Assignment 3.2

American Community Survey Exercise

Gabriel Avinaz

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
setwd("C:\\Users\\desaTuration\\OneDrive - Bellevue University\\DSC520-T301 Statistics for Data Science
community_df <- read_csv("acs-14-1yr-s0201.csv")
```

```
## Rows: 136 Columns: 8
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): Id, Geography, POPGROUP.display-label
## dbl (5): Id2, PopGroupID, RacesReported, HSDegree, BachDegree
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(community_df)
```

```
## Rows: 136
## Columns: 8
## $ Id                       <chr> "0500000US01073", "0500000US04013", "0500000U~
## $ Id2                      <dbl> 1073, 4013, 4019, 6001, 6013, 6019, 6029, 603~
## $ Geography                <chr> "Jefferson County, Alabama", "Maricopa County~
## $ PopGroupID               <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ `POPGROUP.display-label` <chr> "Total population", "Total population", "Tota~
## $ RacesReported            <dbl> 660793, 4087191, 1004516, 1610921, 1111339, 9~
## $ HSDegree                 <dbl> 89.1, 86.8, 88.0, 86.9, 88.8, 73.6, 74.5, 77.~
## $ BachDegree               <dbl> 30.5, 30.2, 30.8, 42.8, 39.7, 19.7, 15.4, 30.~
```

```
str(community_df)
```

```
## spec_tbl_df [136 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                    : chr [1:136] "0500000US01073" "0500000US04013" "0500000US04019" "0500000US
##  $ Id2                   : num [1:136] 1073 4013 4019 6001 6013 ...
##  $ Geography             : chr [1:136] "Jefferson County, Alabama" "Maricopa County, Arizona" "Pima C
##  $ PopGroupID            : num [1:136] 1 1 1 1 1 1 1 1 1 1 ...
##  $ POPGROUP.display-label: chr [1:136] "Total population" "Total population" "Total population" "Tota
##  $ RacesReported         : num [1:136] 660793 4087191 1004516 1610921 1111339 ...
##  $ HSDegree              : num [1:136] 89.1 86.8 88 86.9 88.8 73.6 74.5 77.5 84.6 80.6 ...
##  $ BachDegree            : num [1:136] 30.5 30.2 30.8 42.8 39.7 19.7 15.4 30.3 38 20.7 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_character(),
##   ..   Id2 = col_double(),
##   ..   Geography = col_character(),
##   ..   PopGroupID = col_double(),
##   ..   `POPGROUP.display-label` = col_character(),
##   ..   RacesReported = col_double(),
##   ..   HSDegree = col_double(),
##   ..   BachDegree = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

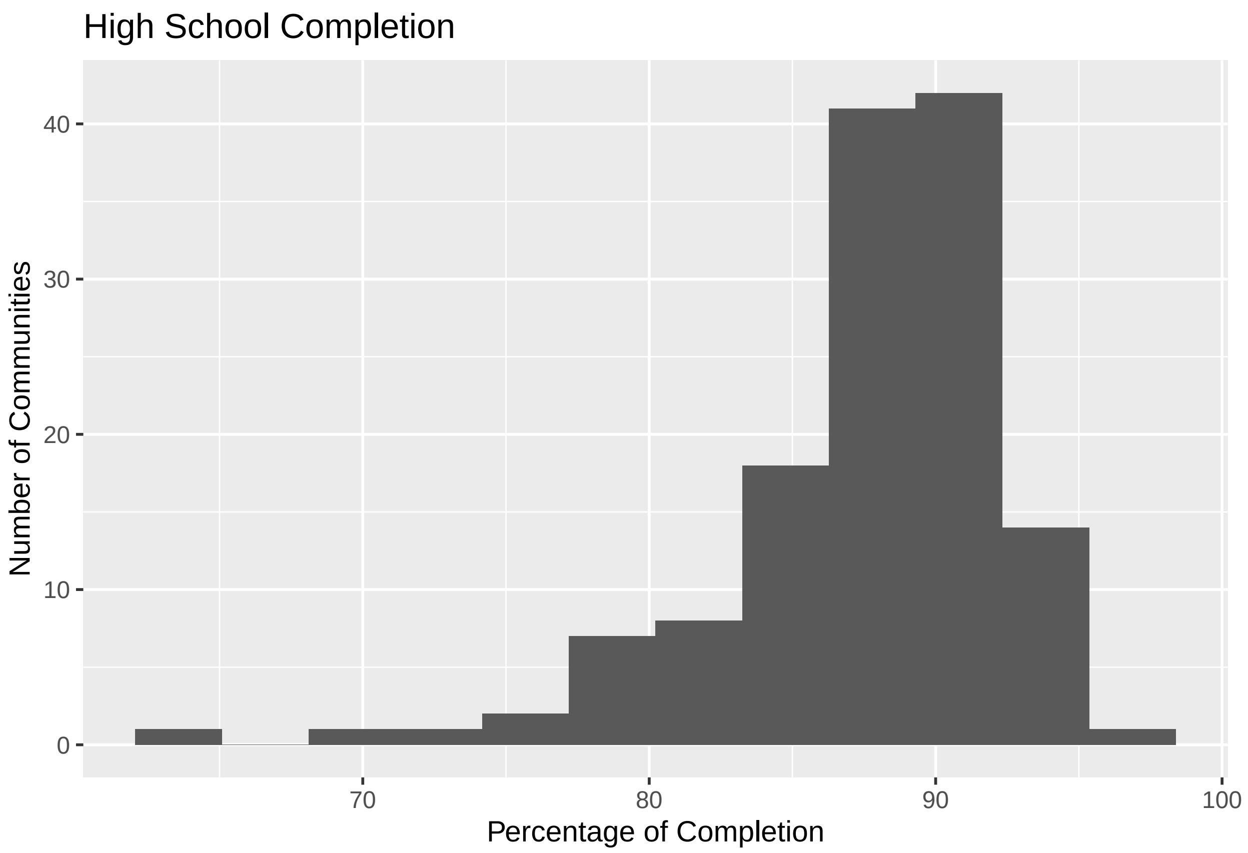
```
nrow(community_df)
```

```
## [1] 136
```

```
ncol(community_df)
```

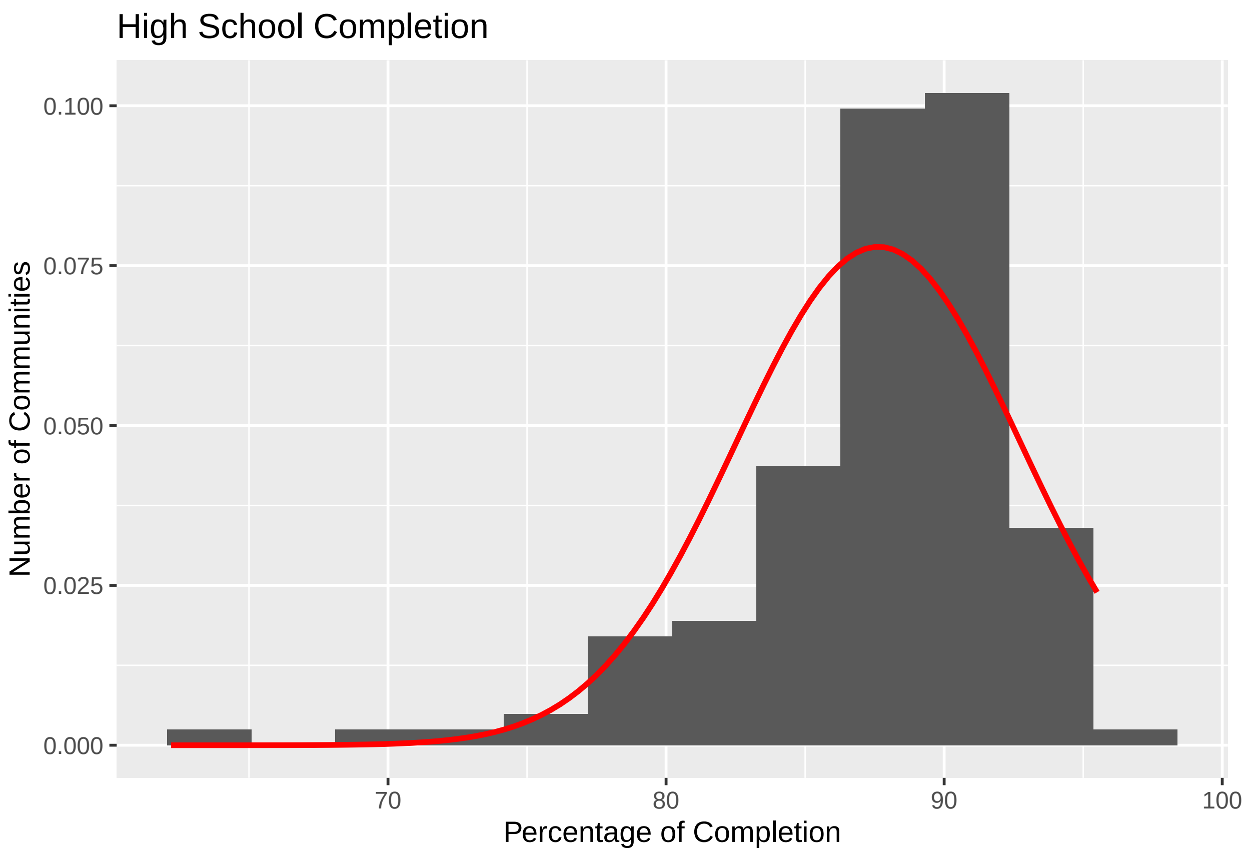
```
## [1] 8
```

```
library(ggplot2)
ggplot(community_df, aes(HSDegree)) + geom_histogram(bins = 12) + ggtitle("High School Completion") + yl
```



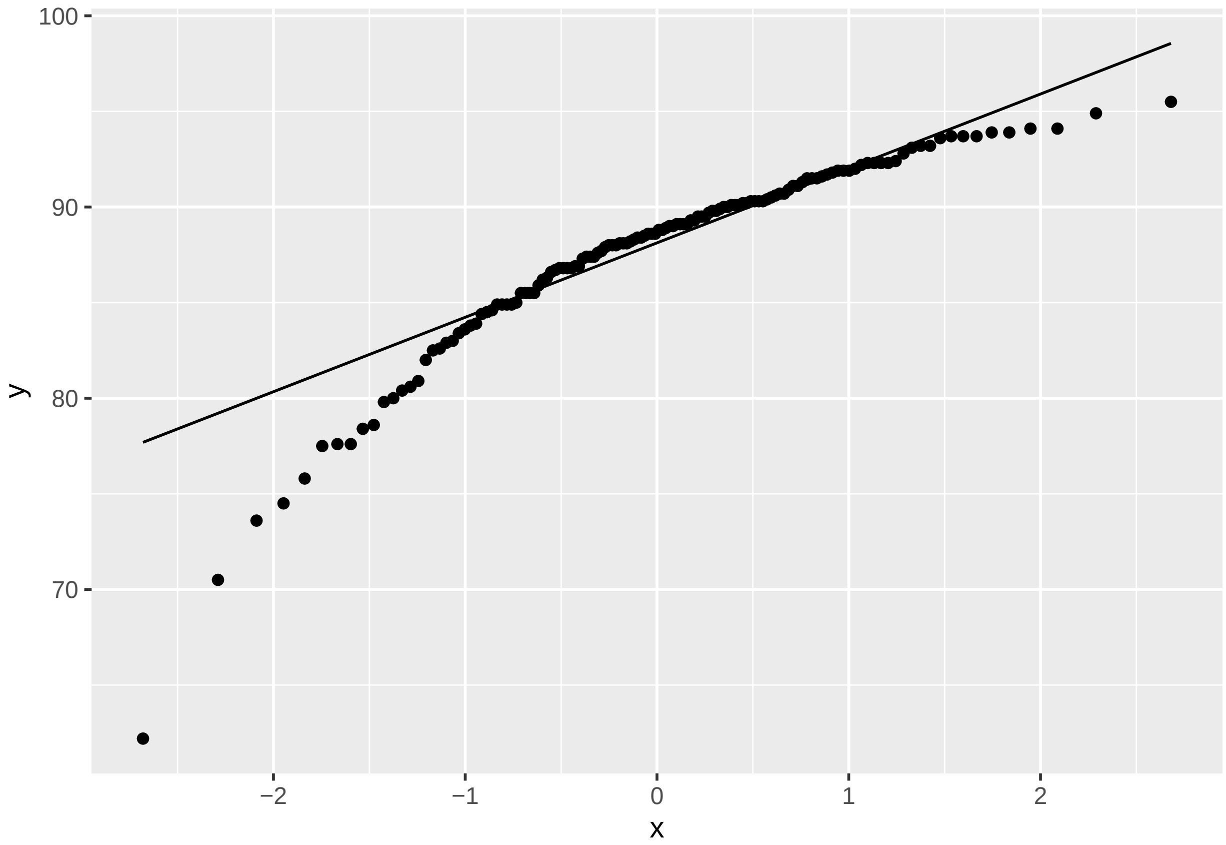
1. Based on what you see in this histogram, is the data distribution unimodal?
Yes, there is a clear singular peak around the 90% mark of the distribution.
2. Is it approximately symmetrical?
No, the distribution isn't similar on both sides of the
3. Is it approximately bell-shaped?
No, since it isn't symmetrical it can't be a bell curve, even if it has one peak.
4. Is it approximately normal?
No, the tails at the end of curve are not even.
5. If not normal, is the distribution skewed? If so, in which direction?
It is skewed Negatively, the distribution's tail on the left is longer.
6. Include a normal curve to the Histogram that you plotted.

```
ggplot(community_df, aes(HSDegree)) + geom_histogram(bins = 12, aes(y = ..density..)) + stat_function(fu
```



7. Explain whether a normal distribution can accurately be used as a model for this data.
A normal distribution cannot be used as a modal for this data because of it's negative skew.

```
ggplot(community_df, aes(sample=HSDegree)) + geom_qq() + geom_qq_line()
```



1. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.
It would appear to not be a normal distribution. There is an apparent curve in the plot
2. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.
It is skewed in a negative direction, this is represented by the downward curve as we move away from the median.

```
library(pastecs)
```

```
##
## Attaching package: 'pastecs'

## The following objects are masked from 'package:dplyr':
##
##     first, last

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
stat.desc(community_df['HSDegree'], norm = TRUE)
```

```
##                    HSDegree
## nbr.val       1.360000e+02
## nbr.null      0.000000e+00
## nbr.na        0.000000e+00
## min           6.220000e+01
## max           9.550000e+01
## range         3.330000e+01
## sum           1.191800e+04
## median        8.870000e+01
## mean          8.763235e+01
## SE.mean       4.388598e-01
## CI.mean.0.95  8.679296e-01
## var           2.619332e+01
## std.dev       5.117941e+00
## coef.var      5.840241e-02
## skewness     -1.674767e+00
## skew.2SE     -4.030254e+00
## kurtosis      4.352856e+00
## kurt.2SE      5.273885e+00
## normtest.W    8.773635e-01
## normtest.p    3.193634e-09
```

Skew and Kurtosis can be determined by the values generated for "Kurtosis" and "Skewness." For kurtosis, a positive number indicates a larger than normal collection of data near the peak. When examining skew, the farther the number deviates from zero, the more skewed the distribution will be. In this case, a negative skew number indicates a larger left tail. We can use the mean and standard deviation to calculate for z-score, which is a measure of the number of standard deviations away from the mean a data point would be. for example the HSDegree z-score for "Jefferson County, Alabama" would be 0.2872249. Population size can have a significant impact on the standard error of the mean, since we do not know the population's standard deviation, the larger our sample, the closer our sample std.dev will be to a population.