Jeffery Bills

D213 Task 1

### **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.

What is the anticipated revenue for the upcoming quarter based on previous revenue?

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

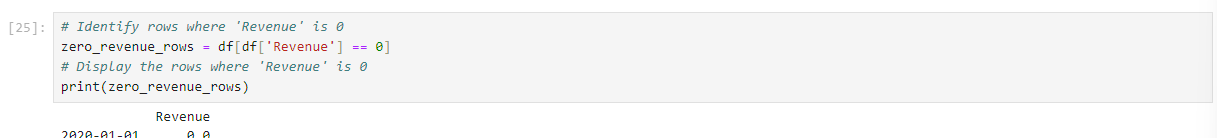
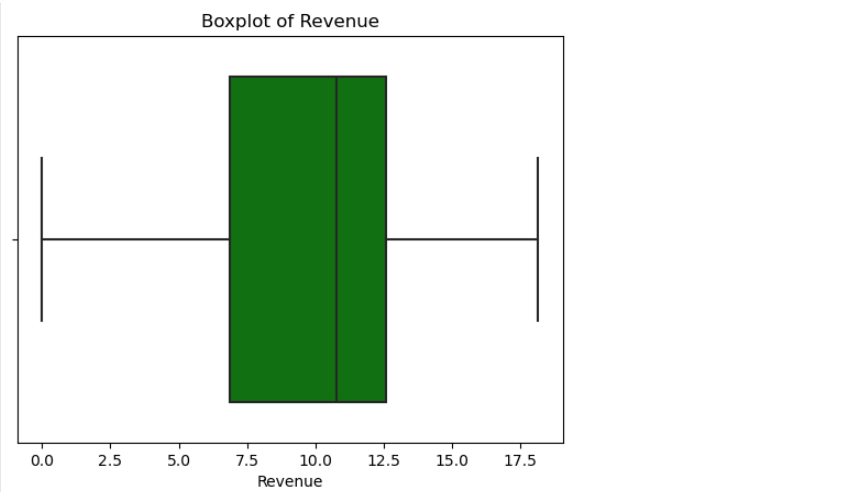
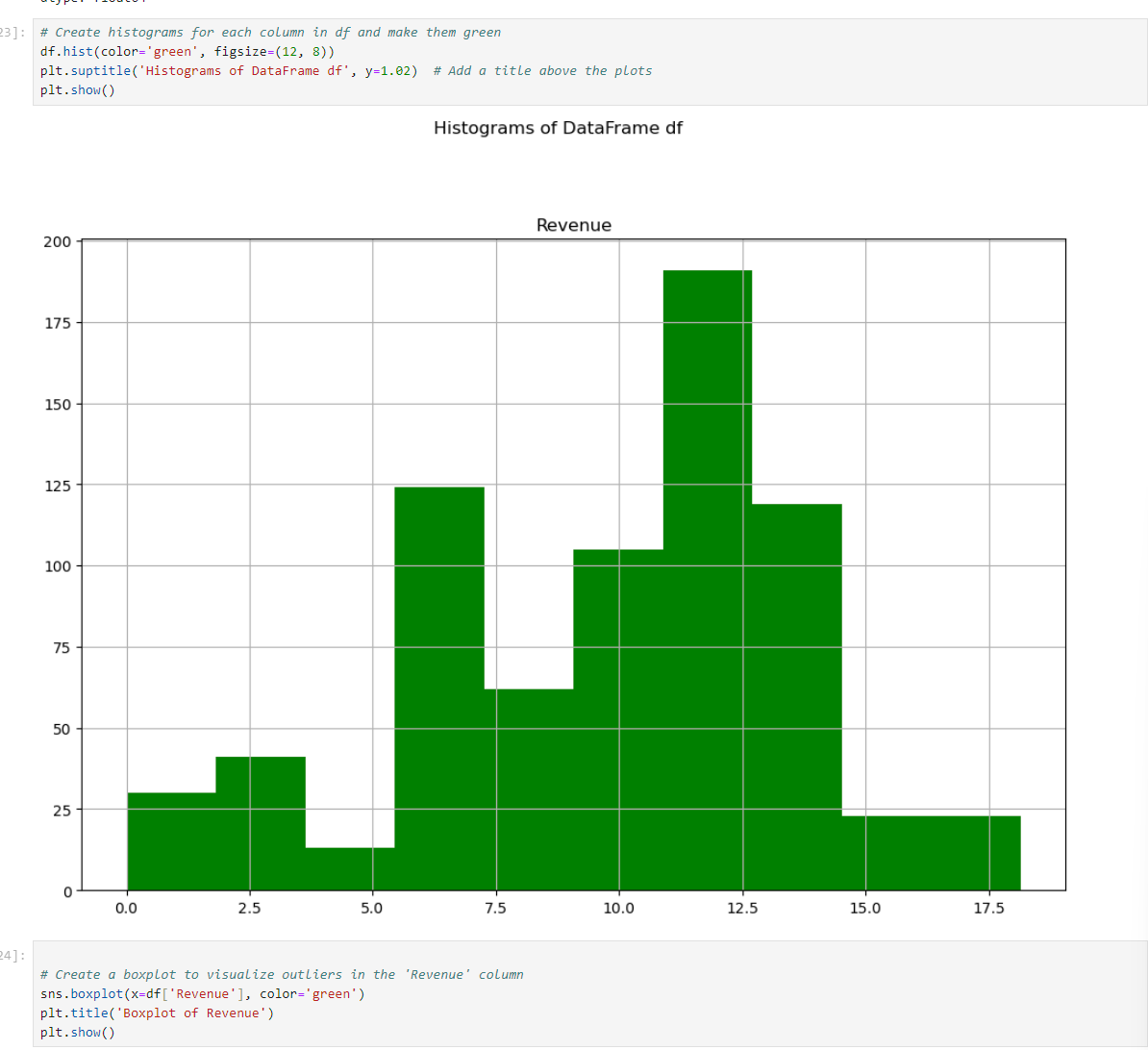
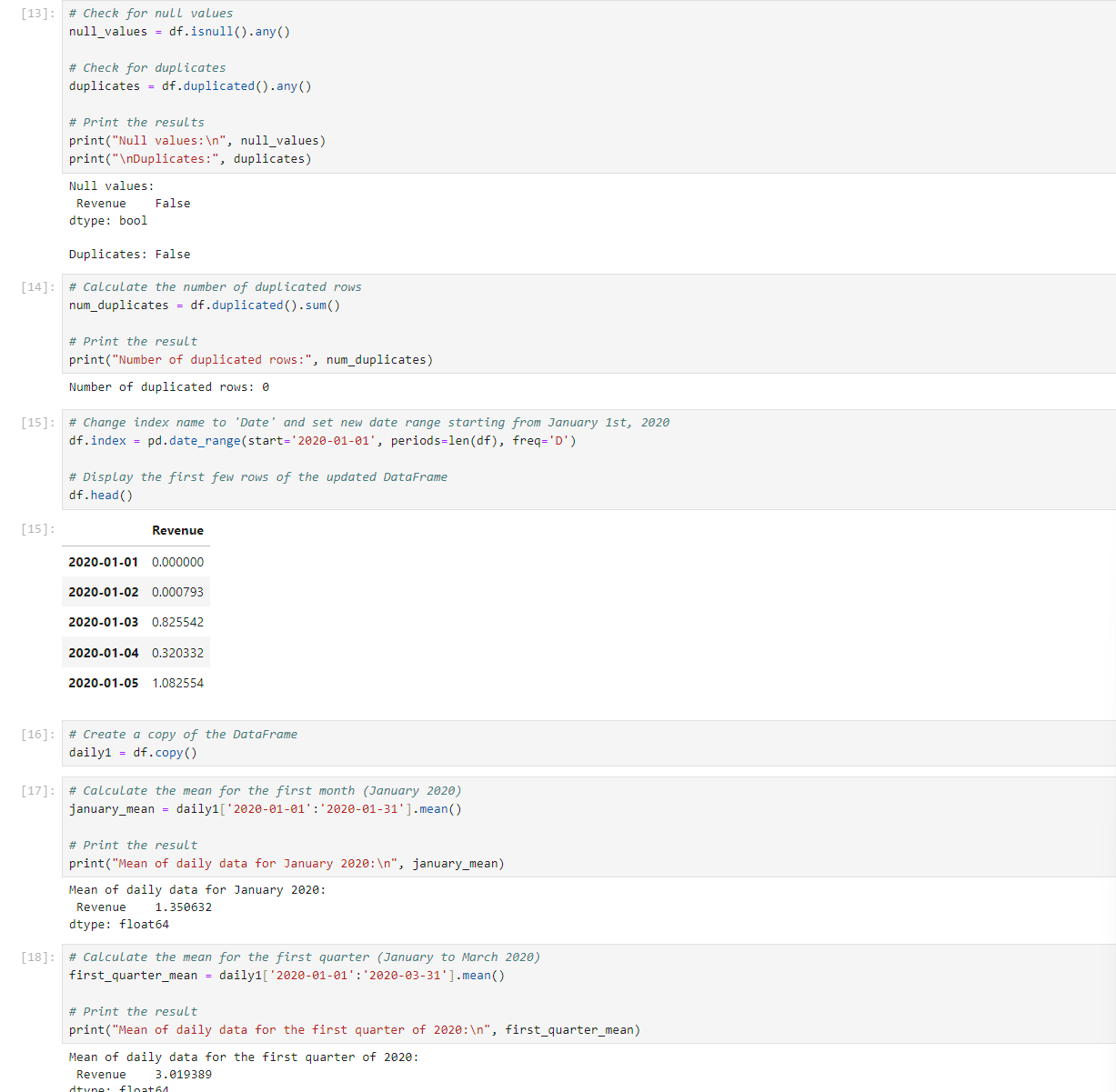
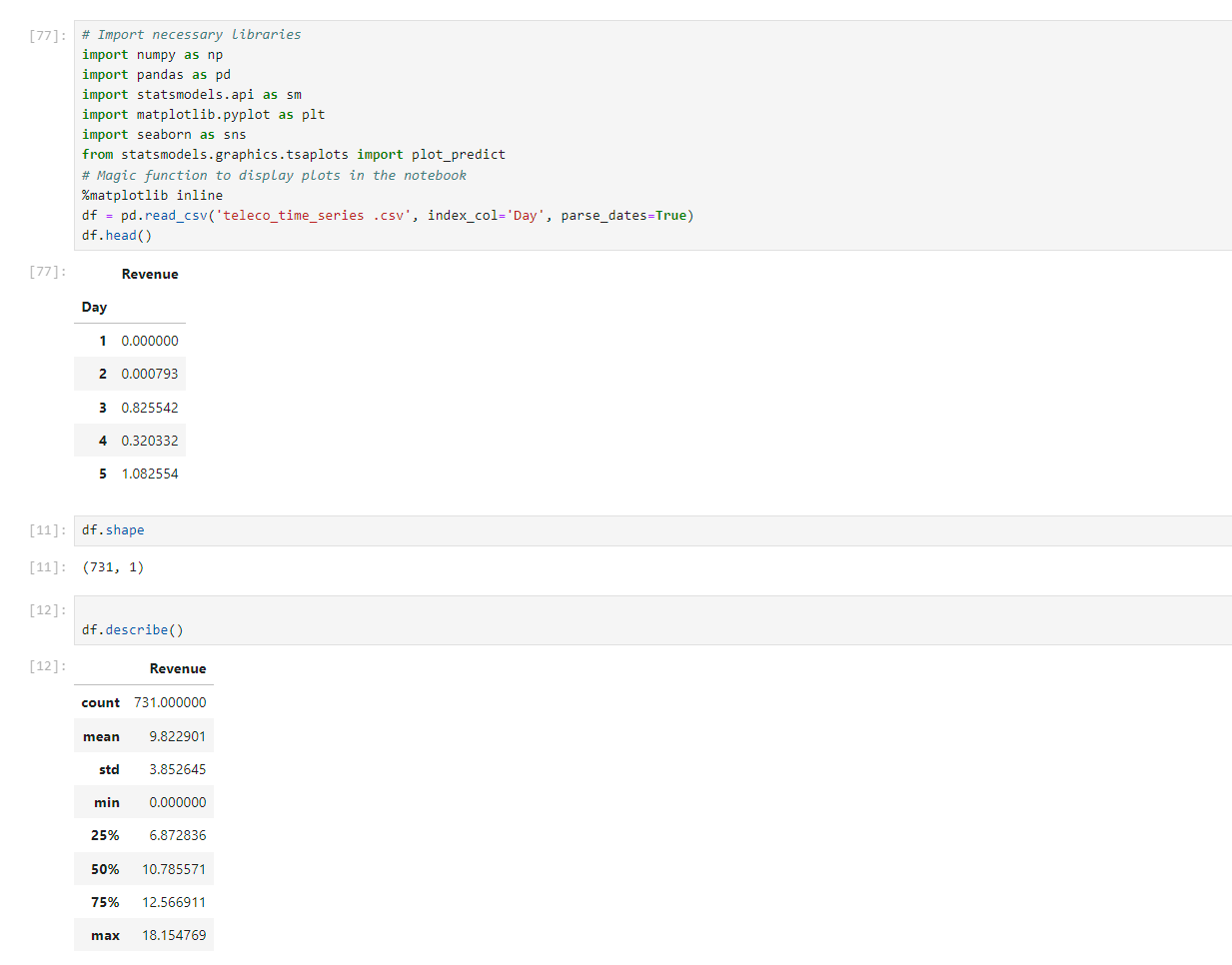
I can forecast next quarter's revenue by leveraging multiple instances of time series data. Ensuring data stationarity allows me to apply the ARIMA model for predictions based on daily data.

### **Part II: Method Justification**

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

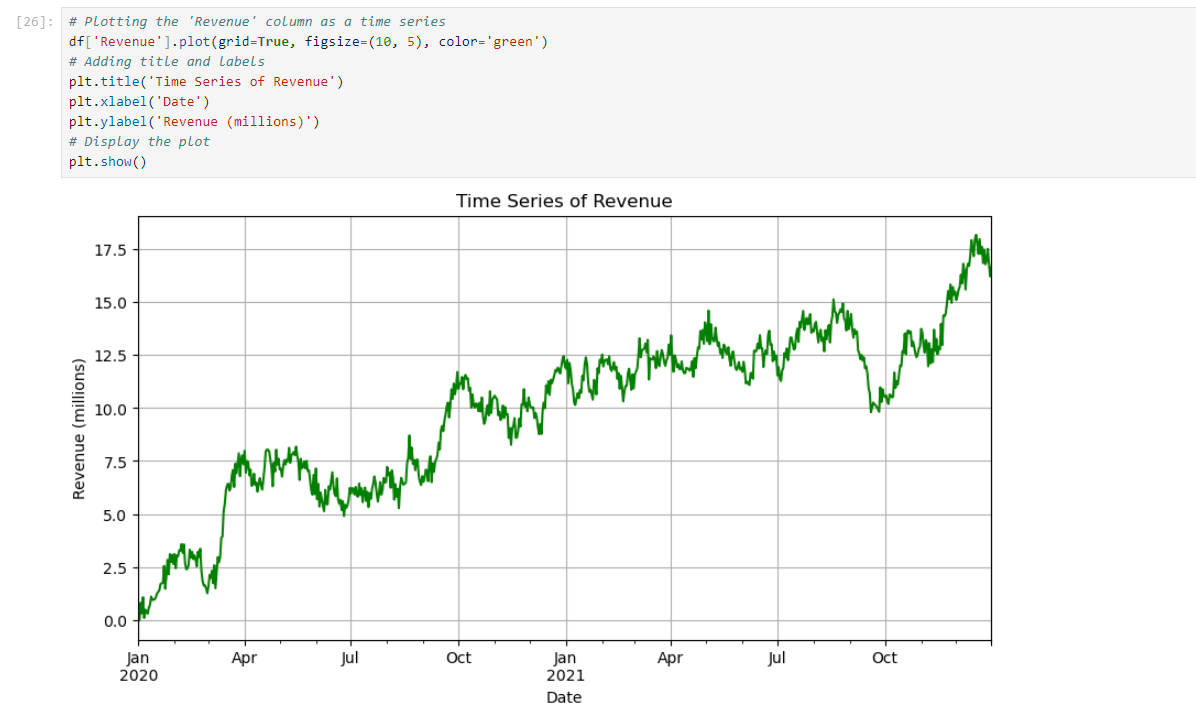
The distribution of time series data is commonly assumed to be Gaussian, or normally distributed. For a time series to be considered stationary, its mean, variance, and autocorrelation structure must remain consistent across selected time intervals (e.g., day, month, year).Autocorrelation refers to the long-term correlation within the data itself. Visualizing this can be achieved through an autocorrelation plot, while quantifying it involves calculating the autocorrelation coefficient. This coefficient measures the correlation between the data and its lagged versions and is interpreted similarly to a Pearson correlation coefficient.

### **Part III: Data Preparation**



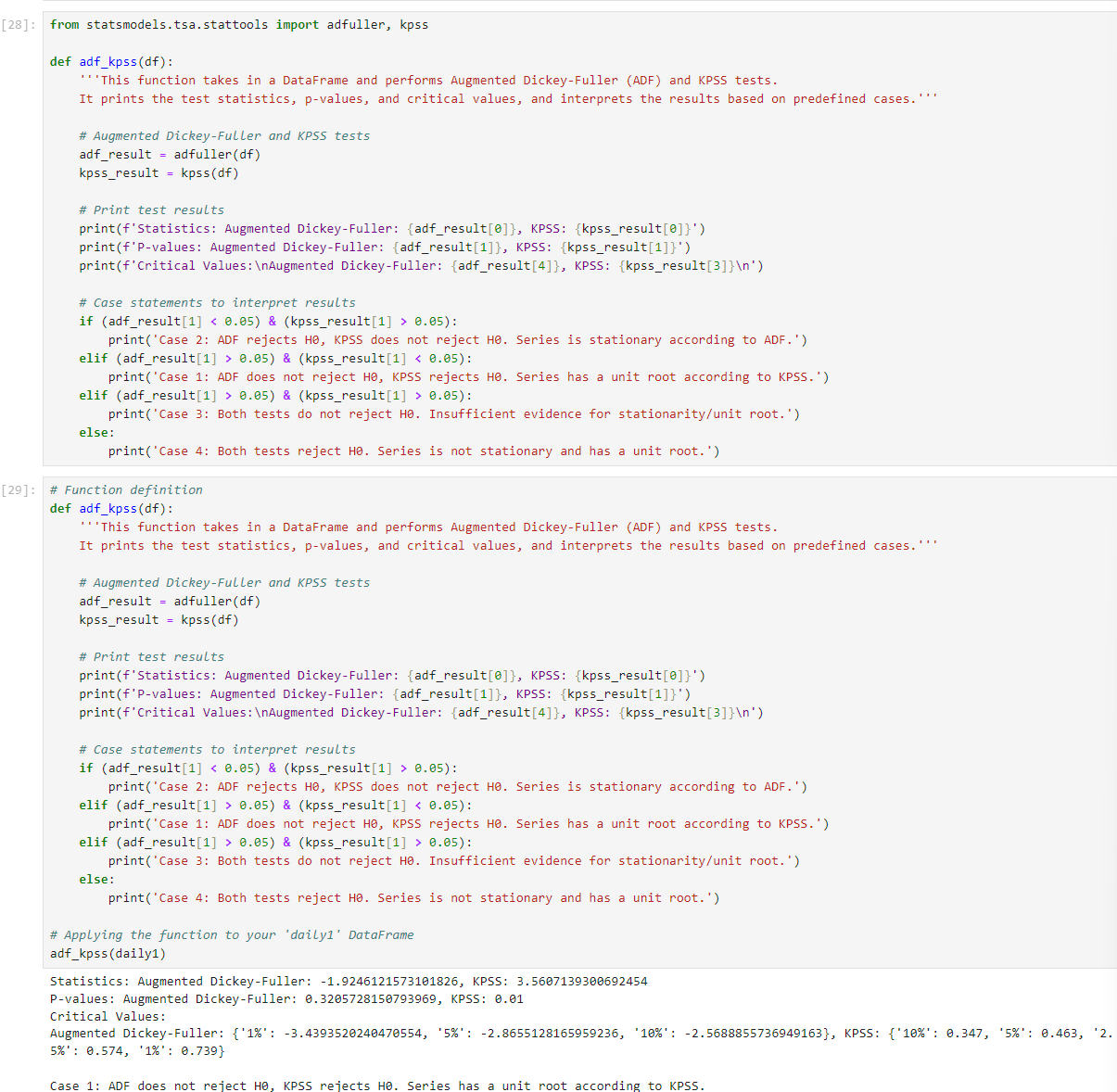
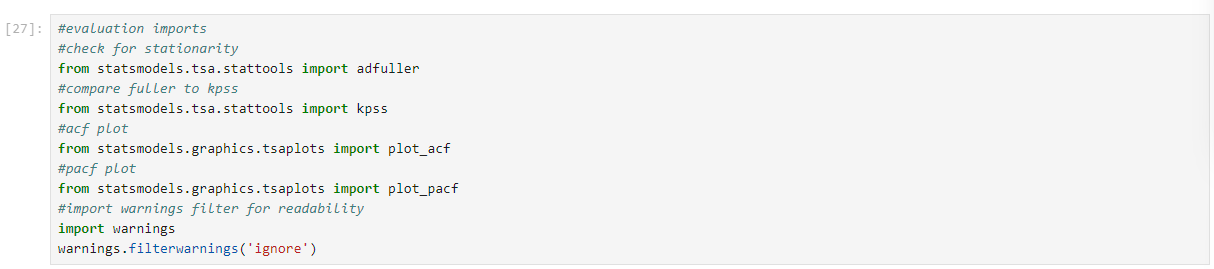
### **C. Summarize the data cleaning process by doing the following:**

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

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

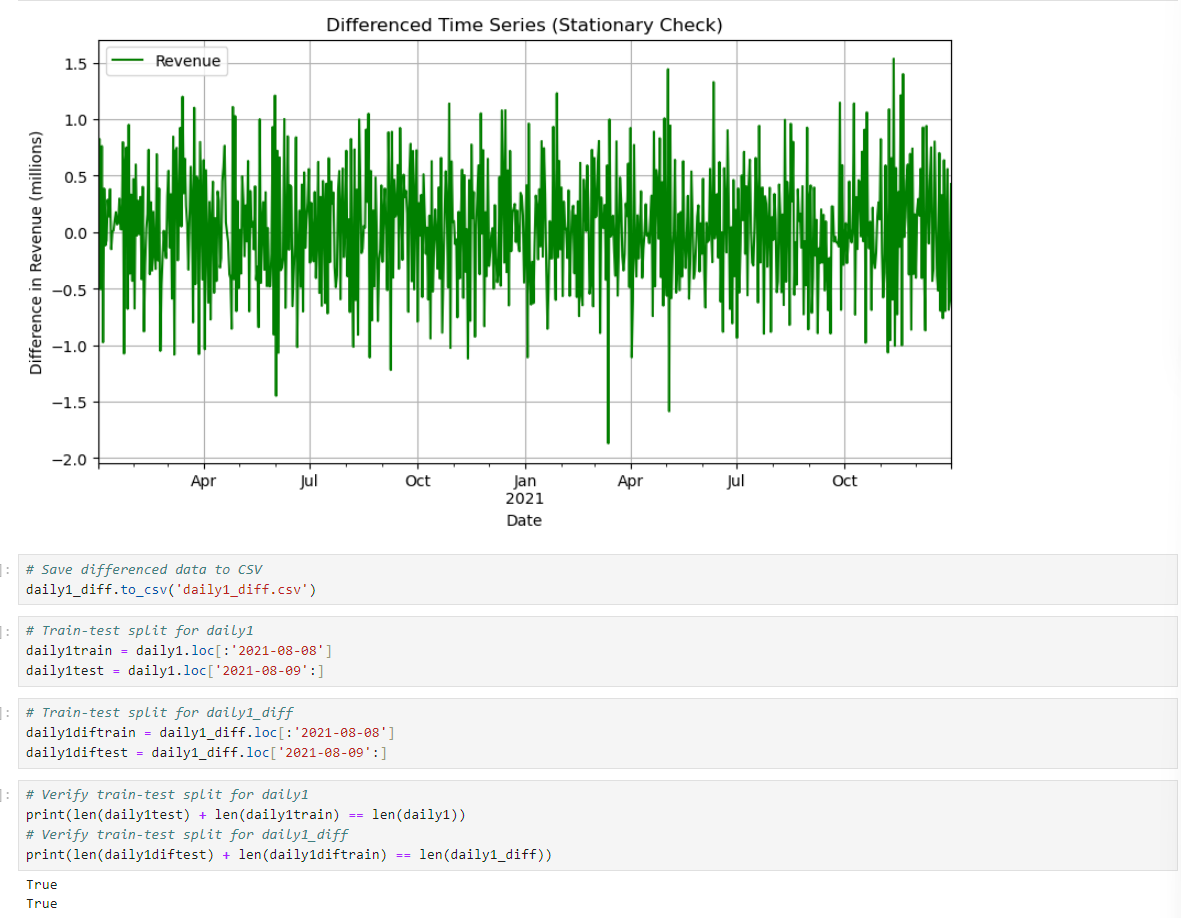
The dataset has no null values or duplicate entries. There is no need for gap treatment or preprocessing measures.

3.Evaluate the stationarity of the time series.

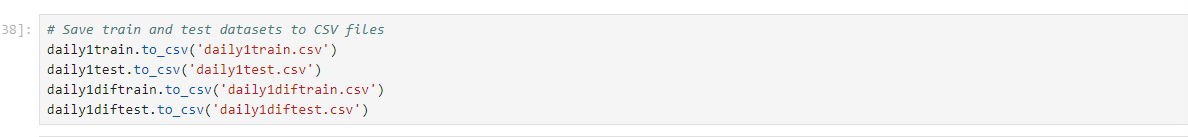


I will be utilizing an Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test at a significance level of α = 0.05.This will allow me to check if both the daily and monthly steps are stationaryAs shown above you can see that daily1 is not stationary. I will now proceed to makew the data stationary and then I will rre run the ADF and KPSS test again to verify its stationary status.

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

After achieving stationarity resembling white noise, the next step involves splitting the data into training and test sets. The training set will use approximately 80% of the data, with the test set using the remaining 20%. The data was initially imported from a CSV file using pandas and converted into datetime format. Subsequently, it was resampled into daily and monthly dataframes. Differencing was applied to ensure stationarity in these datasets. Due to insufficient observations, the quarterly and yearly dataframes were excluded from further analysis. The remaining useful data (daily and monthly) were then partitioned into training and test sets. The training sets will be utilized for model training purposes, while the test sets will assess model performance.

### **5. Provide a copy of the cleaned dataset.**



### **Part IV: Model Identification and Analysis**

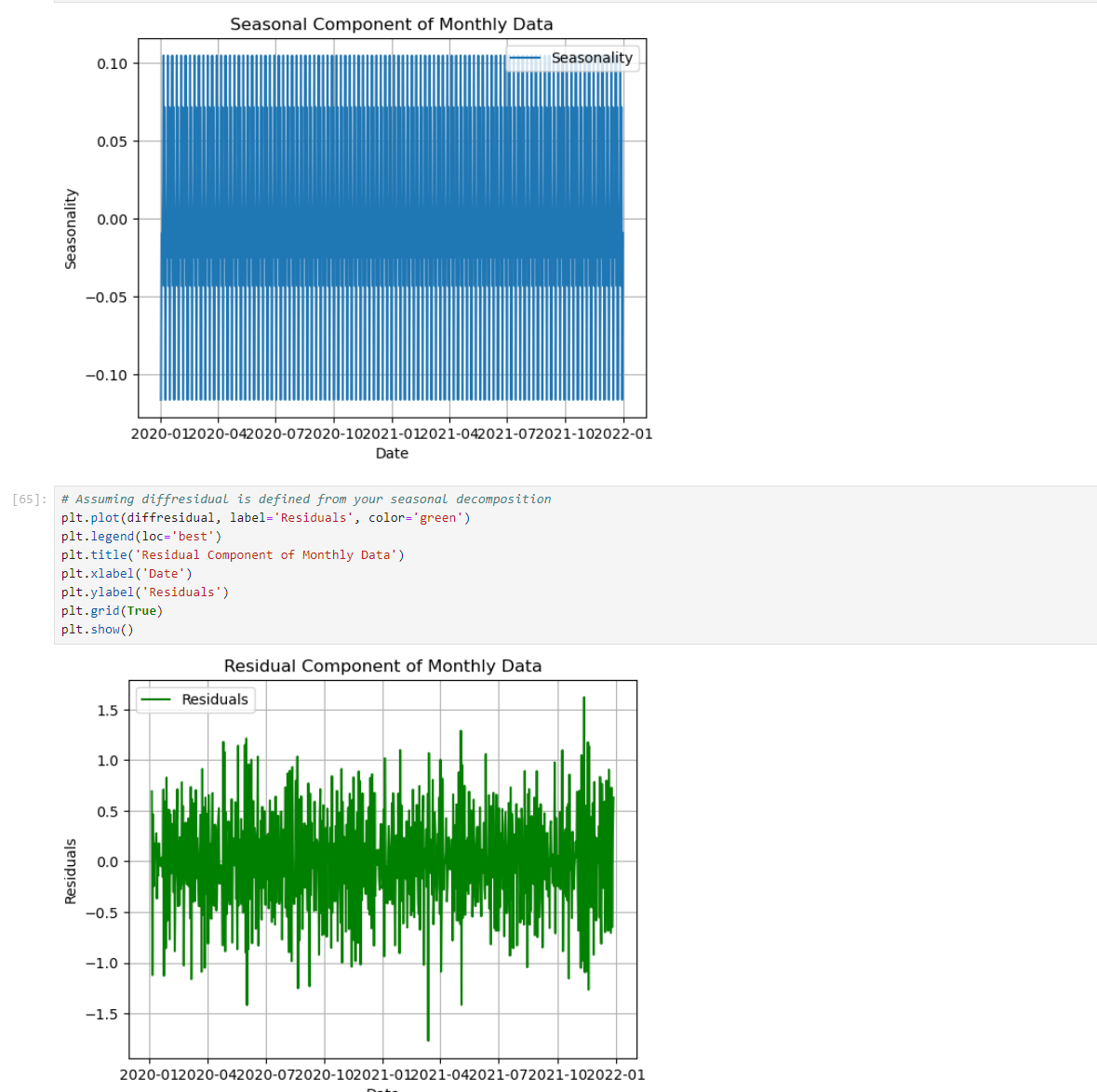
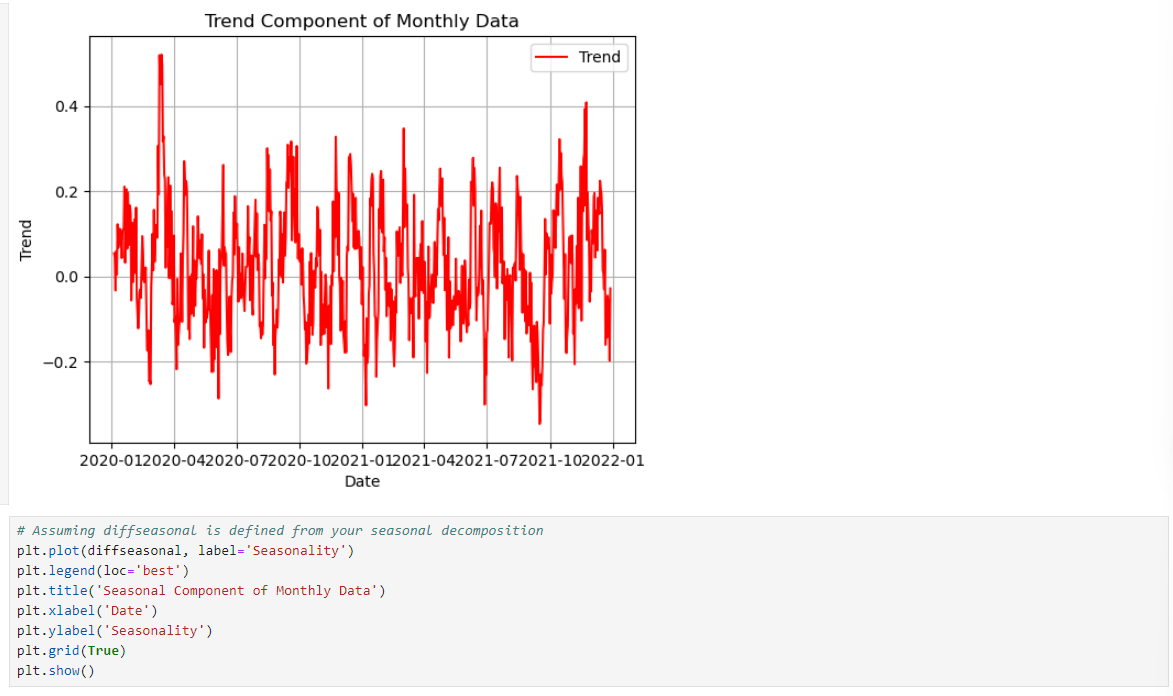
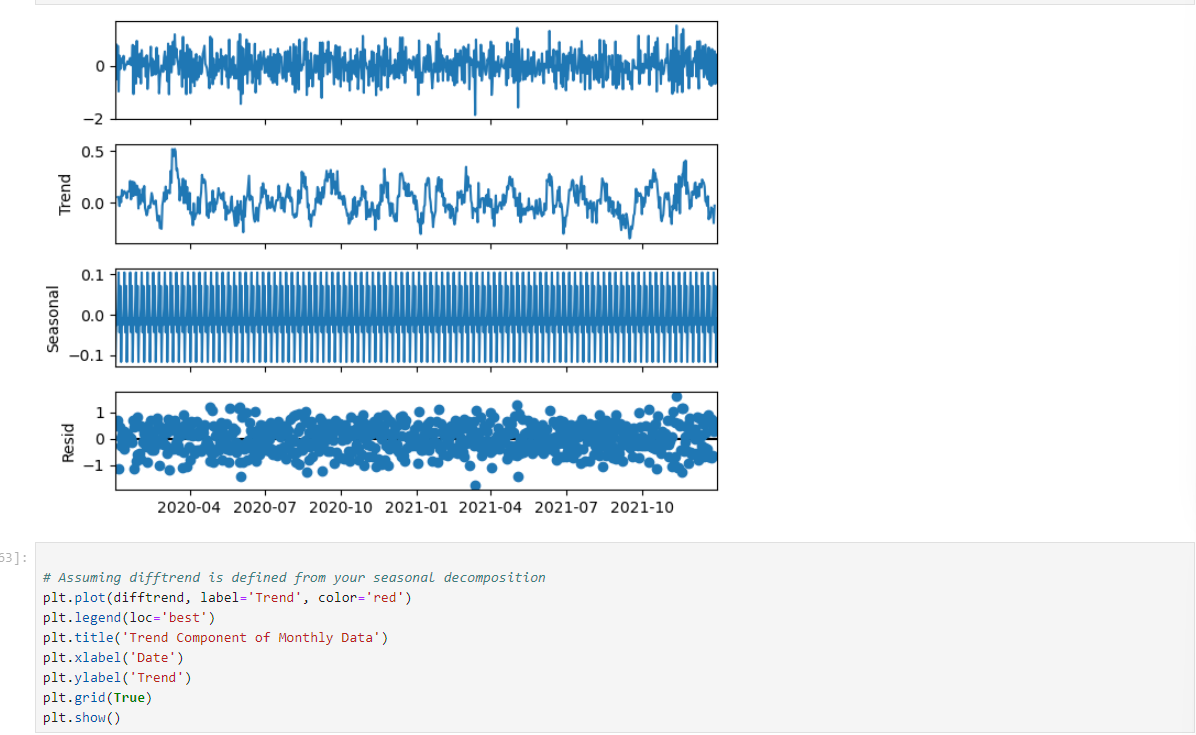
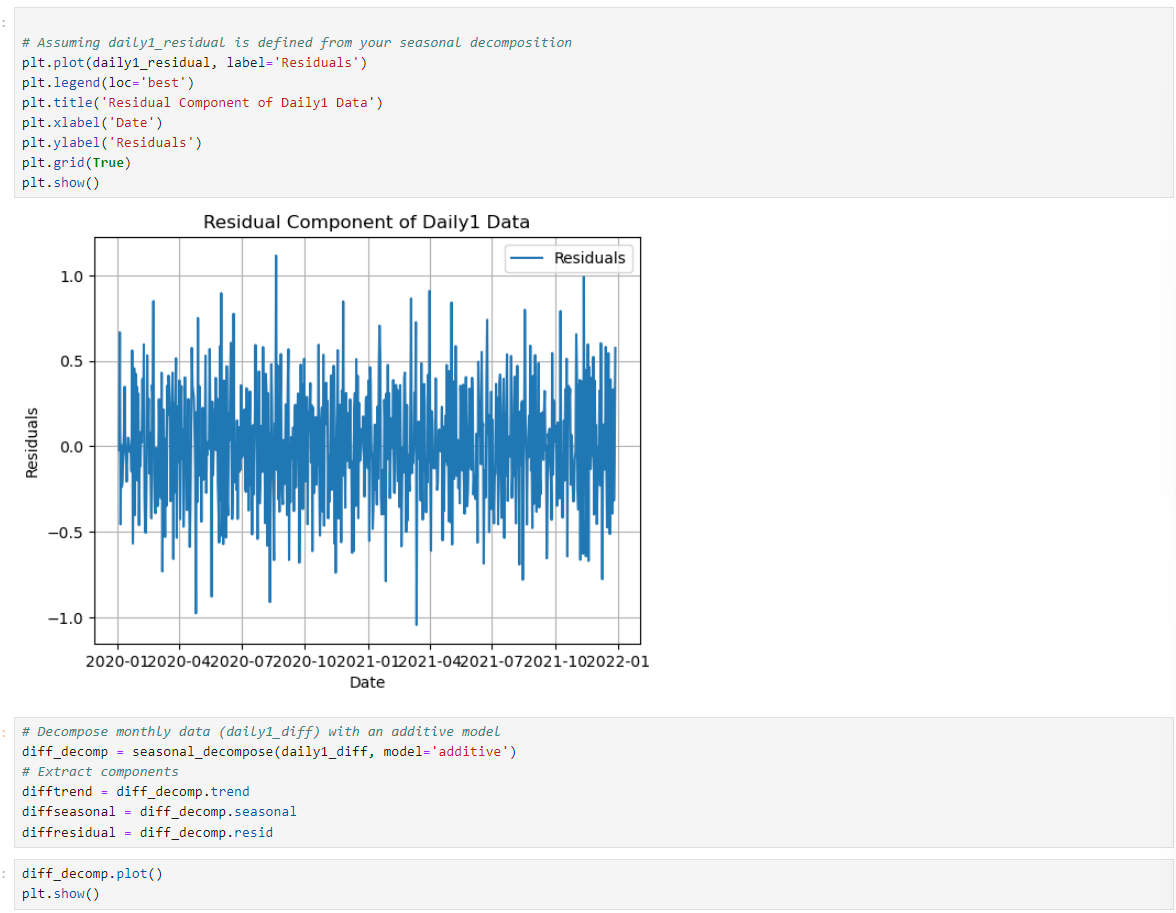
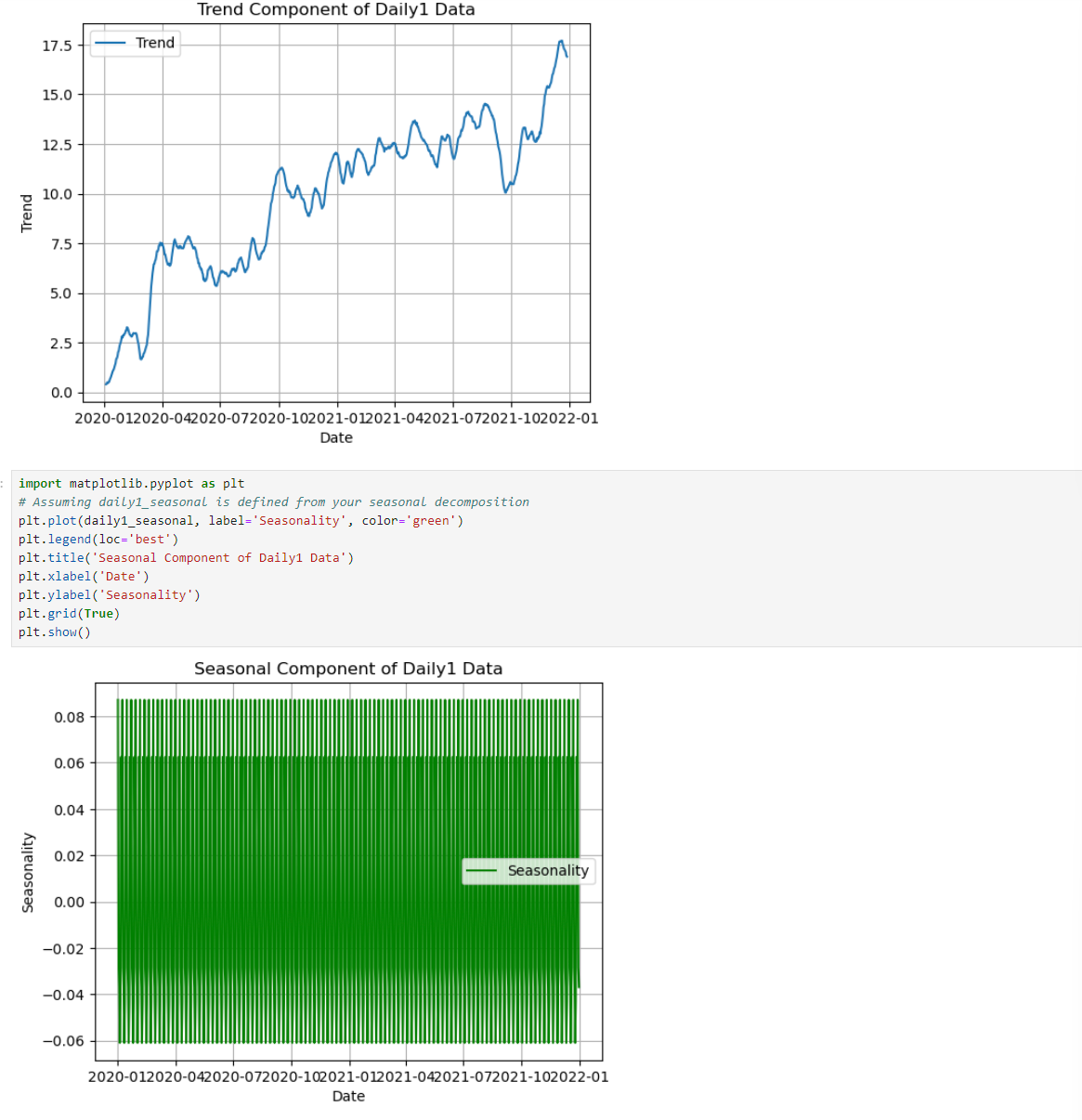
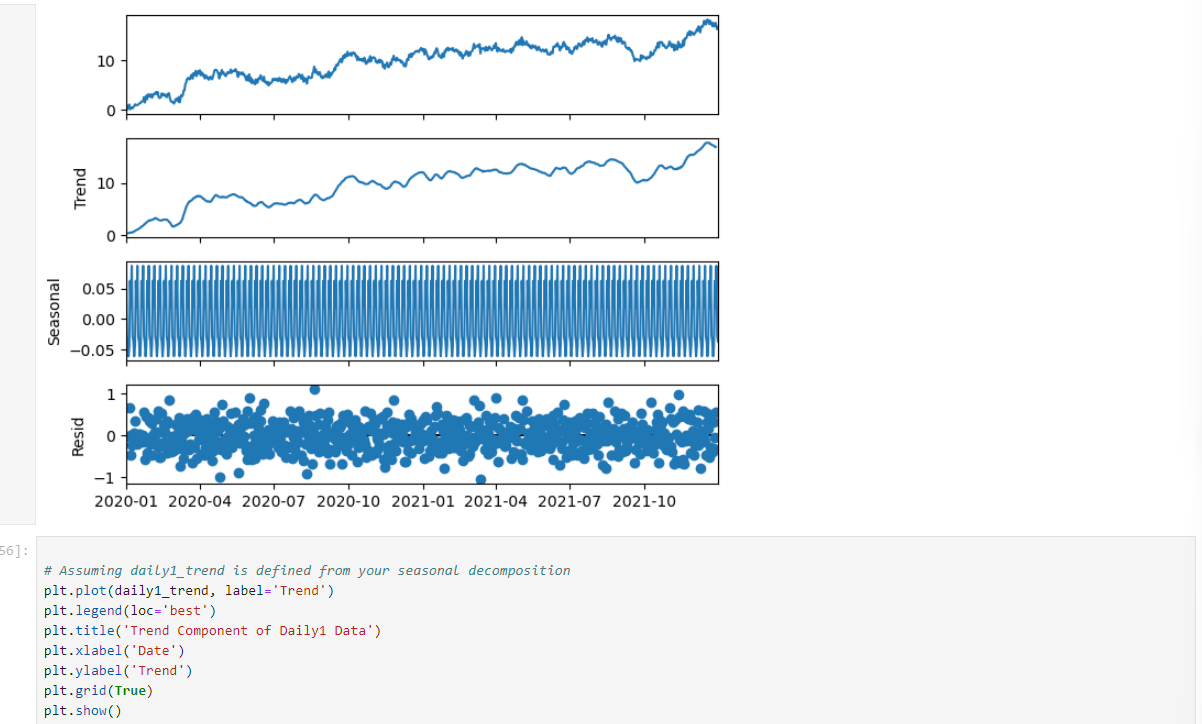
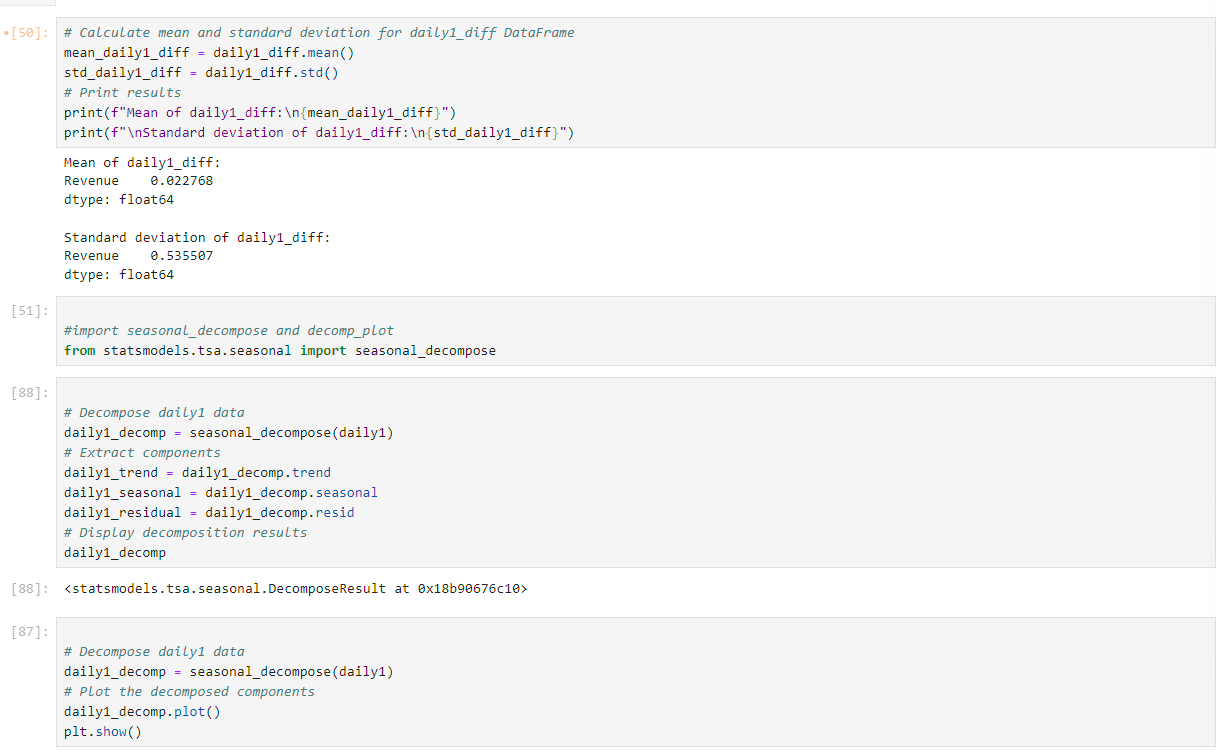
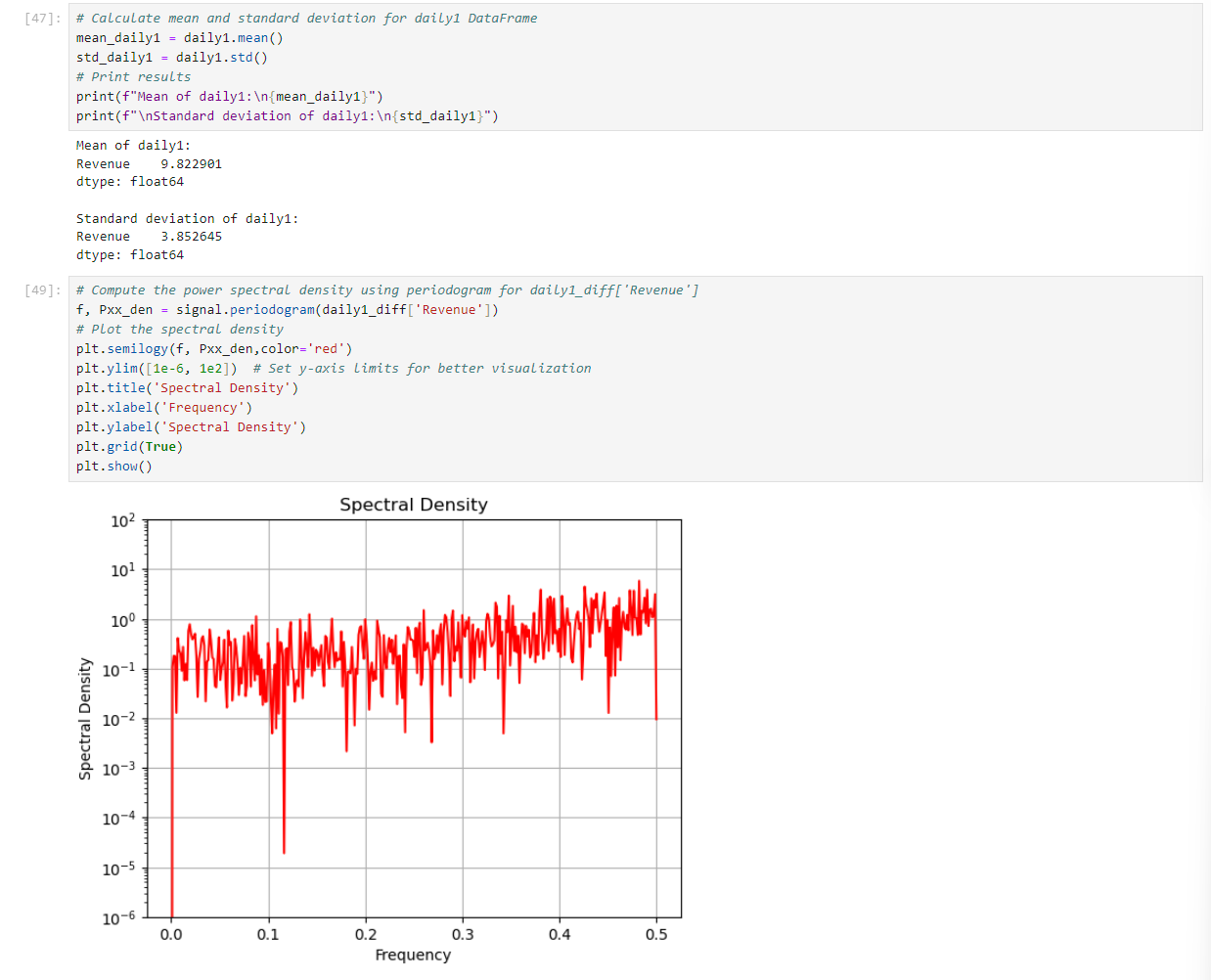
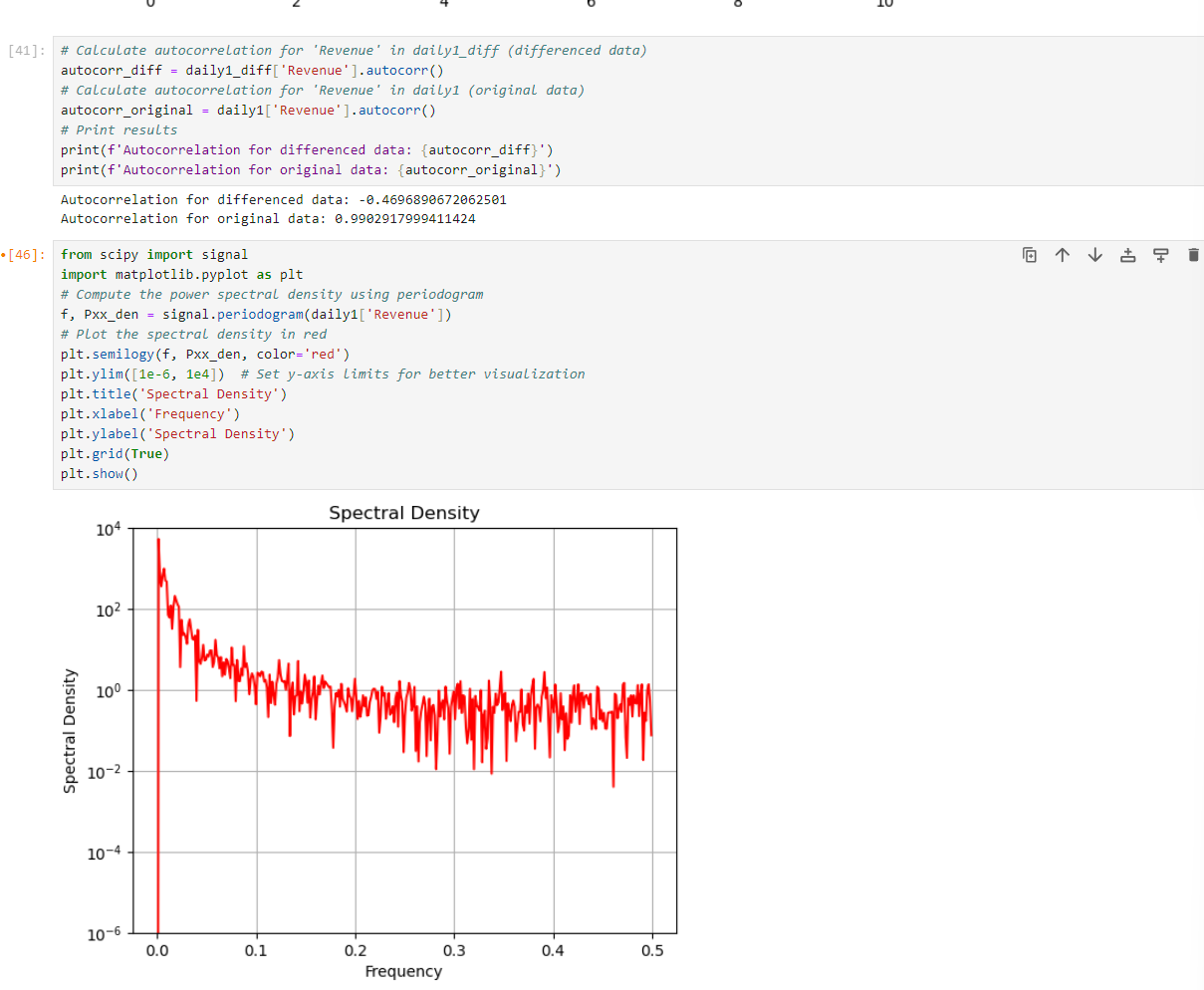
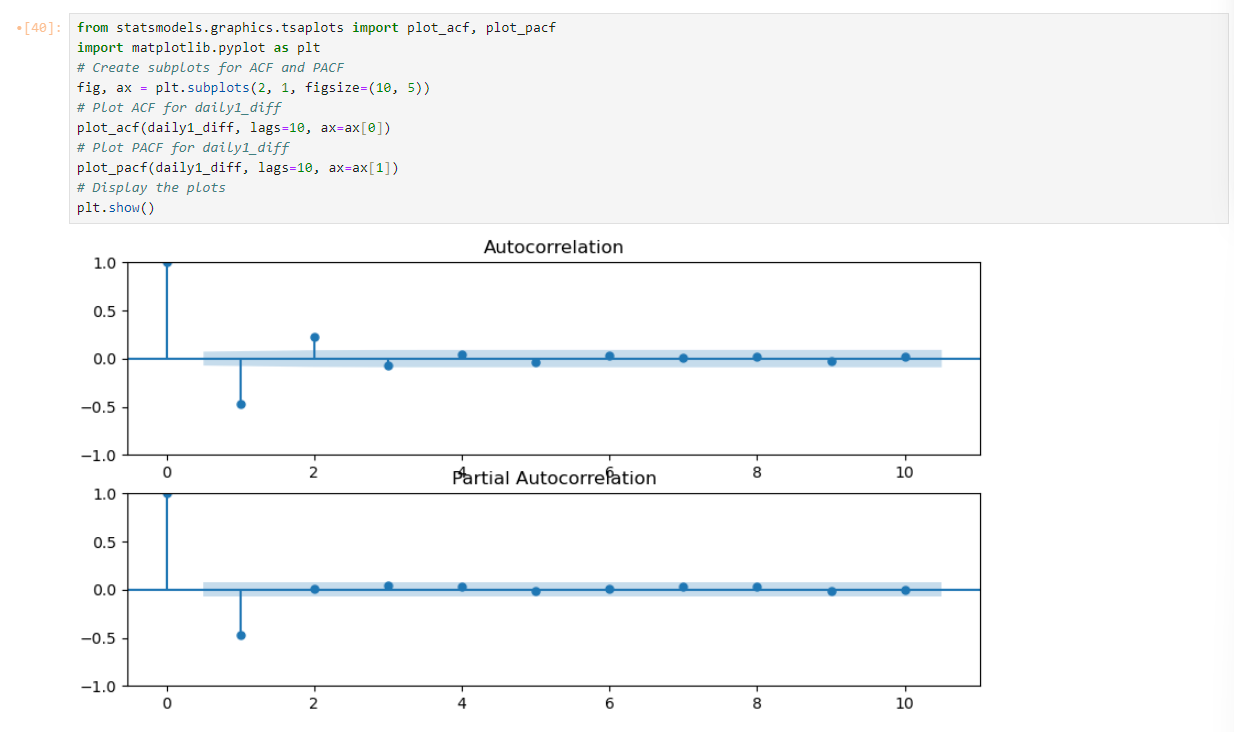
#### **D. Analyze the time series dataset by doing the following:**

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

• the presence or lack of a seasonal component

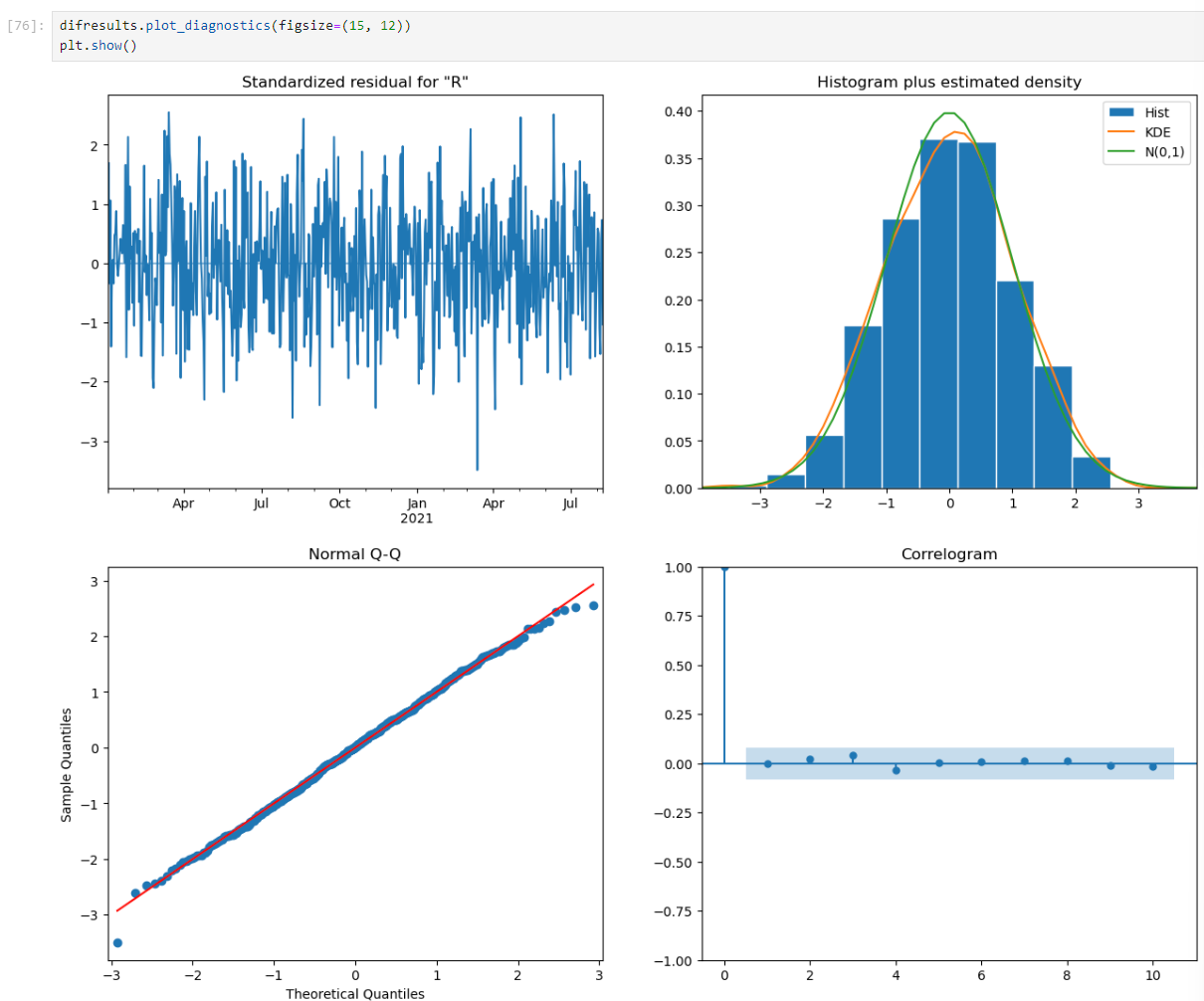
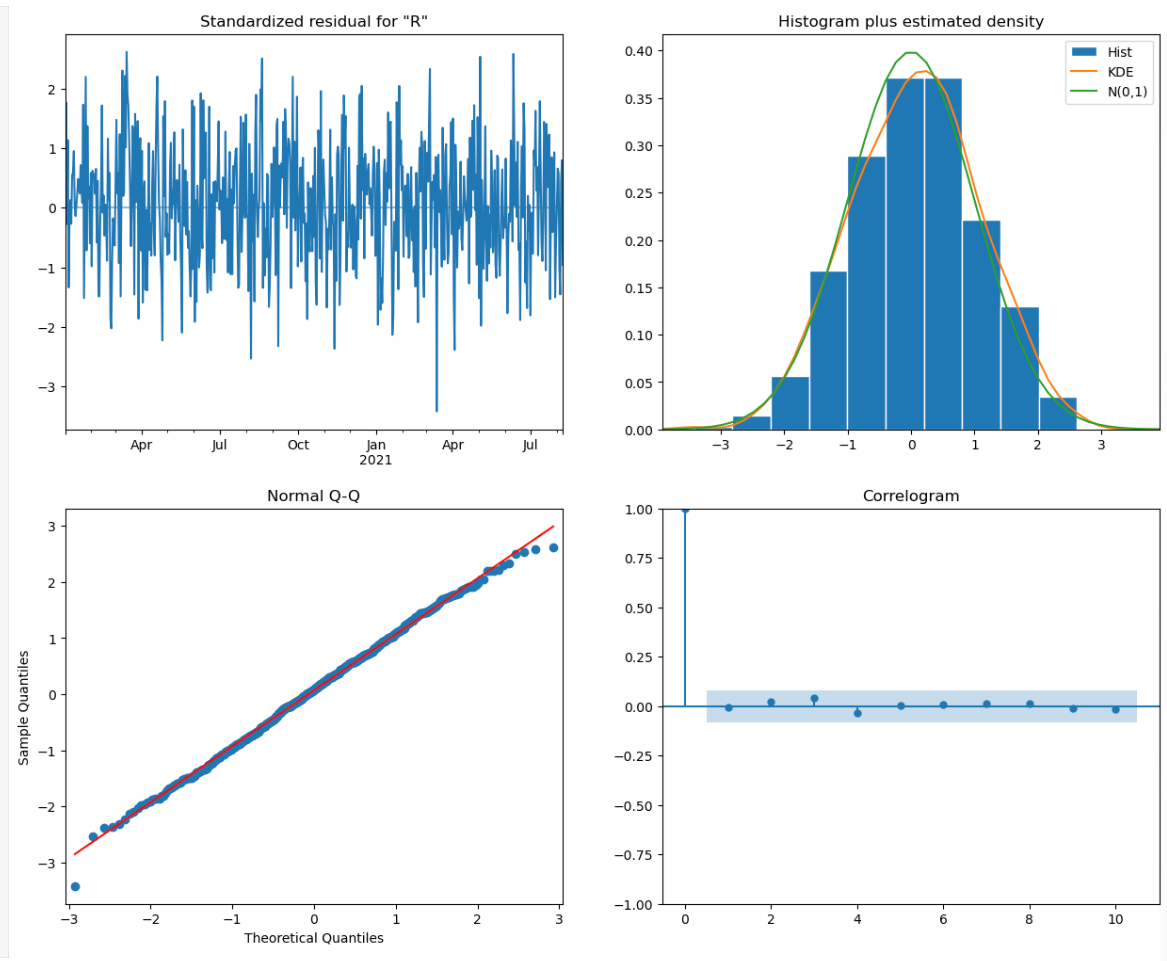
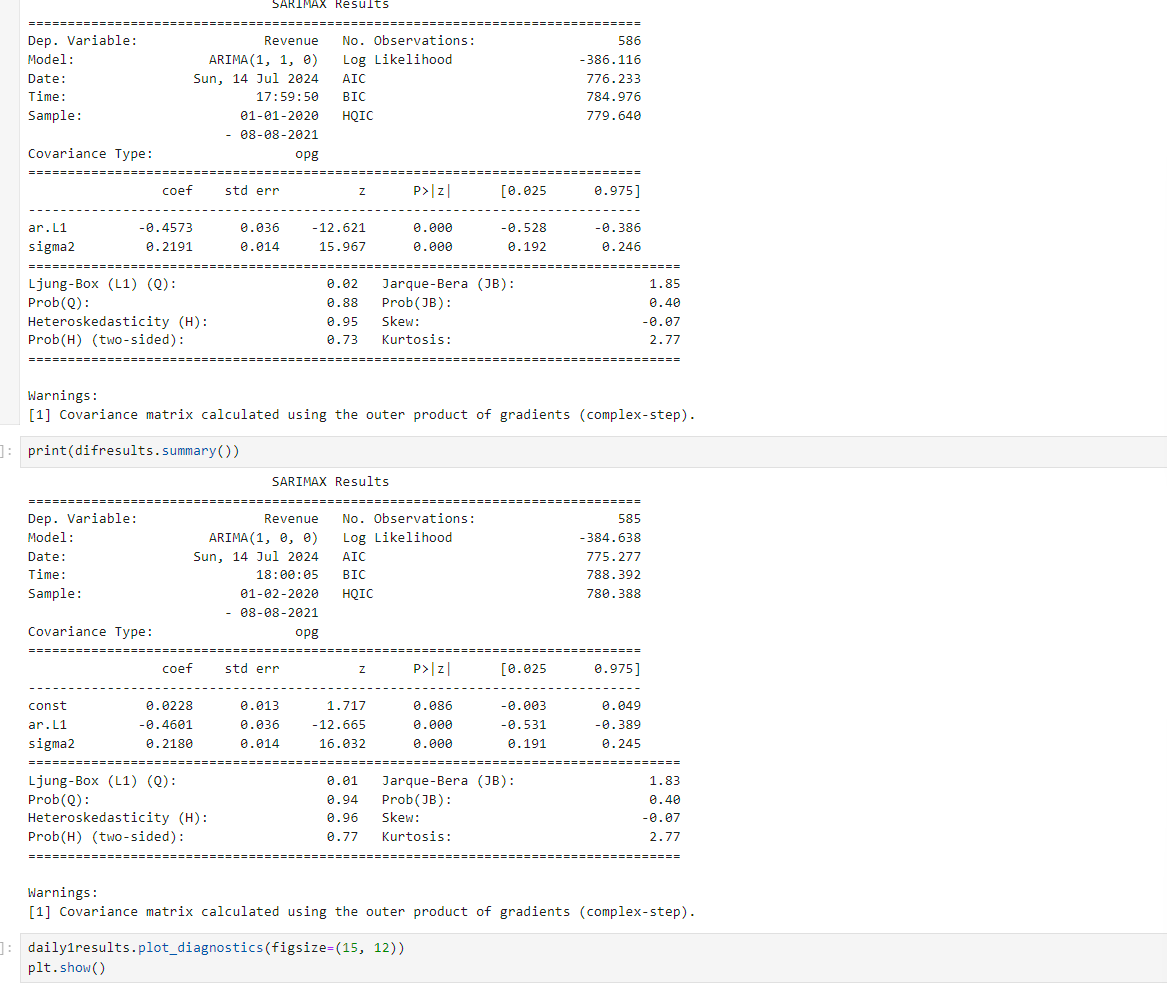
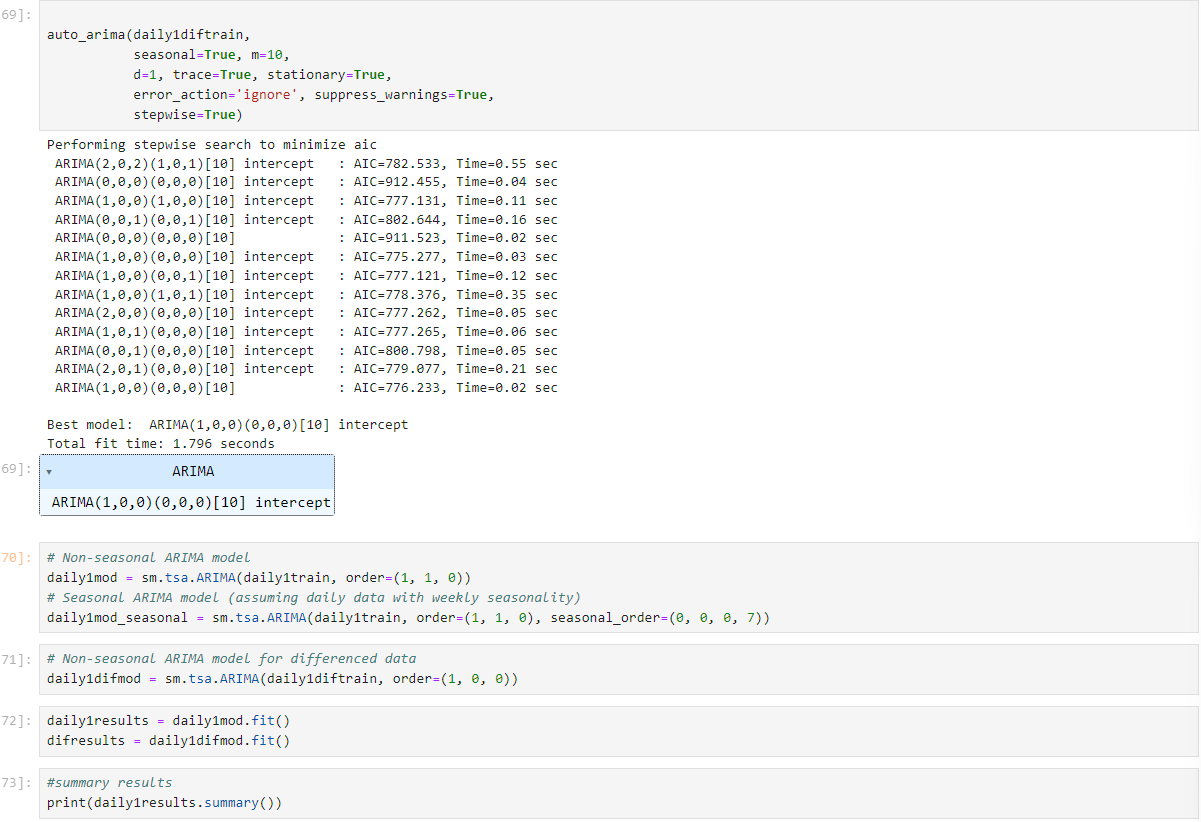
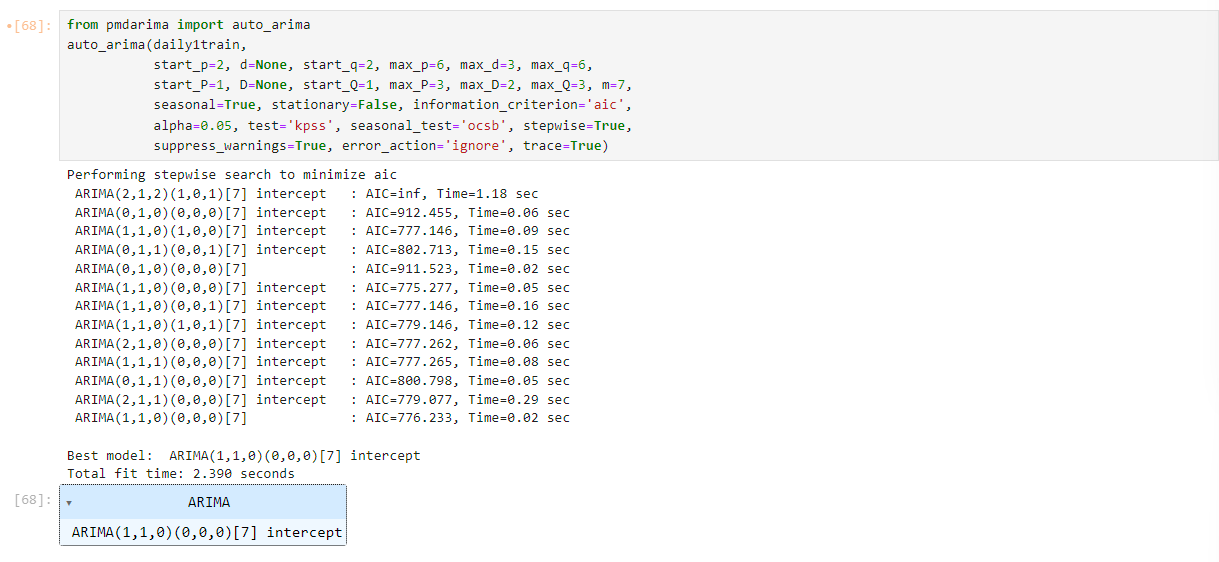
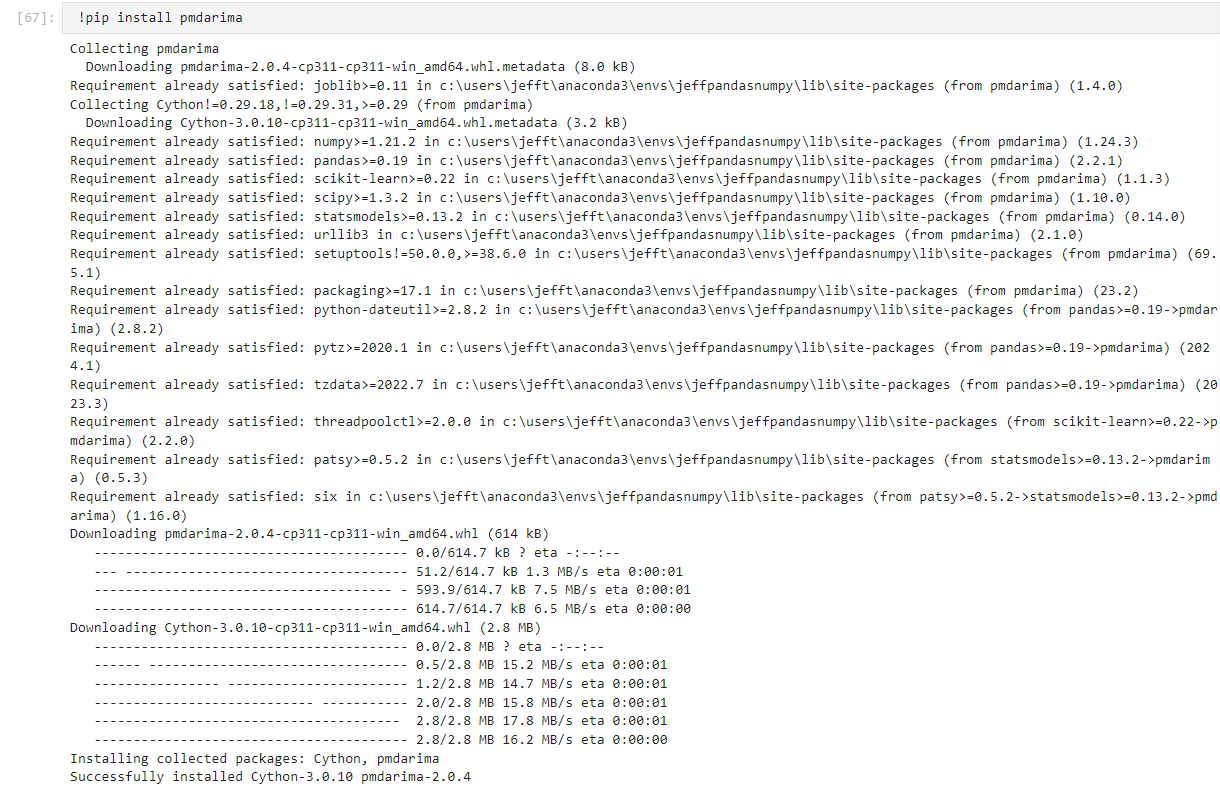
• auto correlation function

There is no seasonality observed in the data, both before and after processing, which I will verify using the Decompose function from statsmodels later in the notebook. All datasets exhibited a noticeable upward trend over time. After transforming the data to achieve stationarity, both the ADF and KPSS tests confirmed its stationary nature. Visually, the trend appears minimal and could potentially be addressed using an ARIMA model.Each autocorrelation coefficient's absolute value is below 0.5, indicating no significant autocorrelation after differencing. Both the ACF and PACF plots confirm the absence of autocorrelation in the data.After differencing the daily data, no dominant frequency is observed, indicating that the ARIMA model should be applied with d=1d=1d=1 for optimal results. I will confirm this using the pdmarima machine learning function. The data exhibits characteristics of white noise.The original dataset exhibits a positive trend, which is effectively removed by applying a differencing approach with diff(1)\text{diff}(1)diff(1).

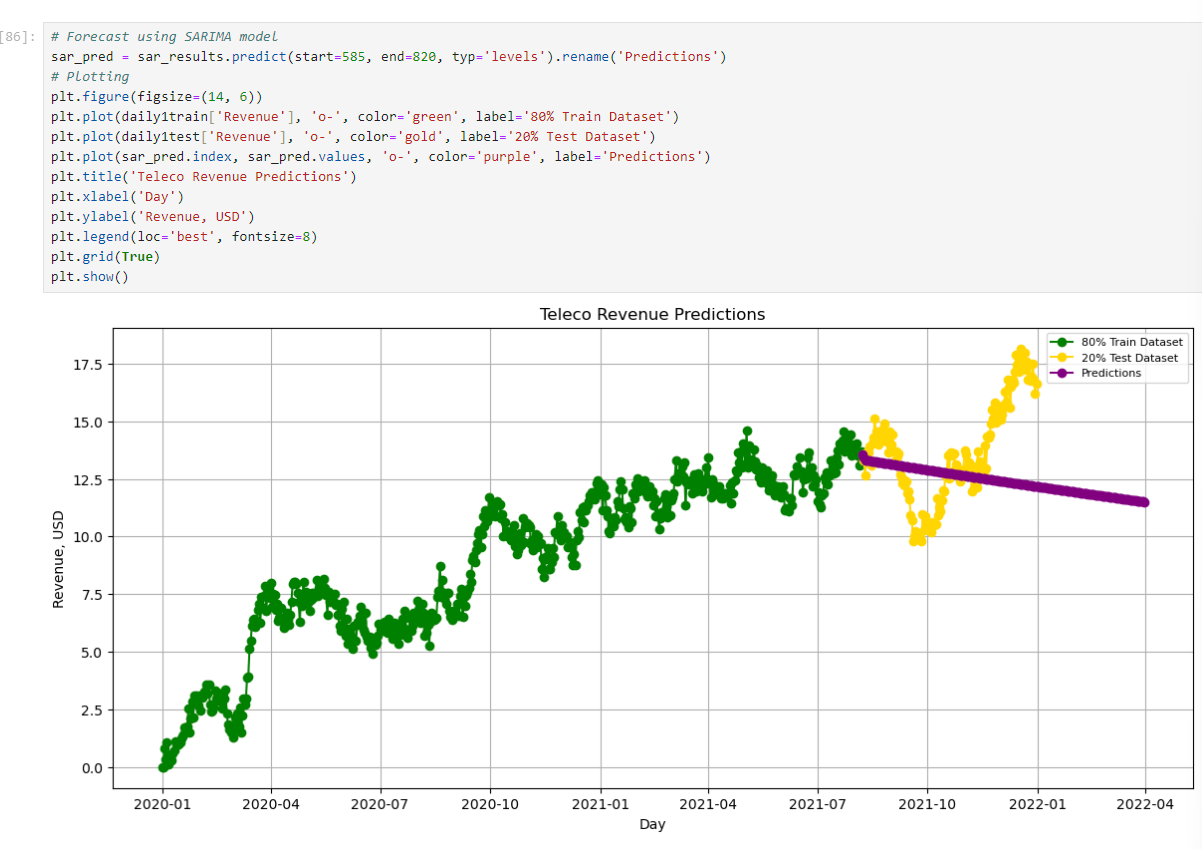
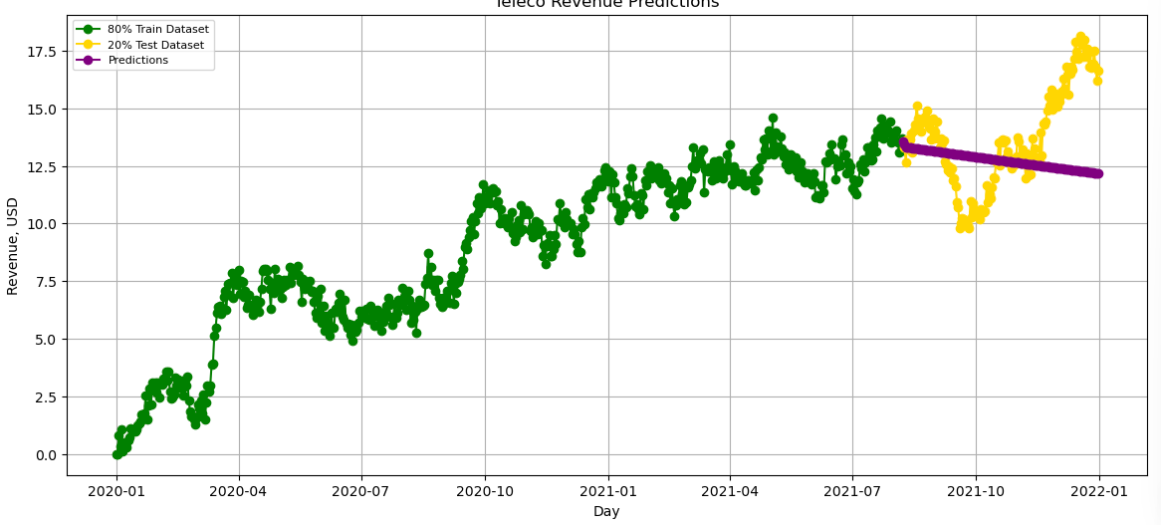
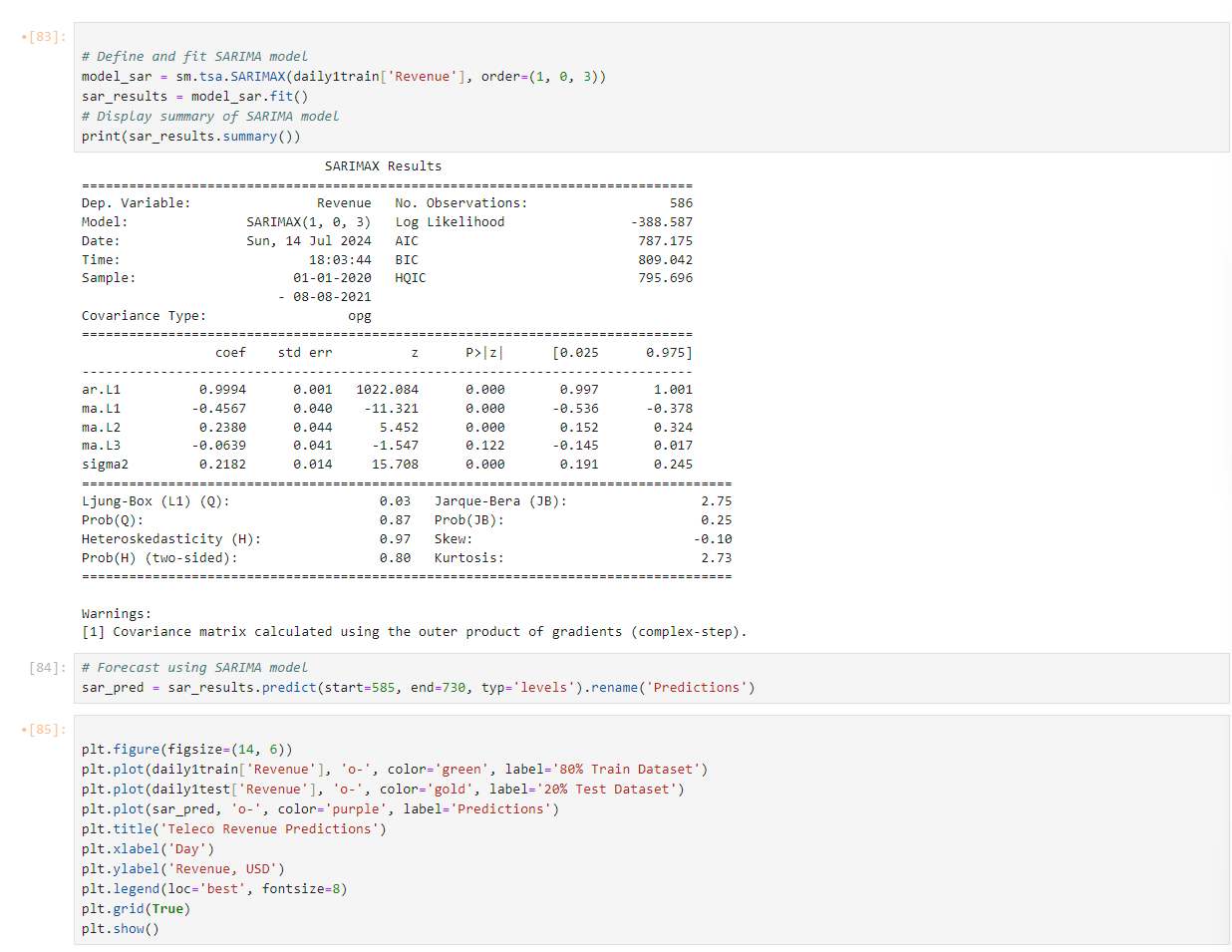
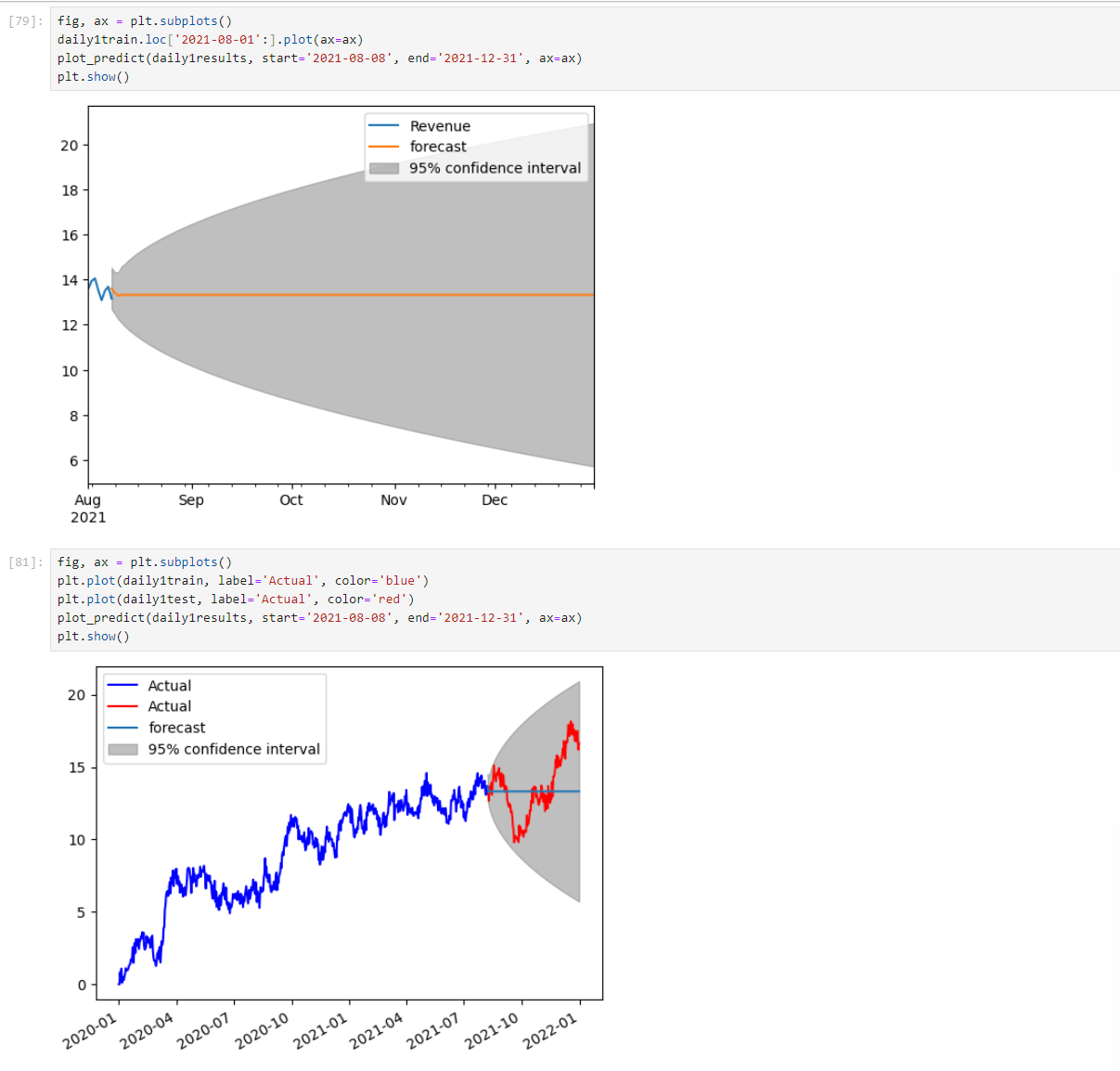


### **2. Identify an autoregressive integrated moving average (ARIMA) model that takes into account the observed trend and seasonality of the time series data.**

PMDArima library has built an autoarima grid-search like optimizer to choose p,q, and d for the Arima model and chooses the ones with the best score. The model is then fit to the data and the summary is printed. The model is then plotted to show the residuals and the model is then evaluated with the AIC and BIC scores.



### **3. Perform a forecast using the derived ARIMA model.**



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

All code provided in ipnyb included in submission

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

All code provided in ipnyb included in submission

# **References**

Built In. (n.d.). *Time series forecasting in Python: A comprehensive guide*. Retrieved July 14, 2024, from<https://builtin.com/data-science/time-series-forecasting-python>

Brownlee, J. (n.d.). *An end-to-end project on time series analysis and forecasting with Python*. Towards Data Science. Retrieved July 14, 2024, from<https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b>

Analytics Vidhya. (2022, June). *Time series forecasting using Python*. Retrieved July 14, 2024, from<https://www.analyticsvidhya.com/blog/2022/06/time-series-forecasting-using-python/>

PraveenPraveen, et al. (2022). *Contradictory results of ADF and KPSS unit root tests*. Cross Validated. Retrieved July 14, 2024, from<https://stats.stackexchange.com/questions/239360/contradictory-results-of-adf-and-kpss-unit-root-tests>

## **Data Summary and Implications**

### **Data Analysis Results**

**1. Selection of an ARIMA Model** To choose the ARIMA model, we used the pmdarima library's auto\_arima function. By experimenting with many combinations and selecting the one with the lowest Akaike Information Criterion (AIC), this function assists in determining the optimal model parameters. For the non-seasonal component, the optimal model's parameters were (p=1, d=1, q=0), and for the seasonal part, they were (P=0, D=0, Q=0, m=7). This indicates that the model has an autoregressive component and employs first-order differencing to make the data stationary.

**2. Prediction Interval of the Forecast** The predict method of the SARIMA model was utilized to compute the prediction intervals. We can see a range where future numbers are probably going to fall from these intervals. A 95% prediction interval, for instance, indicates that there is a 95% likelihood that the actual values will fall within this range.

**3. Justification of the Forecast Length** We chose to forecast from day 585 to day 820 to cover a significant future period and capture potential seasonal patterns. This period aligns well with business planning needs and allows us to observe trends that can inform decision-making.

**4. Model Evaluation Procedure and Error Metric** A train-test split was utilized to assess the model, using 80% of the data for training and 20% for testing. The average absolute difference between the expected and actual values was measured using the Mean Absolute Error (MAE), which served as the error metric. A better match is indicated by lower MAE values.

### **Recommended Course of Action**

**They should Keep an eye on revenue trends because a trustworthy assessment of future revenue trends is provided by the forecast. Keeping a close eye on these patterns can assist in making well-informed decisions regarding resource allocation, marketing, and budgeting.They Consider increasing marketing initiatives if the prediction indicates a possible reduction in sales.**