Introduction to Data

Chapter 1, Lab 1: Solutions

OpenIntro Biostatistics

Topics

- Dataset manipulation in R
- Numerical summaries: mean, SD, median, IQR
- Graphical summaries: boxplots, histograms, scatterplots

The first two sections of this lab introduce basic tools for working with data matrices, as well as the commands for producing numerical and graphical summaries. The last section focuses on data interpretation and reinforces the statistical concepts presented in the text. The material in this lab corresponds to Sections 1.1 - 1.2 and 1.4 - 1.6 of *OpenIntro Biostatistics*.

Section 1: BRFSS.

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States. The BRFSS is designed to identify risk factors in the adult population and report emerging health trends. For example, respondents are asked about their diet, weekly exercise, possible tobacco use, and healthcare coverage.

1. Use the following command to download the dataset cdc from a URL. This dataset is a sample of 20,000 people from the survey conducted in 2000, and contains responses from a subset of the questions asked on the survey.

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

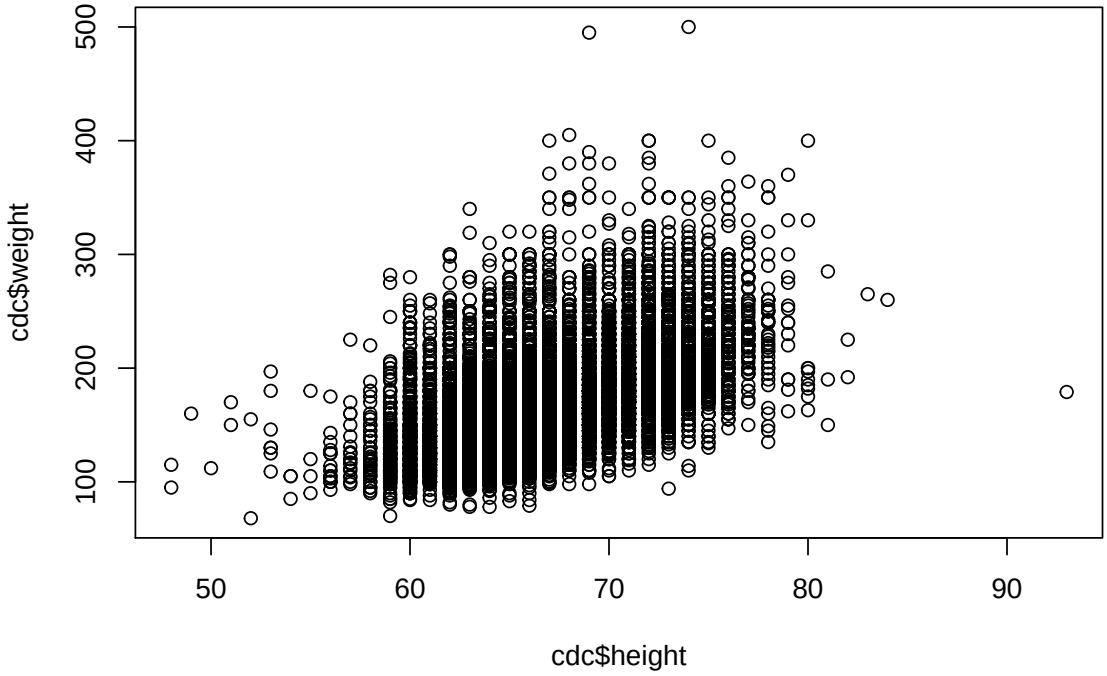
2. Take a look at the Environment tab, where cdc should now be visible. Click the blue button next to the dataset name to view a summary of the 9 variables contained in the data matrix. To view the dataset itself, click on the name of the dataset; alternatively, enter the command

   ```
   View(cdc)
   ```

   Each row of the data matrix represents a case and each column represents a variable. Each variable corresponds to a question that was asked in the survey. For `genhlth`, respondents were asked to evaluate their general health as either "excellent", "very good", "good", "fair", or "poor". The variables `exerany`, `hlthplan`, and `smoke100` are binary variables, with responses recorded as either 0 for "no" and 1 for "yes": whether the respondent exercised in the past month, has health coverage, or has smoked at least 100 cigarettes in their lifetime. The other variables record the respondents' height in inches, weight in pounds, their desired weight (`wtdesire`), age in years, and gender.

3. The $ operator in R is used to access variables within a dataset; for example, `cdc$height` tells R to look in the cdc dataframe for the `weight` variable. Make a scatterplot of `height` and `weight` using the `plot( )` command:

```
plot(cdc$weight ~ cdc$height)
```



Do height and weight appear to be associated?

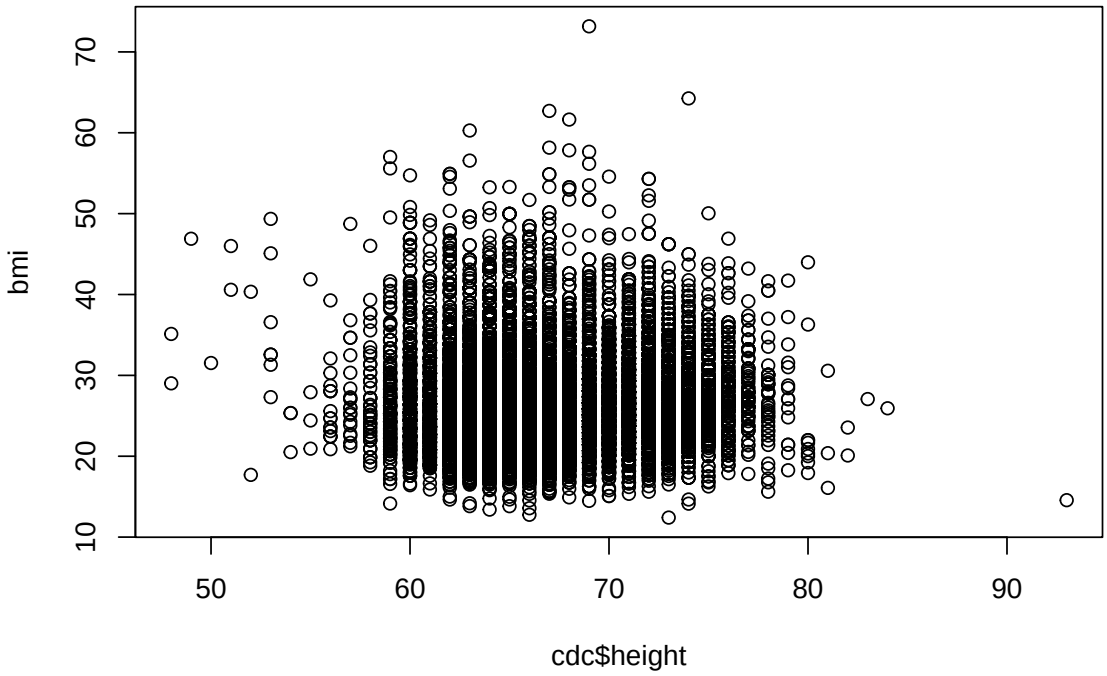
The visible upward trend in the cloud of points shows that height and weight are positively associated; weight tends to increase as height increases.

4. The conversion from inches to meters is 1 in = .0254 m. Create a new variable height.m that records height in meters. Similarly, the conversion from pounds to kilograms is 1 lb = .454 kg. Create a new variable weight.kg that records weight in kilograms.

```
#create height.m
height.m = cdc$height*.0254

#create weight.kg
weight.kg = cdc$weight*.454
```

5. BMI is calculated as weight in kilograms divided by height squared. Create a new variable bmi and make a scatterplot of height and BMI. Do height and BMI seem to be associated?

```
#create bmi
bmi = (weight.kg)/(height.m^2)

#plot height and bmi
plot(cdc$height, bmi)
```



Height and BMI do not appear to be associated.

A BMI of 30 or above is considered overweight. Why might health agencies choose to use BMI as a measure of obesity, rather than weight?

Since `height` and `BMI` have a much weaker association, it is more useful to use BMI as a measure of obesity. Using BMI is one way to account for the fact that taller people tend to have more tissue and thus, weigh more than shorter people.

6. Row-and-column notation in combination with square brackets can be used to access a subset of the data. For example, to access the sixth variable (`weight`) of the 567th respondent, use the command:

```
cdc[567, 6]
```

```
## [1] 160
```

To see the weight for the first ten respondents, use:

```
cdc[1:10, 6]
```

```
##  [1] 175 125 105 132 150 114 194 170 150 180
```

If the column number is omitted, then all the columns will be returned for rows 1 through 10:

```
cdc[1:10, ]
```

```
##      genhlth exerany hlthplan smoke100 height weight wtdesire age gender
```

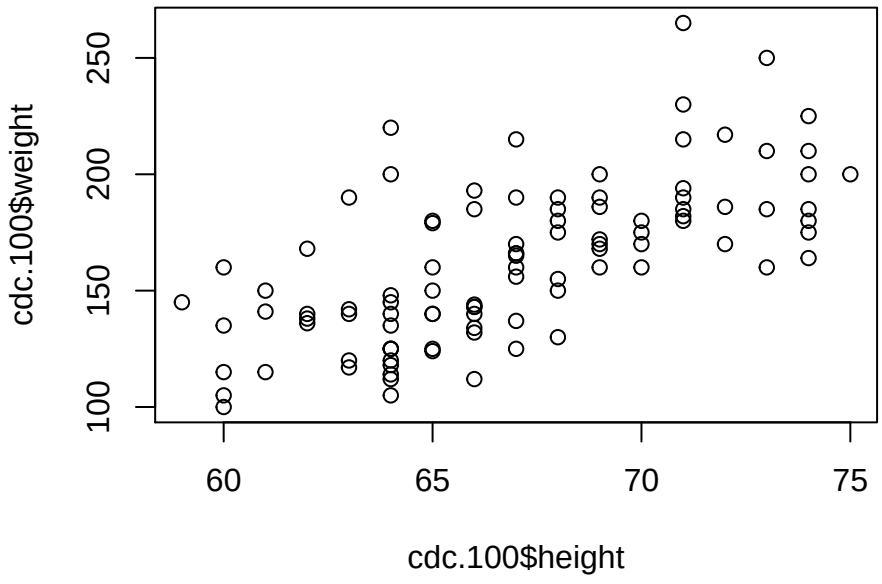
```
## 1       good       0        1        0     70    175       175  77       m
## 2       good       0        1        1     64    125       115  33       f
## 3       good       1        1        1     60    105       105  49       f
## 4       good       1        1        0     66    132       124  42       f
## 5  very good       0        1        0     61    150       130  55       f
## 6  very good       1        1        0     64    114       114  55       f
## 7  very good       1        1        0     71    194       185  31       m
## 8  very good       0        1        0     67    170       160  45       m
## 9       good       0        1        1     65    150       130  27       f
## 10      good       1        1        0     70    180       170  44       m
```

Likewise, omit the range for the rows to access all observations for column 6. The following will return the weight for all 20,000 respondents:

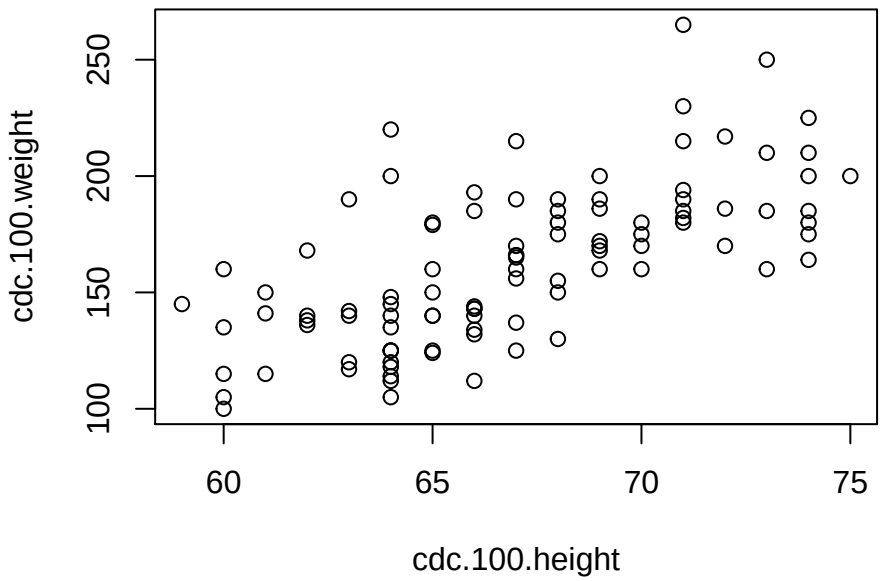
```
cdc[ ,6] #results of this chunk are hidden with eval = FALSE
```

7. Use bracket notation to make a scatterplot of `height` and `weight` for the first 100 respondents. There are multiple ways to do this—find one that works!

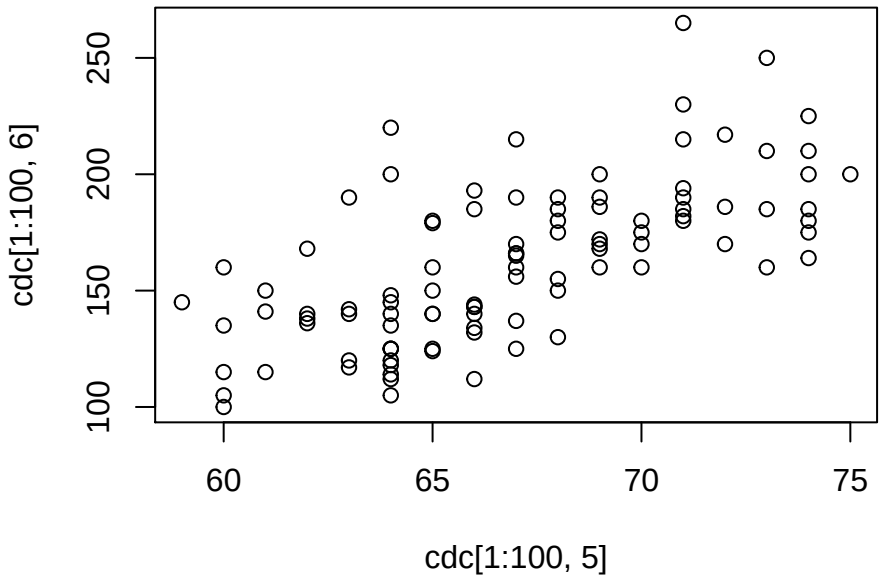
```
#create a new dataset with just 100 observations
cdc.100 = cdc[1:100, ]
plot(cdc.100$height, cdc.100$weight)
```




```
#subset the variables separately
cdc.100.weight = cdc[1:100, 6]
cdc.100.height = cdc[1:100, 5]
plot(cdc.100.height, cdc.100.weight)
```




```
#nest the commands
plot(cdc[1:100, 5], cdc[1:100, 6])
```



Section 2: Gene Transcript Lengths.

Before genes can be translated into proteins, DNA must first be transcribed into RNA. The dataset `coding.mrna` contains the length of known protein-coding transcripts (measured in base pairs). Load the dataset from the `oibiostat` package using the `load()` command.

```
#load the oibiostat package
library(oibiostat)

#load the coding.mrna dataset
data(coding.mrna)
```

1. How many transcripts are represented in this dataset? Use the `nrow()` command to return the number of rows in the dataset.

   ```
   nrow(coding.mrna)
   ```

   ```
   ## [1] 79105
   ```

   Each row corresponds to a transcript; 79,105 transcripts are represented in this dataset.

2. Calculate the 5-number summary for the transcript lengths using the `summary()` command. What striking feature do you notice in the summary?

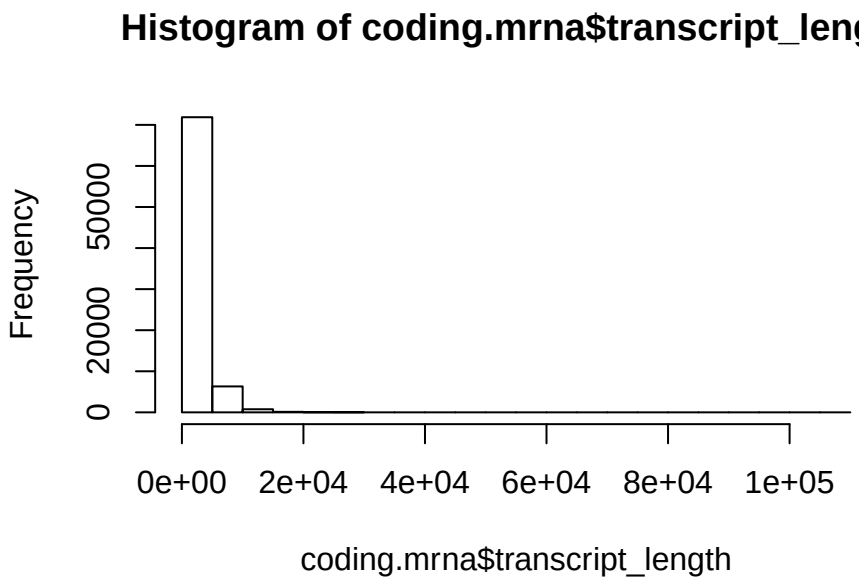
   ```
   #calculate the 5-number summary
   summary(coding.mrna$transcript_length)
   ```

   ```
   ##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   ##      54     707    1547    2219    2913  109224
   ```

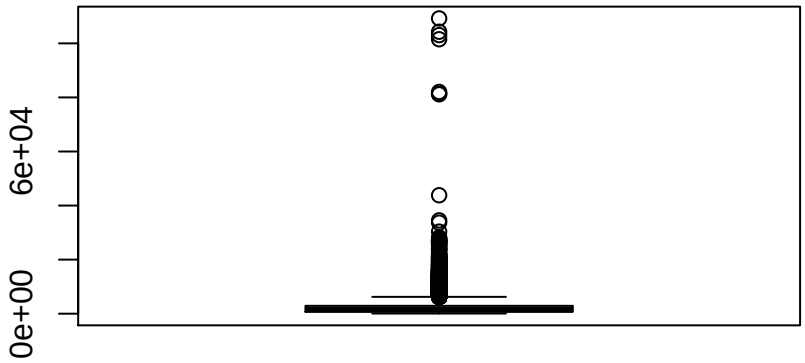
   The maximum is much larger than the other numbers in the five-number summary, which implies that the data is heavily right-skewed. This is also suggested by the fact that the mean is larger than the median. Note that the mean is not part of the 5-number summary as defined in *OpenIntro Biostatistics*.

3. Draw a histogram and a boxplot of the distribution of transcript lengths. When you see them, you will notice that the plots are not particularly informative. Explain why you think that is the case.

   ```
   #create a histogram
   hist(coding.mrna$transcript_length)
   ```




```
#create a boxplot
boxplot(coding.mrna$transcript_length)
```



These plots are not particularly informative because of the extreme skew, as well as the presence of a few very large outliers. This makes it nearly impossible to see features of the distribution around the mode.

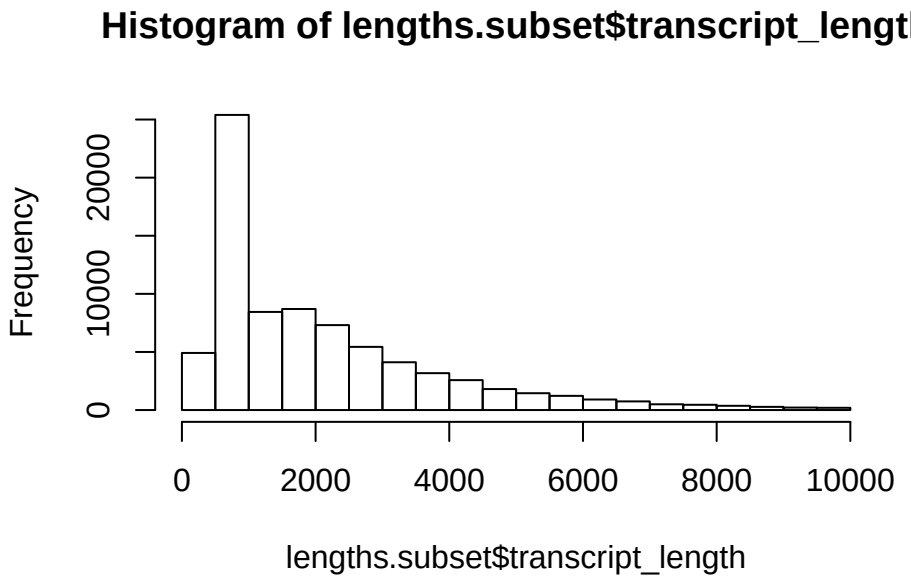
4. For a data item `x`, the notation `x < a` is used to reference the subset of values of `x` that are less than the value `a`. Pick a reasonable length `a` and use the `subset()` command to create a trimmed version of `coding.mrna` called `lengths.subset` that only contains data for transcripts

with length less than a. This is one simple strategy for making the structure of the data easier to view in the plot.

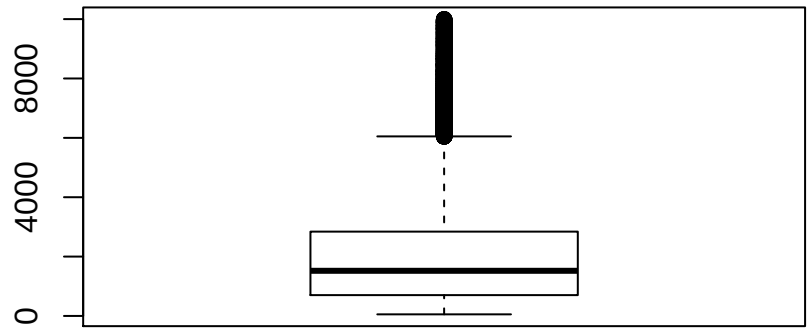
```
lengths.subset = subset(coding.mrna, coding.mrna$transcript_length < 10000)
```

With the trimmed data, draw a histogram and boxplot, and calculate summary statistics. Now describe the shape of the data. Explain your choice of where to trim the data.

```
#create a histogram with lengths.subset
hist(lengths.subset$transcript_length)
```




```
#create a boxplot with lengths.subset
boxplot(lengths.subset$transcript_length)
```




```
#calculate summary statistics
summary(lengths.subset$transcript_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      54     702    1521    2081    2841    9996
```

The following answers will vary depending on the chosen cutoff value. 10,000 bp was chosen as the cutoff value because nearly all of the values appear to fall in that range on the histogram – the goal of trimming is to see the first bin on the original histogram more clearly (instead of having 70,000 transcripts grouped together). The data remains heavily right-skewed, but the new histogram reveals a mode between 500-1000 bp.

5. Use R to find out how many transcripts you have trimmed from the dataset. Hint: this might involve notation used in Questions 1 and 4.

```
#subset the longer transcripts that were trimmed--ones greater than or equal 10000 bp
trimmed.lengths = subset(coding.mrna, coding.mrna$transcript_length >= 10000)
nrow(trimmed.lengths)
```

```
## [1] 934
```

```
#alternatively, subtract the length of the subset from total length
nrow(coding.mrna) - nrow(lengths.subset)
```

```
## [1] 934
```

934 transcripts were trimmed from the original list.

6. One way of manipulating a large dataset is to take a random sample and construct numerical and graphical summaries of the sample. Use the following code to construct a random sample that consists of 10% of the original number of transcripts; the sampling is done without replacement, such that a single transcript cannot be chosen more than once.

Using the `set.seed()` function allows for pseudo-random sampling; that is, a random sample that is reproducible. Replace `xxxx` in the function with four numbers of your choice, then run the code to create `transcript.sample`, a vector of transcript lengths.

```
set.seed(5011)
sample.size = 0.1 * nrow(coding.mrna)
transcript.sample = sample(coding.mrna$transcript_length, size = sample.size,
                           replace = FALSE)
```

Now with `transcript.sample`, calculate the number of transcripts in the dataset, the five-number summary, and draw a histogram and boxplot. Does the sample data more closely resemble the complete version of the data or the trimmed version from Question 4?

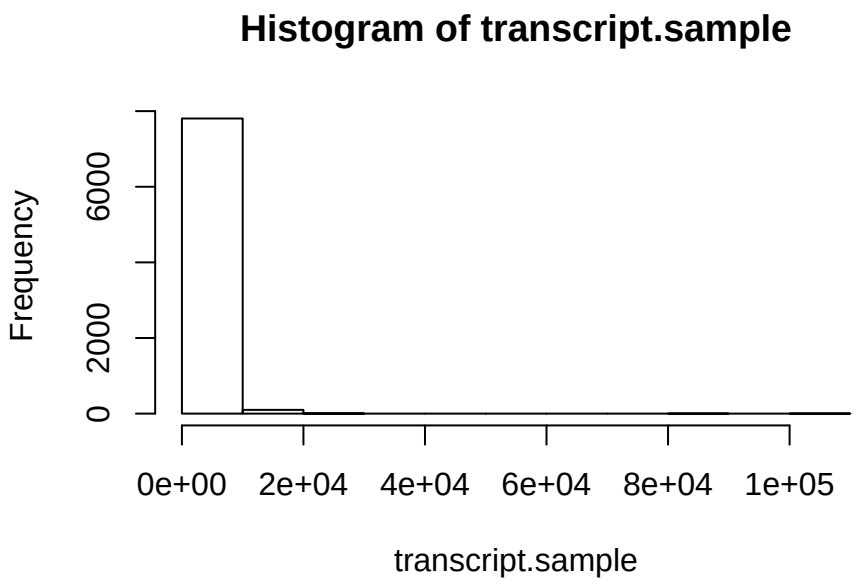
```
#calculate the number of transcripts in transcript.sample
length(transcript.sample)
```

```
## [1] 7910
```

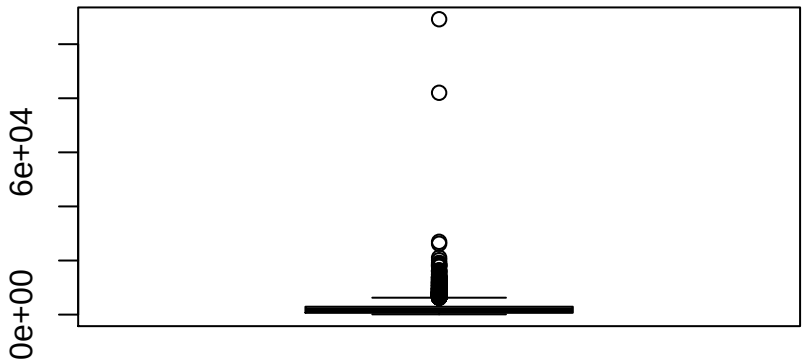
```
#calculate five-number summary
summary(transcript.sample)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      93     701    1550    2252    2921  109224
```

```
#create a histogram
hist(transcript.sample)
```



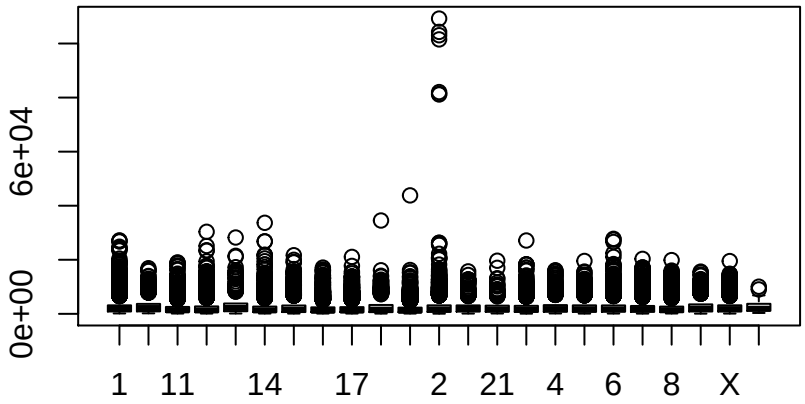

```
#create a boxplot
boxplot(transcript.sample)
```



For this particular sample, some of the most extreme outliers were picked up (such as a transcript of 109,224 bp), so the sample more closely resembles the complete data. All samples most likely will still be skewed right with some large outliers, since there are several in this data.

7. Make side-by-side boxplots of transcript lengths by chromosome. Use the command:

```
boxplot(coding.mrna$transcript_length ~ coding.mrna$chromosome_name)
```



Select "Show in New Window" above the plot and expand the window to be able to see all

the chromosome numbers displayed. Where are the longest transcripts located?

The longest transcripts are on Chromosome 2.

8. Subset `coding.mrna` to only include values from chromosome 2. Repeat for the Y chromosome. Hint: the notation is similar to that used in Question 4.

   Use `nrow()` to compare the number of transcripts on chromosome 2 and the Y chromosome. Are the results what you might expect, based on what you know about the inheritance of human sex chromosomes? Why or why not?

```
#chromosome 2 subset
chr.2 = subset(coding.mrna, coding.mrna$chromosome_name == 2)

#Y chromosome subset
chr.Y = subset(coding.mrna, coding.mrna$chromosome_name == "Y")

#compare the number of transcripts
nrow(chr.2)
```

```
## [1] 5294
```

```
nrow(chr.Y)
```

```
## [1] 135
```

There are many more protein-coding transcripts on Chromosome 2 than the Y chromsome. This is to be expected—the Y chromosome is much smaller than chromosome 2 (59 million bp versus 242 million bp). Additionally, the Y chromosome is only necessary for male sex determination. Genes absolutely essential for general development must be located on either the autosomes or the X chromosome.

Section 3: NHANES.

The National Health and Nutrition Examination Survey (NHANES) is a survey conducted annually by the US National Center for Health Statistics (NCHS). While the original data uses a survey design that oversamples certain subpopulations, the data have been reweighted to undo oversampling effects and can be treated as if it were a simple random sample from the American population.

The following questions will be explored with the NHANES data:

1. At what age do Americans seem to reach full adult height?
2. What proportion of Americans age 25 or older have a college degree?
3. What is the relationship between education level and income?
4. How much more likely is it that someone *not* physically active has diabetes, compared to someone who is active?

The reweighted NHANES data are available from the NHANES package. To view the complete list of study variables and their descriptions, access the NHANES documentation page with `?NHANES`.

For convenience, descriptions of the variables used in this lab exercise are included below.

- `Age`: age in years at screening. Subjects 80 years or older were recorded as 80 years of age.
- `Education`: highest educational level of study participant, reported for participants aged 20 years or older. Recorded as either `8th Grade`, `9 - 11th Grade`, `High School`, `Some College`, or `College Grad`.
- `Poverty`: a ratio of family income to poverty guidelines. Smaller numbers indicate more poverty; i.e., a number below 1 indicates income below the poverty level.
- `Weight`: weight, measured in kilograms.
- `Height`: standing height, measured in centimeters.
- `Diabetes`: `Yes` if the participant was told by a health professional that they have diabetes, `No` otherwise.
- `PhysActive`: coded `Yes` if the participant does moderate or vigorous-intensity sports, fitness, or recreational activities; `No` otherwise. Reported for participants 12 years or older.

Question 1.

a) Describe in words the distribution of ages for the study participants.

The distribution of ages is relatively symmetric. 50% of the respondents are below age 36, while the middle 50% of respondents are between ages 17 and 54. Note that the true maximum respondent age is not 80; subjects 80 years or older were recorded as 80 years of age.

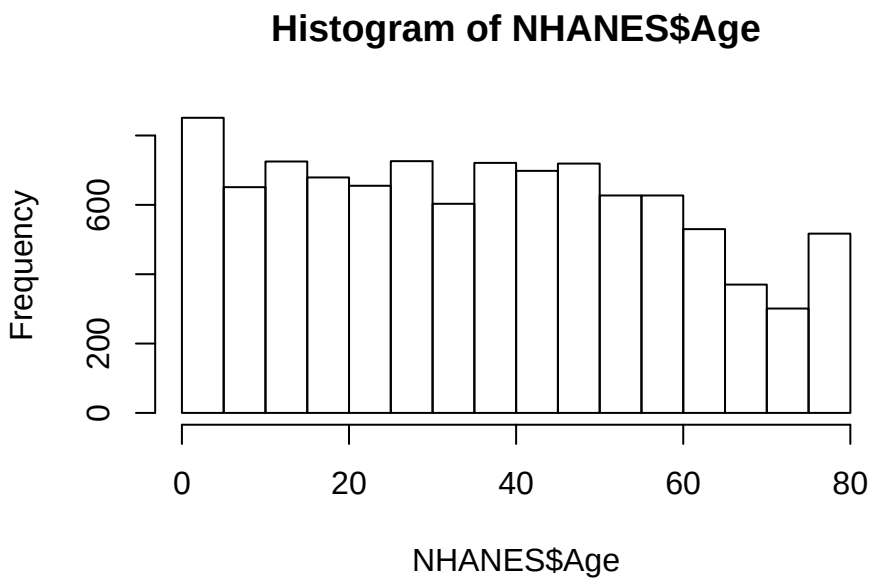
```
#load the NHANES package and dataset
library(NHANES)
data(NHANES)

#numerical summaries
summary(NHANES$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   17.00   36.00   36.74   54.00   80.00
```

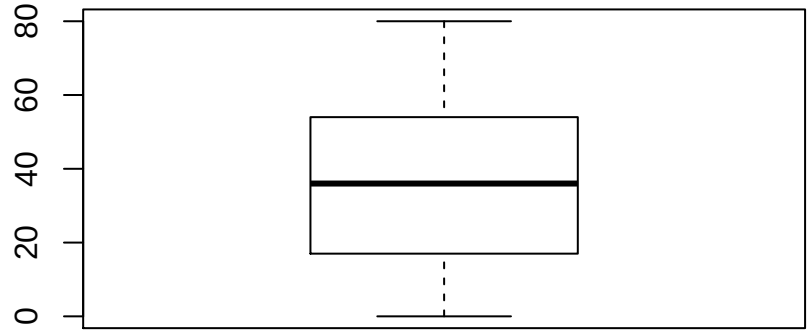
```
sd(NHANES$Age)
```

```
## [1] 22.39757
```

```
#graphical summaries
hist(NHANES$Age)
```




```
boxplot(NHANES$Age)
```



b) Using numerical and graphical summaries, describe the distribution of heights among study participants in terms of inches. Note that 1 centimeter is approximately 0.39 inches.

The distribution of heights among study participants is highly left skewed; there are many more individuals with high values for height than there are for lower values. The median height is about 65 inches (5.5 feet). The boxplot indicates the left skewed distribution as a series of dots on the lower end of the plot.

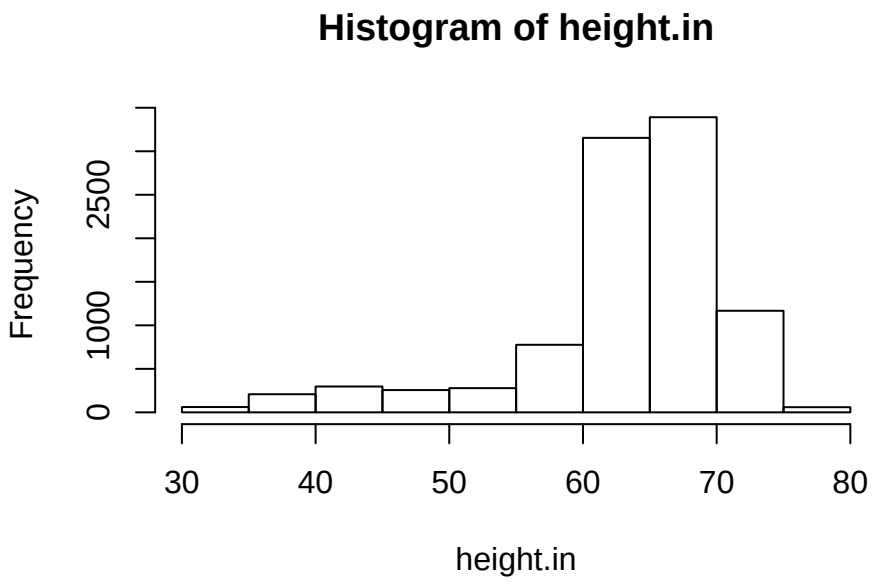
```
#convert to inches
height.in = 0.39*NHANES$Height

#numerical summaries
summary(height.in)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   32.60   61.15   64.74   63.13   68.06   78.16     353
```

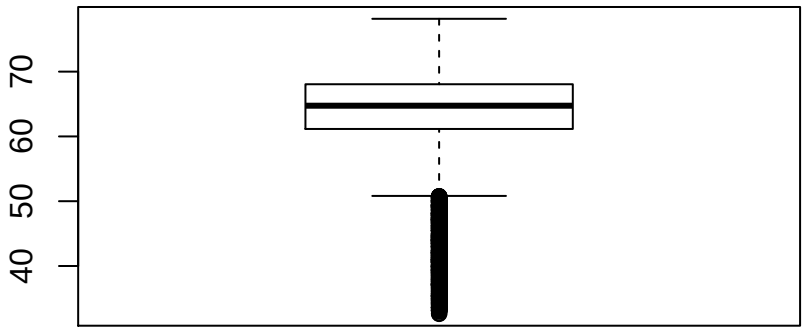
```
sd(height.in, na.rm = TRUE)  #na.rm = TRUE instructs R to ignore missing values (NA's)
```

```
## [1] 7.872761
```

```
#graphical summaries
hist(height.in)
```



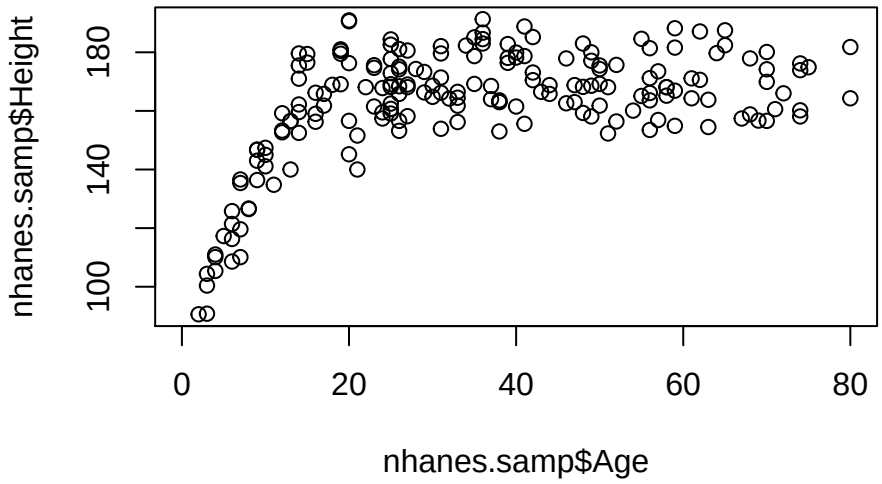

```
boxplot(height.in)
```



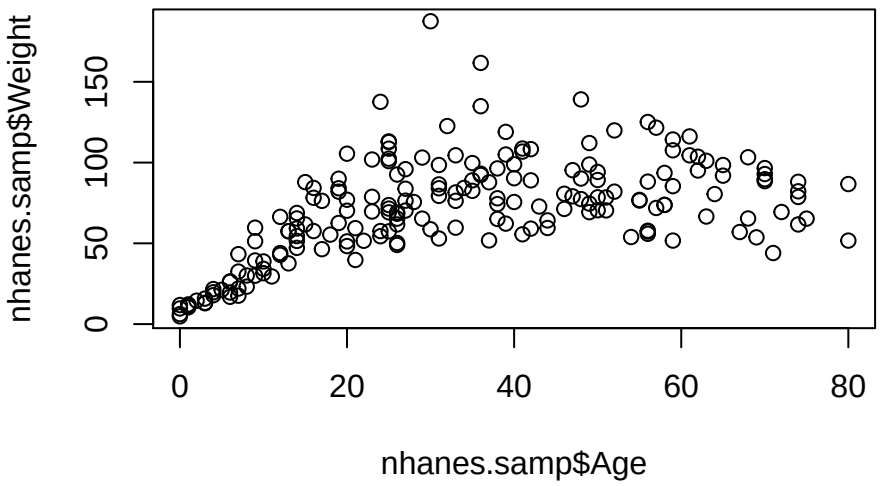
c) Use the following code to draw a random sample of 200 participants from the entire dataset. Using the random sample, `nhanes.samp`, investigate at which age people generally reach their adult height. Is it possible to do the same for weight; why or why not?

The scatterplot shows that people generally reach their adult height around age 20. It is not possible to do the same for weight, since unlike height, weight can fluctuate throughout adulthood.

```r
#draw a random sample
set.seed(5011)
row.num = sample(1:nrow(NHANES), 200, replace = FALSE)
nhanes.samp = NHANES[row.num, ]

#investigate age and height
plot(nhanes.samp$Age, nhanes.samp$Height)
```




```r
#investigate age and weight
plot(nhanes.samp$Age, nhanes.samp$Weight)
```



Question 2.

a) What proportion of Americans at least 25 years of age are college graduates?

0.307 of Americans at least 25 years of age are college graduates.

```
#subset the number of Americans at least 25 years of age
adults = NHANES[NHANES$Age >= 25, ]

#age and education
table(adults$Education) #summary(adults$Education) also works
```

```
##
##      8th Grade 9 - 11th Grade    High School   Some College   College Grad
##            435             814            1345           1951           2016
```

```
total.adults = length(adults$Education)

#calculations
2016/total.adults
```

```
## [1] 0.3068026
```

b) What proportion of Americans with a high school degree are college graduates?

Assuming that all students in college have a high school degree, the proportion of Americans with a high school degree that are college graduates is 0.380.

```
#calculations
(2016)/(1345  + 1951 + 2016)
```

```
## [1] 0.3795181
```

Question 3.

a) Calculate the median and interquartile range of the distribution of the variable `Poverty`. Write a sentence explaining the median in the context of these data.

The median is 2.7, and the IQR is 3.47. The median indicates that 50% of surveyed individuals have a poverty ratio above 2.7, and 50% have a poverty ratio above 2.7; 2.7 indicates an income level that is 2.7 times the poverty level.

```
#numerical summary
summary(NHANES$Poverty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.240   2.700   2.802   4.710   5.000     726
```

```
#alternatively, directly use median() and IQR()
#na.rm = TRUE instructs R to disregard the missing values (NA's)
median(NHANES$Poverty, na.rm = TRUE)
```

```
## [1] 2.7
```

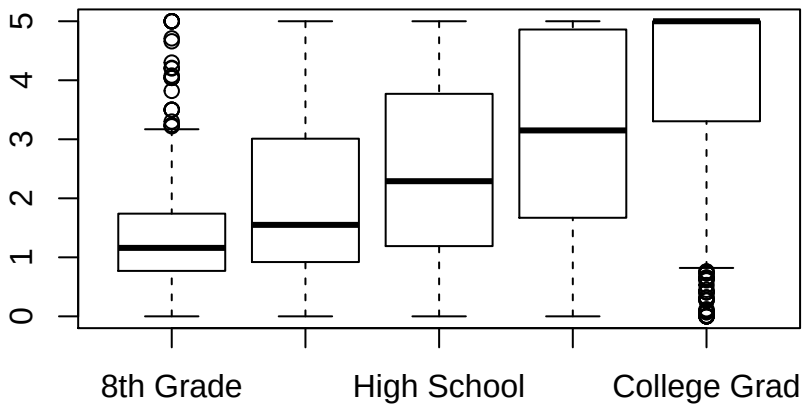
```
IQR(NHANES$Poverty, na.rm = TRUE)
```

```
## [1] 3.47
```

b) Compare the distribution of `Poverty` across each group in `Education` among adults (defined as individuals 25 years of age or older). Describe any trends or interesting observations.

The median level of poverty increases as the highest level of education completed increases. While individuals who only completed 8th grade have median poverty around 1.1, individuals who have college degrees have median poverty at 5. The data also show that some individuals who only completed 8th grade are relatively wealthy, while some individuals with college degrees have incomes below the poverty level.

```
#graphical summary
boxplot(adults$Poverty ~ adults$Education)
```



Question 4.

a) Construct a two-way table, with `PhysActive` as the row variable and `Diabetes` as the column variable. Among participants who are not physically active, what proportion have diabetes? What proportion of physically active participants have diabetes?

Among participants who are not physically active, 0.30 have diabetes; 0.06 of participants who are physically active have diabetes.

```
#create table
addmargins(table(PhysActive=NHANES$PhysActive, Diabetes=NHANES$Diabetes))
```

```
##           Diabetes
## PhysActive   No  Yes  Sum
##        No  3203  472 3675
##        Yes 4361  285 4646
##        Sum 7564  757 8321
```

```
#calculations
diabetes.not.active = 472/3675
diabetes.active = 285/4646

diabetes.not.active
```

```
## [1] 0.1284354
```

```
diabetes.active
```

```
## [1] 0.06134309
```

b) In this context, relative risk is the ratio of the proportion of participants who have diabetes

among those who are not physically active to the proportion of participants with diabetes among those physically active. Relative risks greater than 1 indicate that people who are not physically active seem to be at a higher risk for diabetes than physically active people. Calculate the relative risk of diabetes for the participants.

From these calculations, is it possible to conclude that being physically active reduces one's chance of becoming diabetic?

The relative risk of diabetes is 2.09. From these data, individuals who are not physically active are twice as likely as those who are physically active to have diabetes. However, this is not sufficient to make a causal claim about the relationship between physical activity and diabetes incidence.

```
#calculations
rr.diabetes = diabetes.not.active/diabetes.active
rr.diabetes
```

```
## [1] 2.093722
```