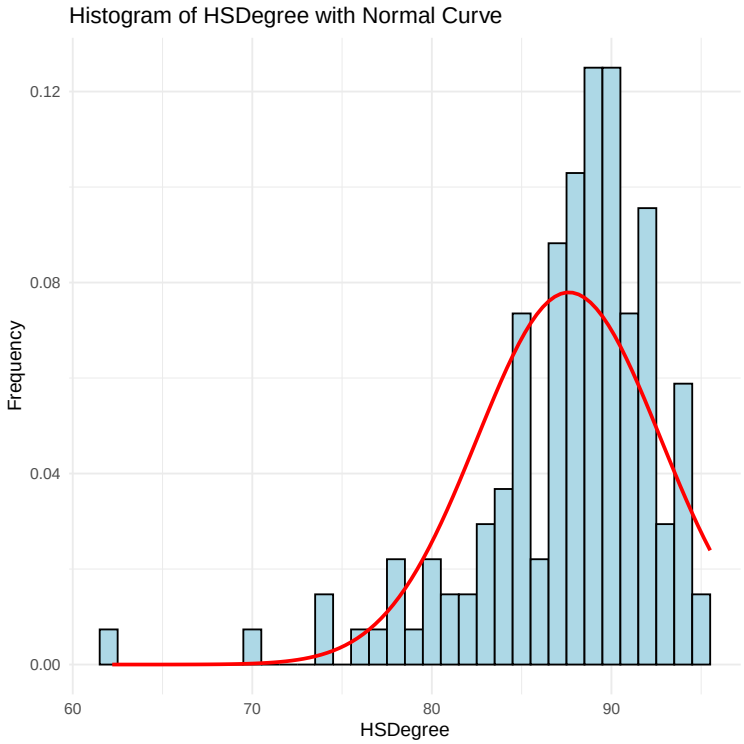
The results below are generated from an R script.

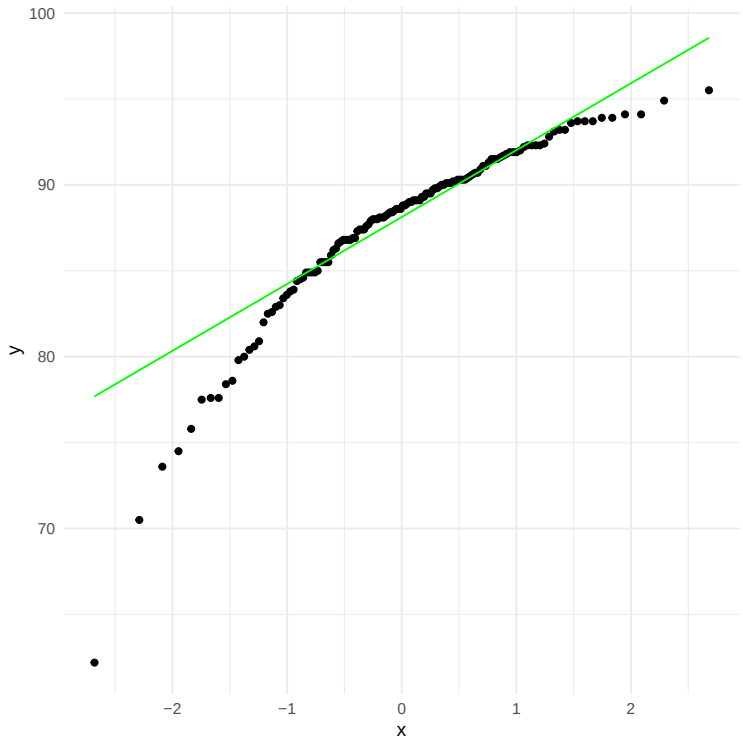
```
# Assignment: ASSIGNMENT 3.1
# Name: Reppeto, Brian
# Date: 2023-06-20

## Load the ggplot2 package
library(ggplot2)
library(psych)
library(pastecs)
theme_set(theme_minimal())

## Set the working directory to the root of your DSC 520 directory
setwd("~/DSC520/Week 3")

## Load the `` to
acs_df <- read.csv("acs-14-1yr-s0201.csv")

## 1.  The data elements are ..............................

#$ Id                    : chr
#$ Id2                   : int
#$ Geography             : chr
#$ PopGroupID            : int
#$ POPGROUP.display.label: chr
#$ RacesReported         : int
#$ HSDegree              : num
#$ BachDegree            : num


## 2.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, ...
##  $ PopGroupID            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ POPGROUP.display.label: chr  "Total population" "Total population" "Total population" "Total popul...
##  $ RacesReported         : int  660793 4087191 1004516 1610921 1111339 965974 874589 10116705 3145515 ...
##  $ 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

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

```r
ggplot(acs_df, aes(x = HSDegree)) +
  geom_histogram(aes(y = ..density..) ,binwidth = 1, color = "black"
  , fill = "lightblue") +
  stat_function(fun = dnorm, args = list(mean = mean(acs_df$HSDegree,
  na.rm =  TRUE), sd = sd(acs_df$HSDegree, na.rm =TRUE)),
  color = "red", size = 1) +
  labs(title = "Histogram of HSDegree with Normal Curve"
  , x = "HSDegree", y = "Frequency")
```




```r
## 4.Answer the following questions based on the Histogram produced:
##    1. Based on what you see in this histogram, is the data
##        distribution unimodal? ----  Yes the distribution is unimodal.
##    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?
```

```r
##    6. Include a normal curve to the Histogram that you plotted.----Done
##    7. Explain whether a normal distribution can accurately be used as a model
##        for this data.-- T



## 5. Create a Probability Plot of the HSDegree variable

ggplot(acs_df, aes(sample = HSDegree)) + stat_qq() + stat_qq_line(color="green")
```




```
## 6.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 distribution is
##      not normal.  This is based on a normal line test.
##    2.If not normal, is the distribution skewed? If so, in which direction?
##      Explain how you know.--------------The plot is left skewed.




## 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.

describe(acs_df$HSDegree)

##    vars   n  mean   sd median trimmed  mad  min  max range  skew kurtosis   se
## X1    1 136 87.63 5.12   88.7   88.28 3.78 62.2 95.5  33.3 -1.67     4.35 0.44

stat.desc(acs_df$HSDegree, basic =TRUE, norm = FALSE)

##      nbr.val     nbr.null       nbr.na          min          max        range
## 1.360000e+02 0.000000e+00 0.000000e+00 6.220000e+01 9.550000e+01 3.330000e+01
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 1.191800e+04 8.870000e+01 8.763235e+01 4.388598e-01 8.679296e-01 2.619332e+01
##      std.dev     coef.var
## 5.117941e+00 5.840241e-02

## 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?

## Skewness, kurtosis, and z-scores are statistical measures used to analyze the
## distribution of a dataset. Skewness measures the asymmetry of the data
## distribution, where positive skewness indicates a longer tail on the
## right side, negative skewness indicates a longer tail on the left side,
## and zero skewness represents a symmetric distribution. Kurtosis measures
## the "tailedness" of the distribution, where positive kurtosis indicates
## heavier tails and a sharper peak, negative kurtosis indicates lighter tails
## and a flatter peak, and zero kurtosis represents a normal distribution.
## Z-scores are a standardized measure that expresses a data point's
## deviation from the mean in terms of standard deviations.  Since the kurtosis
## is greater than 3 this indicates the leptokurtic has lang and skinny tails
## this means there are more chances of outliers.  Additionally, the negative
## skew indicates the data points are more concentrated towards the right
## side of the distribution, and makes the mean bend more toward the right also.
## A change in the sample size can affect the interpretation of these measures.
## Skewness and kurtosis are influenced by extreme values, and as the sample
## size increases, the impact of outliers or extreme values tends to diminish.
##  Therefore, larger sample sizes may result in more accurate estimates of
## skewness and kurtosis, providing a better representation of the underlying
## population. On the other hand, z-scores are not directly affected by sample
## size since they are calculated based on the mean and standard deviation.
## However, a larger sample size can provide more reliable estimates of the
## mean and standard deviation, leading to more precise z-scores.
```

The R session information (including the OS info, R version and all packages used):

```
sessionInfo()

## R version 4.3.0 (2023-04-21)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Ventura 13.4
##
## Matrix products: default
## BLAS:   /System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versio
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  LAPACK ve
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Chicago
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] pastecs_1.3.21 psych_2.3.6    ggplot2_3.4.2
##
## loaded via a namespace (and not attached):
##  [1] crayon_1.5.2     vctrs_0.6.3      knitr_1.43       nlme_3.1-162     cli_3.6.1
##  [6] xfun_0.39        rlang_1.1.1      highr_0.10       generics_0.1.3   glue_1.6.2
## [11] labeling_0.4.2   colorspace_2.1-0 scales_1.2.1     fansi_1.0.4      grid_4.3.0
```

```
## [16] evaluate_0.21    munsell_0.5.0    tibble_3.2.1     lifecycle_1.0.3  compiler_4.3.0
## [21] dplyr_1.1.2      pkgconfig_2.0.3  rstudioapi_0.14  lattice_0.21-8   farver_2.1.1
## [26] R6_2.5.1         tidyselect_1.2.0 utf8_1.2.3       pillar_1.9.0     mnormt_2.1.1
## [31] parallel_4.3.0   magrittr_2.0.3   tools_4.3.0      withr_2.5.0      gtable_0.3.3
## [36] boot_1.3-28.1

Sys.time()

## [1] "2023-06-24 21:07:42 CDT"
```