Homework 2

Tori Bonisese

2/21/2018

The link to this HW on github is https://github.com/vbonise/Homework2.git

Due Date is 21 February 2018

This homework is meant to further your `dplyr` and `ggplot2` skills.

First, install the package

- `car`

Installing the `car` package

We found some hiccups when we were designing this homework. With a little sleuthing, we were able to figure out that some of the issues related to installing the package and dependent package called `quantreg`. So before you install `car` use the following R commands:

- install.packages("quantreg", dependencies=TRUE)
- install.packages("car", dependencies=TRUE)

You might get this question in the console:

"Do you want to install from sources the package which needs compilation" followed by a prompt for you to respond yes or no, which looks like

`y/n:`

Usually when you see this prompt in RStudio, `y` is a good default response. However when installing `quantreg` and `car`, we found that if you answered `n` to the prompts, all will work well. *(answering y here leads to other issues you can avoid for now… we don't want you to descend into R purgatory, LOL)*

The Data - `Davis` dataset in the `car` package

The `Davis` dataset in the `car` package contains data on the measured and reported heights and weights of men and women engagedin regular exercise. *[For more information, type `?car::Davis` in the Console to bring up the HELP pages on the `Davis` dataset in the `car` package.]*

Use tools within the `dplyr` package as much as possible to answer the following questions.

```
library(car)
library(carData)

##
## Attaching package: 'carData'

## The following objects are masked from 'package:car':
##
##     Guyer, UN, Vocab

library(quantreg)

## Loading required package: SparseM

##
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
##
##     backsolve

library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

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

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

library(ggplot2)
data(Davis)
```

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

{r}class(Davis)

data frame

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

{r}str(Davis)

200 observations

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

{r}summary(Davis)

17 observations

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

```
completeObs <- complete.cases(Davis)
summary(completeObs)
```

```
##    Mode   FALSE    TRUE
## logical      19     181
```

181 obs with no missing values

Create a subset containing only females.

Question 5: *How many females are in this subset?*

```
library(dplyr)
datafemale <- Davis %>%
filter(sex == "F")
summary(datafemale)
```

```
##  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
```

OR

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

```
##  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
```

112 Females

That last question was an opportunity for you to show-off your `dplyr` confidence.

Now return to the overall dataset with both males and females.

Body mass index is one way to quantify the amount of tissue mass (muscle, fat, and bone) in an individual, then categorize that person as *underweight*, *normal weight*, *overweight*, or *obese* according to that value.

We calculate the BMI as the **ratio of the weight in kilograms divided by the square of the height in meters**, and the categorization based on BMI is as follows:

BMI Categories

Category	BMI range (kg/m2)
Underweight	<18.5
Normal	18.5 to <25
Overweight	25 to <30
Obese	30 or higher

Create the BMI variable and then a variable to depict BMI category. Note that the `height` variable is in centimeters, and `weight` is in kg. You need to create the BMI variable using the correct formula.

Now answer these questions:

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

Create Variables Needed for BMI *First get rid of missings, could impact BMI calcs*

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

bmidata <- dataComplete%>%
  mutate(height_m = (height/100)) %>%
  mutate(height_m2 = (height_m*height_m)) %>%
  mutate(bmi = (weight/height_m2)) %>%
  mutate(bmicat = if_else(bmi<18.5, "1. Underweight", if_else(bmi<25, "2.
Normal", if_else(bmi<30, "3. Overweight", "4. Obese"))))
summary(bmidata$bmi)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.82   20.24   21.91   25.06   24.16  510.93
```

Average BMI is 25.06

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

Need to convert to numeric? Now after reviewing, I do not think the numeric conversion was necessary. Tried to convert using as.numeric but resulted in many NAs.

```
bmidata2 <- bmidata %>%
mutate(bmicat2 = if_else(bmicat == "1. Underweight", 1, if_else(bmicat == "2.
Normal", 2, if_else(bmicat == "3. Overweight", 3, 4))))

bmistats <- table(bmidata2$bmicat,bmidata2$bmicat2)
bmistats

##
##                    1   2   3   4
##   1. Underweight  15   0   0   0
##   2. Normal        0 130   0   0
##   3. Overweight    0   0  32   0
##   4. Obese         0   0   0   4

prop.table(bmistats)

##
##                            1          2          3          4
##   1. Underweight 0.08287293 0.00000000 0.00000000 0.00000000
##   2. Normal      0.00000000 0.71823204 0.00000000 0.00000000
##   3. Overweight  0.00000000 0.00000000 0.17679558 0.00000000
##   4. Obese       0.00000000 0.00000000 0.00000000 0.02209945
```

15 are underweight, 130 are normal weight, 32 are overweight and 4 are obese

8.29% are underweight, 71.82% are normal weight, 17.68% are overweight and 2.21% are obese

After correction for outlier the stats are as following

```
bmidata3 <- bmidata2[-c(12),]
bmistats3 <- table (bmidata3$bmicat,bmidata3$bmicat)
bmistats3

##
##                  1. Underweight 2. Normal 3. Overweight 4. Obese
##   1. Underweight             15         0             0        0
##   2. Normal                   0       130             0        0
##   3. Overweight               0         0            32        0
##   4. Obese                    0         0             0        3

prop.table(bmistats3)

##
##                  1. Underweight  2. Normal 3. Overweight   4. Obese
##   1. Underweight     0.08333333 0.00000000    0.00000000 0.00000000
##   2. Normal          0.00000000 0.72222222    0.00000000 0.00000000
##   3. Overweight      0.00000000 0.00000000    0.17777778 0.00000000
##   4. Obese           0.00000000 0.00000000    0.00000000 0.01666667
```

15 are underweight, 130 are normal weight, 32 are overweight and 3 are obese

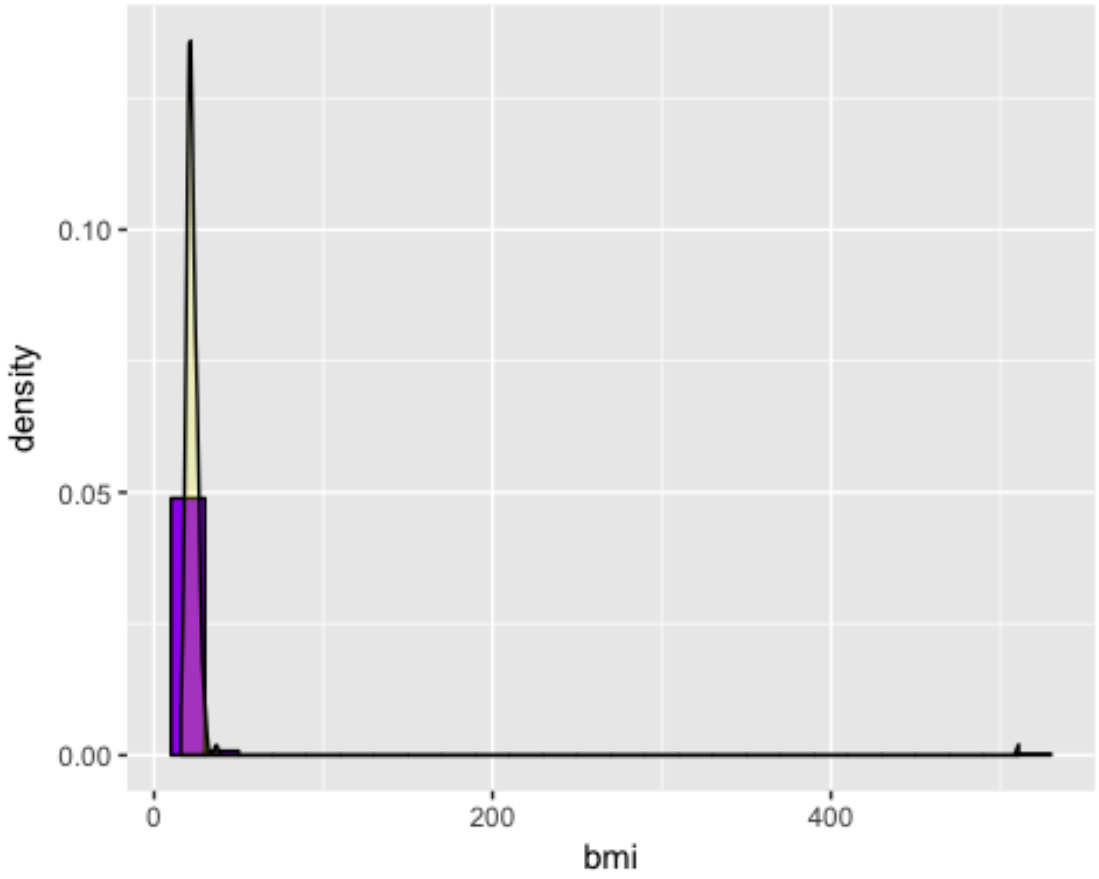
8.33% are underweight, 72.22% are normal weight, 17.78% are overweight and 1.67% are obese

Test your graphing skills using `ggplot2`

Using the `Davis` dataset from the `car` package, create the following graphics/figures using `ggplot()` and associated `geom_xxx()` functions.

Question 8: *Create a histogram of BMI.*

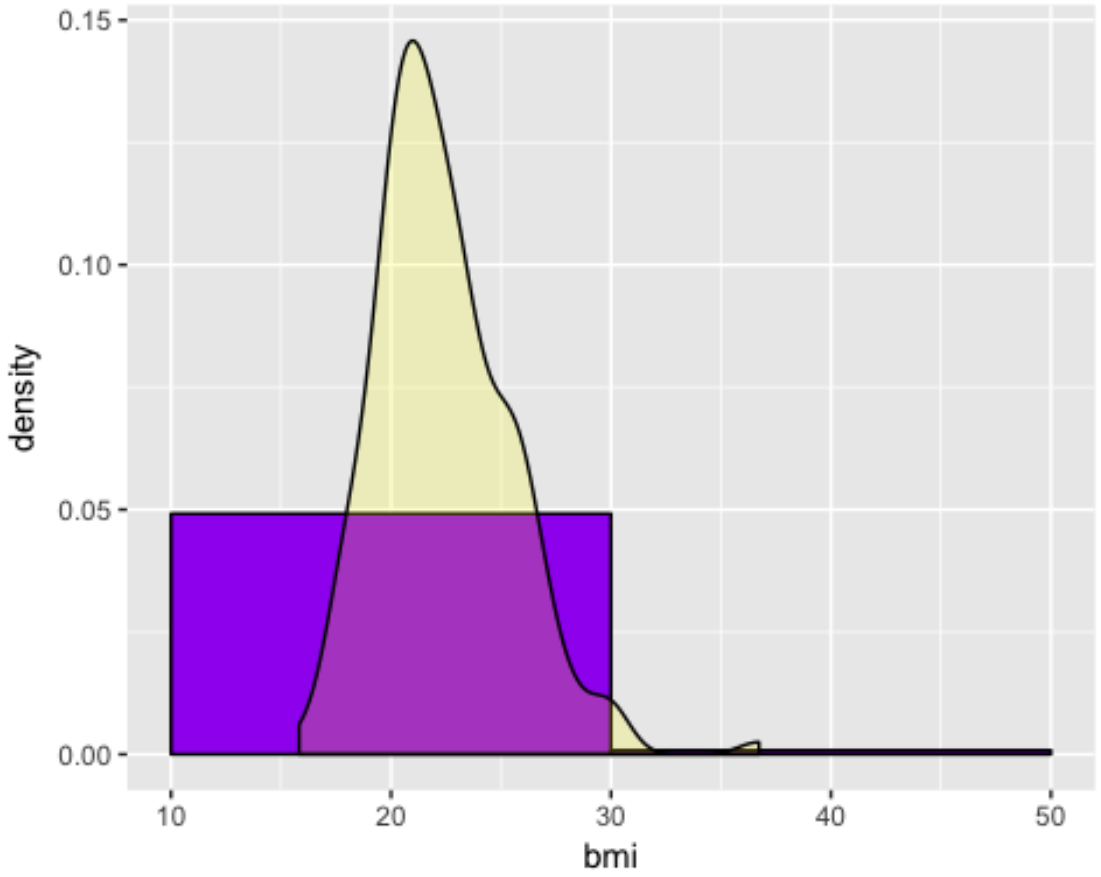
```
ggplot(bmidata2, aes(x=bmi)) +geom_histogram(aes(y=..density..), binwidth=20,
color = "black", fill = "purple") + geom_density(alpha=.2, fill = "yellow")
```



What do you notice about the distribution (any outliers or skewness)?

There is an outlier for obs 12, a BMI of 510. I removed that outlier above, in bmidata3, so will use that dataset.

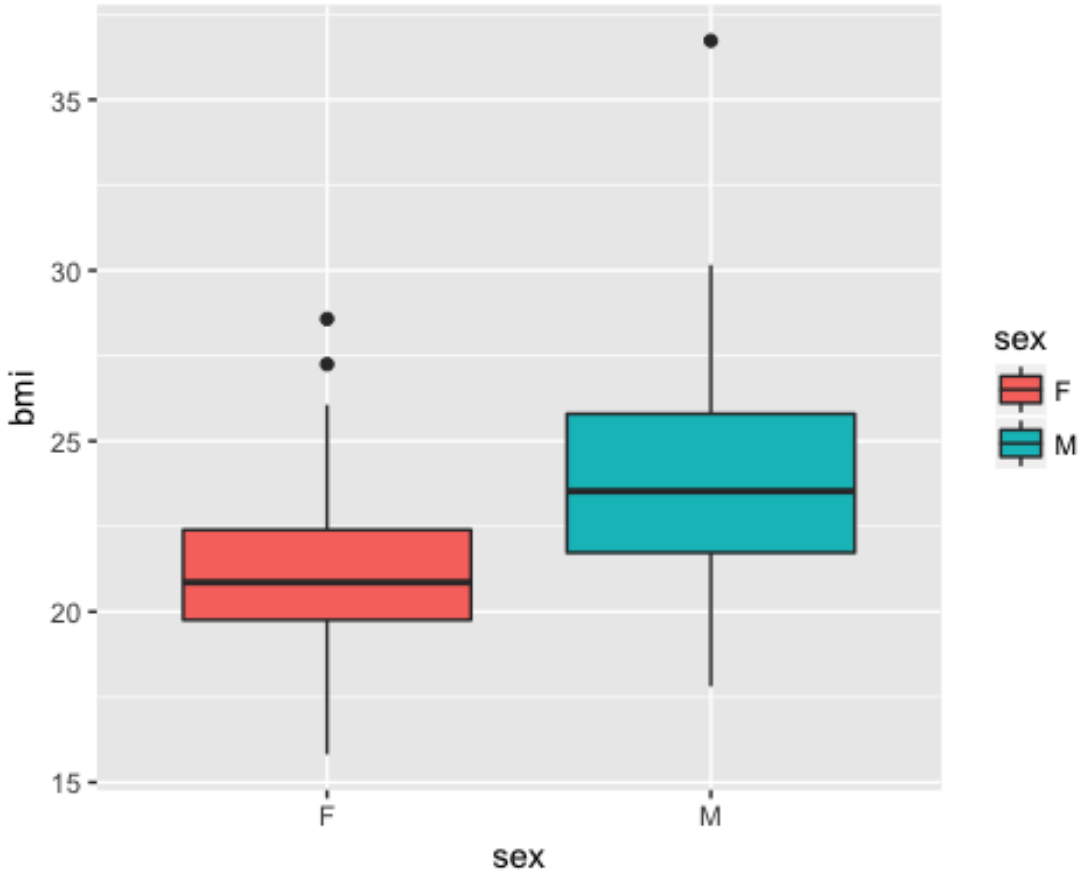
```
ggplot(bmidata3, aes(x=bmi)) +geom_histogram(aes(y=..density..), binwidth=20,
color = "black", fill = "purple") + geom_density(alpha=.2, fill = "yellow")
```



Heavy density in the middle with a slight left skew.

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

```
ggplot(bmidata3, aes(x=sex, y = bmi, fill=sex)) + geom_boxplot()
```

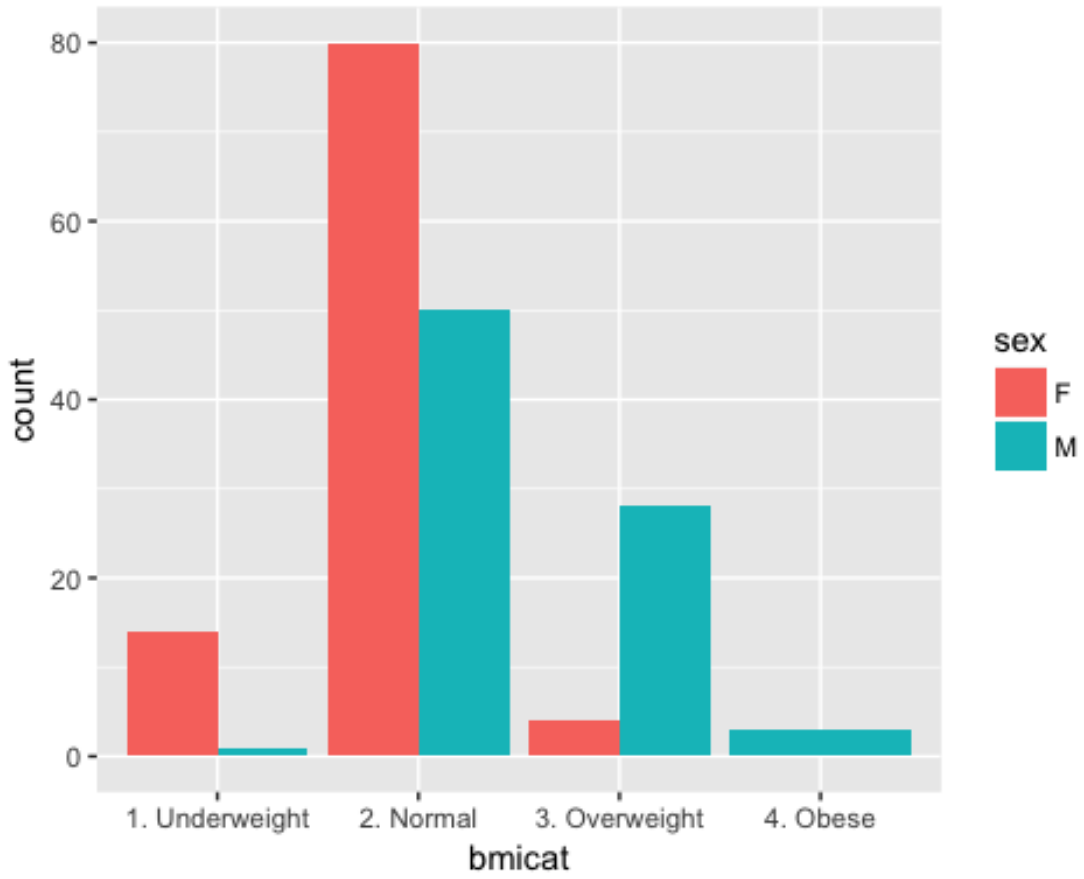


Remember to remove any outliers if needed

There are a few that could be considered outliers, but after googling BMI ranges, they are physically possible

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

```
bmidata3 %>% ggplot(aes(x=bmicat, fill=sex)) + geom_bar(position="dodge")
```



(note: the y-axis should be counts)