American-Community-Survey-Exercise.R

sheetal

2022-12-18

```
# Assignment: ASSIGNMENT 3.1
# Name: Munjewar, Sheetal
# Date: 2022-12-11

## Check your current working directory using `getwd()`
getwd()
```

```
## [1] "E:/Data_Science_DSC510/DSC520-Statistics/dsc520/assignments/assignment03"
```

```
## List the contents of the working directory with the `dir()` function
dir()
```

```
## [1] "American-Community-Survey-Exercise.pdf"
## [2] "American-Community-Survey-Exercise.R"
## [3] "American-Community-Survey-Exercise.spin.R"
## [4] "American-Community-Survey-Exercise.spin.Rmd"
## [5] "American Community Survey Exercise.R"
## [6] "assignment_03_MunjewarSheetal.R"
## [7] "data-visualization-2.1.pdf"
```

```
## If the current directory does not contain the `data` directory, set the
## working directory to project root folder (the folder should contain the `data` directory
## Use `setwd()` if needed
setwd("E:\\Data_Science_DSC510\\DSC520-Statistics\\dsc520")

## Load American Community Survey Exercise survey excel `data/acs-14-1yr-s0201.csv` to `acs_df` using `
## Get summary for data frame 'acs_df'suing summary()`
acs_df <- read.csv("data/acs-14-1yr-s0201.csv")
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.   :62.20   Min.   :15.40
##  Class :character       1st Qu.:  631380   1st Qu.:85.50   1st Qu.:29.65
```

```
##  Mode  :character       Median :  832708   Median :88.70   Median :34.10
##                         Mean   : 1144401   Mean   :87.63   Mean   :35.46
##                         3rd Qu.: 1216862   3rd Qu.:90.75   3rd Qu.:42.08
##                         Max.   :10116705   Max.   :95.50   Max.   :60.30
```

```
##Run the following functions and provide the results: str(); nrow(); ncol()
## Examine the structure of `acs_df` using `str()`
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 314551
##  $ 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 ...
```

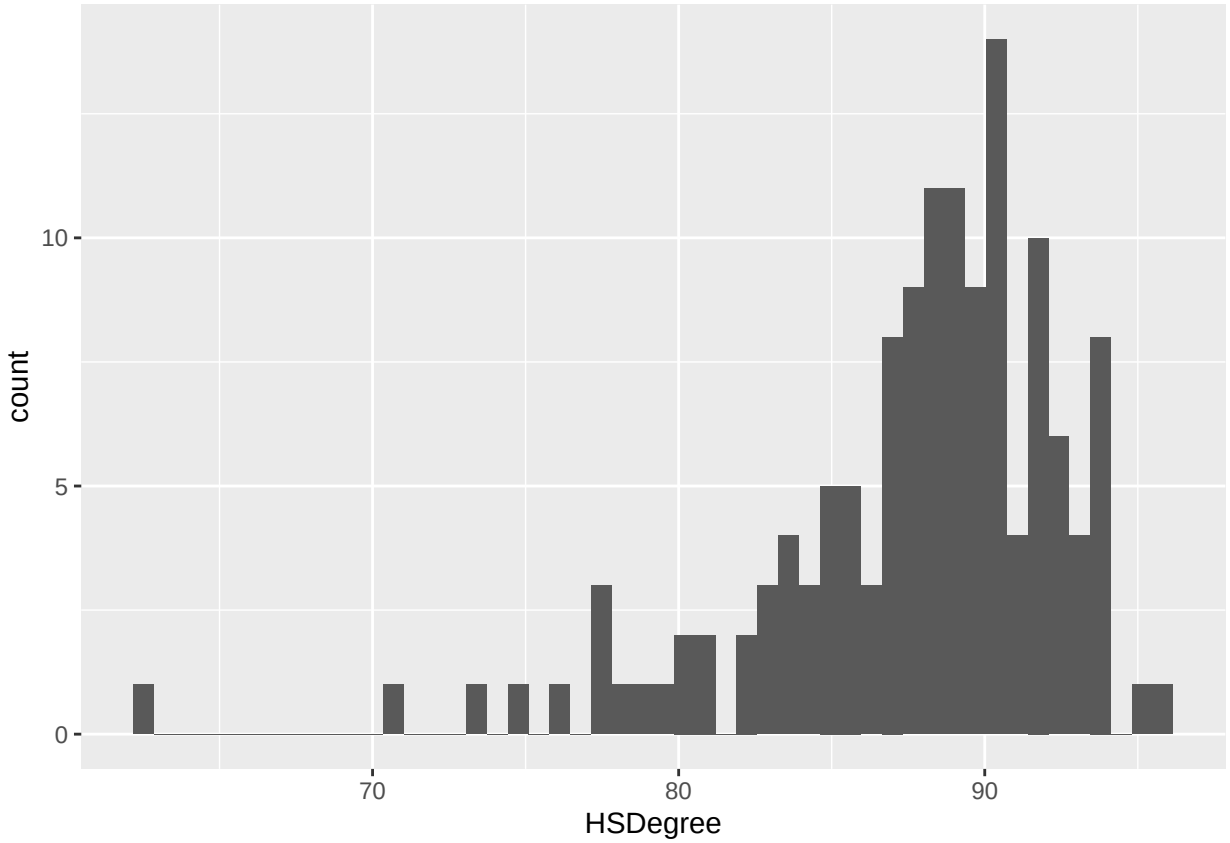
```
nrow(acs_df)
```

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

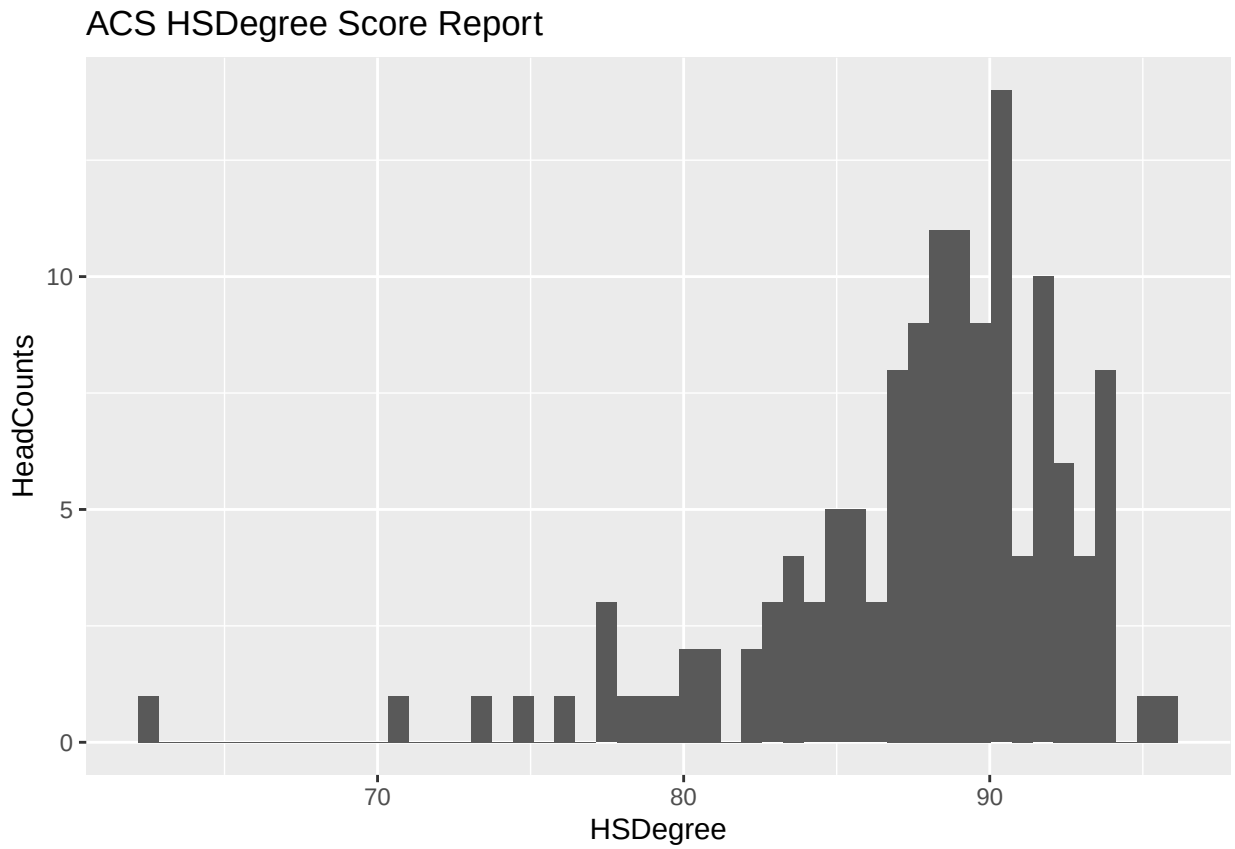
```
ncol(acs_df)
```

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

```
##  Create a Histogram of the HSDegree variable using the ggplot2 package.
##    1.  Set a bin size for the Histogram that you think best visuals the data (the bin size will dete
##        how many bars display and how wide they are)
##    2.  Include a Title and appropriate X/Y axis labels on your Histogram Plot.

library(ggplot2)

ggplot(acs_df, aes(HSDegree)) + geom_histogram(bins = 50)
```

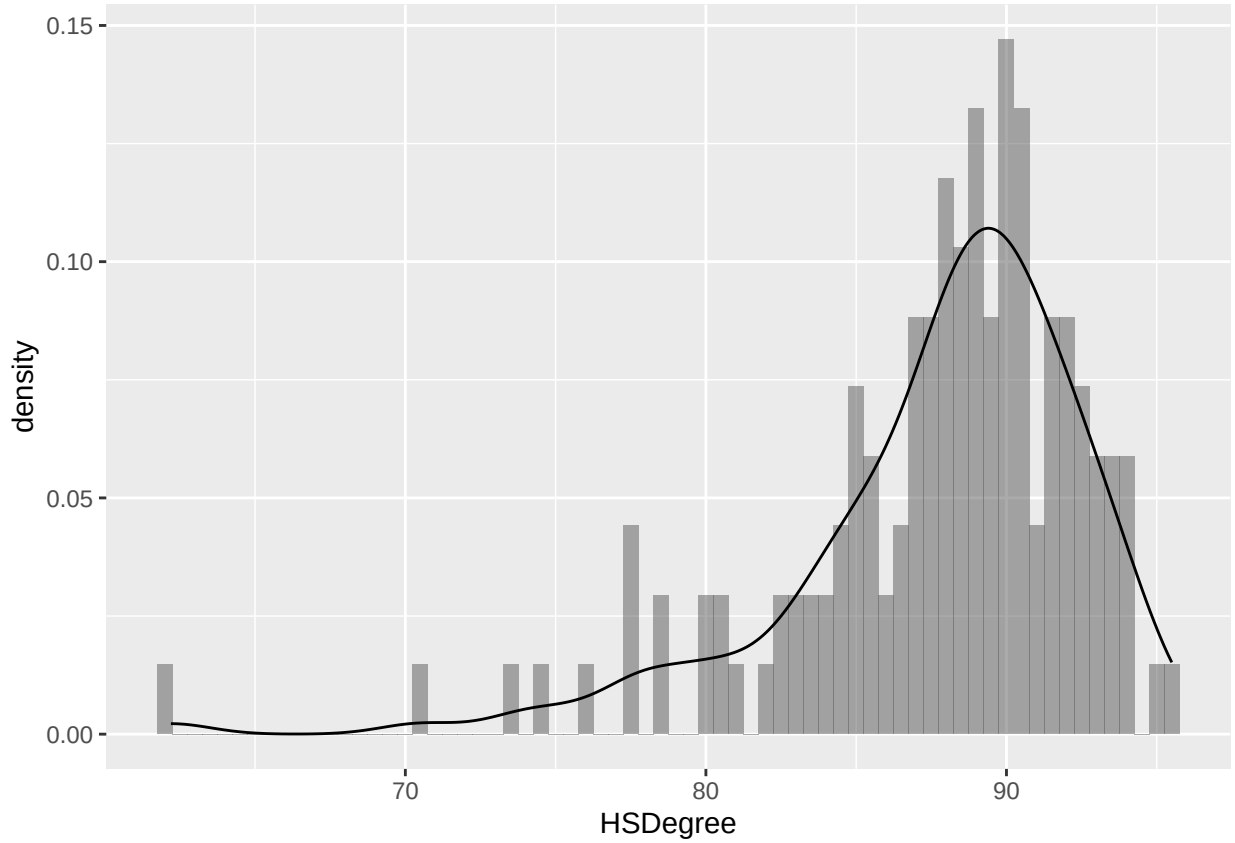



```
ggplot(acs_df, aes(HSDegree)) + geom_histogram(bins = 50) + ggtitle("ACS HSDegree Score Report") + xlab
```

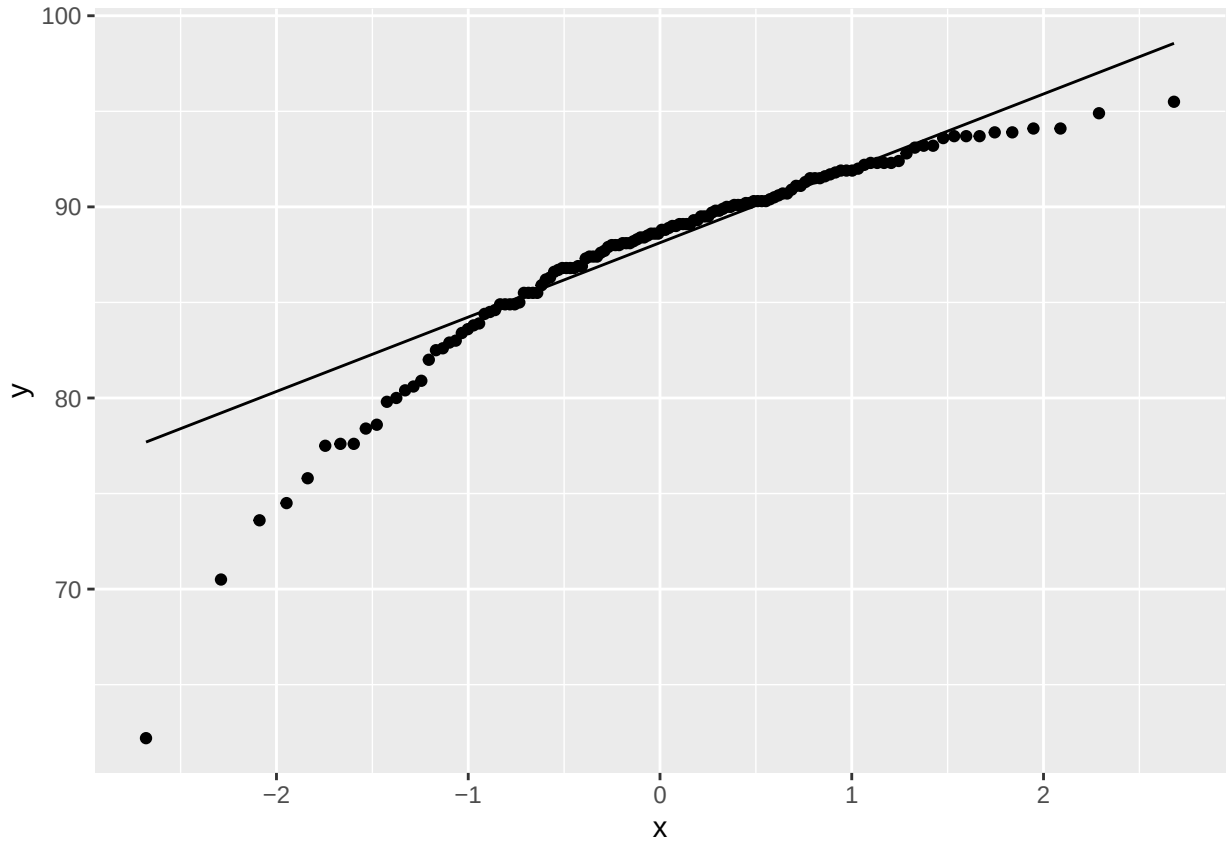



```
## Answer the following questions based on the Histogram produced:
##   Based on what you see in this histogram, is the data distribution unimodal?
#    Answer- Plot is unimodal
##   Is it approximately symmetrical?
#    Answer- Plot is asymmetrical
##   Is it approximately bell-shaped?
#    Answer- Plot is relatively bell shaped
##   Is it approximately normal?
#    Answer- Plot is not relatively normal
##   If not normal, is the distribution skewed? If so, in which direction?
#    Answer- Plot is skewed left ( left skew or Negative skew )
##   Include a normal curve to the Histogram that you plotted.
#    Reference link :
##      https://statisticsglobe.com/normal-density-curve-on-top-of-histogram-ggplot2-r
##      http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-

    ggplot(acs_df, aes(HSDegree)) + geom_histogram(bins=30,binwidth=.5, aes(y=..density..), position="
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
```




```
## Explain whether a normal distribution can accurately be used as a model for this data.
#   Answer -  Normal distribution properties expect symmetrical data distribution and majority of the d
#             within the std. deviation, and data skewness must be zero. Plot drawn is relative to norm
#             but not be accurate.


#     Create a Probability Plot of the HSDegree variable.
    ggplot(acs_df, aes(sample = HSDegree)) +  stat_qq() + stat_qq_line() +  theme(legend.position="top
```




```
    # reference - https://www.geeksforgeeks.org/normal-probability-plot/
    # reference - https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0
#     Answer the following questions based on the Probability Plot:
#       1. Based on what you see in this probability plot, is the distribution approximately normal? Ex
#       Answer - Points plotted on the graph are not perfectly lies on a straight line to indicate dist

#       2. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.
#       Answer - Noticed more deviation at the Bottom end of the QQ plot from straight line, indicate l
#                Left Skewed or Negatively Skewed.

#     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.
    # Reference : https://stats.oarc.ucla.edu/r/faq/how-can-i-get-a-table-of-basic-descriptive-statis
    library(pastecs)
    options(scipen=100)
    options(digits=2)

    #-kurtosis
    stat.desc(acs_df)
```

```
##          Id          Id2 Geography PopGroupID POPGROUP.display.label
## nbr.val  NA       136.00        NA        136                      NA
## nbr.null NA         0.00        NA          0                      NA
## nbr.na   NA         0.00        NA          0                      NA
```

```
## min       NA      1073.00        NA          1                    NA
## max       NA     55079.00        NA          1                    NA
## range     NA     54006.00        NA          0                    NA
## sum       NA   3649306.00        NA        136                    NA
## median    NA     26112.00        NA          1                    NA
## mean      NA     26833.13        NA          1                    NA
## SE.mean   NA      1323.04        NA          0                    NA
## CI.mean   NA      2616.56        NA          0                    NA
## var       NA 238057576.23        NA          0                    NA
## std.dev   NA     15429.11        NA          0                    NA
## coef.var  NA         0.58        NA          0                    NA
##             RacesReported  HSDegree BachDegree
## nbr.val            136.00   136.000     136.00
## nbr.null             0.00     0.000       0.00
## nbr.na               0.00     0.000       0.00
## min             500292.00    62.200      15.40
## max           10116705.00    95.500      60.30
## range          9616413.00    33.300      44.90
## sum          155638535.00 11918.000    4822.70
## median          832707.50    88.700      34.10
## mean           1144400.99    87.632      35.46
## SE.mean          93510.28     0.439       0.82
## CI.mean         184934.56     0.868       1.61
## var      1189207460962.57    26.193      90.43
## std.dev       1090507.89     5.118       9.51
## coef.var            0.95     0.058       0.27
```

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

```
##          nbr.val         nbr.null           nbr.na              min
##   136.0000000000     0.0000000000     0.0000000000    62.2000000000
##              max            range              sum           median
##    95.5000000000    33.3000000000 11918.0000000000    88.7000000000
##             mean          SE.mean     CI.mean.0.95              var
##    87.6323529412     0.4388597852     0.8679296080    26.1933159041
##          std.dev         coef.var         skewness         skew.2SE
##     5.1179405921     0.0584024098    -1.6747666105    -4.0302539978
##         kurtosis         kurt.2SE       normtest.W       normtest.p
##     4.3528564623     5.2738853364     0.8773635436     0.0000000032
```

```
      #-- Finding z-scores
      zscore <- ( acs_df$HSDegree - mean(acs_df$HSDegree)) / sd(acs_df$HSDegree)
      zscore
```

```
##   [1]  0.2868 -0.1626  0.0718 -0.1431  0.2281 -2.7418 -2.5659 -1.9798 -0.5925
##  [10] -1.3741 -0.1626 -1.7648 -0.2017  0.0914 -1.9602  0.0914 -0.0454 -0.0063
##  [19] -1.8039 -0.7879  0.8339 -0.4166  1.0097  1.2637  0.4235  0.3258  0.3649
##  [28]  0.4822  0.5017  0.7752  0.1500  0.2672 -0.0649 -0.2603 -1.3154  0.0523
##  [37]  0.0132  0.4822 -0.5339  0.2477  0.5212  0.1500  0.7166  0.0718  0.8143
##  [46] -0.4166  0.9120 -0.9247  0.5212  0.5994 -0.5143  1.5373  0.2281  0.1695
##  [55]  0.8339  0.5408  0.6385 -0.4166 -0.6316 -1.0028  0.2868  0.9120  1.2637
##  [64]  0.8925 -0.7293  0.4822  0.2868  0.3258  1.1660 -0.5339  1.0879  0.4431
##  [73]  0.4626  1.0879  0.1109 -0.6120  0.7557  0.1305 -0.4166 -0.8270  0.2868
```

```
##  [82]  1.0683  0.7948 -0.7488 -0.2799  0.0718 -3.3475  0.5799 -1.4913  0.5212
##  [91]  0.5994 -0.1626 -1.4131  0.4235 -0.0454  0.2672  0.3649  0.9316  0.0914
## [100]  0.4626  0.5603  0.4040  0.6775 -0.1626  0.1891  0.6775  0.5017  1.2246
## [109]  1.2246  0.9120  0.7557 -0.5339  1.1856 -0.9833 -1.1005 -0.1822 -0.0454
## [118] -0.9051  1.1856 -1.9602  0.8339 -2.3119  0.1891 -1.5304 -4.9693 -0.3385
## [127] -0.5339  0.1891  0.3649  1.1856  0.7557  0.9120  0.5212  0.8534  1.4200
## [136] -0.1431
```

```
      #stat.desc(acs_df, basic=F)
      #stat.desc(acs_df, desc=F)
      #data(acs_df)
      #-kurtosis
      #stat.desc(acs_df$HSDegree, norm = TRUE)
      #or
      #stat.desc(acs_df[,7], norm = TRUE)



#     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 you


#      Results and Explaination :
#
#      Plotted graph for the 2014 American Community Survey dataset is negatively skewed, Mode exceeds
#      visual observation shows the plot tailed at the left; even the probability QQ plot indicates mor
#      the bottom of the graph from a straight line. Another indicator using stat.desc() clearly show s
#
#      Kurtesis measures the degree of peak ness of a frequency distribution, plot derived on HSDegree
#      categorized as MesoKurtic and stat.desc() positive value indicates high peaks.
#
#      z-score is the number of standard deviations a given data point lies above or below the mean, to
#      mean and std deviation need to know for a given data point. z-score values derived using the bel
#      HSDegree data point shows positive and negative values. Results of zero show the point and the m
#      a result of the positive value indicates the deviation above the mean, and negative values indic
#
#      zscore <- ( acs_df$HSDegree - mean(acs_df$HSDegree)) / sd(acs_df$HSDegree)
#
#      Adding,subtracting, multiplying and dividing contant may not impact sample data, however addting
#      impact the balancing point i.e mean. In fact, adding a data point to the set, or taking one away
#
#      example If we add a data point that's above the mean, or take away a data point that's below the
#      will increase. If take away a data point that's above the mean, or add a data point that's below
#      the mean will decrease.
#
#      It's also important that we realize that adding or removing an extreme value from the data set w
#      example : data set (1,2,3) Mean=2 and Medin=2, adding extreme data point like 1000,will change m
```