Team project

Please write down your names on the first line of the report

(12 pts) Q1. Work with the sharing bike dataset: Sharing_Bike.db

In order to encourage people to use shared bicycles and design a reasonable tariff, we shall analyze the data from different perspectives.

The data of sharing bike contains three tables: bike, Calendar, and weather.

- Variables in bike:
  - instant: primary key
  - dteday: date
  - hr: hour (from 0 to 23)
  - casual: count of casual users in that hour
  - registered: count of registered users in that hour
  - cnt: total count in that hour
- Variables in Calendar:
  - cid: primary key
  - detday: date
  - season: season (1: springer, 2: summer, 3: fall, 4: winter)
  - holiday: is holiday or not (0: normal day, 1: holiday)
  - weekday: day of the week
  - workingday: if day is neither weekend nor holiday is 1, otherwise is 0.
  - Freq: this attribute can be ignored.
- Variables in weather:
  - field1: primary key
  - dteday: date
  - hr: hour (from 0 to 23)
  - temp: temperature in Celsius
  - atemp: feeling temperature in Celsius
  - hum: humidity
  - windspeed: windspeed

Work on the following tasks and attach your SQL codes for question A to C into report.

A. (4 pts) Considering the holidays, write a SQL code to calculate the total count of casual users and the total count of registered users in each holiday of 2012.

B. (4 pts) Considering 24 hours in a day, write a SQL code to calculate the average count of casual users and the average count of registered users for each hour.

C. (4 pts) Output the top 20 dates of the highest total counts of casual users in 2012 with a SQL code.

(8 pts) Q2. Work with the real estate dataset, realestate.xlsx or realestate.csv

- Variables – numerical attributes
  - SalePrice: sale price (in thousand)
  - Size: size of the real estate (in square feet)
  - Beds: number of bedrooms
  - Baths: number of bathrooms
  - Num_Garage: number of garages
  - Year: when the real estate was built
- Variables: binary attributes
  - Highway: accessible to the highway in 10 minutes, yes or no
  - Aircondition: with or without air conditioner
  - Swimmingpool: with or without swimming pool

For question A and B, please hand in all source codes and the result.

A. (4 pts) You are asked to predict **Sale Price**, what are your **reasonable choices** from the dataset? Please report the total amounts of reasonable choices in the report, write a python code to express all your reasonable choices in regression formula.
Note: reasonable choices refer to considering all possible combinations of independent variables.

B. (4 pts) To avoid overfitting, we would set some constraints when training the model. Please write a python code to find the best linear regression with the following two constraints and export the regression summary to Q2_b.txt:
   1. **Highway** should be one of the independent variables**.**
   2. The total number of independent variables of the model should not exceed 4.