assignment_04_MunjewarSheetalR.R

sheetal

2023-01-22

```
# Assignment: ASSIGNMENT 4
# Name: Munjewar, Sheetal
# Date: 2023-01-22

# Load the ggplot2 package
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")
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
```

```
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
```

```r
# factor(heights_df$sex)
# To check the structure
str(heights_df)
```

```
## 'data.frame':    1192 obs. of  6 variables:
##  $ earn  : num  50000 60000 30000 50000 51000 9000 29000 32000 2000 27000 ...
##  $ height: num  74.4 65.5 63.6 63.1 63.4 ...
##  $ sex   : chr  "male" "female" "female" "female" ...
##  $ ed    : int  16 16 16 16 17 15 12 17 15 12 ...
##  $ age   : int  45 58 29 91 39 26 49 46 21 26 ...
##  $ race  : chr  "white" "white" "white" "other" ...
```

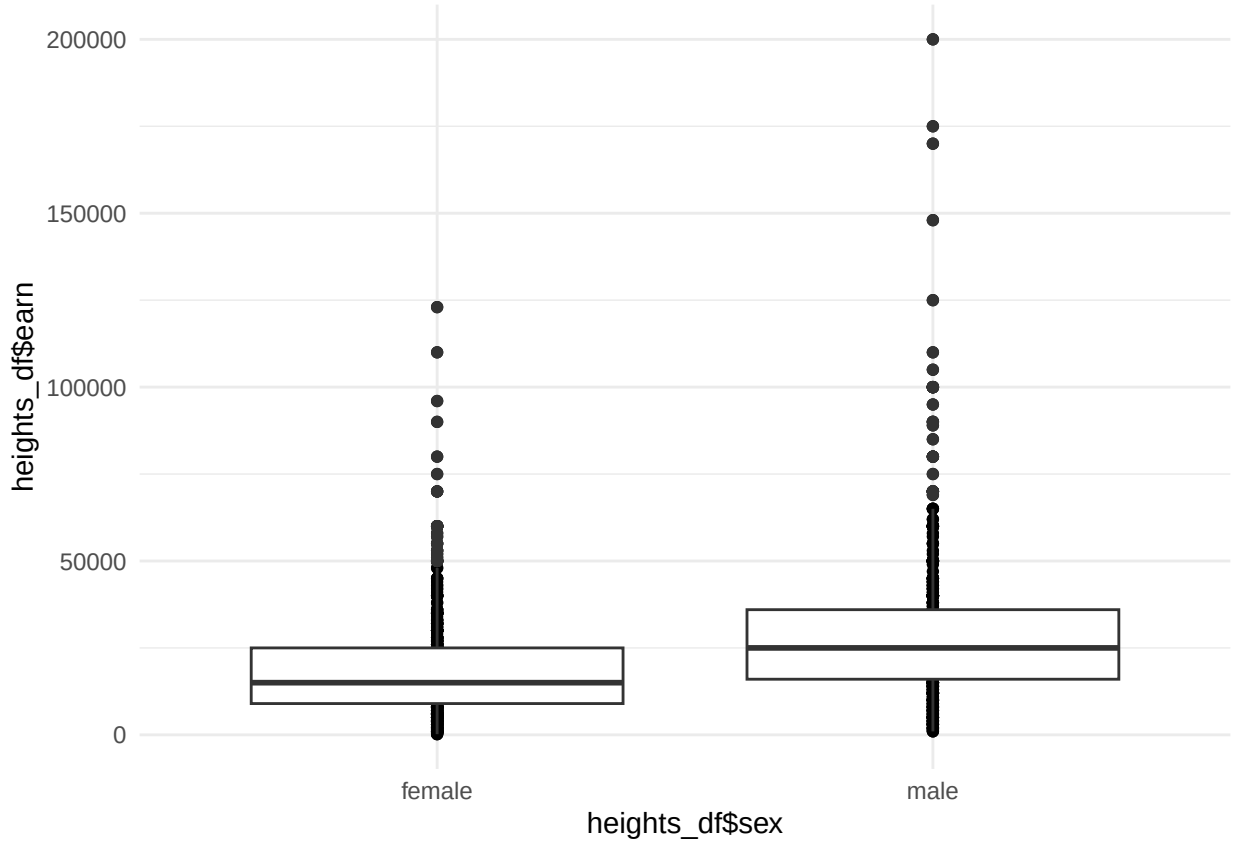
```r
# https://ggplot2.tidyverse.org/reference/geom_boxplot.html
# Create boxplots of sex vs. earn and race vs. earn using `geom_point()` and `geom_boxplot()`
# sex vs. earn
A <- ggplot(heights_df, aes(x=heights_df$sex, y=heights_df$earn))
A + geom_point() + geom_boxplot()
```

```
## Warning: Use of ‘heights_df$sex‘ is discouraged.
## i Use ‘sex‘ instead.
```

```
## Warning: Use of ‘heights_df$earn‘ is discouraged.
## i Use ‘earn‘ instead.
```

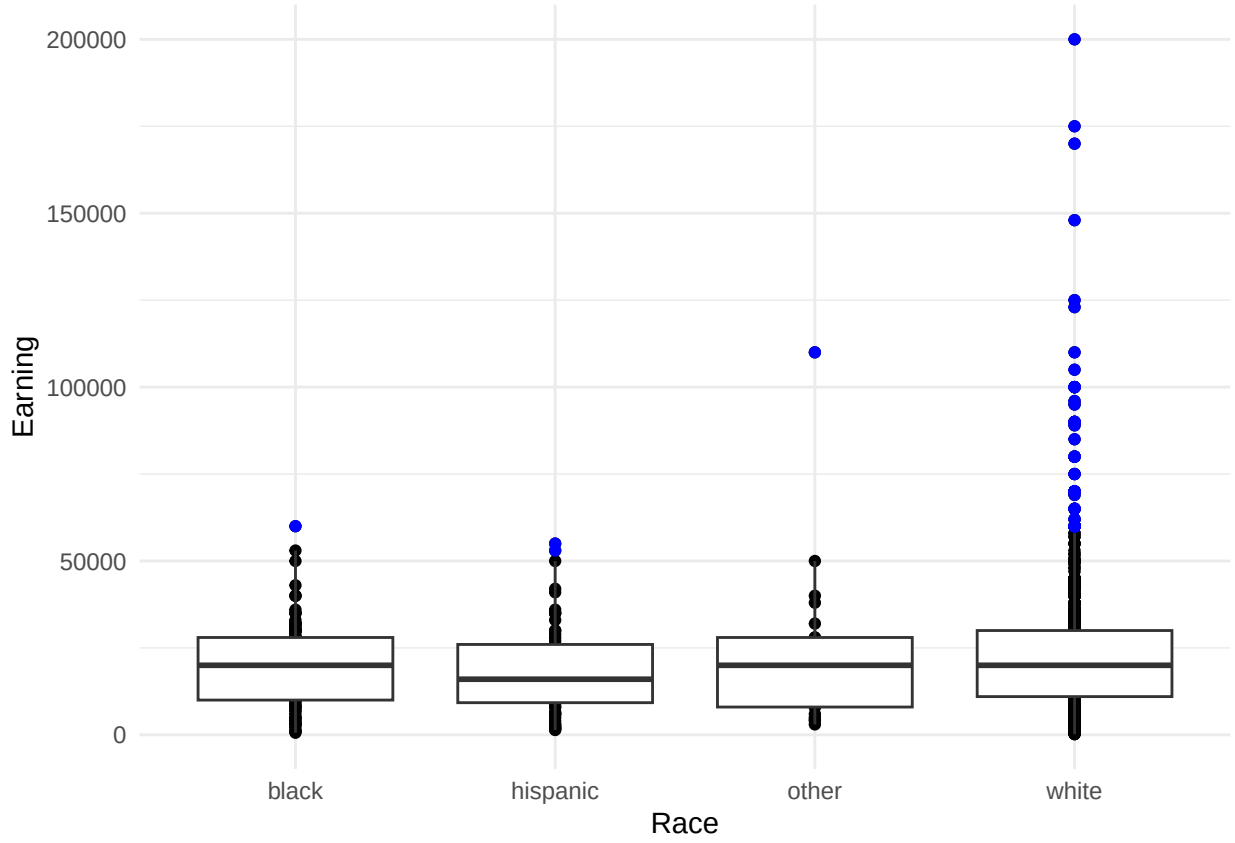
```
## Warning: Use of ‘heights_df$sex‘ is discouraged.
## i Use ‘sex‘ instead.
```

```
## Warning: Use of ‘heights_df$earn‘ is discouraged.
## i Use ‘earn‘ instead.
```



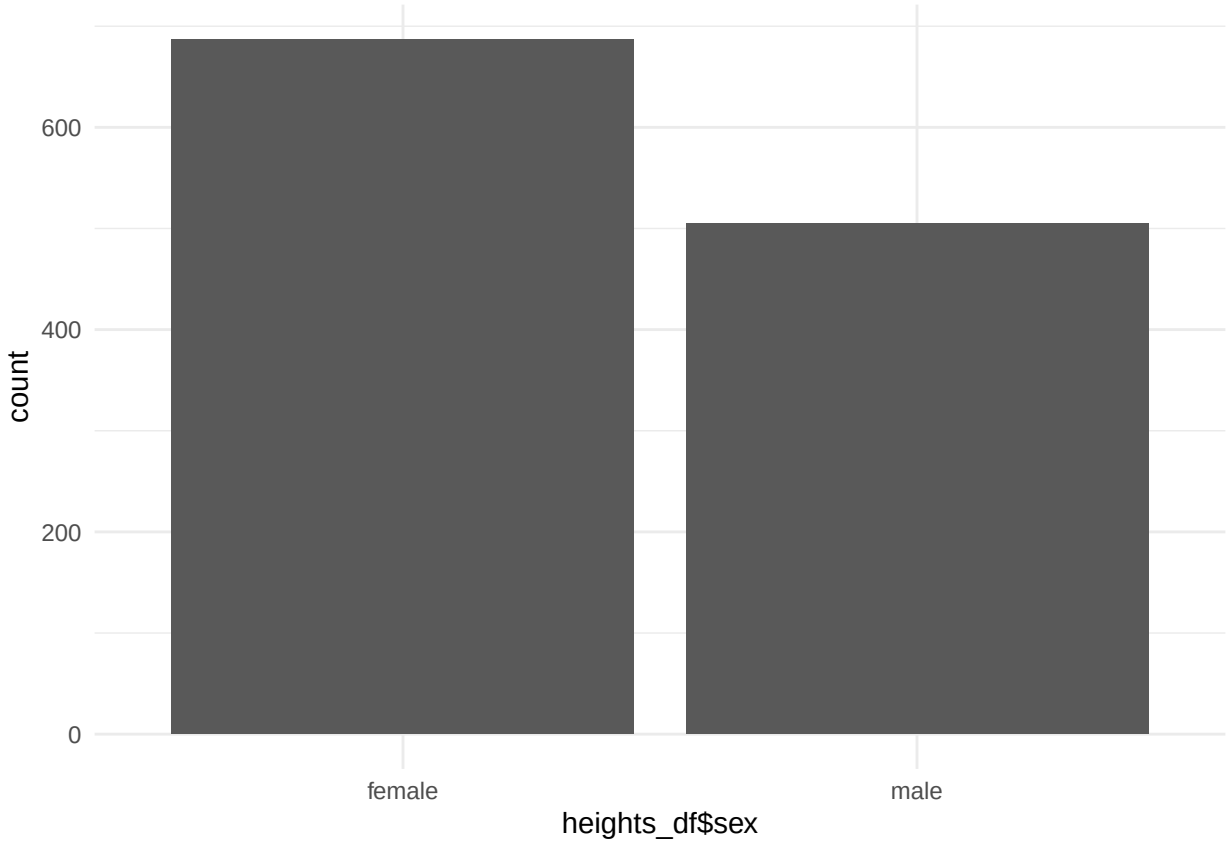

```
# ?geom_point()

# race vs. earn
  ggplot(heights_df, aes(x=heights_df$race, y=heights_df$earn)) +
  geom_point() +
  geom_boxplot(outlier.colour = "Blue", outlier.fill = NULL) +
  xlab("Race") +
  ylab("Earning")
```

```
## Warning: Use of ‘heights_df$race‘ is discouraged.
## i Use ‘race‘ instead.
## Use of ‘heights_df$earn‘ is discouraged.
## i Use ‘earn‘ instead.

## Warning: Use of ‘heights_df$race‘ is discouraged.
## i Use ‘race‘ instead.

## Warning: Use of ‘heights_df$earn‘ is discouraged.
## i Use ‘earn‘ instead.
```



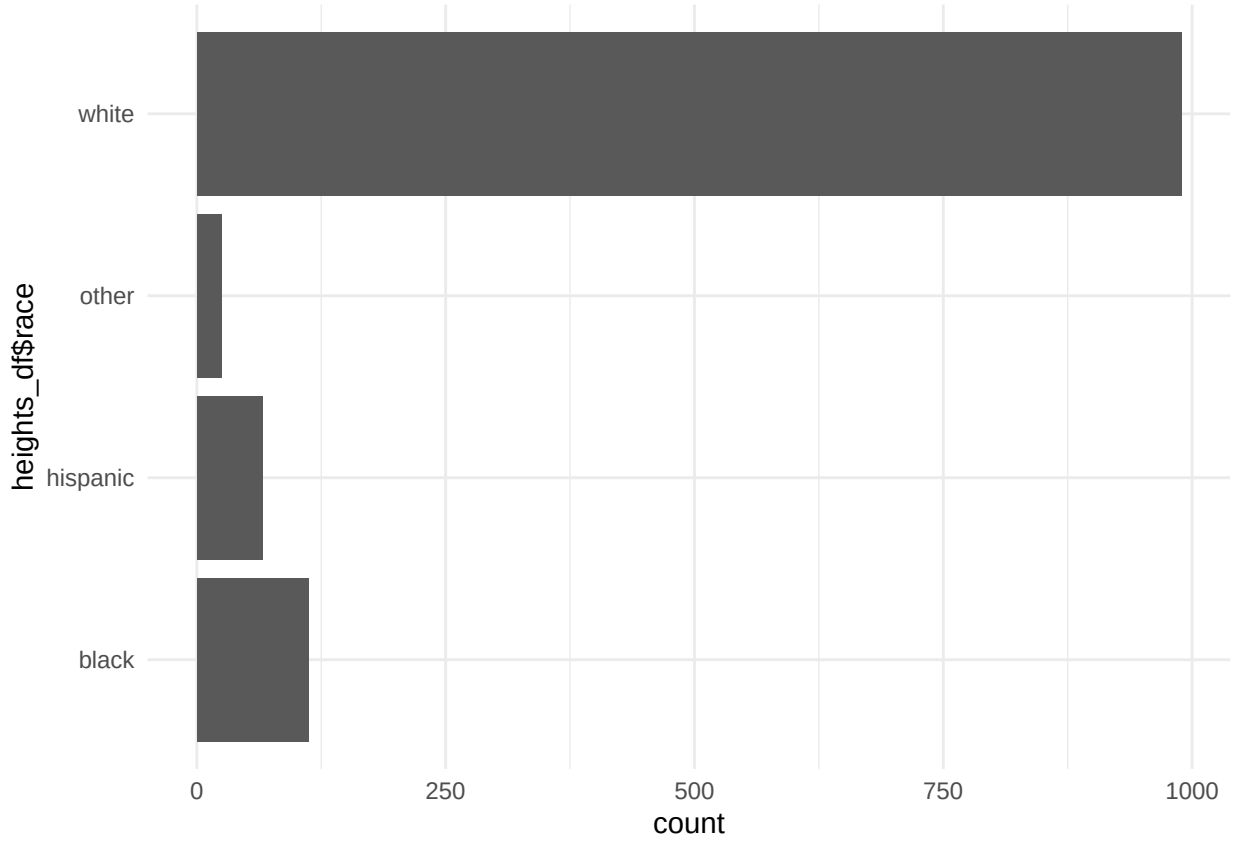

```
# Remove object
# AB <- NULL

# 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(x=heights_df$sex)) + geom_bar()
```

```
## Warning: Use of ‘heights_df$sex‘ is discouraged.
## i Use ‘sex‘ instead.
```



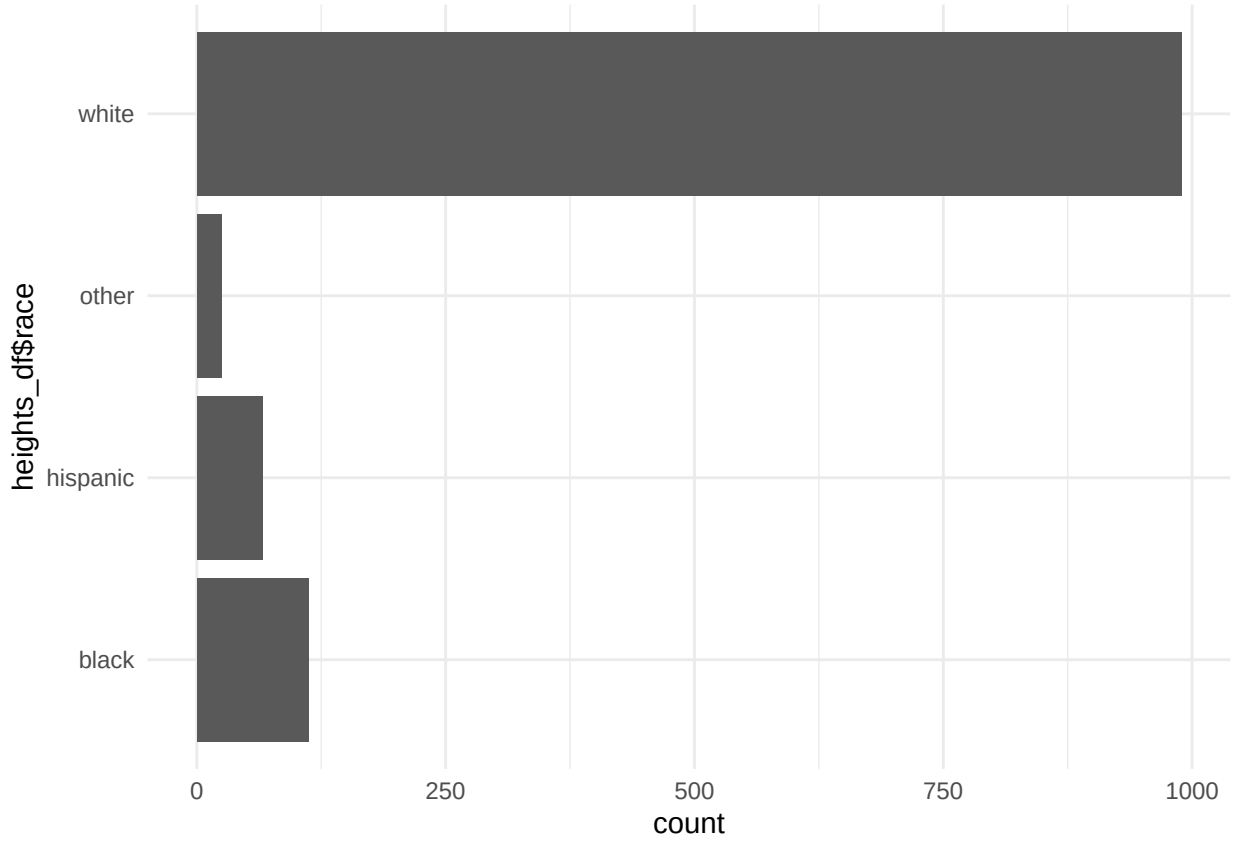

```
# Using `geom_bar()` plot a bar chart of the number of records for each race
ggplot(heights_df, aes(y = heights_df$race )) + geom_bar()
```

```
## Warning: Use of ‘heights_df$race‘ is discouraged.
## i Use ‘race‘ instead.
```




```
# ggplot(heights_df, aes(x = heights_df$race )) + geom_bar()

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

```
## Warning: Use of ‘heights_df$race‘ is discouraged.
## i Use ‘race‘ instead.
```




```
# 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")
head(covid_df)
```

```
##         date      state fips cases deaths
## 1 2020-01-21 Washington   53     1      0
## 2 2020-01-22 Washington   53     1      0
## 3 2020-01-23 Washington   53     1      0
## 4 2020-01-24   Illinois   17     1      0
## 5 2020-01-24 Washington   53     1      0
## 6 2020-01-25 California    6     1      0
```

```
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 ...
```

```
summary(covid_df)
```

```
##      date               state                fips           cases        
##  Length:3039        Length:3039        Min.   : 1.00   Min.   :     1.0  
##  Class :character   Class :character   1st Qu.:17.00   1st Qu.:    25.5  
##  Mode  :character   Mode  :character   Median :31.00   Median :   447.0  
##                                        Mean   :31.31   Mean   :  5425.3  
##                                        3rd Qu.:46.00   3rd Qu.:  2834.0  
##                                        Max.   :78.00   Max.   :288076.0  
##      deaths       
##  Min.   :    0.0  
##  1st Qu.:    0.0  
##  Median :    7.0  
##  Mean   :  228.3  
##  3rd Qu.:   80.0  
##  Max.   :16966.0
```

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

```
##            date          state fips cases deaths
## 3034 2020-04-26 Virgin Islands   78    57      4
## 3035 2020-04-26       Virginia   51 12970    448
## 3036 2020-04-26     Washington   53 13663    757
## 3037 2020-04-26  West Virginia   54  1053     34
## 3038 2020-04-26      Wisconsin   55  5911    274
## 3039 2020-04-26        Wyoming   56   371      7
```

```
summary(covid_df)
```

```
##       date               state                fips           cases        
##  Min.   :2020-01-21   Length:3039        Min.   : 1.00   Min.   :     1.0  
##  1st Qu.:2020-03-16   Class :character   1st Qu.:17.00   1st Qu.:    25.5  
##  Median :2020-03-30   Mode  :character   Median :31.00   Median :   447.0  
##  Mean   :2020-03-28                      Mean   :31.31   Mean   :  5425.3  
##  3rd Qu.:2020-04-13                      3rd Qu.:46.00   3rd Qu.:  2834.0  
##  Max.   :2020-04-26                      Max.   :78.00   Max.   :288076.0  
##      deaths       
##  Min.   :    0.0  
##  1st Qu.:    0.0  
##  Median :    7.0  
##  Mean   :  228.3  
##  3rd Qu.:   80.0  
##  Max.   :16966.0
```

```
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 ...
```

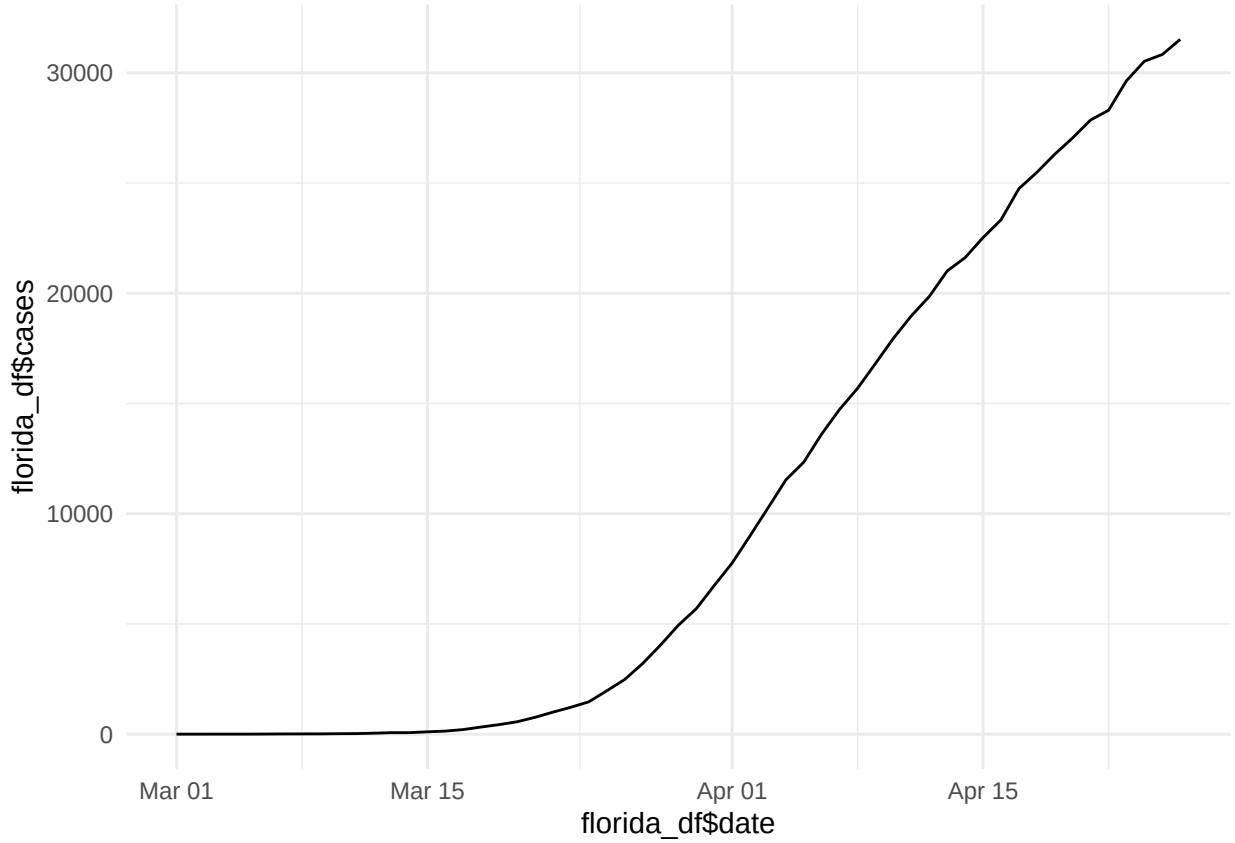
```
## Create three dataframes named `california_df`, `ny_df`, and `florida_df`
## containing the data from California, New York, and Florida
california_df <- covid_df[ which( covid_df$state == "California"), ]
ny_df <- covid_df[ which( covid_df$state == "New York"), ]
florida_df <- covid_df[ which(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

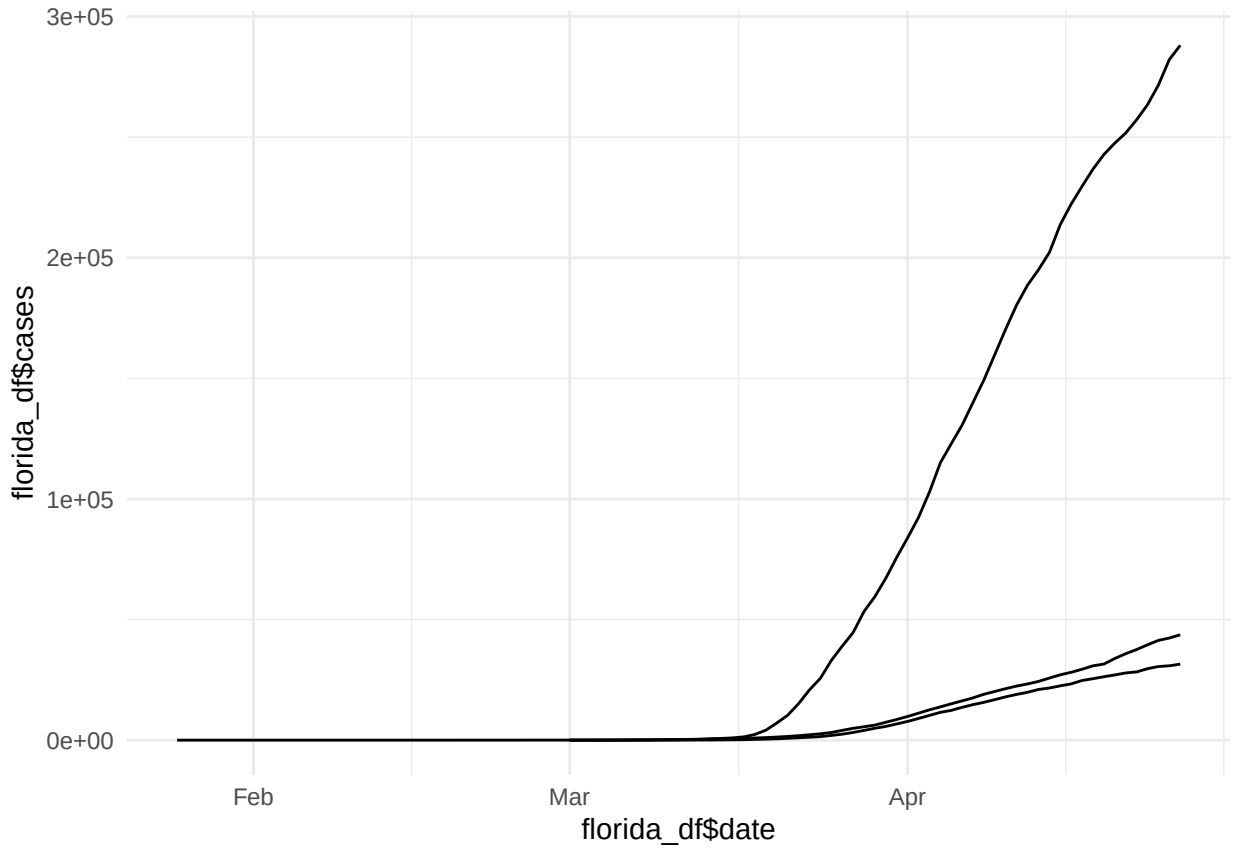
```
tail(california_df)
```

```
##            date      state fips cases deaths
## 2714 2020-04-21 California    6 35844   1316
## 2769 2020-04-22 California    6 37573   1425
## 2824 2020-04-23 California    6 39534   1553
## 2879 2020-04-24 California    6 41368   1619
## 2934 2020-04-25 California    6 42347   1677
## 2989 2020-04-26 California    6 43691   1716
```

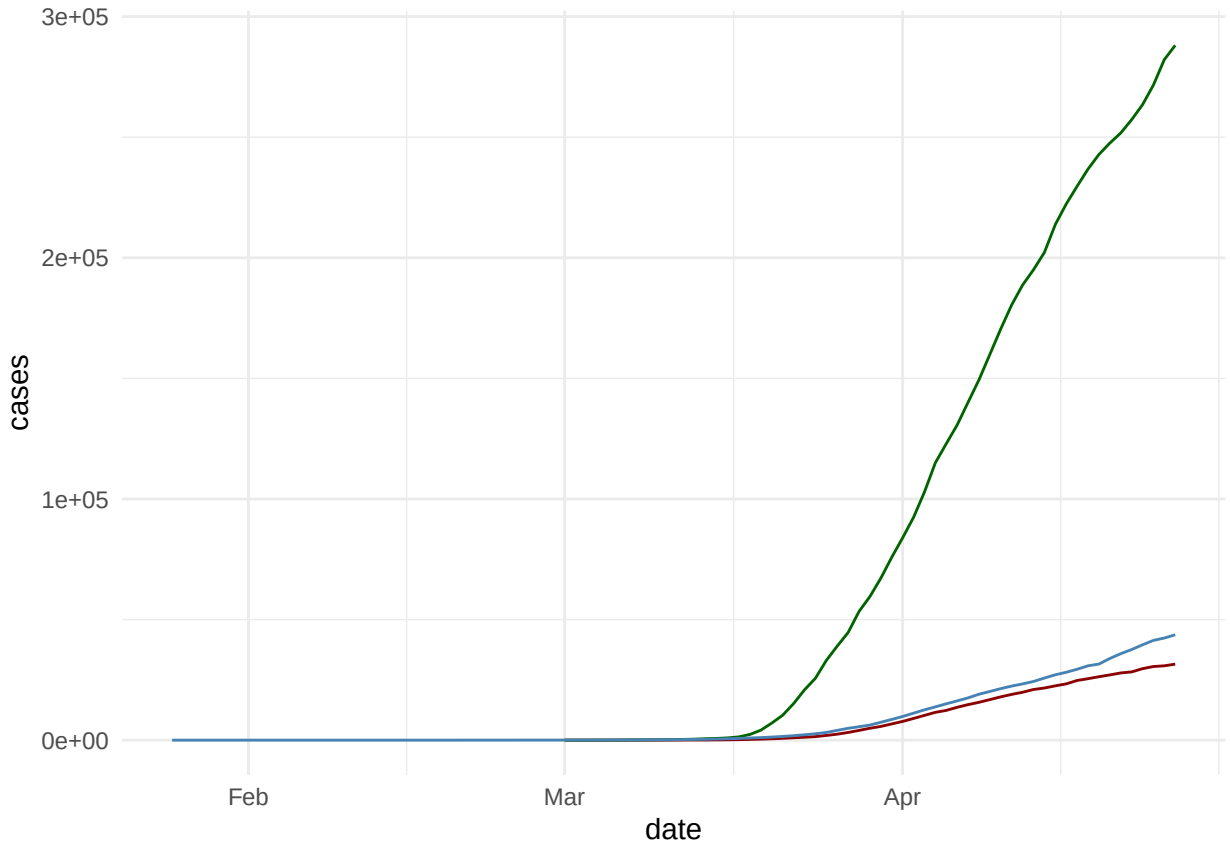
```
## Plot the number of cases in Florida using `geom_line()`
ggplot(data=florida_df, aes(x=florida_df$date, y=florida_df$cases, group=1)) + geom_line()
```



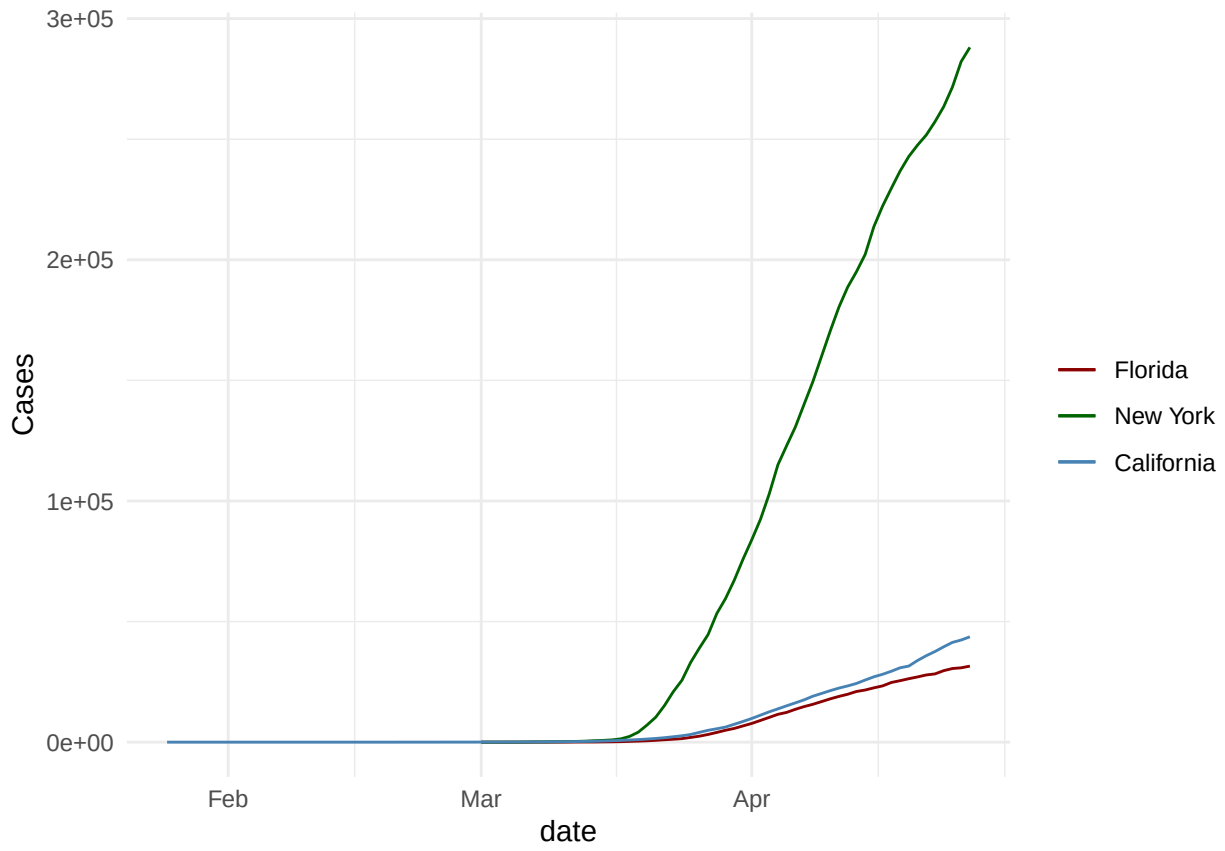

```
## Add lines for New York and California to the plot
ggplot(data=florida_df, aes(x=florida_df$date, group=1)) +
  geom_line(aes(y = florida_df$cases)) +
  geom_line(data=ny_df, aes(y = ny_df$cases)) +
  geom_line(data=california_df, aes(x=california_df$date, y=california_df$cases))
```



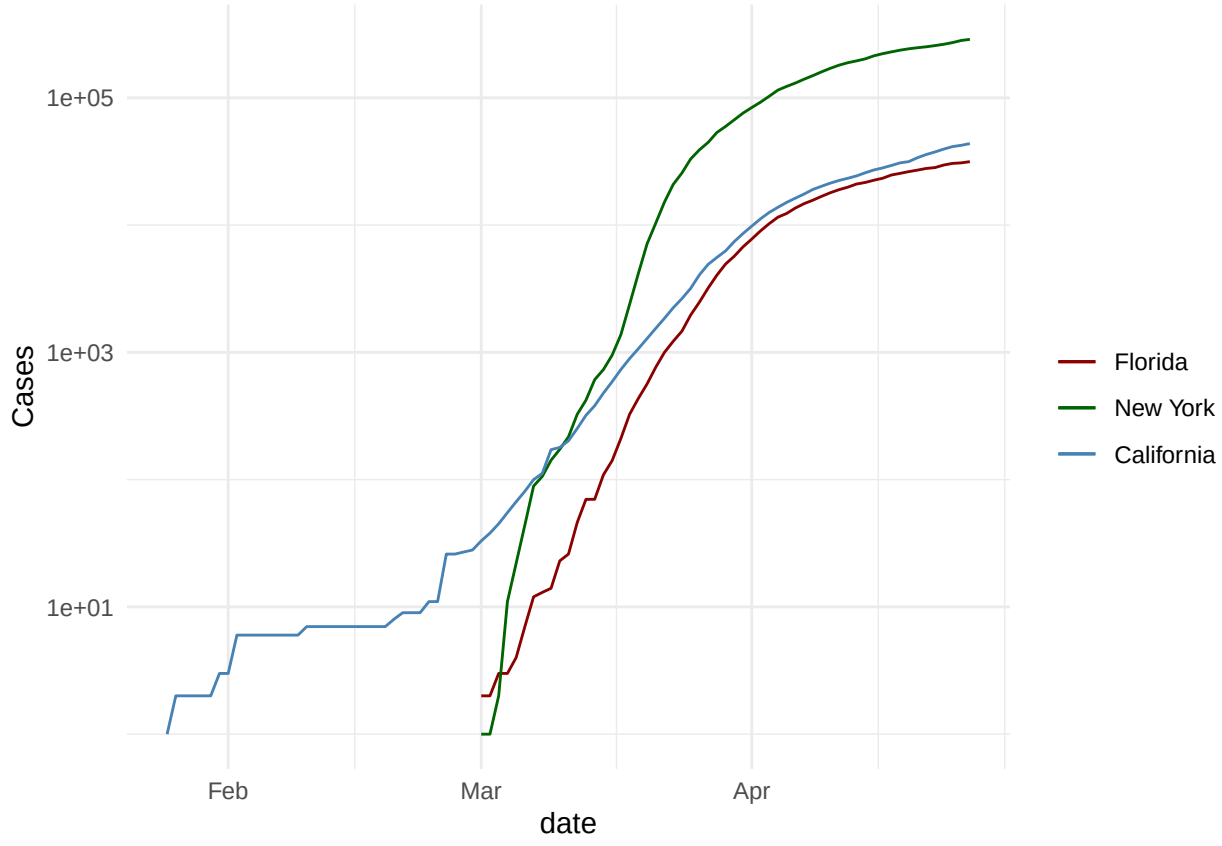

```
##ggplot(data=florida_df, aes(x=florida_df$date, group=1)) +
##  geom_line(data=california_df, aes(x=california_df$date, y=california_df$cases))

##?aes
##?geom_line

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

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_colour_manual
##?scale_y_log10

## 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