Task 1: Data Cleaning & Preprocessing

Objective:

To learn how to clean and prepare raw data for Machine Learning models using Python libraries like Pandas, NumPy, Matplotlib, and Seaborn.

Tools Required:

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn

Step-by-Step Guide:

1. Import Libraries:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

2. Load Dataset:

```
df = pd.read_csv('your_data.csv')
df.head()
```

3. Understand Dataset:

```
df.info()
```

```
df.describe()

df.shape

df.columns
```

4. Handle Missing Values:

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

df.dropna(inplace=True)

df['column_name'].fillna(df['column_name'].mean(), inplace=True)
```

5. Handle Duplicates:

```
df.duplicated().sum()

df.drop_duplicates(inplace=True)
```

6. Fix Data Types:

```
df['date_column'] = pd.to_datetime(df['date_column'])

df['numeric_column'] = pd.to_numeric(df['numeric_column'], errors='coerce')
```

7. Outlier Detection & Removal:

```
Q1 = df['column'].quantile(0.25)

Q3 = df['column'].quantile(0.75)

IQR = Q3 - Q1

df = df[~((df['column'] < (Q1 - 1.5 * IQR)) | (df['column'] > (Q3 + 1.5 * IQR)))]
```

8. Data Encoding (Categorical to Numeric):

```
# Label Encoding

df['category'] = df['category'].astype('category').cat.codes

# One-Hot Encoding
```

```
df = pd.get_dummies(df, columns=['category_column'])
```

9. Feature Scaling:

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

df[['feature1', 'feature2']] = scaler.fit_transform(df[['feature1', 'feature2']])
```

10. Data Visualization:

```
# Missing values heatmap

sns.heatmap(df.isnull(), cbar=False, cmap='viridis')

# Correlation matrix

sns.heatmap(df.corr(), annot=True, cmap='coolwarm')

# Boxplot for outliers

sns.boxplot(x=df['column'])
```

Save Cleaned Data:

```
df.to_csv('cleaned_data.csv', index=False)
```