Homework_2_CARLSON

Nicole Carlson

February 19, 2018

For this homework, as with all of my work in R, I will first run the fix for my virus software trace (utils:::unpackPkgZip, edit=TRUE). Then change it to have a longer lag time (2s).

For the HW2, I will see what is running in my local environment, then add the packages I will need for this HW. I was unable to load the 'dependences=TRUE' argument for car and quantreg, so I loaded them as I would normally.

```
library(tidyverse)
library(ggplot2)
library(quantreg)
library(car)
sessionInfo()
```

```
## R version 3.4.3 (2017-11-30)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 15063)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] car_2.1-6       quantreg_5.35   SparseM_1.77    forcats_0.2.0
##  [5] stringr_1.2.0   dplyr_0.7.4     purrr_0.2.4     readr_1.1.1
##  [9] tidyr_0.7.2     tibble_1.4.2    ggplot2_2.2.1   tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] reshape2_1.4.2     splines_3.4.3      haven_1.1.1
##  [4] lattice_0.20-35    colorspace_1.3-2   htmltools_0.3.6
##  [7] mgcv_1.8-22        yaml_2.1.14        rlang_0.1.6
## [10] nloptr_1.0.4       pillar_1.1.0       foreign_0.8-69
## [13] glue_1.2.0         modelr_0.1.1       readxl_1.0.0
## [16] bindrcpp_0.2       bindr_0.1          plyr_1.8.4
## [19] MatrixModels_0.4-1 munsell_0.4.3      gtable_0.2.0
```

```
## [22] cellranger_1.1.0   rvest_0.3.2         psych_1.7.8
## [25] evaluate_0.10.1    knitr_1.17          pbkrtest_0.4-7
## [28] parallel_3.4.3     broom_0.4.3         Rcpp_0.12.13
## [31] scales_0.5.0       backports_1.1.1     jsonlite_1.5
## [34] lme4_1.1-15        mnormt_1.5-5        hms_0.3
## [37] digest_0.6.12      stringi_1.1.5       grid_3.4.3
## [40] rprojroot_1.2      cli_1.0.0           tools_3.4.3
## [43] magrittr_1.5       lazyeval_0.2.1      crayon_1.3.4
## [46] pkgconfig_2.0.1    MASS_7.3-47         Matrix_1.2-12
## [49] xml2_1.2.0         lubridate_1.7.1     minqa_1.2.4
## [52] assertthat_0.2.0   rmarkdown_1.6       httr_1.3.1
## [55] rstudioapi_0.7     R6_2.2.2            nnet_7.3-12
## [58] nlme_3.1-131       compiler_3.4.3
```

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

I need to load data Davis into my working environment from the car package.

```
data(Davis, package="car")
```

First, I'll run a few commands to see this dataset.

```
head(Davis)
```

```
##   sex weight height repwt repht
## 1   M     77    182    77   180
## 2   F     58    161    51   159
## 3   F     53    161    54   158
## 4   M     68    177    70   175
## 5   F     59    157    59   155
## 6   M     76    170    76   165
```

```
summary(Davis)
```

```
##  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
```

To get the type of R object, I run the class function. I see from this operation that Davis is a data frame in R.

```
class(Davis)
```

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

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

To answer this question, I can run the str function--it will give me the basic details of the data frame. I see from this output that Davis has 200 observations.

```
str(Davis)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ sex   : Factor w/ 2 levels "F","M": 2 1 1 2 1 2 2 2 2 2 ...
##  $ weight: int  77 58 53 68 59 76 76 69 71 65 ...
##  $ height: int  182 161 161 177 157 170 167 186 178 171 ...
##  $ repwt : int  77 51 54 70 59 76 77 73 71 64 ...
##  $ repht : int  180 159 158 175 155 165 165 180 175 170 ...
```

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

To answer this question, I will run a summary for the variable for reported weight, repwt. I see from this report that there are 17 NAs on the variable repwt in the Davis data frame.

```
Davis %>%
  select(repwt) %>%
  summary()
```

```
##      repwt      
##  Min.   : 41.00  
##  1st Qu.: 55.00  
##  Median : 63.00  
##  Mean   : 65.62  
##  3rd Qu.: 73.50  
##  Max.   :124.00  
##  NA's   :17
```

Question 4: How many observations have no missing values? (HINT: find complete cases)

To answer this question, I will create a table showing the tally of complete cases (TRUE). There are 181 complete cases in the Davis dataset.

```
completeDavis <- complete.cases(Davis)
table(completeDavis)
```

```
## completeDavis
## FALSE  TRUE 
##    19   181
```

Question 5: How many females are in this subset (create a subset containing only females)

To answer this question, I use dplyr to link my commands that R first create a new dataset, femaleDavis, with only female participants. Then, I use the summary and dim commands to show details of this new dataset. The dim command shows the number of rows and columns for the new, female-only dataset. Therefore, the number of rows=number of females in the original Davis datset=112.

```
femaleDavis <- Davis %>%
filter(sex == "F")
summary(femaleDavis)
```

```
##  sex         weight           height          repwt           repht
##  F:112   Min.   : 39.00   Min.   : 57.0   Min.   :41.00   Min.   :148.0
##  M:  0   1st Qu.: 52.75   1st Qu.:161.0   1st Qu.:53.00   1st Qu.:159.0
##          Median : 56.00   Median :165.0   Median :56.00   Median :161.0
##          Mean   : 57.87   Mean   :163.7   Mean   :56.74   Mean   :162.2
##          3rd Qu.: 62.00   3rd Qu.:169.0   3rd Qu.:61.00   3rd Qu.:165.0
##          Max.   :166.00   Max.   :178.0   Max.   :77.00   Max.   :176.0
##                                           NA's   :11      NA's   :11
```

```
dim(femaleDavis)
```

```
## [1] 112   5
```

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

I will go ahead and get rid of incomplete cases in this dataset before I proceed with the next questions involving BMI calculations.

```
dataDavisComplete <- Davis %>%
  na.omit()
```

Now, I will create a new variable, BMI, that uses existing variables of weight and height to calculate BMI.

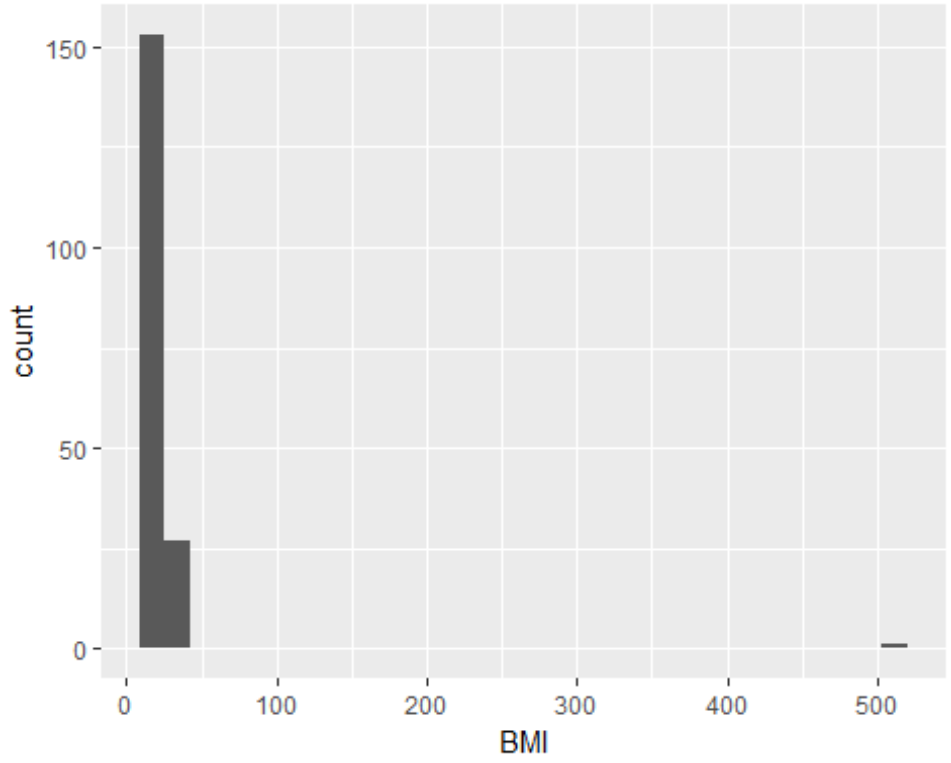
```
dataDavisComplete <- dataDavisComplete %>%
  mutate(BMI = ((weight)/((height/100)^2)))
summary(dataDavisComplete)
```

```
##  sex         weight           height          repwt           repht
##  F:99    Min.   : 39.0   Min.   : 57.0   Min.   : 41.00   Min.   :148.0
##  M:82    1st Qu.: 56.0   1st Qu.:164.0   1st Qu.: 55.00   1st Qu.:161.0
##          Median : 63.0   Median :169.0   Median : 63.00   Median :168.0
##          Mean   : 66.3   Mean   :170.2   Mean   : 65.68   Mean   :168.7
##          3rd Qu.: 75.0   3rd Qu.:178.0   3rd Qu.: 74.00   3rd Qu.:175.0
##          Max.   :166.0   Max.   :197.0   Max.   :124.00   Max.   :200.0
##        BMI
##  Min.   : 15.82
```

```
##  1st Qu.: 20.24
##  Median : 21.91
##  Mean   : 25.06
##  3rd Qu.: 24.16
##  Max.   :510.93
```

It looks like I've got an outlier BMI at 500. I'll run a quick histogram to take a look:

```
ggplot(data=dataDavisComplete) +
  geom_histogram(aes(BMI))
```



I can see on this histogram that I likely have an outlier that would need to be removed fromt the dataset in order for the average BMI to be correct.

```
dataDavisComplete %>%
  arrange(desc(BMI)) %>%
  head()
```

```
##   sex weight height repwt repht       BMI
## 1   F    166     57    56   163 510.92644
## 2   M    119    180   124   178  36.72840
## 3   M    101    183   100   180  30.15916
## 4   M    103    185   101   182  30.09496
## 5   M    102    185   107   185  29.80278
## 6   M     89    173    86   173  29.73704
```

By running the arrange function, I can see that I do have one outlier on BMI--a woman who has a reported height of 163cm, but a recorded height of 57cm. Likely a typo in the data

entry, but for the purposes of calculating the mean BMI as recorded, I will create a new dataset with this outlier deleted:

```
dataDavisNoOutComplete <- dataDavisComplete %>%
  filter(BMI < 500)
dim(dataDavisNoOutComplete)

## [1] 180   6

dim(dataDavisComplete)

## [1] 181   6
```

We can now see that the new dataset, dataDavisNoOutComplete, has one less female, and has one less row. Now, we can ask the question of the mean BMI for all individuals in the Davis men and female, with outliers and incomplete cases removed. The mean = 22.3624574

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

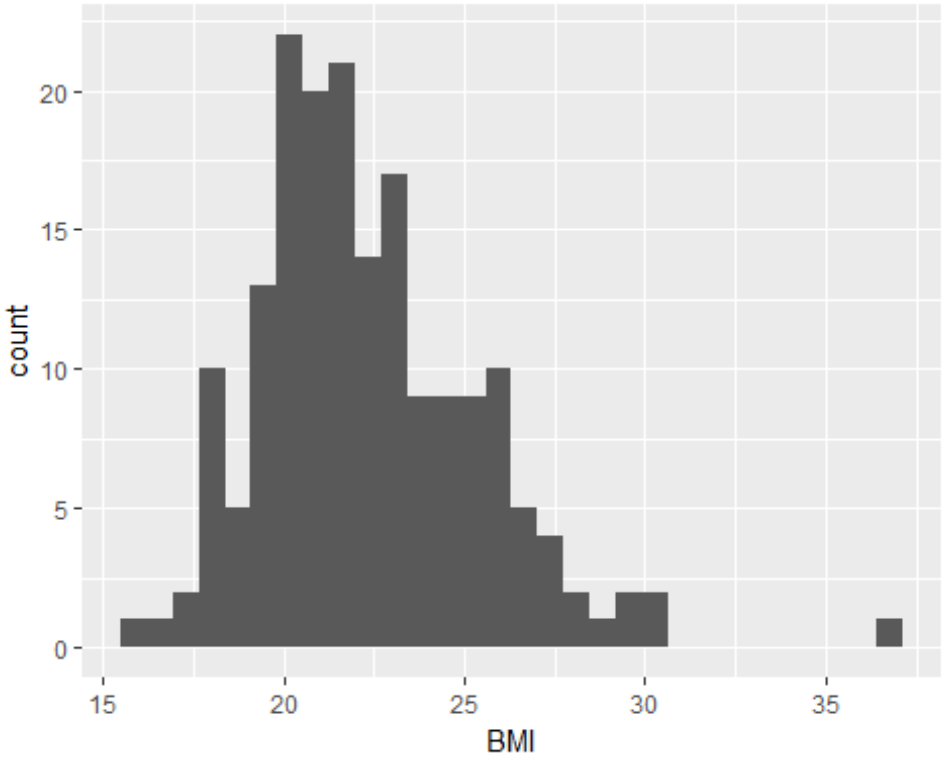
For this question, I will recode the data frame using the mutate function in dplyr to create BMI categories:

```
dataDavisNoOutComplete <- dataDavisNoOutComplete %>%
  mutate(BMIcat = cut(BMI, breaks=c(-Inf, 18.5, 25, 30, Inf),
labels=c("Underweight","Normal","Overweight", "Obese")))
 library(janitor)
dataDavisNoOutComplete %>%
  janitor::tabyl(BMIcat) %>%
  knitr::kable()
```

BMIcat	n	percent
Underweight	15	0.0833333
Normal	130	0.7222222
Overweight	32	0.1777778
Obese	3	0.0166667

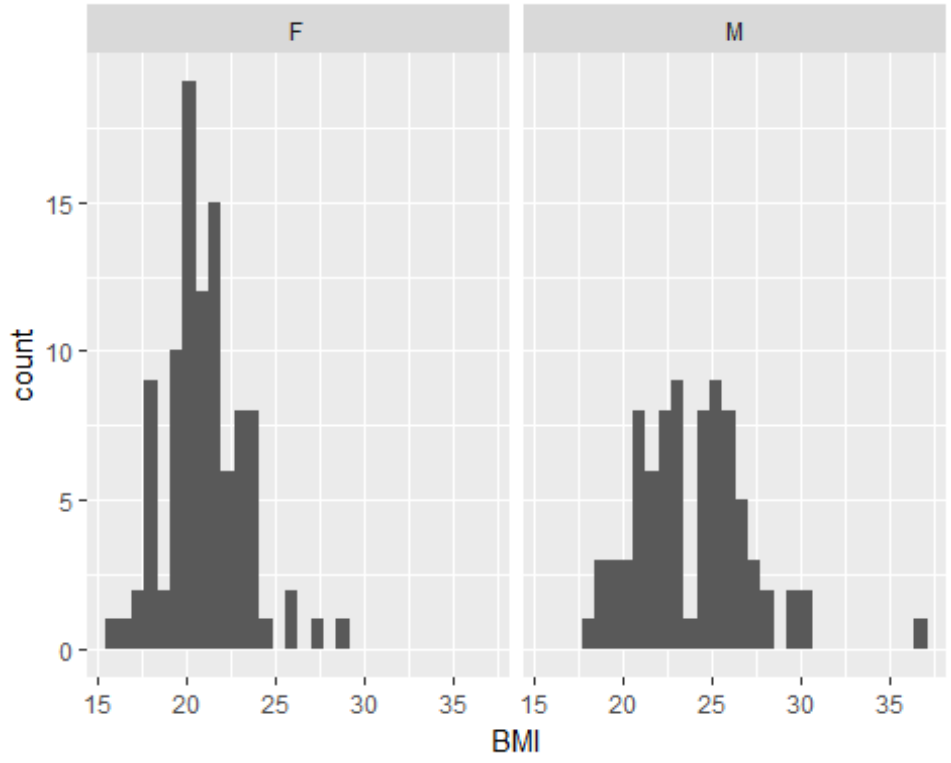
Question 8: Create a histogram of BMI.

```
dataDavisNoOutComplete %>%
  ggplot() +
  geom_histogram(aes(BMI))
```



What do you notice about the distribution (any outliers or skewness)? I notice one outlier, at BMI of around 37.

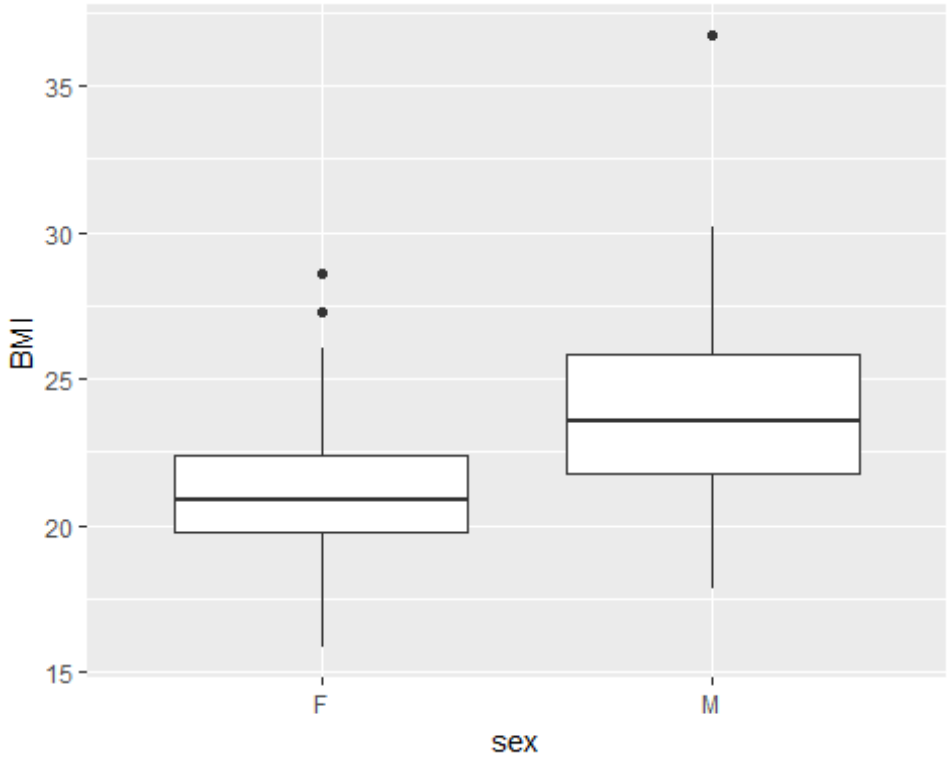
```
dataDavisNoOutComplete %>%
  ggplot() +
  geom_histogram(aes(BMI)) +
  facet_grid(. ~ sex)
```



Now that I can see the plots side by side, I can tell that the questionable outlier is a male, with BMI of around 37. There are a few women with BMIs that are higher than most other females, as well. Are any of these cases an extreme outlier, thereby needing to be removed from the dataset? Let's look at the side-by-side boxplots to see...

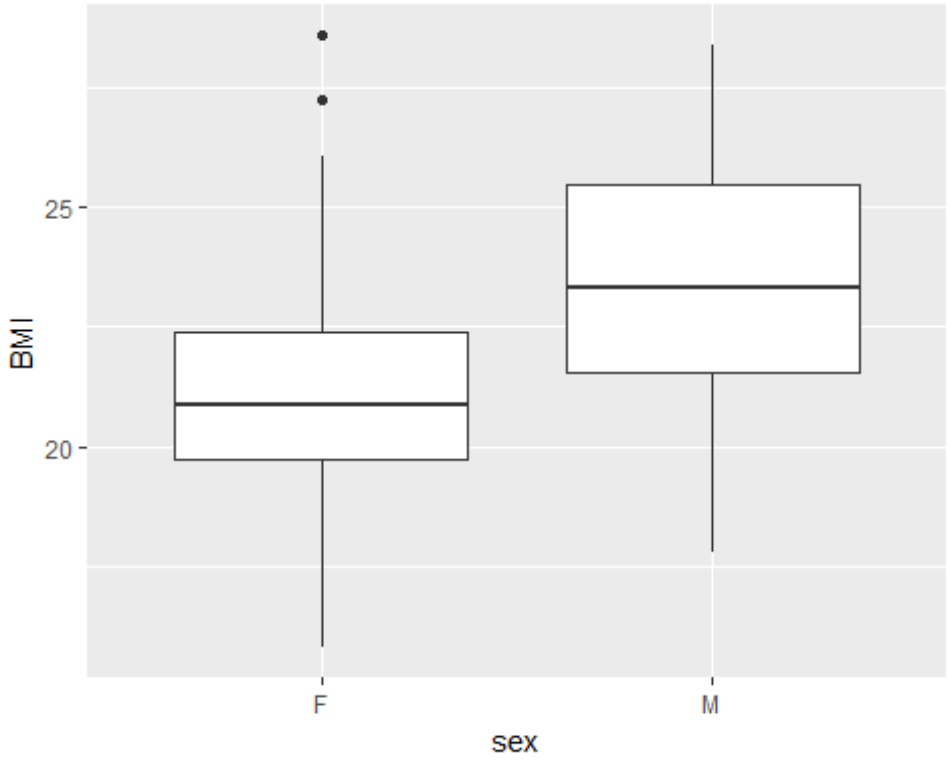
Question 9: Create side-by-side boxplots of the BMI distributions by gender

```
dataDavisNoOutComplete %>%
  ggplot() +
  aes(x=sex, y=BMI) +
geom_boxplot()
```



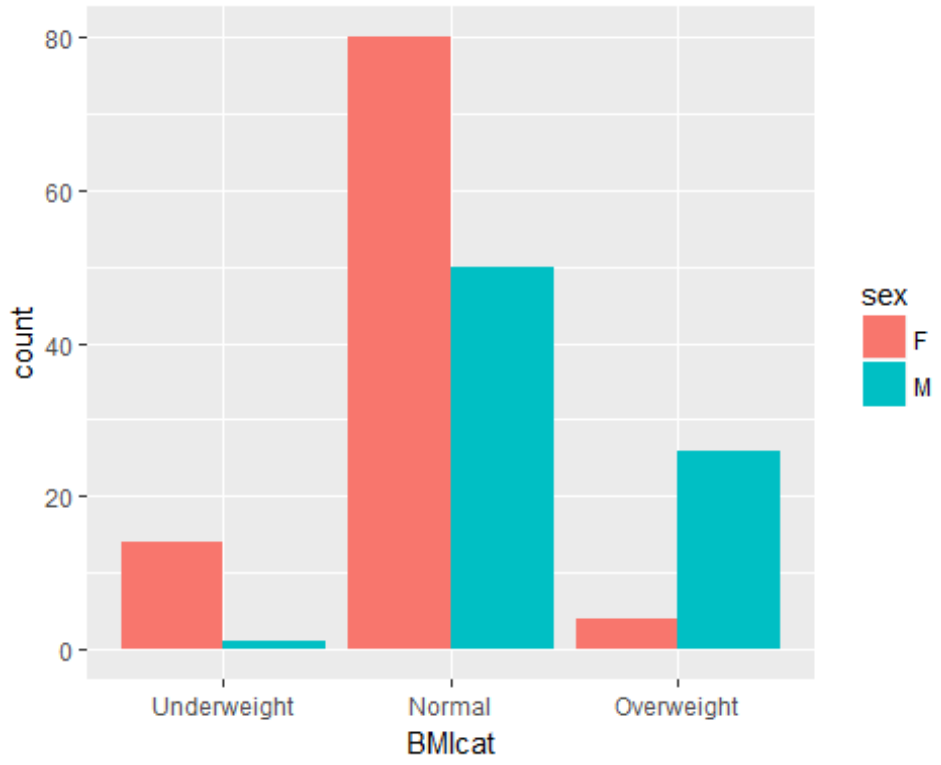
Looks like there is an extreme outlier male BMI (>Q3+1.5IQR). I'll check for outliers (using code I got from this site) with the following code, remove them, then re-plot the boxplots:

```
quantiles <- quantile(dataDavisNoOutComplete$BMI, probs = c(.25, .75))
range <- 1.5 * IQR(dataDavisNoOutComplete$BMI)
normal_Diane <- subset(dataDavisNoOutComplete,
dataDavisNoOutComplete$BMI > (quantiles[1] - range) &
dataDavisNoOutComplete$BMI < (quantiles[2] + range))
normal_Diane %>%
  ggplot() +
  aes(x=sex, y=BMI) +
geom_boxplot()
```



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

```
normal_Diane %>%
  filter(!is.na(BMI)) %>%
  ggplot() +
  aes(x=BMIcat, fill=sex) +
geom_bar(position = "dodge")
```



I don't understand why this bar graph does not show all of the BMI categories I created earlier. To check to make sure that I still have them, I will re-run the table:

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

BMIcat	n	percent
Underweight	15	0.0857143
Normal	130	0.7428571
Overweight	30	0.1714286
Obese	0	0.0000000

I understand. By getting rid of the outliers on BMI, I deleted the 3 cases who had obese BMI in this dataset. Since the obese category was 0, there was no bar plotted.
In real life, I probably would have kept these higher BMI cases unless they would have made some of my planned analyses impossible.

The git hub repository for this Homework 2 can be found at:https://github.com/nicolecarlson/N741Homework2