Michael Herrera

Bellevue University

DSC 520: 2014 American Community Survey

I. What are the elements in your data?
Numerical, Categorical, and Binary.

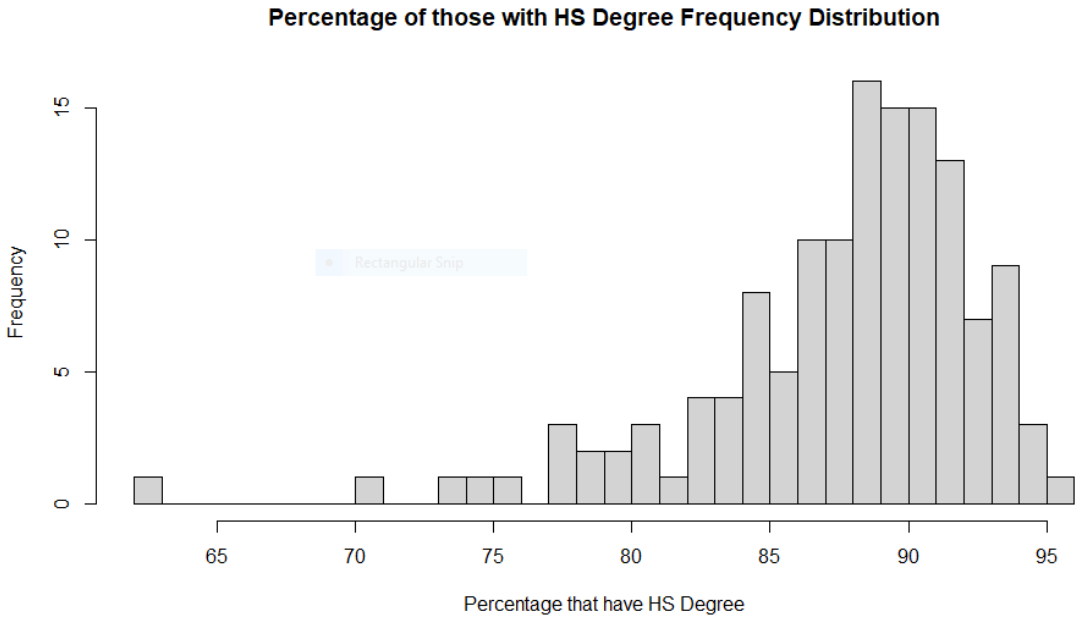
II. Please Provide the output from the following functions: str(); nrow(); ncol()

```
> str(Data1)
tibble [136 x 8] (S3: tbl_df/tbl/data.frame)
 $ Id                : chr [1:136] "0500000US01073" "0500000US04013"
"0500000US04019" "0500000US06001" ...
 $ Id2               : num [1:136] 1073 4013 4019 6001 6013 ...
 $ Geography         : chr [1:136] "Jefferson County, Alabama"
"Maricopa County, Arizona" "Pima County, Arizona" "Alameda
County, California" ...
 $ PopGroupID        : num [1:136] 1 1 1 1 1 1 1 1 1 1 ...
 $ POPGROUP.display-label: chr [1:136] "Total population" "Total
population" "Total population" "Total population" ...
 $ RacesReported     : num [1:136] 660793 4087191 1004516 1610921
1111339 ...
 $ HSDegree          : num [1:136] 89.1 86.8 88 86.9 88.8 73.6 74.5 77.5
84.6 80.6 ...
 $ BachDegree        : num [1:136] 30.5 30.2 30.8 42.8 39.7 19.7 15.4 30.3
38 20.7 ...

> nrow(Data1)
[1] 136

> ncol(Data1)
[1] 8
```

III. 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 Histrogram Plot.

```
Data1 <- read.csv("~/acs-14-1yr-s0201.csv")

x <- Data1$HSDegree
```

hist(x, breaks = 40, xlab = "Percentage that have HS Degree", ylab = "Frequency", main = "Percentage of those with HS Degree Frequency Distribution")



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

1. Based on what you see in this histogram, is the data distribution unimodal?
   No
2. Is it approximately symmetrical?
   No
3. Is it approximately bell-shaped?
   Yes
4. Is it approximately normal?
   No
5. If not normal, is the distribution skewed? If so, in which direction?
   Skewed to the Right

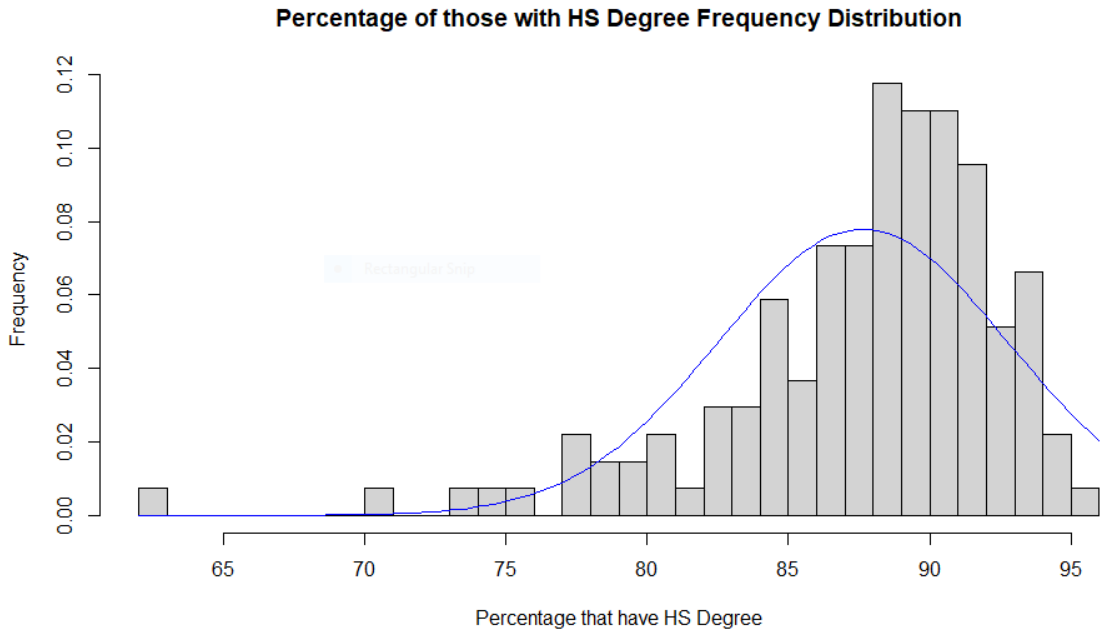
6. Include a normal curve to the Histogram that you plotted.

hist(x, breaks = 40, xlab = "Percentage that have HS Degree", ylab = "Frequency", main = "Percentage of those with HS Degree Frequency Distribution", freq = FALSE)

mean1 <- mean(x)

std1 <- sd(x)

curve(dnorm(x, mean = mean1, sd = std1), add = TRUE, col = "blue")



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

   A normal distribution cannot accurately be used since a HS degree seems to be common; in other words, skewing to the right.

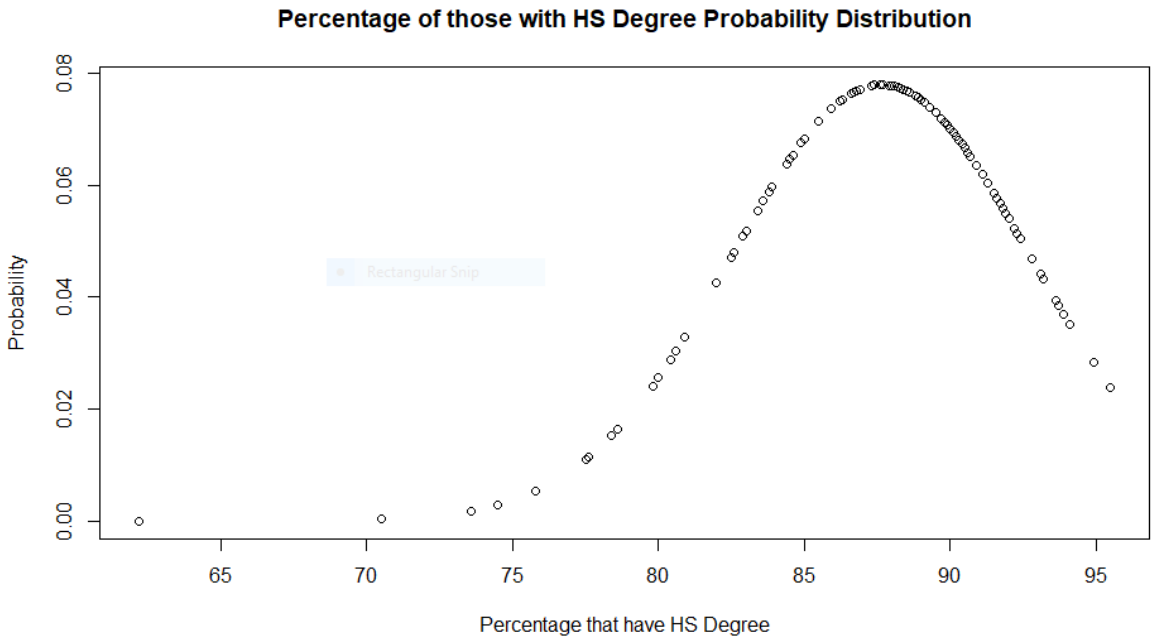
V. Create a Probability Plot of the HSDegree variable.

mean1 <- mean(x)

std1 <- sd(x)

nor <- dnorm(x,mean1,std1)

plot(x, nor, xlab = "Percentage that have HS Degree", ylab = "Probability", main = "Percentage of those with HS Degree Probability Distribution")



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

1. Based on what you see in this probability plot, is the distribution approximately normal?
   It does not appear so due to the fact that the majority of the values are towards the right side.
2. It not normal, is the distribution skewed? If so, in which direction? Explain how you know.

It is not normal and skewed right. The values visually showcase that the majority of them are to the right of the distribution.

VII. 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(x, norm = TRUE)
      nbr.val      nbr.null        nbr.na
 1.360000e+02  0.000000e+00  0.000000e+00
          min           max         range
 6.220000e+01  9.550000e+01  3.330000e+01
          sum        median          mean
 1.191800e+04  8.870000e+01  8.763235e+01
      SE.mean  CI.mean.0.95           var
 4.388598e-01  8.679296e-01  2.619332e+01
      std.dev     coef.var      skewness
 5.117941e+00  5.840241e-02 -1.674767e+00
     skew.2SE      kurtosis     kurt.2SE
-4.030254e+00  4.352856e+00  5.273885e+00
   normtest.W    normtest.p
 8.773635e-01  3.193634e-09
> |
```

VIII. 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 negative skewness shows that the distribution is skewed towards the right. A positive kurtosis would indicate a heavier tail. Z-scores could be generated to see how each value fits under the normal distribution. A change in the sample size with more individuals would most likely skew furthermore as HS degree is the common standard for many individuals in today's day in age.