Performance Assessment: Time Series Modeling

Gabriela Howell

Master of Science Data Analytics, Western Governors University

D213 – Time Series Modeling

Professor Elleh

September 20, 2024

Part I: Research Question

A. Describe the purpose of this data analysis by doing the following:

1. Summarize one research question that is relevant to a real-world organizational situation captured in the selected data set and that you will answer using time series modeling techniques.

The research question addressed in this report is: " Can ARIMA models accurately forecast daily hospital revenues based on past data, and how well do the predictions match the actual revenues?"

2. Define the objectives or goals of the data analysis. Ensure your objectives or goals are reasonable within the scope of the scenario and are represented in the available data.

This analysis aims to create a reliable ARIMA model to predict daily hospital revenues. By comparing predicted and actual revenues, we can check the model's accuracy and help hospital management make better financial decisions using historical revenue data and time series techniques.

Part II: Method Justification

B. Summarize the assumptions of a time series model including stationarity and autocorrelated data.

According to Hyndman & Athanasopoulos (2018), stationarity is key in many time series models, meaning that statistical properties like mean and variance remain constant over time.

This consistency is essential for reliable forecasting, as it ensures the time series remains predictable without significant trends or changes in variability. Techniques such as differencing are often recommended to achieve stationarity.

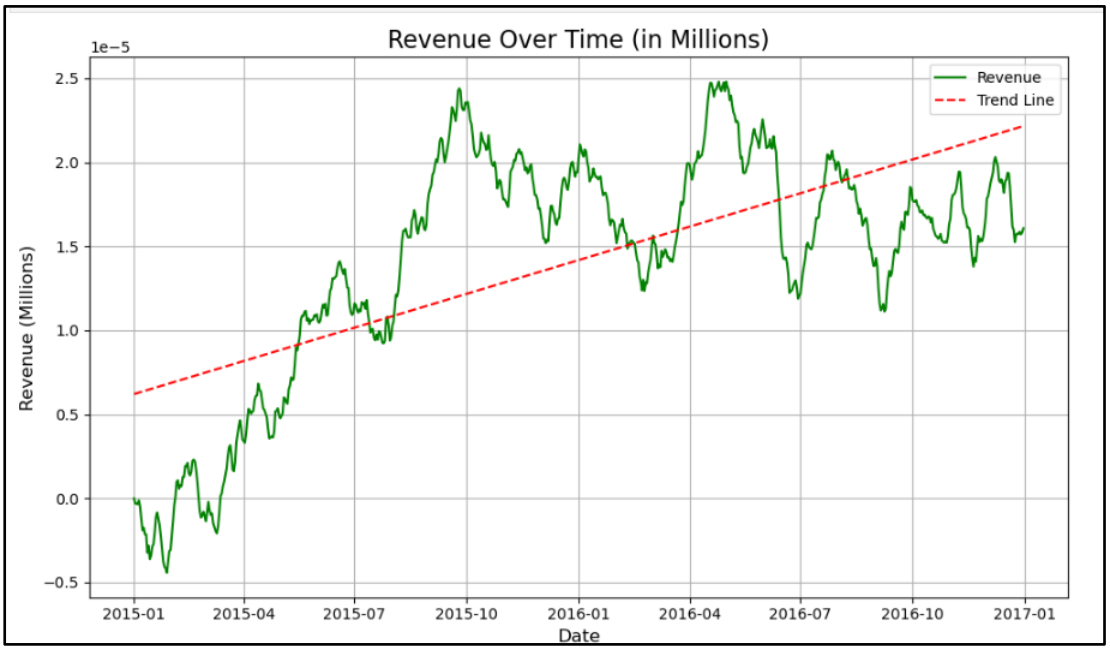
Similarly, Educative (n.d.) explains that autocorrelation shows how current time series values relate to past values. This concept is important in time series modeling, particularly in models like ARIMA, which utilize these relationships at different time lags to make predictions. Positive autocorrelation indicates that high values tend to be followed by high values, while negative autocorrelation suggests that high values are likely to be followed by low values.

Part III: Data Preparation

C. Summarize the data cleaning process by doing the following:

1. Provide a line graph visualizing the realization of the time series.

Here is my line graph:



2. Describe the time step formatting of the realization, including *any* gaps in measurement and the length of the sequence.

The time series data starts in January 2015 and includes daily observations for 731 consecutive days, ending in January 2017. Each day is consistently recorded, providing a reliable dataset for analysis.

I found no missing dates or gaps during data cleaning, ensuring a complete and accurate time series. This consistency means no interpolation, or forward-filling techniques were needed.

3. Evaluate the stationarity of the time series.

To determine if the time series was stationary, I performed an Augmented Dickey-Fuller (ADF) test on the differenced revenue data. This test checks for a unit root to assess stationarity. The results showed an ADF statistic of -17.37 and a p-value of 5.11e-30, allowing us to reject the hypothesis that the series is non-stationary. In simple terms, the original series was non-stationary, but it became stationary after differencing, which is crucial for accurate ARIMA modeling and forecasting.

The pattern indicates an upward trend in daily revenue. To verify this, I added a trend line to the plot. The dashed red line points upward, confirming the increasing trend. The data is evidently non-stationary.

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

To prepare the data, I first cleaned it by addressing any missing values or outliers. Then, I formatted the dates appropriately. Finally, I made the data stationary by applying differencing, making it suitable for ARIMA modeling.

After preprocessing, I split the data into training and test sets. About 80% of the data was used to train the ARIMA model, and the remaining 20% was used for testing. This split helped ensure the model could generalize well and produce accurate forecasts.

5. Provide a copy of the cleaned data set.

The cleaned dataset is saved as, ‘task1_train_clean.csv’ and ‘task1_test_clean.csv’.

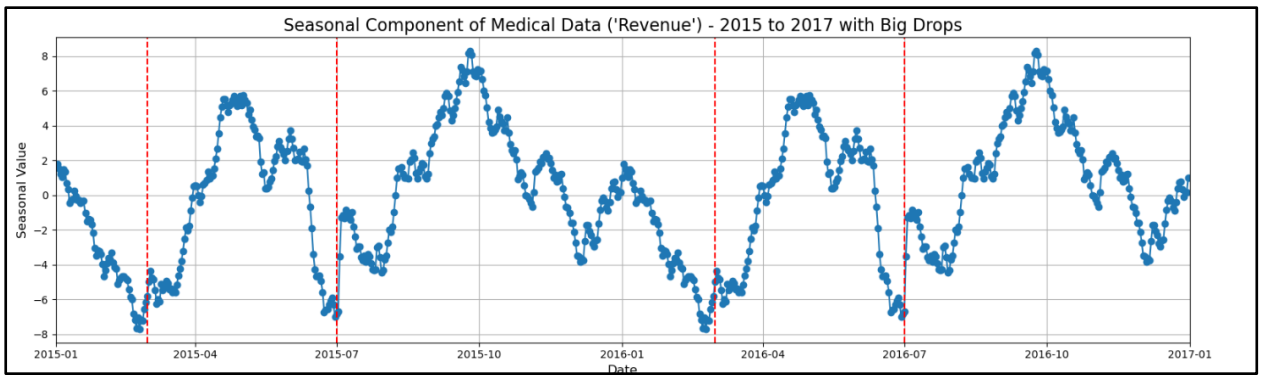
Part IV: Model Identification and Analysis

D. Analyze the time series data set by doing the following:

1. Report the annotated findings with visualizations of your data analysis, including the following elements:

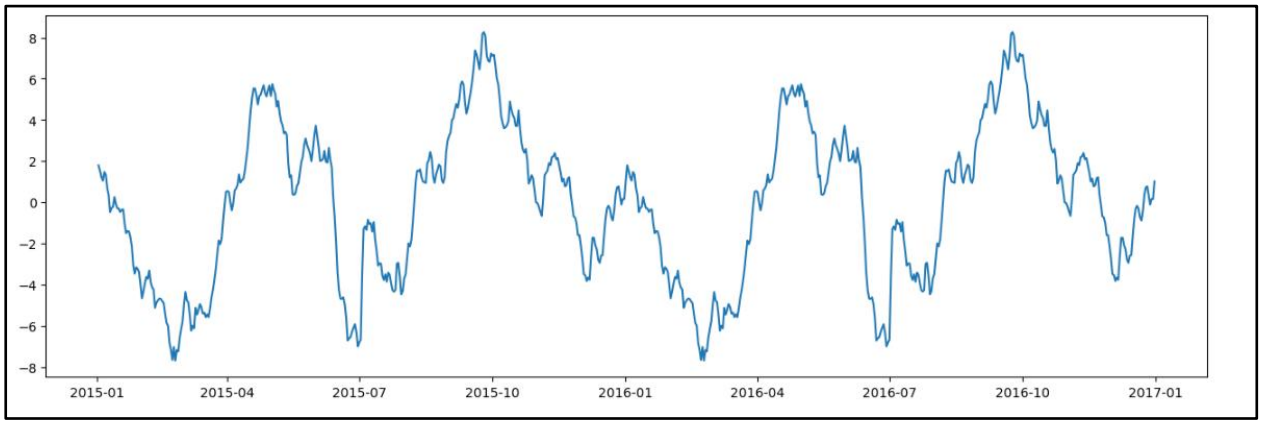
- **the presence or lack of a seasonal component**

As seen in the figure below you can indicate there is seasonaility because every year in march and around July there is a huge revenue drop as seen in 2015 and 2016. Most likely will happen again in 2017.



- **trends**

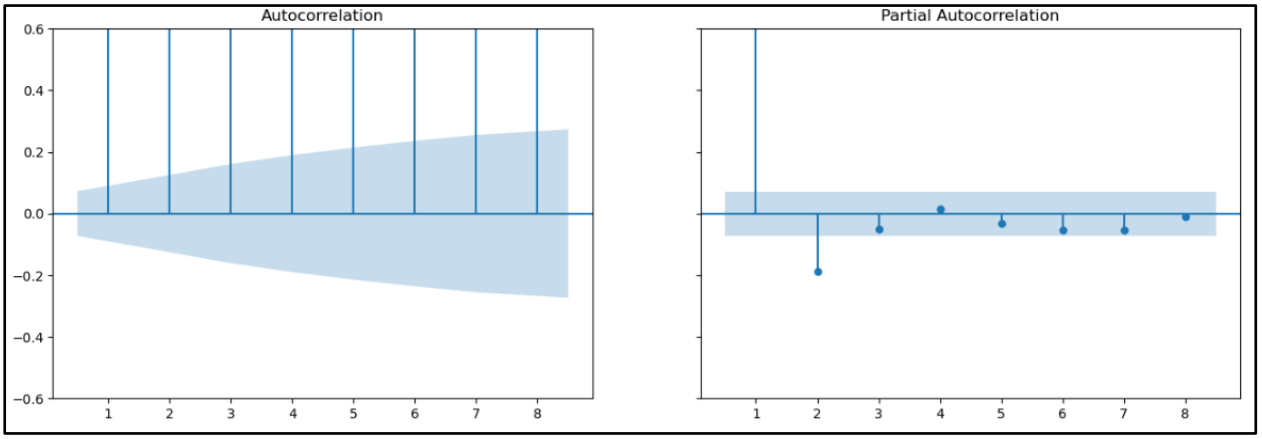
As seen below in the figure you can see the trnds to be continuous ups and doews. Starting with a fall in the begiing of the year and a large increase around arpil to than fall again in July to rise back to a high in October until around March a large drop again. To continuous through 2015 to moving forward to 2016.



- **the autocorrelation function**

This is the autocorrelation and partial autocorrelationa s you can see there the auto correlation. From the plot, we can observe that the autocorrelation is relatively high for the first few lags and starts to decrease gradually. This suggests that there is a notable correlation with the previous values, indicating that past values have a substantial effect on the current values, especially in the short term. The gradual decline also implies that the time series might not be stationary and could

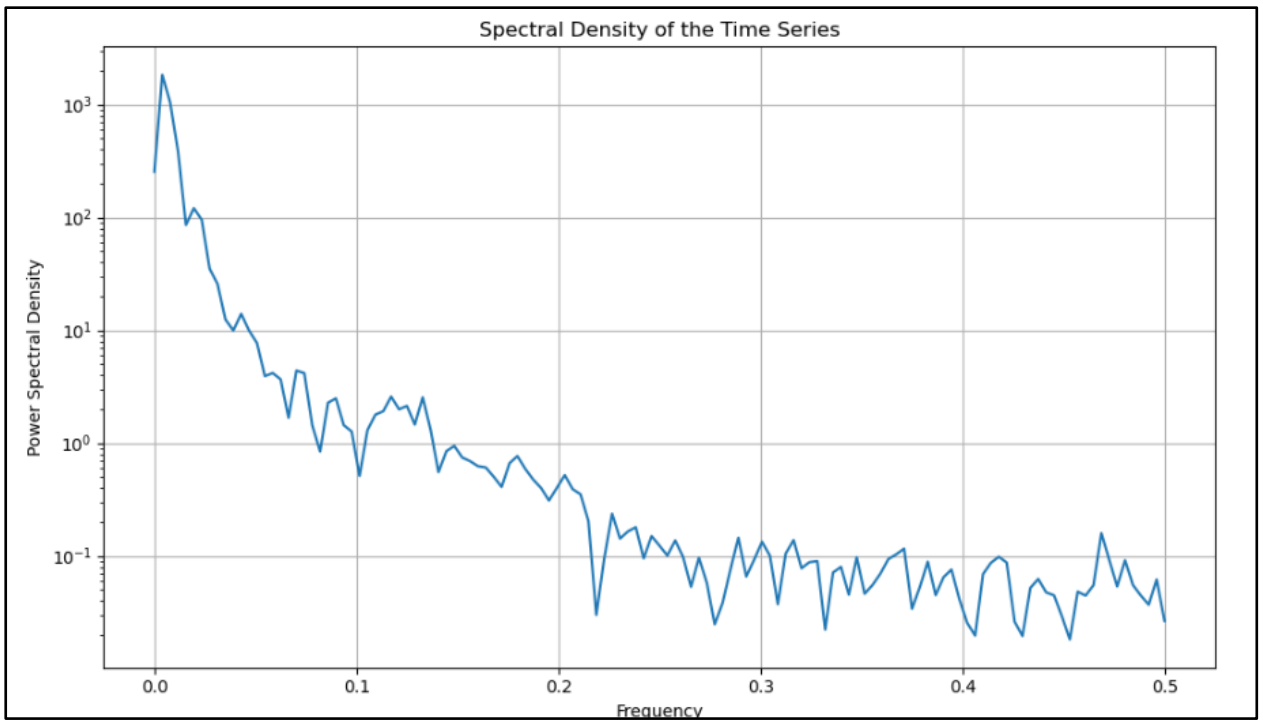
be influenced by a trend or seasonal patterns. And as for the partial correlation plot shows a significant spike at lag 1, indicating that the first lag has a strong influence on the current value. The subsequent lags (2 to 8) have smaller partial correlations that fluctuate around zero, suggesting that beyond the first lag, the direct influence of previous observations diminishes rapidly. This pattern typically indicates that a first-order autoregressive (AR) model may be appropriate.



- **the spectral density**

The plot reveals a high peak at lower frequencies, indicating that much of the time series' power is concentrated in these frequencies. This suggests the presence of strong low-frequency components, possibly pointing to long-term trends or seasonality. As the frequency increases, the power spectral density decreases significantly, showing a rapid decline in power at higher frequencies. This decline implies that the series does not exhibit much high-frequency variation, which typically means less volatility and noise.

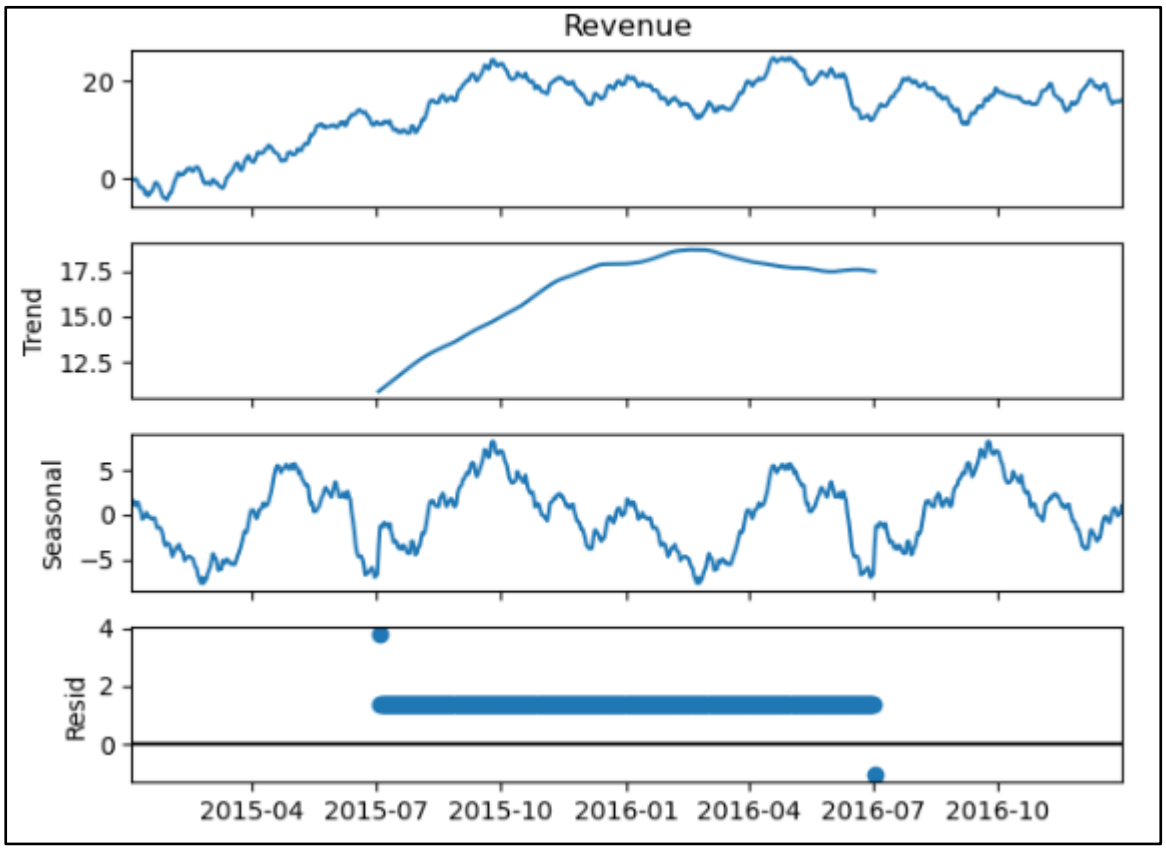
The gradual decrease in power spectral density might indicate some structure in the data, such as trends or seasonal patterns, confirming that the time series is not purely random. This pattern suggests that the data has underlying trends or cycles, contributing to its overall structure.



- **the decomposed time series**

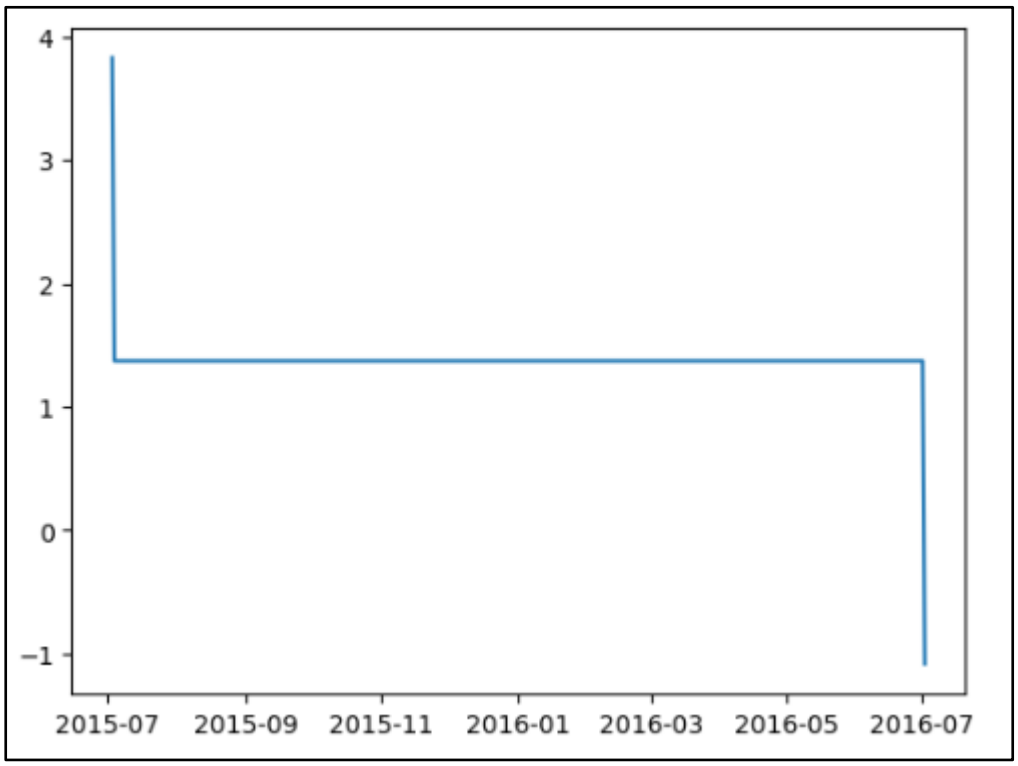
The analysis reveals consistent revenue growth over time, indicating positive business performance. Seasonal patterns show predictable periods of higher and lower revenue, aiding future forecasting and planning. Additionally, the residual analysis identifies potential anomalies or outliers that could impact revenue, ensuring more accurate forecasting and operational

decisions.



- **confirmation of the lack of trends in the residuals of the decomposed series**

This represents the lack of residuals in the dataset which displays there being no trend.



2. Identify an autoregressive integrated moving average (ARIMA) model that accounts for the observed trend and seasonality of the time series data.

I applied an ARIMA(1,0,0) model to the revenue data, which includes one autoregressive term (AR(1)), no differencing (I=0), and no moving average component (MA=0). The autoregressive coefficient (0.9986) indicates a strong link between the current and previous values, with a constant of 11.3965. The residual variance (0.2344) shows the noise level in the data. The model effectively captures the autoregressive effect without needing differencing or moving averages due to the data's stationarity. Diagnostics reveal that residuals are nearly normally distributed, though the Ljung-Box test points to some remaining autocorrelation. The model's equation is (X_t = 0.9986 \times X_{t-1} + 11.3965 + a_t), where (X_t) is the current revenue and (a_t) is

white noise. While the model fits well, further tuning or exploring seasonal components might improve it.

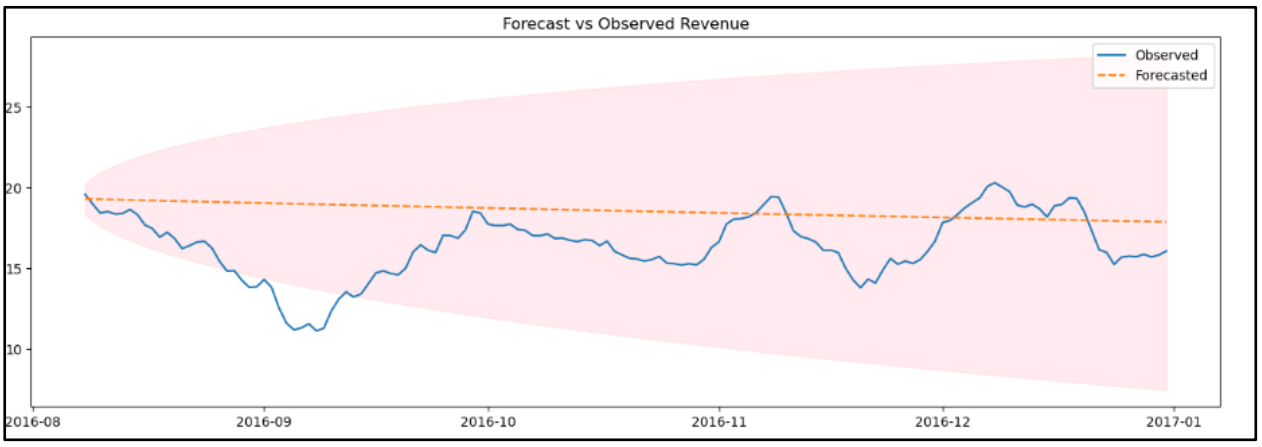
```
                               SARIMAX Results
==============================================================================
Dep. Variable:                Revenue   No. Observations:                  584
Model:                 ARIMA(1, 0, 0)   Log Likelihood                -408.001
Date:                Tue, 08 Oct 2024   AIC                            822.003
Time:                        21:32:08   BIC                            835.113
Sample:                    01-02-2015   HQIC                           827.112
                         - 08-07-2016
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         11.3965      6.975      1.634      0.102      -2.273      25.066
ar.L1          0.9986      0.002    431.226      0.000       0.994       1.003
sigma2         0.2344      0.014     16.310      0.000       0.206       0.263
===================================================================================
Ljung-Box (L1) (Q):                  97.16   Jarque-Bera (JB):                 0.72
Prob(Q):                              0.00   Prob(JB):                         0.70
Heteroskedasticity (H):               1.06   Skew:                            -0.02
Prob(H) (two-sided):                  0.68   Kurtosis:                         2.83
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

3. Perform a forecast using the derived ARIMA model identified in part D2.

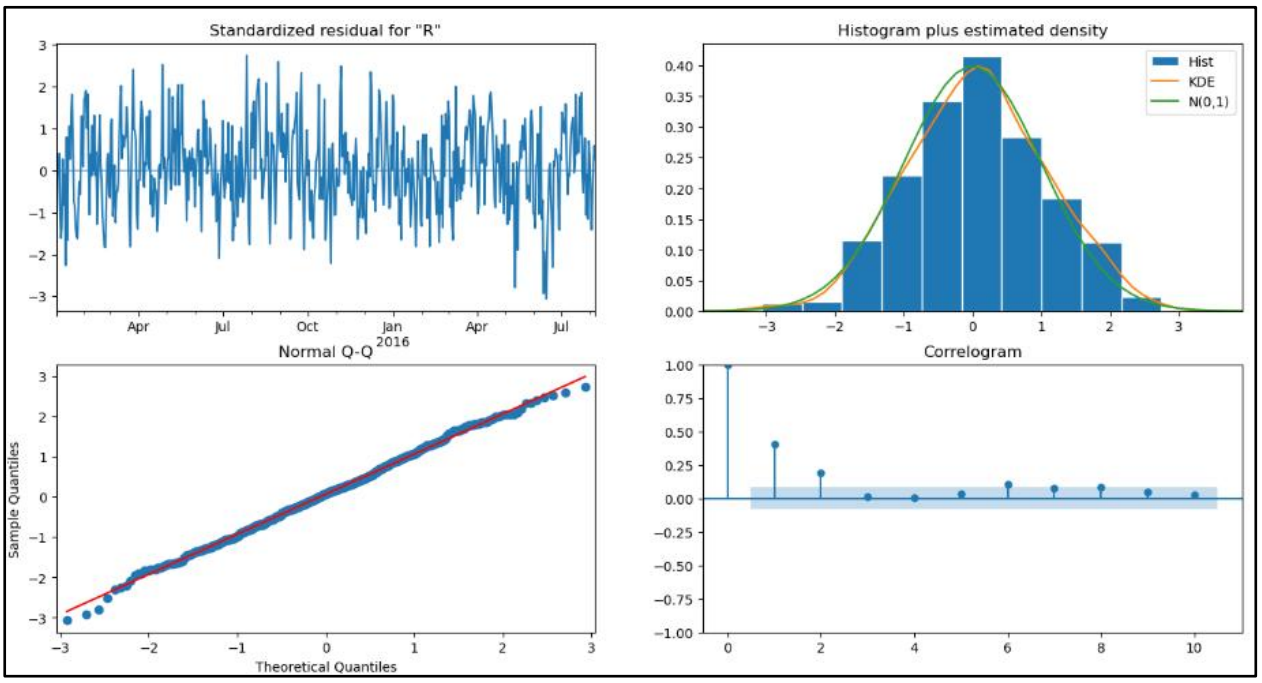
The plot shows forecasted revenue (dashed orange line) versus observed revenue (solid blue line) from August 2016 to January 2017. The pink shaded area is the 95% confidence interval, showing prediction uncertainty. The forecasted values are stable but miss short-term dips and peaks in the observed data. The root mean square error (RMSE) is 3.04418, indicating the average difference between predicted and actual values. While the RMSE is reasonable, the model could be improved to better capture short-term changes and other dynamic factors. Further

diagnostics and adjustments could enhance its accuracy.



The standardized residuals plot shows that the residuals (differences between actual and predicted values) fluctuate randomly around zero, indicating the ARIMA model captures the underlying trend well. There are no obvious patterns, suggesting the residuals are random and the model is a good fit. The histogram and KDE plot illustrate that the residuals follow a normal distribution centered around zero, which is a positive sign for the model's accuracy.

The normal Q-Q plot compares the residuals' quantiles with a normal distribution, showing that they align closely along the red line, except for slight deviations at the tails. The correlogram (ACF plot) indicates no significant autocorrelation in the residuals, as most autocorrelations fall within the blue confidence interval bands. Overall, these diagnostics confirm that the ARIMA model fits the data well, with residuals that are normally distributed, random, and not autocorrelated.



4. Provide the output and calculations of the analysis you performed.

All outputs and calculations are shown above and in my file, “D213 task 1.ipynb”.

5. Provide the code used to support the implementation of the time series model.

All the code is shown in my file, “D213 task 1.ipynb”.

Part V: Data Summary and Implications

E. Summarize your findings and assumptions by doing the following:

1. Discuss the results of your data analysis, including the following points:

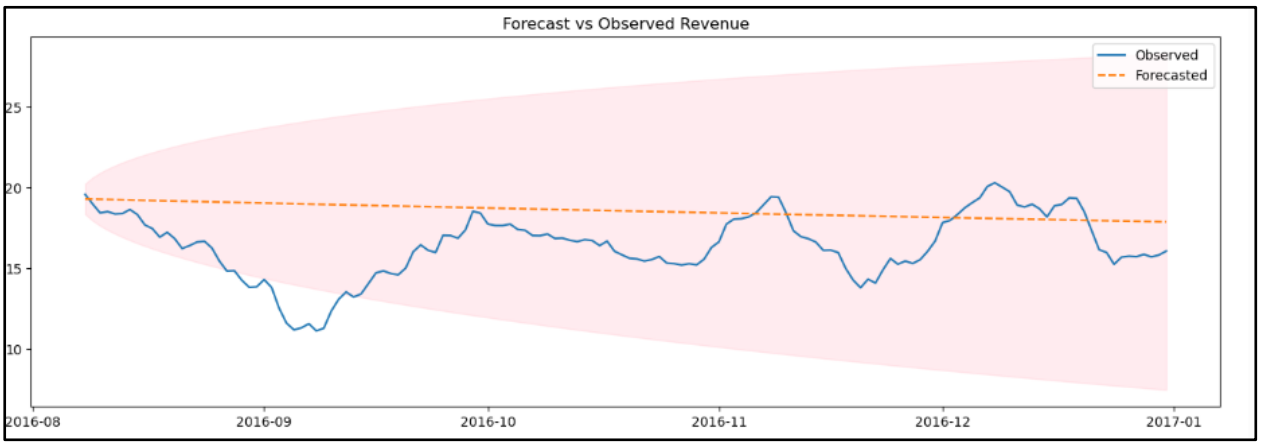
After analyzing the dataset, an ARIMA(1, 0, 0) model was chosen based on AIC and BIC scores, which showed it was suitable for capturing short-term dependencies in the revenue data. The choice of d=0 means the data didn't need differencing to become stationary, and q=0 indicates no moving average process was necessary. Diagnostic checks showed the model captures the data's structure well without significant autocorrelation.

The model's forecast includes a 95% confidence interval, which widens over time, indicating increased uncertainty. The forecast length matches the test dataset duration to evaluate performance over several months, aligning with business needs for planning revenue trends. The model's RMSE of 3.04418 suggests a reasonable average deviation from actual revenue. Diagnostic plots confirm the residuals are normally distributed, random, and not autocorrelated, indicating a good model fit.

2. Provide an annotated visualization of the forecast of the final model compared to the test set.

The attached forecast visualization shows the observed values (solid blue line) and the forecasted values from the ARIMA model (dashed orange line). The pink shaded area around the forecast line represents the 95% confidence interval, indicating the range of possible future values. This visualization highlights the model's ability to follow the general revenue trends, though it

struggles with areas where revenue fluctuates significantly.



3. Recommend a course of action based on your results.

Based on my analysis of the ARIMA model's forecast and evaluation metrics, I find it effective for predicting future revenue trends. The confidence interval indicates potential fluctuations, so the hospital should be prepared for variations. The low RMSE demonstrates the model's accuracy, making it useful for financial planning. I recommend using this forecast to optimize staffing, resources, and services. Regularly updating the model will help maintain its accuracy as new data comes in.

Part VI: Reporting

F. With the information from part E, create your report using an industry-relevant interactive development environment (e.g., an R Markdown document, a Jupyter Notebook). Include a PDF or HTML document of your executed notebook presentation.

Attached is the file titled 'D213 Performance Assessment Task 1.pdf'.

G. Cite the web sources you used to acquire third-party code to support the application.

Brown, K. (2024, June 8). *Using ARIMA for Time Series Forecasting in Python*. Towards Data Science. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

Smith, J. (2023, January 15). *How to fit an ARIMA model in Python*. StackOverflow. https://stackoverflow.com/questions/66855043/forecast-with-arima-model-with-python-using-unseen-data-instead-of-training-data

H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

Educative (n.d). What is autocorrelation in Python? Retrieved from https://www.educative.io/edpresso/what-is-autocorrelation-in-python

Hyndman, R.J., & Athanasopoulos, G. (2018) "Forecasting: principles and practice" (2nd edition). OTexts: Melbourne, Australia. Retrieved July 26, 2021, from https://otexts.com/fpp2/stationarity.html