1. Consider a dataset such as finance, education, marketing, healthcare etc. Load, inspect, and store data in various formats like CSV, Excel, JSON, and SQL using Pandas. Clean and prepare data by handling missing values, duplicates, normalization, and encoding categorical data.

Assignment Steps

1. **Dataset Selection:**

   Choose a dataset relevant to a domain of interest (e.g., finance, education, marketing, or healthcare). Public datasets are available on platforms like Kaggle or UCI Machine Learning Repository.

2. **Tasks:**

   a. Load and Inspect the Data:
   - Load the dataset into a Pandas DataFrame from a CSV file.
   - Display basic statistics using .info() and .describe().
   - Print the first five rows using .head().

   b. Store the Data in Various Formats:
   - Save the dataset in **Excel**, **JSON**, and **SQL** formats using Pandas.
   - Verify each saved format by reloading it back into a DataFrame.

   c. Clean and Prepare the Data:
   - Handle missing values by filling or dropping them using .fillna() or .dropna().
   - Remove duplicate rows using .drop_duplicates().
   - Normalize numerical columns by scaling them between 0 and 1.
   - Encode categorical columns using one-hot encoding or label encoding.

3. **Code Template:**

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler

# Step 1: Load Dataset
file_path = "Heart.csv"  # Replace with your dataset path
df = pd.read_csv(file_path)

# Step 2: Inspect Data
print("Data Info:")
print(df.info())
print("\nData Statistics:")
print(df.describe())
print("\nFirst 5 Rows:")
print(df.head())

# Step 3: Save Data in Various Formats
df.to_excel("dataset.xlsx", index=False)
df.to_json("dataset.json", orient="records", indent=2)
from sqlalchemy import create_engine
engine = create_engine("sqlite:///:memory:")  # In-memory SQL database
df.to_sql("dataset", con=engine, index=False, if_exists="replace")

# Step 4: Reload Data to Verify
```

```
df_excel = pd.read_excel("dataset.xlsx")
df_json = pd.read_json("dataset.json")
df_sql = pd.read_sql("dataset", con=engine)

# Step 5: Clean and Prepare Data
# Handle Missing Values
df.fillna(method="ffill", inplace=True)  # Forward fill missing values

# Remove Duplicates
df.drop_duplicates(inplace=True)

# Normalize Numerical Data
scaler = MinMaxScaler()
numerical_cols = df.select_dtypes(include=["number"]).columns
df[numerical_cols] = scaler.fit_transform(df[numerical_cols])

# Encode Categorical Data
df = pd.get_dummies(df, drop_first=True)  # One-hot encoding

# Final Dataset
print("\nCleaned Data:")
print(df.head())

# Save the cleaned dataset
df.to_csv("cleaned_dataset.csv", index=False)
```

4. **Deliverables:**

- Python code file (.py or .ipynb) implementing the above steps.
- Saved data files in **CSV**, **Excel**, **JSON**, and **SQL** formats.
- A brief report (in .docx or .pdf) summarizing the observations and challenges faced during the assignment.

Evaluation Criteria:

1. Proper loading and inspection of the dataset.
2. Successful storage and reloading of the dataset in all specified formats.
3. Effective handling of missing values and duplicates.
4. Proper normalization and encoding of data.
5. Clarity and efficiency of the Python code.
6. Quality of the report summarizing observations and challenges.

OUTPUT SAMPLE :

```
Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 15 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  303 non-null    int64
 1   Age         303 non-null    int64
 2   Sex         303 non-null    int64
 3   ChestPain   303 non-null    object
 4   RestBP      303 non-null    int64
 5   Chol        303 non-null    int64
 6   Fbs         303 non-null    int64
 7   RestECG     303 non-null    int64
 8   MaxHR       303 non-null    int64
 9   ExAng       303 non-null    int64
 10  Oldpeak     303 non-null    float64
 11  Slope       303 non-null    int64
 12  Ca          299 non-null    float64
 13  Thal        301 non-null    object
 14  AHD         303 non-null    object
dtypes: float64(2), int64(10), object(3)
memory usage: 35.6+ KB
None

Data Statistics:
       Unnamed: 0         Age         Sex      RestBP        Chol         Fbs  \
count  303.000000  303.000000  303.000000  303.000000  303.000000  303.000000
mean   152.000000   54.438944    0.679868  131.689769  246.693069    0.148515
std     87.612784    9.038662    0.467299   17.599748   51.776918    0.356198
min      1.000000   29.000000    0.000000   94.000000  126.000000    0.000000
25%     76.500000   48.000000    0.000000  120.000000  211.000000    0.000000
50%    152.000000   56.000000    1.000000  130.000000  241.000000    0.000000
75%    227.500000   61.000000    1.000000  140.000000  275.000000    0.000000
max    303.000000   77.000000    1.000000  200.000000  564.000000    1.000000

          RestECG       MaxHR       ExAng     Oldpeak       Slope          Ca
count  303.000000  303.000000  303.000000  303.000000  303.000000  299.000000
mean     0.990099  149.607261    0.326733    1.039604    1.600660    0.672241
std      0.994971   22.875003    0.469794    1.161075    0.616226    0.937438
min      0.000000   71.000000    0.000000    0.000000    1.000000    0.000000
25%      0.000000  133.500000    0.000000    0.000000    1.000000    0.000000
50%      1.000000  153.000000    0.000000    0.800000    2.000000    0.000000
75%      2.000000  166.000000    1.000000    1.600000    2.000000    1.000000
max      2.000000  202.000000    1.000000    6.200000    3.000000    3.000000

First 5 Rows:
   Unnamed: 0  Age  Sex     ChestPain  RestBP  Chol  Fbs  RestECG  MaxHR  \
0           1   63    1       typical     145   233    1        2    150
1           2   67    1  asymptomatic     160   286    0        2    108
```

```
2         3  67   1  asymptomatic    120  229   0       2   129
3         4  37   1    nonanginal    130  250   0       0   187
4         5  41   0    nontypical    130  204   0       2   172

   ExAng  Oldpeak  Slope   Ca        Thal  AHD
0      0      2.3      3  0.0       fixed   No
1      1      1.5      2  3.0      normal  Yes
2      1      2.6      2  2.0  reversable  Yes
3      0      3.5      3  0.0      normal   No
4      0      1.4      1  0.0      normal   No

Cleaned Data:
   Unnamed: 0       Age  Sex    RestBP      Chol  Fbs  RestECG     MaxHR  \
0    0.000000  0.708333  1.0  0.481132  0.244292  1.0      1.0  0.603053
1    0.003311  0.791667  1.0  0.622642  0.365297  0.0      1.0  0.282443
2    0.006623  0.791667  1.0  0.245283  0.235160  0.0      1.0  0.442748
3    0.009934  0.166667  1.0  0.339623  0.283105  0.0      0.0  0.885496
4    0.013245  0.250000  0.0  0.339623  0.178082  0.0      1.0  0.770992

   ExAng   Oldpeak  Slope        Ca  ChestPain_nonanginal  \
0    0.0  0.370968    1.0  0.000000                 False
1    1.0  0.241935    0.5  1.000000                 False
2    1.0  0.419355    0.5  0.666667                 False
3    0.0  0.564516    1.0  0.000000                  True
4    0.0  0.225806    0.0  0.000000                 False

   ChestPain_nontypical  ChestPain_typical  Thal_normal  Thal_reversable  \
0                 False               True        False            False
1                 False              False         True            False
2                 False              False        False             True
3                 False              False         True            False
4                  True              False         True            False

   AHD_Yes
0    False
1     True
2     True
3    False
4    False
[ ]:
```