Assignment #	4

By

NAME : Ishmal shahid

ROLL NO #: 1220100981

BS SE (4th)

Fall 2022

COURSE TITLE: ARTIFICAL INTELLIGENCE

SUBMITTED TO : SIR ZUBAR

Department of COMPUTER SCIENCE
International Institute of Science, Arts and Technology (IISAT), Gujranwala

Data Set Selection and Loading

Assume we have a dataset named `house.csv` containing information about house prices based on their area.

```
import pandas as pd

import matplotlib.pyplot as plt

# Read the data

df = pd.read_csv('house.csv')
```

2. Data Exploration

Explore the dataset to understand its structure, features, and statistical summary

```
# Display first few rows of the DataFrame

print("First few rows:")

print(df.head())

# Get information about the DataFrame

print("\nDataFrame info:")

print(df.info())

# Statistical summary of numerical columns

print("\nStatistical summary:")

print(df.describe())

# Check for any missing values

print("\nMissing values:")

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

3. Data Cleaning

Clean the data by handling missing values, duplicates, and performing necessary transformations (if any).

```
# Handle missing values (if any)

df.dropna(inplace=True)

# Handle duplicates (if any)

df.drop_duplicates(inplace=True)

# Confirm changes

print("\nAfter cleaning:")

print(df.info())
```

4. Data Visualization

Use Pandas, Matplotlib, and Seaborn to create various graphs and charts

```
# Visualize the data


plt.figure(figsize=(10, 6))

plt.scatter(df.area, df.price, color='red', marker='+')

plt.xlabel('Area (sqft)')

plt.ylabel('Price ($)')

plt.title('House Prices vs. Area')

plt.grid(True)

plt.show()
```

5. Analysis and Insights

After each visualization, provide analysis and insights derived from it.

Analysis:

- The scatter plot shows a positive correlation between house prices and area, indicating that larger houses tend to have higher prices.

- There are a few outliers where houses with smaller areas have unexpectedly high prices, suggesting other influential factors.

Explanation

- **Data Set Selection and Loading**: We assume `house.csv` contains columns like `area` and `price`.
- **Data Exploration**: Use `head()` to view the first few rows, `info()` for structure, `describe()` for statistics, and `isnull().sum()` for missing values.
- **Data Cleaning**: Drop rows with missing values (`dropna()`) and duplicates (`drop_duplicates()`).
- **Data Visualization**: Plot a scatter plot to visualize the relationship between house prices and area.
- **Analysis and Insights**: Interpret the plot to derive insights about the dataset, such as correlations and potential outliers.

Google Colab link

https://colab.research.google.com/drive/1odVKTWvvhL6uODuls_b-lFw2yK1HGPuI?usp=sharing