For Task 3, I returned to our medical dataset and business problem. For this project, the executives wanted to consider the revenue from prior years to help us understand the impact of readmission at the current time. Knowing our expected revenue will help illustrate how these readmission penalties will impact our profitability. My research question for this real-world organizational situation is, “Where is our forecast projecting us to go based on historical data?”. This time series analysis aims to understand our projected revenue so the stakeholders can understand the impact of these readmission penalties and how we should move forward.

There are assumptions to be made when creating a time series model. First, the autocorrelated data assumption is that the data is linear and dependent on the previous value (*What Are the Assumptions of ARIMA/Box-Jenkins Modeling for Forecasting Time Series?*, n.d.). What this means is that ARIMA modeling further assumes our data is stationary. Stationarity is that the mean and sigma are constant and the series of data has no seasonality (ritvikmath, 2019). We will need to use our ADF test to ensure stationarity. If we do not have stationarity, differencing is required, so we can ensure reliability in our model (Nóbrega, 2024). Another assumption is that there are no level shifts (*What Are the Assumptions of ARIMA/Box-Jenkins Modeling for Forecasting Time Series?*, n.d.), meaning the forecast will be demonstrated with a straight line. We should use confidence intervals to demonstrate that the data can land within the shadowed area (within 95%) while following the trend from our forecast. This ensures that while our forecast shows us the trend, this line can go up and down within its path (Jones, 2025).

After understanding these assumptions, I moved into the data-cleaning process. It first started by posting the line graph of the visualization of the historical dataset as seen below, so I could understand where we were starting. From there it was important to use time step formatting. From what I could tell from the data dictionary, we had two years of data, and each data point was separated by day. The gap we had was that we did not have months or years accessible. Understanding this and the fact that the last year mentioned in the data dictionary was 2015, I thought an educated assumption would be this data started in 2014 and ended at the end of 2015. This is a gap in measurement, but we can easily update it with the code I used from the Medium resource (Nóbrega, 2024).

A graph showing a line

AI-generated content may be incorrect.

Moving forward in the data-cleaning process, it was important to evaluate stationarity. When plotting the autocorrelation (ACF) plot, we could see there were no peaks or spikes at regular intervals, demonstrating no seasonality (GeeksforGeeks, 2025). The gradual decline suggests a long-term dependency and that the data is non-stationary. The partial (PACF) shows the only significant lag is 1, but we must not move forward yet, until we have differenced our data (DataMites, 2024).

A graph of a graph of a graph

AI-generated content may be incorrect.

After differencing the data, we had better results as seen below. We can see from the ACF plot that our autocorrelation lag is 2 (MA(2)) (Monigatti, 2024). From the ADF statistic, our p-value is significant, under 0.05. With these findings, I will move forward with the differenced dataframe as it is stationary (ritvikmath, 2020). I will take this dataframe and split it train/test 80/20. From there, I can create my ARIMA using the differenced dataframe, as I have done my differencing manually, and set (p, d, q) = (0, 0, 2) (Sik Flow Analytics, 2019).



A graph of a graph of a graph

AI-generated content may be incorrect.

**Annotated Findings**

**A graph showing a graph of a company

AI-generated content may be incorrect.**

The annotated findings are printed above in my forecast. As I said before, I took the educated assumption on the date as the dictionary referenced 2015 being the last year. We can see the positive trend this forecast demonstrates over the next year. The autocorrelation function post-differencing and creating my model can be seen below. The results demonstrate no significant correlations, demonstrating that the model has captured significant autocorrelations. The spectral density of the differenced model is visualized below, and we can see the time series by frequency. We can interpret that there is not any significant seasonality as the frequency goes down quickly and levels out (Tom, 2024). Now I have decomposed my time series prior to differencing it and I have it pictured below, as well. The trend here demonstrates what we can also see in the forecast which is that positive trend. The dataframe before differencing, did show some variation in the seasonality. While it shows short-term fluctuations caused by seasons, this changed after differencing. The residuals did provide mostly a straight line and demonstrated outside of trend and seasonality, that the variability mostly stays the same and confirmed lack of trends (GeeksforGeeks, 2023).

A graph with blue dots

AI-generated content may be incorrect.

A graph of a power spectrum

AI-generated content may be incorrect.

A graph of growth and revenue

AI-generated content may be incorrect.

As mentioned, I identified the ARIMA model for observed trends and seasonality. I manually differenced the data and the results can be seen below for the latest ACF and PACF. On further analysis, I decided to use this after the ADF statistic showed this was a significant result with stationary data. The moving average was 2. I performed the differencing manually, so I set it to d=0 as it didn’t need to be differenced again. Lastly, the PCAF was insignificant, and the model performed better with p = 0. My model was set as (p,d,q)=(0,0,2) with a manual differencing.

A graph of a graph of a graph

AI-generated content may be incorrect.

A graph with blue and orange lines

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

I selected to use a regular ARIMA model. Before creating the model, I differenced the data, so when creating the forecast as seen above, I had to inverse the data back. The prediction interval I used in the forecast was the 95% confidence interval so that we could cover 95% of future responses. I chose a year's forecast length due to having two prior years. The stakeholders wanted to have a year's worth of data, and with two times the historical data to the forecast, this was perfect.

A graph showing a line graph

AI-generated content may be incorrect.



I was happy with the performance of my model. As seen above, when forecasting over the test data, the forecast fits the test data with 95% confidence and follows the upward trend of the historical data. The confidence interval area has room for improvement, but this can be expected with ARIMA models with complex historical data. The MAE showed a low magnitude of errors at 0.375104 and RMSE is 0.4876186, demonstrating only a low magnitude of errors for the predictions (GeekforGeeks, 2024).

My analysis has answered our business question. We now know this next year we are forecasted to make more than $20 million, but less than $25 million. Now that we have this information, I suggest further analysis to project how much the readmission penalties will hurt our projected revenue. With profitability in mind, it is important to compare this cost to others and ensure it is significant. When we know it is significant for us, we will know it is one of the costs we should focus on stemming.

A graph showing a graph of a company

AI-generated content may be incorrect.

References:

DataMites. (2023, March 15). *Introduction to ACF and PACF | Uses of ACF and PACF plots | Time Series Forecasting* [Video]. YouTube. <https://www.youtube.com/watch?v=fLWhySbphxM>

GeeksforGeeks. (2025, January 22). *Time Series Analysis & Visualization in Python.* GeeksforGeeks. <https://www.geeksforgeeks.org/time-series-data-visualization-in-python/>

GeeksforGeeks. (2024b, August 13). *Time series analysis and forecasting*. GeeksforGeeks. <https://www.geeksforgeeks.org/time-series-analysis-and-forecasting/>

GeeksforGeeks. (2023, October 20). *Time Series Decomposition techniques*. GeeksforGeeks. <https://www.geeksforgeeks.org/time-series-decomposition-techniques/>

Jones, K. (2025, February 26). *Confidence Intervals for Time Series Forecasts with Python*. Medium. <https://medium.com/@kylejones_47003/confidence-intervals-for-time-series-forecasts-95d2d3f81415>

Monigatti, L. (2024, October 21). *Interpreting ACF and PACF plots for time series Forecasting | TDS Archive. Medium*. <https://medium.com/towards-data-science/interpreting-acf-and-pacf-plots-for-time-series-forecasting-af0d6db4061c>

Nóbrega, S. (2024, November 18). *5 Must-Know Techniques for Mastering Time-Series Analysis.* Medium. <https://medium.com/towards-data-science/5-must-know-techniques-for-mastering-time-series-analysis-a23ccf4d053a>

ritvikmath. (2019, July 9). *Time series Talk : Stationarity* [Video]. YouTube. <https://www.youtube.com/watch?v=oY-j2Wof51c>

ritvikmath. (2020, September 7). *Time Series Talk : Augmented Dickey Fuller Test + Code* [Video]. YouTube. <https://www.youtube.com/watch?v=1opjnegd_hA>

Sik Flow Analytics. (2019, October 25). *ARIMA Predictions when Differencing* [Video]. YouTube. <https://www.youtube.com/watch?v=-71NKFfNQDA>

Tom. (2024, November 16). *SciPY for Signal Processing - TomTalksPython - Medium*. Medium. <https://medium.com/tomtalkspython/scipy-for-signal-processing-2fa2c8a290c4>

*What are the assumptions of ARIMA/Box-Jenkins modeling for forecasting time series?* (n.d.). Cross Validated. <https://stats.stackexchange.com/questions/77374/what-are-the-assumptions-of-arima-box-jenkins-modeling-for-forecasting-time-seri>