Name: Connor Lauerman Collaborators: (Any one you worked with)

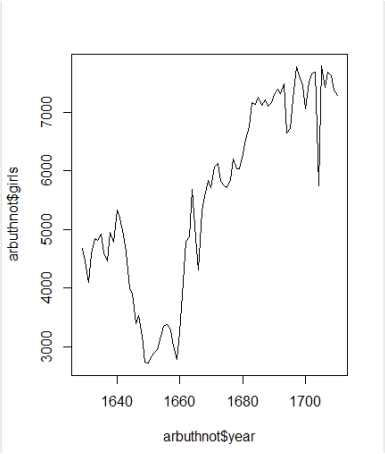
R Lab 1: Introduction to R and RStudio

Please answer all the **Exercises** and **the questions from the "On Your Own" section**. If you use any graphs or charts to justify your answer, please include them.

Exercise 1: What command would you use to extract just the counts of girls baptized?
arbuthnot$girls

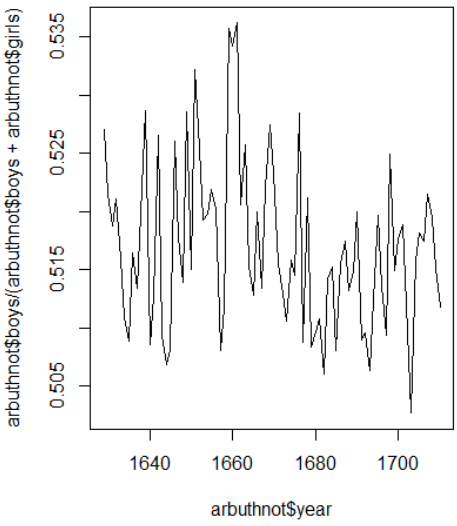
Exercise 2: Is there an apparent trend in the number of girls baptized over the years? How would you describe it? (Include plot)

The trend appears to be increasing over the years after a dip in the data



Exercise 3: Now, make a plot of the proportion of boys over time. What do you see? (Include plot)

The proportion oscillates from a high point to a low point over a few years with an overall trend of decreasing towards .5 but in total are more than .5



On Your Own:

1) What years are included in this data set? What are the dimensions of the data frame and what are the variable or column names?

Years = 1940-2002

Dimensions: [1] 63 rows 3 columns

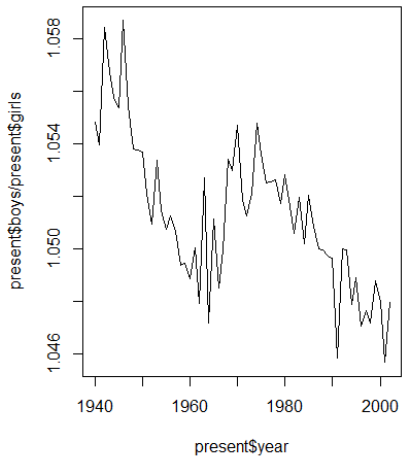
Variables: Years, boys, girls

2) How do these counts compare to Arbuthnot's? Are they on a similar scale?

The present counts are about 100 times larger than Arbuthnot's count

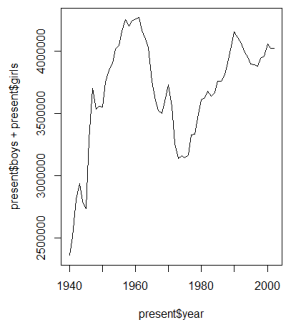
3) Make a plot that displays the boy-to-girl ratio for every year in the data set. What do you see? Does Arbuthnot's observation about boys being born in greater proportion than girls hold up in the U.S.? Include the plot in your response.

Arbuthnot's observation does hold true by a small amount since For every girl there is more than 1 boy born



4) In what year did we see the most total number of births in the U.S.?

In 1961 the us saw 4268326 births



R Lab 2: Introduction to data

Please answer all the Exercises and the questions from the "On Your Own" section. If you use any graphs or charts to justify your answer, please include them.

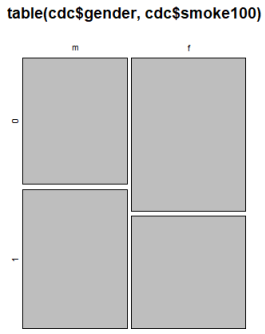
Exercise 1: How many cases are there in this data set? How many variables? For each variable, identify its data type (e.g. categorical, discrete).
Cases: There are 20000 with 9 variables.
Genhealth: Qualitative variable
Exerany: Qualitative variable
Hlthplan: Qualitative variable
Smoke100: Qualitative variable
Height: Quantitative continuous
Weight: Quantitative continuous
Wtdesire: Quantitative continuous
Age: Quantitative discrete
Gender: Qualitative variable

Exercise 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?

IQR for height: 70-64=6
IQR for age: 57-31=26
Relative frequency male: .47845 or 47.8%
Relative frequency female: .52155 or 52.2%
Relative frequency yes exercise: .7457 or 74.57%
Relative frequency no exercise: .2543 or 25.43%
# of males: 9569 males
.23285 of the sample reported being in excellent health

Exercise 3: What does the mosaic plot reveal about smoking habits and gender? (Include plot)

More males in the sample have smoked more than 100 cigarettes than hadn't while more females had smoke less than 100 cigarettes than had. This statistics suggest males are more likely to smoke than females

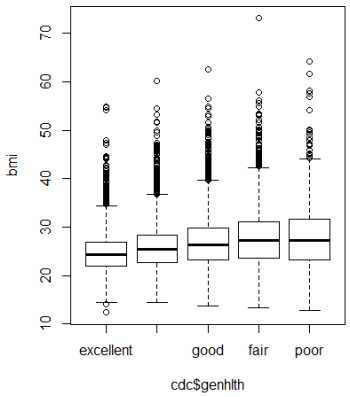


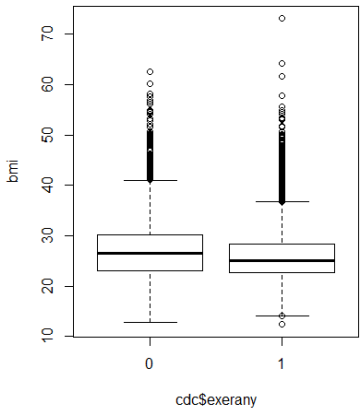
Exercise 4: Create a new object called `under23_and_smoke` that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime. Write the command you used to create the new object as the answer to this exercise.

```
under23_and_smoke <- subset(cdc,cdc$age < 24 & cdc$smoke100 == 1)
```

Exercise 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. (Include plot)

The box plot shows that when respondents said they had better health the lower there bmi mean and quartiles were.
I chose exerany I think it would relate to bmi because exercising can make people lose wight and weight is an input for bmi. The figure suggests that when the respondents had exercises on average they had a lower bmi.

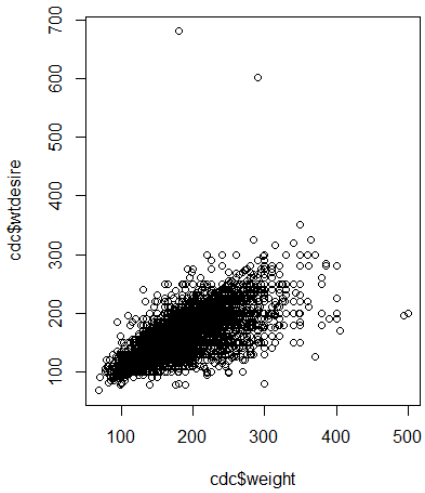




On Your Own:

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

These two variables seem to have a close to linear relationship but with more data lower than the a y=x line suggesting more people are at a greater wight then desired



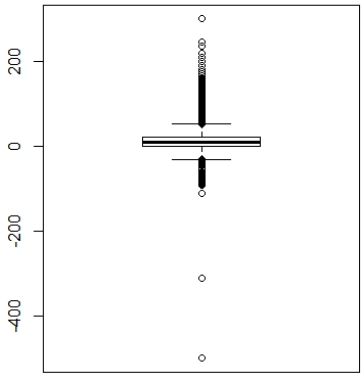
2) Let's consider a new variable: the difference between desired weight (`wtdesire`) and current weight (`weight`). Create this new variable by subtracting the two columns in the data frame and assigning them to a new object called `wdiff`.

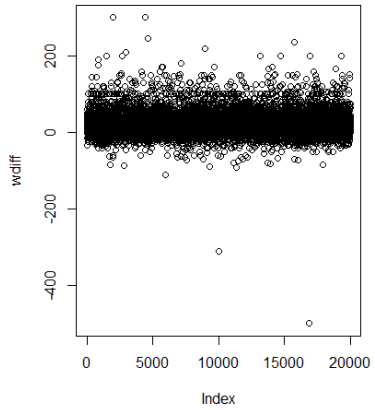
3) What type of data is `wdiff`? If an observation `wdiff` is 0, what does this mean about the person's weight and desired weight. What if `wdiff` is positive or negative?

The data is an int variable. If the wdiff = 0 they are the weight they desire to be. If wdiff is positive they weigh more than their desired weight and if wdiff is negative they weigh less than their desired weight.

4) 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?

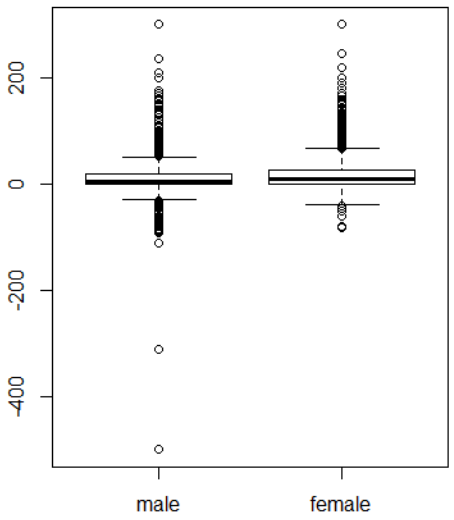
The center of in terms of median is 10 while in terms of mean is 14.59 the spread of the data in terms of range is 800 but in terms of IQR is 21. The quartiles create a small box since most of the data is in the small range from 0 – 21. This tells me that most people are above their desired weight.





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

Both men and women weigh more than they're desired weight however women on average are farther way to their desired weight than men are



6) 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 = 169.7

Std = 40.07

1std range=129.621-209.779

in range = 14152

.7076 or 70.76% of the sample are within one standard deviation of weight