American community survey exercise

What are the elements in your data (including the categories and data types)?

To answer this question, I checked structure of my data frame.

```
# Load data into data frame
data <- read.csv("acs-14-1yr-s0201.csv")

# check structure of the data
> str(data)
'data.frame':   136 obs. of  8 variables:
 $ Id                : Factor w/ 136 levels "0500000US01073",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Id2               : int  1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...
 $ Geography         : Factor w/ 136 levels "Alameda County, California",..: 56 70 98 1 20 43 62
68 92 106 ...
 $ PopGroupID        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ POPGROUP.display.label: Factor w/ 1 level "Total population": 1 1 1 1 1 1 1 1 1 1 ...
 $ 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 ...
```

Observations

1. Total number of rows or observations = 136
2. Total number of variables of columns = 8
3. Column / Field details
    a. Id is a category/factor with 136 levels i.e., unique values. Because this is alphanumeric field it's assumed as factor by R. This can be changed to character or string as this appears to be unique id.
    b. Id2 contains numeric data and is rightly interpreted as int
    c. Geography is interpreted as factor and contains 136 unique values of County name, State name in United States. Splitting this column will help exploring data distributions at state level as well.
    d. PopGroupID is numeric field, and it has 1 for all the observations.
    e. POPGROUP.display.label is interpreted as category/factor and contains "Total Population" for all records. Level is 1 and thus there exits only 1 unique value.
    f. RacesReported contains numeric values, and it appears to be population corresponding to certain races in each county, state.

g. HSDegree is numeric field, and it appears to be percentage of population who has high school degree in each county, state. As there is no % sign it is rightly interpreted as integer.
h. BachDegree is numeric field, and it appears to be percentage of population who has bachelor's degree in each county, state. As there is no % sign it is rightly interpreted as integer.

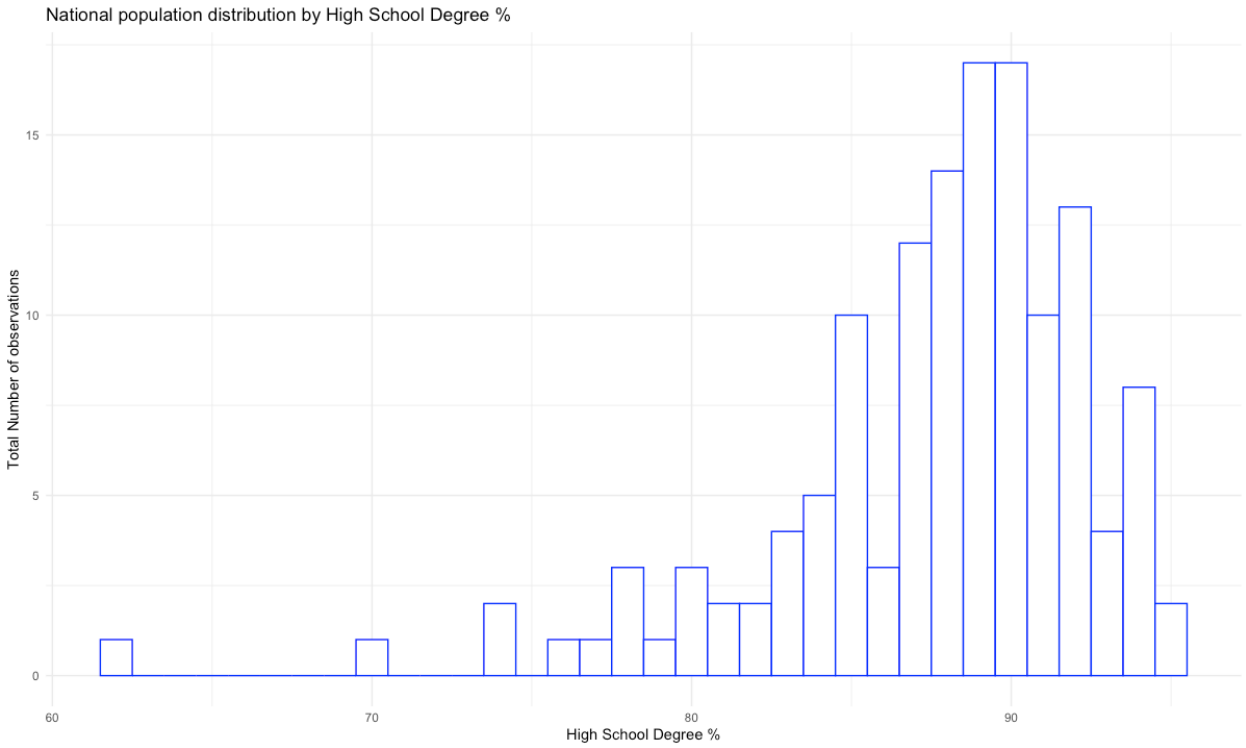
Please provide the output from the following functions: str(); nrow(); ncol()

```
> str(data)
'data.frame':   136 obs. of  8 variables:
 $ Id                : Factor w/ 136 levels "0500000US01073",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Id2               : int  1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...
 $ Geography         : Factor w/ 136 levels "Alameda County, California",..: 56 70 98 1 20 43 62
68 92 106 ...
 $ PopGroupID        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ POPGROUP.display.label: Factor w/ 1 level "Total population": 1 1 1 1 1 1 1 1 1 1 ...
 $ 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 ...

> nrow(data)
[1] 136

> ncol(data)
[1] 8
```

Create a Histogram of the HSDegree variable using the ggplot2 package.

1. Set a bin size for the Histogram.
2. Include a Title and appropriate X/Y axis labels on your Histogram Plot.

```
ggplot(data = data, aes(x=HSDegree)) +
  geom_histogram(bins = 20, color = "Blue", fill = "White", binwidth = 1) +
  ggtitle("National population distribution by High School Degree %") +
  xlab("High School Degree %") +
  ylab("Total Number of observations")
```



Answer the following questions based on the Histogram produced:

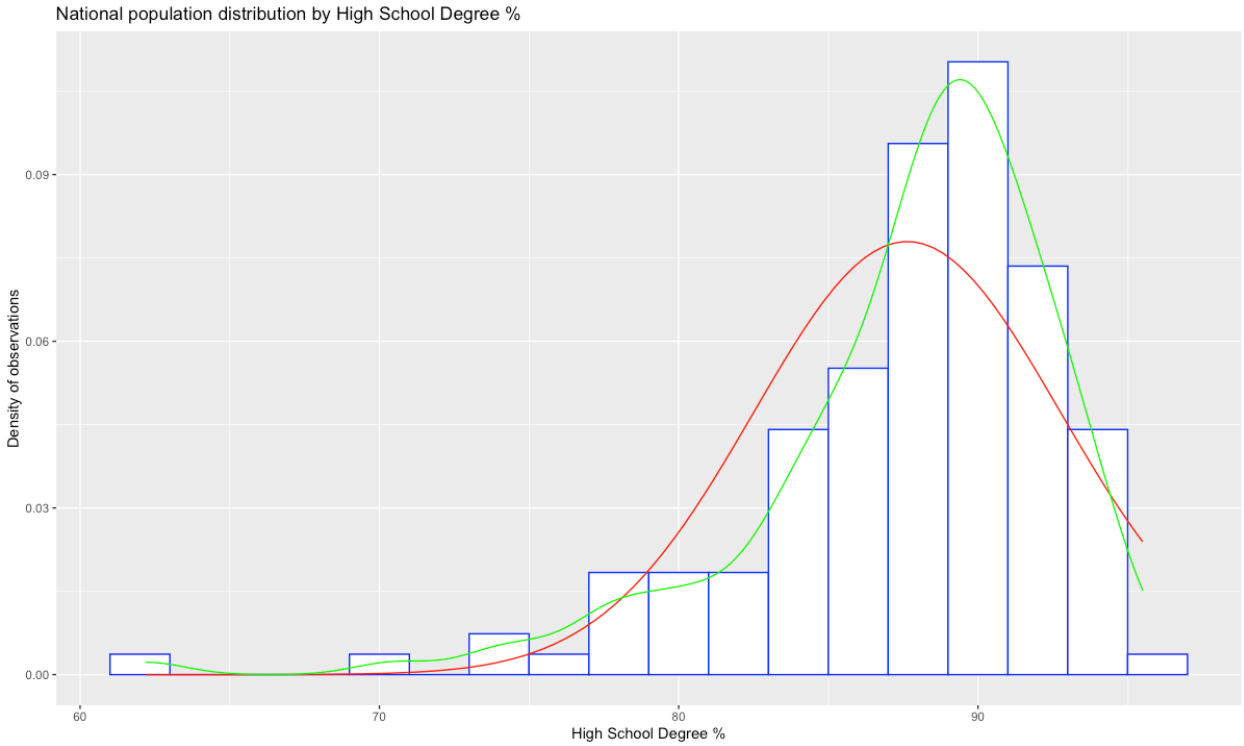
1. Based on what you see in this histogram, is the data distribution unimodal?
2. Is it approximately symmetrical?
3. Is it approximately bell-shaped?
4. Is it approximately normal?
5. If not normal, is the distribution skewed? If so, in which direction?
6. Include a normal curve to the Histogram that you plotted.
7. Explain whether a normal distribution can accurately be used as a model for this data.

1. No, data is not unimodal, as we have most data points at 89 as well as 90.
2. Data looks unsymmetrical with negative skew (long left tail). Thus, most observations are concentrated to the right.
3. Data cannot be considered approximately bell shaped as it has negative skew.
4. Data cannot be considered approximately normal as it has negative skew.
5. Data is skewed to the left and thus it's negatively skewed. Most observations are towards right and thus mean, median and mode are towards right as well.
6. To include normal curve to the histogram, had to do following –
    a. By default, histogram is plotting counts / frequency of observations on the y-axis. We need to change that to probability density instead.
    b. Use stat_function() to plot normal curve with mean & standard deviation of HSDegree.
    c. Use geom_density() to plot the probability density curve representing the data at hand. This will help eye further to see how far data is from is normal plot.

Code -

```
ggplot(data = data, aes(x=HSDegree)) +
 geom_histogram(bins = 20,
                color = "Blue",
                fill = "White",
                binwidth = 2,
                aes(y=..density..)) +
 ggtitle("National population distribution by High School Degree %") +
 xlab("High School Degree %") +
 ylab("Density of observations") +
 stat_function(fun = dnorm,
               color = "Red",
               args = list(mean=mean(data$HSDegree, na.rm = TRUE),sd=sd(data$HSDegree, na.rm =
TRUE))) +
 geom_density(color="Green")
```

Plot –



7. Normal distribution cannot be used as a model for this data, as clearly in the plot above we can see that density curve of data at hand (green) is different from corresponding normal density plot (red). Density curve is negatively skewed with fatter tail on left side and observations concentrated to the right. Mean, Median, Mode of density curve are all seems to appear on the right-side normal curve's mean. Also, density curve has higher kurtosis and thus a pointy curve with the flatter tail in comparison to normal curve. For a normal distribution both skewness and kurtosis should be 0.

   I installed moments library to calculate exact kurtosis and skewness also looked up to ranges that defines them.

   > skewness(data$HSDegree)
   [1] -1.69341

   Usually below rules are considered to interpret skewness.
   - Perfect normal: Value is 0
   - Approx. Symmetric: Values between -0.5 to 0.5
   - Moderated Skewed data: Values between -1 and -0.5 or between 0.5 and 1
   - Highly Skewed data: Values less than -1 or greater than 1

   As skew for our data field is less than -1, it is highly negatively skewed with most observations concentrated to the right and tail on left side.

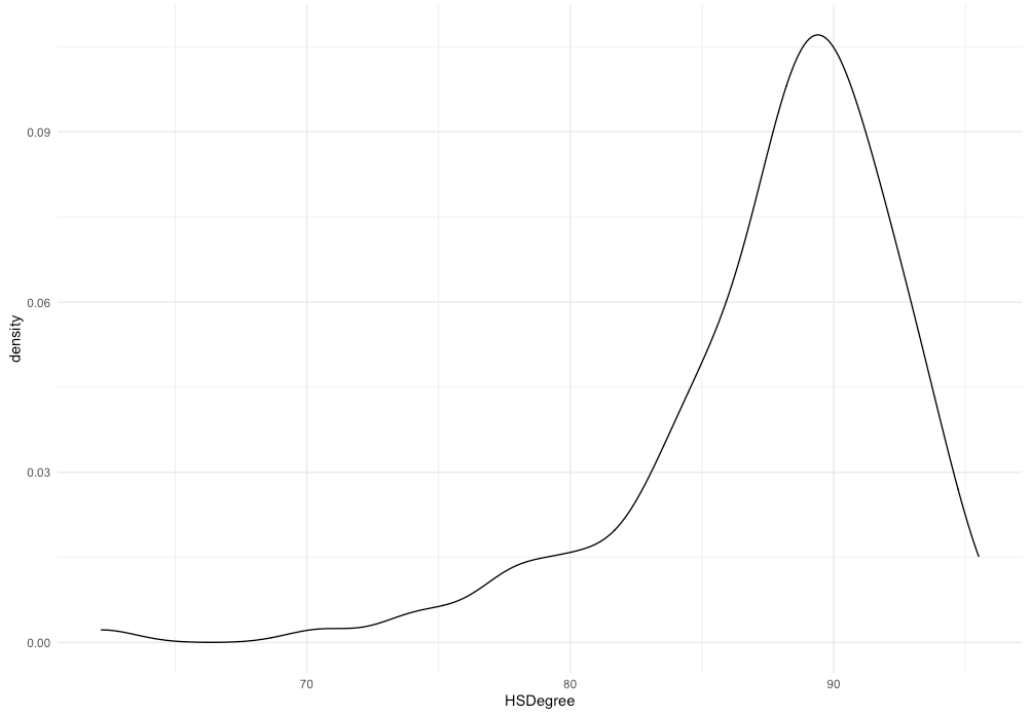
> kurtosis(data$HSDegree)
[1] 7.462191

Usually below rules are considered to interpret kurtosis.

- Mesokurtic: This is the normal distribution with Kurtosis value 0.
- Leptokurtic: This distribution has fatter tails and a sharper peak. The kurtosis is "positive" with a value greater than 3
- Platykurtic: The distribution has a lower and wider peak and thinner tail. The kurtosis is "negative" with a value greater than 3

Kurtosis value of our data is > 3 and as we can see it peaked and thus represents leptokurtic.

Create a Probability Plot of the HSDegree variable.



Answer the following questions based on the Probability Plot:

1. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.
2. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.

1. Distribution as seen in the probability plot above is not normal as left tail is fatter and data is negatively skewed. It is not symmetric like it would be in normal distribution curve. Most of data points are concentrated to the right. It is also pretty sharply peaked which means we have high kurtosis or leptokurtic.
2. Distribution is skewed. By looking at the density curve we can say that it's not symmetrical and it has fatter left tail. Data is negatively skewed. Most of the data points are concentrated to right with mean, median, and mode.

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.

```
19 library(pastecs)
20 # change the format of display by setting the options
21 options(scipen=100)
22 options(digits=2)
23 stat.desc(data$HSDegree)
24 stat.desc(data$HSDegree, basic = FALSE, norm = TRUE)
25
28:1  (Top Level)

Console  Terminal  Jobs
~/Desktop/R Programming/DSC520/stats_for_data_science/stats_for_data_science/
> options(digits=2)
> stat.desc(data$HSDegree)
     nbr.val     nbr.null       nbr.na          min          max        range
     136.000        0.000        0.000       62.200       95.500       33.300
         sum       median         mean      SE.mean CI.mean.0.95          var
   11918.000       88.700       87.632        0.439        0.868       26.193
     std.dev     coef.var
       5.118        0.058
> stat.desc(data$HSDegree, basic = FALSE, norm = TRUE)
       median          mean       SE.mean  CI.mean.0.95           var
88.7000000000 87.6323529412  0.4388597852  0.8679296080 26.1933159041
      std.dev      coef.var      skewness      skew.2SE      kurtosis
 5.1179405921  0.0584024098 -1.6747666105 -4.0302539978  4.3528564623
     kurt.2SE    normtest.W    normtest.p
 5.2738853364  0.8773635436  0.0000000032
>
```

So, looking at numbers above we can see that distribution ranges from 62.2 (min) to 95.5(max). Both mean (87.6) and, median (88.7) are closer to max value and thus away from min.

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?

1. Standard deviation = 5.118
2. Mean = 87.6
3. Skewness = -1.67
4. Kurtosis = 4.35
5. Skew.2SE = -4.03. This is Z-score corresponding to the skew.
6. Kurtosis.2SE = 5.27. This is Z-score corresponding to the kurtosis.

As a rule of thumb, for skew –

- For perfectly normal distribution: Value should be 0
- Symmetric: Values between -0.5 to 0.5
- Moderated Skewed data: Values between -1 and -0.5 or between 0.5 and 1
- Highly Skewed data: Values less than -1 or greater than 1

Skew in our data is -1.67 and thus it is highly skewed negatively i.e., long tail on left with most data points concentrated on right.

As a rule of thumb, for kurtosis –

- Mesokurtic: This is the normal distribution; value should be 0.
- Leptokurtic: This distribution has fatter tails and a sharper peak. The kurtosis is "positive" with a value greater than 3
- Platykurtic: The distribution has a lower and wider peak and thinner tail. The kurtosis is "negative" with a value greater than 3

Kurtosis in our data is 4.35 which is > 3 and thus distribution has fatter tails and a sharper peak and is Leptokurtic.

Now, both skewness and kurtosis values can be converted to z-scores. Z-score is simply the score from a distribution that has mean of 0 and standard deviation of 1. By converting skew and kurtosis to their z-score values we would know how likely the values are to occur.

Z score for skew = Skew.2SE
Z score for kurtosis = Kurtosis.2SE

If absolute values of Skew.2SE and Kurtosis.2SE are > 1 then they are considered significant with p value>0.05. In our case both these values > 1 and thus both skew and kurtosis are significant. As we have small sample of data (136 observations) looking at these stats helps. If sample size is large it is recommended to look visuals as well i.e., histograms, density charts etc to know about normality.