code

Yue Wu

May 9, 2018

Data

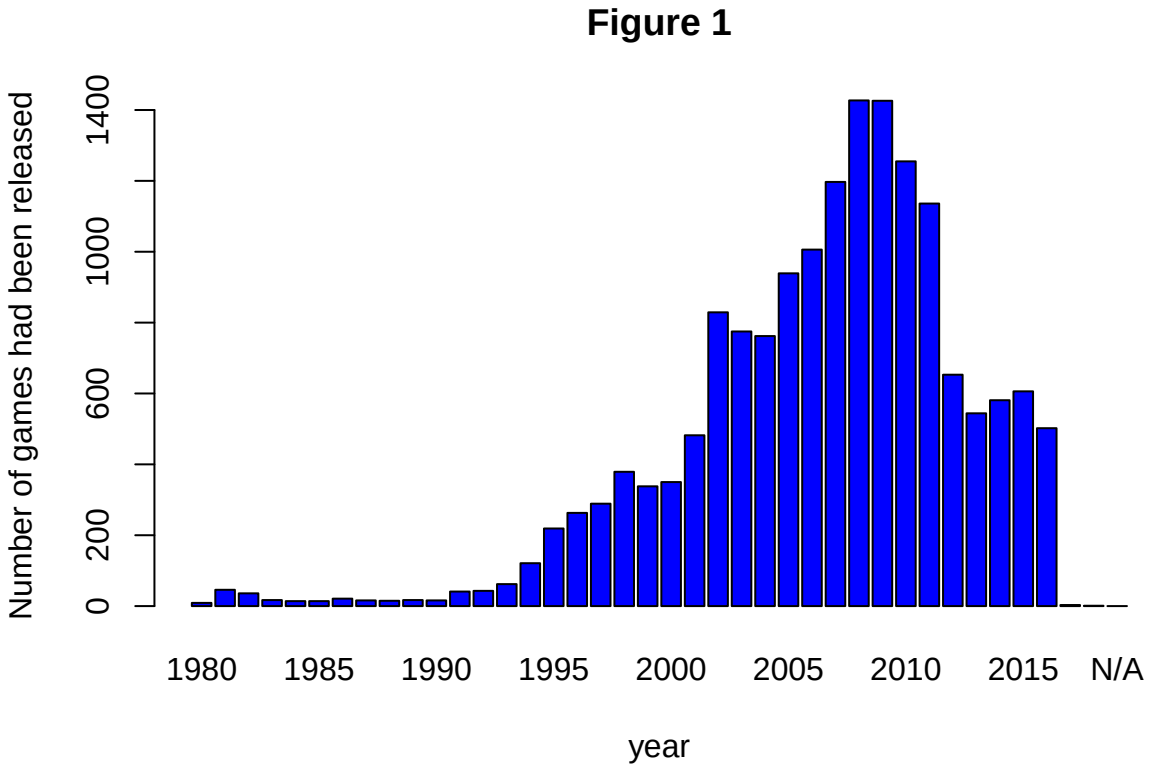
The data is from kaggle: https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings/downloads/Video_Games_Sales_as_at_22_Dec_2016.csv/

First, let's load the data into a vector called "gamedata"

```
gamedata<-read.csv('Video_Games_Sales_as_at_22_Dec_2016.csv') #Format into data frame
```

Before we clean the data, I would like to plot the number of games released by defferent years. If I do the plot after clean data, the plot will be biased because some games were deleted.

```
gamedata$Year_of_Release[gamedata$Year_of_Release=="N/A"]<-NA
plot(gamedata$Year_of_Release,col = "blue",
     ylab = "Number of games had been released",
     xlab = "year",
     main = "Figure 1"
)
```



Second I deleted some variables because we will not use them in our analysis.

```
gamedata<-gamedata[,-6]
gamedata<-gamedata[,-6]
gamedata<-gamedata[,-6]
gamedata<-gamedata[,-6]
gamedata<-gamedata[,-8]
gamedata<-gamedata[,-9]
gamedata<-gamedata[,-10]
head(gamedata)
```

```
##                       Name Platform Year_of_Release         Genre Publisher
## 1                Wii Sports      Wii            2006        Sports  Nintendo
## 2         Super Mario Bros.      NES            1985      Platform  Nintendo
## 3            Mario Kart Wii      Wii            2008        Racing  Nintendo
## 4         Wii Sports Resort      Wii            2009        Sports  Nintendo
## 5 Pokemon Red/Pokemon Blue       GB            1996  Role-Playing  Nintendo
## 6                    Tetris       GB            1989        Puzzle  Nintendo
##   Global_Sales Critic_Score User_Score Developer
```

```
## 1        82.53           76          8  Nintendo
## 2        40.24           NA
## 3        35.52           82        8.3  Nintendo
## 4        32.77           80          8  Nintendo
## 5        31.37           NA
## 6        30.26           NA
```

Then we have do deal with the missing value in the critic score. I consider the missing values are completey at ramdom(MCAR).

```
gamedata<-gamedata[!is.na(gamedata$Critic_Score),]
```

Because the variable global sales is highly skewed. I have to use "log" to transform this variable.

```
gamedata$logGlobal_sale<-log(gamedata$Global_Sales)
```

Model

First, I have to pick one between user score and critic score to measure the quality of games.
To do that, I am giong to find the relationships between user and critic score and global sales.

```
gamedata$Critic_Score<-as.numeric(gamedata$Critic_Score)
gamedata$User_Score<-as.numeric(gamedata$User_Score)
fit1<-lm(gamedata$Global_Sales ~ gamedata$User_Score + gamedata$Critic_Score)
summary(fit1)
```

```
##
## Call:
## lm(formula = gamedata$Global_Sales ~ gamedata$User_Score + gamedata$Critic_Score)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.877 -0.615 -0.273  0.162 81.633
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
```

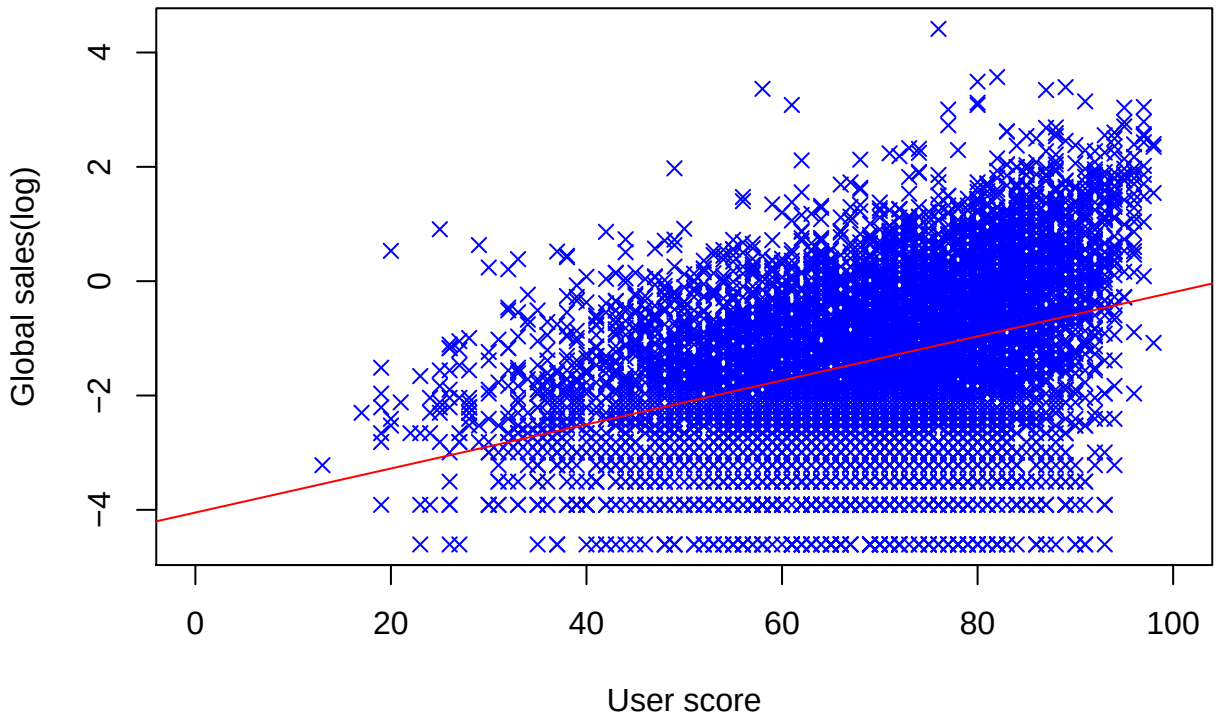
```
## (Intercept)           -1.175853   0.114675 -10.254  < 2e-16 ***
## gamedata$User_Score   -0.007058   0.001217  -5.798 6.95e-09 ***
## gamedata$Critic_Score  0.034611   0.001469  23.559  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.758 on 8134 degrees of freedom
## Multiple R-squared:  0.06412,    Adjusted R-squared:  0.06389
## F-statistic: 278.7 on 2 and 8134 DF,  p-value: < 2.2e-16
```

Because the result shows that the user-score and global sales have nagetive relationship, it doesn't make sense. So i chose critic score to measure the quality.

Here is the plot of the critic score and global sales. Make sure they have positive relationship.

```
par(mar=c(4,4,4,1),cex=1)
plot(gamedata$Critic_Score,gamedata$logGlobal_sale,
     pch=4,
     ylab = "Global sales(log)",
     xlab = "User score",
     col = 'blue',
     main = 'Figure 2',
     xlim = c(0,100))
abline(lm(gamedata$logGlobal_sale~gamedata$Critic_Score),col = "red")
```

Figure 2



My equation is Global sales(log) = b0 + b1 * Critic score + b2 * Genre + b3 * platform + e

```
fit<-lm(data = gamedata, logGlobal_sale ~ Genre + Platform + Critic_Score )
#summary(fit)
```

The result from the summary of the equation (I don't print the summary because it is really long list):

—Critic_score have positive relationship with global sales. Each point increases in critic score will lead to 0.04% increases in global sales.

—Some polular genres have positive relationship with global sales.(Shooters,Sport,Fighting)

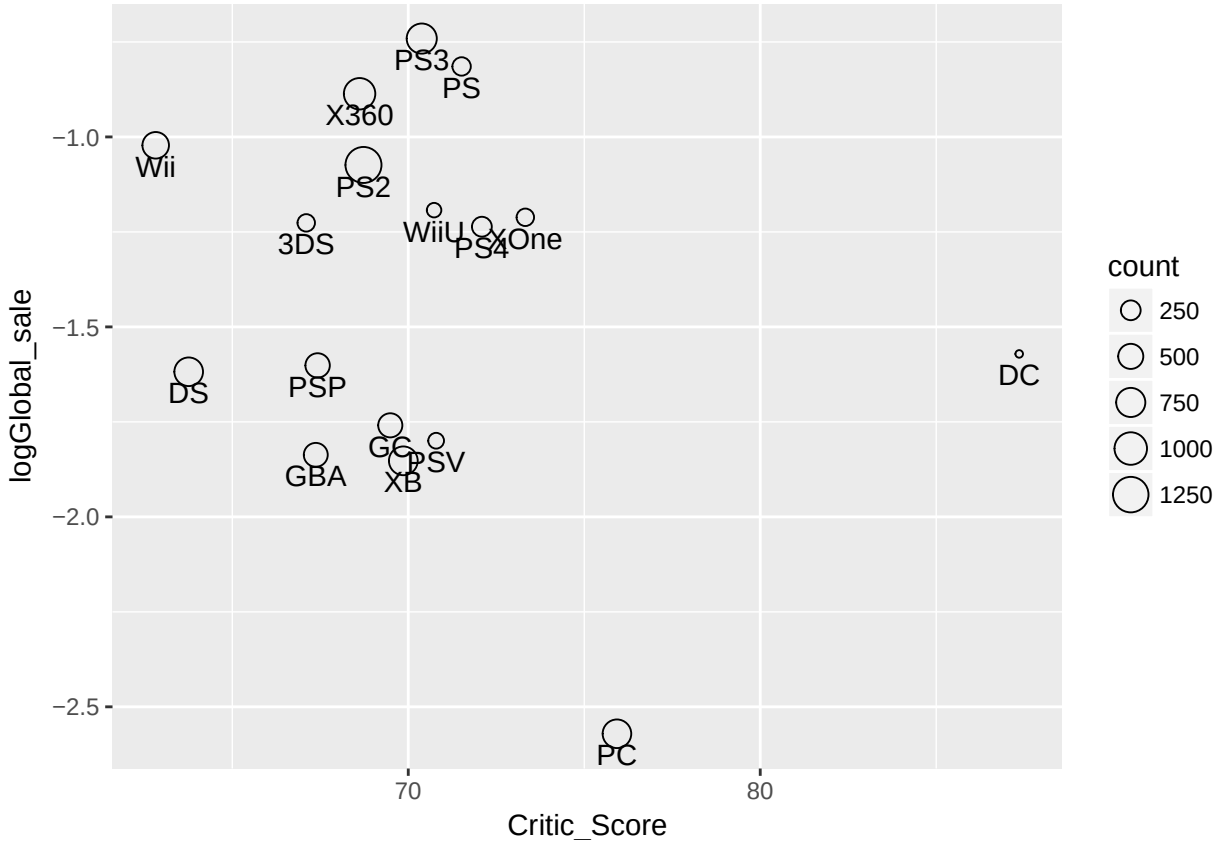
—Some good device have positive relationship with global sales.(PS,XBOX)

Findings

To find out what device is polular, I draw the graphy below:

```
Platform<-aggregate(gamedata[,],list(gamedata$Platform),mean)
count<-count(gamedata$Platform)
Platform$count<-count$freq
```

```
Platform$device<-Platform$Group.1
ggplot(Platform,aes(x=Critic_Score,y=logGlobal_sale,size=count))+
  geom_point(shape=21)+
  geom_text(data=Platform,
            aes(x=Critic_Score,
                y=logGlobal_sale,
                label=device),
            vjust=1.5,size=4)
```



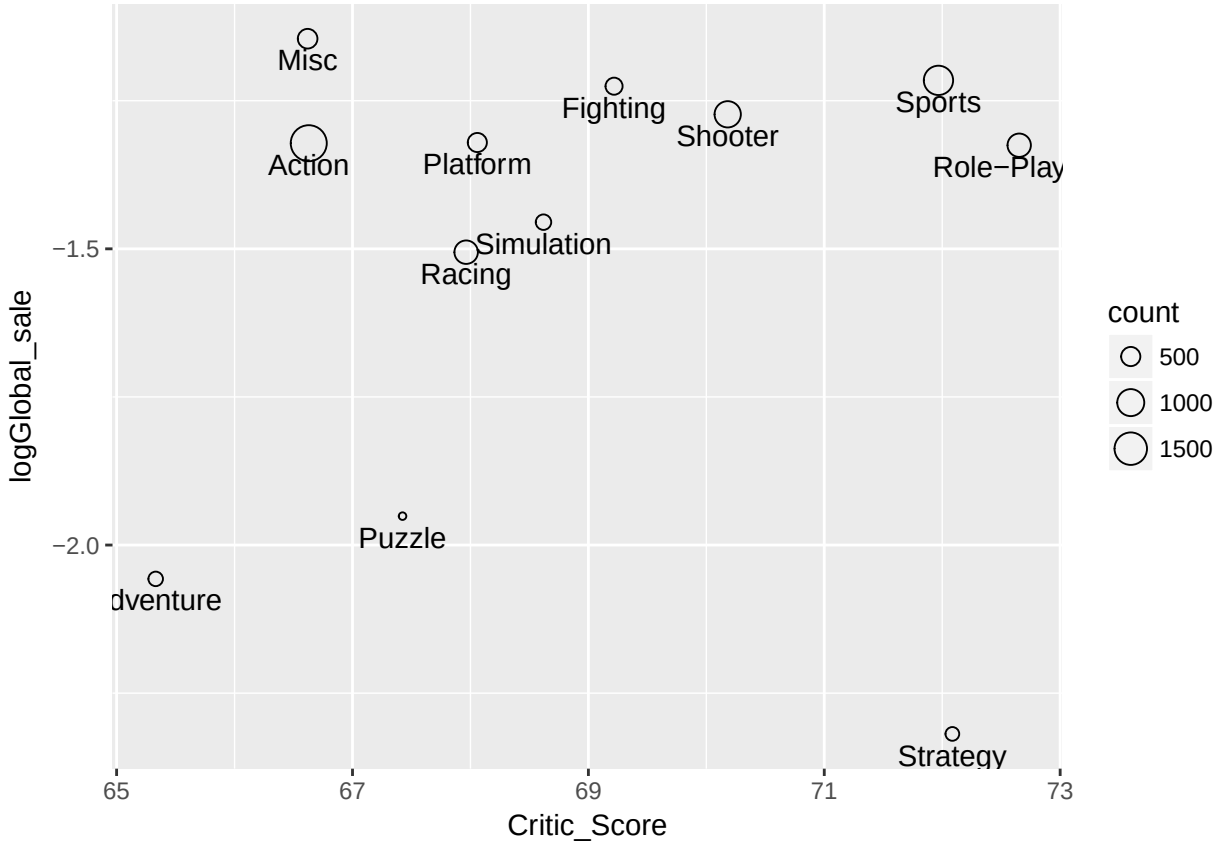
This graphy shows that the mean global sales and mean critic scores of each device in the data, the circle size measures the numbers of games.

Devices with higher technology have more games, and thier games have more global sales.

To find out what genre of game is polular, I drwa the graphy below:

```
Genre<-aggregate(gamedata[,],list(gamedata$Genre),mean)
count<-count(gamedata$Genre)
```

```
Genre$count<-count$freq
Genre$type<-Genre$Group.1
ggplot(Genre,aes(x=Critic_Score,y=logGlobal_sale,size=count))+
  geom_point(shape=21)+
  geom_text(data=Genre,
            aes(x=Critic_Score,
                y=logGlobal_sale,
                label=type),
            vjust=1.5,size=4)
```



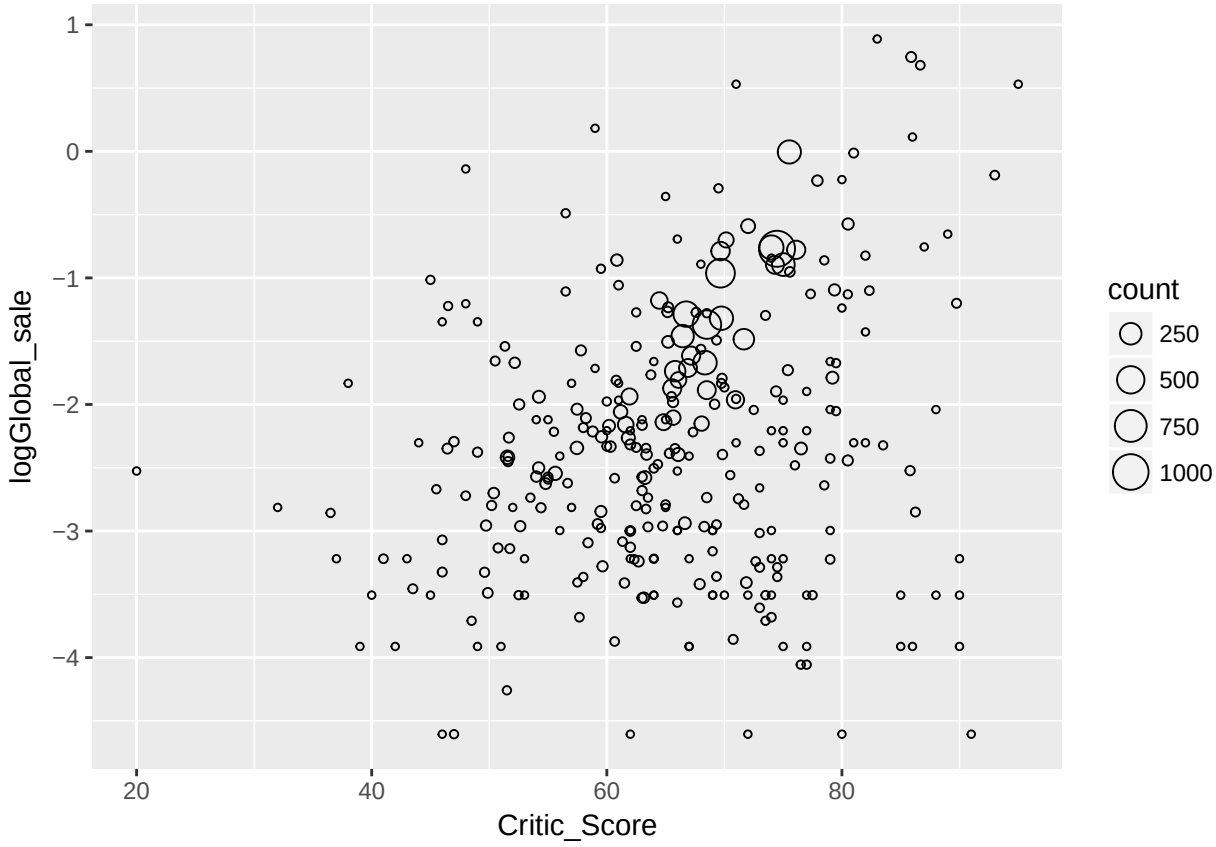
This graphy shows the mean global sales and critic score for each genre of game. The size of circle measures the number of games for certain gerne.

Shtooer, Fighting and Sports game have more global sales because those games are stimulatte people's brain.

To find out what companies are more successful, I draw the praphy below:

```
Publisher<-aggregate(gamedata[,],list(gamedata$Publisher),mean)
count<-count(gamedata$Publisher)
```

```
Publisher$count<-count$freq
Publisher$company<-Publisher$Group.1
ggplot(Publisher,aes(x=Critic_Score,y=logGlobal_sale,size=count))+
  geom_point(shape=21)
```



This graphy shows the mean global sales and critic score for each company. The size of circle measures the number of games for certain company.

I consider the companies with large circle are more successful because they produce more games.

Successful companies have more resources and better developer team, so their games have better quality(critic score) and more global sales. Also, successful companies have better advertisment.