DSC520_Week3_Guruprasad_Velikadu_Assignment3.2

Guruprasad Velikadu Krishnamoorthy

2022-12-18

```
# Including the libraries and setting the global theme
library(ggplot2)
theme_set(theme_minimal())
library(formatR)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

Set the working directory to the root of your DSC 520 directory

```
# setting working directory
knitr::opts_knit$set(root.dir = "C:/Users/Gurup/GURU/Learning/Masters/Term_2/DSC520_T302_Statistics_for_
```

1. 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 numbers) Intent: unique identifier for each row)

Solution

i. Id - DataType: varchar (contains text and numbers); Intent: unique identifier for each row

ii. Id2 - DataType: Integer (contains only numbers); Intent: Looks like 5 digit zip code, hence another unique identifier for each county

iii. Geography - DataType: Character; Intent:Description of name of the county

iv. PopGroupID - DataType: Integer; Intent: The dataset has one value of 1. Hence it may be a batchID to representset of population groups

v. POPGROUP.display-label- DataType: Character; Intent:This has the value of "Total population" for all the rows.Hence this may be description column.

vi. RacesReported- Datatype:Integer; Intent: This represents the entire population of the county.

vii. HSDegree - Datatype:Numeric; Intent: This represents the percentage of population of county that has High SchoolDegree.

viii. BachDegree - Datatype:Numeric; Intent: This represents the percentage of population of county that has Bachelor's Degree.

2.Run the following functions and provide the results: str(); nrow(); ncol()

```
population_by_county_df1 <- read.csv(file = "acs-14-1yr-s0201.csv",
    header = TRUE, stringsAsFactors = FALSE)
str(population_by_county_df1)
```

```
## '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(population_by_county_df1)
```

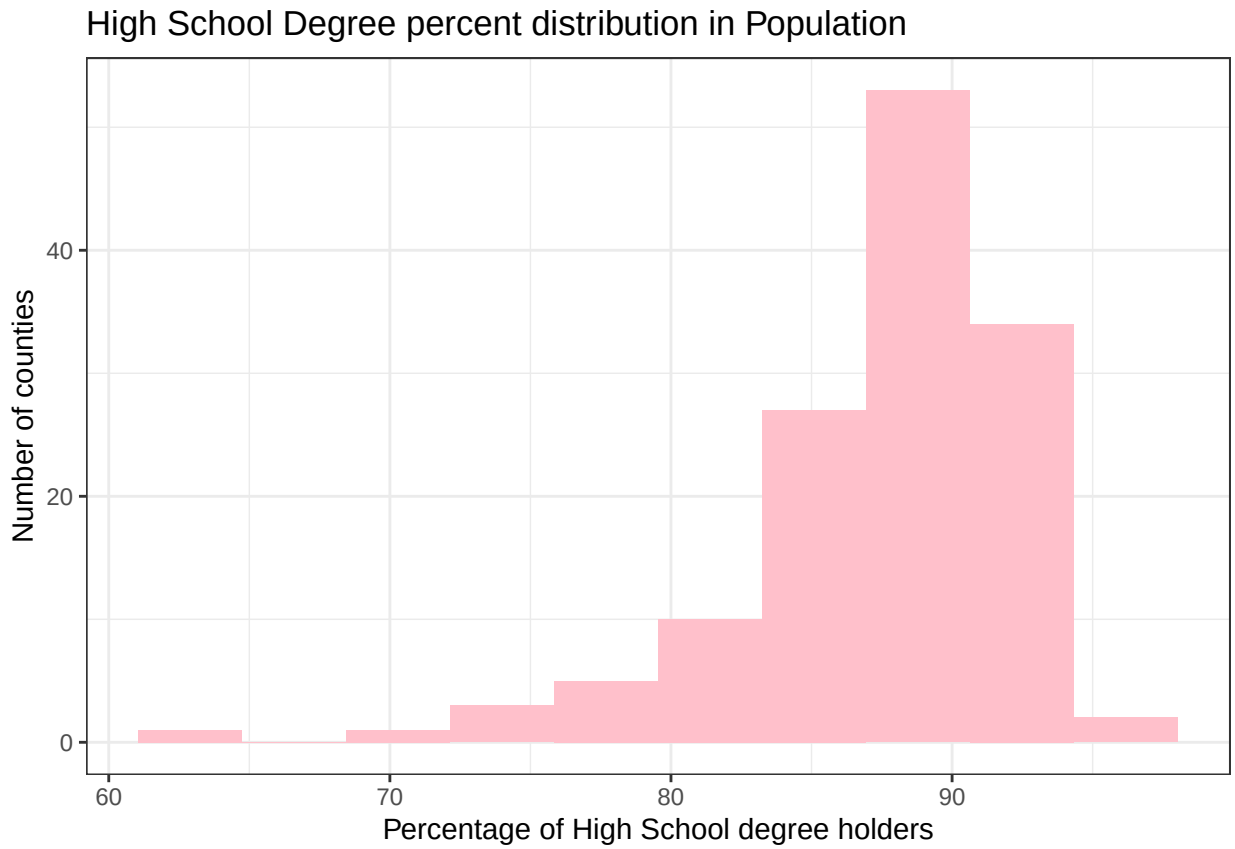
```
## [1] 136
```

```
ncol(population_by_county_df1)
```

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

3.Create a Histogram of the HSDegree variable using the ggplot2 package. 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). Include a Title and appropriate X/Y axis labels on your Histogram Plot.

```
ggplot(data = population_by_county_df1, aes(x = HSDegree)) +
    geom_histogram(bins = 10, fill = "pink") + labs(title = "High School Degree percent distribution in
    x = "Percentage of High School degree holders", y = "Number of counties") +
    theme_bw()
```



4.Answer the following questions based on the Histogram produced:

4.1 Based on what you see in this histogram, is the data distribution unimodal?

Solution: The data distribution is unimodal with one peak

4.2 Is it approximately symmetrical?

Solution: The distribution is NOT symmetrical, because it is skewed and it is asymmetrical.

4.3 Is it approximately bell-shaped?

Solution: Yes it is approximately bell shaped but not exactly bell shaped.

4.4 Is it approximately normal?

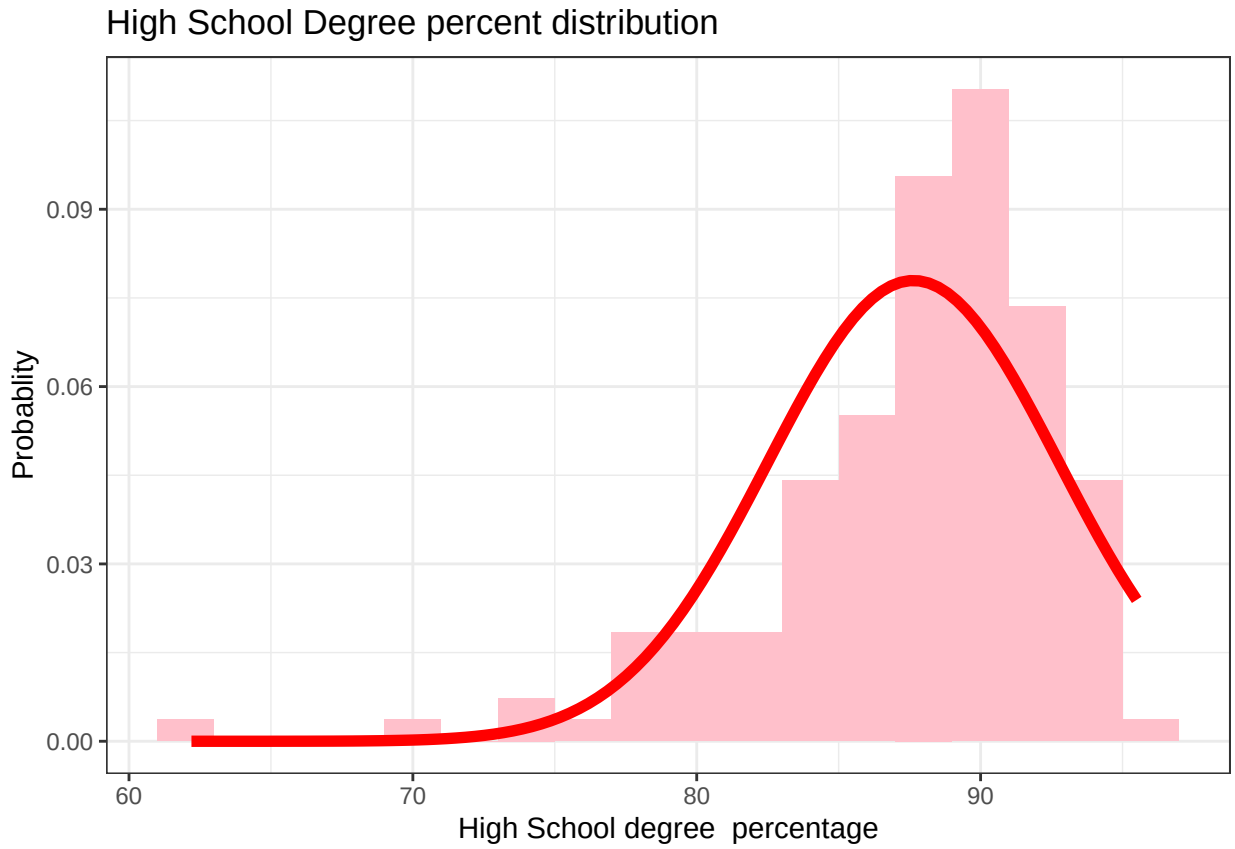
Solution: The distribution is not Normal distribution.

4.5 If not normal, is the distribution skewed? If so, in which direction?

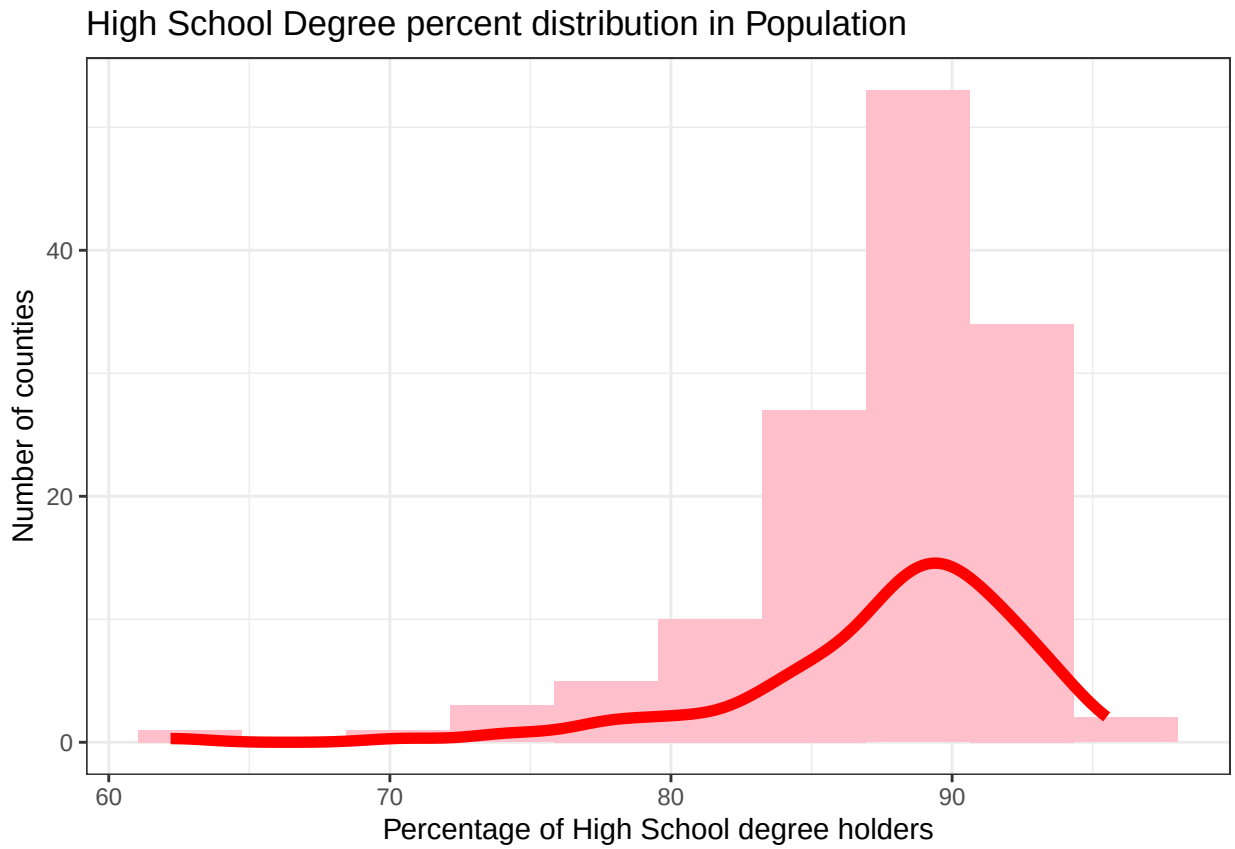
Solution: The distribution is skewed to the left(negative skew)

4.6 Include a normal curve to the Histogram that you plotted.

```
# Plotting curve using stat_function
ggplot(data = population_by_county_df1, aes(x = HSDegree)) +
    geom_histogram(aes(y = after_stat(density)), binwidth = 2,
        fill = "pink") + labs(title = "High School Degree percent distribution",
    x = "High School degree  percentage", y = "Probablity") +
    stat_function(fun = dnorm, args = list(mean = mean(population_by_county_df1$HSDegree),
        sd = sd(population_by_county_df1$HSDegree)), col = "red",
        linewidth = 2) + theme_bw()
```

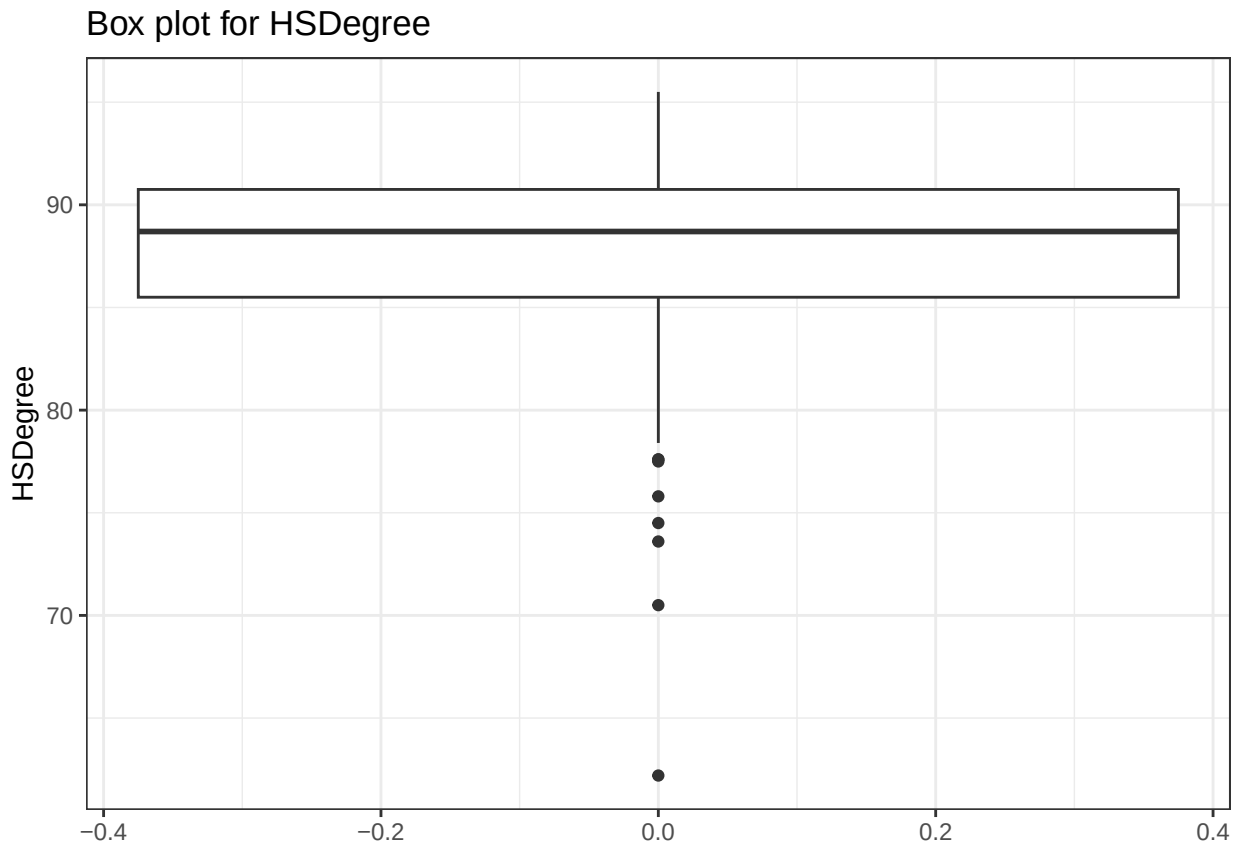



```
# Plotting curve using geom_density
ggplot(data = population_by_county_df1, aes(x = HSDegree)) +
    geom_histogram(bins = 10, fill = "pink") + labs(title = "High School Degree percent distribution in
    x = "Percentage of High School degree holders", y = "Number of counties") +
    geom_density(aes(y = after_stat(count)), col = "red", linewidth = 2) +
    theme_bw()
```



4.7 Explain whether a normal distribution can accurately be used as a model for this data.

```
ggplot(data = population_by_county_df1, aes(y = HSDegree)) +
    geom_boxplot() + theme_bw() + labs(title = "Box plot for HSDegree")
```




```
shapiro.test(population_by_county_df1$HSDegree)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  population_by_county_df1$HSDegree
## W = 0.87736, p-value = 3.194e-09
```

Solution: The distribution deviates from the normal distribution as it is skewed and hence it is not a normal distribution.In the boxplot, the whiskers are of varying length and the lower whisker is longer than the upper one. Also there are many outliers below the lower whisker. All these explains the model is skewed to the left and asymmetrical and deviates from the normal distribution.Also from Shapiro-Wilk normality test, the p value is 0.000000003 which is less than 0.05 which indicates the model is Not a normal distribution

5. Create a Probability Plot of the HSDegree variable.

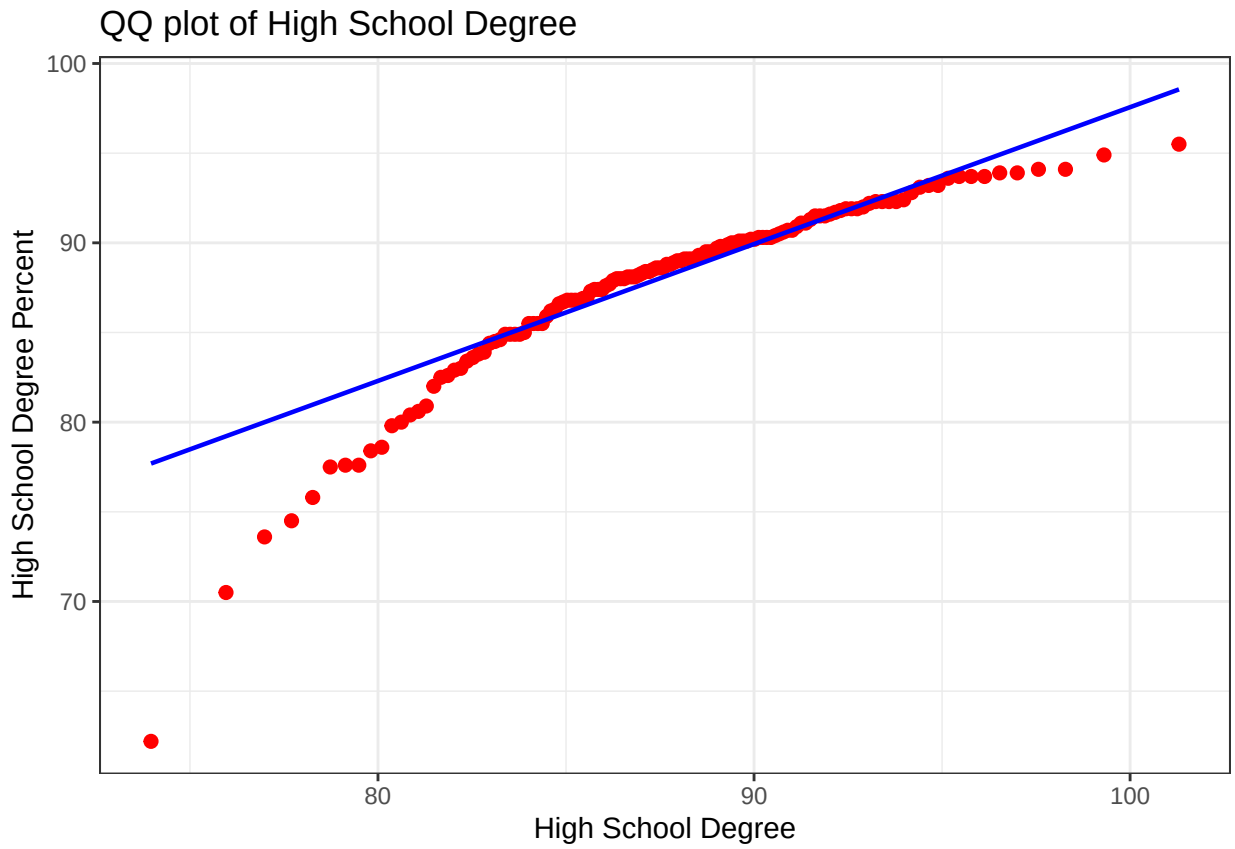
```
library(qqplotr)
```

```
##
## Attaching package: 'qqplotr'
```

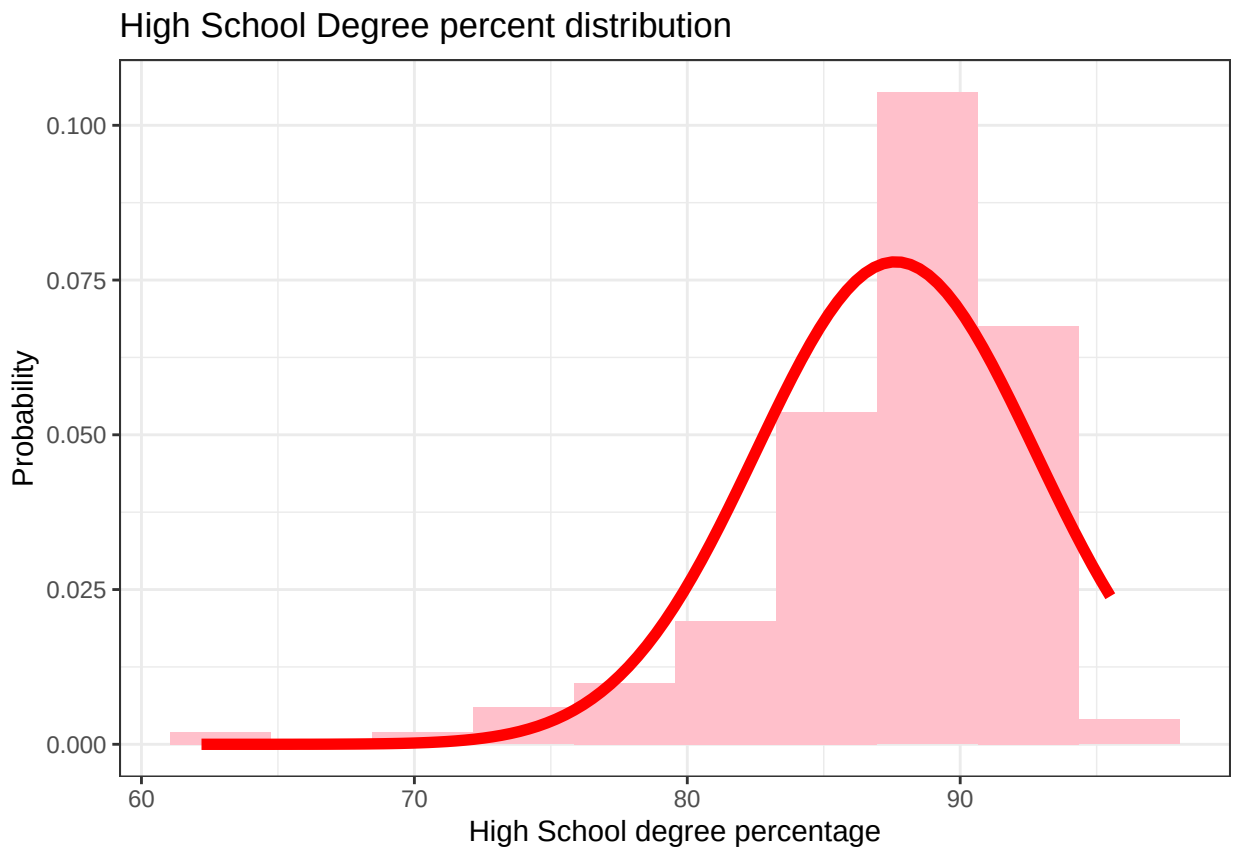
```
## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine
```

```
ggplot(data = population_by_county_df1["HSDegree"], aes(sample = population_by_county_df1$HSDegree)) +
    stat_qq_point(size = 2, color = "red", distribution = "norm") +
    stat_qq_line(color = "blue") + labs(x = "High School Degree",
    y = "High School Degree Percent", title = "QQ plot of High School Degree") +
    theme_bw()
```

```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
# Using stat_function
ggplot(data = population_by_county_df1, aes(x = HSDegree)) +
    geom_histogram(aes(y = after_stat(density)), bins = 10, fill = "pink") +
    labs(title = "High School Degree percent distribution", x = "High School degree percentage",
        y = "Probability") + stat_function(fun = dnorm, args = list(mean = mean(population_by_county_df
    sd = sd(population_by_county_df1$HSDegree)), col = "red",
    linewidth = 2) + theme_bw()
```



6 Answer the following questions based on the Probability Plot:

6.1 Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.

Solution: The distribution is not approximately normal, the reason being in the QQ plot, the data points clearly deviate from the straight line and don't follow along the line.

6.2 If not normal, is the distribution skewed? If so, in which direction? Explain how you know.

Solution: The distribution is definitely skewed and towards the left mostly(Negative skew). In the bottom end of the QQ Plot, the data points deviate from the straight line which clearly states that the distribution has longer tail towards the left. Also there are many outliers in bottom end of the plot. Though some points deviate in the upper end, it is more pronounced in the bottom end.

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.

```
library(pastecs)
options(scipen = 100)
options(digits = 2)
stat.desc(population_by_county_df1["HSDegree"], norm = TRUE)
```

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

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?

```
## Creating a new field for Z Score
population_by_county_df1$zScores <- (population_by_county_df1$HSDegree -
    mean(population_by_county_df1$HSDegree))/sd(population_by_county_df1$HSDegree)
```

Solution: The Skewness of the distribution is -1.67 which indicates it deviates from the normal distribution skewness of 0. Also the number indicates it is negatively skewed.The kurtosis of the distribution is 4.35 which is more than the normal distribution range of 3 and it indicates the distribution is leptokurtic as it has a big tail on the negative end of the distribution.

```
head(population_by_county_df1[(order(population_by_county_df1$zScores)),
    ][c("HSDegree", "zScores")])
```

```
##     HSDegree zScores
## 125       62    -5.0
## 87        70    -3.3
## 6         74    -2.7
## 7         74    -2.6
## 122       76    -2.3
## 8         78    -2.0
```

```
tail(population_by_county_df1[(order(population_by_county_df1$zScores)),
    ][c("HSDegree", "zScores")])
```

```
##     HSDegree zScores
## 108       94     1.2
## 109       94     1.2
## 24        94     1.3
## 63        94     1.3
## 135       95     1.4
## 52        96     1.5
```

```
## Creating a new dataframe df2 with only 40 samples
population_by_county_df2 <- head(population_by_county_df1, n = 40)
stat.desc(population_by_county_df2["HSDegree"], norm = TRUE)
```

The results from the tail command above indicates that the negative scores of HSDegree values such as 62,70,74 are outliers which have probability percentage lesser than 0.34% according to table of standard normal distribution.The results from head command indicates that the z-scores ranging from 1.5 to 1.2 are not too much skewed.

```
##                HSDegree
## nbr.val        40.00000
## nbr.null        0.00000
## nbr.na          0.00000
## min            73.60000
## max            94.10000
## range          20.50000
## sum          3449.40000
## median         87.65000
## mean           86.23500
## SE.mean         0.77585
## CI.mean.0.95    1.56930
## var            24.07772
## std.dev         4.90691
## coef.var        0.05690
## skewness       -0.99673
## skew.2SE       -1.33330
## kurtosis        0.18115
## kurt.2SE        0.12363
## normtest.W      0.88871
## normtest.p      0.00091
```

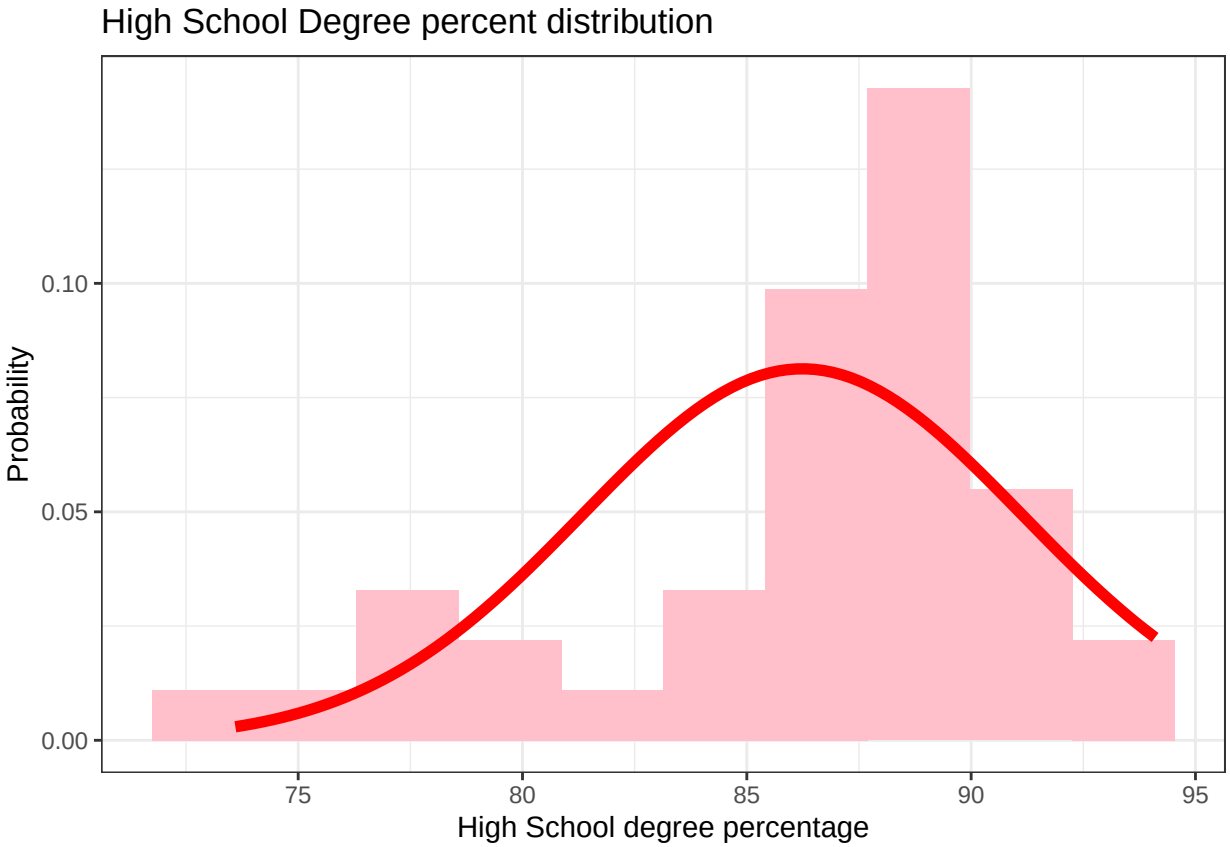
```r
# creating a new field called new z-score in the df2
# dataframe
population_by_county_df2$new_zScores <- (population_by_county_df2$HSDegree -
    mean(population_by_county_df2$HSDegree))/sd(population_by_county_df2$HSDegree)
head(population_by_county_df2[(order(population_by_county_df2$new_zScores)),
    ][c("HSDegree", "zScores", "new_zScores")])
```

```
##    HSDegree zScores new_zScores
## 6        74    -2.7        -2.6
## 7        74    -2.6        -2.4
## 8        78    -2.0        -1.8
## 15       78    -2.0        -1.8
## 19       78    -1.8        -1.6
## 12       79    -1.8        -1.6
```

```r
tail(population_by_county_df2[(order(population_by_county_df2$new_zScores)),
    ][c("HSDegree", "zScores", "new_zScores")])
```

```
##    HSDegree zScores new_zScores
## 38       90    0.48        0.79
## 29       90    0.50        0.81
## 30       92    0.78        1.09
## 21       92    0.83        1.15
## 23       93    1.01        1.34
## 24       94    1.26        1.60
```

```r
ggplot(data = population_by_county_df2, aes(x = HSDegree)) +
    geom_histogram(aes(y = after_stat(density)), bins = 10, fill = "pink") +
    labs(title = "High School Degree percent distribution", x = "High School degree percentage",
        y = "Probability") + stat_function(fun = dnorm, args = list(mean = mean(population_by_county_df2
    sd = sd(population_by_county_df2$HSDegree)), col = "red",
    linewidth = 2) + theme_bw()
```



By restricting the sample size to just 40, kurtosis, skewness and z-scores are impacted.The Kurtosis is now 0.18 which indicates model is platykurtic and the skewness is now reduced to -0.99673 which still indicates negative skew. Also the new z-score is calculated in the dataframe population_by_county_df2.These results indicate that sample size and also the values in the samples hugely impact the statistical results