Math208-A1

MATH 208 - Assignment 1 - Christopher Zheng - 206760794

Question 1:

```
data(ToothGrowth)
# (a) Determine the (i) mode and (ii) class of the ToothGrowth data object.
mode(ToothGrowth)
```

```
## [1] "list"
```

```
class(ToothGrowth)
```

```
## [1] "data.frame"
```

```
#(b) Determine how many rows and columns the object has by using R functions.

nrow(ToothGrowth)
```

```
## [1] 60
```

```
ncol(ToothGrowth)
```

```
## [1] 3
```

```
#(c) Using boxplots, histograms, and density plots to describe the distribution of odontoblast l
engths by supplement type. Does one supplement seem to be associated with greater lengths? Expla
in your answer.

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

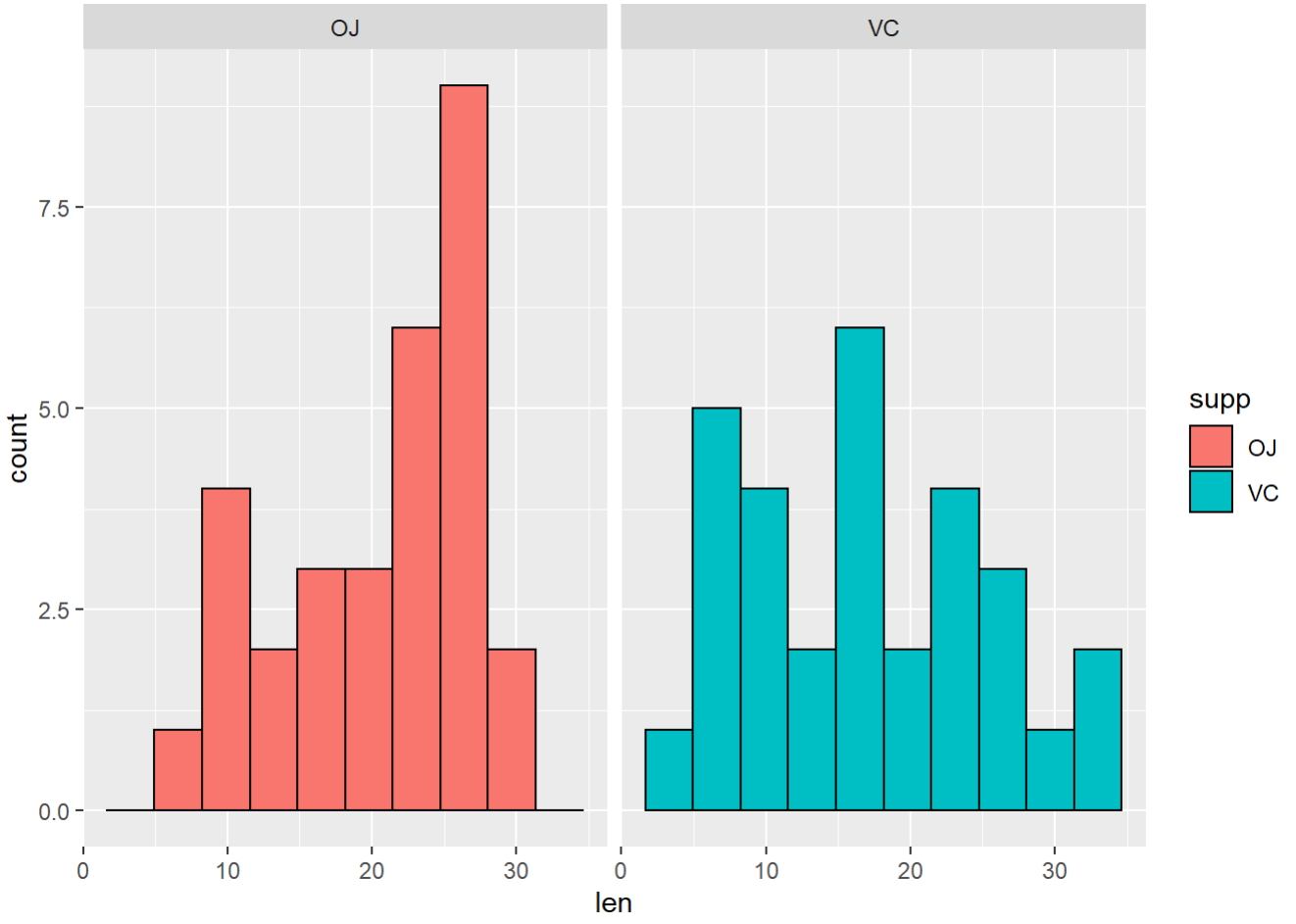
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

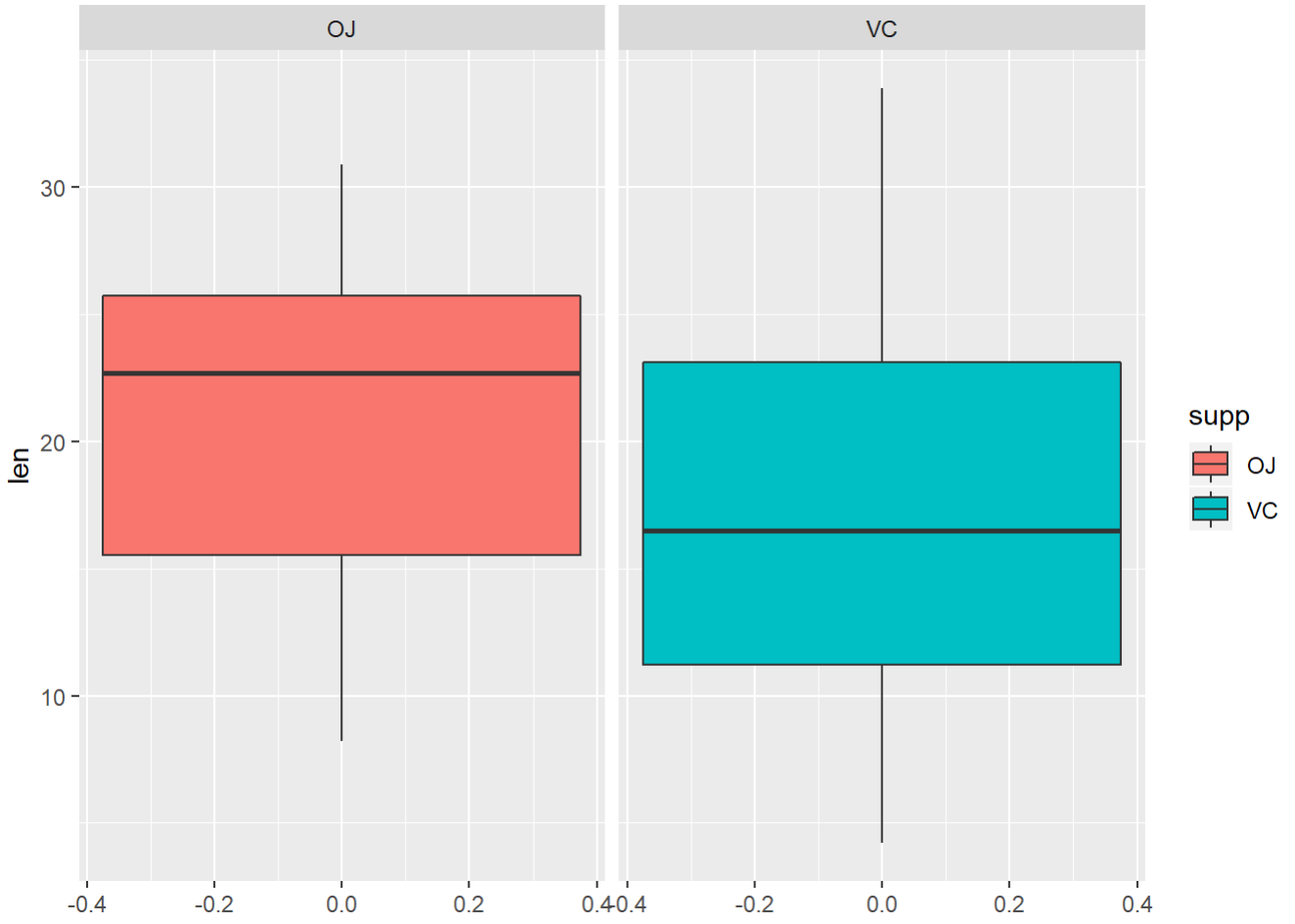
```
library(ggplot2)
tooth_df <- tbl_df(ToothGrowth)
class(tooth_df)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

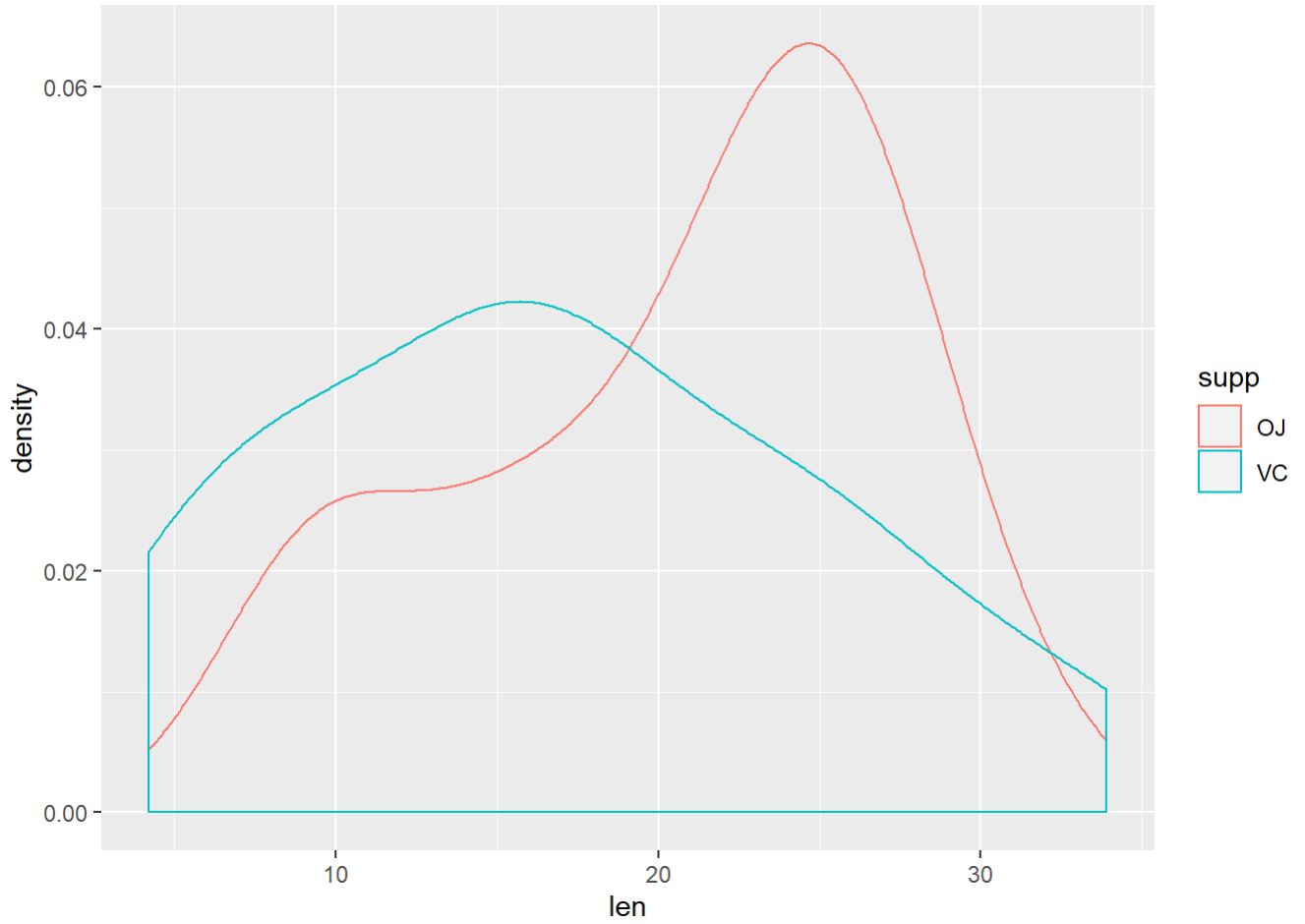
```
odon_lengths <- group_by(tooth_df,supp)

#histogram
ggplot(odon_lengths, aes(x=len,group=supp,fill=supp)) +
  geom_histogram(bins=10,col="black") +
  facet_wrap(~supp)
```




```
#boxplot
ggplot(odon_lengths, aes(y=len,group=supp,fill=supp)) +
  geom_boxplot() +
  facet_wrap(~supp)
```




```
#density plot
ggplot(odon_lengths, aes(x=len,col=supp)) +
  geom_density() + xlab("len")
```

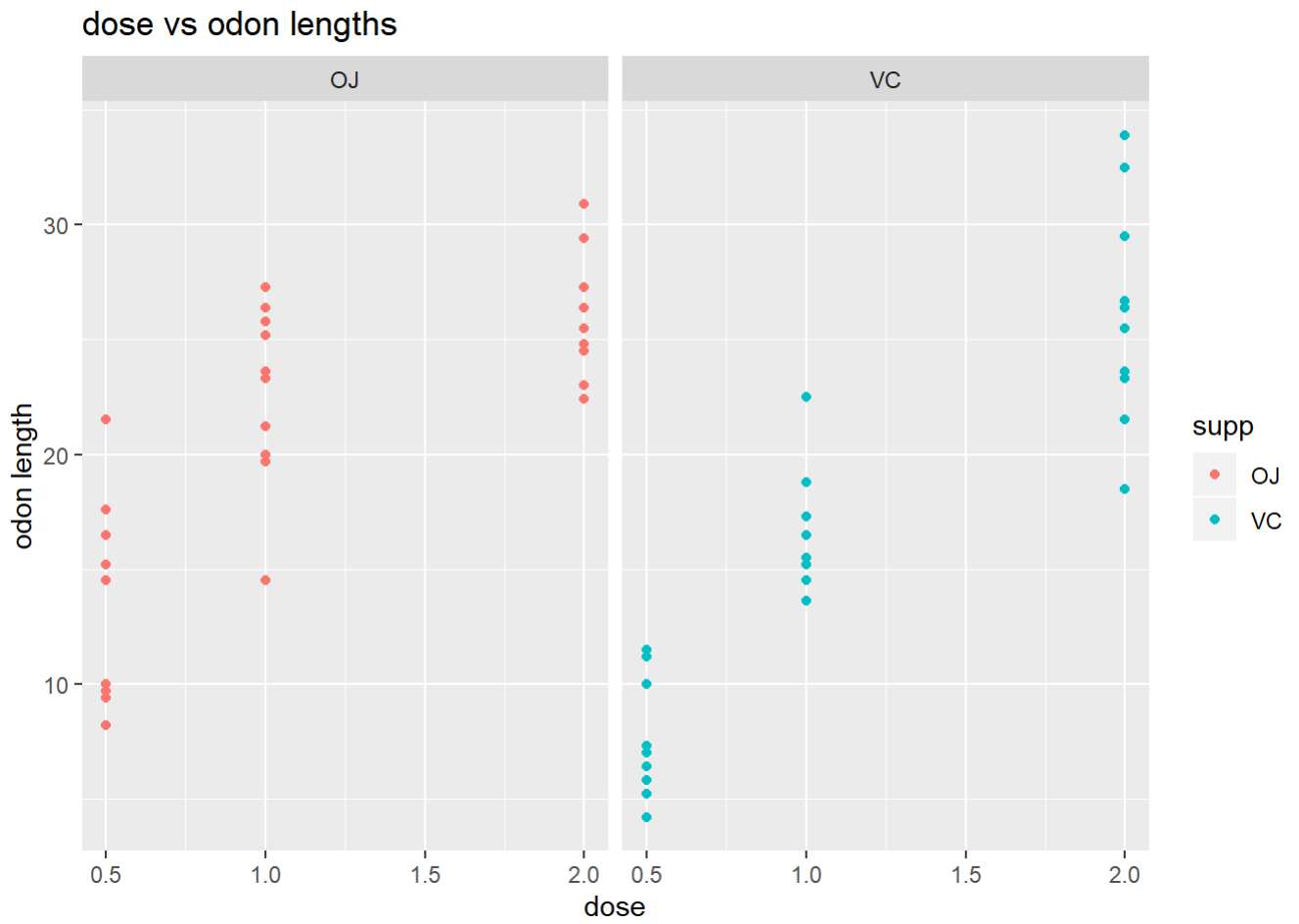



```
  #facet_wrap(~supp)


#Supplement OJ seems to be associated with greater lengths because, in the density plot, OJ's ma
ss centers more towards the greater lengths than that of VC.
```

d. Based on your output from part (c), which plot do you think is most effective for assessing whether there is a difference in distribution of lengths between the two groups? Explain your answer.

A: Density plot is the most suggestive plot for telling the distribution difference because we can directly compare and contrast the two distributions w.r.t two groups.

```
#(e) Create an appropriate scatterplot to assess the association betweeen the dose of the supple
ment and the lengths and to determine whether the nature of the association depends on the type
 of supplement. Does the association between length and dose seem to depend on the type of suppl
ement? Explain your answer.

#scatterplot
ggplot(odon_lengths, aes(x=dose,y=len,col=supp)) +
  geom_point() + labs(x="dose", y="odon length",title="dose vs odon lengths") +
  facet_wrap(~supp)
```



Regradless the supplement type, the lengths increase as the doses increase. If we drew two lines of trend, we might notice that the two very similar increasing patterns indicate that this association does NOT depend on specific suuplement types.

```
#(f) Generate a summary table that contains the mean, median, and standard deviation of the leng
ths for each supplement type.

#mySummary <- function(vector, na.rm = FALSE, round = 2){
#  results <- c('mean' = round(mean(vector),2), 'median' = round(median(vector, #na.rm), 2), 'st
d dev' = round(sd(vector, na.rm), 2))
#  return(results)
#}

#tapply(tooth_df$len, tooth_df$supp, mySummary)

ToothGrowth %>% group_by(supp) %>%
  summarize(Mean=mean(len),
            Median=median(len),
            Std=sd(len))
```

```
## # A tibble: 2 x 4
##   supp   Mean Median   Std
##   <fct> <dbl>  <dbl> <dbl>
## 1 OJ     20.7   22.7  6.61
## 2 VC     17.0   16.5  8.27
```

Question 2:

```
#(a) Read in the data directly to a tibble object from the URL (https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data) by using the read_csv() function (note: the column names are NOT included in the dataset).
library(tibble)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------------------- tidyverse 1.2.1 --
```

```
## v tidyr   1.0.0     v purrr   0.3.2
## v readr   1.3.1     v stringr 1.4.0
## v tidyr   1.0.0     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data <- read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data", col_names = FALSE)
```

```
## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_double(),
##   X6 = col_double(),
##   X7 = col_double(),
##   X8 = col_double(),
##   X9 = col_double()
## )
```

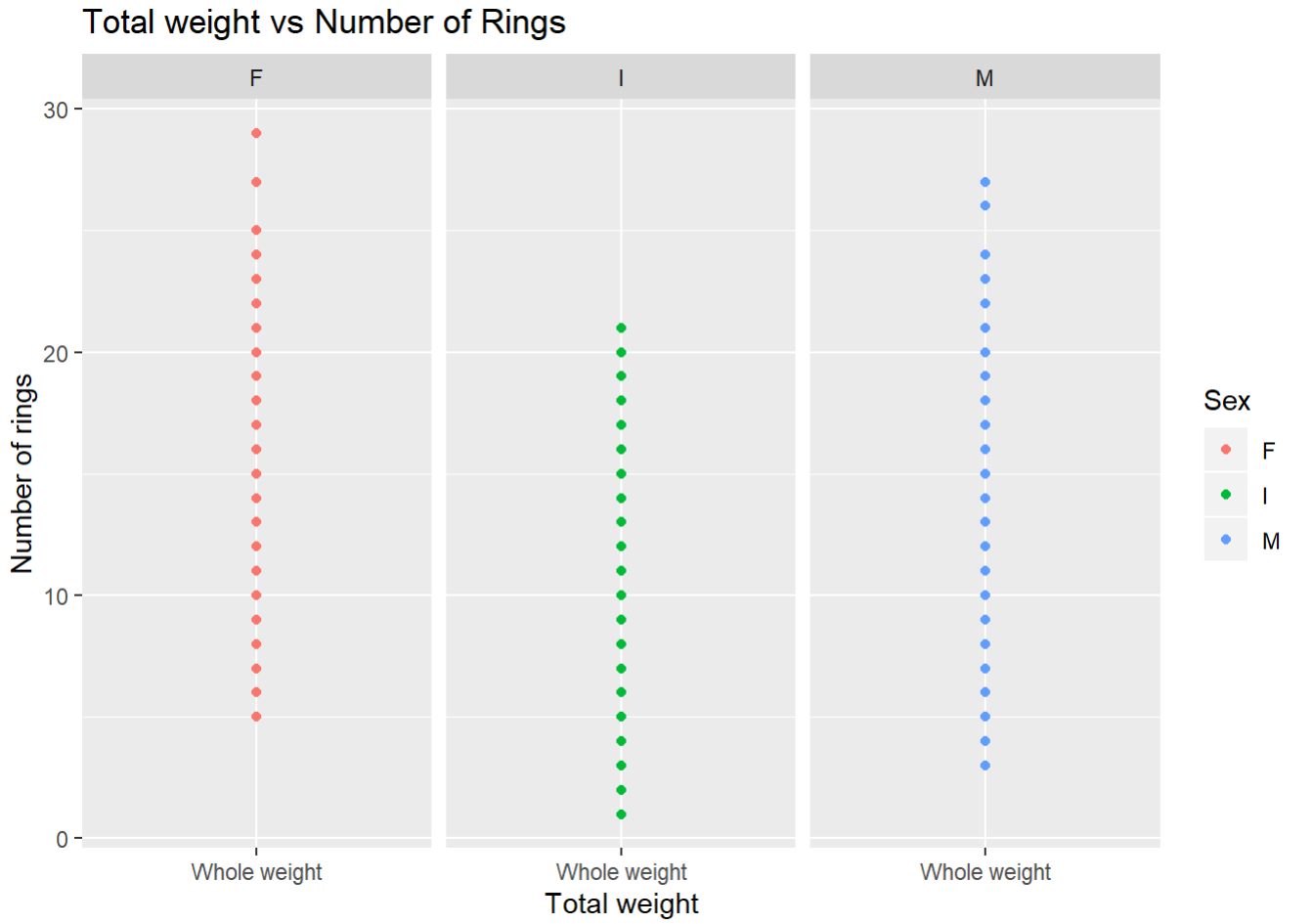
```
#(b) Assign names to the columns of the tibble. The columns are in order of the measurements given in the table above.
names(data) = c("Sex","Length","Diameter","Height","Whole weight","Shucked weight","Viscera weight","Shell weight","Rings")
#attributes(data)$names
```

```
#(c) Create a new column for the radius of the abalone shell by using the diameter.
data <- mutate(data, Radius = Diameter/2)
```

```
#(d) Find the maximum and minimum number of rings for each value of the Sex variable by using R
 functions.
data %>%
  group_by(Sex) %>%
  summarize(max_number = max(Rings), min_number = min(Rings))
```

```
## # A tibble: 3 x 3
##   Sex   max_number min_number
##   <chr>      <dbl>      <dbl>
## 1 F             29          5
## 2 I             21          1
## 3 M             27          3
```

```
#(e) Using only plots, explain whether you think the association between total weight and the nu
mber of rings depends on the value for Sex.
#scatterplot
ggplot(data, aes(x=Rings,y="Whole weight",col=Sex)) +
  geom_point() + labs(x="Number of rings", y="Total weight",title="Total weight vs Number of Rin
gs") + #coord_flip()
  facet_wrap(~Sex) + coord_flip()
```




```
# The association does depend on the Sex as we can see both males and females have greater numbe
rs of rings than infants do given a relatively fixed value of the total weight.
```

Question 3:

```
#Assume that Prof. Steele creates the following list in R to help manage his life:
shopping_list <- list(
  Grocery = list(
    Dairy = c("Milk","Cheese"),
    Meat = c("Chicken","Sausage","Bacon"),
    Spices = c("Cinnamon")
  ),
  Pharmacy = c("Soap","Toothpaste","Toilet Paper")
)
```

```
#(a) What objects (or values) are returned by the following lines of R code?

#shopping_list$Pharmacy
#[1] "Soap"         "Toothpaste"   "Toilet Paper"

#shopping_list[1][[2]]
#Error in shopping_list[1][[2]] : subscript out of bounds

#shopping_list[[1]][[3]]
#[1] "Cinnamon"


#shopping_list$Grocery[2][1]
#$Meat
#[1] "Chicken" "Sausage" "Bacon"
```

```
#(b) Using R code, show which statement yield the following three results:

shopping_list$Pharmacy
```

```
## [1] "Soap"         "Toothpaste"   "Toilet Paper"
```

```
shopping_list[2]
```

```
## $Pharmacy
## [1] "Soap"         "Toothpaste"   "Toilet Paper"
```

```
shopping_list[[1]][[2]][[2]]
```

```
## [1] "Sausage"
```