Lab 1

1) How many cases are there in this data set? How many variables? For each variable, identify its data type (e.g. categorical, discrete).

Cases: 20,000 observations and 9 variables.

Genhlth: categorical , exerany: categorical, hlthplan: categorical, smoke100: categorical, height: numerical discrete, weight: numerical continuous, wtdesire: numerical discrete, age: numerical continuous, and gender: categorical

2) Create a numerical summary for height and age, and compute the interquartile range for each. Compute the relative frequency distribution for gender and exerany. How many males are in the sample? What proportion of the sample reports being in excellent health?

Summary(cdc$height) = 1st quartile = 64, 3rd quartile = 70

Height IQR= 6

Summary(cdc$age) = 1st quartile = 31, 3rd quartile = 57

Age IQR= 26

Table(cdc$gender) = .478 males, .522 females, and 9569 males total

Table(cdc$exerany)/20000 = .254 no exerecise, .746 exercise

Table(cdc$genhlth)/20000 = .233 excellent health

3) What does the mosaic plot reveal about smoking habits and gender?

Males have slightly more proportion of smokers to non-smokers, while women have noticeably less smokers to nonsmokers. Males have a higher frequency of smokers than females.

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 <- subset(cdc, age < 23 & smoke100 == 1)

5) What does this box plot show? 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.

The box plot shows how BMI relates to the cases’ perceived opinion of general health of themselves. The boxplot does show that the healthier a person views themselves, the lower their Q1-Q3 is for BMI. The median rises from excellent to poor for BMI. I chose another categorical variable called exerany, which asks people if they exercised over the last month. The box plot shows that people who said they didn’t had a higher BMI median, Q1, and Q3 than those that did exercise. I believed this would have a relationship to BMI, as exercising can help people lose weight, and weight is a part of the BMI variable.

**On Your Own**

1. Make a scatterplot of weight versus desired weight. Describe the relationship between these two variables.

Plot(cdc$height, cdc$weight)

There is a slight positive trend between height and weight. The taller a person is, the stronger chance they will have a higher weight than a shorter person.

1. Let’s consider a new variable: the difference between desired weight (wtdesire) and current weight (weight). Create this new variable by subtracting the two columns in the data frame and assigning them to a new object called wdiff.

Wdiff <- (cdc$wtdesire – cdc$weight)

1. What type of data is wdiff? If an observation wdiff is 0, what does this mean about the person’s weight and desired weight. What if wdiff is positive or negative?

Wdiff is numerical discrete. If the observation is 0, then the person’s weight is not different from their desired weight. If the value is positive, then the person wants to gain weight, and if negative, then they want to lose weight.

1. Describe the distribution of wdiff in terms of its center, shape, and spread, including any plots you use. What does this tell us about how people feel about their current weight?

Summary(wdiff) = mean is -14.59, meaning people on average wanted to lose 14.59 pounds. The data is slightly left skewed when viewing a histogram, as more people feel the need to lose weight than gain weight, with a median of losing 10 pounds. The 3rd quartile is 0 pounds, so that means 75% of the data is less than 0 pounds and 75% of people want to lose more than 0 pounds. The histogram appears unimodal with a peak around 0 pounds. The variable wdiff shows that the majority want to lose weight more than they want to gain weight.

1. Using numerical summaries and a side-by-side box plot, determine if men tend to view their weight differently than women.

summary(subset(wdiff, cdc$gender == 'm'))

Min. 1st Qu. Median Mean 3rd Qu. Max.

-300.00 -20.00 -5.00 -10.71 0.00 500.00

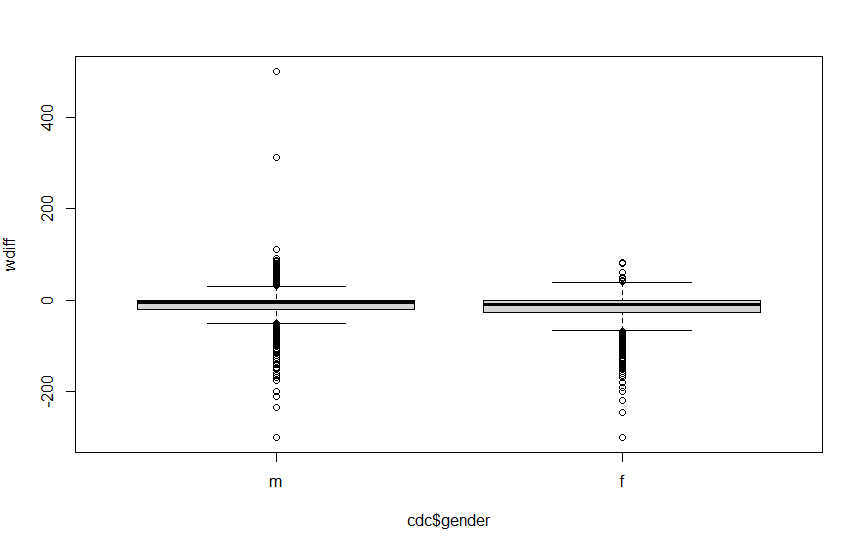
summary(subset(wdiff, cdc$gender == 'f'))

Min. 1st Qu. Median Mean 3rd Qu. Max.

-300.00 -27.00 -10.00 -18.15 0.00 83.00

According to the subset data separating the wdiff by gender, men appear to want to lose weight just like women, but they are a little less likely to want to lose as much weight as women on average. The median for men’s desire to lose weight was 5 pounds, and women wanted to lose 10 pounds from the median. The 3rd quartile for both was 0 pounds, so 75% of people wanted to lose weight or stay at the same weight for both genders. Women wanted to lose 7 more pounds on average from the 1st quartile, so 25% of women want to lose 27 pounds compared to 20 pounds for 25% of men.

boxplot(wdiff ~ cdc$gender)



The boxplot shows this relationship as well, as men have a higher median, Q1, and Q3 than women. In addition, there are more men that want to gain weight than women as shown on the top of each graph.

1. Now it’s time to get creative. Find the mean and standard deviation of weight and determine what proportion of the weights are within one standard deviation of the mean.

Mean = mean(cdc$weight) = 169.683 pounds

Standard deviation = sd(cdc$weight) = 40.08

One standard deviation means that people who fall into this category are 40.08 pounds above or below the mean. In a normal distribution, about 2/3 of people fall within one standard deviation. To calculate the actual proportion of people in this category, we would need to find the people whose weight is less than (weight plus 1 stdev) AND greater than (weight minus 1 stdev). To do this, we would have to declare a variable that utilizes the AND function we learned. I turned the mean and stdev to variables called avg and stdev, by assigning them to the formulas shown above. I made a variable called onesd and assigned it to a subset which included the entire cdc data set and included the two conditions of the AND I talked about. It looks like this:

onesd <- subset(cdc, weight < (avg + stdev) & weight > (avg-stdev))

The variable onesd shows 14,152 observations and dividing this by 20,000 total gives a proportion of 0.7076 people falling within one standard deviation.