Saemi Ramirez

011926418

D213 PA 1

Times Series Modeling

10/11/2024

WGU

1. Purpose of this data analysis
   1. Research Question

What is the WGU medical organization’s revenue in the following quarter?

* 1. Define objectives or goals

The objective is to predict the future revenue of WGU medical organization using ARIMA time series modeling. ARIMA time series modeling can capture the patterns in the past revenue data. It also can help the management to make critical decisions about budgeting, resource allocation, and or financial goals by providing the accurate revenue forecasting.

1. Summarize assumptions of a time series modeling including stationarity and autocorrelated data
   1. The assumption for the stationary time series modeling implies that its statistical properties, such as mean, variance, and autocorrelation, remain constant over time. A stationary time series is one where the mean, variance, and autocorrelation structure remain consistent over time. This assumption is crucial because many time series analysis methods, particularly those used for forecasting, rely on the premise that the patterns and relationships in the data are stable over time. (The Stationary)
   2. Autocorrelation measures how related the current data points are to past values, which can have significant implications across various industries. For instance, if our passenger data exhibits strong autocorrelation, we can infer that high passenger numbers today indicate a strong likelihood of similarly high numbers tomorrow. (Pierre)
2. Data cleaning process
   1. Provide the line graph visualizing the realization of the time series

Following screenshot is the plotting code and the line graph of the WGU medical revenue with the trend line starting January 1, 2020 to December 31, 2021.

A screen shot of a computer code

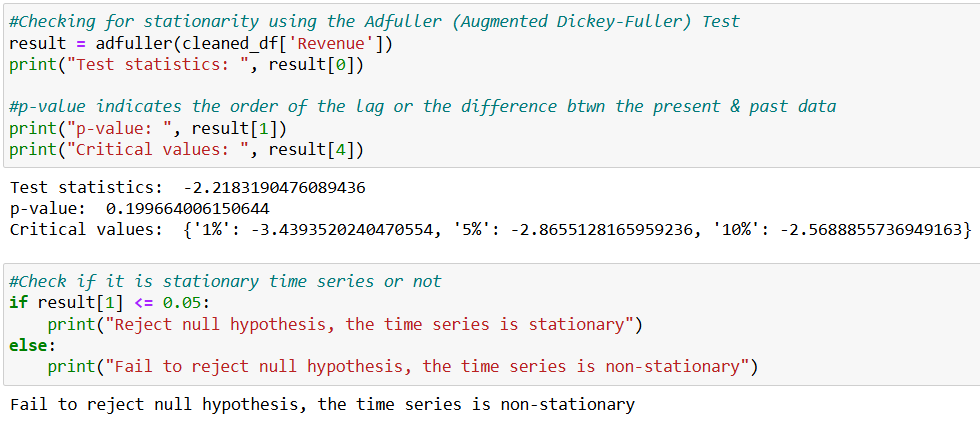
Description automatically generated

A graph with lines and numbers

Description automatically generated

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

The Day column in the provided medical data was indexed from 1 to 731 in index datatype. The new index was given with the column name ‘Date’ from 2020-01-01 ended on 2021-12-31. The data was for 2-year worth of information. There was no gap in both columns to fill in or to drop.

* 1. Evaluate the stationarity of the time series  
     The given WGU medical dataset was not the stationarity of the time series. After running the adfuller (Augmented Dickey-Fuller) function on the cleaned data, the p-value was close to 0.2, so differentiated function was applied to lower the p-value. Then the p-value lowered to close to 0, which means the data is now stationary.

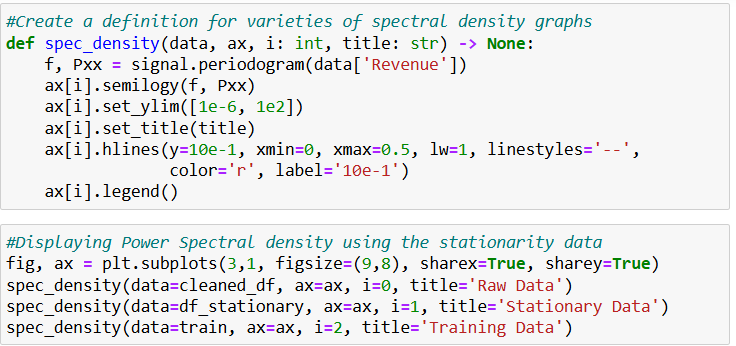
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

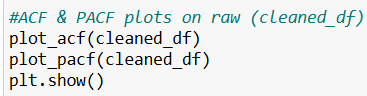
The following visualizations are the Spectral Density graphs on the raw (cleaned\_df), the differentiated (df\_stationary), and the test set (test).



A graph of data with blue lines

Description automatically generated with medium confidence

Autocorrelaton and partial autocorrelation visualizations are compared between the raw (cleaned\_df, on the left) and the train data (on the right).

 A white background with black text

Description automatically generated

A graph with a line

Description automatically generatedA graph with a blue line

Description automatically generated

A graph of a graph

Description automatically generated with medium confidenceA graph of a train

Description automatically generated

A graph with blue dots and numbers

Description automatically generatedA graph with blue dots and numbers

Description automatically generated

* 1. Explain the steps you used to prepare the data for analysis including the training and test set split
     1. Read the WGU medical dataset using read\_csv function
     2. Checked the shape, head, info, and any null values and dropped null values if there was any in the dataset
     3. Checked the info and description again
     4. Created the ‘Date’ column and added the date starting January 1, 2020 in yyyy-mm-dd format to the end
     5. Set the ‘Date’ column as index and dropped ‘Day’ column
     6. Ran the adfuller function if the dataset is stationary
     7. The p-value was over 0.05, so ran the differentiated function
     8. Utilized pandas loc function to separate the test and the training set
  2. Copy of a cleaned data set

‘cleaned\_data.csv’, ‘stationary\_data.csv’, ‘test\_data.csv’, and ‘train\_data.csv’ are provided

1. Analyze the time series data set
   1. Report the annotated finding with visualizations of your data analysis

Presence or lack of a seasonal component  
The following screenshot is a seasonal graph which shows that it repeats the same behavior approximately every 3-4 months. Raw (cleaned\_df), which is non-stationary, was used.

* + 1. A screen shot of a graph

       Description automatically generated
    2. Trends

Trend graphs shows that the revenue was reaching higher until the October of the first year. There are some ups and downs after October 2020. Raw (cleaned\_df) was used.

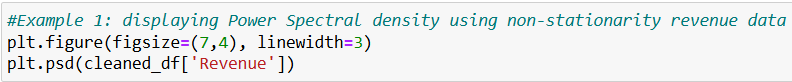
A graph with a line

Description automatically generated

* + 1. Autocorrelation function  
       Following screenshot is the autocorrelation graph on raw (cleaned\_df) data.  
       A graph with lines and numbers

       Description automatically generated
    2. Spectral density

Following screenshot is a plot of power spectral density of the raw (cleaned\_df) data.



A graph with a line graph

Description automatically generated

Following screenshot is a plot of power spectral density of the non-stationary (df\_stationary) data.

A screen shot of a graph

Description automatically generated

* + 1. Decomposed time series  
       The following screenshot is the full seasonal decomposition graph including the raw data, trendline, seasonal, and the residual of the cleaned\_df non-stationary data.

A graph of sales and revenue

Description automatically generated with medium confidence

* + 1. Confirmation of the lack of trends in the residuals of the decomposed series

Residual graph in the following screenshot of the non-stationary raw (cleaned\_df) data  
A graph with blue lines

Description automatically generated

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

ARIMA function was utilized with train dataset with order of (1, 1, 0) as a best model fit which will be discussed in section E1i. Summary of the ARIMA shows as follows.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated



* 1. Perform a forecast using the derived ARIMA model

Added the forecast values for the next 92 days

A number with numbers on it

Description automatically generated with medium confidence

Created the date list from 2021-10-01 to 2021-12-31

A screenshot of a computer code

Description automatically generated

Assigned the start and end positions with index

A screenshot of a computer code

Description automatically generated

Converted the Series to the DataFrame

A screenshot of a computer

Description automatically generated

Plotted all 3 (train, test, and prediction) in one visualization

A screen shot of a computer code

Description automatically generated

A graph with blue lines

Description automatically generated

* 1. Provide the output and calculations of the analysis  
     The root mean squared error of this forecasting of Q4 2021 is approximately 2.08

A close-up of a white background

Description automatically generated

* 1. Provide the code

‘Saemi Ramirez D213 PA1 – Time Series Modeling.ipynp’ is submitted

1. Summarize your findings and assumptions
   1. Discuss the results of your data analysis
      1. Selection of an ARIMA model  
         In order to find the best model for ARIMA, auto-ARIMA function was conducted using the train data. As it is indicated in the middle of the screenshots, the best ARIMA function model is (1,1,0)

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* + 1. Prediction interval of the forecast and justification  
       The prediction interval of the forecast was set to 90 days using the forecast() method because it is a balance between meaningful foresight and reasonable prediction accuracy.
    2. Model evaluation procedure and error metric  
       A screenshot of a computer code

       Description automatically generated

A collage of graphs and diagrams

Description automatically generated

* 1. Provide an annotated visualization of the forecast of the final model compared to the test set  
     The final forecast model compared to the test set is provided in the following screenshots.  
     A screen shot of a computer code

     Description automatically generated

A graph with blue lines

Description automatically generated

* 1. Recommend a course of action  
     With the given forecast of next 90 days, the organization can set the short-term financial and operational goals such as improving cash flow, managing operational costs, or optimizing resource allocation. They can make an adjustment to the quarterly budget as necessary with increase or decrease staffing levels or launching initiatives based on expected revenue.

1. Create your report using an industry-relevant interactive development environment  
   ‘Saemi Ramirez (011926418) D213 PA 1 Time Series Modeling.pdf’ is submitted
2. List 3rd party code used  
   N/A
3. List 3rd party citations and references
   1. Pierre, Sandrach. *A Guide to Time Series Analysis in Python.* Builtin. (May 10, 2024). <https://builtin.com/data-science/time-series-python>.
   2. *The Stationary Data Assumption in Time Series Analysis*. Complete Dissertation. <https://www.statisticssolutions.com/stationary-data-assumption-in-time-series-analysis>.