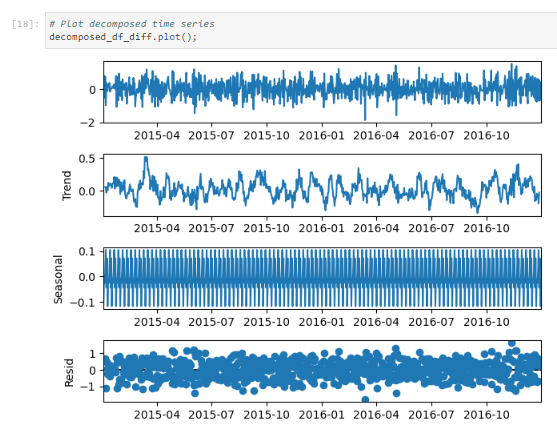
**4.1.4 Spectral Density**

Spectral density does not discern any obvious trends or seasonality.



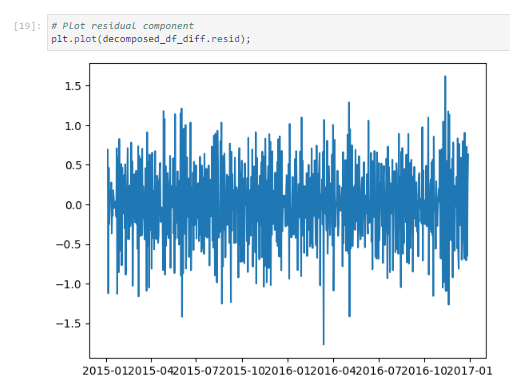
**4.1.5 Decomposed Time Series**

Decomposed time series does not discern any obvious trends or seasonality.

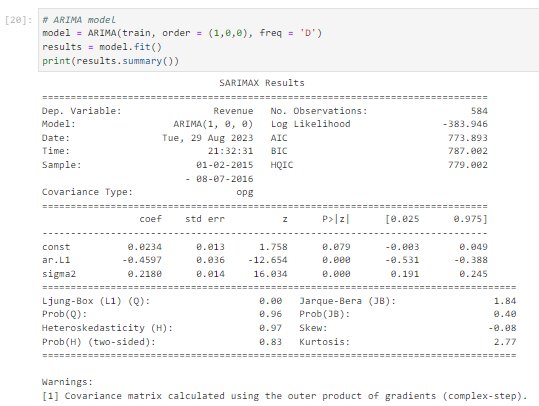


**4.1.6 Confirm Lack of Trends in Residuals of Decomposition**

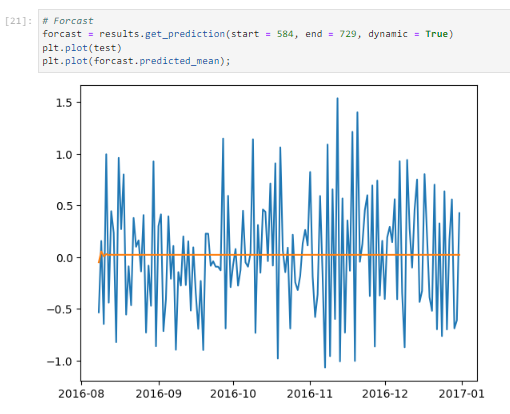
The residual of decomposition does not discern any obvious trends or seasonailty.

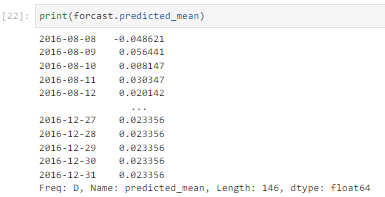


**4.2 Arima Model**



**4.3 Forecasting Using Arima Model**





A screenshot of a computer

Description automatically generated

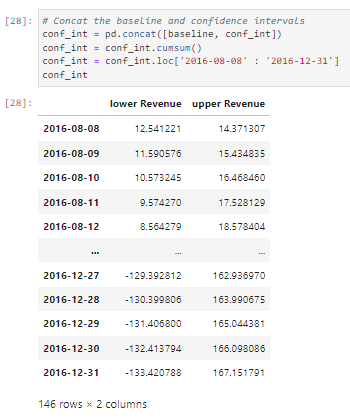


A screenshot of a computer screen

Description automatically generated

A screen shot of a computer

Description automatically generated



A screen shot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**4.4 Output and Calculations**

All outputs and calculations can be found in previous sections.

**4.5 Code**

All code can be found in previous sections.

**5 Data Summary and Implications**

**5.1 Results**

The selection of the ARIMA model was determined based on the ACF and PACF results, which have been provided once more below. The shaded area within the graphs represent statistically insignificant values, meaning those datapoints can be ignored. The ACF seems to taper off at 2, while the PACF cuts off at 1. According to the course material, it is indictive of an AR model order ‘p’ when the ACF tapers off and the PACF cuts off after lag ‘p’ which is why the dimensions (1,0,0) were chosen.

A screenshot of a computer screen

Description automatically generated

The final prediction interval of the forecast is 146 days, or 20% of the dataset. This is greater than the initial prediction interval of the research question, 120 days. The justification of the forecast length would be that typically when splitting a dataset into test/train sets, an 80/20 split is a reasonable choice.

The model can be evaluated by using the root mean squared error. The root mean squared error of this model is 2.473594. A lower rMSE is better and is indictive of a more accurate model. This measure would be a point of comparison against other models. The model’s performance can also be evaluated by the output of diagnostic\_plots(). For your convenience, both results have been provided again below.

A screenshot of a computer screen

Description automatically generated

**5.2 Annotated Visualization**

A screen shot of a graph

Description automatically generated

**5.3 Recommendations**

Having more details regarding the dataset, including its timeframe and additional entries, would be advantageous. Nevertheless, the outcomes of this analysis appear to have yielded a satisfactory model for the requested results.

**6 Reporting**

**6.1 Reporting**

The report is provided as “D213 Part 1 Performance Assessment Report Lora Milam.pdf”.

**7 Supporting Documentation**

**7.1 Sources**

Hirst, A. T. (2021, September 14). *Fragment: Loading data into pandas DataFrames in jupyerlite*. OUseful.Info, the blog... <https://blog.ouseful.info/2021/09/14/fragment-loading-data-into-pandas-dataframes-in-jupyerlite/>

*How to split one column into multiple columns in pandas DataFrame*. Saturn Cloud Blog. (2023, August 8). https://saturncloud.io/blog/how-to-split-one-column-into-multiple-columns-in-pandas-dataframe/#:~:text=Series.-,str.,using%20the%20double%20bracket%20notation.

Western Governors University. (n.d.). D213 Advanced Data Analytics. Salt Lake City.

Why the P-value is significant. Tidsskrift for Den norske legeforening. (2015, September 8). https://tidsskriftet.no/en/2015/09/why-p-value-significant-0#:~:text=The%20p%2Dvalue%20can%20be,be%20passed%20over%20in%20silence.