Data Science - Homework #3

Raymond Ogunjimi

Completed at 19:48:39 on 07 February, 2022

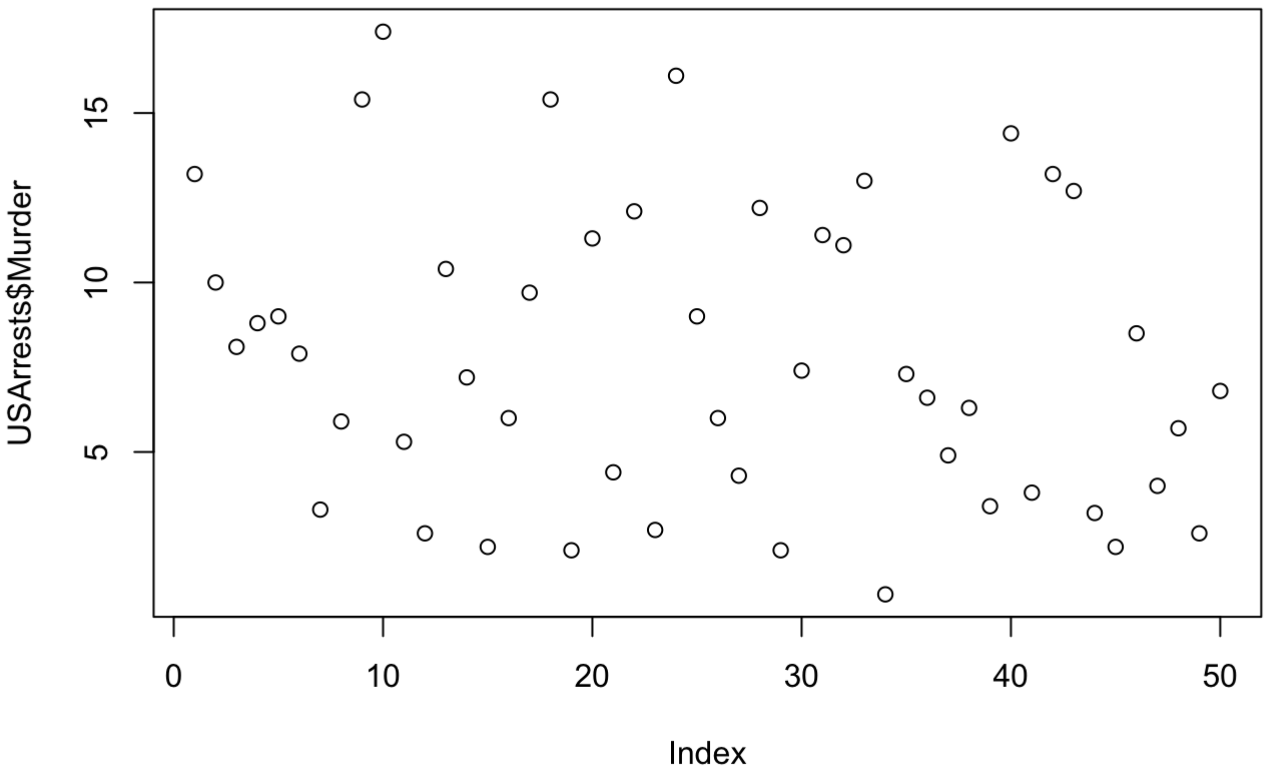
```
library(tidyverse)
library(mosaicData) # For problem 2
library(babynames) # For problem 3
library(mdsr) # For problem 4
```

Problem 1) Arrests Per State in the US

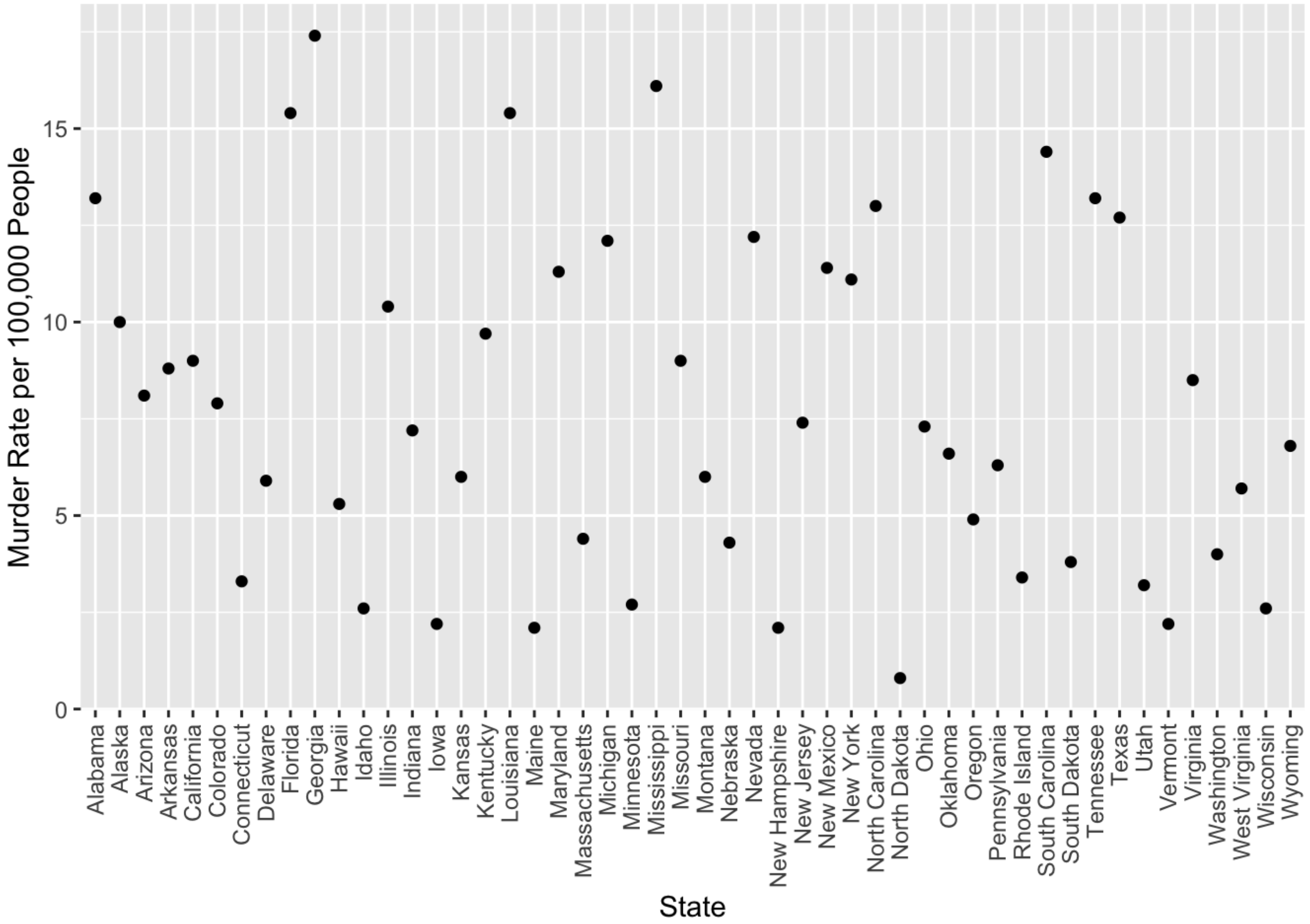
The `USArrests` dataset contains one observation per US State. The variables are the murder, assault, and rape rates per 100,000 people as well as the percent of the population living in urban areas.

Problem 1a) Create a graph of murder rate.

```
plot(USArrests$Murder)
```



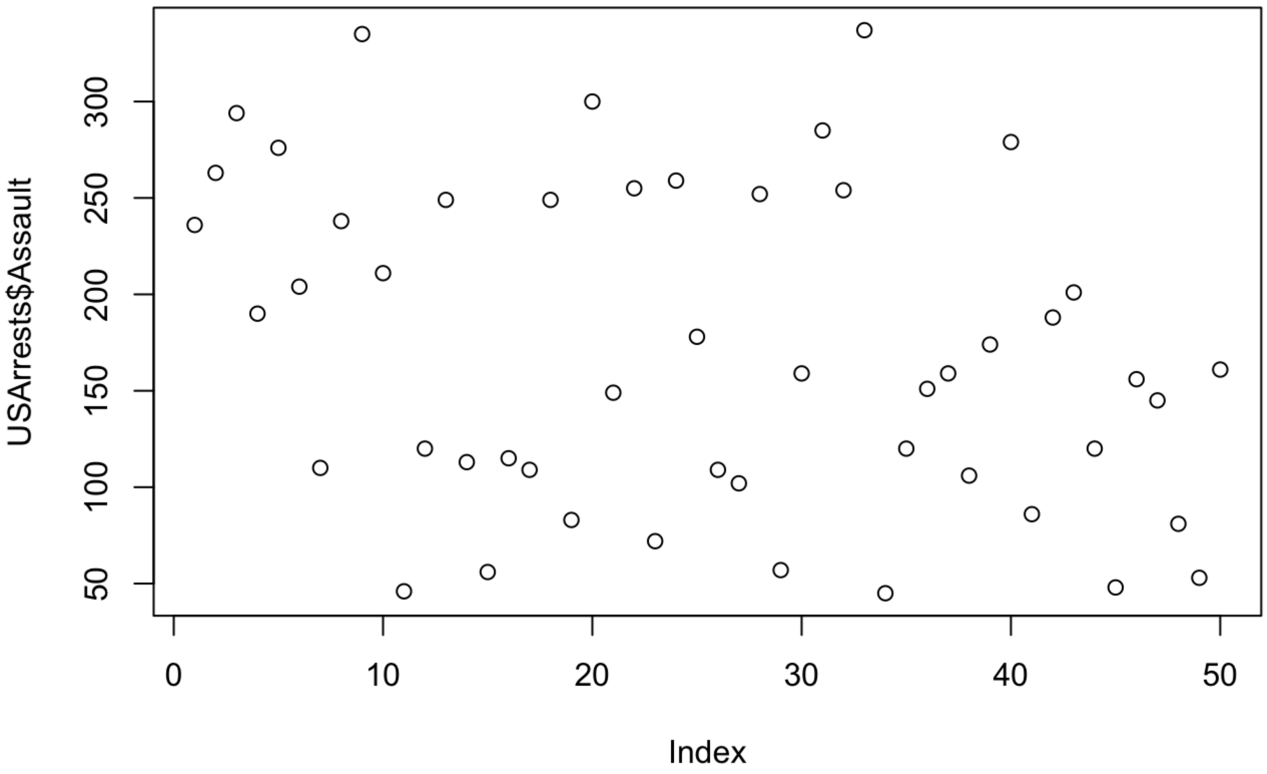
```
ggplot(data = USArrests) +
    geom_point(aes(x = attributes(USArrests)$row.names, y = Murder)) +
    theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
    xlab("State") +
    ylab("Murder Rate per 100,000 People")
```



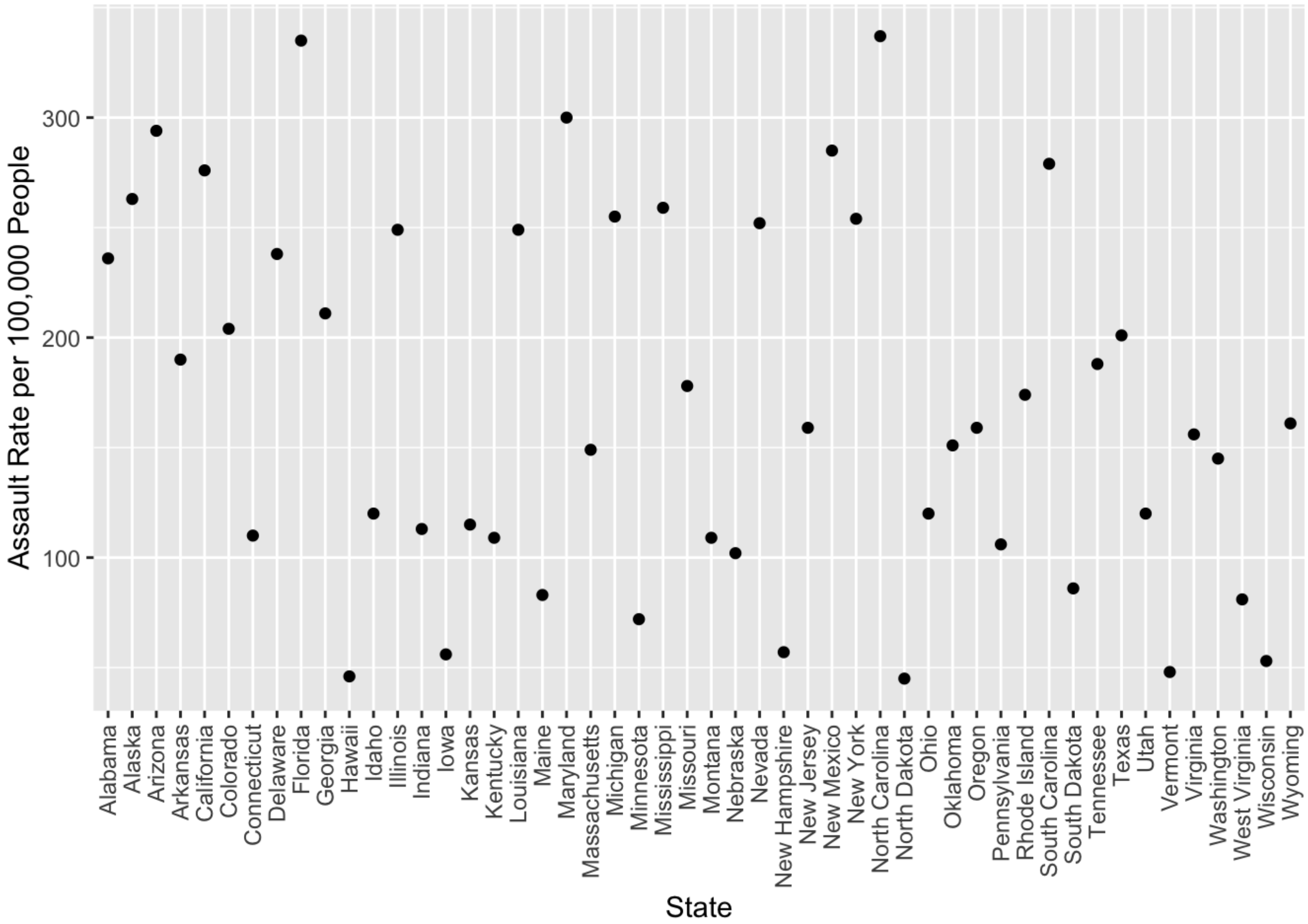
Seems to be a true catter with very little to no correlation.

Problem 1b) Create a graph of assault rate.

```
plot(USArrests$Assault)
```



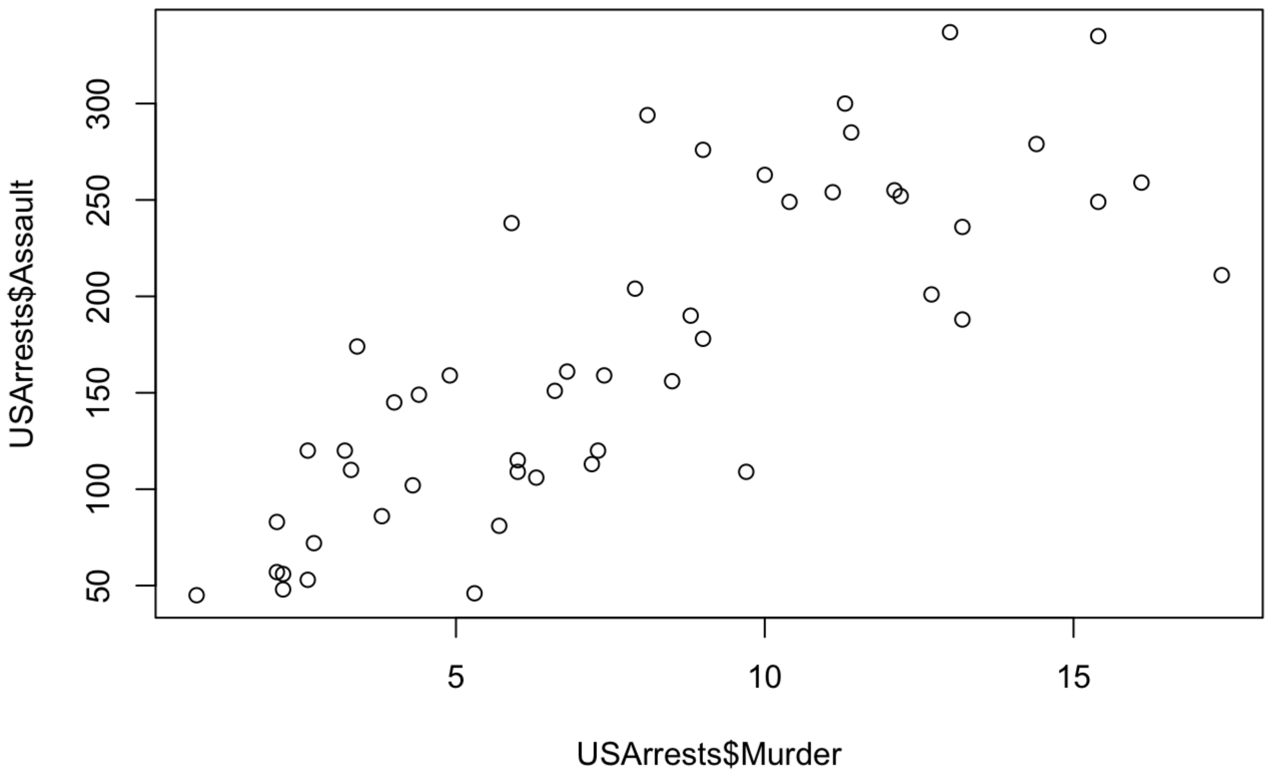
```
ggplot(data = USArrests) +
    geom_point(aes(x = attributes(USArrests)$row.names, y = Assault)) +
    theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
    xlab("State") +
    ylab("Assault Rate per 100,000 People")
```



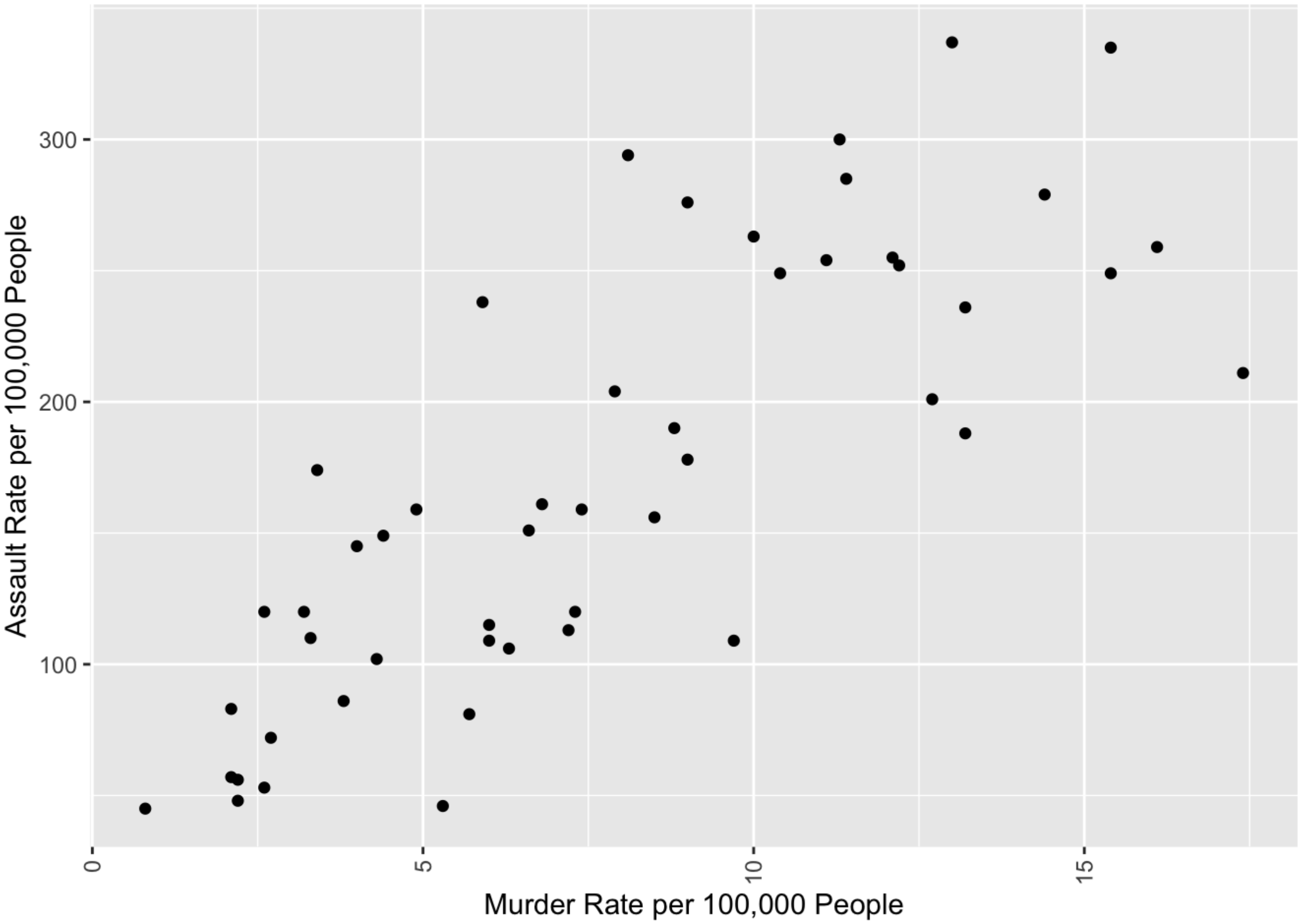
Seems to be a true catter with very little to no correlation.

Problem 1c) Create a scatterplot of murder rate (`x`) vs. assault rate (`y`).

```
plot(USArrests$Murder, USArrests$Assault)
```



```
ggplot(data = USArrests) +
    geom_point(aes(x = Murder, y = Assault)) +
    theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
    xlab("Murder Rate per 100,000 People") +
    ylab("Assault Rate per 100,000 People")
```

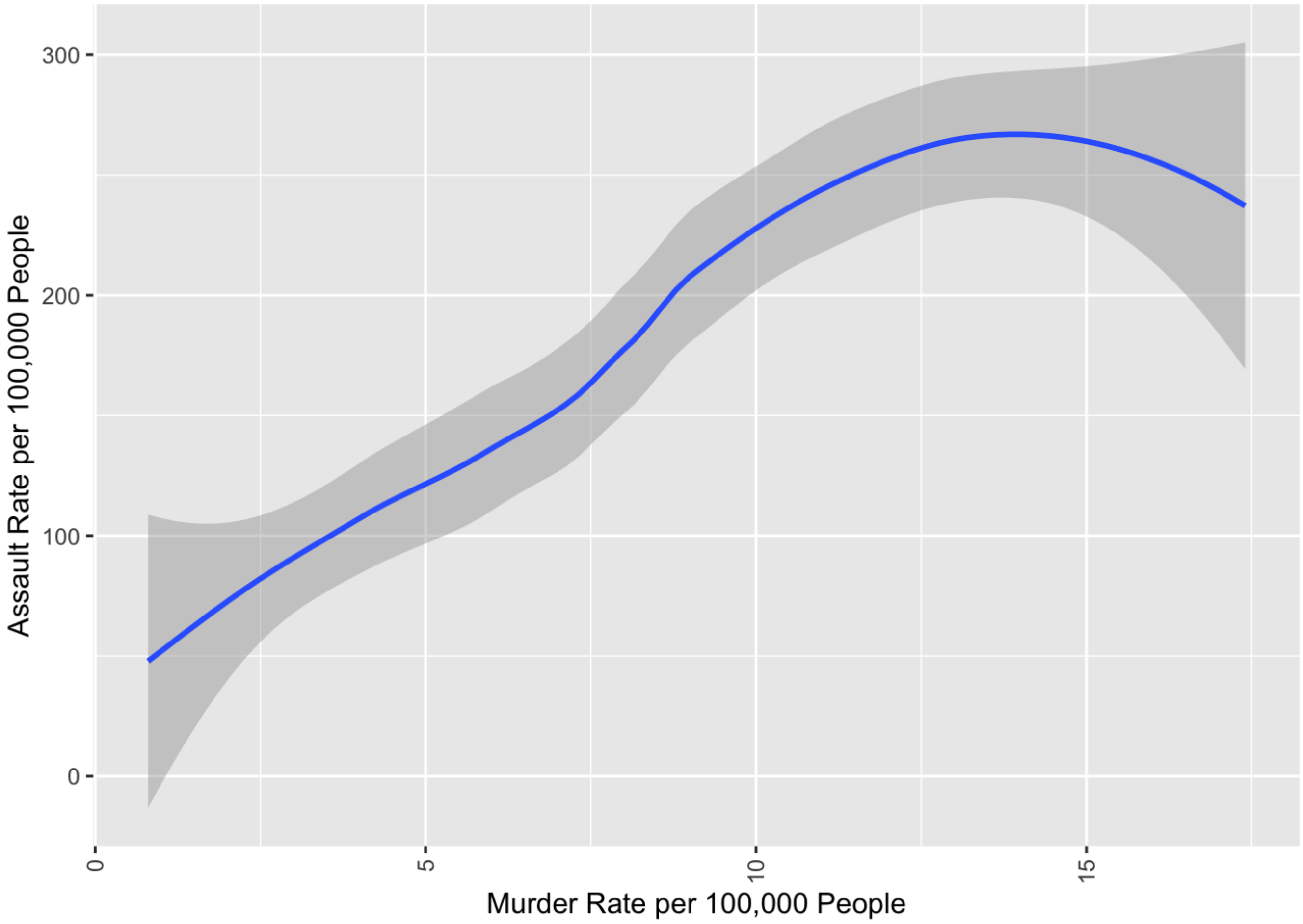


Seems to be a close to linear correlation between murder and assault rates

Problem 1d) Add a `geom_smooth` to the previous plot.

```
ggplot(data = USArrests) +
    geom_smooth(aes(x = Murder, y = Assault)) +
    theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
    xlab("Murder Rate per 100,000 People") +
    ylab("Assault Rate per 100,000 People")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

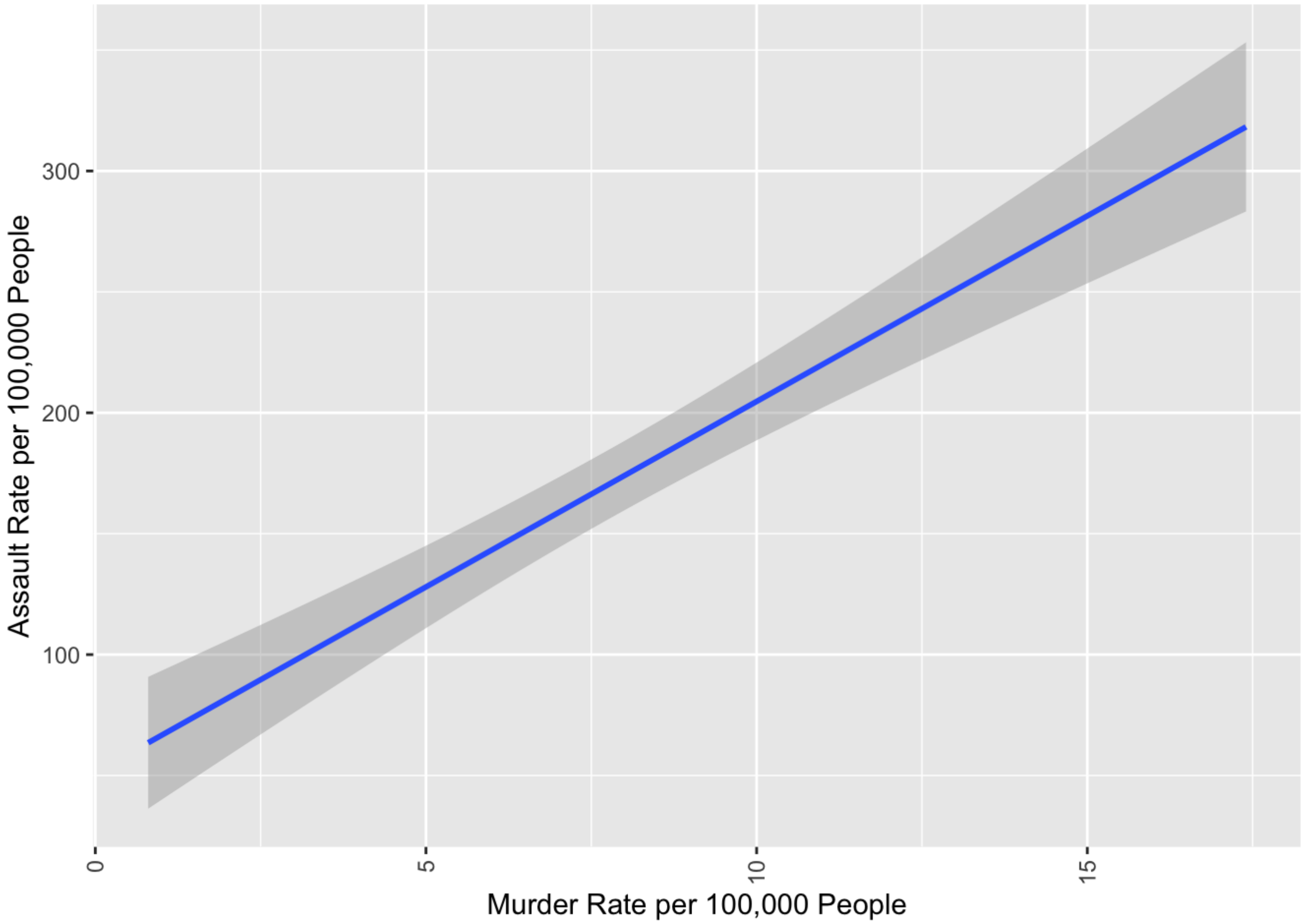


Seems that at very high murder rates the murder rates seem to still be increasing even when the assault rates begin to drop

Problem 1e) Add another `geom_smooth` with a `method = "lm"` argument to the previous plot. (You might want to make it a different color to distinguish it from the previous one.)

```
ggplot(data = USArrests) +
    geom_smooth(aes(x = Murder, y = Assault), method = "lm") +
    theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
    xlab("Murder Rate per 100,000 People") +
    ylab("Assault Rate per 100,000 People")
```

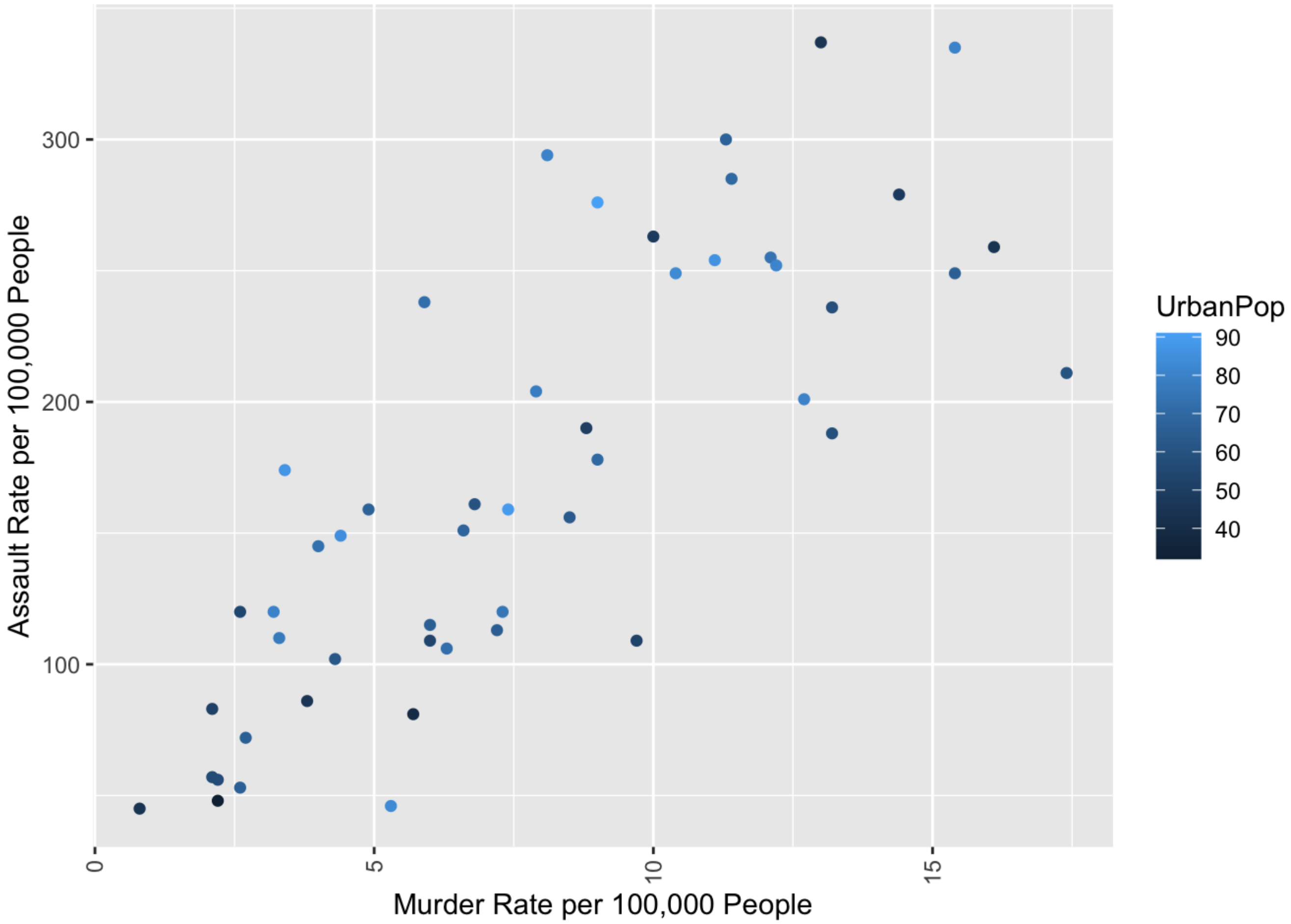
```
## `geom_smooth()` using formula 'y ~ x'
```



I do not believe that the linear model is appropriate for this data becasue at high murder rates, murder rates may not be a good predictor of the number of assaults that happened witha linear model

Problem 1d) Add `UrbanPop` as a color aesthetic to the scatterplot you created in 1c) above.

```
ggplot(data = USArrests) +
    geom_point(aes(x = Murder, y = Assault, color = UrbanPop)) +
    theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
    xlab("Murder Rate per 100,000 People") +
    ylab("Assault Rate per 100,000 People")
```



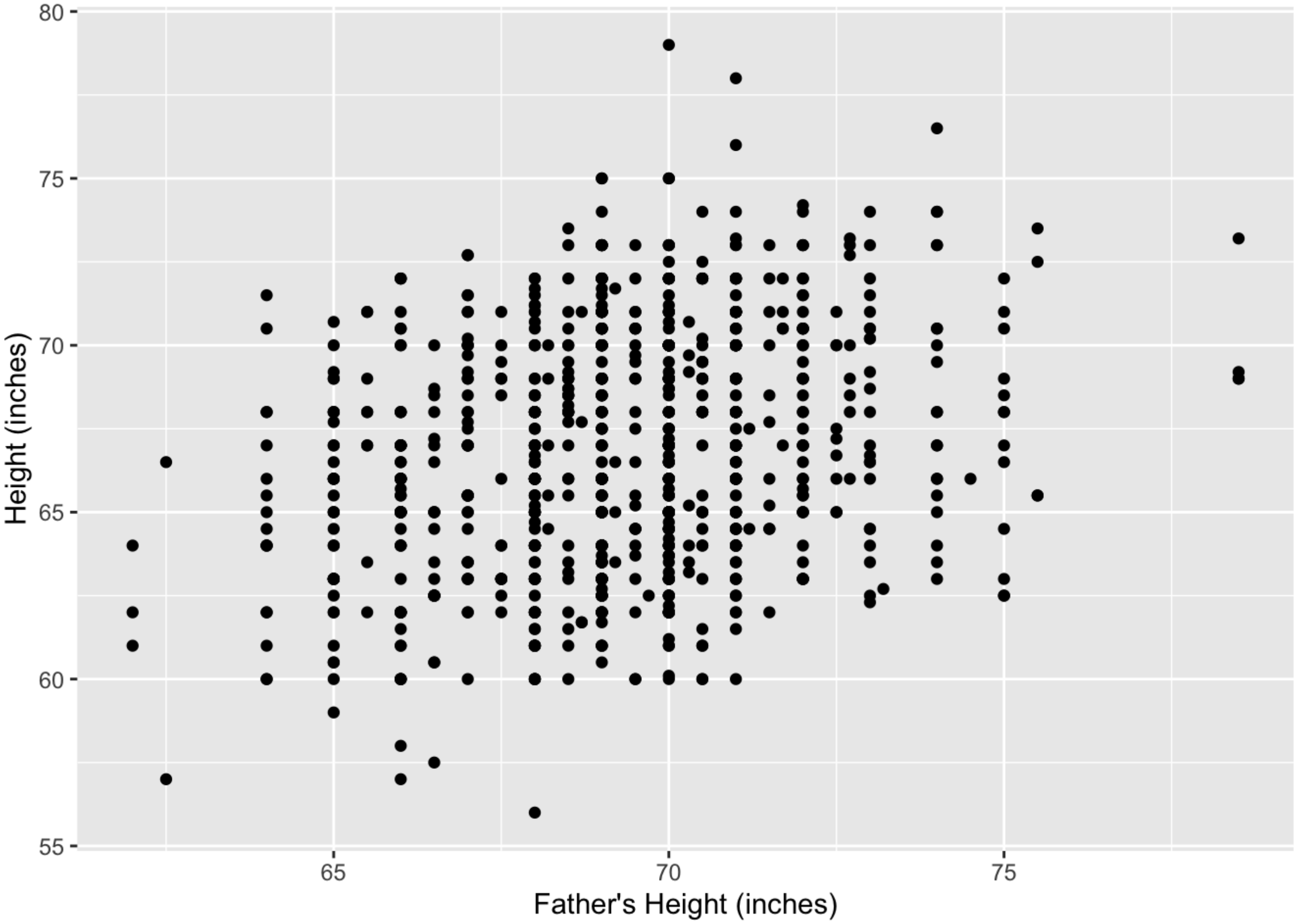
Generally for low murder and assault rates the urban population seems to be low.

Problem 2) Galton Height Data

The Galton data can be found in the `mosaicData` package. The data contains `height` information on all children within a `family`, including the child's `sex` (gender), the `mother`'s height, the `father`'s height, and the number of kids (`nkids`) in the family. Height is measured in inches.

Problem 2a) Create a scatterplot of each person's height against their Father's height. (As typically done, please put the independent variable on the x-axis)

```
ggplot(data = Galton) +
    geom_point(aes(x = father, y = height)) +
    xlab("Father's Height (inches)") +
    ylab("Height (inches)")
```

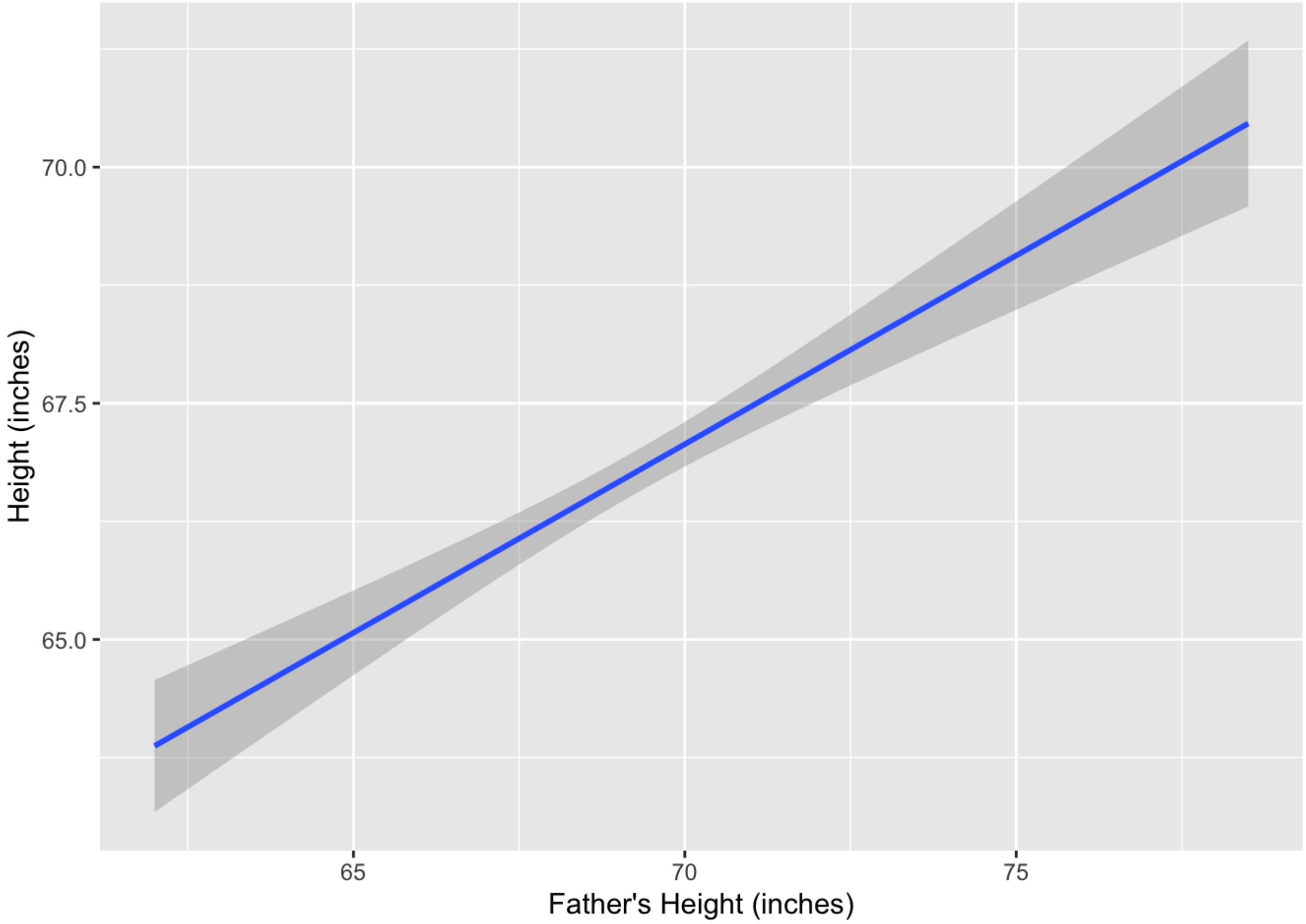


There seems to be weak correaltion but generally a taller father seems to have taller children

Problem 2b) Add a regression line to the previous plot by using `geom_smooth(method = "lm")`.

```
ggplot(data = Galton) +
    geom_smooth(aes(x = father, y = height), method = "lm") +
    xlab("Father's Height (inches)") +
    ylab("Height (inches)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

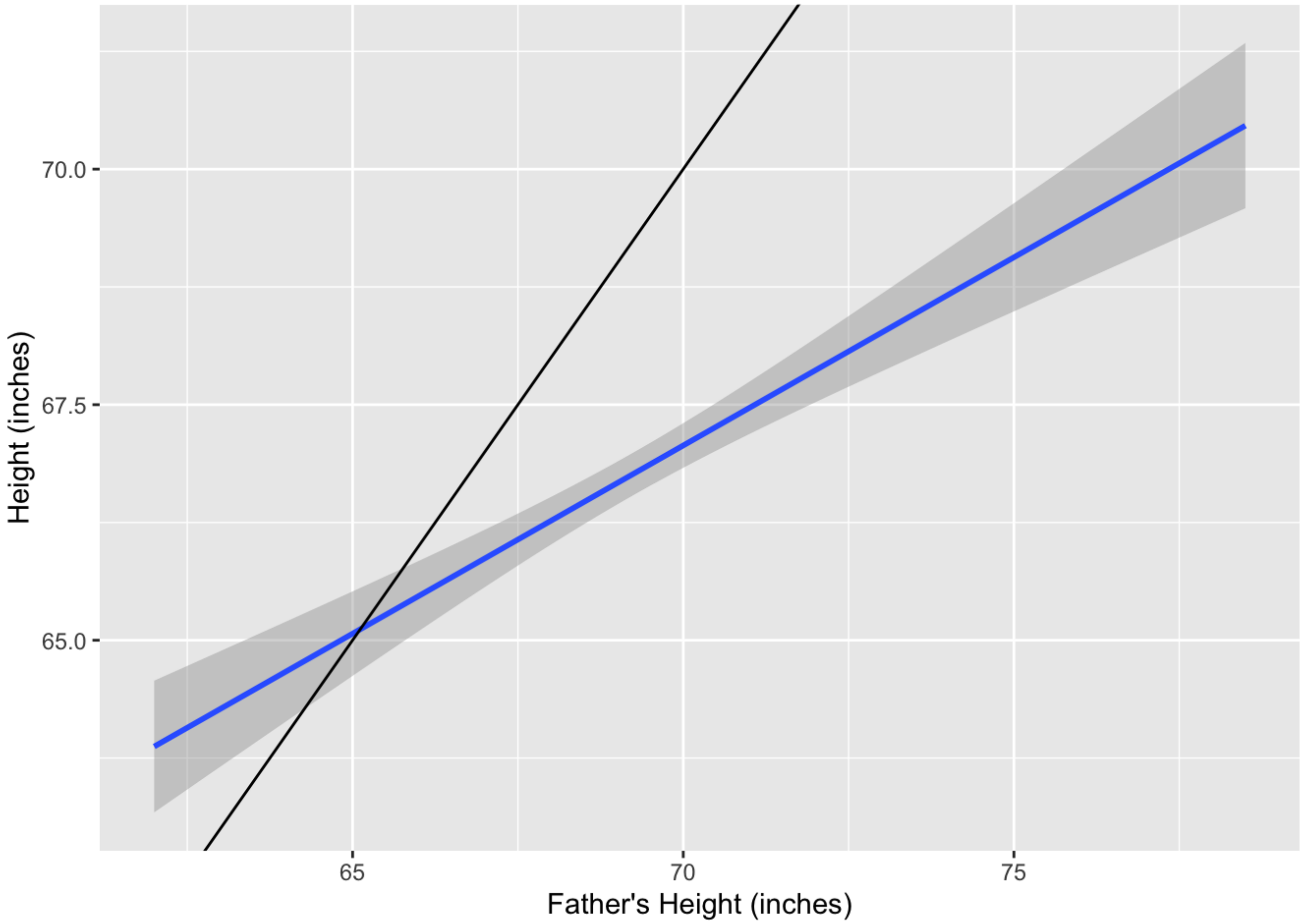


Seems to be a positive correlation where childrean are shorther than ther fathers.

Problem 2c) Add a y = x line to the previous plot using `geom_abline(intercept = 0,slope = 1)`

```
ggplot(data = Galton) +
    geom_smooth(aes(x = father, y = height), method = "lm") +
    xlab("Father's Height (inches)") +
    ylab("Height (inches)")+
    geom_abline(intercept = 0,slope = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

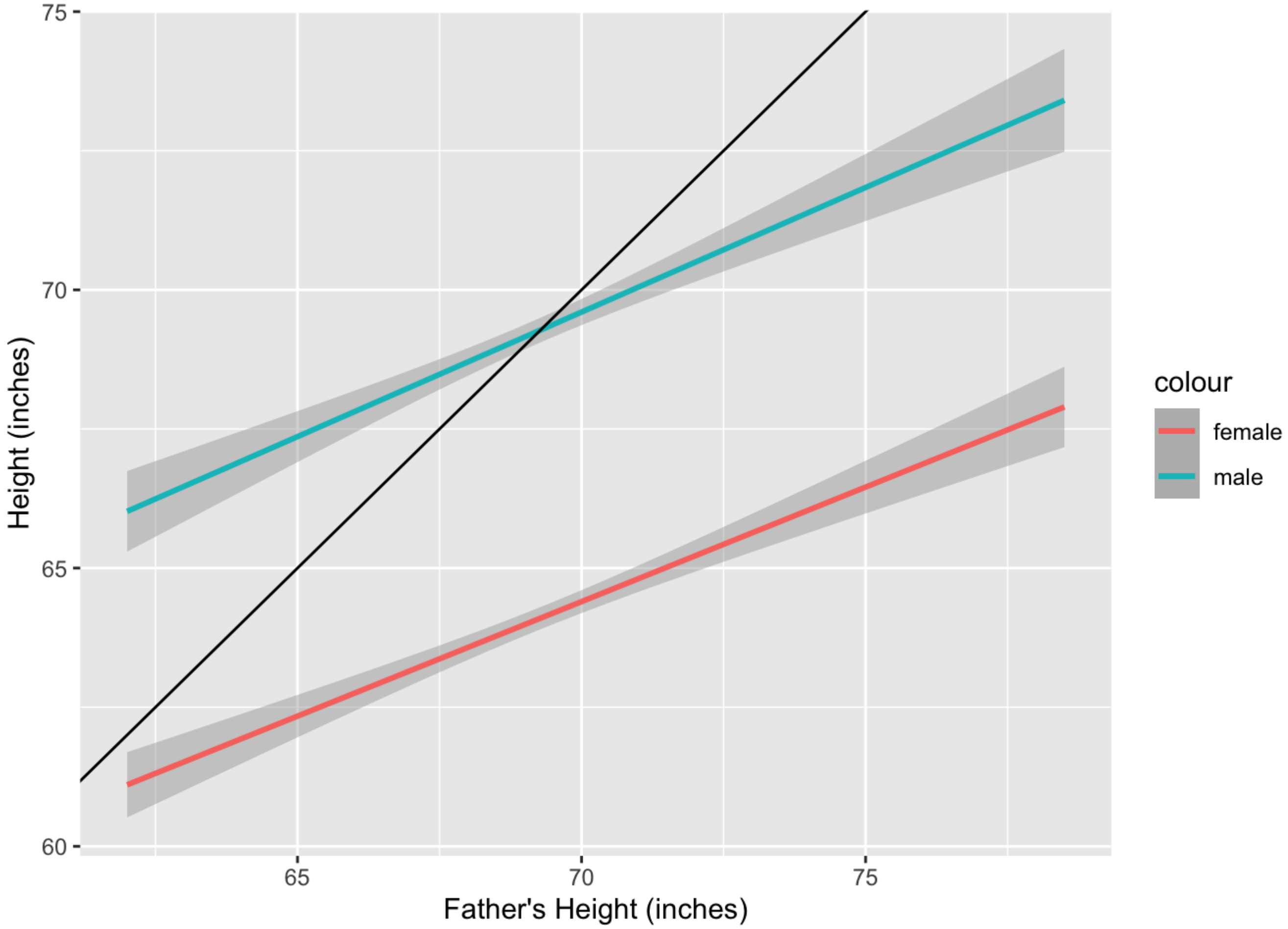


The slope of the regression line is lower than the slope of y = x

Problem 2d) Add `sex` as an aesthethic to the previous plot. Make sure you have two separate regression lines, one per each `sex`.

```
ggplot() +
    geom_smooth(data = Galton[Galton$sex == "M",], aes(x = father, y = height, color =
"male"), method = "lm") +
    geom_smooth(data = Galton[Galton$sex == "F",], aes(x = father, y = height, color =
"female"), method = "lm") +
    xlab("Father's Height (inches)") +
    ylab("Height (inches)")+
    geom_abline(intercept = 0,slope = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



The correlation for both seems to be the same slope however female heights are generally lower than male heights.