**EXERCISE 5**

This is your second exercise with real data. This time, instead of a bank of test scores, we will use the 2014 American Community Survey. These data are 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:

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

**ANSWER)** Below is the List of Elements of Data along with their datatype

ID - String Type(Categorical) - Holds a specific string for each location, kind of Location ID

ID2 - Int (Categorical) - It looks like the last 5 digits of the ID. We can say it as Location ID

Geography - String (Categorical)- Name of the place/county along with State information

PopGroupDisplayLabel (Categorical)- String - Seems the County are grouped based on population

PopGroupID (Categorical)- int - Number specifying the population group

RacesReported (Quantitative)- int - seems like the total number of population

HSDegree (Quantitative)- int - seems like % age

BachelorsDegree (Quantitative)- int - seems like %age

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

**ANSWER)**

str() -  ran with stringsAsFactor = TRUE to check, if any string attributes are having some specific structure of values. However it does not look like that, as there are 136 possible values for each String Values

'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(community\_df)

[1] 136

> ncol(community\_df)

[1] 8

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

**ANSWER)** R File and Histogram is attached

a. Set a bin size for the Histogram.

**ANSWER)** (see attached - R File and Histogram - set bin to 16. Based on dividing range of 60 -100, by 16 makes bw = 2.5)

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

**ANSWER)(**see attached - R File and Histogram)

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, It looks like a uni-modal with a peak around 90 approx.

b. Is it approximately symmetrical?

**ANSWER)** It is Asymmetric and Skewed towards Left

c. Is it approximately bell-shaped?

**ANSWER)** Not clearly Bell-shaped.

d. Is it approximately normal?

**ANSWER)** Not clearly Normal

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

**ANSWER)** Skewed towards Left or Negatively Skewed

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

**ANSWER)** Attached

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

**ANSWER)** The distribution is somewhat skewed towards the left and not a symmetric bell curve. Mean and Median are towards Left of the peak of the curve.[Median :88.70, Mean   :87.63]

So based on the reference article below: We may use it as there is slight skewness. However if there was too much skewness logarithmic or quantile regression techniques could be used.

<https://www.statisticshowto.com/probability-and-statistics/skewed-distribution/>

5. Create a Probability Plot of the HSDegree variable.

**ANSWER)** Attached the QQPlot.

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? **ANSWER)**Based on the probability or qq plot, as I read through articles it shows it's negatively skewed.

If the qqplot and qqline are almost similar or somewhat overlap, then we can say that the distribution is approximately normal. If the curve goes down it's negatively skewed, if upwards it's positively skewed and if the curve is in the opposite direction to the qq line, it has outliers.

Reference : <https://data.library.virginia.edu/understanding-q-q-plots/#:~:text=If%20both%20sets%20of%20quantiles,truly%20come%20from%20Normal%20distributions>

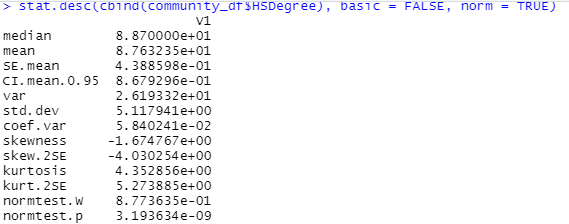
b. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.

**ANSWER)** Distribution is negatively skewed.

If the qqplot and qqline are almost similar or somewhat overlap, then we can say that the distribution is approximately normal. If the curve goes down it's negatively skewed, if upwards it's positively skewed and if the curve is in the opposite direction to the qq line, it has outliers.

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.

**ANSWER)**



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)**

A general guideline for **skewness** is that if the number is greater than +1 or lower than –1, this is an indication of a substantially **skewed** distribution. Our value comes out to be -1.675 approx.

For **kurtosis**, the general guideline is that if the number is greater than +1, the distribution is too peaked. For our case, this value is 4.352 approx.

Based on the book, skew.2SE and kurt.2SE are skewness and kurtosis by 2SE(Standard Error)

So, as Z Scores are

equation image

Z Skew  = 2X-4.03 = -8.06approx

Z Kurt = 2X5.27 = 10.54 approx

So as per reference statement shown below - since the sample size is 136 and our values are over 3.29, we should conclude the distribution is not normal

*“*

1. *For small samples (n < 50), if absolute z-scores for either skewness or kurtosis are larger than 1.96, which corresponds with a alpha level 0.05, then reject the null hypothesis and conclude the distribution of the sample is non-normal.*
2. *For medium-sized samples (50 < n < 300), reject the null hypothesis at absolute z-value over 3.29, which corresponds with a alpha level 0.05, and conclude the distribution of the sample is non-normal.*
3. *For sample sizes greater than 300, depend on the histograms and the absolute values of skewness and kurtosis without considering z-values. Either an absolute skew value larger than 2 or an absolute kurtosis (proper) larger than 7 may be used as reference values for determining substantial non-normality.*

*”*

So if sample size is greater than 300, our ranges change, based on the above statement. Then we can directly decide based on skew and kurtosis values, without looking for z-scores. Absolute Skew value more than 2 and Absolute Kurtosis value of 7 are references for determining substantial non-normality.