assignment_03_MunjewarSheetal.R

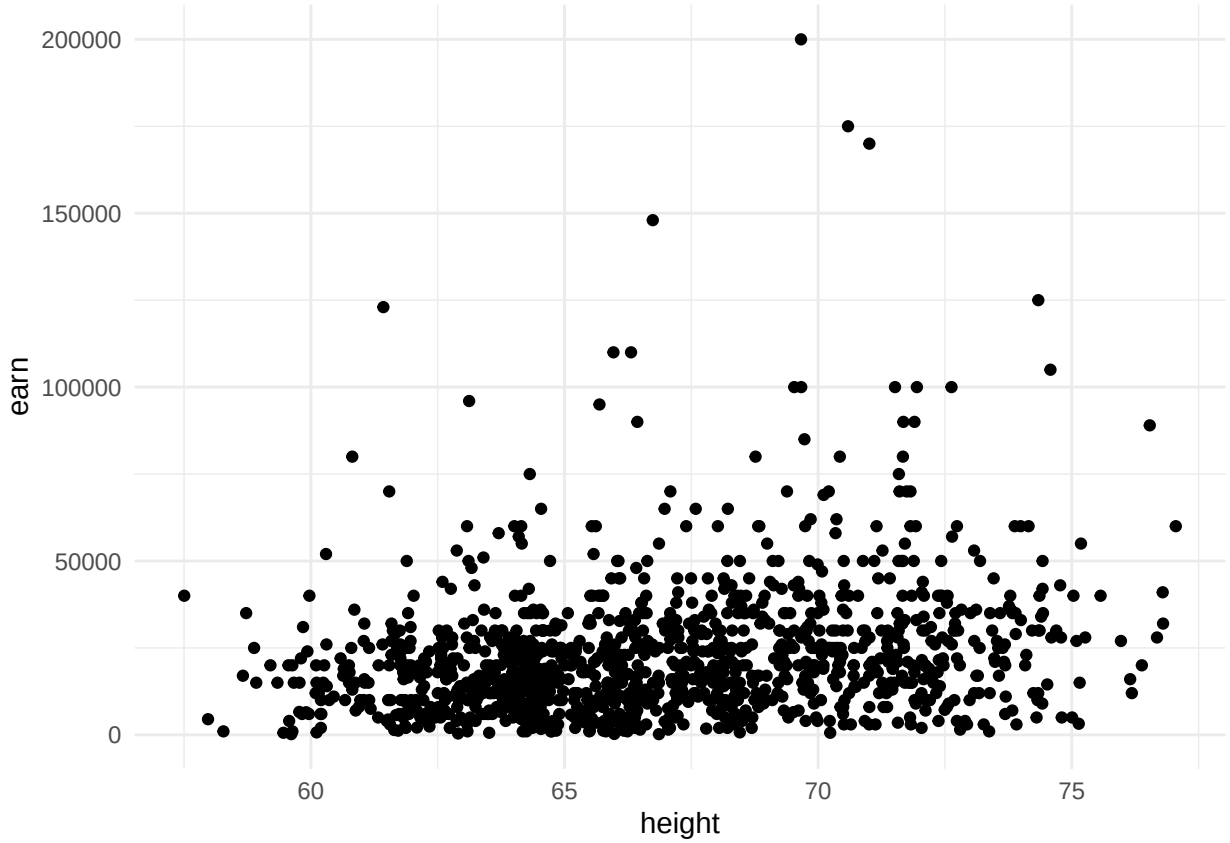
sheetal

2022-12-18

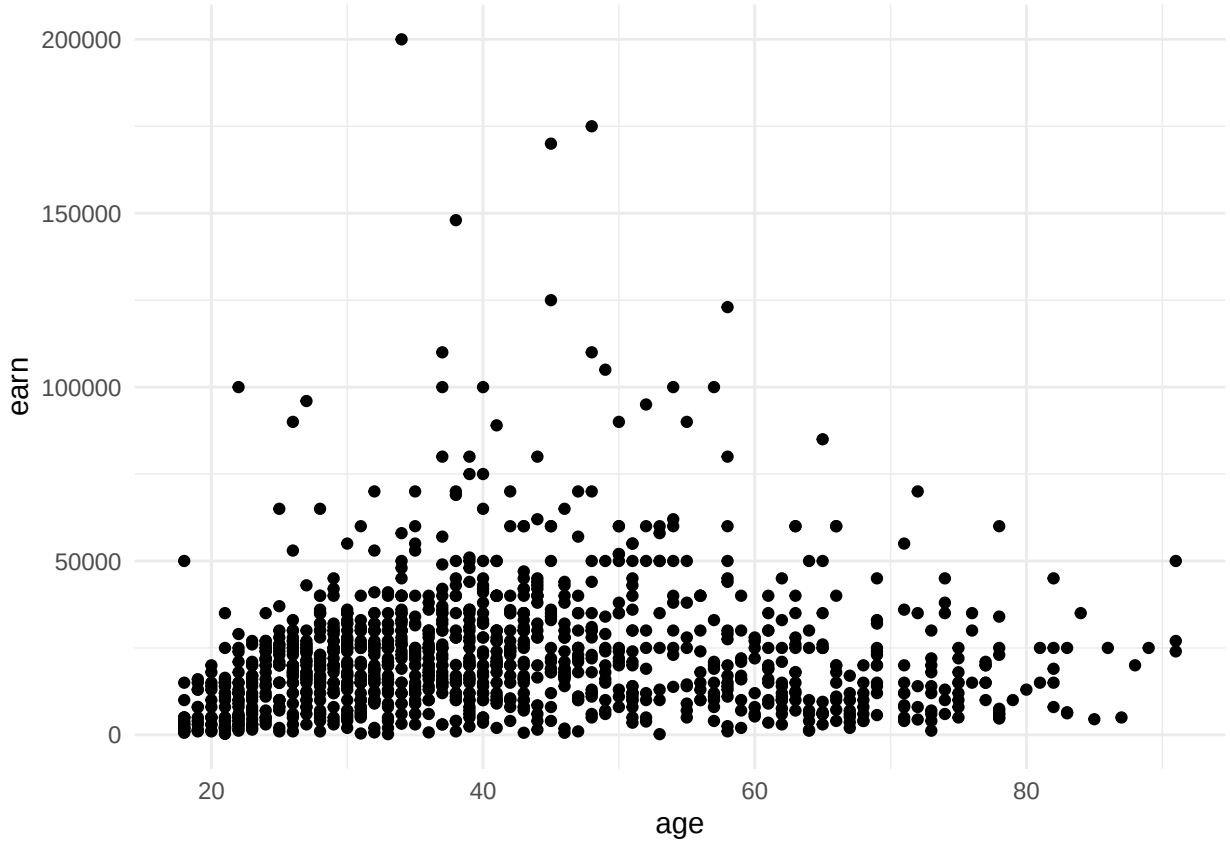
```
# Assignment: ASSIGNMENT 3
# Name: Munjewar, Sheetal
# Date: 2022-12-18

## Load the ggplot2 package
#install.packages("ggplot2")
library(ggplot2)
theme_set(theme_minimal())

## Set the working directory to the root of your DSC 520 directory
#setwd("/home/jdoe/Workspaces/dsc520")
setwd("E:\\Data_Science_DSC510\\DSC520-Statistics\\dsc520")

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

```
##       earn            height          sex                 ed
##  Min.   :   200   Min.   :57.50   Length:1192        Min.   : 3.0
##  1st Qu.: 10000   1st Qu.:64.01   Class :character   1st Qu.:12.0
##  Median : 20000   Median :66.45   Mode  :character   Median :13.0
##  Mean   : 23155   Mean   :66.92                      Mean   :13.5
##  3rd Qu.: 30000   3rd Qu.:69.85                      3rd Qu.:16.0
##  Max.   :200000   Max.   :77.05                      Max.   :18.0
##       age            race
##  Min.   :18.00   Length:1192
##  1st Qu.:29.00   Class :character
##  Median :38.00   Mode  :character
##  Mean   :41.38
##  3rd Qu.:51.00
##  Max.   :91.00
```

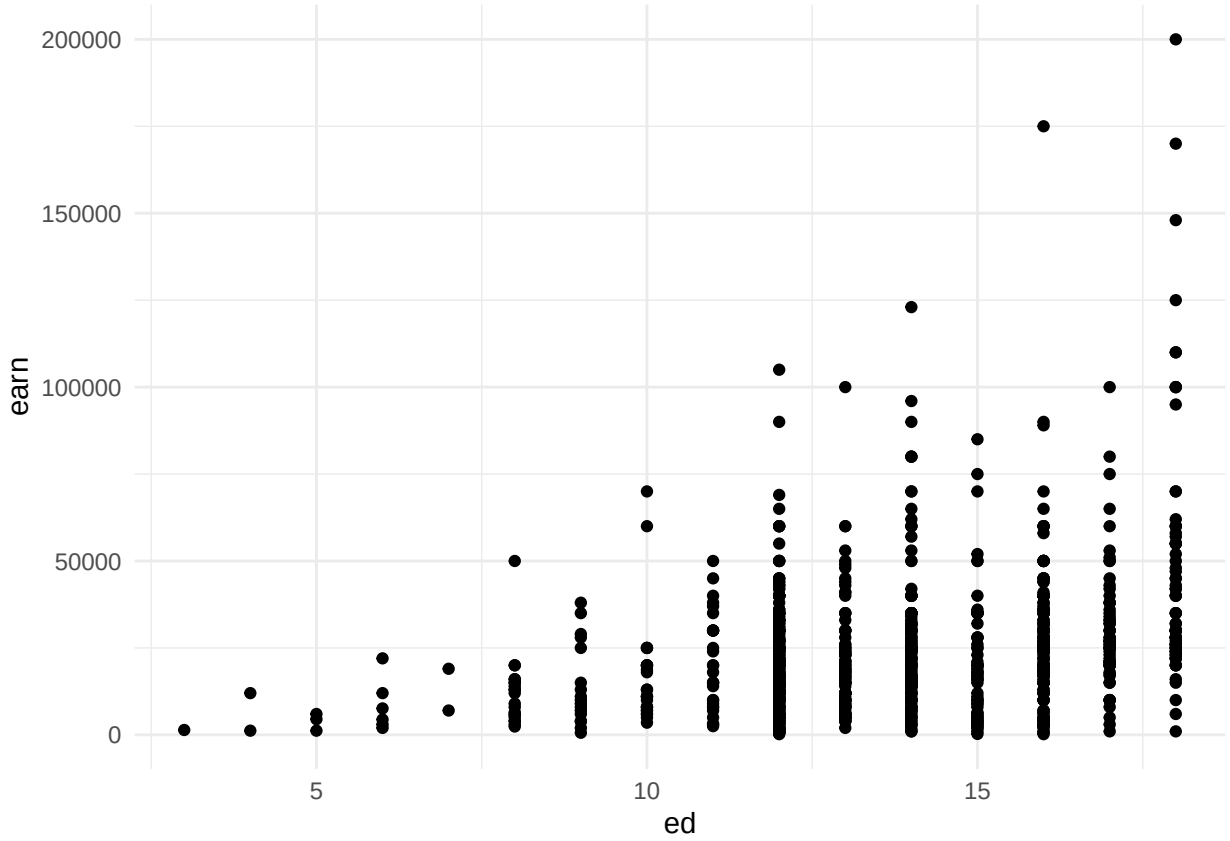
```
# https://ggplot2.tidyverse.org/reference/geom_point.html
## Using `geom_point()` create three scatterplots for
## `height` vs. `earn`
ggplot(heights_df, aes(x=height, y=earn)) + geom_point()
```




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

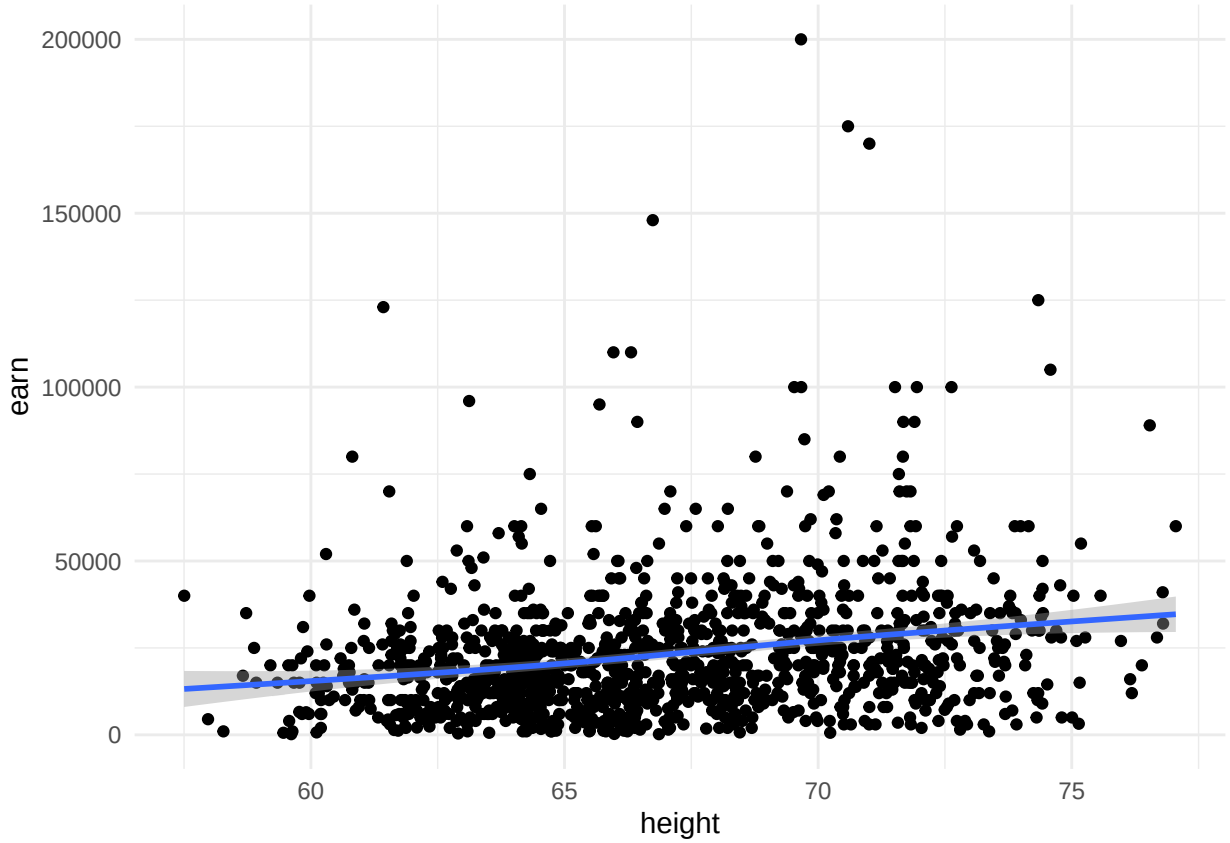



```
## `ed` vs. `earn`
#Actual question.
#ggplot(heights_df) + __
ggplot(heights_df, aes(x=ed, y=earn)) + geom_point()
```



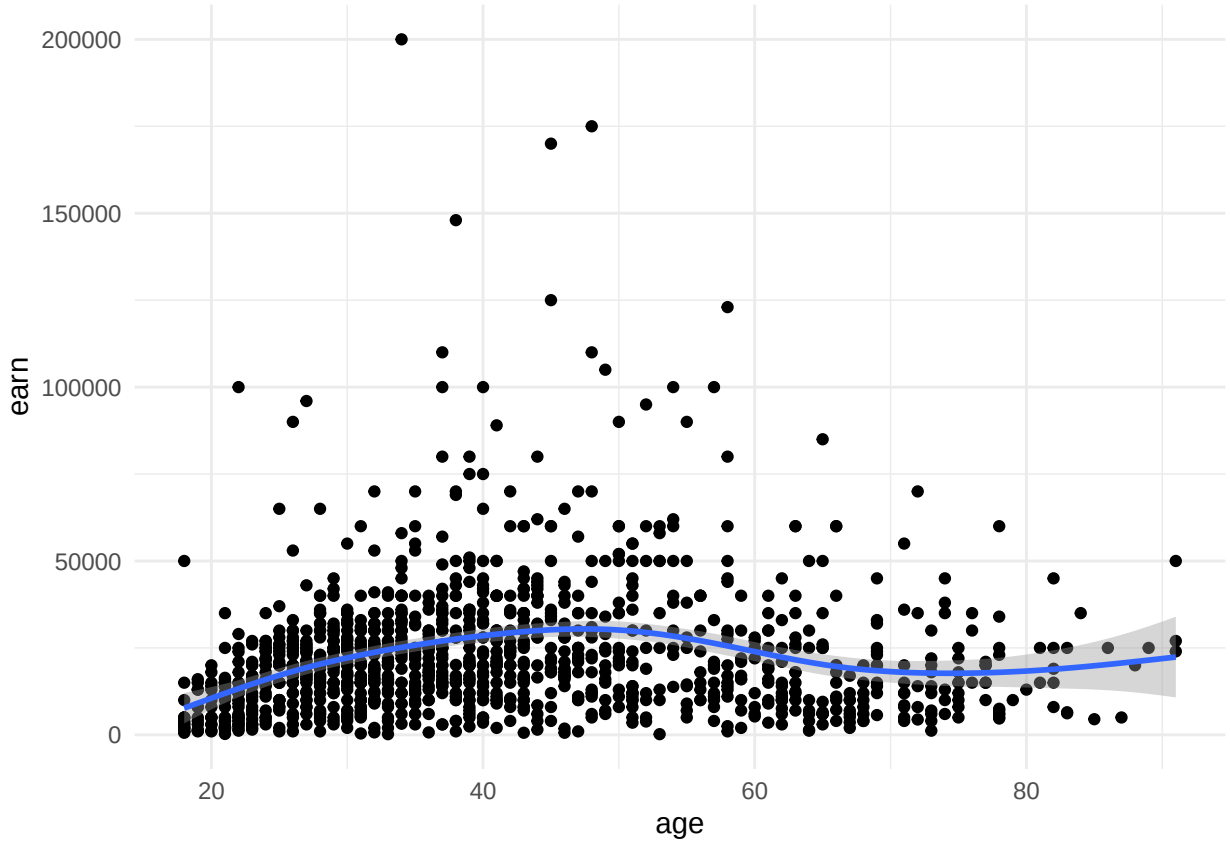

```
## Re-create the three scatterplots and add a regression trend line using
## the `geom_smooth()` function
## `height` vs. `earn`
ggplot(heights_df, aes(x=height, y=earn)) + geom_point() + geom_smooth()
```

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



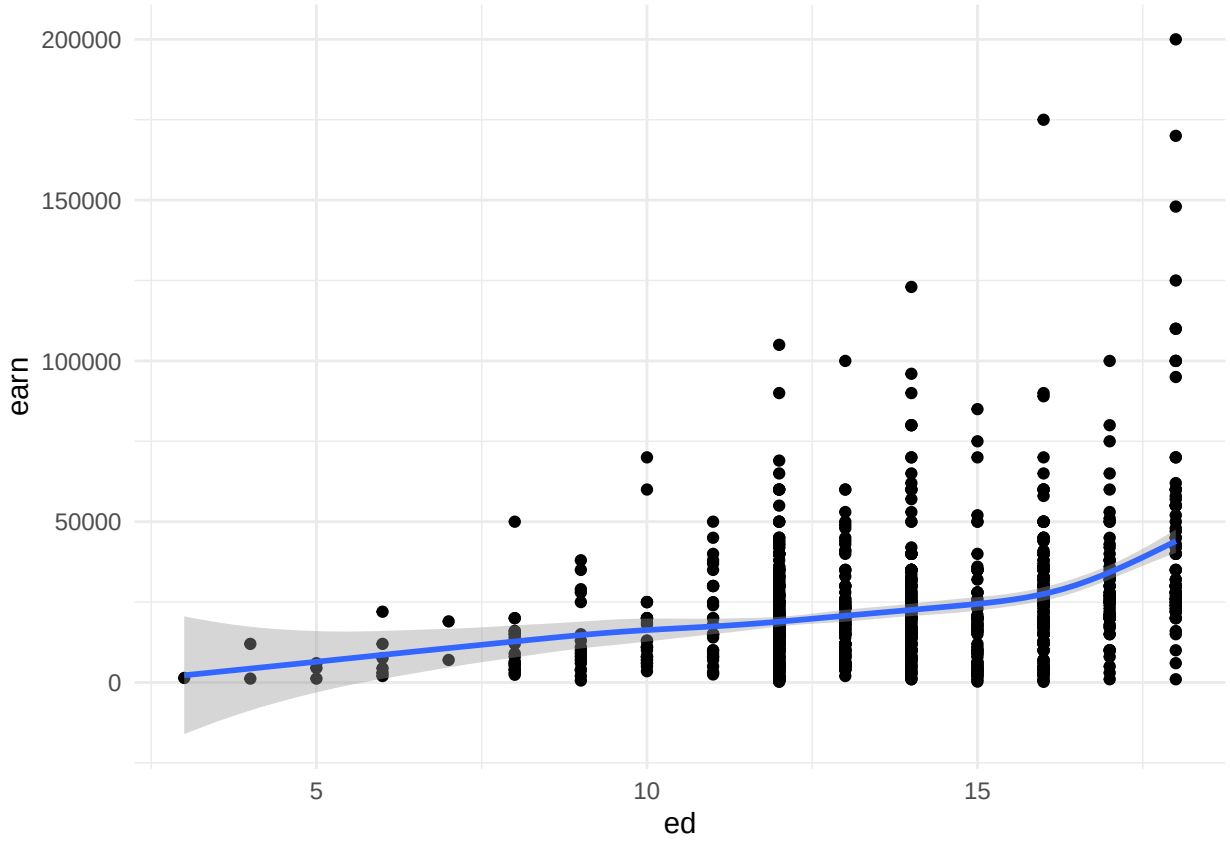

```
## `age` vs. `earn`
ggplot(heights_df, aes(x=age, y=earn)) + geom_point() + geom_smooth()
```

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

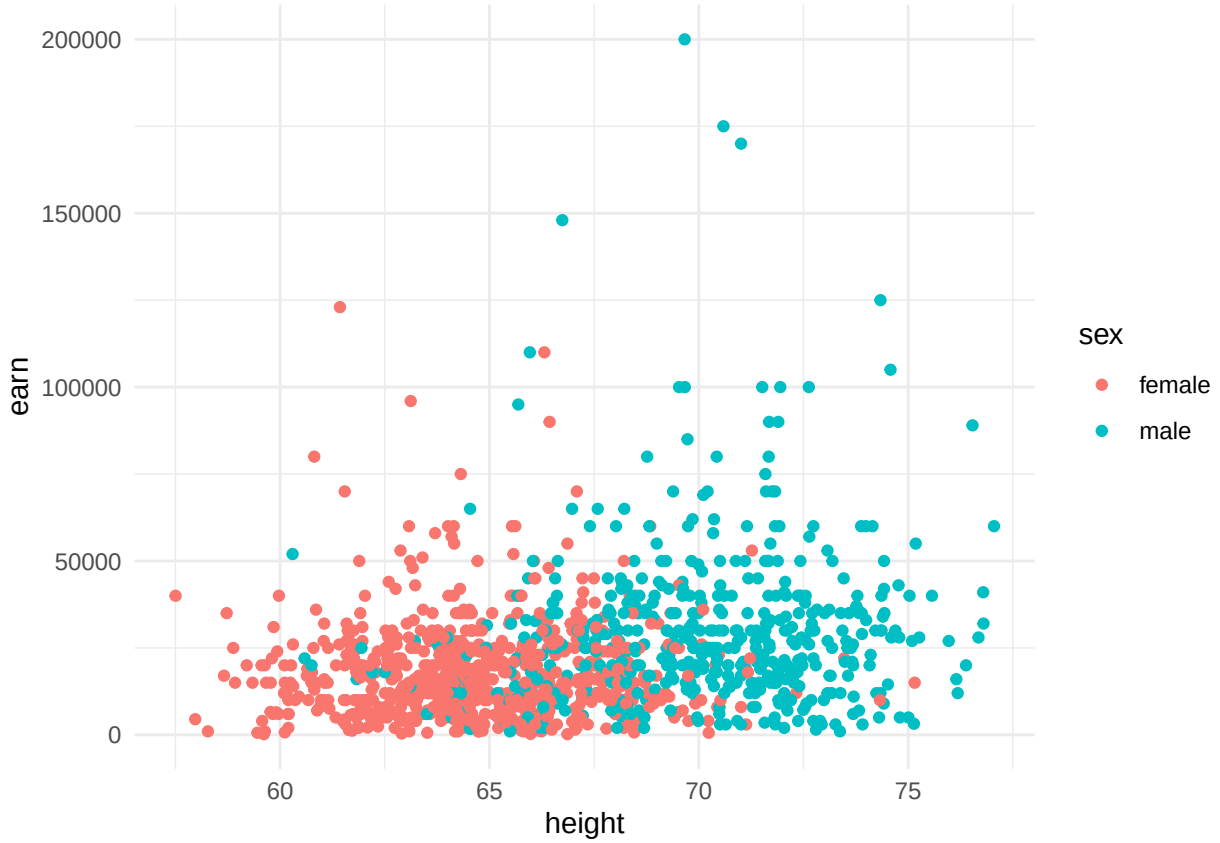



```
## `ed` vs. `earn`
ggplot(heights_df, aes(x=ed, y=earn)) + geom_point() + geom_smooth()
```

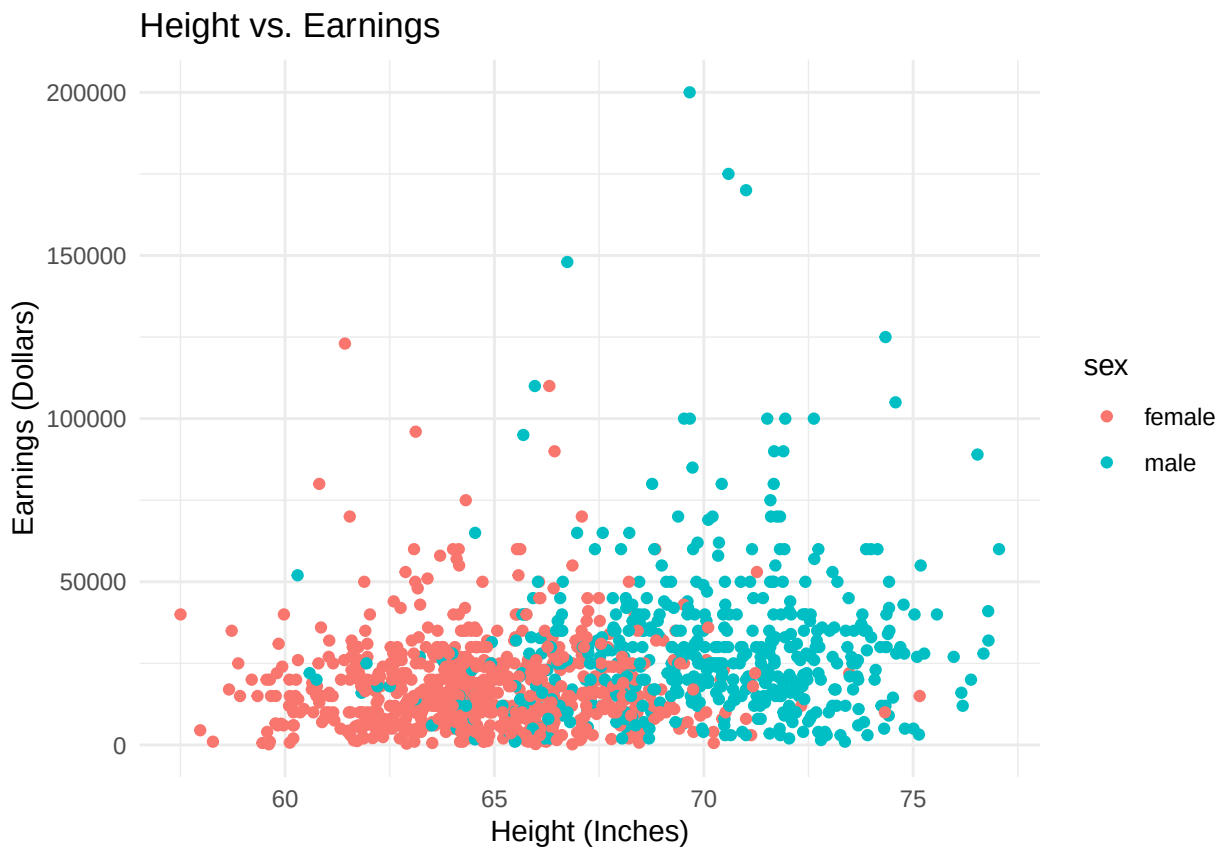
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
## 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()
```

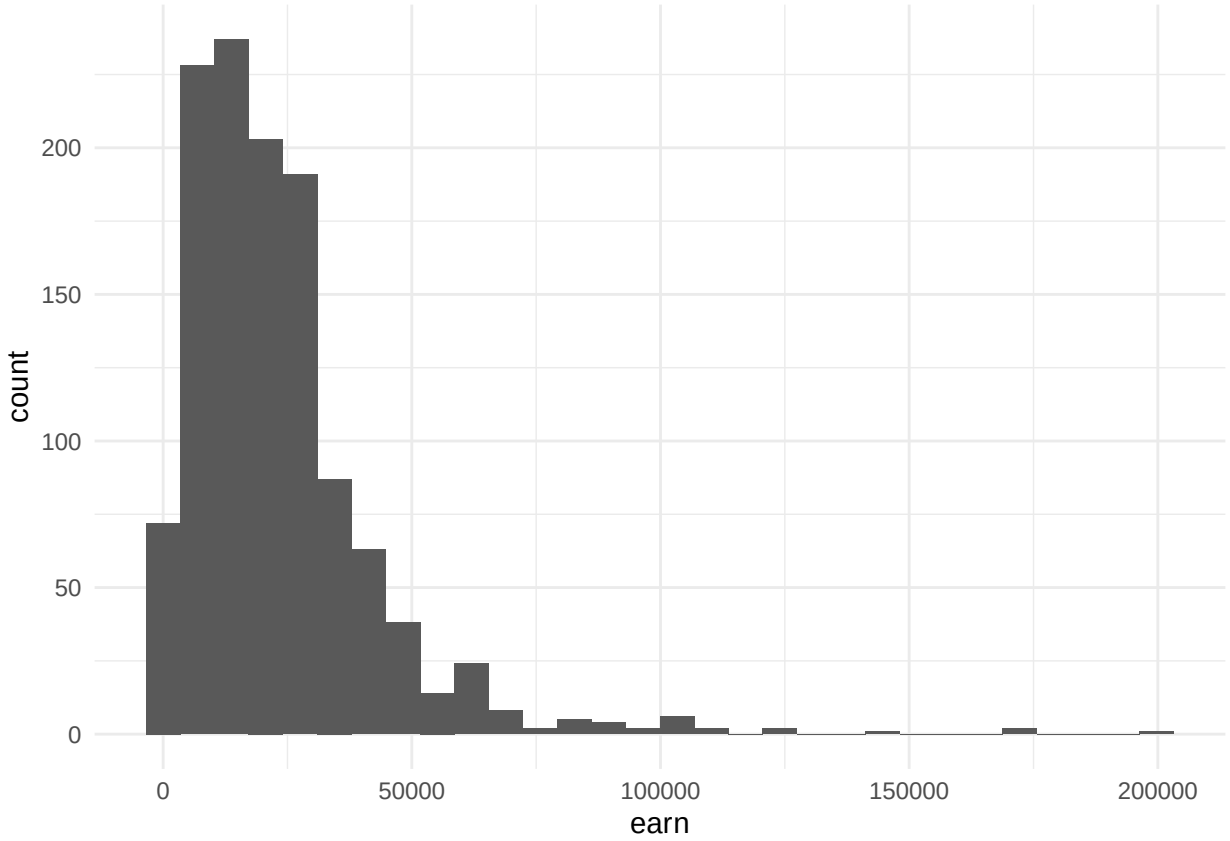



```
## Using `ggtitle()`, `xlab()`, and `ylab()` to add a title, x label, and y label to the previous plot
## Title: Height vs. Earnings
## X label: Height (Inches)
## Y Label: Earnings (Dollars)
ggplot(heights_df, aes(x=height, y=earn, col=sex)) + geom_point() + ggtitle("Height vs. Earnings") + xla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

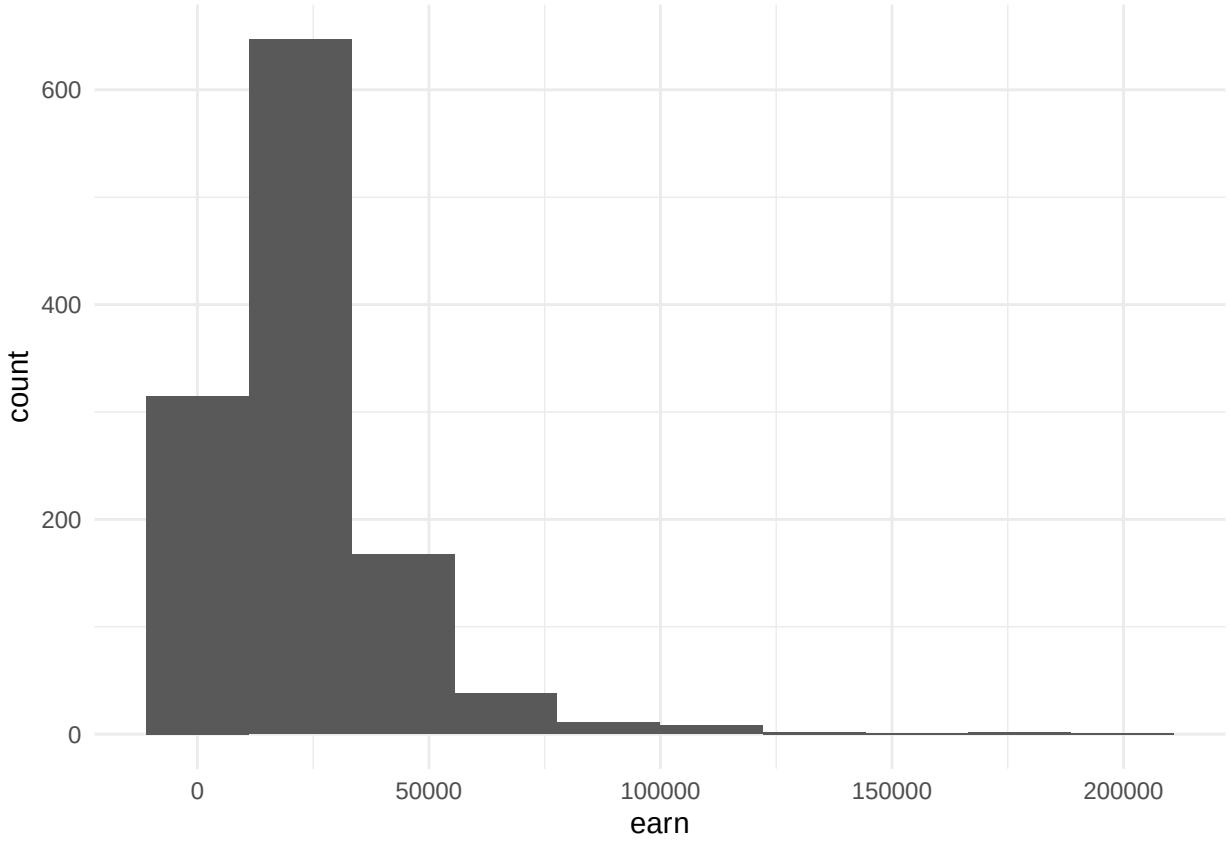



```
# https://ggplot2.tidyverse.org/reference/geom_histogram.html
## Create a histogram of the `earn` variable using `geom_histogram()`
ggplot(heights_df, aes(earn)) + geom_histogram()
```

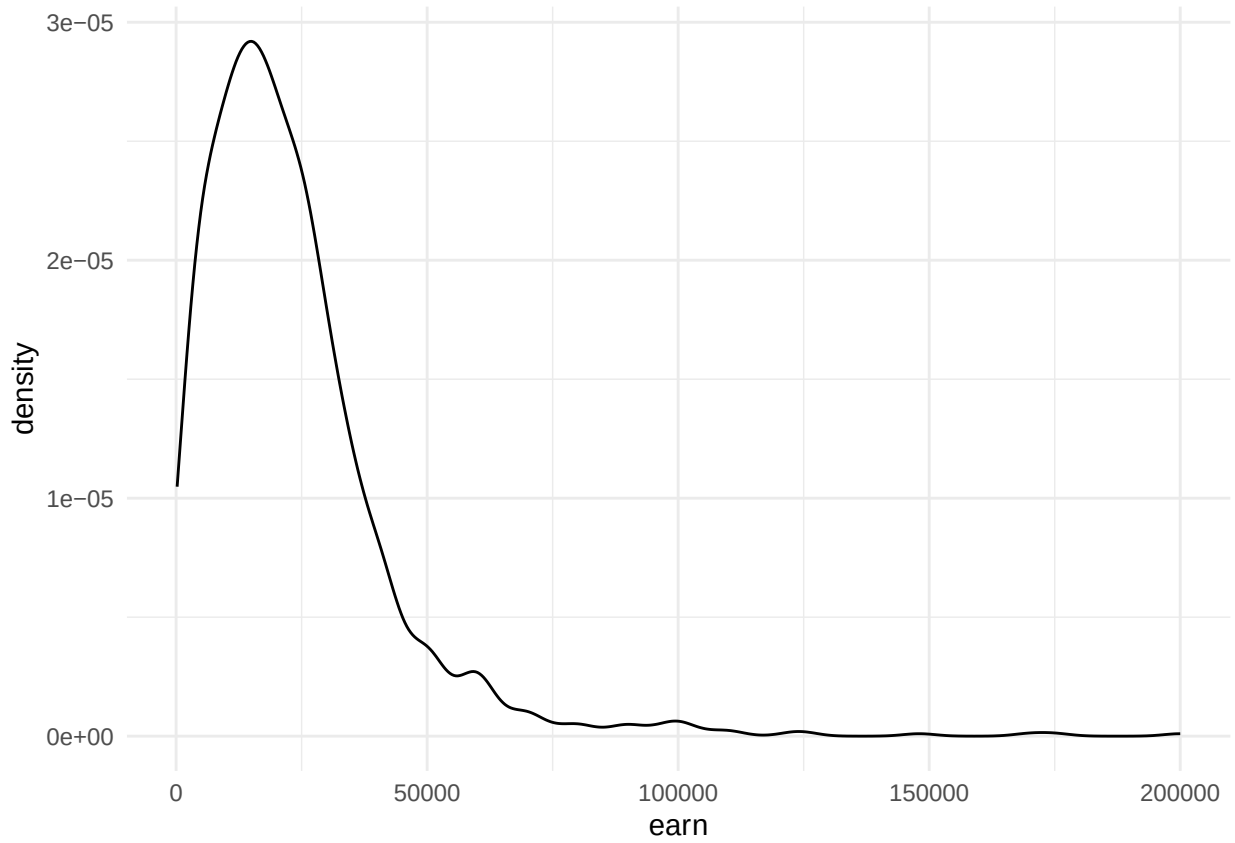
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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


3e−05
2e−05
1e−05
0e+00
density
0
50000
100000
150000
200000
earn