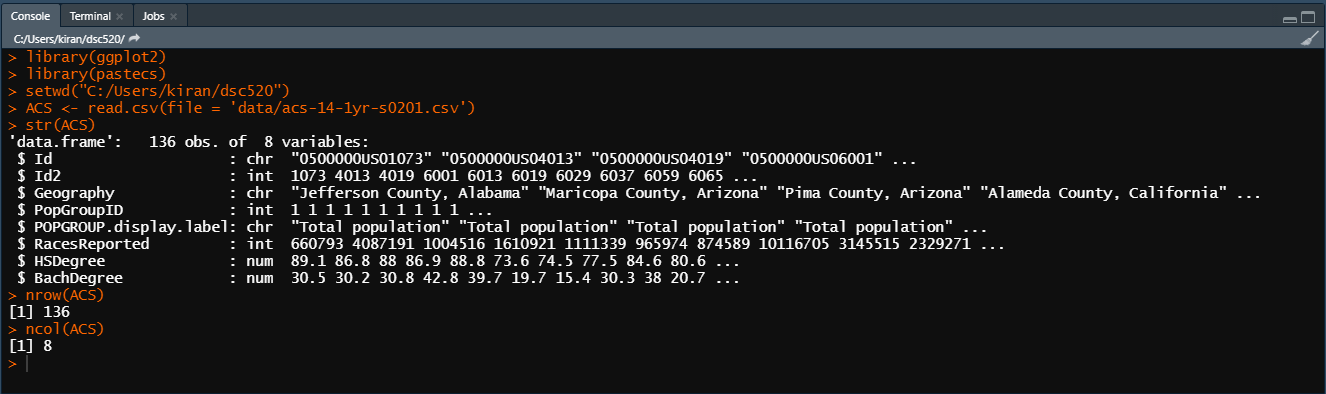
**Exercise 2: American Community Survey Exercise**

**i)**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Category** | **Data Type** |
| Id | nominal | string |
| Id2 | nominal | integer |
| Geography | categorical | string |
| PopGroupID | nominal | integer |
| POPGROUP.display-label | categorical | string |
| RacesReported | discrete and ratio | integer |
| HSDegree | continuous and ratio | numeric/double |
| BachDegree | continuous and ratio | numeric/double |

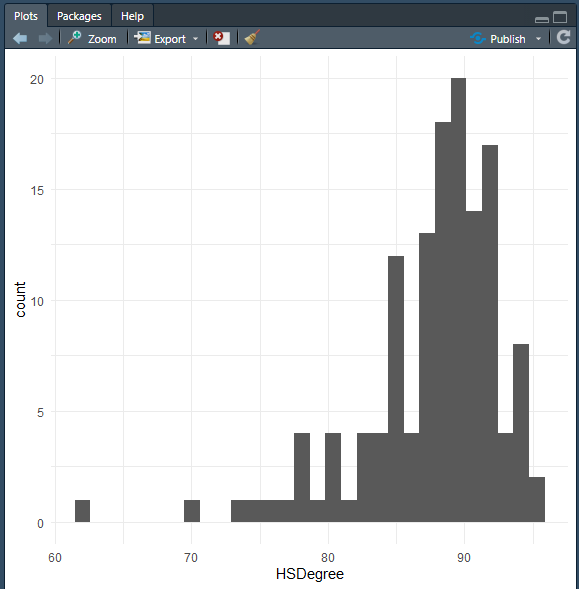
**ii)**

****

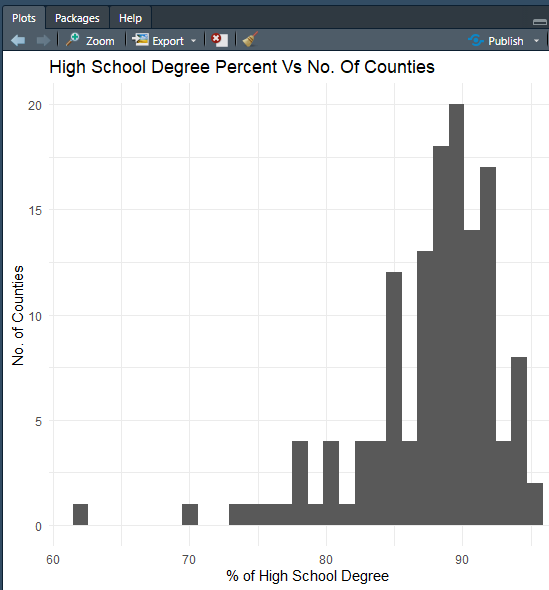
**iii)**

1)

ggplot(ACS,aes(HSDegree)) + geom\_histogram(bins=30)



2) ggplot(ACS,aes(HSDegree)) + geom\_histogram(bins=30) + ggtitle('High School Degree Percent Vs No. Of Counties') + xlab('% of High School Degree')+ylab('No. of Counties')



iv)

1.Based on what you see in this histogram, is the data distribution unimodal? - yes. It has only one bump

2.Is it approximately symmetrical? - No. it is not symmetrical

3.Is it approximately bell-shaped? - Yes. It is bell shaped.

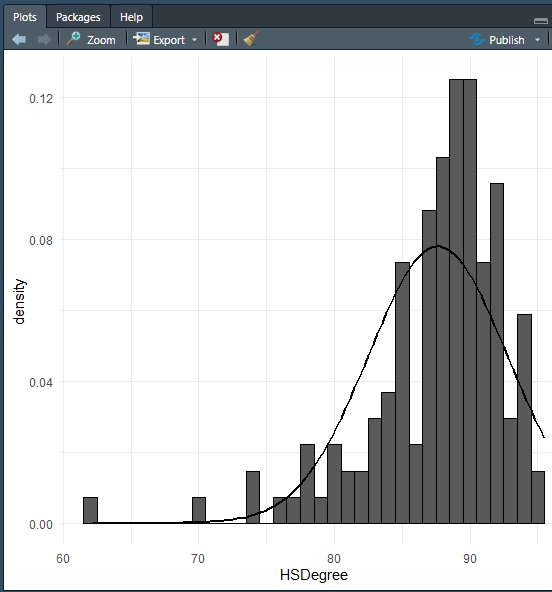
4.Is it approximately normal? - No. It is not normal.

5.If not normal, is the distribution skewed? If so, in which direction? - It is left skewed which means that the values are heavily distributed on the right side.

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

ACS\_HIST <- ggplot(ACS,aes(HSDegree)) + geom\_histogram(aes(y = ..density..),binwidth = 1, colour = "black")

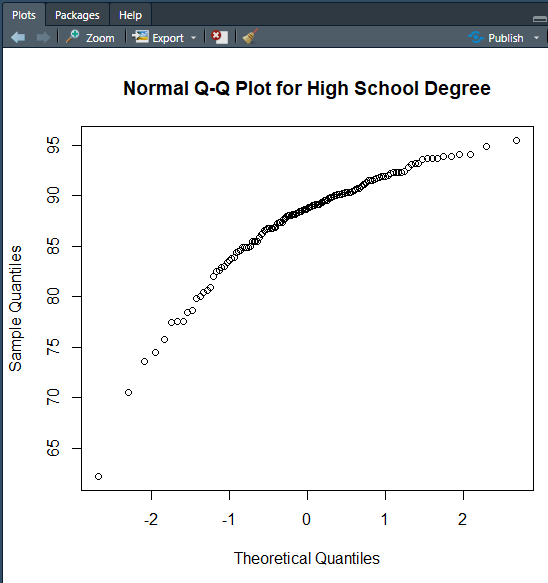
ACS\_HIST + stat\_function(fun = dnorm,args=list(mean=mean(ACS$HSDegree),sd=sd(ACS$HSDegree)),color="black",size=1)



7. Explain whether a normal distribution can accurately be used as a model for this data. – The output clearly shows that it is left skewed and hence normal distribution cannot be used as a model.

v)

qqnorm(ACS$HSDegree, main = "Normal Q-Q Plot for High School Degree")



vi)

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

The distribution is not normal as the output of probability plot is not a straight line.

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

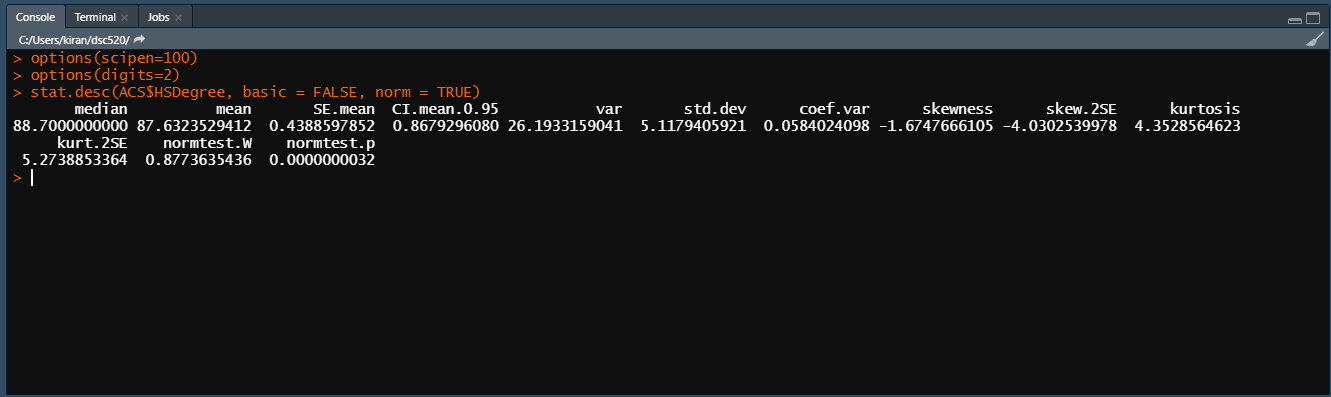
The distribution is left skewed as the bottom-end of the output deviates from the straight line but not the upper end.

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.

options(scipen=100)

options(digits=2)

stat.desc(ACS$HSDegree, basic = FALSE, norm = TRUE)



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?

In our output, the skewness is negative which means that it is skewed to the right. Also, kurtosis is positive which means that it is pointy and heavy-tailed. Also, the absolute values of skew.2se and kurt.2se which are the skew and kurtosis value divided by 2 standard errors are greater than 1 suggests that the distribution is not normal and p<0.001

They are very dependent on the sample size. As the sample size increases, they decrease. Smaller sample size can give misleading results.