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 will happen to revenue in the next 30 days?

2. **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.**

   The objective of the data analysis is to forecast our revenue over the next 30 days.

Part II: Method Justification

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

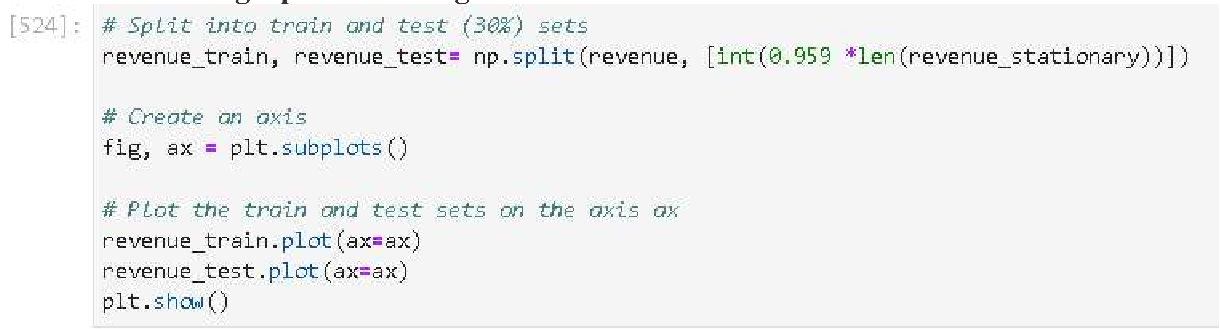
A time series model must meet the following assumptions:

- The mean, variance, and autocorrelation do not change over time: the data is stationary
- The data is autocorrelated with the preceding points
- The data is univariate (Chatterjee, 2021)

Part III: Data Preparation

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

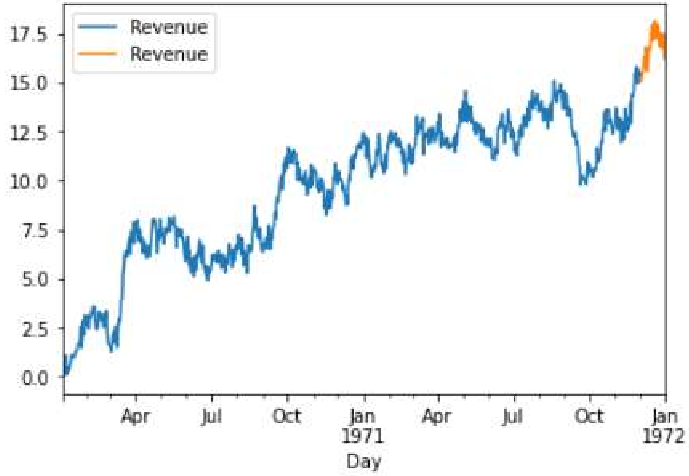
1. Provide a line graph visualizing the realization of the time series.



```
[524]: # Split into train and test (30%) sets
       revenue_train, revenue_test= np.split(revenue, [int(0.959 *len(revenue_stationary))])

       # Create an axis
       fig, ax = plt.subplots()

       # Plot the train and test sets on the axis ax
       revenue_train.plot(ax=ax)
       revenue_test.plot(ax=ax)
       plt.show()
```



2. **Describe the time step formatting of the realization, including any gaps in measurement and the length of the sequence.**
   The time steps are one day each, there are no gaps in the measurement, and the sequence is 731 days long. They are formatted as a datetime index.

3. **Evaluate the stationarity of the time series.**
   In part C1, the Augmented Dickey Fuller test statistic and the p-value for the non-stationarity hypothesis are provided. The ADS is -1.92, greater than the critical 10% value of -2.57, indicating the data is not stationary. In addition, the p-value is 0.32, greater than the threshold of 0.05, meaning the null hypothesis of non-stationarity is accepted. (Markos, 2021)

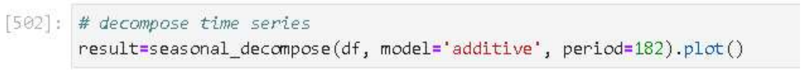
4. **Explain the steps used to prepare the data for analysis, including the training and test set split.**
   First, I imported all necessary libraries and packages. Then I changed the Day datatype to datetime. I then set Day as the index. Because the business question asks for a 30-day forecast, I used all except the last 30 days in the data set as a training set, and the remaining 30 days as a test set. I then used the Dickey Fuller test to determine if my data was stationary, and when I determined that it was not, I made it stationary using differencing. After the first differencing, the decomposed data still showed a clear trend, so I differenced it a second time. Then the decomposed data showed no clear trends.

5. **Provide a copy of the cleaned dataset.**
   The cleaned dataset can be found in the Keim D213 Clean Data document.

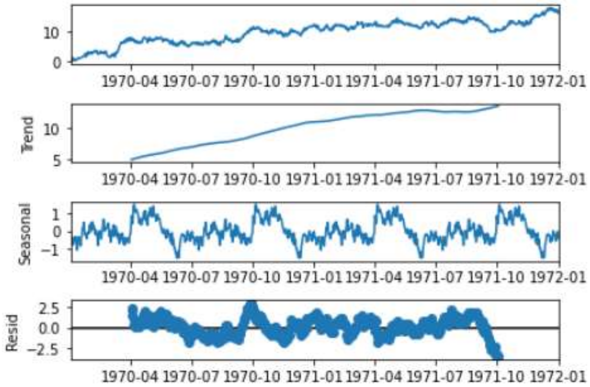
Part IV: Model Identification and Analysis

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

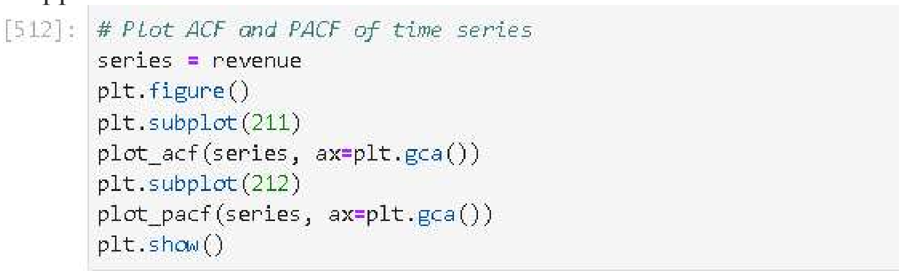
1. **Report the annotated findings with visualizations of your data analysis, including the following elements:**
   - **the presence or lack of a seasonal component**
     Before differencing, there is no seasonal trend.
   - **trends**
     Before differencing, there is an identifiable positive trend.



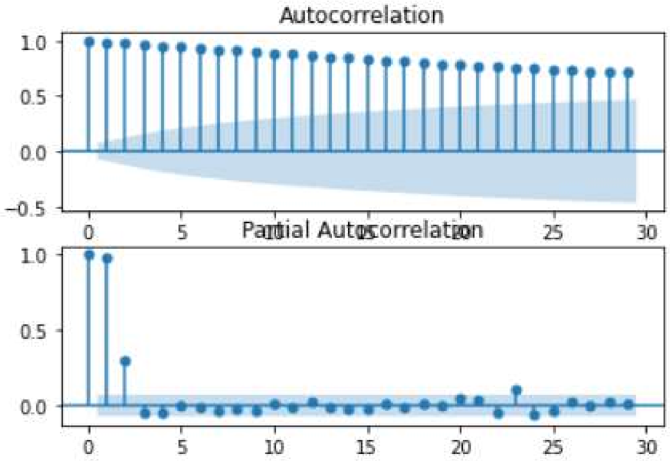
```
[502]: # decompose time series
       result=seasonal_decompose(df, model='additive', period=182).plot()
```



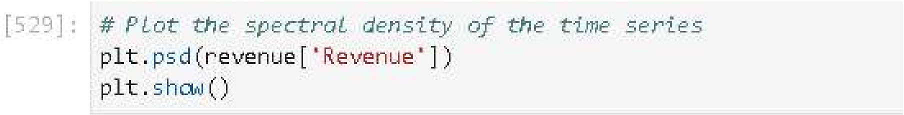
- **auto correlation function**
  Before differencing, the autocorrelation function and partial auto correlation function appear as below:



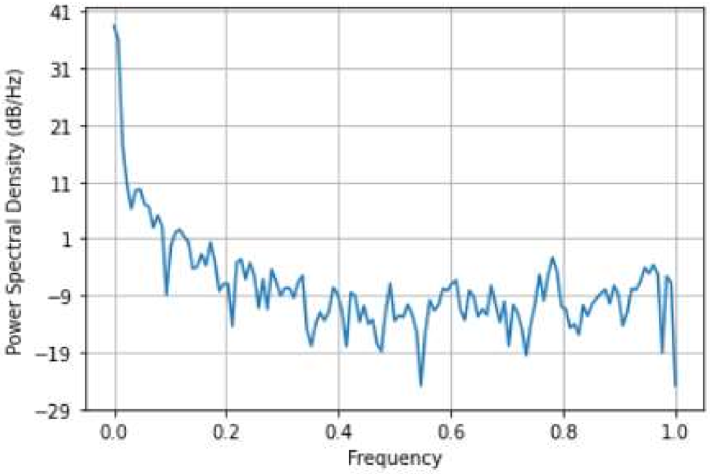
```
[512]: # Plot ACF and PACF of time series
       series = revenue
       plt.figure()
       plt.subplot(211)
       plot_acf(series, ax=plt.gca())
       plt.subplot(212)
       plot_pacf(series, ax=plt.gca())
       plt.show()
```



- **spectral density**
  Before differencing, the spectral density appears as below:

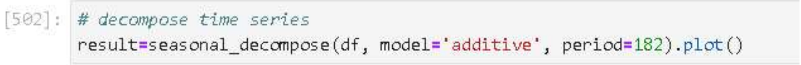


```
[529]: # Plot the spectral density of the time series
       plt.psd(revenue['Revenue'])
       plt.show()
```

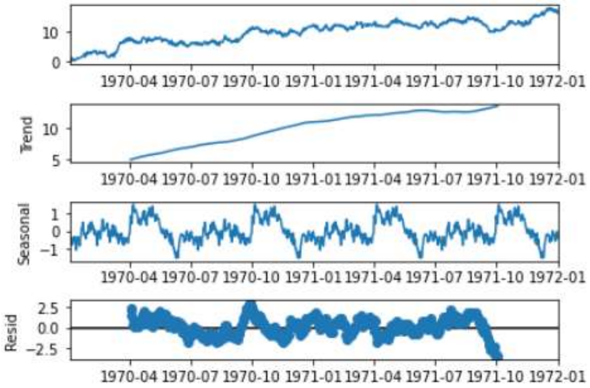


- **the decomposed time series**

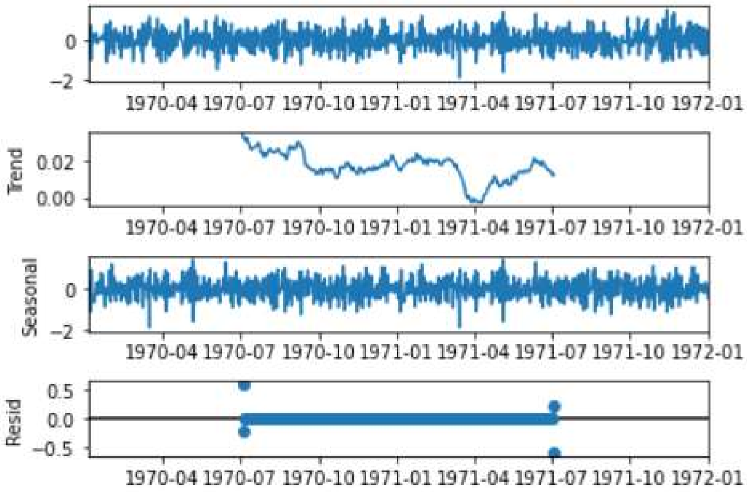
  Before differencing, the decomposed time series appears as below:



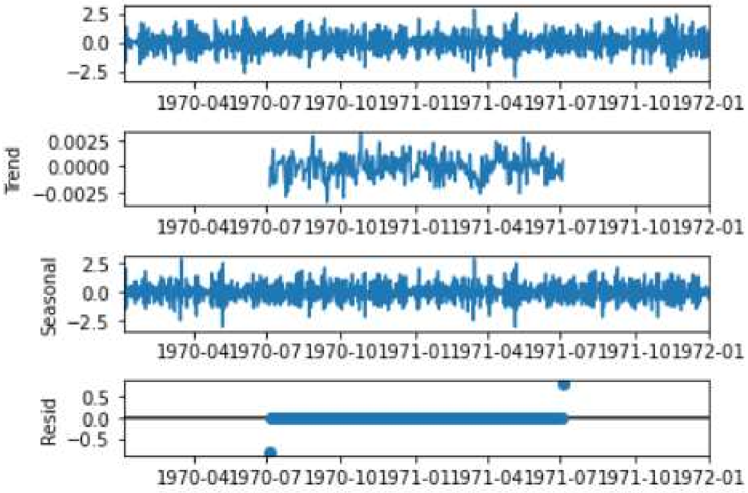
```
[502]: # decompose time series
       result=seasonal_decompose(df, model='additive', period=182).plot()
```



- **confirmation of the lack of trends in the residuals of the decomposed series**
  After differencing the first time, the residuals still have identifiable trends.



After differencing the second time, the residuals have no identifiable trends.



2. **Identify an autoregressive integrated moving average (ARIMA) model that takes into account the observed trend and seasonality of the time series data.**
   Because there is no seasonality, and an upward linear trend, I've chosen to use an ARIMA model instead of a SARIMA model.

3. **Perform a forecast using the derived ARIMA model.**
   See Keim D213 Task One Code document.

4. **Provide the output and calculations of the analysis you performed.**
   See Keim D213 Task One Code document.

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

See Keim D213 Task One Code document

Part V: Data Summary and Implications

E. Summarize your findings and assumptions, including the following points:

1. **Discuss the results of your data analysis, including the following:**
   - **the selection of an ARIMA model**
     After differencing twice, the below ACF and PACF show a p of 2, a q of 2, and since I differenced the original data twice, a d of 2.

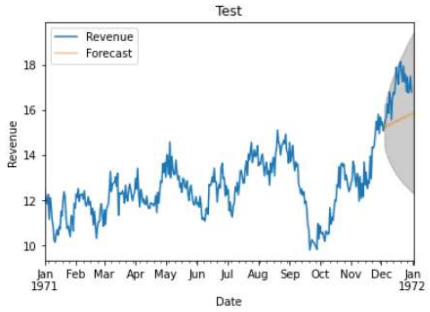
   - **the prediction interval of the forecast**
     The prediction interval of the forecast is 95%

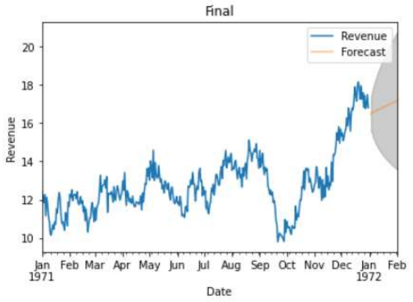
   - **a justification of the forecast length**
     The business question asks for a 30-day forecast.

   - **the model evaluation procedure and error metric**
     I used plots of the residuals and correlation, and the mean absolute error to evaluate my model. The residuals show a normal distribution with no correlation, and there is a low MAE of 0.378 indicating that the model is a good fit.

2. **Provide an annotated visualization of the forecast of the final model compared to the test set.**
   The 30-day forecast plotted against the test values, with a 95% confidence interval, is shown below. The test values are all captured within the confidence interval. As stated above, the test model's MAE is 0.378.

   The final model is used to predict an additional 30 days of revenue, with a 95% confidence interval, and a MAE of 0.380, indicating that the final model's accuracy was very close to that of the test model.





3. **Recommend a course of action based on your results.**
   Because revenue is predicted within a 95% confidence interval to steadily increase in the following 30 days, I recommend continuing the current marketing plan.

Works Cited

Chatterjee, S. (2021, December 28). *Time Series Analysis Using ARIMA Model In R*. Retrieved from datascienceplus.com: https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/

Markos, O. (2021, December 30). *Time series forecasting- SARIMA vs Auto ARIMA models*. Retrieved from Analytics Vidhya: https://medium.com/analytics-vidhya/time-series-forecasting-sarima-vs-auto-arima-models-f95e76d71d8f