# Assignment: ASSIGNMENT 5 Student Survey

# Name: Karna, RajasekharReddy

# Date: 2020-10-02

***R-Programming:***

library(ggplot2)

library(pastecs)

library(ggfortify)

library(moments)

theme\_set(theme\_minimal())

setwd("C:/Users/vahin/Documents/GitHub/dsc520/")

Comm\_Survey <- read.csv("data/acs-14-1yr-s0201.csv")

str(Comm\_Survey)

summary(Comm\_Survey)

head(Comm\_Survey, 2)

**1. What are the elements in your data (including the categories and data types)?**

Data contains following elements:

* ID: a combination of numbers and letters to find area. Not primary or unique parameter.
* ID2: A numerical value to find area. Again not primary or unique parameter.
* Geography: Location parameter and can be considered as unique parameter.
* PopGroupID: A numeric field with static data value as = 1.
* POPGROUP.display.label: A letters combination with static data value as Total Population
* RacesReported: Numeric field holding number of races reported.
* HSDegree: Numeric percentage field describes the percentage of HS degree holders per County
* BachDegree: Numeric percentage field describes the percentage of Bachelor degree holders per County

***R-Programming:***

str(Comm\_Survey)

paste("Number of rows =", nrow(Comm\_Survey), "; Number of columns = ", ncol(Comm\_Survey), ";")

**2. Please provide the output from the following functions: str(); nrow(); ncol()**

**str():**

'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" ...

$ 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 ...

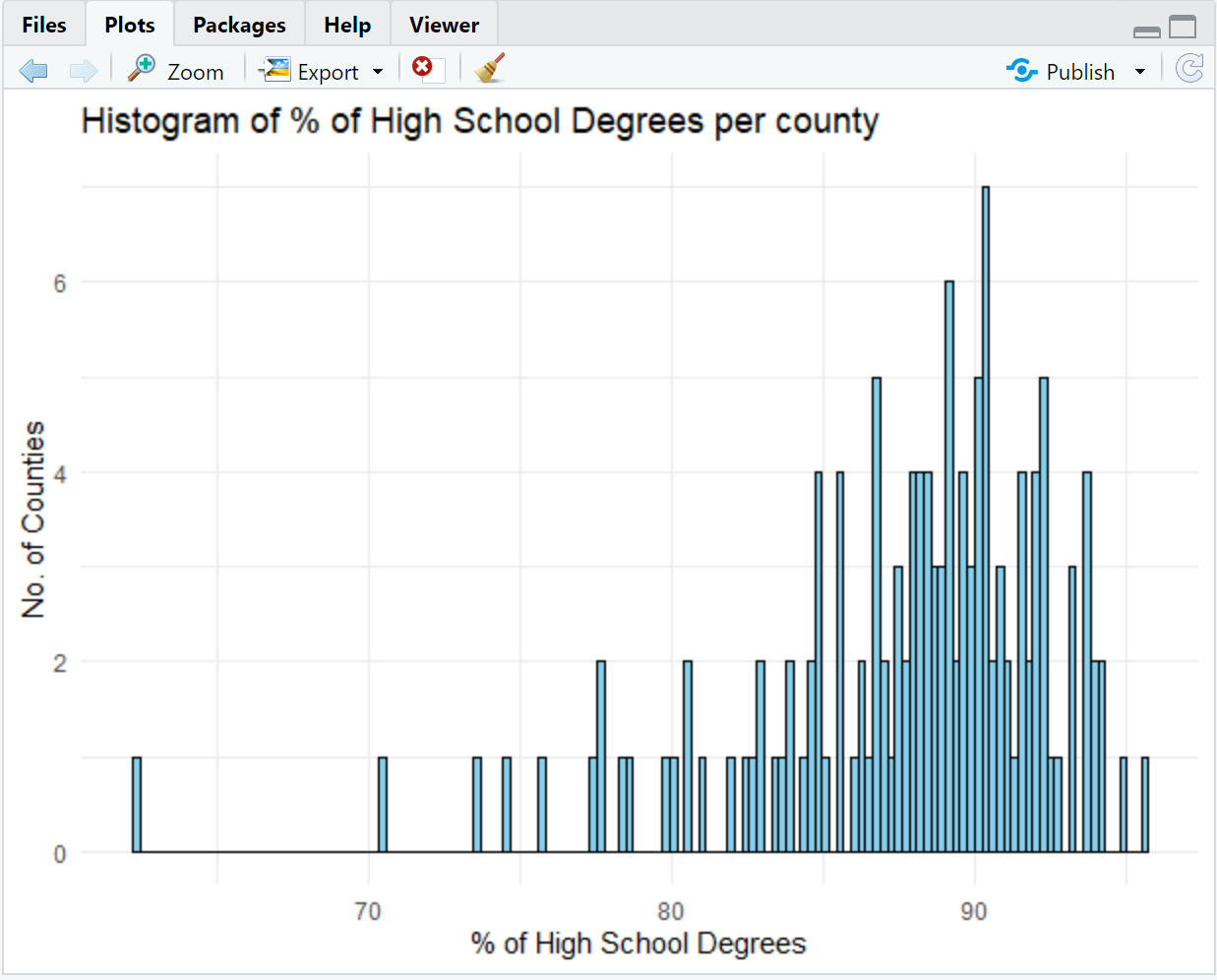
**nrow(); ncol():**

[1] "Number of rows = 136 ; Number of columns = 8 ;"

**3. Create a Histogram of the HSDegree variable using the ggplot2 package.**

**a. Set a bin size for the Histogram.**

**b. Include a Title and appropriate X/Y axis labels on your Histogram Plot.**

****

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

**a. Based on what you see in this histogram, is the data distribution unimodal?**

Definition for Unimodel Distribution:

A histogram is unimodal if there is one hump, bimodal if there are two humps and multimodal if there are many humps.

Observation:

Since there is a single spike / bump in the above histogram plot, Yes the data distribution model is unimodal.

**b. Is it approximately symmetrical?**

Definition of Symmetric Distribution:

A symmetric distribution is one in which the 2 "halves" of the histogram appear as mirror-images of one another.

Observation:

Since there are couple of spike/bumps are mirror image of another county spike/bump, Yes the data distribution model has symmetric distribution. Example: San Joaquin County, California & Dallas County, Texas has same distribution.

Also at the same time, all data distribution models are not symmetric distribution. So at few combinations data is symmetric distribution and few cases data is non-symmetric distribution. Overall perspective, data is non-symmetric.

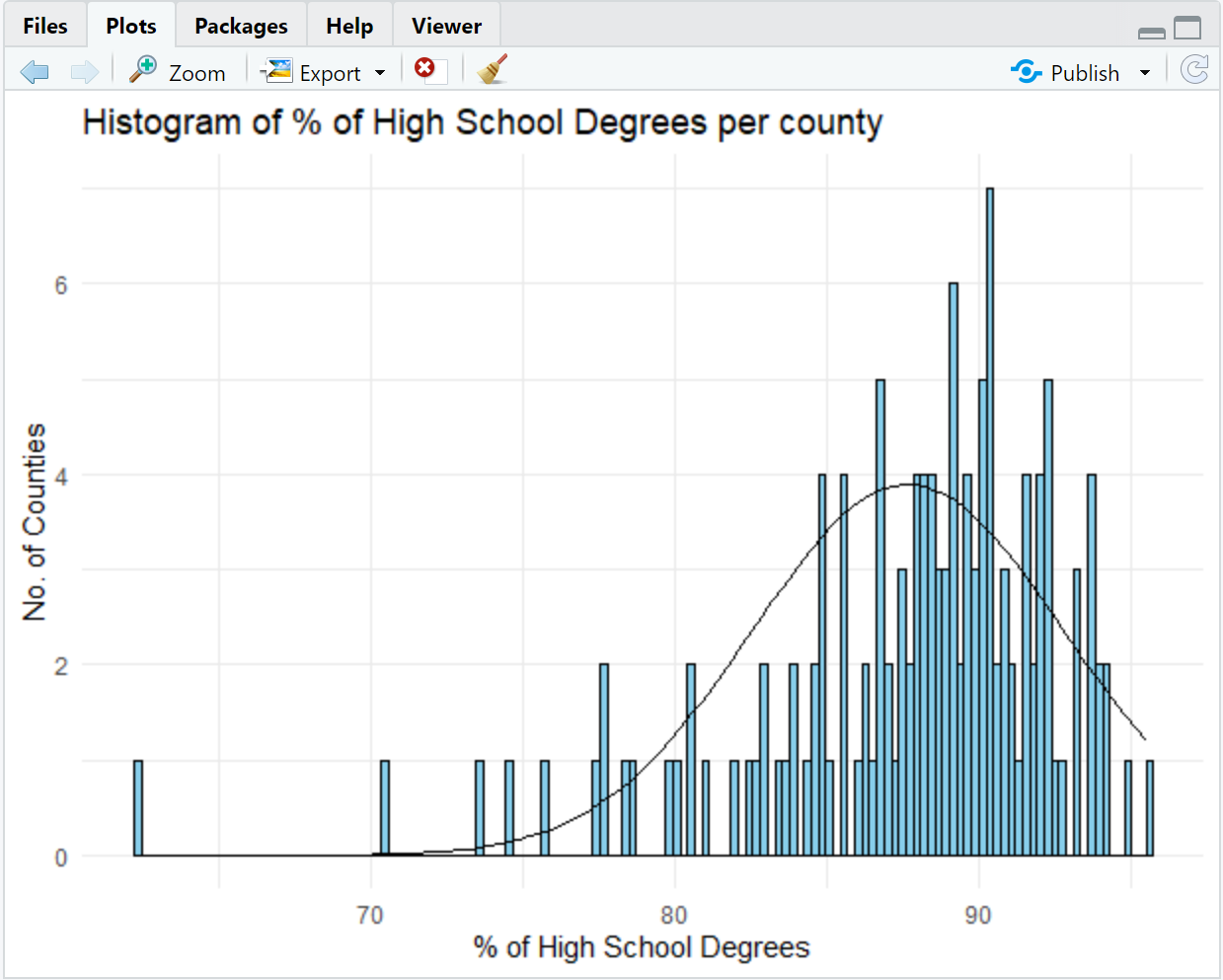
***R-Programming:***

print(mean(Comm\_Survey$HSDegree))

print(median(Comm\_Survey$HSDegree))

print(sd(Comm\_Survey$HSDegree))

ggplot(Comm\_Survey, aes(x=HSDegree)) + geom\_histogram(bins = 140, color = "black", fill = "skyblue") + xlab("% of High School Degrees") + ylab("No. of Counties") + ggtitle("Histogram of % of High School Degrees per county") + stat\_function(fun = function(x) dnorm(x, mean(Comm\_Survey$HSDegree), sd(Comm\_Survey$HSDegree))\* 50)



**c. Is it approximately bell-shaped?**

Definition of Bell Shaped Distribution:

A bell-shaped picture usually presents a normal distribution.

Observation:

Since the overall data distribution is non-symmetric distribution and not evenly normal distribution happened, so we could consider the provided data is not 100% bell shaped distributed. But at overall, it has edge of normalizing the data. Possibility to become bell shaped with more sample data, would give opportunity to confirm is provided data is bell shaped or not. But certainly looking at normalization curve, we can say data is Bell-Shaped distribution.

**d. Is it approximately normal?**

Definition for Normal Distribution:

A common pattern is the bell-shaped curve known as the "normal distribution."

Observation:

Though the provided sample data is not 100% bell-shared curve, based on overall lean curve stand point, data is leading more of normal distribution to confirm if we get more sample data. If provided data is complete data, we can say, it is not normally distributed.

**e. If not normal, is the distribution skewed? If so, in which direction?**

Definition for Skewed:

A "skewed left" distribution is one in which the tail is on the left side. A "skewed right" distribution is one in which the tail is on the right side.

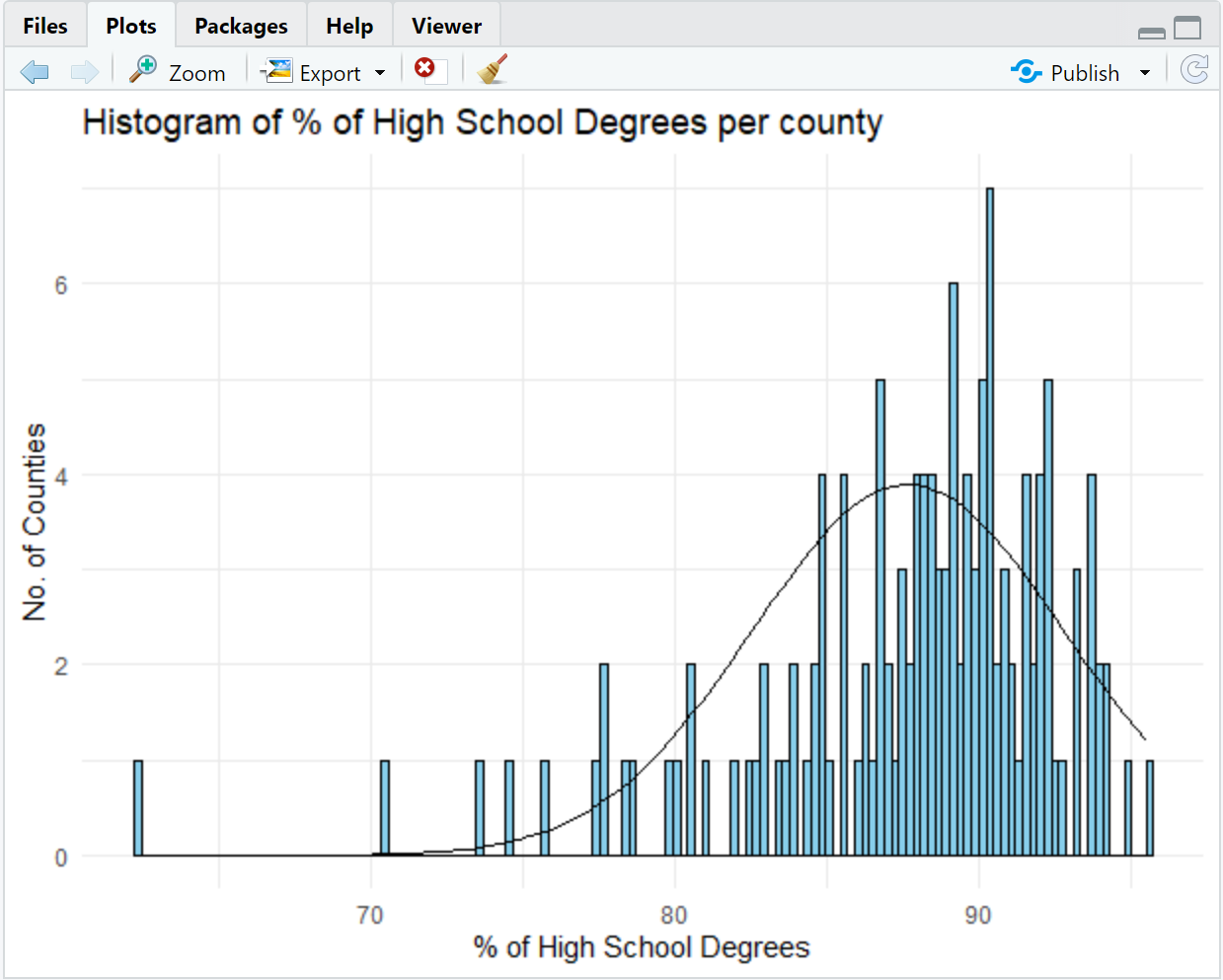
Observation:

Since the data is not spiked / bump either at left most or at right most and the overall data is leading towards normal distributed, we could say the provided data distribution is not 100% skewed. But with the sample data provided, we could see data is leading towards skewed right.

***R-Programming:***

ggplot(Comm\_Survey, aes(x=HSDegree)) + geom\_histogram(bins = 140, color = "black", fill = "skyblue") + xlab("% of High School Degrees") + ylab("No. of Counties") + ggtitle("Histogram of % of High School Degrees per county") + stat\_function(fun = function(x) dnorm(x, mean(Comm\_Survey$HSDegree), sd(Comm\_Survey$HSDegree))\* 50)

**f. Include a normal curve to the Histogram that you plotted.**



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

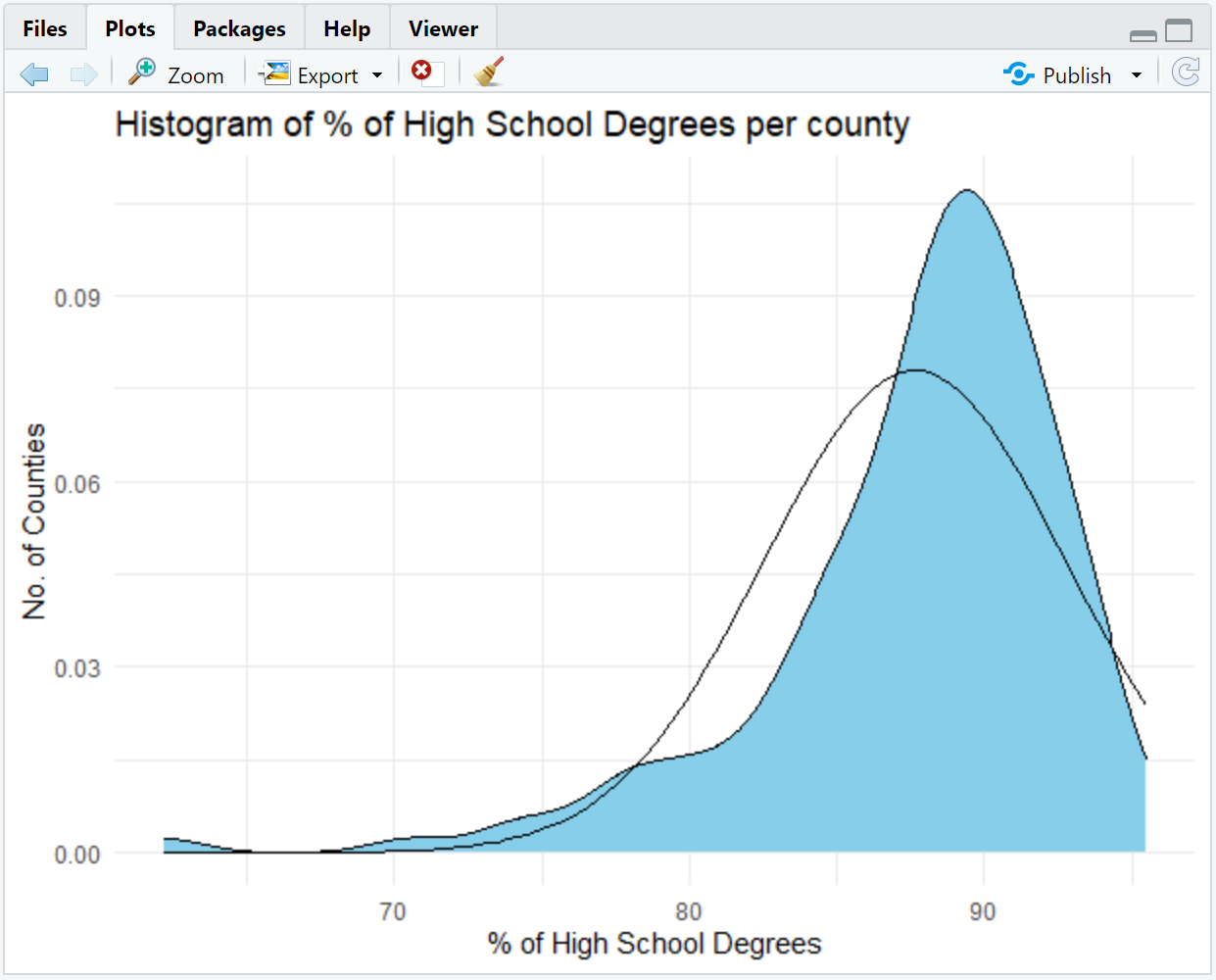
Since the provided data is not 100% normalized, though it is leading towards high percentage of normal distribution, we could say normal distribution could be an option to start with data modeling.

***R-Programming:***

ggplot(Comm\_Survey, aes(x=HSDegree)) + geom\_density(bins = 140, color = "black", fill = "skyblue") + xlab("% of High School Degrees") + ylab("No. of Counties") + ggtitle("Histogram of % of High School Degrees per county")

ggplot(Comm\_Survey, aes(x=HSDegree)) + geom\_density(color = "black", fill = "skyblue") + xlab("% of High School Degrees") + ylab("No. of Counties") + ggtitle("Histogram of % of High School Degrees per county") + stat\_function(fun = function(z) dnorm(z, mean(Comm\_Survey$HSDegree), sd(Comm\_Survey$HSDegree)))

**5. Create a Probability Plot of the HSDegree variable.**

****

**6. Answer the following questions based on the Probability Plot:**

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

Though the provided sample data is not 100% normal distribution curve, data is leading more of normal distribution only to confirm if we get more sample data. If provided data is complete data, we can say, it is not normally distributed as data is skewed most at the right and it has more tail.

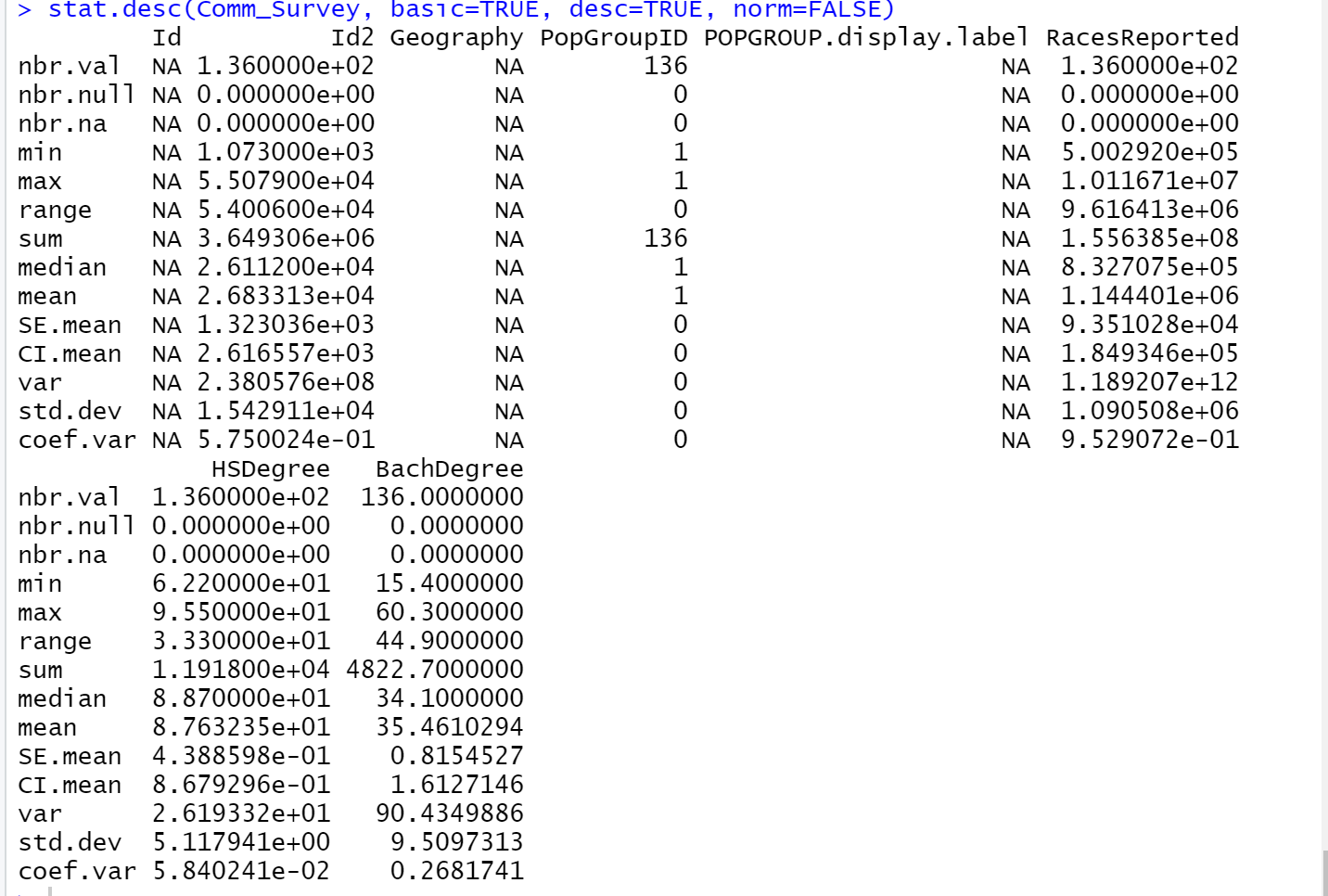
**b. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.**

Since the data is not spiked / bump either at left most or at right most and the overall data is leading towards normal distributed, we could say the provided data distribution is not 100% skewed too. But with the sample data provided, we could see data is leading towards skewed right.

***R-Programming:***

stat.desc(Comm\_Survey, basic=TRUE, desc=TRUE, norm=FALSE)

**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.**

****

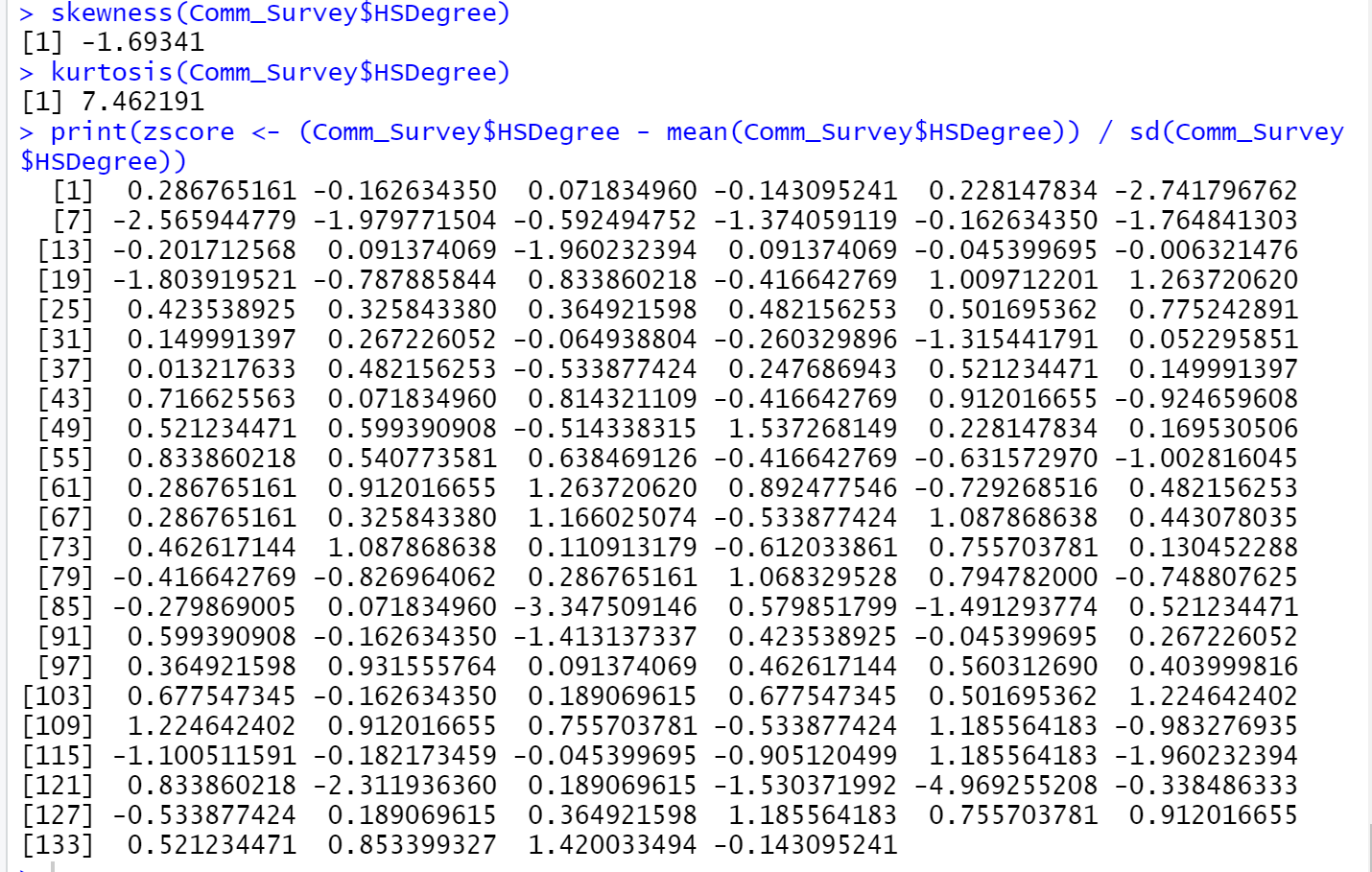
***R-Programming:***

skewness(Comm\_Survey$HSDegree)

kurtosis(Comm\_Survey$HSDegree)

print(zscore <- (Comm\_Survey$HSDegree - mean(Comm\_Survey$HSDegree)) / sd(Comm\_Survey$HSDegree))

**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?**



Definition for Skew, Kurtosis, z-score:

Skewness tells user the amount and direction of skew (departure from horizontal symmetry), and kurtosis tells you how tall and sharp the central peak is, relative to a standard bell curve. When a frequency distribution is normally distributed, we can find out the probability of a score occurring by standardizing the scores, known as standard scores (or z scores).

* A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈3 (excess ≈0) is called mesokurtic.
* A distribution with kurtosis <3 (excess kurtosis <0) is called platykurtic. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.
* A distribution with kurtosis >3 (excess kurtosis >0) is called leptokurtic. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.

Reference: <https://statistics.laerd.com/statistical-guides/standard-score-4.php>

<https://brownmath.com/stat/shape.htm>

Observation:

Since the provided data is leading towards more of normal distribution (again not 100% normal distribution), we could say the data result provided neither skewness left or skewness right. But at high level data is leading towards skewness right, compared with skewness left. Again the provided sample data not having accurate normal distribution, the provided can be viewed with kurtosis distribution >3 and we can call leptokurtic, since its central peak is higher and sharper. Again the provided data is not 100%.