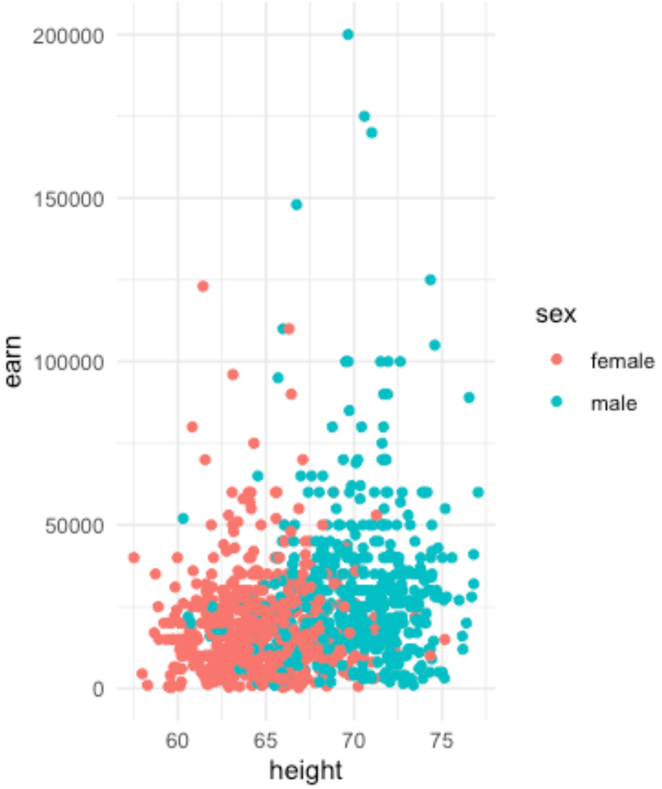
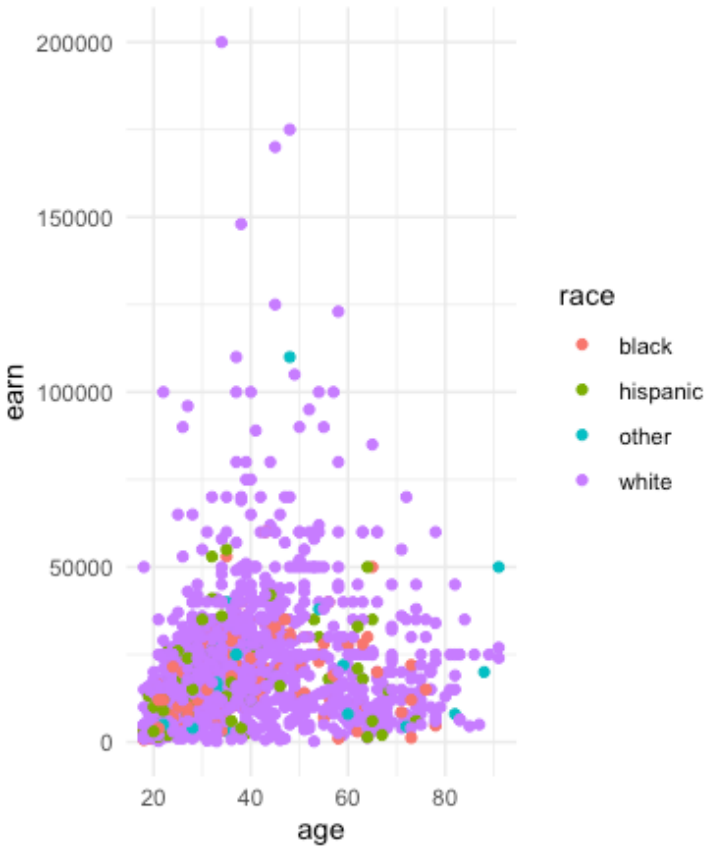
# Assignment: ASSIGNMENT 3
# Name: Koby-Hercsky, Theodore
# Date: 2021-04-18

1. Load the ggplot2 package
   a. library(ggplot2)
   b. theme_set(theme_minimal())
2. Set the working directory to the root of your DSC 520 directory
   a. setwd("/home/downloads/DSC520/dsc520")
3. Load the `data/r4ds/heights.csv` to
   a. heights_df <- read_csv("r4ds/heights.csv")
   b. https://ggplot2.tidyverse.org/reference/geom_point.html
4. Using `geom_point()` create three scatterplots for
   a. ## `height` vs. `earn`
      i. ggplot(heights_df, aes(x=height, y=earn)) + geom_point(aes(color=sex))

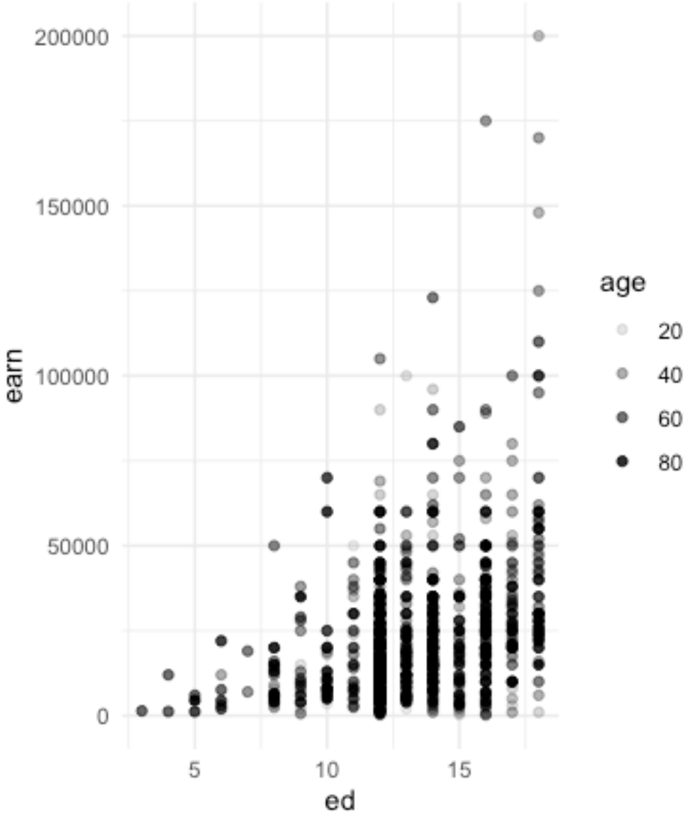


   b. `age` vs. `earn`
      i. ggplot(heights_df, aes(x=age, y=earn)) + geom_point(aes(color=race))

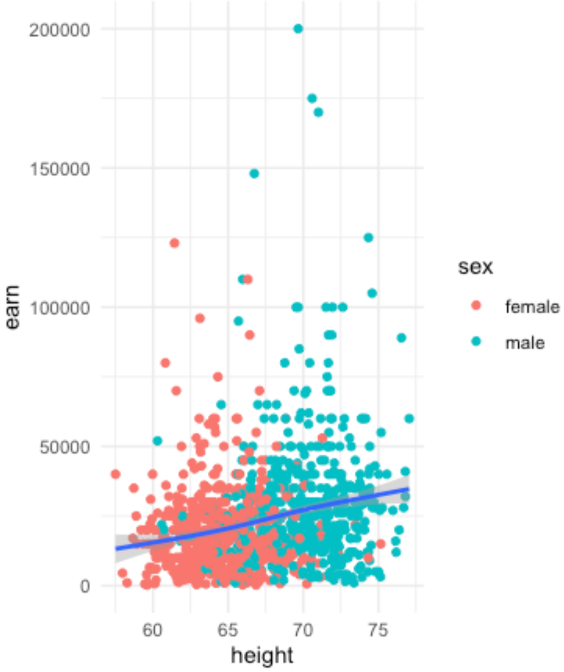


c. `ed` vs. `earn`

i. ggplot(heights_df, aes(x=ed, y=earn)) + geom_point(aes(alpha=age))

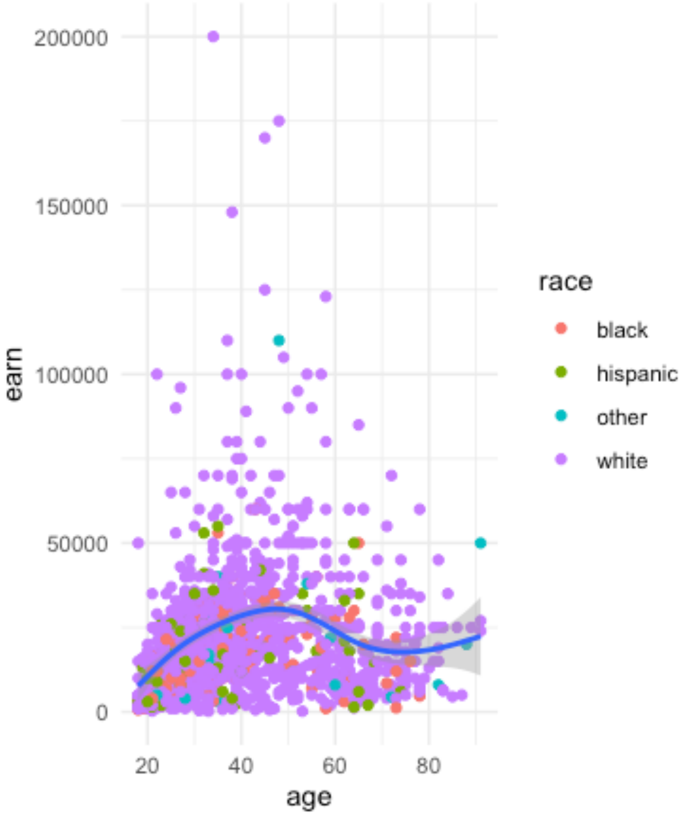


5. Re-create the three scatterplots and add a regression trend line using the `geom_smooth()` function
   a. **`height` vs. `earn`**
      i. ggplot(heights_df, aes(x=height, y=earn)) + geom_point(aes(color=sex)) + geom_smooth(orientation = "x")



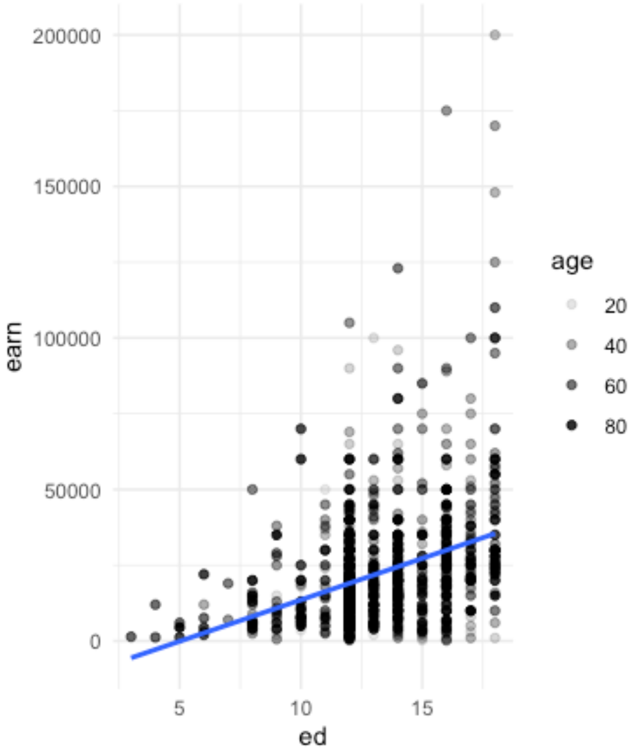
b. `age` vs. `earn`

   i. ggplot(heights_df, aes(x=age, y=earn)) + geom_point(aes(color=race)) + geom_smooth(span = 0.5)

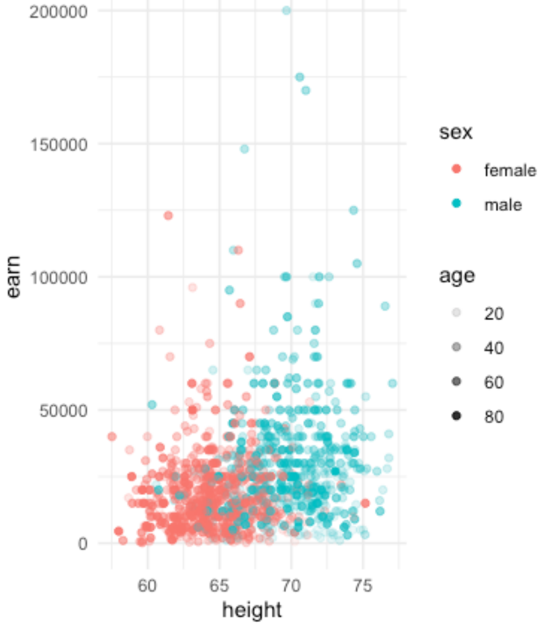


c. `ed` vs. `earn`

   i. ggplot(heights_df, aes(x=ed, y=earn)) + geom_point(aes(alpha=age)) + geom_smooth(se = FALSE, method = lm)

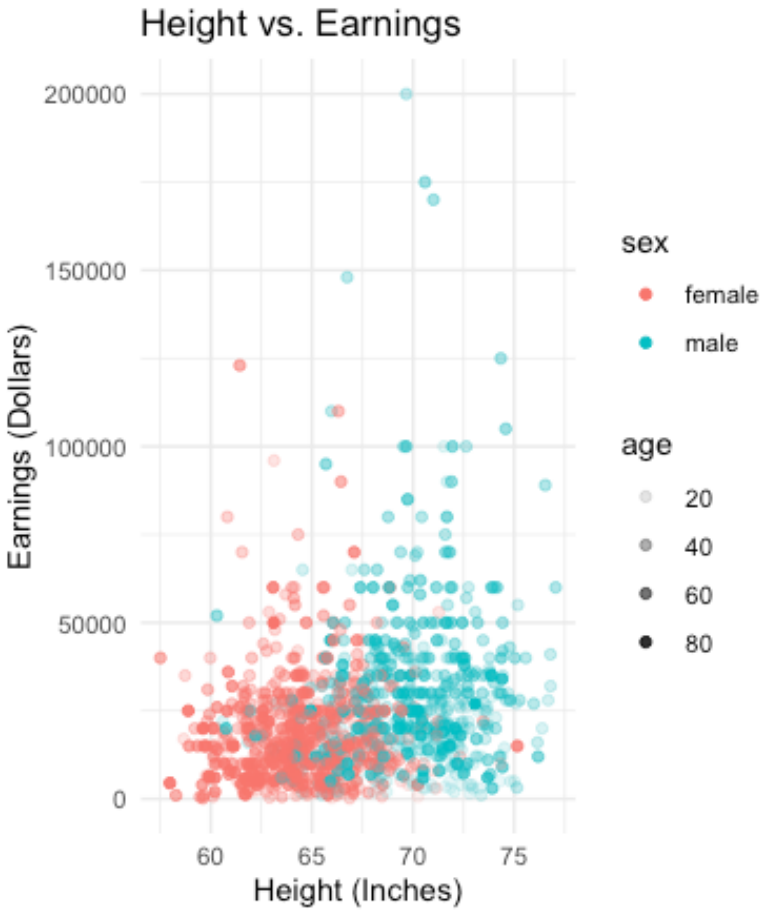


6. Create a scatterplot of `height`` vs. `earn`. Use `sex` as the `col` (color) attribute
    a. ggplot(heights_df, aes(x=height, y=earn, col=sex)) + geom_point(aes(alpha=age))

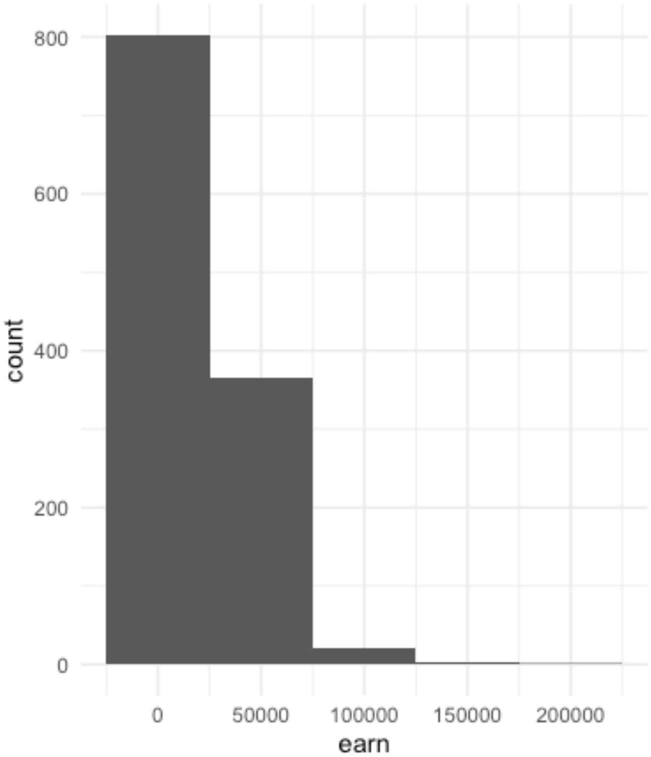


7. Using `ggtitle()`, `xlab()`, and `ylab()` to add a title, x label, and y label to the previous plot
    a. Title: Height vs. Earnings
    b. X label: Height (Inches)
    c. Y Label: Earnings (Dollars)

i. ggplot(heights_df, aes(x=height, y=earn, col=sex)) + geom_point(aes(alpha=age)) + ggtitle("Height vs. Earnings") + xlab("Height (Inches)") + ylab("Earnings (Dollars)")

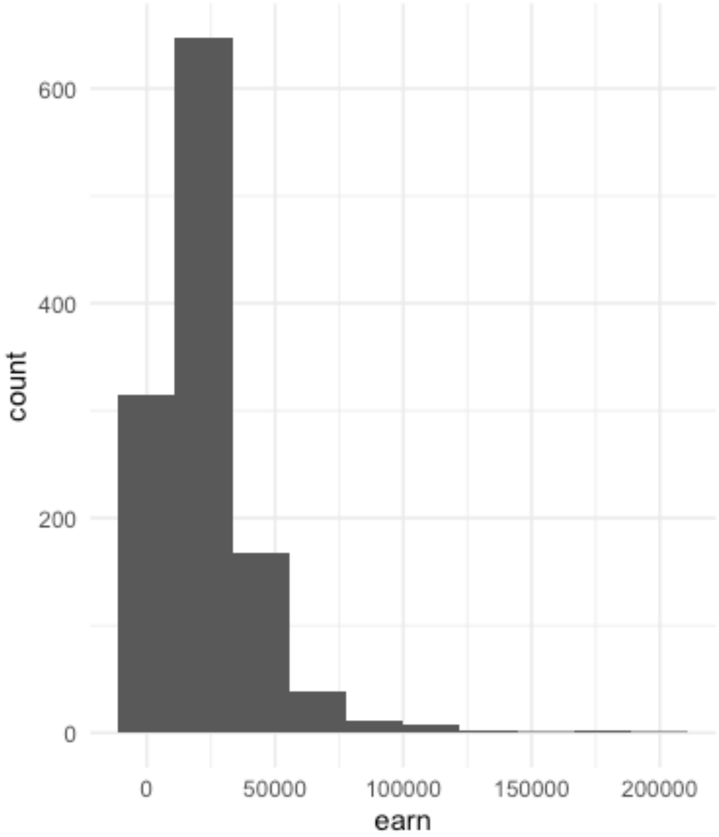


8. Create a histogram of the `earn` variable using `geom_histogram() `https://ggplot2.tidyverse.org/reference/geom_histogram.html
    a. ggplot(heights_df, aes(earn)) + geom_histogram(binwidth = 50000)

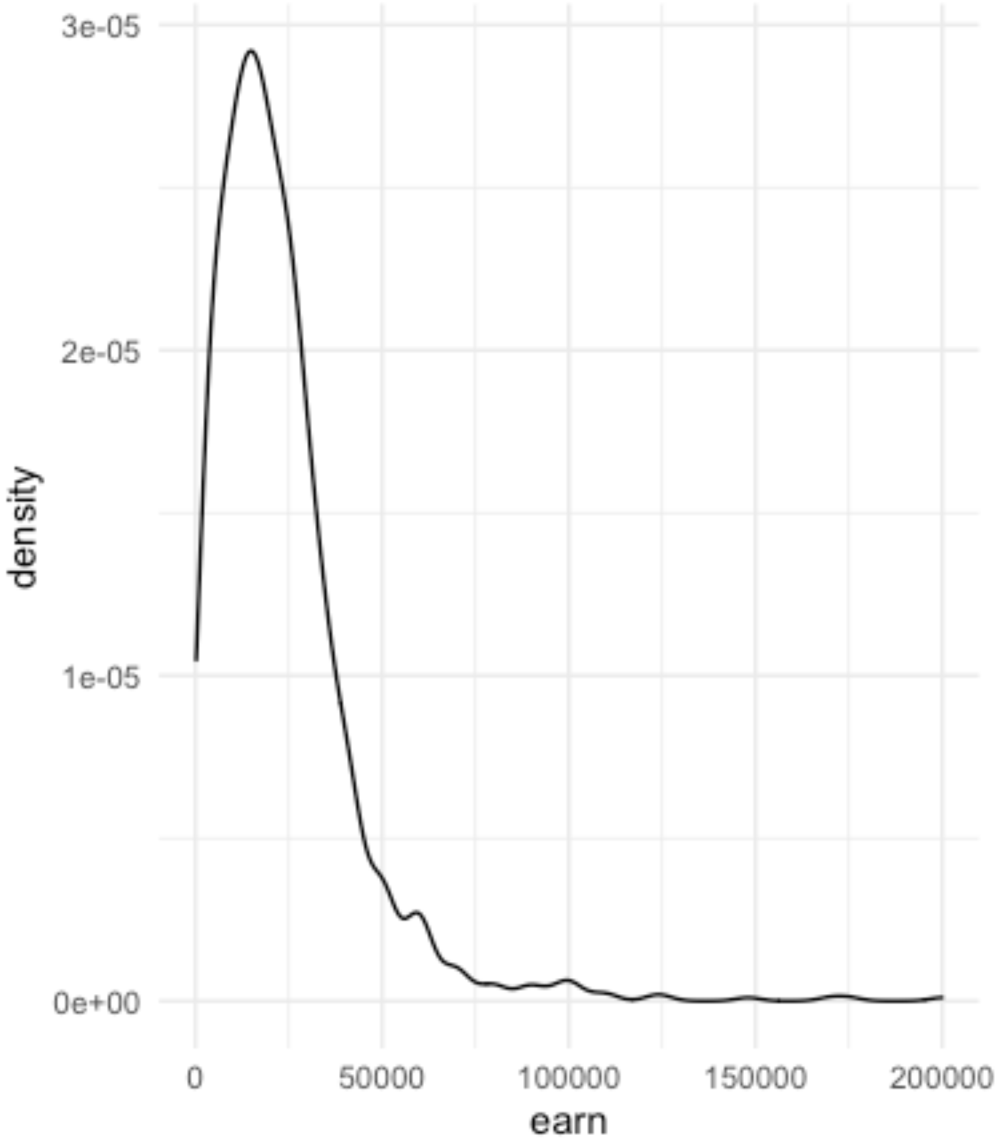


Create a histogram of the `earn` variable using `geom_histogram()` Use 10 bins

b. ggplot(heights_df, aes(earn)) + geom_histogram(bins = 10)

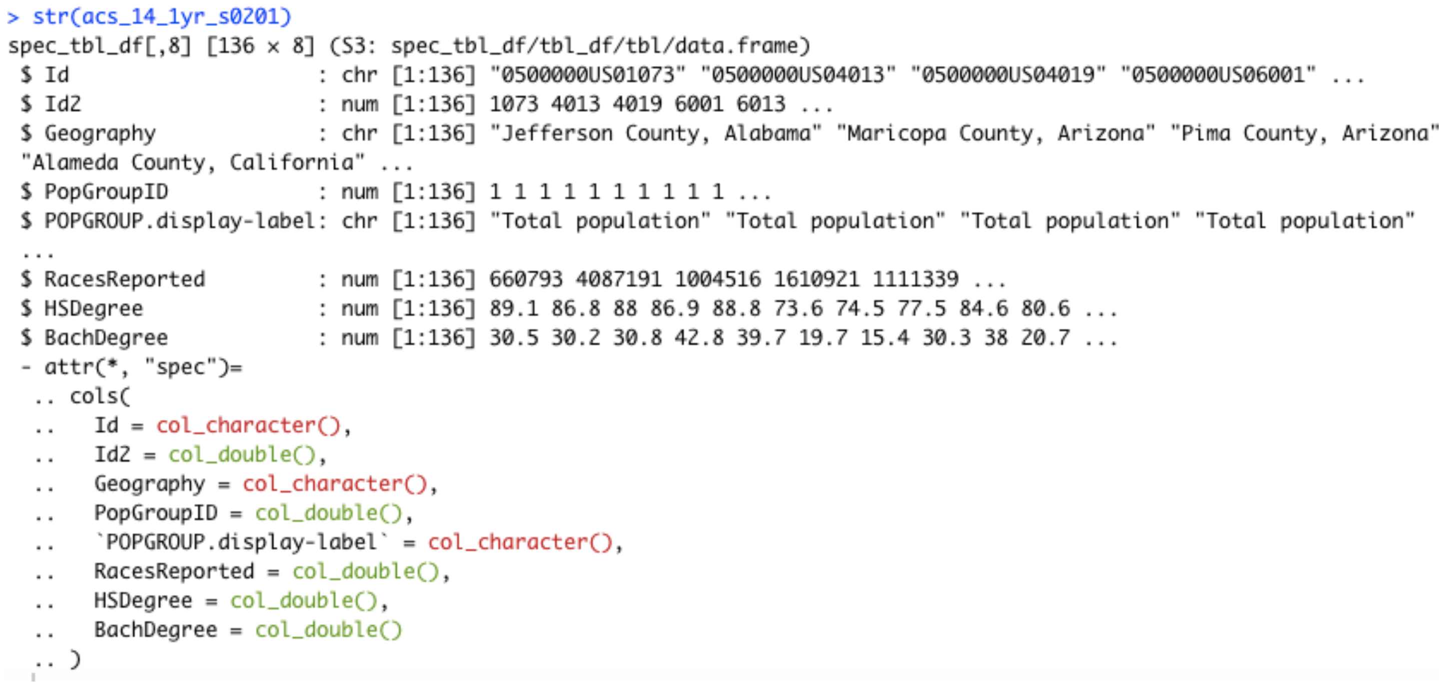


9. Create a kernel density plot of `earn` using `geom_density()` https://ggplot2.tidyverse.org/reference/geom_density.html
    a. ggplot(heights_df, aes(earn)) + geom_density(kernel = "gaussian")



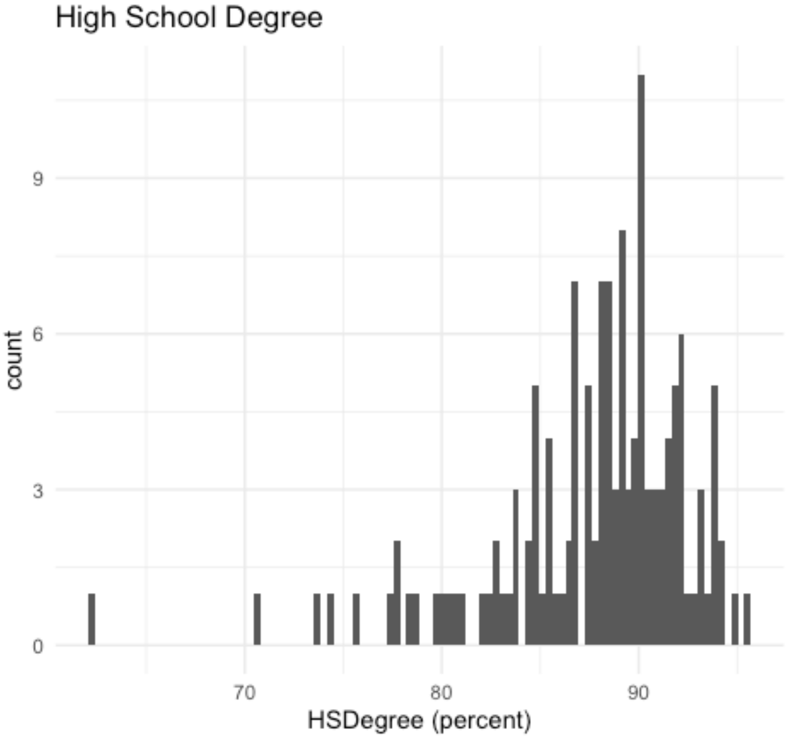
# Assignment: 2014 American Community Survey Data # Name: Koby-Hercsky, Theodore
# Date: 2021-03-18

1. Please provide the output from the following functions: str(); nrow(); ncol()
   a. str(acs_14_1yr_s0201)

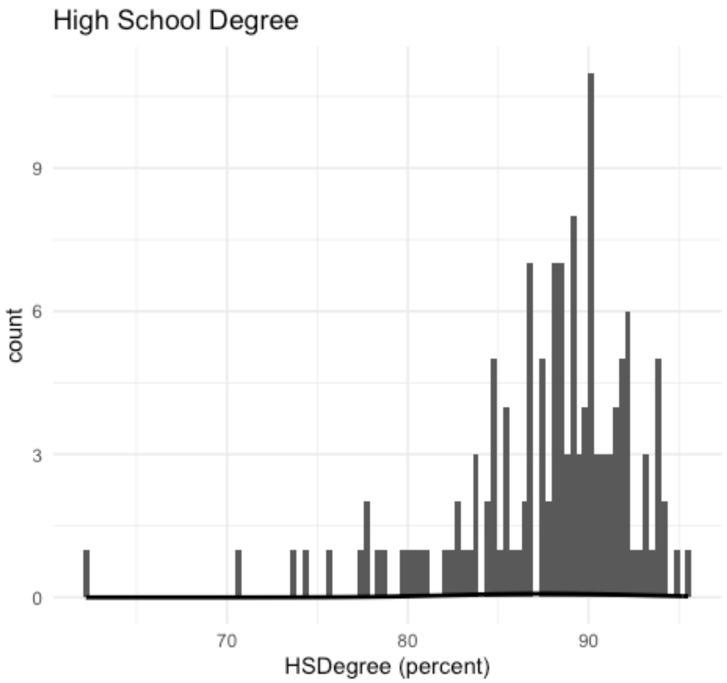


```
> str(acs_14_1yr_s0201)
spec_tbl_df[,8] [136 × 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Id                   : chr [1:136] "0500000US01073" "0500000US04013" "0500000US04019" "0500000US06001" ...
 $ Id2                  : num [1:136] 1073 4013 4019 6001 6013 ...
 $ Geography            : chr [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()
  .. )
```

   b. nrow(acs_14_1yr_s0201)
      i. [1] 136
   c. ncol(acs_14_1yr_s0201)
      i. [1] 8
2. Create a Histogram of the HSDegree variable using the ggplot2 package.
   a. Set a bin size for the Histogram.
   b. Include a Title and appropriate X/Y axis labels on your Histogram Plot.
      i. hist.HSDegree <- ggplot(acs_14_1yr_s0201, aes(HSDegree)) + geom_histogram(bins = 100) + ggtitle("High School Degree") + xlab("HSDegree (percent)")

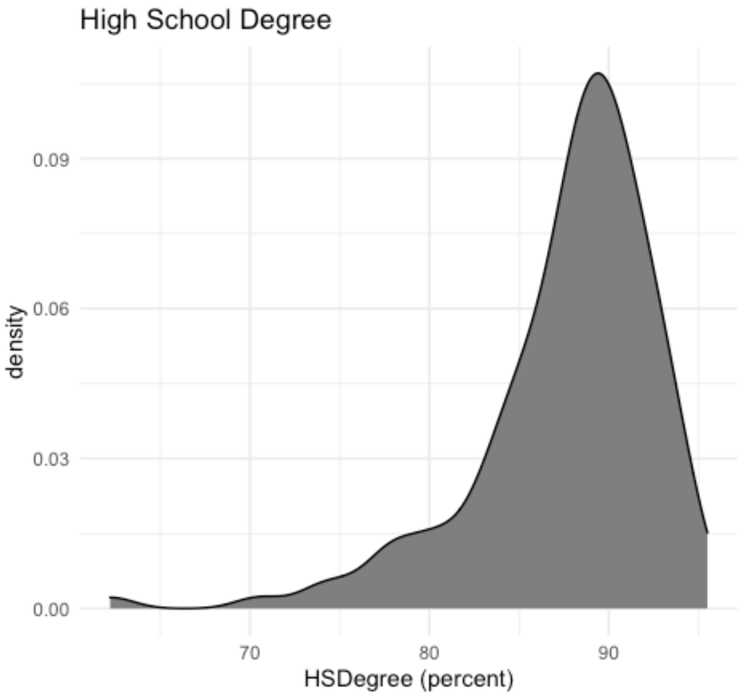


3. Answer the following questions based on the Histogram produced:
   a. Based on what you see in this histogram, is the data distribution unimodal?
      i. answer: Yes from what is seen on this Histogram we see a single peak which justifies a unimodal distribution
   b. Is it approximately symmetrical?
      i. answer: No this is not approximately symmetrical as if a line is drawn at the peak it would not mirror the right and left side
   c. Is it approximately bell-shaped?
      i. Yes it can be considered bell-shaped as it has a prominent mound but is skewed to the left and is unimodal
   d. Is it approximately normal?
      i. answer: No this histogram is left-skewed
   e. If not normal, is the distribution skewed? If so, in which direction?
      i. answer: Yes the distribution is skewed to the left as this indicates the peak is to the right of the center.
4. Include a normal curve to the Histogram that you plotted.
   a. hist.HSDegree + stat_function(fun = dnorm, args = list(mean = mean(acs_14_1yr_s0201$HSDegree, na.rm = TRUE), sd = sd(acs_14_1yr_s0201$HSDegree, na.rm = TRUE)), colour = "black", size = 1)



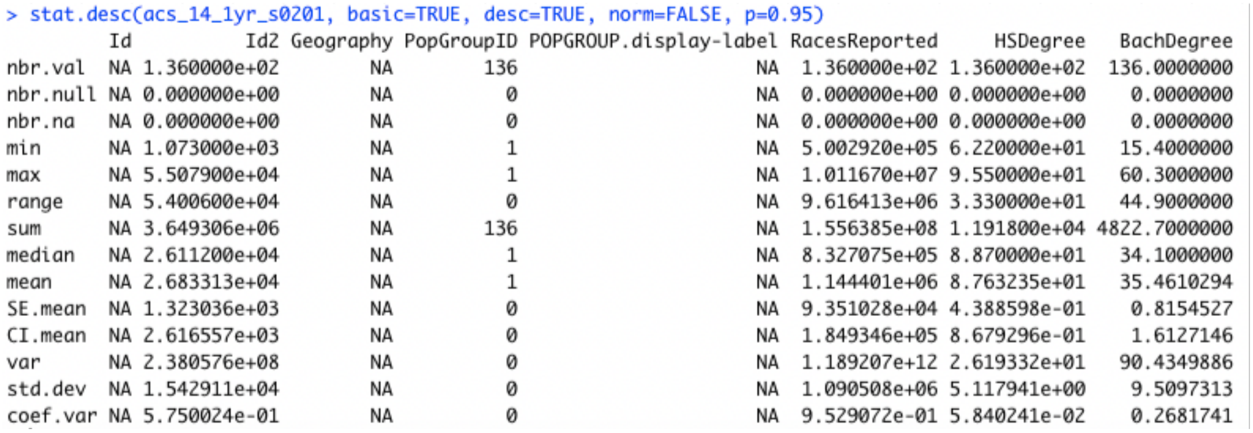
b. Explain whether a normal distribution can accurately be used as a model for this data
   i. answer: No a normal distribution could not be used to model this data accurately as our data is skewes to the left

5. Create a Probability Plot of the HSDegree variable.
   a. ggplot(acs_14_1yr_s0201, aes(HSDegree)) + geom_density(aes(HSDegree), fill="grey50") + ggtitle("High School Degree") + xlab("HSDegree (percent)")



6. Answer the following questions based on the Probability Plot:
   a. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.
      i. Answer: No the distribution is not normal as it is skewed to the left and a normal plot indicates that this probability plot is not normally distributed.
   b. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.
      i. Answer: Yes this distribution is skewed to the left due to the peak being to the right of the center

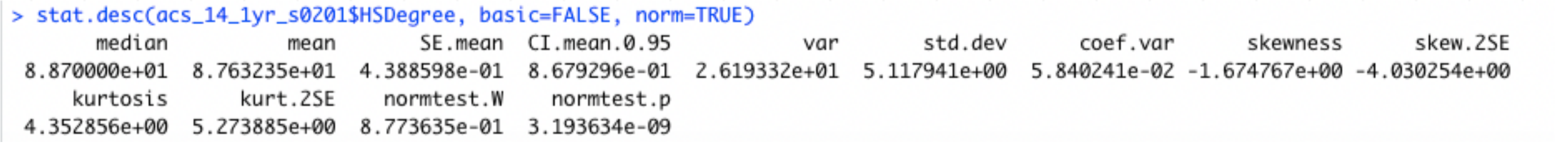
7. Now that you have looked at this data visually for normality, you will now quantify normality with numbers using the stat.desc() function. Include a screen capture of the results produced.
   a. stat.desc(acs_14_1yr_s0201, basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)



```
> stat.desc(acs_14_1yr_s0201, basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)
         Id          Id2 Geography PopGroupID POPGROUP.display-label RacesReported     HSDegree   BachDegree
nbr.val  NA 1.360000e+02        NA        136                     NA  1.360000e+02 1.360000e+02  136.0000000
nbr.null NA 0.000000e+00        NA          0                     NA  0.000000e+00 0.000000e+00    0.0000000
nbr.na   NA 0.000000e+00        NA          0                     NA  0.000000e+00 0.000000e+00    0.0000000
min      NA 1.073000e+03        NA          1                     NA  5.002920e+05 6.220000e+01   15.4000000
max      NA 5.507900e+04        NA          1                     NA  1.011670e+07 9.550000e+01   60.3000000
range    NA 5.400600e+04        NA          0                     NA  9.616413e+06 3.330000e+01   44.9000000
sum      NA 3.649306e+06        NA        136                     NA  1.556385e+08 1.191800e+04 4822.7000000
median   NA 2.611200e+04        NA          1                     NA  8.327075e+05 8.870000e+01   34.1000000
mean     NA 2.683313e+04        NA          1                     NA  1.144401e+06 8.763235e+01   35.4610294
SE.mean  NA 1.323036e+03        NA          0                     NA  9.351028e+04 4.388598e-01    0.8154527
CI.mean  NA 2.616557e+03        NA          0                     NA  1.849346e+05 8.679296e-01    1.6127146
var      NA 2.380576e+08        NA          0                     NA  1.189207e+12 2.619332e+01   90.4349886
std.dev  NA 1.542911e+04        NA          0                     NA  1.090508e+06 5.117941e+00    9.5097313
coef.var NA 5.750024e-01        NA          0                     NA  9.529072e-01 5.840241e-02    0.2681741
```

8.

   a. stat.desc(acs_14_1yr_s0201$HSDegree, basic=FALSE, norm=TRUE)



```
> stat.desc(acs_14_1yr_s0201$HSDegree, basic=FALSE, 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
```

9. 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?
   a. Answer: Seen in the results produced for the skew of the HSDegree we see that the skewness is at -1.674767e+00 and skew.2SE is -4.030254e+00 that shows the measurement of the asymmetry of the distribution of the HSDegree data set that is negatively skewed and with the majority of data values greater than mean.
   b. Answer: Seen in the results produced for the kurtosis of the HSDegree we see that it is 4.352856e+00 and kurt. 2SE is 5.273885e+00 that measures the sharpness of the peak in the data distribution as the HSDegree is Leptokurtic as it is greater than 3 so it shows a sharp peak on the graph.
   c. Answer: The Z-Score is calculated by subtracting the mean of the distribution and then divide by the standard deviation of the distribution such as the standard error. This is calculated by -1.674767e+00 - 0 / 5.117941e+00 = -0.327234 meaning that the z-score is negative, so its' corresponding raw score is below the mean.

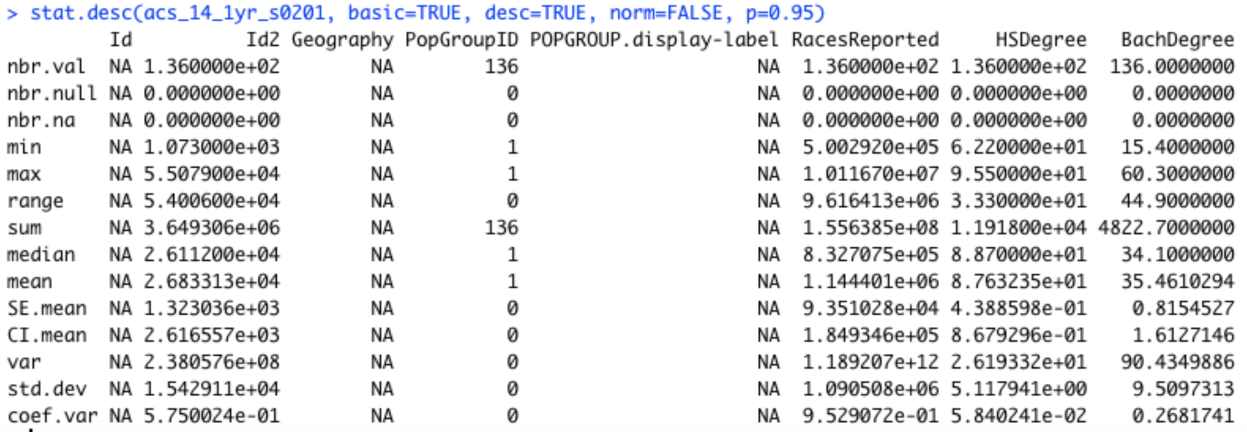
```
# Assignment: ASSIGNMENT 3
# Name: Koby-Hercsky, Theodore
# Date: 2021-03-31

## Load the ggplot2 package
library(ggplot2)
theme_set(theme_minimal())

## Set the working directory to the root of your DSC 520 directory
setwd("/home/downloads/DSC520/dsc520")

## Load the `data/r4ds/heights.csv` to
heights_df <- read_csv("r4ds/heights.csv")

# https://ggplot2.tidyverse.org/reference/geom_point.html
## Using `geom_point()` create three scatterplots for
## `height` vs. `earn`
ggplot(heights_df, aes(x=height, y=earn)) + geom_point(aes(color=sex))
## `age` vs. `earn`
ggplot(heights_df, aes(x=age, y=earn)) + geom_point(aes(color=race))
## `ed` vs. `earn`
ggplot(heights_df, aes(x=ed, y=earn)) + geom_point(aes(alpha=age))

## Re-create the three scatterplots and add a regression trend line using
## the `geom_smooth()` function
## `height` vs. `earn`
ggplot(heights_df, aes(x=height, y=earn)) + geom_point(aes(color=sex)) + geom_smooth(orientation = "x")
## `age` vs. `earn`
ggplot(heights_df, aes(x=age, y=earn)) + geom_point(aes(color=race)) + geom_smooth(span = 0.5)
## `ed` vs. `earn`
ggplot(heights_df, aes(x=ed, y=earn)) + geom_point(aes(alpha=age)) + geom_smooth(se = FALSE, method = lm)

## Create a scatterplot of `height`` vs. `earn`.  Use `sex` as the `col` (color) attribute
ggplot(heights_df, aes(x=height, y=earn, col=sex)) + geom_point(aes(alpha=age))

## Using `ggtitle()`, `xlab()`, and `ylab()` to add a title, x label, and y label to the previous plot
## Title: Height vs. Earnings
## X label: Height (Inches)
## Y Label: Earnings (Dollars)
ggplot(heights_df, aes(x=height, y=earn, col=sex)) + geom_point(aes(alpha=age)) + ggtitle("Height vs. Earnings") +
xlab("Height (Inches)") + ylab("Earnings (Dollars)")

# https://ggplot2.tidyverse.org/reference/geom_histogram.html
## Create a histogram of the `earn` variable using `geom_histogram()`
ggplot(heights_df, aes(earn)) + geom_histogram(binwidth = 50000)

## Create a histogram of the `earn` variable using `geom_histogram()`
## Use 10 bins
ggplot(heights_df, aes(earn)) + geom_histogram(bins = 10)

# https://ggplot2.tidyverse.org/reference/geom_density.html
## Create a kernel density plot of `earn` using `geom_density()`
ggplot(heights_df, aes(earn)) + geom_density(kernel = "gaussian")
```

```
# Assignment: 2014 American Community Survey Data
# Name: Koby-Hercsky, Theodore
# Date: 2021-03-31

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

#Please provide the output from the following functions: str(); nrow(); ncol()
str(acs_14_1yr_s0201)
nrow(acs_14_1yr_s0201)
# [1] 136
ncol(acs_14_1yr_s0201)
# [1] 8

# Create a Histogram of the HSDegree variable using the ggplot2 package.
# Set a bin size for the Histogram.
# Include a Title and appropriate X/Y axis labels on your Histogram Plot.
hist.HSDegree <- ggplot(acs_14_1yr_s0201, aes(HSDegree)) + geom_histogram(bins = 100) + ggtitle("High School Degree") + xlab("HSDegree (percent)")

#Answer the following questions based on the Histogram produced:
#  Based on what you see in this histogram, is the data distribution unimodal?
# answer: Yes from what is seen on this Histogram we see a single peak which justifies a unimodal distribution
#  Is it approximately symmetrical?
# answer: No this is not approximately symmetrical as if a line is drawn at the peak it would not mirror the right and left side
#  Is it approximately bell-shaped?
# answer: Yes it can be considered bell-shaped as it has a prominent mound but is skewed to the left and is unimodal
#  Is it approximately normal?
# answer: No this histogram is left-skewed
#  If not normal, is the distribution skewed? If so, in which direction?
# answer: Yes the distribution is skewed to the left as this indicates the peak is to the right of the center.
#  Include a normal curve to the Histogram that you plotted.
hist.HSDegree + stat_function(fun = dnorm, args = list(mean = mean(acs_14_1yr_s0201$HSDegree, na.rm = TRUE), sd = sd(acs_14_1yr_s0201$HSDegree, na.rm = TRUE)), colour = "black", size = 1)
#Explain whether a normal distribution can accurately be used as a model for this data.
# answer: No a normal distribution could not be used to model this data accurately as our data is skewes to the left.

# Create a Probability Plot of the HSDegree variable.
ggplot(acs_14_1yr_s0201, aes(HSDegree)) + geom_density(aes(HSDegree), fill="grey50") + ggtitle("High School Degree") + xlab("HSDegree (percent)")
# 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.
# Answer: No the distribution is not normal as it is skewed to the left and a normal plot indicates that this probability plot is not normally distributed.
#If not normal, is the distribution skewed? If so, in which direction? Explain how you know.
# Answer: Yes this distribution is skewed to the left due to the peak being to the right of the center

# 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(acs_14_1yr_s0201, basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)
stat.desc(acs_14_1yr_s0201$HSDegree, basic=FALSE, norm=TRUE)
# 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?
# Answer: Seen in the results produced for the skew of the HSDegree we see that the skewness is at -1.674767e+00 and skew.2SE is -4.030254e+00 that shows the measurement of the asymmetry of the distribution of the HSDegree data set that is negatively skewed and with the majority of data values greater than mean.
# Answer: Seen in the results produced for the kurtosis of the HSDegree we see that it is 4.352856e+00 and kurt.2SE is 5.273885e+00 that measures the sharpness of the peak in the data distribution as the HSDegree is Leptokurtic as it is greater than 3 so it shows a sharp peak on the graph.
# Answer: The Z-Score is calculated by subtracting the mean of the distribution and then divide by the standard deviation of the distribution such as the standard error. This is calculated by -1.674767e+00 - 0 / 5.117941e+00 = -0.327234 meaning that the z-score is negative so its' corresponding raw score is below the mean.
```

Screen Shots from 2014 American Community Survey Data Assignment

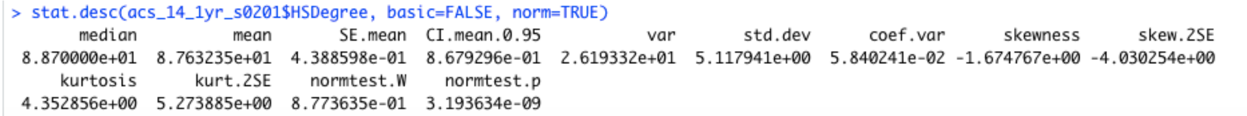
# Now that you have looked at this data visually for normality, you will now quantify normality with numbers using the stat.desc() function. Include a screen capture of the results produced.

stat.desc(acs_14_1yr_s0201, basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)



```
> stat.desc(acs_14_1yr_s0201, basic=TRUE, desc=TRUE, norm=FALSE, p=0.95)
         Id          Id2 Geography PopGroupID POPGROUP.display-label RacesReported     HSDegree   BachDegree
nbr.val  NA 1.360000e+02        NA        136                     NA  1.360000e+02 1.360000e+02  136.0000000
nbr.null NA 0.000000e+00        NA          0                     NA  0.000000e+00 0.000000e+00    0.0000000
nbr.na   NA 0.000000e+00        NA          0                     NA  0.000000e+00 0.000000e+00    0.0000000
min      NA 1.073000e+03        NA          1                     NA  5.002920e+05 6.220000e+01   15.4000000
max      NA 5.507900e+04        NA          1                     NA  1.011670e+07 9.550000e+01   60.3000000
range    NA 5.400600e+04        NA          0                     NA  9.616413e+06 3.330000e+01   44.9000000
sum      NA 3.649306e+06        NA        136                     NA  1.556385e+08 1.191800e+04 4822.7000000
median   NA 2.611200e+04        NA          1                     NA  8.327075e+05 8.870000e+01   34.1000000
mean     NA 2.683313e+04        NA          1                     NA  1.144401e+06 8.763235e+01   35.4610294
SE.mean  NA 1.323036e+03        NA          0                     NA  9.351028e+04 4.388598e-01    0.8154527
CI.mean  NA 2.616557e+03        NA          0                     NA  1.849346e+05 8.679296e-01    1.6127146
var      NA 2.380576e+08        NA          0                     NA  1.189207e+12 2.619332e+01   90.4349886
std.dev  NA 1.542911e+04        NA          0                     NA  1.090508e+06 5.117941e+00    9.5097313
coef.var NA 5.750024e-01        NA          0                     NA  9.529072e-01 5.840241e-02    0.2681741
```

stat.desc(acs_14_1yr_s0201$HSDegree, basic=FALSE, norm=TRUE)



```
> stat.desc(acs_14_1yr_s0201$HSDegree, basic=FALSE, 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 
```