Biostats 521: Lab #0 (EPID)
Brian Kim

https://github.com/kimqi/biostat521

1. Each dataset contains a variable for self-reported race. Rename that variable to simply race. Is there any need to order the levels of the race variable? Why?
Technically, there is no need to order the levels of the race variable because it does not have a natural ordering. For conventional purposes we can reorder to have the other category last…

```
DATA = rename(DATA, Race = RIDRETH1)
DATA$Race = ordered(DATA$Race, levels = c("Black", "White", "MexicanAmerican",
"OtherHispanic", "Other"))
```

2. How many levels are included in the race variable? What are the proportions for each group in your dataset?
There are 5 levels.

```
length(table(DATA$Race))
prop.table(table(DATA$Race))
```

Black	White	MexicanAmerican	OtherHispanic	Other
22.92%	45.01%	23.95%	3.89%	4.23%

3. Compute the mean, standard deviation and Five-Number Summary for the BMI variable in your dataset.

```
mean(DATA$BMXBMI, na.rm = TRUE)
sd(DATA$BMXBMI, na.rm = TRUE)
summary(DATA$BMXBMI)
```

Mean: 29.25
Standard Deviation: 7.20

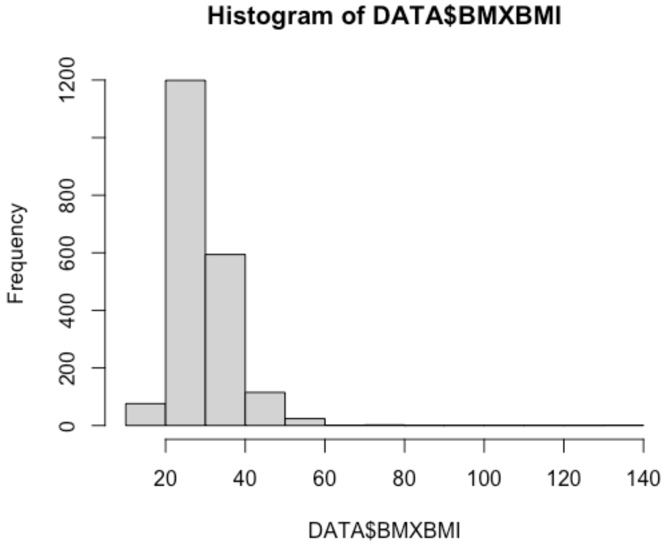
Min	**1st Q**	**Median**	**Mean**	**3rd Q**	**Max**	**NAs**
16.71	**24.59**	**27.83**	**29.25**	**32.28**	**130.21(?)**	**21**

4. Based on the descriptive statistics for BMI, do you have any concerns about potential outlier values?
The Max value is most likely a data entry error, as although there have been records of BMI's reaching the 100s, other health metrics seem to point to less of a systemic health issue.

5. Based on the descriptive statistics for BMI, do you think the distribution for BMI is most likely to be left skewed, right skewed or symmetric? Why?
It is likely right skewed because there is a slight floor effect in that it becomes hazardous under a certain BMI range, while being above average in weight is not as immediately life threatening.

6. Confirm your answer to the above question by creating an appropriate plot to visualize the distribution of BMI.

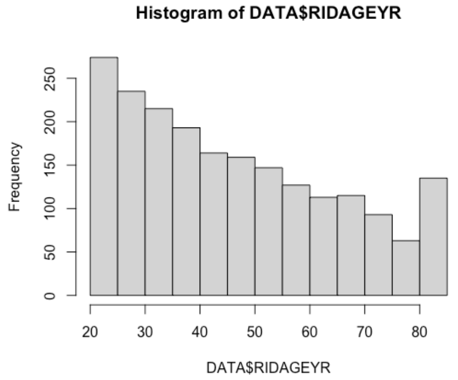
```
hist(DATA$BMXBMI)
```

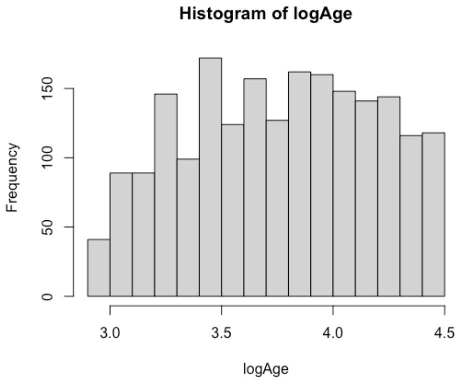

Histogram of DATA$BMXBMI

7. Create a new variable called LogAge containing the natural logarithm of the Age variable. (HINT: The log function computes the natural log, logAge=log(Age))

```
logAge = log(DATA$RIDAGEYR)
hist(DATA$RIDAGEYR)
hist(logAge)
```

8. What does the distribution of your new variable logAge? (HINT: Create a histogram or boxplot of the variable you created in the previous step.) Compare this to the shape of the original Age variable? That is, how did applying the natural log function change the shape of the distribution of ages in the dataset.


Histogram of DATA$RIDAGEYR


Histogram of logAge

The distribution became more uniform with the log transform.

9. Suppose that you are interested in designing a study with individuals aged 65 and above. How many such samples are in your dataset? (HINT: create a new variable to identify samples 65+ and use a table to count.)

There are 427 subjects aged 65 and above.

```
DATA = mutate(DATA, Age65 = ifelse(RIDAGEYR >= 65, 1, 0))
table(DATA$Age65)
```