Introduction to Data Science

Instructor: Daniel D. Gutierrez

HOMEWORK 2

NOTE: for this homework assignment, please use the airquality, CO2 and mtcars data sets from base R where appropriate. Please submit a single R script that includes the R code as well as output inserted as comments. There's no need to submit the plots.

Question 1

Using the sqldf() function found in the sqldf package to select data from the CO2 data set, execute the SQL statement required to calculate the average value for uptake grouped by Type [Hint: use Google to find tutorials for using SQL with this package.]

Question 2

Use the following vector assignments to provide data content for a new data frame:

```
Died.At <- c(22,40,72,41)
Writer.At <- c(16, 18, 36, 36)
First.Name <- c("John", "Edgar", "Walt", "Jane")
Second.Name <- c("Doe", "Poe", "Whitman", "Austen")
Sex <- c("MALE", "MALE", "MALE", "FEMALE")
Date.Of.Death <- c("2015-05-10", "1849-10-07", "1892-03-
26","1817-07-18")
```

Write some data munging code to performing the following operations:

- Create a new data frame df with the above data for each of six columns. Make sure that character values are NOT converted to factors.
- Use the appropriate as.() function to coerce the Sex variable to a factor.
- The variable names are inconvenient so write R code to change them to: age_at_death, age_as_writer, first_name, surname, gender, date_died [Hint: remember the names() function for data frames.]
- Say "John Doe" died on his birthday, calculate and display the birthdate value based on the variables date_died and age_at_death

Question 3

When recording experimental observations, there are two general formats – “long” and “wide.” The long format for recording observations is when there is one observation row per variable. A lot of statistical tests favor this format. Here is an example of wide format:

product	height	width	weight
A	10	5	2
B	20	10	NA

The following R code generates the wide format with some simulated data:

```
> product <- c("A", "B")

> height <- c(10,20)

> width <- c(5,10)

> weight <- c(2,NA)

> observations_wide <- data.frame(product, height, width,
weight)

> observations_wide

  product height width weight
1       A     10     5      2
2       B     20    10     NA
```

The wide format for recording observations is when you have multiple values, spread out over multiple columns, for the same observations. Since different functions may require you to input your data either in long or wide format, you might need to reshape your data set. Write a data transformation R script to take the `observations_wide` data frame above and convert it to long format. Here is what the output should look like below (make sure you order the rows to match the results shown).

product	variable	value
A	Height	10
A	Width	5
A	Weight	2
B	Height	20
B	Width	10

[Hint: take a look at `reshape2` package and the `melt()` function, which should also handle the removal of NAs]

Question 4

Let’s take a look at the `mtcars` data set that comes in Base R. The data can be loaded with the code:

```
library(datasets)
data(mtcars)
? mtcars     # View a description of the data set
```

You will now see an object called `mtcars` in your workspace. Which of the following R code statements calculates the average miles per gallon (`mpg`) by number of cylinders in the car (`cyl`)?

(a) `sapply(mtcars, cyl, mean)`

(b) `lapply(mtcars, mean)`

(c) `sapply(split(mtcars$mpg, mtcars$cyl), mean)`

(d) `tapply(mtcars$cyl, mtcars$mpg, mean)`

Question 5

Using the `mtcars` data set, what is the absolute difference between the average horsepower of 4-cylinder cars and the average horsepower of 8-cylinder cars? [Hint: remember the `abs()` function for calculating the absolute value of a number.]

Question 6

What is the mean of the `Ozone` column in the `airquality` data set? Exclude missing values (coded as NA) from this calculation.

(a) 42.1

(b) 53.2

(c) 31.5

(d) 18.0

Question 7

Using the `airquality` data set, what is the mean value of the `Temp` variable when the `Month` variable is equal to 6?

(a) `85.6`

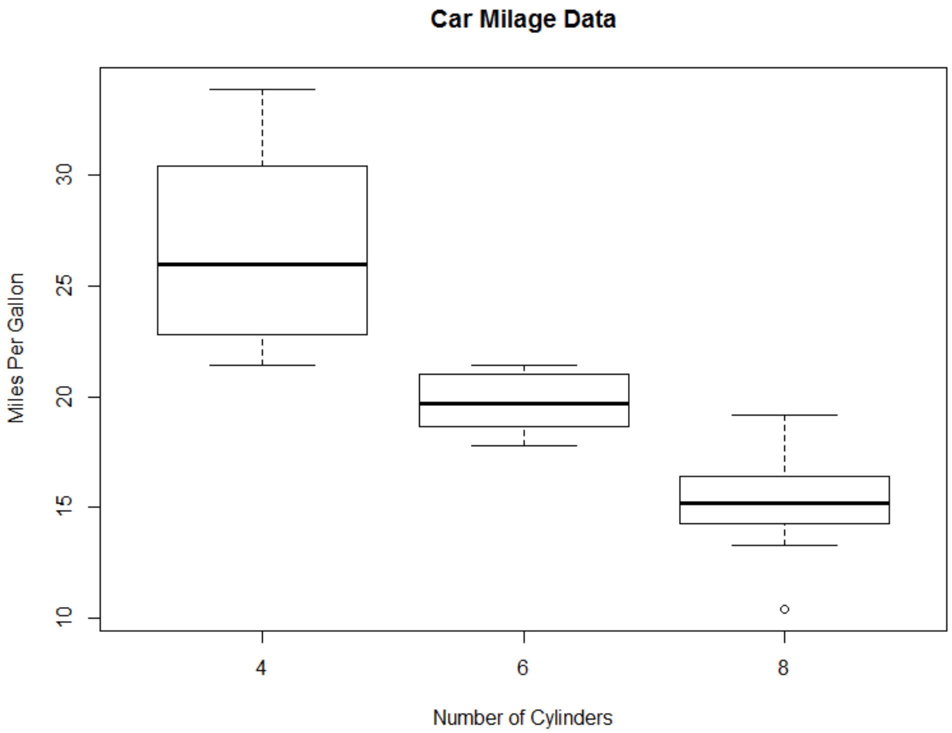
(b) `90.2`

(c) `75.3`

(d) `79.1`

Question 8

Provide the R code necessary to reproduce the boxplot data visualization below using the `mtcars` data set.



Question 9

In this question, you'll need to install, load and learn about a new package called `scatterplot3d` in order to render a 3D visualization for the `mtcars` data set. You'll need to use the `scatterplot3d()` function in this package. Please provide the R code that produces a data visualization with the following requirements:

- Use the variables for the X, Y and Z axis respectively: `wt, disp, mpg`
- Include an appropriate title for the plot
- Include labels for each axis and include the units for the variables
- Configure the appropriate argument that specifies that a grid should be drawn on the plot.
- Add a fourth data point to the plot, i.e. the `am` variable (transmission), by using the PCH argument.

Question 10

The `airquality` data set contains data on different measures of air quality in New York City. Please provide the R code that produces the following data visualizations:

- Produce a scatterplot of the ozone level versus the temperature for the complete set of observations.
- Next, produce a scatterplot of ozone less than 100 versus temperatures less than 80 (Hint: there are a number of approaches to this problem including – subsetting, or the `plot()` function's `xlim` and `ylim` arguments along with the `min()` statistical function).