LAB-1 :

write python code, consider filename as "housing.csv"

i) To load .csv file into the dataframe

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

ii) To display information of all columns

```
print(df.info())
```

iii) To display statistical information of all numerical

```
print(df.describe())
```

iv) To display the count of unique labels for "Ocean Proximity" column

```
if "Ocean Proximity" in df.columns:
    print(df["Ocean Proximity"].value_counts())
else:
    print("Not found")
```

v) To display which attributes in dataset have missing values count greater than zero

```
m_values = df.isnull().sum()
m_columns = m_values[m_values > 0]

if not m_columns.empty:
    print(m_columns)
else:
    print("No missing values found)
```

Output:

ii)

0	longitude	3000	non-null	float64
1	latitude	3000	non-null	float64
2	housing_median_age	3000	non-null	float64
3	total_rooms	3000	non-null	float64
4	total_bedrooms	3000	non-null	float64
5	population	3000	non-null	float64
6	households	3000	non-null	float64
7	median_income	3000	non-null	float64
8	median_house_value	3000	non-null	float64

iii)

	longitude	latitude
count	3000	3000.0
mean	-119.5	35.6
std	1.99	2.12
min	-124.18	32.5
25%	-121.81	33.9
50%	-118.48	34.2
75%	-118.02	37.6
max	-114.49	41.9

iv) 'ocean Proximity' column not found in the dataset

v) No missing values found in the dataset

Questions:

Q1) Which columns in the dataset 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.

Q2) 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.

Q3) What is the difference between Min-Max scaling and standardization? When would you use one over the other?

Soln:

Min-Max scaling:

$$\rightarrow X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

→ scales values between 0 & 1.

→ sensitive to outliers

Standardization

$$\rightarrow X' = \frac{X - \mu}{\sigma}$$

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

→ less affected 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.