Data Source

The data explored was taken from Kaggle: Video Game Sales with Ratings.

Reading the Data

```
#Importing libraries.
library(tidyverse)
library(tidytext)
library(dplyr)
library(data.table)
library(ggplot2)
library(knitr)
library(stringr)

#Loading the dataset from local storage.
vgsale <- read_csv("C:/Users/achu3/Documents/datasets/Video_Games_Sales_as_at_22_Dec_2016.csv/Video_Game
```

Data Cleaning & Manipulation

```
str(vgsale)
```

```
## spec_tbl_df [16,719 x 16] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Name           : chr [1:16719] "Wii Sports" "Super Mario Bros." "Mario Kart Wii" "Wii Sports Reso
##  $ Platform       : chr [1:16719] "Wii" "NES" "Wii" "Wii" ...
##  $ Year_of_Release: chr [1:16719] "2006" "1985" "2008" "2009" ...
##  $ Genre          : chr [1:16719] "Sports" "Platform" "Racing" "Sports" ...
##  $ Publisher      : chr [1:16719] "Nintendo" "Nintendo" "Nintendo" "Nintendo" ...
##  $ NA_Sales       : num [1:16719] 41.4 29.1 15.7 15.6 11.3 ...
##  $ EU_Sales       : num [1:16719] 28.96 3.58 12.76 10.93 8.89 ...
##  $ JP_Sales       : num [1:16719] 3.77 6.81 3.79 3.28 10.22 ...
##  $ Other_Sales    : num [1:16719] 8.45 0.77 3.29 2.95 1 0.58 2.88 2.84 2.24 0.47 ...
##  $ Global_Sales   : num [1:16719] 82.5 40.2 35.5 32.8 31.4 ...
##  $ Critic_Score   : num [1:16719] 76 NA 82 80 NA NA 89 58 87 NA ...
##  $ Critic_Count   : num [1:16719] 51 NA 73 73 NA NA 65 41 80 NA ...
##  $ User_Score     : chr [1:16719] "8" NA "8.3" "8" ...
##  $ User_Count     : num [1:16719] 322 NA 709 192 NA NA 431 129 594 NA ...
##  $ Developer      : chr [1:16719] "Nintendo" NA "Nintendo" "Nintendo" ...
##  $ Rating         : chr [1:16719] "E" NA "E" "E" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Name = col_character(),
##   ..   Platform = col_character(),
##   ..   Year_of_Release = col_character(),
##   ..   Genre = col_character(),
##   ..   Publisher = col_character(),
##   ..   NA_Sales = col_double(),
##   ..   EU_Sales = col_double(),
```

```
##   ..   JP_Sales = col_double(),
##   ..   Other_Sales = col_double(),
##   ..   Global_Sales = col_double(),
##   ..   Critic_Score = col_double(),
##   ..   Critic_Count = col_double(),
##   ..   User_Score = col_character(),
##   ..   User_Count = col_double(),
##   ..   Developer = col_character(),
##   ..   Rating = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

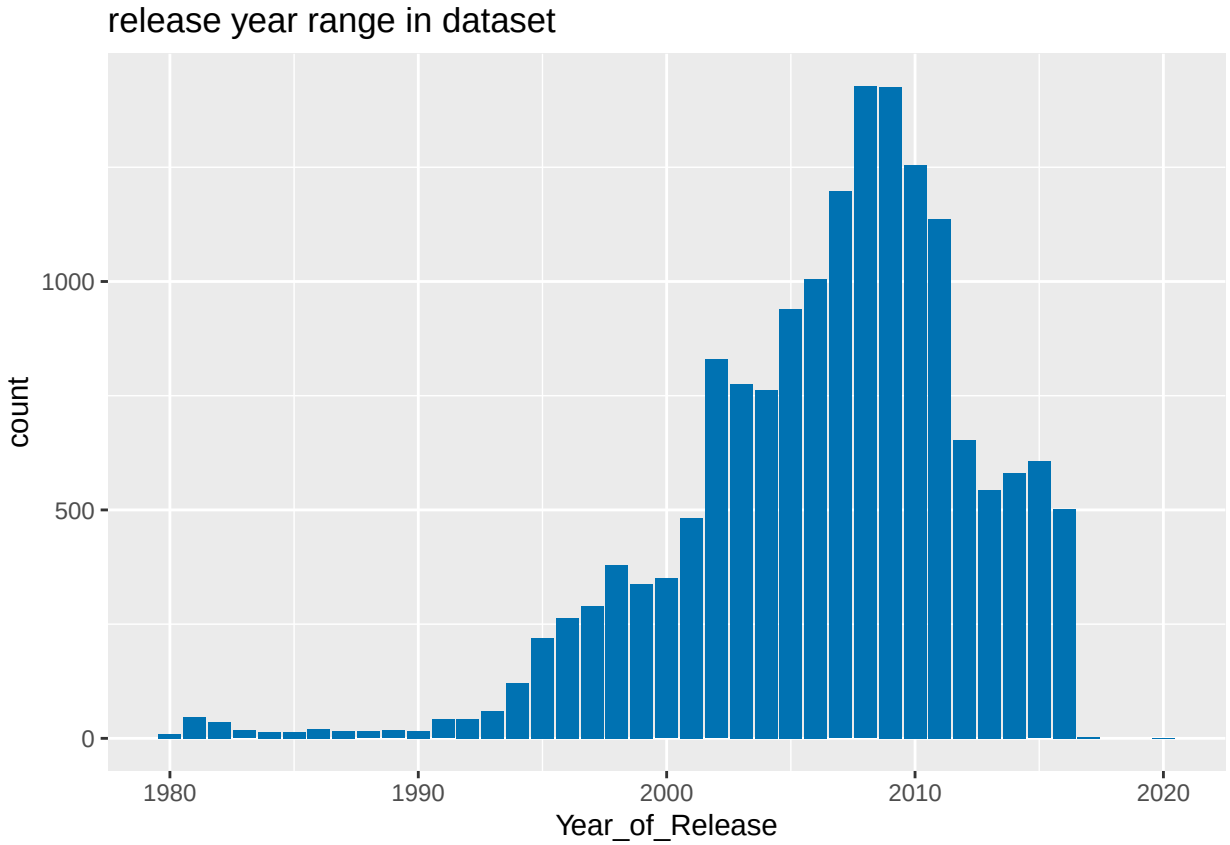
```
summary(vgsale)
```

```
##      Name              Platform         Year_of_Release       Genre
##  Length:16719       Length:16719       Length:16719       Length:16719
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Publisher            NA_Sales          EU_Sales         JP_Sales
##  Length:16719       Min.   : 0.0000   Min.   : 0.000   Min.   : 0.0000
##  Class :character   1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.0000
##  Mode  :character   Median : 0.0800   Median : 0.020   Median : 0.0000
##                     Mean   : 0.2633   Mean   : 0.145   Mean   : 0.0776
##                     3rd Qu.: 0.2400   3rd Qu.: 0.110   3rd Qu.: 0.0400
##                     Max.   :41.3600   Max.   :28.960   Max.   :10.2200
##
##   Other_Sales        Global_Sales      Critic_Score    Critic_Count
##  Min.   : 0.00000   Min.   : 0.0100   Min.   :13.00   Min.   :  3.00
##  1st Qu.: 0.00000   1st Qu.: 0.0600   1st Qu.:60.00   1st Qu.: 12.00
##  Median : 0.01000   Median : 0.1700   Median :71.00   Median : 21.00
##  Mean   : 0.04733   Mean   : 0.5335   Mean   :68.97   Mean   : 26.36
##  3rd Qu.: 0.03000   3rd Qu.: 0.4700   3rd Qu.:79.00   3rd Qu.: 36.00
##  Max.   :10.57000   Max.   :82.5300   Max.   :98.00   Max.   :113.00
##                                       NA's   :8582    NA's   :8582
##   User_Score          User_Count       Developer            Rating
##  Length:16719       Min.   :    4.0   Length:16719       Length:16719
##  Class :character   1st Qu.:   10.0   Class :character   Class :character
##  Mode  :character   Median :   24.0   Mode  :character   Mode  :character
##                     Mean   :  162.2
##                     3rd Qu.:   81.0
##                     Max.   :10665.0
##                     NA's   :9129
```

```
#Missing Value Inspection & Filter
vgsale <- vgsale%>%filter(!is.na(Name)) #only clearing character NAs

#Transforming user/critic scores, release years, and similar to numeric format for use.
vgsale1 <- transform(vgsale,User_Score=as.numeric(User_Score),Year_of_Release=as.numeric(Year_of_Releas
```

```r
vgsale1%>%ggplot(aes(x=Year_of_Release))+
  geom_histogram(stat="count",fill="#0072B2")+ #I like #0072b2 as a colorblind-friendly option.
  labs(title="release year range in dataset")
```



This dataset includes data from Metacritic, such as critic and user scores, which was launched in 1999. Since we will be using its metrics to explore publishers' performance, we will filter out games from 1999 to the most recent year covered in the dataset (2016).

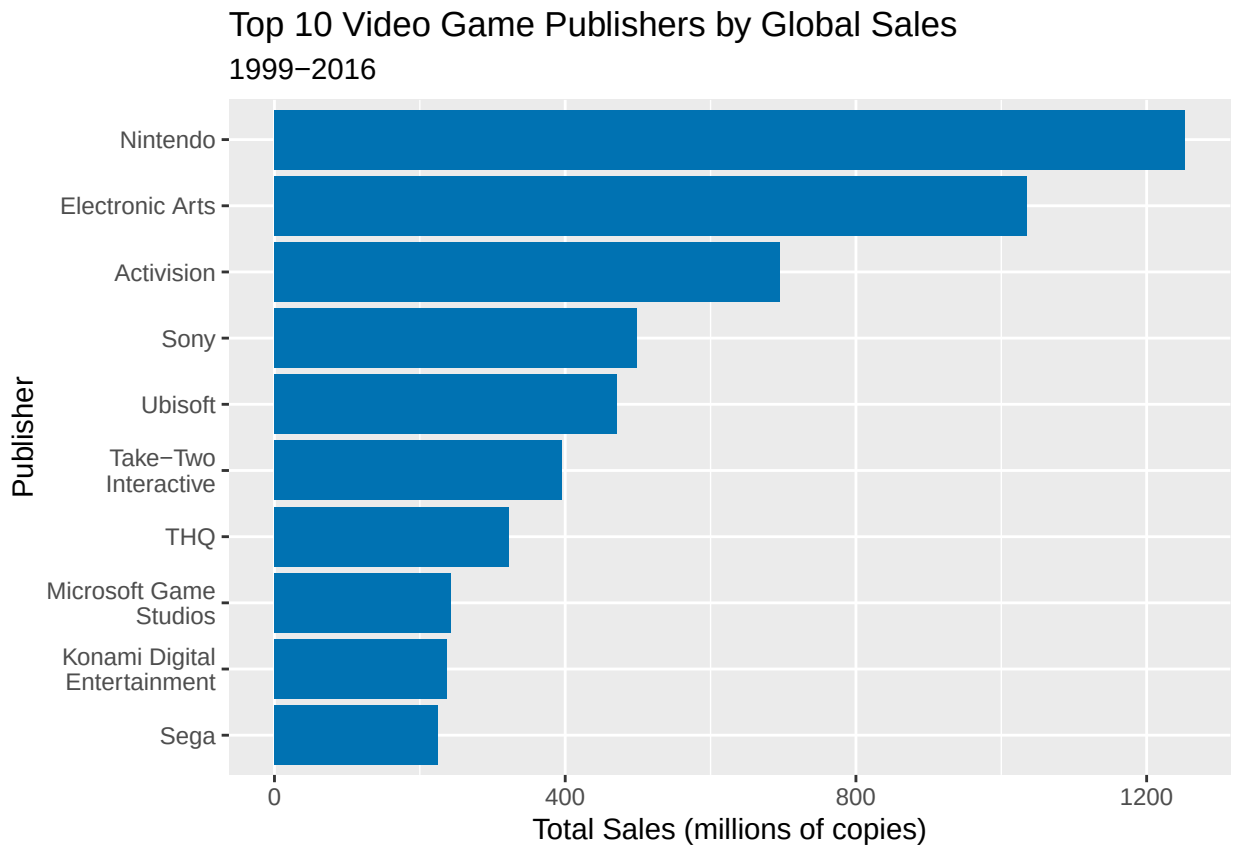
```r
vgsale2 <- vgsale1%>%filter(Year_of_Release>=1999) #I prefer creating new datasets for bigger manipulat
```

```r
#Publisher subsidiaries are merged & renamed with parent companies.
levels(factor(vgsale2$Publisher)) #ls unique names;brief googling if unsure about similar names
#Activision*;Codemasters & Codemasters Online appear to be different;EA Games, Electronic Arts*;Enix Co
#Unknown = Unknown Worlds? check game of obsv
vgsale2$Publisher <- sub("^Activision.*","Activision",vgsale2$Publisher)
vgsale2$Publisher <- sub("^Electronic Arts.*","Electronic Arts",vgsale2$Publisher)
vgsale2$Publisher <- sub("EA Games","Electronic Arts",vgsale2$Publisher)
vgsale2$Publisher <- sub("^Square.*","Square Enix",vgsale2$Publisher)
vgsale2$Publisher <- sub("^Enix.*","Square Enix",vgsale2$Publisher)
vgsale2$Publisher <- sub("^Idea Factory.*","Idea Factory",vgsale2$Publisher)
vgsale2$Publisher <- sub("^Marvelous.*","Marvelous Inc",vgsale2$Publisher)
vgsale2$Publisher <- sub("^Rebellion.*","Rebellion",vgsale2$Publisher)
vgsale2$Publisher <- sub("^Sony.*","Sony",vgsale2$Publisher)
vgsale2$Publisher <- sub("^Ubisoft.*","Ubisoft",vgsale2$Publisher)
vgsale2$Publisher <- sub("^Zoo.*","Zoo Games",vgsale2$Publisher)
vgsale2$Publisher <- sub("^Zushi Games.*","Zoo Games",vgsale2$Publisher)
```

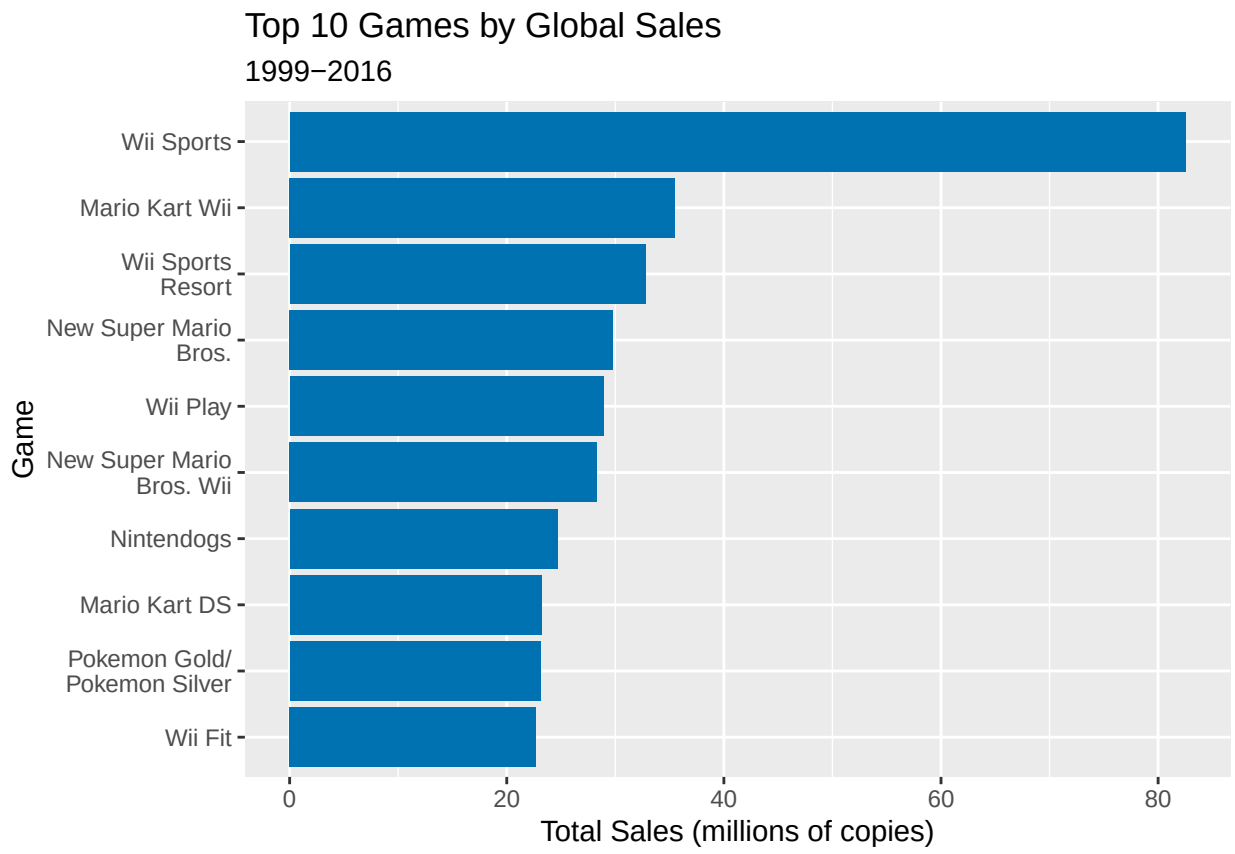
```
#Sort multiples (ex. Activision - CoD:BOII) by platform.
vgsale2$Name_Platform <- paste(vgsale2$Name,vgsale2$Platform,sep=", ")
```

Exploring - publisher focus.

```
#What are the top ten publishing companies? by global sale
pubsales <- vgsale2%>%group_by(Publisher)%>%
  summarise(pubttlsale=sum(Global_Sales))%>%
  arrange(desc(pubttlsale))%>%
  top_n(10)%>%
  mutate(wrapname=str_wrap(Publisher,width=15))
pubsales%>%ggplot(aes(x=reorder(wrapname,pubttlsale),y=pubttlsale))+
  geom_bar(stat="identity",fill="#0072B2")+
  coord_flip()+
  labs(x="Publisher",y="Total Sales (millions of copies)",
       title="Top 10 Video Game Publishers by Global Sales",subtitle="1999-2016")
```



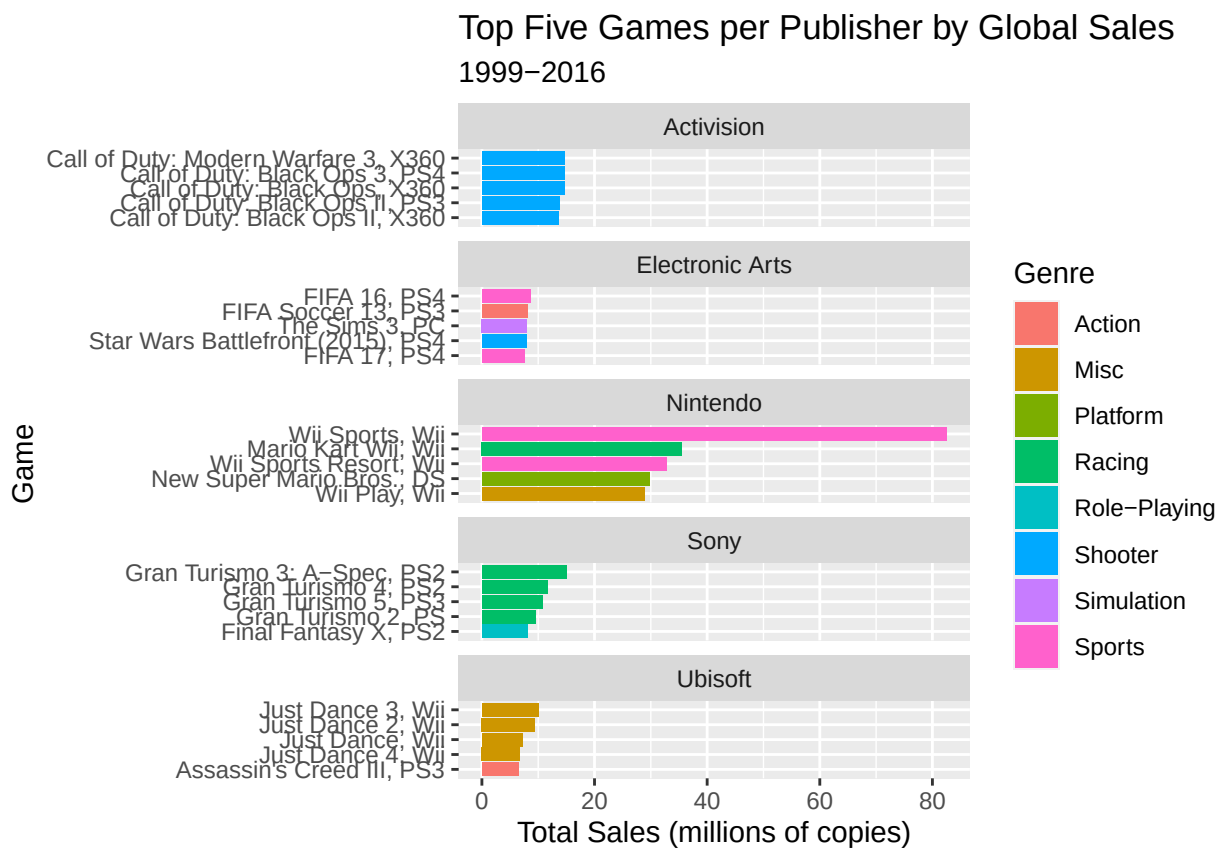

```
#What are the top ten games by publishing companies? by global sale
topgamesale <- vgsale2%>%top_n(10,Global_Sales)%>%mutate(wrapname=str_wrap(Name,width=15))
topgamesale%>%ggplot(aes(x=reorder(wrapname,Global_Sales),y=Global_Sales))+
  geom_bar(stat="identity",fill="#0072B2",position="dodge")+
  coord_flip()+
  labs(x="Game",y="Total Sales (millions of copies)",
       title="Top 10 Games by Global Sales",subtitle="1999-2016")
```



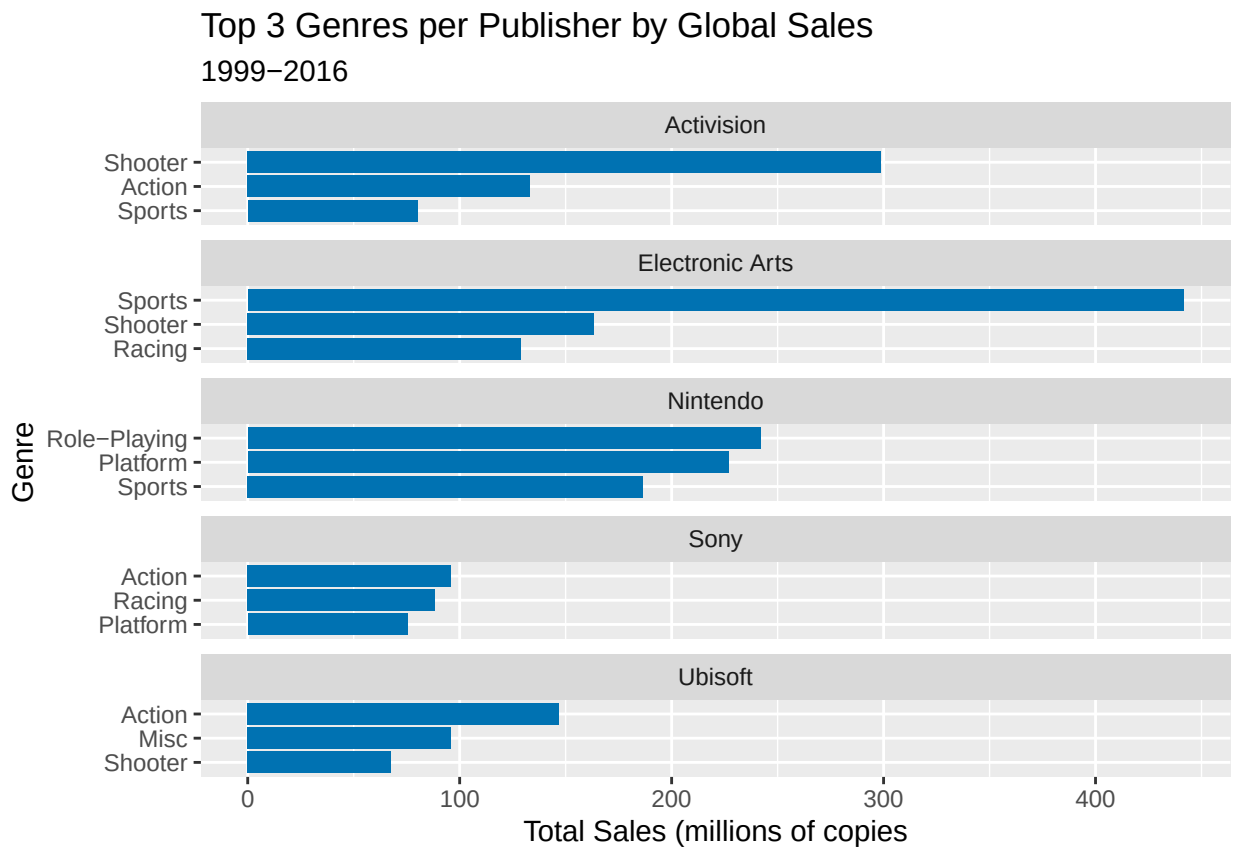

```
#Nintendo dominates in sales.
#Why is that?
```

```
#What genres do publishers create best?

#top 5 games(sales)/top 5 publisher (w/Genre) (top 5 to narrow focus)
#This gives us a brief look as to what type of games do well per publisher.
#ex. Activation is pure shooter, while Nintendo has the most variety.
pubsales_5 <- vgsale2%>%group_by(Publisher)%>%
  summarise(pubttlsale=sum(Global_Sales))%>%
  arrange(desc(pubttlsale))%>%
  top_n(5) #top five publishers from prev top pub subset
topgamesale2 <- vgsale2%>%group_by(Publisher)%>%filter(Publisher%in%pubsales_5$Publisher)%>%
  select(Name_Platform,Publisher,Genre,Global_Sales)%>%
  arrange(desc(Global_Sales))%>%
  top_n(5) #top five games/publisher by sale
topgamesale2 <- topgamesale2%>%mutate(Publisher=reorder(Publisher,-Global_Sales),#order pub by sales
                                      Name_Platform=reorder_within(
                                        Name_Platform,Global_Sales,Publisher,fun=sum)) #order name by s
topgamesale2%>%ggplot(aes(x=Global_Sales,y=Name_Platform,fill=Genre))+
  geom_col()+ #functions the same as flipped geombar(identity)
  #geom_text(aes(label=Global_Sales),hjust=1)+
  facet_wrap(~Publisher,ncol=1,scales="free_y")+
  scale_y_reordered()+
  labs(x="Total Sales (millions of copies)",y="Game",
       title="Top Five Games per Publisher by Global Sales",subtitle="1999-2016")
```



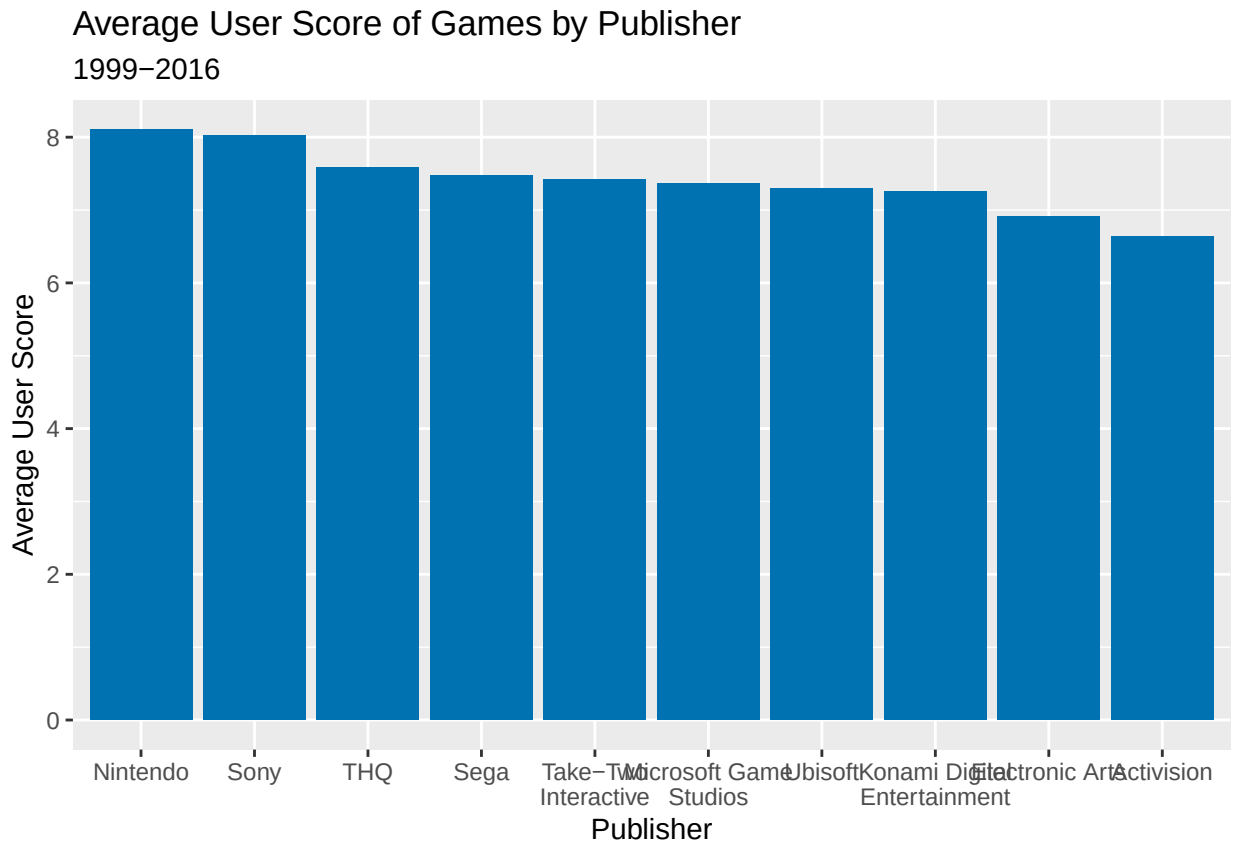

```
#top 3 genres(sales)/top 5 publisher
#A more pure/diverse look than before.
topgenrepub <- vgsale2%>%filter(Publisher%in%pubsales_5$Publisher)%>%
  select(Publisher,Genre,Global_Sales)%>%
  group_by(Genre,Publisher)%>%
  summarise(genrettlsale=sum(Global_Sales))%>%
  arrange(desc(genrettlsale))%>%
  group_by(Publisher)%>%
  top_n(3)%>%
  mutate(Publisher=reorder(Publisher,-genrettlsale),
         Genre=reorder_within(Genre,genrettlsale,Publisher,fun=sum))
topgenrepub%>%ggplot(aes(x=genrettlsale,y=Genre))+
  geom_col(fill="#0072B2")+
  facet_wrap(~Publisher,ncol=1,scales="free_y")+
  scale_y_reordered()+
  labs(x="Total Sales (millions of copies",y="Genre",
       title="Top 3 Genres per Publisher by Global Sales",subtitle="1999-2016")
```



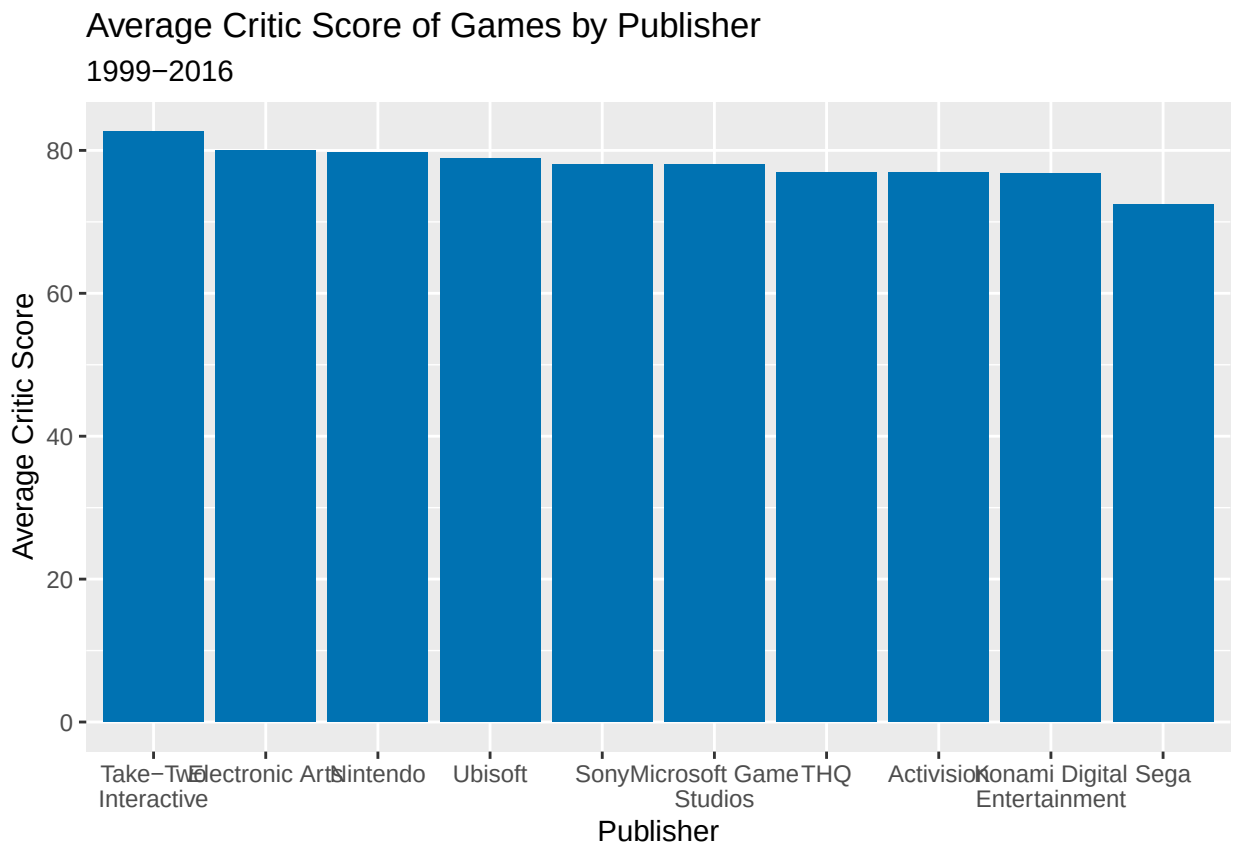

```
#most frequently - sports, shooter, action;
#Nintendo seems to specialize with platforming games; has unique range compared to others
#confirms Activision & shooters; EA dominates sports (in terms of sales)
```

```
####look into the percentage sales per genre per publisher

#Do these sales match with score?
#average user score of games/publisher
topuserscorepub <- vgsale2%>%filter(Publisher%in%pubsales$Publisher)%>%filter(User_Count>=40)%>%
  select(Publisher,User_Score)%>%
  group_by(Publisher)%>%
  summarise(avguserscore=mean(User_Score))%>%
  mutate(wrappub=str_wrap(Publisher,width=15))
topuserscorepub%>%ggplot(aes(x=reorder(wrappub,-avguserscore),y=avguserscore))+
  geom_bar(stat="identity",fill="#0072B2")+
  labs(x="Publisher",y="Average User Score",title="Average User Score of Games by Publisher",subtitle="
```




```
#average critic score of games/publisher
topcriticscorepub <- vgsale2%>%filter(Publisher%in%pubsales$Publisher)%>%filter(Critic_Count>=40)%>%
  group_by(Publisher)%>%
  summarise(avgcriticscore=mean(Critic_Score))%>%
  mutate(wrappub=str_wrap(Publisher,width=15))
topcriticscorepub%>%ggplot(aes(x=reorder(wrappub,-avgcriticscore),y=avgcriticscore))+
  geom_bar(stat="identity",fill="#0072B2")+
  labs(x="Publisher",y="Average Critic Score",title="Average Critic Score of Games by Publisher",subtit
```

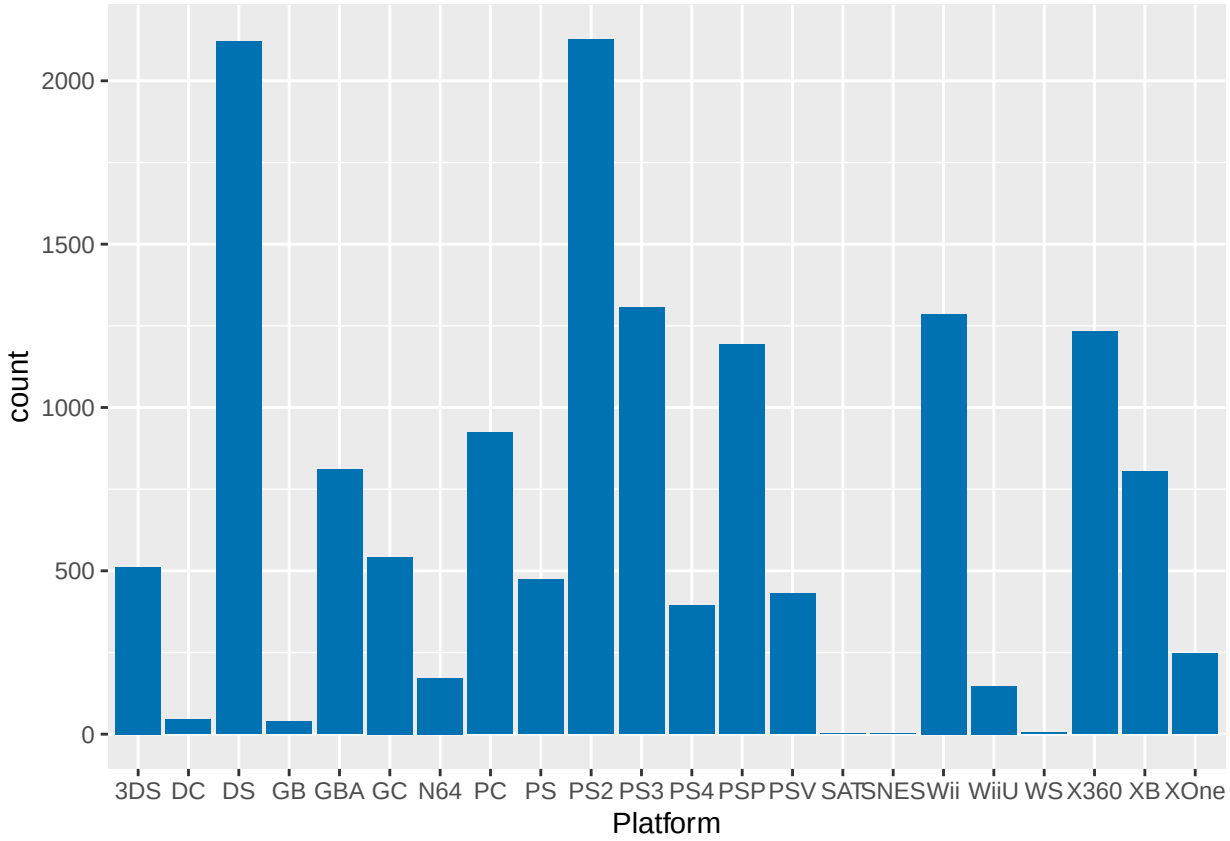



```
#Not much to see in terms of publisher score - scores are fairly close together.
```

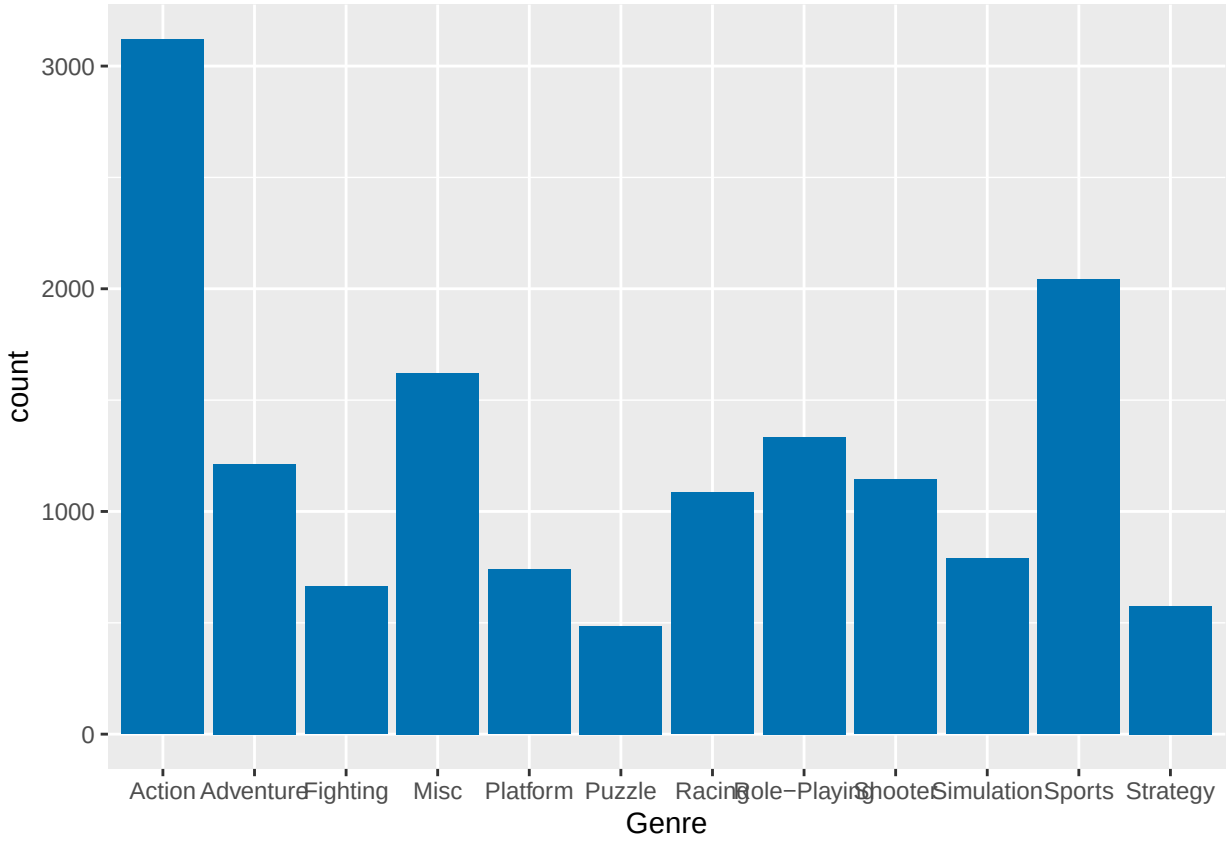
```
##Platform
#most games on a platform
levels(factor(vgsale2$Platform))
```

```
##  [1] "3DS"  "DC"   "DS"   "GB"   "GBA"  "GC"   "N64"  "PC"   "PS"   "PS2"
## [11] "PS3"  "PS4"  "PSP"  "PSV"  "SAT"  "SNES" "Wii"  "WiiU" "WS"   "X360"
## [21] "XB"   "XOne"
```

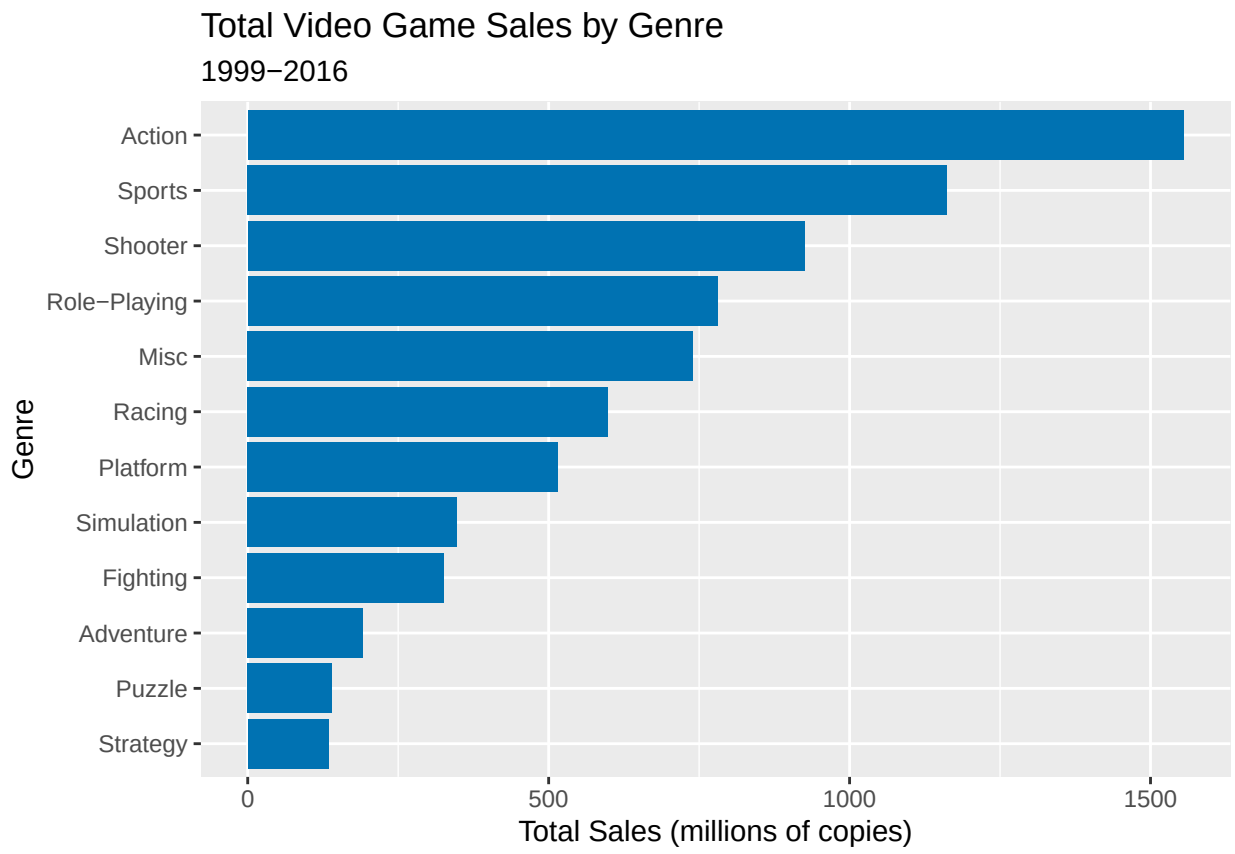
```
vgsale2%>%ggplot(aes(x=Platform))+
  geom_histogram(stat="count",fill="#0072B2")
```




```
##Genre
vgsale2%>%ggplot(aes(x=Genre))+
  geom_histogram(stat="count",fill="#0072B2") #game/genre
```

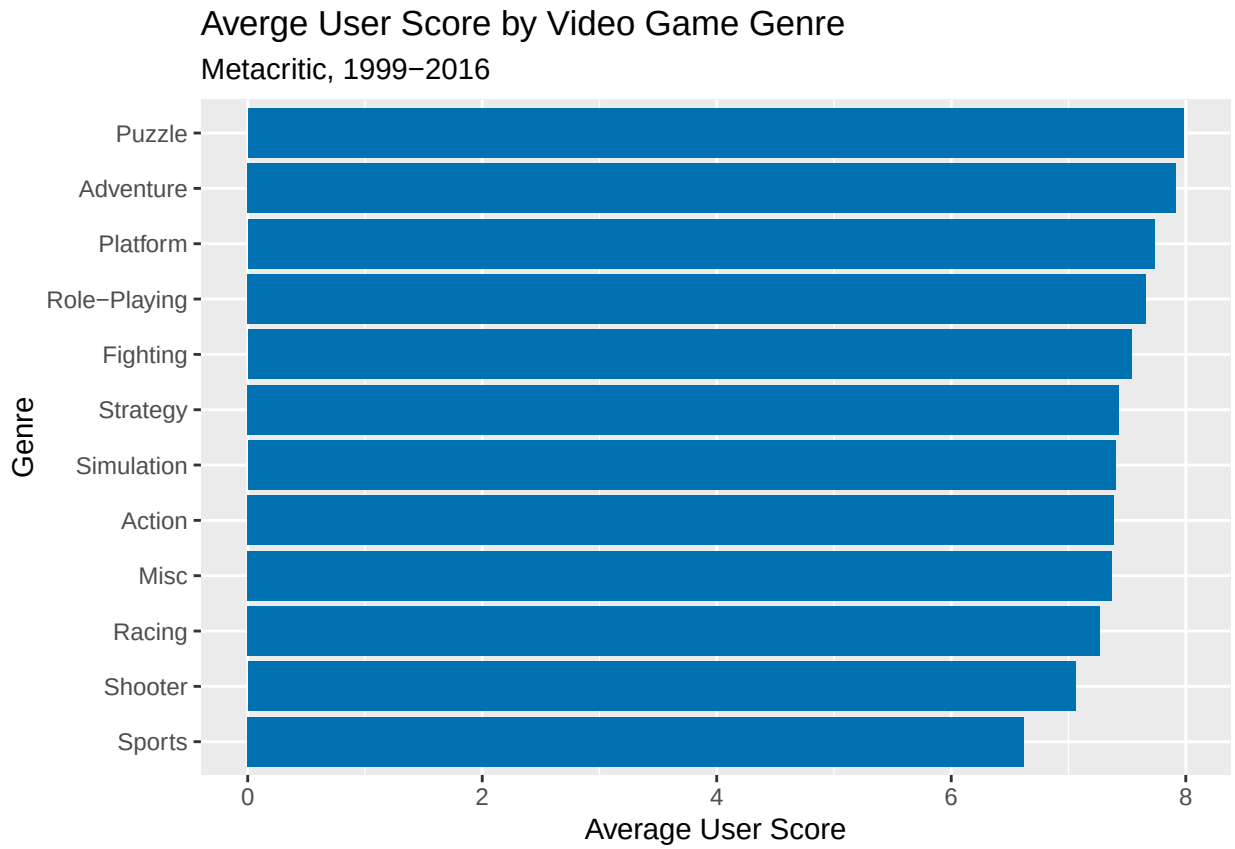



```
genresale <- vgsale2%>%group_by(Genre)%>%
  summarise(genrettlsale=sum(Global_Sales))%>%
  arrange(desc(genrettlsale))
genresale%>%ggplot(aes(x=reorder(Genre,genrettlsale),y=genrettlsale))+
  geom_bar(stat="identity",fill="#0072B2")+
  coord_flip()+
  labs(x="Genre",y="Total Sales (millions of copies)",title="Total Video Game Sales by Genre",subtitle=
```

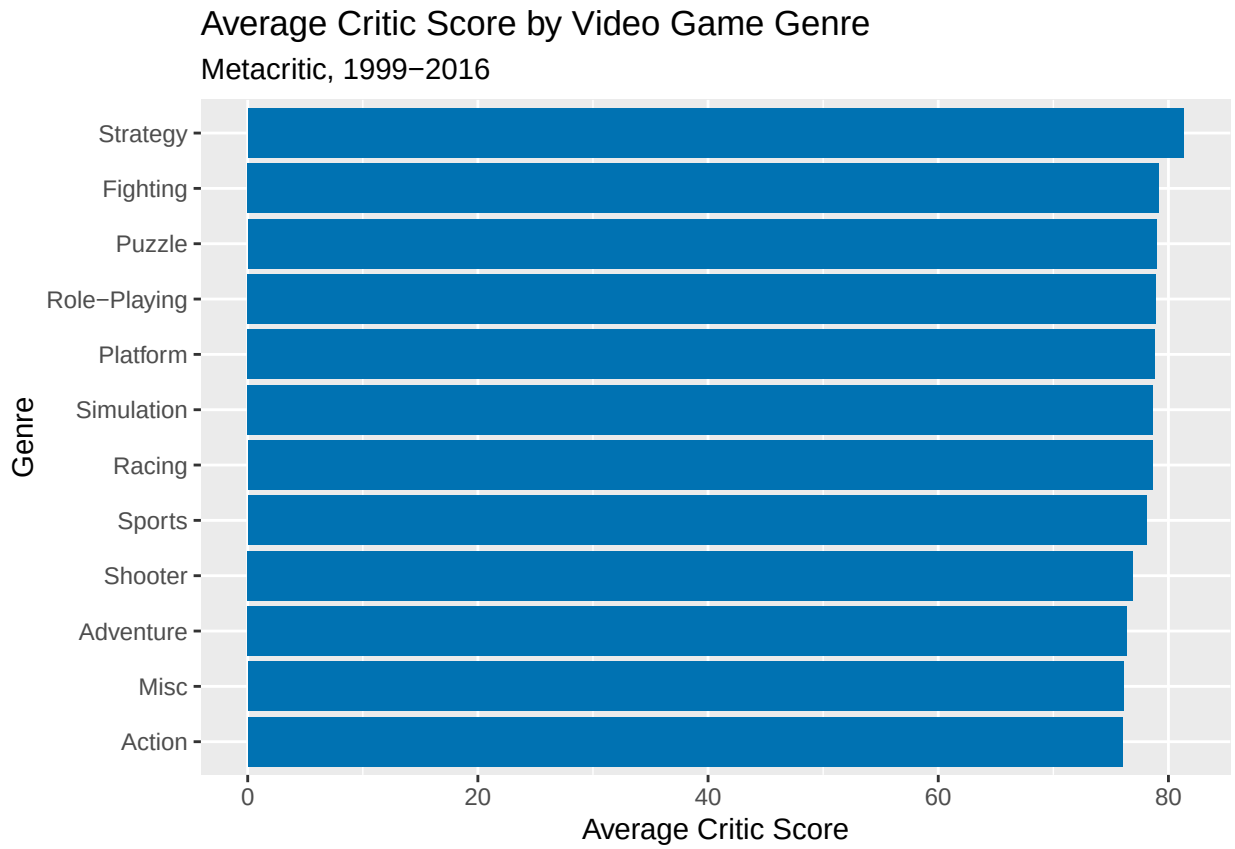



```
#but is action, sports rated highly?
#who buys these? what region? region&score genre
```

```
genreuser <- vgsale2%>%filter(User_Count>=40)%>%
  group_by(Genre)%>%
  summarise(avgscore_user=mean(User_Score))%>%
  arrange(desc(avgscore_user))
genreuser%>%ggplot(aes(x=reorder(Genre,avgscore_user),y=avgscore_user))+
  geom_bar(stat="identity",fill="#0072B2")+
  coord_flip()+
  labs(x="Genre",y="Average User Score",title="Averge User Score by Video Game Genre",subtitle="Metacri
```

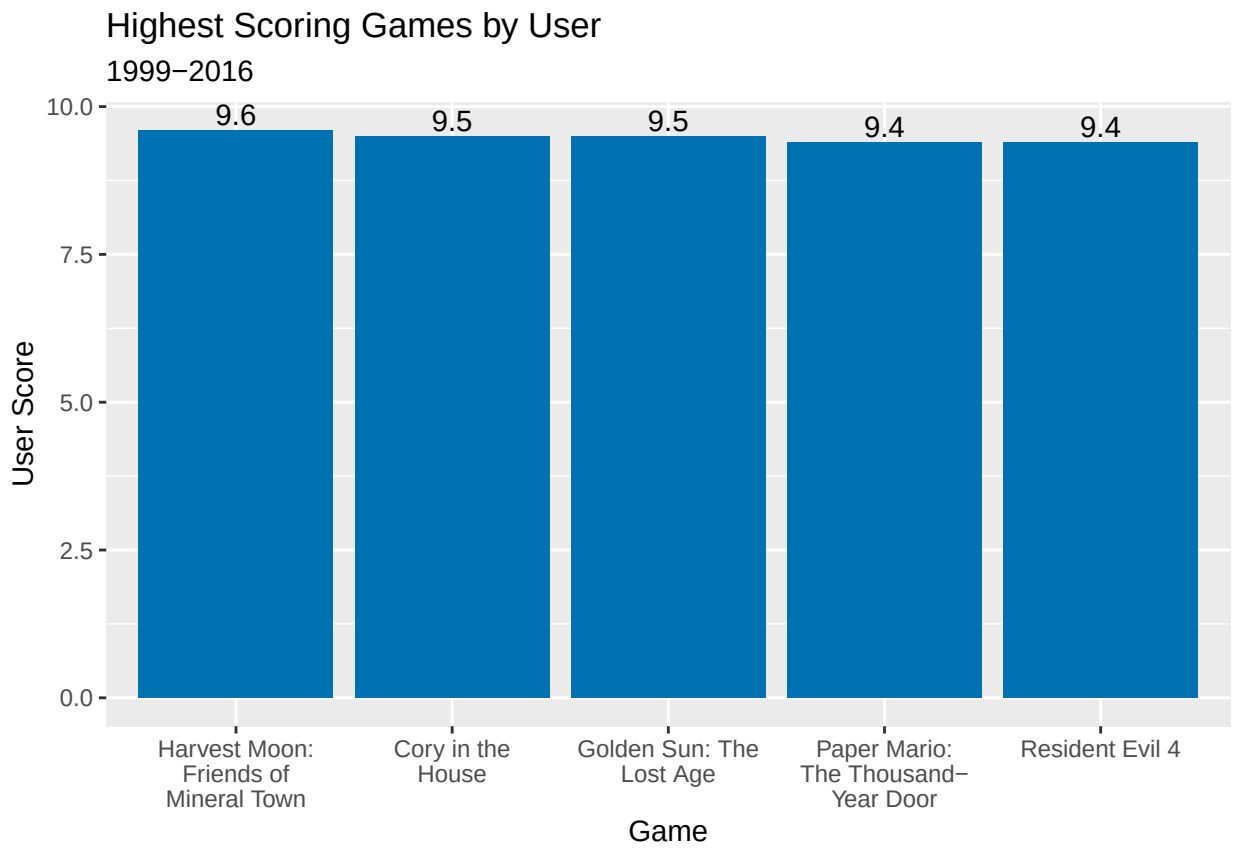



```
criticuser <- vgsale2%>%filter(Critic_Count>=40)%>%
  group_by(Genre)%>%
  summarise(avgscore_critic=mean(Critic_Score))%>%
  arrange(desc(avgscore_critic))
criticuser%>%ggplot(aes(x=reorder(Genre,avgscore_critic),y=avgscore_critic))+
  geom_bar(stat="identity",fill="#0072B2")+
  coord_flip()+
  labs(x="Genre",y="Average Critic Score",title="Average Critic Score by Video Game Genre",subtitle="Me
```

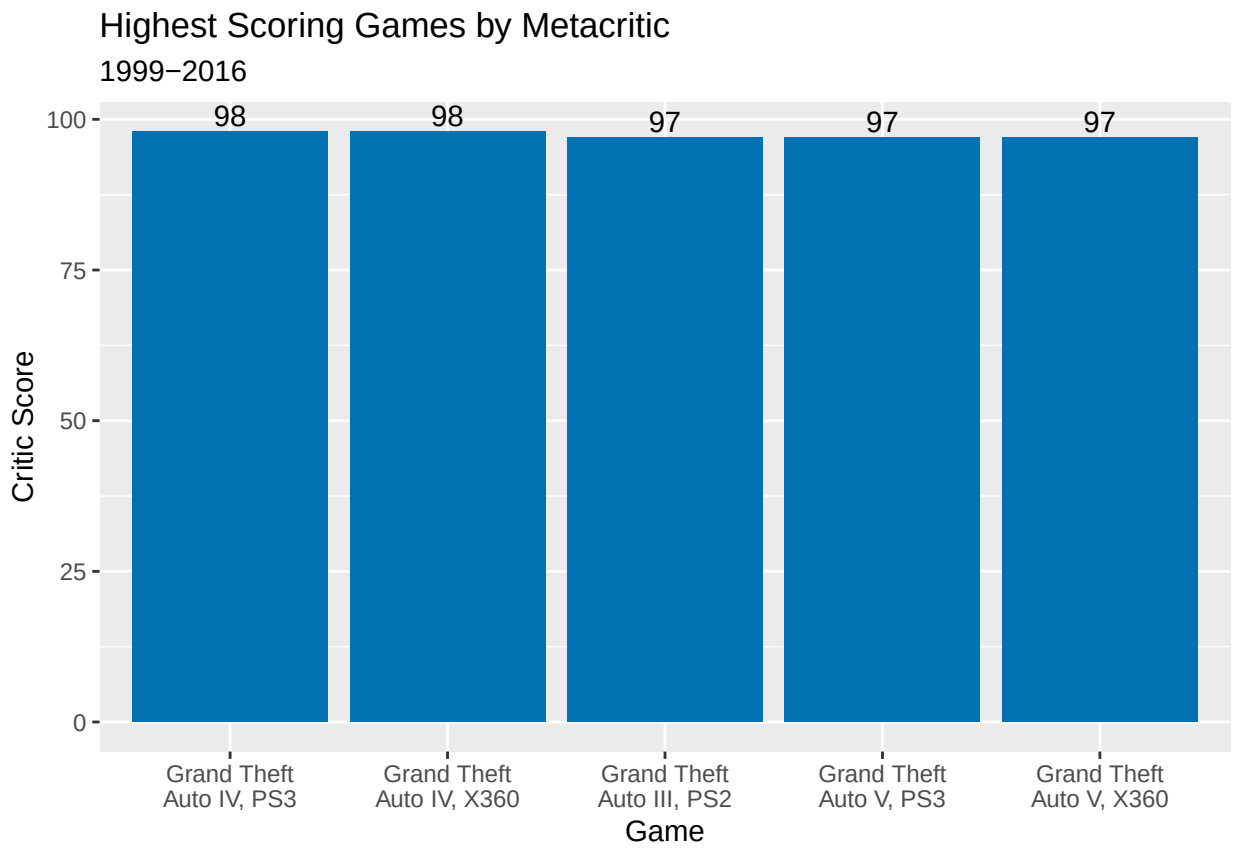



```
#seemingly not too much difference w/in groups, although top differs
```

```
#Highest Rated Games (User Score)
hiscore_user <- vgsale2%>%filter(User_Count>=40)%>%
  select(Name,User_Score)%>%
  arrange(desc(User_Score))%>%
  #top_n(5)
  slice(1:5)
hiscore_user$wrapname <- str_wrap(hiscore_user$Name,width=15)
hiscore_user%>%ggplot(aes(reorder(wrapname,-User_Score),y=User_Score))+
  geom_bar(stat="identity",fill="#0072B2")+
  geom_text(aes(label=User_Score),vjust=-.25)+
  #scale_x_discrete(guide=guide_axis(n.dodge=3))+
  #scale_x_discrete(labels=abbreviate)+
  labs(x="Game",y="User Score",title="Highest Scoring Games by User",subtitle="1999-2016")
```




```
#Highest Rated Games (Critic)
hiscore_critic <- vgsale2%>%filter(Critic_Count>=40)%>%
  select(Name_Platform,Critic_Score)%>%
  arrange(desc(Critic_Score))%>%
  slice(1:5)
hiscore_critic$wrapname <- str_wrap(hiscore_critic$Name_Platform,width=15)
hiscore_critic%>%ggplot(aes(reorder(wrapname,-Critic_Score),y=Critic_Score))+
  geom_bar(stat="identity",fill="#0072B2")+
  geom_text(aes(label=Critic_Score),vjust=-.25)+
  #scale_x_discrete(guide=guide_axis(n.dodge=3))+
  #scale_x_discrete(labels=abbreviate)+
  labs(x="Game",y="Critic Score",title="Highest Scoring Games by Metacritic",subtitle="1999-2016")
```




```
#GTA on multiple platforms...maybe instead of combining names, average the
#scores for multiples?
```