Problem Set 1

Problem 1

```
# environment set up
setwd("/Users/aryansingh/Desktop/stats_506_git/hw1")
pacman::p_load(tidyverse)
```

Part A

```
# data
abalone_data <- read.csv("abalone/abalone.data", header = F) %>% `colnames<-`(c(
    "sex", "length", "diameter", "height", "whole_weight", "shucked_weight", "viscera_weight'
))
```

Part B

```
table(abalone_data$sex)
```

```

   F    I    M
1307 1342 1528
```

Part C

1.

```r
weight_names <- c("whole_weight", "shucked_weight", "viscera_weight", "shell_weight")
weights <- abalone_data[, weight_names]
cor(weights, abalone_data$rings)
```

```
                    [,1]
whole_weight   0.5403897
shucked_weight 0.4208837
viscera_weight 0.5038192
shell_weight   0.6275740
```

The weight with the highest correlation with the rings is Shell Weight (R = 0.6275740).

2.

From cor_table (see below, Part D). For Shell Weight, the sex of I has the highest correlation (0.725).

3.

```r
abalone_data[abalone_data$rings == max(abalone_data$rings), c(weight_names, "rings")]
```

```
    whole_weight shucked_weight viscera_weight shell_weight rings
481       1.8075         0.7055         0.3215        0.475    29
```

The abalone with the most rings (29) had the following weights:

- whole: 1.8075
- shucked: 0.7055
- viscera: 0.3215
- shell: 0.475

4.

```r
mean(abalone_data$viscera_weight > abalone_data$shell_weight) * 100
```

```
[1] 6.511851
```

```r
# we can use mean here as the input is a vector of TRUE, FALSE which R will treat as 1, 0
# thus mean is equivalent the summing the TRUE entries over all observations
```

About 6.51% of abalones have a viscera weight larger than shell weight.

Part D

```r
cor_table <- abalone_data %>% group_by(sex) %>% summarise(across(all_of(weight_names), ~ cor
cor_table
```

```
# A tibble: 3 x 5
  sex   whole_weight shucked_weight viscera_weight shell_weight
  <chr>        <dbl>          <dbl>          <dbl>        <dbl>
1 F            0.267         0.0948          0.212        0.406
2 I            0.696         0.620           0.673        0.725
3 M            0.372         0.222           0.321        0.511
```

Part E

```r
# M vs F
t.test(rings ~ sex, data = subset(abalone_data, sex %in% c("M", "F")))
```

```

    Welch Two Sample t-test

data:  rings by sex
t = 3.6657, df = 2742.4, p-value = 0.0002514
alternative hypothesis: true difference in means between group F and group M is not equal to
95 percent confidence interval:
 0.1971045 0.6505082
sample estimates:
mean in group F mean in group M
        11.1293         10.7055
```

```r
# M vs I
t.test(rings ~ sex, data = subset(abalone_data, sex %in% c("M", "I")))
```

```
	Welch Two Sample t-test

data:  rings by sex
t = -27.221, df = 2859, p-value < 2.2e-16
alternative hypothesis: true difference in means between group I and group M is not equal to
95 percent confidence interval:
 -3.017808 -2.612263
sample estimates:
mean in group I mean in group M 
       7.890462       10.705497 
```

```r
# F vs I
t.test(rings ~ sex, data = subset(abalone_data, sex %in% c("F", "I")))
```

```
	Welch Two Sample t-test

data:  rings by sex
t = 29.477, df = 2508.9, p-value < 2.2e-16
alternative hypothesis: true difference in means between group F and group I is not equal to
95 percent confidence interval:
 3.023380 3.454304
sample estimates:
mean in group F mean in group I 
      11.129304        7.890462 
```