Ggplot2 in R programming

Kesav Adithya Venkidusamy

Bellevue university - Master of Science in Data Science

Course Name: DSC520-T301 Statistics for Data Science (2221-1)

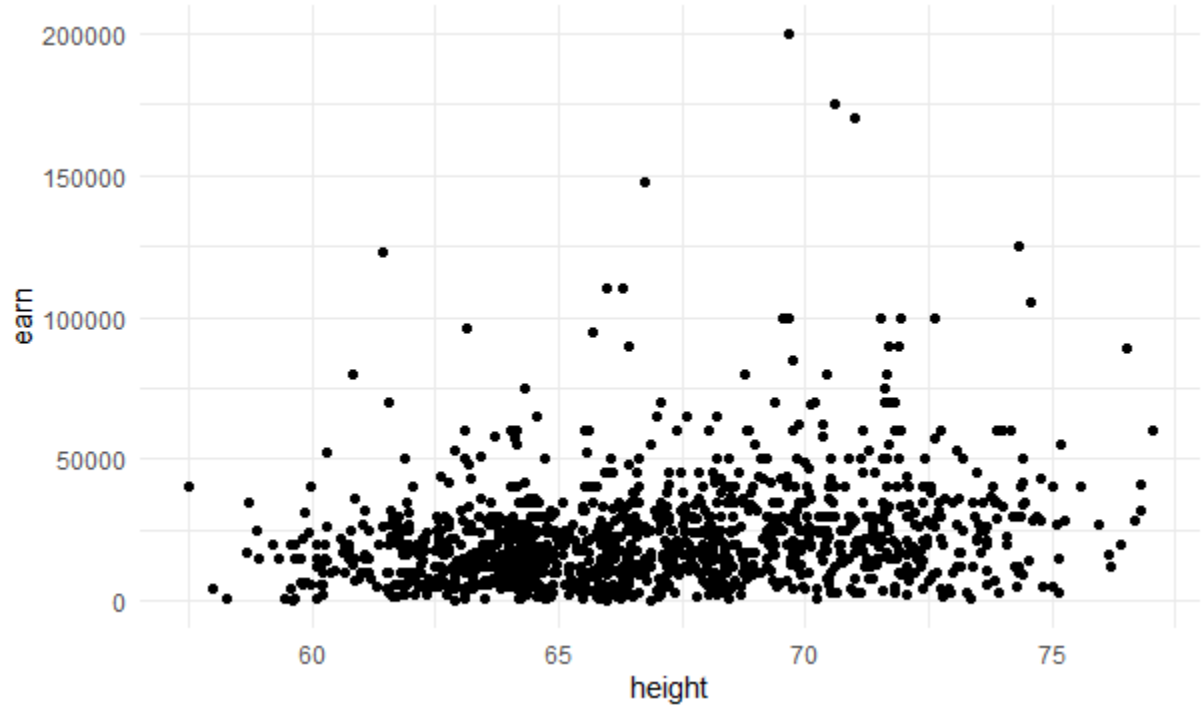
Assignment: Week 3.2 Assignment

Instructor: Dr Richard Bushart

Due Date: 09/19/2021

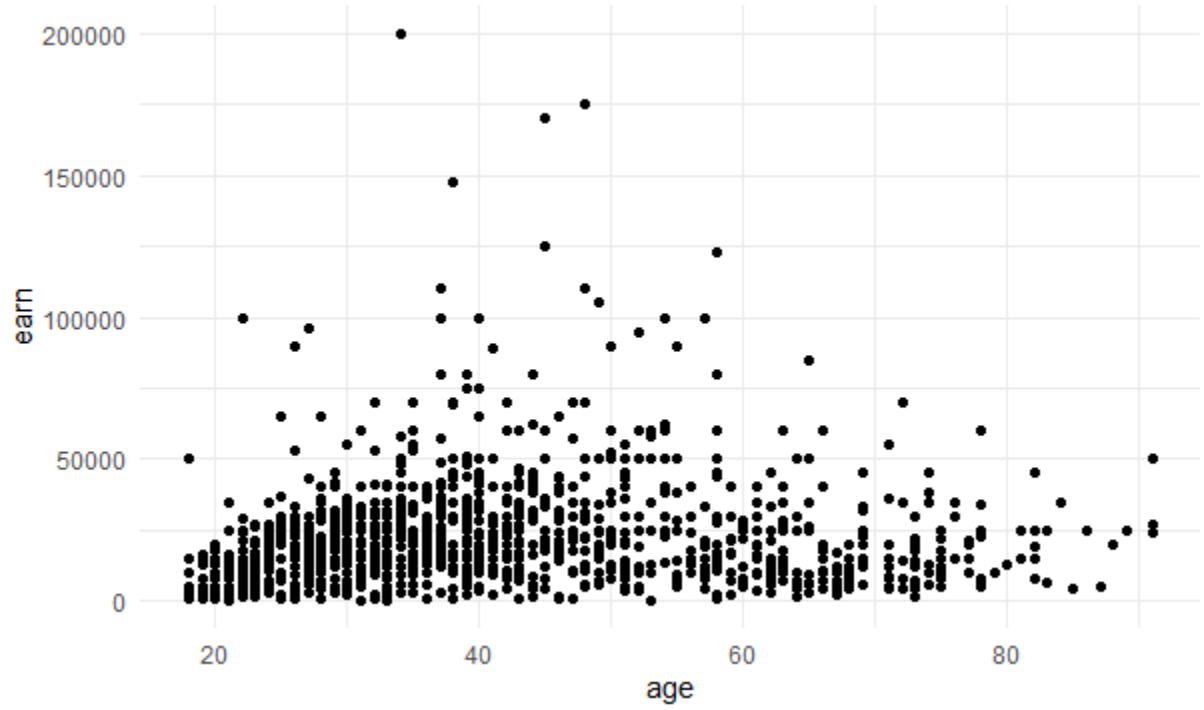
Assignment 1

```
## Load the ggplot2 package

library(ggplot2)

theme_set(theme_minimal())



## Set the working directory to the root of your DSC 520 directory

setwd("E:/Personal/Bellevue University/Course/github/dsc520")



## Load the `data/r4ds/heights.csv` to

heights_df <- read.csv("data/r4ds/heights.csv")



# https://ggplot2.tidyverse.org/reference/geom_point.html

## Using `geom_point()` create three scatterplots for

## `height` vs. `earn`
```



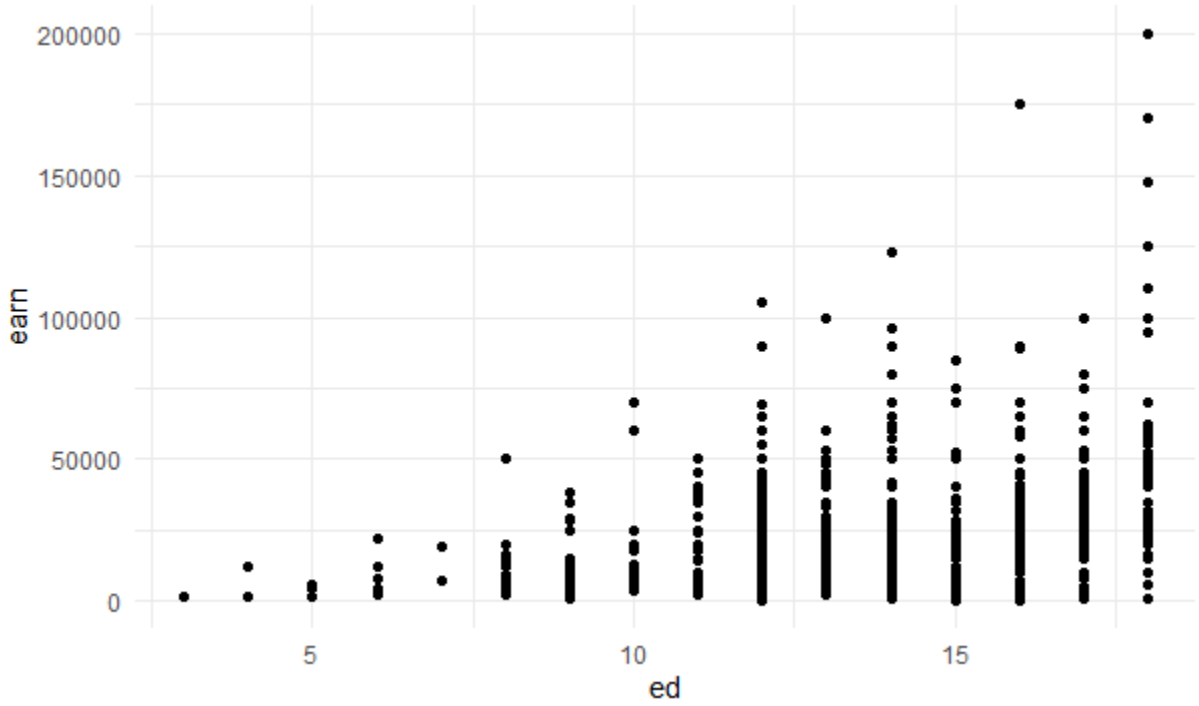
`age` vs. `earn`

ggplot(heights_df, aes(x=age, y=earn)) + geom_point()



`ed` vs. `earn`

```
ggplot(heights_df, aes(x=ed, y=earn)) + geom_point()
```



Re-create the three scatterplots and add a regression trend line using

the `geom_smooth()` function

`height` vs. `earn`

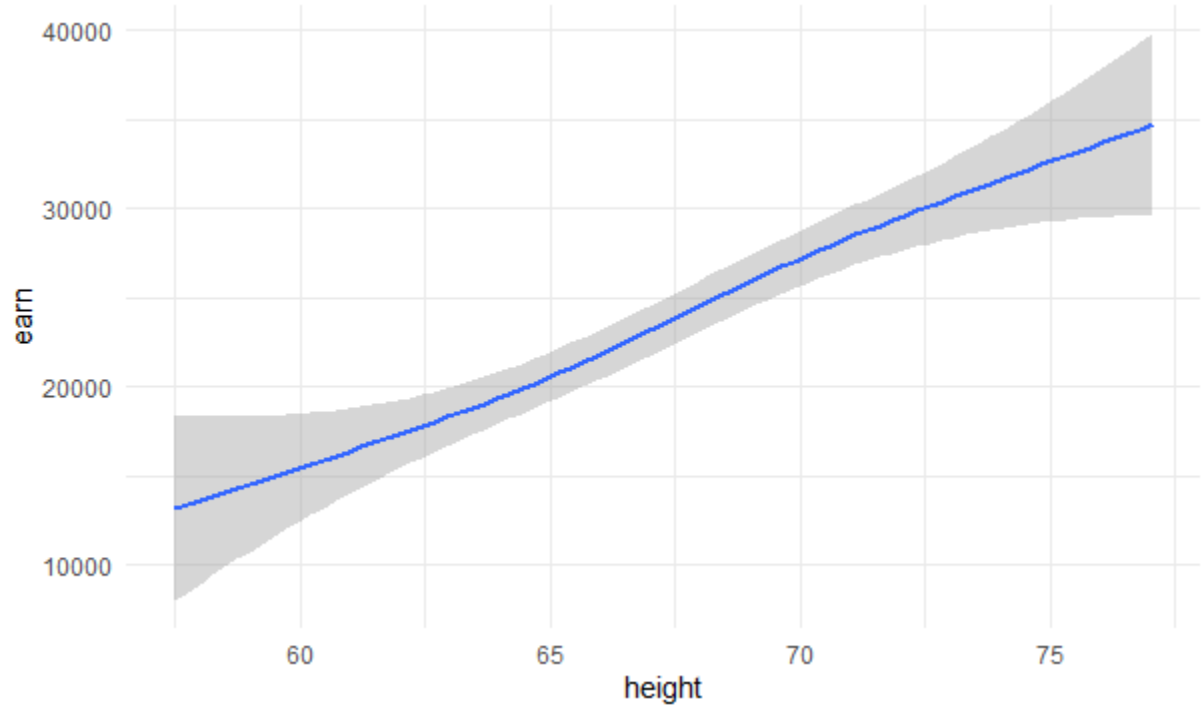
```
ggplot(heights_df, aes(x=height, y=earn)) + geom_smooth(method='gam', formula = y ~ s(x,
bs="cs"))
```

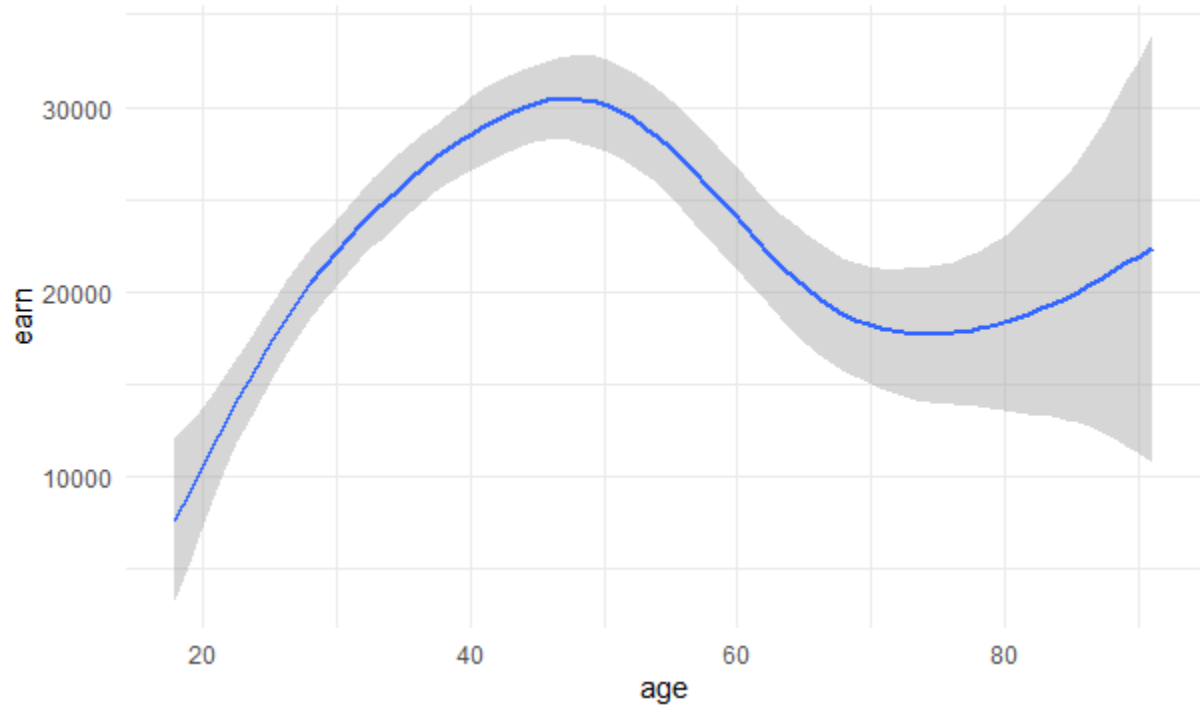
`age` vs. `earn`

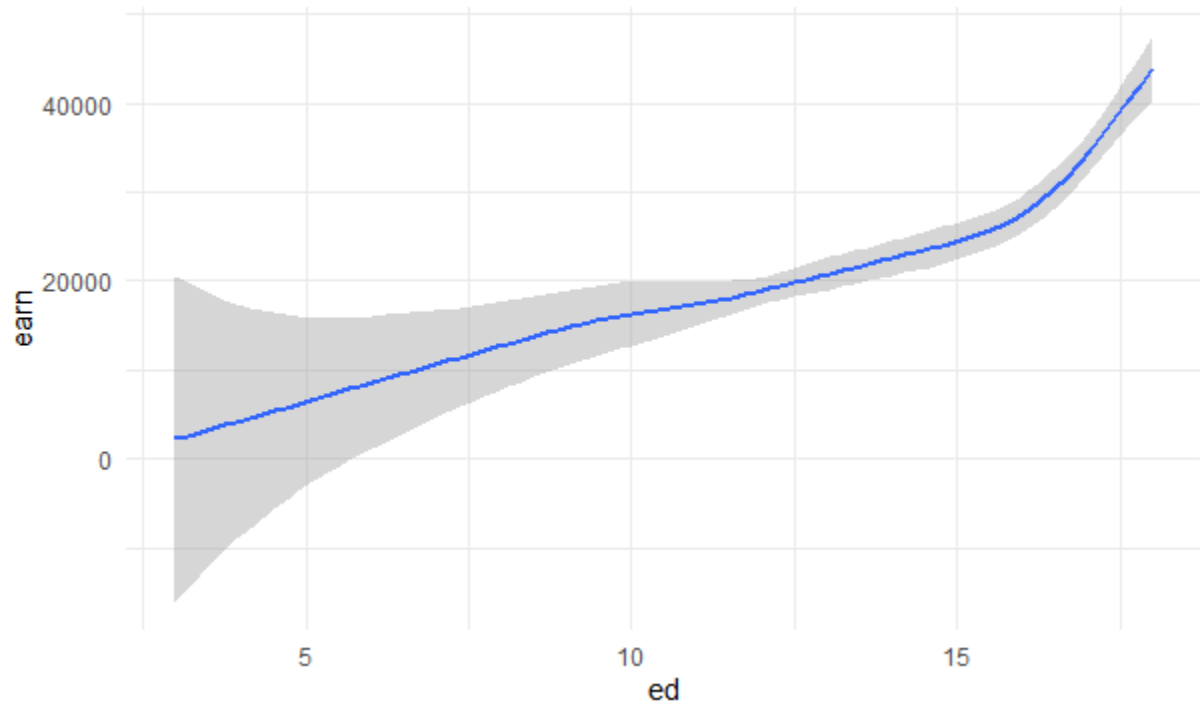
```
ggplot(heights_df, aes(x=age, y=earn))  + geom_smooth(method='gam', formula = y ~ s(x,
bs="cs"))
```

`ed` vs. `earn`

```
ggplot(heights_df, aes(x=ed, y=earn))  + geom_smooth(method='gam', formula = y ~ s(x,
bs="cs"))
```

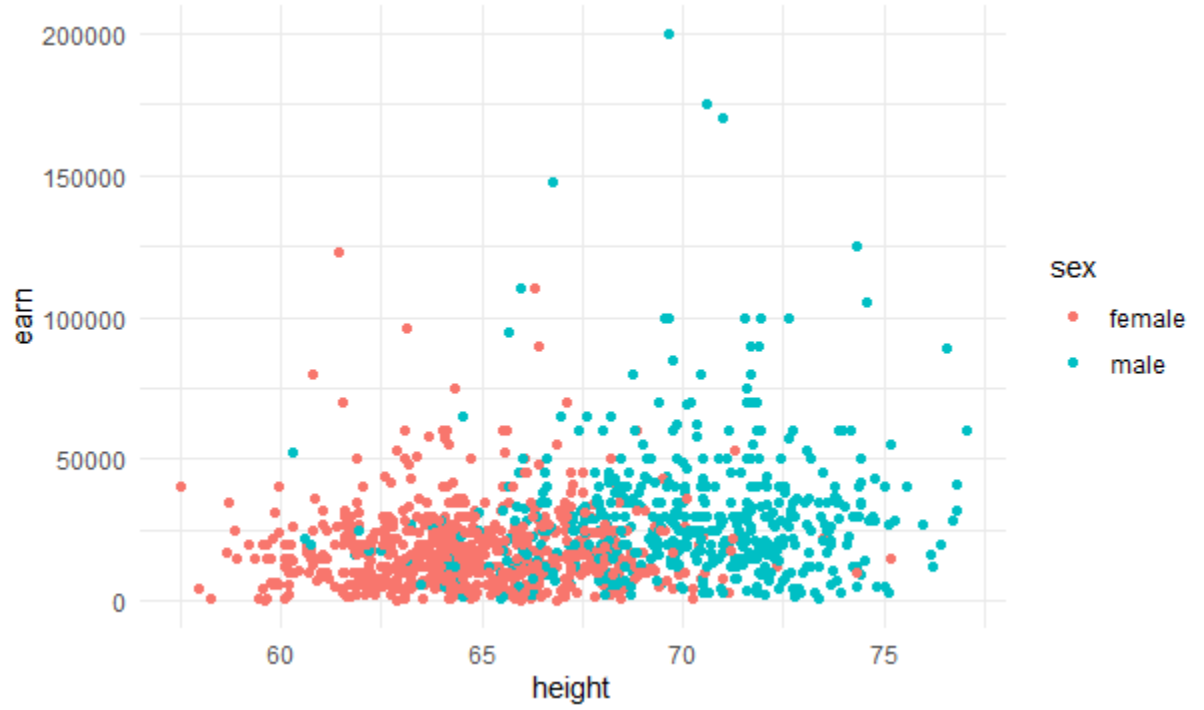






Create a scatterplot of `height`` vs. `earn`. Use `sex` as the `col` (color) attribute

ggplot(heights_df, aes(x=height, y=earn, col=sex)) + geom_point()



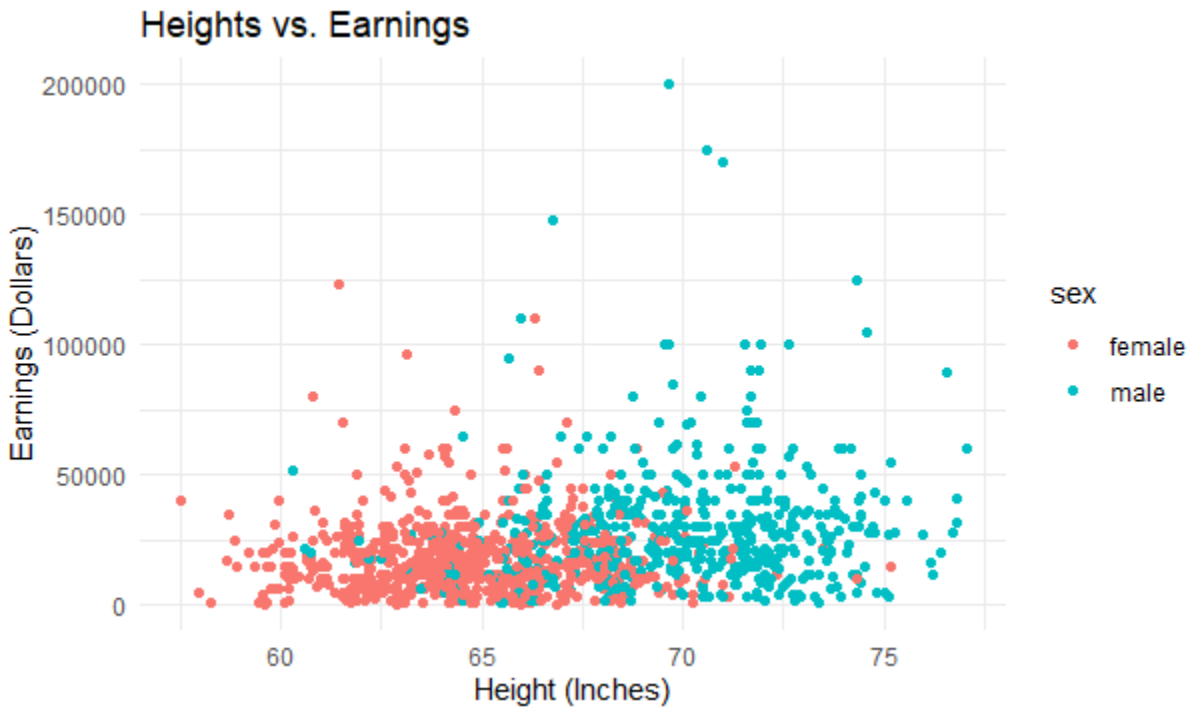
Using `ggtitle()`, `xlab()`, and `ylab()` to add a title, x label, and y label to the previous plot

Title: Height vs. Earnings

X label: Height (Inches)

Y Label: Earnings (Dollars)

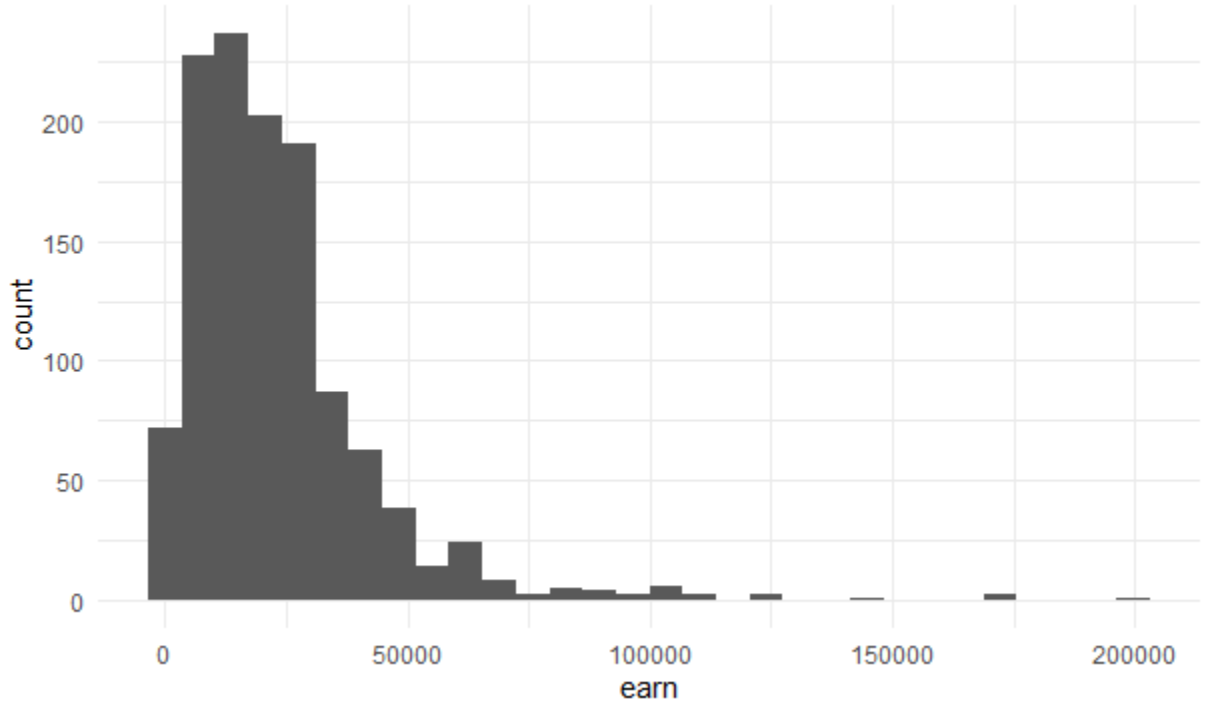
```
ggplot(heights_df, aes(x=height, y=earn, col=sex)) + geom_point() + ggtitle("Heights vs. Earnings") + xlab("Height (Inches)") + ylab("Earnings (Dollars)")
```



https://ggplot2.tidyverse.org/reference/geom_histogram.html

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

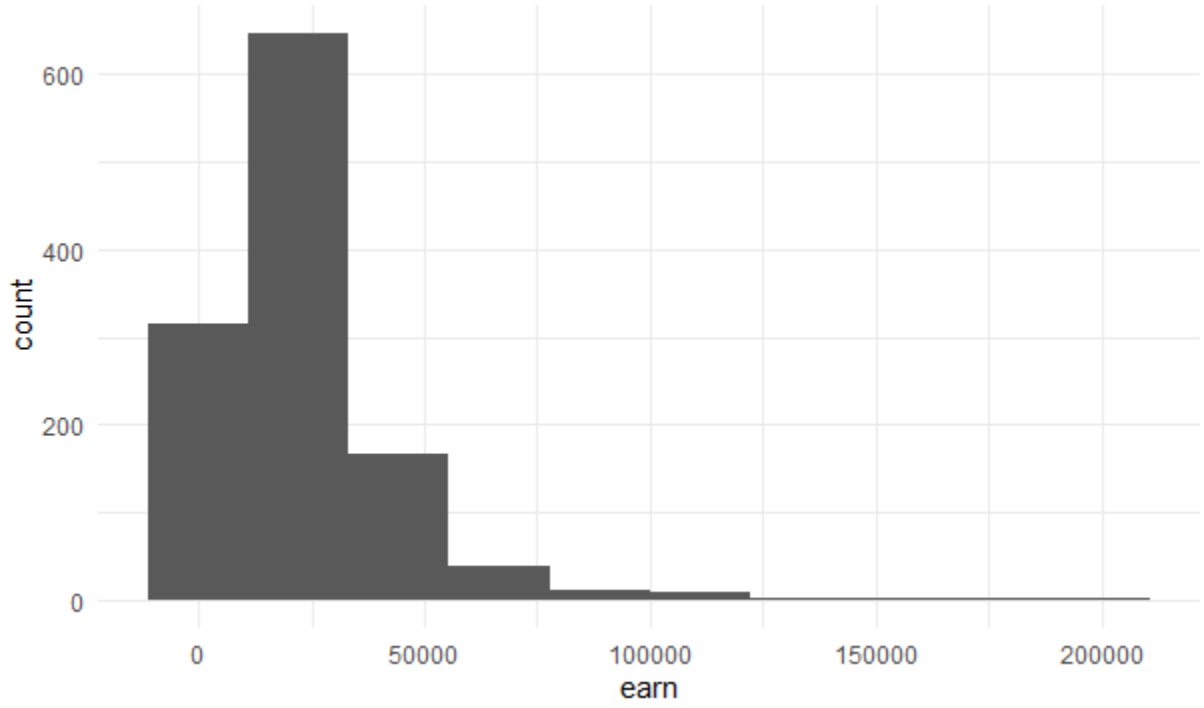
```
ggplot(heights_df, aes(earn)) + geom_histogram()
```



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

Use 10 bins

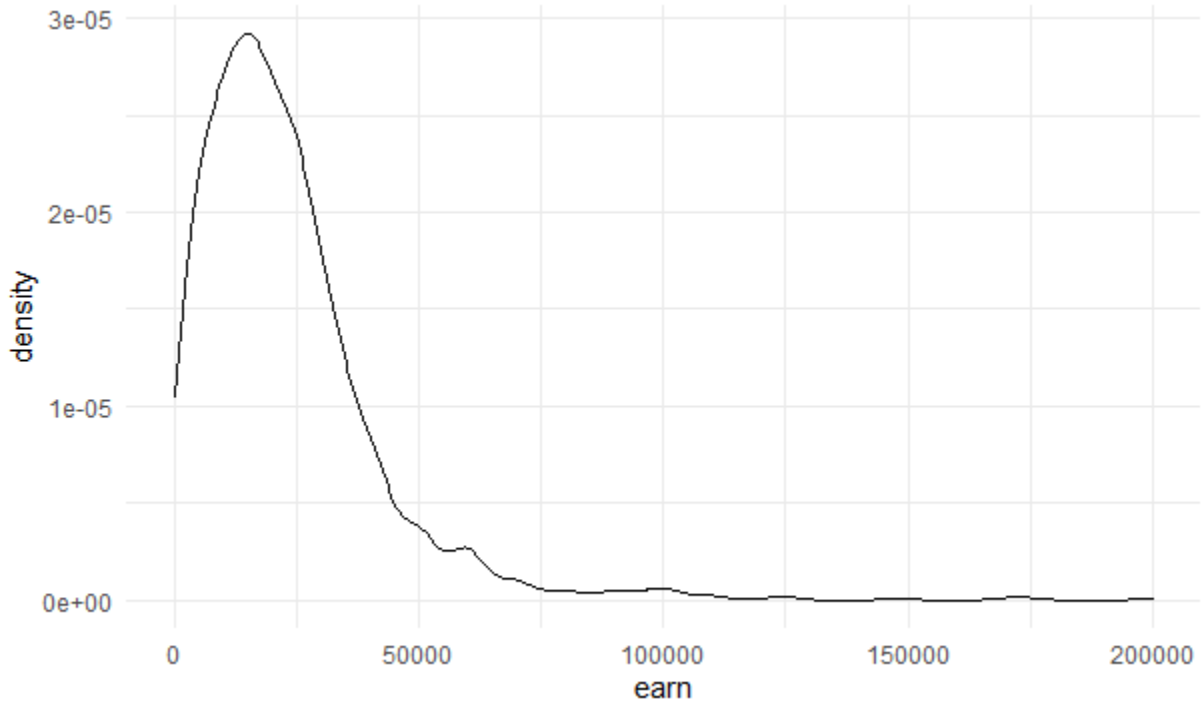
```
ggplot(heights_df, aes(earn)) + geom_histogram(bins=10)
```



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()



Assignment 2

Set the working directory to the root of your DSC 520 directory

> setwd("E:/Personal/Bellevue University/Course/github/dsc520")

> ## Load the `data/r4ds/heights.csv` to

> amer_survey <- read.csv("data/acs-14-1yr-s0201.csv")

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

> typeof(amer_survey)

[1] "list"

> attributes(amer_survey)

```
$names

[1] "Id"                  "Id2"                 "Geography"

[4] "PopGroupID"          "POPGROUP.display.label" "RacesReported"

[7] "HSDegree"            "BachDegree"


$class

[1] "data.frame"


$row.names

  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19

 [20]  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38

 [39]  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57

 [58]  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76

 [77]  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95

 [96]  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114

[115] 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133

[134] 134 135 136
```

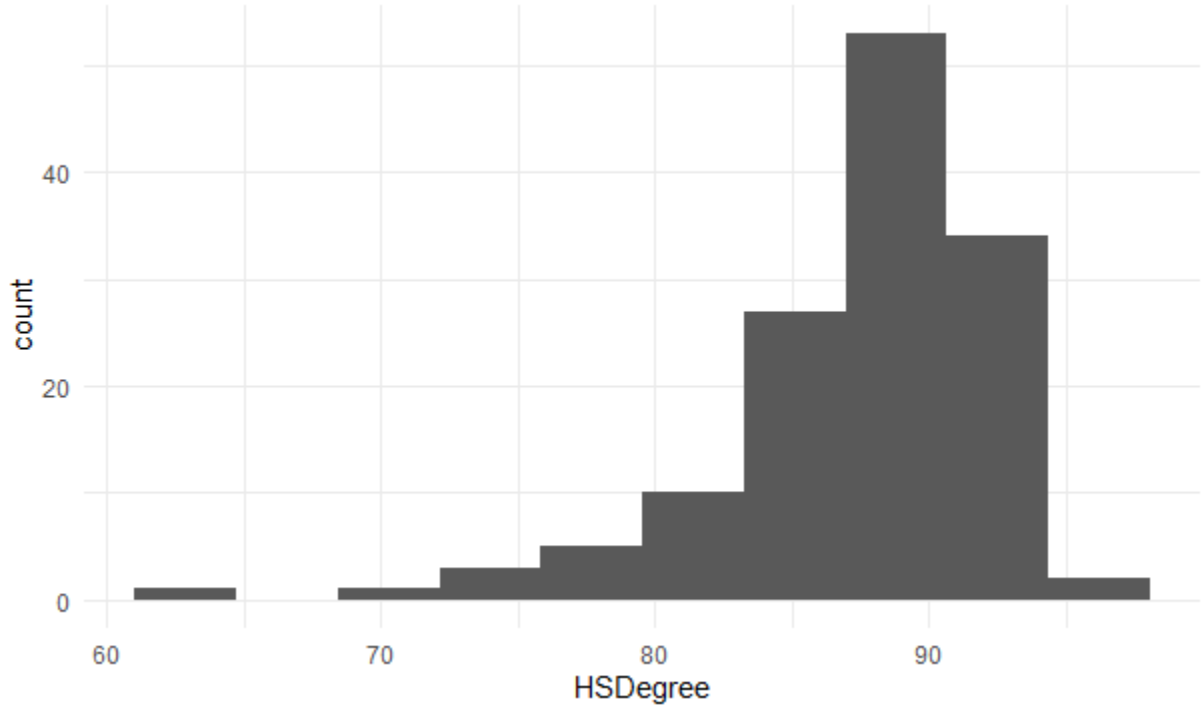
```
> ##Please provide the output from the following functions: str(); nrow(); ncol()

> str(amer_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 ...

> nrow(amer_survey)

[1] 136

> ncol(amer_survey)

[1] 8
```

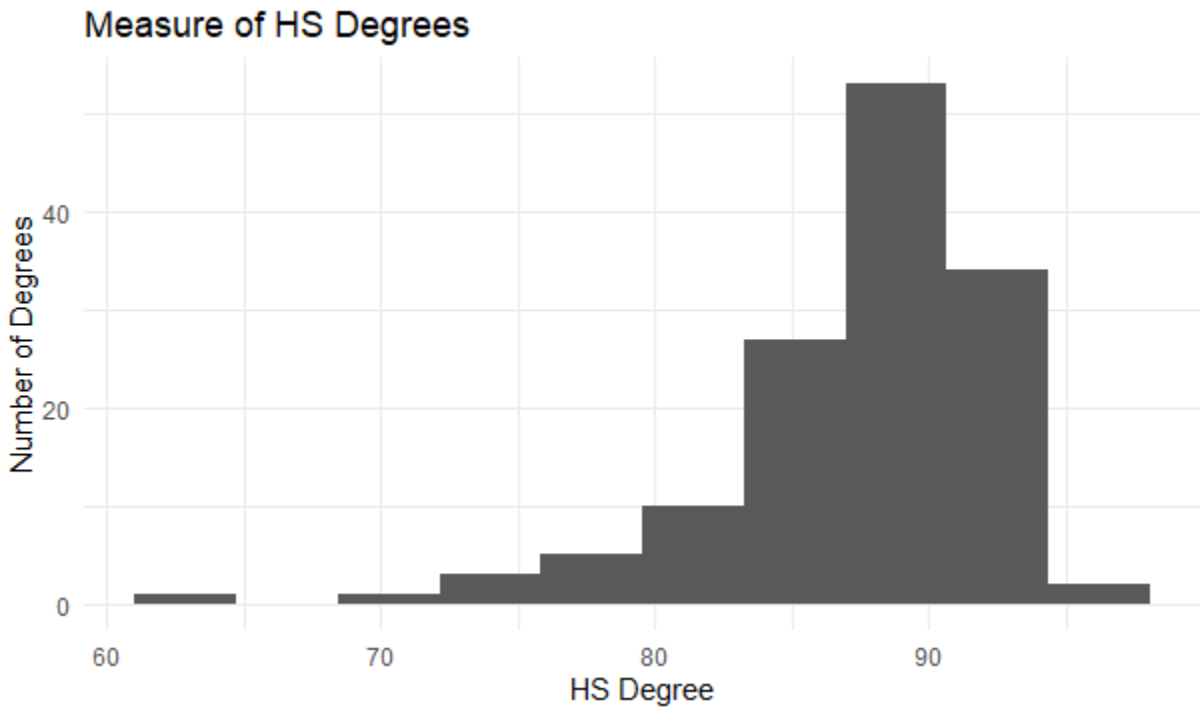
##Create a Histogram of the HSDegree variable using the ggplot2 package.

> ##Set a bin size for the Histogram.

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

> ggplot(amer_survey, aes(HSDegree)) + geom_histogram(bins=10)





Answer the following questions based on the Histogram produced:

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

Yes, the data distribution is unimodal as it has only one hump.

Is it approximately symmetrical?

2 halves present in the histogram is not mirror image of one another. So, the histogram is asymmetric.

Is it approximately bell-shaped?

Yes. It is approximately bell-shaped.

Is it approximately normal?

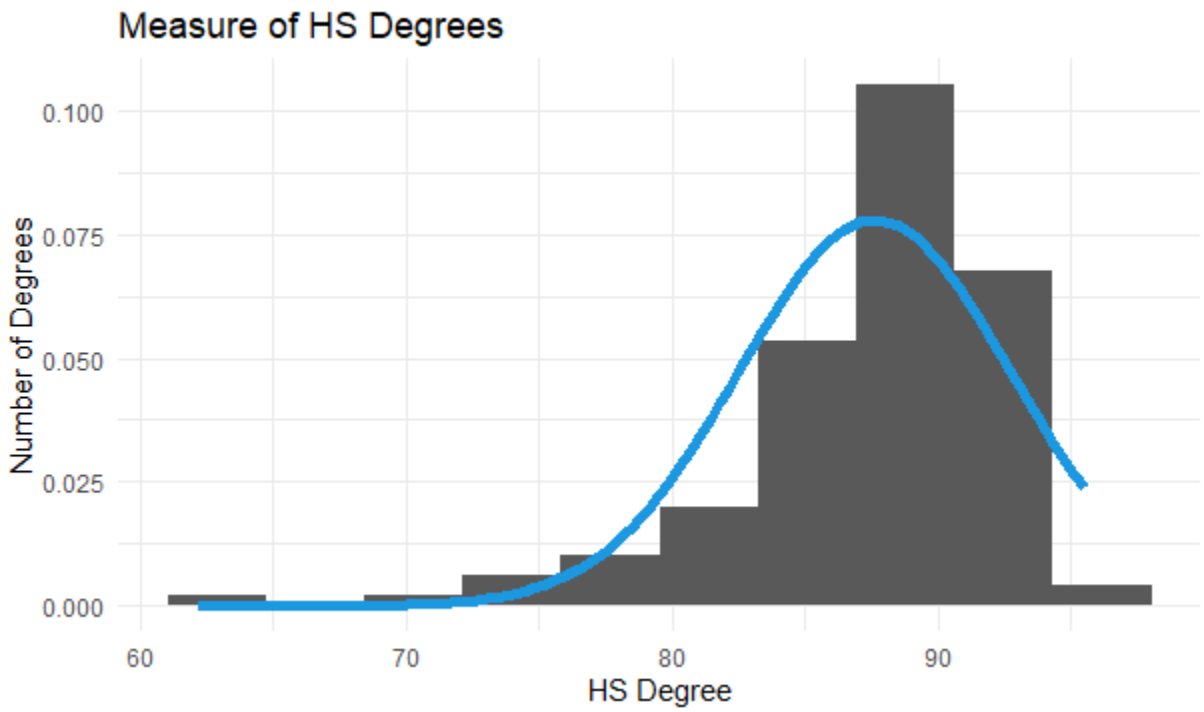
The distribution is not normal. Though the graph is bell-shaped, it follows asymmetric about the mean. So, it is not normal.

If not normal, is the distribution skewed? If so, in which direction?

The graph is skewed left as left side is longer than its right.

Include a normal curve to the Histogram that you plotted.

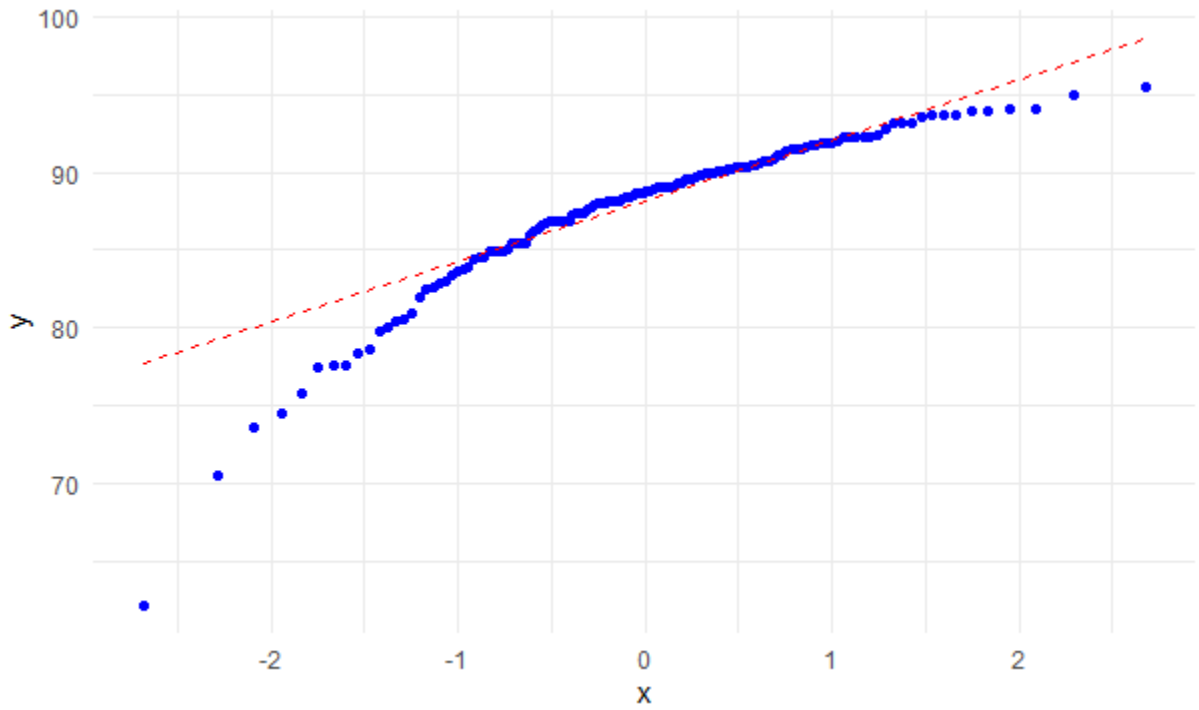
ggplot(amer_survey, aes(HSDegree)) + geom_histogram(bins=10, aes(y = ..density..)) + stat_function(fun = dnorm, args = list(mean = mean(amer_survey$HSDegree), sd = sd(amer_survey$HSDegree)), col = "#1b98e0", size = 2)



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

No, normal distribution cannot be used as a model for this data. Though the graph is bell shaped, it is skewed left compared to right. So, it is asymmetric in nature.

Create a Probability Plot of the HSDegree variable.



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.

The graph is not normally distributed. A straight, diagonal line means that you have normally distributed data. But here it is not straight diagonal line.

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

The graph is having skewed distribution as plotted points bend down and to the right of the normal line that indicates a long tail to the left.

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(amer_survey$HSDegree, basic = FALSE, norm = TRUE)
      median          mean       SE.mean  CI.mean.0.95           var       std.dev
8.870000e+01  8.763235e+01  4.388598e-01  8.679296e-01  2.619332e+01  5.117941e+00
    coef.var      skewness      skew.2SE      kurtosis      kurt.2SE    normtest.W
5.840241e-02 -1.674767e+00 -4.030254e+00  4.352856e+00  5.273885e+00  8.773635e-01
  normtest.p
```

3.193634e-09

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?

The first thing you usually notice about a distribution's shape is whether it has one mode (peak) or more than one. Since the given chart has only one peak, it unimodal (has just one peak).

The next thing you notice is whether it's symmetric or skewed to one side. If the bulk of the data is at the left and the right tail is longer, we say that the distribution is skewed right or positively skewed; if the peak is toward the right and the left tail is longer, we say that the distribution is skewed left or negatively skewed. Here, the data is skewed left or negatively skewed as left tail is longer compared to right.

The other common measure of shape is called the kurtosis. As skewness involves the third moment of the distribution, kurtosis involves the fourth moment. The outliers in a sample, therefore, have even more effect on the kurtosis than they do on the skewness and in a symmetric distribution both tails increase the kurtosis, unlike skewness where they offset each other.

Z-score is a statistical measure that tells you how far is a data point from the rest of the dataset. In a more technical term, Z-score tells how many standard deviations away a given observation is from the mean.

Z= (value – mean)/ (Standard Deviation)

We have taken sample size of 136 records. If the sample size, the distribution can change from skewed to the normal or right skewed. The graph would also show equal partition or right tail is longer than left. It depends on how much sample data we consider for our analysis.