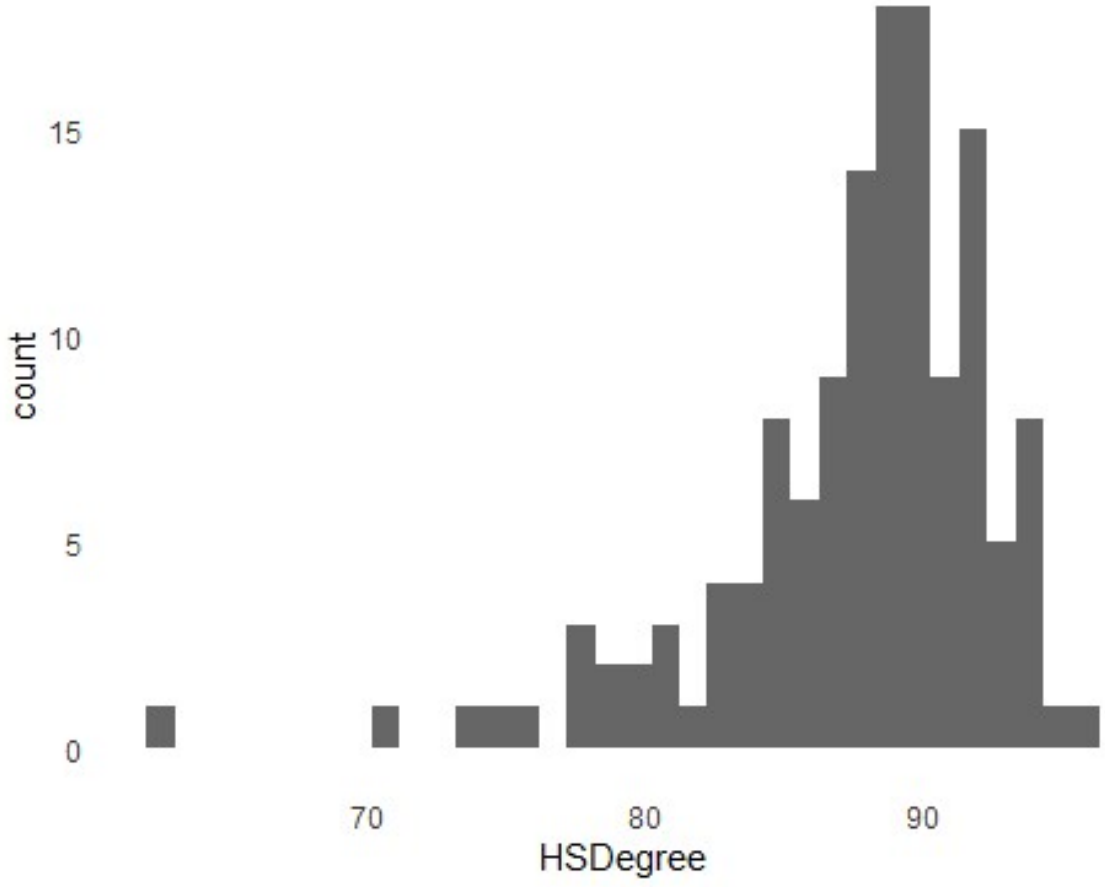
assignment_3_2_MukherjeeChitramoy.R

chitro

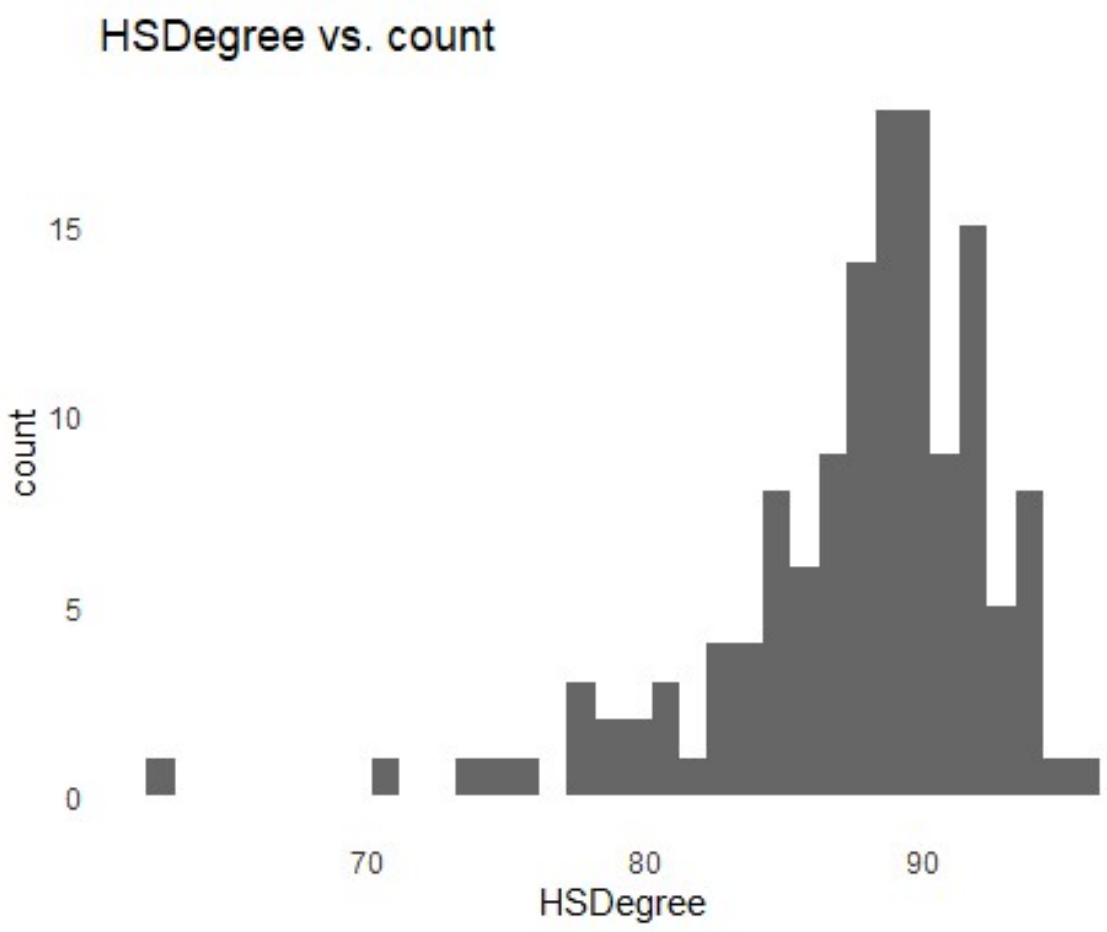
2022-12-19

```
# Assignment: ASSIGNMENT 3.2
# Name: Mukherjee, Chitramoy
# Date: 2022-12-16

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

## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/chitro/Desktop/dsc520-fork-chitro")

## Load the `data/r4ds/heights.csv` to
american_cs <- read.csv("data/acs-14-1yr-s0201.csv")

#Id : Datatype - varchar(contains text and number) Intent: unique
identifier for each row
#Id2 : Dadatype : Integer (contains last 4 byte of Id) Intent : unique
identifier for each
#Geography: Varchar(contains text) Intent : County and State
information
#PopGroupID : Integer(contains number) : Intent : Numeric 1 value for
all rows
#POPGROUP.display.label : varchar, Intent : Total population for all
rows
#RacesReported : Integer(contains numbers) Intent : total population
count
#HSDegree : Numeric(contains number with one decimal point) Intent :
marks obtained in HS Degree
#BachDegree : Numeric(contains number with one decimal point) Intent :
marks obtained in Bachelore Degree

#Display structure of american_cs
str(american_cs)

## '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" "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 ...

#Display no. of rows in american_cs
nrow(american_cs)

## [1] 136

#Display no. of columns in american_cs
ncol(american_cs)

## [1] 8

#Create a Histogram of the HSDegree variable using the ggplot2 package
ggplot(american_cs, aes(x=HSDegree)) + geom_histogram(bins=34)
```




```
ggplot(american_cs, aes(x=HSDegree)) + geom_histogram(bins=34) +
ggtitle("HSDegree vs. count") + xlab("HSDegree") + ylab("count")
```



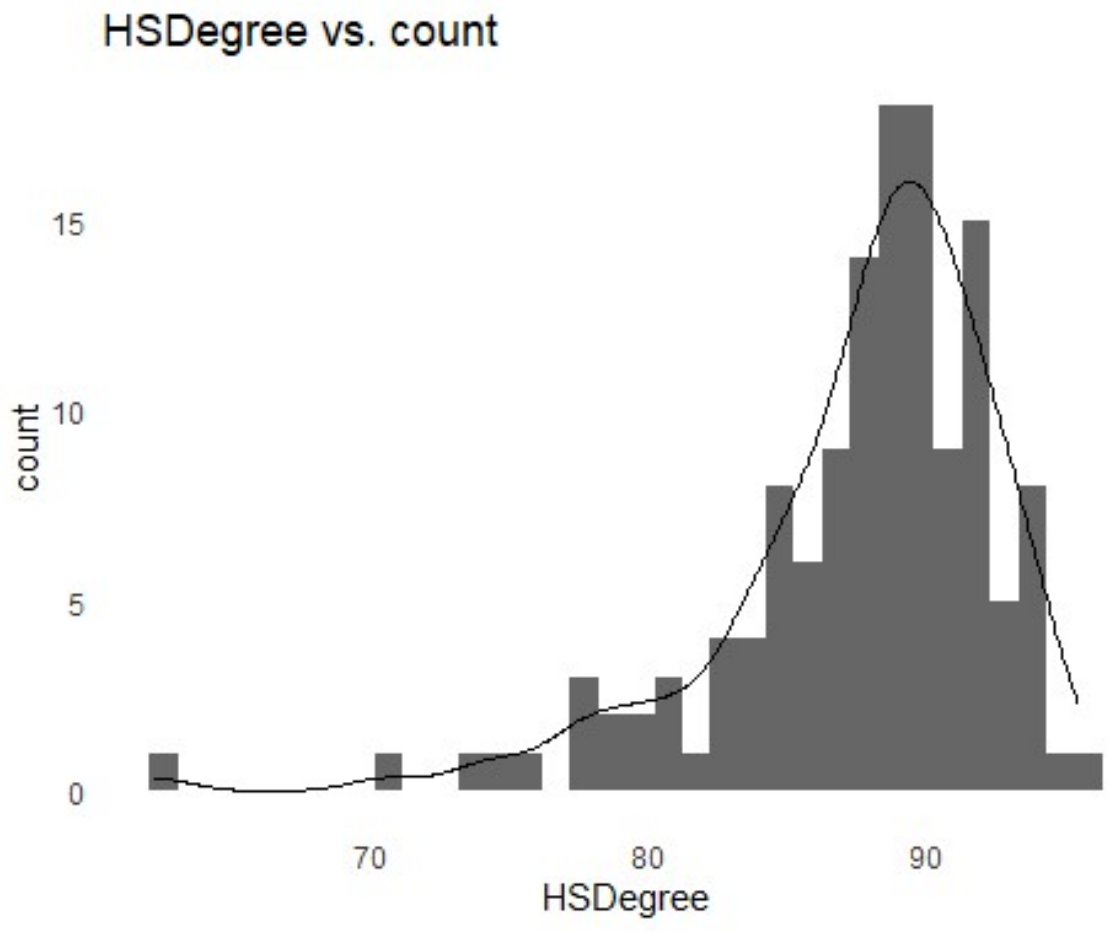

```
sapply(american_cs[,7:8], sd)
```

```
##   HSDegree BachDegree
##   5.117941   9.509731
```

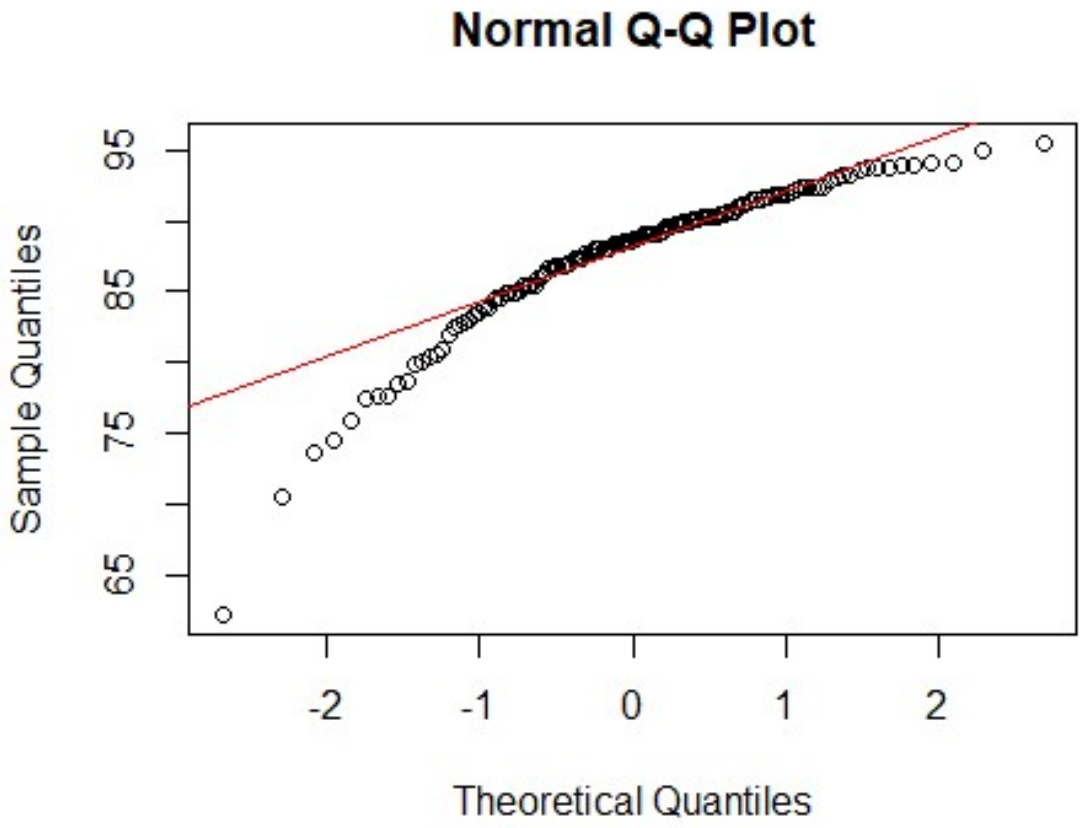
IV.1 As the distribution has 2 peak, it's bimodal distribution.
IV.2 the histogram is not approximately symmetrical as we cant draw a vertical line at some point in the histogram such that
the shape to the left and the right of the vertical line are mirror images of each other.
IV.3 It is not approximately bell-shaped as 68% of the data is not within 1 standard deviation of the mean.
IV.4 It is Not approximately normal as the data is skewed towards the left side of the plotting.
IV.5 It is skewed towards the left side as per the histogram.
IV.6 Include a normal curve to the Histogram that you plotted :

```
ggplot(american_cs, aes(x=HSDegree)) + geom_histogram(bins=34) +
ggtitle("HSDegree vs. count") + xlab("HSDegree") + ylab("count") +
geom_density(aes(y=1.1*..count..))
```

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




```
# Normal distribution can be used as a model for this data as it
replicates the pattern shown in histogram.

qqnorm(american_cs$HSDegree)
qqline(american_cs$HSDegree, col = "red")
```



VI.1 Distribution is not approximately normal distribution has more values in the tails compared to a normal distribution.
VI.2 Distribution is left skewed as per the plotting.

```
library (pastecs)
stat.desc (american_cs)
```

```
##          Id          Id2 Geography PopGroupID
POPGROUP.display.label
## nbr.val  NA 1.360000e+02        NA         136
NA
## nbr.null NA 0.000000e+00        NA           0
NA
## nbr.na   NA 0.000000e+00        NA           0
NA
## min      NA 1.073000e+03        NA           1
NA
## max      NA 5.507900e+04        NA           1
```

```
NA
## range    NA 5.400600e+04        NA          0
NA
## sum      NA 3.649306e+06        NA        136
NA
## median   NA 2.611200e+04        NA          1
NA
## mean     NA 2.683313e+04        NA          1
NA
## SE.mean  NA 1.323036e+03        NA          0
NA
## CI.mean  NA 2.616557e+03        NA          0
NA
## var      NA 2.380576e+08        NA          0
NA
## std.dev  NA 1.542911e+04        NA          0
NA
## coef.var NA 5.750024e-01        NA          0
NA
##          RacesReported     HSDegree   BachDegree
## nbr.val   1.360000e+02 1.360000e+02  136.0000000
## nbr.null  0.000000e+00 0.000000e+00    0.0000000
## nbr.na    0.000000e+00 0.000000e+00    0.0000000
## min       5.002920e+05 6.220000e+01   15.4000000
## max       1.011671e+07 9.550000e+01   60.3000000
## range     9.616413e+06 3.330000e+01   44.9000000
## sum       1.556385e+08 1.191800e+04 4822.7000000
## median    8.327075e+05 8.870000e+01   34.1000000
## mean      1.144401e+06 8.763235e+01   35.4610294
## SE.mean   9.351028e+04 4.388598e-01    0.8154527
## CI.mean   1.849346e+05 8.679296e-01    1.6127146
## var       1.189207e+12 2.619332e+01   90.4349886
## std.dev   1.090508e+06 5.117941e+00    9.5097313
## coef.var  9.529072e-01 5.840241e-02    0.2681741
```

```
library(moments)
skewness(american_cs$HSDegree)
```

```
## [1] -1.69341
```

```
# Since the skewness is negative, this indicates that the distribution
is left-skewed. This confirms what we saw in the histogram.
kurtosis(american_cs$HSDegree)
```

```
## [1] 7.462191
```

```
# Since the kurtosis is greater than 3, this indicates that the
distribution has more values in the tails compared to a normal
distribution.
```

```
a<- (american_cs$HSDegree)
head(a)
```

```
## [1] 89.1 86.8 88.0 86.9 88.8 73.6
```

```
mean(a)
```

```
## [1] 87.63235
```

```
sd(a)
```

```
## [1] 5.117941
```

```
a.z <- (a -mean(a)) / sd(a)
a.z
```

```
##   [1]  0.286765161 -0.162634350  0.071834960 -0.143095241
0.228147834
##
[6] -2.741796762 -2.565944779 -1.979771504 -0.592494752 -1.374059119
##  [11] -0.162634350 -1.764841303 -0.201712568
0.091374069 -1.960232394
##  [16]
0.091374069 -0.045399695 -0.006321476 -1.803919521 -0.787885844
##  [21]  0.833860218 -0.416642769  1.009712201  1.263720620
0.423538925
##  [26]  0.325843380  0.364921598  0.482156253  0.501695362
0.775242891
##  [31]  0.149991397
0.267226052 -0.064938804 -0.260329896 -1.315441791
##  [36]  0.052295851  0.013217633  0.482156253 -0.533877424
0.247686943
##  [41]  0.521234471  0.149991397  0.716625563  0.071834960
0.814321109
##  [46] -0.416642769  0.912016655 -0.924659608  0.521234471
0.599390908
##  [51] -0.514338315  1.537268149  0.228147834  0.169530506
0.833860218
##  [56]  0.540773581
0.638469126 -0.416642769 -0.631572970 -1.002816045
##  [61]  0.286765161  0.912016655  1.263720620
0.892477546 -0.729268516
##  [66]  0.482156253  0.286765161  0.325843380
1.166025074 -0.533877424
##  [71]  1.087868638  0.443078035  0.462617144  1.087868638
0.110913179
##  [76] -0.612033861  0.755703781
0.130452288 -0.416642769 -0.826964062
##  [81]  0.286765161  1.068329528
0.794782000 -0.748807625 -0.279869005
```

```
##  [86]  0.071834960 -3.347509146  0.579851799 -1.491293774
0.521234471
##  [91]  0.599390908 -0.162634350 -1.413137337
0.423538925 -0.045399695
##  [96]  0.267226052  0.364921598  0.931555764  0.091374069
0.462617144
## [101]  0.560312690  0.403999816  0.677547345 -0.162634350
0.189069615
## [106]  0.677547345  0.501695362  1.224642402  1.224642402
0.912016655
## [111]  0.755703781 -0.533877424
1.185564183 -0.983276935 -1.100511591
## [116] -0.182173459 -0.045399695 -0.905120499
1.185564183 -1.960232394
## [121]  0.833860218 -2.311936360
0.189069615 -1.530371992 -4.969255208
## [126] -0.338486333 -0.533877424  0.189069615  0.364921598
1.185564183
## [131]  0.755703781  0.912016655  0.521234471  0.853399327
1.420033494
## [136] -0.143095241
```

 # Z-score is the distance of raw score value from the mean in terms of standard deviation.Raw scores above the mean have a positive Z-score value, while a raw score below the mean has a negative z-score value.

Introduction of more sample data in the dataset having HSDegree value less than the mean value will reduce the meand and will bring the data more towards normal distribution instead of left skew.