assignment_04_Chattapadhyay_Kausik.R

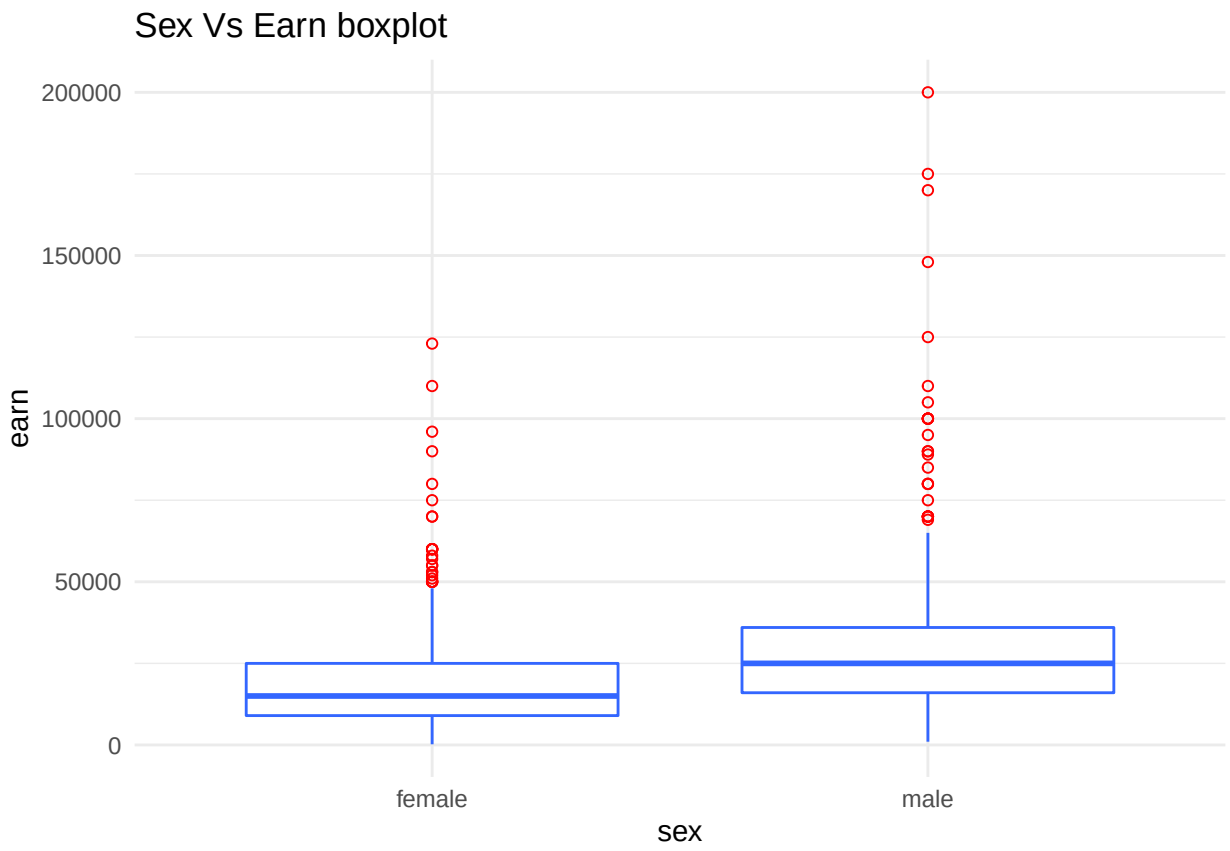
kausik

2022-10-04

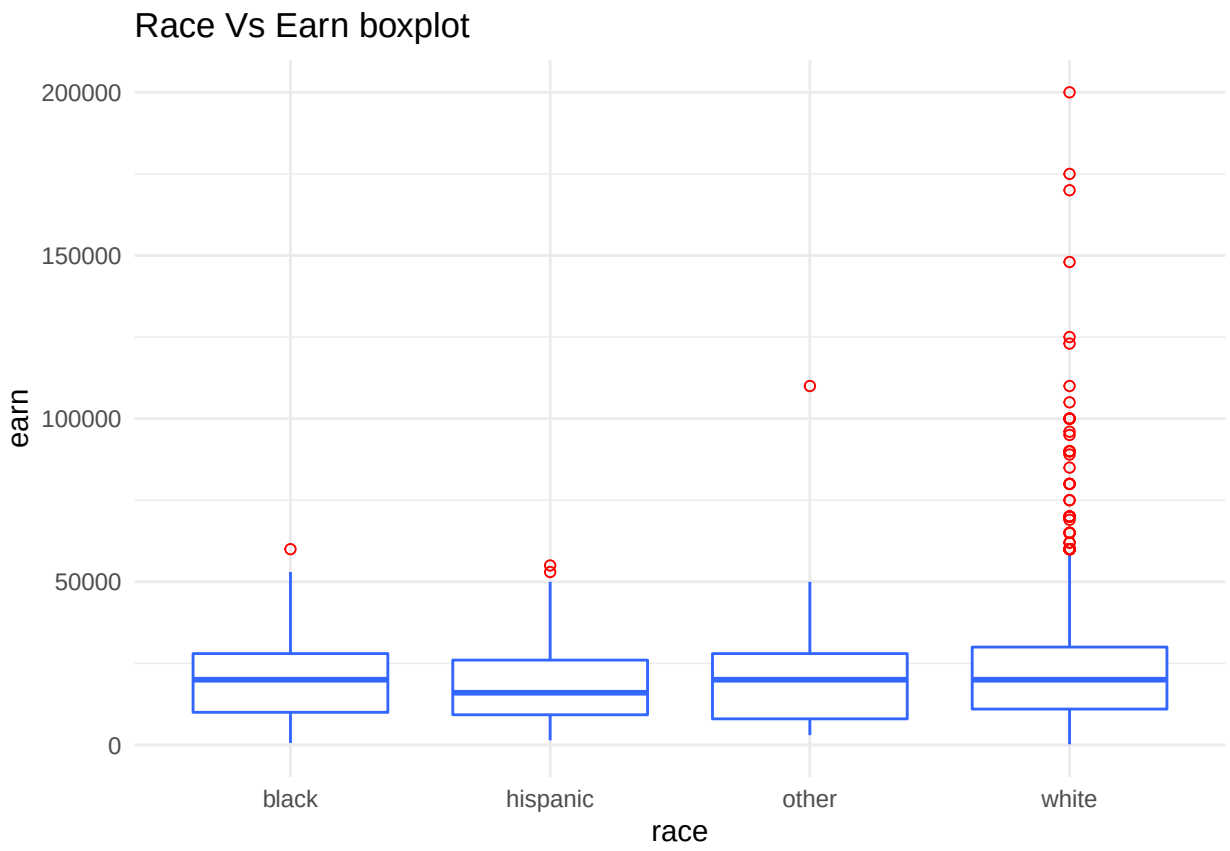
```
# Assignment: ASSIGNMENT 4
# Name: Chattapadhyay, Kausik
# Date: 2022-10-04

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

## Set the working directory to the root of your DSC 520 directory
setwd("/Users/kausik/desktop/MS Data Science/DSC 520/dsc520-stats-r-assignments")

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

```
##    earn   height    sex ed age  race
## 1 50000 74.42444   male 16  45 white
## 2 60000 65.53754 female 16  58 white
## 3 30000 63.62920 female 16  29 white
## 4 50000 63.10856 female 16  91 other
## 5 51000 63.40248 female 17  39 white
## 6  9000 64.39951 female 15  26 white
```

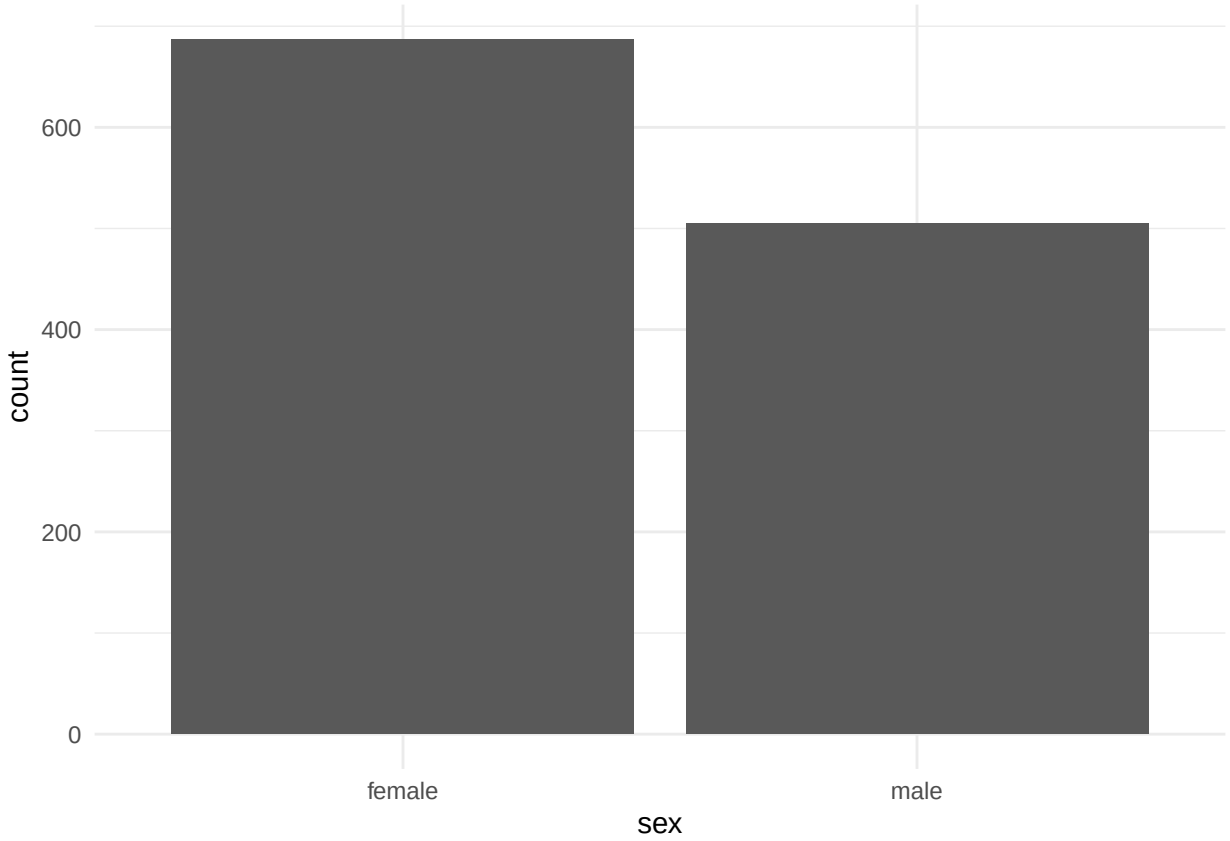
```
# https://ggplot2.tidyverse.org/reference/geom_boxplot.html
## Create boxplots of sex vs. earn and race vs. earn using `geom_point()` and `geom_boxplot()`
## sex vs. earn
ggplot(heights_df, aes(x=sex, y=earn)) +
  geom_boxplot(outlier.colour = "red",outlier.shape = 1, fill = "white",
               colour = "#3366FF") + labs(title="Sex Vs Earn boxplot")
```



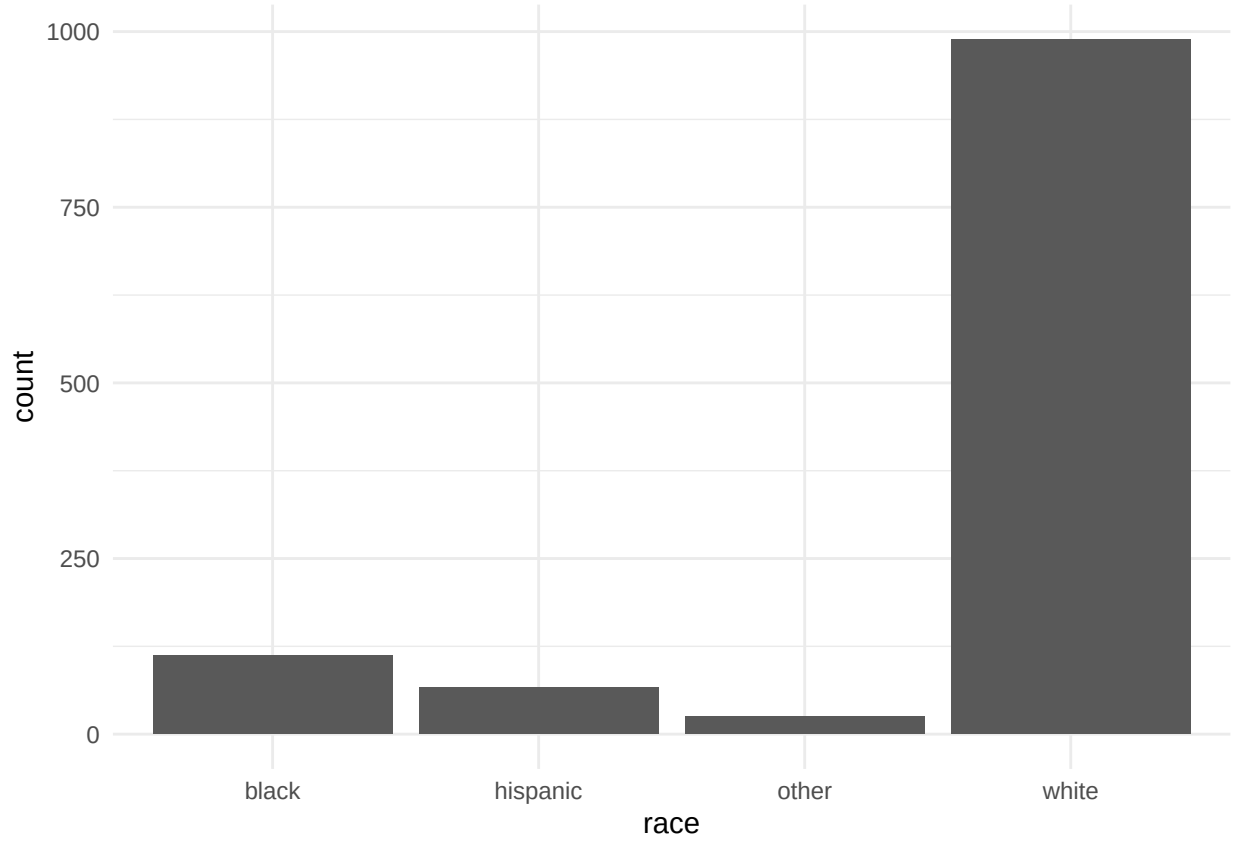

```
## race vs. earn
ggplot(heights_df, aes(x=race, y=earn)) + geom_boxplot(outlier.colour = "red",
                                    outlier.shape = 1, fill = "white",
                                    colour = "#3366FF") +
  labs(title="Race Vs Earn boxplot")
```



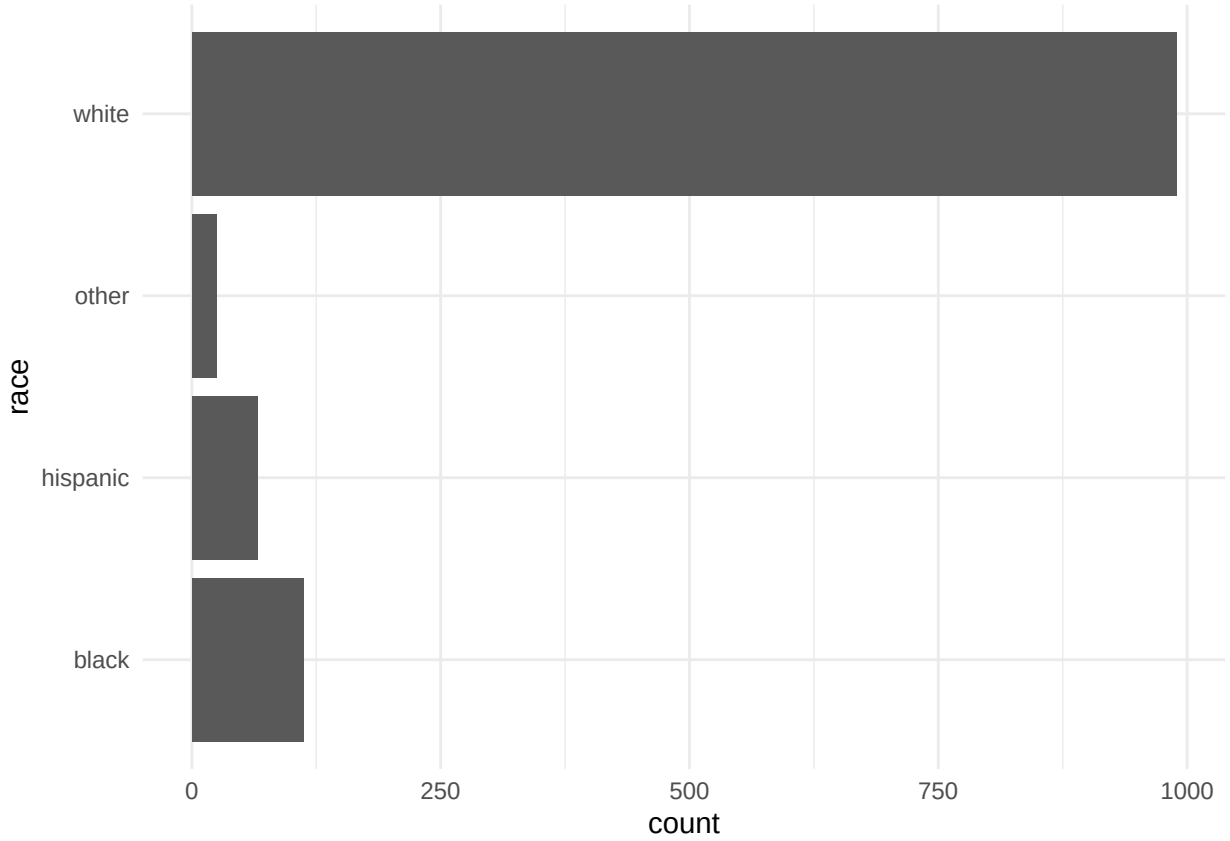

```
# https://ggplot2.tidyverse.org/reference/geom_bar.html
## Using `geom_bar()` plot a bar chart of the number of records for each `sex`
ggplot(heights_df, aes(sex)) + geom_bar()
```




```
## Using `geom_bar()` plot a bar chart of the number of records for each race
ggplot(heights_df, aes(race)) + geom_bar()
```




```
## Create a horizontal bar chart by adding `coord_flip()` to the previous plot
ggplot(heights_df, aes(race)) + geom_bar() + coord_flip()
```




```
# https://www.rdocumentation.org/packages/ggplot2/versions/3.3.0/topics/geom_path
## Load the file `"data/nytimes/covid-19-data/us-states.csv"` and
## assign it to the `covid_df` dataframe
covid_df <- read.csv("data/nytimes/covid-19-data/us-states.csv")
str(covid_df)
```

```
## 'data.frame':    3039 obs. of  5 variables:
##  $ date  : chr  "2020-01-21" "2020-01-22" "2020-01-23" "2020-01-24" ...
##  $ state : chr  "Washington" "Washington" "Washington" "Illinois" ...
##  $ fips  : int  53 53 53 17 53 6 17 53 4 6 ...
##  $ cases : int  1 1 1 1 1 1 1 1 1 2 ...
##  $ deaths: int  0 0 0 0 0 0 0 0 0 0 ...
```

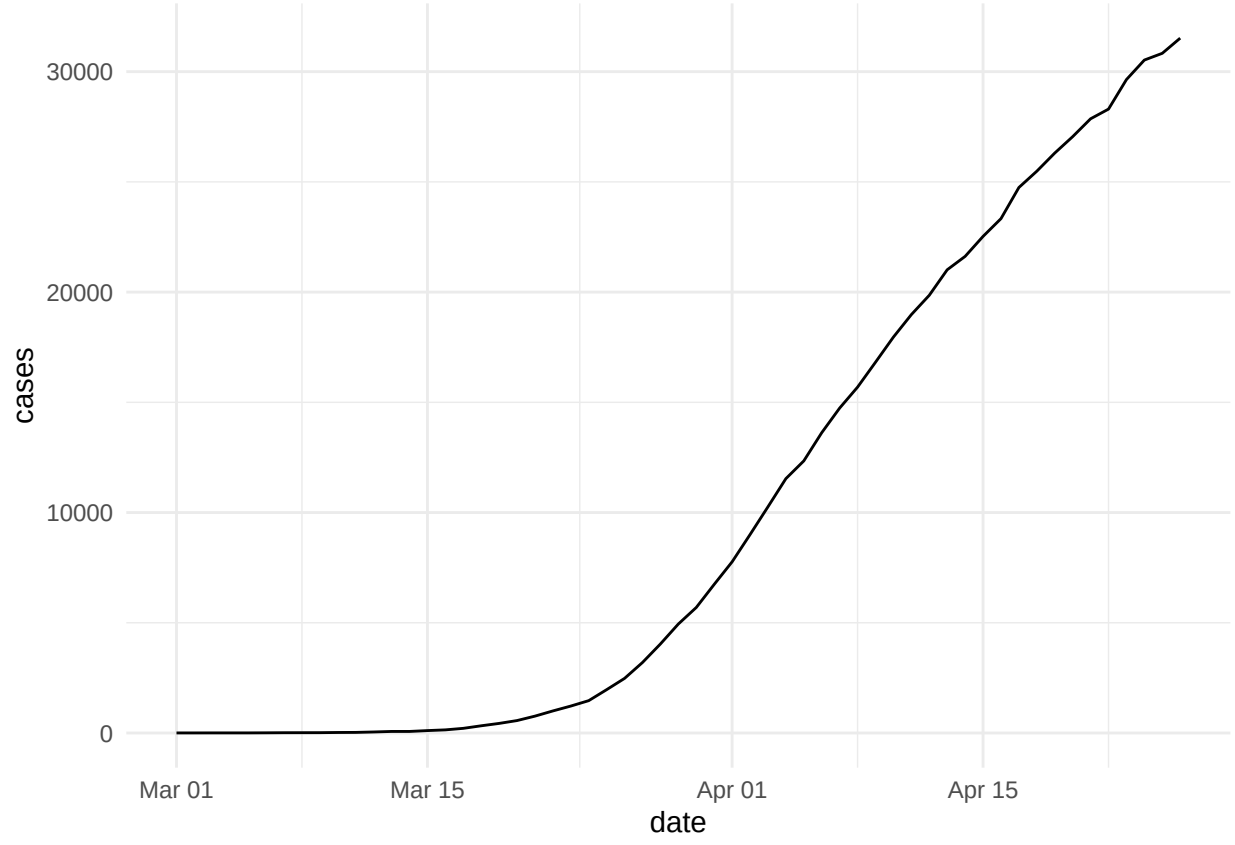
```
## Parse the date column using `as.Date()``
covid_df$date <- as.Date(covid_df$date)
str(covid_df)
```

```
## 'data.frame':    3039 obs. of  5 variables:
##  $ date  : Date, format: "2020-01-21" "2020-01-22" ...
##  $ state : chr  "Washington" "Washington" "Washington" "Illinois" ...
##  $ fips  : int  53 53 53 17 53 6 17 53 4 6 ...
##  $ cases : int  1 1 1 1 1 1 1 1 1 2 ...
##  $ deaths: int  0 0 0 0 0 0 0 0 0 0 ...
```

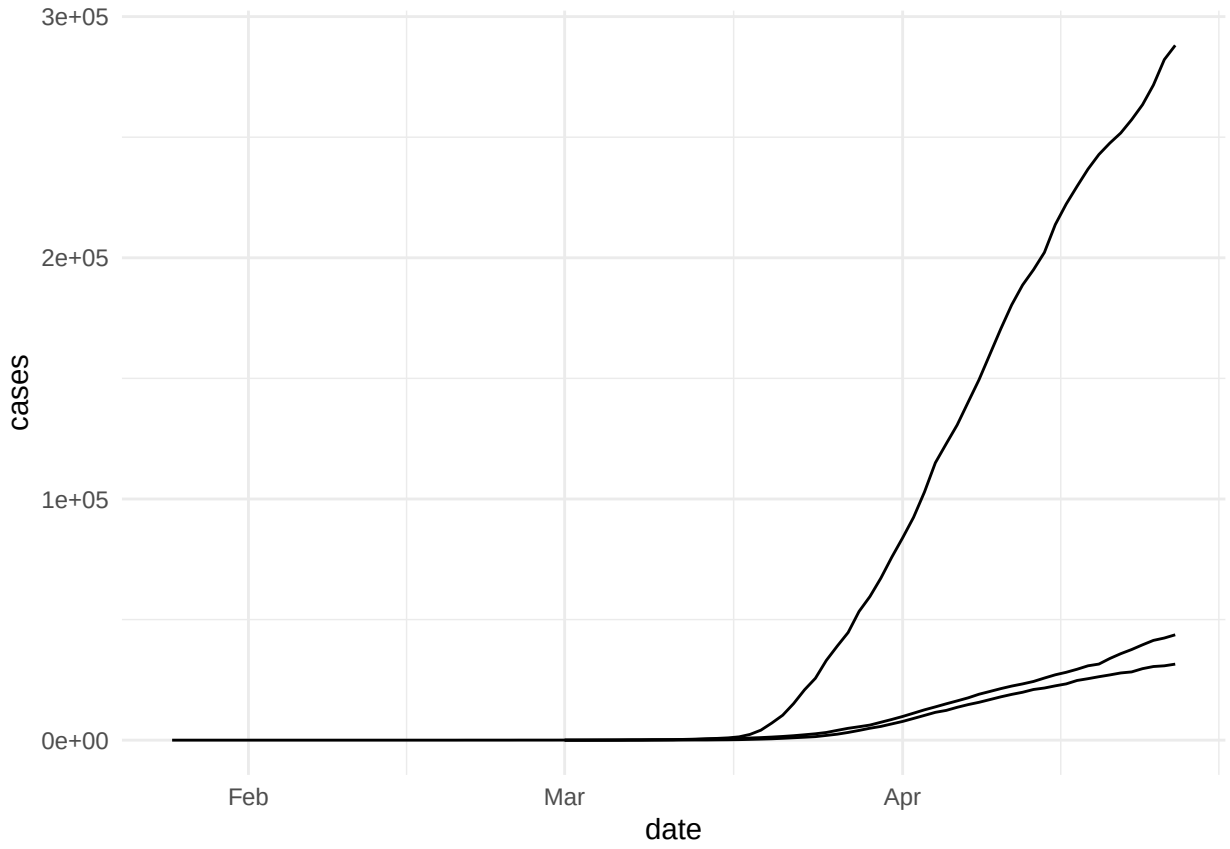
```
## Create three dataframes named `california_df`, `ny_df`, and `florida_df`
## containing the data from California, New York, and Florida
california_df <- covid_df[ which( covid_df$state == "California"), ]
ny_df <- covid_df[which(covid_df$state == "New York"), ]
florida_df <- covid_df[which(covid_df$state == "Florida"), ]
head(florida_df)
```

```
##           date   state fips cases deaths
## 243 2020-03-01 Florida   12     2      0
## 256 2020-03-02 Florida   12     2      0
## 271 2020-03-03 Florida   12     3      0
## 287 2020-03-04 Florida   12     3      0
## 305 2020-03-05 Florida   12     4      0
## 326 2020-03-06 Florida   12     7      2
```

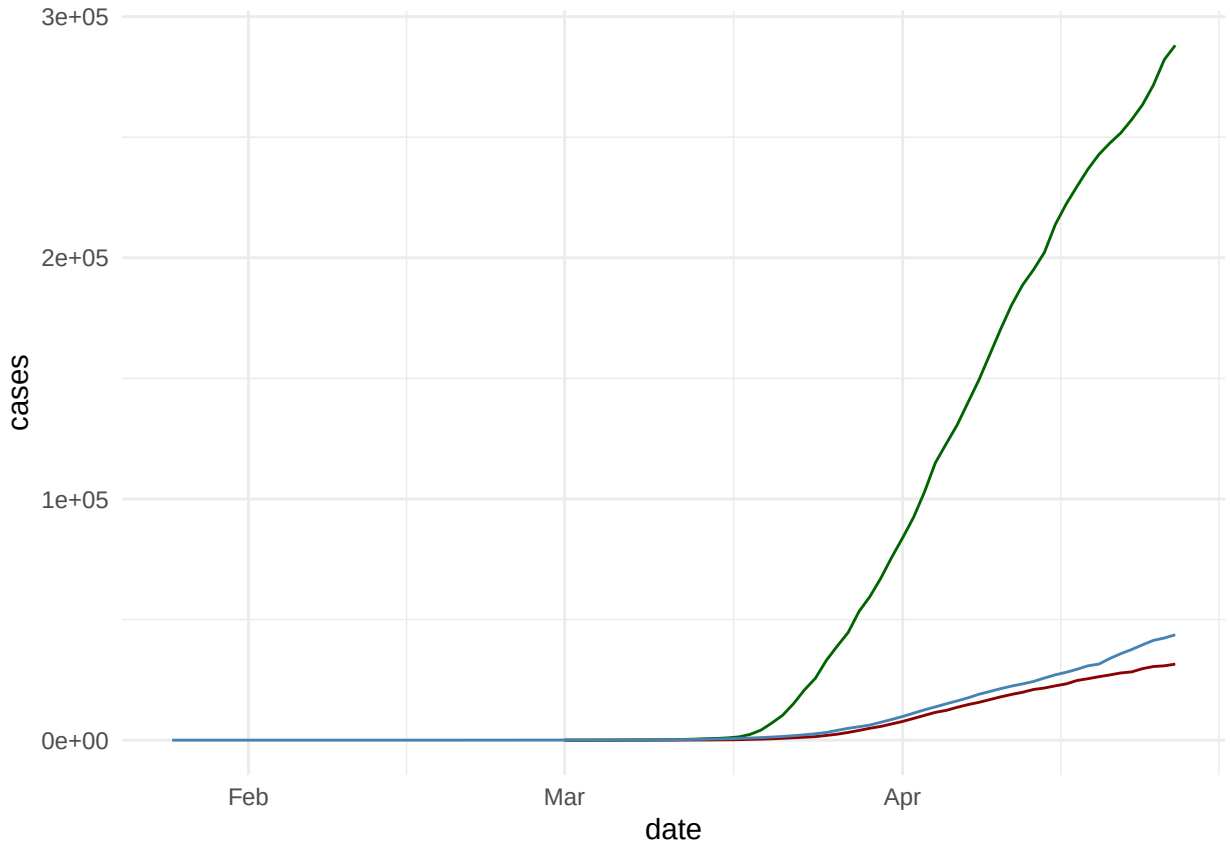
```
## Plot the number of cases in Florida using `geom_line()`
ggplot(data=florida_df, aes(x=date, y=cases, group=1)) + geom_line()
```



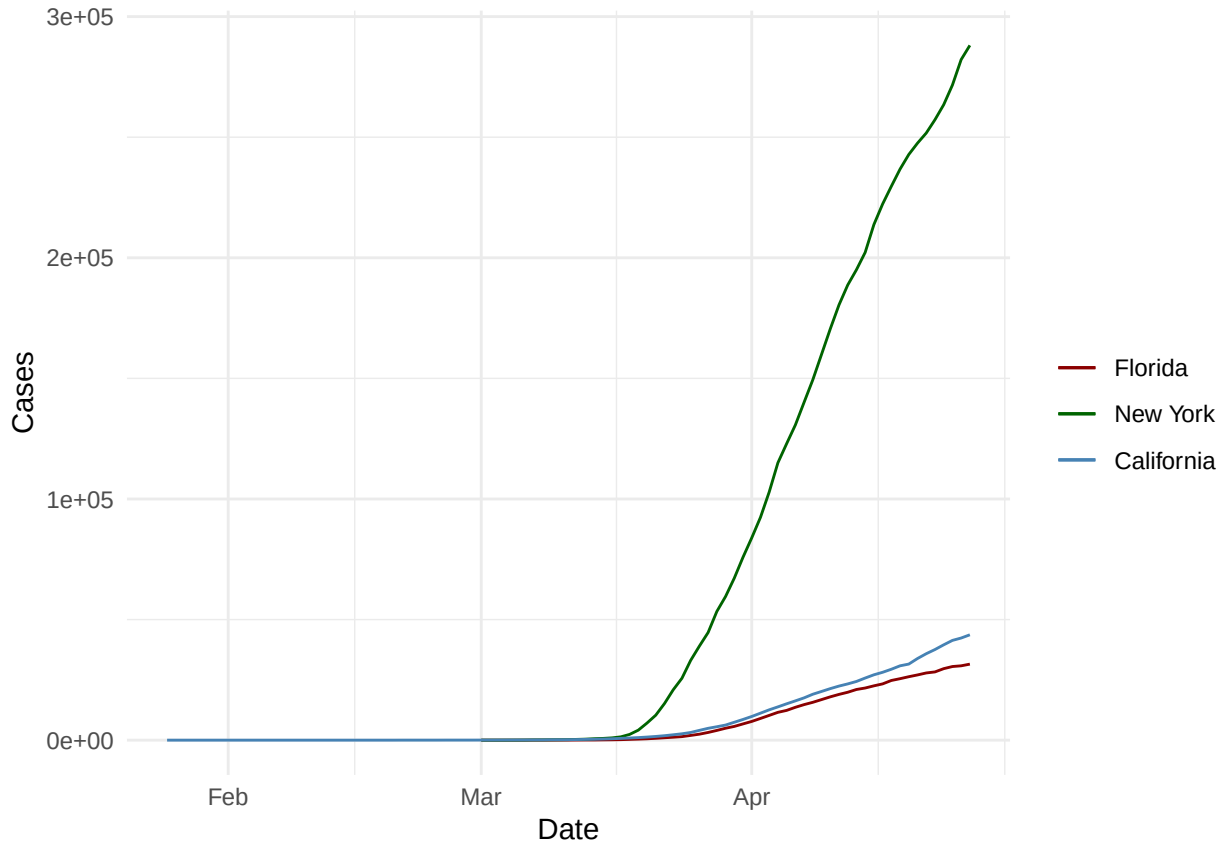

```
## Add lines for New York and California to the plot
ggplot(data=florida_df, aes(x=date, group=1)) +
  geom_line(aes(y = cases)) +
  geom_line(data=california_df, aes(y = cases)) +
  geom_line(data=ny_df, aes(y = cases))
```



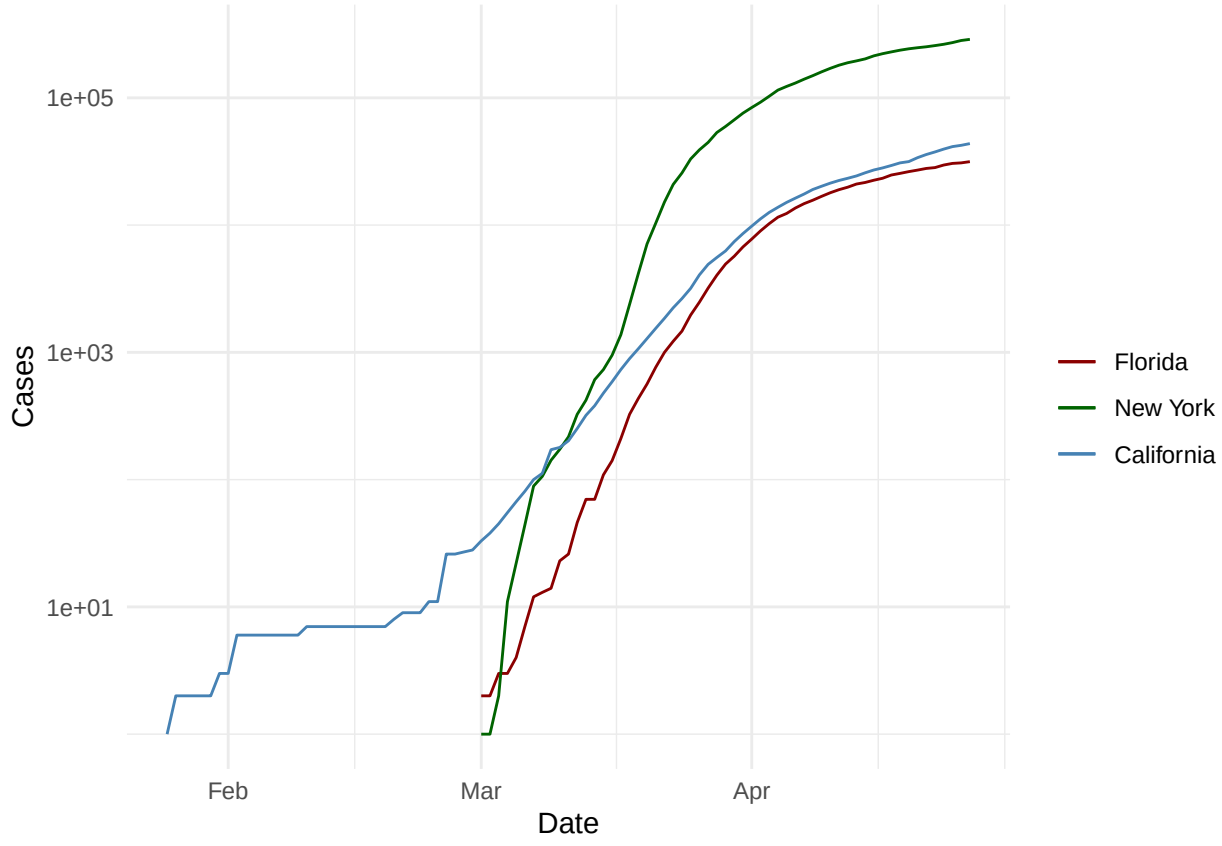

```
## Use the colors "darkred", "darkgreen", and "steelblue" for Florida, New York, and California
ggplot(data=florida_df, aes(x=date, group=1)) +
  geom_line(aes(y = cases), color = "darkred") +
  geom_line(data=ny_df, aes(y = cases), color="darkgreen") +
  geom_line(data=california_df, aes(y = cases), color="steelblue")
```




```
## Add a legend to the plot using `scale_colour_manual`
## Add a blank (" ") label to the x-axis and the label "Cases" to the y axis
ggplot(data=florida_df, aes(x=date, group=1)) +
  geom_line(aes(y = cases, colour = "Florida")) +
  geom_line(data=ny_df, aes(y = cases,colour="New York")) +
  geom_line(data=california_df, aes(y = cases, colour="California")) +
  scale_colour_manual("",
                      breaks = c("Florida", "New York", "California"),
                      values = c("darkred", "darkgreen", "steelblue")) +
  xlab("Date") + ylab("Cases")
```




```
## Scale the y axis using `scale_y_log10()`
ggplot(data=florida_df, aes(x=date, group=1)) +
  geom_line(aes(y = cases, colour = "Florida")) +
  geom_line(data=ny_df, aes(y = cases,colour="New York")) +
  geom_line(data=california_df, aes(y = cases, colour="California")) +
  scale_colour_manual("",
                      breaks = c("Florida", "New York", "California"),
                      values = c("darkred", "darkgreen", "steelblue")) +
  xlab("Date") + ylab("Cases") + scale_y_log10()
```


1e+05
1e+03
1e+01
Cases
Feb
Mar
Apr
Date
Florida
New York
California