Lab 3

Kent Codding

2023-09-14

# Set up

setwd("C:/Users/Kent Codding/Desktop/Adv Biostat")  
covid\_df <- read.csv("Lab 2/all\_data.csv")  
vote\_df <- read.csv("Lab 3/vote.csv")

# clean data if necessary

str(vote\_df)

## 'data.frame': 50 obs. of 13 variables:  
## $ state : chr "Alabama" "Alaska" "Arizona" "Arkansas" ...  
## $ called : chr "R" "R" "D" "R" ...  
## $ dem\_votes : int 849624 153778 1672143 423932 11110250 1804352 1080831 296268 5297045 2473633 ...  
## $ rep\_votes : int 1441170 189951 1661686 760647 6006429 1364607 714717 200603 5668731 2461854 ...  
## $ other\_votes : int 32488 15801 53497 34490 384192 87993 28309 7475 101680 62229 ...  
## $ dem\_percent : num 36.6 42.8 49.4 34.8 63.5 55.4 59.3 58.7 47.9 49.5 ...  
## $ rep\_percent : num 62 52.8 49.1 62.4 34.3 41.9 39.2 39.8 51.2 49.3 ...  
## $ other\_percent : num 1.4 4.4 1.6 2.8 2.2 2.7 1.6 1.5 0.9 1.2 ...  
## $ dem\_this\_margin : num -25.5 -10.1 0.3 -27.6 29.2 13.5 20.1 19 -3.4 0.2 ...  
## $ margin\_shift : num 2.3 4.7 3.9 -0.7 -0.9 8.6 6.4 7.6 -2.2 5.4 ...  
## $ vote\_change : num 9.4 12.8 31.6 7.8 23.4 17.1 10.9 14.2 17.5 22.1 ...  
## $ stateid : chr "AL" "AK" "AZ" "AR" ...  
## $ Total.2016.Votes: num 2123372 318608 2573165 1130635 14181595 ...

str(covid\_df)

## 'data.frame': 50 obs. of 31 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Row.names : chr "Alabama" "Alaska" "Arizona" "Arkansas" ...  
## $ perc\_one\_dose : num 65.1 73.2 78.4 70.1 85.1 84 95 88.5 82.9 68.6 ...  
## $ perc\_fully\_vax : num 53.3 65.3 66.2 57 74.9 73.8 83.2 73.5 69.7 57.5 ...  
## $ total\_deaths : int 22342 1514 29685 12531 108732 14953 12284 3453 79794 36119 ...  
## $ death\_rate\_3\_months : num 2 NA 1.8 1.9 1.7 1.6 1 1.6 2.8 1.8 ...  
## $ total\_death\_rate : num 359 227 324 338 245 ...  
## $ new\_case\_rate\_this\_week: num 7.63 2.87 3.48 6.86 6.56 3.32 5.16 3.7 11.2 7.27 ...  
## $ abbr : chr "AL" "AK" "AZ" "AR" ...  
## $ fips : int 1 2 4 5 6 8 9 10 12 13 ...  
## $ pop2010 : int 4779736 710231 6392017 2915918 37253956 5029196 3574097 897934 18801310 9687653 ...  
## $ pop2000 : int 4447100 626932 5130632 2673400 33871648 4301261 3405565 783600 15982378 8186453 ...  
## $ homeownership : num 71.1 64.7 67.4 67.7 57.4 67.6 69.2 73.6 69.7 67.2 ...  
## $ multiunit : num 15.5 24.6 20.7 15.2 30.7 25.6 34.6 17.7 30 20.5 ...  
## $ income : int 22984 30726 25680 21274 29188 30151 36775 29007 26551 25134 ...  
## $ med\_income : int 42081 66521 50448 39267 60883 56456 67740 57599 47661 49347 ...  
## $ poverty : num 17.1 9.5 15.3 18 13.7 12.2 9.2 11 13.8 15.7 ...  
## $ fed\_spend : num 11.73 16.79 9.85 9.61 8.89 ...  
## $ land\_area : num 50645 570641 113594 52035 155779 ...  
## $ smoke : num 24.8 25 20.4 23.5 15.2 19.9 16.5 20.7 21.6 22.2 ...  
## $ murder : num 8.2 4.8 7.5 6.7 6.9 3.7 2.9 4.4 5 6.2 ...  
## $ robbery : num 141.4 80.9 144.4 91.1 176.1 ...  
## $ agg\_assault : num 5.2 65.5 5.1 13.3 0.9 5.3 3.9 47.7 2.6 2.7 ...  
## $ larceny : num 55.4 366 46.4 93 5.1 ...  
## $ motor\_theft : num 6 55.1 14.5 9 1.9 11.1 8.3 31 2.3 5.1 ...  
## $ soc\_sec : num 18.5 8.9 13.9 18.7 11.8 11.4 16.3 16.6 18 12.3 ...  
## $ nuclear : num 16.3 0 15.6 13 7 0 25.6 0 7.3 11.1 ...  
## $ coal : num 56.6 9.5 39.6 48.2 1.1 71.7 11.9 59.4 28.4 63.8 ...  
## $ tr\_deaths : num 23.7 10.1 18.4 22.2 11.6 12 7.7 14.9 18.8 17.8 ...  
## $ tr\_deaths\_no\_alc : num 1.3 7.3 0.9 2.2 0.2 1.2 1.6 5.7 0.3 0.7 ...  
## $ unempl : num 7.6 7.1 8.7 7.6 10.9 7.8 7.8 7 9.4 9.1 ...

library(magrittr) # for pipe operator  
any(vote\_df %>% is.na() == T) # checks if any NA values exist => none exist within DF

## [1] FALSE

any(covid\_df %>% is.na() == T) # some exist within COVID DF

## [1] TRUE

indices = which(covid\_df %>% is.na() == T)   
covid\_df[indices,] #can see all indices are indeed NA

## X Row.names perc\_one\_dose perc\_fully\_vax total\_deaths  
## NA NA <NA> NA NA NA  
## NA.1 NA <NA> NA NA NA  
## NA.2 NA <NA> NA NA NA  
## NA.3 NA <NA> NA NA NA  
## NA.4 NA <NA> NA NA NA  
## NA.5 NA <NA> NA NA NA  
## NA.6 NA <NA> NA NA NA  
## NA.7 NA <NA> NA NA NA  
## NA.8 NA <NA> NA NA NA  
## NA.9 NA <NA> NA NA NA  
## NA.10 NA <NA> NA NA NA  
## NA.11 NA <NA> NA NA NA  
## NA.12 NA <NA> NA NA NA  
## NA.13 NA <NA> NA NA NA  
## NA.14 NA <NA> NA NA NA  
## NA.15 NA <NA> NA NA NA  
## NA.16 NA <NA> NA NA NA  
## NA.17 NA <NA> NA NA NA  
## NA.18 NA <NA> NA NA NA  
## NA.19 NA <NA> NA NA NA  
## NA.20 NA <NA> NA NA NA  
## death\_rate\_3\_months total\_death\_rate new\_case\_rate\_this\_week abbr fips  
## NA NA NA NA <NA> NA  
## NA.1 NA NA NA <NA> NA  
## NA.2 NA NA NA <NA> NA  
## NA.3 NA NA NA <NA> NA  
## NA.4 NA NA NA <NA> NA  
## NA.5 NA NA NA <NA> NA  
## NA.6 NA NA NA <NA> NA  
## NA.7 NA NA NA <NA> NA  
## NA.8 NA NA NA <NA> NA  
## NA.9 NA NA NA <NA> NA  
## NA.10 NA NA NA <NA> NA  
## NA.11 NA NA NA <NA> NA  
## NA.12 NA NA NA <NA> NA  
## NA.13 NA NA NA <NA> NA  
## NA.14 NA NA NA <NA> NA  
## NA.15 NA NA NA <NA> NA  
## NA.16 NA NA NA <NA> NA  
## NA.17 NA NA NA <NA> NA  
## NA.18 NA NA NA <NA> NA  
## NA.19 NA NA NA <NA> NA  
## NA.20 NA NA NA <NA> NA  
## pop2010 pop2000 homeownership multiunit income med\_income poverty  
## NA NA NA NA NA NA NA NA  
## NA.1 NA NA NA NA NA NA NA  
## NA.2 NA NA NA NA NA NA NA  
## NA.3 NA NA NA NA NA NA NA  
## NA.4 NA NA NA NA NA NA NA  
## NA.5 NA NA NA NA NA NA NA  
## NA.6 NA NA NA NA NA NA NA  
## NA.7 NA NA NA NA NA NA NA  
## NA.8 NA NA NA NA NA NA NA  
## NA.9 NA NA NA NA NA NA NA  
## NA.10 NA NA NA NA NA NA NA  
## NA.11 NA NA NA NA NA NA NA  
## NA.12 NA NA NA NA NA NA NA  
## NA.13 NA NA NA NA NA NA NA  
## NA.14 NA NA NA NA NA NA NA  
## NA.15 NA NA NA NA NA NA NA  
## NA.16 NA NA NA NA NA NA NA  
## NA.17 NA NA NA NA NA NA NA  
## NA.18 NA NA NA NA NA NA NA  
## NA.19 NA NA NA NA NA NA NA  
## NA.20 NA NA NA NA NA NA NA  
## fed\_spend land\_area smoke murder robbery agg\_assault larceny motor\_theft  
## NA NA NA NA NA NA NA NA NA  
## NA.1 NA NA NA NA NA NA NA NA  
## NA.2 NA NA NA NA NA NA NA NA  
## NA.3 NA NA NA NA NA NA NA NA  
## NA.4 NA NA NA NA NA NA NA NA  
## NA.5 NA NA NA NA NA NA NA NA  
## NA.6 NA NA NA NA NA NA NA NA  
## NA.7 NA NA NA NA NA NA NA NA  
## NA.8 NA NA NA NA NA NA NA NA  
## NA.9 NA NA NA NA NA NA NA NA  
## NA.10 NA NA NA NA NA NA NA NA  
## NA.11 NA NA NA NA NA NA NA NA  
## NA.12 NA NA NA NA NA NA NA NA  
## NA.13 NA NA NA NA NA NA NA NA  
## NA.14 NA NA NA NA NA NA NA NA  
## NA.15 NA NA NA NA NA NA NA NA  
## NA.16 NA NA NA NA NA NA NA NA  
## NA.17 NA NA NA NA NA NA NA NA  
## NA.18 NA NA NA NA NA NA NA NA  
## NA.19 NA NA NA NA NA NA NA NA  
## NA.20 NA NA NA NA NA NA NA NA  
## soc\_sec nuclear coal tr\_deaths tr\_deaths\_no\_alc unempl  
## NA NA NA NA NA NA NA  
## NA.1 NA NA NA NA NA NA  
## NA.2 NA NA NA NA NA NA  
## NA.3 NA NA NA NA NA NA  
## NA.4 NA NA NA NA NA NA  
## NA.5 NA NA NA NA NA NA  
## NA.6 NA NA NA NA NA NA  
## NA.7 NA NA NA NA NA NA  
## NA.8 NA NA NA NA NA NA  
## NA.9 NA NA NA NA NA NA  
## NA.10 NA NA NA NA NA NA  
## NA.11 NA NA NA NA NA NA  
## NA.12 NA NA NA NA NA NA  
## NA.13 NA NA NA NA NA NA  
## NA.14 NA NA NA NA NA NA  
## NA.15 NA NA NA NA NA NA  
## NA.16 NA NA NA NA NA NA  
## NA.17 NA NA NA NA NA NA  
## NA.18 NA NA NA NA NA NA  
## NA.19 NA NA NA NA NA NA  
## NA.20 NA NA NA NA NA NA

# Load packages

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.3

library(usmap)

## check if rownames match up in both dfs

rownames(vote\_df) == rownames(covid\_df)

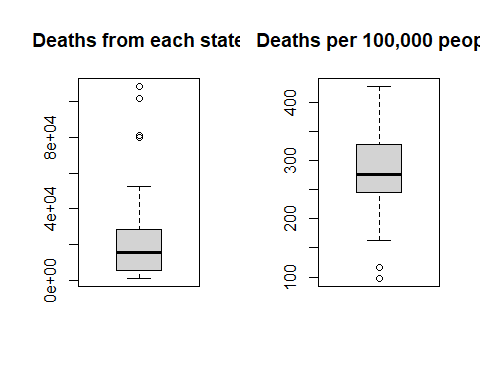
## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [46] TRUE TRUE TRUE TRUE TRUE

# Examine response variables

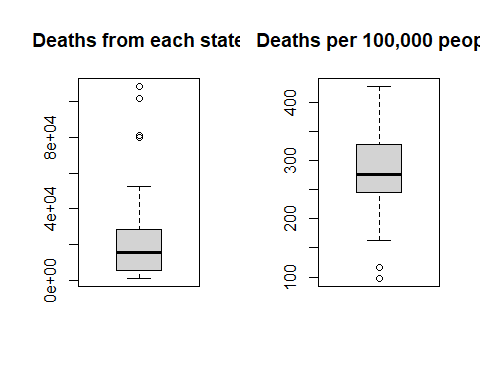
**Question 1:** Analyzing case rates reduces bias between states with high population and low population. For example, the most populated state, California, will almost undoubtedly have more cases than Wyoming, the least populated state. A better metric to compare would be the case rate: Covid deaths per 100,000 people. This metric tells us whether a person is more likely to die in California or Wyoming.

## Graph it

par(mfrow = c(1,2)) #mfrow set up like matrix c(numrows, numcols) to create side by side charts  
boxplot(covid\_df$total\_deaths,  
 main = "Deaths from each state"  
 )  
boxplot(covid\_df$total\_death\_rate,  
 main = "Deaths per 100,000 people"  
 )

 ## Graph it to pdf

par(mfrow = c(1,2)) #mfrow set up like matrix c(numrows, numcols) to create side by side charts  
boxplot(covid\_df$total\_deaths,  
 main = "Deaths from each state"  
 )  
boxplot(covid\_df$total\_death\_rate,  
 main = "Deaths per 100,000 people"  
 )



#pdf(file = "Lab 3/Covid death rate by state.pdf", width = 8, height = 6)  
#Note: I commented this out after running this code chunk once to save to pdf.

**Question 2:**

#Find the outliers!!  
summary(covid\_df$total\_deaths)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1016 5580 15266 22752 27954 108732

#get IQR  
IQR(covid\_df$total\_deaths)

## [1] 22374.5

# get threshold values for outliers  
Tmin = 5580-(1.5\*22374.5) #first quartile minus 1.5\*IQR  
Tmax = 27954+(1.5\*22374.5) ##third quartile plus 1.5\*IQR  
  
subset(covid\_df, total\_deaths > Tmax)[,2]

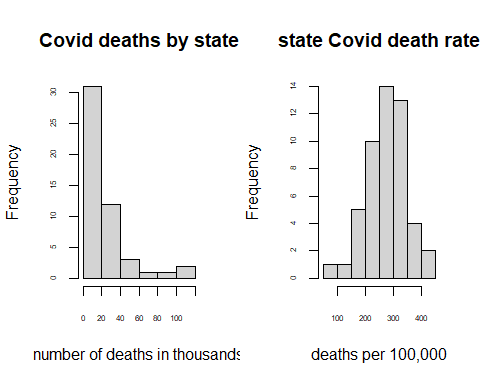
## [1] "California" "Florida" "New York" "Texas"

#the output shows states with total deaths greater than the Tmax threshold calculated above. These states are also the four most populated in the country, which explains why the total is higher than the less populated states.

# Histograms and data distributions

## Graph it

par(mfrow = c(1,2), cex.axis = 0.5) #mfrow set up like matrix c(numrows, numcols) to create side by side charts  
  
hist(covid\_df$total\_deaths,  
 xlab = "number of deaths in thousands",  
 main = "Covid deaths by state",  
 breaks = 6,  
 xlim = c(0,120000),  
 axes = F) #specify breaks, xlim, turn axes off to modify  
axis(side = 1, at = c(0,20000, 40000, 60000, 80000, 100000, 120000), labels = c(0,20,40,60,80,100,120)) #creates shorter labels that represent same numeric values as before  
axis(side = 2, at = c(0,5, 10, 15, 20, 25, 30))  
hist(covid\_df$total\_death\_rate,  
 xlab = "deaths per 100,000",  
 main = "state Covid death rate")

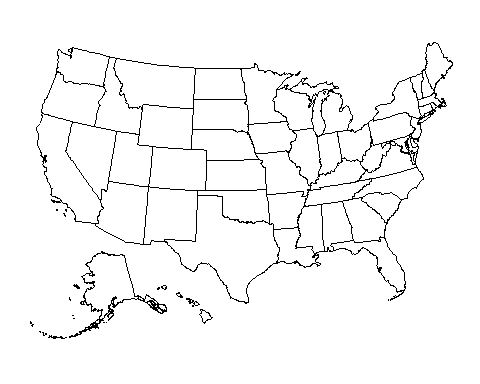


**Question 3:** The total number of deaths does not show a normal distribution bell curve. This is because total number of deaths for each state correlates with the total population of each state, which does not show a normal distribution as some states are much more populated than others like California and Texas.

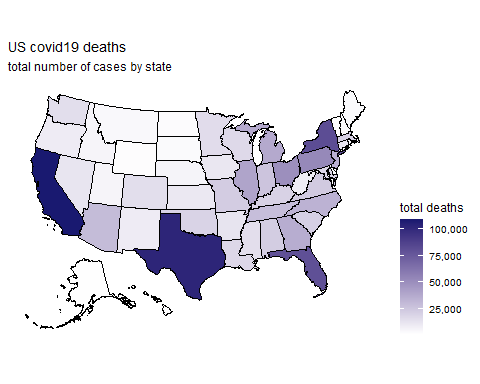
# Map plotting

## Spatial correlation

plot\_usmap()

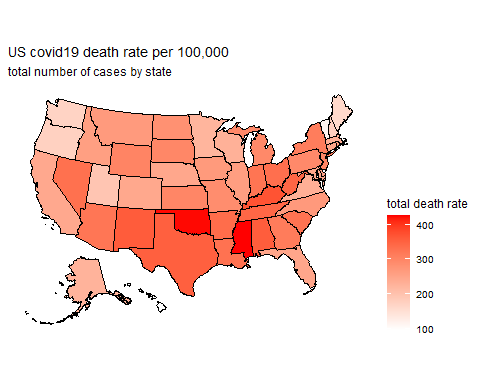


#first we have to add a column to our covid\_df called "state".  
#the plot\_usmap() function requires this column to work  
covid\_df$state <- rownames(covid\_df)  
  
plot\_usmap(data = covid\_df, values = "total\_deaths") +  
 scale\_fill\_continuous(low = "white", high = "midnightblue",   
 label = scales::comma, name = "total deaths") +  
 theme(legend.position = "right") +  
 labs(title = "US covid19 deaths",   
 subtitle = "total number of cases by state")



## Graph it

plot\_usmap(data = covid\_df, values = "total\_death\_rate") +  
 scale\_fill\_continuous(low = "white", high = "red",   
 label = scales::comma, name = "total death rate") +  
 theme(legend.position = "right") +  
 labs(title = "US covid19 death rate per 100,000",   
 subtitle = "total number of cases by state")



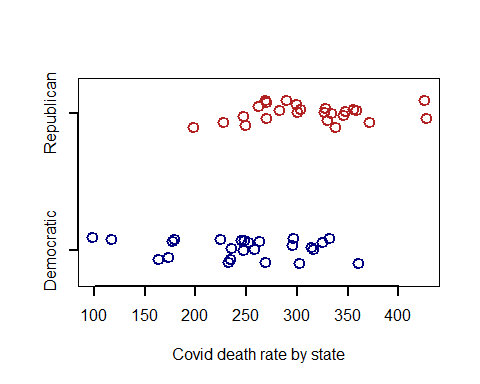
**Question 4:** The case rate map eliminates the confounding variable of total state population. This map shows the states where people are most likely to die from COVID.

# Strip charts for correlations between numeric and categorical variables

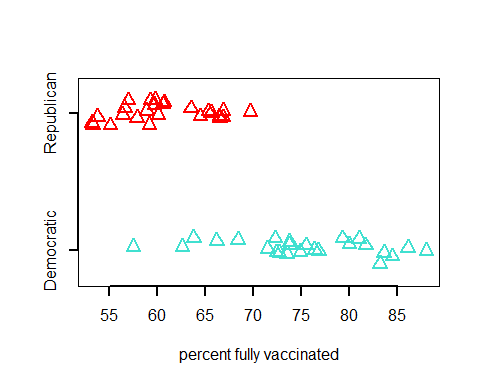
## Political correlation

## Graph it

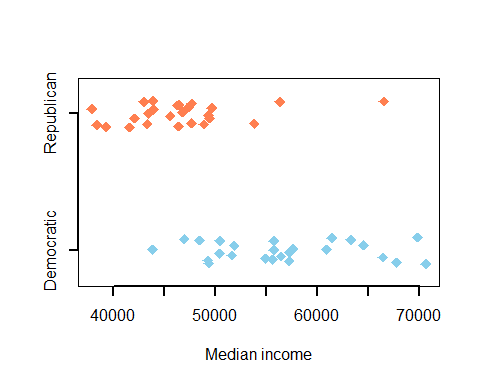
stripchart(covid\_df$total\_death\_rate ~ vote\_df$called,   
 col = c("navyblue", "firebrick"), # choose colors  
 method = "jitter", #adds minor variation   
 #in data to limit overlapping of points  
 pch = 1, # controls the shape of the data points  
 xlab = "Covid death rate by state",  
 group.names = c("Democratic","Republican"),  
 cex = 1.5, lwd = 2 #controls size and line thickness of points  
 )



stripchart(covid\_df$perc\_fully\_vax ~ vote\_df$called,   
 col = c("turquoise", "red"), # choose colors  
 method = "jitter", #adds minor variation   
 #in data to limit overlapping of points  
 pch = 2, # controls the shape of the data points  
 xlab = "percent fully vaccinated",  
 group.names = c("Democratic","Republican"),  
 cex = 1.5, lwd = 2 #controls size and line thickness of points  
 )



stripchart(covid\_df$med\_income ~ vote\_df$called,   
 col = c("skyblue", "coral"), # choose colors  
 method = "jitter", #adds minor variation   
 #in data to limit overlapping of points  
 pch = 18, # controls the shape of the data points  
 xlab = "Median income",  
 group.names = c("Democratic","Republican"),  
 cex = 1.5, lwd = 2 #controls size and line thickness of points  
 )

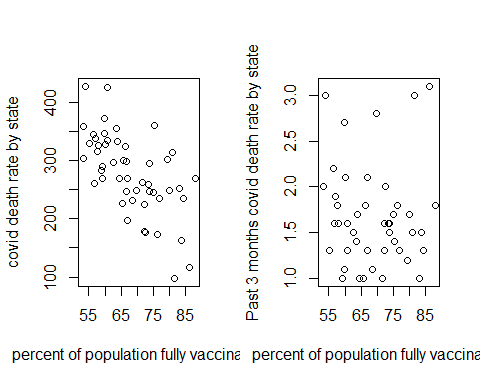


**Question 5:** Yes, states that voted Democrat generally had a lower death rate. Additionally, states that voted Republican generally had a lower vaccination rate. Hence, vaccination rate and other variables like regulations and mask wearing likely affected the overall death rate per state. Evidently, political party affiliation explains businesses and individual decisions behind vaccination and masking up…

# Scatter plots of correlations between numerical variables

## Graph it

par(mfrow = c(1,2)) #mfrow set up like matrix c(numrows, numcols) to create side by side charts  
plot(total\_death\_rate ~ perc\_fully\_vax, data = covid\_df, xlab = "percent of population fully vaccinated", ylab = "covid death rate by state")  
plot(death\_rate\_3\_months ~ perc\_fully\_vax, data = covid\_df, xlab = "percent of population fully vaccinated", ylab = "Past 3 months covid death rate by state")



# Fancy up that plot

## Graph it

#make a custom color pallet called "clr"  
clr <- c("navy", "firebrick4")   
  
#set up the plotting space  
plot(total\_death\_rate ~ perc\_fully\_vax, data = covid\_df,  
 type = "n",   
 xlab = "percent of state population that is fully vaccinated",  
 ylab = "death rate by state"  
 )  
  
# The argument type = "n" tells R to make a blank plotting space  
# now we can use the "text" function to plot our data using the   
# state name abbreviations, which are in the vote\_df dataframe under "stateid"  
# first we'll only plot the states that voted democratic by subsetting  
# and we'll make them blue (i.e. the first color in our pallet "clr")  
  
text(total\_death\_rate ~ perc\_fully\_vax, #gives the coordinates to plot the stat abbr.  
 lab = vote\_df$stateid[vote\_df$called == "D"], #tells which text to plot  
 data = subset(covid\_df, vote\_df$called == "D"), #subsets only the D states  
 col = clr[1]) #color of the text  
  
#now we'll plot the states that voted republican  
#and make them red  
  
text(total\_death\_rate ~ perc\_fully\_vax,   
 lab = vote\_df$stateid[vote\_df$called == "R"],   
 data = subset(covid\_df, vote\_df$called == "R"),  
 col = clr[2])  
  
#Add a legend  
legend("topright", #puts legend in the top right corner  
 legend = c("voted dem.","voted rep."), #list of things in legend  
 text.col = clr, #colors to use for the text  
 bty = 'n', #this tells R to not draw a box around the legend  
 cex = 1.5) # this increases the font size by 50%  
  
#Add a trend line using abline  
#this is a function for adding lines to plots  
#we can use a simple linear model (lm()) to define our line  
#we'll talk much more about lm() next time,  
#so don't worry if it is new to you now  
abline(  
 lm(total\_death\_rate ~ perc\_fully\_vax, data = covid\_df),  
 lwd = 2, col = "purple") #set line width and color

