Assignment\_06\_Karthikeyan Chellamuthu

KarthikeyanChellamuthu

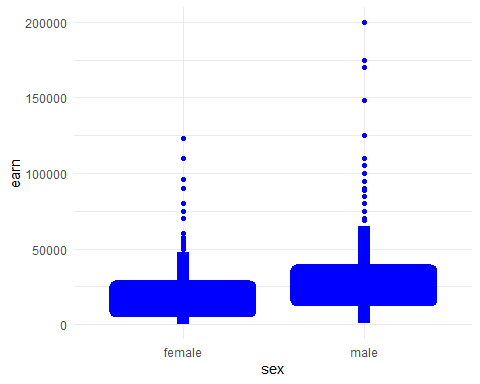
23/01/2022

## Load the ggplot2 package  
library(ggplot2)  
theme\_set(theme\_minimal())

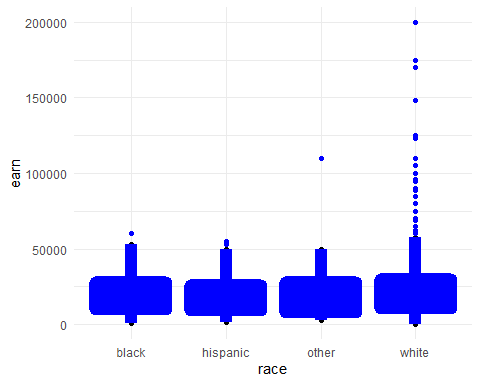
## Set the working directory to the root of your DSC 520 directory  
  
setwd("C:/Users/LENOVO/Desktop/BU/DSC 520 T302-2221 winter 2021-22/GIT-Hub/dsc520-master/data")  
heights\_df <- read.csv("r4ds/heights.csv")  
str(heights\_df)

## '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" ...

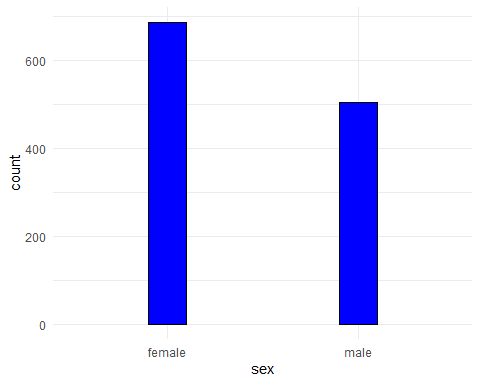
# https://ggplot2.tidyverse.org/reference/geom\_boxplot.html  
## Create boxplots of sex vs. earn and race vs. earn using `geom\_point()` and `geom\_boxplot()`  
## sex vs. earn  
  
ggplot(heights\_df, aes(sex, earn))+geom\_boxplot(colour = "blue", size = 4.5)



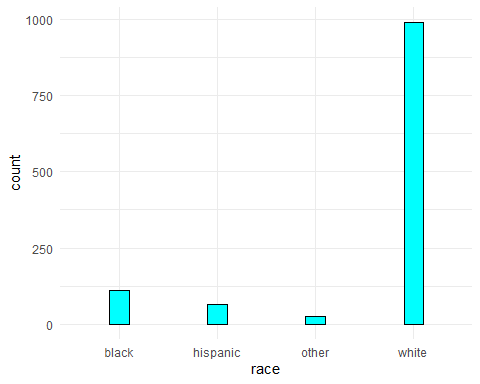
ggplot(heights\_df, aes(race, earn))+geom\_point()+geom\_boxplot(colour = "blue", size = 4.5)



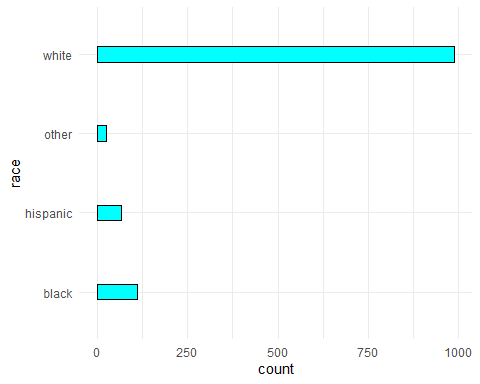
# https://ggplot2.tidyverse.org/reference/geom\_bar.html  
## Using `geom\_bar()` plot a bar chart of the number of records for each `sex`  
ggplot(heights\_df, aes(sex)) + geom\_bar(colour = "black",fill="blue", width = 0.2)



## Using `geom\_bar()` plot a bar chart of the number of records for each race  
ggplot(heights\_df, aes(race)) + geom\_bar(colour = "black",fill="cyan", width = 0.2)



## Create a horizontal bar chart by adding `coord\_flip()` to the previous plot  
ggplot(heights\_df, aes(race)) + geom\_bar(colour = "black",fill="cyan", width = 0.2)+ coord\_flip()



# https://www.rdocumentation.org/packages/ggplot2/versions/3.3.0/topics/geom\_path  
## Load the file `"nytimes/covid-19-data/us-states.csv"` and  
## assign it to the `covid\_df` dataframe  
setwd("C:/Users/LENOVO/Desktop/BU/DSC 520 T302-2221 winter 2021-22/GIT-Hub/dsc520-master/data")  
covid\_df <- read.csv("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\_dt <- as.Date(covid\_df$date)  
  
head(covid\_df\_dt)

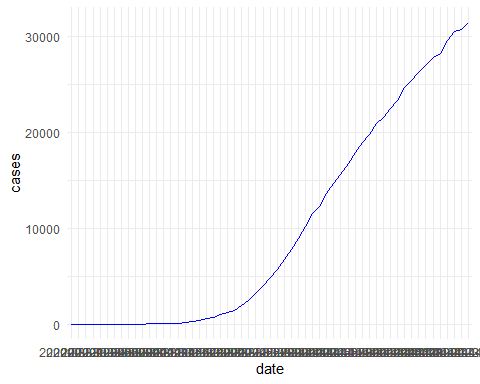
## [1] "2020-01-21" "2020-01-22" "2020-01-23" "2020-01-24" "2020-01-24"  
## [6] "2020-01-25"

## 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"), ]

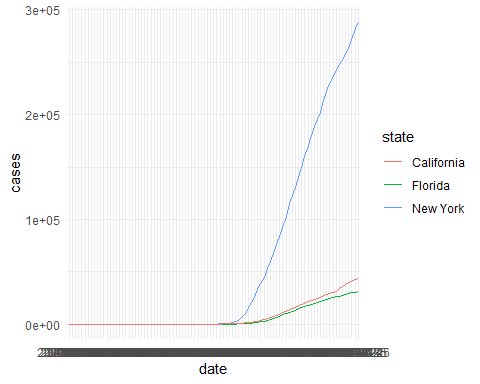
## Unique values to obtain the values New York and Florida  
unique(covid\_df[c("state")])

## state  
## 1 Washington  
## 4 Illinois  
## 6 California  
## 9 Arizona  
## 36 Massachusetts  
## 58 Wisconsin  
## 99 Texas  
## 134 Nebraska  
## 200 Utah  
## 226 Oregon  
## 243 Florida  
## 247 New York  
## 249 Rhode Island  
## 257 Georgia  
## 261 New Hampshire  
## 278 North Carolina  
## 293 New Jersey  
## 304 Colorado  
## 308 Maryland  
## 311 Nevada  
## 318 Tennessee  
## 328 Hawaii  
## 330 Indiana  
## 331 Kentucky  
## 334 Minnesota  
## 341 Oklahoma  
## 343 Pennsylvania  
## 345 South Carolina  
## 354 District of Columbia  
## 360 Kansas  
## 365 Missouri  
## 380 Vermont  
## 381 Virginia  
## 387 Connecticut  
## 394 Iowa  
## 432 Louisiana  
## 443 Ohio  
## 472 Michigan  
## 487 South Dakota  
## 496 Arkansas  
## 500 Delaware  
## 515 Mississippi  
## 521 New Mexico  
## 524 North Dakota  
## 539 Wyoming  
## 540 Alaska  
## 557 Maine  
## 587 Alabama  
## 599 Idaho  
## 613 Montana  
## 626 Puerto Rico  
## 685 Virgin Islands  
## 701 Guam  
## 847 West Virginia  
## 1426 Northern Mariana Islands

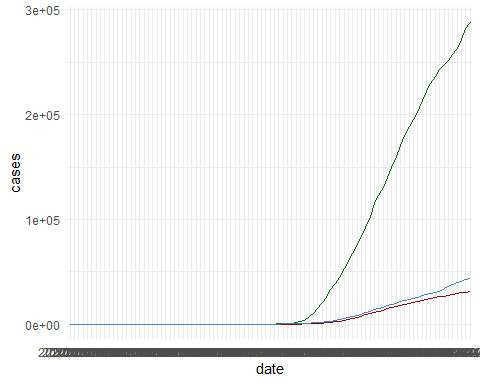
## Plot the number of cases in Florida using `geom\_line()`  
ggplot(data=florida\_df, aes(x=date, y=cases, group=2)) + geom\_line(color='blue')



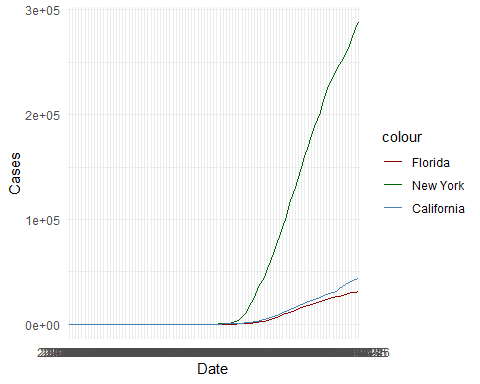
## Add lines for New York and California to the plot  
  
ggplot(data=florida\_df, aes(x=date, group=1)) +  
 geom\_line(aes(y = cases, col=state)) +  
 geom\_line(data=ny\_df, aes(y = cases, col=state)) +  
 geom\_line(data=california\_df, aes(y =cases, col=state))



## 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, col="Florida")) +  
 geom\_line(data=ny\_df, aes(y = cases, col="New York")) +  
 geom\_line(data=california\_df, aes(y =cases,col="California")) +  
 scale\_colour\_manual(values= c('Florida'='darkred',  
 'New York'='darkgreen',   
 'California'='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, col="Florida")) +  
 geom\_line(data=ny\_df, aes(y = cases, col="New York")) +  
 geom\_line(data=california\_df, aes(y =cases,col="California")) +  
 scale\_colour\_manual(values= c('Florida'='darkred',  
 'New York'='darkgreen',   
 'California'='steelblue' ))+  
 xlab("Date") + ylab("Cases")+scale\_y\_log10()

