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.

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`

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 change them to: `age_at_death`, `age_as_writer`, `first_name`, `surname`, `gender`, `date_died`
- 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:

Product	Attribute	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]

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?

Question 6

What is the mean of the Ozone column in this dataset? 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`

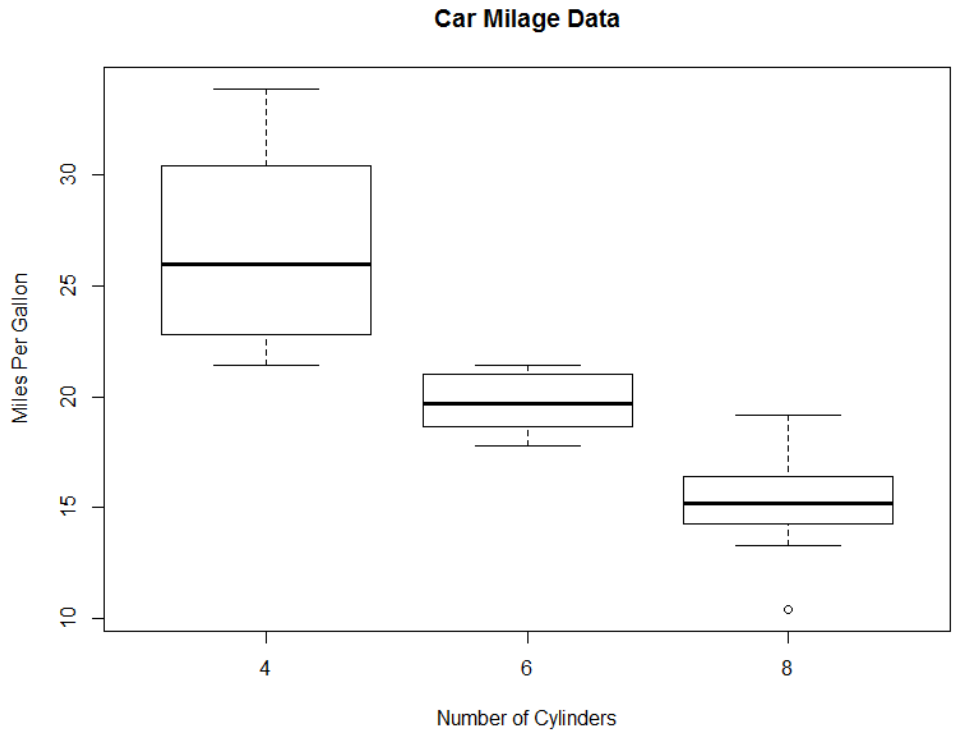
(b) `90.2`

(c) `75.3`

(d) `79.1`

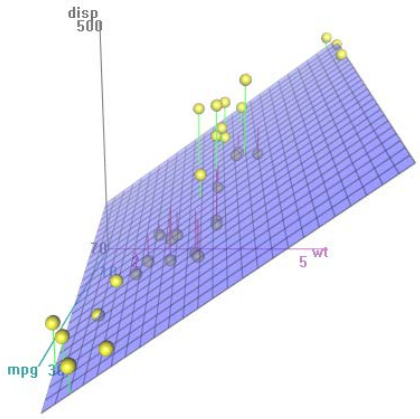
Question 8

Write the R script code necessary to reproduce the boxplot 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 `Rcmdr` in order to render an interactive 3D visualization for the `mtcars` data set that appears like the image below. You'll need the `scatter3d()` function in this package. We need to visualize the variables: `wt, disp, mpg`. Hand in your R script code that produces this data visualization.



Question 10

The `airquality` data set contains data on different measures of the air quality in New York City. Produce a plot of the ozone level versus the temperature for the complete set of observations.

Next, produce a plot to display temperatures less than 80 and ozone less than 100 (Hint: one way would be to use the `min()` function along with the `plot()` function's `xlim` and `ylim` arguments).