June 25, 2023

The results below are generated from an R script.

```
# Assignment: American Community Survey
# Name: Smith, David
# Date: 2023-06-25

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

## Set the working directory to the root of your DSC 520 directory
setwd("F:\\GitLab-Projects\\Bellevue\\SMITH-DSC520")

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

# What are the elements in your data (including the categories and data types)?
summary(acs_df)
```

```
##       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    BachDegree
##  Length:136             Min.   :  500292   Min.   :62   Min.   :15
##  Class :character       1st Qu.:  631380   1st Qu.:86   1st Qu.:30
##  Mode  :character       Median :  832708   Median :89   Median :34
##                         Mean   : 1144401   Mean   :88   Mean   :35
##                         3rd Qu.: 1216862   3rd Qu.:91   3rd Qu.:42
##                         Max.   :10116705   Max.   :96   Max.   :60
```

```
# 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 ...
```

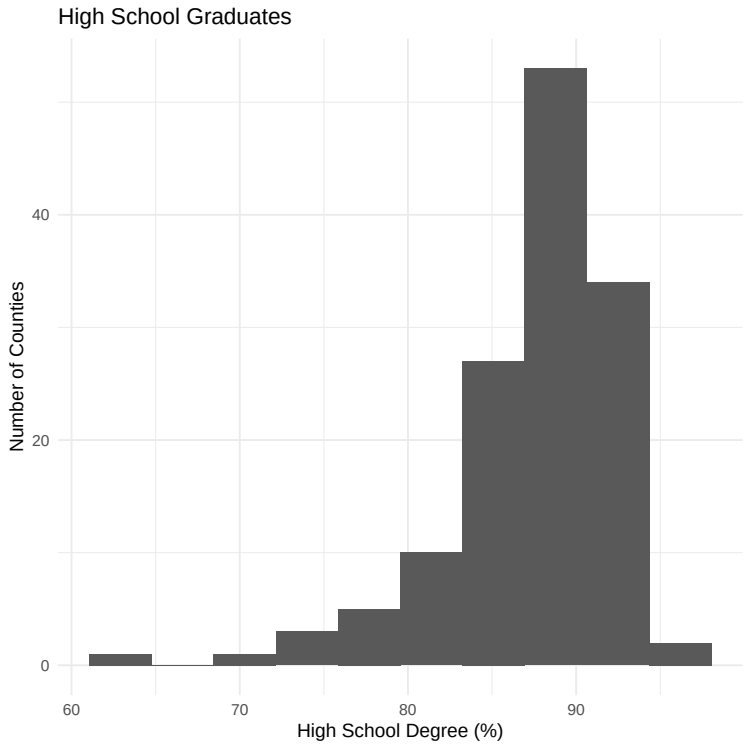
```r
nrow(acs_df)
```

```
## [1] 136
```

```r
ncol(acs_df)
```

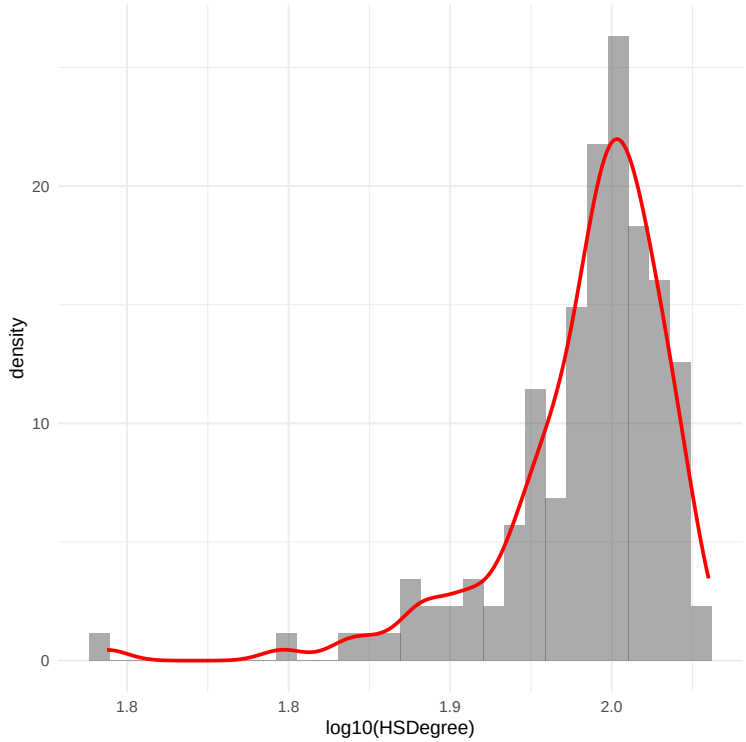
```
## [1] 8
```

```r
## Create a histogram of the 'HSDegree' variable using 'geom_histogram()'
## Use 10 bins
hist1_HSDegree <- ggplot(acs_df, aes(HSDegree)) + geom_histogram(bins=10)+ ggtitle("High School Graduate
hist1_HSDegree
```

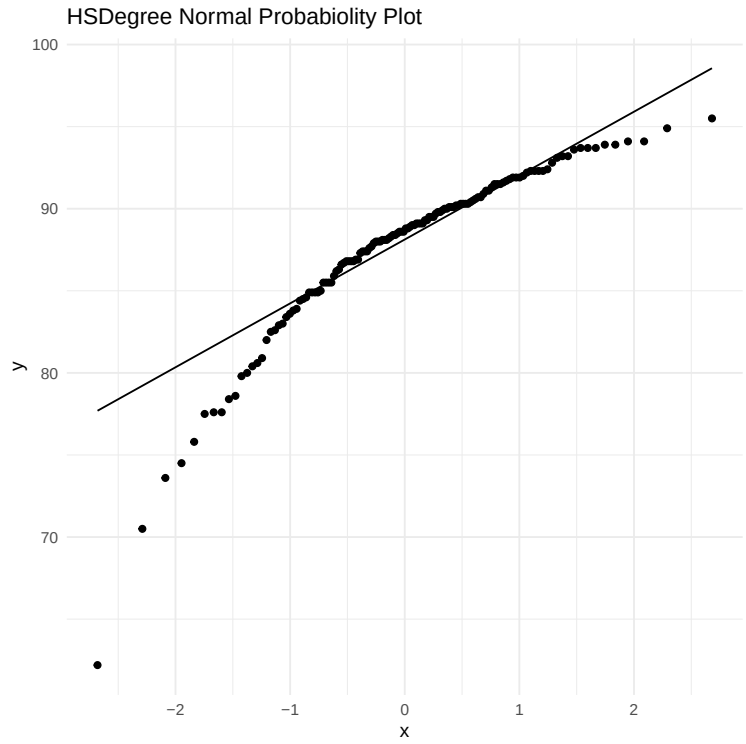



```r
# hist2_HSDegree <- hist1_HSDegree + stat_function(data = acs_df, fun = dnorm, args = list(mean = mean(
hist2_HSDegree <- ggplot(acs_df, aes(x = log10(HSDegree),y = after_stat(density)))   +  geom_histogram(a
hist2_HSDegree
```

```
## 'stat_bin()' using 'bins = 30'.  Pick better value with 'binwidth'.
```




```
## Create a Probability Plot of the HSDegree variable.
prob_HSDegree <- ggplot(acs_df, aes(sample=HSDegree)) + stat_qq() + stat_qq_line() + labs(title = "HSDeg
prob_HSDegree
```



```
# Now that you have looked at this data visually for normality, you will now quantify normality with nu
library(pastecs)
options(scipen=100)
options(digits=2)
stat.desc(acs_df$HSDegree)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
##      136.000        0.000        0.000       62.200       95.500       33.300
##          sum       median         mean      SE.mean CI.mean.0.95          var
##    11918.000       88.700       87.632        0.439        0.868       26.193
##      std.dev     coef.var
##        5.118        0.058
```

```
## Supportive metrics
## In several sentences provide an explanation of the result produced for skew, kurtosis, and z-scores.
## the sample size may change your explanation?

library(moments)
skew_HSDegree <- skewness(acs_df$HSDegree)
skew_HSDegree
```

```
## [1] -1.7
```

```
kurtosis_HSDegree <- kurtosis(acs_df$HSDegree)
kurtosis_HSDegree
```

```
## [1] 7.5
```

```
test_HSDegree <- jarque.test(acs_df$HSDegree)
test_HSDegree
```

```
##
##  Jarque-Bera Normality Test
##
## data:  acs_df$HSDegree
## JB = 178, p-value <0.0000000000000002
## alternative hypothesis: greater
```

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

```
sessionInfo()
```

```
## R version 4.3.1 (2023-06-16 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_United States.utf8  LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8 LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: America/New_York
```

```
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] RSQLite_2.3.1  moments_0.14.1 pastecs_1.3.21 ggplot2_3.4.2
##
## loaded via a namespace (and not attached):
##  [1] bit_4.0.5        gtable_0.3.3     dplyr_1.1.2      compiler_4.3.1   highr_0.10
##  [6] tidyselect_1.2.0 blob_1.2.4       scales_1.2.1     boot_1.3-28.1    fastmap_1.1.1
## [11] R6_2.5.1         labeling_0.4.2   generics_0.1.3   knitr_1.43       tibble_3.2.1
## [16] munsell_0.5.0    DBI_1.1.3        pillar_1.9.0     rlang_1.1.1      utf8_1.2.3
## [21] cachem_1.0.8     xfun_0.39        bit64_4.0.5      memoise_2.0.1    cli_3.6.1
## [26] withr_2.5.0      magrittr_2.0.3   grid_4.3.1       rstudioapi_0.14  lifecycle_1.0.3
## [31] vctrs_0.6.3      evaluate_0.21    glue_1.6.2       farver_2.1.1     fansi_1.0.4
## [36] colorspace_2.1-0 tools_4.3.1      pkgconfig_2.0.3

Sys.time()

## [1] "2023-06-25 20:08:43 EDT"
```