|  |
| --- |
| Western Governors University |
| Advanced Data Analytics |
| D213 TASK 1: TIME SERIES MODELING |

|  |
| --- |
| Allison Casey  1-4-2025 |

**Part I: Research Question**

**A1: RESEARCH QUESTION**

Are there trends impacting revenue that could be used to help predict future revenue?

**A2: OBJECTIVES OR GOALS**   
The goal of the analysis is to identify significant patterns in revenue to help predict future revenue.

**Part II: Method Justification**

**B: SUMMARY OF ASSUMPTIONS**

The first assumption of a time series model is stationarity which means that over time the properties of the time series are consistent which includes the mean, variance, and autocorrelation so the series doesn’t have a clear upwards or downwards trend. The second assumption is that data will not be highly correlated with itself over time. Autocorrelation represents the similarity between a time series and an older version of the time series.

**Part III: Data Preparation**

**C1: LINE GRAPH VISUALIZATION**

A graph showing the growth of the stock market

Description automatically generated

**C2: TIME STEP FORMATTING**

It was initially confirmed that there was no missing data or gaps based on there being no null data present as well as there being an adequate amount of data points for 2 years. The days for the sequence were then converted to a time series object and indexed appropriately.

A screenshot of a computer

Description automatically generated

**C3: STATIONARITY**

To check for stationarity the ADFuller test was run on the data. This produced a p-value of higher than .05 which indicates that the data is not stationary and needs to be differenced to make it stationary. Once this was done the p-value was brought below .05.

A screenshot of a computer

Description automatically generated

**C4: STEPS TO PREPARE THE DATA**

Explain the steps you used to prepare the data for analysis, including the training and test set split.

The submission explains the steps used to prepare the data for analysis. Each step is complete and includes the training and test set split and relates to preparing for time series modeling.

1. Data was imported into a pandas data frame
2. Explore the data and check for missing and null values to see if the data needs to be cleaned
   1. Confirmed there are no null or missing values to treat
   2. Reviewed descriptive statistics to further explore the data
3. Converted the data frame into a time series object so that it will be usable for the modeling techniques. Through the first step it was confirmed that there are no missing values or gaps, so the length of the sequence is complete.
4. Evaluated the data for stationarity using the ADFuller test. The data used to build the ARIMA model must be stationary so this needs to be checked ahead of time.
   1. The result of the test produced a p-value greater than .05 which indicated the data was not stationary, so the data was differenced to stabilize the mean of the time series. After differencing once and re-running the ADFuller test the p-value was acceptable to reject the null hypothesis and consider the data stationary.
5. Prepared data was split into training and test sets using a standard 80/20 split so that the model can be trained and evaluated.

See D213.ipynb for the code relating to each of the steps

**C5: PREPARED DATA SET**

train.csv

test.csv

**Part IV: Model Identification and Analysis**

**D1: REPORT FINDINGS AND VISUALIZATIONS**

1. There doesn’t appear to be seasonality as there are no discernable cycles that repeat over time at the same frequency.
   1. A blue line graph with numbers

      Description automatically generated
2. The overall trend of the time series is moving downward
   1. A graph showing a line of blue lines

      Description automatically generated with medium confidence
3. The ACF and PACF show relationships between observations and lag. The majority of these values are not statistically significant since they are so small lie within the blue region which shows that the data is stationary.
   1. A graph with blue dots and numbers

      Description automatically generatedA graph with a line and dots

      Description automatically generated with medium confidence
4. The spectral density plot’s spikes do not appear to follow any sort of pattern or reoccur which indicates that the time series data does not have seasonality or periodicity.
   1. A graph of a power spectrum

      Description automatically generated
5. The decomposed time series chart allows the components of trend, seasonality and residual to be compared and viewed all at once.
   1. A graph of blue lines

      Description automatically generated with medium confidence
6. The residual plot also shows no trends in the residuals
   1. A graph with numbers and a line

      Description automatically generated

**D2: ARIMA MODEL**

The auto\_arima function provides the optimal model to be 1,0,0 with no seasonality 0,0,0, and and no periodicity 0. Using this the model was then created using the training data. See D213\_Task1.ipynb for full code.

A black background with white text

Description automatically generated



A computer screen shot of text

Description automatically generated

**D3: FORECASTING USING ARIMA MODEL**

See D213\_Task1.ipynb for full code.

A screen shot of a computer

Description automatically generated

A computer screen with text on it

Description automatically generated

**D4: OUTPUT AND CALCULATIONS**

See D213\_Task1.ipynb for full code.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A blue and orange sound wave

Description automatically generatedA graph of a graph

Description automatically generated with medium confidence

A number on a black background

Description automatically generated

**D5: CODE**

See D213\_Task1.ipynb for full code.

**Part V: Data Summary and Implications**

**E1: RESULTS**

The submission accurately discusses the results of the data analysis and includes each of the 4 given points. The discussion aligns with the research question and the data analysis from part A.

The ARIMA model was used because it is useful for predicting future trends using time series data. The model’s p, d and q properties were selected using the auto\_arima function to get optimal values based on the AIC score which resulted in 1,0,0. The data set is in daily intervals for two years, so the model uses that to produce predictions at a daily interval. The prediction intervals chosen were first one to match the test data to be able to evaluate the model and the second interval that was chosen was one year because that is the maximum interval that can be predicted for this data (Elleh, 2025). The model was evaluated by first selecting the model with the best AIC then the error metrics calculated for the final model were MSE, RMSE, and MAE:

MAE: 0.4670346782998644

MSE: 0.32462756295272455

RMSE: 0.5697609700152552

These low values close to 0 indicate that the model is usable. As a result, this indicates that there are trends in the data that can be used to predict future revenue.

**E2: ANNOTATED VISUALIZATION**

A graph of a graph

Description automatically generated with medium confidence

**E3: RECOMMENDATION**

Recommend a course of action based on your results.

The submission recommends an appropriate course of action based on the results as they relate to the research question.

The time series data has good performance accuracy to forecast 90-180 days of future revenue. Using the model the company can forecast revenue to plan for necessary spending as well as upgrades and marketing campaigns. As more data is gathered over time the model can also be improved. Since churn likely has the greatest impacts overall on revenue the company should continue to attempt to mitigate this and they can also use the models to anticipate churn and further investigate the causes as well as other causes associated with the major dips in revenue.

**Part VI: Reporting**

**F: REPORTING**

D213\_Task1.html

**Sources**

Brownlee, Jason. “How to Use and Remove Trend Information from Time Series Data in Python.” *MachineLearningMastery.Com*, 14 Aug. 2020, machinelearningmastery.com/time-series-trends-in-python/#:~:text=A%20time%20series%20with%20a%20trend%20is%20called,the%20dataset%20is%20said%20to%20be%20trend%20stationary. Accessed 21 Jan. 2025.

Elleh, Festus. “Advanced Data Analytics - Task 1.” D213 Task 1 Cohort Webinar. 2025

“Forecasting Time Series with Decomposition.” *PyQuant News*, www.pyquantnews.com/the-pyquant-newsletter/forecasting-time-series-with-decomposition. Accessed 21 Jan. 2025.

“Periodogram#.” *Periodogram - SciPy v1.15.1 Manual*, docs.scipy.org/doc/scipy/reference/generated/scipy.signal.periodogram.html. Accessed 21 Jan. 2025.

Verma, Yugesh. “Quick Way to Find P, D and Q Values for Arima.” *Analytics India Magazine*, 30 Dec. 2024, analyticsindiamag.com/quick-way-to-find-p-d-and-q-values-for-arima/. Accessed 21 Jan. 2025.