library(readr)

acs\_14\_1yr\_s0201 <- read\_csv("completed/assignment03/acs-14-1yr-s0201.csv")

table1 <- acs\_14\_1yr\_s0201

library(ggplot2)

I: List the name of each field and what you believe the data type

and intent is of the data included in each field

(Example: Id - Data Type: varchar (contains text and numbers)

Intent: unique identifier for each row)

spec(table1)

cols(

Id = col\_character(), -- contains numbers and characters

Id2 = col\_double(), -- includes numbers only

Geography = col\_character(), -- characters describing the observation location by county

PopGroupID = col\_double(), -- numeric (value of 1), ID denoting the population group

`POPGROUP.display-label` = col\_character(), -- character, denoting the label of the population group

RacesReported = col\_double(), -- number, a count of the reported races in the survey

HSDegree = col\_double(), -- number (float), presumably a percentage of the population with HS Degrees

BachDegree = col\_double()) -- number (float), presumably a percentage of the population with Bachelor's Degrees

II: Run the following functions and provide the results: str(); nrow(); ncol()

str(table1)

spc\_tbl\_ [136 × 8] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ Id : chr [1:136] "0500000US01073" "0500000US04013" "0500000US04019" "0500000US06001" ...

$ Id2 : num [1:136] 1073 4013 4019 6001 6013 ...

$ Geography : chr [1:136] "Jefferson County, Alabama" "Maricopa County, Arizona" "Pima County, Arizona" "Alameda County, California" ...

$ 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" "Total population" ...

$ 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(table1) -- 136

ncol(table1) -- 8

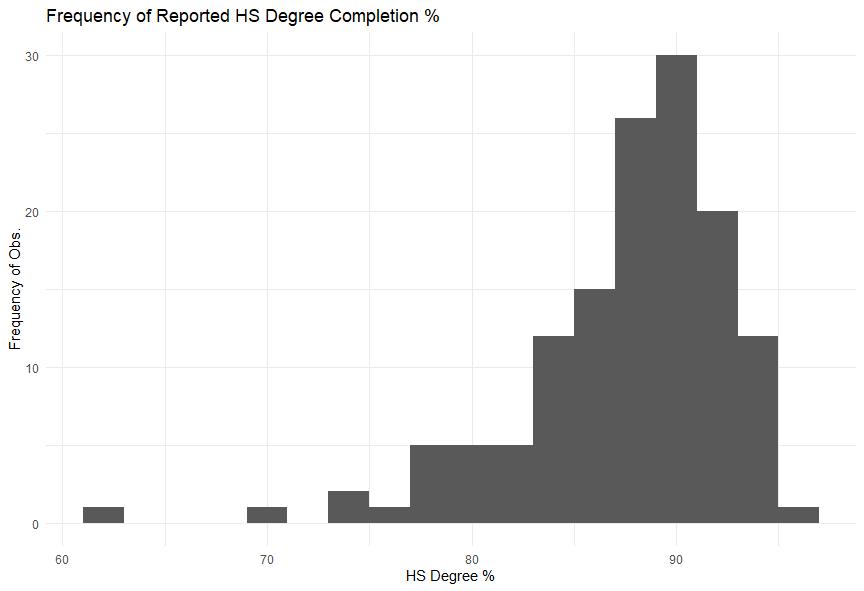
III: Create a Histogram of the HSDegree variable using the ggplot2 package.

Set a bin size for the Histogram that you think best visuals the data

(the bin size will determine how many bars display and how wide they are)

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

ggplot(table1, aes(HSDegree)) + geom\_histogram(binwidth = 2) + xlab("HS Degree %") + ylab("Frequency of Obs.") + ggtitle("Frequency of Reported HS Degree Completion %")



IV: Answer the following questions based on the Histogram produced:

Based on what you see in this histogram, is the data distribution unimodal? -- The data is unimodal.

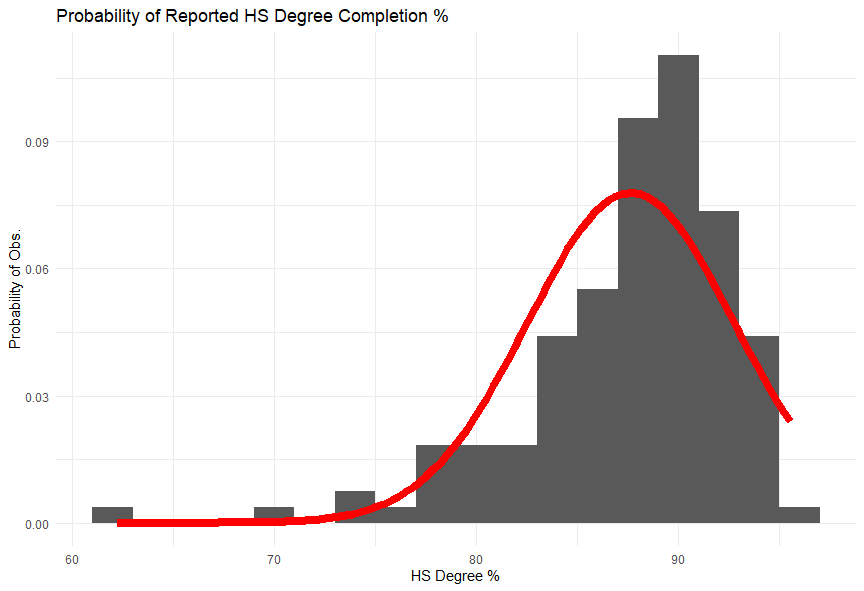
Is it approximately symmetrical? -- The data is not symetrical per its left-skewdness.

Is it approximately bell-shaped? -- The data is shaped like a bell, but is not normally distributed.

Is it approximately normal? -- The data isn't normal, since the median & mode are offset right from the mean.

If not normal, is the distribution skewed? If so, in which direction? -- The distribution is left-skewed.

Include a normal curve to the Histogram that you plotted.



ggplot(table1, aes(HSDegree)) + geom\_histogram(aes(y = after\_stat(density)), binwidth = 2) +

xlab("HS Degree %") + ylab("Probability of Obs.") +

ggtitle("Probability of Reported HS Degree Completion %") +

stat\_function(fun = dnorm,

args = list(mean = mean(table1$HSDegree),

sd = sd(table1$HSDegree)),

col = "red",

size = 3)

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

An offset normal distribution would be a low-accuracy predictor of the data, per the skewness and slightly positive kurtosis.

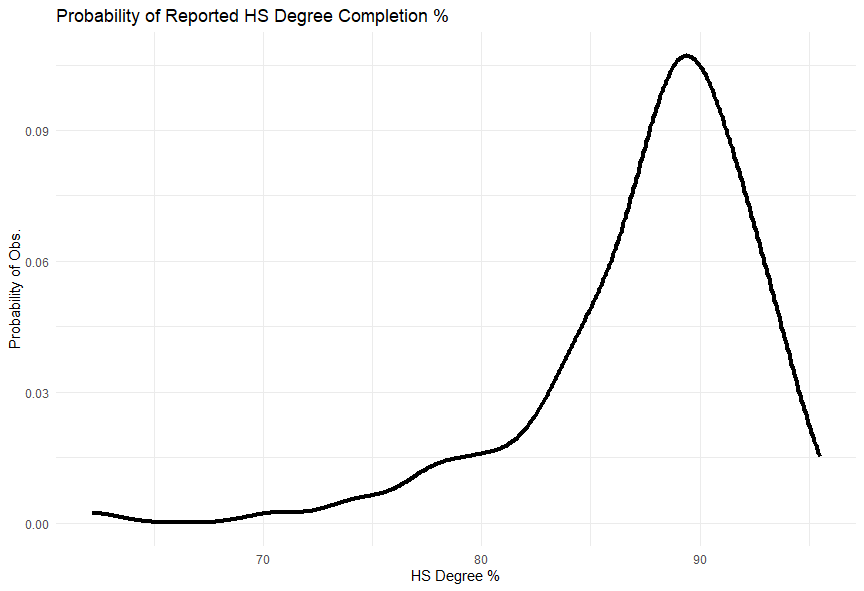
Values left of the center would be over-predicted, and the right, under-predicted.

V: Create a Probability Plot of the HSDegree variable.

ggplot(table1, aes(HSDegree)) + geom\_density(aes(HSDegree), size = 1.5) +

xlab("HS Degree %") + ylab("Probability of Obs.") +

ggtitle("Probability of Reported HS Degree Completion %")



VI: Answer the following questions based on the Probability Plot:

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

One could argue that the distribution resembles normality, but a normal distribution would be a poor predictor of this distribution per:

the left skewdness coupled with the positive kurtosis; therefore, I would state that the distribution is not 'approximately normal'.

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

The distribution is left-skewed because: mean < median < mode.

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.

table2 <- subset(table1, select = -c(PopGroupID))

stat.desc(table2, norm = TRUE)

Id Id2 Geography POPGROUP.display-label RacesReported HSDegree BachDegree

nbr.val NA 1.360000e+02 NA NA 1.360000e+02 1.360000e+02 136.00000000

nbr.null NA 0.000000e+00 NA NA 0.000000e+00 0.000000e+00 0.00000000

nbr.na NA 0.000000e+00 NA NA 0.000000e+00 0.000000e+00 0.00000000

min NA 1.073000e+03 NA NA 5.002920e+05 6.220000e+01 15.40000000

max NA 5.507900e+04 NA NA 1.011671e+07 9.550000e+01 60.30000000

range NA 5.400600e+04 NA NA 9.616413e+06 3.330000e+01 44.90000000

sum NA 3.649306e+06 NA NA 1.556385e+08 1.191800e+04 4822.70000000

median NA 2.611200e+04 NA NA 8.327075e+05 8.870000e+01 34.10000000

mean NA 2.683313e+04 NA NA 1.144401e+06 8.763235e+01 35.46102941

SE.mean NA 1.323036e+03 NA NA 9.351028e+04 4.388598e-01 0.81545273

CI.mean NA 2.616557e+03 NA NA 1.849346e+05 8.679296e-01 1.61271456

var NA 2.380576e+08 NA NA 1.189207e+12 2.619332e+01 90.43498856

std.dev NA 1.542911e+04 NA NA 1.090508e+06 5.117941e+00 9.50973126

coef.var NA 5.750024e-01 NA NA 9.529072e-01 5.840241e-02 0.26817415

skewness NA 4.793197e-02 NA NA 4.976198e+00 -1.674767e+00 0.32843046

skew.2SE NA 1.153462e-01 NA NA 1.197501e+01 -4.030254e+00 0.79035382

kurtosis NA -1.335207e+00 NA NA 3.349995e+01 4.352856e+00 -0.27742492

kurt.2SE NA -1.617726e+00 NA NA 4.058826e+01 5.273885e+00 -0.33612576

normtest.W NA 9.314546e-01 NA NA 5.185873e-01 8.773635e-01 0.98316075

normtest.p NA 3.490602e-06 NA NA 3.040373e-19 3.193634e-09 0.09206162

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?

Negative skewness is described by this relationship between summary statistics: mean < median < mode.

Positive kurtosis means the distribution is "taller" and "thinner" than a normal distribution,

meaning that the data is relatively more dense around the mode, and less dense away from the mode.

The p value from the Shapiro-Wilk test is also less than 0.05, rejecting the null that the distribution is normal.

I'm not sure what is meant regarding z scores since a particular point isn't mentioned, but I'll say

that a coefficient of variation of ~5 implies high variance and "instability" regarding the

prediction of dependent variables using simple approximations.

A larger sample size would potentially normalize the distribution, reducing all metrics deviating from the norm in proportion to the amount of samples added.

A smaller sample size would serve to exacerbate the lack of normality in the dataset, and provide more variance.