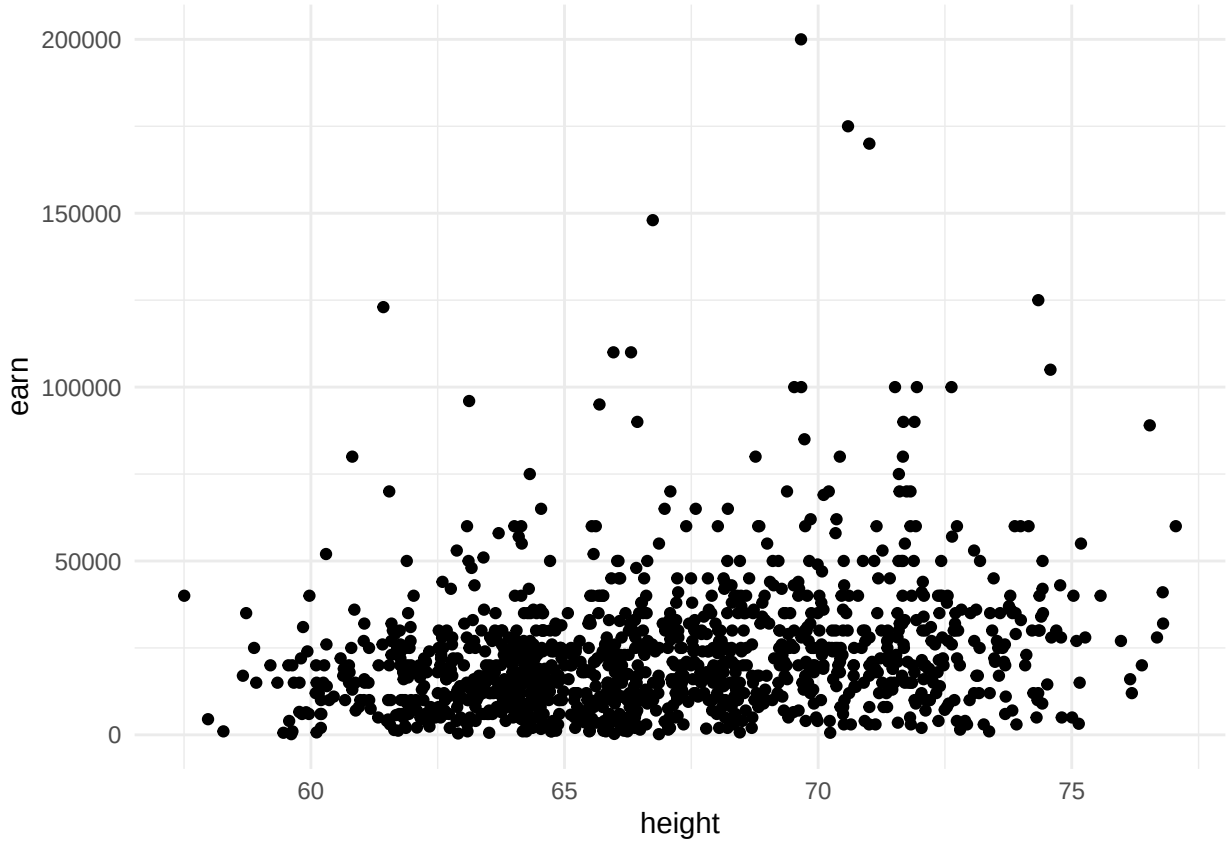
3.2Exercises.R

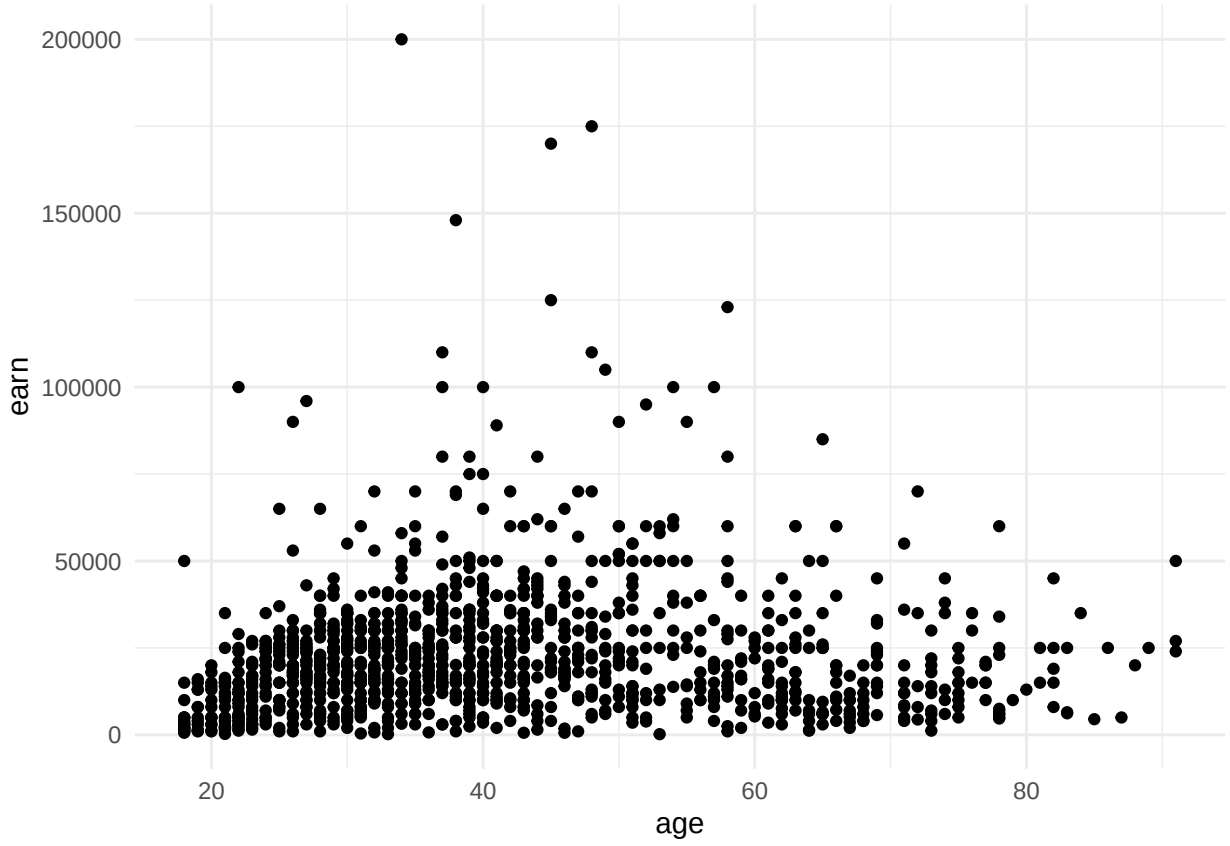
Rahul Rajeev

2022-12-14

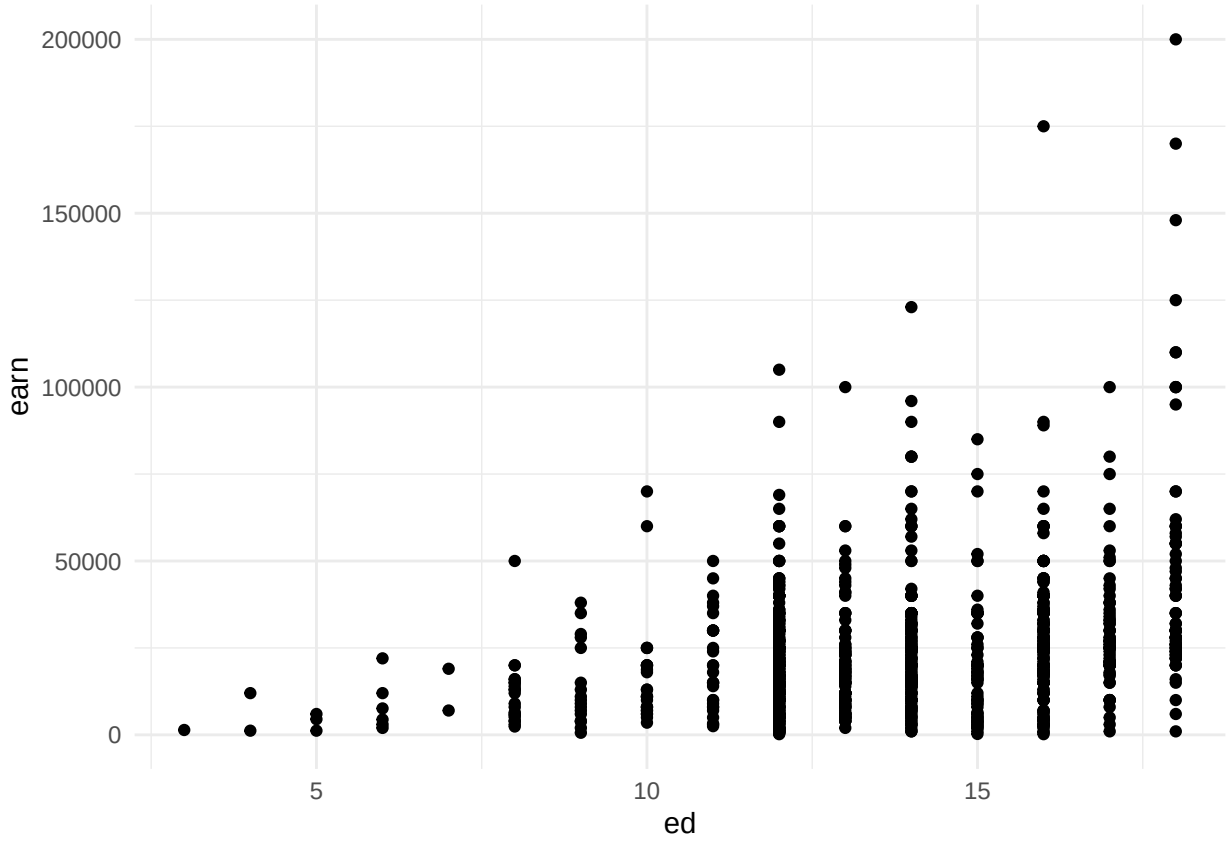
```
# Assignment: ASSIGNMENT 3
# Name: Rajeev, Rahul
# Date: 2022-12-12

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

## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/rahul/Documents/Bellevue/DSC 520")

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




```
## `age` vs. `earn`
ggplot(heights_df, aes(x= age, y= earn)) + geom_point()
```

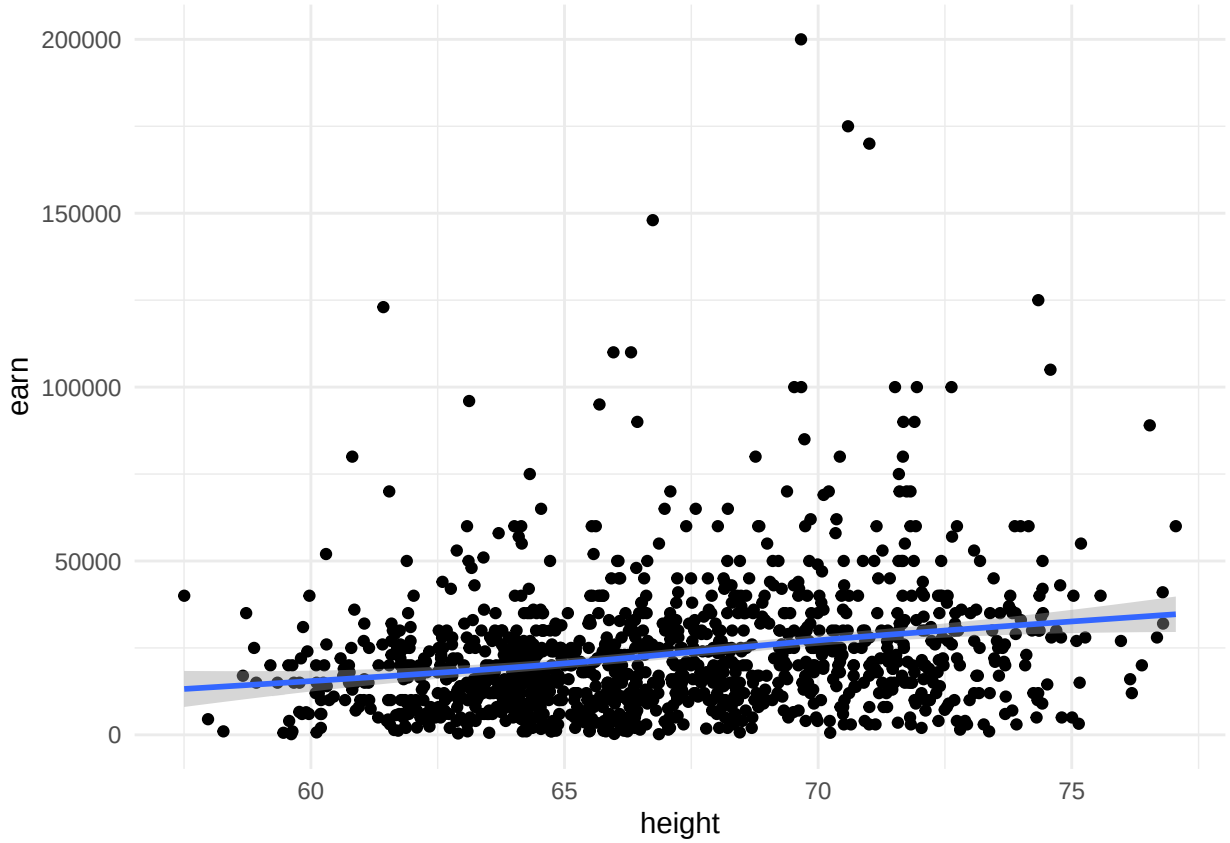



```
## `ed` vs. `earn`
ggplot(heights_df, aes(x= ed, y=earn)) + geom_point()
```



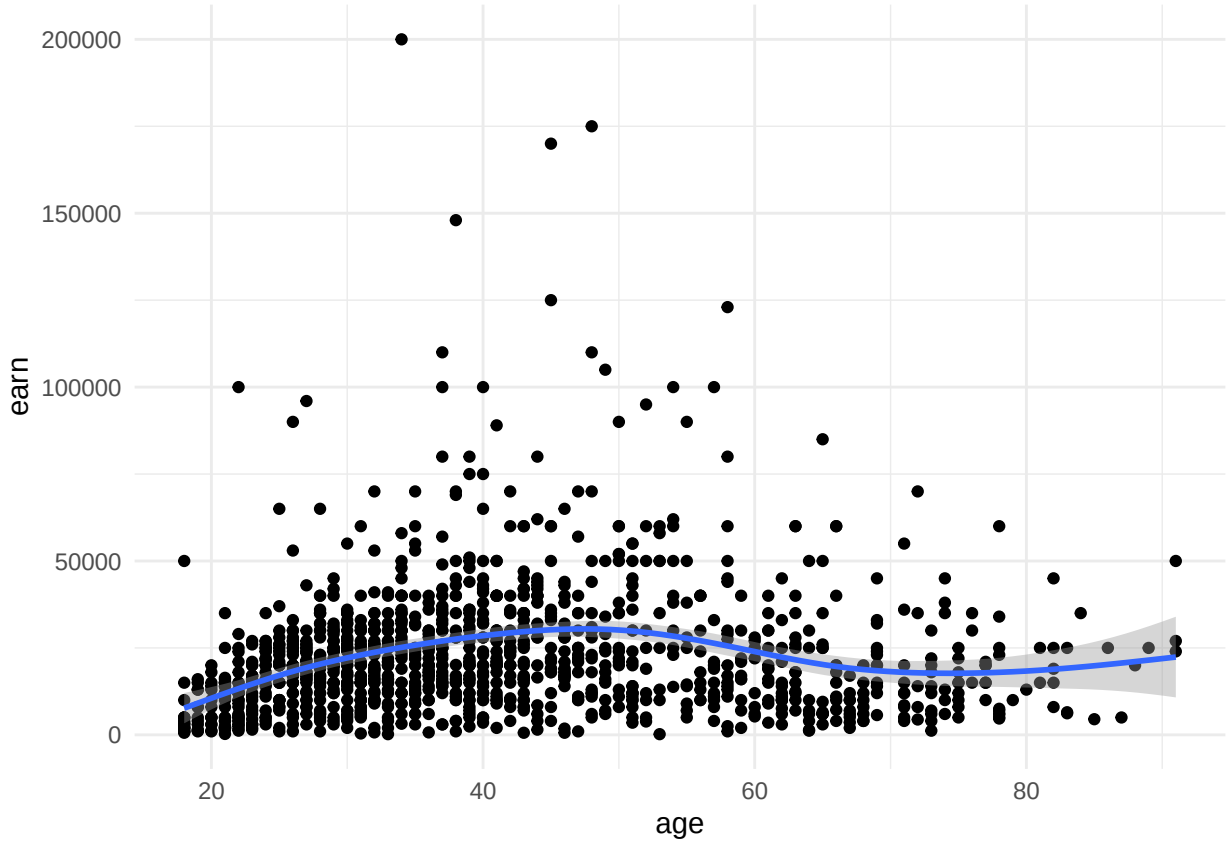

```
## Re-create the three scatterplots and add a regression trend line using
## the `geom_smooth()` function
## `height` vs. `earn`
ggplot(heights_df, aes(x= height, y= earn)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



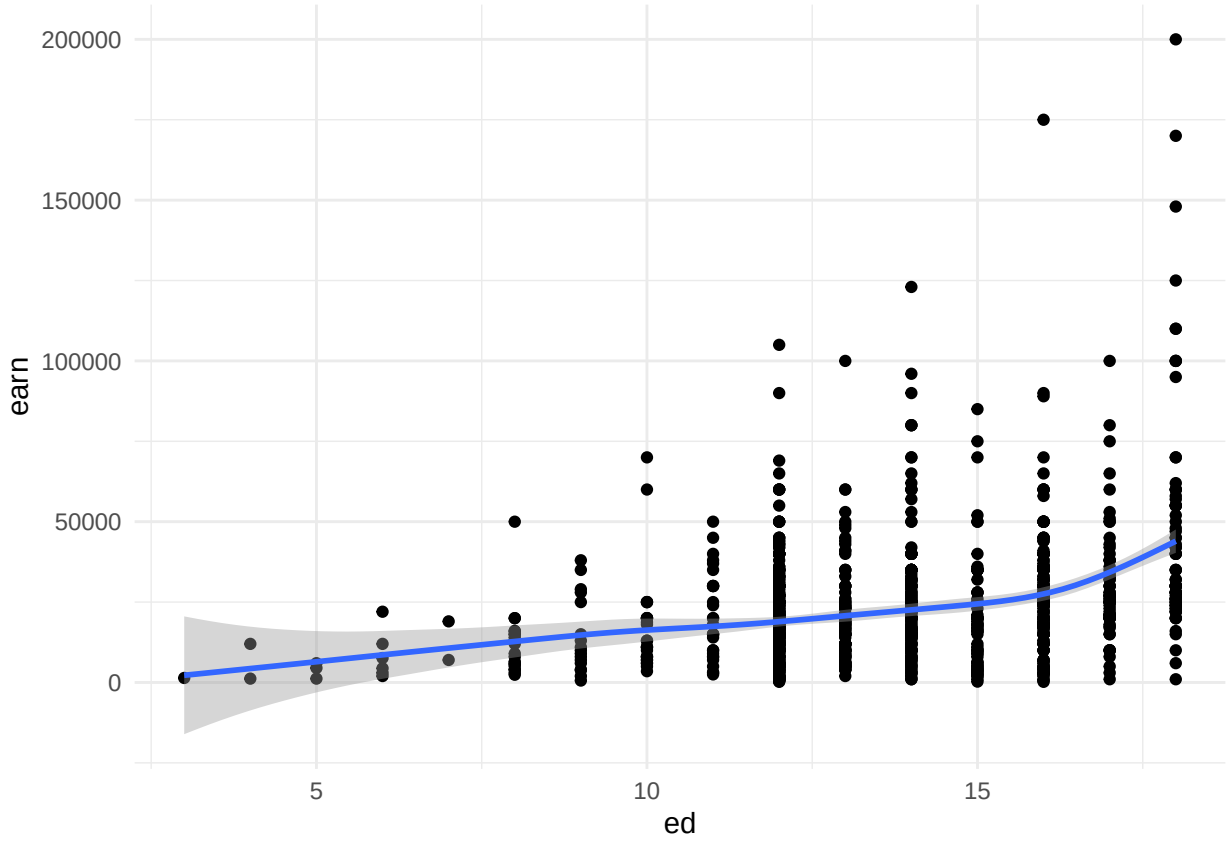

```
## `age` vs. `earn`
ggplot(heights_df, aes(x= age, y= earn)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

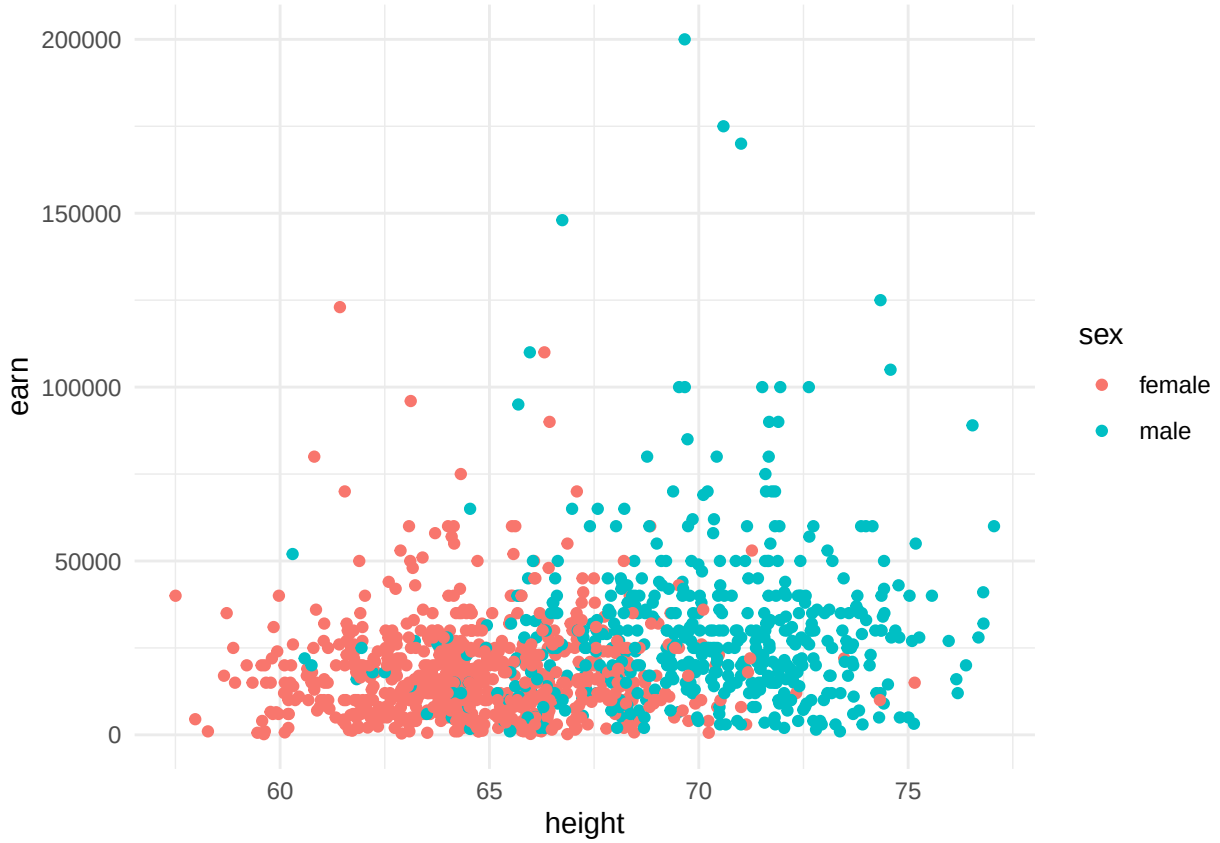



```
## `ed` vs. `earn`
ggplot(heights_df, aes(x= ed, y=earn)) + geom_point() + geom_smooth()
```

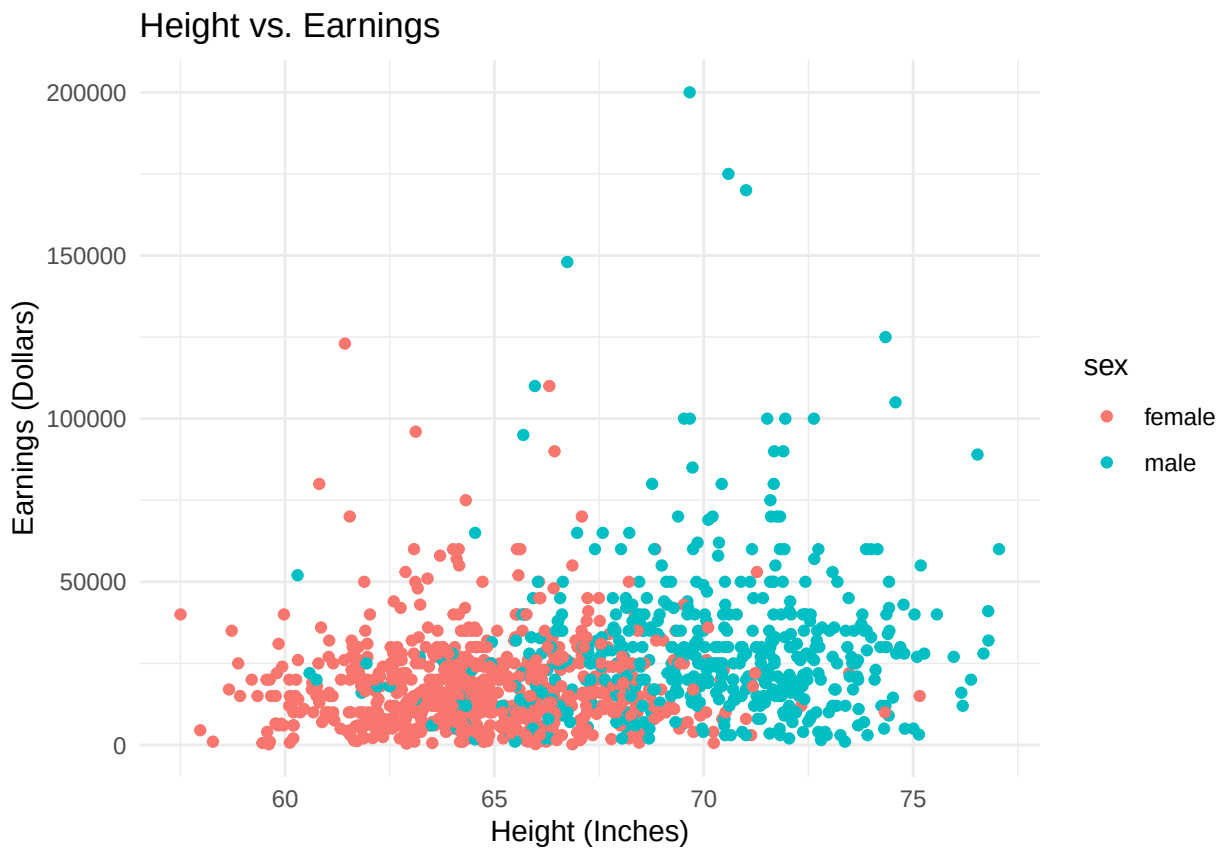
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
## 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()
```

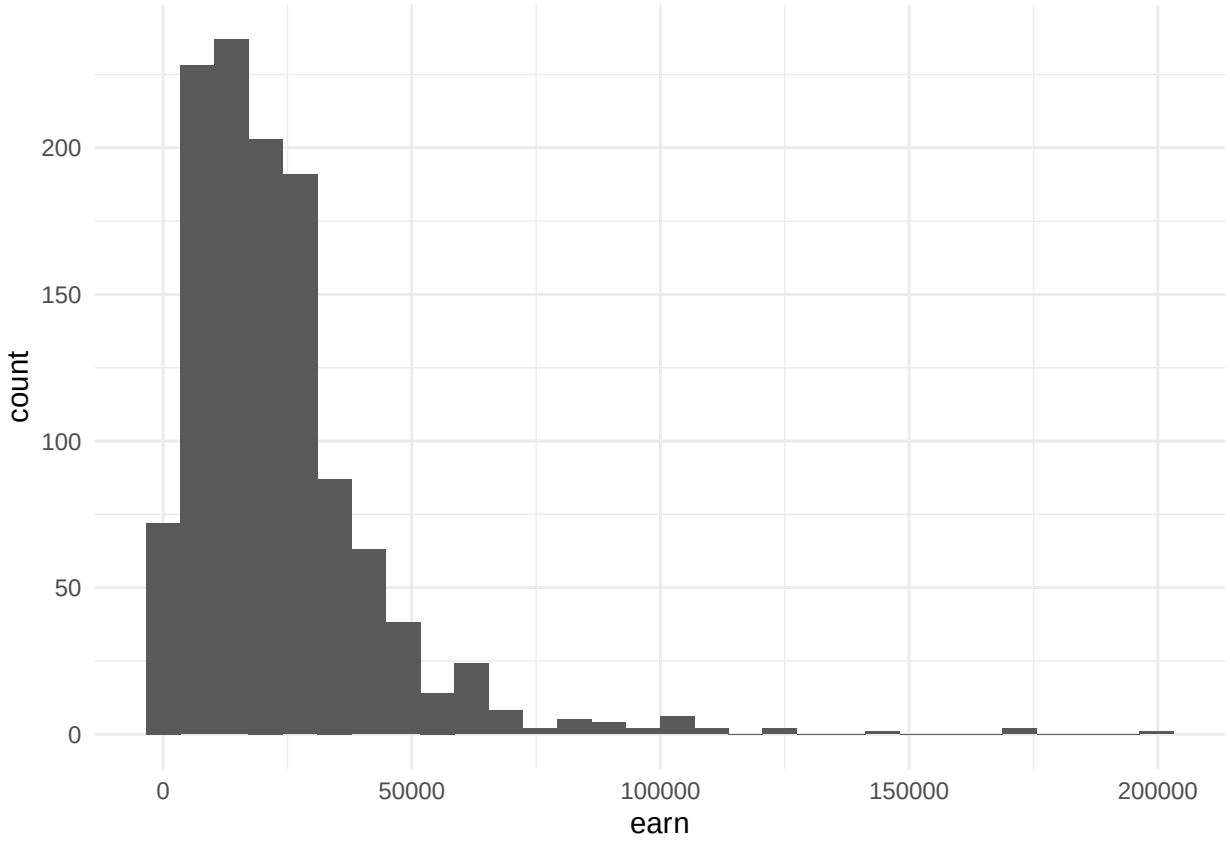



```
## Using `ggtitle()`, `xlab()`, and `ylab()` to add a title,
## x label, and y label to the previous plot
## Title: Height vs. Earnings
## X label: Height (Inches)
## Y Label: Earnings (Dollars)
ggplot(heights_df, aes(x= height, y= earn, col = sex)) +
  ggtitle('Height vs. Earnings') + xlab('Height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

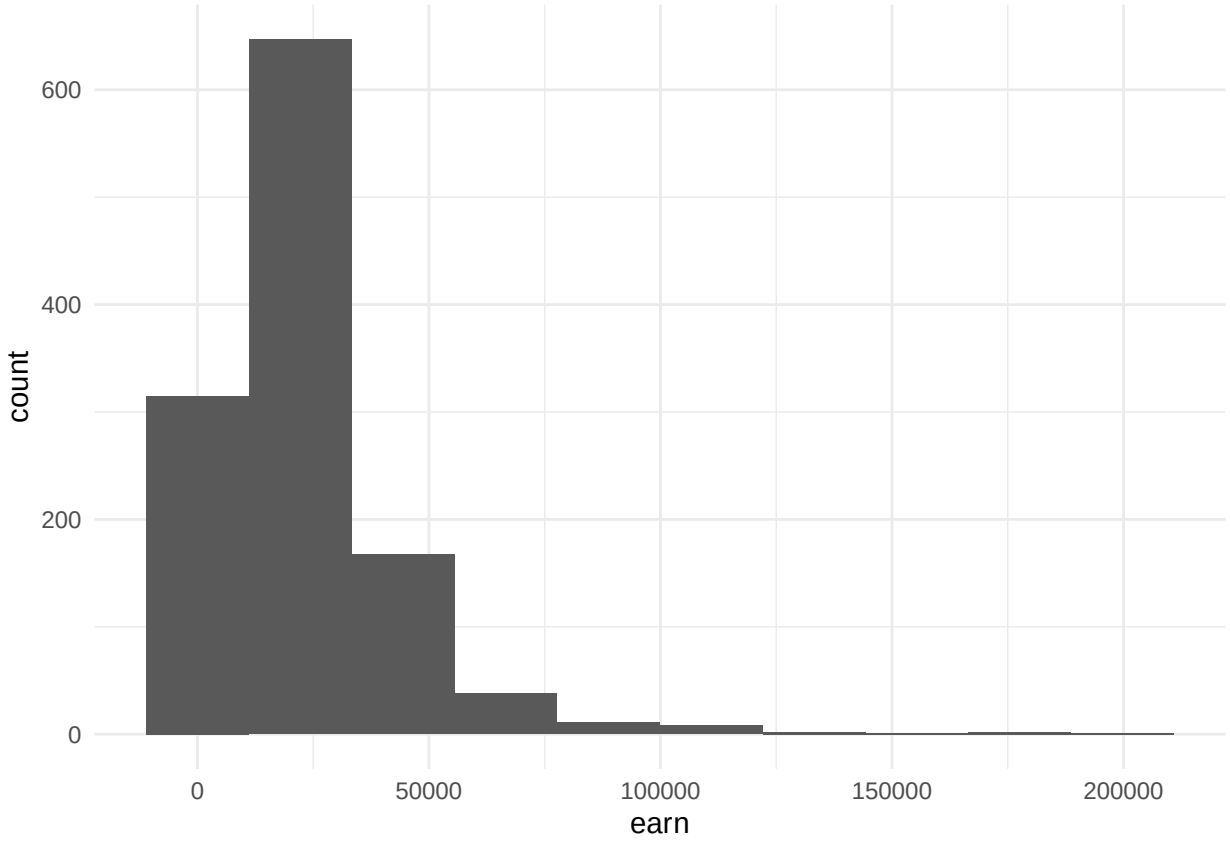



```
# https://ggplot2.tidyverse.org/reference/geom_histogram.html
## Create a histogram of the `earn` variable using `geom_histogram()`
ggplot(heights_df, aes(x=earn)) + geom_histogram()
```

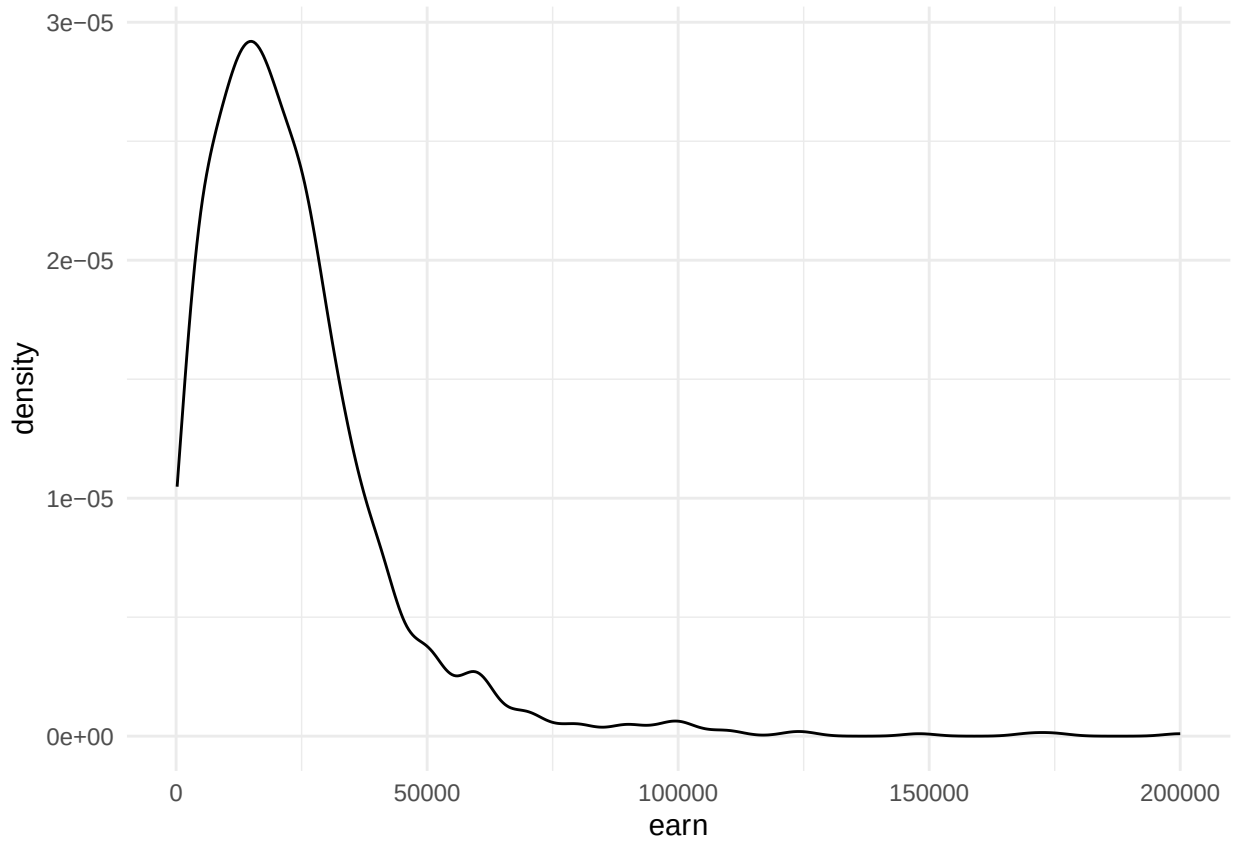
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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




```
#==============================================================================
# Assignment: American Community Survey Exercise
# Name: Rajeev, Rahul
# Date: 2022-12-12

library(ggplot2)
theme_set(theme_minimal())

# loading American Community Survey dataset
setwd("C:/Users/rahul/Documents/Bellevue/DSC 520")

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

# i) listing name of each field, data type, and intent of data
# Name: Id, Data type: varchar, intent: unique identifier")
# Name: Id2, Data type: int, intent: another identifier?")
# Name: Geography, Data type: char, intent: name of county")
# Name: PopGroupID, Data type: int, intent: I'm not sure about this one")
# Name: POPGROUP.display.label, Data type: chr, intent: a label for the type of data")
# Name: Races Reported, Data type: int, intent: population amount that took survey")
# Name: HSDegree, Data type: num, intent: percent of population with HS degree")
# Name: BSDegree, Data type: num, intent: percent of population with BS degree")

# ii) running str(), nrow(), and ncol()
str(acs_df)
```

```
## '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,
##  $ PopGroupID            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ POPGROUP.display.label: chr  "Total population" "Total population" "Total population" "Total popul
##  $ RacesReported         : int  660793 4087191 1004516 1610921 1111339 965974 874589 10116705 314551
##  $ 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 ...
```

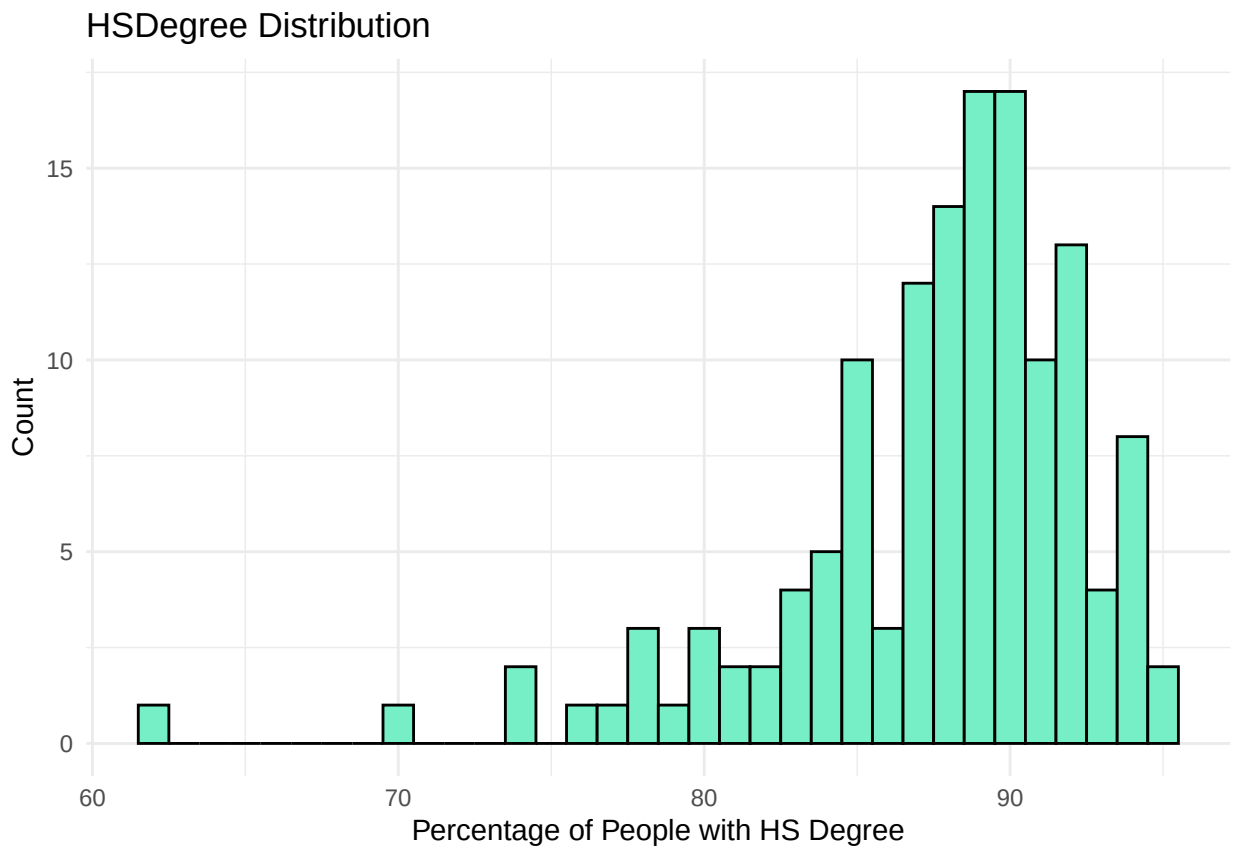
```
nrow(acs_df)
```

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

```
ncol(acs_df)
```

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

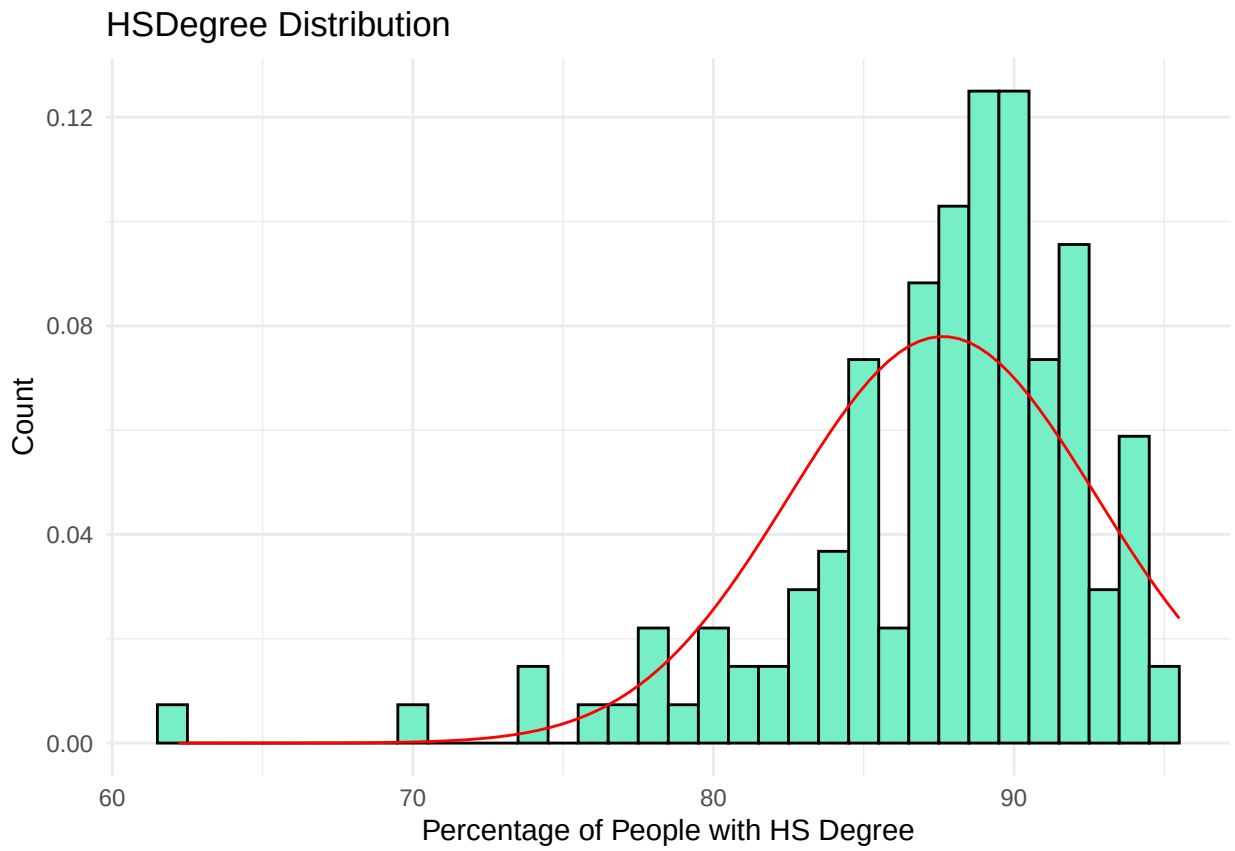
```
# iii) create histogram of HSDegree variable using ggplot2 package
ggplot(acs_df, aes(x=HSDegree)) +
  geom_histogram(fill="aquamarine2", colour="black",binwidth=1, bins=50) +
  ggtitle('HSDegree Distribution') + ylab("Count") +
  xlab('Percentage of People with HS Degree')
```




```
# iv) Answering questions based on Histogram
# 1. The data distribution appears to be unimodal
# 2. It's not symmetrical, it appears to be skew right.
# 3. It appears to be bell-shaped with a peak.
```

```
# 4. It isn't normal, because the distribution isn't centered.
# 5. As said in number 2, the distribution appears to be skewed right.

# 6. include a normal distribution curve to the histogram plotted
ggplot(data = acs_df) +
  geom_histogram(mapping = aes(x = HSDegree, y=..density..),
                 fill="aquamarine2", colour="black", binwidth = 1,bins = 50) +
  ggtitle('HSDegree Distribution') + ylab("Count") +
  xlab('Percentage of People with HS Degree') +
  stat_function(fun = dnorm, colour = "red",
                args = list(mean = mean(acs_df$HSDegree),
                            sd = sd(acs_df$HSDegree)))
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
```



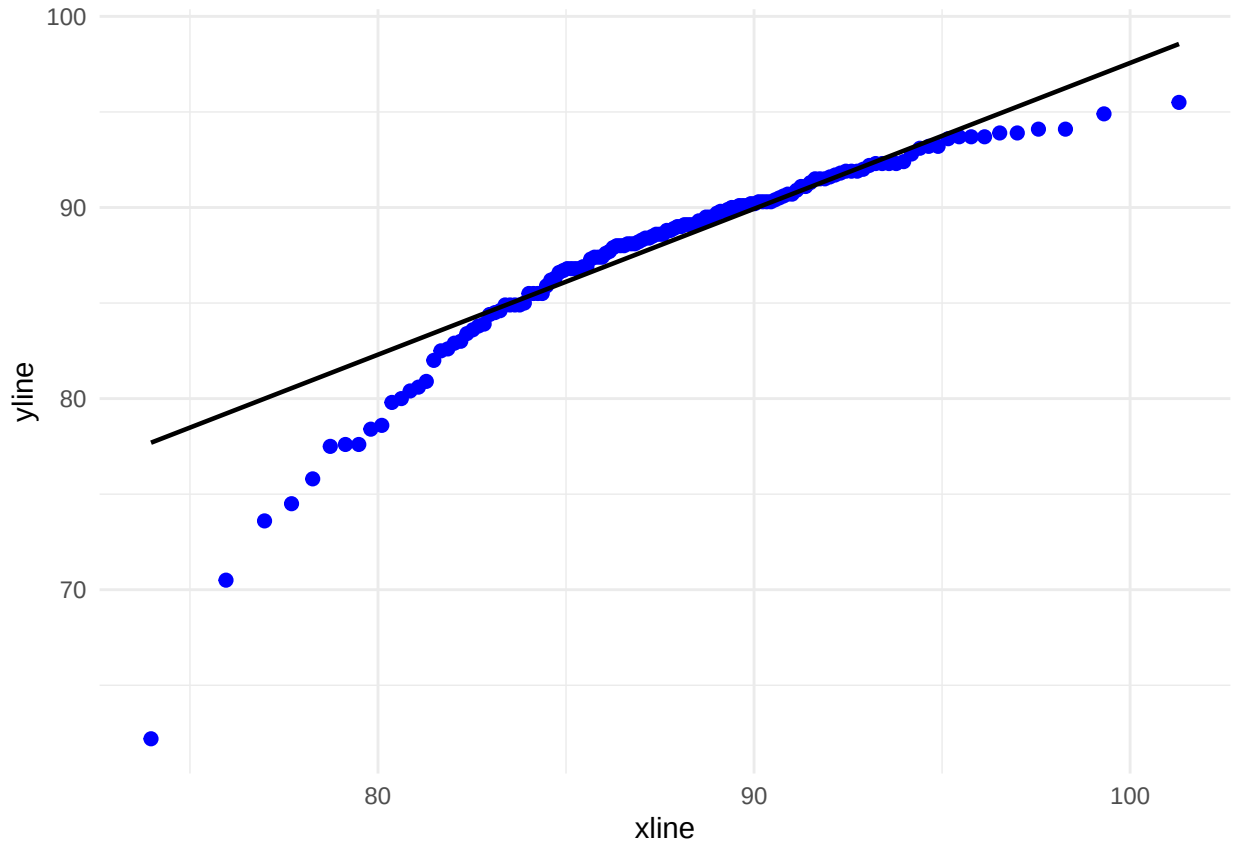

```
# 7. The curve used to represent the data isn't a normal curve
# because it is skewed.

# v) create a probability plot of the HSDegree variable
library(qqplotr)
```

```
##
## Attaching package: 'qqplotr'
##
## The following objects are masked from 'package:ggplot2':
##
```

```
##     stat_qq_line, StatQqLine
```

```
ggplot(acs_df, aes(sample=HSDegree)) + stat_qq_point(size=2, color = 'blue') +
  stat_qq_line(color = 'black')
```

```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
# vi) answer questions about probability plot
# The distribution is not normal because it does not fitthe normal distributed
# line on the probability plot.
# The distribution must be skewed right because the graph curves above the line.

# vii) quantify normality using stat.desc() and screenshot results
library(pastecs)
stat.desc(acs_df$HSDegree, basic = FALSE, norm = TRUE)
```

```
##        median          mean       SE.mean  CI.mean.0.95           var
##  8.870000e+01  8.763235e+01  4.388598e-01  8.679296e-01  2.619332e+01
##       std.dev      coef.var      skewness      skew.2SE      kurtosis
##  5.117941e+00  5.840241e-02 -1.674767e+00 -4.030254e+00  4.352856e+00
##      kurt.2SE    normtest.W    normtest.p
##  5.273885e+00  8.773635e-01  3.193634e-09
```

```
# the results are on the next page of the pdf.

# viii) in several sentences provide an explanation of the result
# produced for skew, kurtosis, and z-scores.

# The mean is about 87.6, with a median of 88.7 and a standard deviation of 5.
# Skewness is less than -1 which implies that the data is highly negatively
# skewed, or skew right.
# Kurtosis value is 4.32 which is higher than 3, implying positive kurtosis.
# The graph is heavy-tailed and the top of the curve is steep, which pulls up
# the distribution. The z-score is used to do a test and the p-value is
# 3.19e-09 which is much smaller than 0.05 for a normal distribution.
# If the sample size is increased, then the absolute value of the skewness and
# kurtosis decreases. The skewness would start to disappear, act more normal,
# and the positive kurtosis will be brought down into a normal bell curve.
```