Part I: Research Question

A.1. One question relevant to a real world organization that I can answer using time series modeling is “With what accuracy can we predict future hospital revenue?”

A.2. The objectives of this analysis are

* To prepare the data
* To predict revenue for the hospital for the chosen amount of time (249 days)
* To determine the accuracy of the predictions

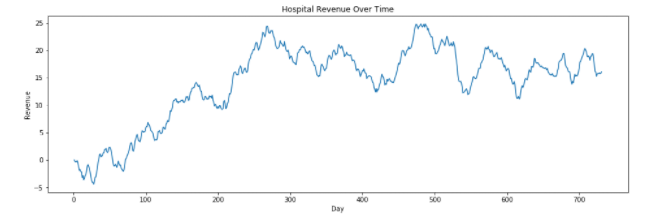
Part II: Method Justification

B. The assumptions of a time series model are as follows:

* The data are stationary (the mean, variance, and autocorrelation do not change over time (Stationarity).
* Autocorrelation does not change over time but decreases with lags.
* Future trends will be similar to those in the past.
* Data are affected by three components: trend, seasonality, and irregularity (Rajbhoj).

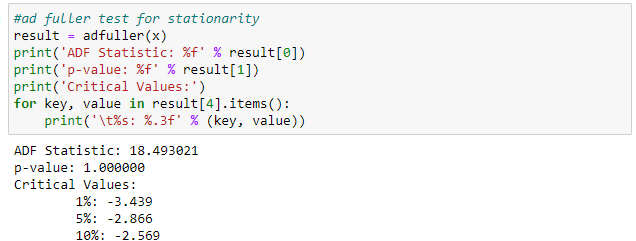
Part III: Data Preparation

C.1. A line graph visualizing the realization of the time series is displayed below.



C.2. The time step formatting for the realization is in days. The length of the sequence is 731 days. There are no gaps in measurement, assuming the data was collected on consecutive days.

C.3. To evaluate the stationarity of the time series I ran the Augmented Dickey-Fuller test. Due to the fact that the p-value is 1.0 and the test statistic is above the critical value, I conclude that the data are not stationary.



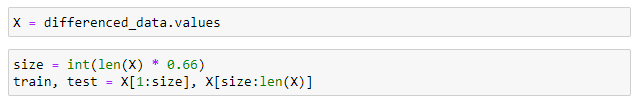
C.4. Because the data was not stationarity, I took the first difference of the data and re-ran the Ad-Fuller test.



With that, the ADF statistic came back below the critical value, showing that the data is now stationary.



Following that, I created the variable size to divide the set of differenced data into training and testing sets, and then created those sets, using 66% of the data for training and 34% for testing.

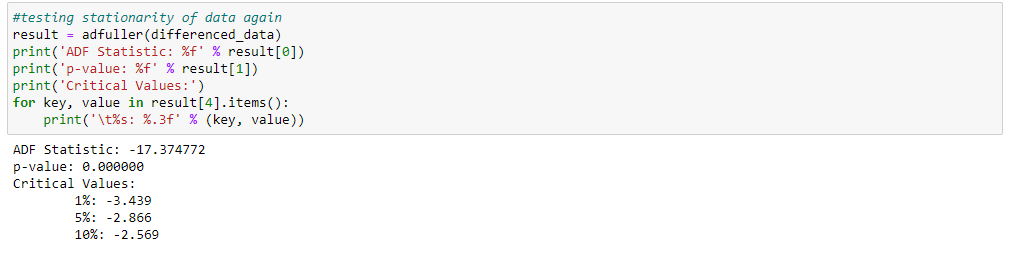


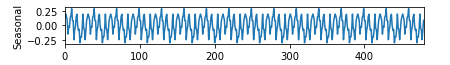
C.5. A cleaned copy of this dataset is included in this submission in the file ‘time\_series\_data.csv’.

Part IV: Model Identification and Analysis

D.1. The findings from this analysis are that we can use ARIMA to predict future hospital revenue with over 95% accuracy.

* The line plot of realization and the ADF test indicated initially that there was seasonality in the given data set. Upon discovering this we calculated the 1st difference for the model and removed the seasonality, as indicated by the post difference ADF test and the consistent pattern in the seasonality plot.

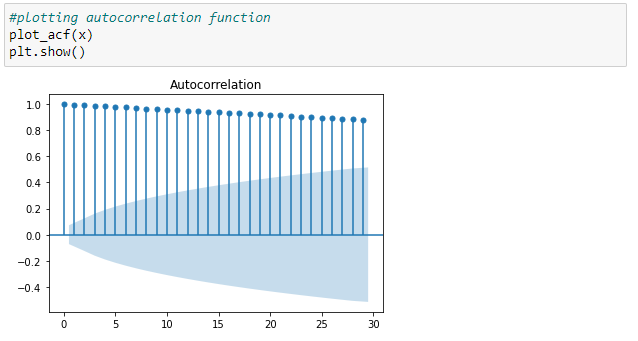




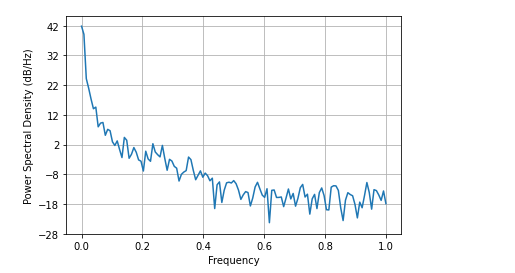
* The trend for revenue is a positive one, with some fluctuation present. There are valleys and troughs with a strong upward trend at the end of the series.



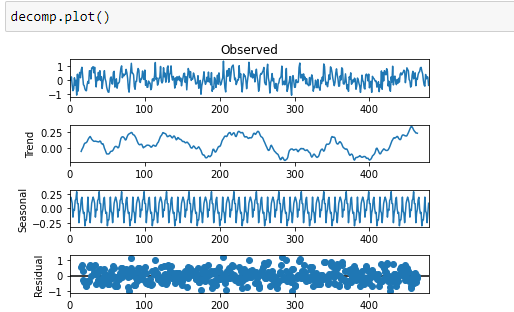
* The autocorrelation plot for the testing data is pictured below. At 30 lags, the ACF is still over 80%. Because the ACF is significantly lower than at 0 lags, and 30 is also what we’re using for the period, we will use 30 time lags for the ARIMA model.



* The power spectral density for the data is plotted below. We have significantly more power at the 0.0+ frequency, with the largest decrease in power happening before the 0.2 mark.

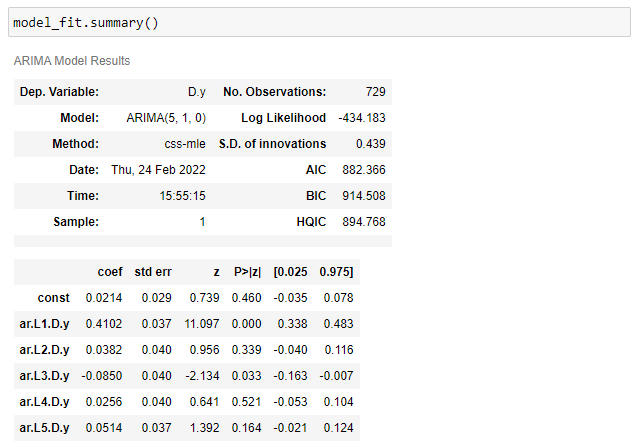


* The complete decomposed time series plot is included below. The first plot is of the original data. X is the number of day in the series and Y is revenue in millions of dollars. The trend line is a lot smoother than our line plot of realization, and still shows a strong upward trend. The seasonality as discussed above has been removed as shown by the consistent pattern and lack of variation in the plot. The residuals are well distributed as well.

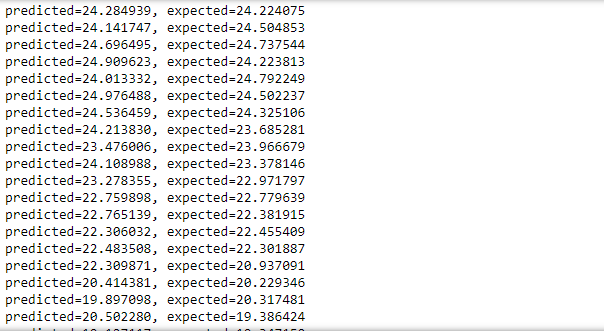


* With this plot, we can confirm there isn’t a trend in the residuals. They are random, indicating a lack of bias in the adjusted data.

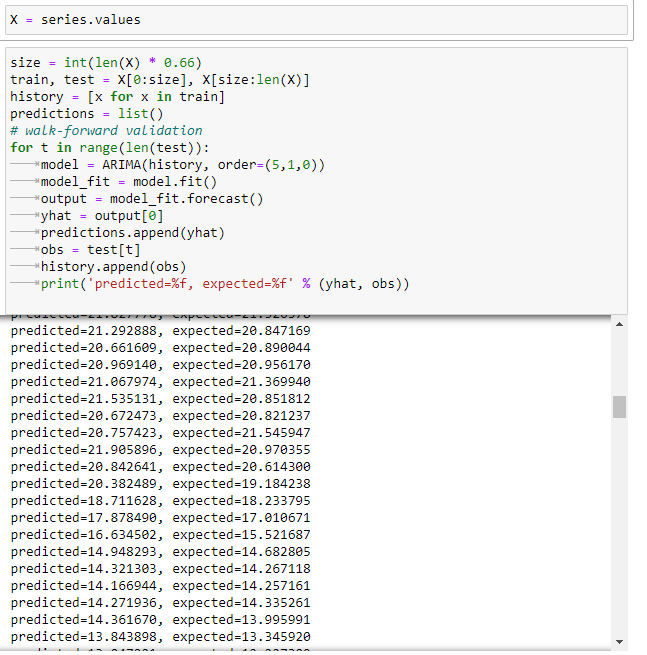
D.2. The ARIMA model I’ve chosen that takes into account the observed trend is pictured below. It uses parameters 5, 1, and 0 for the number of time lags, the degree of differencing, and the order of the moving average, respectively. The output is pictured below.

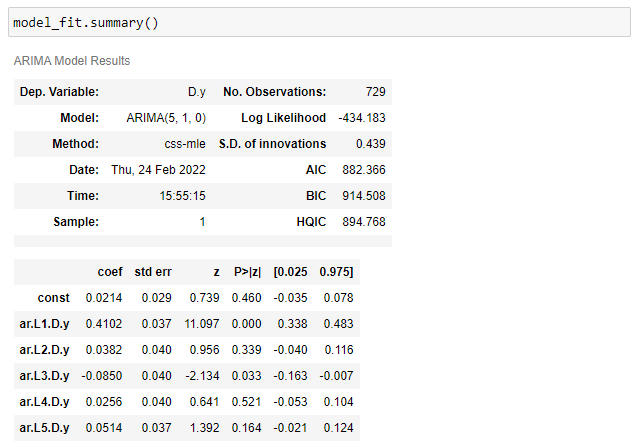


D.3. The beginning of the forecast using the model is pictured below. The complete list of predictions is included in the file predictions.csv.



D.4. The calculations and output for the analysis are pictured below.





D.5. The code used to support the implementation of the model is included in the .ipynb file of this submission.

Part V: Data Summary and Implications

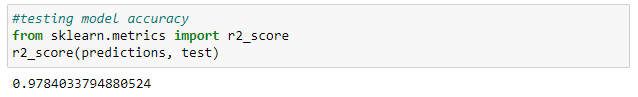
E.1. In an ARIMA model there are three parameters.

p = the number of time lags

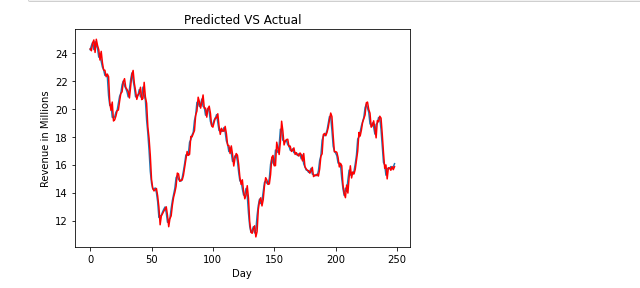
d = the degree of differencing

q = the order of the moving average

* The ARIMA model that takes into account the observed trend and seasonality of the time series data is using parameters (5,1,1). The number of time lags used here is 5. I chose to use a small value because although they do decrease autocorrelation slightly, they do not have an affect on the overall accuracy of the model. Increasing time lags enough to significantly decrease autocorrelation causes a large increase in the time it takes to use the program. For these reasons, I used p = 5 in my model. The degree of differencing (d) is 1. I found this parameter by using the ndiffs package from pmdarima. Finally, for order of the moving average I used q = 1 which is standard for stationary data. After choosing my parameters and calculating accuracy, I got a score of about 0.9782. This is above my goal of p = 0.05 or 95%, so I selected this model.
* The forecast is 249 predictions long.
* This was the length of the testing data set that was created with 34% of the provided data.
* The model evaluation and error metric for this project was the r2 score, for which we received a value of 0.9784033794880524.



E.2. An annotated visualization of the predictions compared to the data is pictured below. The differences are visible, but very minute, confirming our accuracy of about 97%.



E.3. Based on my results, I recommend using this model to predict future hospital revenue. It can be reliably used to make business and staffing decisions. Our accuracy is about 97% and should continue to rise with more data collected every day.

Part VI: Reporting

F. My report was created in Jupyter Notebook and is in the .ipynb file of this submission. A copy in html format was submitted as well.

G. Sources of third party code

Bonaros, B. (2021, April 22). *Time series decomposition in Python*. Medium. Retrieved January 10, 2022, from https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2

Brownlee, J. (2020, August 14). *A gentle introduction to autocorrelation and partial autocorrelation*. Machine Learning Mastery. Retrieved January 10, 2022, from https://machinelearningmastery.com/gentle-introduction-autocorrelation-partial-autocorrelation/

Brownlee, J. (2020, December 9). *How to create an Arima model for time series forecasting in Python*. Machine Learning Mastery. Retrieved February 25, 2022, from https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

khuyentran1476. (2021, January 7). *Take the difference between rows within a column in pandas*. Data Science Simplified. Retrieved February 17, 2022, from https://mathdatasimplified.com/2021/01/07/dataframe-diff-and-dataframe-shift-take-the-difference-between-rows-within-a-column-in-pandas/

H. Works cited

Rajbhoj, A. (n.d.). *ARIMA Simplified*. Towards Data Science. Retrieved January 25, 2022, from https://towardsdatascience.com/arima-simplified-b63315f27cbc

*Stationarity*. NIST. (n.d.). Retrieved October 1, 2021, from https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm#:~:text=A%20common%20assumption%20in%20many,do%20not%20change%20over%20time.&text=If%20the%20data%20contain%20a,the%20residuals%20from%20that%20fit.