Problem Set 1

Problem Set 1
Francisco Brady 2023-09-09

Question 1

The `anes2020` dataset has 8280 records.

```
nrow(anes2020)
```

```
## [1] 8280
```

Question 2

Below is the output of `summary(anes2020$SciImptCovid)`.

```
summary(anes2020$SciImptCovid)
```

```
## Not at all important   A little important Moderately important
##                  118                  341                 1183
##       Very important  Extremely important                 NA's
##                 1883                 3858                  897
```

Below is the output of `table(anes2020$SciImptCovid)`

```
table(anes2020$SciImptCovid)
```

```
##
## Not at all important   A little important Moderately important
##                  118                  341                 1183
##       Very important  Extremely important
##                 1883                 3858
```

(a) The difference between the two outputs is that the `summary()` command identified that the `SciImptCovid` variable has 897 missing values or NA's.

(b) It's important to take account of the missing values. We should be concerned about missing values because if there are a lot of missing values, we may introduce **nonresponse bias** into our understanding of the variable. It might not be possible to say anything meaningful about the data if there are a sufficient number of missing values. In the prompt it was mentioned that the `ScienceExperts` question was only asked in the post-election survey. The addition of that question, along with the `SciImpCovid` question may have impacted responses.

(c) The most common response to `SciImptCovid` is "Extremely Important".

(d) The `SciImptCovid` is a categorical variable who's responses form an ordinal scale. The "lowest" response is "Not at all important", and the "highest" value is "Extremely important".

Question 3

```
table(anes2020$EconWorse)
```

```
##
##     Much better Somewhat better  About the same  Somewhat worse      Much worse
##             750             800            1704            1764            3222
```

```
table(anes2020$DiscussPol)
```

```
##
##    0    1    2    3    4    5    6    7
##  245  933 1123 1011  664  714  285 1838
```

```
table(anes2020$HomeOwnership)
```

```
##
##              Homeowner                 Renter Some other arrangement
##                   5650                   2071                    442
```

```
table(anes2020$VotedBiden)
```

```
##
## Other Biden
##  2633  3270
```

```
table(anes2020$Empathy)
```

```
##
## Not often at all    Not too often   Somewhat often       Very often
##              313              778             2580             2484
##  Extremely often
##             1184
```

a) EconWorse – Categorical, Ordinal, Discrete

b) DiscussPol – Quantitative, Interval, Discrete

c) HomeOwnership – Categorical, Nominal, Discrete

d) VotedBiden – Categorical, Nominal, Discrete

e) Empathy – Categorical, Ordinal, Discrete

Question 4

a) EconWorse – Mode, Median

b) DiscussPol – Mean, Median, Mode

c) HomeOwnership – Mode

d) VotedBiden – Mode

e) Empathy – Median, Mode

Question 5

```
summary(anes2020$SCOTUStherm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   50.00   60.00   60.67   75.00  100.00     909
```

```
table(anes2020$SCOTUStherm)
```

```
##
##    0    1    2    3    5    6    7    8    9   10   15   20   25   28   30   33
##  144    2    3    2   10    2    2    2    3   14  166   27   24    1  312    1
##   35   40   43   44   45   47   48   49   50   51   52   55   56   59   60   61
##   30  756    1    1   43    1    1    5 1454    2    1   45    1    3 1082    1
##   62   64   65   66   67   68   69   70   71   75   77   78   80   83   85   86
##    2    1   75    4    5    4    2 1223    1  135    3    4  158    2  998   11
##   87   88   89   90   92   94   95   98  100
##    1    4    3  116    1    1   29    2  444
```

```
#max(table(anes2020$SCOTUStherm))
```

a) Mean: 60.67, Median: 60, Mode: 50

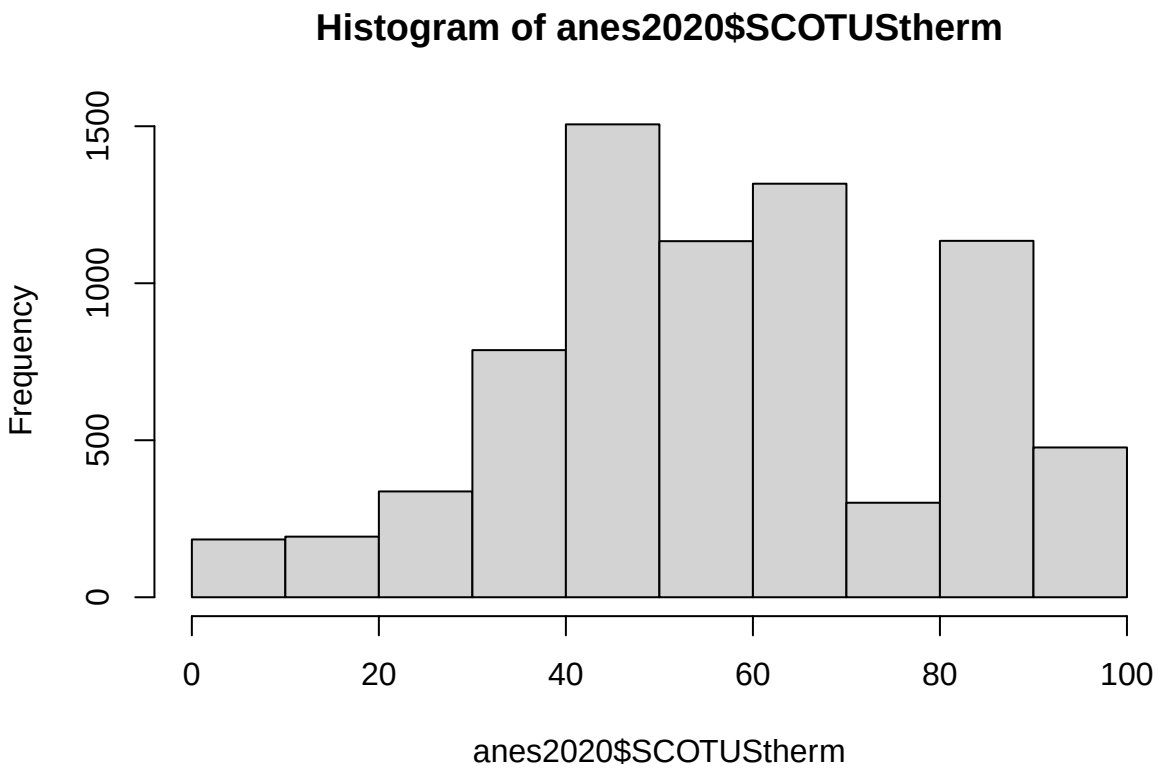
```
# using sd function to print standard deviation
sd(anes2020$SCOTUStherm, na.rm = T)
```

```
## [1] 21.83507
```

b) The standard deviation of this variable is 21.83507

c) A Histogram

```
hist(anes2020$SCOTUStherm)
```



Question 6

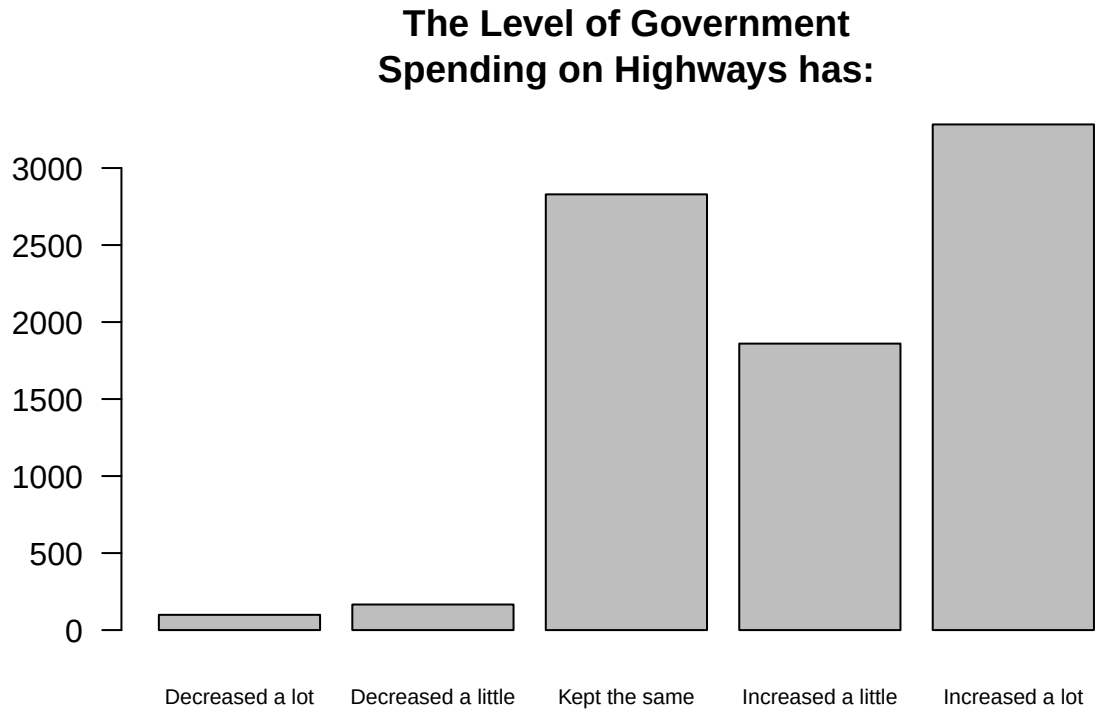
a) A frequency table for `SpendHighways`. The measurement level is categorical, ordinal, discrete.

```
table(anes2020$SpendHighways)
```

```
##
##    Decreased a lot Decreased a little      Kept the same Increased a little
##                 99                166               2829               1860
##    Increased a lot
##               3283
```

b) A bar graph for `SpendHighways`

```
barplot(table(anes2020$SpendHighways),
        las=1, cex.names=.7,
        main = 'The Level of Government\nSpending on Highways has:')
```



c) Treating the categorical variable as numeric, removes the labels from the variable. The counts correspond to the categories (Decreased a lot = 1, Decreased a little = 2, Kept the same = 3, Increased a little = 4, Increased a lot = 5). The ordering of the number does have meaning in that the response choices imply an increase in government spending. It's not possible to say whether each step on the scale is the same amount of change, so it would be difficult to interpret as an interval level variable.

```
table(as.numeric(anes2020$SpendHighways))
```

```
##
##    1    2    3    4    5
##   99  166 2829 1860 3283
```

d) Reporting the `SpendHighways` as a numeric variable:

```
summary(as.numeric(anes2020$SpendHighways))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   3.000   4.000   3.979   5.000   5.000      43
```

The calculation here is using the response as it is stored by R to calculate the summary statistics. Calculating the mean or standard deviation for these numbers would not be meaningful.

Question 7

a) For the first method, the democracy variable would be categorical, nominal, and discrete. For the second method, the democracy variable would be categorical, ordinal, and discrete.

b) It depends heavily on the context which measurement strategy is more reliable. Reliability is about how consistent the measure is. Since the first method relies on events that happen (elections), it could be more reliable. The second method may be less reliable because the values are contingent on a set of experts (humans) whose value assignment might be affected by many other circumstances or contexts.

c) Again this depends on the context and the aims of the measurement. Validity of a measurement is dependent on how well it represents the concept it is measuring. A rating from experts on civil liberties and political rights could be a good indicator of how democratic a country is. A binary measurement of whether a country is a democracy or not based solely on elections may not be as valid of a representation.

Question 8

a) For the first method, the measurement levels would be categorical ("favorable", "neutral", "unfavorable"), ordinal, and discrete. For the second method the measurement levels would be quantitative (count of "negative" key words), interval, and discrete.

b) Reliability might be a concern when using the first method, but could be addressed by having the graduate students overlap the news articles they read and assign codes to, since re-testing and getting the same grade would increase confidence in the reliability of the students' assessments. In the second method, since the measure is simply a count of key words, it seems like it would be more reliable.

c) In terms of validity, depending on the key words picked, I think the second method would be more valid, since it is relying on opinions a bit less. The fact that the measurement is being done by computer software also increases the ease of validation by other researchers.