STUDENT’S NAME: EBENEZER ESSILFIE WILSON

COURSE INSTRUCTOR’S NAME: DR. SHERIN ALY

COURSE TITLE: ADVANCED DATA ANALYTICS TASK 1 (TIME SERIES MODELING)

APRIL 1, 2025

A1.

Based on the past daily revenues generated, can we predict the revenues for the next three months using time series modeling techniques?

The dataset consists of 731 days of revenue, so I am assuming it is two years of revenue including a leap year. If that is the case, then I will begin the year 2015 through to 2016 ending.

A2.

The main objective of this project is to use time series modeling techniques like ARIMA to better understand trends within the dataset to make informed projections about revenue in the next three months.

B.

Time series models rely on several key assumptions to ensure accurate forecasting and meaningful insights. These assumptions include:

1. Stationarity

A time series is stationary if its statistical properties, such as mean, variance, and autocorrelation, remain constant over time. Many time series models, such as Autoregressive Integrated Moving Average (ARIMA), assume stationarity because non-stationary data can lead to unreliable predictions (Hyndman & Athanasopoulos, 2018). Stationarity can be tested using statistical methods like the Augmented Dickey-Fuller (ADF) test and the KPSS test. If a series is non-stationary, transformations such as differencing, log transformations, or seasonal adjustments can help stabilize it.

2. Autocorrelation

Autocorrelation measures the relationship between a time series observation and its past values. Many time series models, such as Autoregressive (AR) and ARIMA models, assume that past values influence future values (Box et al., 2015). The presence of significant autocorrelation can be detected using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots. If high autocorrelation is detected, methods such as differencing, adding lagged variables, or using appropriate models like ARIMA can address it.

3. Linearity

Most traditional time series models assume a linear relationship between past and future values. However, if the relationship is nonlinear, more advanced models such as Long Short-Term Memory (LSTM) networks or Random Forest Regression may be more appropriate (Hyndman & Athanasopoulos, 2018).

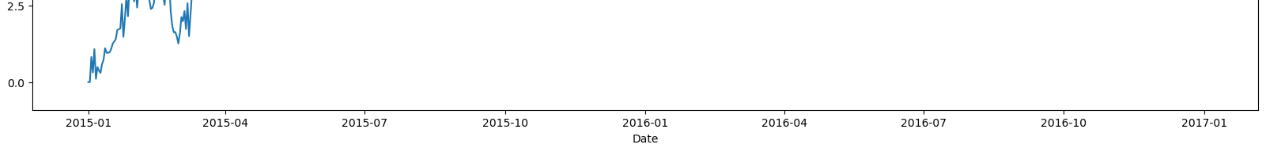
4. Homoscedasticity (Constant Variance of Errors)

A time series should have a constant variance of residuals (errors) over time. If the variance changes, the data exhibits heteroscedasticity, which can lead to biased model results. This issue is often addressed using log transformations or by applying ARCH/GARCH models for volatility forecasting (Engle, 1982).

C1.

A graph with lines and a red line

AI-generated content may be incorrect.



C2.

The dataset consists of two columns: Day and Revenue. The Day column consists of 731 days. 731 days divided by 2 gives us 365 twice plus an extra day. Ironically, this could mean a regular year (365 days) and a leap year (366 days). I am starting the year from January 1, 2015, through to the ending of the leap year 2016 with the help of .to\_datetime. A new column for the date is created. With the help of df. drop and df.set\_index, the Day column was dropped, and the dataset was sorted out nicely. Lest I forget, the revenue was in million dollars.

A screenshot of a calculator

AI-generated content may be incorrect.

A table with numbers and numbers

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

C3.

To evaluate the stationarity of the time series shown in the image, we need to check if the statistical properties (mean, variance, and autocorrelation) remain constant over time. Based on the plot in C1., here’s an initial assessment:

Upward Trend: The time series appears to have a clear upward trend over time, which suggests non-stationarity. Increasing Variability: The fluctuations in the revenue seem to become larger as the values increase, indicating possible heteroscedasticity (changing variance over time). Mean Shift: The red dashed line (trend line) confirms that the mean of the series is changing, violating the assumption of stationarity. I am going to use the ADF Test to assess the stationarity of the dataset.

**Augmented Dickey-Fuller (ADF) Test:**

* Null Hypothesis: The time series has a unit root (i.e., it is non-stationary).
* Alternative Hypothesis: The time series is stationary.
* If the p-value is greater than 0.05, we fail to reject H0H\_0H0​, meaning the series is likely non-stationary

A screenshot of a computer

AI-generated content may be incorrect.

Since the p-value (0.3206) is greater than 0.05, we fail to reject the null hypothesis, meaning the data is likely non-stationary. Additionally, the ADF statistic is not lower than the critical values at any significance level, further confirming non-stationarity.

C4.

First, we read the dataset into a Pandas DataFrame. We assume the dataset contains a time series with a date column. Before splitting the data, we check for missing values and handle them appropriately. To ensure reliable forecasting, we check whether the data is stationary using the ADF test. To evaluate the model’s performance, we divide the dataset into training (80%) and test (20%) sets. If the ADF test indicates non-stationarity, we apply differencing.

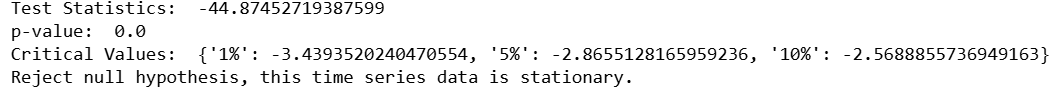
A screenshot of a computer

AI-generated content may be incorrect.

p-value before differencing

A close-up of a computer code

AI-generated content may be incorrect.

****

This is the p-value after I chose to difference the data

C5. See attached

D1.

A screen shot of a graph

AI-generated content may be incorrect.

A screen shot of a graph

AI-generated content may be incorrect.

This graph shows that the data is volatile, fluctuating around a mean of zero. This indicates a stationary series.

A graph with blue lines

AI-generated content may be incorrect.

A graph showing a number of numbers

AI-generated content may be incorrect.

Initial observations reveal frequent fluctuations suggesting natural irregularities in the datasets. A significant upward trend emerges towards the end indicating a potential shift or anomaly. The pattern implies possible seasonal behavior.

**A screen shot of a graph

AI-generated content may be incorrect.**

**A graph with blue dots

AI-generated content may be incorrect.**

The ACF plot has negative spike at lag 1 which indicates a strong negative autocorrelation at that lag. The quick decay to near-zero values in subsequent lags suggest the series is stationary or close to it. This pattern aligns well with a Moving Average (MA) process, likely of order 1 (MA(1)).

The PACF plot has spike at lag 1 and near-zero values afterward and this suggests an autoregressive relationship. The rapid cutoff after lag indicates an AR process of order 1 (AR(1)).

These plots suggest the series could be modeled using an ARMA(1,1) or ARIMA(1,0,1) model if differencing was applied already

A screen shot of a computer

AI-generated content may be incorrect.

A graph of a frequency

AI-generated content may be incorrect.

The plot shows high spectral density values at a wide range of frequencies, indicating the presence of high-frequency components. This suggests rapid fluctuations in the data, which could signify noise or periodic variations with short cycles. The density remains relatively consistent, with minor peaks and troughs throughout. This is characteristic of a process with a lot of random variability or white noise. The spectral density plot reveals a dataset dominated by high-frequency components and potential noise. There is no clear, strong cyclical pattern, suggesting a lack of consistent seasonality. Further decomposition or filtering techniques may be needed to extract meaningful patterns if present.

**A screenshot of a graph

AI-generated content may be incorrect.**

**A blue and white graph

AI-generated content may be incorrect.**

The decomposition reveals a weak trend, a distinct seasonal component, and well-behaved residuals. The seasonal pattern confirms periodicity in the dataset, while the trend component suggests minor long-term changes. The residuals indicate successful decomposition, with minimal remaining structure or autocorrelation. Overall, the series is likely suitable for forecasting with time series models like SARIMA or Exponential Smoothing**.**

**A screen shot of a computer screen

AI-generated content may be incorrect.**

**A graph with numbers and lines

AI-generated content may be incorrect.**

The residual plot indicates a well-fitted model, with errors behaving as expected. However, a formal residual analysis can confirm this observation.

D2.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a data sheet

AI-generated content may be incorrect.

A white background with black text

AI-generated content may be incorrect.

The stepwise search process tested different ARIMA models and selected ARIMA(1,0,0)(0,0,0)[0] as the best model based on the lowest Akaike Information Criterion (AIC = 983.122). This model is a pure autoregressive (AR) model with one lag (AR(1)) and no differencing (I=0) or moving average components (MA=0). The selected model does not incorporate explicit seasonality (SARIMA components are all 0). The decomposition plots show some trend in the data, indicating possible non-stationarity. The seasonal component suggests periodic fluctuations, but the best model found does not explicitly account for seasonality.

If further diagnostics indicate non-stationarity, adding differencing (I=1) could improve performance (e.g., ARIMA(1,1,0)). If significant seasonal patterns exist, a SARIMA model (e.g., ARIMA(1,0,0)(1,0,0)[s]) may be more appropriate. There needs to be verification if residuals are white noise (uncorrelated and normally distributed) to ensure model adequacy.

D3.

A screenshot of a computer

AI-generated content may be incorrect.

A close-up of a white background

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A computer screen shot of a person

AI-generated content may be incorrect.

A graph with a line

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A green line graph on a white background

AI-generated content may be incorrect.

D4.

A screenshot of a computer screen

AI-generated content may be incorrect.

A graph of a graph with a red line

AI-generated content may be incorrect.

D5.

See attached in the submissions

E1.

Model Selection: ARIMA(1,0,0)

I selected an ARIMA(1,0,0) model based on your analysis—likely using autocorrelation and partial autocorrelation plots or an auto ARIMA function to identify optimal values for (p, d, q):

p = 1: The model includes one autoregressive term.

d = 0: No differencing was needed, indicating the data was stationary.

q = 0: No moving average component was included.

This model choice seems justified given the low AIC (983.122) and BIC (996.901), which balance goodness of fit and complexity.

2. Prediction Interval of the Forecast

My forecast includes 95% confidence intervals:

These are represented by mean\_ci\_lower and mean\_ci\_upper bounds in the forecast output. The shaded green region in my plot visualizes uncertainty around the point forecasts. This helps communicate forecast reliability—narrower bands indicate more confident predictions.

3. Forecast Length Justification

The model forecasts 30 days ahead initially and then over a test set matching the holdout period. This forecast window is appropriate for short-term revenue forecasting in daily time series. It maintains accuracy without extrapolating too far into uncertain territory.

4. Model Evaluation Procedure and Error Metric

Evaluation Method:

I split the data into training and testing subsets. The model was fitted on training data and tested on unseen data to assess performance.

Error Metric:

I used Mean Absolute Error (MAE):

A black and yellow numbers and a black line

AI-generated content may be incorrect.

This indicates the average daily forecast error is about 0.38 million units of revenue, which seems reasonable based on the scale of the data.

Residual Diagnostics:

Residuals appear randomly scattered around zero (as seen in the standardized residual plot). The histogram of residuals closely follows a normal distribution. These diagnostics support that the residuals are roughly white noise → the model is well-fitted.

E2.

A screenshot of a computer code

AI-generated content may be incorrect.

A graph with blue and red lines

AI-generated content may be incorrect.

E3.

. Use the Forecast to Drive Short-Term Financial Planning

The ARIMA model provides reliable 30-day revenue projections. Use these projections for Cash flow planning, Inventory management, Staffing and resource allocation.

Monitor Forecast Accuracy Weekly

Continuously compare forecasted vs. actual revenue to monitor performance. If forecast errors increase, consider retraining or re-evaluating the model (Hodson, 2022).

Incorporate Seasonality for Medium-Term Forecasts

While ARIMA(1,0,0) works well short term, it doesn’t capture seasonality. For quarterly planning or marketing campaigns, consider upgrading to SARIMA or Prophet, especially if: There are weekly/monthly trends and Events or promotions drive revenue spikes.

F. See attached files.

G. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. Wiley.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*(4), 987-1007.

H.

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development, 15, 5481–5487. <https://doi.org/10.5194/gmd-15-5481-2022>

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.

I.