Time Series Forecasting Project Report

~ By Mayank Bharti

Dataset: Monthly International Airline Passengers (1949–1960)

Models Used: ARIMA and SARIMA

1. Introduction

This project aims to perform time series analysis and forecasting on historical airline passenger data using classical statistical models. The dataset represents the number of international airline passengers per month from 1949 to 1960. The primary goal is to analyze the data for underlying trends and seasonality and apply forecasting models to predict future values.

2. Dataset Overview

A. Background :

The **Airline Passengers dataset** is one of the most widely used and classic examples in time series analysis. It consists of **monthly totals of international airline passengers** (in thousands) from **January 1949 to December 1960**, spanning a total of **144 observations**. This dataset was originally published by the U.S. Bureau of Transportation Statistics and was later popularized in several statistical and machine learning textbooks, including works by Rob J. Hyndman and George E. P. Box.

B. Source: AirPassengers dataset (Monthly total of international airline Passengers) AirPassengers.csv

Sample of dataset

Month	Passengers
1949-01	112
1949-02	118
1949-03	132
1949-04	129
1949-05	121
1949-06	135
1949-07	148
1949-08	148
1949-09	136

• Time Period: January 1949 – December 1960
• Frequency: Monthly
• Target Variable: Passengers (Number of passengers per month)

Analysis through excel

- **Dataset overview values**

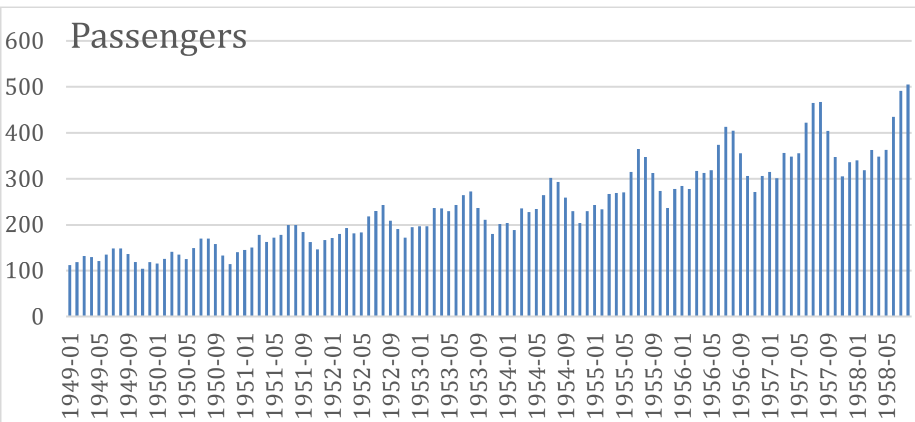
[33]:
```python
# Data Overview

print(df.describe())
print(df.isnull().sum())
```

- Result :
  count = 144.000000
  mean of Data = 280.298611
  standard Deviation = 119.966317
  minimum value in data = 104.000000
  Maximum value in data = 622.000000

Methodology

1. Data Loading and Preparation

- The dataset was imported from a CSV file containing monthly totals of international airline passengers.
- Dates were converted to datetime format and set as the index to facilitate time series analysis.
- Basic exploratory analysis through excel was conducted to understand the structure, shape, and missing values in the data.



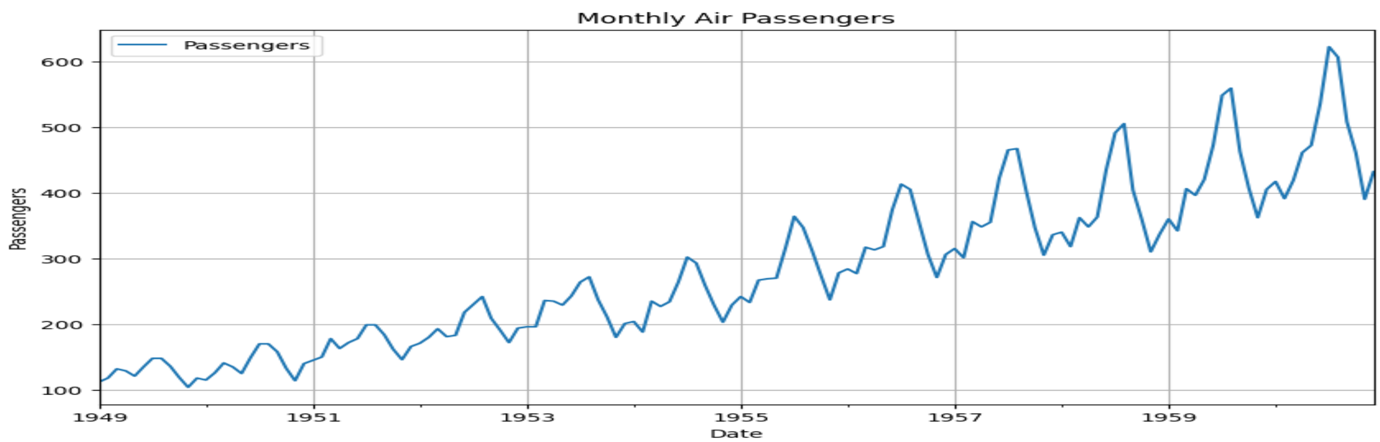
Conclusion : There is no missing or duplicate values in the data. Also by excel we can observe that the data has an upward trend. I used line plot also to get detailed analysis of data

2. **2. Data Visualization**
   - Line plots were used to visualize the passenger counts over time.
   - This revealed a clear **upward trend** and **annual seasonality**, indicating the need for seasonal modelling

```python
#loading file

df = pd.read_csv(r'C:\Users\richh\Downloads\airline-passengers.csv')
df['Month'] = pd.to_datetime(df['Month'])
df.set_index('Month', inplace=True)
```

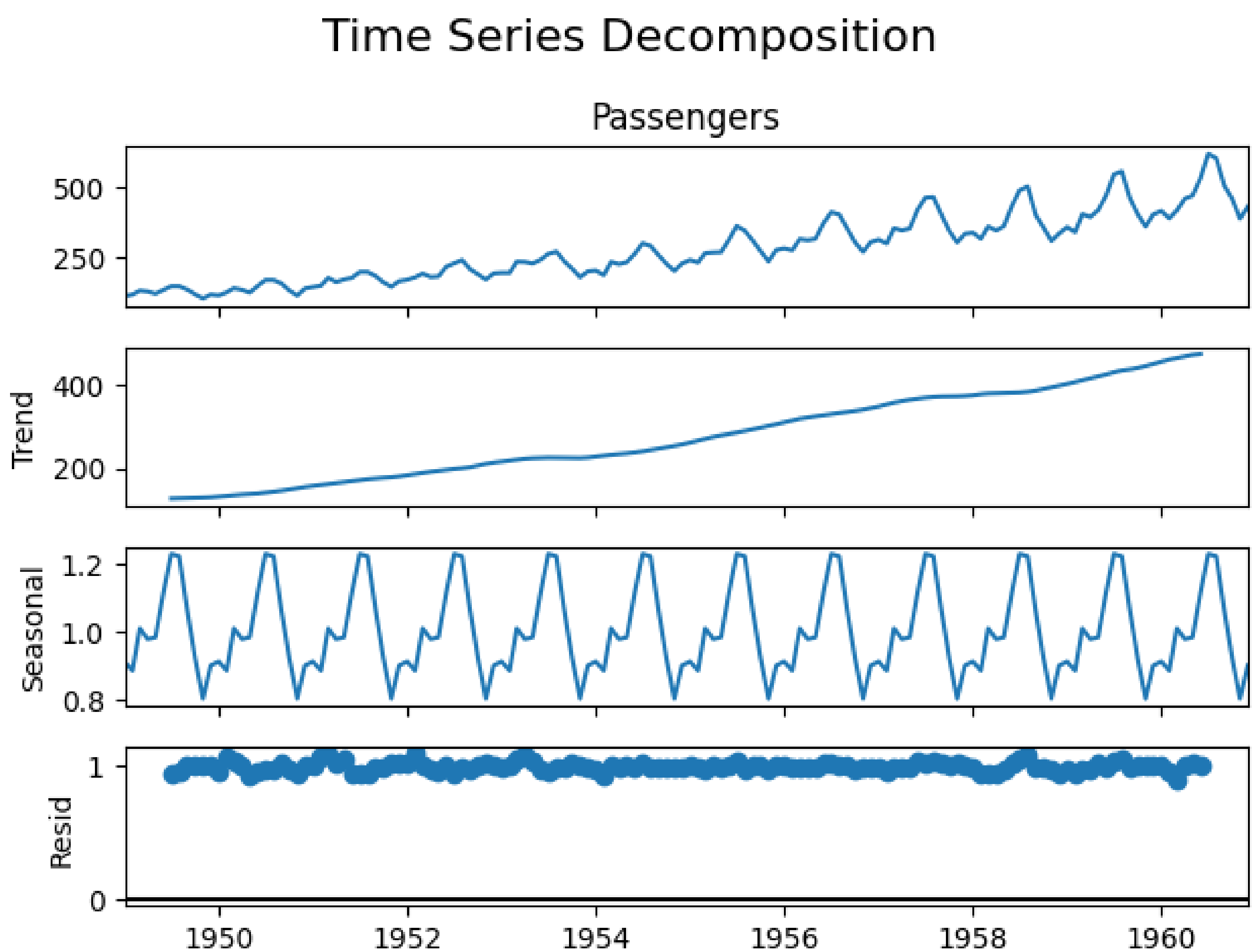
```python
# Exploratory Data Analysis

df.plot(figsize=(10, 6), title="Monthly Air Passengers")
plt.xlabel("Date")
plt.ylabel("Passengers")
plt.grid()
plt.show()
```

Result :



This shows that data has a clear upward trend.

4. **Time Series Decomposition**
   - Decomposition using a **multiplicative model** with a 12-month seasonal cycle was performed.
   - This separated the series into **trend**, **seasonal**, and **residual** components.
   - This step confirmed the presence of consistent yearly seasonal patterns.



5. **Stationarity Testing**
    - The **Augmented Dickey-Fuller (ADF) test** was applied to check for stationarity.
    - Since the ADF Statistic value is smaller than p-value then data is Stationary according to test of hypothesis.

[34]:
```
#Calculating ADF statistic

result = adfuller(df['Passengers'])
print(f'ADF Statistic: {result[0]}')
print(f'p-value: {result[1]}')
```

ADF Statistic: 0.8153688792060482

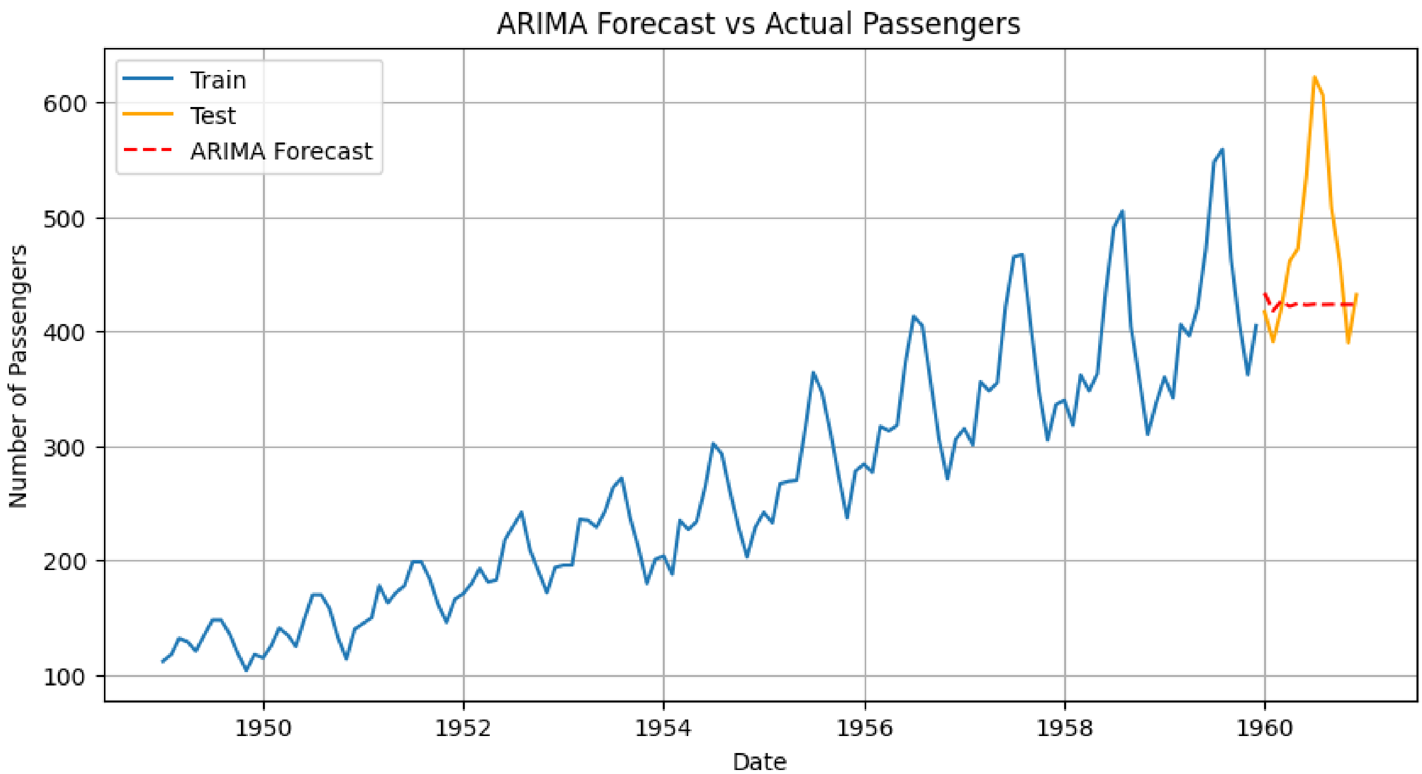
p-value: 0.991880243437641

6. **Fitting ARIMA Model**

```
#Fitting ARIMA MODEL

model = ARIMA(train['Passengers'], order=(1,1,1))
model_fit = model.fit()
forecast = model_fit.forecast(steps=12)
print(forecast)
```

```
1960-01-01    433.451927
1960-02-01    417.984221
1960-03-01    426.393139
1960-04-01    421.821685
1960-05-01    424.306927
1960-06-01    422.955841
1960-07-01    423.690350
1960-08-01    423.291039
1960-09-01    423.508122
1960-10-01    423.390106
1960-11-01    423.454265
1960-12-01    423.419385
Freq: MS, Name: predicted_mean, dtype: float64
```

Plotting forecast in ARIMA model



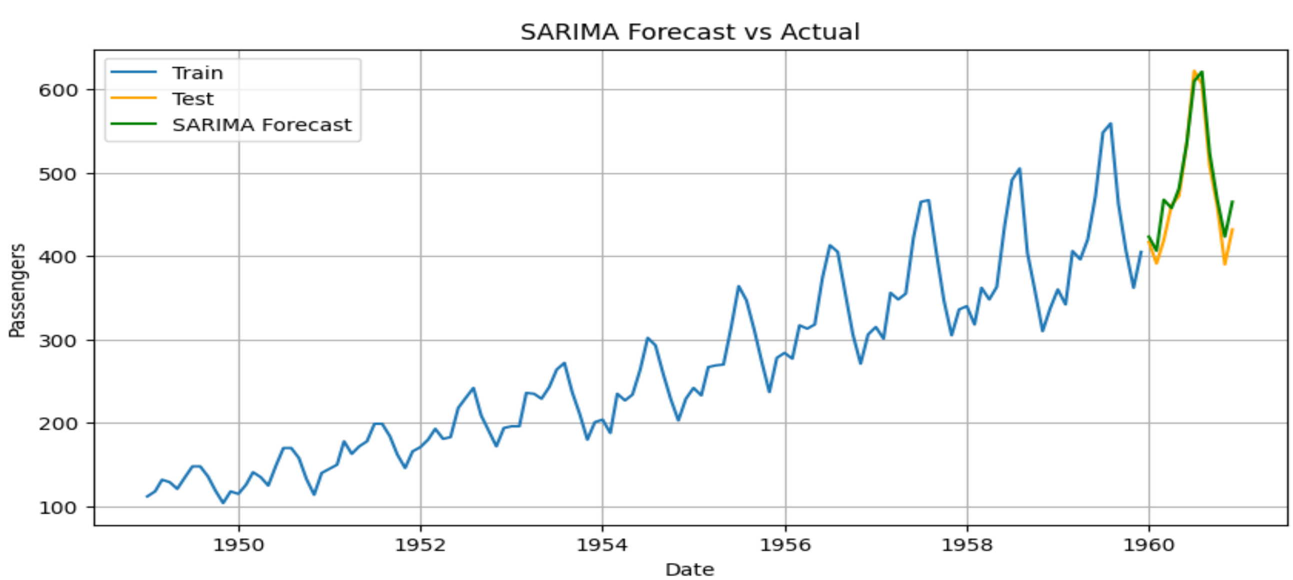
Conclusion: Now according to this graph, we can see that there is a difference between test values and forecast value which tell us that ARIMA forecasting is not suitable for this. Therefore, we will now try testing SARIMA model for this time series data.

Value of MSE and RMSE for ARIMA

- ***MSE: 8322.70,***
- ***RMSE: 91.23***

7. Fitting SARIMA Model

- An extension of ARIMA that includes **seasonal differencing** and seasonal terms.
- Model parameters were set to account for the 12-month seasonal cycle.



Conclusion: Now according to this graph, we can see that there is a LESS difference between test values and forecast value which tell us that SARIMA forecasting is best for this.

- ***Value of MSE and RMSE for SARIMA***
  MSE: 467.48,
  RMSE: 21.62

FORECASTING AND EVALUATION

MODEL	MSE	RMSE
ARIMA	8332.7	91.23
SARIMA	467.48	21.62

Both models were used to forecast the final 12 months of the dataset. Their performance was evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

8. Conclusion

- The time series data exhibits strong seasonality and a long-term trend.
- Here SARIMA model is better than ARIMA model on the basis of error metric
- Model performance was assessed based on forecast accuracy using error metrics.