Lab 0

Francisco Brady

2024-09-05

Load Data

```
setwd(here::here())
epid <- read_csv('EPID_521_lab_data.csv', na = 'NA')
```

```
## Rows: 2033 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): RIAGENDR, RIDRETH1, DMDEDUC2, DBQ700, DIQ010
## dbl (8): SEQN, LBXHCY, LBXCOT, RIDAGEYR, BMXBMI, BPXSY1, LBXTC, WkAlcDays
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

View Data

```
#View(epid)
glimpse(epid)
```

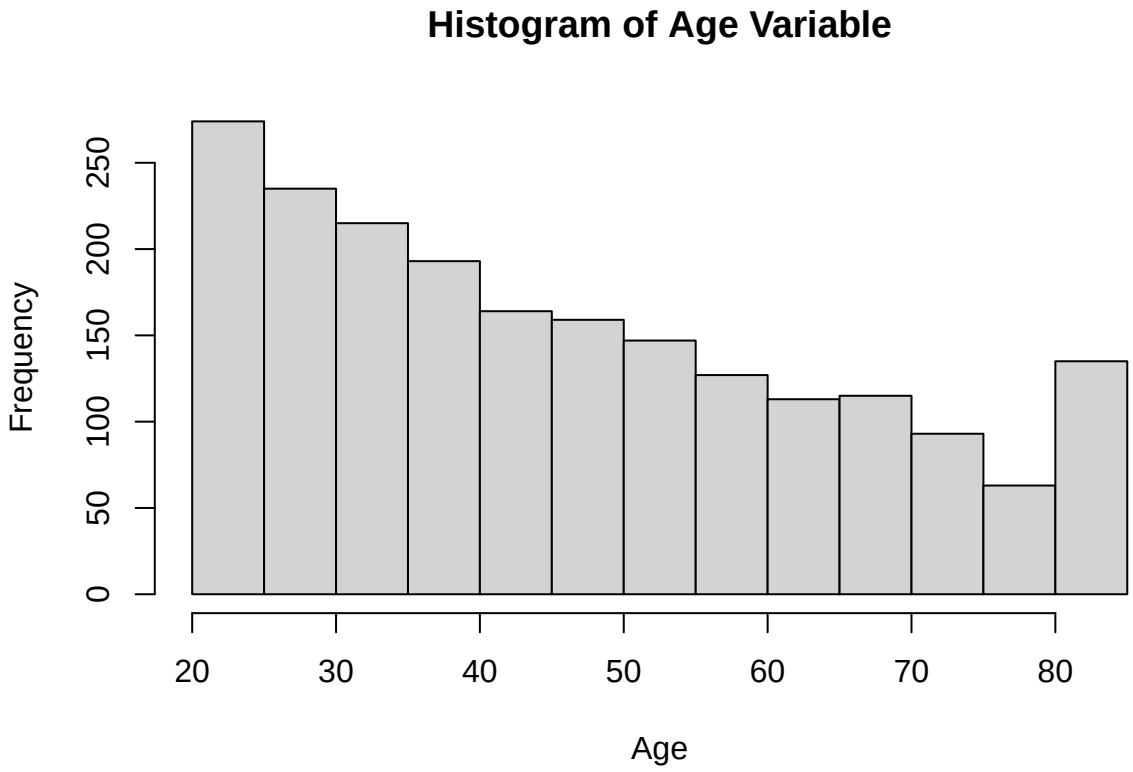
```
## Rows: 2,033
## Columns: 13
## $ SEQN      <dbl> 31131, 31132, 31134, 31144, 31149, 31152, 31155, 31160, 3116~
## $ LBXHCY    <dbl> 9.33, 8.96, 8.20, 7.97, 10.29, 3.97, 6.75, 7.98, 8.18, 9.68,~
## $ LBXCOT    <dbl> 0.035, 0.021, 0.065, 0.125, 0.011, 0.018, 0.030, 0.020, 0.01~
## $ RIAGENDR  <chr> "Female", "Male", "Male", "Male", "Female", "Female", "Male"~
## $ RIDAGEYR  <dbl> 44, 70, 73, 21, 85, 27, 38, 39, 71, 54, 33, 22, 22, 47, 25, ~
## $ RIDRETH1  <chr> "Black", "White", "White", "OtherHispanic", "White", "Mexica~
## $ DMDEDUC2  <chr> "SomeCollege", "College", "HighSchool", "HighSchool", "SomeH~
## $ BMXBMI    <dbl> 30.90, 24.74, 30.63, 25.03, 21.63, 39.88, 25.61, 35.19, 29.6~
## $ DBQ700    <chr> "Good", "VeryGood", "Good", "Excellent", "Good", "Good", "Go~
## $ DIQ010    <chr> "No", "Yes", "No", "No", "No", "No", "No", "No", "Yes", "No"~
## $ BPXSY1    <dbl> 144, 138, 130, 116, 110, 94, 126, 134, 106, NA, 114, 108, NA~
## $ LBXTC     <dbl> 105, 147, 186, 207, 121, 243, 170, 197, 224, 159, 167, 181, ~
## $ WkAlcDays <dbl> 0, 4, 2, 0, NA, NA, NA, 1, 0, 0, 0, NA, 1, 0, 1, NA, 0, NA, ~
```

```
dim(epid)
```

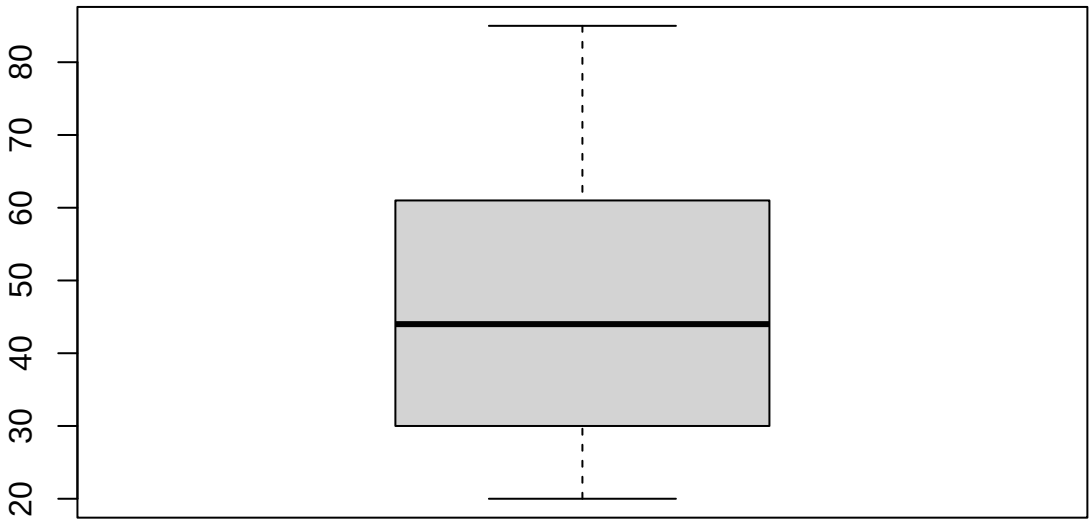
```
## [1] 2033   13
```

Numerical Variables

```
hist(epid$RIDAGEYR, main = "Histogram of Age Variable", xlab = "Age")
```




```
boxplot(epid$RIDAGEYR)
```



Five-Number Summary

```
mean(epid$RIDAGEYR, na.rm = TRUE)
```

```
## [1] 46.71864
```

```
sd(epid$RIDAGEYR, na.rm = TRUE)
```

```
## [1] 18.78938
```

```
summary(epid$RIDAGEYR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   30.00   44.00   46.72   61.00   85.00
```

Categorical Variables

Table

```
table(epid$RIAGENDR)
```

```
##
## Female   Male
##   1291    742
```

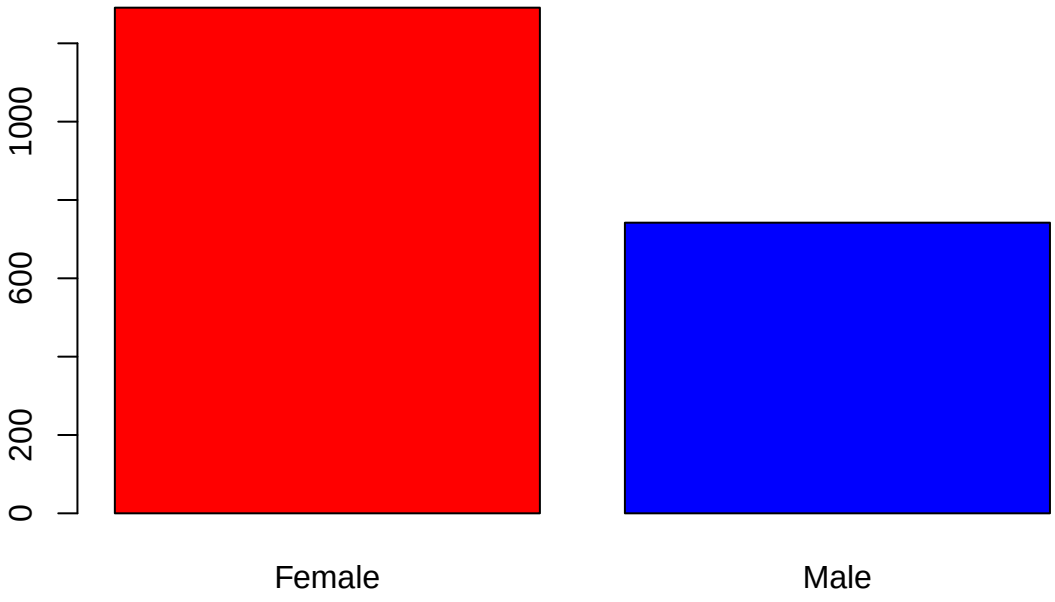
Table of Proportions

```
prop.table(table(epid$RIAGENDR))
```

```
##
##    Female      Male
## 0.6350221 0.3649779
```

Barplot

```
barplot(table(epid$RIAGENDR), col = c("red", "blue"))
```



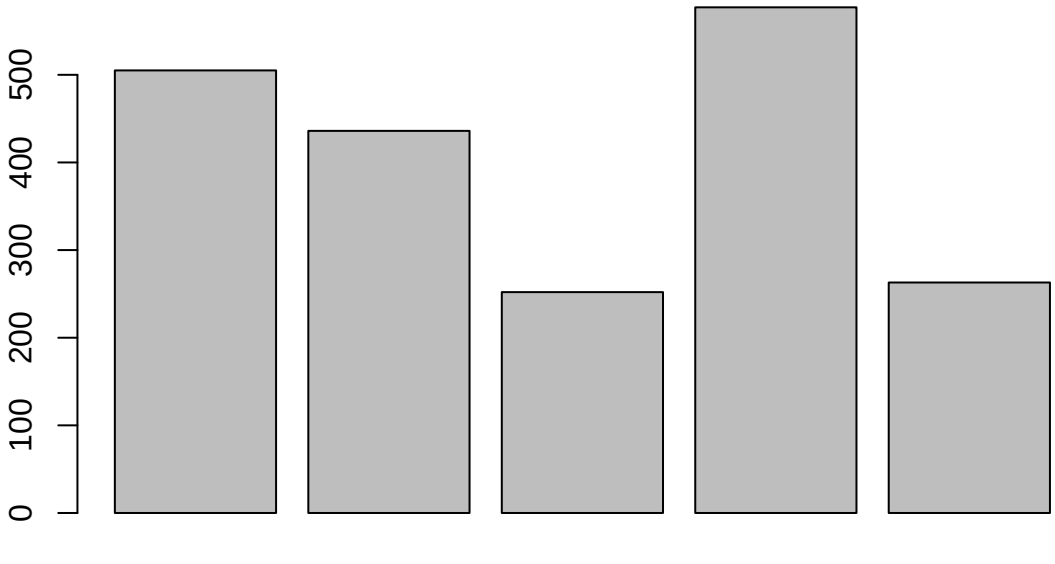
Education

```
table(epid$DMDEDUC2)
```

```
##
##        College     HighSchool   NoHighSchool    SomeCollege SomeHighSchool
##            505            436            252            577            263
```

Bar plot

```
barplot(table(epid$DMDEDUC2), cex.names = 0.9)
```



Ordered

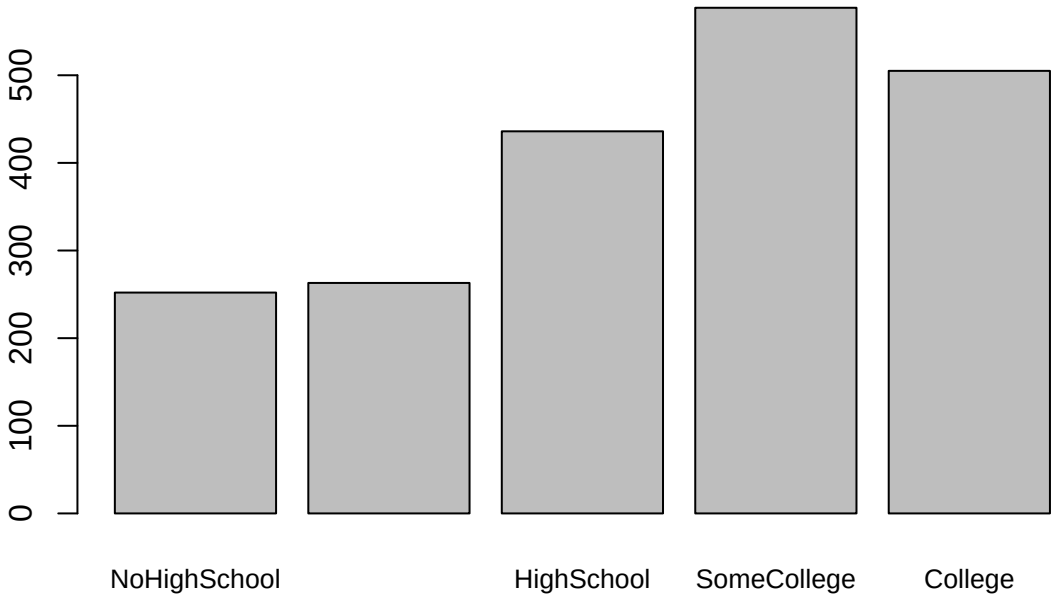
```
# order the levels of the education variable
epid$DMDEDUC2 <- ordered(epid$DMDEDUC2,
                         levels = c("NoHighSchool", "SomeHighSchool","HighSchool", "SomeCollege", "Colle
```

Barplot

```
table(epid$DMDEDUC2)
```

```
##
##   NoHighSchool SomeHighSchool     HighSchool    SomeCollege        College
##            252            263            436            577            505
```

```
barplot(table(epid$DMDEDUC2), cex.names = 0.8)
```



Basic Data Manipulation

```
epid <- rename(epid, Age = RIDAGEYR)
epid <- mutate(epid, Age50 = ifelse(Age >= 50, 1, 0))
```

```
table(epid$Age50)
```

```
##
##    0    1
## 1206  827
```

Questions

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?

No there is no need to order the race variable because there is no implied ordering to the variable, it does not make sense.

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

```
prop.table(table(epid$RIDRETH1))
```

```
##
##           Black MexicanAmerican           Other   OtherHispanic           White
##      0.22921790      0.23954747      0.04230202      0.03885883      0.45007378
```

There are 5 different levels for the race variable.

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

```
mean(epid$BMXBMI, na.rm = T)
```

```
## [1] 29.25318
```

```r
sd(epid$BMXBMI, na.rm = T)
```

```
## [1] 7.204896
```

```r
summary(epid$BMXBMI)
```

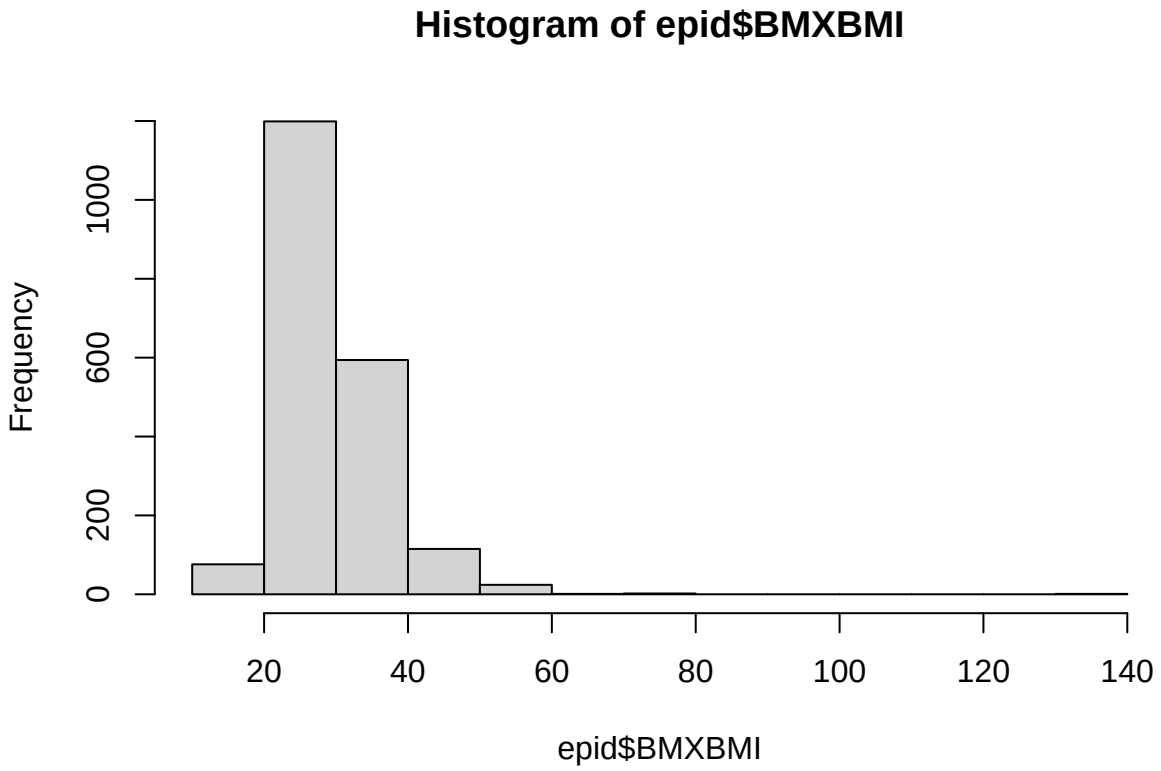
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   16.71   24.59   27.83   29.25   32.28  130.21      21
```

4. Based on the descriptive statistics for BMI, do you have any concerns about potential outlier values?
   There is one very high BMI value (130). The rest of the data is closer to the mean. It is possible that this is an error.

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?
   The distribution is most likely right-skewed because the mean is larger than the median.

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

```r
hist(epid$BMXBMI)
```

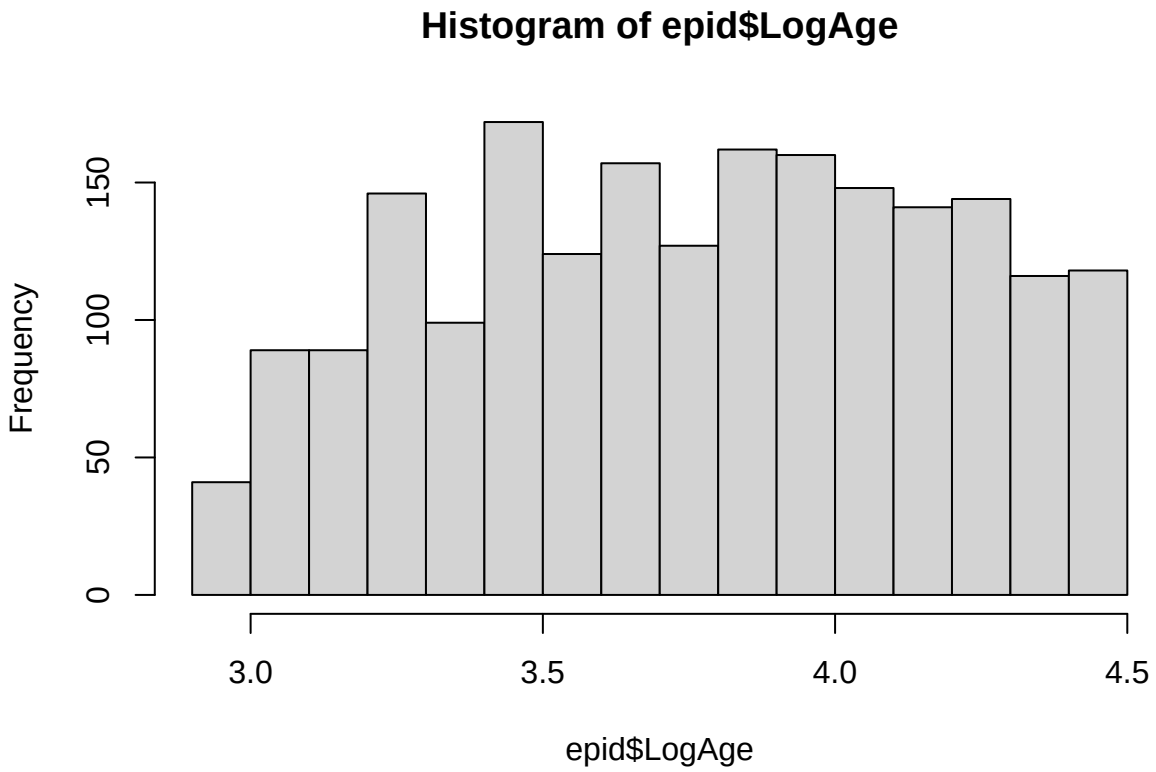


7. Create a new variable called LogAge containing the natural logarithm of the Age variable. (HINT: The log function computes the natural log, logAge=log(Age))

```r
epid <- mutate(epid, LogAge = log(Age))
```

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.

```r
# hist(epid$Age)
hist(epid$LogAge)
```



Taking the log of age changed the distribution of the Age variable from being more right-slewed to being more evenly distributed.

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

```
epid <- mutate(epid, over65 = ifelse(Age >= 65, 1, 0))
table(epid$over65)
```

```
##
##    0    1
## 1606  427
```

There are 427 individuals over 65 in the dataset.