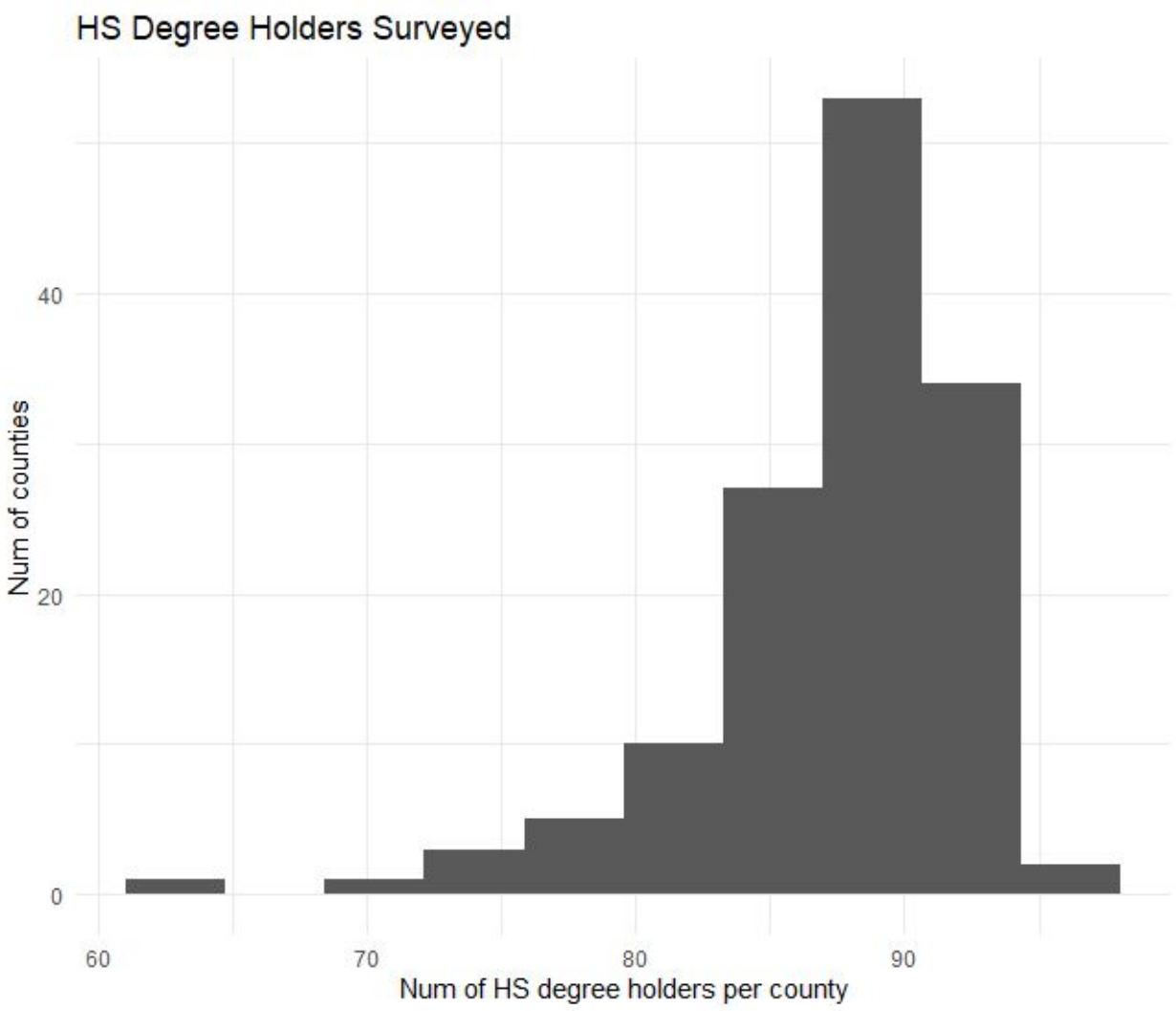
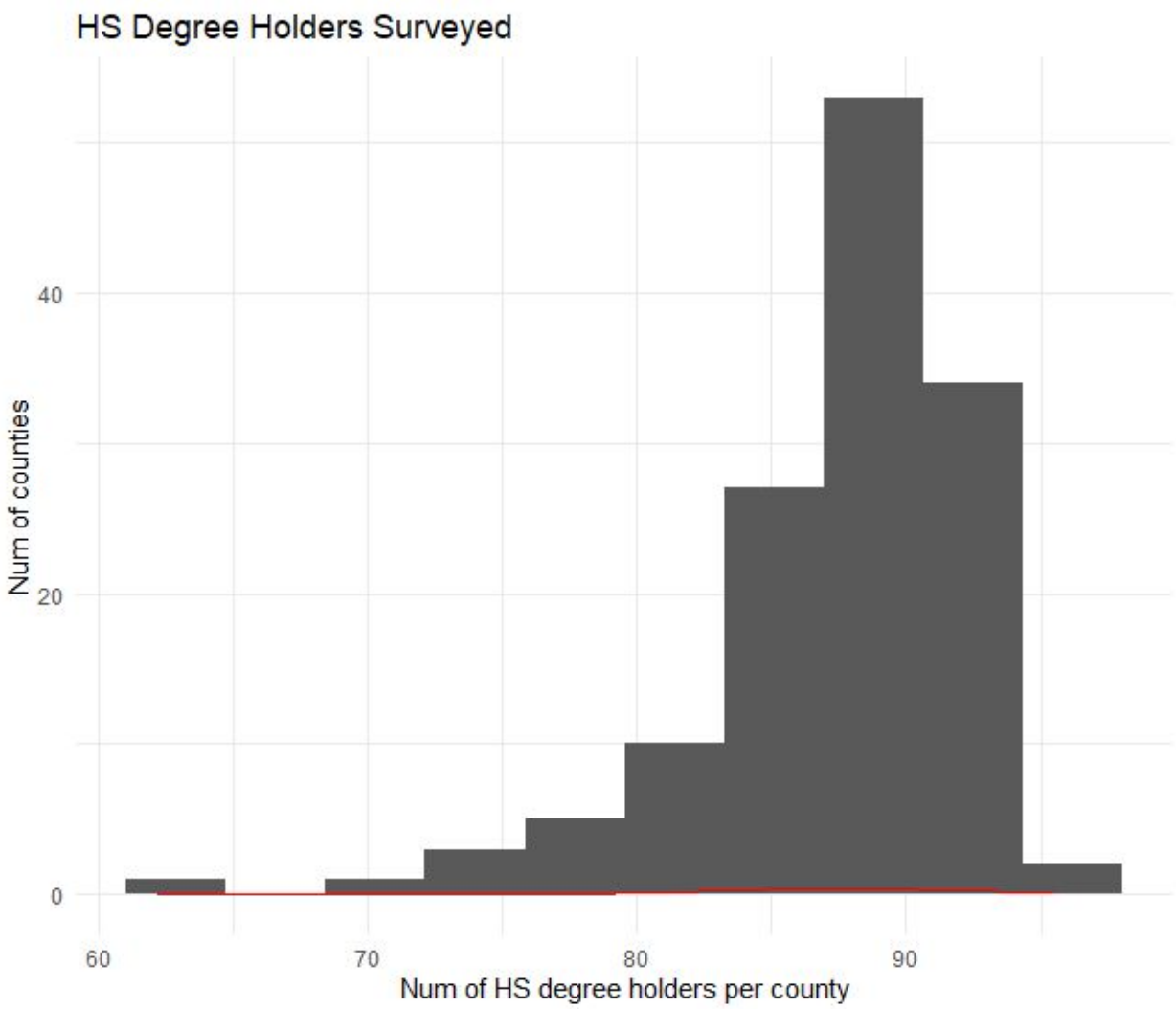
```
> # Assignment: American Survey Assignment
> # Name: Kline, Matthew
> # Date: 2020-09-20
>
> ## Load the ggplot2 package
> library(ggplot2)
> theme_set(theme_minimal())
>
> ## Set the working directory to the root of your DSC 520 directory
> setwd("C:/Users/Matt Kline/Documents/GitHub/dsc520")
>
> ## Load the data
> survey <- read.csv("data/acs-14-1yr-s0201.csv")
>
> summary(survey)
      Id                 Id2          Geography           PopGroupID POPGROUP.display.label
RacesReported         HSDegree       BachDegree
 Length:136         Min.   : 1073   Length:136         Min.   :1    Length:136           Min.   :  500292
Min.   :62.20   Min.   :15.40
 Class :character   1st Qu.:12082   Class :character   1st Qu.:1    Class :character     1st Qu.:
631380   1st Qu.:85.50   1st Qu.:29.65
 Mode  :character   Median :26112   Mode  :character   Median :1    Mode  :character
Median :  832708   Median :88.70   Median :34.10
                    Mean   :26833                      Mean   :1                       Mean   : 1144401   Mean
:87.63   Mean   :35.46
                    3rd Qu.:39123                      3rd Qu.:1                       3rd Qu.: 1216862   3rd
Qu.:90.75   3rd Qu.:42.08
                    Max.   :55079                      Max.   :1                       Max.   :10116705   Max.
:95.50   Max.   :60.30
> head(survey)
            Id  Id2                      Geography PopGroupID POPGROUP.display.label
RacesReported HSDegree BachDegree
1 0500000US01073 1073       Jefferson County, Alabama          1       Total population
660793     89.1       30.5
2 0500000US04013 4013        Maricopa County, Arizona          1       Total population
4087191     86.8       30.2
3 0500000US04019 4019           Pima County, Arizona          1       Total population
1004516     88.0       30.8
4 0500000US06001 6001      Alameda County, California          1       Total population
1610921     86.9       42.8
5 0500000US06013 6013 Contra Costa County, California          1       Total population
1111339     88.8       39.7
```

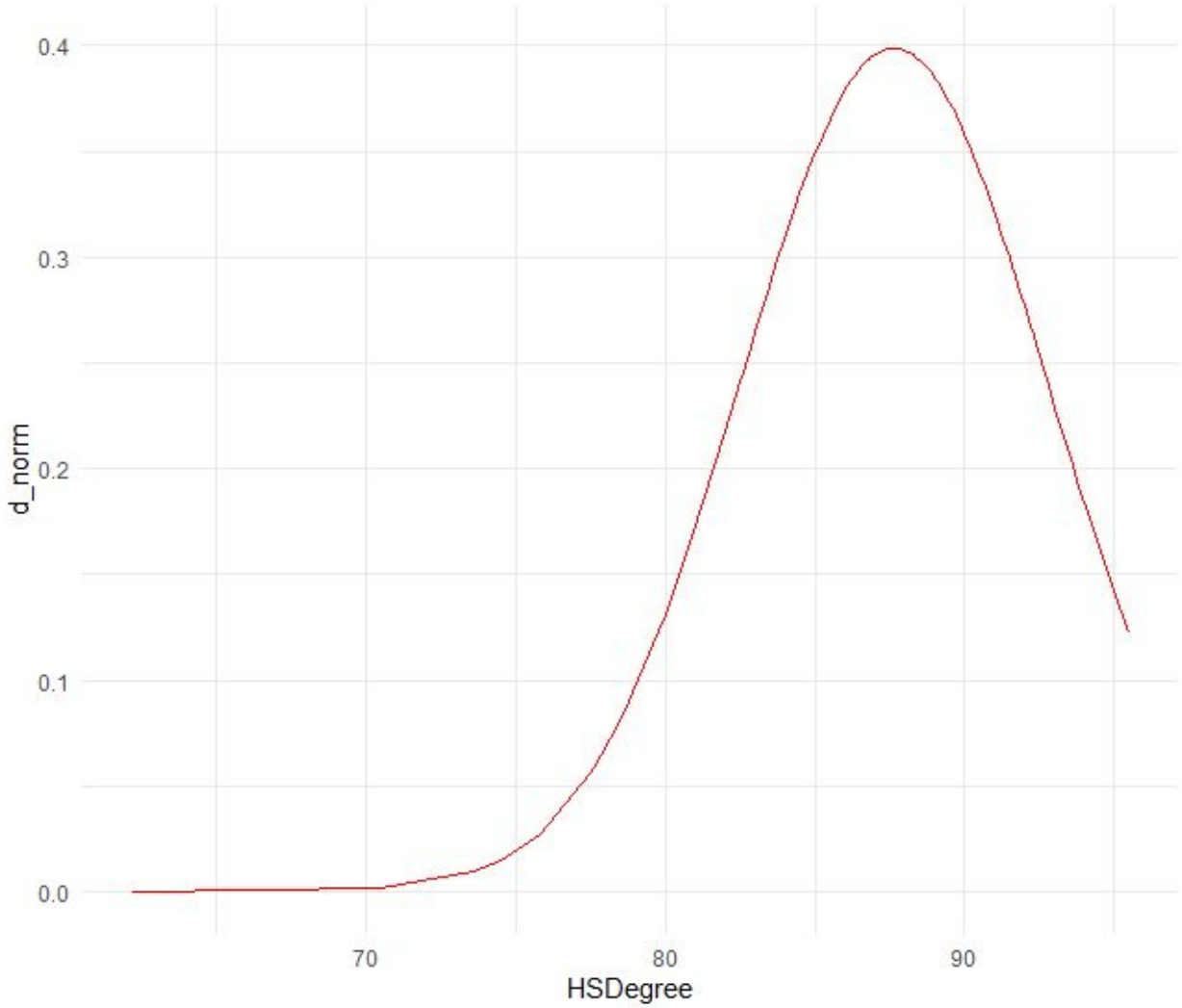
```
6 0500000US06019 6019       Fresno County, California          1       Total population
965974     73.6       19.7
>
> str(survey)
'data.frame':    136 obs. of  8 variables:
 $ Id                  : chr  "0500000US01073" "0500000US04013" "0500000US04019"
"0500000US06001" ...
 $ Id2                 : 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 ...
> ##Question 1: Based off of our str output:
> # Id                  : chr
> # Id2                 : int
> # Geography           : chr
> # PopGroupID          : int
> # POPGROUP.display.label: chr
> # RacesReported       : int
> # HSDegree            : num
> # BachDegree          : num
>
> ##Question 2:
> nrow(survey)
[1] 136
> ncol(survey)
[1] 8
>
> ##Question 3:
> ggplot(survey, aes(HSDegree)) + geom_histogram(bins=10)+ xlab("Num of HS degree
holders per county") + ylab("Num of counties") + ggtitle("HS Degree Holders Surveyed")
>
> ##Qustion 4:
> #A: Yes the data is unimodal, with the peak being between 85-90.
> mean(survey$HSDegree, na.rm = TRUE)
[1] 87.63235
> median(survey$HSDegree, na.rm = TRUE)
[1] 88.7
```

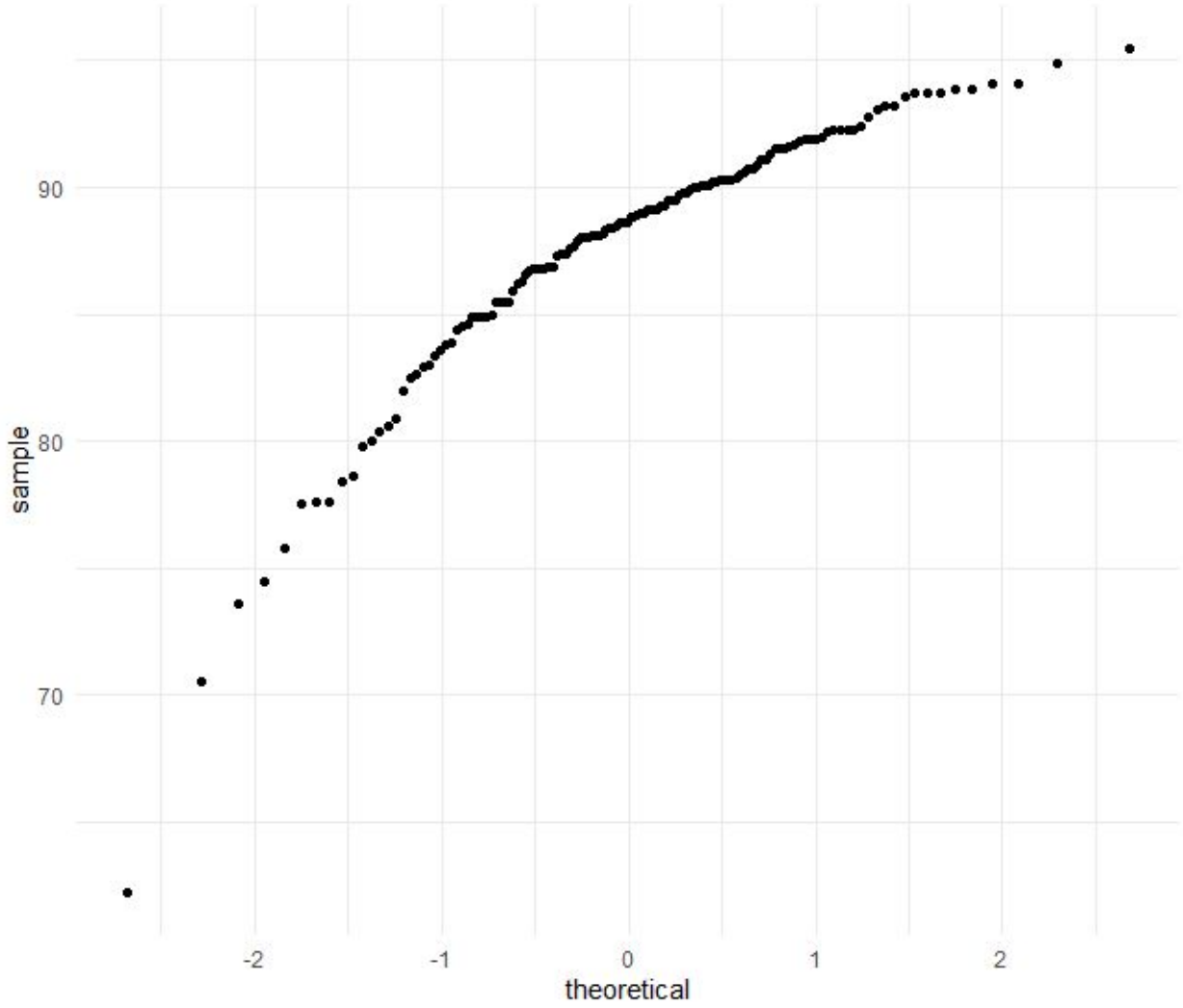
```
> # mode function found at : https://www.tutorialspoint.com/r/r_mean_median_mode.htm
> getmode <- function(v) {
+   uniqv <- unique(v)
+   uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
> print(getmode(survey$HSDegree))
[1] 89.1
> #B: The graph is not symmetrical because the mode, mean, and median do not all occur at
the same point.
> #C: The graph does have a bell shape with only one peak.
> #D: No the graph is not normal.
> #E: The graph is skewed to the right.
> #F:
> z<- (survey$HSDegree - mean(survey$HSDegree ))/sd(survey$HSDegree)
> data <- cbind(survey, z)
> d_norm <- dnorm(data$z)
> data <- cbind(data, d_norm)
> ggplot(data = data, aes(x = HSDegree)) + geom_histogram(bins=10) + geom_line(aes(
y=d_norm), colour="red") + xlab("Num of HS degree holders per county") + ylab("Num of
counties") + ggtitle("HS Degree Holders Surveyed")
> ggplot(data = data, aes(x = HSDegree)) + geom_line(aes( y=d_norm), colour="red")
> #G: Yes because a normal distribuition shows the prbability at what value might occur and
with the propper z scores we would see an accurate calculation.
>
> ##Question 5:
> ggplot(survey, aes(sample = HSDegree)) + stat_qq()
> ##Question 6:
> #A: No, because the data is curved slightly on the graph instead of being a straight line.
> #B: The plot is skewed to the left since the data points downs and to the right after the bend.
>
> ##Question 7:
> library(pastecs)
> stat.desc(survey$HSDegree)
    nbr.val     nbr.null      nbr.na         min         max       range         sum
median        mean
SE.mean CI.mean.0.95
1.360000e+02 0.000000e+00 0.000000e+00 6.220000e+01 9.550000e+01 3.330000e+01
1.191800e+04 8.870000e+01 8.763235e+01 4.388598e-01 8.679296e-01
         var      std.dev     coef.var
2.619332e+01 5.117941e+00 5.840241e-02
> library(e1071)
> kurtosis(survey$HSDegree)
[1] 4.352856
```

> ##Question 8: When looking at the graphs of the data we can see how the data is skewed. Based on the kurtosis value we can have a better understanding of the data's peak.
> #This data was skewed to the right side of the graph, showing the majority of the data was around the mean. The kurtosis value showed that this data set was too peaked due
> #to the bunching of the data around the mean. Based off of our distribution graph, we can tell that the zscores fit with a normal distribution. If we had a different data size,
> #this could change all these factors, our data may not have the same peak, as well as we may see a change in the normality of the distribution. This would result in changing the kurtosis, skew, and zscores.
>




HS Degree Holders Surveyed
Num of counties
0
20
40
60
70
80
90
Num of HS degree holders per county


0.4
0.3
0.2
0.1
0.0
d_norm
70
80
90
HSDegree


90
80
70
sample
-2
-1
0
1
2
theoretical