Assignment: ASSIGNMENT 3.2

Name: Anjale, Jiteshwar

Date: 2021-03-30

#American Community Survey Exercise

> ## Load the ggplot2 package

> library(ggplot2)

> theme_set(theme_minimal())

>

> ## Load the pastecs package

> library(pastecs)

>

> ## Set the working directory to the root of your DSC 520 directory

> setwd('C:/Users/anjal/OneDrive/Desktop/MS/DSC520/dsc520')

>

> ## Load the `data/acs-14-1yr-s0201.csv` to

> acs_df <- read.csv("data/acs-14-1yr-s0201.csv")

>

>

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

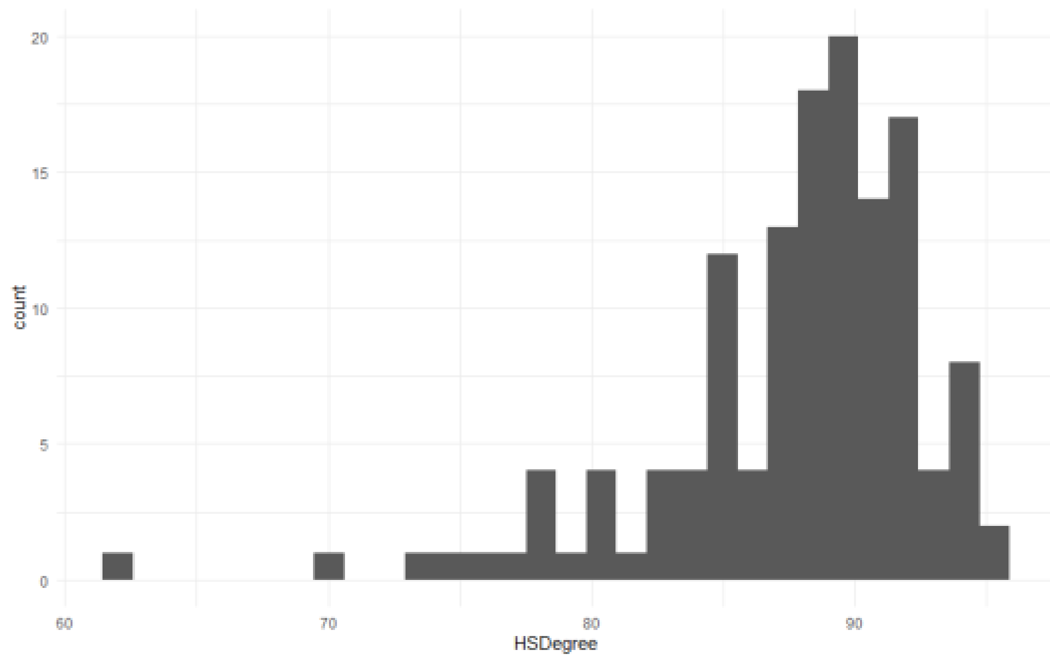
> summary(acs_df)

```
      Id                 Id2          Geography           PopGroupID POPGROUP.display.label RacesReported
 Length:136         Min.   : 1073   Length:136         Min.   :1    Length:136             Min.   :  500292
 Class :character   1st Qu.:12082   Class :character   1st Qu.:1    Class :character       1st Qu.:  631380
 Mode  :character   Median :26112   Mode  :character   Median :1    Mode  :character       Median :  832708
                    Mean   :26833                      Mean   :1                           Mean   : 1144401
                    3rd Qu.:39123                      3rd Qu.:1                           3rd Qu.: 1216862
                    Max.   :55079                      Max.   :1                           Max.   :10116705
    HSDegree       BachDegree
 Min.   :62.20   Min.   :15.40
 1st Qu.:85.50   1st Qu.:29.65
 Median :88.70   Median :34.10
 Mean   :87.63   Mean   :35.46
 3rd Qu.:90.75   3rd Qu.:42.08
 Max.   :95.50   Max.   :60.30
```

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

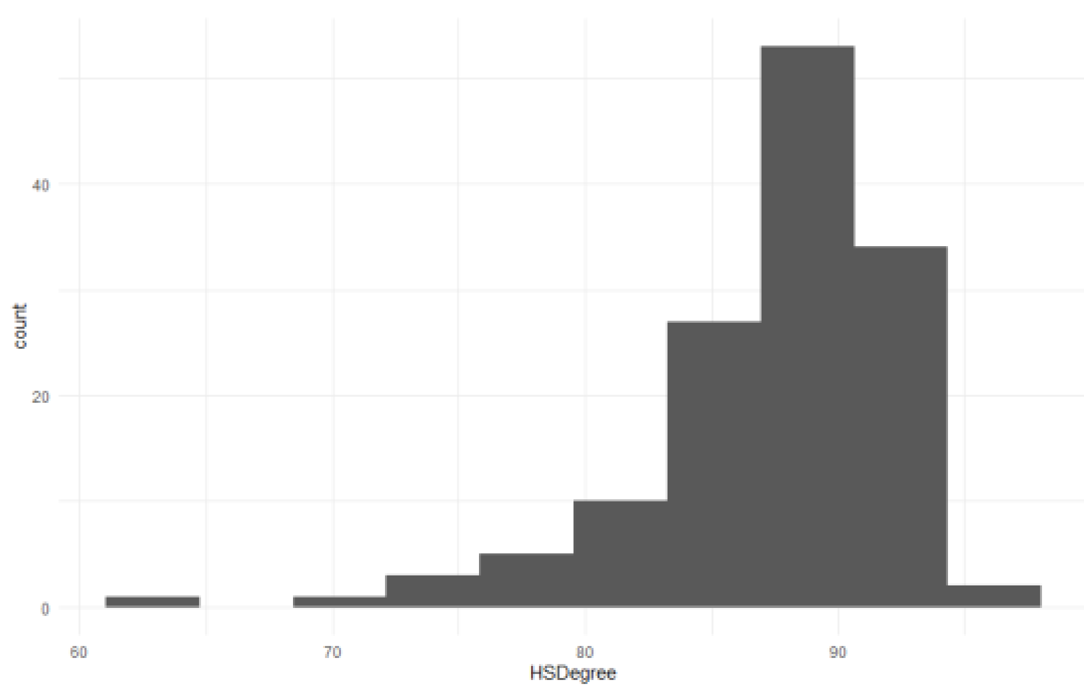
```
> str(acs_df)
'data.frame':	136 obs. of  8 variables:
 $ Id                    : chr  "0500000US01073" "0500000US04013" "0500000US04019" "0500000US06001" ...
 $ Id2                   : int  1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...
 $ Geography             : chr  "Jefferson County, Alabama" "Maricopa County, Arizona" "Pima County, Arizona" "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(acs_df)
[1] 136
> ncol(acs_df)
[1] 8
```

```
## iii.Create a Histogram of the HSDegree variable using the ggplot2 package.
ggplot(acs_df, aes(x=HSDegree)) + geom_histogram()
```



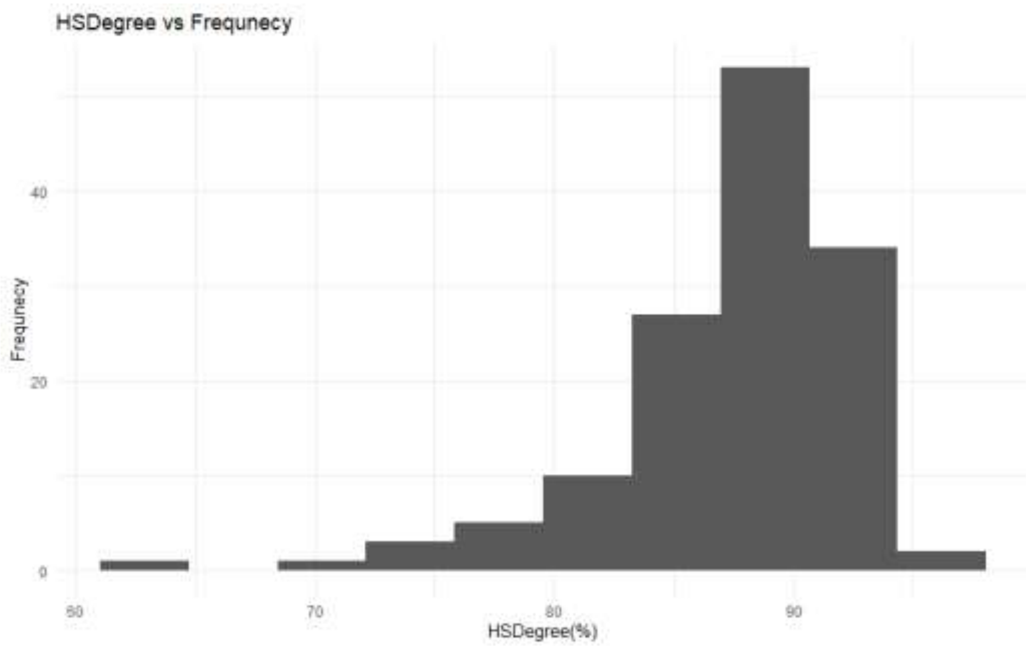
1.Set a bin size for the Histogram.

ggplot(acs_df, aes(HSDegree)) + geom_histogram(bins = 10)



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

ggplot(acs_df, aes(HSDegree)) + geom_histogram(bins = 10) + ggtitle("HSDegree vs Frequnecy") + xlab("HSDegree(%)") +ylab("Frequnecy")



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

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

As the probability distribution in given histogram has a single peak, the data distribution unimodal.

2.Is it approximately symmetrical?

The distribution is not symmetrical

3.Is it approximately bell-shaped?

The distribution is not bell-shaped

4.Is it approximately normal?

The distribution is not normal.

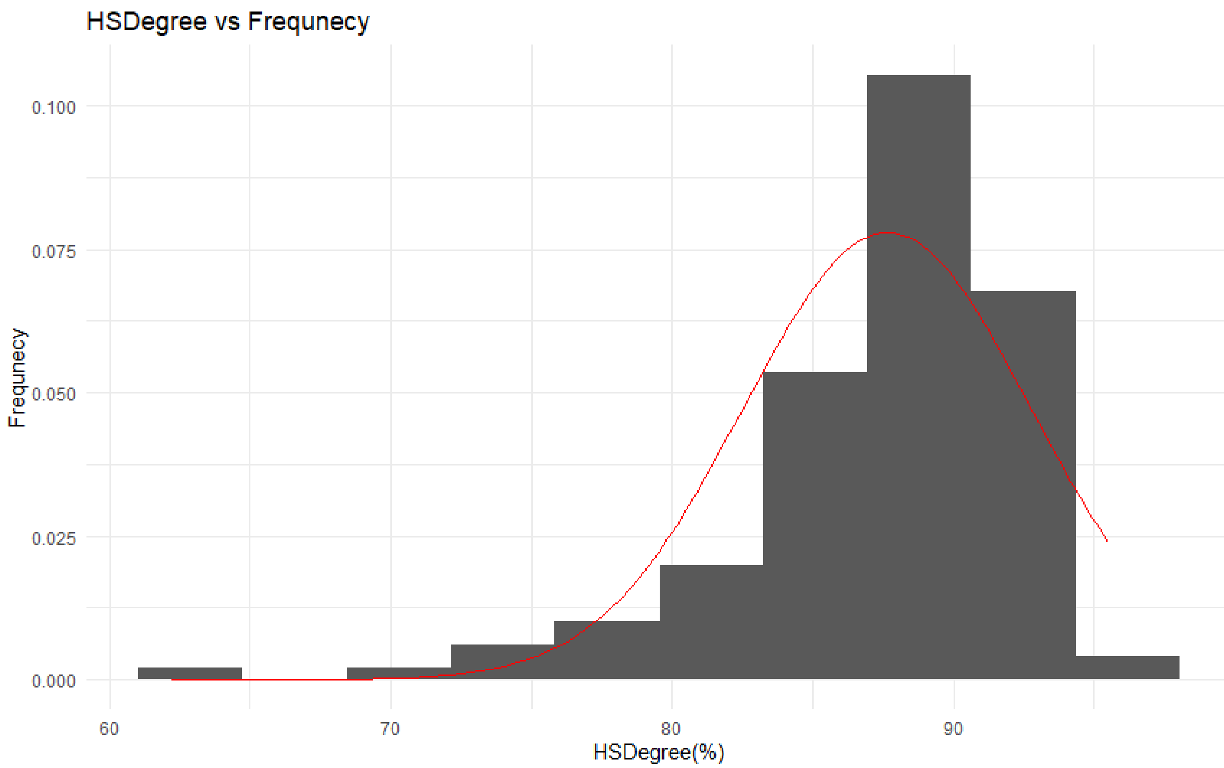
5.If not normal, is the distribution skewed? If so, in which direction?

The distribution skewed in left (hence it is negatively Skewed)

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

ggplot(acs_df, aes(x=HSDegree)) + geom_histogram(aes(y = ..density..),bins = 10)+ stat_function(fun = dnorm, colour = "red",args = list(mean = mean(acs_df$HSDegree, na.rm = TRUE), sd =

sd(acs_df$HSDegree, na.rm = TRUE))) + ggtitle("HSDegree vs Frequnecy") + xlab("HSDegree(%)") +ylab("Frequnecy")



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

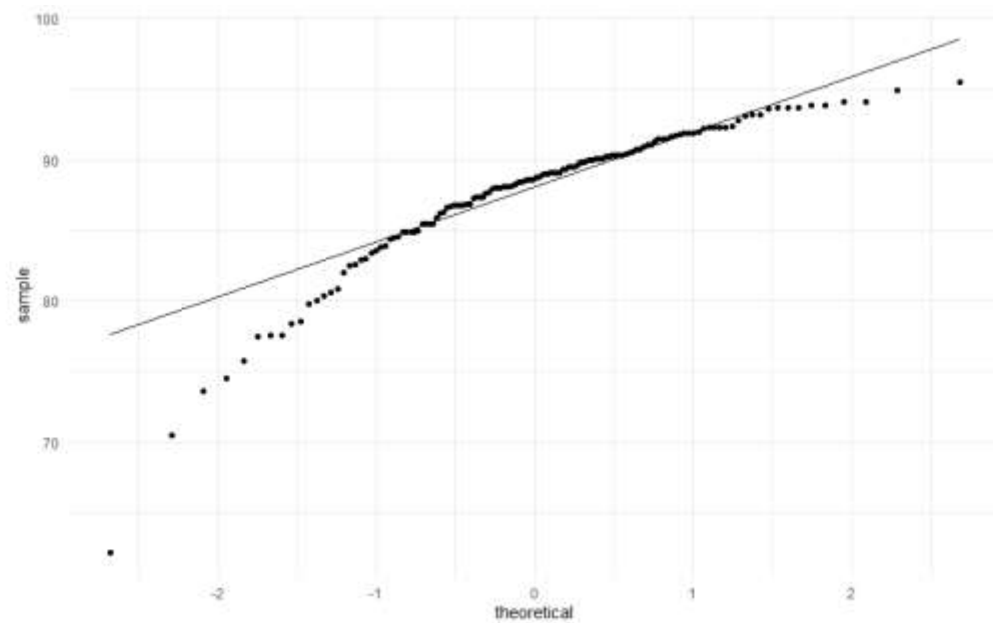
A normal distribution can not accurately be used as a model for this data

##because the histogram does not have shape of normal curve and

##The histogram shows that the distribution is negatively skewed.

V.Create a Probability Plot of the HSDegree variable.

ggplot(acs_df, aes(sample=HSDegree)) + stat_qq() + stat_qq_line()



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.

The straight line in this plot represents

a normal distribution, and the points represent the observed residuals. Therefore, in

a perfectly normally distributed data set, all points will lie on the line.

As the points are not on line, it is not normal distribution.

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

##In this plot, the lower end of QQ plot deviates from straight line

##then we can clearly say that the distribution has a longer tail to its left or simply it is left-skewed (or negatively skewed)

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(acs_df, basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)

```
> stat.desc(acs_df, basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)
          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
```

stat.desc(acs_df["HSDegree"], basic=FALSE, desc=TRUE, norm=TRUE)

```
> stat.desc(acs_df["HSDegree"], basic=FALSE, desc=TRUE, norm=TRUE)
                 HSDegree
median       8.870000e+01
mean         8.763235e+01
SE.mean      4.388598e-01
CI.mean.0.95 8.679296e-01
var          2.619332e+01
std.dev      5.117941e+00
coef.var     5.840241e-02
skewness    -1.674767e+00
skew.2SE    -4.030254e+00
kurtosis     4.352856e+00
kurt.2SE     5.273885e+00
normtest.W   8.773635e-01
normtest.p   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?

The negative value for skewness indicates that the data are skewed left.

The positive value of of kurtosis indicates a "heavy-tailed" distribution.

Z score for skew (skew.2SE) and Z score for kurtosis(kurt.2SE) are greater than 1 and hence both skew and kurtosis are significant.

##Large samples will give rise to small standard errors and so when sample sizes are big, significant values arise from even small deviations from normality. Also for larger sample size we need to use visualizations for data analysis.