2.American Community Survey Exercise

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

```
> asc_2014 <- read.csv("data/acs-14-1yr-s0201.csv")
> summary(asc_2014)
  Id              Id2        Geography          PopGroupID
 Length:136        Min.   : 1073   Length:136        Min.   :1
 Class :character   1st Qu.:12082   Class :character   1st Qu.:1
 Mode  :character   Median :26112   Mode  :character   Median :1
Mean   :26833                     Mean   :1
3rd Qu.:39123                     3rd Qu.:1
Max.   :55079                    Max.   :1
 POPGROUP.display.label RacesReported        HSDegree
 Length:136            Min.   :  500292   Min.   :62.20
 Class :character       1st Qu.:  631380   1st Qu.:85.50
 Mode  :character       Median :  832708   Median :88.70
Mean   : 1144401   Mean   :87.63
3rd Qu.: 1216862   3rd Qu.:90.75
Max.   :10116705   Max.   :95.50
   BachDegree
 Min.   :15.40
 1st Qu.:29.65
 Median :34.10
 Mean   :35.46
 3rd Qu.:42.08
 Max.   :60.30
```

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

```
    > str(asc_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(asc_2014)

[1] 136

     > ncol(asc_2014)

[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.

```
> ggplot(asc_2014, aes(HSDegree) ) + geom_histogram(bins = 30) + ggtitle("HSDegree vs. Count")
```

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

2) Is it approximately symmetrical?

No, based on the Histogram, it's not Symmetrical

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?

it is a Negative Skewed (left skewed)

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

```
> ggplot(asc_2014, aes(HSDegree) ) + geom_histogram(bins = 30, aes(y = ..density..)) + ggtitle("HSDegree vs. Count") + stat_function(fun = dnorm, args = list(mean = mean(asc_2014$HSDegree), sd = sd(asc_2014$HSDegree)), col = "#1b98e0", size = 2)
```

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

No, we cannot use this data for Normal distribution, for Normal distribution the mean and medium should be same

V. Create a Probability Plot of the HSDegree variable.

```
> qqnorm(asc_2014$HSDegree)
> qqline(asc_2014$HSDegree)
```

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

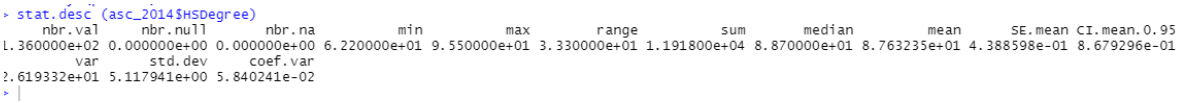
1. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.

Based on the Probability plot the distribution is not normal as the line did not fall on to the plots(dots)

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

It's not normal, yes, it is skewed, it is left skewed , based on the data plots the data is not normally distributed.

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 (asc_2014$HSDegree)
     nbr.val     nbr.null       nbr.na          min          max        range          sum       median         mean      SE.mean CI.mean.0.95 
L.360000e+02 0.000000e+00 0.000000e+00 6.220000e+01 9.550000e+01 3.330000e+01 1.191800e+04 8.870000e+01 8.763235e+01 4.388598e-01 8.679296e-01 
         var      std.dev     coef.var 
2.619332e+01 5.117941e+00 5.840241e-02 
> |
```

viii. In several sentences provide an explanation of the result produced for skew, kurtosis, and z-scores. In addition, explain

How can a change in the sample size change your explanation?

In probability theory and statistics, skewness is a measure of the extent to which a probability distribution of a real-valued random variable "leans" to one side of the mean.

Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution in the HDegree data.

in the data the distribution extreme is more on to the right side,so the kurtosis identifies whether the tails of a given distribution contain extreme values.

Z-score is also known as standard score gives us an idea of how far a data point is from the mean