HW 1, STAT 450
Due: Wednesday, September 8

Directions: Please submit your completed assignment to Blackboard. The assignment should be completed using R Markdown and rendered to an HTML or PDF format. Note that Blackboard will not accept HTML files. One workaround is to first zip your HTML file, and then submit the zipped file to Blackboard.

Exercise 1.

(a) What are four common data types for vectors in R?

(b) Determine the data type of each vector below. Print out the values of each vector and use the `class()` function.

```
a <- c(1, 2, 3, 4, 5)
b <- 1:5
c <- c("blue", "orange", "red")
d <- c(T, T, T, T)
e <- c("1", 2, 3)
f <- c(7, NA, NA, 5 , 3)
g <- c(7, "NA", "NA", 5 , 3)
h <- c()
```

Exercises 2 and 3 will use 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. Enter `help(airquality)` into the console to read about this data in the help menu.

Exercise 2. Run the following code to subset the `Ozone` column and assign it to a variable called `Ozone1`.

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

(a) 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?

(b) Compute the min, median, mean, max, and standard deviation of the numeric vector `Ozone2`.

(c) Run the following commands, and explain how each command handles missing data.

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

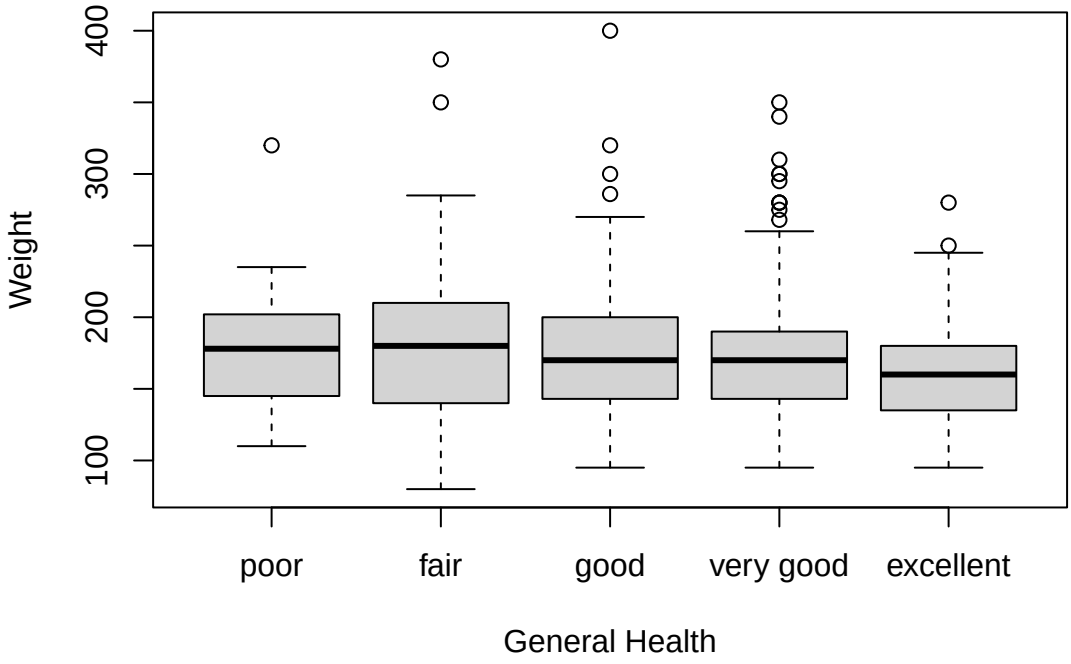
Exercise 3. Make a scatter plot with `Temp` on x-axis and `Ozone` on the y-axis. Label the x-axis "Temperature (degrees F)"and the y-axis "Ozone (ppb)". Describe the association between the two variables.

The following exercises will use the CDC data set discussed in lecture 5. Run the following command to load this data set into R:

```
cdc <- readRDS(url("https://ericwfox.github.io/data/cdc1000.rds"))
```

Exercise 4. Make a bar plot for the variable `exerany`. Label the x-axis "Exercised in the last month" and the label bars "no" and "yes".

Exercise 5. Make side-by-side box plots of `weight` across the categories for `genhlth`. In the plot the categories of `genhlth` should be ordered as poor, fair, good, very good, excellent (hint: use the `factor()` function to specify the correct ordering). This is what the plot should look like:



Exercise 6. Make a new variable called `wtdiff` which is the difference between each person's desired weight, `wtdesire`, and current weight, `weight` (that is, desired weight − current weight). Plot a histogram for `wtdiff` and describe the shape of the distribution.