DATA SCIENCE AND AI MENTORSHIP PROGRAM

Week One Learning Task

Practice task to reinforce the concepts learned in the "Data Science with Python" course:

1. Exploratory Data Analysis (EDA):
 - Using any of the following datasets (e.g., Iris dataset, Titanic dataset, or a dataset related to your area of interest), perform exploratory data analysis, including:
    - Loading and inspecting the data
    - Handling missing values
    - Summarizing numerical and categorical variables
    - Visualizing the distributions and relationships between variables

Dataset
- Iris dataset (from UCI Machine Learning Repository)
- Titanic dataset (from Kaggle)
- NYC OpenData datasets (e.g., NYC Airbnb listings, NYC Motor Vehicle Crashes)

2. Data Cleaning and Preprocessing:
 - You have a messy dataset (e.g., with inconsistent data formats, missing values, duplicates).
- Write Python scripts to clean and preprocess the data, including:
    - Handling missing values (e.g., imputation, dropping rows/columns)
    - Removing duplicates
    - Converting data types
    - Handling categorical variables (e.g., one-hot encoding, label encoding)
    - Scaling numerical features

Dataset
 - Any dataset that is messy or contains inconsistencies, missing values, or formatting issues would be suitable for this assignment. You could intentionally introduce some issues in a clean dataset for practice purposes.
 - Potential datasets: NYC OpenData datasets (e.g., NYC Airbnb listings, NYC Motor Vehicle Crashes), Kaggle datasets (e.g., Used Cars Dataset, Adult Census Income Dataset).

3. Data Visualization:
 - You have a dataset with multiple variables (numerical and categorical).
 - Create various visualizations using Matplotlib or other libraries (e.g., scatter plots, histograms, bar charts, box plots, heatmaps).
 - Experiment with different plot types and customizations.

Dataset
 - Iris dataset (from UCI Machine Learning Repository)
 - Titanic dataset (from Kaggle)

- FiveThirtyEight datasets (e.g., Comic Book Characters, Airline Safety)
- Gapminder datasets (e.g., life expectancy, income, population data)

4. Data Analysis and Modeling:
- Perform data analysis and build a predictive model using techniques like linear regression, logistic regression, or decision trees
- Evaluate the model's performance using appropriate metrics

Dataset
- Boston Housing dataset (from UCI Machine Learning Repository)
- Loan Default dataset (from Kaggle)
- Bank Customer Churn dataset (from Kaggle)
- Student Performance dataset (from UCI Machine Learning Repository)

Some Datasets description:
1. UCI Machine Learning Repository: This repository (https://archive.ics.uci.edu/ml/datasets.php) provides a wide range of datasets for various machine learning tasks, including classification, regression, and clustering.

2. Kaggle Datasets: Kaggle (https://www.kaggle.com/datasets) is a popular platform for data science competitions and offers a vast collection of datasets across various domains, such as finance, healthcare, sports, and more.

3. NYC OpenData: The NYC OpenData portal (https://opendata.cityofnewyork.us/) provides open datasets related to New York City, covering topics like transportation, housing, education, and more.

4. FiveThirtyEight Data Repository: This repository (https://data.fivethirtyeight.com/) contains datasets used in data journalism articles published by FiveThirtyEight, covering topics like politics, sports, and economics.

5. Gapminder Data: The Gapminder Foundation (https://www.gapminder.org/data/) offers datasets related to global development, including indicators like life expectancy, income, and population.

6. NOAA Climate Data: The National Oceanic and Atmospheric Administration (NOAA) provides various climate and weather-related datasets (https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00861).