Time-Series Analysis Performance Assessment

Ednalyn C. De Dios

D213

September 7, 2023

Western Governors University

Time-Series Analysis Performance Assessment

This analysis utilizes Python 3.9.9 and Jupyter 7.0.2.

Part I. Research Question

A1. 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 revenue forecast for the next month?***

A2. Define the objectives or goals of the data analysis. Ensure your objectives or goals are reasonable within the scope of the scenario and are represented in the available data.

Analyze two years' worth of daily revenue data of the organization and create a predictive model that will forecast the next 30 days of future revenue.

Part II. Method Justification

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

Two assumptions of time series analysis include stationarity and autocorrelation. Stationarity means that "the mean, variance, and autocorrelation structure are constant over time" (Statisticssolutions.com, n.d.). In other words, "the statistical properties of a time series do not change over time" (Statisticssolutions.com, n.d.). The other assumption is no autocorrelation. "Autocorrelation occurs when future values in a time series linearly depend on past values" (Pierre, 2021).

Part III. Data Preparation

C1. Provide a line graph visualizing the realization of the time series.

C2. Describe the time step formatting of the realization, including any gaps in measurement and the length of the sequence.

There are no gaps in measurement and the sequence is at 700 days.

C3. Evaluate the stationarity of the time series.

The dataset is stationary and as such no further action is necessary to make it so.

```
result = adfuller(df['Revenue'])
print('Test statistics: ', result[0])
print('P-value: ', result[1])
print('Critical value: ', result[4])
print('-----------------------------------------------------------------------------------------------')

if result[1] >= 0.05:
    print('Reject the null hypothesis. The time series is stationary. No further action required.')
else:
    print('Fail to reject the null hypothesis. The time series is not stationary. You must make it so.')
```

```
Test statistics:  -1.7746383121968732
P-value:  0.3931237595029723
Critical value:  {'1%': -3.4393644334758475, '5%': -2.8655182850048306, '10%': -2.568888486973192}
-----------------------------------------------------------------------------------------------
Reject the null hypothesis. The time series is stationary. No further action required.
```

C4. Explain the steps you used to prepare the data for analysis, including the training and test set split.

To prepare the data I dropped 0 values of Revenue. I also ensured there are no missing values. I then converted the sequential number of days into proper date formatting and set it as the index. For splitting the train and test sets, I took the last 30 days of the dataset and set it as the test set and assigned the rest as the training set.

C5. Provide a copy of the cleaned data set.

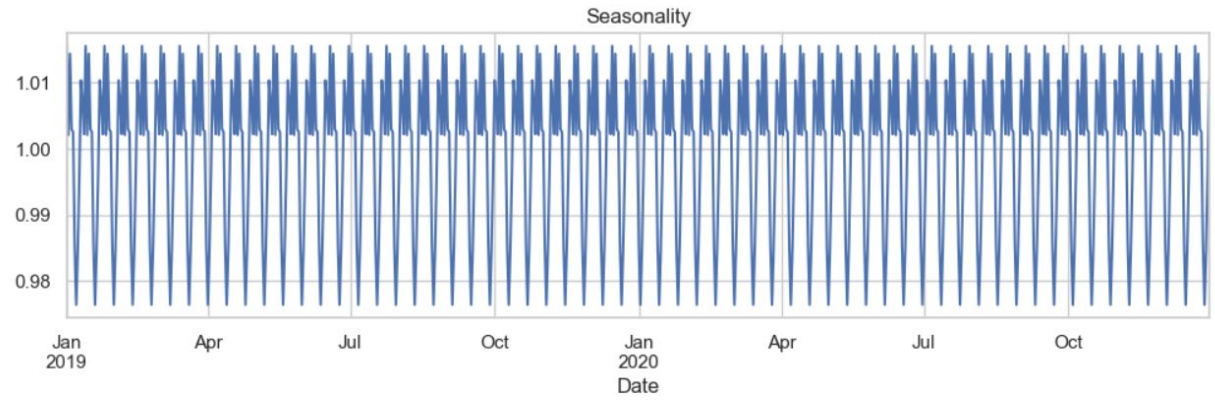
Filenames:

- df.to_csv('../data/teleco_cleaned1.csv', index=False)
- train.to_csv('../data/teleco_cleaned1_train.csv', index=False)
- test.to_csv('../data/teleco_cleaned1_test.csv', index=False)

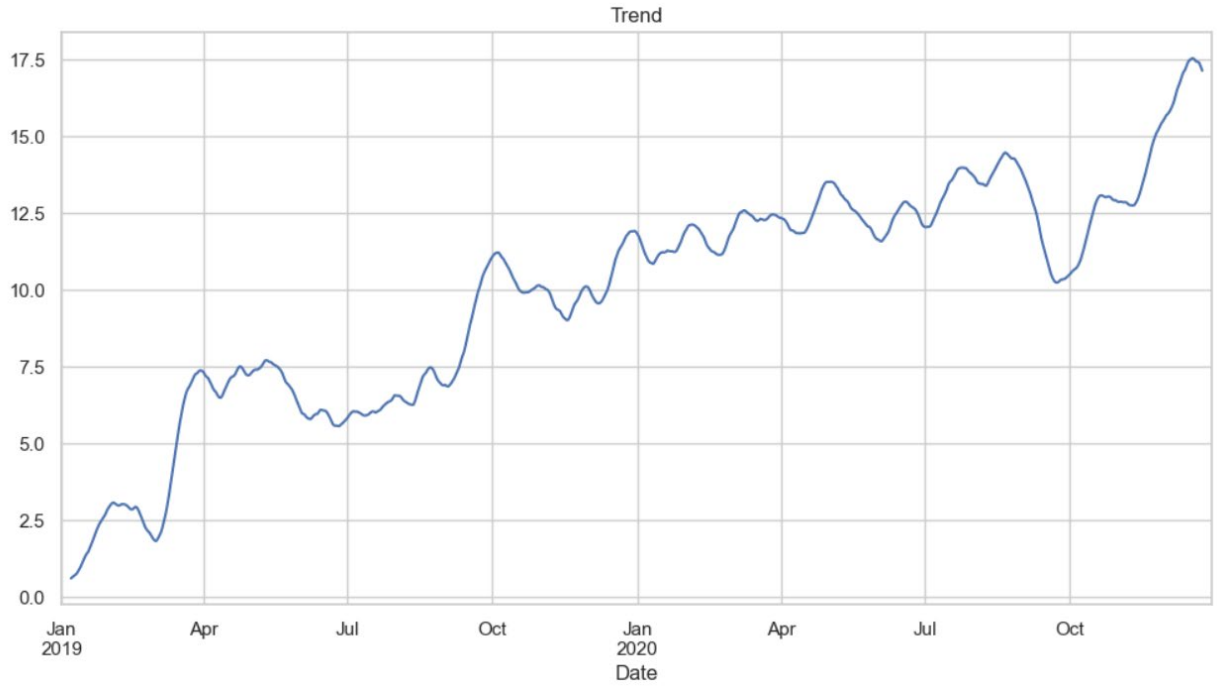
Part IV. Model Identification and Analysis

D1. Report the annotated findings with visualizations of your data analysis, including the following elements:

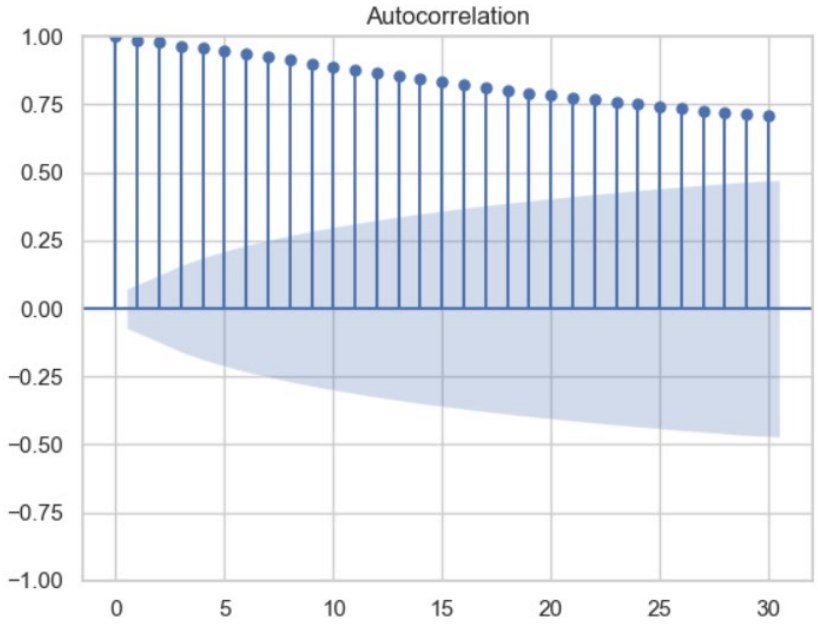
- the presence or lack of a seasonal component



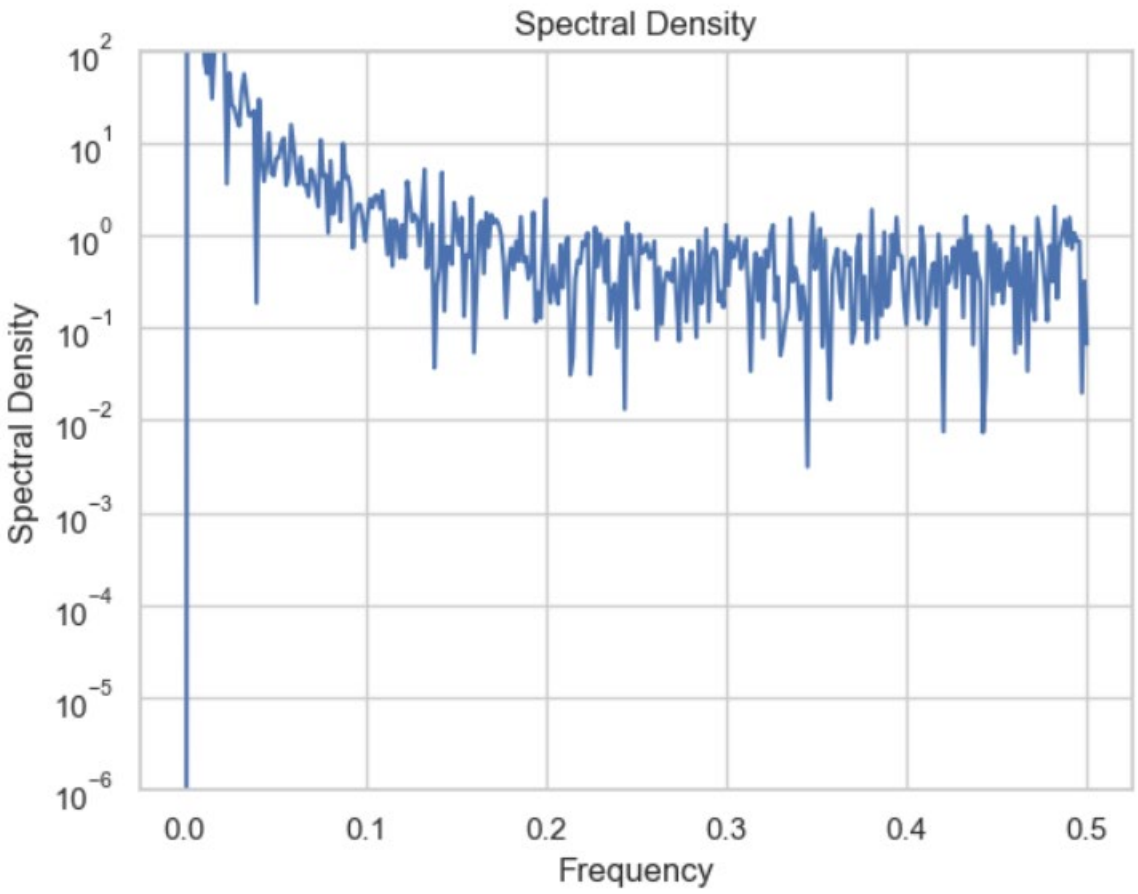
- trends



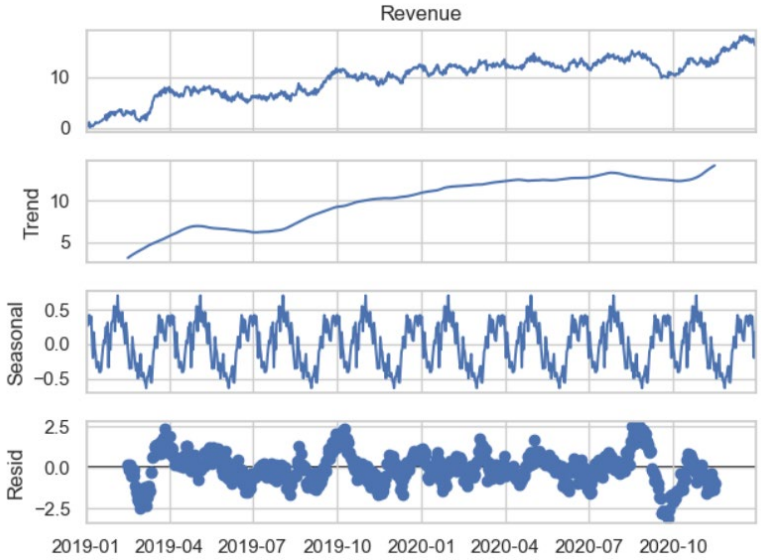
- the autocorrelation function



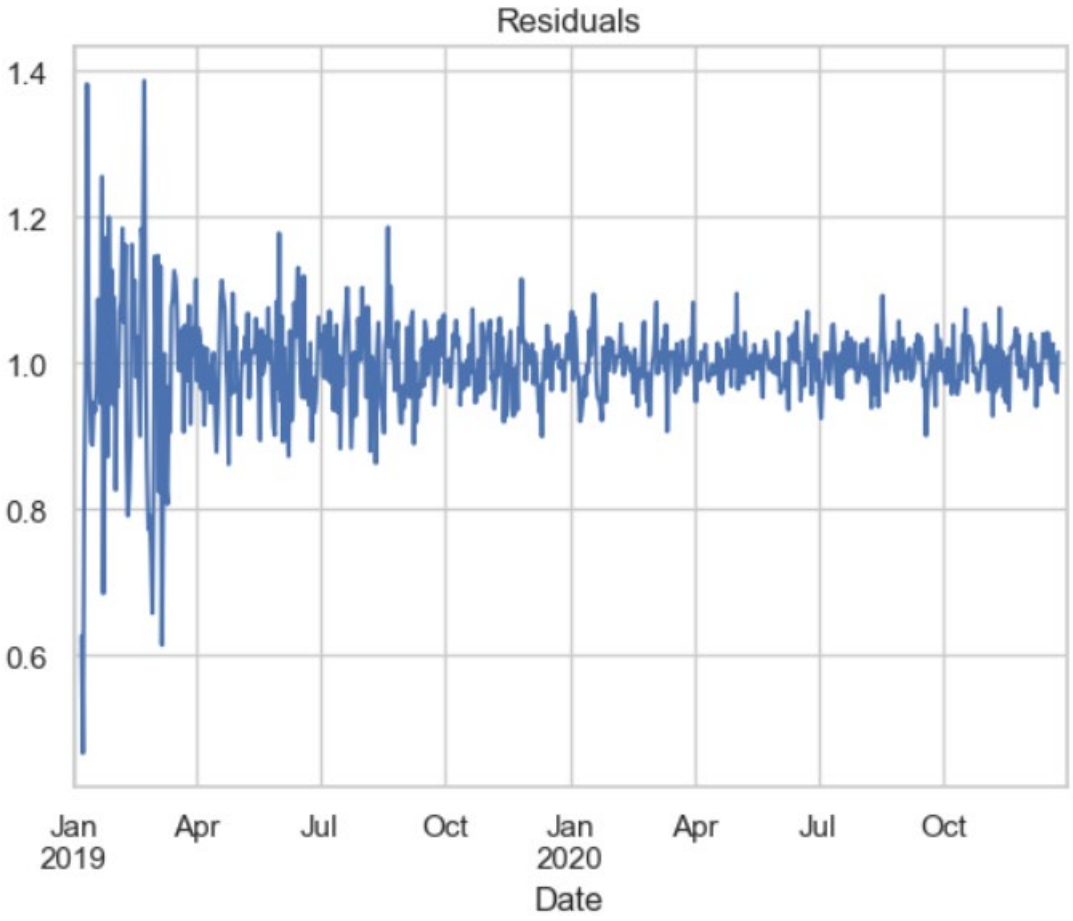
- the spectral density



- the decomposed time series



- confirmation of the lack of trends in the residuals of the decomposed series



D2. Identify an autoregressive integrated moving average (ARIMA) model that accounts for the observed trend and seasonality of the time series data.

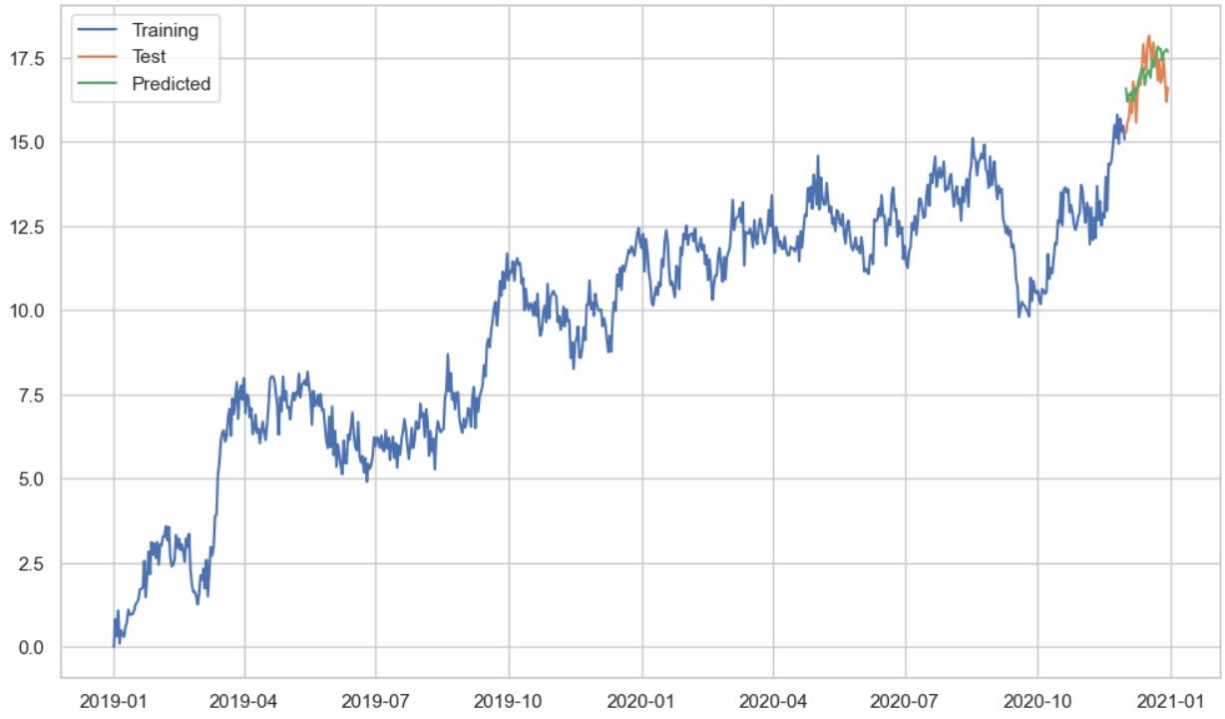
SARIMAX Results

Dep. Variable:	Revenue	**No. Observations:**	730
Model:	ARIMA(1, 1, 0)x(5, 1, 0, 12)	**Log Likelihood**	-535.993
Date:	Wed, 30 Aug 2023	**AIC**	1085.987
Time:	20:49:43	**BIC**	1118.012
Sample:	01-01-2019	**HQIC**	1098.353
	- 12-30-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.4781	0.034	-14.173	0.000	-0.544	-0.412
ar.S.L12	-0.8591	0.039	-22.126	0.000	-0.935	-0.783
ar.S.L24	-0.7058	0.052	-13.577	0.000	-0.808	-0.604
ar.S.L36	-0.4672	0.058	-8.071	0.000	-0.581	-0.354
ar.S.L48	-0.3017	0.050	-6.017	0.000	-0.400	-0.203
ar.S.L60	-0.1672	0.039	-4.304	0.000	-0.243	-0.091
sigma2	0.2565	0.015	17.414	0.000	0.228	0.285

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	2.12
Prob(Q):	0.95	**Prob(JB):**	0.35
Heteroskedasticity (H):	1.08	**Skew:**	0.00
Prob(H) (two-sided):	0.54	**Kurtosis:**	2.73

D3. Perform a forecast using the derived ARIMA model identified in part D2.



D4. Provide the output and calculations of the analysis you performed.

Filename: " D213 Performance Assessment Task 1 (Rev. 4) Notebook.pdf"

D5. Provide the code used to support the implementation of the time series model.

Filename: " D213 Performance Assessment Task 1 (Rev. 4) Notebook.pdf"

Part V. Data Summary and Implications

E1. Discuss the results of your data analysis, including the following points:

- the selection of an ARIMA model
- the prediction interval of the forecast
- a justification of the forecast length
- the model evaluation procedure and error metric

The final ARIMA model was based on the results of Auto ARIMA (Best model: ARIMA(1,1,0)(5,1,0)[12]) which considers the trend and seasonality of the data set. The prediction interval of the forecast is 30 days and can be made using the .predict() or .forecast() methods. Forecast length of 30 is enough information to make changes in preparation for the next month. The final model was evaluated with R2. It "measures the strength of the relationship between your model and the dependent variable" (Frost, 2018). Although 11.77 is a low result, a low R2 doesn't necessarily mean the model is bad (Frost, 2018).

E2. Provide an annotated visualization of the forecast of the final model compared to the test set.

E3. Recommend a course of action based on your results.

The forecast data estimates that revenue will be at $17.69 million. Visual inspection of the plot also reveals an upward trend for the forecasted months. As such, I recommend a conservative approach to configuring organization operations for this quarter.

Part VI. Reporting

F. With the information from part E, create your report using an industry-relevant interactive development environment (e.g., an R Markdown document, a Jupyter Notebook). Include a PDF or HTML document of your executed notebook presentation.

Filename: "D213 Performance Assessment Task 1 (Rev. 4) Notebook.pdf"

G. Cite the web sources you used to acquire third-party code to support the application.

- https://github.com/ecdedios/code-snippets/blob/main/notebooks/master.ipynb
- https://www.datacamp.com/tutorial/matplotlib-time-series-line-plot
- https://towardsdatascience.com/finding-seasonal-trends-in-time-series-data-with-python-ce10c37aa861
- https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a5b2
- https://analyticsindiamag.com/what-are-autocorrelation-and-partial-autocorrelation-in-time-series-data/
- https://github.com/mkosaka1/AirPassengers_TimeSeries/blob/master/Time_Series.ipynb

H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

Anonymous. (2023, August 30). The Stationary Data Assumption in Time Series Analysis. Statistic Solutions. https://www.statisticssolutions.com/stationary-data-assumption-in-time-series-analysis

Pierre, Sadrach. (2022, October 12). A Guide to Time Series Analysis in Python. Builtin. https://builtin.com/data-science/time-series-python

Frost, Jim. (2018, February 24). How To Interpret R-squared in Regression Analysis. Statistics By Jim. https://statisticsbyjim.com/regression/interpret-r-squared-regression