DSC520 – Week 3 Assignment (Part2): Janine Par

Github:
https://github.com/Tutay0913JP/dsc520/blob/master/completed_janinepar/assignment_03_2_ParJanine.R

American Community Survey Exercise

For this exercise, you will use the following dataset, 2014 American Community Survey. This data is maintained by the US Census Bureau and are designed to show how communities are changing. Through asking questions of a sample of the population, it produces national data on more than 35 categories of information, such as education, income, housing, and employment. For this assignment, you will need to load and activate the ggplot2 package. For this deliverable, you should provide the following:

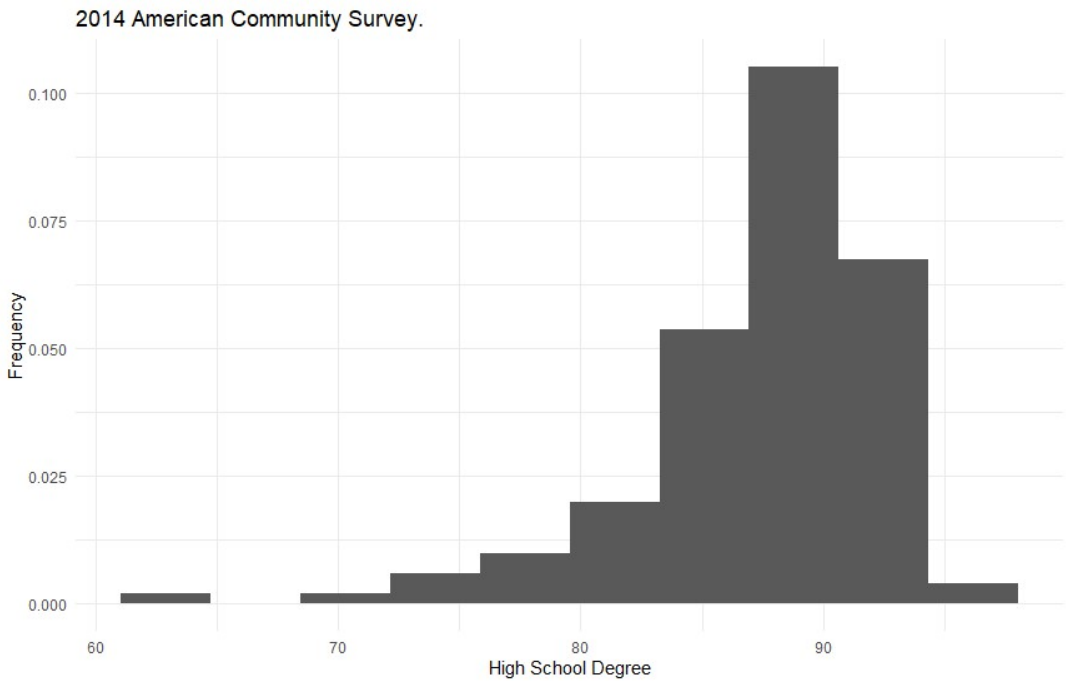
I. What are the elements in your data (including the categories and data types?

Variable	Category	Data type
Id	Categorical Nominal	Character
Id2	Categorical Nominal	Integer
Geography	Categorical Nominal	Character
PopGroupID	Categorical Nominal	Integer
POPGROUP	Categorical Nominal	Character
RacesReported	Categorical Nominal	Integer
HSDegree	Discrete Ratio	Number
BachDegree	Discrete Ratio	Number

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

```
> str(acse_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" "Alam
y, 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(acse_df)
[1] 136
> ncol(acse_df)
[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 Histogram Plot.

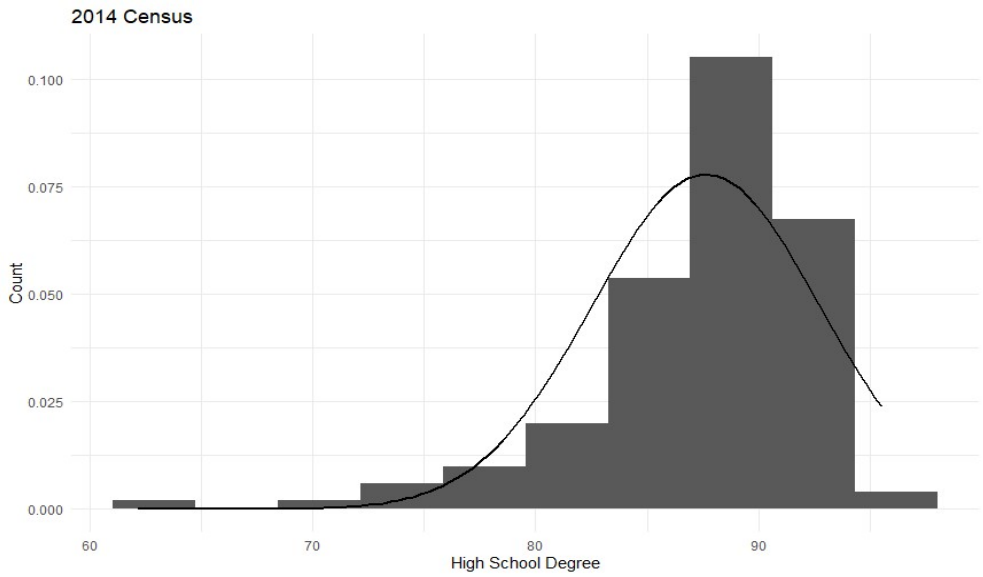


IV. Answer the following questions based on the Histogram produced:

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

   Yes, it is Unimodal because there is only one hump.

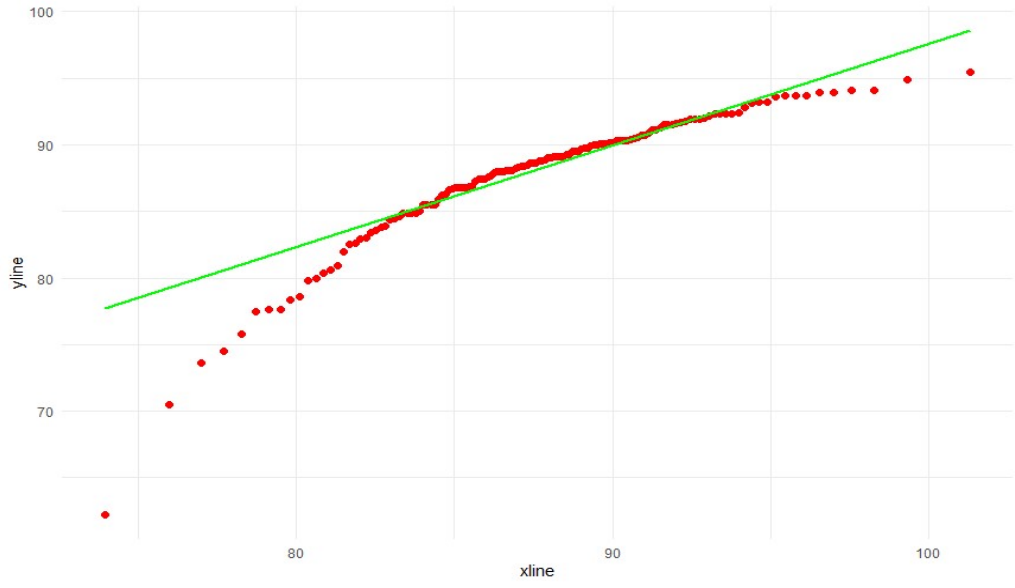
2. Is it approximately symmetrical? No
3. Is it approximately bell-shaped? No
4. Is it approximately normal? No
5. If not normal, is the distribution skewed? If so, in which direction?
   The histogram is left skewed because it has a tail on the left side of the distribution
6. Include a normal curve to the Histogram that you plotted.

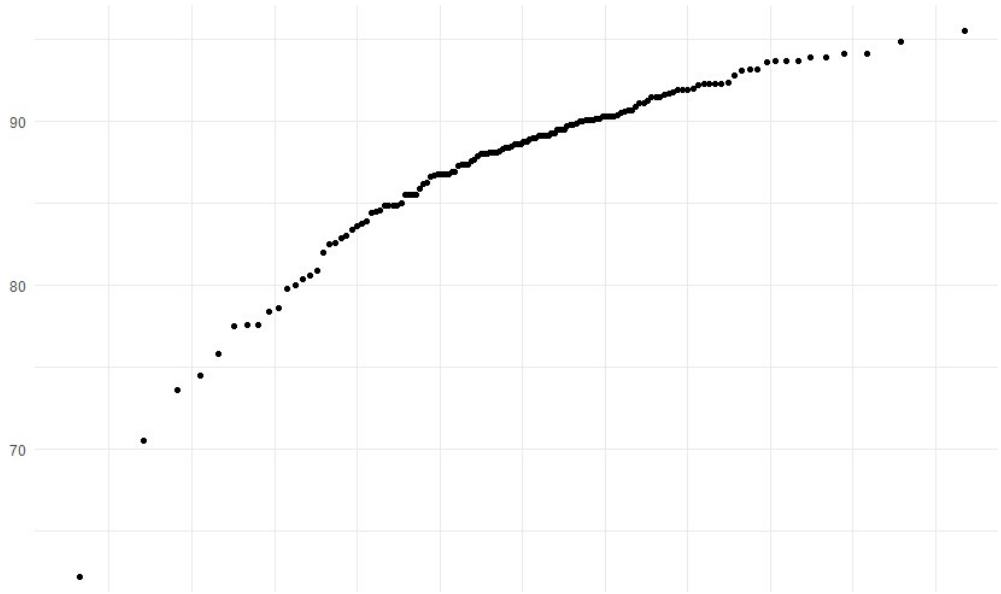


7. Explain whether a normal distribution can accurately be used as a model for this data.

   No, this is because the data shows skewed frequency distribution where the frequent scores are clustered at the higher end and tail points toward the lower scores.

V. Create a Probability Plot of the HSDegree variable.





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? Explain how you know.

No distribution is not normal. If the data are normally distributed, then data points will be plotted on the straight diagonal line.

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

Yes, the distribution is skewed. Our data point shows deviation from the diagonal line and both ends curve below the line.

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(acse_df$HSDegree)
     nbr.val     nbr.null       nbr.na          min          max        range          sum       median         mean
     136.000        0.000        0.000       62.200       95.500       33.300    11918.000       88.700       87.632
     SE.mean CI.mean.0.95          var      std.dev     coef.var
       0.439        0.868       26.193        5.118        0.058
```

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?

✓ The frequency distribution of HSDegree shows Negative Skewness where most frequent scores are clustered showed on the right.

- ✓ The histogram shows positive kurtosis or leptokurtic distribution where there are many scores distributed on the tail and is pointy
- ✓ Z-scores is the value of observation and calculated by taking the score minus the mean of all the scores then divide by the standard deviation of all the scores. Using R, result is

  z_scores <- (acse_df$HSDegree-mean(acse_df$HSDegree,))/sd(acse_df$HSDegree)

```
> z_scores
  [1]  0.2868 -0.1626  0.0718 -0.1431  0.2281 -2.7418 -2.5659 -1.9798 -0.5925 -1.3741 -0.1626 -1.7648 -0.2017  0.0914
 [15] -1.9602  0.0914 -0.0454 -0.0063 -1.8039 -0.7879  0.8339 -0.4166  1.0097  1.2637  0.4235  0.3258  0.3649  0.4822
 [29]  0.5017  0.7752  0.1500  0.2672 -0.0649 -0.2603 -1.3154  0.0523  0.0132  0.4822 -0.5339  0.2477  0.5212  0.1500
 [43]  0.7166  0.0718  0.8143 -0.4166  0.9120 -0.9247  0.5212  0.5994 -0.5143  1.5373  0.2281  0.1695  0.8339  0.5408
 [57]  0.6385 -0.4166 -0.6316 -1.0028  0.2868  0.9120  1.2637  0.8925 -0.7293  0.4822  0.2868  0.3258  1.1660 -0.5339
 [71]  1.0879  0.4431  0.4626  1.0879  0.1109 -0.6120  0.7557  0.1305 -0.4166 -0.8270  0.2868  1.0683  0.7948 -0.7488
 [85] -0.2799  0.0718 -3.3475  0.5799 -1.4913  0.5212  0.5994 -0.1626 -1.4131  0.4235 -0.0454  0.2672  0.3649  0.9316
 [99]  0.0914  0.4626  0.5603  0.4040  0.6775 -0.1626  0.1891  0.6775  0.5017  1.2246  1.2246  0.9120  0.7557 -0.5339
[113]  1.1856 -0.9833 -1.1005 -0.1822 -0.0454 -0.9051  1.1856 -1.9602  0.8339 -2.3119  0.1891 -1.5304 -4.9693 -0.3385
[127] -0.5339  0.1891  0.3649  1.1856  0.7557  0.9120  0.5212  0.8534  1.4200 -0.1431
```

- ✓ As the sample size increases or gets larger then the distribution of the sample approaches a normal distribution.