HW#2

Vazgen Tadevosyan

June 29, 2018

In this homework you will work on video games dataset containing information about popular video games, their sales in North America, Europe, Japan and globally in the world. In the dataset ratings by critics and users are presented and ratings of the games.

Solve the problems and submit the .Rmd file.

WARNINGS!!! (If not done you will lose points.) 1) Make sure to put titles on the plots and texts on axes. 2) If the plot is not interpretable, zoom on "x" or "y" axes to make the graph more interpretable (P4,P5, P7 and P8).

P1)

Import the dataframe in R and with the use of dplyr subset it using the following information.

-remove columns Publisher, JP_Sales (Sales in Japan), Critic_Count, User_Count and Developer. (1p) -Multiply the numbers in NA_Sales, EU_Sales and GP_Sales by 1 million as they are given in millions of sales. (1p) -include only those for which NA_Sales>=20000, EU_Sales>=20000 and Ranking is among Everyone("E"),Mature("M"), Teen("T"), Everyone 10+("E10+") and Adults Only ("AO"). (1p)

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
video_games<-read.csv("Video_Games.csv",stringsAsFactors = F)
str(video_games)
```

```
## 'data.frame':    16719 obs. of  15 variables:
##  $ Name          : chr  "Wii Sports" "Super Mario Bros." "Mario Kart Wii" "Wii Sports Resort" ...
##  $ Platform      : chr  "Wii" "NES" "Wii" "Wii" ...
##  $ Year          : chr  "2006" "1985" "2008" "2009" ...
##  $ Genre         : chr  "Sports" "Platform" "Racing" "Sports" ...
##  $ Publisher     : chr  "Nintendo" "Nintendo" "Nintendo" "Nintendo" ...
##  $ NA_Sales      : num  41.4 29.1 15.7 15.6 11.3 ...
##  $ EU_Sales      : num  28.96 3.58 12.76 10.93 8.89 ...
##  $ JP_Sales      : num  3.77 6.81 3.79 3.28 10.22 ...
##  $ Global_Sales: num  82.5 40.2 35.5 32.8 31.4 ...
##  $ Critic_Score: int  76 NA 82 80 NA NA 89 58 87 NA ...
##  $ Critic_Count: int  51 NA 73 73 NA NA 65 41 80 NA ...
##  $ User_Score  : chr  "8" "" "8.3" "8" ...
##  $ User_Count  : int  322 NA 709 192 NA NA 431 129 594 NA ...
##  $ Developer   : chr  "Nintendo" "" "Nintendo" "Nintendo" ...
##  $ Rating      : chr  "E" "" "E" "E" ...
```

```
#1
video_games<-video_games%>%
  select(-c("Publisher","JP_Sales","Critic_Count","User_Count","Developer"))%>%
  mutate(NA_Sales=NA_Sales*1000000,
         EU_Sales=EU_Sales*1000000,
         Global_Sales=Global_Sales*1000000)%>%
  filter(NA_Sales>=20000,EU_Sales>=20000,Rating %in% c("E","M","T","E10+","AO"))
#
```

P2)

Use data cleaning tools to clean the data.

(a) Look at the columns which are either numeric or integer. Make sure they contain only numbers or NA's (nothing else). (1p)

(b) Critic scores can be from 0 to 100 and users scores from 0 to 10. If there are values not from these intervals clean that observations using ifelse statement. (2p)

(c) Look at the Genres: check if all categories are unique and if not, clean them so that there are no duplicate names. (2p)

```
sapply(video_games,class)
```

```
##         Name     Platform         Year        Genre     NA_Sales 
##  "character"  "character"  "character"  "character"    "numeric" 
##     EU_Sales Global_Sales Critic_Score   User_Score       Rating 
##    "numeric"    "numeric"    "integer"  "character"  "character" 
```

```
#as we see we should make year and user_score as numeric
unique(video_games$Year)
```

```
##  [1] "2006" "2008" "2009" "2005" "2007" "2010" "2013" "2004" "2002" "2001"
## [11] "2011" "2012" "2014" "1997" "1999" "2015" "2016" "2003" "1998" "1996"
## [21] "2000" "N/A"  "1994" "1992"
```

```
video_games$Year<-gsub("N/A",NA,video_games$Year)
video_games$Year<-as.numeric(video_games$Year)

#for user_score
unique(video_games$User_Score)
```

```
##  [1] "8"   "8.3" "8.5" "6.6" "8.4" "8.6" "7.7" "6.3" "7.4" "8.2" "9"  
## [12] "7.9" "8.1" "8.7" "7.1" "3.4" "5.3" "4.8" "3.2" "8.9" "6.4" "7.8"
## [23] "7.5" "2.6" "7.2" "9.2" "7"   "7.3" "4.3" "7.6" "5.7" "5"   "9.1"
## [34] "6.5" "tbd" "8.8" "6.9" "9.4" "6.8" "6.1" "6.7" "5.4" ""    "4"  
## [45] "4.9" "4.5" "6.2" "4.2" "6"   "3.7" "4.1" "5.8" "5.6" "5.5" "4.4"
## [56] "4.6" "5.9" "3.9" "9.3" "3.1" "2.9" "5.2" "3.3" "4.7" "5.1" "3.5"
## [67] "2.5" "1.9" "3"   "2.7" "2.2" "2"   "9.5" "2.1" "3.6" "2.8" "1.8"
## [78] "3.8" "1.6" "9.6" "2.4" "1.7" "1.5" "999" "0.7" "1.2" "0.2" "0.5"
```

```r
video_games$User_Score<-ifelse(video_games$User_Score %in% c("tbd",""),NA,video_games$User_Score)
video_games$User_Score<-as.numeric(video_games$User_Score)


##Critic score and User score
video_games$Critic_Score<-ifelse(video_games$Critic_Score  %in% c(0:100),video_games$Critic_Score,NA)
unique(video_games$Critic_Score)
```

```
##  [1] 76 82 80 89 58 87 91 61 97 95 77 88 83 94 93 85 86 98 96 90 84 73 74
## [24] 78 92 71 72 68 62 49 NA 67 81 66 56 79 70 59 64 75 60 63 69 50 25 42
## [47] 44 55 48 57 29 47 65 54 20 53 37 38 33 52 30 32 43 45 51 40 46 34 39
## [70] 35 41 36 28 31 26 27 19 23 24 21 17
```

```r
video_games$User_Score<-ifelse(video_games$User_Score >=0 & video_games$User_Score <=10,video_games$User
unique(video_games$User_Score)
```

```
##  [1] 8.0 8.3 8.5 6.6 8.4 8.6 7.7 6.3 7.4 8.2 9.0 7.9 8.1 8.7 7.1 3.4 5.3
## [18] 4.8 3.2 8.9 6.4 7.8 7.5 2.6 7.2 9.2 7.0 7.3 4.3 7.6 5.7 5.0 9.1 6.5
## [35]  NA 8.8 6.9 9.4 6.8 6.1 6.7 5.4 4.0 4.9 4.5 6.2 4.2 6.0 3.7 4.1 5.8
## [52] 5.6 5.5 4.4 4.6 5.9 3.9 9.3 3.1 2.9 5.2 3.3 4.7 5.1 3.5 2.5 1.9 3.0
## [69] 2.7 2.2 2.0 9.5 2.1 3.6 2.8 1.8 3.8 1.6 9.6 2.4 1.7 1.5 0.7 1.2 0.2
## [86] 0.5
```

```r
#Genres
unique(video_games$Genre)
```

```
##  [1] "Sports"       "Racing"       "Platform"     "Misc"
##  [5] "Action"       "Puzzle"       "Shooter"      "Fighting"
##  [9] "Simulation"   "Role-Playing" "SHooter"      "   Sports"
## [13] "Adventure"    "Strategy"     "ACTION"       "SPORTS"
```

```r
table(video_games$Genre)
```

```
## 
##       Sports       Action       ACTION    Adventure     Fighting 
##            1         1507            1          224          315 
##         Misc     Platform       Puzzle       Racing Role-Playing 
##          484          406          107          571          471 
##      Shooter      SHooter   Simulation       Sports       SPORTS 
##          713            1          275          919            1 
##     Strategy 
##          149
```

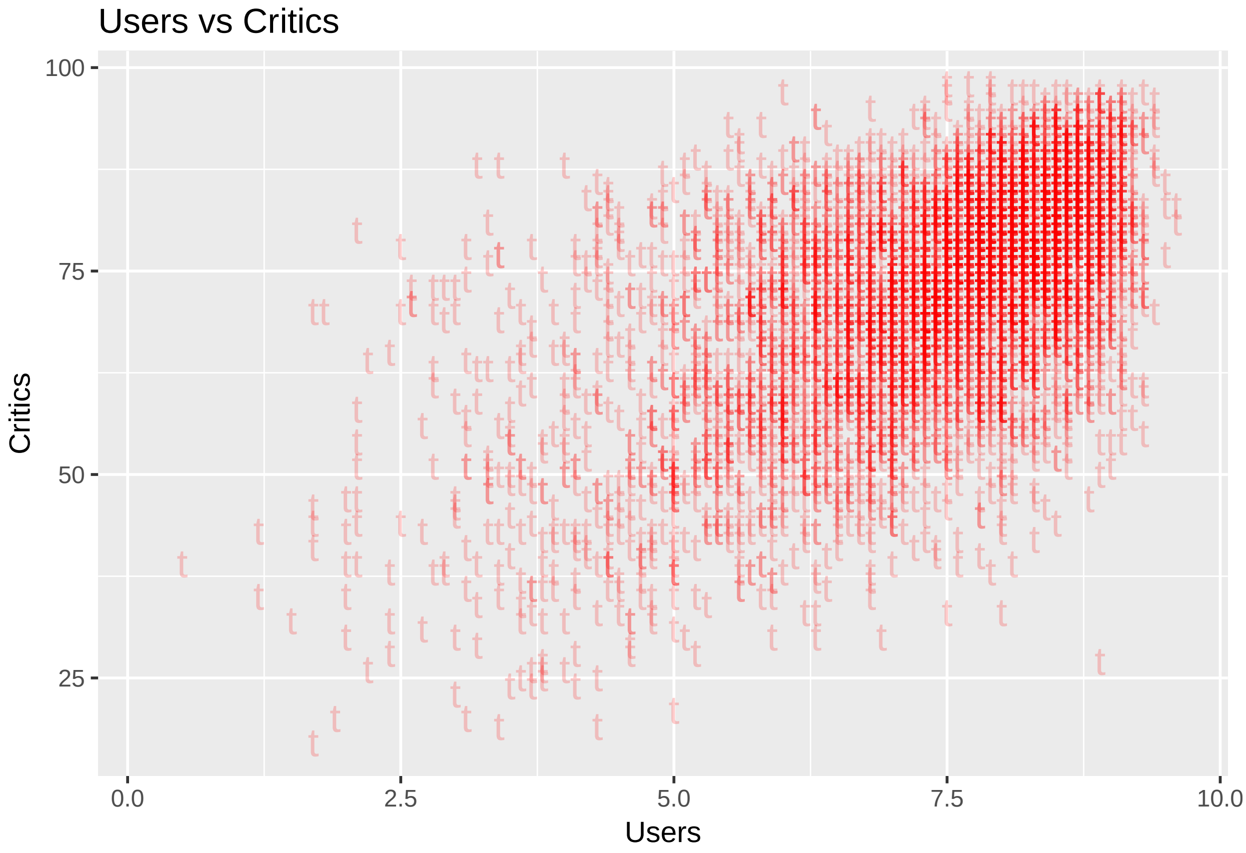
```r
video_games$Genre<-toupper(trimws(video_games$Genre))
table(video_games$Genre)
```

```
## 
##       ACTION    ADVENTURE     FIGHTING         MISC     PLATFORM 
##         1508          224          315          484          406 
##       PUZZLE       RACING ROLE-PLAYING      SHOOTER   SIMULATION 
##          107          571          471          714          275 
##       SPORTS     STRATEGY 
##          921          149
```

P3)

Create a scatterplot displaying how User scores and Critics score are interconnected -make the point shape triangle, color red and transparency 20%. Explain what you see in the graph. (1p)

```
library(ggplot2)
ggplot(video_games,aes(User_Score,Critic_Score))+
  geom_point(col="red",shape='triangle',alpha=0.2,size=5)+labs(title="Users vs Critics",x="Users",y="Cr
```

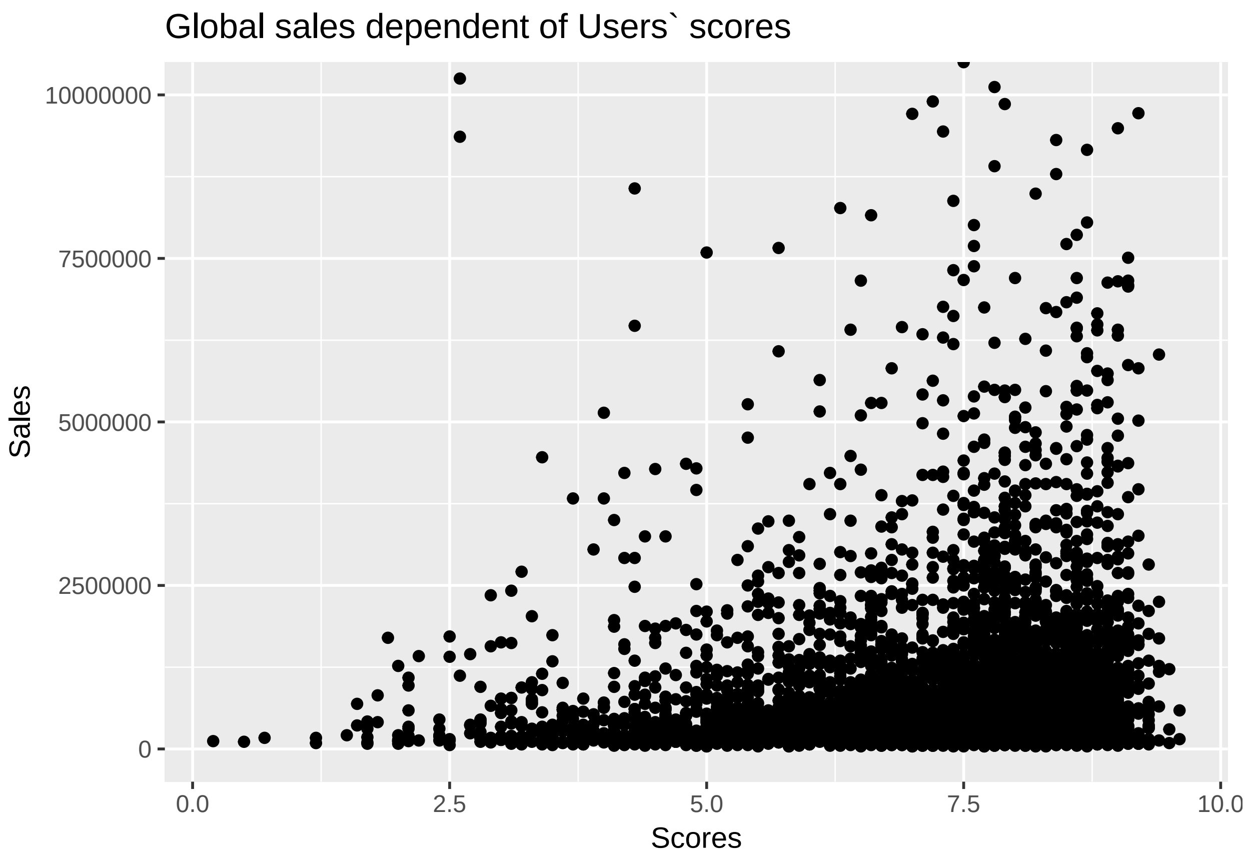


```
##from graph we can state that  most of games have a positive feedback because
#there is density in right above part of the plot.Big part of the cases games have more or less the sam
#overall there is linear regression and we can see decreasing variance
```

P4)

Construct a graph showing how the global sales of the game is dependent of a score given by the user and explain what you see in the graph. (Hint! ?options to display values without "e" short notation) (1p)

```
options(scipen=999)
ggplot(video_games,aes(User_Score,Global_Sales))+geom_point()+
  coord_cartesian(ylim = c(0,10000000))+
  labs(title="Global sales dependent of Users` scores",x="Scores",y="Sales")
```



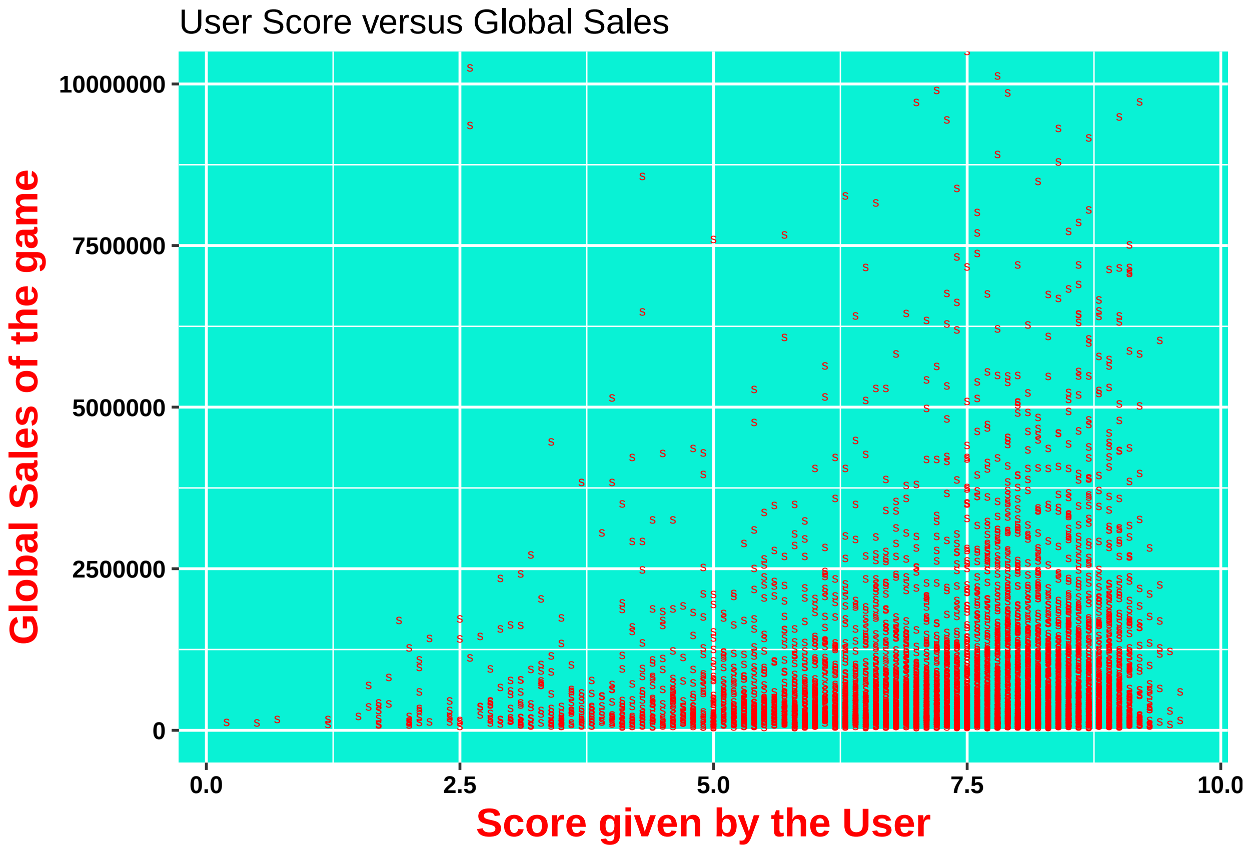
```
##we can conclude that sales  are high when it has good positive feedback ,
#when score is increasing sales increases as well
```

P5)

Make previous plot more appealing using the following. (1p) -x axis name – “Score given by the User” color red, bold size=15 -y axis name – “Global Sales of the game” color red, bold size=15 -points (shape - square, color-red, size- 1.5) -title of the plot – “User Score versus Global Sales” - Make panel background color #09f2d5 - axis texts bold black

---

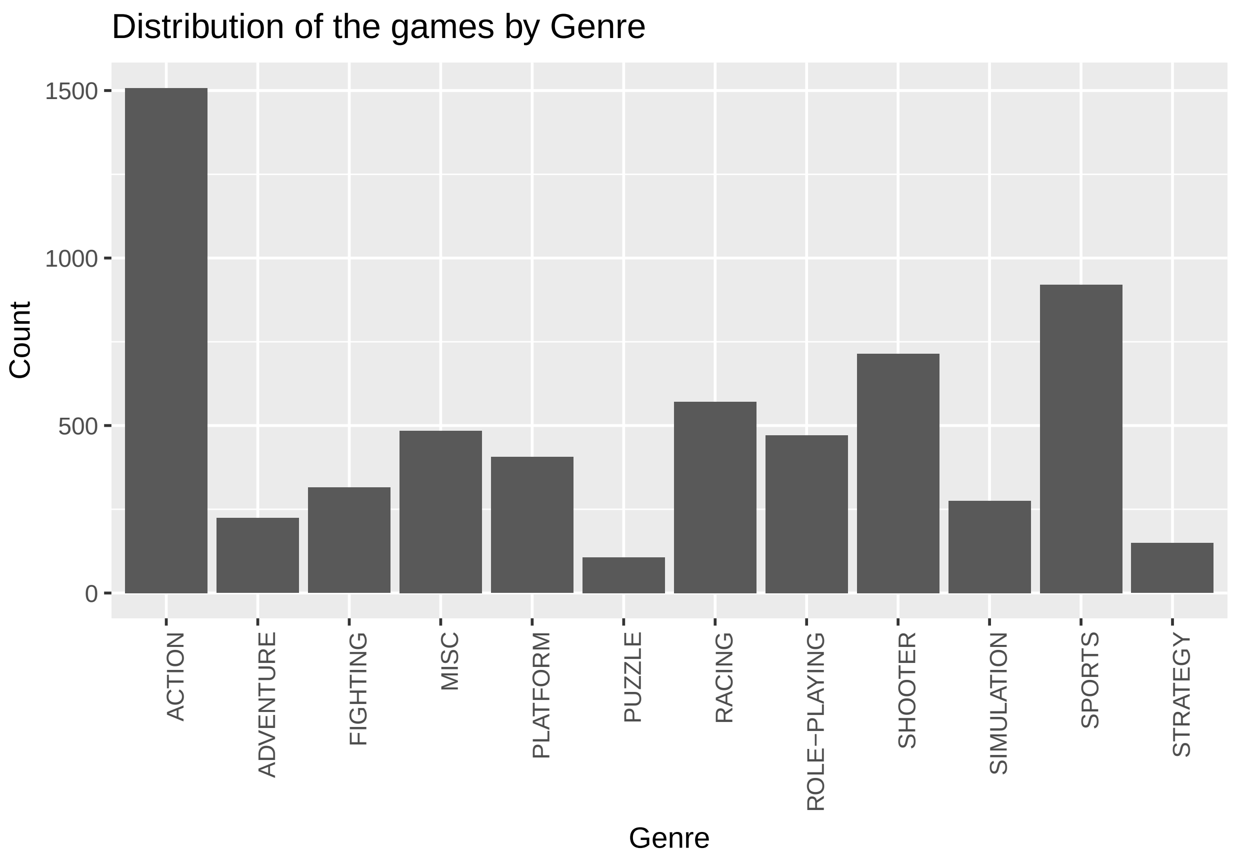
```
options(scipen=999)
ggplot(video_games,aes(User_Score,Global_Sales))+geom_point(shape="square",size=1.5,
col="red")+
  coord_cartesian(ylim = c(0,10000000))+
  labs(title="User Score versus Global Sales",x="Score given by the User",y="Global Sales of the game")+
  theme(axis.title = element_text(size = 15,face = "bold",color = "red"),
        panel.background = element_rect(fill="#09f2d5"),
        axis.text = element_text(color ="black",face="bold"))
```



P6)

Create a histogram to find the distribution of the games by Genre. What are the top 3 Genres. Rotate Genre names on “x” axis to avoid overlapping text (Hint! ?element_text, ?theme) (2p)

```
ggplot(video_games,aes(Genre))+geom_histogram(stat = "count")+
  theme(axis.text.x =element_text(angle =90,hjust = 1))+
  labs(title="Distribution of the games by Genre",x="Genre",y="Count")
```

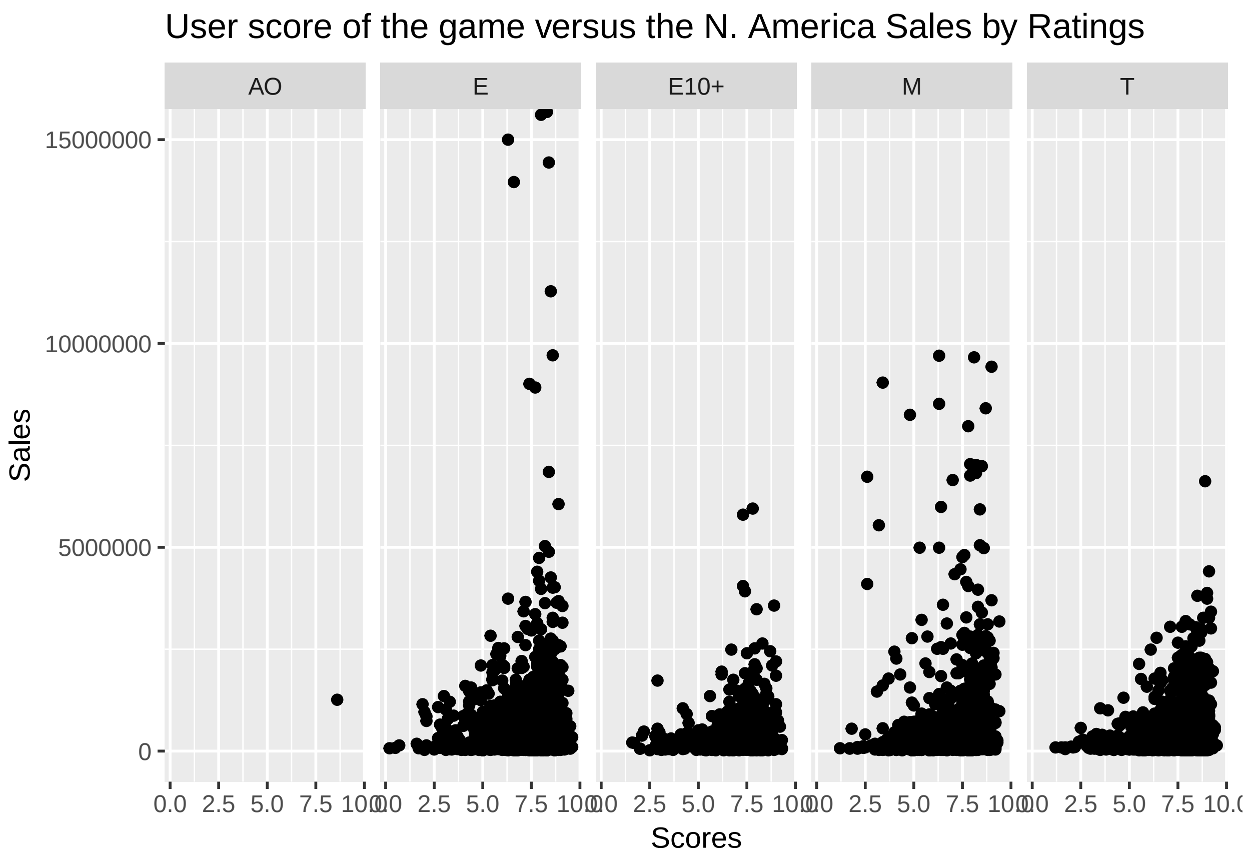


```
#1-Action-about 1500
#2-Sport -about 800
#3-Shooter-about 600
```

P7)

Define the Rating as Factor and use faceting to plot the User score of the game versus the North America Sales for different Ratings. Make comment about the results.(2p)

```
video_games$Rating<-factor(video_games$Rating)
ggplot(video_games,aes(User_Score,NA_Sales))+geom_point()+
  facet_grid(.~Rating)+coord_cartesian(ylim = c(0,15000000))+
  labs(title="User score of the game versus the N. America Sales by Ratings",x="Scores",y="Sales")
```

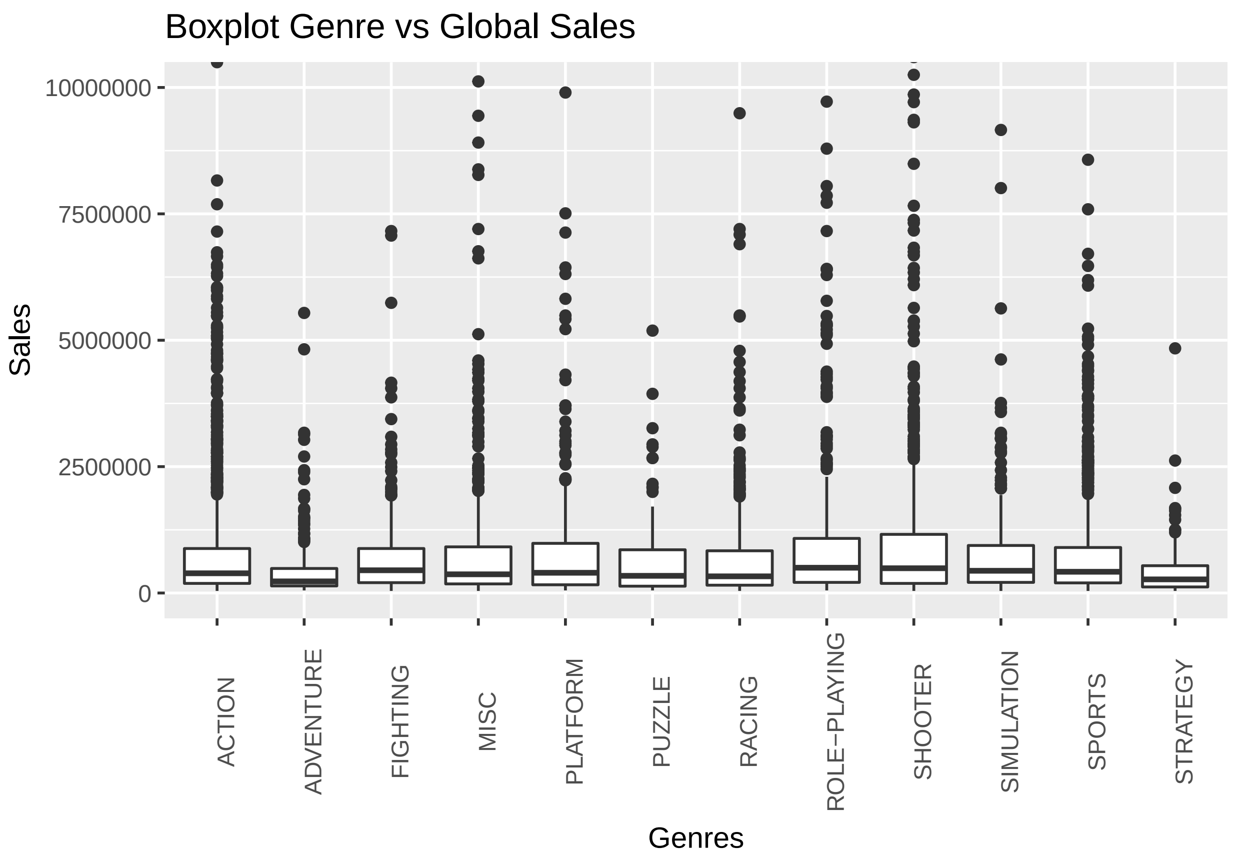


```
#we see in  Rating group M variance is violated and there are games with high sales even
# if they low user score,in rating group E we have outlighers with extremely high sales.
```

P8)

Create a boxplot where "x axis" represents the Genre and "y axis" the Global Sales of the video game for a particular Genre. Make the text on "x" axis vertical (Hint! ?theme, ?element_text). Make some comments.(2p)

```
ggplot(video_games,aes(Genre,Global_Sales))+geom_boxplot()+
  coord_cartesian(ylim = c(0,10000000))+theme(axis.text.x =element_text(angle =90,vjust = 1))+
  labs(title="Boxplot Genre vs Global Sales",x="Genres",y="Sales")
```

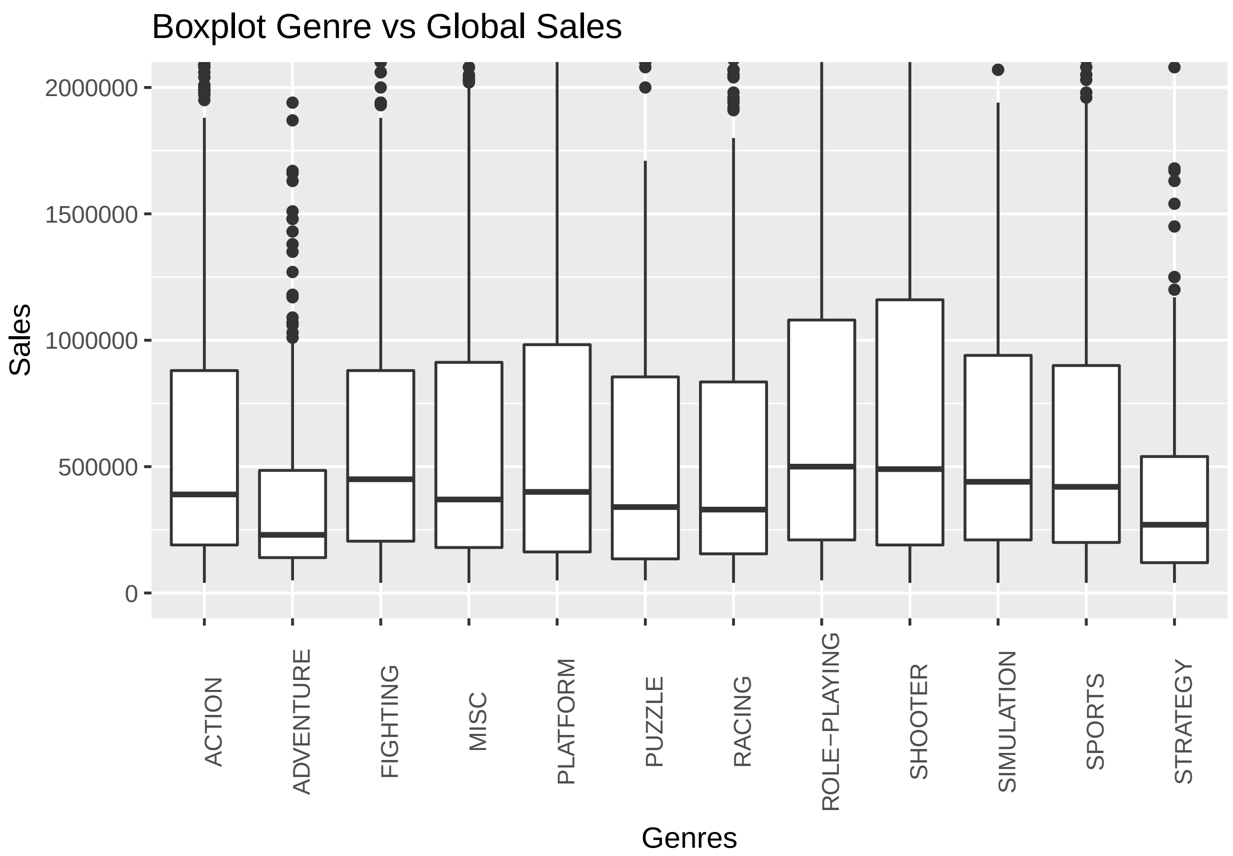


```
##we zoomed graph in order to see boxplots but however we cannot conclude any valuable information from
```

P9)

Zoom the previous plot (Numbers on "y" axis (0,2million)) to clearly see the boxplots for each Genre and make comments. (1p)

```
ggplot(video_games,aes(Genre,Global_Sales))+geom_boxplot()+
  coord_cartesian(ylim = c(0,2000000))+theme(axis.text.x =element_text(angle =90,vjust = 1))+
  labs(title="Boxplot Genre vs Global Sales",x="Genres",y="Sales")
```



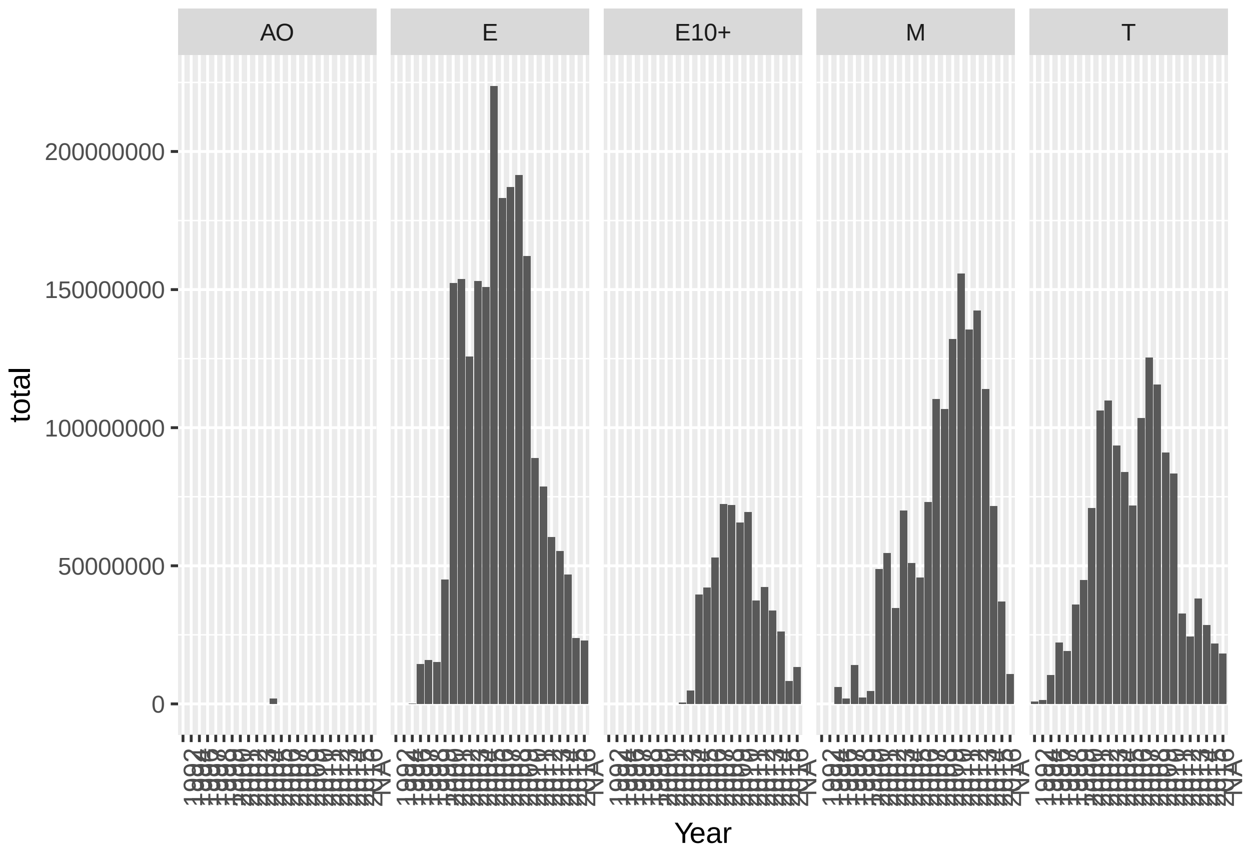
```
##we see that most of sales for each movie is higher than its median and almost all medians are equal
```

```
#P10)
Create a barplot using dyplr functionalities and faceting to show the total Global Sales for each year

-----------------------------------------------------------------------
```

```r
library(dplyr)

video_games1<-video_games%>% select("Year","Rating","Global_Sales")%>%
  group_by(Year,Rating)%>%
  summarise(total=sum(Global_Sales))
video_games1<-as.data.frame(video_games1)
video_games1$Year<-factor(video_games1$Year)
ggplot(video_games1,aes(Year,total))+geom_bar(stat = "identity")+
  theme(axis.text.x =element_text(angle =90,vjust = 1,size = 10))+
  facet_grid(.~Rating)
```



```
#P11)
Use the pipe operator and functions from dplyr package and show the number of video games in each genre
```

```
video_games%>%
  group_by(Genre)%>%
  summarise(Count=n())%>%
  arrange(desc(Count))
```

```
## # A tibble: 12 x 2
##    Genre        Count
##    <chr>        <int>
##  1 ACTION        1508
##  2 SPORTS         921
##  3 SHOOTER        714
##  4 RACING         571
##  5 MISC           484
##  6 ROLE-PLAYING   471
##  7 PLATFORM       406
##  8 FIGHTING       315
##  9 SIMULATION     275
## 10 ADVENTURE      224
## 11 STRATEGY       149
## 12 PUZZLE         107
```

P12)

Use dplyr to create a new variable (CU_Score) in Video dataset which for each video game will show the average of Critic score and 10* User Score. (2p)

```r
CU_Score<-video_games%>%select("Name","Critic_Score","User_Score")%>%
  group_by(Name)%>%
  mutate(AVG=rowMeans(data.frame(Critic_Score,User_Score*10)))
```

P13)

Use the pipe operator and functions from dplyr package to find the top 3 platforms and the number of video games developed for each of them. (2p)

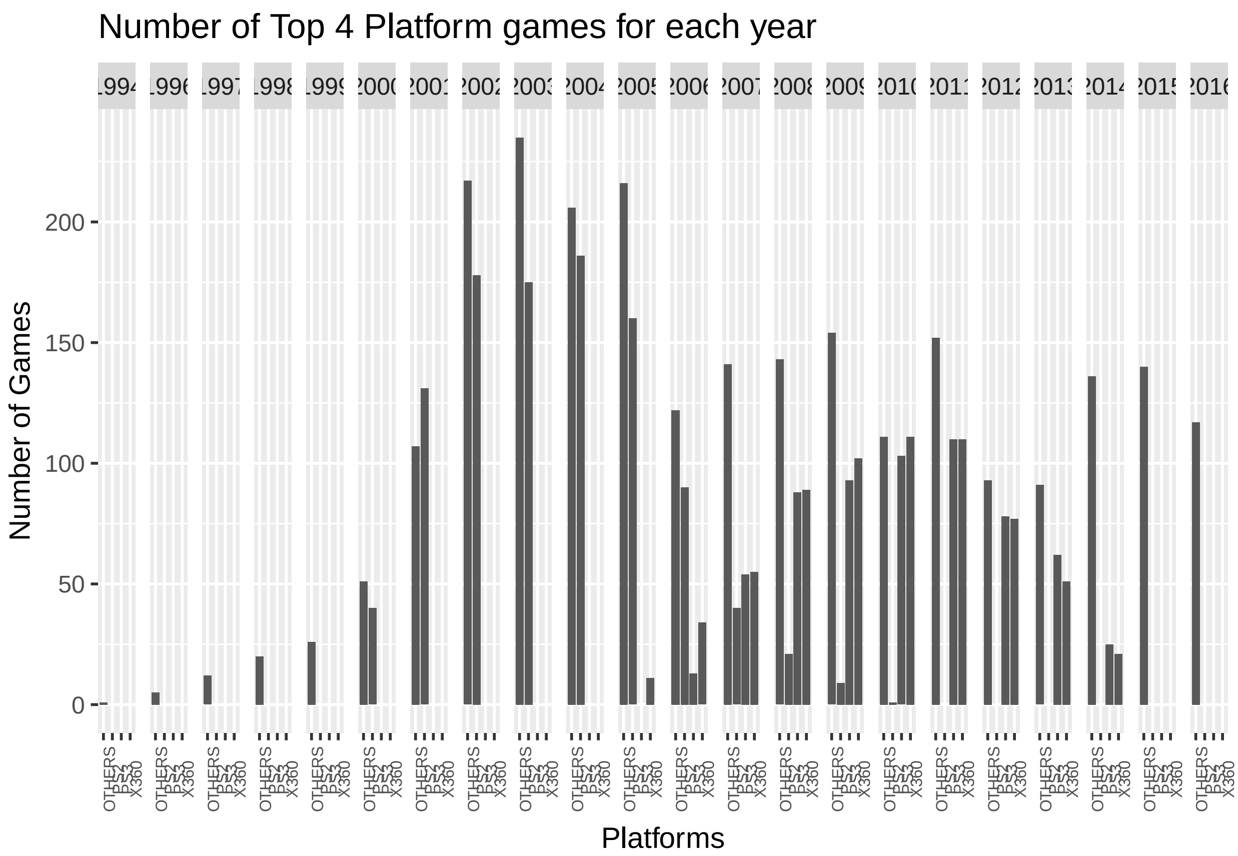
```r
Top3<-video_games%>%
  group_by(Platform)%>%
  summarise(Count=n())%>%
  arrange(desc(Count))%>%top_n(3)
```

```
## Selecting by Count
```

P14)

We are interested in the number of video games developed for top platforms for different years. Pick the top 3 platforms from previous problem and make other platforms as "Other" using dplyr (Hint! ifelse statement). Thereafter remove observations from dataframe which have NA values (Hint! ?complete.cases).Now use faceting to draw the distribution of games for each year for each platform. Make text on "x" axis vertical and size=6. Make comments how the number of video games changed for each platform for different years.(4p)

```r
video_games$Platform<-ifelse(video_games$Platform %in% Top3$Platform,video_games$Platform,"OTHERS")
video_games<-video_games[complete.cases(video_games),]
ggplot(video_games,aes(x=Platform))+geom_bar()+facet_grid(.~Year)+
  theme(axis.text.x = element_text(angle = 90,vjust = 1,size = 6))+
  labs(title="Number of Top 4 Platform games for each year",x="Platforms",y="Number of Games")
```



```
##we can esily say that Year 200o was breakthrough for game industry because the PlayStation 2
#was released in 2000 since then it reached
#popularity,so there were released a lot of games for this platform.X360(The second version of XBOX)
#was officially unveiled on MTV on May 12, 2005, with detailed launch and game information
#announced later that month at the 2005 E3 expo. So it has become worthy competitor for playstation.
```