Homework 2 - Answer Key

Vicki Hertzberg and Melinda Higgins

2/14/2018

# *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)*

# load packages car and tidyverse  
# which loads dplyr and ggplot2  
library(car)  
library(tidyverse)

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

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

## -- Conflicts --------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::recode() masks car::recode()  
## x purrr::some() masks car::some()

## 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.

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

class(car::Davis)

## [1] "data.frame"

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

# base r approach  
dim(car::Davis)

## [1] 200 5

# dplyr approach  
car::Davis %>%  
 summarise(n = n())

## n  
## 1 200

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

# base r approach  
summary(car::Davis$repwt)

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

# dplyr approach  
# this is challenging  
car::Davis %>%  
 summarise\_all(funs(sum(is.na(.))))

## sex weight height repwt repht  
## 1 0 0 0 17 17

# purrr approach  
# not covered in class yet  
car::Davis %>% map(~ sum(is.na(.)))

## $sex  
## [1] 0  
##   
## $weight  
## [1] 0  
##   
## $height  
## [1] 0  
##   
## $repwt  
## [1] 17  
##   
## $repht  
## [1] 17

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

# Base r approach  
nomiss <- complete.cases(car::Davis)  
sum(nomiss)

## [1] 181

# dplyr approach  
davisComplete <- car::Davis %>%   
 filter(complete.cases(.))  
  
davisComplete %>%  
 summarise(n = n())

## n  
## 1 181

# all together - dplyr approach  
car::Davis %>%   
 filter(complete.cases(.)) %>%  
 summarise(n = n())

## n  
## 1 181

Create a subset containing only females.

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

# dplyr approach  
davisFonly <- car::Davis %>%  
 filter(sex == "F")  
  
davisFonly %>%   
 summarise(n = n())

## n  
## 1 112

# base r approach to get number F and M  
table(car::Davis$sex)

##   
## F M   
## 112 88

summary(car::Davis$sex)

## F M   
## 112 88

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?

# dplyr approach to compute bmi  
Davis <- car::Davis %>%  
 mutate(bmi = weight/(height/100)^2)   
  
# dplyr approach to get mean bmi  
Davis %>%  
 summarise(meanbmi = mean(bmi))

## meanbmi  
## 1 24.70096

# base r approach  
mean(Davis$bmi)

## [1] 24.70096

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

# dplyr approach for recoding  
# using mutate() and if\_else() functions  
Davis <- Davis %>%  
 mutate(bmicat = if\_else(bmi<18.5,  
 "1. underweight",  
 if\_else(bmi<25,  
 "2. normal",  
 if\_else(bmi<30,  
 "3. overweight",  
 "4. obese",  
 "missing"),  
 "missing"),  
 "missing"))  
  
# dplyr approach to get counts  
# of bmi categories  
Davis %>%   
 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

# base r approach for counts  
table(Davis$bmicat)

##   
## 1. underweight 2. normal 3. overweight 4. obese   
## 18 143 35 4

summary(as.factor(Davis$bmicat))

## 1. underweight 2. normal 3. overweight 4. obese   
## 18 143 35 4

# optional - alternate ways to get   
# frequency summary tables  
# this uses the gmodels package  
library(gmodels)  
gmodels::CrossTable(x=Davis$bmicat, y=Davis$sex)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 200   
##   
##   
## | Davis$sex   
## Davis$bmicat | F | M | Row Total |   
## ---------------|-----------|-----------|-----------|  
## 1. underweight | 17 | 1 | 18 |   
## | 4.751 | 6.046 | |   
## | 0.944 | 0.056 | 0.090 |   
## | 0.152 | 0.011 | |   
## | 0.085 | 0.005 | |   
## ---------------|-----------|-----------|-----------|  
## 2. normal | 90 | 53 | 143 |   
## | 1.229 | 1.564 | |   
## | 0.629 | 0.371 | 0.715 |   
## | 0.804 | 0.602 | |   
## | 0.450 | 0.265 | |   
## ---------------|-----------|-----------|-----------|  
## 3. overweight | 4 | 31 | 35 |   
## | 12.416 | 15.803 | |   
## | 0.114 | 0.886 | 0.175 |   
## | 0.036 | 0.352 | |   
## | 0.020 | 0.155 | |   
## ---------------|-----------|-----------|-----------|  
## 4. obese | 1 | 3 | 4 |   
## | 0.686 | 0.874 | |   
## | 0.250 | 0.750 | 0.020 |   
## | 0.009 | 0.034 | |   
## | 0.005 | 0.015 | |   
## ---------------|-----------|-----------|-----------|  
## Column Total | 112 | 88 | 200 |   
## | 0.560 | 0.440 | |   
## ---------------|-----------|-----------|-----------|  
##   
##

# optional - a nice way to get a formatted  
# table of the counts for bmi categories  
library(janitor)  
Davis %>%  
 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

# keep the janitor::tabyl output  
# and make a table using knitr::kable()  
t1 <- Davis %>%  
 janitor::tabyl(bmicat)  
  
knitr::kable(t1)

|  |  |  |
| --- | --- | --- |
| bmicat | n | percent |
| 1. underweight | 18 | 0.090 |
| 2. normal | 143 | 0.715 |
| 3. overweight | 35 | 0.175 |
| 4. obese | 4 | 0.020 |

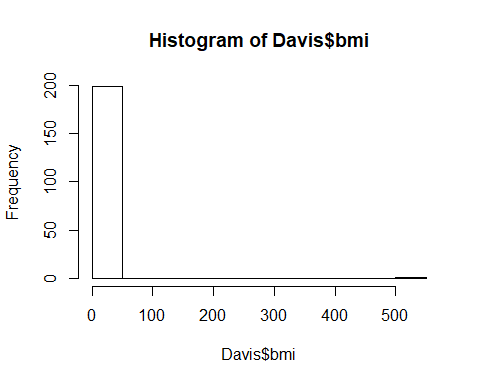
## Test your graphing skills using ggplot2

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.

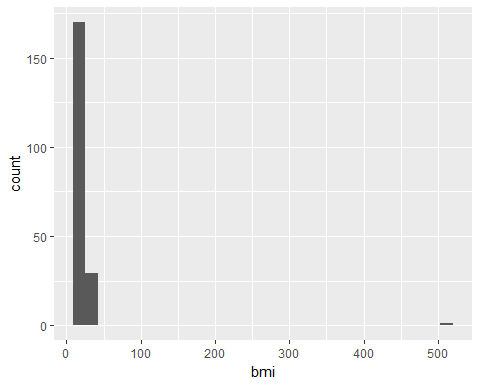
*What do you notice about the distribution (any outliers or skewness)?*

# base r  
hist(Davis$bmi)



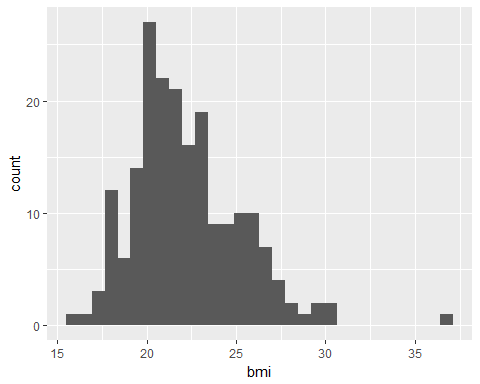
# dplyr approach  
ggplot(Davis, aes(bmi)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

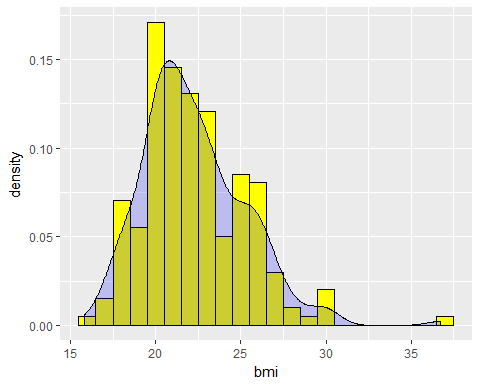


# there is an outlier - update plot or   
# data to exclude outlier in plot  
  
Davis2 <- Davis %>%  
 filter(bmi < 100)  
  
ggplot(Davis2, aes(bmi)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



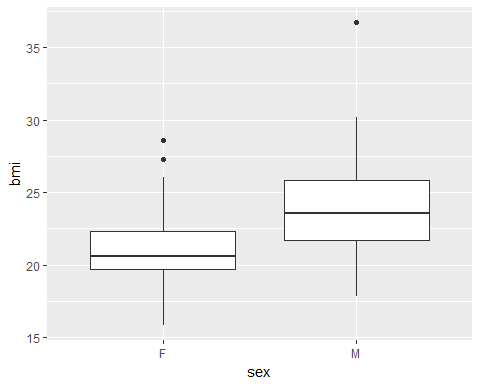
# add density curve overlay to histogram  
Davis2 %>%  
 ggplot(aes(bmi)) +  
 geom\_histogram(aes(y=..density..),  
 colour="black",fill="yellow",  
 binwidth=1) +  
 geom\_density(alpha=.2, fill="blue")



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

*Remember to remove any outliers if needed*

# boxplots of bmi by gender  
Davis2 %>% ggplot(aes(x=sex, y=bmi)) +  
 geom\_boxplot()



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

*(note: the y-axis should be counts)*

# cluster barchart of bmi categories by gender  
Davis2 %>% ggplot(aes(x=bmicat, fill=sex)) +  
 geom\_bar(position = "dodge")

