Data Analysis and Visualization Lab (Final Term)

Total marks: 100

Instructions:

1- **Attempt all questions.**
2- **Take a screen shot with output and past in the provide space for each question.**
3- **You will also submit the original code file with this file.**
4- **You will make a folder and place the code file and word file. Folder name is your Roll Number**

Part1: (60 marks)

Read the CSV file name **kc_house_data.csv** and do the following operation on this dataset.

Question 1 (5 marks)

Display the data types of each column. Take a screenshot of code with output and paste here.

Question 2 (5 marks)

Drop the columns "id" and "Unnamed: 0" from axis, then use the method to obtain a statistical summary of the data. Take a screenshot of code with output and paste here.

Question 3 (20 marks)

Apply Data Wrangling processes on the given dataset (missing values, redundant features etc) as discussed in class. Provide visual evidence (plots, charts etc...) for each process you are going to apply on and why? Take a screenshot of code with output and paste here.

Question 4 (10 marks)

Use the method to count the number of houses with unique floor values, use the method to convert it to a dataframe. Take a screenshot of code with output and paste here.

Question 5 (10 marks)

Fit a linear regression model to predict the 'price' using the feature 'sqft_living'. Take a screenshot of code with output and paste here.

Question 6 (10 marks)

Fit a linear regression model to predict the 'price' using the list of features:

features =["floors", "waterfront","lat" ,"bedrooms" ,"sqft_basement" ,"view" ,"bathrooms","sqft_living15","sqft_above","grade","sqft_living"]

Take a screenshot of code with output and paste here.

Part2:

Read the CSV file name **spam-RAW.csv** and do the following operation on this dataset.

Question 1 (10 marks)

Apply Data Wrangling processes on the given dataset (missing values, redundant features etc) as discussed in class. Which are the top 11 features/predictors that appear to vary the most between spam and non-spam emails? Provide visual evidence (plots, charts etc...) for each process you are going to apply on and why? Take a screenshot of code with output and paste here. Take a screenshot of code with output and paste here.

Question 2 (10 marks)

Fit a logistic regression model to predict the spam. Take a screenshot of code with output and paste here.

Part3:

Read the CSV file name **bird-window-collision-death.csv** and do the following operation on this dataset.

Question 1 (20 marks)

Using Plotly express shows Building Number, Percentage on Pie Chart along with on Hover shows each species death count? Change Color with respect to Building Sides? Take a screenshot of code with output and paste here.