**D213 PA**

Jon Isak

D213: Data Mining II

Contents

[**D212 Task 3 3**](#_3xni6yjcaxld)

[**Part I: Research Question 3**](#_on428cwqsz8k)

[**Part II: Market Basket Justification 3**](#_mytzr5pv706l)

[**Part III: Data Preparation and Analysis 4**](#_103541cwhku6)

[**Part IV: Data Summary and Implications 6**](#_gbpyfqak8j0a)

[**Part V: Attachments 7**](#_ixwt1kkbz9xk)

## 

## D212 Task 1

## Part I: Research Question

A1. This data analysis will look into the question, how does daily revenue vary over the first two years of operation for the company, and can we predict future revenue?

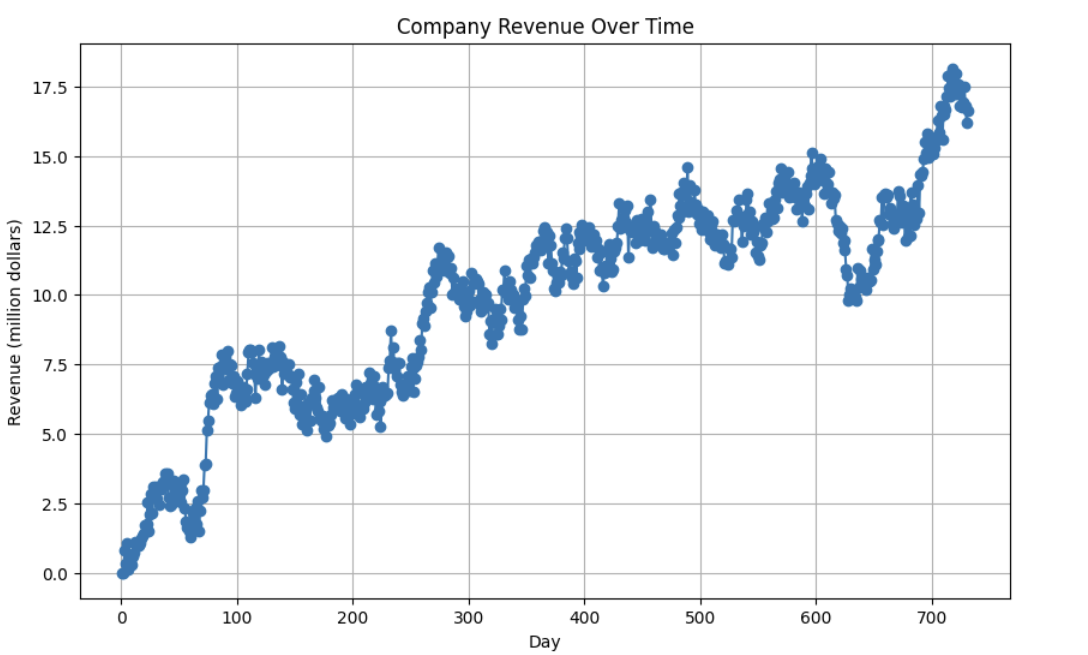
A2. The goal of the data analysis is to use time series modeling techniques to determine the trends within the data and develop a model to forecast future revenue.

## Part II: Method Justification

B. One assumption of a time series model is that it is stationary. Stationary means that statistical properties of a time series do not change over time (Statisticssolutions, n.d.). Autocorrelation refers to the degree of similarity between a given time series and a lagged version of itself over successive time intervals (Taylor, n.d.).

## Part III: Data Preparation

C1. The graph shows the revenue increasing as time progresses.



C2. The data includes data 731 days. Revenue is recorded for each day. The data is continuous with no gaps.

C3. The Augmented Dickey-Fuller test was used to assess the data. The statistic was -1.9, the p-value was .3, and critical values were -3.4 at 1% and -2.86 at 5%, and -2.56 at 10%. The results indicate that the statistic is more than the critical values and the p-value is higher than the .05 threshold. These results indicate that the data is not stationary.

C4. The data checked for nulls and data types. Since the data is non-stationary, the first step was to transform the data to make it stationary using differencing. The transformed data showed a p-value and statistic less than the threshold and critical values, suggesting that the data is now stationary. The data was then split into a training set and a test set.

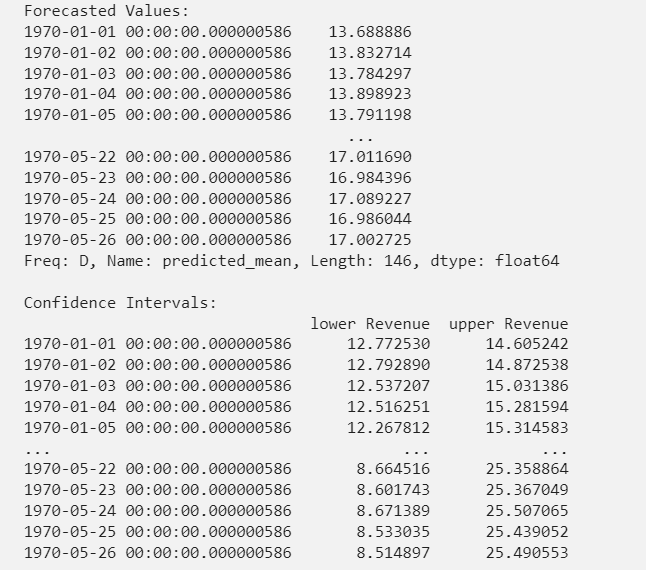
C5. The data is attached.

## Part IV: Model Identification and Analysis

D1. The presence of a seasonal component in the time series data was observed through the seasonal decomposition analysis. The trend line showed the general direction of the data, indicating an overall increasing pattern. The ACF plot showed the correlation of the time series data with its own lagged values at different time lags. Peaks outside the confidence interval in the ACF plot indicated significant autocorrelation at those lags. The spectral density plot displayed the frequency components present in the time series data. It showcased the distribution of power across different frequencies within the data. Peaks in the spectral density indicated dominant frequencies in the data, highlighting potential periodic behaviors or significant changes over time. The decomposed time series data exhibited a clear seasonal pattern along with a discernible trend.

D2. The identification of the seasonal component in the time series data involved employing both the PACF and ACF techniques. Additionally, a seasonal decomposition analysis using the seasonal\_decompose method from statsmodels was performed to explicitly identify the seasonal patterns. The decomposition revealed clear seasonal fluctuations in the data. Autocorrelation analysis was extended to include investigation of seasonal lags, emphasizing the examination of significant peaks or patterns at multiples of the seasonal frequency. Peaks at these lags confirmed the presence of pronounced seasonal effects, further supporting the inclusion of a seasonal component in the ARIMA mode.

D3. This forecast estimate revenue on each day from 1/1/1970 to 5/26/1970. The confidence intervals indicate the range within the true revenue likely falls.

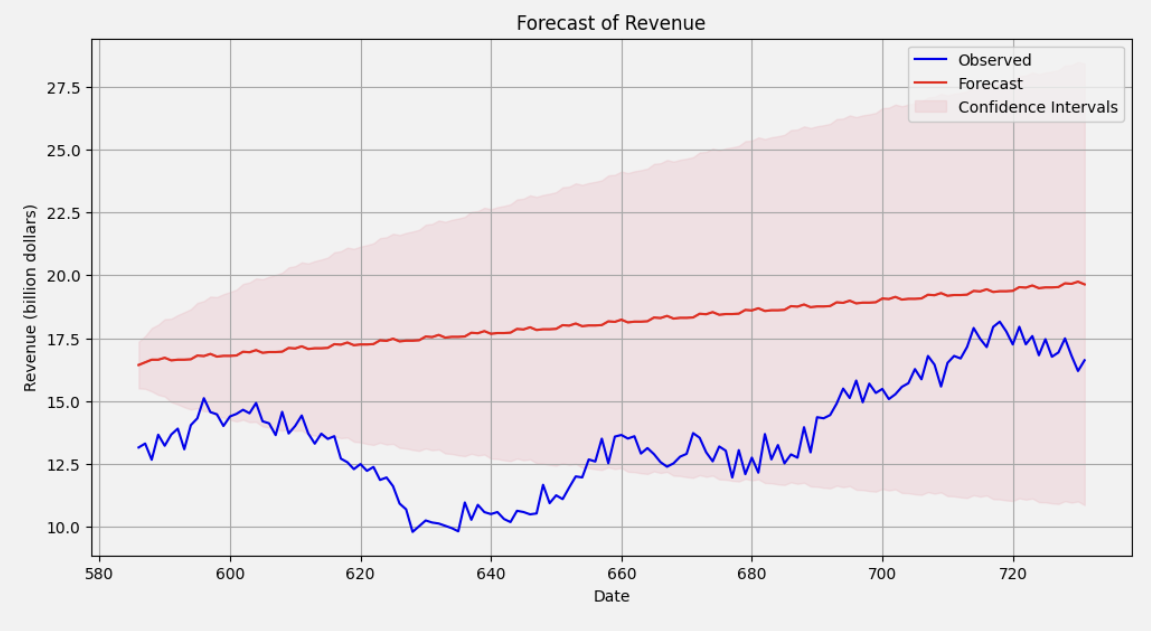


D5. The code is attached.

## Part V: Data Summary and Implications

E1. The selection of the ARIMA model involved a comprehensive analysis of the time series data to identify the most suitable parameters. By examining the autocorrelation and partial autocorrelation functions, alongside seasonal decomposition techniques, it was determined that the non-seasonal and seasonal orders to be 1, 1, 0, and 0, 1, 1, 7, respectively. This approach showed trends and seasonality effectively within the model. Regarding model evaluation, the forecast length was aligned with the test dataset's duration of 146 sequential days. This choice facilitated an extensive assessment of the model's performance on data that wasn't utilized for training. The forecast's prediction intervals were calculated, offering an understanding of the uncertainty surrounding the predictions. These intervals portray the potential range within which the actual revenue values might fall, indicating the forecast's reliability and possible variability. Assessment of the model's predictive performance was conducted using various error metrics. The mean absolute error of 0.161 million signifies the average magnitude of errors between the model's predictions and the actual values. Similarly, the mean squared error MSE of 0.032 provides insight into the average squared deviation of predictions from the observed data. The root mean squared error of 0.179 serves as an indicator of the average magnitude of prediction errors by the model. Overall, these evaluation metrics collectively suggest that the ARIMA model performed well. The low values of MAE, MSE, and RMSE indicate a high level of accuracy in the model's forecasting capabilities. However, it's important to note the wider prediction intervals, emphasizing the inherent uncertainty in predicting future revenue, despite the model's overall good performance.

E2. The graph shows the forecasted daily revenue against the observed revenue. The blue line represents the observed revenue data. As the graph progresses, the red line for forecasted revenue gradually diverges from the observed data. The shaded pink area shows the confidence intervals. The widening of confidence intervals towards the end indicates an increase in uncertainty as the forecast extends further.



E3. Based on the results, the company can continue investigating what factors led to the decline in revenue and rise in revenue. Furthermore, the company can look into how they can maximize revenue during peaks and mitigate revenue dips. They can also continue monitoring and refining the model to get more accurate predictions as more data is collected. The company can also create goals based on the revenue forecast to stay within the budget.

## Part VI: Reporting

F. The file is attached.

G. Kang, C. (2020, June 15). *ARMA Models*. Chan`s Jupyter. Retrieved December 18, 2023, from https://goodboychan.github.io/python/datacamp/time\_series\_analysis/2020/06/15/01-ARMA-Models.html

H.

Minita. (n.d.). *Interpret the partial autocorrelation function (PACF) - Minitab*. Support. Retrieved December 18, 2023, from https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/time-series/how-to/partial-autocorrelation/interpret-the-results/partial-autocorrelation-function-pacf/

Statisticssolutions. (n.d.). *The Stationary Data Assumption in Time Series Analysis*. Statistics Solutions. Retrieved December 14, 2023, from https://www.statisticssolutions.com/stationary-data-assumption-in-time-series-analysis/

Taylor, S. (n.d.). *Autocorrelation - Overview, How It Works, and Tests*. Corporate Finance Institute. Retrieved December 14, 2023, from https://corporatefinanceinstitute.com/resources/data-science/autocorrelation/