**Time Series Analysis and Forecasting for Air Passengers Data**

**Introduction:**

This project focuses on the analysis and forecasting of the number of international airline passengers over time. The data spans from January 1949 to December 1960. We aim to transform the time series into a stationary form, apply ARIMA modeling, and make future predictions.

**Data Preparation:**

In this project, the "AirPassengers.csv" dataset was loaded into a pandas DataFrame with the 'Month' column set as the index, allowing for time series analysis. The `parse\_dates=True` parameter ensures the 'Month' column is parsed as date objects for easier manipulation. The '#Passengers' column, which holds the monthly passenger count, was extracted into a separate pandas Series named `ts`. This step is essential for further analysis, forecasting, and identifying trends in the passenger data over time.

**Exploratory Data Analysis:**

**Plotting the Original Time Series:**

A plot of the original time series was created to visualize the trend and seasonality.

A graph showing a line

Description automatically generated with medium confidence

**Checking Stationarity:**

We performed the Dickey-Fuller test to check for stationarity. The null hypothesis of the test states that the time series is non-stationary. The test statistic and p-value were:

Test Statistic: 0.815369

p-value: 0.991880

A graph with a line and a red line

Description automatically generated

Since the test statistic is greater than the critical values at 1%, 5%, and 10%, and the p-value is much higher than 0.05, we fail to reject the null hypothesis, indicating the series is non-stationary.

**Making the Series Stationary:**

**Log Transformation:**

To stabilize the variance, a log transformation was applied to the series.

A graph showing a line

Description automatically generated

**Moving Average:**

A rolling mean was calculated and subtracted from the log-transformed series to remove the trend.

Test Statistic: -3.162908

p-value: 0.022235

A graph showing the value of a stock market

Description automatically generated

The p-value is less than 0.05, so we reject the null hypothesis, indicating that the series is stationary after removing the trend.

**Exponentially Weighted Moving Average:**

An exponentially weighted moving average was used to remove the trend.

Note: that here the parameter ‘halflife’ is used to define the amount of exponential decay. This is just an assumption here and would depend largely on the business domain.

Test Statistic: -3.601262

p-value: 0.005737

A graph showing a red line and blue line

Description automatically generated

The p-value is less than 0.05, indicating stationarity.

**Eliminating Trend and Seasonality:**

The simple trend reduction techniques discussed before don’t work in all cases, particularly the ones with high seasonality. Lets discuss two ways of removing trend and seasonality:

**Differencing:**

Differencing was used to remove both trend and seasonality.

Test Statistic: -2.717131

p-value: 0.071121

A graph showing a number of different levels of data

Description automatically generated with medium confidence

The p-value is slightly higher than 0.05, suggesting marginal non-stationarity.

**Decomposition:**

Seasonal decomposition was performed to separate the series into trend, seasonal, and residual components.

Test Statistic: -6.332387e+00

p-value: 2.885059e-08

A graph showing a line graph

Description automatically generated with medium confidence

The p-value is significantly less than 0.05, confirming the series is stationary.

**ARIMA Modeling:**

The Auto ARIMA function was used to determine the best parameters (p, d, q) for the ARIMA model.

A screenshot of a computer code

Description automatically generated

The best model selected was ARIMA(0, 1, 5).

A close-up of a computer code

Description automatically generated

A graph of a graph

Description automatically generated

**Model Evaluation:**

**A close-up of a computer code

Description automatically generated**

RMSE: 0.09278036898977059

**Conclusion:**

The time series analysis and forecasting for the airline passengers dataset involved several steps to ensure the data was stationary. Various transformations and statistical tests were applied to achieve this. The ARIMA model was then used to forecast future values, with an RMSE indicating a reasonable fit for the model. Further steps could include refining the model and incorporating external variables to improve prediction accuracy.