American Community Survey - Exercise

Akila Selvaraj

Masters of Data Science,

Bellevue University

DSC 520 T301

December 18,2021

i) What are the elements in your data (including the categories and data types)?

```
install.packages('rlang')
install.packages('tidyr')
install.packages("ggplot2")
install.packages('psych')
install.packages('pastecs')
install.packages('probplot')

library("tidyr")
library("ggplot2")
library("pastecs")
library("probplot")

getwd()
setwd("G:/Users/a162940/Akila/Work/R/Projects/dsc520-master")
slist.files (path = "G:/Users/a162940/Akila/Work/R/Projects/dsc520-master/data")
survey_df <- read.csv("data/acs-14-1yr-s0201.csv")
survey_df
```

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

categories and data types

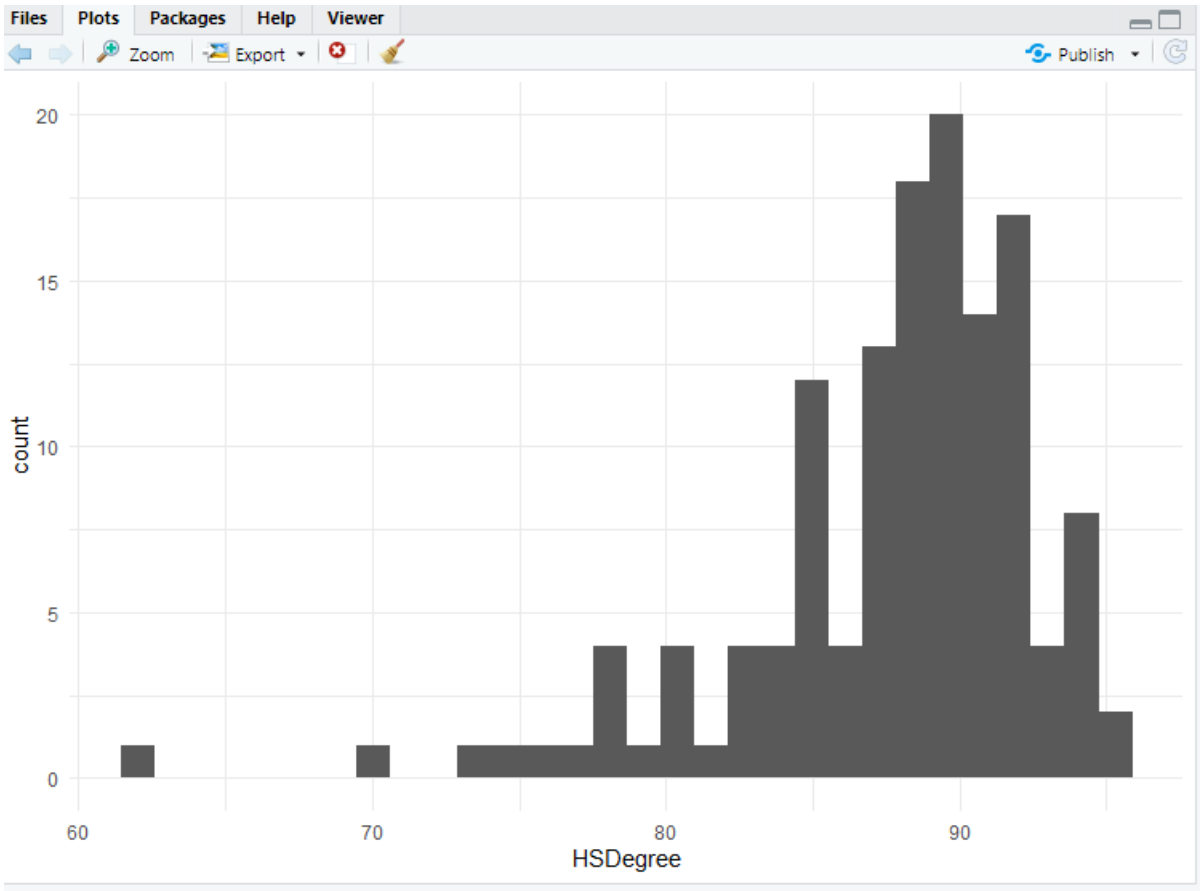
```
> str(survey_df)
'data.frame':           136 obs. of  8 variables:
 $ Id                : chr
 $ Id2               : int
 $ Geography         : chr
 $ PopGroupID        : int
 $ POPGROUP.display.label: chr
 $ RacesReported     : int
 $ HSDegree          : num
 $ BachDegree        : num

> nrow(survey_df)
[1] 136
> ncol(survey_df)
```

[1] 8

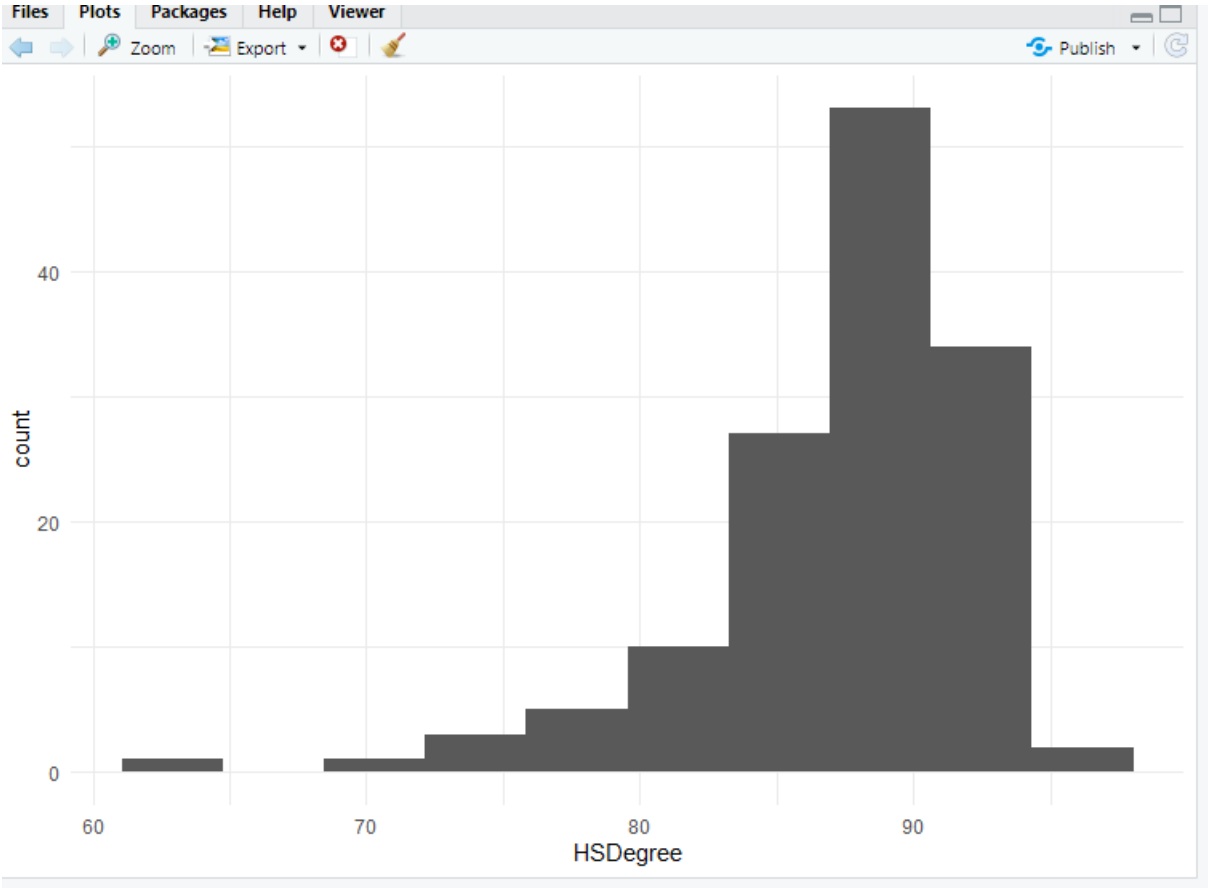
iii. Create a Histogram of the HSDegree variable using the ggplot2 package.

ggplot(survey_df, aes(HSDegree)) + geom_histogram()



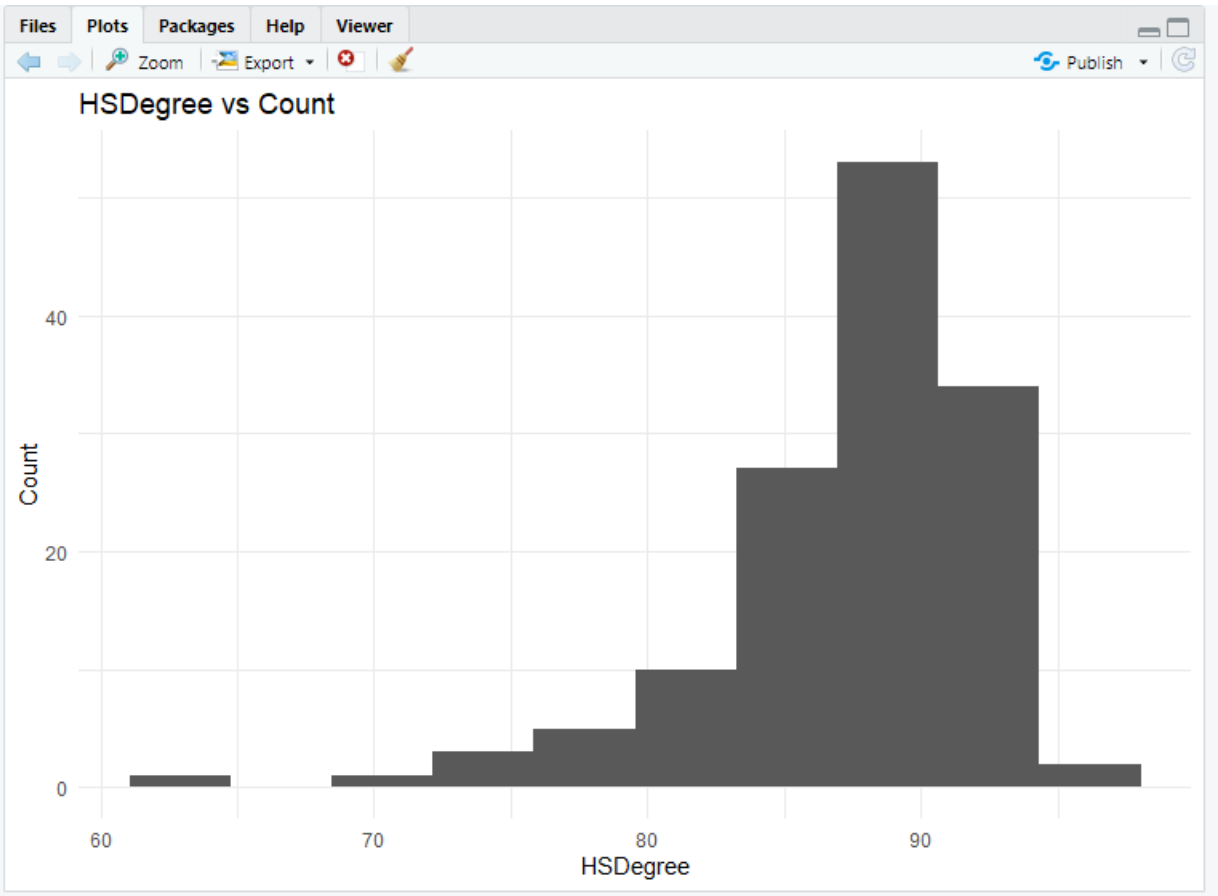
1. Set a bin size for the Histogram.

ggplot(survey_df, aes(HSDegree)) + geom_histogram(bins = 10)



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

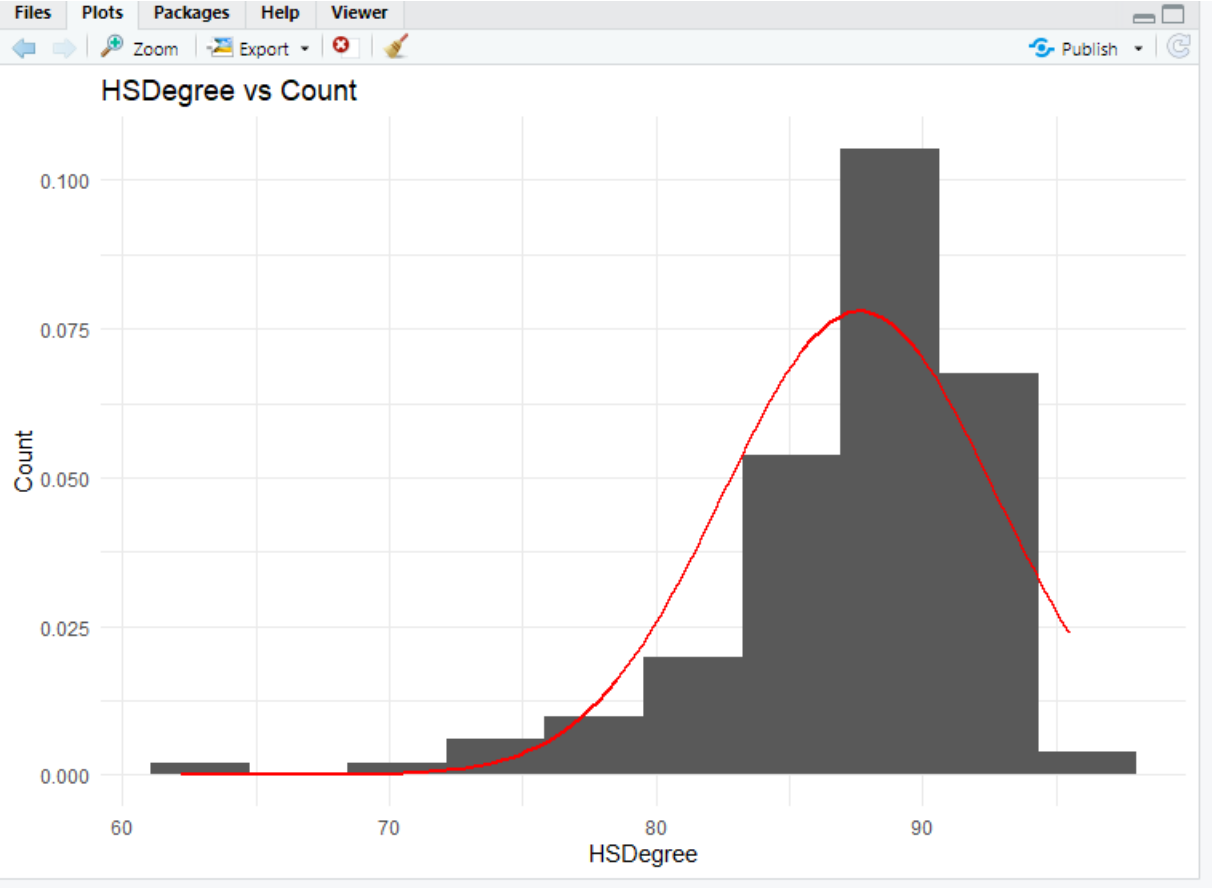
ggplot(survey_df, aes(HSDegree)) + geom_histogram(bins = 10) + ggtitle("HSDegree vs Count") + xlab("HSDegree") + ylab("Count")



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

1. Based on what you see in this histogram, is the data distribution unimodal?
   Yes, Data distribution is unimodal as there is one clear peak in the histogram.
2. Is it approximately symmetrical?
   No, it is not symmetrical
3. Is it approximately bell-shaped?
   Yes, it is approximately 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?
   Yes, the distribution is skewed. It is left skewed.
6. Include a normal curve to the Histogram that you plotted.

```
ggplot(survey_df, aes(HSDegree)) +  geom_histogram(aes(y= ..density..),bins = 10) + stat_function(fun= dnorm,args=list(mean= mean(survey_Degree,na.rm = TRUE),sd=sd(survey_Degree,na.rm = TRUE)),color="red",size=1) + ggtitle("HSDegree vs Count") + xlab("HSDegree")  + ylab("Count")
```



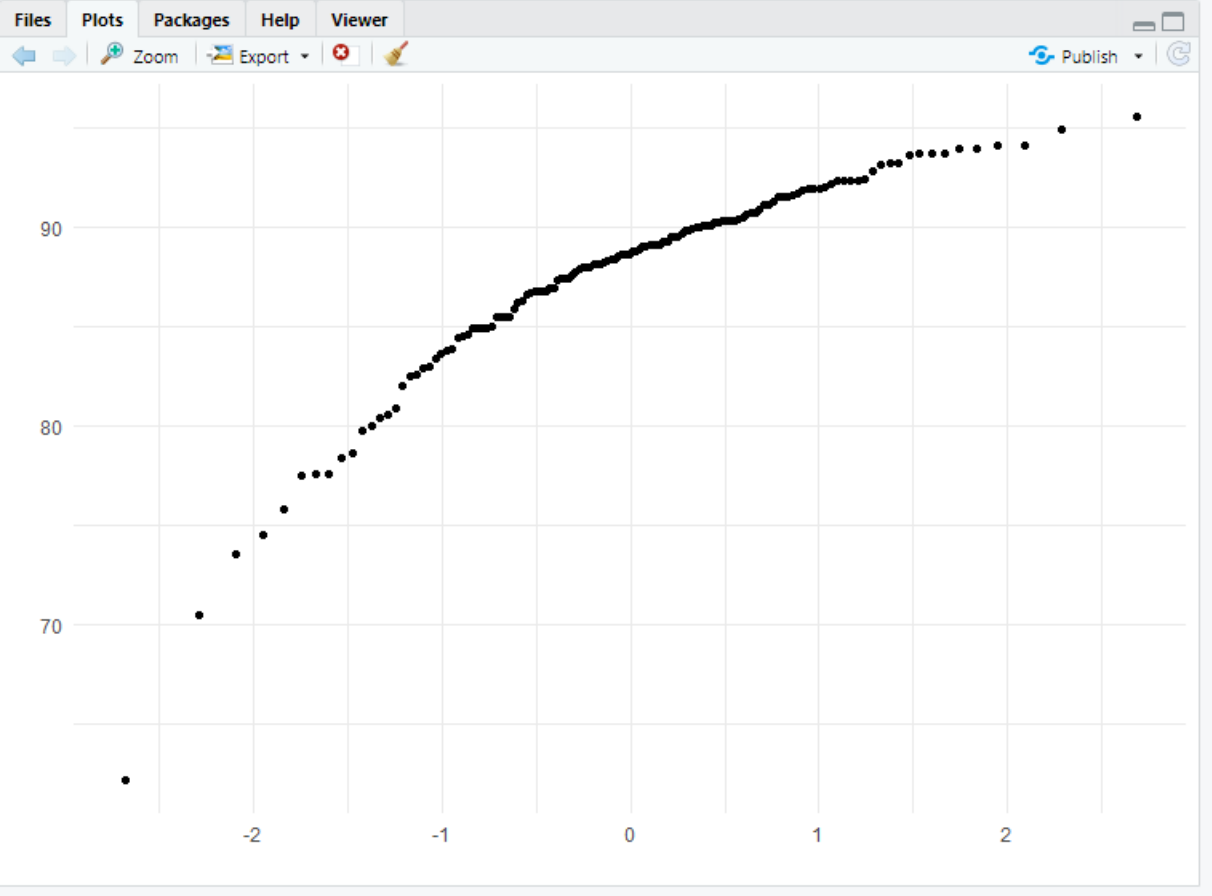
7. Explain whether a normal distribution can accurately be used as a model for this data.

If the distribution is normal without being skewed, then the top of the curve will align perfectly with the shape of the histogram. As normal curve doesn't align perfectly with the histogram, normal distribution can't be accurately used as a model for this data.

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

```
survey_Degree <- survey_df$HSDegree

qplot(sample = survey_Degree, stat="qq")
```



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.

   The distribution is not approximately normal. If data are normally distributed, then the actual HSDegree count will have the same distribution as normal distribution and we would have got a straight line. As we didn't get a straight line out of probability plot, the distribution is not normal.

2. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.
   The bottom end of the plot deviates from the straight line but the upper end is not. By this, we can clearly say that the distribution has a longer tail to its left or simply 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.

stat.desc(survey_Degree, basic = FALSE, desc=TRUE,norm = TRUE)

```
> stat.desc(survey_Degree, basic = FALSE, desc=TRUE,norm = TRUE)
      median         mean      SE.mean  CI.mean.0.95          var       std.dev      coef.var      skewness      skew.2SE
 8.870000e+01  8.763235e+01  4.388598e-01  8.679296e-01  2.619332e+01  5.117941e+00  5.840241e-02 -1.674767e+00 -4.030254e+00
     kurtosis      kurt.2SE    normtest.W    normtest.p
 4.352856e+00  5.273885e+00  8.773635e-01  3.193634e-09
>
```

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?

The value of skew and kurtosis will be zero if the distribution is normal. Non-zero values of skew and kurtosis implies that the data is skewed. Skew value we got as -1.674767 is a negative value which indicates that the data points are piled up more on the right than on the left. When it comes to kurtosis, we obtained a positive value 4.352856 which indicates that the distribution is pointy and heavy-tailed.

```
> zscore <- (survey_Degree - mean(survey_Degree, na.rm = TRUE)) / sd(survey_Degree, na.rm = TRUE)
> zscore
  [1]  0.286765161 -0.162634350  0.071834960 -0.143095241  0.228147834 -2.741796762 -2.565944779 -1.979771504 -0.592494752 -1.374059119
 [11] -0.162634350 -1.764841303 -0.201712568  0.091374069 -1.960232394  0.091374069 -0.045399695 -0.006321476 -1.803919521 -0.787885844
 [21]  0.833860218 -0.416642769  1.009712201  1.263720620  0.423538925  0.325843380  0.364921598  0.482156253  0.501695362  0.775242891
 [31]  0.149991397  0.267226052 -0.064938804 -0.260329896 -1.315441791  0.052295851  0.013217633  0.482156253 -0.533877424  0.247686943
 [41]  0.521234471  0.149991397  0.716625563  0.071834960  0.814321109 -0.416642769  0.912016655 -0.924659608  0.521234471  0.599390908
 [51] -0.514338315  1.537268149  0.228147834  0.169530506  0.833860218  0.540773581  0.638469126 -0.416642769 -0.631572970 -1.002816045
 [61]  0.286765161  0.912016655  1.263720620  0.892477546 -0.729268516  0.482156253  0.286765161  0.325843380  1.166025074 -0.533877424
 [71]  1.087868638  0.443078035  0.462617144  1.087868638  0.110913179 -0.612033861  0.755703781  0.130452288 -0.416642769 -0.826964062
 [81]  0.286765161  1.068329528  0.794782000 -0.748807625 -0.279869005  0.071834960 -3.347509146  0.579851799 -1.491293774  0.521234471
 [91]  0.599390908 -0.162634350 -1.413137337  0.423538925 -0.045399695  0.267226052  0.364921598  0.931555764  0.091374069  0.462617144
[101]  0.560312690  0.403999816  0.677547345 -0.162634350  0.189069615  0.677547345  0.501695362  1.224642402  1.224642402  0.912016655
[111]  0.755703781 -0.533877424  1.185564183 -0.983276935 -1.100511591 -0.182173459 -0.045399695 -0.905120499  1.185564183 -1.960232394
[121]  0.833860218 -2.311936360  0.189069615 -1.530371992 -4.969255208 -0.338486333 -0.533877424  0.189069615  0.364921598  1.185564183
[131]  0.755703781  0.912016655  0.521234471  0.853399327  1.420033494 -0.143095241
>
```

This z-score tells us how many standard deviations away a HSDegree is from the mean value of 87.63

Change in sample size will impact the variability of distribution. If sample size increases, the shape of distribution will become more becomes more similar to the normal distribution. With the current sample size, the standard error of mean is 0.438. When sample size increases, standard error will also decrease.

With the current data, the distribution is skewed to left. With the larger sample size, distribution will be approximately normal.

Sample size change will also affect the shape of distribution. with smaller sample size, large gaps between each proportion are possible whereas the gap will decrease with large sample.