Nicole Burns

D214—Data Analytics Capstone

**A. Research Question (justification, context, and hypothesis):**

“Can a SARIMA model effectively forecast Apple Inc’s quarterly revenue with a model accuracy of >80%?”

Since its inception in 1976, Apple has grown to become the first company with a 3 trillion-dollar market value (Vlastelica, 2023). Its continuing growth makes Apple of keen interest to stakeholders such as investors, analysts, and competitors. These stakeholders make critical decisions based on the company’s financial health and growth trajectory. The ability to predict future sales with a high degree of accuracy can inform a range of strategic decisions, from investment strategies to marketing and resource allocation. This study seeks to evaluate whether a SARIMA model can effectively forecast Apple’s quarterly revenue with an accuracy of >80%.

The following are the hypotheses:

**Null hypothesis:** H0—A SARIMA model cannot effectively forecast Apple’s quarterly revenue at a model accuracy of > 80%.

**Alternative hypothesis:** H1—A SARIMA model can effectively forecast Apple’s quarterly revenue at a model accuracy of > 80%.

**B. Data Collection Process:**

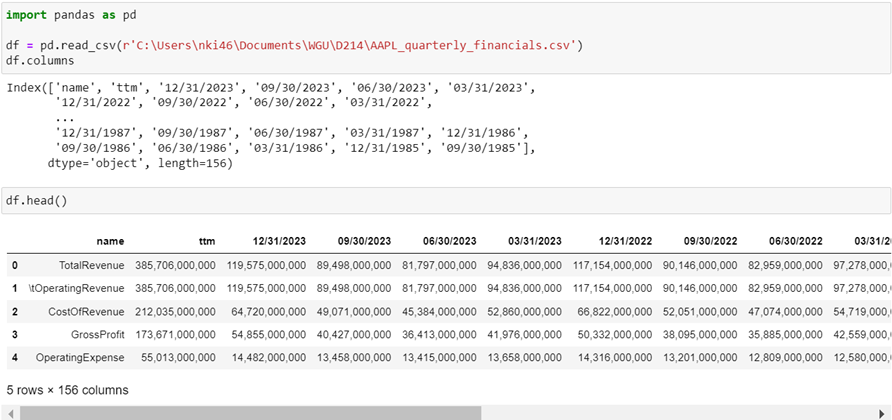
The data needed for the analysis is available via the Yahoo Finance website. The data is found under the ‘Income Statement’ and ‘Quarterly’ tabs. The data available spans from quarters ending in 1985 to 2023. Prior to data cleaning and transformation, the dataset contained 51 rows representing financial data names and 156 columns representing the dates each quarter ended. Names include the following:

* Total Revenue
* Cost of Revenue
* Gross Profit
* Operating Expense
* Operating Income
* Net Non-Operating Interest Income Expense
* Interest Income Non-Operating
* Interest Expense Non-Operating
* Total Other Finance Cost
* Pretax Income
* Tax Provision
* Net Income Common Stockholders
* Net Income
* Diluted NI Available to Com Stockholders
* Basic EPS
* Diluted EPS
* Total Operating Income as Reported
* Total Expenses
* Net Income from Continuing & Discontinued Operation
* Normalized Income
* Interest Income
* Interest Expense
* Net Interest Income
* EBIT Quantitative
* Reconciled Cost of Revenue
* Reconciled Depreciation
* Net Income from Continuing Operation Net Minority Interest
* Total Unusual Items Excluding Goodwill
* Total Unusual Items
* Normalized EBITDA
* Tax Rate for Calcs
* Tax Effect of Unusual Items

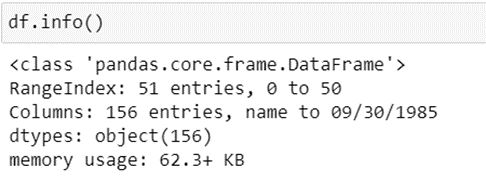
One of the advantages of this data-gathering methodology includes the dataset being publicly available. However, the disadvantage includes accessibility and exportation being limited. Individuals will need a subscription to view all dates available or download the dataset. I circumvented this issue by signing up for a free trial.

**C. Data Extraction and Preparation:**

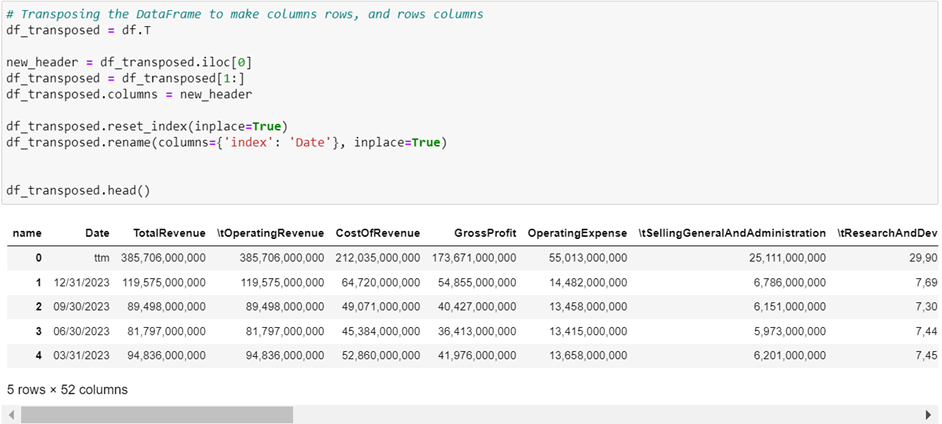
The analysis will be performed in Jupyter Notebook using Python. The advantages of using Python include it being a very versatile programming language that can be used for a variety of tasks. Such includes machine learning techniques, the method being used for this analysis (SudoPurge, 2021). Python also has libraries specifically designed for data analysis, and time series forecasting, such as Pandas, NumPy, Statsmodels, and Scikit-learn, among others. It also is widely used in the industry and is supported by many organizations (GFG, 2023). While I didn’t encounter any issues using Python, one disadvantage includes its consumption of a considerable amount of memory, causing performance issues when handling large datasets.



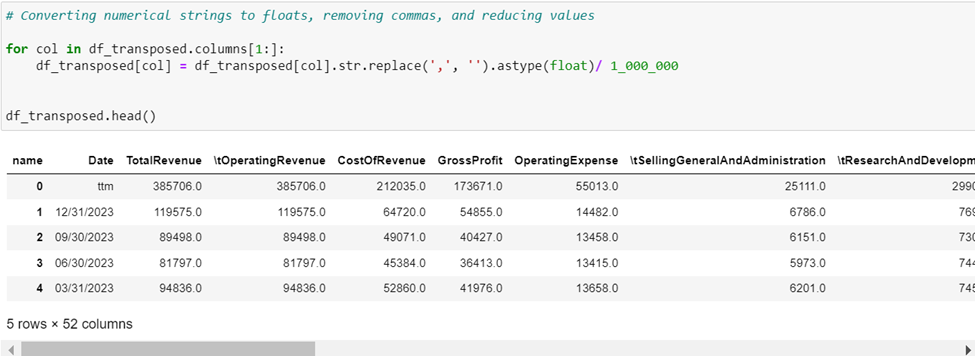
The data preparation process included the importation of the CSV file mentioned in the previous section.



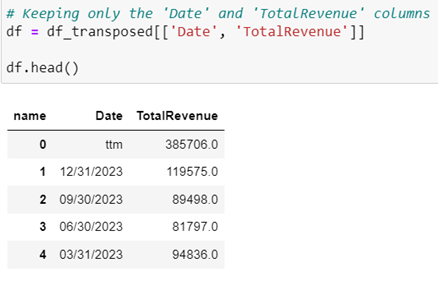
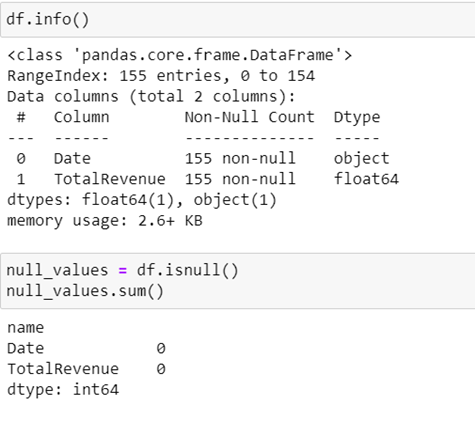
Upon inspection to verify that the CSV was properly imported, .info() was used to inspect the number of rows, columns, and their respective datatypes.



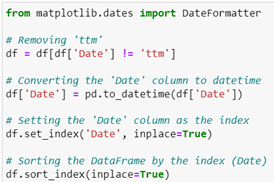
The dataset needed to be transposed, making the dates the row values, and the names the column values.



Upon verifying that the transposing was completed successfully, the numerical values were converted from objects to floats, commas were removed, and the values were reduced to improve readability.

All columns not needed for the analysis were removed and I verified whether any null values were present.



Lastly, row ‘ttm’ (trailing twelve months) was removed. The ‘Date’ column’s format was changed from ‘object’ to ‘datetime’, used as the index, and sorted.

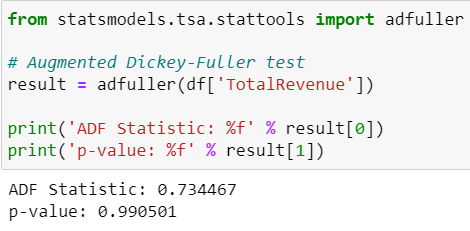
**D. Data Analysis (advantage & disadvantage):**

My decision to use a SARIMA model was based on the characteristics of the data, which displayed a clear trend and seasonality, both of which are well-handled by SARIMA models. A SARIMA model can account for seasonality within the data and an ARIMA cannot (Patra, 2023). The initial ADF test confirmed the need for differencing, indicating that SARIMA was a fitting choice. An advantage of using a SARIMA model includes it being well-suited for the dataset because it can model both the non-stationary trend and the seasonality present in Apple’s revenue data. One disadvantage of using a SARIMA model is it can become quite complex and may not capture sudden changes in the trend or seasonality or external factors which might limit their forecasting accuracy in the case of market shifts or new product introductions.

My analysis process included:

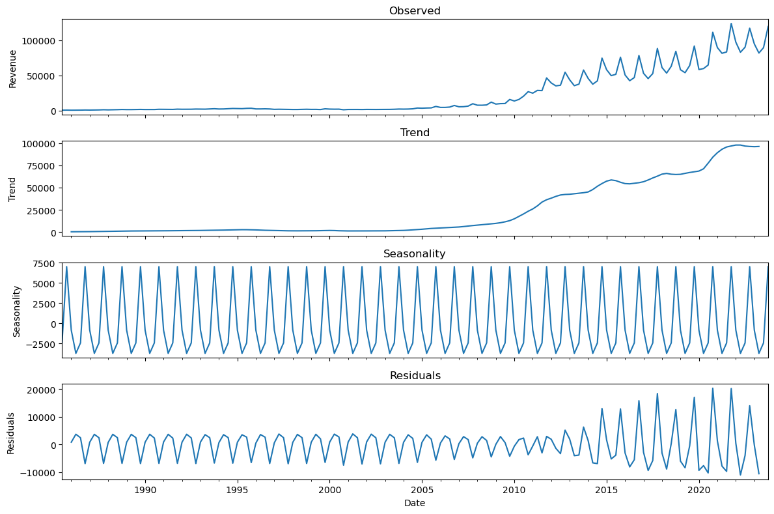
1. Stationarity Testing:

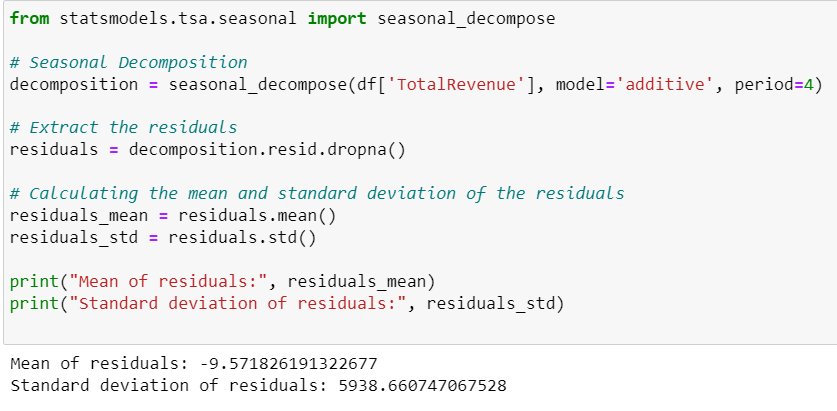
* I started by using the Augmented Dickey-Fuller (ADF) test to check for stationarity in the time series data. This is an essential step in time series analysis because most models require the data to be stationary.
* The high p-value suggested that the series is non-stationary.



1. Seasonal Decomposition:

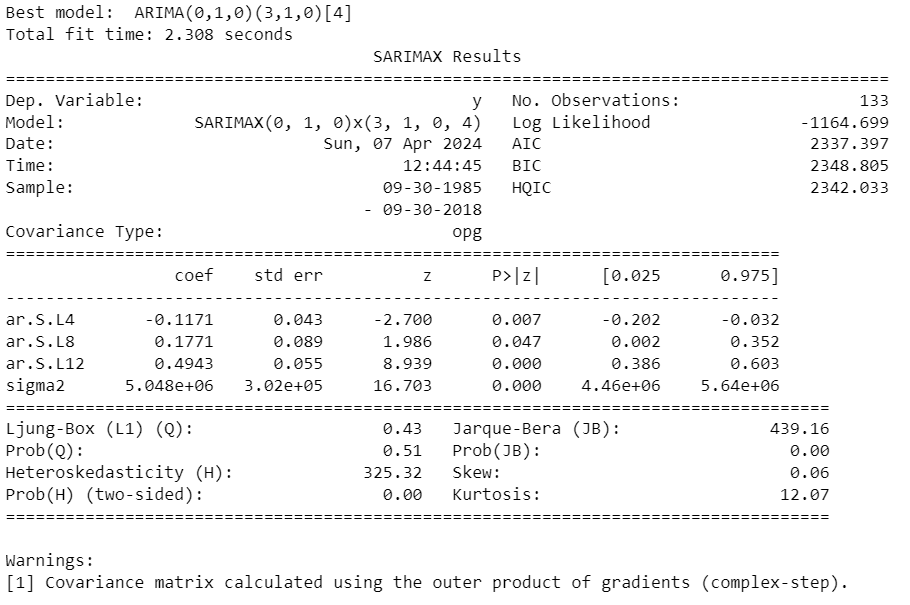
* I then applied seasonal decomposition to separate the time series into trend, seasonal, and residual components. This helps to understand underlying patterns in the data.
* I also calculated the mean and standard deviation of the residuals, which aids in understanding the noise in the data after accounting for trend and seasonality.





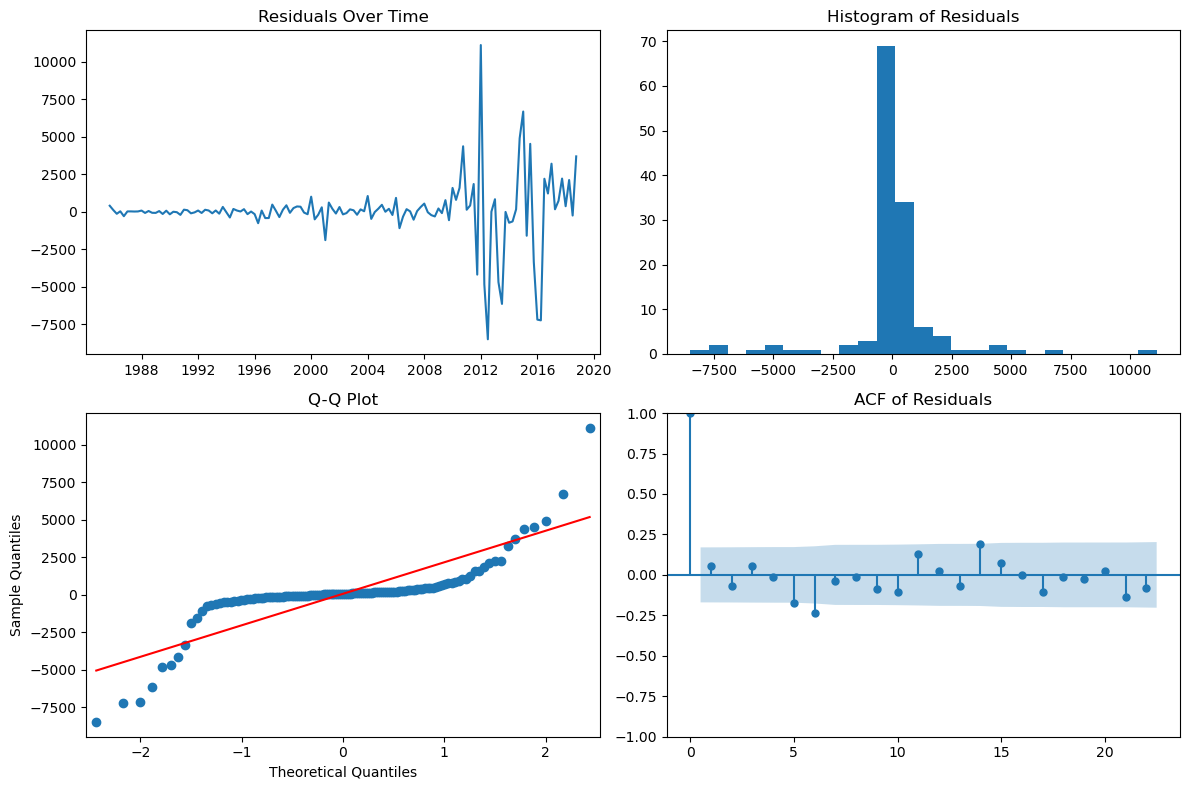
1. The SARIMA model:

* Given the seasonality and non-stationarity in the data, I chose a Seasonal Autoregressive Integrated Moving Average (SARIMA) model, which is designed to handle both seasonality and non-stationarity.
* I used the ‘auto\_arima’ function to obtain the best parameters for the SARIMA model.



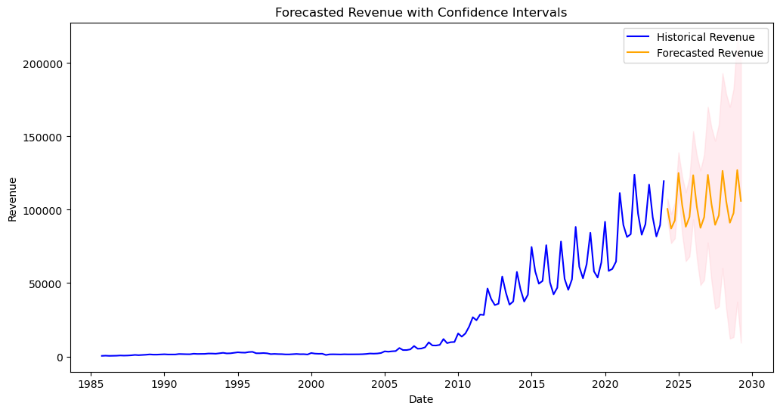
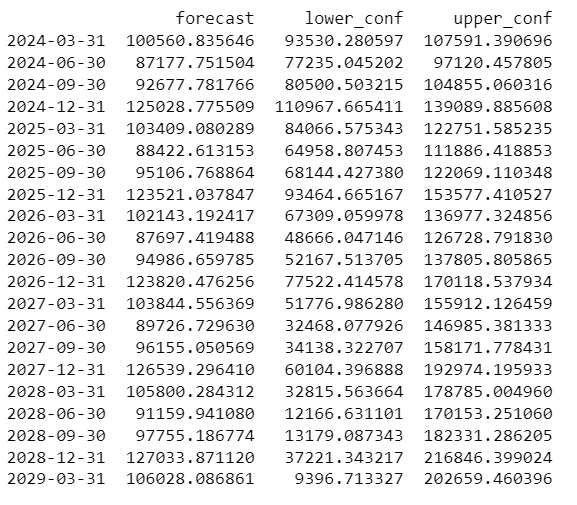
1. Model Fitting:

* After the SARIMA model was fitted to the data, the model parameters suggest that both seasonal and non-seasonal components are significant. The Ljung-Box test and the histogram of residuals suggest that the residuals are random, which is a good sign. However, the Q-Q plot indicates some deviation from normality, especially in the tails.



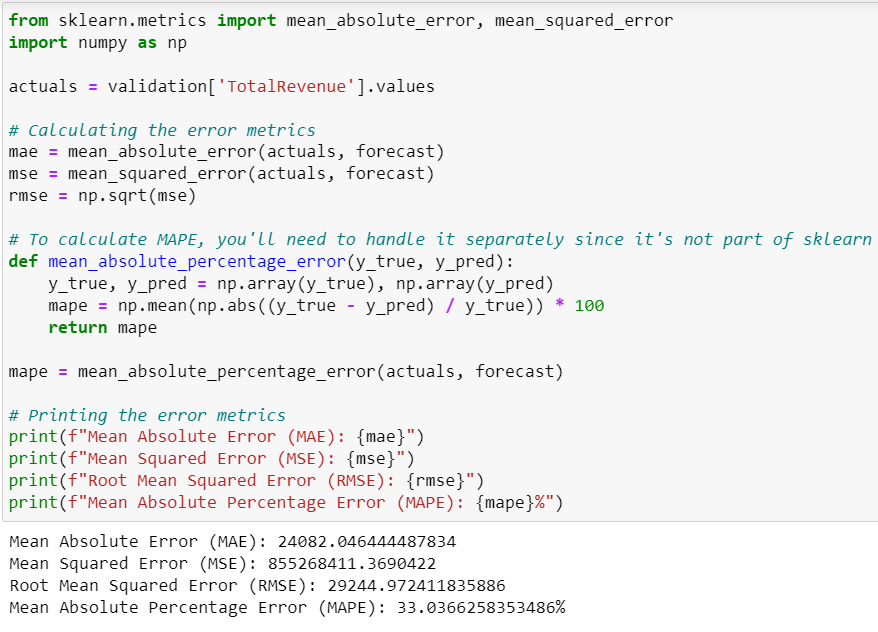
1. Forecasting:

* Using the SARIMA model, I forecasted future revenues and visualized these with historical data, showing confidence intervals.
* I created forecasts and accompanying confidence intervals for 21 quarters into the future.



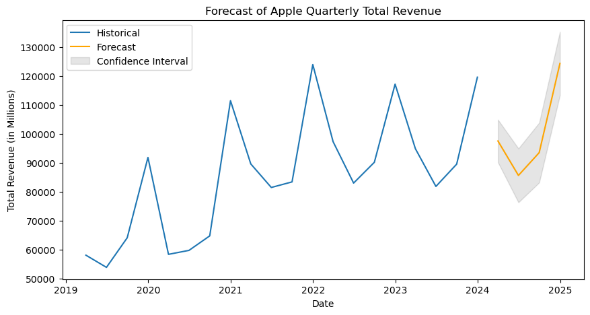
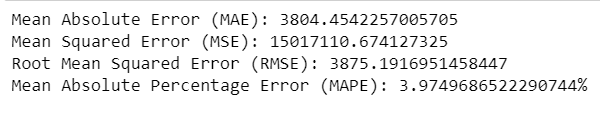
1. Error Metric Calculations:

* To evaluate the accuracy of my model, I calculated the error metrics such as MSE, RMSE, MAE, and MAPE.
* MAPE or the percentage of error resulted in 33.0%, accuracy is inferred as ~67%



**E. Results & Recommendations:**

The forecasting plot showing the predicted revenue for future quarters along with confidence intervals displays wide confidence intervals suggesting considerable uncertainty in the forecasts, especially as the forecast increases. Because of this, I decided to try experimenting with focusing on short-term historical values and projections. Taking a look at the last five years and projecting a year into the future yielded the following results:

The new SARIMA model results show an error percentage of 3.97% or ~96% accuracy. These results suggest the model has more proficiency in short-term forecasting, which can be a limitation. Forecasting too far into the future creates uncertainty that can be due to unexpected changes in the market. A recommended course of action would be to use the SARIMA model as a starting point for forecasting revenue, forecasting mainly into the near future. However, the model should always only be used as a guide, not a prediction, and combined with business intuition and other market insights. Keeping the previous statements in mind, two directions for future study are as follows:

1. Incorporate external factors such as consumer sentiment, competitor strategies, economic indicators, or product launch impact to create a more powerful model in accounting for external shocks or trends.
2. Consider pairing other forecasting techniques, or machine learning models, with the SARIMA model to leverage the strengths of different approaches and potentially improve accuracy.

**F. Sources:**

GfG. (2023, June 12). SAS vs R vs Python. GeeksforGeeks. Retrieved April 3, 2024, from <https://www.geeksforgeeks.org/sas-vs-r-vs-python/>

Patra, T. D.-D. (2023, September 24). *Sarima vs Arima for Timeseries Analysis Model*. Medium. Retrieved April 3, 2024, from <https://dhirajpatra.medium.com/sarima-vs-arima-for-timeseries-analysis-model-a600ab544b1f>

SudoPurge. (2021, March 16). *Python vs. R for data science*. Medium. Retrieved April 3, 2024, from <https://towardsdatascience.com/python-vs-r-for-data-science-cf2699dfff4b>

Vlastelica, R., & Bloomberg. (2023, June 30). *Apple just made history by becoming the first company with a $3 trillion market value-’all powered by a device people look at 4 hours a day’*. Fortune. Retrieved April 3, 2024, from <https://fortune.com/2023/06/30/apple-history-3-trillion-market-value/>

Yahoo! (2024, April 1). *Apple Inc. (AAPL) income statement*. Yahoo! Finance. Retrieved April 1, 2024, from <https://finance.yahoo.com/quote/AAPL/financials>