GTSC2143 Machine Learning for Business
Tutorial 4

Please write down your answers in this document and submit it at iSpace by the end of this tutorial.

Activity 1.　　Data Loading and Preprocessing

1. Load the Dataset

　　a)　Load the house prices dataset using pandas:

```
1. import pandas as pd
2. import numpy as np
3. from sklearn.model_selection import train_test_split
4. from sklearn.linear_model import LinearRegression
5. from sklearn.metrics import mean_squared_error, r2_score
6. import matplotlib.pyplot as plt
7. import seaborn as sns
8.
9. # Load the house prices dataset
10. data = pd.read_csv("GTSC2143-Lecture 4 predicting-house-prices-assignment_home_data.csv")
```

　　b)　Display basic information:

- Dataset shape
- First 5 rows
- Column names and data types

　　c)　Check for any missing values in the dataset

2. Data Filtering

　　a)　Filter the data to exclude the house whose id is '1925069082'

　　b)　Display the shape of the filtered dataset

3. Train/Test Split

　　a)　Split the filtered data into training (80%) and testing (20%) sets using `train_test_split`

　　b)　Use `random_state=42` for reproducible results

　　c)　Display the shapes of training and testing sets

　　d)　Analysis: Write 2-3 sentences explaining the purpose of train/test split in machine learning.

Activity 2.　　Model Training, Evaluation and Prediction

1. Feature Selection and Model Training

　　a)　Select these features for your regression model:

- 'bedrooms'
- 'bathrooms'
- 'sqft_living'
- 'sqft_lot'
- 'floors'
- 'zipcode'

　　b)　Create and train a Linear Regression model using the selected features

　　c)　Display the model coefficients and intercept

　　d)　Analysis: Write 2-3 sentences interpreting what the coefficients tell us about each feature's impact on house prices.

2. Evaluate Model Quality

a) Make predictions on the test set
b) Calculate and display the following metrics:
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score (coefficient of determination)
c) Analysis: Write 2-3 sentences interpreting these evaluation metrics and assessing the model's performance.

3. Predict for the Excluded House
a) Find the house with id '1925069082' in the original dataset
b) Extract the feature values for this house
c) Use your trained model to predict the price of this house
d) Display the actual price of this house
e) Calculate the prediction error:
- Absolute error
- Percentage error
f) Analysis: Write 2-3 sentences evaluating how accurate your model's prediction was for this specific house.

- End of Tutorial 4 -