EXP 5

Aim : Implementation of ARIMA model in python

Theory:

The ARIMA (AutoRegressive Integrated Moving Average) model is a popular statistical model used for time series forecasting. It consists of three components:

1. **AR (AutoRegressive)**: The relationship between an observation and a number of lagged observations.
2. **I (Integrated)**: The differencing of raw observations to make the time series stationary.
3. **MA (Moving Average)**: The relationship between an observation and a residual error from a moving average model applied to lagged observations.

1. AutoRegressive (AR) Component:

The AR part of the ARIMA model represents the relationship between the current value of the time series and its past values (lags). The AR model uses the dependency between an observation and several lagged observations to predict future values.

Mathematically, the AR model of order **p** can be written as:

$X_t=\phi_1 X_{t-1}+\phi_2 X_{t-2}+\ldots+\phi_p X_{t-p}+\epsilon_t$

Where:

- X_t is the value of the time series at time t.
- $\phi_1,\phi_2,\ldots,\phi_p$ are the coefficients (parameters) of the lagged values.
- ϵ_t is the white noise (random error term).

2. Integrated (I) Component:

The Integrated component deals with making a time series stationary. A stationary time series has a constant mean, variance, and autocovariance over time. Many real-world time series data are non-stationary, meaning their statistical properties change over time (e.g., trends or seasonality).

The differencing operation is used to make the time series stationary. The first differencing is computed as:

$X_t'=X_t-X_{t-1}$

If the series is still non-stationary after first differencing, further differencing might be applied (second differencing, etc.).

3. Moving Average (MA) Component:

The MA part of the ARIMA model models the relationship between the current value of the series and the residual errors (or shocks) from past observations. In other words, the forecast depends not only on past observations but also on the errors made by previous predictions.

Mathematically, the MA model of order **q** can be written as:

$X_t=\mu+\epsilon_t+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}+\ldots+\theta_q\epsilon_{t-q}$

Where:

- μ is the mean of the series.

- ϵt is the white noise (random error term) at time t.
- $\theta 1, \theta 2, \ldots, \theta q$ are the coefficients for the lagged error terms.

Steps to Implement ARIMA Model in Python:

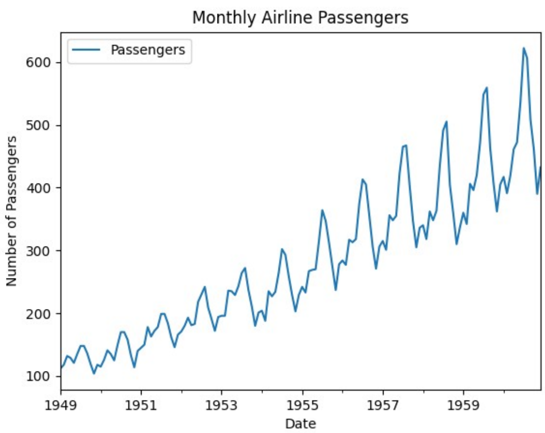
1. Load the dataset (time series data).
2. Preprocess and visualize the data.
3. Check for stationarity.
4. Make the series stationary (if needed).
5. Fit the ARIMA model.
6. Forecast future values.
7. Evaluate model performance.

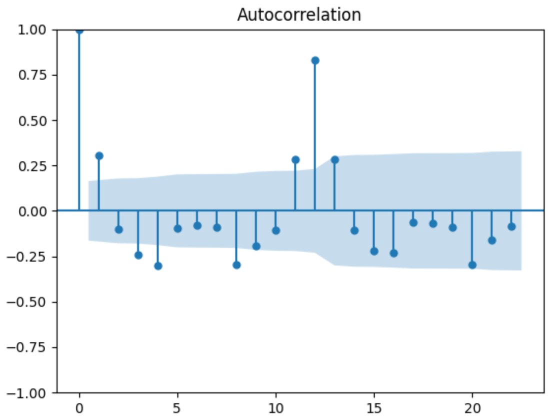
Code:

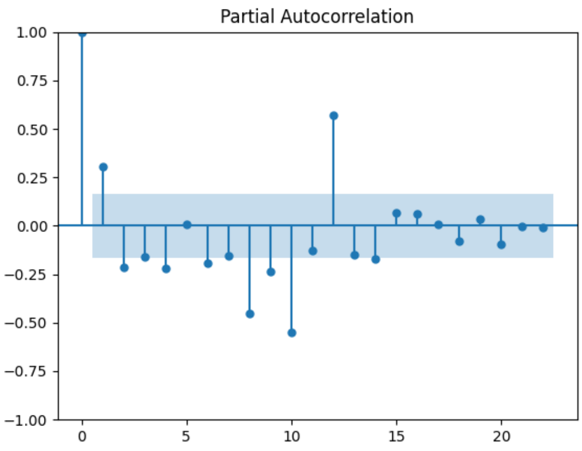
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/airline-
passengers.csv"
data = pd.read_csv(url, header=0, parse_dates=[0], index_col=0,
date_parser=pd.to_datetime)

data.plot()
plt.title("Monthly Airline Passengers")
plt.xlabel("Date")
plt.ylabel("Number of Passengers")
plt.show()

def adf_test(series):
result = adfuller(series, autolag='AIC')
print(f"ADF Statistic: {result[0]}")
print(f"p-value: {result[1]}")
if result[1] <= 0.05:
print("Series is stationary")
else:
print("Series is not stationary")

adf_test(data)

data_diff = data.diff().dropna()

plot_acf(data_diff)
plot_pacf(data_diff)
plt.show()
```

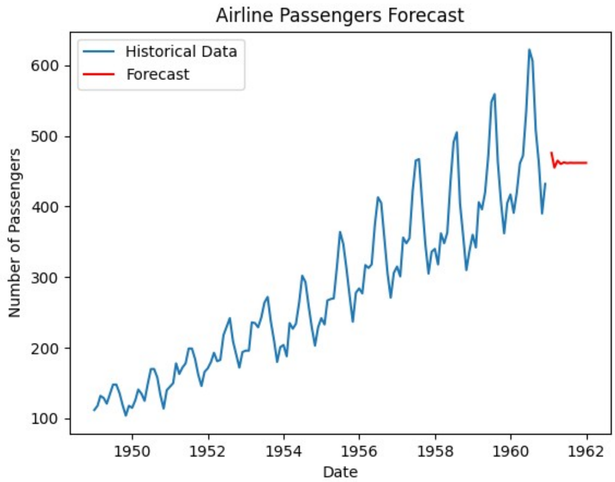
```
model = ARIMA(data, order=(1, 1, 1))
model_fit = model.fit()

print(model_fit.summary())

forecast = model_fit.forecast(steps=12)

plt.plot(data, label='Historical Data')
plt.plot(pd.date_range(start=data.index[-1], periods=13, freq='M')[1:], forecast,
label='Forecast', color='red')
plt.title("Airline Passengers Forecast")
plt.xlabel("Date")
plt.ylabel("Number of Passengers")
plt.legend()
plt.show()
```

Output:









```
                               SARIMAX Results
==============================================================================
Dep. Variable:             Passengers   No. Observations:                  144
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -694.341
Date:                Tue, 04 Mar 2025   AIC                           1394.683
Time:                        10:22:17   BIC                           1403.571
Sample:                    01-01-1949   HQIC                          1398.294
                         - 12-01-1960
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4742      0.123     -3.847      0.000      -0.716      -0.233
ma.L1          0.8635      0.078     11.051      0.000       0.710       1.017
sigma2       961.9270    107.433      8.954      0.000     751.362    1172.492
===================================================================================
Ljung-Box (L1) (Q):                   0.21   Jarque-Bera (JB):                 2.14
Prob(Q):                              0.65   Prob(JB):                         0.34
Heteroskedasticity (H):               7.00   Skew:                            -0.21
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.43
===================================================================================
```

Conclusion:

The ARIMA model effectively captures patterns in time series data, such as trends and autocorrelations, for forecasting future values. By selecting the appropriate order of parameters (p, d, q), the model can provide accurate predictions. Evaluating model performance and ensuring stationarity are crucial for achieving reliable forecasts.