University of Hertfordshire UH

School of Physics, Engineering and Computer Science

MSc Advanced Research Topic in Data Science

7PAM2016-0105-2024

Department of Physics, Astronomy and Mathematics

Assignment Title:

Time Series Modelling Case Study

Student Name and SRN:

Hari Bahadur Gharti Magar (22075765)

GitHub address: https://github.com/harimagar/Advanced-Research-Topics-in-Data-Science-

Time Series Analysis

Time series analysis is a set of observations taken sequentially in time, typically in chronological order (Rocha et al., 2023). It helps to recognize patterns in data over a certain time period, recognize structures and also helps to forecast future values.

Plotting of the Data

The line plot and histogram are created to visualize the time series and to view the distribution of values of the dataset. The line plot and histogram plot of the jj.csv dataset are shown below:

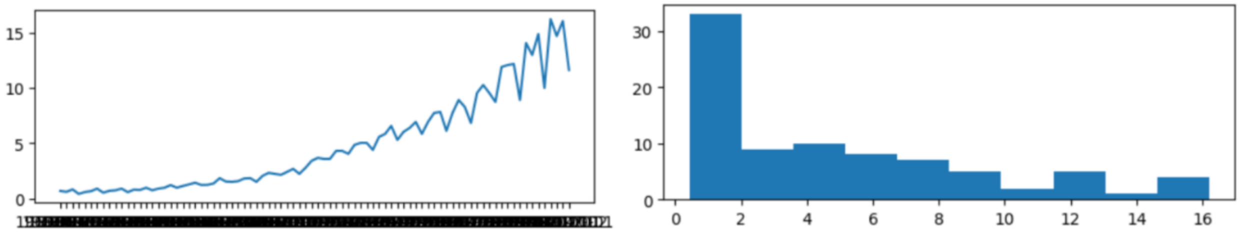


Figure 1 Line Plot and Histogram of jj.csv Dataset

The line plot does not look as good as it can be, therefore, a dataframe is created, and the line plot is constructed again for perfect representations of the data. Further, the growth is shown using linear form where quadratic growth is reduced by using square root transform, making the observations nearly Gaussian. After that, Box-Cox transformation is applied to convert distribution into a more normal distribution using lambda value as 0 to represent log transformation. Also, using the 6-month and 12-month SMA (Simple Moving Averages), the trend is identified, which is shown in the below diagram:

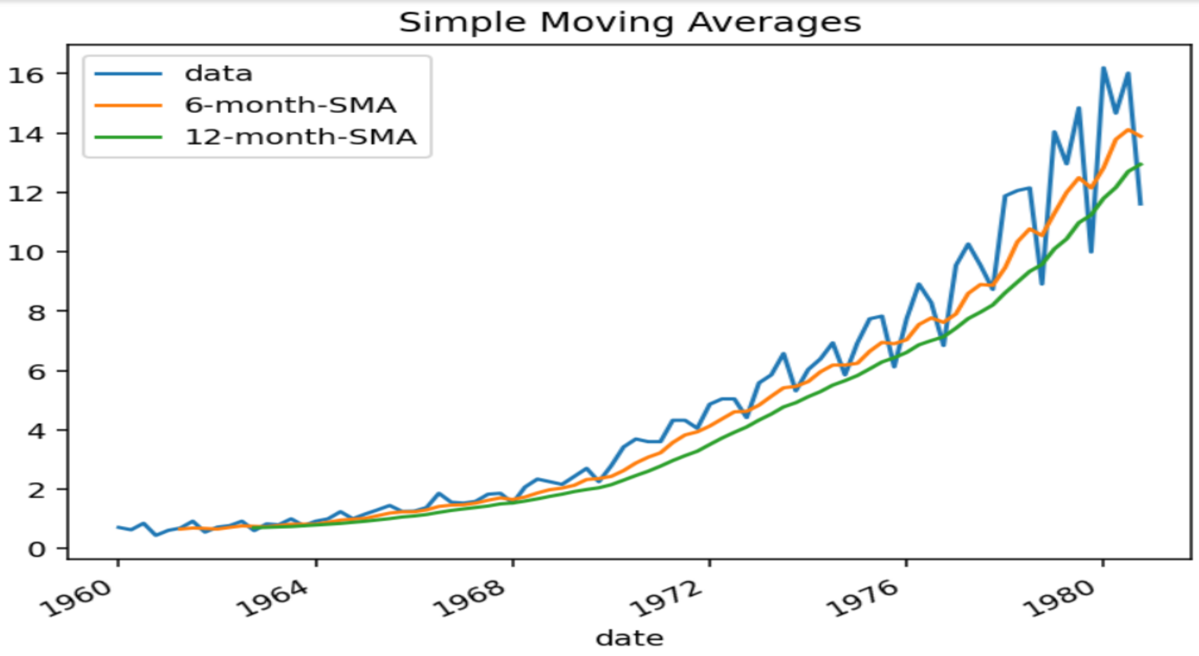


Figure 2 Trend Identification using 6-Month and 12-Month SMA

The trend, seasonality and residual of the jj.csv dataset is shown below, which shows an increasing trend, recurring seasonal patterns and random fluctuations after removing trend and seasonality:

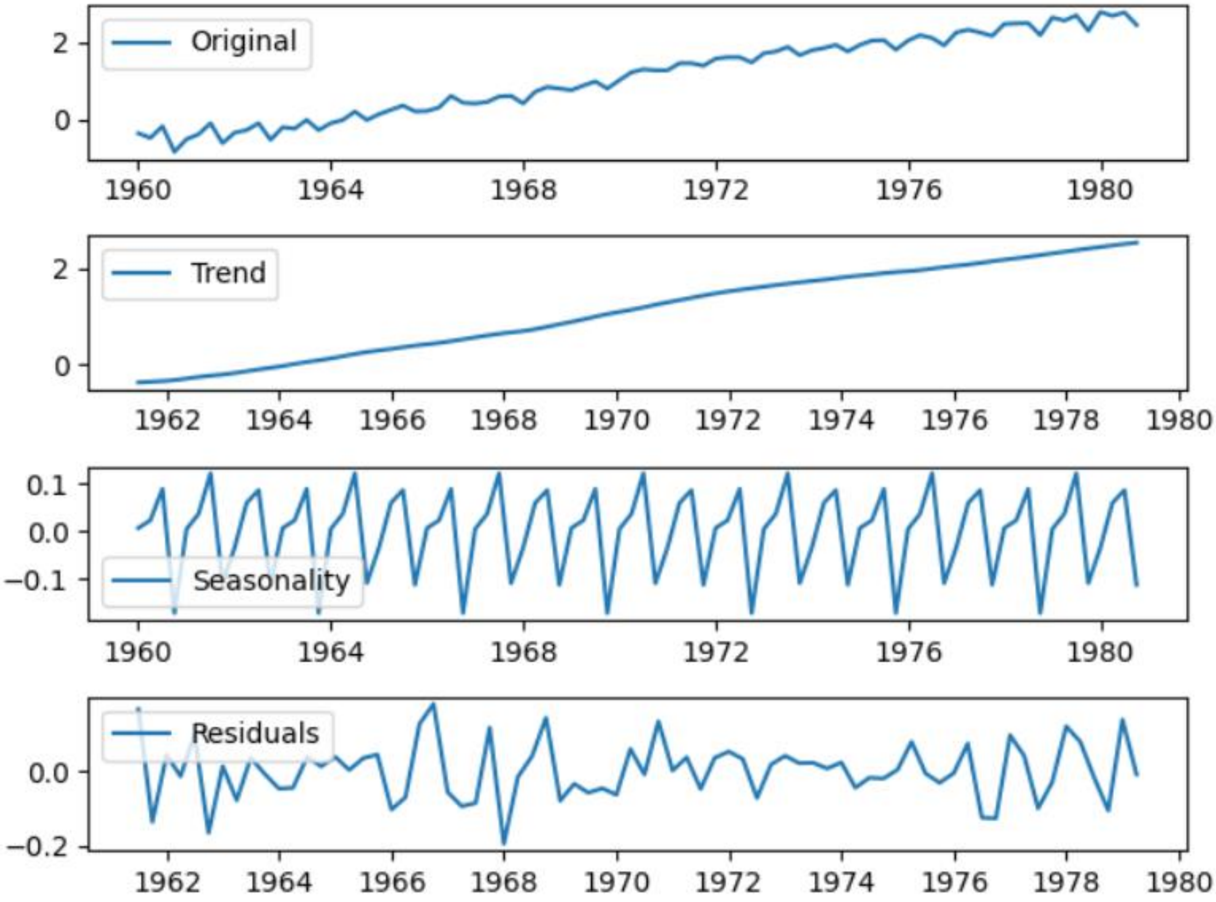


Figure 3 Trend, Seasonality and Residuals of jj.csv Dataset

The coding can be improved through the use of Python functions for different repetitive tasks, including data loading, using exception handling to remove the chance of getting run time errors and also checking for missing values in the dataset.

ACF (Autocorrelation Function)

ACF is used to measure the correlation of a time series against a time-shifted version of itself. It is used to define the correlation of the current value with its past values of a time series (Ye, 2013). It is used to recognize dependencies, detect seasonality, and identify patterns and non-stationarity in the time series. The result of ACF, while applied to the jj.csv dataset, which is shown below, shows that there is no white noise and the time series is not random, there is a persistence or trend over time suggesting non-stationarity along with strong positive dependence.

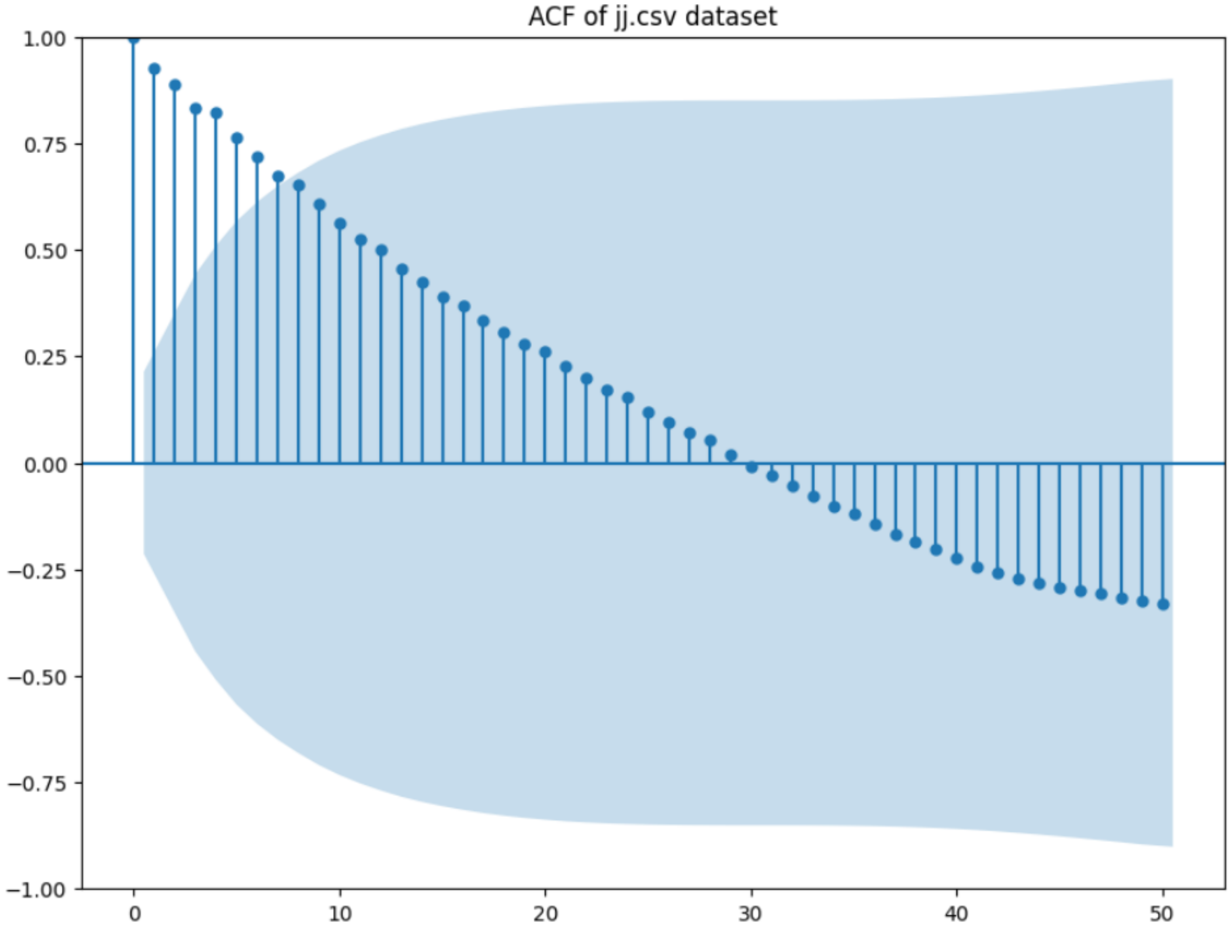


Figure 4 ACF of jj.csv Dataset

Random Walk Time Series

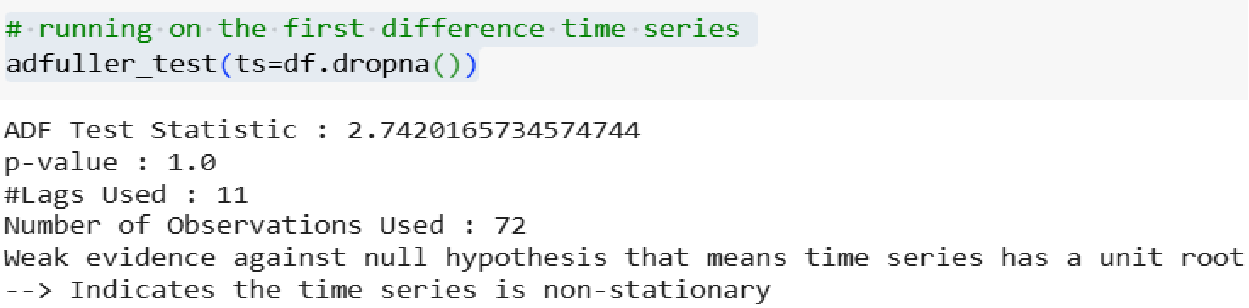
It is a kind of non-stationary time series where each observation is equal to the previous observation plus a random noise or error term. It is particularly used for modelling stock prices (Sarala et al., 2023). In the code, random walk simulation is performed using the UP and DOWN discrete method and normal distribution method, calculated ACF of random walk of non-stationary time series and made the series stationary.

Autoregressive (AR), Moving Average (MA) and ARMA Models

All these models are fundamental tools for time series analysis to forecast data based on past observations (Zhang, 2018). Autoregressive expresses the current value as a linear combination of p previous values plus a random error and predicts future values based on them. The moving average expresses the current value as a linear combination of q previous error terms plus a constant, and it predicts future values based on the noise in the data. ARMA is best applied for stationary time series, which combines both AR and MA models to predict future values. In the code, AR (2), AR (3), MA (2), ARMA (1, 1) and ARMA (2, 2) processes are performed using statsmodels and predefined coefficients.

ADF (Augmented Dicky-Fuller Test)

It is a statistical test in time series analysis to determine whether a given time series is stationary or not (Gianfreda et al., 2023). Stationary time series remain constant over time, and it is important to determine because models like ARIMA assume time series are stationary by default. After performing the ADF test to both the dataset, it is found that both datasets are non-stationary by default. The series is converted to stationary after performing differencing, which is then made suitable for the ARIMA technique that assumes stationarity by default.



```
# running on the first difference time series
adfuller_test(ts=df.dropna())
```

```
ADF Test Statistic : 2.7420165734574744
p-value : 1.0
#Lags Used : 11
Number of Observations Used : 72
Weak evidence against null hypothesis that means time series has a unit root
--> Indicates the time series is non-stationary
```

Figure 5 ADF Test Result of jj.csv Dataset

EWMA (Exponential Weighted Moving Average)

In the code, SMA is calculated for 6 months and 12 months to identify trends. After that, EWMA is applied to provide more weight to recent data. The dataset is then converted into training and testing sets. The training data is then used in Holt-Winters exponential smoothing to identify trends and seasonality. After that, the forecast is done, which is plotted along with training and testing data. The result is shown in the following plot:

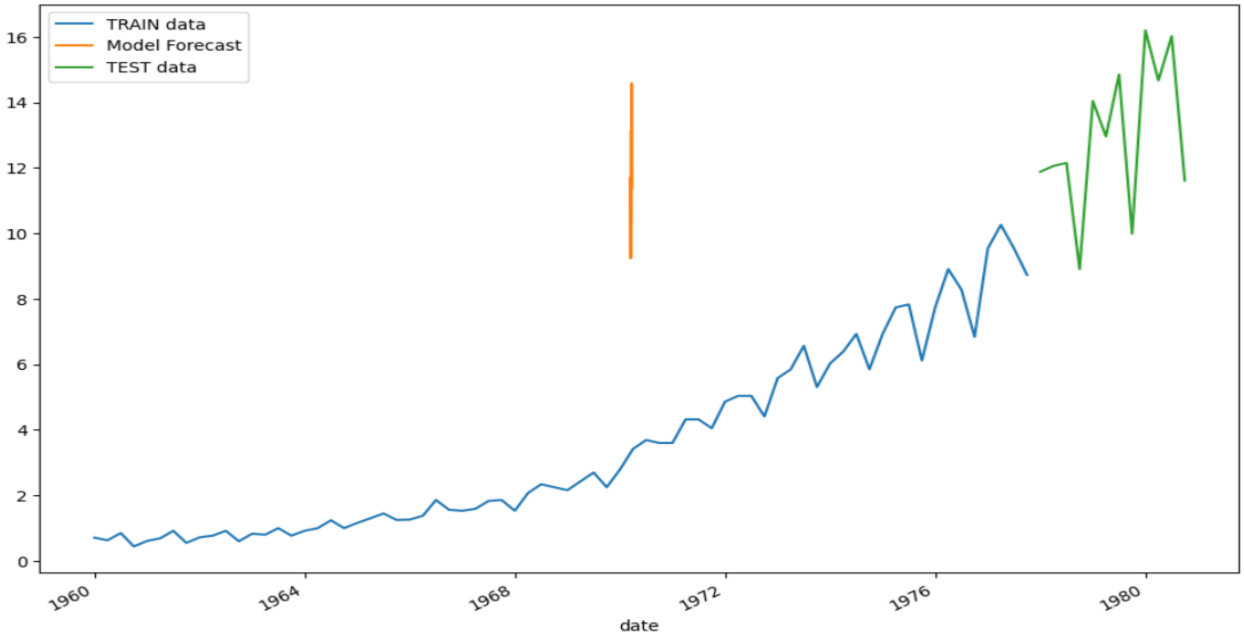


Figure 6 Forecast Result

The train data shows an upward trend with some cyclical or seasonal variations. The model forecast shows a short segment around 1970, which is unusual, and it can be predicted that there is an issue with forecast implementation or visualization. The forecast is not perfectly accurate. The model might be overestimating or underestimating the actual values at certain points. The forecast does not align well with either the training or test data patterns. This can be improved by using an accurate and recent portion of the dataset for training and applying more advanced models.

ARIMA (AutoRegressive Integrated Moving Average) Model

It is a versatile statistical model for forecasting and analyzing data and predicting future values (Arumugam & Natarajan, 2023). It is implemented if the dataset is non-stationary but shows a pattern over time. In the code, the p, d and q are selected as 6, 1 and 3, respectively. The data shows an upward trend with some cyclical or seasonal variations, as shown in the below diagram. The black line shows the forecast in future for 24 periods. It shows the implementation error due to how the time indices were handled, issues with model specification or problems with prediction period alignment. It can be improved by using auto ARIMA instead of manually providing the value, improving model specification and addressing non-stationary.

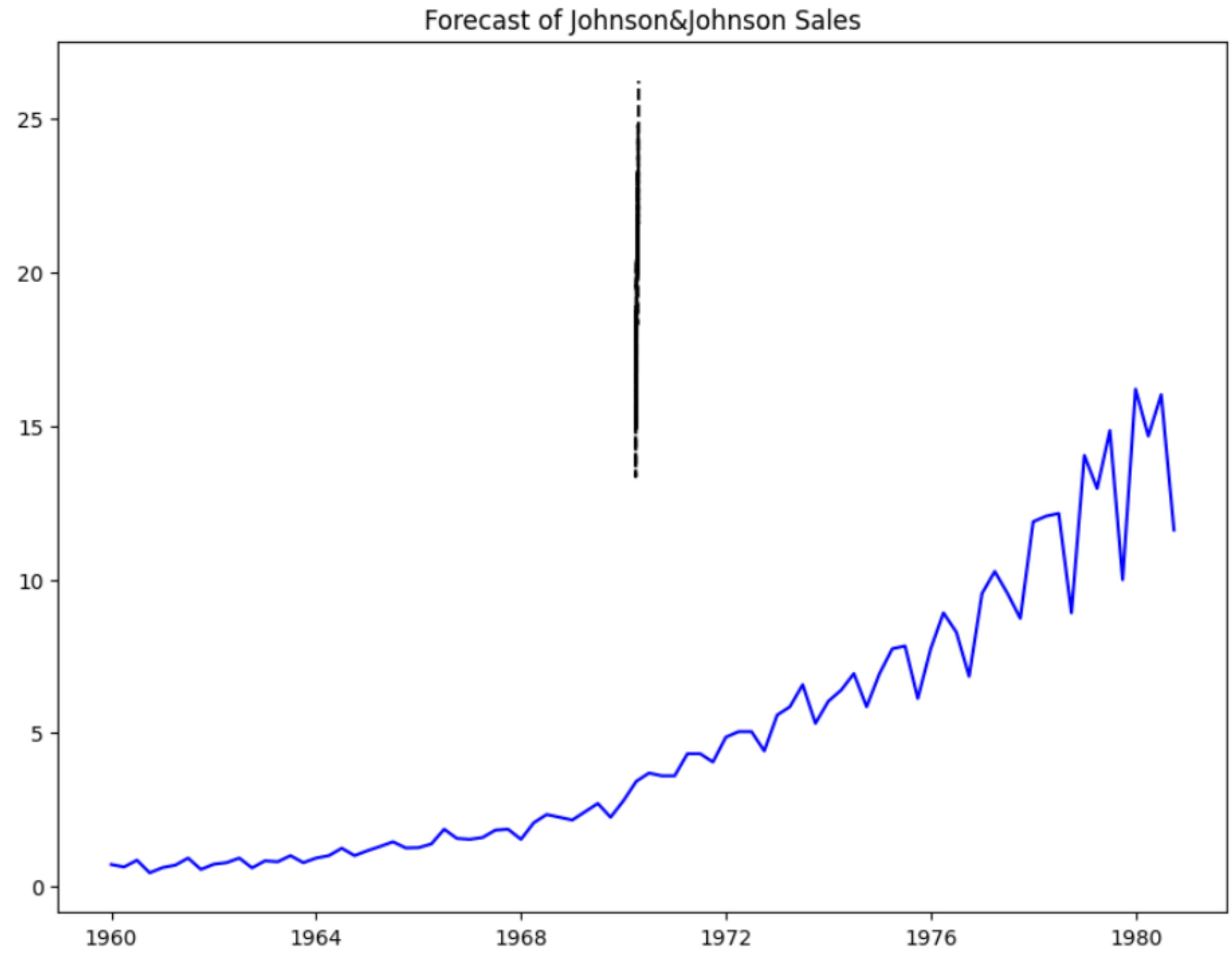


Figure 7 Forecast of JJ Sales

LSTM (Long Short-Term Memory)

LSTM is a type of RNN (Recurrent Neural Network) that learns from the sequences and remembers them for a long period (Staudemeyer and Morris, 2019). In the code, both datasets are used to train the LSTM neural network and predict RMSE. While comparing the result of both datasets, it can be found that the model created by the AMZN.csv dataset learns more quickly and achieves a better fit to the data while overfitting less than the jj.csv dataset, which means the model created using AMZN.csv creates a more successful training process.

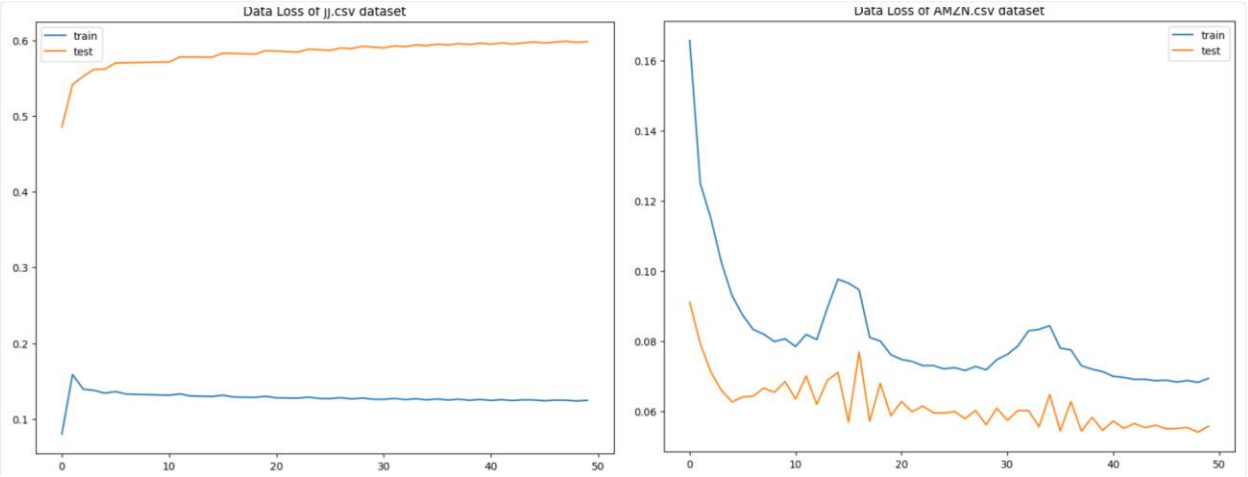


Figure 8 Result of Both Datasets

The analysis and prediction can be further improved by using an improved and updated dataset with consistency data, including data of regular time intervals without errors, using the latest and larger dataset and using a consistent and relevant dataset. The other advanced machine learning and deep learning models can be applied for the advanced analysis and prediction of data.

References

Arumugam, V., & Natarajan, V. (2023). Time series modeling and forecasting using autoregressive integrated moving average and seasonal autoregressive integrated moving average models. Instrumentation Mesure Métrologie, 22(4), 161–168. https://doi.org/10.18280/i2m.220404

Gianfreda, A., Maranzano, P., Parisio, L., & Pelagatti, M. (2023). Testing for integration and cointegration when time series are observed with noise. Economic Modelling, 125, 106352. https://doi.org/10.1016/j.econmod.2023.106352

Rocha, J., Oliveira, S., & Viana, C. (2023). Time Series Analysis - Recent advances, new perspectives and applications [Working title]. In IntechOpen eBooks. https://doi.org/10.5772/intechopen.111223

Sarala, O., Pyhäjärvi, T., & Sillanpää, M. J. (2023). BELMM: Bayesian model selection and random walk smoothing in time-series clustering. Bioinformatics, 39(11). https://doi.org/10.1093/bioinformatics/btad686

Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586

Ye, N. (2013). Autocorrelation and time series analysis. In CRC Press eBooks (pp. 277–285). https://doi.org/10.1201/b15288-24

Zhang, M. (2018). Time series: Autoregressive models ar, ma, arma, arima. University of Pittsburgh.