Homework 2

Rosemary Kinuthia

2/20/2018

```
knitr::opts_chunk$set(echo = TRUE)
library(car)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
```

```
## ✔ ggplot2 2.2.1     ✔ purrr   0.2.4
## ✔ tibble  1.4.2     ✔ dplyr   0.7.4
## ✔ tidyr   0.8.0     ✔ stringr 1.3.0
## ✔ readr   1.1.1     ✔ forcats 0.3.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ dplyr::recode() masks car::recode()
## ✖ purrr::some()   masks car::some()
```

Question 1: What kind of R object is the Davis dataset?

```
#Load Davis dataset
dataDavis <-car::Davis
```

```
#Look at type of object
class(dataDavis)
```

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

Question 2: How many observations are in the Davis dataset?

```
#Summary of Data
summary(dataDavis)
```

```
##  sex         weight          height          repwt            repht
##  F:112   Min.   : 39.0   Min.   : 57.0   Min.   : 41.00   Min.   :148.0
##  M: 88   1st Qu.: 55.0   1st Qu.:164.0   1st Qu.: 55.00   1st Qu.:160.5
##          Median : 63.0   Median :169.5   Median : 63.00   Median :168.0
##          Mean   : 65.8   Mean   :170.0   Mean   : 65.62   Mean   :168.5
##          3rd Qu.: 74.0   3rd Qu.:177.2   3rd Qu.: 73.50   3rd Qu.:175.0
##          Max.   :166.0   Max.   :197.0   Max.   :124.00   Max.   :200.0
##                                          NA's   :17       NA's   :17
```

Question 3: For reported weight, how many observations have a missing value?

```
#Complete cases
completeObs <- complete.cases(dataDavis)
table(completeObs)
```

```
## completeObs
## FALSE  TRUE
##    19   181
```

Question 4: How many observations have no missing values?

```
#Keep only the complete data
dataComplete <- dataDavis %>%
  filter(complete.cases(.))
```

Question 5: How many females are in this subset?

```
#Create subset containing only sex
dataDavis_subset <- select(dataDavis,
                           sex)
dim(dataDavis_subset)
```

```
## [1] 200   1
```

```
#Create subset containing only females
Female_subset <- filter(dataDavis_subset,
                    sex == "F")
dim(Female_subset)
```

```
## [1] 112   1
```

Question 6: What is the average BMI for these individuals?

```
#Compute a new variable (Height in Meters)
dataDavis <-dataDavis %>%
  mutate(HeightMeters= height/100)
```

```
#Compute a new variable (Height in Meters Squared)
dataDavis <- dataDavis%>%
  mutate(HeightMeters_Squared= HeightMeters*HeightMeters)
```

```
#Create a new variable (BMI)
dataDavis <- dataDavis%>%
  mutate(BMI=weight/HeightMeters_Squared)
```

```
#Calculate average BMI (Look at summary statistics)
summary(dataDavis)
```

```
##  sex         weight          height          repwt            repht
##  F:112   Min.   : 39.0   Min.   : 57.0   Min.   : 41.00   Min.   :148.0
##  M: 88   1st Qu.: 55.0   1st Qu.:164.0   1st Qu.: 55.00   1st Qu.:160.5
##          Median : 63.0   Median :169.5   Median : 63.00   Median :168.0
##          Mean   : 65.8   Mean   :170.0   Mean   : 65.62   Mean   :168.5
##          3rd Qu.: 74.0   3rd Qu.:177.2   3rd Qu.: 73.50   3rd Qu.:175.0
##          Max.   :166.0   Max.   :197.0   Max.   :124.00   Max.   :200.0
##                                          NA's   :17       NA's   :17
##   HeightMeters   HeightMeters_Squared      BMI
##  Min.   :0.570   Min.   :0.3249       Min.   : 15.82
##  1st Qu.:1.640   1st Qu.:2.6896       1st Qu.: 20.23
##  Median :1.695   Median :2.8731       Median : 21.84
##  Mean   :1.700   Mean   :2.9050       Mean   : 24.70
##  3rd Qu.:1.772   3rd Qu.:3.1418       3rd Qu.: 23.94
##  Max.   :1.970   Max.   :3.8809       Max.   :510.93
##
```

```
#Mean BMI is 24.7
```

Question 7: How do these individuals fall into the BMI categories (what are the frequencies and relative %'s)?

```
#Recode data
dataDavis <- dataDavis %>%
  mutate(BMIcat = if_else (BMI<18.5, "1. Underweight",
                  if_else (BMI<25,
                          "2. Normal",
                  if_else (BMI<30,
                          "3. Overweight",
                          "4. Obese"),
                  "4. missing")))
```

```
#Count the number of individuals that fall into the BMI categories
dataDavis %>%
  count(BMIcat)
```

```
## # A tibble: 4 x 2
##   BMIcat              n
##   <chr>           <int>
## 1 1. Underweight    18
## 2 2. Normal        143
## 3 3. Overweight     35
## 4 4. Obese           4
```

```
#Formatted table of contents
library(janitor)
dataDavis %>%
  janitor::tabyl(BMIcat)
```

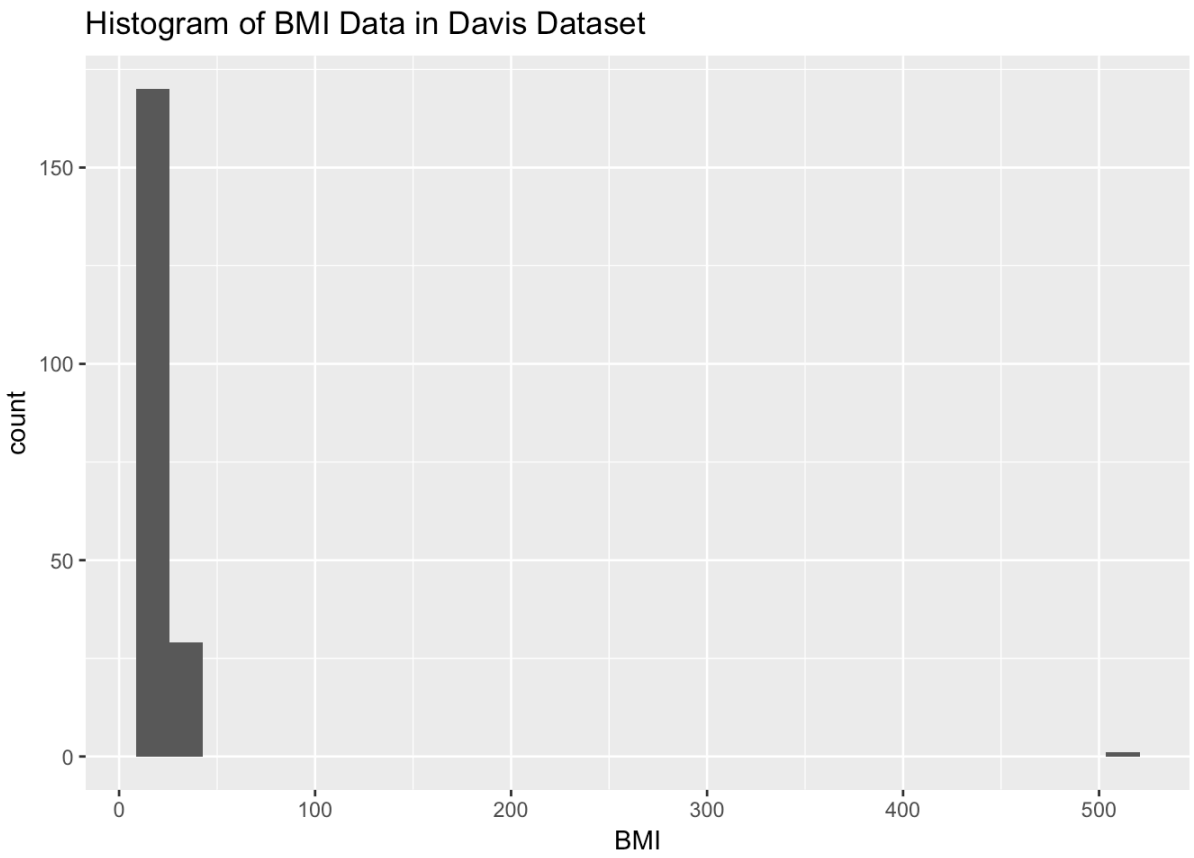
```
##           BMIcat   n percent
## 1 1. Underweight  18   0.090
## 2      2. Normal 143   0.715
## 3  3. Overweight  35   0.175
## 4       4. Obese   4   0.020
```

```
dataDavis %>%
  janitor::tabyl(BMIcat) %>%
  knitr::kable()
```

BMIcat	n	percent
1. Underweight	18	0.090
2. Normal	143	0.715
3. Overweight	35	0.175
4. Obese	4	0.020
#Question 8: Crea	te a h	istogram of BMI.

```
#Histogram of BMI
dataDavis%>%
  ggplot()+
  geom_histogram(aes(BMI))+
  ggtitle("Histogram of BMI Data in Davis Dataset")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



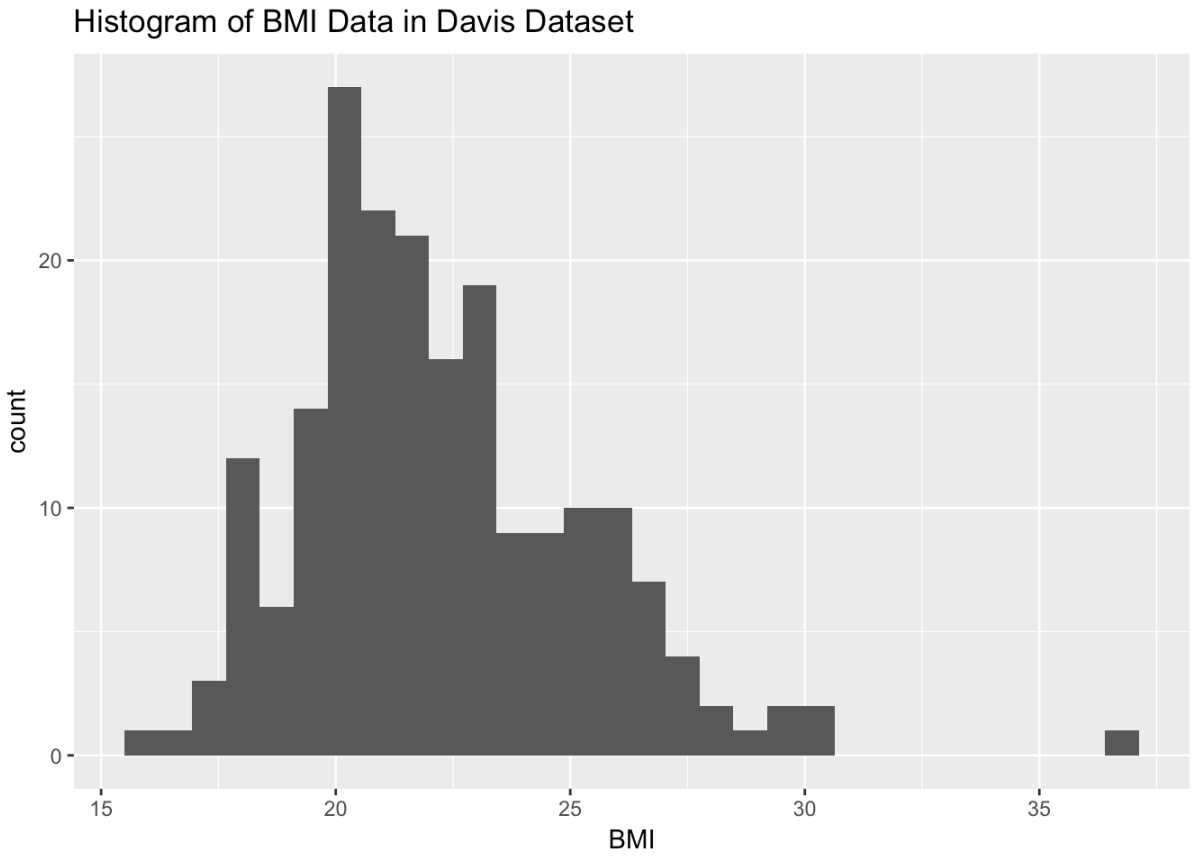
#I noticed that there is an outlier of BMI around 500. I will delete the outlier below.

```
#Remove outlier (BMI of 500)
dataDavis_OutlierDeleted <- dataDavis %>%
  filter (BMI<100)
dim(dataDavis_OutlierDeleted)
```

```
## [1] 199   9
```

```
#New histogram of BMI with outlier removed
dataDavis_OutlierDeleted%>%
  ggplot()+
  geom_histogram(aes(BMI))+
  ggtitle("Histogram of BMI Data in Davis Dataset")
```

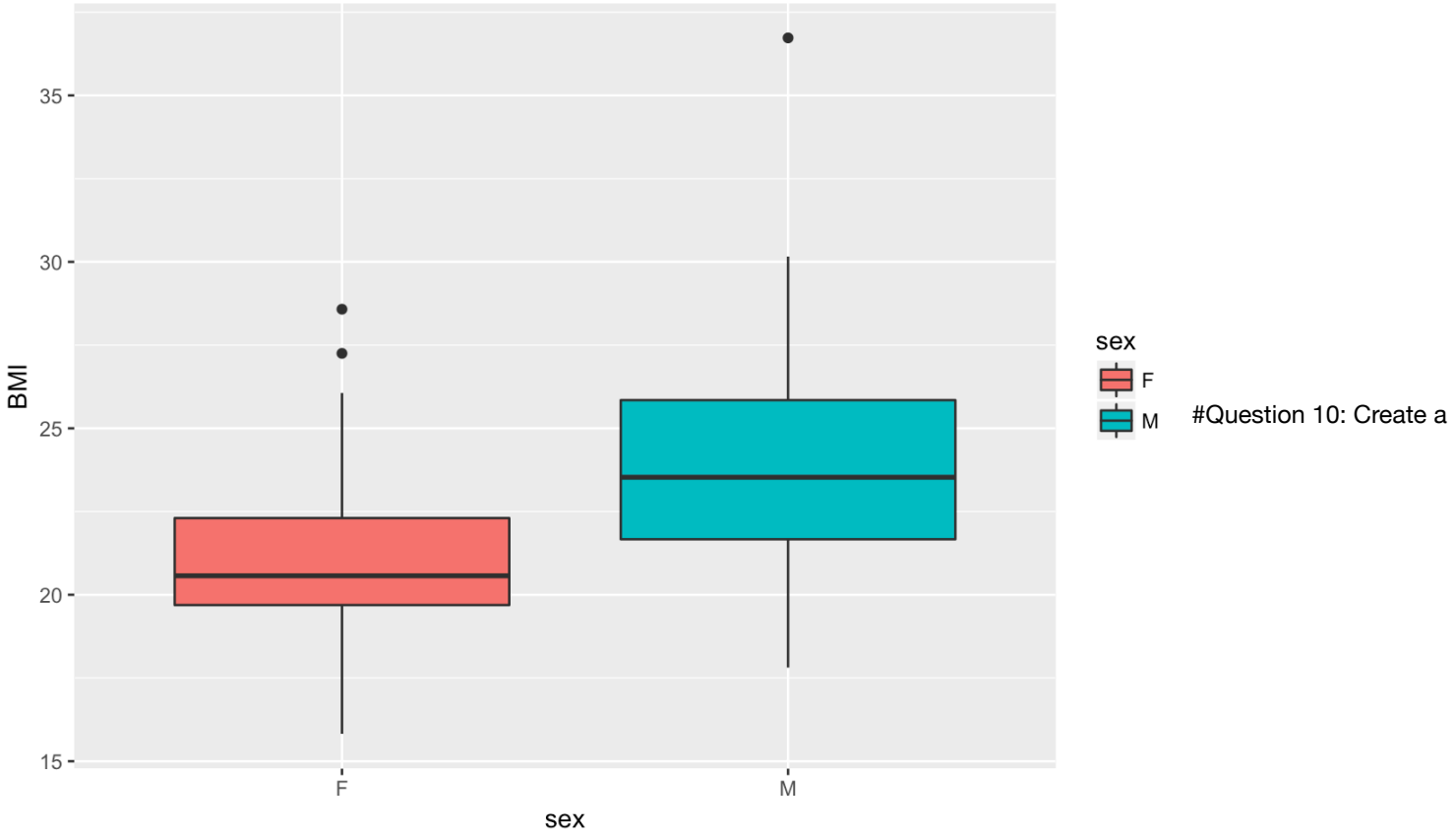
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#After deleting the outlier I noticed that the distribution is skewed to the right.
```

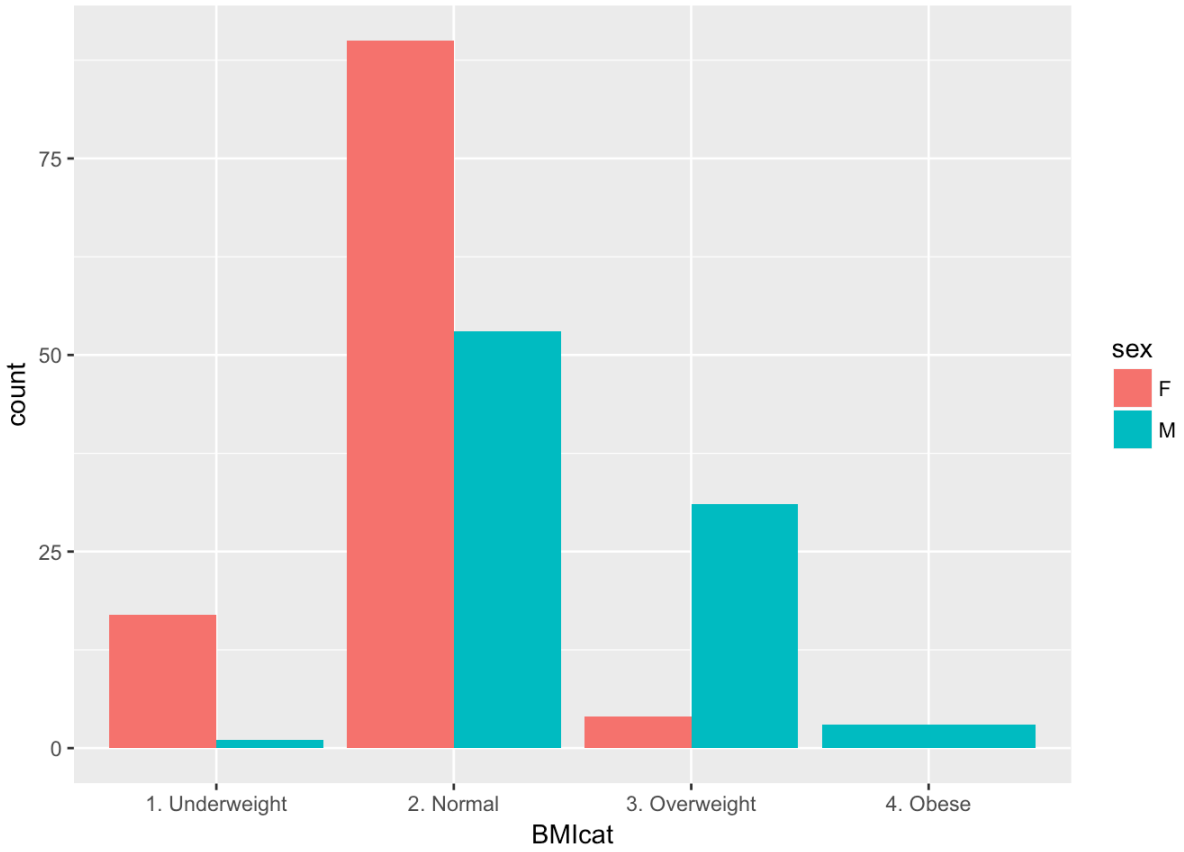
Question 9: Create side-by-side boxplots of the BMI distributions by gender

```
#Side-by-side boxplots of the BMI distribution by gender
ggplot(dataDavis_OutlierDeleted,
       aes(x=sex, y=BMI, fill=sex)) +
  geom_boxplot()
```



#Question 10: Create a clustered bar chart of the BMI categories by gender

```
#Clustered barchart of the BMI categories by gender
dataDavis_OutlierDeleted %>% ggplot(aes(x=BMIcat, fill=sex)) +
  geom_bar(position = "dodge")
```



The link to this assignment can be found at [https://github.com/RosemaryKinuthia/N741Homework2.git (https://github.com/RosemaryKinuthia/N741Homework2.git)] (https://github.com/RosemaryKinuthia/N741Homework2.git (https://github.com/RosemaryKinuthia/N741Homework2.git))