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

Data Category	Data Types
Id	Varchar
Id2	Number
Geography	Varchar
PopGroupID	Boolean
POPGROUP.display-label	Varchar
RacesReported	NUmber
HSDegree	Decimal
BachDegree	Decimal

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

Please Refer git hub links

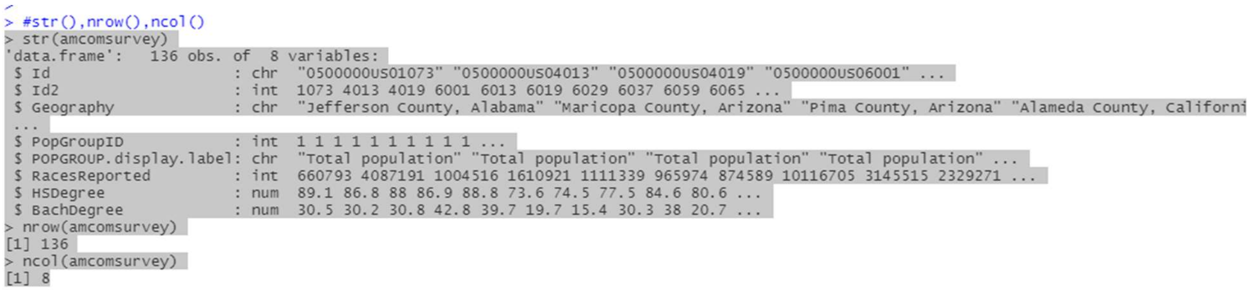
Code :

```
library (pastecs)
library(ggplot2)


theme_set(theme_minimal())
amcomsurvey <- read.csv("data/acs-14-1yr-s0201.csv")

#str(),nrow(),ncol()
str(amcomsurvey)
nrow(amcomsurvey)
ncol(amcomsurvey)
```

Ouptut



```
> #str(),nrow(),ncol()
> str(amcomsurvey)
'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, Californi
 ...
 $ 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(amcomsurvey)
[1] 136
> ncol(amcomsurvey)
[1] 8
```

3. Create a Histogram of the HSDegree variable using the ggplot2 package.
   a. Set a bin size for the Histogram.
   b. Include a Title and appropriate X/Y axis labels on your Histogram Plot.

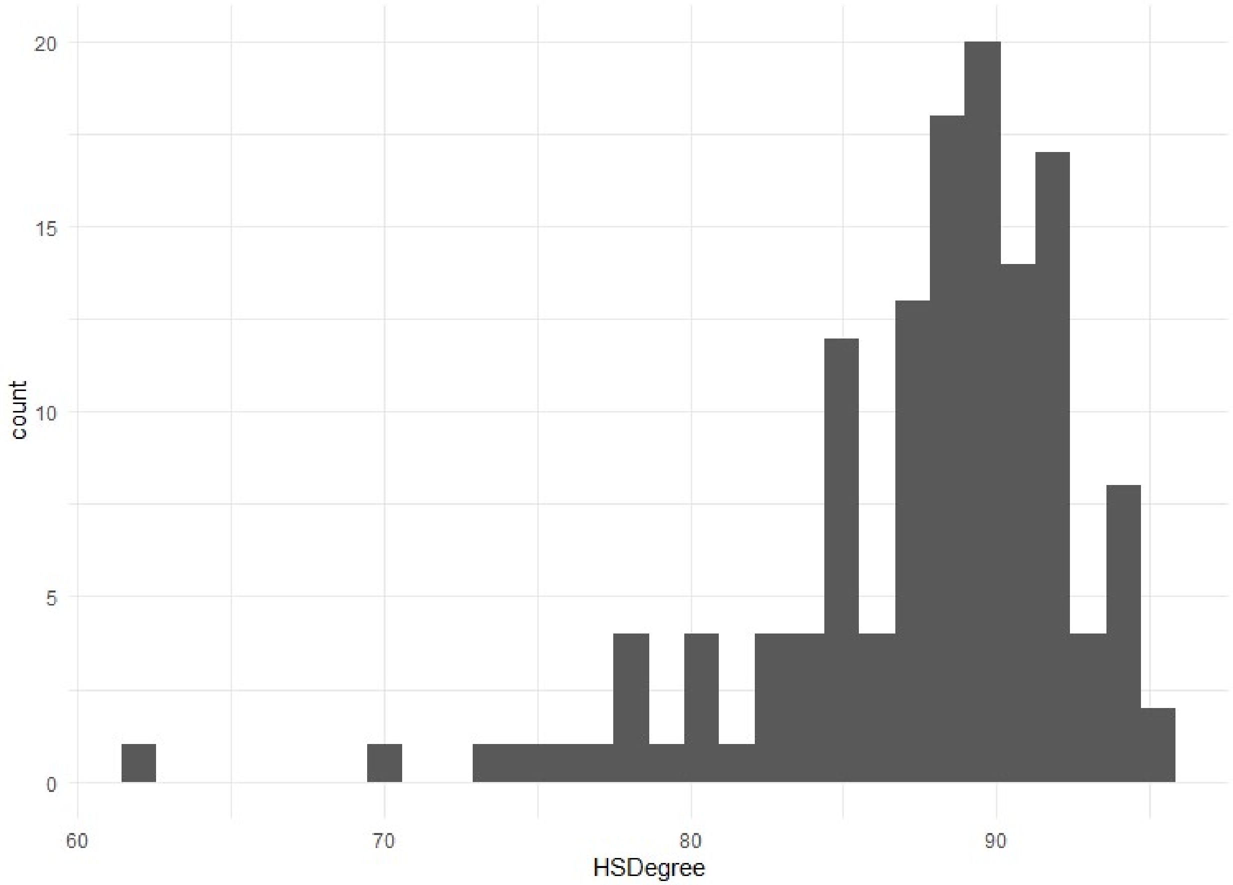
Code :

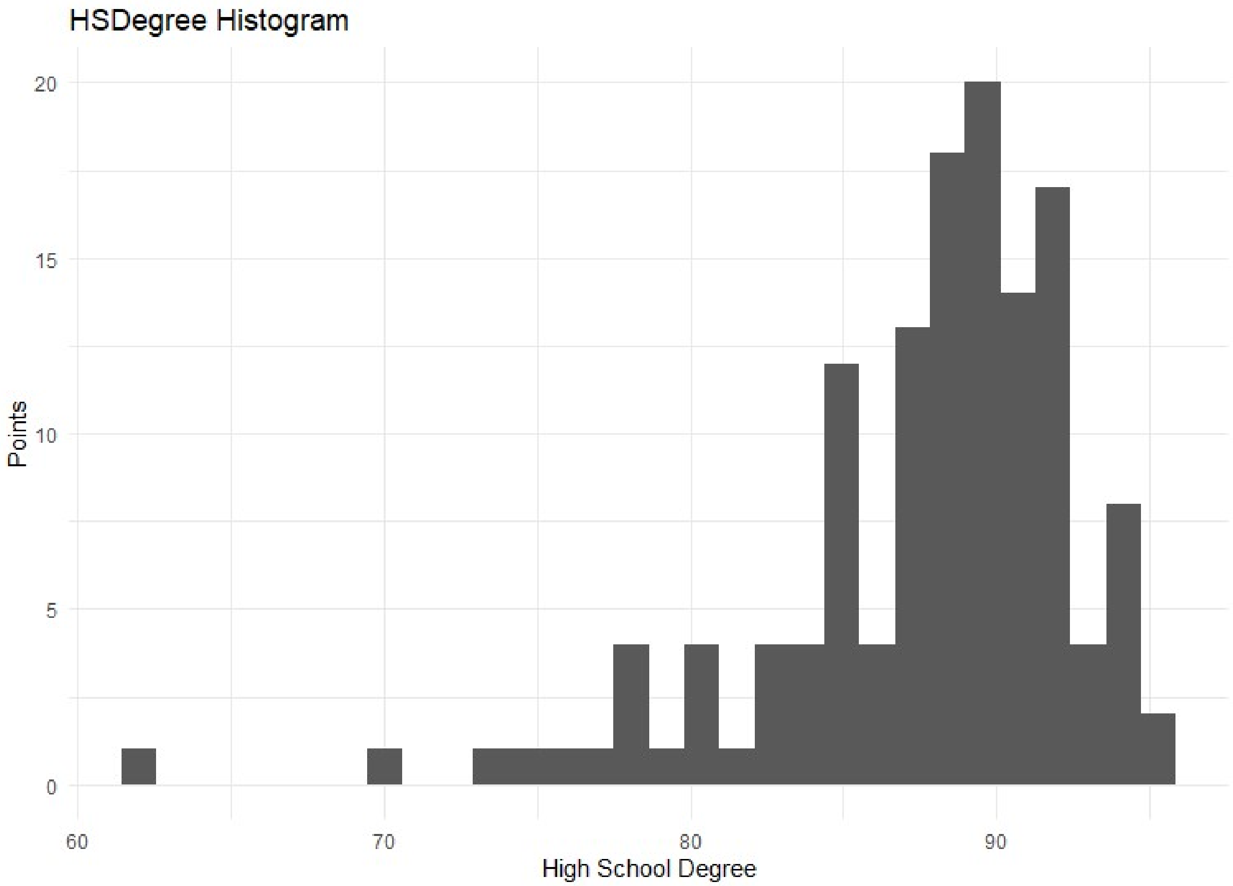
```
ggplot(amcomsurvey,aes(HSDegree)) + geom_histogram(bins=30)

#Histogram with bin size + titles
```

```
ggplot(amcomsurvey,aes(HSDegree)) + geom_histogram(bins=30) + ggtitle('HSDegree
Histogram') + xlab('High School Degree') + ylab('Points')
```

Output:





4. Answer the following questions based on the Histogram produced:
   a. Based on what you see in this histogram, is the data distribution unimodal?
      Ans: As there is only one peak, it is a unimodal distribution.

   b. Is it approximately symmetrical?
      Ans : The distribution doesn't seem to be a symmetrical.

   c. Is it approximately bell-shaped?
      Ans : The Distribution doesn't seem to be a bell shaped.

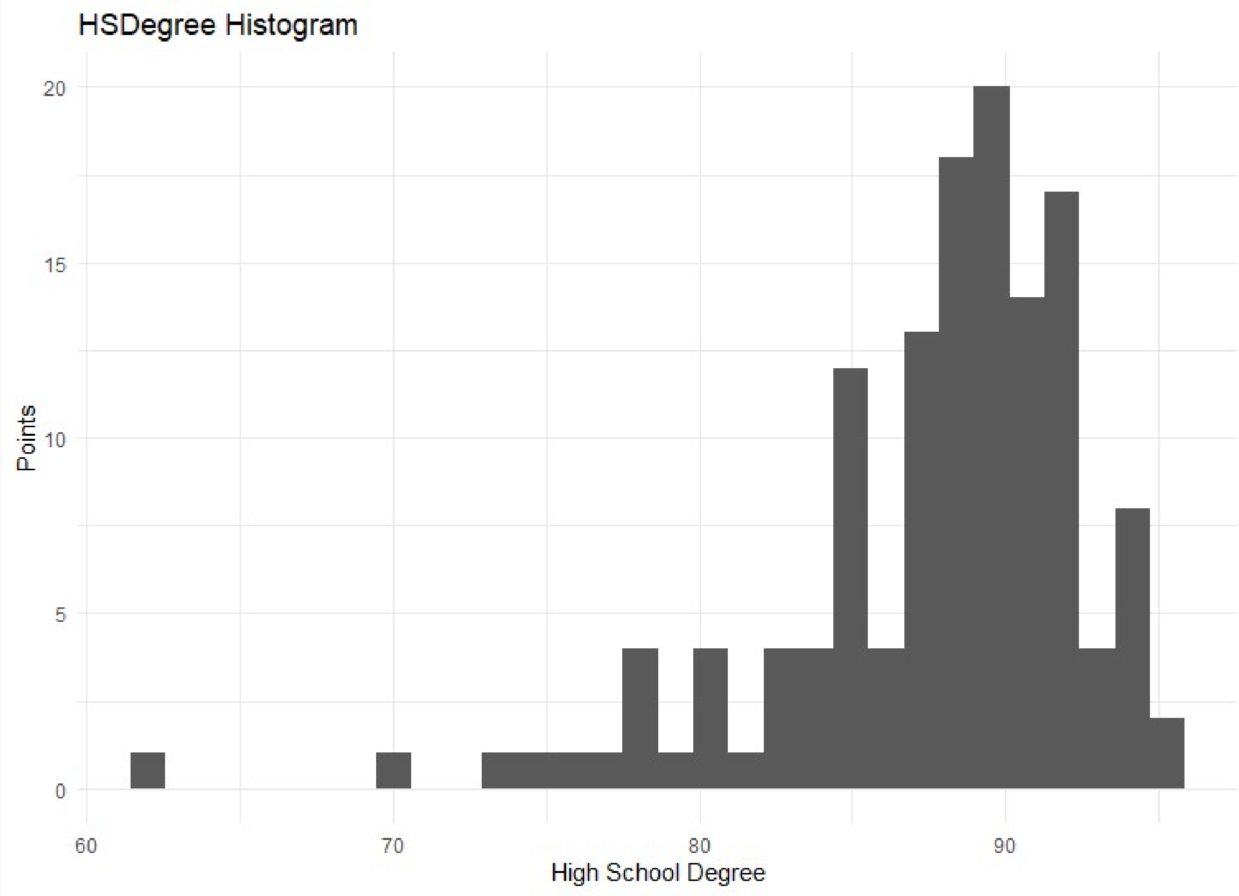
   d. Is it approximately normal?
      ns: It is not a normal distribution exactly, but we can say it is approximately normal
   e. If not normal, is the distribution skewed? If so, in which direction?
      Ans: It is a negatively skewed distribution because the tail is pointing towards the lower values of HSDegree.

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

Code :

```
        ggplot(amcomsurvey,aes(HSDegree)) + geom_histogram(aes(y = ..density..), bins
= 30) + geom_function(fun = dnorm, args = list(mean = mean(amcomsurvey$HSDegree),
sd = sd(amcomsurvey$HSDegree)), size=3, colour="Red")
```

Output:



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

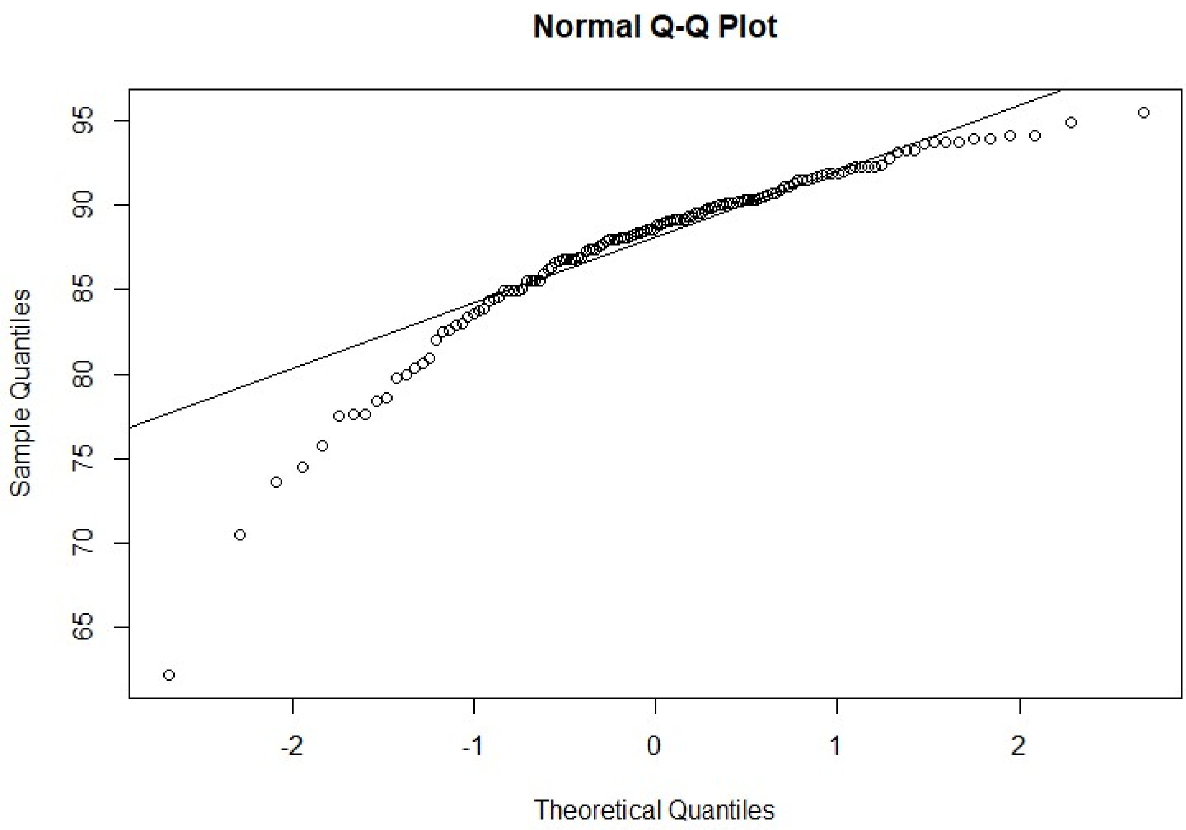
Ans: normal probability model can be used even if the distribution of the continuous outcome is not perfectly symmetrical; it just has to be reasonably close to a normal

5. Create a Probability Plot of the HSDegree variable.

Code :

```
qqnorm(amcomsurvey$HSDegree)
qqline(amcomsurvey$HSDegree)
```

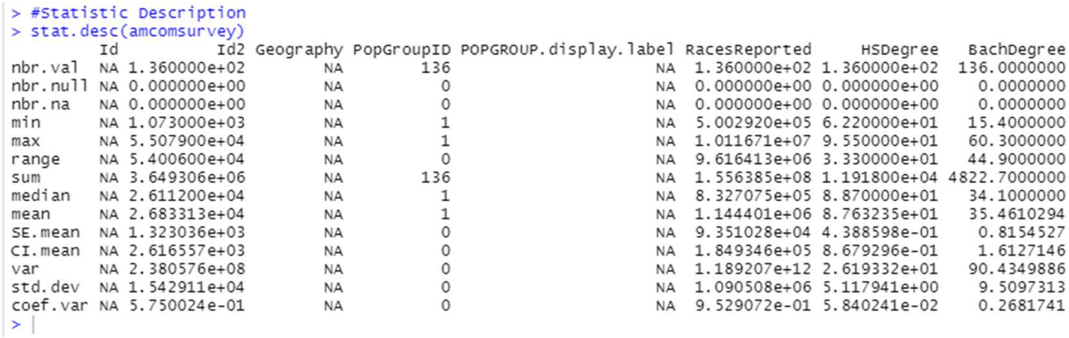
Output:



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.
      Ans: The plotted points bend down and to the right of the normal line that indicates a long tail to the left. It indicates that the distribution is not normal.

   b. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.
      Ans: The distribution is left skewed. The plotted points bend down and to the right of the normal line that indicates a long tail to the left.

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.
   Ans: Please find screen capture below.



```
> #Statistic Description
> stat.desc(amcomsurvey)
         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?

Skew:

Skew is the lack of symmetry in the data. As we can see the plot is not exactly normally distributed and it's tail is pointing towards the lower numbers, it is negatively skewed data.

Kurtosis:

Kurtosis is the pointiness of the data. The plot seems to be heavily tailed; it is leptokurtosis.

Z Scores:

Z score is the relationship of a score with the mean of the score and the. To Calculate Z score of X the formula is

Z = x – mean / standard deviation.

From the given dataset lets find the Z score of value 88.1

Mean of HSDegree Values is 87.63235

Standard deviation of HSDegree is 5.117941

Z score = (88.1 -87.63) /5.11

= 0.09

It means that the value 88.1 is 0.09 Standard Deviation away from the mean. It implies that the value is pretty much close to the mean.

How The Sample Size impacts the Skew, Kurtosis and Z score.

I am not sure how to explain this or do we need to explain with an example. Here is what I came up with.

The Skew, Kurtosis and Z score are mainly dependent on the mean and standard deviation. If we take a smaller sample size it will be very misleading as the mean value will not be far from the mean of the sample. Bigger sample size leads to the mean value which is close to the mean of the population and increases the possibility of model to be a good fit.