744 Homework 2

Siyi Wang

2021/9/22

Statement of Jia You's online graphic

This graph is trying to show how effective the vaccine works for infectious diseases.

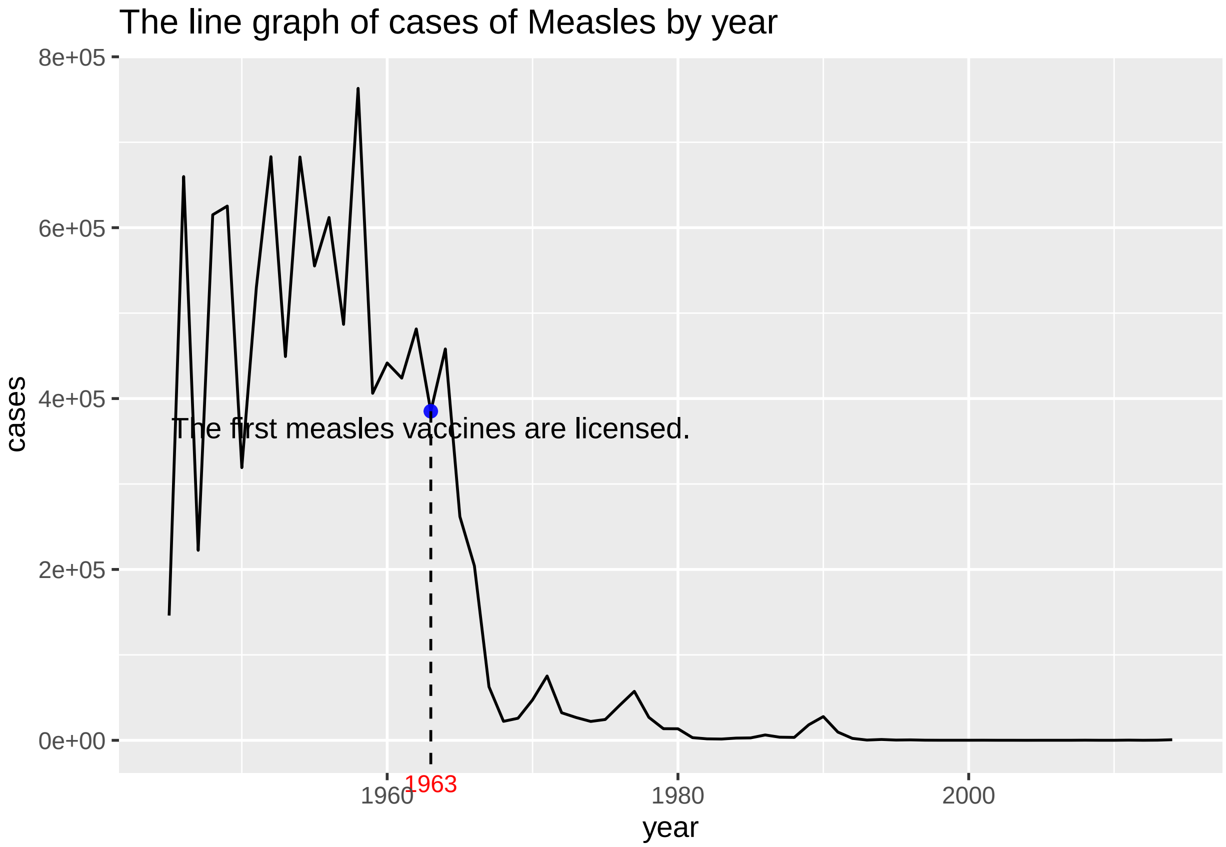
The example disease is Measles, and circles in this graph represent the number of reported cases of this disease.

It is easy to notice that, after 1963, the Measles vaccine introduced, the number of reported cases showed a general decreasing trend and reached a very low level after 1980, which proves the importance of the vaccine.

```
library(ggplot2)
data_dis<-read.csv("https://mac-theobio.github.io/DataViz/data/vaccine_data_online.csv")
```

First GG plot

```
Measles_data<-data_dis[which(data_dis$disease=="Measles"),]
Measles_vaccine<-Measles_data[which(Measles_data$vaccine!=FALSE),]
ggplot(data=Measles_data, aes(x=year, y=cases))+
  geom_line()+ geom_point(data=Measles_vaccine,aes(x=year,y=cases),
                          size=3,shape=20, color="blue",alpha=0.9)+
   geom_text(data = Measles_vaccine, aes(x =year, y = cases*0.95, label = vaccine))+
   annotate( geom = "segment",x=Measles_vaccine$year,
          xend = Measles_vaccine$year,y=Measles_vaccine$cases,yend = -Inf,
          lty="dashed")+
  annotate(geom = "text",x=Measles_vaccine$year,y=10,
          label=Measles_vaccine$year,
          vjust=3,color="red",size=3)+
  coord_cartesian(clip = "off")+labs(title = "The line graph of cases of Measles by year")
```



Generally, for time series data, line graph is a better choice for showing changes over time. Compared with Jia You's graph, it is easy to understand that the cases dropped dramatically after 1964 from the line graph. Moreover, the line graph shows a clear comparison about cases number before and after vaccines licensed. The drawbacks of Jia You's graph are those circles are too large for representing number of cases, the overlaps of circles make the graph become a little bit difficult to read and it is not precise enough for indicating the number of cases by the area of circles.

Second GG plot

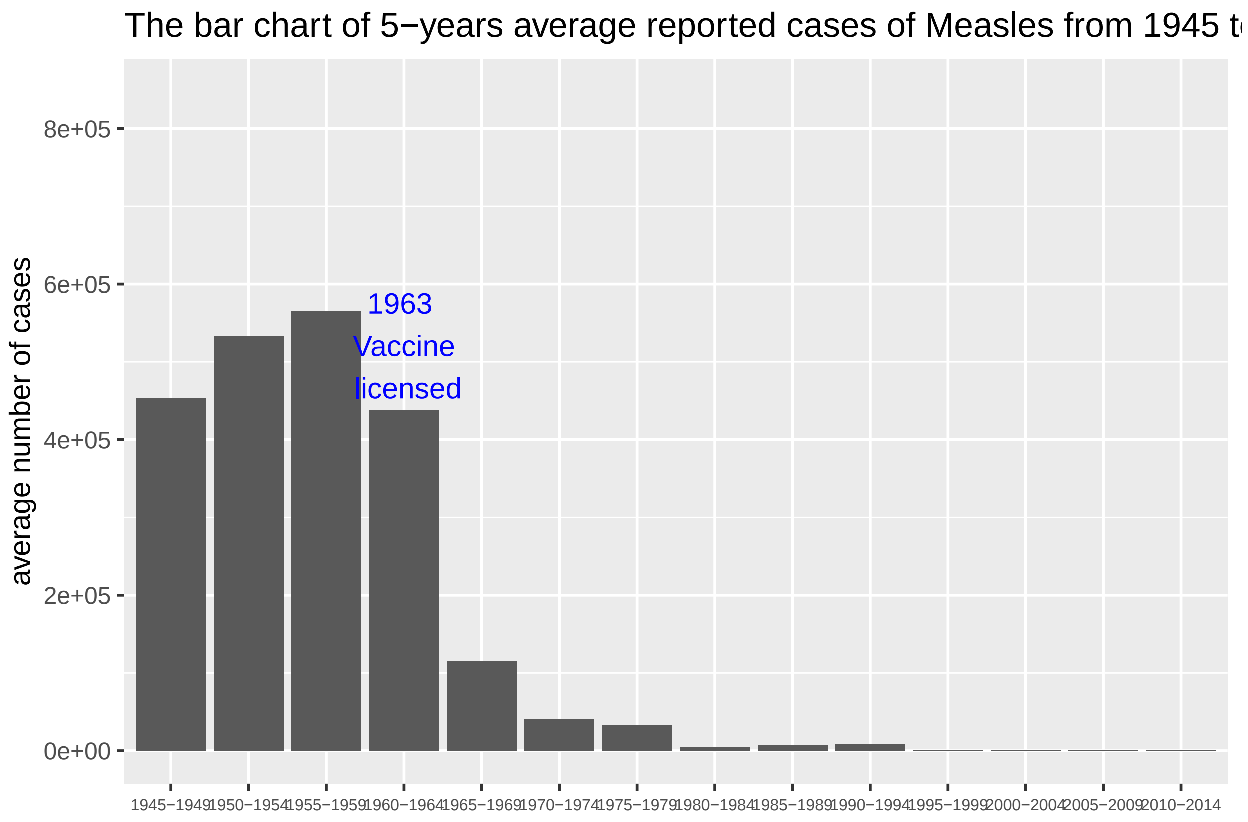
```
dim(Measles_data)
```

```
## [1] 70  8
```

```
cases_meanbyfive=c()
year_label=c()

for (i in 1:14) {
    cases_meanbyfive[i]=mean(Measles_data$cases[(5*i-4):(5*i)])
    year_label[i]<-paste(1944+(5*i-4), 1944+(5*i), sep = "-")
}

Measles_data_rough=data.frame(cbind(as.numeric(cases_meanbyfive),year_label))
area_text <- rep(NA, 14)
area_text[4]="1963 \n Vaccine \n licensed"
ggplot(Measles_data_rough, aes(x=year_label, y=cases_meanbyfive))+
    geom_bar(stat = "identity")+
    theme(axis.text.x = element_text(angle=0, vjust=0.5,size=6))+
    labs(y="average number of cases",x='',
         title = "The bar chart of 5-years average reported cases of Measles from 1945 to 2014")+
    geom_text(label=area_text, colour = "blue",
                   position=position_stack(1.5), vjust=1.5)
```



If our goal is to show the effectiveness of vaccines, it is not necessary for us to draw data points by year. Here, the histogram shows the 5-years average reported cases of Measles, which sacrifices the data accuracy, but shows the effectiveness of vaccines in a simple and efficient approach, and weakens the effect of data fluctuations. In the first four periods, the number of cases fluctuates in a certain level, but after the period contains the event vaccine licensed, the height of the bar shrinks to nearly $\frac{1}{4}$ of previous periods. And then though some fluctuations still happen, the number of cases shrinks in general and after the period of 1990 to 1994, it is nearly disappeared.