da420\_lab1\_grahn

jason grahn

9/23/2018

## Loading data

source('http://www.openintro.org/stat/data/cdc.R')

# Exercise 1:

## 1. how many cases are there in this dataset?

nrow(cdc)

## [1] 20000

## 2. How many variables in this dataset?

ncol(cdc)

## [1] 9

## 3. For each variable, identify its data type.

#but even better than running nrow() and ncol() is glimpse() because it shows both of those AND information about each of the variables  
glimpse(cdc)

## Observations: 20,000  
## Variables: 9  
## $ genhlth <fct> good, good, good, good, very good, very good, very go...  
## $ exerany <dbl> 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0,...  
## $ hlthplan <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1,...  
## $ smoke100 <dbl> 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1,...  
## $ height <dbl> 70, 64, 60, 66, 61, 64, 71, 67, 65, 70, 69, 69, 66, 7...  
## $ weight <int> 175, 125, 105, 132, 150, 114, 194, 170, 150, 180, 186...  
## $ wtdesire <int> 175, 115, 105, 124, 130, 114, 185, 160, 130, 170, 175...  
## $ age <int> 77, 33, 49, 42, 55, 55, 31, 45, 27, 44, 46, 62, 21, 6...  
## $ gender <fct> m, f, f, f, f, f, m, m, f, m, m, m, m, m, m, m, m, m,...

genhlth is categorical exerany is categorical (binary) hlthplan iscategorical (binary) smoke100 is categorical (binary) height is continuous and numerical weight is continuous and numerical wtdesire is continuous and numerical age is continuous and numerical gender is categorical

# Exercise 2

## Create a numerical summary for height and age, and compute the interquartile range for each.

#the summary of height using base-R  
summary(cdc$height)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 48.00 64.00 67.00 67.18 70.00 93.00

The IQR is 9.

#the summary of age using base-R  
summary(cdc$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 31.00 43.00 45.07 57.00 99.00

the IQR is 39.

#or if we want to gather summary stats for both at once, we can use tidyverse dplyr select with psych::describe()   
cdc %>%   
 select(height, age) %>%   
 describe(IQR = TRUE)

## vars n mean sd median trimmed mad min max range skew  
## height 1 20000 67.18 4.13 67 67.13 4.45 48 93 45 0.10  
## age 2 20000 45.07 17.19 43 44.07 19.27 18 99 81 0.45  
## kurtosis se IQR  
## height -0.38 0.03 6  
## age -0.66 0.12 26

## Compute the relative frequency distribution for gender and exerany.

table(cdc$gender)/nrow(cdc)

##   
## m f   
## 0.47845 0.52155

## How many males are in the sample?

cdc %>%   
 select(gender) %>%   
 filter(gender == "m") %>%   
 summarize(count\_of\_males = n())

## count\_of\_males  
## 1 9569

#or  
table(cdc$gender)

##   
## m f   
## 9569 10431

#which answers for both genders

## What proportion of the sample reports being in *excellent* health?

cdc %>%   
 select(genhlth) %>%   
 group\_by(genhlth) %>%   
 summarise(health\_count = n()) %>%  
 mutate(freq = health\_count / nrow(cdc) ) %>%   
 filter(genhlth == "excellent")

## # A tibble: 1 x 3  
## genhlth health\_count freq  
## <fct> <int> <dbl>  
## 1 excellent 4657 0.233

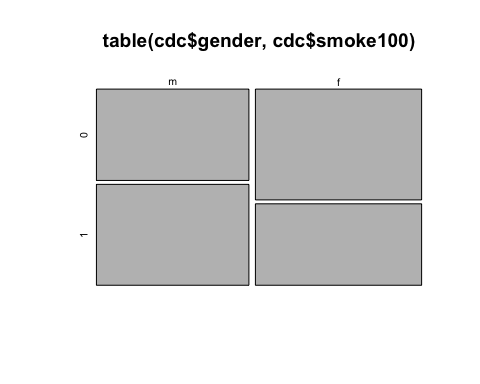
#which is maybe a little bit long winded, so in base-R we can create the same using a couple of commands  
table(cdc$genhlth)/nrow(cdc)

##   
## excellent very good good fair poor   
## 0.23285 0.34860 0.28375 0.10095 0.03385

# Exercise 3

## What does the mosiac plot reveal about smoking habits and gender?

mosaicplot(table(cdc$gender,cdc$smoke100))

 Mosaic plots aren’t great for this because the perception of sizes are hard to identify. This would be better as a clustered bar graph. That said, it appears that Men more often report having smoked at least 100 cigarettes.

# Exercise 4

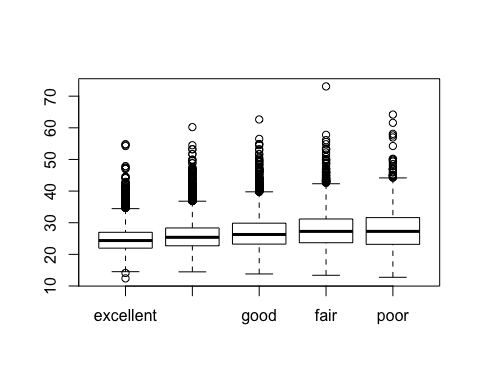
## Create a new object called under23\_and\_smoke that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime. Write the command you used to create the new object as the answer to this exercise.

under23\_and\_smoke\_tidy <-   
 cdc %>%   
 filter(age < 23,  
 smoke100 == 1)   
  
#or we could do this with base-R like   
under23\_and\_smoke <-   
 subset(cdc, age < 23 & smoke100 == 1)

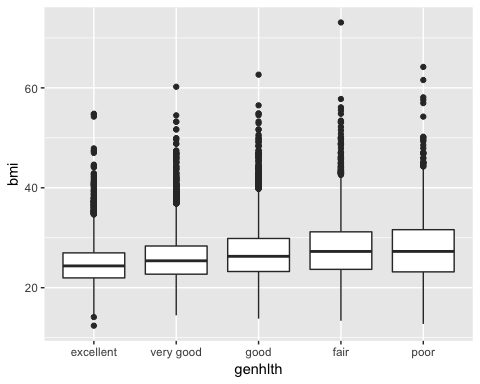
# Exercise 5

## What does this box plot show?

bmi <- (cdc$weight / cdc$height^2) \* 703  
boxplot(bmi ~ cdc$genhlth)

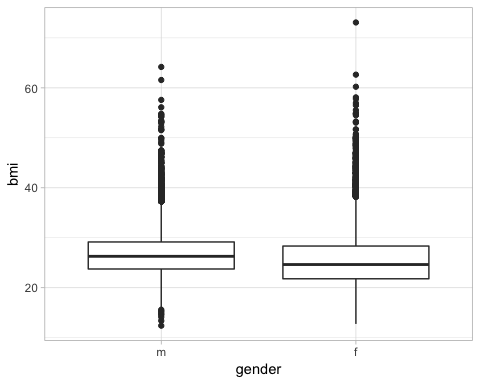


#or better, we add the BMI variable into the original dataset  
cdc %>%   
 mutate(bmi = weight / height^2 \* 703) %>%   
 #then add the plot right to the data  
 ggplot() +   
 geom\_boxplot(aes(x = genhlth, y = bmi))

 The initial boxplot shows a relationship between the general feeling of health and the distribution of those people. We see the median BMI appears to be rising as the health decreases.

## Pick another categorical variable from the data set and see how it relates to BMI. List the variable you chose, why you might think it would have a relationship to BMI, and indicate what the figure seems to suggest.

cdc %>%   
 mutate(bmi = weight / height^2 \* 703) %>%   
 #then add the plot right to the data  
 ggplot(aes(x = gender, y = bmi)) +   
 geom\_boxplot() +  
 theme\_light()



The variable I chose was Gender. I selected this on the expectation that there would be a difference between the median and spread of values of the two provided genders. When we review the boxplot, we can see yes, there is a difference between the median and spread, with males having a higher median with a smaller spread. At this level of analysis this might indicate that males, while having a higher BMI, stay more stable as BMI changes relative to age.