from google.colab import drive

drive.mount('/content/drive')

import pandas as pd

import numpy as np

from sklearn.preprocessing import StandardScaler

file\_path = '/content/drive/My Drive/CSE422 LAB/Assignment 4 (ML)/Dataset/Housing Price.xlsx'

df = pd.read\_excel(file\_path)

print("Initial Dataset:")

print(df.head())

# Task 1: Remove null values

df\_cleaned = df.dropna()

print("\nDataset after removing null values:")

print(df\_cleaned.head())

# Task 2: Remove duplicate rows

df\_cleaned = df\_cleaned.drop\_duplicates()

print("\nDataset after removing duplicate rows:")

print(df\_cleaned.head())

# Task 3: Handle categorical variables (Binary encoding and One-Hot Encoding)

binary\_columns = ['mainroad', 'guestroom', 'basement', 'hotwaterheating', 'airconditioning', 'prefarea']

df\_cleaned[binary\_columns] = df\_cleaned[binary\_columns].replace({'yes': 1, 'no': 0})

# One-Hot Encoding for 'furnishingstatus'

df\_cleaned = pd.get\_dummies(df\_cleaned, columns=['furnishingstatus'], drop\_first=True)

print("\nDataset after handling categorical variables:")

print(df\_cleaned.head())

# Task 4: Feature scaling for continuous variables

scaler = StandardScaler()

df\_cleaned[['price', 'area', 'parking']] = scaler.fit\_transform(df\_cleaned[['price', 'area', 'parking']])

print("\nDataset after feature scaling:")

print(df\_cleaned[['price', 'area', 'parking']].head())

# Task 5: Remove variables with high correlation (threshold > 0.8)

corr\_matrix = df\_cleaned.corr().abs()

upper = corr\_matrix.where(np.triu(np.ones(corr\_matrix.shape), k=1).astype(bool))

to\_drop = [column for column in upper.columns if any(upper[column] > 0.8)]

df\_cleaned.drop(columns=to\_drop, inplace=True)

print("\nDataset after removing highly correlated features:")

print(df\_cleaned.head())