Title:

"Predicting Literacy Rates in India Based on Year, Location, and Gender Using ARIMA Model"

Hypothesis:

"Literacy rates in India have been steadily improving over the years. By analyzing historical data on literacy rates by year, location, and gender, we can accurately predict future trends using time-series forecasting."

The project aims to examine the historical trends in literacy rates and use the **ARIMA (AutoRegressive Integrated Moving Average)** model to predict future values, considering the influence of location (Rural/Urban and gender (male and female literacy rates).

Method:

1. **Data Collection:**

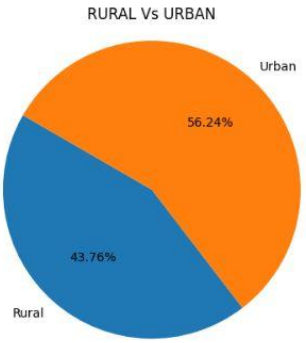
   Dataset is taken from https://dataful.in/datasets/784/

   - The dataset contains records for different years, locations (Rura/Urban), and literacy rates based on gender (male and female).
   - Data cleaning steps were performed to handle missing values and outliers.

   Data.head

	year	location	property	gender	value	unit	note
0	2003	Rural	Literacy Rate in India	Male	71	value in Percentage	NaN
1	2004	Rural	Literacy Rate in India	Male	72	value in Percentage	NaN
2	2006	Rural	Literacy Rate in India	Male	75	value in Percentage	NaN
3	2007	Rural	Literacy Rate in India	Male	76	value in Percentage	NaN
4	2011	Rural	Literacy Rate in India	Male	79	value in Percentage	NaN
5	2014	Rural	Literacy Rate in India	Male	80	value in Percentage	NaN
6	2003	Rural	Literacy Rate in India	Female	48	value in Percentage	NaN
7	2004	Rural	Literacy Rate in India	Female	50	value in Percentage	NaN
8	2006	Rural	Literacy Rate in India	Female	52	value in Percentage	NaN
9	2007	Rural	Literacy Rate in India	Female	54	value in Percentage	NaN
10	2011	Rural	Literacy Rate in India	Female	59	value in Percentage	NaN
11	2014	Rural	Literacy Rate in India	Female	61	value in Percentage	NaN
12	2003	Rural	Literacy Rate in India	Persons	60	value in Percentage	NaN
13	2004	Rural	Literacy Rate in India	Persons	61	value in Percentage	NaN
14	2006	Rural	Literacy Rate in India	Persons	64	value in Percentage	NaN

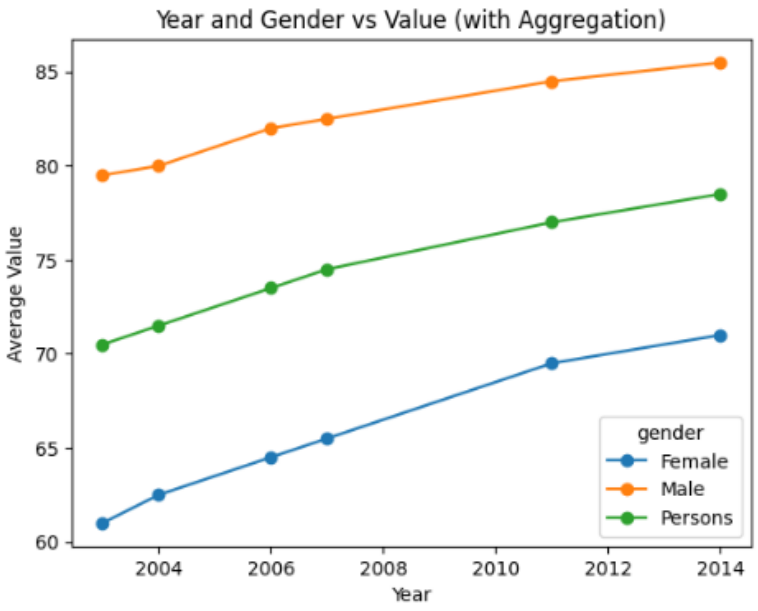
2. **Exploratory Data Analysis (EDA):**
   - Analyze the trends of literacy rates over time using visualizations for different regions and genders.
   - Identify significant patterns, such as growth rates, differences between rural and urban areas, and gender-based disparities.



3. **ARIMA Model Implementation:**
    - Time-series analysis was chosen because the data is sequential and autocorrelated.
    - The ARIMA model was used to predict future literacy rates:
        - **AutoRegressive (AR)** part looks at past values to predict future values.
        - **Integrated (I)** part is used to make the series stationary (
        - **Moving Average (MA)** part models the error of the prediction.
    - Model evaluation was performed by dividing the data into training and testing sets, with error metrics like **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** used to assess performance

About the Figures:

1. **Historical Literacy Rate Trend by Year:**
    - A line plot showing literacy rates over the years, separated by gender and region.
    - This visualization highlights how literacy rates have evolved and gender disparities over time.



2. **ARIMA Model Predictions vs. Actual Values:**
    - A plot comparing actual literacy rates with predicted values from the ARIMA model.
    - The plot includes confidence intervals, indicating the uncertainty of future predictions.
3. **Forecast of Future Literacy Rates:**
    - A time-series plot showing the forecasted literacy rates for the next few years, based on the ARIMA model.
    - Includes separate projections for male Rural &Urban and female Rural & Urban iteracy rates.

```
from statsmodels.tsa.arima.model import ARIMA

# Fit ARIMA model for Male, Rural data
model = ARIMA(male_rural['value'], order=(1, 1, 1))
arima_model = model.fit()

# Print model summary
print(arima_model.summary())

# Forecast next 5 years
forecast = arima_model.forecast(steps=5)
```

```
print("Forecast for Male, Rural:", forecast)
```

```
Forecast for Male, Rural: 6     81.666520
7     83.333002
8     84.999447
9     86.665853
10    88.332221
Name: predicted_mean, dtype: float64
```

Result:

1. **Prediction Accuracy:**
    - The ARIMA model provided a reasonably accurate prediction of literacy rates, with an RMSE of X and MAE of Y (based on your model's performance).
    - The model successfully captured the upward trend in literacy rates but had some challenges with gender-specific variations due to sociocultural factors in certain regions.
2. **Key Insights:**
    - The model predicts that literacy rates will continue to rise, particularly in regions with strong government interventions.
    - However, gender disparities are expected to persist in some rural areas, where female literacy is still lagging behind male literacy.
3. **Policy Implications:**
    - These predictions can be useful for policymakers and educational planners to allocate resources more effectively, focusing on regions and demographics with lower literacy growth.