assignment3.2_ChattapadhyayKausik.R

kausik

2022-09-13

```
# Assignment: ASSIGNMENT 3.2
# Name: Chattapadhyay, Kausik
# Date: 2022-09-13

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

## Set the working directory to the root of your DSC 520 directory
setwd("/Users/kausik/desktop/MS Data Science/DSC 520/dsc520-stats-r-assignments")

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

```
##                Id  Id2                       Geography PopGroupID
## 1 0500000US01073 1073       Jefferson County, Alabama          1
## 2 0500000US04013 4013        Maricopa County, Arizona          1
## 3 0500000US04019 4019            Pima County, Arizona          1
## 4 0500000US06001 6001      Alameda County, California          1
## 5 0500000US06013 6013 Contra Costa County, California          1
## 6 0500000US06019 6019       Fresno County, California          1
##   POPGROUP.display.label RacesReported HSDegree BachDegree
## 1       Total population        660793     89.1       30.5
## 2       Total population       4087191     86.8       30.2
## 3       Total population       1004516     88.0       30.8
## 4       Total population       1610921     86.9       42.8
## 5       Total population       1111339     88.8       39.7
## 6       Total population        965974     73.6       19.7
```

```
## 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 numbers) Intent: unique identifier for each row).

# Id  - Data Type: character (contains text and numbers) Intent: unique identifier for each row
# Id2 - Data Type: integer (contains whole integer) Intent: Unique integer identifier for each row
# Geography - Data Type: character (contains characters) Intent: Location name
# PopGroupID - Data Type: Integer (contains integer value) Intent: Population group id.
# POPGROUP.display.label - Data Type: Character (contain characters) Intent: population group label
# RacesReported - Data Type: integer (contains integer number) Intent: Total Population
# HSDegree - Data Type: float (contains numbers with decimals) Intent: Percentage of HS pass
# BachDegree - Data Type: float (contains numbers with decimals) Intent: Percentage of Bachelors degree

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

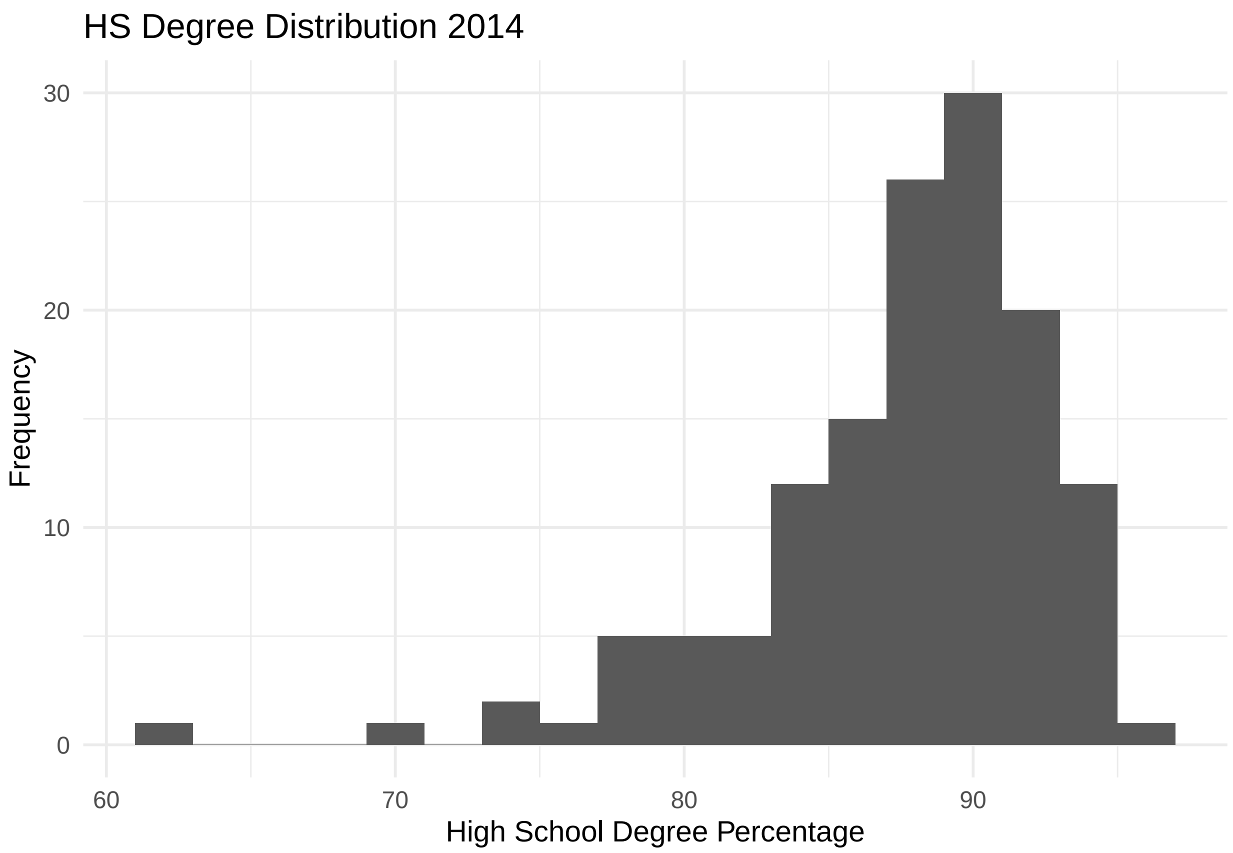
```
## [1] 136
```

```
ncol(survey_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 and appropriate X/Y axis labels on your Histogram Plot.

ggplot(data=survey_df, aes(x=HSDegree)) + geom_histogram(bins=25, binwidth = 2) +
    labs(title="HS Degree Distribution 2014", x = "High School Degree Percentage",
         y= "Frequency")
```



```
## iv. Answer the following questions based on the Histogram produced:
##      1. Based on what you see in this histogram, is the data distribution unimodal?
names(table(survey_df$HSDegree))[table(survey_df$HSDegree) == max(table(survey_df$HSDegree))]
```

```
## [1] "84.9" "85.5" "86.8" "89.1" "90.3" "92.3"
```

```
# Multimodal as there are 4 occurences of "84.9" "85.5" "86.8" "89.1" "90.3" "92.3".

##      2. Is it approximately symmetrical?
mean(survey_df$HSDegree)
```

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

```
median(survey_df$HSDegree)
```

```
## [1] 88.7
```

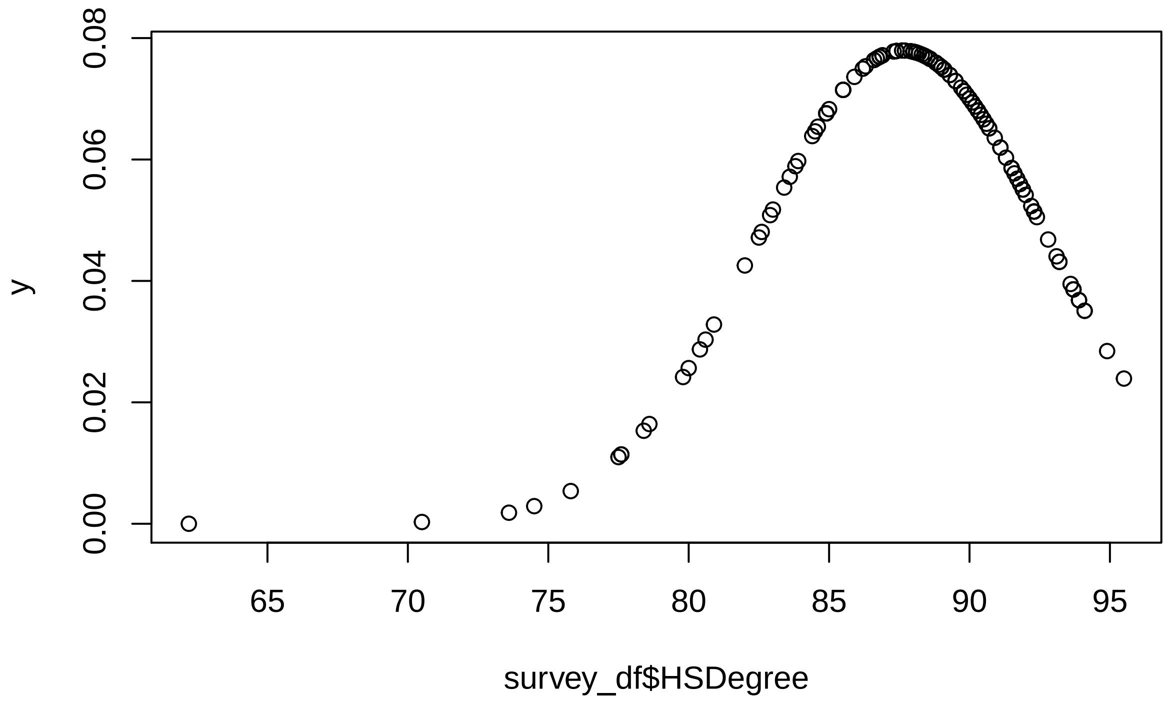
```
# Mean is 87.63 and median is 88.7 so it is not symmetrical.

##      3. Is it approximately bell-shaped?
sd(survey_df$HSDegree)
```

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

```
# Not bell-shaped.

##      4. Is it approximately normal?
y <- dnorm(survey_df$HSDegree, mean = mean(survey_df$HSDegree),
           sd = sd(survey_df$HSDegree))
plot(survey_df$HSDegree, y)
```



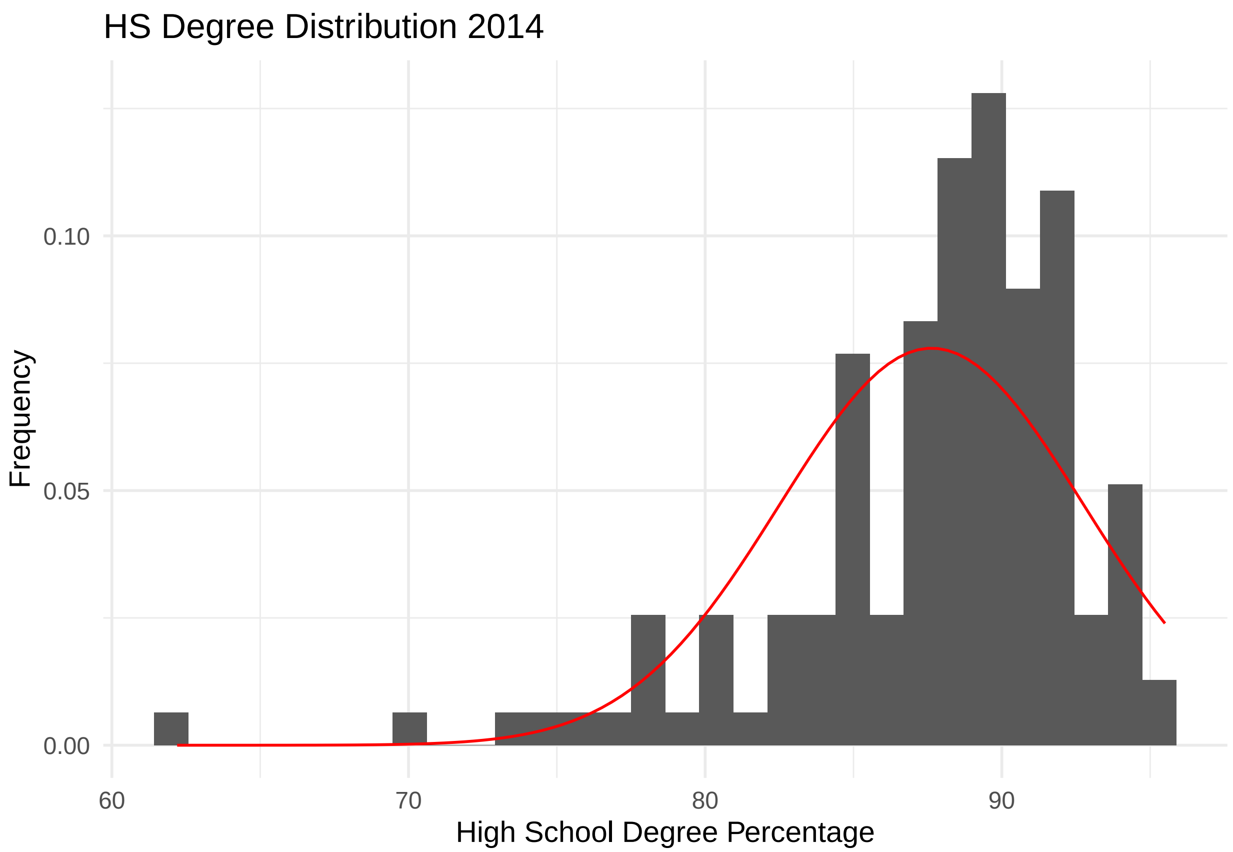
```
# Not normal distribution.

##      5. If not normal, is the distribution skewed? If so, in which direction?
# Negatively skewed distribution

##      6. Include a normal curve to the Histogram that you plotted.
ggplot(survey_df, aes(x=HSDegree)) + geom_histogram(aes(y=..density.., bins=25)) +
    labs(title="HS Degree Distribution 2014", x = "High School Degree Percentage",
         y= "Frequency") + stat_function(fun=dnorm, color="red",
        args=list(mean = mean(survey_df$HSDegree), sd = sd(survey_df$HSDegree)))
```

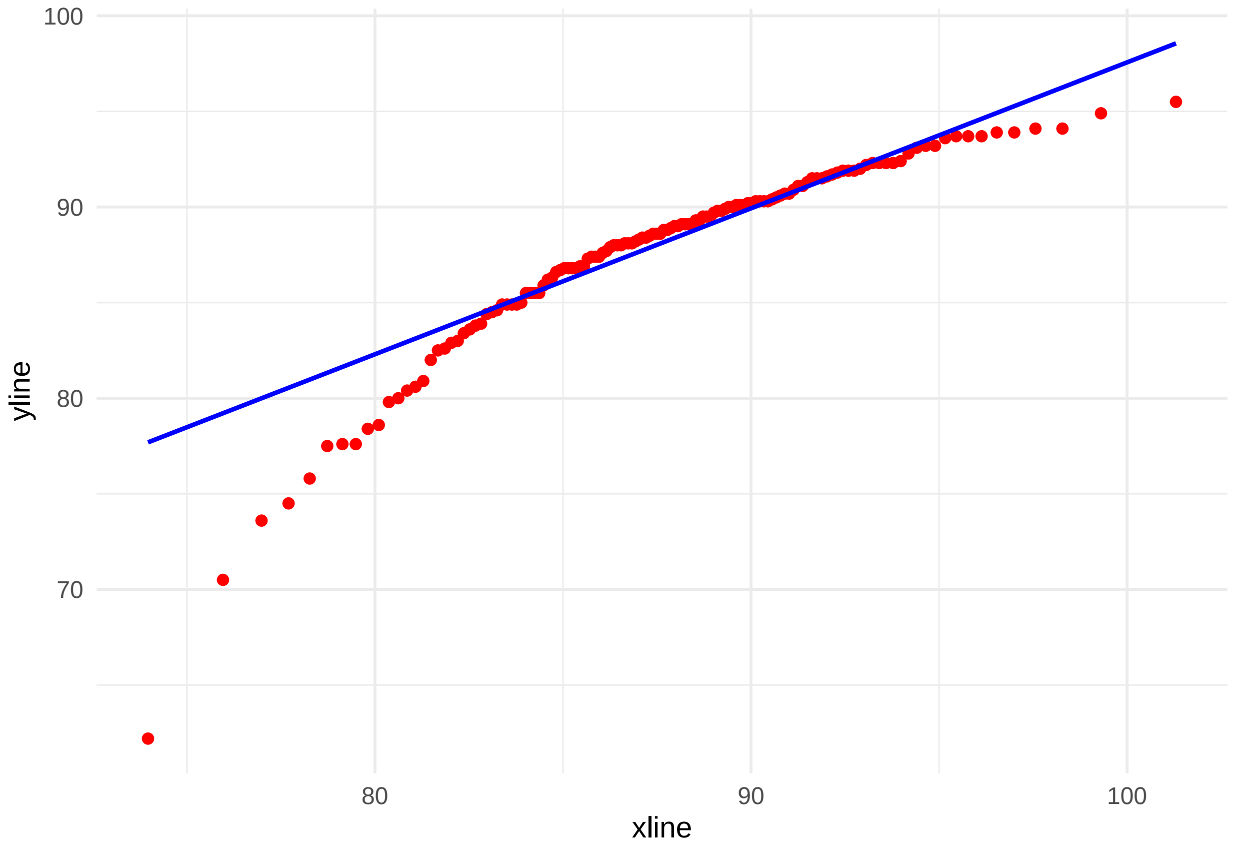
```
## Warning: Ignoring unknown aesthetics: bins
```

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



```
## Explain whether a normal distribution can accurately be used as a model for this data.
# Since it is a negatively skewed distribution, normal distribution can not be used.

## v. Create a Probability Plot of the HSDegree variable.
ggplot(survey_df, aes(sample=HSDegree)) + stat_qq_point(color="red") +
    stat_qq_line(color="blue")
```



```
## 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.
#          It is not a normal distribution as it is not a straight line and curved.
##      2. If not normal, is the distribution skewed? If so, in which direction?
##          Explain how you know.
#           This is a negatively skewed distribution as the plot bends down and
#           to the right of normal line.

## 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(survey_df$HSDegree)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
## 1.360000e+02 0.000000e+00 0.000000e+00 6.220000e+01 9.550000e+01 3.330000e+01
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 1.191800e+04 8.870000e+01 8.763235e+01 4.388598e-01 8.679296e-01 2.619332e+01
##      std.dev     coef.var
## 5.117941e+00 5.840241e-02
```

```
## viii. 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?
```

```
#           skew - Negative
#           kurtosis - Platykurtic
#           z-scores - Positive Value
#           When the sample is changed to add new values in the lower side of the
#           curve then we can get a normal distribution.
```