

Name - NetID

## Homework 2

### Homework Instructions

For questions that require code, please create a code chunk directly below the question and type your code there. Your knitted pdf will show both your code and your output. You are encouraged to knit your file as you work to check that your coding and formatting is done so appropriately.

For written responses or multiple choice questions, please bold your (selected) answer.

### Grading Details

All questions will be graded full credit (1 point), half credit (0.5 point) or no credit (0 points).

Full credit responses should have the correct response and appropriate code (if applicable). Half credit responses will have a reasonable attempt (typically no more than one small error or oversight), and no credit responses will be either non-attempts or attempts with significant errors.

### Exercise 1

**General Note** For each dataset we explore, you might want to View the data using `View(data_name)` in the starter code. However if you can, delete these codes (or place `#` in front) before knitting your pdf for submission to avoid unnecessary clutter in your final report when grading.

**Also** Every time you knit your file, you'll notice that the documentation for each `?data` entry opens a webpage. If that bothers you, you may delete that code or place `#` in front to prevent it from running automatically.

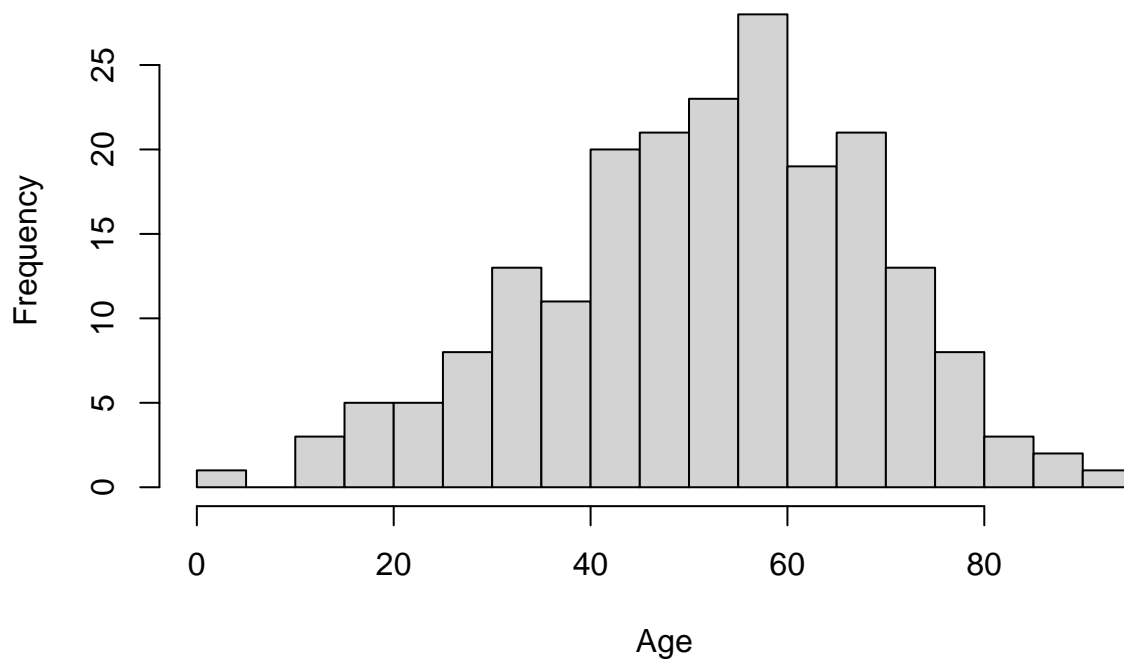
```
# starter
library(MASS)
#?Melanoma
```

For the next several questions, we will be working with the `Melanoma` dataset from the `MASS` package.

Create a histogram of `age` in the `Melanoma` dataset. Please give your histogram 20 bins, label the x axis "Age", and provide the following plot title: "Age of Melanoma Patients"

```
# solution
hist(Melanoma$age,
     breaks = 20,
     xlab = "Age",
     main = "Age of Melanoma Patients")
```

## Age of Melanoma Patients



### Exercise 2

Create a table that shows what proportion of patients in the dataset died of a melanoma, remained alive, or died from other causes.

*Use an R command to produce this number.*

After running this result, directly answer what percent of patients in this dataset died of a melanoma (you don't need to extract this with R code—just answer directly as a statement below your code chunk.)

*You can find information on what each status value means from the documentation!*

```
# solution
table(Melanoma$status)/nrow(Melanoma)
```

```
##
##           1           2           3
## 0.27804878 0.65365854 0.06829268
```

**27.8% of these patients died of a melanoma**

### Exercise 3

What is the average age of individuals in the `Melanoma` dataset who remained alive?

*Hint: First, create a subset of patients who are alive. Then take the mean of the age variable from this subset. You can do it in two lines, or if you want to challenge yourself, do it in just one line!*

```
# solution
```

```
mean(Melanoma[Melanoma$status == 2, ]$age)
```

```
## [1] 50.00746
```

```
#or alternatively
```

```
mean(subset(Melanoma, status == 2)$age)
```

```
## [1] 50.00746
```

#### Exercise 4

```
#starter
```

```
library(MASS)
```

```
##mammals
```

We will now be using the `mammals` dataset from the `MASS` package.

Which animal in the `mammals` dataset has the largest brain weight relative to its body weight? (That is, the largest brain weight to body weight ratio.)

*Hint: which.max might be helpful here.*

*Hint: Notice that names is not exactly a variable, but instead represent the actual row names. To extract the animal name only, use the `rownames` function around your subsetted row.*

```
#solution
```

```
rownames(mammals[which.max(mammals$brain/mammals$body), ])
```

```
## [1] "Ground squirrel"
```

#### Exercise 5

```
#starter
```

```
##iris
```

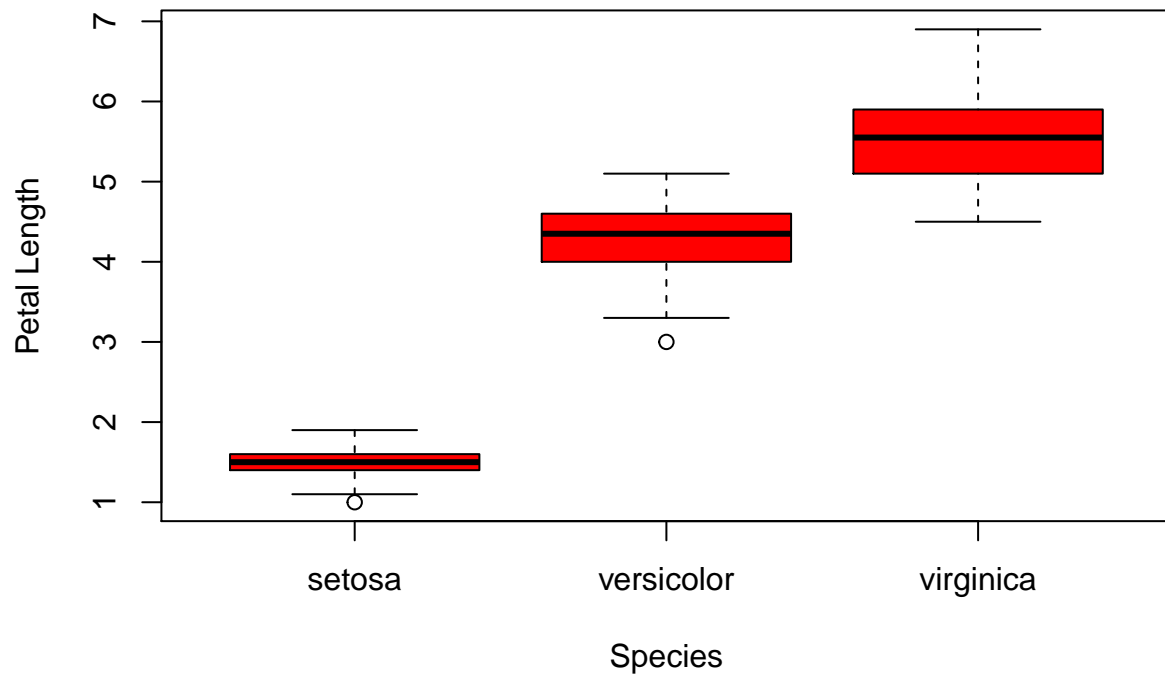
The `iris` dataset is built into base `R` and thus requires no package to load.

Create side by side boxplots that compare `Petal.Length` of the three different species of iris plants in the dataset. Color the fill of your boxplots red, re-label the y axis as “Petal Length”, and title the plot “Petal Length of Species”

```
#solution
```

```
boxplot(Petal.Length ~ Species, data = iris,  
        col = "red",  
        ylab = "Petal Length",  
        main = "Petal Length of Species")
```

## Petal Length of Species



### Exercise 6

```
#starter  
#?airquality
```

Using the `airquality` dataset, what is the average wind speed in May ?

```
# solution  
mean(subset(airquality, Month == 5)$Wind)
```

```
## [1] 11.62258
```

```
#or alternatively  
mean(airquality[airquality$Month == 5, ]$Wind)
```

```
## [1] 11.62258
```

### Exercise 7

Using the `airquality` dataset, what is the average ozone measurement?

*Hint: Whenever you use a function that doesn't produce a result as expected, try reading the documentation for that function. You might find something helpful that addresses the issue!*

```
# solution
mean(airquality$Ozone, na.rm = TRUE)
```

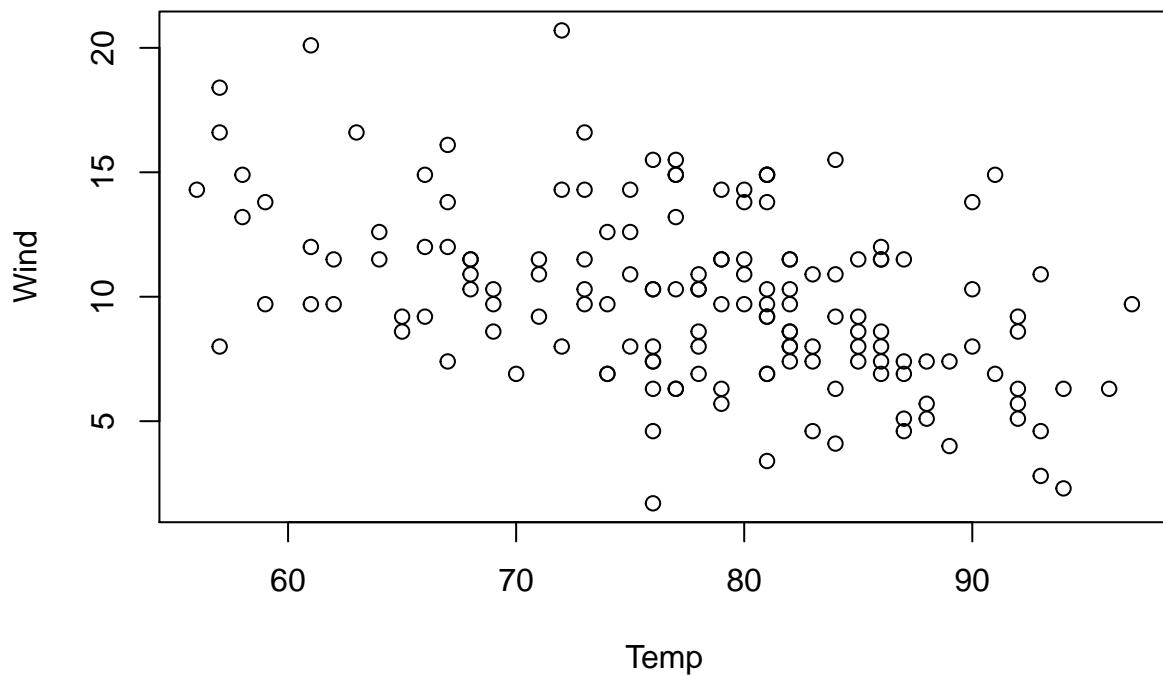
```
## [1] 42.12931
```

### Exercise 8

Using the `airquality` dataset, create a scatter plot to compare windspeed and temperature. Based on this plot, what appears to be true?

- Wind speed and temperature have no relationship.
- As temperature increases, wind speed slightly increases on average.
- **As temperature increases, wind speed slightly decreases on average**

```
# solution
plot(Wind ~ Temp, data = airquality)
```



### Exercise 9

Consider a random variable  $X$  that has a normal distribution with a **mean** of 5 and a **variance** of 9. Calculate  $P[X > 4]$ .

*Hint: Remember that the norm commands use standard deviations as inputs, not variance.*

*Hint: By default `pnorm()` considers area under the curve to the left of the given value.*

```
#solution
1 - pnorm(4, mean = 5, sd = 3)
```

```
## [1] 0.6305587
```

```
# or alternatively
pnorm(4, mean = 5, sd = 3, lower.tail = FALSE)
```

```
## [1] 0.6305587
```

## Exercise 10

```
#starter
set.seed(42)
```

Let's assume that the number of minutes between calls at a help center follow a poisson distribution with  $\lambda = 6$ . Run the seed code provided above, and then create a dataset titled `calls` that represents minutes between calls to that help center for a random sample of 500 calls.

It typically takes the caller just over 2 minutes to receive the call, take down the message, and resolve the issue or send the caller to someone who can. Using the simulated data, calculate the percentage of times that a call came in within 2 minutes of the previous call.

*Hint: Remember that calls will be a vector, not a data frame. Therefore, `nrow` will not work here. You'll want to use `length`.*

Now, use the appropriate probability function to calculate the expected percentage of the time the call center should expect calls within 2 minutes or less

```
#solution
calls = rpois(500, lambda = 6)
length(calls[calls <= 2])/length(calls)
```

```
## [1] 0.064
```

```
ppois(2, lambda = 6)
```

```
## [1] 0.0619688
```

## Exercise 11 (Bonus)

Let's return to the `airquality` dataset again. Might there be a difference in the average windspeed in the month of May (month 5) as compared to the month of September (month 9)?

First, create a subset of the `airquality` data that only includes these two months. Name it `Air_Comparison`

*Don't worry about the warning: "longer object length is not a multiple of shorter object length"*

Second, let's assume this data represents a random sample of air data. Run a two-sample t-test to compare the means and determine if there is evidence that May and September have different wind speeds on average.

*#solution*

```
Air_Comparison = subset(airquality, Month == 5 | Month == 9)
t.test(Wind ~ Month, data = Air_Comparison)
```

```
##
## Welch Two Sample t-test
##
## data: Wind by Month
## t = 1.6112, df = 58.99, p-value = 0.1125
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.3490012 3.2341625
## sample estimates:
## mean in group 5 mean in group 9
## 11.62258 10.18000
```

## Exercise 12 (Bonus)

**Continued from Exercise 11:** Which conclusion is most appropriate to make regarding our t-test comparison in the previous exercise?

- We have found evidence that May and September have equal wind speeds on average
- We have found evidence that May and September have non-equal wind speeds on average - **There is not enough evidence to conclude a difference in average wind speeds between May and September**