HW-2

P.Kohli

Due Date: 2/13/25 by 10PM

Problem 1: [20 points]

Using Ozone hourly 2024 data compare the EDA findings with those done in class for the 2020 data.

Problem 2: [5 points]

Collecting data is often a messy process resulting in multiple errors in the data. Consider the following small vector representing the weights of 10 adults in pounds.

```
my.weights <- c(150, 138, 289, 239, 12, 103, 310, 200, 218, 178)
```

As far as we know, it's not possible for an adult to weigh 12 pounds, so that is most likely an error. Change this value to NA, and then find the standard deviation of the weights after removing the NA value.

Problem 3: [10 points]

Consider the following variables: age and income:

```
age <- c("middle age", "senior", "middle age", "senior",
         "senior", "senior", "senior", "middle age")
income <- c("lower", "lower", "upper", "middle", "upper",
            "lower", "lower", "middle")
```

(a) what is the class of each variable?

(b) change the age variable to a factor with levels for age as: youth, young adult, middle age, and senior.

(c) change the income variable to a factor with levels as lower, middle, and upper.

Problem 4: [5 points]

Suppose you keep track of the mileage each time you fill up. At your last 8 fill-ups the mileage was: 65311, 65624, 65908, 66219, 66499, 66821, 67145, 67447.

Enter this data into R. Use the function diff on the data (use ?diff). Use the documentation for diff function to learn about it and then explain briefly what does this function give you?

Problem 5: [15 points]

Create the following data frame:

```
my_data <- data.frame(student_id = c(100234, 132454, 453123),
                      test_1_grade = c(82, 93, 87),
                      hw_1_grade = c(92, 89, 98),
                      session = c("7 AM", "7 PM", "7 AM"))
my_data
```

```
##   student_id test_1_grade hw_1_grade session
## 1     100234           82         92    7 AM
## 2     132454           93         89    7 PM
## 3     453123           87         98    7 AM
```

Obtain the column names of our data frame.

Get the number of rows or columns in a data frame, try nrow(), ncol(), or dim() functions.

To subset rows and columns of a data frame we can use the following syntax: my_data_frame[row condition, column condition]. The row/column conditions may be either numeric indexes, logical expressions, or vectors

Explain the subset you get from the following code:

```
my_data_frame <- data.frame(make = c("Toyota","Honda","Ford", "Toyota",
                                     "Ford", "Honda"),
                            mpg = c(34, 33, 22, 32, 29, 27),
                            cylinders = c(4, 4, 8, 6, 6, 8))
```

Explain the subset you get from the following code:

```
a.)     my_data_frame[1:3,1:2]

b.)     my_data_frame[c(1, 2, 3), c(1, 2)]

c.)     my_data_frame[2, c(1, 3)]

d.)     my_data_frame[1:2, ]

e.)     logical_condition <- my_data_frame$mpg >= 30
   my_data_frame[logical_condition, ]

f.)     my_data_frame[my_data_frame$mpg >= 30, ]

g.)     my_data_frame[my_data_frame$mpg >= 32, c(2, 3)]

h.)     my_data_frame[my_data_frame$mpg >= 32, c("mpg", "cylinders")]
```

i.) Now try to subset my_data_frame to only include rows that have a cylinders value of 4.

Problem 6: [10 points]

Create data as the following list:

```
my_list <- list(classes_offered = c("MIS 431", "MIS 310", "MIS 410", "MIS
412"),student_data = data.frame(student_id = c(54, 100, 32, 423,
                                                2, 19, 39),
                                age = c(18, 22, 27, 18, 29,
                                        22, 20),
                                gpa = c(3.1, 2.8, 3.7, 3.4, 3.2,
                                        3.4, 3.2),
                                stringsAsFactors = FALSE))
```

Write the R code that calculates the median value (use the median() function) of the gpa variable in student_data. All you need to do is pass the student_id vector into the median() function.

Problem 7: [10 points]

Let us first create dataframes.

```
Feature1A <- c("A", "B", "C", "D")
Feature2A <- c(1000, 2000, 3000, 4000)
Feature3A <- c(25.5, 35.5, 45.5, 55.5)
Feature4A <- c(10, 34, 78, 3)
Dataframe1 <- data.frame(Feature1A, Feature2A, Feature3A, Feature4A)
colnames(Dataframe1) <- c("Feature1", "Feature2", "Feature3", "Feature4")
Dataframe1

##   Feature1 Feature2 Feature3 Feature4
## 1        A     1000     25.5       10
## 2        B     2000     35.5       34
## 3        C     3000     45.5       78
## 4        D     4000     55.5        3

# creating Dataframe2
Feature1B <- c("E", "F", "G", "H")
Feature2B <- c(5000, 6000, 7000, 8000)
Feature3B <- c(65.5, 75.5, 85.5, 95.5)
Dataframe2 <- data.frame(Feature1B, Feature2B, Feature3B)
colnames(Dataframe2) <- c("Feature1", "Feature2", "Feature3")
Dataframe2

##   Feature1 Feature2 Feature3
## 1        E     5000     65.5
## 2        F     6000     75.5
```

```
## 3        G     7000     85.5
## 4        H     8000     95.5
```

Merge merges Features 1-3 of the two data frames and called the resulting dataframe as Output. Use function merge().

Problem 8: [25 points]

(a)

Import the data from Moodle or shared Google drive, it is called pima.csv. Change the name of the nine columns to preg_times, glucose_test, blood_press, tsk_thickness, serum, bm_index, pedigree_fun, age, class.

(b)

All patients (768 Observations) in this dataset contains are females at least 21 years old of Pima Indian heritage. All zero values for the biological variables other than number of times pregnant should be treated as missing values. Count how many zeros are there in each variable (column). For any 0 in the data (except for class and preg_times) assign it as an NA.

(c)

For class variable, check if it is a factor and if not, then make it a factor with levels 0 replaced with neg (for negative diabetic) and 1 replicated with pos (for positive diabetic).

(d)

Make data subsets for four age groups: 21-36, 37-51, 52-66 and 67-81. ## (e) Create a new factor vector called age.factor, with age in pima data replaced with the age group.

Project Problem: [20 points]

Do the following for the approved dataset(s). ## (a) Read the data here in R. ## (b) Show the structure of data. ## (c) What is the dimension of your data? ## (d) Show names of variables in the data. ## (e) Find easy answers to your research question (one of them) using the data