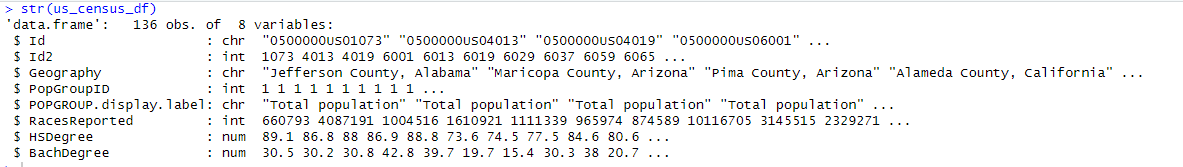
**Week-3: Assignment\_03**

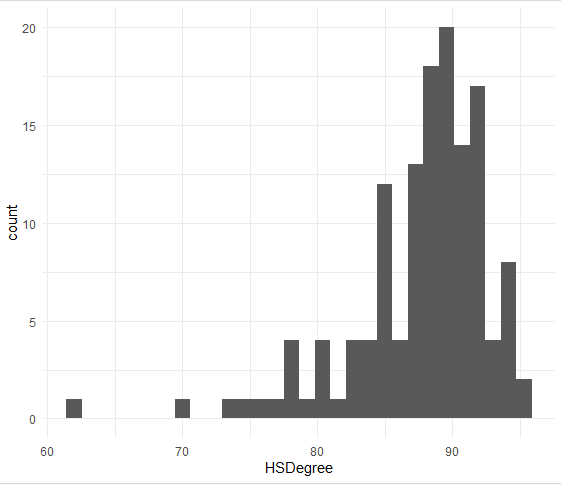
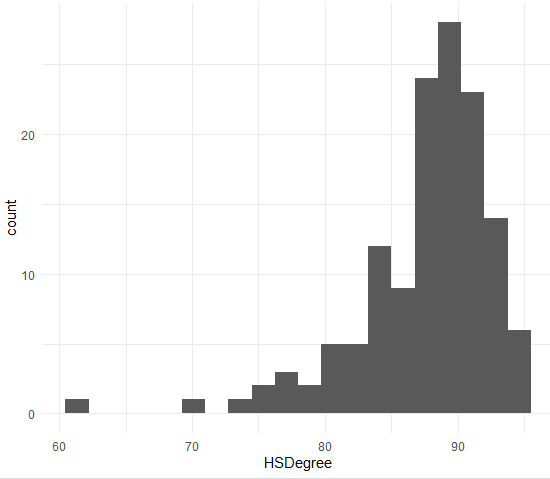
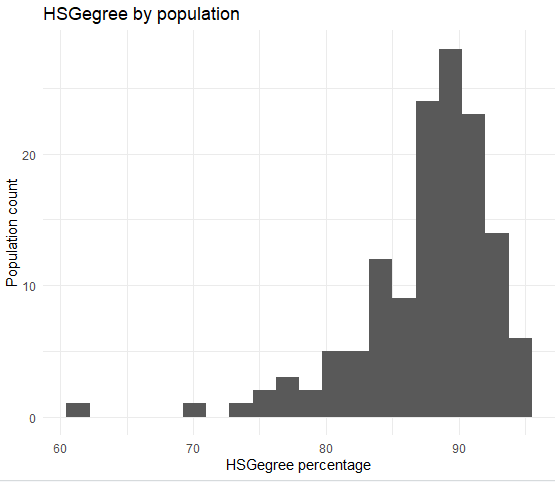
# Kalyan Pothineni

1. List the name of each field and what you believe the data type and intent is of the data included in each field (Example: Id - Data Type: varchar (contains text and numbers) Intent: unique identifier for each row)

* Id – Data Type: varchar (contains text and numbers) Intent: unique identifier for each row
* Id2 – Data Type: integer (contains numbers) Intent: unique identifier for each row
* Geography – Data Type: char (contains characters) Intent: geography location – county and state
* PopGroupID – Data Type: integer (contains numbers) Intent: how the population is grouped linked to population group description
* POPGROUP.display-label – Data Type: char (contains numbers) Intent: how the population is grouped description
* RacesReported – Data Type: integer (contains numbers) Intent: total count of population reported
* HSDegree – Data Type: numeric (contains numbers) Intent: percentage of population with high school degree
* BachDegree – Data Type: numeric (contains numbers) Intent: percentage of population with bachelor’s degree

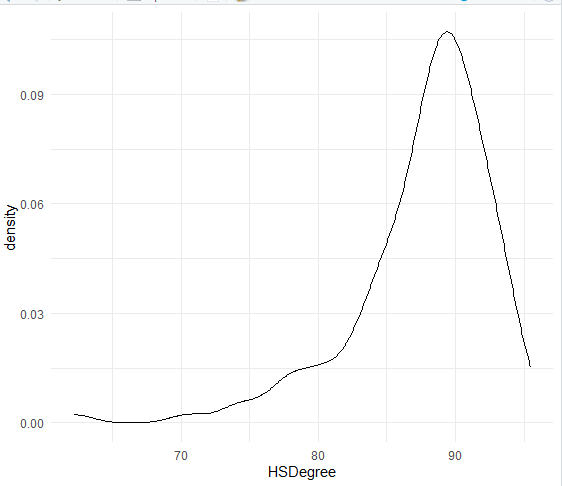
1. Run the following functions and provide the results: str(); nrow(); ncol()  
   str()

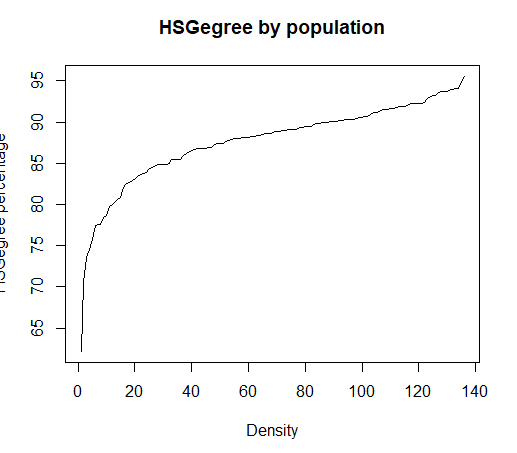
  
nrow()  
  
ncol()  


1. Create a Histogram of the HSDegree variable using the ggplot2 package.  
   
2. Set a bin size for the Histogram that you think best visuals the data (the bin size will determine how many bars display and how wide they are) – bin size of 20  
   
3. Include a Title and appropriate X/Y axis labels on your Histogram Plot.  
   
4. Answer the following questions based on the Histogram produced:

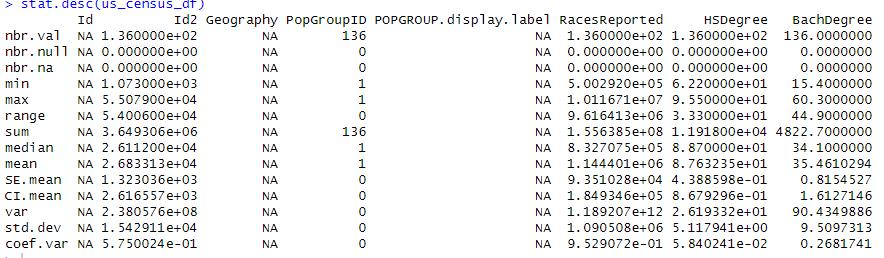
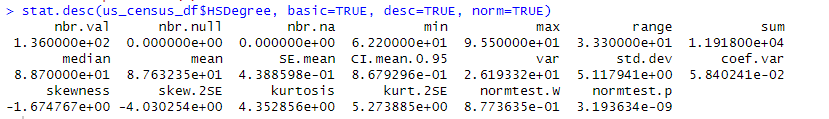
* Based on what you see in this histogram, is the data distribution unimodal?

No, the data is not distributed unimodal. There are skews in between

* Is it approximately symmetrical? No, it is not approximately symmetrical, as it is not balanced on left or right tails
* Is it approximately bell-shaped? No, it is not approximately bell shaped
* Is it approximately normal? No, as the bars roughly did not follow a bell shape
* If not normal, is the distribution skewed? If so, in which direction? Yes, in left tail direction
* Include a normal curve to the Histogram that you plotted.
* Explain whether a normal distribution can accurately be used as a model for this data.  
  Cannot use normal distribution, as the plot is not symmetrical around its mean value. It is not visually depicted as the bell curve  
  

1. Create a Probability Plot of the HSDegree variable.  
   
2. Answer the following questions based on the Probability Plot:

* Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know. No, it is not normal, as the plot is not symmetrical at the mean value
* If not normal, is the distribution skewed? If so, in which direction? Explain how you know. It is distributed skewed, it is the left tail, as it is much longer than right

1. 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.  
     
   
2. 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?  
     
   **Skewness:** As a measure of the asymmetry of the probability distribution of a random variable about its mean. Skewness tells you the amount and direction of skew. The skewness value can be positive or negative, or even undefined. If skewness is 0, the data are perfectly symmetrical, although it is quite unlikely for real-world data. In the above data, we can see skewness is -1.6747, according to the distribution this is a highly skewed scenario

As a rule of thumb:

* If skewness is less than -1 or greater than 1, the distribution is highly skewed.
* If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
* If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

**Kurtosis:** As per definition, if a positive value for the kurtosis indicates a distribution more peaked than normal. In contrast, a negative kurtosis indicates a shape flatter than normal. Analogous to the skewness, the general guideline is that if the kurtosis is greater than +2, the distribution is too peaked. In the above case, Kurtosis is 4.352, the distribution is too peaked  
  
A **z score** is applied for normality test using skewness and kurtosis. A z-score could be obtained by dividing the skew values or excess kurtosis by their standard errors

**Z = Skew Value/SE (Skewness)**Z = -1.6747/-4.0302 = 0.415  
  
Also, for medium-sized samples (50 < n < 300), reject the null hypothesis at absolute z-value and conclude the distribution of the sample is non-normal