assignment_03-2_RathShakti.R

shakr

2022-09-18

```
# Assignment: ASSIGNMENT 3
# Name: Rath, Shakti
# Date: 2022-09-18

## 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/shakr/OneDrive/Desktop/shakti-data/shakti/Rcode/dsc520")

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

## i. List the name of each field and what you believe the data type and
##intent is of the data included in each field (Example: Id - Data Type: varchar (contains text and numb
##Intent: unique identifier for each row)
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 ...
```

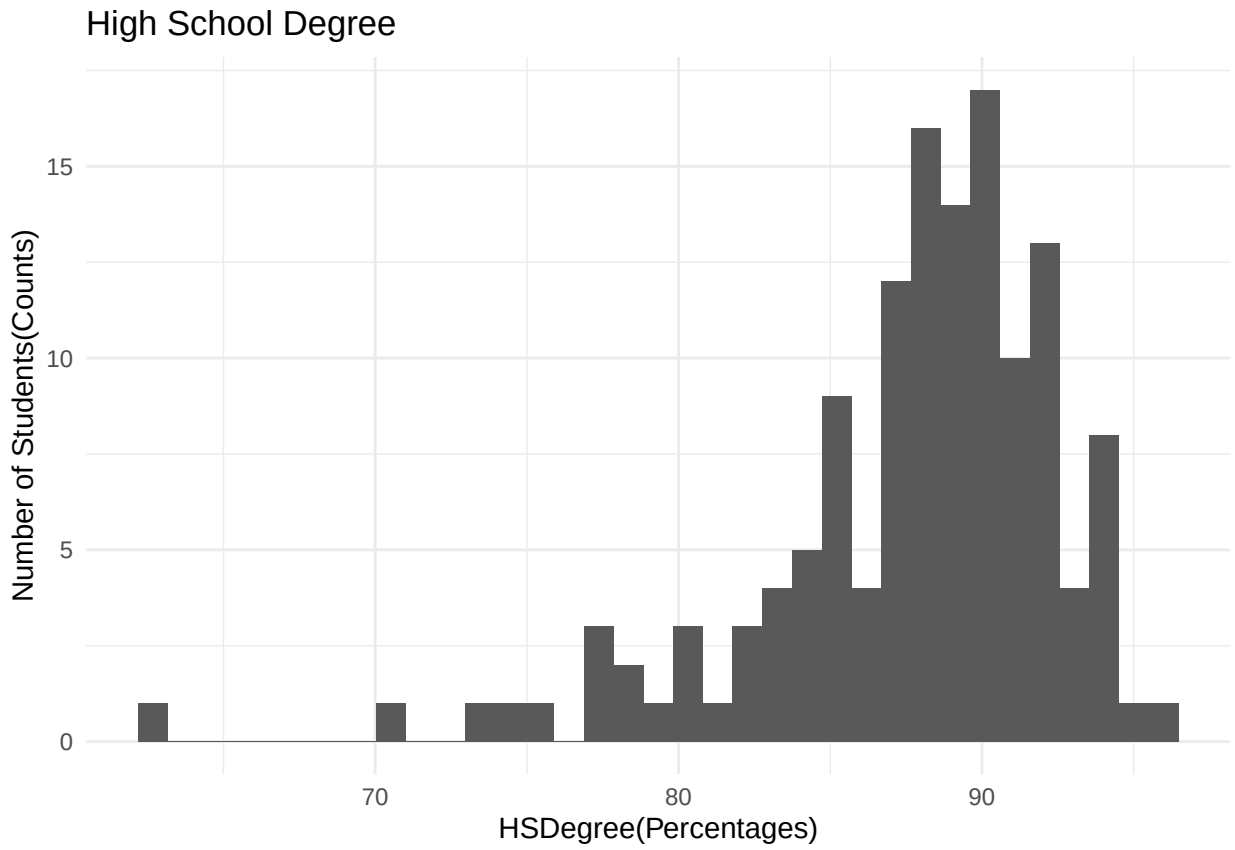
```
## ii. Run the following functions and provide the results: str(); nrow(); ncol()
nrow(acs_df)
```

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

```
ncol(acs_df)
```

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

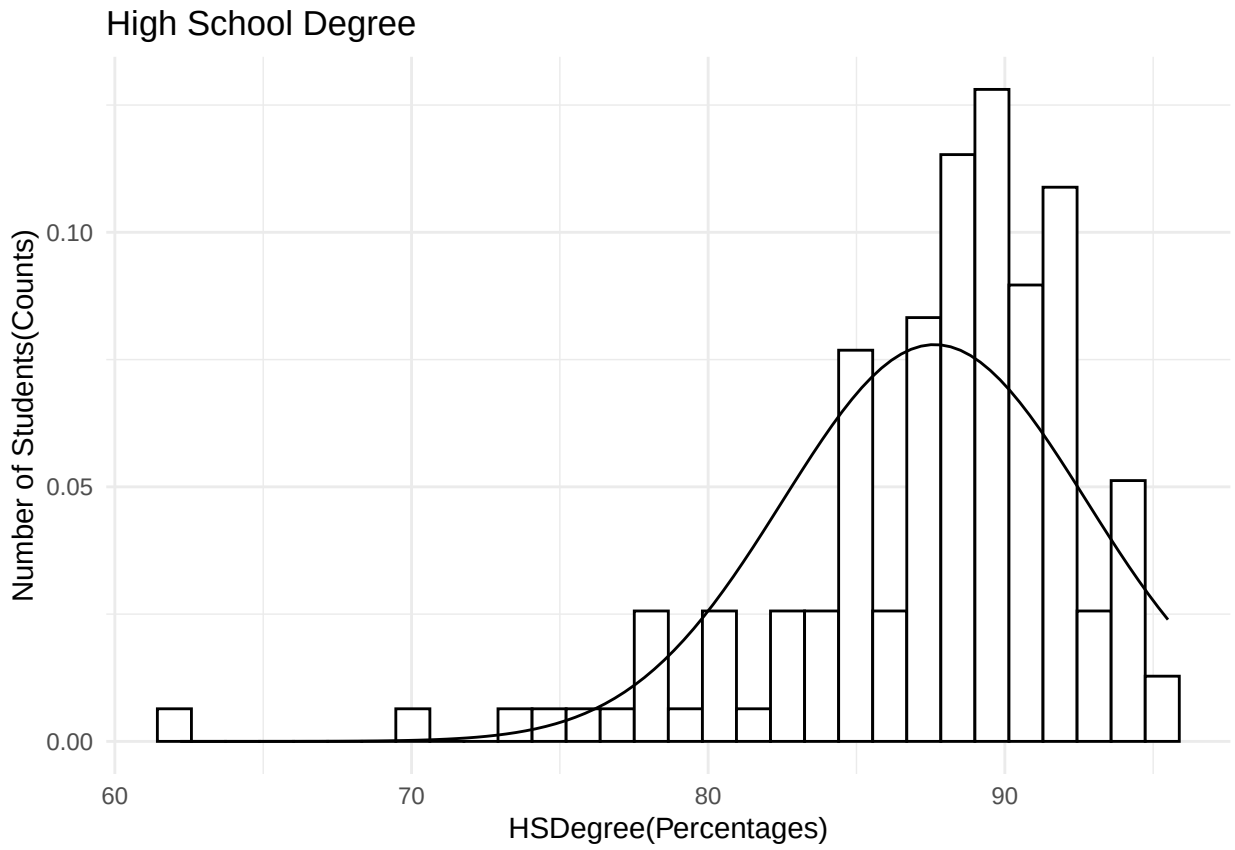
```
## iii. Create a Histogram of the HSDegree variable using the ggplot2 package.
## 1.Set a bin size for the Histogram that you think best visuals the
##data (the bin size will determine how many bars display and how wide they are)
## 2.Include a Title and appropriate X/Y axis labels on your Histogram Plot.
ggplot(acs_df, aes(x = HSDegree)) +
  geom_histogram(bins = 35) +
  ggtitle("High School Degree") +
  xlab("HSDegree(Percentages)") +
  ylab("Number of Students(Counts)")
```



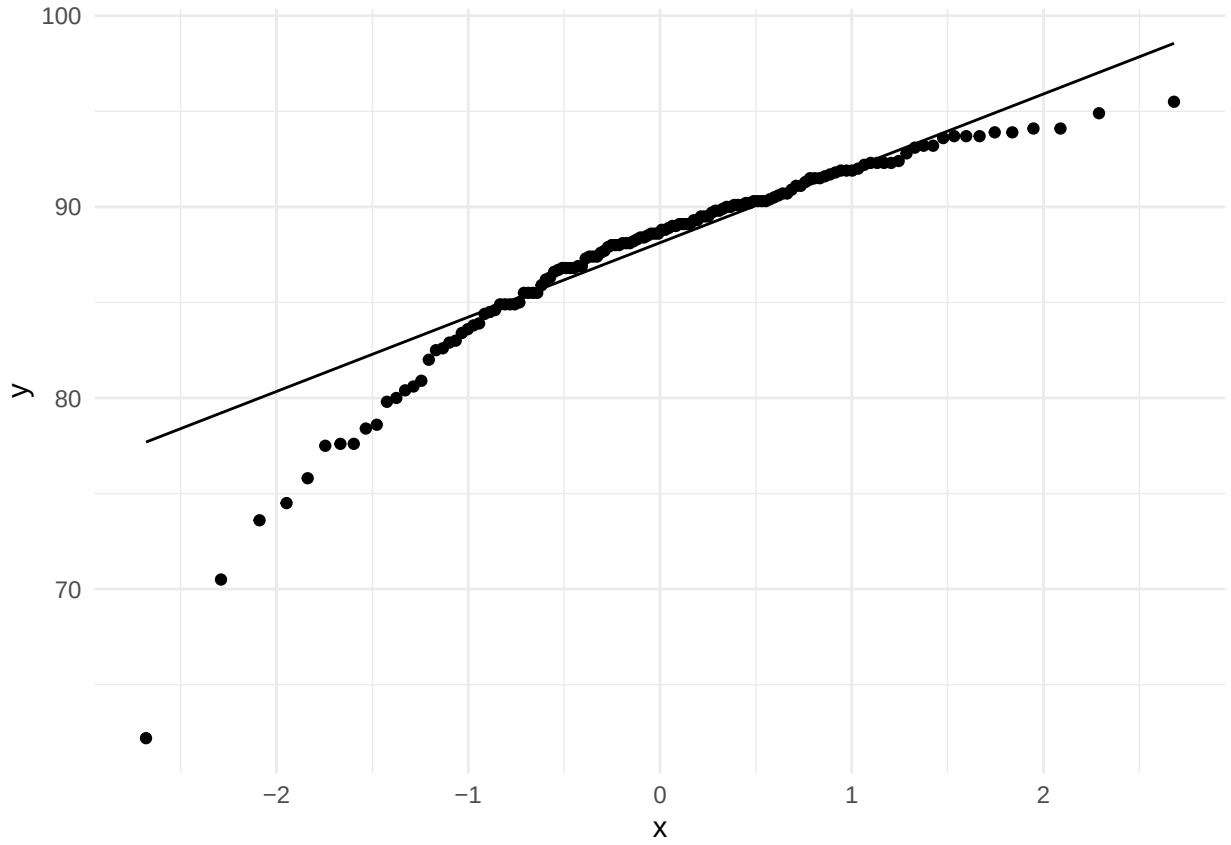

```
## iv. Answer the following questions based on the Histogram produced:
## 1.Based on what you see in this histogram, is the data distribution unimodal?
##[Answer]: No

## 2.Is it approximately symmetrical?
##[Answer]: No

## 3.Is it approximately bell-shaped?
##[Answer]: No

## 4.Is it approximately normal?
##[Answer]: No

## 5.If not normal, is the distribution skewed? If so, in which direction?
##[Answer]: It is Left Skewed distribution.
```

```
## 6.Include a normal curve to the Histogram that you plotted.
ggplot(acs_df, aes(x = HSDegree)) +
  geom_histogram(aes(y = ..density..), colour="black", fill="white") +
  ggtitle("High School Degree") +
  xlab("HSDegree(Percentages)") +
  ylab("Number of Students(Counts)") +
  stat_function(fun = dnorm, args = list(mean = mean(acs_df$HSDegree), sd = sd(acs_df$HSDegree)))
```

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




```
## 7.Explain whether a normal distribution can accurately be used as a model for this data
##[Answer]: Yes. A normal distribution can be accurately used. The normal distribution is a mount-shape

## 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
##[Answer]: It is not normal distribution. The data provided in this datset is not forming a straight.
## So, it is not a normal distribution

## 2. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.
##[Answer]: Yes it is Left 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.
library(pastecs)
stat.desc(acs_df)
```

```
##          Id          Id2 Geography PopGroupID POPGROUP.display.label RacesReported     HSDegree    Ba
## nbr.val  NA 1.360000e+02        NA        136                     NA  1.360000e+02 1.360000e+02   136
## nbr.null NA 0.000000e+00        NA          0                     NA  0.000000e+00 0.000000e+00     0
## nbr.na   NA 0.000000e+00        NA          0                     NA  0.000000e+00 0.000000e+00     0
## min      NA 1.073000e+03        NA          1                     NA  5.002920e+05 6.220000e+01    15
## max      NA 5.507900e+04        NA          1                     NA  1.011671e+07 9.550000e+01    60
## range    NA 5.400600e+04        NA          0                     NA  9.616413e+06 3.330000e+01    44
## sum      NA 3.649306e+06        NA        136                     NA  1.556385e+08 1.191800e+04 4822
## median   NA 2.611200e+04        NA          1                     NA  8.327075e+05 8.870000e+01    34
## mean     NA 2.683313e+04        NA          1                     NA  1.144401e+06 8.763235e+01    35
```

```
## SE.mean  NA 1.323036e+03        NA          0                    NA  9.351028e+04 4.388598e-01    0
## CI.mean  NA 2.616557e+03        NA          0                    NA  1.849346e+05 8.679296e-01    1
## var      NA 2.380576e+08        NA          0                    NA  1.189207e+12 2.619332e+01   90
## std.dev  NA 1.542911e+04        NA          0                    NA  1.090508e+06 5.117941e+00    9
## coef.var NA 5.750024e-01        NA          0                    NA  9.529072e-01 5.840241e-02    0
```

```
## viii. In several sentences provide an explanation of the result produced for skew, kurtosis, and z-s
## In addition, explain how a change in the sample size may change your explanation?
library(moments)
skewness(x = acs_df$HSDegree)
```

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

```
##[Answer]: skewness is negative, this indicates that the distribution is left-skewed.
##This confirms what we saw in the histogram

kurtosis(acs_df$HSDegree)
```

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

```
##[Answer]: kurtosis is greater than 3, this indicates that the distribution has more values in the tai
##compared to a normal distribution.
mdata<-mean(acs_df$HSDegree)
sdata<-sd(acs_df$HSDegree)
z_score<-(acs_df$HSDegree-mdata)/sdata
print(z_score)
```

```
##   [1]  0.286765161 -0.162634350  0.071834960 -0.143095241  0.228147834 -2.741796762 -2.565944779 -1.9
##   [9] -0.592494752 -1.374059119 -0.162634350 -1.764841303 -0.201712568  0.091374069 -1.960232394  0.0
##  [17] -0.045399695 -0.006321476 -1.803919521 -0.787885844  0.833860218 -0.416642769  1.009712201  1.1
##  [25]  0.423538925  0.325843380  0.364921598  0.482156253  0.501695362  0.775242891  0.149991397  0.1
##  [33] -0.064938804 -0.260329896 -1.315441791  0.052295851  0.013217633  0.482156253 -0.533877424  0.1
##  [41]  0.521234471  0.149991397  0.716625563  0.071834960  0.814321109 -0.416642769  0.912016655 -0.9
##  [49]  0.521234471  0.599390908 -0.514338315  1.537268149  0.228147834  0.169530506  0.833860218  0.5
##  [57]  0.638469126 -0.416642769 -0.631572970 -1.002816045  0.286765161  0.912016655  1.263720620  0.8
##  [65] -0.729268516  0.482156253  0.286765161  0.325843380  1.166025074 -0.533877424  1.087868638  0.4
##  [73]  0.462617144  1.087868638  0.110913179 -0.612033861  0.755703781  0.130452288 -0.416642769 -0.8
##  [81]  0.286765161  1.068329528  0.794782000 -0.748807625 -0.279869005  0.071834960 -3.347509146  0.5
##  [89] -1.491293774  0.521234471  0.599390908 -0.162634350 -1.413137337  0.423538925 -0.045399695  0.1
##  [97]  0.364921598  0.931555764  0.091374069  0.462617144  0.560312690  0.403999816  0.677547345 -0.1
## [105]  0.189069615  0.677547345  0.501695362  1.224642402  1.224642402  0.912016655  0.755703781 -0.5
## [113]  1.185564183 -0.983276935 -1.100511591 -0.182173459 -0.045399695 -0.905120499  1.185564183 -1.9
## [121]  0.833860218 -2.311936360  0.189069615 -1.530371992 -4.969255208 -0.338486333 -0.533877424  0.1
## [129]  0.364921598  1.185564183  0.755703781  0.912016655  0.521234471  0.853399327  1.420033494 -0.1
```