Lab-1

Write python code for all considering "housing.csv"

① To load .csv file into dataframe

```
import pandas as pd
filename = "/content/housing.csv"
df = pd.read_csv(filename)
```

② To display info of all columns

```
print("Dataset Info:")
print(df.info())
```

Dataset Info:

#	Column	Non-Null Count	Dtype
0	longitude	20640	float64
1	latitude	20640	— " —
2	housing_median_age	20640	— " —
3	total_rooms	20640	— " —
4	total_bedrooms	20640	— " —
5	population	— " —	— " —
6	households	— " —	— " —
7	median_income	— " —	— " —
8	median_house_value	— " —	— " —
9	ocean_proximity	— " —	object

③ To display statistical info of all numerical columns

```
print("\n Statistical summary of Numerical columns :")
print(df.describe())
```

④ To display count of unique labels for "ocean proximity" columns

```
print("\n Unique Value counts for 'Ocean Proximity':")
print(df["ocean_proximity"].value_counts())
```

⑤ To display which columns in a dataset have missing values count greater than zero

```
missing_values = df.isnull().sum()
missing_columns = missing_values[missing_values > 0]
print("\nColumns with missing values:")
print(missing_columns)
```

o/p :-

Unique Value Counts for 'Ocean Proximity':

ocean_proximity	
<1H OCEAN	9136
INLAND	6551
NEAR OCEAN	2658
NEAR BAY	2290
ISLAND	5

Columns with Missing Values:

total_bedrooms : 207

Questions:-

1) Which columns in the datasets had missing values? How did you handle them?

Ans: Missing value columns:

Adult income dataset → Age, Salary

Diabetes dataset → Glucose, BMI

Handling approach:

Adult income dataset
- for age → used median since its less sensitive to outliers
- for salary → used mean as salaries typically follow normal distribution

Diabetes dataset
- Glucose → used median since glucose levels may have outliers
- BMI → used mean assuming normal distribution.

2) Which categorical columns did you identify in the dataset? How did you encode them?

Ans:- Adult Income dataset:

Categorical columns: Gender → Original encoding
City → One-Hot encoding

Diabetes dataset:

categorical columns:- Gender → Original encoding
Outcome → one-hot encoding

3) What is the difference b/n Min-Max scaling and standardization? When would you use one over another?

Ans: Min Max Scaling:

→ $X' = \frac{x - x_{min}}{x_{max} - x_{min}}$

→ scales values b/n 0 & 1

→ sensitive to outliers

Standardization:

→ $x' = \frac{x - \mu}{\sigma}$

→ Transforms data to have mean = 0 and variance = 1

→ less effected by outliers

When data is not normally distributed and has known bounds, min-max scaling is used.

When data follows a normal distribution, standardization is used

Sneha S B
10/3/25