Community Survey

Answers:

a)

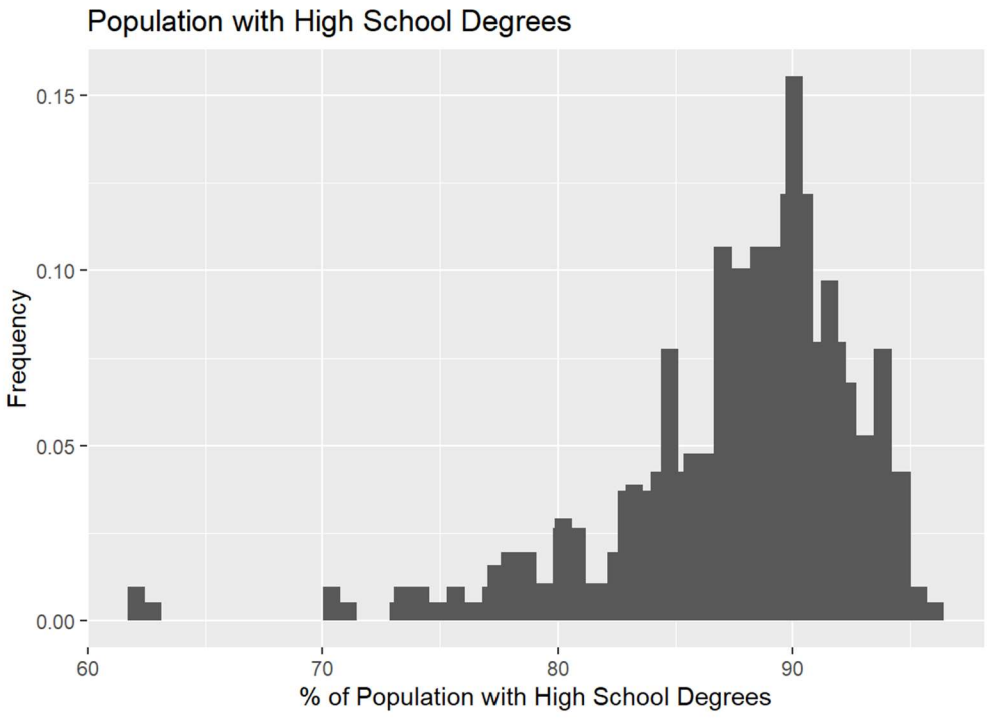
i)

Id:	chr
Id2:	int
Geography:	chr
PopGroupID:	int
POPGROUP.displat.label:	chr
RacesReported:	int
HSDegree:	num
BachDegree:	num

ii)

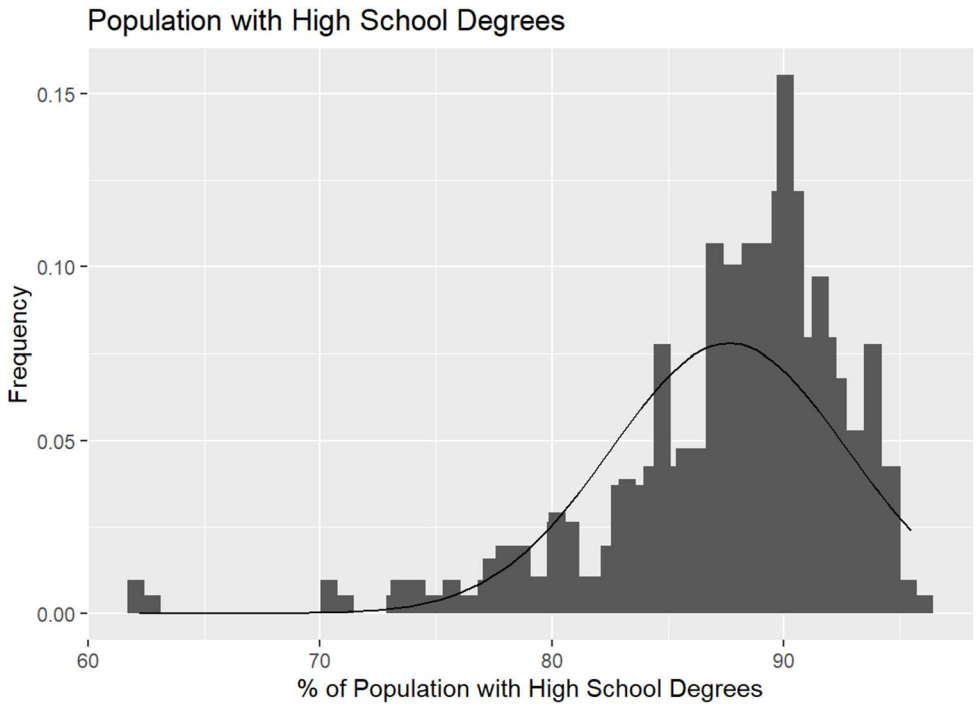
```
> str(survey)
'data.frame':   136 obs. of  8 variables:
 $ Id                    : chr  "0500000US01073" "0500000US04013" "0500000US04019" "050000
0US06001" ...
 $ Id2                   : int  1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...
 $ Geography             : chr  "Jefferson County, Alabama" "Maricopa County, Arizona" "Pi
ma 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 10116
705 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 ...
```

iii)



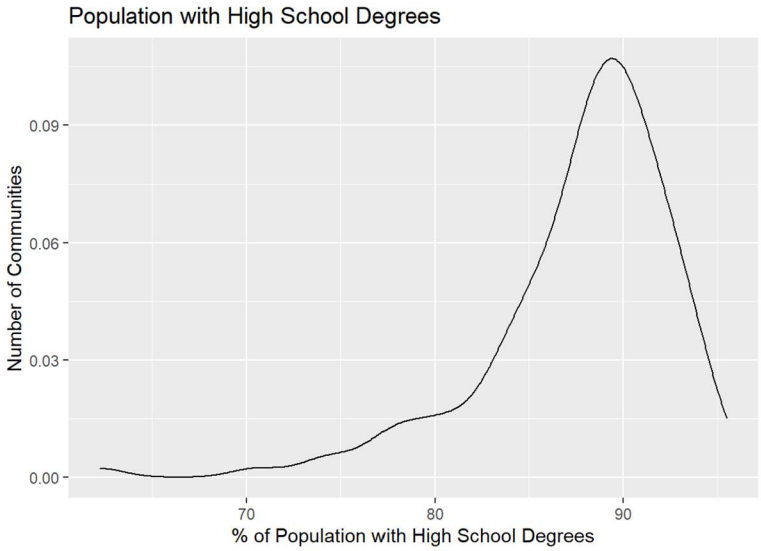
iv)

1) Yes, based on this graph the data appears to be unimodal because there is only one highest peak
2) No, the histogram is not symmetrical. The data is skewed to one side.
3) No, it is not bell shaped because it is skewed to one side.
4) No, the histogram is not normally distributed
5) Yes, the histogram is skewed. It is negatively skewed.
6)



7) No, a normal distribution would not be an accurate model for this data as it can be seen the data is far from the normal curve.

v)



vi)

1) No, the distribution is not normal. The data is heavily skewed to one end of the graph.
2) Yes, the data is skewed. It is negatively skewed because the graph has its peak to the right side and a long tail to the left.

vii)

```
> round(stat.desc(survey$HSDegree, basic=FALSE, norm=TRUE), 2)
     median        mean     SE.mean CI.mean.0.95         var
      88.70       87.63        0.44        0.87       26.19
    std.dev    coef.var    skewness    skew.2SE    kurtosis
       5.12        0.06       -1.67       -4.03        4.35
  kurt.2SE  normtest.W  normtest.p
       5.27        0.88        0.00
```

viii)

- The skew being a negative number suggests that there is a build up of high scores
- The positive value for kurtosis indicates that the data set is pointy with a heavily-tailed distribution
- The skew.2SE and kurt.2SE being larger than 1 indicates that they are significant
- As the sample size increases the kurtosis and skew numbers could appear to be significant even if they are not and these numbers should not be used for analysis

Code:

```
survey <- read.csv("data\\acs-14-1yr-s0201.csv")

str(survey)

nrow(survey)

ncol(survey)

sur_histo <- ggplot(survey, aes(HSDegree))


sur_histo + geom_histogram(aes(y=..density..), bins = 45) +
  labs(x="% of Population with High School Degrees", y="Frequency") +
  ggtitle("Population with High School Degrees") +
  stat_function(fun=dnorm, args=list(mean=mean(survey$HSDegree, na.rm=TRUE),
                                     sd=sd(survey$HSDegree, na.rm=TRUE)))

sur_histo + geom_density() +
  labs(x="% of Population with High School Degrees", y="Number of Communities") +
  ggtitle("Population with High School Degrees")

round(stat.desc(survey$HSDegree, basic=FALSE, norm=TRUE), 2)
```