D213 – Task 1: Time Series Modeling

Part 1: Research Question

1. Can any meaningful patterns be revealed from the company's time series data of the profits over two years? Can these patterns be used to forecast future profits?

2. The object of this analysis is to build a model that can be used to predict future profits. This allows decision makers in the company to plan for periods of growth and recession.

Part 2: Method Justification

1. An assumption for time series analysis is that the data is stationary. Being stationary refers to having a mean, variance, and autocorrelation that remains constant. (Stationarity, n.d.) Autocorrelation is the correlation between two values in a time series. The time series is correlated with itself at intervals. These intervals are called lags. Lags are numbered by the count of intervals between data points. (Frost, 2022)

Part 3: Data Preparation

1. Line graph visualizing the realization of the time series:

2. ARIMA models require date format to run properly. To format this correctly, the day column was set to the index. It was then converted to datetime format using 'day' as the unit and starting on 2020-01-02. The length of the sequence is 731 days. It contains no gaps, missing data, or NAs.

3. The stationarity of the time series was evaluated using the augmented Dickey-Fuller test. The test returned a test statistic of -1.92 and a p-value of 0.322. The test statistic is higher the critical values implying that the time series is not stationary. Using a null hypothesis that the time series is not stationary and an alpha of 0.05, we fail to reject the null hypothesis because

the p-value (0.322) is greater than the alpha. We can conclude that the time series is not stationary.

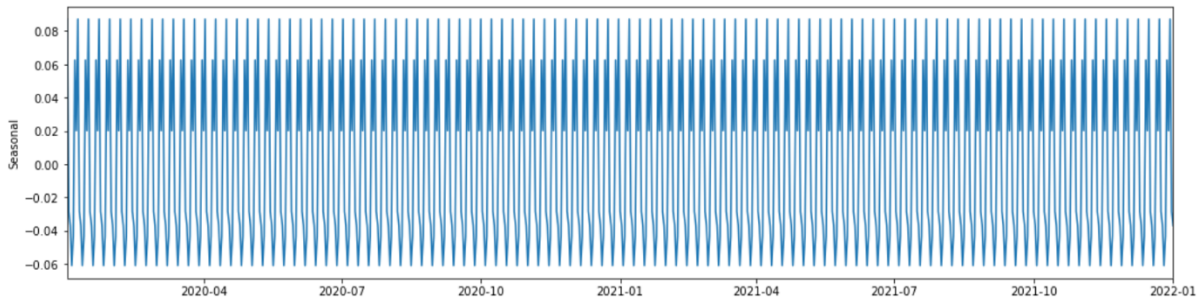
4. The steps used to prepare the data are as follows:

- Load the dataset in to the Jupyter notebook
- Add a Date column to the dataset and set it as the index. ARIMA models require dates to function properly.
- Graph the realization of the time series
- Look at the basic structure of the dataset, check datatypes, look for nulls, check for duplicated rows
- Look at the descriptive statistics
- Use the augmented Dickey-Fuller test to evaluate the stationarity of the time series
- Make the series stationary by taking the first difference. Recheck with augmented Dickey-Fuller. Time series is now stationary.
- Create a training set and test set using a 70-30 split.

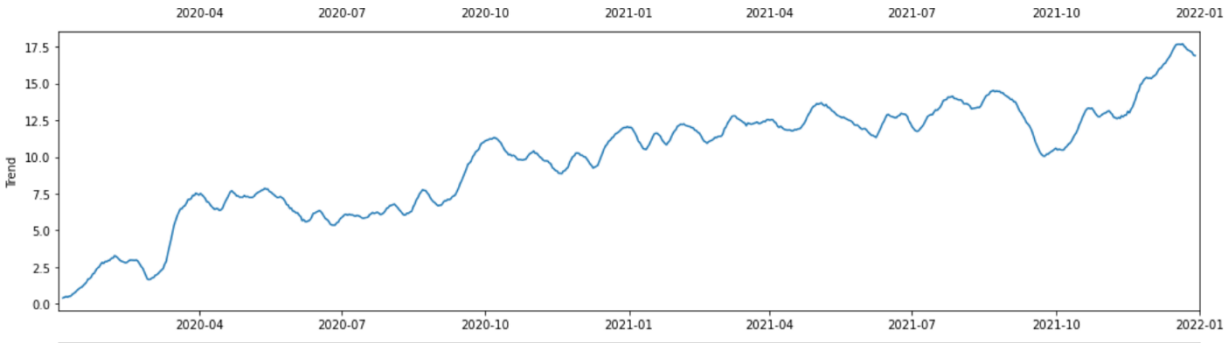
5. Prepared data attached: time_series_prepared_data.csv

Part IV: Model and Identification and Analysis

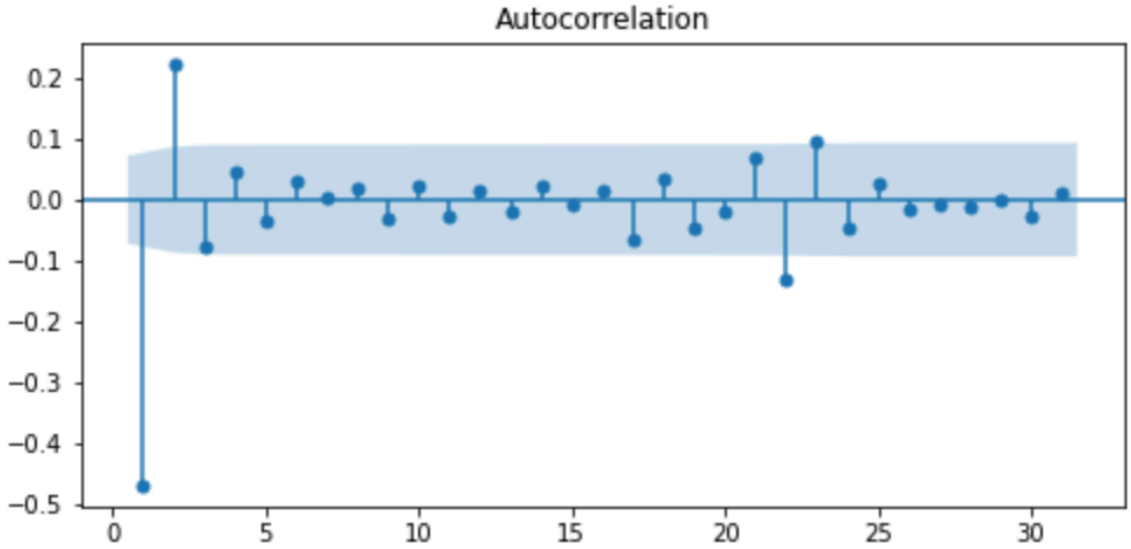
1. There does appear to be a seasonal component to the time series by visual inspection of the decomposition. The period looks to be 1 week.



The data has a positive trend over time.

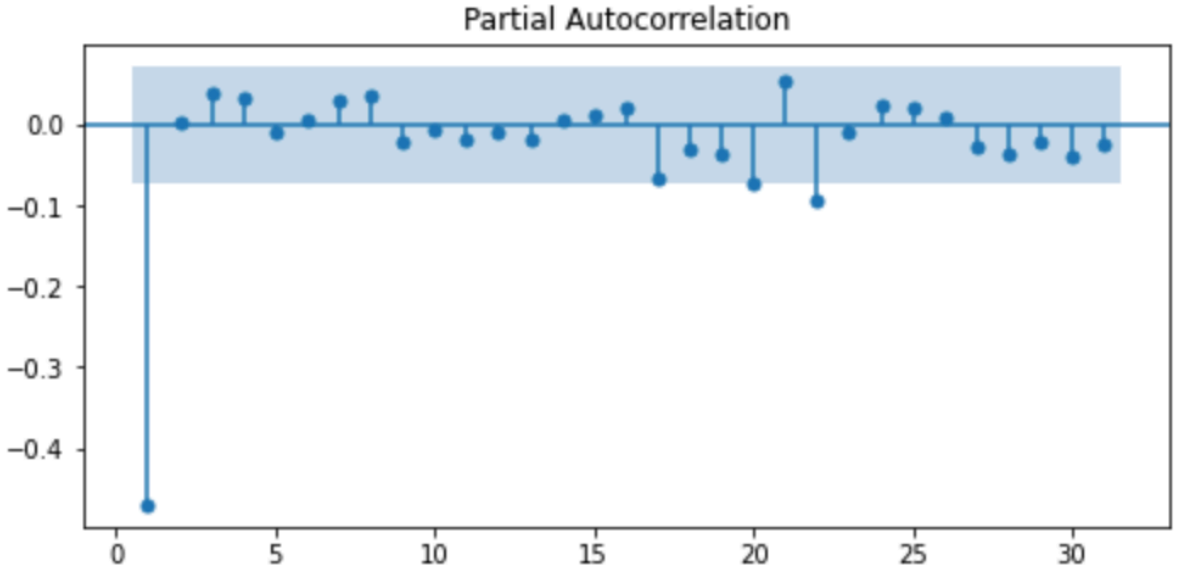


Autocorrelation displays the correlation of a time series with itself lagged by a set number of time units. (Radecic, 2021) This model used 31 days as the lag. The correlation values drop off after two lags. There is an increase in the correlation around the twenty-first lag, dropping off again after the twenty- third lag. It is difficult to judge visually if these points have significant value. Further testing will be needed.



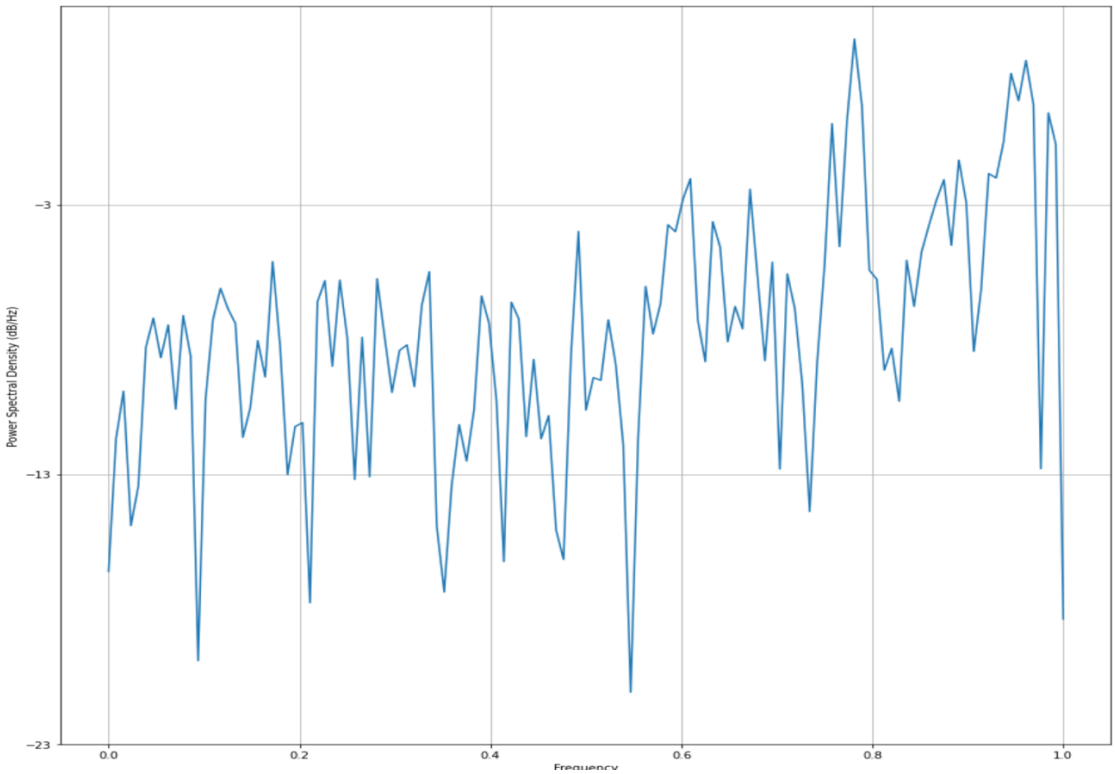
Partial Autocorrelation shows the same correlation as autocorrelation but only the direct effects are shown. All intermediary effects are removed. (Radecic, 2021) The model used 31

days as the lag. The correlation value drops off after the first lag.

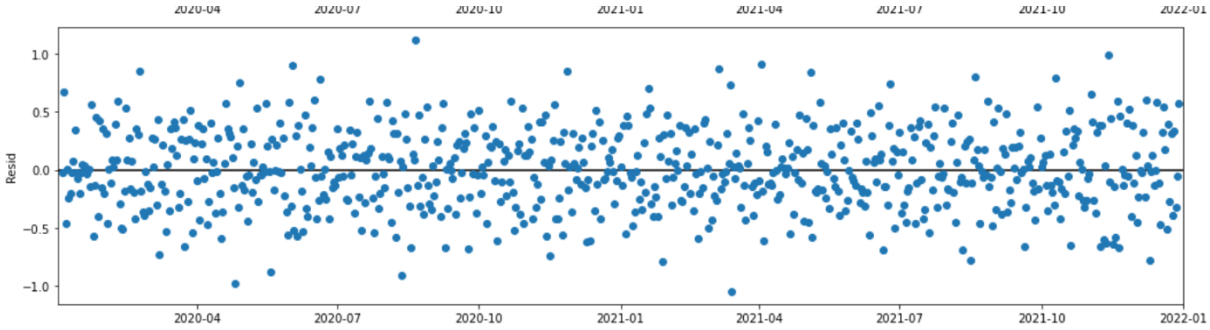


The spectral density plot shows a frequency domain representation of a time series. If the spectral density is constant, the dataset is considered white noise. (12.1 Estimating the Spectral Density, n.d.) The graph is not constant so therefore it is not white noise. This means that it can

be used to make predictions.



The residuals appear to be evenly distributed across the graph. There are no clear patterns and there are an even number of points above and below zero.



2. Using auto_arima from the pmdarima package, the best model was determined to be SARIMAX (2,0,0) (2,1,0,12). A summary and diagnostic plot were created to further confirm the model. The summary is shown below:

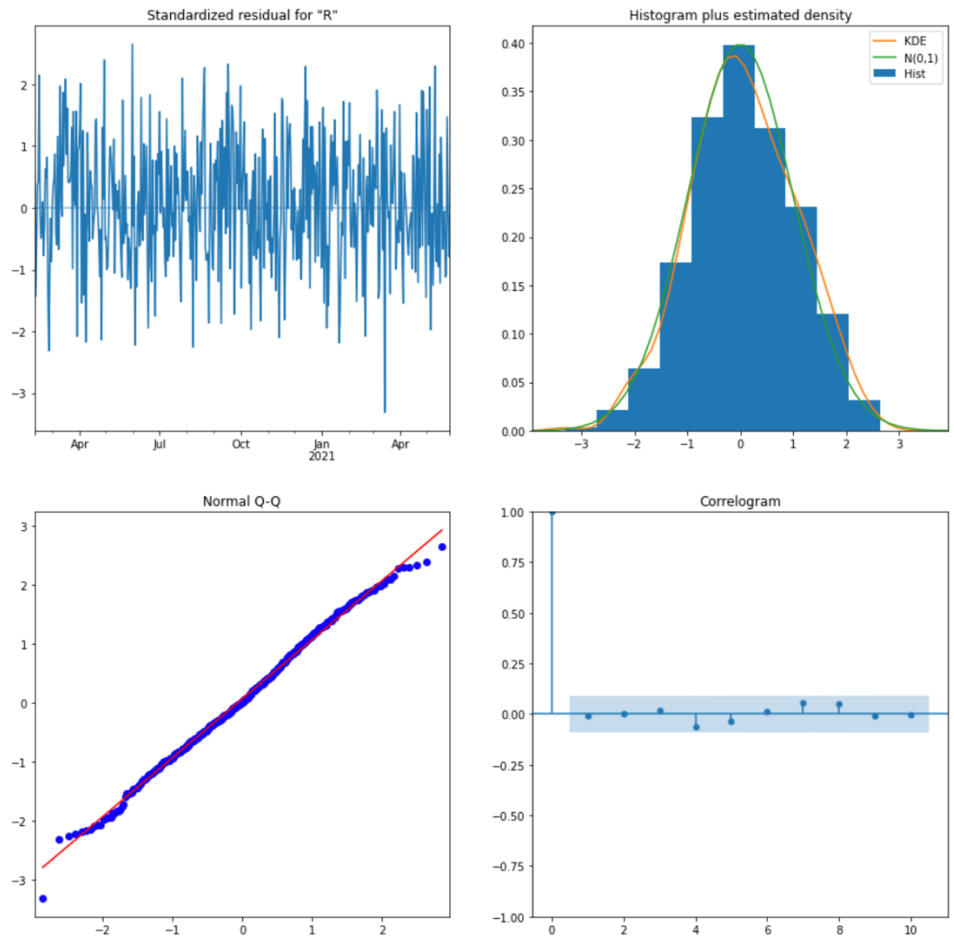
SARIMAX Results

Dep. Variable:	Revenue	**No. Observations:**	512
Model:	SARIMAX(3, 0, 2)x(2, 1, [], 12)	**Log Likelihood**	-364.295
Date:	Wed, 02 Feb 2022	**AIC**	744.589
Time:	10:12:34	**BIC**	777.862
Sample:	01-02-2020	**HQIC**	757.676
	- 05-27-2021		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.3543	0.136	9.922	0.000	1.087	1.622
ar.L2	-0.0399	0.157	-0.255	0.799	-0.347	0.267
ar.L3	-0.3323	0.111	-2.998	0.003	-0.550	-0.115
ma.L1	-0.8597	0.140	-6.124	0.000	-1.135	-0.585
ma.L2	0.1297	0.105	1.235	0.217	-0.076	0.336
ar.S.L12	-0.7545	0.048	-15.698	0.000	-0.849	-0.660
ar.S.L24	-0.4066	0.048	-8.469	0.000	-0.501	-0.313
sigma2	0.2728	0.019	14.191	0.000	0.235	0.311

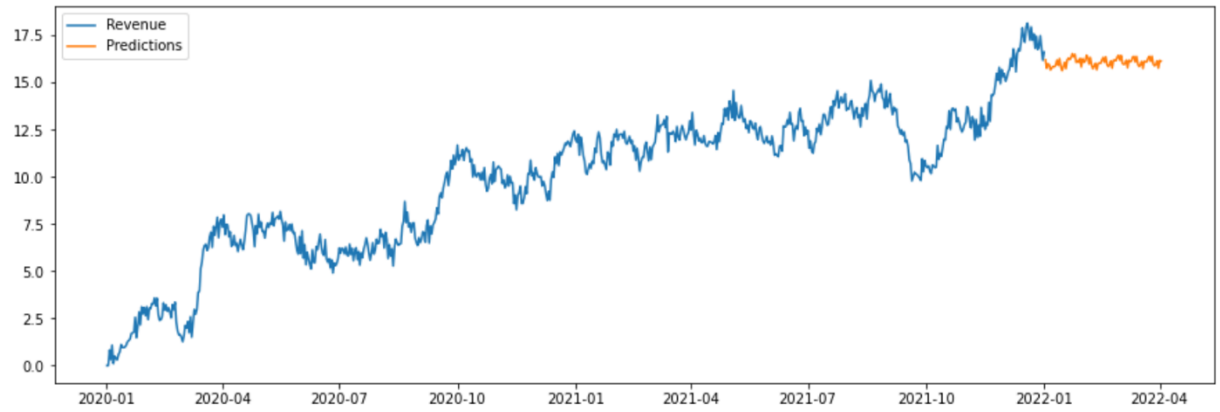
Ljung-Box (L1) (Q):	0.02	**Jarque-Bera (JB):**	1.18
Prob(Q):	0.88	**Prob(JB):**	0.55
Heteroskedasticity (H):	0.99	**Skew:**	-0.03
Prob(H) (two-sided):	0.95	**Kurtosis:**	2.76

Since the Prob(Q) and the Prob (JB) are but larger than 0.05, we can conclude that the residuals are both normal and uncorrelated. Next, a diagnostics plot was created (shown below). The standardized residual plot has no obvious pattern. The lines of the histogram are close to each other showing that the data distribution is almost normal. The data points on the normal Q-Q are close to the red line. The correlogram shows that no lags lie outside of the confidence intervals.



Since all these factors confirm that the model is good, the test data was used to make a forecast.

3. Forecast using the model for the next quarter.



4. The mean squared error was 4.28 million and the root mean squared error was 2.07 million. The forecast bias was 0.79 million.

5. Code to support the implementation of the time series model attached as: D213 Task 1 Time Series Modeling.pdf

Part V: Data Summary and Implications

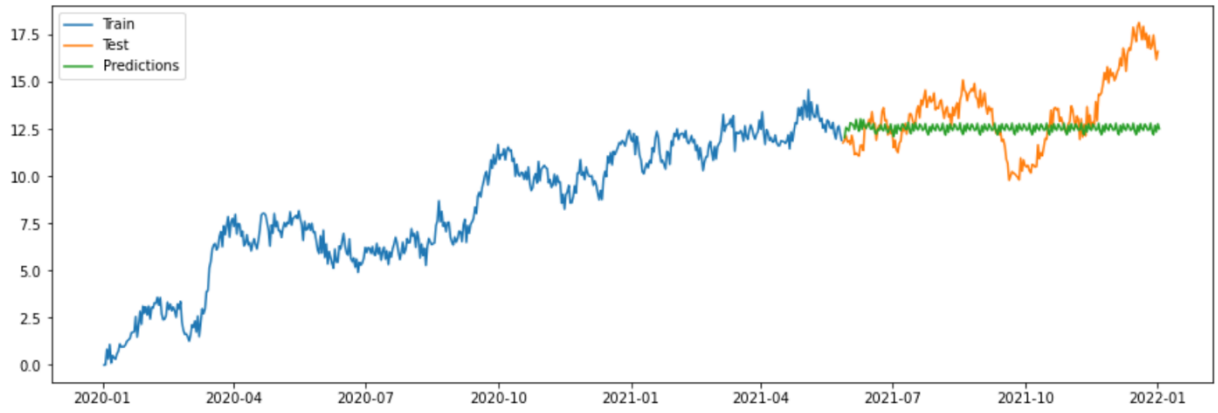
1. After fitting the training data using auto_arima, the best model was determined to be a SARIMAX model with values of (3,0,2) (2,1,0) (12). This model had the smallest AIC. The summary and a diagnostics plot were reviewed to analyze the model. The Prob(Q) and Prob (JB) were both greater than 0.5. All plots in the diagnostics plot are well within reason.

Based on a 95% confidence interval, there is a 5% chance that the true value will fall outside the range of 11.61 to 13.65. Both the true value and the forecast value fall within this range.

A period of three months was chosen because it allows the decision makers time to implement any changes that are needed. While the predictions do show the ups and downs of the pervious data, the upwards trend is not reflected. This could affect the decisions that are made based on these predictions. Further study may be needed to find if some outside factor, not accounted for in this analysis, affects the model.

The metrics used to evaluate the models error metric were the mean square error (MSE), the root mean squares error (RMSE), and the mean forecast error. The MSE was 4.28 and the RMSE was 2.07 both of which are low values of errors. The mean forecast error was 0.79 meaning that this model tends to under forecast. The forecasted values will be less than the actual values. (Browniee, 2017)

2. Annotated visualization of the forecast of the final model compared to the test set:



3. The time series analysis shows that there is a periodic fluctuation in revenue. Further study could reveal why this is occurring. The model was able to forecast future profits. Given that past trend was positive, it is reasonable to think that it will remain so for the future. More investigation is needed to identify what is affecting the trend. Based on this model, I would suggest that the company maintain the current course.

Works Cited

12.1 Estimating the Spectral Density. (n.d.). Retrieved January 2022, from PennState: https://online.stat.psu.edu/stat510/lesson/12/12.1

Browniee, J. (2017, February 1). *Times Series Forecasting Preformance Measures with Python*. Retrieved February 2022, from Machine Learning Mastery: https://machinelearningmastery.com/time-series-forecasting-performance-measures-with-python/

Frost, J. (2022). *Autocorrelation and Partial Autocorrelation in Time Series Data*. Retrieved January 2022, from Statistics By Jim: https://statisticsbyjim.com/time-series/autocorrelation-partial-autocorrelation/

Radecic, D. (2021, July 19). *Time Series From Scratch - Autocorrelation and Partial Autocorrelation Explained*. Retrieved January 2022, from Towards Data Science: https://towardsdatascience.com/time-series-from-scratch-autocorrelation-and-partial-autocorrelation-explained-1dd641e3076f

Stationarity. (n.d.). Retrieved January 2022, from Engineering Statistics Handbook: https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm