

HW 1, STAT 650

Due: Friday, September 6

Directions: The assignment should be completed using Quarto and submitted to Canvas as a self-contained HTML or PDF file.

Reading: Chapter 1 and Appendix B from *Modern Data Science with R*

Exercise 1

- (a) What are the four common data types of vectors in R?
- (b) How is a matrix different from a data frame?
- (c) What does `as.integer()` do when applied to a logical vector?
- (d) What does `as.matrix()` do when applied to a data frame with columns of different types?

Exercise 2

Determine the type of each vector below. Print out the value of each vector and use the `typeof()` function.

```
a <- c(1, 2, 3, 4, 5)
b <- 1:5
c <- c(sqrt(2), 4.7e4, 1/0)
d <- c(T, T, T, T)
e <- c("1", 2, 3)
f <- c(7L, NA, NA, 5L, 3L)
g <- c(7L, "NA", "NA", 5L, 3L)
h <- c()
```

Exercise 3

This exercise uses the `airquality` data frame, which is already loaded into R.

```
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67      5    1
## 2    36     118  8.0   72      5    2
## 3    12     149 12.6   74      5    3
## 4    18     313 11.5   62      5    4
## 5    NA      NA 14.3   56      5    5
## 6    28      NA 14.9   66      5    6
```

The data frame contains daily air quality measurements in New York from May to September 1973. Type `help(airquality)` in the console to read about this data in the help menu.

- (a) Run the following code to subset the `Ozone` column and assign it to a variable called `Ozone1`.

```
Ozone1 <- airquality$Ozone
```

- (b) Use `is.na()` to remove the missing data (`NA` values) from the vector `Ozone1`. Assign the vector with the missing values removed to a variable called `Ozone2`. How many `NA` values were removed?
- (c) Compute the min, median, mean, max, and standard deviation of the numeric vector `Ozone2`.
- (d) Run the following commands, and explain how each command handles missing data.

```
summary(airquality$Ozone)
sd(airquality$Ozone)
sd(airquality$Ozone, na.rm = TRUE)
```

- (e) How many rows of the data frame `airquality` have one or more missing values? [Hint: use the `na.omit()` function]

Exercise 4

In this exercise, you will graph the function $f(p) = p(1 - p)$ for $p \in [0, 1]$.

- (a) Use `seq()` to create a vector `p` of numbers from 0 to 1 going in increments of 0.2.
- (b) Use `plot()` to plot `p` in the x coordinate and `p(1-p)` in the y coordinate. Read the help page for `plot` and experiment with the `type` argument to find a good choice for this graph.
- (c) Repeat, but with creating a vector `p` of numbers from 0 to 1 going in increments of 0.01.

Exercise 5

This exercise uses the `movies` data frame from Lecture 3.

```
movies <- read.csv("https://ericwfox.github.io/data/movies.csv")
```

- (a) Make a histogram and box plot of the variable `runtime`. Describe the distribution in terms of its center, shape, and spread. Are there any potential outliers?
- (b) Make a scatter plot using two numerical variables of your choosing. Add the least square line, and describe the association in the scatter plot.
- (c) Make side-by-side box plots with `mpaa_rating` on the x -axis and `imdb_rating` on the y -axis. In the plot the categories (i.e., levels) of `mpaa_rating` should have the following ordering: G, PG, PG-13, R, NC-17, Unrated.