R Assignment 1

Exploratory Data Analsys

name

2024-12-30

Introduction / Goal:

In this assignment you will get yourself acquainted with new data. When we study data we are unfamiliar with we need to explore it. We do this by making graphs, calculating statistics and looking at the documentation. You will do all of these things for the yrbss data. Of course you don't know the yrbss data well yet, but by the end of this assignment you will be able to tell me where it originated, and some aspects of the age variable.

If you would like you can watch me do a similar analysis on another data set by watching the video on canvas.

Assignment:

You will do an exploratory data analysis on the yrbss survey data, specifically focusing on the age variable.

You will turn in a pdf output of your work by the due date.

There are 4 parts to complete and 8 questions to answer in this assignment. These will be peer graded, I will double check your peers' grading.

Familiarize yourself with your data.

You should make a chunk below and use the ? operator to pull up the yrbss data. Answer the questions below to introduce your data:

Make a chunk below (See R assignment 0).

Data Questions 1. How many variables are in the data?

Answer: 13

2. As listed in the documentation list the first three the variables that are categorical.

Answer: gender, grade, hispanic, race (grade is categorized as a character varible even though it is a number. So if it is included instead of race, that's ok.)

3. How many youth are present in this sample?

Answer: 13,583 youth

4. Which organization collected the data? The link in the documentation doesn't work well. This one is better

Answer: The cdc collects this data.

Calculate two Summary Statistics

Let us focus on one variable. Let's consider the age variable, which is numeric. Calculate the mean and the five number summary of the age variable. If you get NA for either one, you need to add the argument na.rm=TRUE as we did in the R Assignment 0.

```
# The code for the mean goes here
mean(yrbss$age, na.rm = TRUE )
```

```
## [1] 16.15704
```

```
# The code for the five number summary goes here
fivenum(yrbss$age)
```

```
## [1] 12 15 16 17 18
```

```
# The summary function works too
summary(yrbss$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   12.00   15.00   16.00   16.16   17.00   18.00      77
```

Recall from the reading that the five number summary breaks data up into quartiles, or chunks of 25%. You may want to reread chapter 5.5.

Questions:

5. What percent of the students are over 17? **Answer:** 25%
6. What are the ages of the youngest and oldest students in our data? **Answer:** 12 and 18
7. How many students did not report their age? (Hint: use the summary() function)

Answer: There were 77 students that did not report their age.

Make two plots

age is a numeric variable. We saw how to make a histogram for a numeric variable in R Assignment 0. We will do that again below. We will also add a title and site the source of our data. The outline for a histogram is completed for you

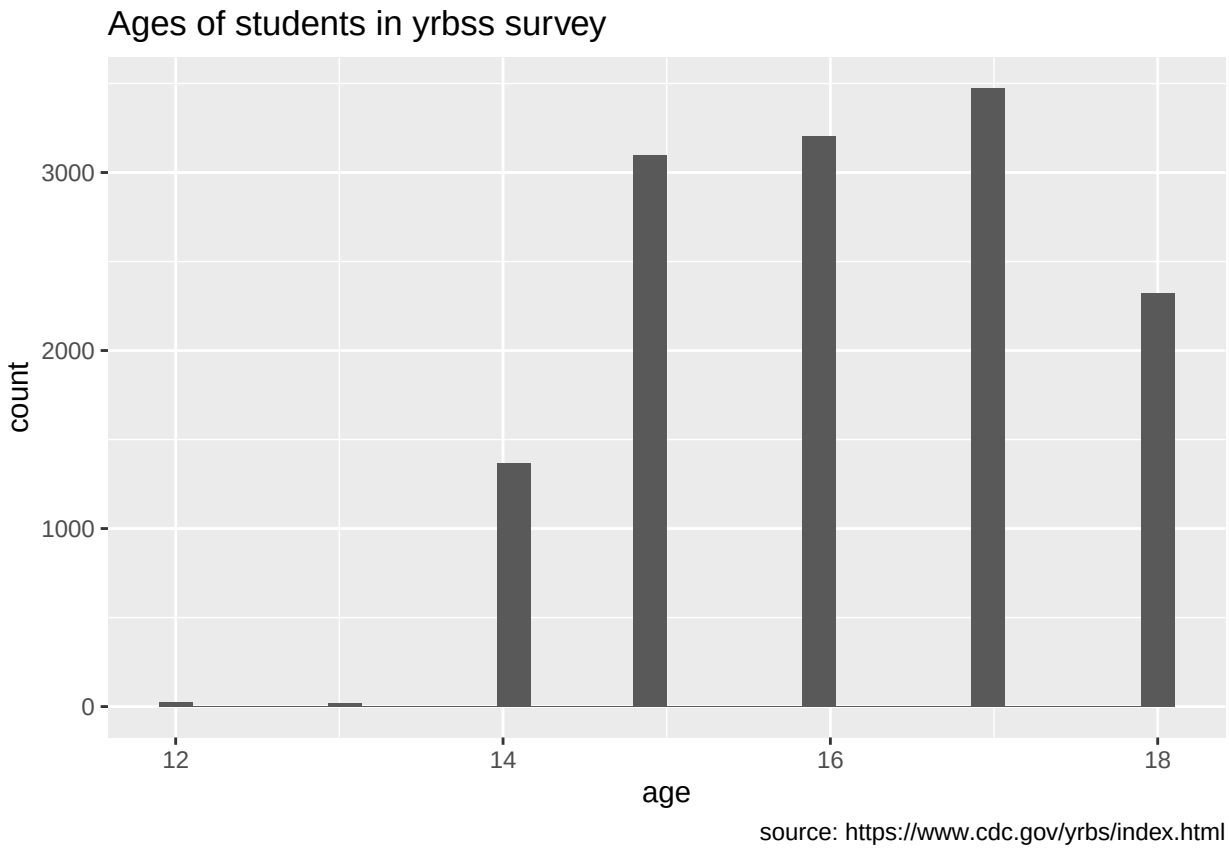
8. Make a histogram of age, include a title and caption.

```
# Write some ggplot() code for a histogram here.

ggplot(data =   yrbss )+   # Add the data
  geom_histogram(aes(x=  age ))+   # Add the age variable
  labs(
    title = "Ages of students in yrbss survey",
    caption = "source: https://www.cdc.gov/yrbs/index.html"
  )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 77 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

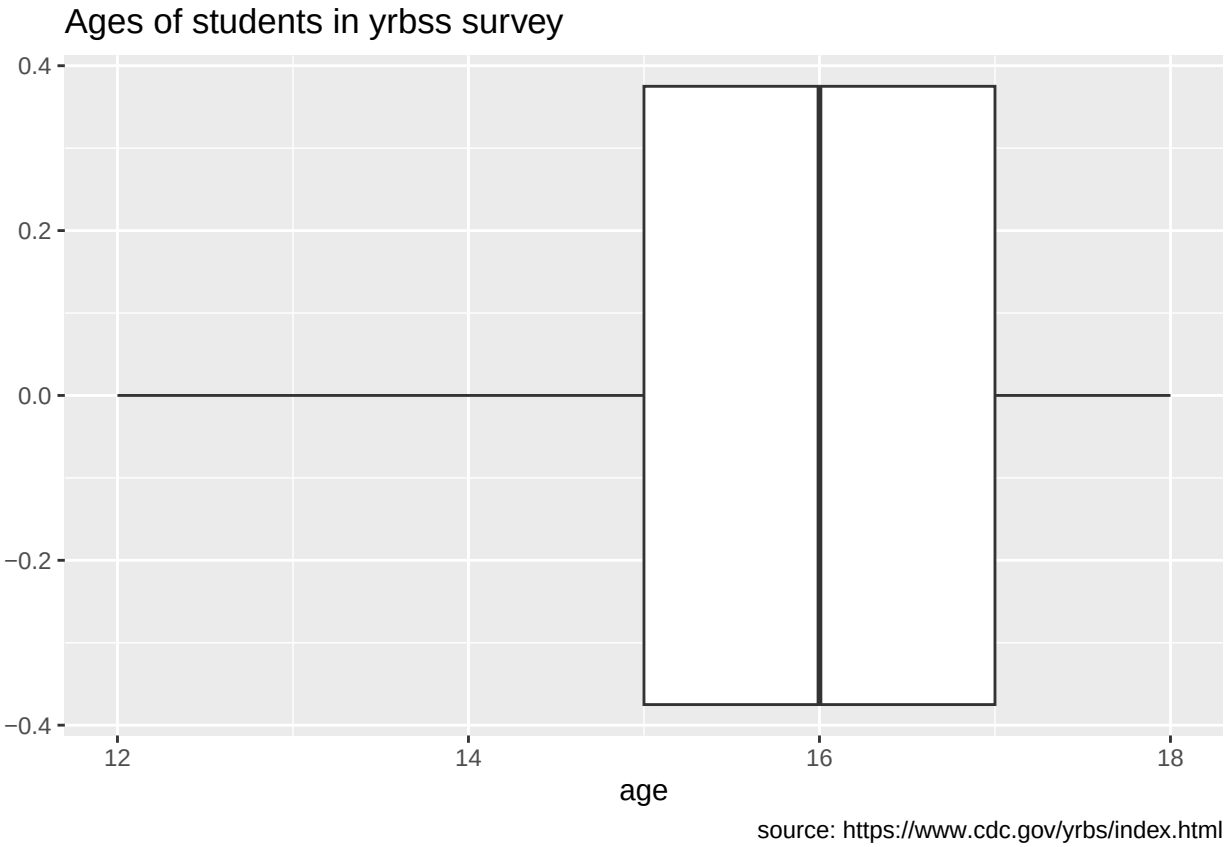


9. Now make a bloxplot for the data. The boxplot is a visual representation of the five number summary. Hint: You can recycle the code for your histogram just change the geom from histogram to boxplot.

```
# Write some ggplot() code for a boxplot here.

ggplot(data =   yrbss )+   # Add the data
  geom_boxplot(aes(x=  age ))+   # Add the age variable
  labs(
    title = "Ages of students in yrbss survey",
    caption = "source: https://www.cdc.gov/yrbs/index.html"
  )
```

```
## Warning: Removed 77 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



Conclusion

10. You've explored your data and gotten a better idea of the age variable in yrbss. Describe the ages of the population included in our sample. Tell me who they are, and who they are not. What age is most represented in our sample?

Answer: These are high school age students. Some of them are 12 or 13, but most of them are over 13 and younger than 18. 17 year olds are most represented in the data.

Finally knit and upload this document to canvas. If your document doesn't knit correctly post a screen shot to the forums.