Rancel Hernandez

Feb 4, 2025

D213 Advanced Data Analytics

NLM3 TASK 1: TIME SERIES MODELING

A1: RESEARCH QUESTION

What patterns exist in hospital revenue, and how accurately can an autoregressive integrated moving average (ARIMA) model forecast the hospital revenue?

A2: OBJECTIVES OR GOALS

The goals of this time series analysis include exploring revenue patterns, preparing the data for forecasting, selecting and training the optimal model, evaluating its performance, and interpreting the results to provide actionable insights. By examining the revenue data for trends and seasonality, the analysis can determine how best to prepare and transform it for modeling. Then once the data is properly formatted and split into the training and test sets, it can be used to train and compare different models to identify one that reliably forecasts future revenue. Finally, interpreting these forecasts could lead to deriving meaningful conclusions and recommending informed courses of action. These goals structure the analysis into its core components and align with the scope of the dataset.

B: SUMMARY OF ASSUMPTIONS

Time series models have several assumptions that must be satisfied to ensure reliable results. These assumptions include:

1. Linearity: The ARIMA model assumes a linear relationship between past values and future values, meaning that the time series can be modeled as a linear combination of its past observations and errors [1]. As a result, future values are forecasted as a weighted sum of past data and residual terms. Due to this assumption, the ARIMA model requires the data to be transformed to fit within its rigid linear structure, often involving the removal of complex and dynamic patterns such as trends and seasonality. However, if these nonlinear patterns persist, the model may fail to adequately fit the data and produce accurate forecasts, making alternative nonlinear time series models a more suitable choice.
2. Sequential, Complete, and Equal Time Steps: The data must be sequential to prevent distortion of the natural fluctuations and patterns. In other words, it must be ordered chronologically because unordered data can lead to disorganized training and test sets, resulting in inaccurate predictions. Additionally, the time series must be complete and have equal time steps, meaning there should be no gaps and the distance between observations must remain constant [2]. If uneven gaps occur, they can negatively impact the model's ability to learn the underlying structure and may result in poor forecasts.
3. Stationarity: The time series must be stationary, meaning its statistical properties do not change over time [3]. These properties can include the mean, variance, and

autocorrelation. Although minor fluctuations around a baseline level are acceptable, there should not be any strong shifts in these properties. The ARIMA model is a linear regression model on lagged values, so the data is assumed to have a certain level of linearity and stationarity. If there are significant changes in these properties over time, then the series becomes non-stationary. As a result, the model will likely fail to capture the structure, resulting in poor fit and unreliable forecasts.

4. Autocorrelated Data: For a time series model, the data must be autocorrelated, which means that observations at different points in time are related to one another [4]. This creates a dependency where past values influence future ones, including dependencies between the observations themselves and their errors or residuals. The strength of the autocorrelation tends to weaken over time, so values that are closer are more correlated than those farther apart. Additionally, this relationship can also be positive if the values move in the same direction or negative if they move in opposite directions.

    Without any significant autocorrelation, the time series analysis would not be necessary because the past values would not provide meaningful information for predicting the future ones [5]. Alternatively, a more simpler approach would be to just predict the mean of data.

5. No Seasonality: The time series should not contain seasonality because not all time series models explicitly account for it [6]. Seasonality refers to recurring patterns at fixed intervals, such as a decrease in revenue every seven days or an increase in sales every December. ARIMA models are linear, so they can only capture patterns that follow a consistent relationship over time. However, seasonal fluctuations involve more complex and repeating variations that do not fit well within ARIMA's linear structure. Because of

this limitation, ARIMA cannot effectively model seasonality, so it should be decomposed and removed to create a more stable time series for the analysis.

6. White Noise Residuals: After fitting a time series model, the residuals should resemble white noise, meaning they should be randomly scattered, show no discernible patterns over time, and follow a normal distribution [7].Since the ARIMA model is linear, it assumes that all patterns in the data have been captured and modeled to stabilize the series. These patterns include trends, seasonality, and autocorrelations. Properly accounting for these components should leave behind a stationary time series with only random fluctuations. If the residuals still show patterns, correlations, or structure, this suggests that the model has not fully captured the data, indicating that additional transformations or model adjustments may be needed.

C1: LINE GRAPH VISUALIZATION

The line graph visualizing the realization of the time series, along with the rolling mean and rolling standard deviation, is shown in *Figure 1*.

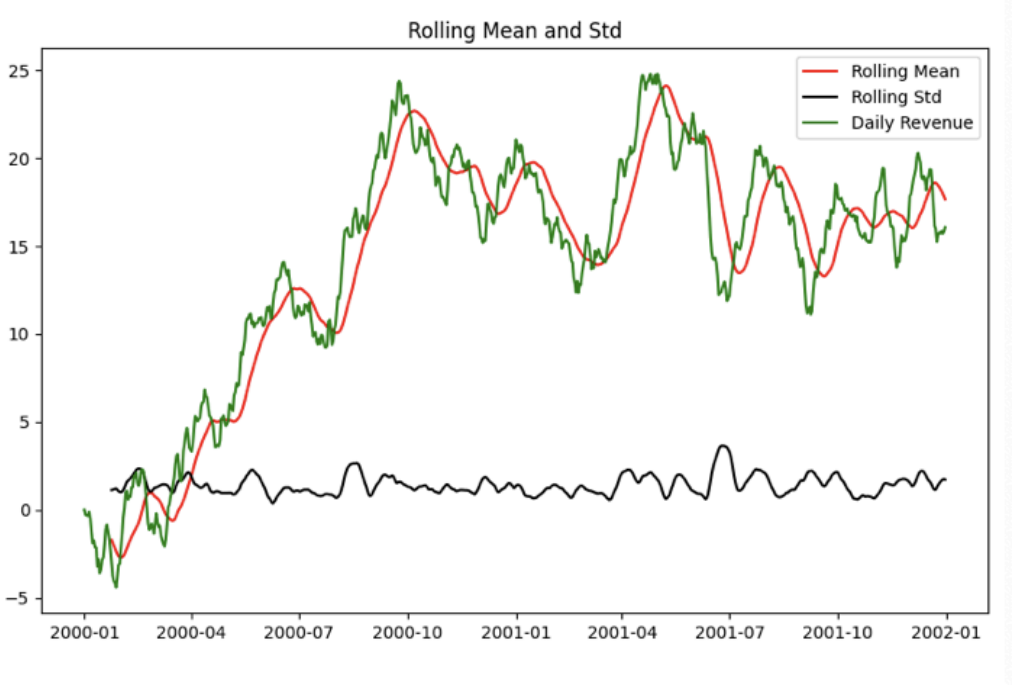

Figure 1: A visualization of the time series realization.

C2: TIME STEP FORMATTING

The time series spans two years for a total of 731 days, beginning on '2000-01-01' and ending on '2001-12-31.' However, these start and end dates were artificially set, so the actual date range of the time series is unknown. Each time step represents one day, and all entries are ordered chronologically. There are no missing values, meaning the sequence is continuous with no gaps in the measurement. The time series appears to exhibit some mean reversion but follows a random walk with drift because the revenue shows a slightly inconsistent upward trend over

time. A random walk is a time series in which each value is determined by the previous value plus a random fluctuation, and it typically appears to have meaningful patterns [8]. However, its oscillations are unstructured and result in an unpredictable path. Despite this randomness, a random walk can still follow a long-term trend that leads to an overall increase or decrease over time.

C3: STATIONARITY

There were non-stationary visual cues present in the revenue plot in *Figure 1* because the rolling mean changed over time while maintaining a stable standard deviation, indicating unstable statistical properties. To further assess the stationarity, the Augmented Dickey-Fuller (ADF) test was conducted, and the results are shown in *Figure 2*. In simple terms, the null hypothesis for this test states that the time series is non-stationary [9]. The p-value of approximately 0.2 exceeds the 0.05 threshold, so we fail to reject the null hypothesis that the data is non-stationary. Additionally, the ADF statistic can indicate whether the time series requires differencing to become stationary, and the critical values provide thresholds at 1%, 5%, and 10% significance levels. The 1% threshold represents the strongest evidence against non-stationarity because the null hypothesis can be rejected with a 99% confidence if the ADF statistic is lower than this threshold. However, the ADF statistic of approximately -2.22 is higher than all critical value thresholds, strongly suggesting that the data is non-stationary and requires differencing.

```
The ADF statistc is -2.22 > -2.87 (5%)
The p-value: 0.2
The lags used: 1
The number of observations used: 729
Critical Values: {'1%': -3.44, '5%': -2.87, '10%': -2.57}
```

Figure 2: The results of the ADF test.

C4: STEPS TO PREPARE THE DATA

To prepare the data for analysis, it was first explored and then transformed to ensure it was compatible with the ARIMA model and time series tests. The ARIMA model requires a complete and sequential time series. Thus, the dataset was checked for missing values in the 'Revenue' column, and none were found. The Day column represents the number of days since hospital operations began and should span two full years. To verify completeness, missing dates were identified by comparing the existing index to a complete two-year range, and no gaps in measurement were found. Duplicate values were also checked because multiple entries for the same day could artificially inflate fluctuations in the time series, and no duplicated dates were identified. The start of the range was arbitrarily set to '2000-01-01' because the specific start date does not impact the analysis. This transformation was applied to correct the date indexing for the ARIMA model because Statsmodels cannot interpret the time intervals correctly without a proper datetime index [10]. Finally, the dataset was split into an 80% training set and a 20% test set to provide sufficient data for training and evaluating the model. The data was chronologically ordered to ensure that all training data preceded the test data to allow for a proper evaluation of the test and forecasted values.

C5: PREPARED DATA SET

The cleaned dataset will be submitted separately in a CSV file.

D1: REPORT FINDINGS AND VISUALIZATIONS

The data was analyzed through a time series decomposition, and the autocorrelation function (ACF) and spectral density plots were examined. A time series decomposition separates the data into its key components: trend, seasonality, and residuals [11]. A period length must be defined to assess how strong a seasonal component at this rate is present. In the revenue plot, there appeared to be a slight seasonal component in the form of mean reversion and cyclical fluctuations, but the period of these patterns was unknown due to their inconsistency. The ACF and spectral density plots were examined to identify the potential period length for the decomposition and will be discussed in detail later.

The ACF plot further confirmed the absence of a seasonal component, so determining a definitive period length for the decomposition was difficult. However, the spectral density plot identified a slight peak in variance around a frequency of 0.12. The frequency in the spectral density plot is normalized, so dividing 1 by the frequency converts it into the time steps, resulting in an estimated period of approximately 8 days [12]. Therefore, the decomposition was performed using a period length of 8 days because this was the rate of the most significant spike in variance, potentially reflecting the duration of the mean reversion cycles. That said, the seasonal decomposition component follows a white noise distribution with no apparent structure, as shown in *Figure 3*. This absence of any patterns and presence of random fluctuations provides strong evidence that no seasonal component exists in the time series [13].

The trend indicates if the data is increasing or decreasing over time, but these trends can be deterministic or stochastic. A deterministic trend will show a steady increase or decrease over time, but a stochastic trend fluctuates unpredictably and is likely influenced by external factors and not time [14]. The trend decomposition component shows an upward trend in revenue that plateaus midway through the series, as shown in *Figure 3*. This trend was also noticeable in the revenue plot but appears to resemble a random walk with drift because it does not persist throughout the entire series. Therefore, it is likely that this trend is influenced by external factors. While the revenue eventually seems to stabilize, this temporary trend signifies that the data is non-stationary and may require differencing.

The residual decomposition component shows that the residuals are randomly scattered with no consistent fluctuations. These residuals represent the remaining structure of the time series after removing the trend and seasonal components [15]. The absence of clear patterns in the remaining decomposed series suggests that the model should fit the data well, assuming that the trend component is addressed through differencing the series to ensure stationarity. Additionally, the randomness of the residuals indicates that no further patterns are present, suggesting the decomposition captured the underlying structure of the time series.

The ADF test was performed on the residuals to provide further evidence for the absence of patterns and presence of stable statistical properties, with the results shown in *Figure 4*. The p-value was near 0 and well below the 0.05 threshold, so we can reject the null hypothesis that the residuals are non-stationary. Moreover, the ADF statistic of approximately -12.87 was significantly lower than all the critical values, providing a 99% confidence that the residuals are stationary. Therefore, this is strong evidence to support the absence of trends in the residuals of the decomposed series.

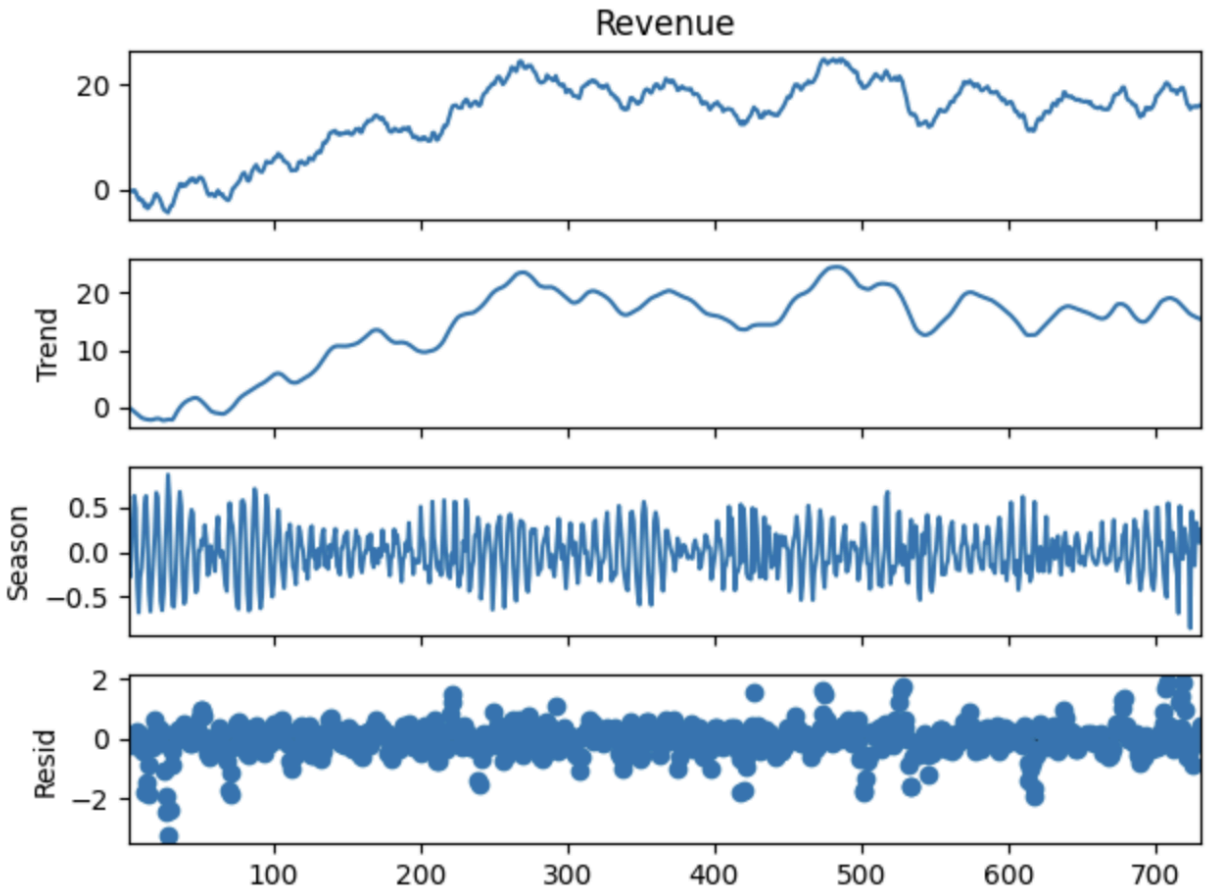

Figure 3: The time series decomposition components.

```
The ADF statistc is -12.87 > -2.87 (5%)
The p-value: 0.0
The lags used: 2
The number of observations used: 728
Critical Values: {'1%': -3.44, '5%': -2.87, '10%': -2.57}
```

Figure 4: The ADF test results of the residuals.

The ACF plot visualizes the significance of correlations between a time series and its lagged values, where each bar represents the correlation coefficient between a lag and the time series itself [16]. A lag value indicates how many past observations relate to the current time series as a whole. The coefficients can vary in magnitude, either positively or negatively, and their significance tends to diminish as the number of lags increases. Seasonality can be detected in the ACF plot when spikes in the coefficients appear at consistent intervals because they reflect the cyclical patterns in a time series. However, the lags displayed a gradual decrease in significance rather than periodic spikes, suggesting no seasonal component is present. The ACF plot is shown in *Figure 5*.

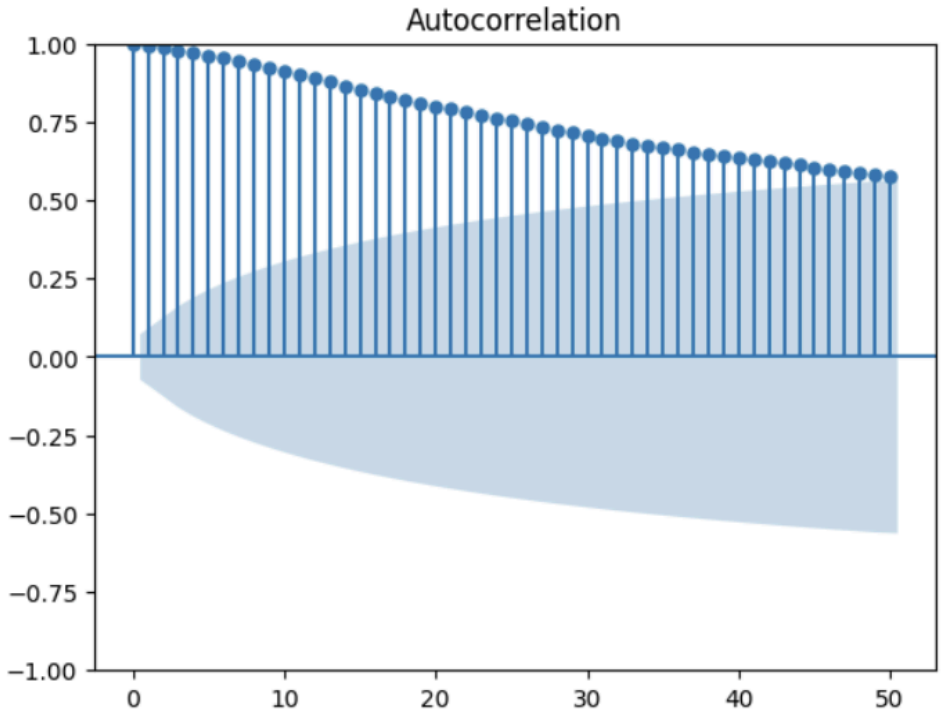

Figure 5: The autocorrelation function plot.

The spectral density plot visualizes the distribution of variance across different frequencies in a time series [12]. In simpler terms, the spectral density measures the strength of each frequency as a portion of the total variance. A high spectral density at a given frequency indicates that a significant portion of the variance is concentrated in patterns at that frequency, while a low spectral density at that frequency means its contribution is minimal to the overall structure of the time series.

Each frequency represents a repeating pattern at a certain rate of change [17]. Low-frequency patterns correspond to slow-moving, long-term trends, while high-frequency patterns represent rapid fluctuations. Seasonality is not restricted to a specific frequency range. In this series, long-term seasonal patterns, such as yearly fluctuations, would appear at low frequencies because they represent gradual changes over time. Mid-range frequencies would represent a moderate change over time, such as monthly or quarterly fluctuations. Higher-frequency seasonality may appear as daily or weekly fluctuations that occur at shorter intervals.

Noise can appear at any frequency but is most common at higher frequencies in the form of rapid, short-lived fluctuations [18]. However, noise typically has low spectral density because its random fluctuations contribute less variance than more consistent and structured patterns. Thus, the spectral density plot can help determine whether a time series is dominated by trends, seasonality, or noise.

The spectral density plot shows a high concentration of spectral density at lower frequencies that gradually decreases as the frequency increases, with no significant spikes. Seasonality corresponds to structured patterns at a specific frequency, so the absence of any distinct spikes supports the lack of a seasonal component in the time series. This plot shows that

the series is dominated by a long-term trend, exhibits some mean reversion at low-to-mid frequencies, and contains noise in the form of rapid fluctuations that contribute minimal variance. This observation aligns with the revenue patterns seen in *Figure 1*. The spectral density plot further supports the idea that the time series follows a random walk with drift. The mean reversion of the random walk is reflected in the low-to-mid frequencies, while the drift corresponds to the concentrated spectral density at low frequencies that gradually decreases. The spectral density plot is shown in *Figure 6*.

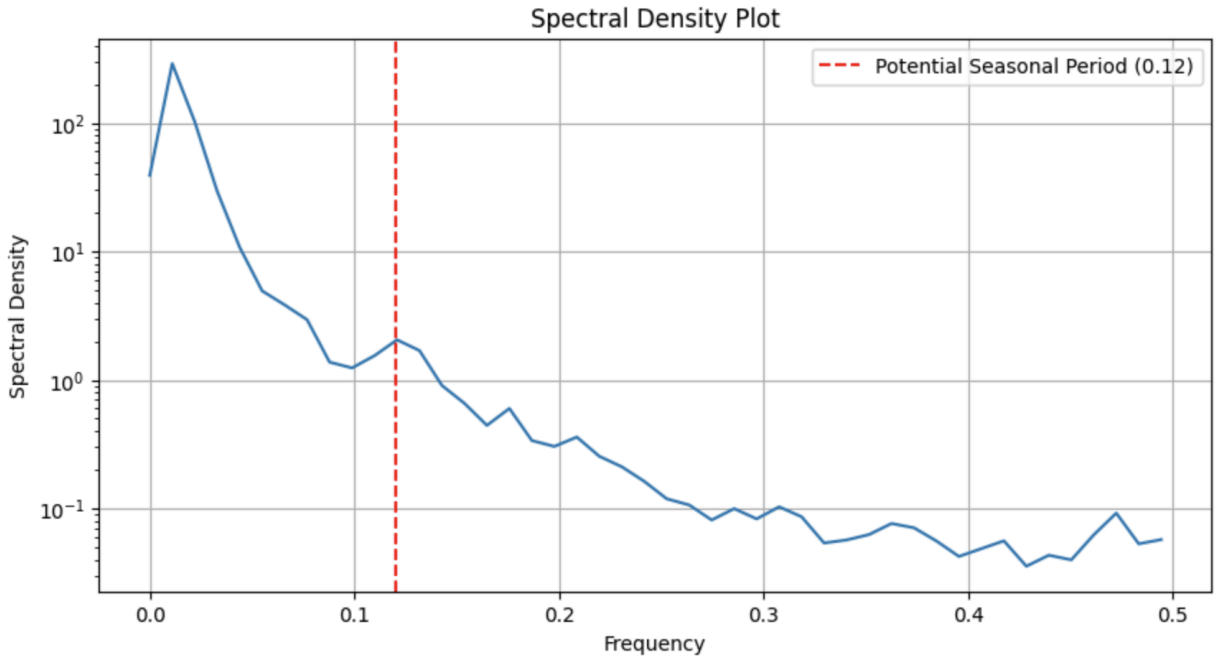

Figure 6: The spectral density plot.

D2: ARIMA MODEL

The ARIMA model was selected by comparing the p-values, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) scores across models of different orders. These models were determined by identifying the range of significant terms based on the structure of the time series. The p-values of the model's terms help assess their significance because the null hypothesis states that the coefficients are equal to zero, meaning they do not contribute meaningfully to the model [19]. Once the meaningful sets of terms are identified, they can be used to initialize models of different orders to select the optimal one with significant coefficients and low AIC and BIC scores. Including too many terms can lead to overfitting, so AIC and BIC balance model fit and complexity by penalizing excessive parameters [20].

The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots were used to determine the autoregressive (AR) and moving average (MA) terms [16]. The ACF depicts the direct and indirect relationships between lags and the current value, where each lag represents a range of previous values. Conversely, the PACF focuses on the direct relationship between a specific past value and the current value by removing the influence of intermediate lags. The strength of the relationship between each lag and the current value is measured using the correlation coefficient. To determine the significance of the lags, the plot includes a shaded confidence interval, and lags within this area are considered near zero and insignificant with a 95% confidence.

The AR term represents how many lags are used to predict the current value, while the MA term represents how many past forecast errors are considered in making the prediction [21]. The ACF and PACF reveal the underlying structure of the time series and help identify the significant lags to select the range of AR and MA terms. If the ACF plot shows significant

correlations at early lags but then sharply cuts off, it suggests that past forecast errors had a temporary influence on the current values. Their impact quickly diminishes due to short-term fluctuations, and these errors do not persist over the following time steps. The MA term corrects this by incorporating past errors into the prediction equation, allowing the model to adjust for these brief dependencies before their influence fades.

However, the ACF plot showed that the significance of the lags gradually decreased, so the past errors persist over time because there are no major disruptions affecting their influence. This suggests the time series follows an autoregressive process and past values consistently influence the current value, making AR terms more appropriate and MA terms likely unnecessary [21]. The ACF plot can be seen in *Figure 5*.

The PACF plot further confirmed the autoregressive nature of the time series because only the first two lags were significant, as shown in *Figure 7*. This suggests that past errors do not directly influence the current value. Instead, only the past two values contribute to its predictions. The first lag had a high correlation coefficient of approximately 0.99 with the current value, while the second lag had a much lower value around -0.25. Although the second lag showed some significance, it may not be meaningful for the model.

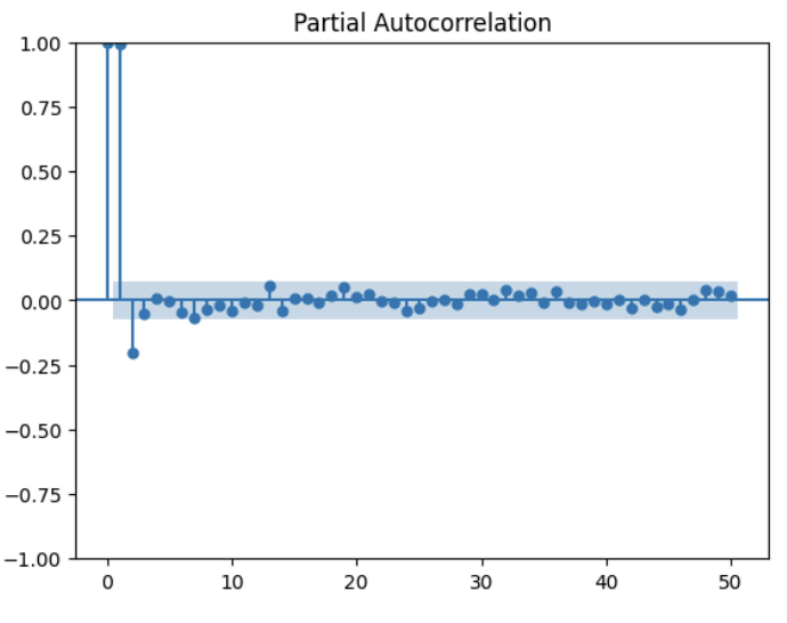

Figure 7: The partial autocorrelation plot.

The final term that will need to be considered is the differencing term, and it determines how many differencing steps should be applied to the time series to become stationary [22]. This term is selected by performing an ADF test on the differenced data until stationarity is achieved. The results of the ADF test on the original time series indicated non-stationarity, so first-order differencing was applied. After one differencing, the ADF statistic was approximately -17.4 and lower than all critical values, providing 99% confidence that the series is now stationary. Additionally, the p-value was near 0 and well below the 0.05 threshold, so we reject the null hypothesis that the series is non-stationary. Thus, the model will likely require only one differencing step, and the results can be seen in *Figure 8*.

```
The ADF statistc is -17.37 > -2.87 (5%)
The p-value: 0.0
The lags used: 0
The number of observations used: 729
Critical Values: {'1%': -3.44, '5%': -2.87, '10%': -2.57}
```

Figure 8: The results of the ADF test on time series with one differencing step.

The parameter selection process resulted in the two AR terms, zero MA terms, and one differencing step. However, an MA term and zero differencing will still be tested to validate the selection process. The ARIMA model follows the parameter order (p, d, q), where p is the AR term, d is the difference term, and q is the MA term [1]. The full set of model parameters that were tested includes (2, 1, 0), (1, 1, 0), (1, 1, 1), and (1, 0, 0). The optimal model will be selected based on the significance of the parameters and the lowest AIC and BIC scores.

The second AR term in the model (2, 1, 0) had a p-value of approximately 0.43, and the MA term in the model (1, 1, 1) had a p-value of approximately 0.59. The high p-values well above the 0.05 threshold indicate that these extra terms were not statistically significant. Thus, we fail to reject the null hypothesis that the coefficients of these terms are zero. As a result, only one AR term is required for the model.

The (1,0,0) model with zero differencing produced an AIC score of approximately 822 and a BIC score of 835. These were the highest among all tested models, despite using the fewest terms, indicating the poorest performance. Conversely, the (1,1,0) model achieved the lowest AIC and BIC scores, around 705 and 714, while also having statistically significant parameters. Thus, the (1,1,0) model was determined to be the optimal model in the selection process. The results of the model comparisons can be seen in *Figure 9*.

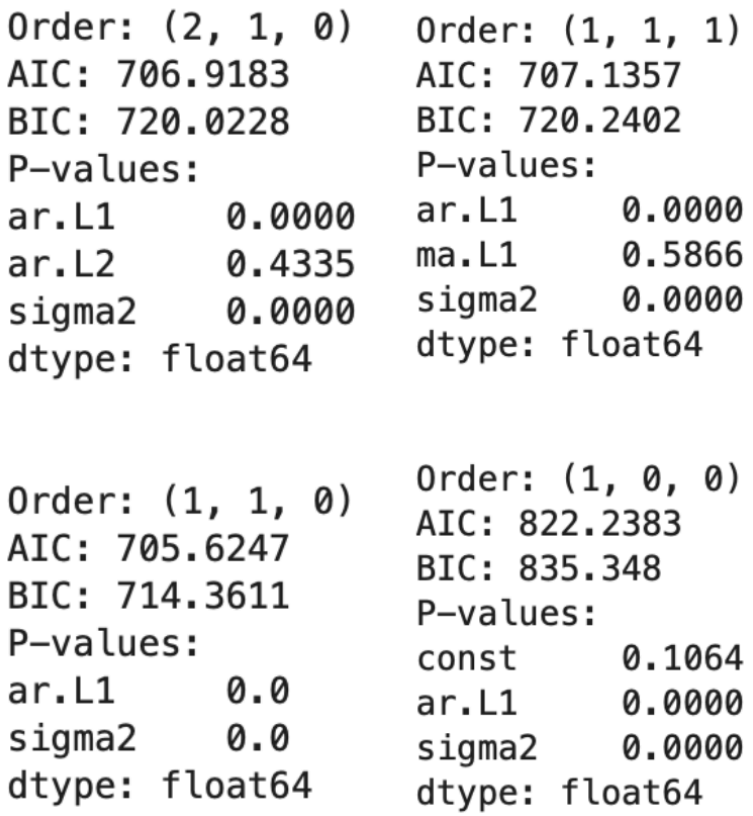

Figure 9: The results of the model selection process.

D3: FORECASTING USING ARIMA MODEL

A forecast was performed using the optimal ARIMA model (1,1,0) from the selection process to predict hospital revenue for the next 90 days. The forecast output includes the revenue data from the training and test sets, along with the predicted values. The results are shown in *Figure 10*.

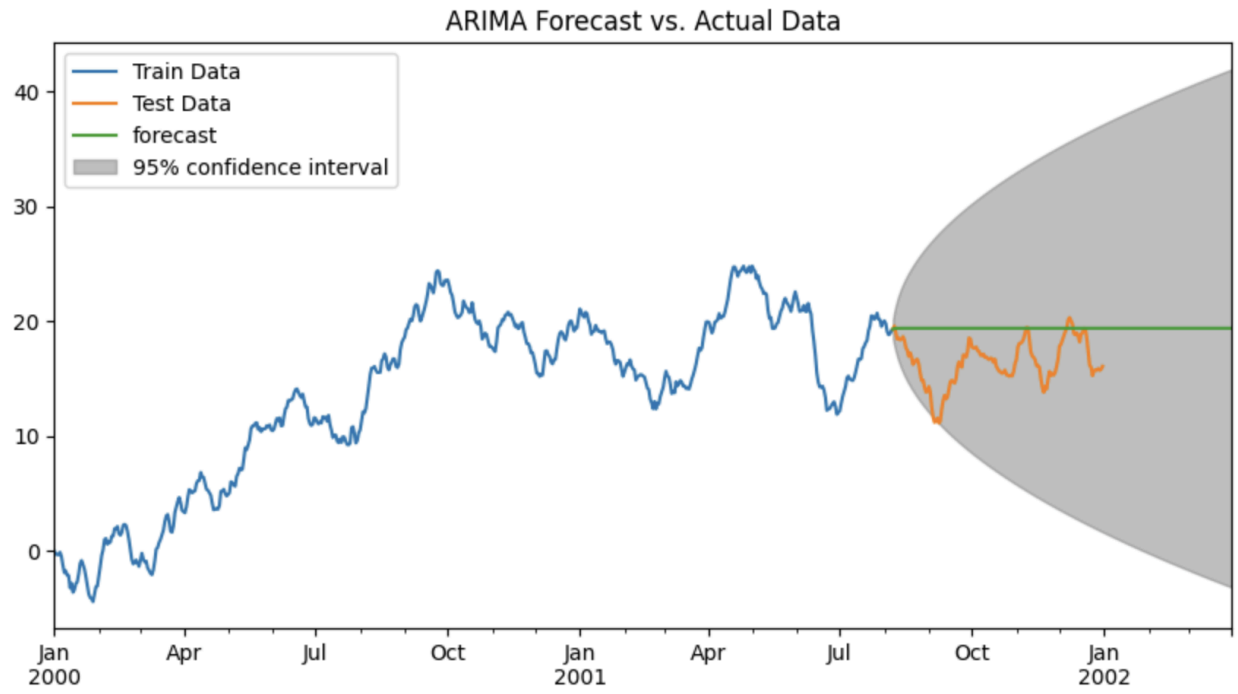

Figure 10: The ARIMA forecasts compared to the actual data.

D4: OUTPUT AND CALCULATIONS

The summary of the ARIMA model (1,1,0) is shown in *Figure 11*, which includes the model coefficients, standard errors, and statistical significance measures. The Root Mean Squared Error (RMSE) for the model was approximately 3.6. The RMSE measures the average deviation between the predicted values and the actual test data [23]. It is calculated by squaring each error to penalize larger deviations and remove the negative signs, taking the average of these squared errors, and then applying the square root to return the error to the original scale. This value provides an average margin of error in the model's predictions. The first five predictions are shown in *Figure 12*, and the range of values for the original, test, and predicted sets is shown in *Figure 13*.

SARIMAX Results

Dep. Variable:	Train Data	No. Observations:	584
Model:	ARIMA(1, 1, 0)	Log Likelihood	-350.812
Date:	Wed, 12 Feb 2025	AIC	705.625
Time:	11:31:20	BIC	714.361
Sample:	01-01-2000	HQIC	709.030
	- 08-06-2001		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4106	0.038	10.837	0.000	0.336	0.485
sigma2	0.1950	0.012	15.888	0.000	0.171	0.219

Ljung-Box (L1) (Q):	0.15	Jarque-Bera (JB):	1.86
Prob(Q):	0.70	Prob(JB):	0.39
Heteroskedasticity (H):	1.04	Skew:	-0.05
Prob(H) (two-sided):	0.78	Kurtosis:	2.74

Figure 11: The summary output of the ARIMA model (1, 1, 0).

2001-08-07	19.299360
2001-08-08	19.345870
2001-08-09	19.364967
2001-08-10	19.372809
2001-08-11	19.376029

FIgure 12: The first 5 predictions.

Total Revenue Range: -4.4233 to 24.7922
Test Data Range: 11.1192 to 20.3071
Prediction Range: 19.2994 to 19.3783
1000-day forecast Range: 19.2994 to 19.3783

Figure 13: The range of the original, test, and prediction sets, including the range for the 1000-day forecasts.

D5: CODE

The code for the implementation of the time series model will be submitted in a separate Python file.

E1: RESULTS

The results of the analysis focus on the selection of an ARIMA model and its forecasting performance. The optimal ARIMA model (1, 1, 0) was selected based on the statistical significance of its coefficients and low AIC/BIC scores to ensure a balance between model complexity and overfitting. This model includes one AR term, one differencing step, and no MA terms. The ACF plot showed a gradual decline in the significance across the lags, indicating that the past values continue to influence the current ones and suggests an autoregressive process. The PACF plot confirmed this relationship by showing a sharp drop in significance after two lags, implying that the past errors do not directly impact the current values beyond the two most recent ones. Thus, two AR terms were evaluated, with the first lag exhibiting a strong correlation and the second lag showing minimal but potential significance. The original time series was non-stationary, but achieved stationarity after one differencing step. Several alternative models were tested but rejected due to poor performance because their coefficients were not significant and they produced higher AIC and BIC scores.

The prediction interval is shown in *Figure 10* as the shaded area on the right side of the plot. This interval represents the range within which future actual values are expected to fall with a 95% confidence, and it accounts for model uncertainty and potential future variability [24]. The prediction interval is narrow near the test data range of 147 days but starts to expand further into the future. This suggests the model has higher confidence in short-term predictions with minimal

variability because over 95% of the test data falls within the shaded region, with only a few points slightly outside its border. However, the prediction interval starts to widen over time, reflecting a greater uncertainty and increased variability in long-term forecasts [25].

The length of the forecast was determined by analyzing the behaviour of the model's predictions. The range of the predictions were limited to approximately 19.3 to 19.4, regardless of the date. To further validate this observation, the forecasts were projected 1,000 days into the future, yet the range remained consistent, as shown in *Figure 13*. This suggests the model assumes the revenue will plateau and stabilize over time. This outcome could be due to the absence of complex patterns and the minimal influence from past errors in the time series, which could have introduced greater variability in the predictions. Additionally, the model includes only one AR term and no MA terms, so it only relies on a single previous value for its forecasts. This low model complexity may explain the limited variability in the predictions because the simplicity of the model could have led to underfitting, preventing it from capturing more complex patterns in the data [26]. While the length of the forecast is somewhat negligible, its reliability is at its highest around the range of the test data because the prediction interval remains narrow, providing greater certainty for short-term predictions.

The error metric to evaluate the model was the RMSE, and the model produced a score of approximately 3.6. This metric is moderately low compared to the full range of the time series of approximately -4.4 to 24.8. However, the range of the test data was significantly reduced, at approximately 11.1 to 20.3, making the RMSE relatively high within this subset. Given the narrow range of the predictions within the test data range, the RMSE suggests that short-term forecasts remain within a reasonable margin of error. That said, the model is less reliable for

long-term predictions because the margin of error is likely to increase due to expected fluctuations over time.

To validate the model's performance, the residuals were tested for stationarity, and the residual, ACF, and PACF plots were analyzed. The ADF statistic was approximately -14.46 and well below all the critical values, indicating the residuals were stationary. Moreover, the p-value was near zero and below the 0.05 threshold, so we reject the null hypothesis that the residuals are non-stationary. Therefore, the ADF test indicated that the residuals are stable and no additional patterns remained. The results of the ADF test are shown in *Figure 14*.

```
The ADF statistc is -14.46 > -2.87 (5%)
The p-value: 0.0
The lags used: 2
The number of observations used: 581
Critical Values: {'1%': -3.44, '5%': -2.87, '10%': -2.57}
```

Figure 14: The results of the ADF test on the model's residuals.

The residual plot provides further evidence for the results of the ADF test because the residuals resemble white noise, as shown in *Figure 15*. These fluctuations appear to be random, indicating that no consistent patterns remain. The ACF plot shows a rapid decrease in significance across the lags, suggesting that only the most recent residuals influence the current ones and no long-term effects persist. While most lags fall within the confidence interval, the few exceptions between lags 6 and 9 are likely due to noise and are insignificant. The PACF further confirms the randomness of the residuals because only the first two lags show some significance, meaning that past errors do not persist over time. Therefore, the residuals are stable and random,

indicating that the model has successfully captured the patterns in the time series and is a good fit [7]. The ACF and PACF plots of the residuals are shown in *Figures 16 and 17.*

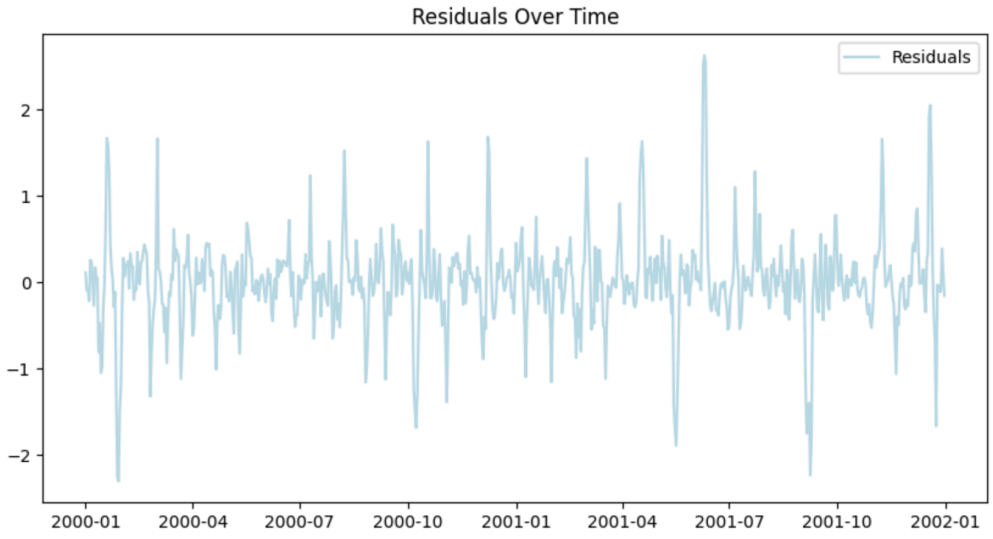

Figure 15: The residual plot.

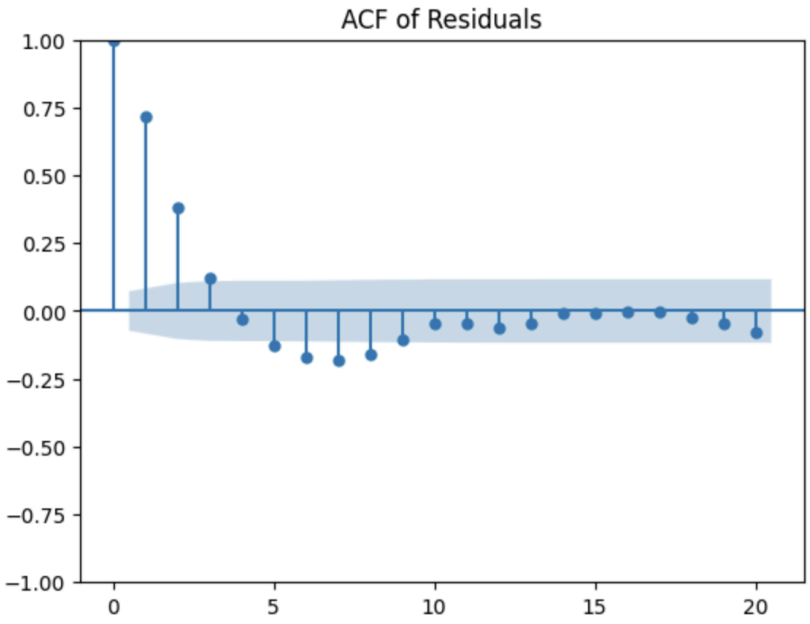

Figure 16: The ACF plot of the model's residuals.

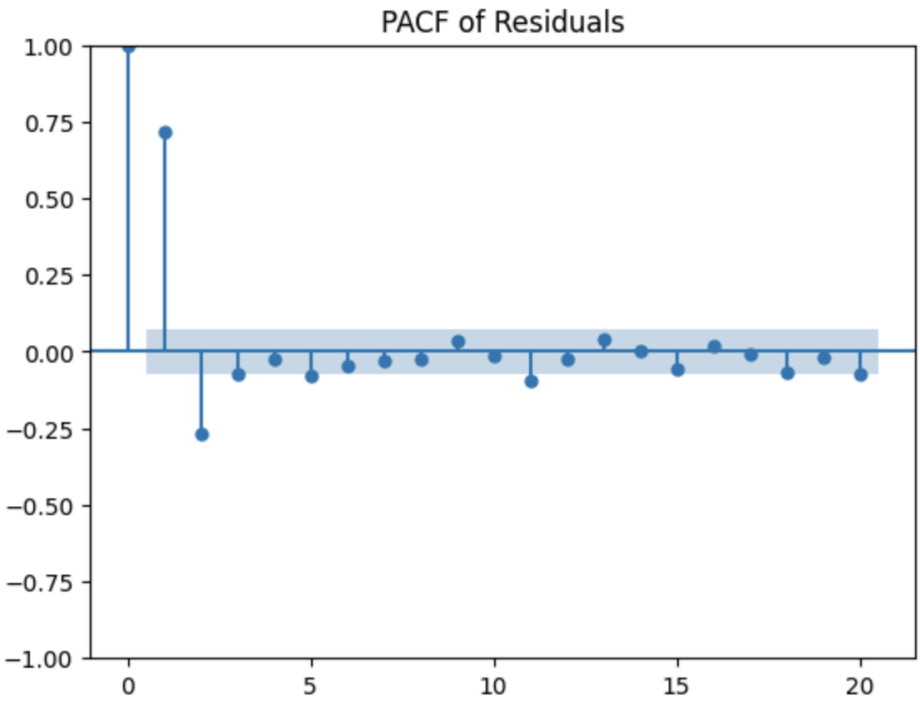

Figure 17: The PACF plot of the model's residuals.

The results of this time series analysis identified an optimal ARIMA model (1,1,0) with statistically significant coefficients and low AIC/BIC scores, ensuring a balance between model complexity and overfitting. The model's forecasts remained within a narrow range and produced a moderately low RMSE of 3.6, indicating a reasonable accuracy for short-term predictions within the test data range of approximately 147 days. The model's performance was validated through a residual analysis, which confirmed that the residuals were stationary, random, and resembled white noise. The ACF and PACF plots of residuals showed minimal short-term dependencies that were likely due to noise rather than consistent patterns. These outcomes suggest that the model captured the underlying structure of the data.

While the model performed well for short-term forecasting, it is not reliable for long-term projections. The prediction interval widened over time, and forecasts remained constrained within a narrow range, even when projected 1,000 days into the future. This limitation suggests that the model assumes the revenue will plateau and stabilize, potentially overlooking long-term variability. Therefore, the ARIMA model is reliable for short-term forecasting due to its low margin of error, narrow prediction interval, and uncorrelated residuals.

E2: ANNOTATED VISUALIZATION

The annotated visualization of the forecast of the final model compared to the test set is shown in *Figure 18*. The dashed red line represents the model's predictions, the oscillating green line depicts the actual revenue from the test set, and the shaded area indicates the RMSE margin of error. This visualization highlights how well the model's forecasts align with the actual values for short-term projections.

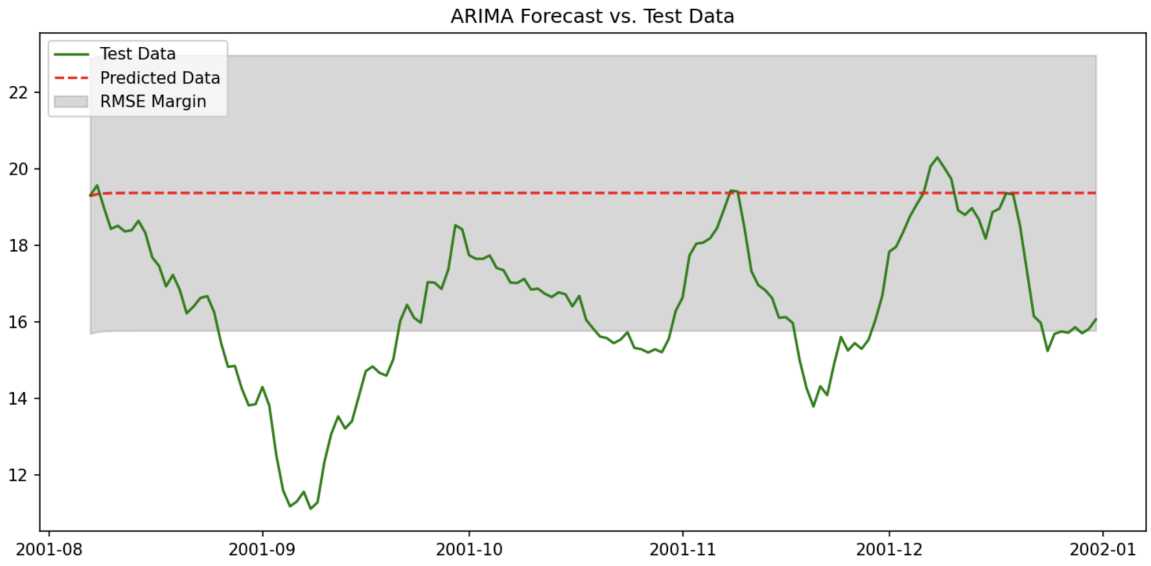

Figure 18: The annotated visualization of the forecast of the final model compared to the test set.

E3: RECOMMENDATION

The time series analysis identified a long-term upward trend in revenue but no consistent patterns beyond that. The series appears to follow a random walk with drift, suggesting that its fluctuations are likely influenced by external factors. Therefore, future analyses should incorporate additional variables to better understand these influences.

The model's low RMSE of approximately 3.6 indicates a moderately low margin of error, making it reliable for short-term predictions. The forecast length should remain within the test data range of 147 days because the prediction interval was narrow in this range, suggesting that 95% of actual values should confidently fall within it. However, the model is not recommended for long-term projections because it assumes the series will plateau and stabilize, failing to account for any future variability.

Additionally, the narrow forecast range fails to capture the natural oscillations in revenue. This is likely due to the model's simplicity because it includes only one AR term and relies only on the previous value for its predictions. Alternatively, the time series may follow a random walk with drift, so long-term predictions are inherently difficult due to the increasing random and unstructured variability [8]. Therefore, more flexible models should be explored to better capture the underlying structure of the data. Moreover, the model's effectiveness depends on the original revenue scale, and the standardization makes it difficult to interpret the significance of the error metric. The next steps should include investigating external revenue factors, incorporating additional features, destandardizing the data and RMSE for better interpretability, and exploring alternative time series models to improve the forecasting accuracy.

G: SOURCES FOR THIRD-PARTY CODE

Statsmodels, “Augmented Dickey-Fuller Unit Root Test,” *Statsmodels Documentation*, 2025. [Online]. Available: https://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.adfuller.html. [Accessed: Feb. 6, 2025].

SciPy, “Welch’s Method for Power Spectral Density Estimation,” *SciPy Documentation*, 2025. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.welch.html. [Accessed: Feb. 6, 2025].

Statsmodels, “Seasonal-Trend Decomposition using LOESS (STL),” *Statsmodels Documentation*, 2025. [Online]. Available: https://www.statsmodels.org/dev/generated/statsmodels.tsa.seasonal.STL.html. [Accessed: Feb. 7, 2025].

Statsmodels, “Autoregressive Integrated Moving Average (ARIMA) Model,” *Statsmodels Documentation*, 2025. [Online]. Available: https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html. [Accessed: Feb. 7, 2025].

Statsmodels, “Plot Predict Function for ARIMA Results,” *Statsmodels Documentation*, 2025. [Online]. Available: https://www.statsmodels.org/v0.11.1/generated/statsmodels.tsa.arima_model.ARIMAResults.plot_predict.html. [Accessed: Feb. 8, 2025].

H: SOURCES

Reference List

[1] Data Overload, “Understanding ARIMA Models: A Comprehensive Guide to Time Series Forecasting,” *Medium*, 2025. [Online]. Available: https://medium.com/@data-overload/understanding-arima-models-a-comprehensive-guide-to-time-series-forecasting-dfc7207f2406#:~:text=Assumption%20of%20Linearity%3A%20ARIMA%20models,models%20may%20be%20more%20suitable.. [Accessed: Feb. 4, 2025].

[2] R. J. Hyndman and G. Athanasopoulos, “Dealing with missing values and outliers,” *Forecasting: Principles and Practice*, OTexts, 2025. [Online]. Available: https://otexts.com/fpp3/missing-outliers.html. [Accessed: Feb. 4, 2025].

[3] R. J. Hyndman and G. Athanasopoulos, “Stationarity and differencing,” *Forecasting: Principles and Practice*, OTexts, 2025. [Online]. Available: https://otexts.com/fpp2/stationarity.html. [Accessed: Feb. 4, 2025].

[4] J. Noble and E. Kavlakoglu, “Autocorrelation,” *IBM Think*, 2025. [Online]. Available: https://www.ibm.com/think/topics/autocorrelation#:~:text=Autocorrelation%2C%20or%20serial%20correlation%2C%20analyzes,a%20value%20correlates%20with%20itself.. [Accessed: Feb. 4, 2025].

[5] GeeksforGeeks, “Autocorrelation,” *GeeksforGeeks*, 2025. [Online]. Available: https://www.geeksforgeeks.org/autocorrelation/. [Accessed: Feb. 4, 2025].

[6] R. Nau, “Nonseasonal ARIMA models,” *Duke University*, 2025. [Online]. Available: https://people.duke.edu/rnau/411arim.htm#:~:text=A%20nonseasonal%20ARIMA%20model%20is,errors%20in%20the%20prediction%20equation.. [Accessed: Feb. 5, 2025].

[7] S. Zanwar, “Residual analysis in time series,” *Medium*, 2025. [Online]. Available: https://medium.com/@ShankiiZ_/residual-analysis-in-time-series-612a450b08f5#:~:text=Residuals%20are%20simply%20the%20difference,no%20discernible%20patterns%20or%20trends.. [Accessed: Feb. 5, 2025].

[8] R. Nau, “Slides on the random walk model,” *Duke University*, 2025. [Online]. Available: https://people.duke.edu/~rnau/Slides_on_the_random_walk_model--Robert_Nau.pdf. [Accessed: Feb. 5, 2025].

[9] V., “Statistical tests to check stationarity in time series – Part 1,” *Analytics Vidhya*, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/statistical-tests-to-check-stationarity-in-time-series-part-1/. [Accessed: Feb. 5, 2025].

[10] Statsmodels, “State space models - forecasting,” *Statsmodels Documentation*, 2025. [Online]. Available: https://www.statsmodels.org/dev/examples/notebooks/generated/statespace_forecasting.html. [Accessed: Feb. 5, 2025].

[11] R. J. Hyndman and G. Athanasopoulos, “Time series decomposition,” in *Forecasting: Principles and Practice*, OTexts, 2025. [Online]. Available: https://otexts.com/fpp2/decomposition.html. [Accessed: Feb. 6, 2025].

[12] IBM, “Understanding the periodogram and spectral density,” *IBM Documentation*, 2025. [Online]. Available: https://www.ibm.com/docs/en/spss-statistics/saas?topic=periodicity-understanding-periodogram-spectral-density. [Accessed: Feb. 6, 2025].

[13] ArcGIS, “Seasonal-Trend decomposition using LOESS (STL),” *ArcGIS Insights Documentation*, 2025. [Online]. Available: https://doc.arcgis.com/en/insights/latest/analyze/stl.htm. [Accessed: Feb. 6, 2025].

[14] R. J. Hyndman and G. Athanasopoulos, “Stochastic and deterministic trends,” *Forecasting: Principles and Practice*, OTexts, 2025. [Online]. Available: https://otexts.com/fpp3/stochastic-and-deterministic-trends.html. [Accessed: Feb. 6, 2025].

[15] A. N. Kis, “Demystifying STL: Understanding seasonal decomposition of time series,” *Medium*, 2025. [Online]. Available: https://medium.com/@kis.andras.nandor/demystifying-stl-understanding-seasonal-decomposition-of-time-series-d3c50150ec12. [Accessed: Feb. 6, 2025].

[16] A. N. Kis, “Understanding autocorrelation and partial autocorrelation functions (ACF and PACF),” *Medium*, 2025. [Online]. Available: https://medium.com/@kis.andras.nandor/understanding-autocorrelation-and-partial-autocorrelation-functions-acf-and-pacf-2998e7e1bcb5. [Accessed: Feb. 6, 2025].

[17] IBM, “Spectral plots in forecasting,” *IBM Documentation*, 2025. [Online]. Available: https://www.ibm.com/docs/en/spss-statistics/saas?topic=forecasting-spectral-plots. [Accessed: Feb. 6, 2025].

[18] D. Galar and U. Kumar, “Noise,” in *eMaintenance*, ScienceDirect, 2017. [Online]. Available: https://www.sciencedirect.com/topics/engineering/spectral-noise#:~:text=Noise%20is%20generally%20distributed%20across,the%20same%20at%20all%20frequencies.. [Accessed: Feb. 7, 2025].

[19] Minitab, “Key results: ARIMA model interpretation,” *Minitab Support Documentation*, 2025. [Online]. Available: https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/time-series/how-to/arima/interpret-the-results/key-results/#:~:text=The%20null%20hypothesis%20is%20that,alpha)%20of%200.05%20works%20well.%20of%200.05%20works%20well.). [Accessed: Feb. 7, 2025].

[20] S. Banerjee, “Model magic: AIC, BIC, MDL – navigating fit and elegance,” *Medium*, 2025. [Online]. Available: https://shekhar-banerjee96.medium.com/model-magic-aic-bic-mdl-navigating-fit-and-elegance-726c784edf9b. [Accessed: Feb. 8, 2025].

[21] R. Nau, “AR and MA terms in ARIMA models,” *Duke University*, 2025. [Online]. Available: https://people.duke.edu/~rnau/411arim3.htm. [Accessed: Feb. 10, 2025].

[22] D. Abugaber, “Differencing a time series,” *University of Illinois Chicago*, 2025. [Online]. Available: https://ademos.people.uic.edu/Chapter23.html. [Accessed: Feb. 10, 2025].

[23] J. Frost, “Root Mean Square Error (RMSE),” *Statistics by Jim*, 2025. [Online]. Available: https://statisticsbyjim.com/regression/root-mean-square-error-rmse/. [Accessed: Feb. 11, 2025].

[24] E. Miyakawa, J. W. Dennis, J. Bishop, and A. Gelder, “Quantifying and visualizing forecast uncertainty with the FIFE,” *Institute for Defense Analyses*, 2025. [Online]. Available: https://www.ida.org/-/media/feature/publications/q/qu/quantifying-and-visualizing-forecast-uncertainty-with-the-fife/p-31857.ashx. [Accessed: Feb. 12, 2025].

[25] R. J. Hyndman and G. Athanasopoulos, “Prediction intervals,” *Forecasting: Principles and Practice*, 2nd ed., OTexts, 2025. [Online]. Available: https://otexts.com/fpp2/prediction-intervals.html. [Accessed: Feb. 14, 2025].

[26] N. Gupta, “Bias-variance tradeoff in time series,” *Towards Data Science*, 2025. [Online]. Available: https://towardsdatascience.com/bias-variance-tradeoff-in-time-series-8434f536387a/. [Accessed: Feb. 14, 2025].