MATH 208 Assignment1

Yan Miao 260711311

2019-09-21

Load libraries and data:

```
library(ggplot2)
library(gridExtra)
library(magrittr)
library(dplyr)
library(tidyverse)
library(readr)
data(ToothGrowth)
```

Data exploration:

```
summary(ToothGrowth)
```

```
##       len        supp         dose      
##  Min.   : 4.20   OJ:30   Min.   :0.500  
##  1st Qu.:13.07   VC:30   1st Qu.:0.500  
##  Median :19.25           Median :1.000  
##  Mean   :18.81           Mean   :1.167  
##  3rd Qu.:25.27           3rd Qu.:2.000  
##  Max.   :33.90           Max.   :2.000
```

```
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

Question 1

(a)

```
mode(ToothGrowth)
```

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

```
class(ToothGrowth)
```

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

(b)

```
dim(ToothGrowth)
```

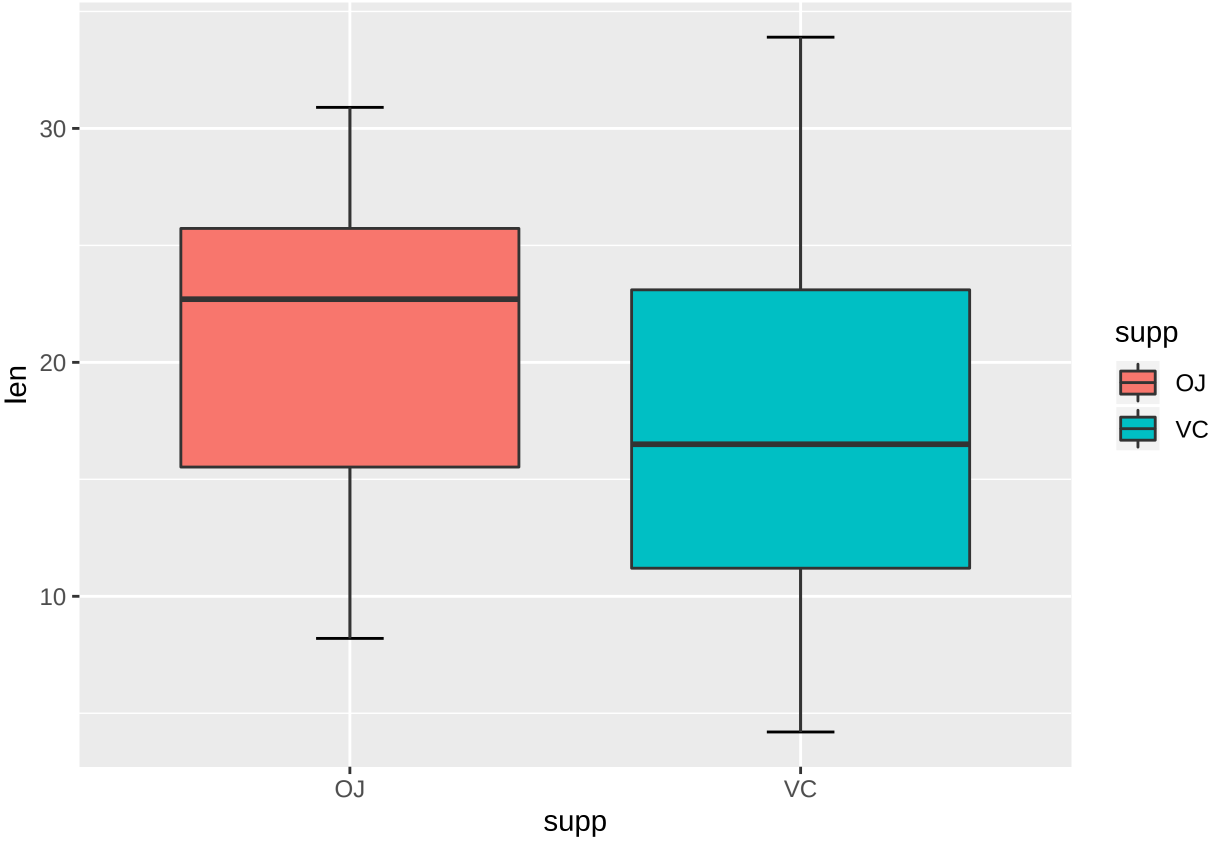
```
## [1] 60  3
```

So the number of rows is 60 and the number of columns is 3.

(c)

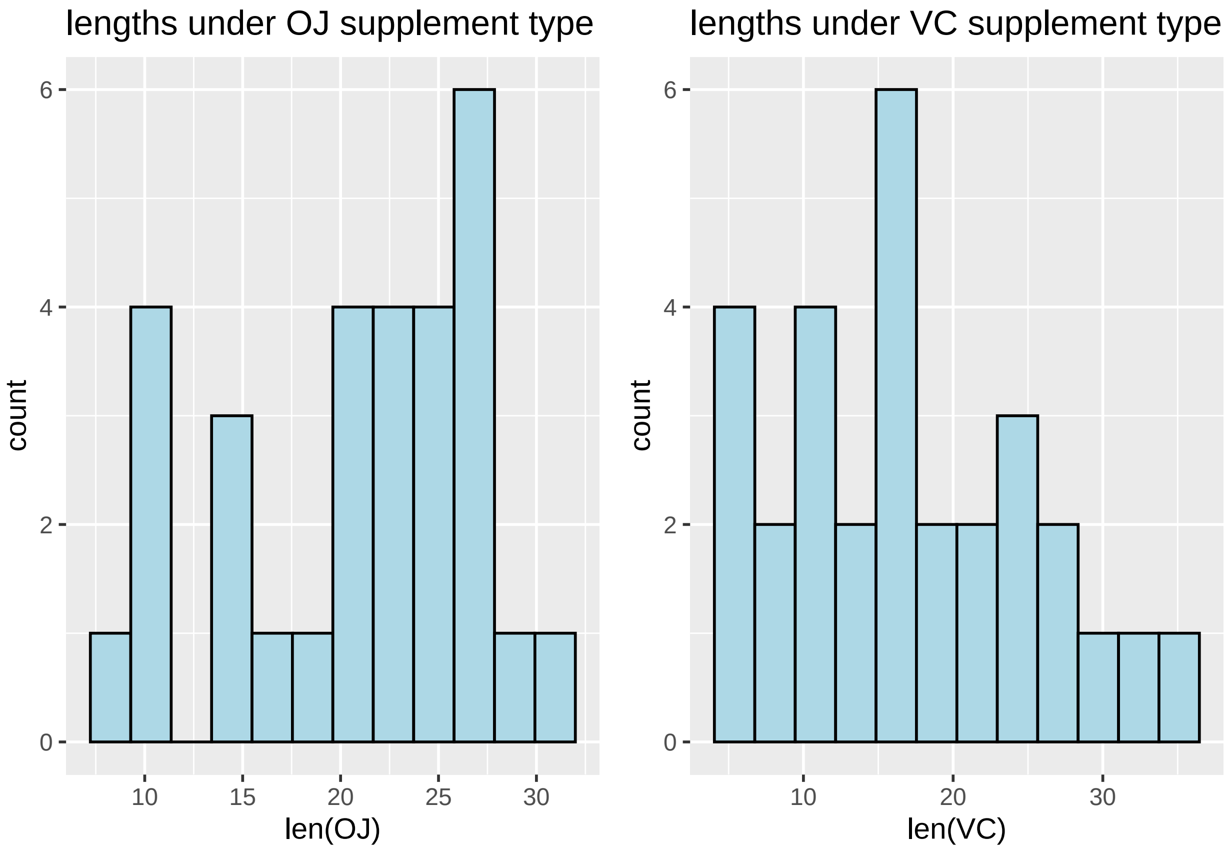
Boxplot:

```
ggplot(ToothGrowth,aes(x = supp,y = len, fill = supp)) +
  stat_boxplot(geom = 'errorbar', width = 0.15) +
  geom_boxplot() + xlab('supp') + ylab('len')
```

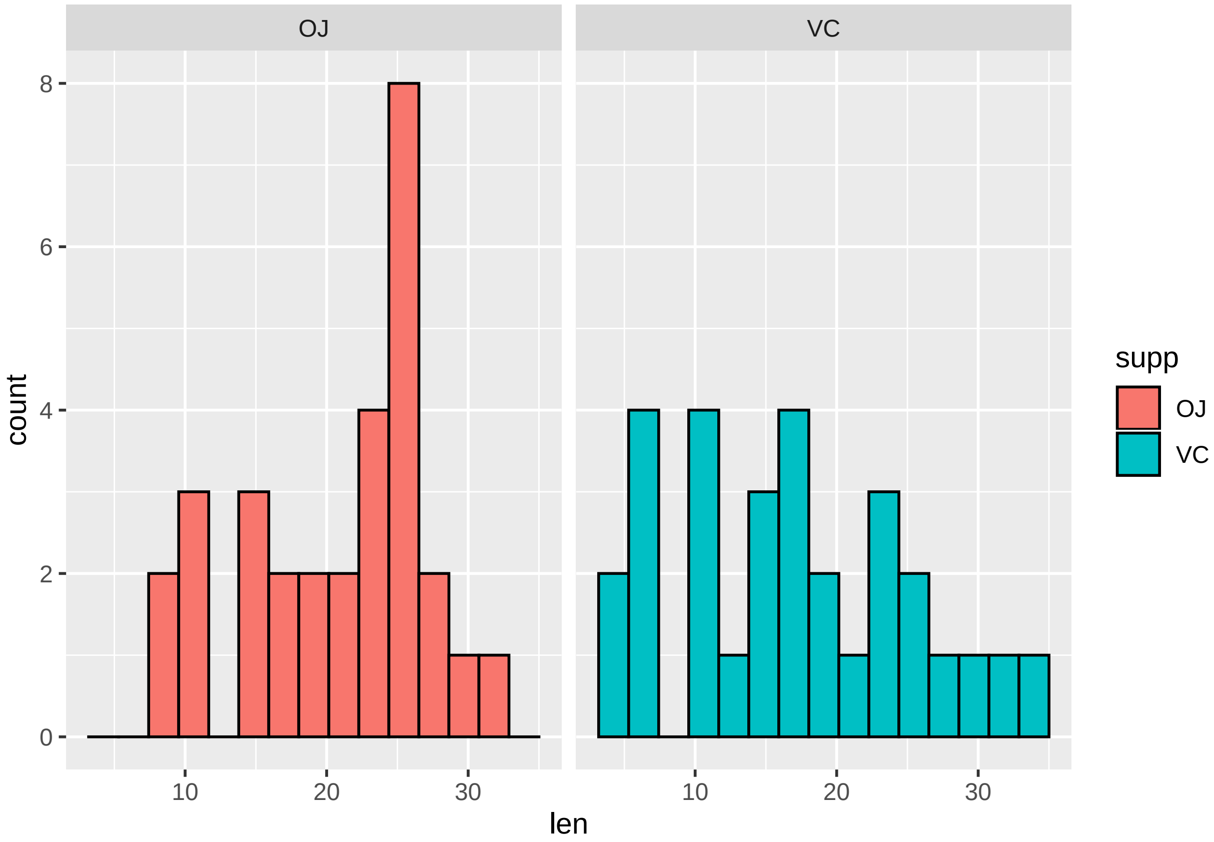


Histograms:

```
p1 = ggplot(ToothGrowth[ToothGrowth$supp %in% c('VC'),],
            aes(x = ToothGrowth[ToothGrowth$supp %in% c('VC'),]$len)) +
  geom_histogram(bins = 12, col = 'black', fill = 'lightblue') +
  xlab('len(VC)') + ggtitle("lengths under VC supplement type")
p2 = ggplot(ToothGrowth[ToothGrowth$supp %in% c('OJ'),],
            aes(x = ToothGrowth[ToothGrowth$supp %in% c('OJ'),]$len)) +
  geom_histogram(bins = 12, col = 'black', fill = 'lightblue') +
  xlab('len(OJ)') + ggtitle("lengths under OJ supplement type")
grid.arrange(p2,p1, layout_matrix = rbind(c(1,2)))
```

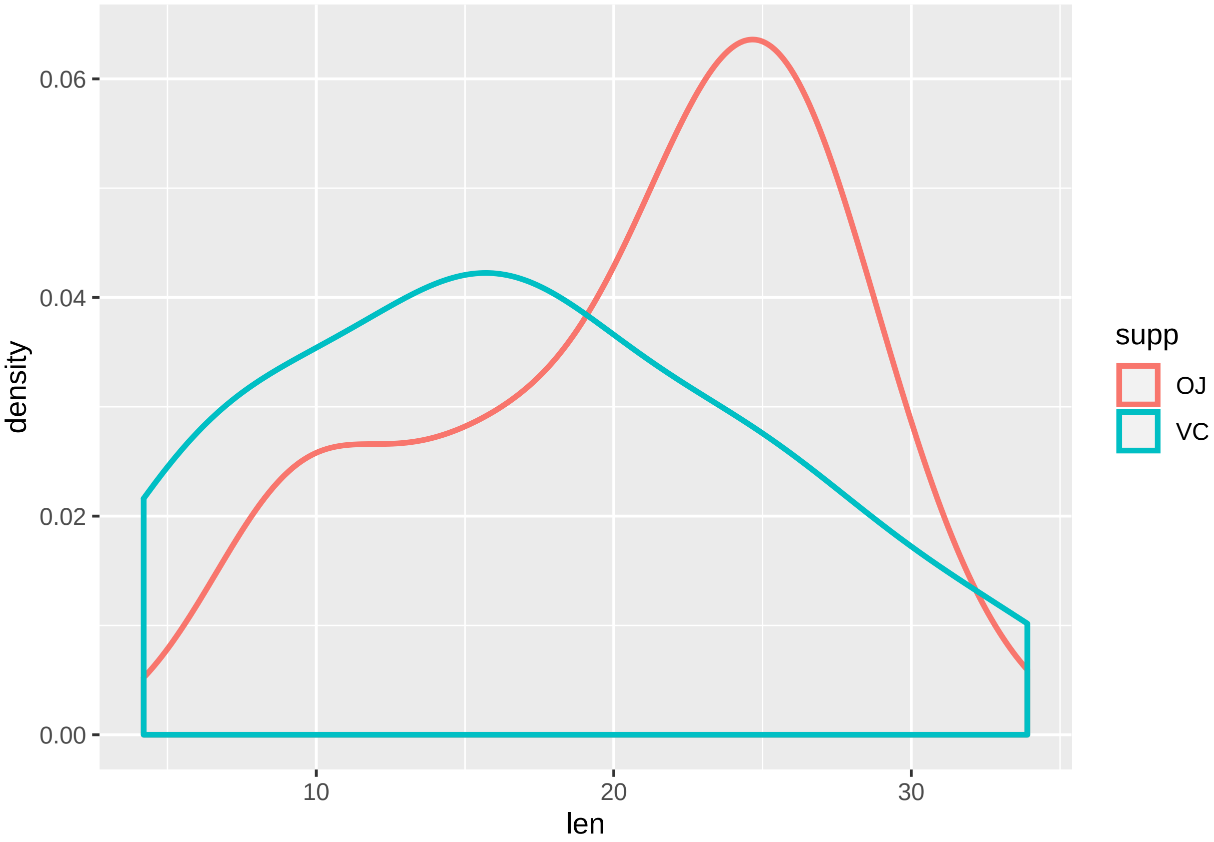


```r
ggplot(ToothGrowth, aes(x = len, group = supp, fill = supp)) +
  geom_histogram(bins = 15, col = 'black') +
  facet_wrap(~supp)
```



Density plot:

```
ggplot(ToothGrowth, aes(x = len, col = supp)) + geom_density(size = 1)
```



Based on the plots above, "OJ" seems to be associated with greater lengths.

From the boxplot, the red box representing distribution of odontoblast lengths with type "OJ" has a larger portion on the greater values compared to the blue box.

From the histograms, more bins with greater height are associated with greater lengths for the type "OJ" than the type "VC".

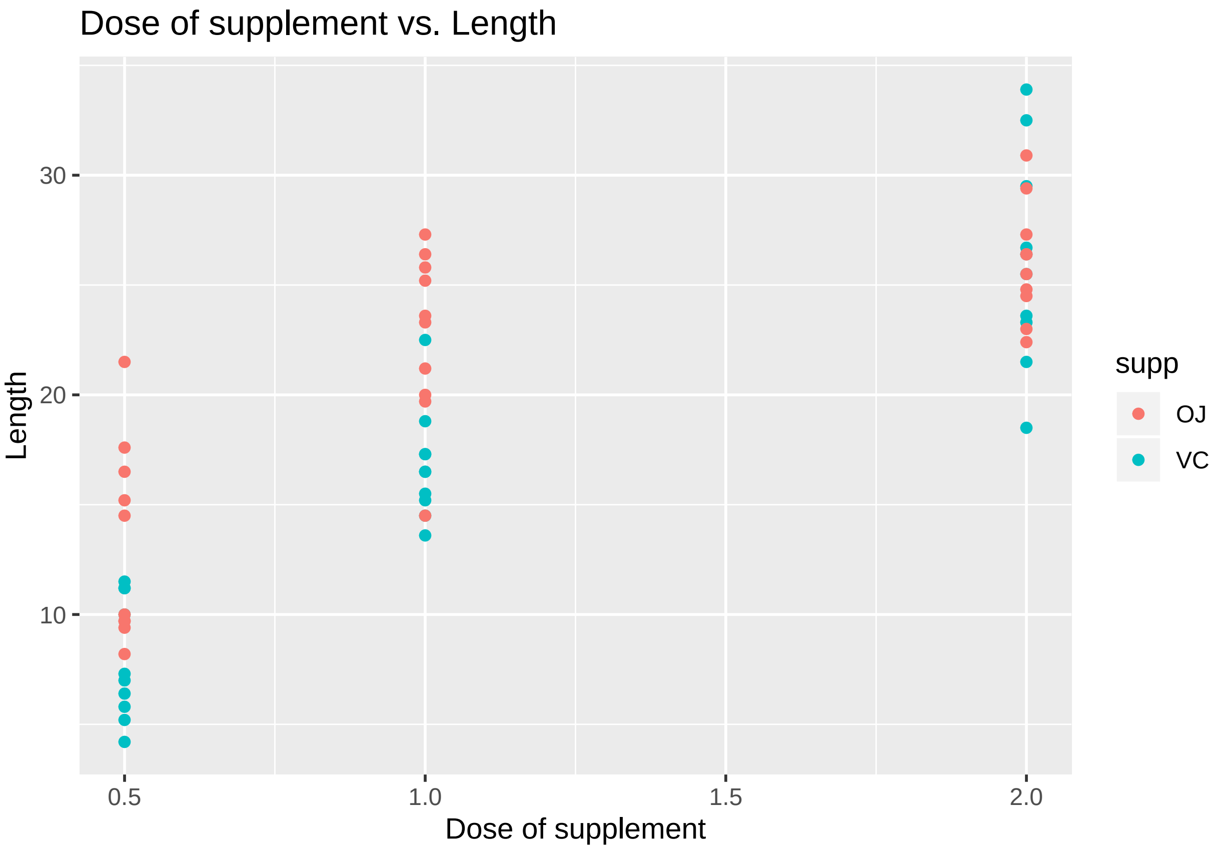
From the density plot, the peak of the red curve (type "OJ") lies on the right side of that of the blue curve (type "VC"), which also shows that "OJ" is associated with greater lengths.

(d)

To check whether there is a difference in distribution of lengths between the two groups, it is better to use density plots over the others. By plotting out the two distributions, we can clearly visualize the overlapping region of the distributions. And also since the curves are continuous, their shape can be easily determined so that we are able to tell if the difference exists immediately.

(e)

```
ggplot(ToothGrowth, aes(x = dose, y = len, col = supp)) +
  geom_point() +
  labs(x = "Dose of supplement", y = "Length", title = "Dose of supplement vs. Length")
```



Yes. From the plot above, when the dose of supplement is relatively low, supplement with type “OJ” leads to greater length in general. However, if we increase the dose to 2.0, the supplement with type “VC” can result in both greater and lower length than that with type “OJ”. That is to say, the possible consequences using a larger amount of supplement with type “VC” has a higher variation.

(f)

```
ToothGrowth %>% group_by(supp) %>% summarise(Avg = mean(len),
                                             Med = median(len),
                                             Std = sd(len))
```

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

Question 2

(a)

```
ab <- read_csv("http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data")
```

(b)

```
names(ab) <- c("sex", "length", "diameter", "height", "whole.weight",
               "shucked.weight", "viscera.weight", "shell.weight", "rings")
```

Data exploration:

```r
summary(ab)
```

```
##      sex                length         diameter          height      
##  Length:4176        Min.   :0.075   Min.   :0.0550   Min.   :0.0000  
##  Class :character   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150  
##  Mode  :character   Median :0.545   Median :0.4250   Median :0.1400  
##                     Mean   :0.524   Mean   :0.4079   Mean   :0.1395  
##                     3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650  
##                     Max.   :0.815   Max.   :0.6500   Max.   :1.1300  
##   whole.weight    shucked.weight   viscera.weight     shell.weight   
##  Min.   :0.0020   Min.   :0.0010   Min.   :0.00050   Min.   :0.0015  
##  1st Qu.:0.4415   1st Qu.:0.1860   1st Qu.:0.09337   1st Qu.:0.1300  
##  Median :0.7997   Median :0.3360   Median :0.17100   Median :0.2340  
##  Mean   :0.8288   Mean   :0.3594   Mean   :0.18061   Mean   :0.2389  
##  3rd Qu.:1.1533   3rd Qu.:0.5020   3rd Qu.:0.25300   3rd Qu.:0.3290  
##  Max.   :2.8255   Max.   :1.4880   Max.   :0.76000   Max.   :1.0050  
##      rings       
##  Min.   : 1.000  
##  1st Qu.: 8.000  
##  Median : 9.000  
##  Mean   : 9.932  
##  3rd Qu.:11.000  
##  Max.   :29.000
```

```r
head(ab)
```

```
## # A tibble: 6 x 9
##   sex   length diameter height whole.weight shucked.weight viscera.weight
##   <chr>  <dbl>    <dbl>  <dbl>        <dbl>          <dbl>          <dbl>
## 1 M      0.35     0.265  0.09         0.226         0.0995         0.0485
## 2 F      0.53     0.42   0.135        0.677         0.256          0.142 
## 3 M      0.44     0.365  0.125        0.516         0.216          0.114 
## 4 I      0.33     0.255  0.08         0.205         0.0895         0.0395
## 5 I      0.425    0.3    0.095        0.352         0.141          0.0775
## 6 F      0.53     0.415  0.15         0.778         0.237          0.142 
## # ... with 2 more variables: shell.weight <dbl>, rings <dbl>
```

(c)

```r
ab_new <- mutate(ab, radius = diameter/2)
head(ab_new)
```

```
## # A tibble: 6 x 10
##   sex   length diameter height whole.weight shucked.weight viscera.weight
##   <chr>  <dbl>    <dbl>  <dbl>        <dbl>          <dbl>          <dbl>
## 1 M      0.35     0.265  0.09         0.226         0.0995         0.0485
## 2 F      0.53     0.42   0.135        0.677         0.256          0.142 
## 3 M      0.44     0.365  0.125        0.516         0.216          0.114 
## 4 I      0.33     0.255  0.08         0.205         0.0895         0.0395
## 5 I      0.425    0.3    0.095        0.352         0.141          0.0775
## 6 F      0.53     0.415  0.15         0.778         0.237          0.142 
## # ... with 3 more variables: shell.weight <dbl>, rings <dbl>, radius <dbl>
```

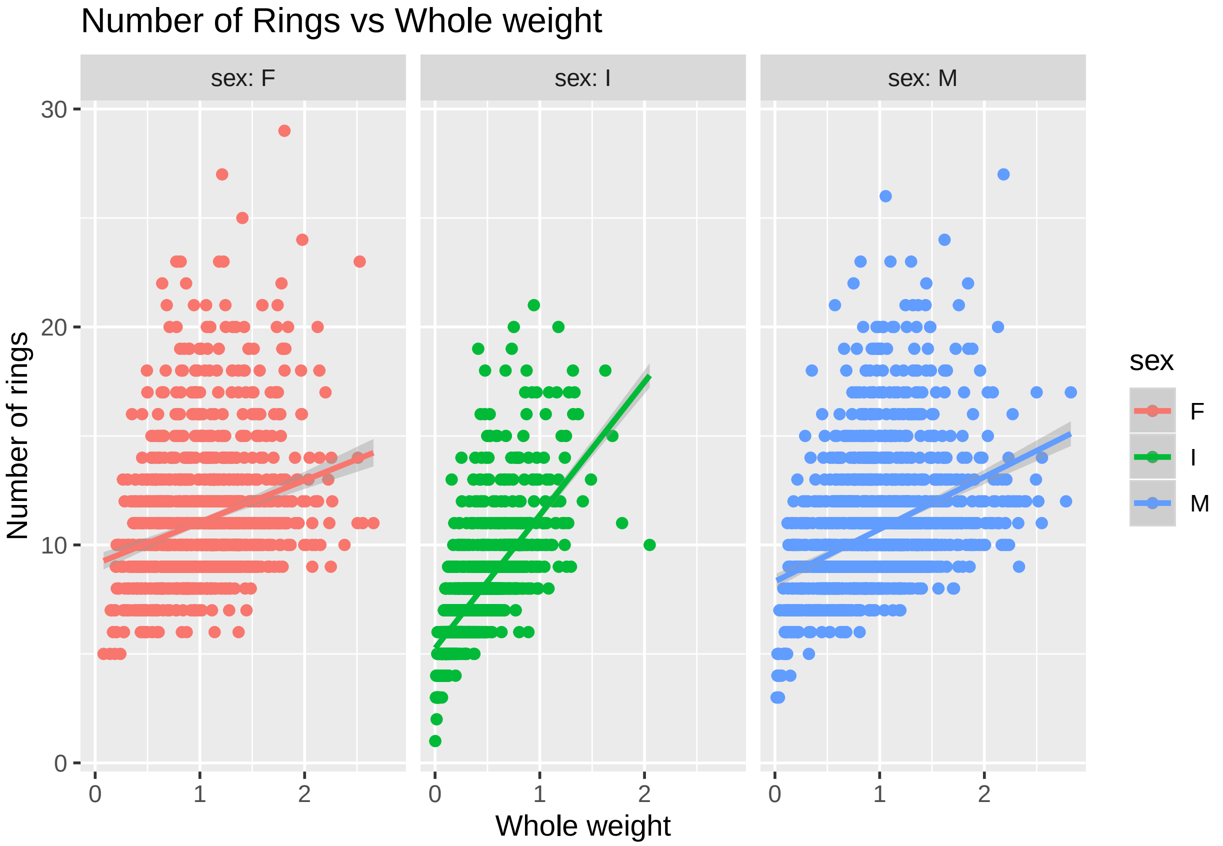
(d)

```r
ab_new %>% group_by(sex) %>% summarise(Max = max(rings), Min = min(rings))
```

```
## # A tibble: 3 x 3
##   sex      Max   Min
##   <chr> <dbl> <dbl>
## 1 F        29     5
## 2 I        21     1
## 3 M        27     3
```

(e)

```
ggplot(ab_new) + aes(x = whole.weight, y = rings, color = sex) +
  geom_point() + labs(x = "Whole weight",
    y = "Number of rings", title = "Number of Rings vs Whole weight") +
  facet_grid(. ~ sex, labeller = label_both) +
  geom_smooth(method = 'lm')
```



From the plots above, we can see that the association between total weight and the number of rings depends on the value for Sex. The increase of the number of rings is the most drastic as the total weight goes up under the sex “I” while it is the slowest under the sex “F”.

Question 3

```
shopping_list <- list( Grocery = list(
Dairy = c("Milk","Cheese"),
Meat = c("Chicken","Sausage","Bacon"),
Spices = c("Cinnamon")
),
Pharmacy = c("Soap","Toothpaste","Toilet Paper") )
```

(a)

```
shopping_list$Pharmacy
```

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

“shopping_list[1][[2]]” returns a “subscript out of bounds” error.

```
shopping_list[[1]][[3]]
```

```
## [1] "Cinnamon"
```

```
shopping_list$Grocery[2][1]
```

```
## $Meat
## [1] "Chicken" "Sausage" "Bacon"
```

(b)

```
shopping_list$Pharmacy
```

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

```
shopping_list[2]
```

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

```
shopping_list$Grocery[[2]][[2]]
```

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