For this assignment, you will need to load and activate the ggplot2 package. For this deliverable, you should provide the following:

1. **What are the elements in your data (including the categories and data types)?**
2. Please provide the output from the following functions: str(); nrow(); ncol()

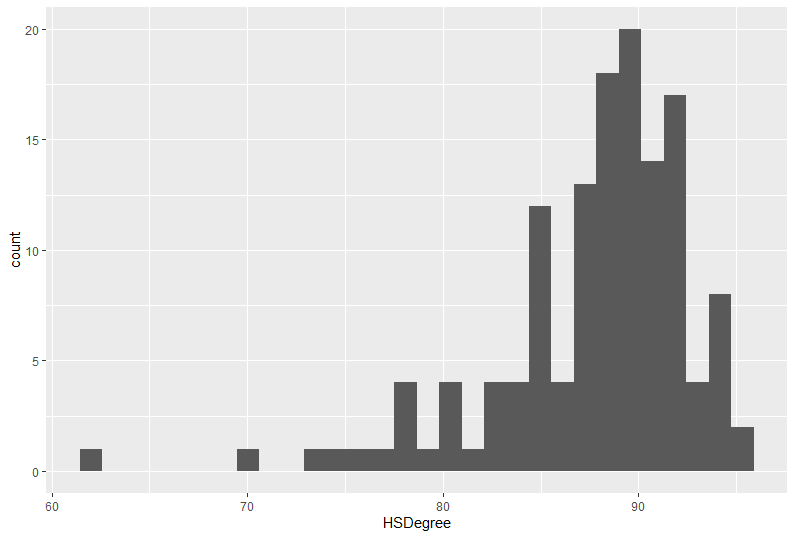
**Answer for Question 1 & 2 :**

8 elements in our data which including Id, Geography, Population Group Id, Label, Races Reported id, HS and bachelor’s degree. Please find the below details for data type and values.

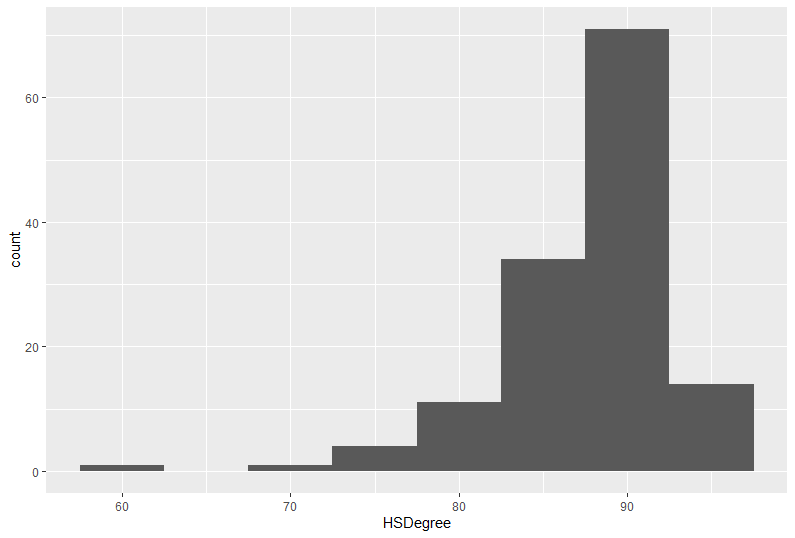
|  |
| --- |
| > setwd("C:/Users/ragun/Documents/GitHub/dsc520-master/DSC520-new")  > acs\_2014 <- read.csv("Data/acs-14-1yr-s0201.csv")  > str(acs\_2014)  'data.frame': 136 obs. of 8 variables:  $ Id : chr "0500000US01073" "0500000US04013" "0500000US04019" "0500000US06001" ...  $ Id2 : int 1073 4013 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 ...  > nrow(acs\_2014)  [1] 136  > ncol(acs\_2014)  [1] 8 |

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

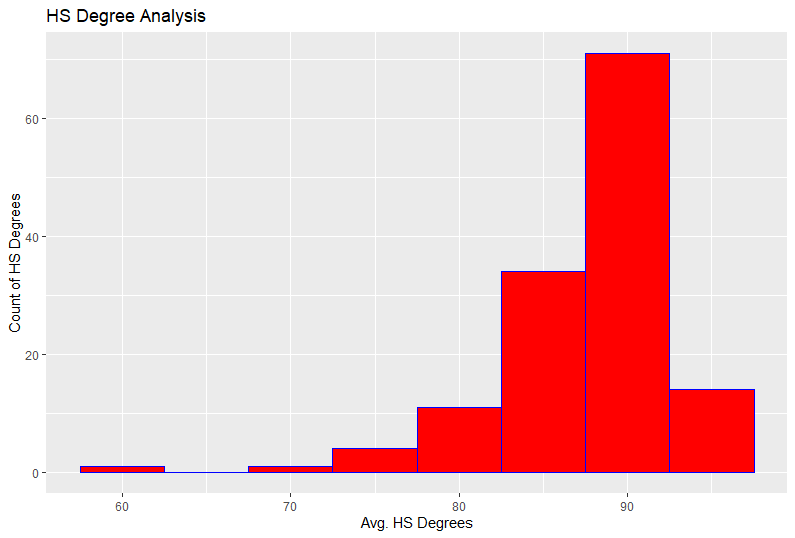
|  |
| --- |
| library(ggplot2)  ggplot(acs\_2014, aes(x = HSDegree)) + geom\_histogram()  # a. Set a bin size for the Histogram  ggplot(acs\_2014, aes(x = HSDegree)) + geom\_histogram(binwidth = 5)  # b. Include a Title and appropriate X/Y axis labels on your Histogram Plot.  ggplot(acs\_2014, aes(x = HSDegree)) + geom\_histogram(binwidth = 5, color="blue", fill="red") +  labs(title = "HS Degree Analysis", x = "Avg. HS Degrees", y="Count of HS Degrees") |



**a. Set a bin size for the Histogram.**



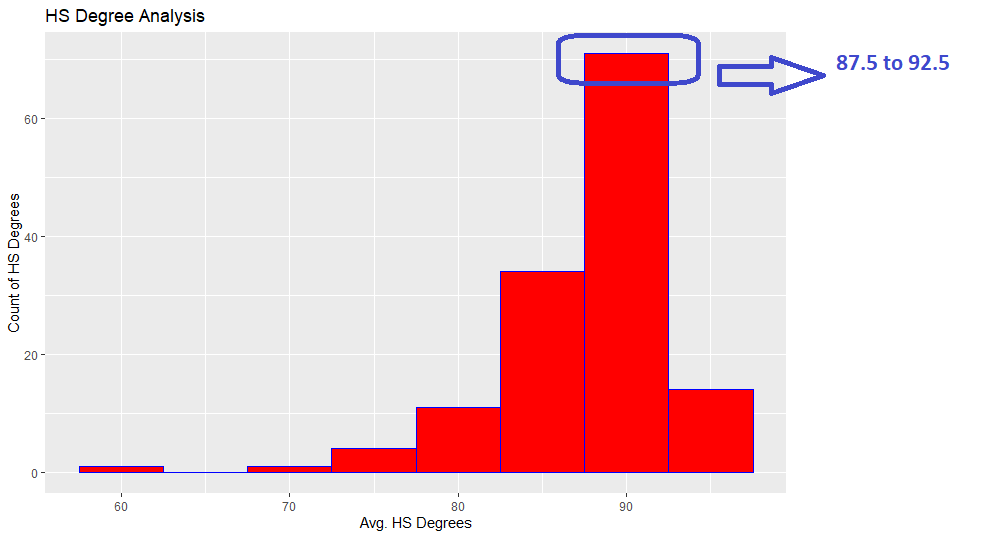
**b. Include a Title and appropriate X/Y axis labels on your Histogram Plot.**



4. Answer the following questions based on the Histogram produced:

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

Answer : Yes



b. Is it approximately symmetrical?

Answer : No, It is not symmetrical

c. Is it approximately bell-shaped?

Answer : No, it’s not bell Shaped

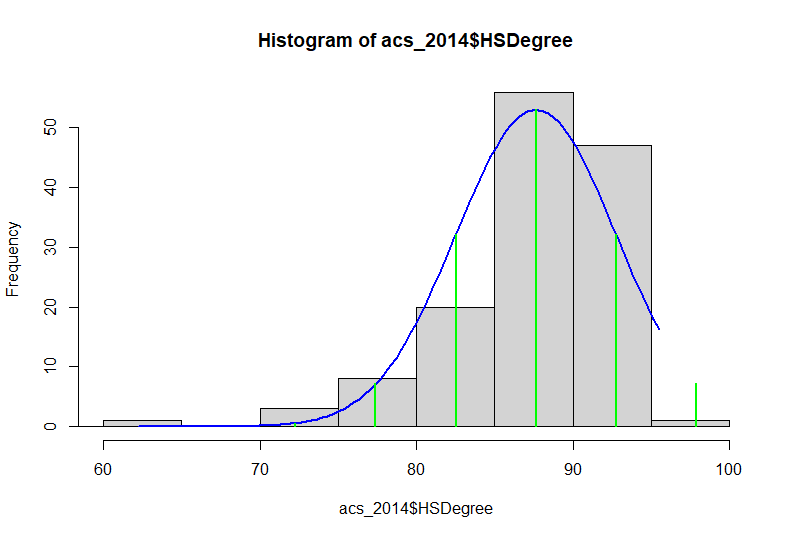
d. Is it approximately normal?

Answer : No, It’s not symmetrical. Hence not normal too

e. If not normal, is the distribution skewed? If so, in which direction?

Answer : Yes, the distribution is skewed. Negative Skewed

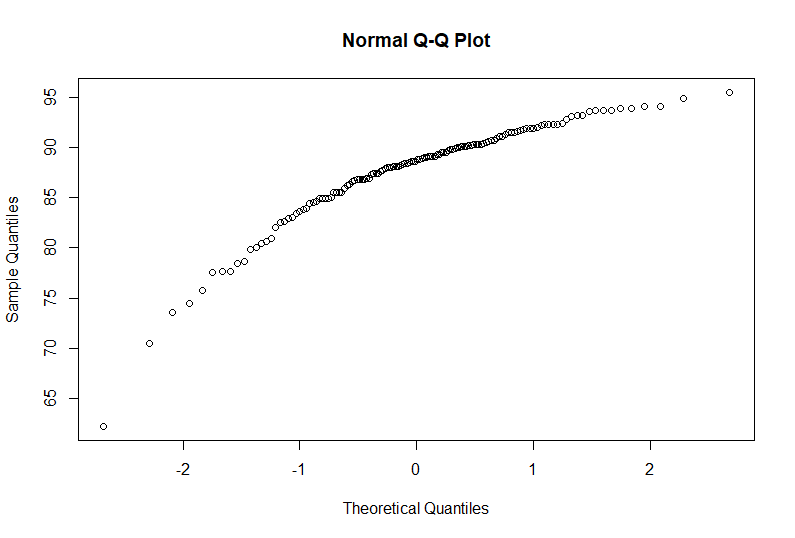
f. Include a normal curve to the Histogram that you plotted.



g. Explain whether a normal distribution can accurately be used as a model for this data.

Answer : As per the graph, we can not say that normal distribution since its negatively skewed. We can see the graph and histogram are not in sync.

1. Create a Probability Plot of the HSDegree variable.



6. Answer the following questions based on the Probability Plot:

a. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.

Answer : No, It is not normal

b. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.

Answer : Negative Skewed

7. 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.

|  |
| --- |
| > library(pastecs)  > stat.desc(acs\_2014)  Id Id2 Geography PopGroupID POPGROUP.display.label RacesReported HSDegree BachDegree  nbr.val NA 1.360000e+02 NA 136 NA 1.360000e+02 1.360000e+02 136.0000000  nbr.null NA 0.000000e+00 NA 0 NA 0.000000e+00 0.000000e+00 0.0000000  nbr.na NA 0.000000e+00 NA 0 NA 0.000000e+00 0.000000e+00 0.0000000  min NA 1.073000e+03 NA 1 NA 5.002920e+05 6.220000e+01 15.4000000  max NA 5.507900e+04 NA 1 NA 1.011671e+07 9.550000e+01 60.3000000  range NA 5.400600e+04 NA 0 NA 9.616413e+06 3.330000e+01 44.9000000  sum NA 3.649306e+06 NA 136 NA 1.556385e+08 1.191800e+04 4822.7000000  median NA 2.611200e+04 NA 1 NA 8.327075e+05 8.870000e+01 34.1000000  mean NA 2.683313e+04 NA 1 NA 1.144401e+06 8.763235e+01 35.4610294  SE.mean NA 1.323036e+03 NA 0 NA 9.351028e+04 4.388598e-01 0.8154527  CI.mean NA 2.616557e+03 NA 0 NA 1.849346e+05 8.679296e-01 1.6127146  var NA 2.380576e+08 NA 0 NA 1.189207e+12 2.619332e+01 90.4349886  std.dev NA 1.542911e+04 NA 0 NA 1.090508e+06 5.117941e+00 9.5097313  coef.var NA 5.750024e-01 NA 0 NA 9.529072e-01 5.840241e-02 0.2681741 |

8. 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?

Answer : This data is negative Skewed. As we know, the sample size increases, the margin of error started decreases. ( inverse because the two moves in opposite directions )