data\_viz

Md Easin Hasan

4/20/2021

# Visualizing Amounts:

# Import the dataset:

data<-read.csv(file="owid-covid-data425.csv",header=TRUE)  
dim(data)

## [1] 83862 59

We have obtained the data-set on COVID-19 (coronavirus) by Our World in Data. They have up-to-date data on confirmed cases, deaths, hospitalizations, testing, and vaccinations, throughout the duration of the COVID-19 pandemic.

data2 <- data[data$iso\_code == "USA", ]   
data3 <- data[79847:79960,]   
data3 <-data3[,c(1,4,35,36,37,38,39,40,41,42,43)]

library(lubridate)

library(tibble)

## Warning: package 'tibble' was built under R version 4.0.5

dates <- data3$date  
date\_dis <- tibble(date = dates,   
 month = month(dates),  
 day = day(dates),  
 year = year(dates))  
data3$date <- date\_dis$month

Since we will visualize the number of people vaccinated in USA, here we kept the observations only on USA. The covid-19 vaccinations start from Dec 20, 2020 but we considered the observations from Jan 1st, 2021 to April 24, 2021 in USA.

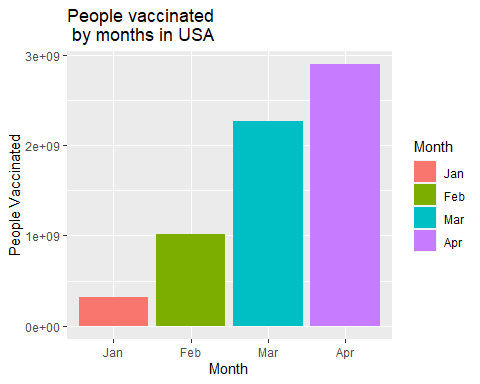
# people vaccinated

library(ggplot2)  
library(reshape2)

## Warning: package 'reshape2' was built under R version 4.0.4

library(scales)  
  
ggplot(data3, aes(x = date, y = people\_vaccinated, fill=as.factor(date))) +  
 xlab("Month") + ylab("People Vaccinated")+  
 geom\_bar(stat="identity")+  
 ggtitle("People vaccinated \n by months in USA")+  
 scale\_fill\_discrete(name="Month",  
 breaks=c("1", "2", "3", "4"),  
 labels= c("Jan","Feb","Mar", "Apr"))+  
 scale\_x\_discrete(limit = c("1", "2", "3", "4"),  
 labels = c("Jan","Feb","Mar", "Apr"))

## Warning: Removed 9 rows containing missing values (position\_stack).

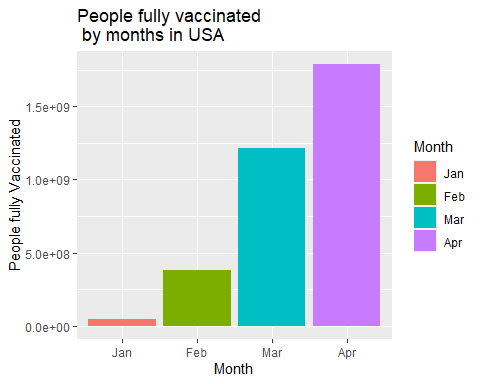


From this graph we can see that number of people vaccinated is increasing gradually. This graph follows the Tufte’s Principle 5: Provide the user with an overview and details on demand. So, we can get an overall overview of what amount of people in each month received the 1 dose of vaccine. This graph follows the Tufte’s Principle 6: Utilize Layering & Separation since using different colors we can figure out in which months the people get vaccinated more. Hence, we can easily compare between months to figure out the vaccination rate in USA.

# people fully vaccinated

ggplot(data3, aes(x = date, y = people\_fully\_vaccinated, fill=as.factor(date))) +  
 xlab("Month") + ylab("People fully Vaccinated")+  
 geom\_bar(stat="identity")+  
 ggtitle("People fully vaccinated \n by months in USA")+  
 scale\_fill\_discrete(name="Month",  
 breaks=c("1", "2", "3", "4"),  
 labels= c("Jan","Feb","Mar", "Apr"))+  
 scale\_x\_discrete(limit = c("1", "2", "3", "4"),  
 labels = c("Jan","Feb","Mar", "Apr"))

## Warning: Removed 17 rows containing missing values (position\_stack).

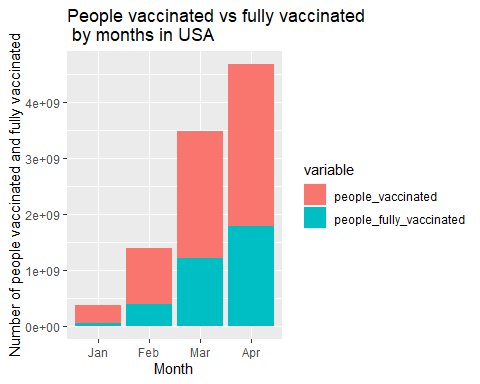


The number of people fully vaccinated also increasing gradually. This graph follows the Tufte’s Principle 5: Provide the user with an overview and details on demand. So, we can get an overall overview of what amount of people in each month getting fully vaccinated. This graph follows the Tufte’s Principle 6: Utilize Layering & Separation since using different colors we can figure out in which months the amount of people getting fully vaccinated. Hence, we can easily compare between months to figure out the fully vaccination rate in USA.

# Visualizing Amounts Improvement

dfm <- melt(data3[,c('date','people\_vaccinated','people\_fully\_vaccinated')],id.vars = 1)  
  
ggplot(dfm,aes(x = date,y = value)) +   
 geom\_bar(aes(fill = variable),stat = "identity")+  
 labs(x = "Month")+  
 labs(y = "Number of people vaccinated and fully vaccinated")+  
 ggtitle("People vaccinated vs fully vaccinated \n by months in USA")+  
 scale\_x\_discrete(limit = c("1", "2", "3", "4"),  
 labels = c("Jan","Feb","Mar", "Apr"))

## Warning: Removed 26 rows containing missing values (position\_stack).



From this visualization we can very easily compare the number of people vaccinated vs number of people fully vaccinated each month. According to Cole Nussbaummer presentation on “Story Telling with Data”, graphics play a vital role in data visualization. The graphics of this diagram is very clear since the color, font, shapes are used accurately. I think this diagram follow the 5 seconds rule since the color differentiation makes it easy to understand the story of the data within few seconds. This graph explains the real scenarios behind the covid-19 vaccination in USA.

# Visualizing Distributions

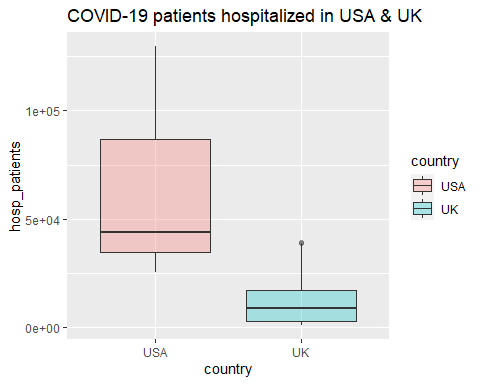
data1 <- data[data$iso\_code == "USA", ]   
data2 <- data[data$iso\_code == "GBR", ]   
data3<-rbind(data1,data2)  
country <- ifelse(data3$iso\_code == "USA",0,1)  
data4 <- cbind(country,data3)  
data4$country<-as.factor(data4$country)  
data4$country<-factor(data4$country,levels=c(0,1),labels=c("USA", "UK"))  
data5 <- data4[,-c(2)]  
dat2 <-data5[,c(1,6,8,11,18,20)]  
colnames(dat2)

## [1] "country" "new\_cases"   
## [3] "total\_deaths" "total\_cases\_per\_million"  
## [5] "icu\_patients" "hosp\_patients"

#Data visualization distribution

ggplot(dat2, aes(x = country, y = hosp\_patients, fill = country)) +  
 geom\_boxplot(alpha=0.3) +  
 theme(legend.position="right")+  
 ggtitle("COVID-19 patients hospitalized in USA & UK")

## Warning: Removed 241 rows containing non-finite values (stat\_boxplot).

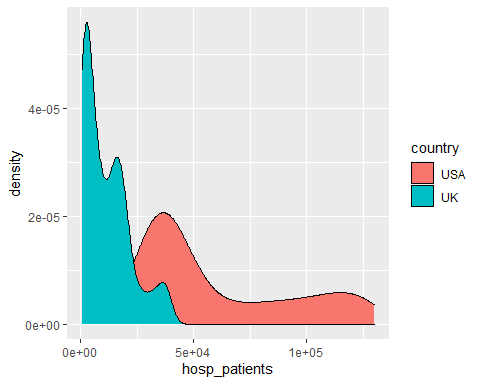


From this visualization we can see the distributions of the hosp\_patients in USA and UK. This graph follows the Tufte’s Principle 1: Maximizing the data-ink ratio, within reason because the non-redundant ink arranged in response to variation in the numbers represented. It’s also follow Principle 4: Escape flatland - small multiples, parallel sequencing because of the parallel plot of box-plots. This graph follows the Tufte’s Principle 6: Utilize Layering & Separation since using different colors we can figure out in which countries the hosp\_patients are too high.

The story behind this graph is comparing the distributions of hosp-patients in USA and UK and I think this graph tell the comparison between hosp\_patients rate for both countries during this COVID-19 pandemic.

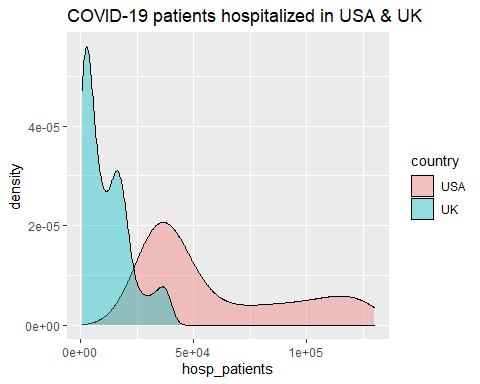
ggplot(dat2, aes(x = hosp\_patients, fill = country)) +  
 geom\_density()

## Warning: Removed 241 rows containing non-finite values (stat\_density).



# Use semi-transparent fill  
p<-ggplot(dat2, aes(x=hosp\_patients, fill = country)) +  
 geom\_density(alpha=0.4)+  
 ggtitle("COVID-19 patients hospitalized in USA & UK")  
p

## Warning: Removed 241 rows containing non-finite values (stat\_density).



From this density plot, we can see the hosp\_patients in USA and UK. This graph follows Tufte’s Principle 1: Maximizing the data-ink ratio, within reason and Principle 6: Utilize Layering & Separation.

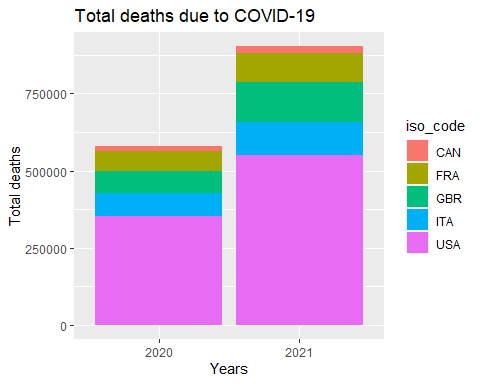
# Data visualization proportion:

data<-read.csv(file="owid-covid-data\_new.csv",header=TRUE)  
dim(data)

## [1] 78245 59

USA\_20 <- data[74514:74514, ]  
UK\_20 <- data[74081:74081, ]  
CAN\_20 <- data[13100:13100, ]  
FRA\_20 <- data[25899:25899, ]  
ITA\_20 <- data[35349:35349, ]  
data\_20<-rbind(USA\_20,UK\_20,CAN\_20,FRA\_20,ITA\_20)  
data\_f20 <- data\_20[,c(1,4,8)]  
  
USA\_21 <- data[74602:74602, ]  
UK\_21 <- data[74169:74169, ]  
CAN\_21 <- data[13188:13188, ]  
FRA\_21 <- data[25987:25987, ]  
ITA\_21 <- data[35437:35437, ]  
data\_21<-rbind(USA\_21,UK\_21,CAN\_21,FRA\_21,ITA\_21)  
data\_f21 <- data\_21[,c(1,4,8)]  
data\_final <- rbind(data\_f20,data\_f21)

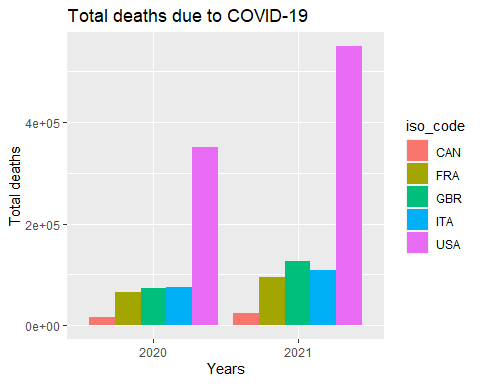
#library(tidyverse)  
#library(magrittr)  
#library(ggplot2)  
#library(reshape2)  
#library(scales)  
  
ggplot(data\_final, aes(x = date, y = total\_deaths, fill = iso\_code)) +  
 xlab("Years") + ylab("Total deaths")+  
 geom\_bar(stat="identity", position = 'dodge')+  
 geom\_col()+  
 scale\_x\_discrete(name = "Years", labels=c("2020-12-31" = "2020", "2021-03-29" = "2021"))+  
 scale\_y\_continuous(name="Total deaths")+  
 ggtitle("Total deaths due to COVID-19")



From this data visualization, we can see the total deaths due to Covid-19 in Canada, France, Great Britain, Italy, and United States respectively. This graph follows the Tufte’s Principle 5: Provide the user with an overview and details on demand. So, we can compare what proportion of people died in each countries during this pandemic. This graph follows the Tufte’s Principle 6: Utilize Layering & Separation since using different colors. ## Data

**visualization proportion improvement:**

plot <- ggplot(data\_final, aes(date, total\_deaths, fill=iso\_code))  
plot <- plot + geom\_bar(stat = "identity", position = 'dodge')+  
 scale\_x\_discrete(name = "Years", labels=c("2020-12-31" = "2020", "2021-03-29" = "2021"))+  
 scale\_y\_continuous(name="Total deaths")+  
 ggtitle("Total deaths due to COVID-19")  
plot



From this data visualization, we can see the total deaths due to COVID-19 in Canada, France, Great Britain, Italy, and United States in 2020 and 2021 respectively. The number of total deaths very higher in USA compare to other countries. This graph follows the Tufte’s Principle 5: Provide the user with an overview and details on demand. So, we can compare what proportion of people died in each countries during this pandemic. This graph follows the Tufte’s Principle 6: Utilize Layering & Separation since using different colors. It’s also follows Tufte’s Principle 1: Maximizing the data-ink ratio, within reason.

The story behind this graph is comparing total deaths due to COVID-19 in the above mentioned countries. It’s clearly visible that the total deaths in USA is very high both in 2020 and 2021 compare to other countries. I think this diagram follow the 5 seconds rule since the color differentiation makes it easy to understand the story of the data within few seconds.

## Visualizing Associations:

data1 <- data[data$iso\_code == "USA", ]   
dat2 <-data1[,c(1,6,8,11,18,20)]  
colnames(dat2)

## [1] "iso\_code" "new\_cases"   
## [3] "total\_deaths" "total\_cases\_per\_million"  
## [5] "icu\_patients" "hosp\_patients"

data2 <- data[data$iso\_code == "GBR", ]   
data3<-rbind(data1,data2)  
country <- ifelse(data3$iso\_code == "USA",0,1)  
data4 <- cbind(country,data3)  
data4$country<-as.factor(data4$country)  
data4$country<-factor(data4$country,levels=c(0,1),labels=c("USA", "UK"))  
data5 <- data4[,-c(2)]  
dat\_2 <-data5[,c(1,6,8,9,11,18,20)]

miss.info<- function(dat, filename=NULL){   
   
 vnames <- colnames(dat); vnames   
   
 n <- nrow(dat)   
   
 out <- NULL   
   
 for (j in 1: ncol(dat)){   
   
 vname <- colnames(dat)[j]   
   
 x <- as.vector(dat[,j])   
   
 n1 <- sum(is.na(x), na.rm=T)   
   
 n2 <- sum(x=="NA", na.rm=T)   
   
 n3 <- sum(x=="", na.rm=T)   
   
 nmiss <- n1 + n2 + n3   
   
 ncomplete <- n-nmiss   
   
 out <- rbind(out, c(col.number=j, vname=vname, mode=mode(x), n.levels=length(unique(x)), ncomplete=ncomplete, miss.perc=nmiss/n)) }   
   
 out <- as.data.frame(out)   
   
 row.names(out) <- NULL   
   
 return(out)   
   
}   
  
miss.info(dat2)

## col.number vname mode n.levels ncomplete  
## 1 1 iso\_code character 1 433  
## 2 2 new\_cases numeric 400 432  
## 3 3 total\_deaths numeric 395 395  
## 4 4 total\_cases\_per\_million numeric 406 433  
## 5 5 icu\_patients numeric 248 249  
## 6 6 hosp\_patients numeric 250 249  
## miss.perc  
## 1 0  
## 2 0.0023094688221709  
## 3 0.0877598152424942  
## 4 0  
## 5 0.424942263279446  
## 6 0.424942263279446

library(mice)

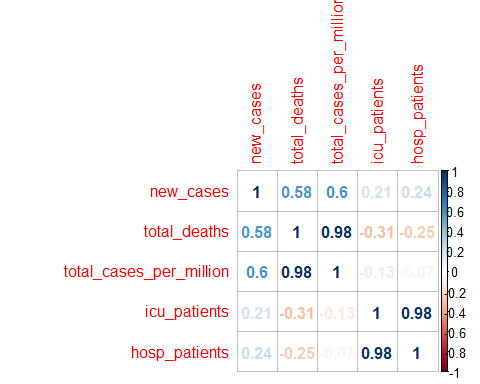
fit.mice.dat\_f <- mice(dat2, m = 1, maxit = 50, method = 'pmm', seed = 500)

dat\_f.imputed <- complete(fit.mice.dat\_f,1)  
datf <- as.data.frame(dat\_f.imputed)

library("corrplot")

## corrplot 0.84 loaded

dat3 <- datf[,-c(1)]  
M <- cor(dat3)  
corrplot(M, method="number")



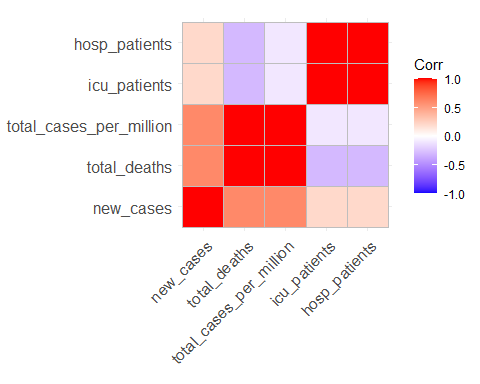
From this correlation plot among the selected variables, we can see that total\_deaths and total+cases\_per\_million, and icu\_patients and hosp\_patients in USA are very strongly correlated. This graph follows the Tufte"s Principle 1: Maximizing the data-ink ratio, within reason because it’s has less redundant data-ink. This graph follows the Tufte’s Principle 5: Provide the user with an overview and details on demand since we can easily figure out which variables are strongly correlated.

# Association Imporvements:

library(ggcorrplot)

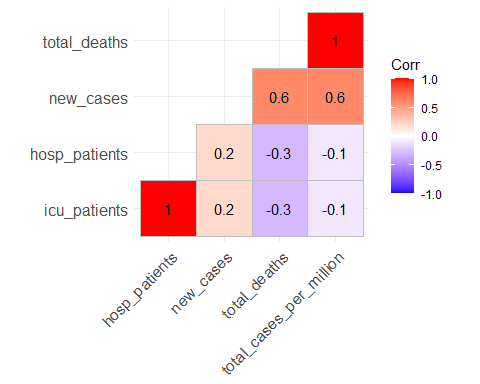
## Warning: package 'ggcorrplot' was built under R version 4.0.4

corr <- round(cor(dat3), 1)  
ggcorrplot(corr)



Here the red color is indicating the variables are very strongly correlated and blue color is indicating very weekly correlated. This graph follows the Tufte’s Principle 6: Utilize Layering & Separation since using different colors we can figure out the associations between the variables.

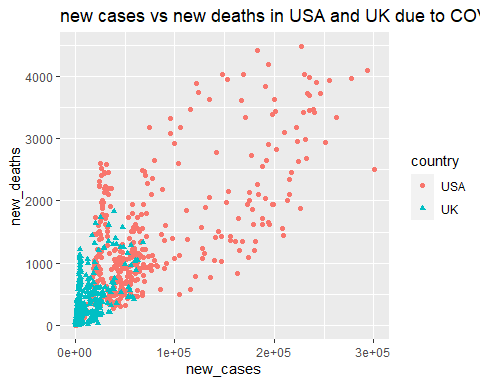
ggcorrplot(corr, hc.order = TRUE, type = "lower",  
 lab = TRUE)



From this correlation plot, we can very easily visualize that total\_deaths and total+cases\_per\_million, and icu\_patients and hosp\_patients in USA are very strongly correlated. Hence, when the hospitalized patients were high the icu\_patients also high in USA during the covid-19 pandemic. This graph follows the Tufte’s Principle 5: Provide the user with an overview and details on demand. and the Tufte’s Principle 6: Utilize Layering & Separation since using different colors we can figure out the associations between the variables and I think this diagram follow the 5 seconds rule since the color differentiation with number makes it easy to understand the story of the data within few seconds.

ggplot(dat\_2, aes(x=new\_cases, y=new\_deaths, shape = country, color = country)) +  
 geom\_point()+  
 ggtitle("new cases vs new deaths in USA and UK due to COVID-19")

## Warning: Removed 73 rows containing missing values (geom\_point).



From this graph, we can visualize that when the new\_cases is high the new\_deaths also high both in USA and UK. This graph follows the Tufte’s Principle 3: Maximize data density and the size of the data matrix, within reason and the Tufte’s Principle 6: Utilize Layering & Separation. The story behind this graph is understanding the relation between new\_deaths and new\_cases both in USA and UK and it’s obvious that when the new\_cases are high the new\_deaths also high.

# Visualizing Uncertainty

data<-read.csv(file="owid-covid-data\_new1.csv",header=TRUE)  
dim(data)

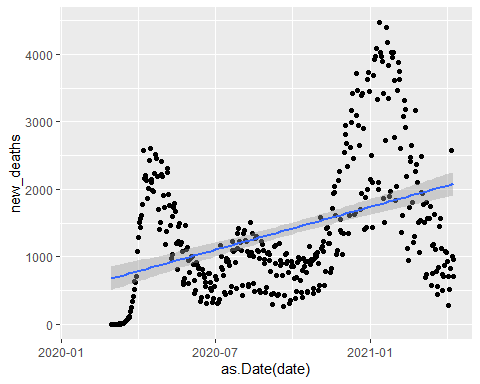
## [1] 80745 59

data1 <- data[data$iso\_code == "USA", ]   
dim(data1)

## [1] 445 59

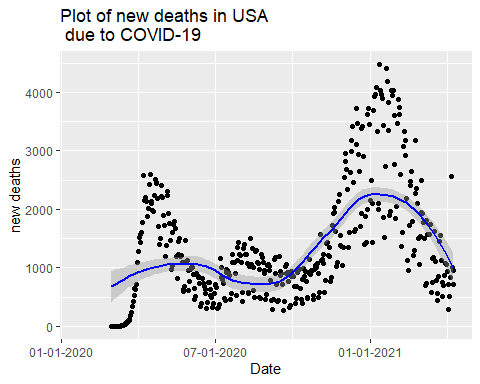
#library(ggplot2)  
library(reshape2)  
library(scales)  
ggplot(data1, aes(x = as.Date(date), y = new\_deaths)) +   
 geom\_point() +   
 geom\_smooth(method = "lm")

## `geom\_smooth()` using formula 'y ~ x'



ggplot(data1, aes(x = as.Date(date), y = new\_deaths)) +   
 geom\_point()+  
 geom\_smooth(method = loess, se=TRUE, color='blue')+  
 scale\_x\_date(labels = date\_format("%m-%d-%Y"))+  
 labs(x = "Date")+  
 labs(y = "new deaths")+  
 ggtitle("Plot of new deaths in USA \n due to COVID-19")

## `geom\_smooth()` using formula 'y ~ x'



This graph follows the Tufte’s Principle 3: Maximize data density and the size of the data matrix, within reason and the Tufte’s Principle 6: Utilize Layering & Separation and the Tufte’s Principle 5: Provide the user with an overview and details on demand. From this visualization we can see the trends and uncertainty of new deaths in USA due to COVID-19. Here we used locally estimated scatter-plot smoothing (LOESS) instead of lm method. It’s pretty clear that the new death rate is decreasing in USA last couple of days since the LOESS line is decreasing and the higher death rate happened at the beginning of 2021. So. it’s clearly tell the stories of how the new-deaths rate is either increasing or decreasing during this COVID-19 pandemic.

data<-read.csv(file="owid-covid-data417.csv",header=TRUE)  
dim(data)

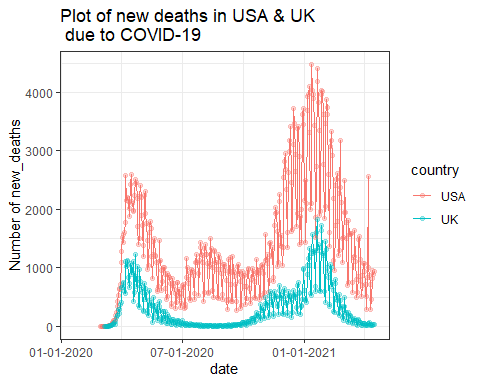
## [1] 82087 59

data1 <- data[data$iso\_code == "USA", ]   
data2 <- data[data$iso\_code == "GBR", ]   
data3<-rbind(data1,data2)  
country <- ifelse(data3$iso\_code == "USA",0,1)  
data4 <- cbind(country,data3)  
data4$country<-as.factor(data4$country)  
data4$country<-factor(data4$country,levels=c(0,1),labels=c("USA", "UK"))  
data5 <- data4[,-c(2)]  
dat2 <-data5[,c(1,4,9)]  
colnames(dat2)

## [1] "country" "date" "new\_deaths"

#library(ggplot2)  
library(gganimate)  
library(viridis)

library(scales)  
theme\_set(theme\_bw())  
plot <- ggplot(dat2, aes(x = as.Date(date), y = new\_deaths, color = country)) +  
 geom\_line()+  
 geom\_point(alpha = 0.3)+   
 scale\_x\_date(labels = date\_format("%m-%d-%Y"))+  
 labs(x = "date")+  
 labs(y = "Number of new\_deaths")+  
 ggtitle("Plot of new deaths in USA & UK \n due to COVID-19")  
plot



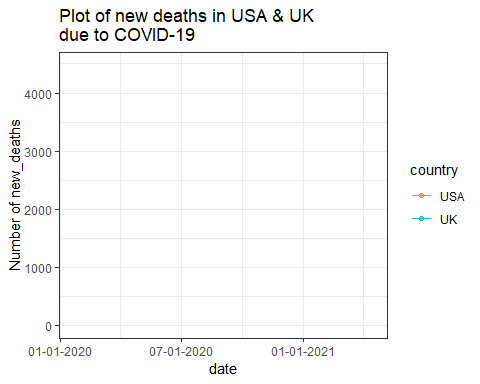
This graph follows the Tufte’s Principle 3: Maximize data density and the size of the data matrix, within reason and the Tufte’s Principle 6: Utilize Layering & Separation and the Tufte’s Principle 5: Provide the user with an overview and details on demand. From this visualization, we can compare the new\_deaths in USA & UK due to COVID-19. In addition, we can figure out the number of new\_deaths.So, the story behind this graph is telling the new\_deaths rate both in USA and UK during this COVID-19 pandemic.

# 

# Data Visualization improvement:

library(magick)

plot <- ggplot(dat2, aes(x = as.Date(date), y = new\_deaths, color = country)) +  
 geom\_line()+  
 geom\_point(alpha = 0.7)+  
 transition\_reveal(as.Date(date))+   
 scale\_x\_date(labels = date\_format("%m-%d-%Y"))+  
 labs(x = "date")+  
 labs(y = "Number of new\_deaths")+  
 ggtitle("Plot of new deaths in USA & UK \n due to COVID-19")  
animation <- animate(plot, nframes=100, renderer=magick\_renderer())   
animation



#image\_write\_gif(animation, 'animation.gif')

This graph follows the Tufte’s Principle 6: Utilize Layering & Separation and the Tufte’s Principle 5: Provide the user with an overview and details on demand. The story behind this animated graph is letting us understand how the new deaths rate increasing and decreasing over the time both in USA and UK during this COVID-19 pandemic.