Myranda Swartzwelter
Week 3
DSC-520 Stats for DS

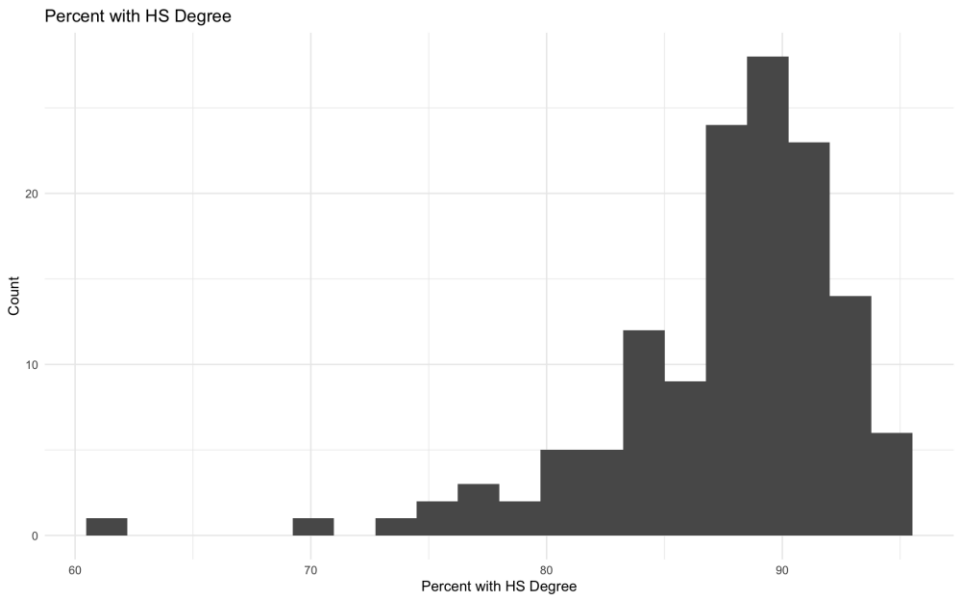
American Community Survey Exercise

1. What are the elements in your data? There are 136 observations of 8 variables.
    a. Id: Character
    b. Id2: integer
    c. Geography: character
    d. PopGroupID: integer
    e. POPGROUP.display.label: character
    f. RacesReported: integer
    g. HSDegree: number
    h. BachDegree: number
2. Provide output of str(), nrow(), ncol()
    i. str()

```
> str(comm_df)
'data.frame':   136 obs. of  8 variables:
 $ Id                  : chr  "0500000US01073" "0500000US04013" "0500000US04019" "0500000US06001" ...
 $ Id2                 : int  1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...
 $ Geography           : chr  "Jefferson County, Alabama" "Maricopa County, Arizona" "Pima County, Arizona" "Alameda County, California" ...
 $ PopGroupID          : int  1 1 1 1 1 1 1 1 1 1 ...
 $ POPGROUP.display.label: chr  "Total population" "Total population" "Total population" "Total population" ...
 $ RacesReported       : int  660793 4087191 1004516 1610921 1111339 965974 874589 10116705 3145515 2329271 ...
 $ HSDegree            : num  89.1 86.8 88 86.9 88.8 73.6 74.5 77.5 84.6 80.6 ...
 $ BachDegree          : num  30.5 30.2 30.8 42.8 39.7 19.7 15.4 30.3 38 20.7 ...
```

    j. nrow(): 136
    k. ncol(): 8
3. Create a historgram with a set bin size axis labels and a title
    l. R code:

```
gplot(comm_df, aes(HSDegree)) + geom_histogram(bins = 20)  + ggtitle('Percent with HS Degree') + xlab('Percent with HS Degree') + ylab('Count')
```



4. Answer the following questions based on the Histogram produced:
    m. Based on what you see in this histogram, is the data distribution unimodal?

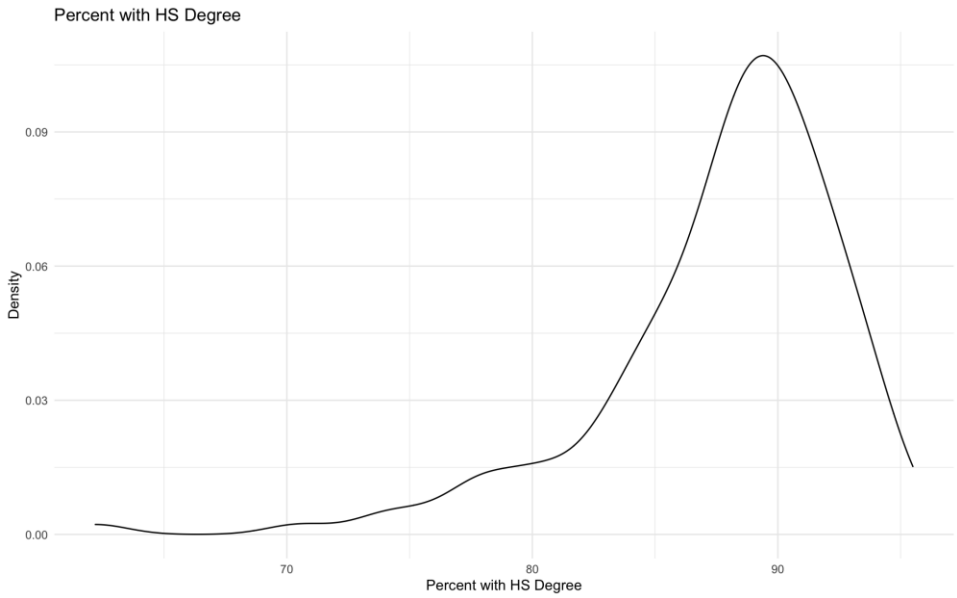
   i. Yes the data distribution has only one peak so the distribution is unimodal
2. Is it approximately symmetrical?
   i. The distribution is negatively skewed, so it is not symmetrical.
3. Is it approximately bell-shaped?
   i. The distribution is roughly bellshape but the tail extends farther left than it does right.
4. Is it approximately normal?
   i. No the distribution is skewed, so it is not normal.
5. If not normal, is the distribution skewed? If so, in which direction?
   i. Yes the distribution is skewed. It is left skewed (negatively skewed).
6. Include a normal curve to the Histogram that you plotted.

7. Explain whether a normal distribution can accurately be used as a model for this data.
   i. No a normal distribution can not accurately be used as a model for this data as the data is skewed negatively. It is a left skewed distribution, and the tails are not the same size, so it can not be viewed as a normal distribution.

5. Create a Probability Plot of the HSDegree variable.

R code:

```
ggplot(comm_df, aes(HSDegree)) + geom_density() + ggtitle('Percent with HS Degree') +
xlab('Percent with HS Degree') + ylab(Density)
```



6. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.
   a. No the probability plot is not approximately normal, although it is bell shaped, the left tail is significantly longer than the right tail.

7. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.
    a. Yes this distribution is left-skewed because the left tail is longer. It is negatively skewed.
8. Now that you have looked at this data visually for normality, you will now quantify normality with numbers using the stat.desc() function. Include a screen capture of the results produced.

```
> stat.desc(comm_df['HSDegree'])
                  HSDegree
nbr.val      1.360000e+02
nbr.null     0.000000e+00
nbr.na       0.000000e+00
min          6.220000e+01
max          9.550000e+01
range        3.330000e+01
sum          1.191800e+04
median       8.870000e+01
mean         8.763235e+01
SE.mean      4.388598e-01
CI.mean.0.95 8.679296e-01
var          2.619332e+01
std.dev      5.117941e+00
coef.var     5.840241e-02
```

9. In several sentences provide an explanation of the result produced for skew, kurtosis, and z-scores. In addition, explain how a change in the sample size may change your explanation?
    a. The skew of HSDegree is -1.69341. Since it is negative, it confirms what I saw in the histogram, and the distribution is left-skewed. The kurtosis is 7.462191. Since this is higher than 3, this indicates there are more values in the tails than we would find in a normal distribution. Since zscores are a shift of the data, but don't change the position of the data points, we would expect to see more negative zscores than positive. A change in sample size may change the data set if the new data points created a normal distribution, and we saw the skewedness and kurtosis decrease.