Problem Statement and Background

Give a clear and complete statement of the problem. Don't describe methods or tools yet. Where does the data come from, what are its characteristics? Include background material as appropriate: who cares about this problem, what impact it has? Include a brief summary of any related work you know about.

My dataset is a video game dataset from Kaggle, which describes information on individual video games (Name, Platform, Release Year, Genre, Publisher, Rating), their physical sales from Vgchartz (regionally and globally) and their scores from users and critics on Metacritic, covering video games released between 1980 and 2016.The problem I am trying to solve is finding interesting statistics and relationships bewtween the attributes from my dataset, which revolves around the topic of video games, and visualizing them with various graphs. I am also trying to cluster (group) similar video games and see the differences between each cluster. Third, I am trying to find a way to predict whether a video game's Global sales is Low or High. The problem I am trying to solve is important because by finding interesting statistics and relationships bewtween the attributes of my video game dataset and visualizing them with various graphs, the findings and results can be used to analyze the video game industry since its beginning. I have no knowledghe of any other related work in relation to my project.

Methods

Describe the tools and methods that you used and the reasons for their choice. Justify them in terms of the problem statement. This section will include data wrangling and visualization, machine learning. Please explain your code and discuss all methods you used.

The tools and methods I used for my project include descriptive statistics, visualization, clustering, and decision tree analysis. More specifically, I used the dplyr, ggplot2, kmeans, rpart, and rpart.plot libraries. I used descriptive statistics (dplyr) for to perform data wrangling and find interesting summary statistics bewtween the attributes of my dataset, displaying them using tables. I used visualizaion (ggplot2) to visualize the interesting summary statistics between the attributes of my dataset using graphs, plots, etc. I used clustering (kmeans) to cluster (group) similar video games. I used Decision Tree Analysis (rpart, rpart.plot) to predict whether a video game's Global sales is Low or High by building a classification model.

Data Processing/Cleaning

In [1]:
```
library(dplyr)
library(ggplot2)
library(rpart)
library(rpart.plot)
```

```
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
```

In [2]:
```
vg<-read.csv("Video_Games_Sales_as_at_22_Dec_2016.csv")
head(vg)
```

Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_S
Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81
Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79
Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28
Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22
Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22

Changing data type of columns

When initially looking at the structure of the dataset, I found that the User_Score column was a character field, which shouldn't be the case. Below, I changed the User_Score column to a numeric field, an appropriate data type for the column.

In [3]:
```
vg$User_Score=as.numeric(vg$User_Score)
user_score<-ifelse(vg$User_Score!=1,vg$User_Score,NA)
vg$User_Score<-user_score
user_score<-ifelse(vg$User_Score!=97,vg$User_Score,NA)
vg$User_Score<-user_score
head(vg)
```

Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_S
Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81
Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79
Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28
Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22
Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22

From analyzing the dataset, I found there to be many missing values, especially in the Rating, Critic_Score, Critic_Count, User_Score, User_Count, Developer, and Rating columns of the dataset. There are many missing values in these columns because the data for these columns were brought from Metacritic. Metacritic doesn't cover all the video games and video game platforms in this dataset, leading to there being many missing values in the above columns. For the video games with missing values in these columns, I will leave and use them as is. Removing the video games with missing values in these columns would remove a large chunk of my dataset, leading to less meaningful and accurate results. I will work around the video games with missing values in these columns when needed.

Data Wrangling and Visualization

Top 10 Best-Selling games by Region

The table below summarizes the top 10 best-selling games globally, in North America, in Europe, in Japan, and in other regions. Wii Sports, Grand Theft Auto 5, Super Mario Bros, Tetris, and Mario Kart Wii are the best-selling games globally. Wii Sports, Super Mairo Bros, Cuck Hunt, Tetris, and Grand Theft Auto 5 are the best-selling games in North America. Wii Sports, Grand Theft Auto 5, Mario Kart Wii, FIFA 15, and Call of Duty: Modern Warfare 3 are the best-selling games in Europe. Pokemon Red/Blue, Pokemon Gold/Silver, Super Mario Bros, New Super Mario Bros, and Pokemon Diamond/Pearl are the best-selling games in Japan. Grand Theft Auto: San Andreas, Wii Sports, Grand Theft Auto 5, Gran Turismo 4, and Call of Duty: Black Ops 2 are the best-selling games in other regions. Pokemon and Super Mario games are very popular in Japan.

In [4]:
```
sales_global<-vg %>% group_by(Name) %>% summarize(gsales=sum(Global_Sales))
%>% arrange(desc(gsales))
t10_global<-sales_global[1:10,]
sales_NA<-vg %>% group_by(Name) %>% summarize(nsales=sum(NA_Sales)) %>% arra
nge(desc(nsales))
t10_NA<-sales_NA[1:10,]
sales_EU<-vg %>% group_by(Name) %>% summarize(esales=sum(EU_Sales)) %>% arra
nge(desc(esales))
t10_EU<-sales_EU[1:10,]
sales_JP<-vg %>% group_by(Name) %>% summarize(jsales=sum(JP_Sales)) %>% arra
nge(desc(jsales))
t10_JP<-sales_JP[1:10,]
sales_other<-vg %>% group_by(Name) %>% summarize(osales=sum(Other_Sales)) %
>% arrange(desc(osales))
t10_other<-sales_other[1:10,]
cbind(t10_global,t10_NA,t10_EU,t10_JP,t10_other)
```

Name	gsales	Name	nsales	Name	esales	Name	jsales	Nam
Wii Sports	82.53	Wii Sports	41.36	Wii Sports	28.96	Pokemon Red/Pokemon Blue	10.22	Gran Thef Auto San Andr
Grand Theft Auto V	56.57	Super Mario Bros.	32.48	Grand Theft Auto V	23.42	Pokemon Gold/Pokemon Silver	7.20	Wii Spor
Super Mario Bros.	45.31	Duck Hunt	26.93	Mario Kart Wii	12.76	Super Mario Bros.	6.96	Gran Thef Auto
Tetris	35.84	Tetris	26.17	FIFA 15	12.02	New Super Mario Bros.	6.50	Gran Turis 4
Mario Kart Wii	35.52	Grand Theft Auto V	23.84	Call of Duty: Modern Warfare 3	11.15	Pokemon Diamond/Pokemon Pearl	6.04	Call Duty Blac Ops
Wii Sports Resort	32.77	Call of Duty: Black Ops	17.57	FIFA 16	11.09	Tetris	6.03	FIFA Socc 08
Pokemon Red/Pokemon Blue	31.37	Super Mario World	15.99	FIFA 14	10.96	Pokemon Black/Pokemon White	5.65	Pro Evol Socc 2008
Call of Duty: Black Ops	30.82	Mario Kart Wii	15.68	Nintendogs	10.95	Dragon Quest VII: Warriors of Eden	5.40	Call Duty Blac Ops
Call of Duty: Modern Warfare 3	30.59	Wii Sports Resort	15.61	Wii Sports Resort	10.93	Pokemon Ruby/Pokemon Sapphire	5.38	Call Duty Blac Ops
New Super Mario Bros.	29.80	Call of Duty: Modern Warfare 3	15.54	Call of Duty: Black Ops II	10.84	Animal Crossing: Wild World	5.33	Call Duty Mod Warf 3

Top 10 Video Games with Highest Critic Scores/User Scores

The table below summarizes the top 10 video games with the highest critic and user scores. SoulCalibur, Metroid Prime, NFL 2K1, Super Mario Galaxy, and Grand Theft Auto 5 are the highest rated video games by critics. Boktai:The Sun is in Your Hand, Harvest Moon: Friends of Mineral Town, Cory in the House, Golden Sun: The Lost Age, and Karnaaj Rally are the highest rated games by users. From this, there appears to be a huge difference between the scores critics and users give toward video games.

In [5]:
```
top_cscore<-vg %>% group_by(Name) %>% summarize(cscore=mean(Critic_Score)) %>%
 arrange(desc(cscore))
t10_cscore<-top_cscore[1:10,]
top_uscore<-vg %>% group_by(Name) %>% summarize(uscore=mean(User_Score)) %>% a
rrange(desc(uscore))
t10_uscore<-top_uscore[1:10,]
cbind(t10_cscore,t10_uscore)
```

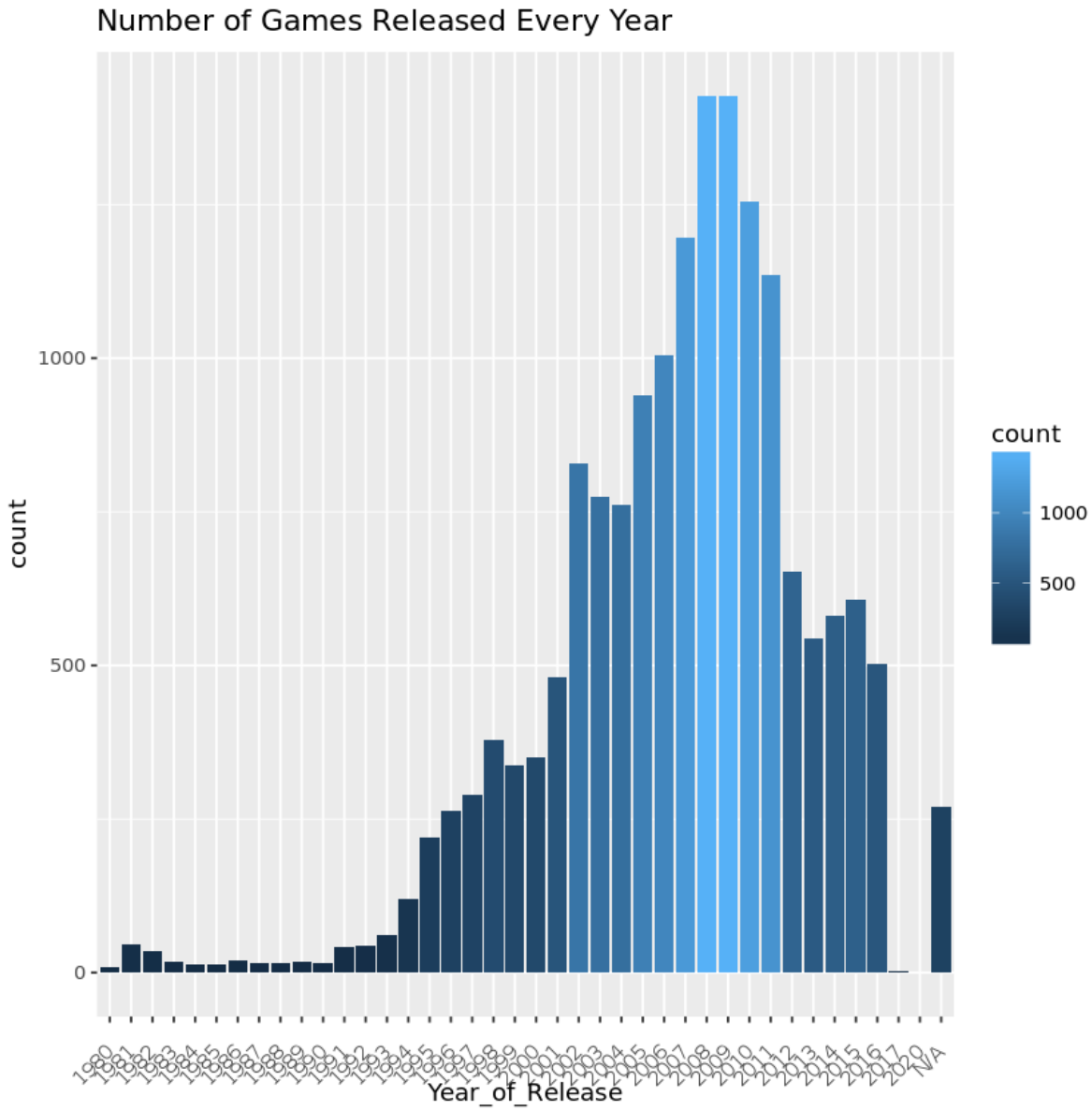
Name	cscore	Name	uscore
SoulCalibur	98.0	Boktai: The Sun is in Your Hand	95
Metroid Prime	97.0	Harvest Moon: Friends of Mineral Town	95
NFL 2K1	97.0	Cory in the House	94
Super Mario Galaxy	97.0	Golden Sun: The Lost Age	94
Super Mario Galaxy 2	97.0	Karnaaj Rally	94
Grand Theft Auto V	96.8	Super Puzzle Fighter II	94
Gran Turismo	96.0	Wade Hixton's Counter Punch	94
Metal Gear Solid 2: Sons of Liberty	96.0	Advance Wars 2: Black Hole Rising	93
Tekken 3	96.0	Castlevania: Symphony of the Night	93
Uncharted 2: Among Thieves	96.0	Monster Rancher Advance 2	93

The Number of Video Games Released per Year

The table below summarizes the number of video games that have been released each year. The graph visualizes the number of video games released each year. Overall, since 1980, the number of video games released per year has increased. However, the number of video released in a year reached its peak in 2008, and has started to decrease ever since.

In [6]:
```
vg_year<-vg %>% group_by(Year_of_Release) %>% tally() %>% arrange(desc(n))
head(vg_year)
ggplot(vg,aes(x=Year_of_Release,fill=..count..))+
geom_bar()+
labs(title="Number of Games Released Every Year")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Year_of_Release	n
2008	1427
2009	1426
2010	1255
2007	1197
2011	1136
2006	1006

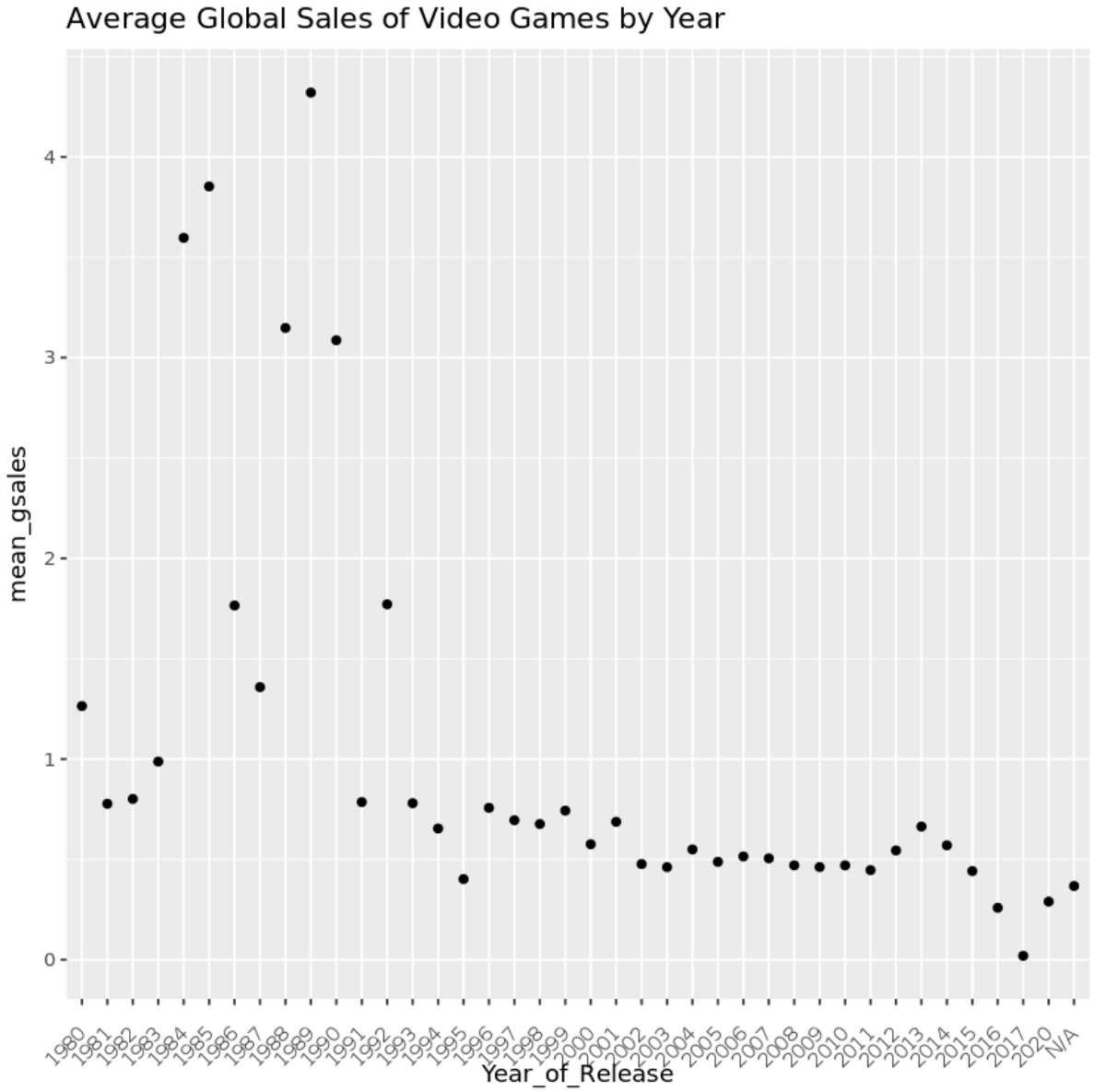


Average Global Sales of Video Games by Year

The table below shows the average global sales for a video game in a given year. The scatterplot visualizes the average global sales for video games in a given year. In 1989, video games had the highest average global sales. Overall, as the years have gone by, individual video games have tended to have less global sales.

In [7]:
```
gsales_year<-vg %>% group_by(Year_of_Release) %>% summarize(mean_gsales=mean(G
lobal_Sales)) %>% arrange(desc(mean_gsales))
head(gsales_year)
ggplot(gsales_year, aes(x=Year_of_Release, y=mean_gsales))+
geom_point()+
labs(title="Average Global Sales of Video Games by Year")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Year_of_Release	mean_gsales
1989	4.320588
1985	3.852857
1984	3.597143
1988	3.148000
1990	3.086875
1992	1.771395

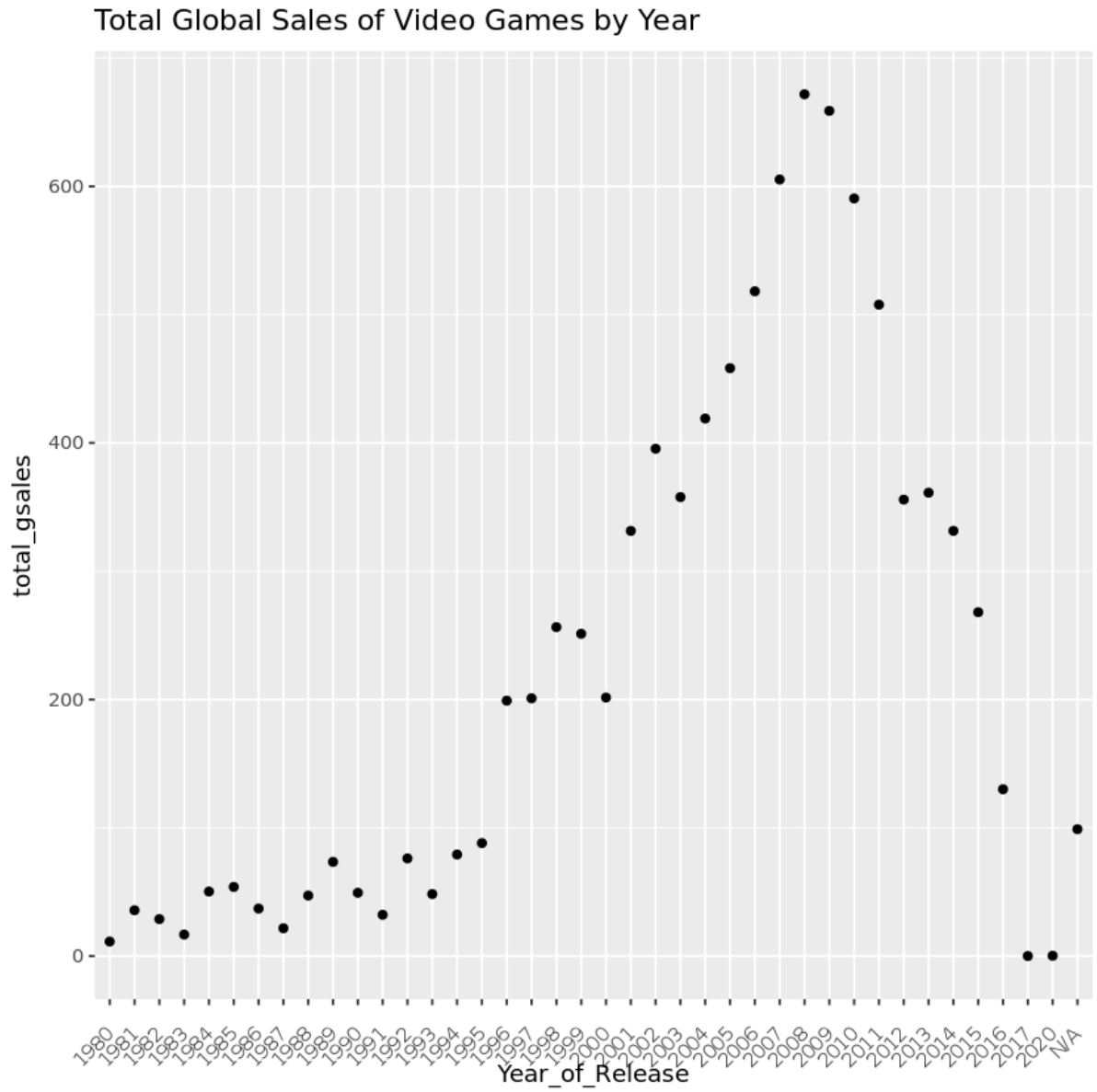


Total Global Sales of Video Games by Year

The table below shows the total global sales for video games in a given year. The scatterplot visualizes the total global sales for video games in a given year. From 1980 to 2008, the total global sales of video games continued to increase and reached its peak in 2008. Since 2008, the total global sales of video games has been decreasing.

In [9]:
```
gsales_year<-vg %>% group_by(Year_of_Release) %>% summarize(total_gsales=sum(G
lobal_Sales)) %>% arrange(desc(total_gsales))
head(gsales_year)
ggplot(gsales_year, aes(x=Year_of_Release, y=total_gsales))+
geom_point()+
labs(title="Total Global Sales of Video Games by Year")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Year_of_Release	total_gsales
2008	671.79
2009	658.88
2007	605.37
2010	590.59
2006	518.22
2011	507.79

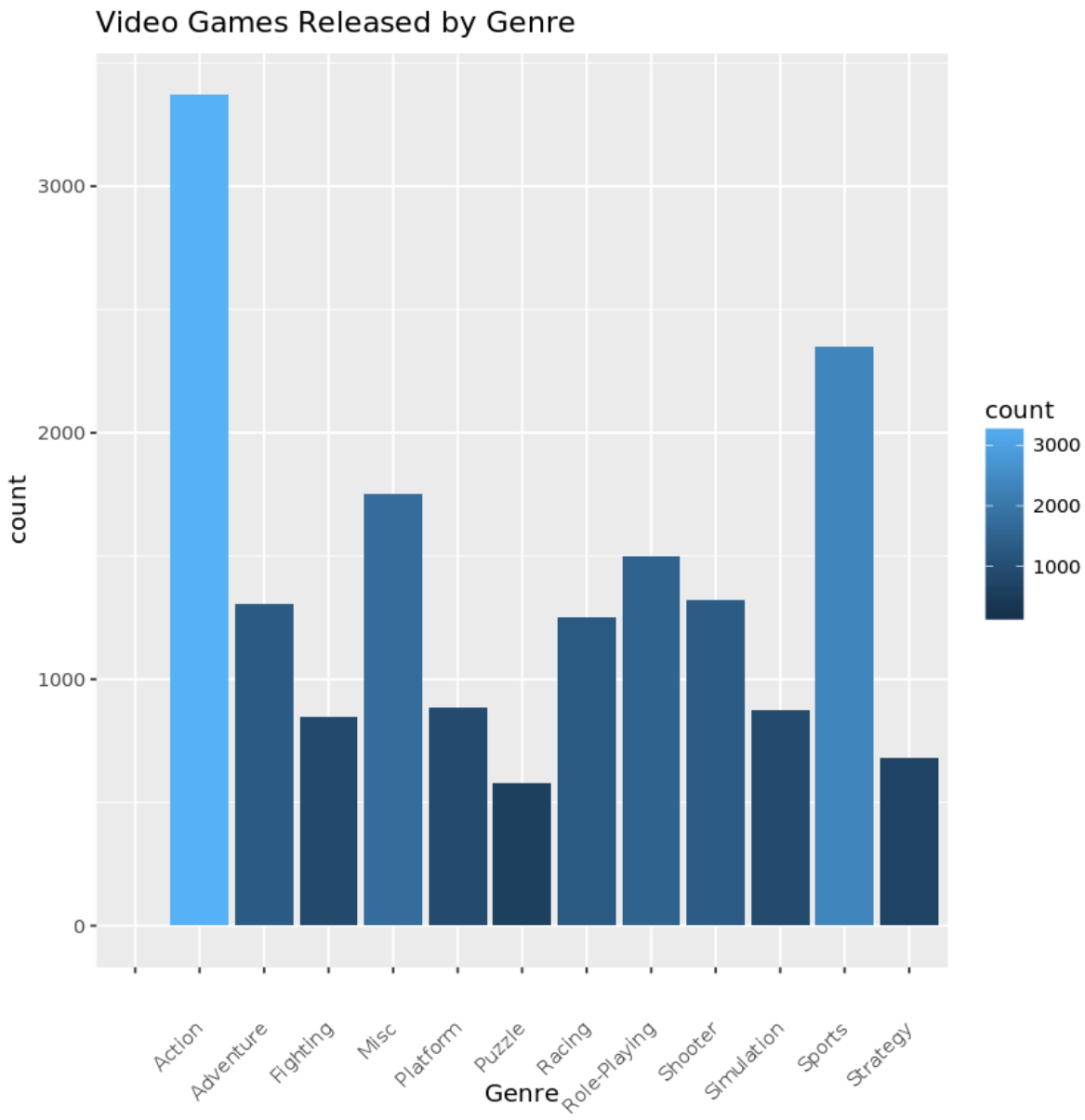


Video Games Released by Genre

The table below summarizes the number of video games that have been released per genre. The graph visualizes the number of video games that have been released per genre. Overall, video games are most likely to be part of the action genre or sports genre, while least likely to be part of the puzzle or strategy genre. This implies that people tend to like playing action and sports games the most.

In [10]:
```
vg_genre<-vg %>% group_by(Genre) %>% tally() %>% arrange(desc(n)) %>% mutate(p
roportion=n/sum(n))
vg_genre
ggplot(vg,aes(x=Genre,fill=..count..))+
geom_bar()+
labs(title="Video Games Released by Genre")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Genre	n	proportion
Action	3370	0.2015670794
Sports	2348	0.1404390215
Misc	1750	0.1046713320
Role-Playing	1500	0.0897182846
Shooter	1323	0.0791315270
Adventure	1303	0.0779352832
Racing	1249	0.0747054250
Platform	888	0.0531132245
Simulation	874	0.0522758538
Fighting	849	0.0507805491
Strategy	683	0.0408517256
Puzzle	580	0.0346910700
	2	0.0001196244



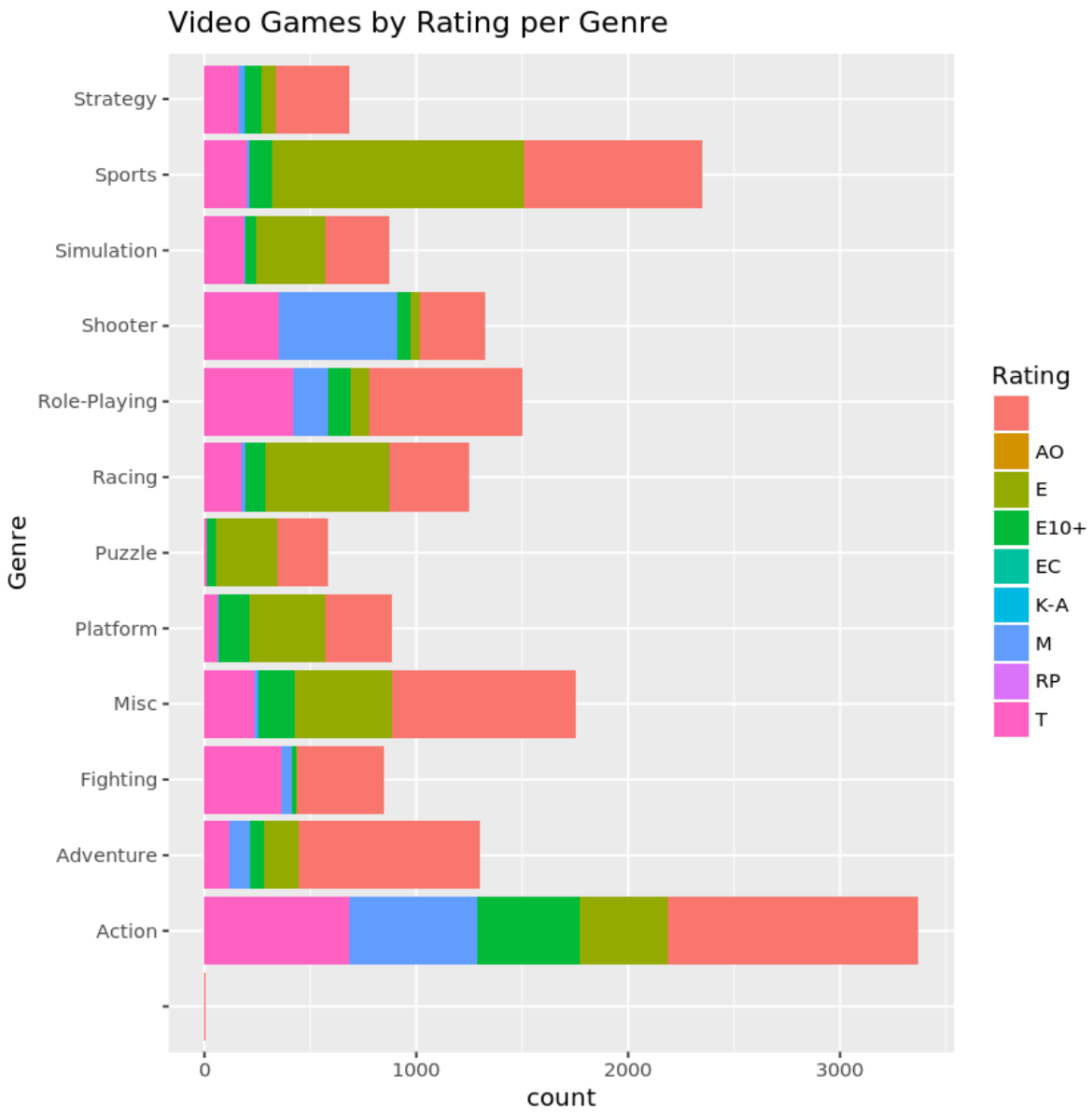
Video Games by Rating per Genre

The table below summarizes the number and proportion of video games by rating per genre. The graph visualizes the proportion of video games by rating per genre. Overall, Sports, Racing, and Puzzle games tend to be rated Eveyone (E), Fighting games tend to be rated Teen (T), and Shooter games tend to be rated Mature (M).

In [11]:
```
rating_genre<-vg %>% group_by(Genre,Rating) %>% tally() %>% mutate(proportion=
n/sum(n))
rating_genre
ggplot(vg)+
geom_bar(aes(x=Genre,fill=Rating))+
labs(title="Video Games by Rating per Genre")+
coord_flip()
```

Genre	Rating	n	proportion
		2	1.0000000000
Action		1182	0.3507418398
Action	AO	1	0.0002967359
Action	E	416	0.1234421365
Action	E10+	481	0.1427299703
Action	EC	1	0.0002967359
Action	M	608	0.1804154303
Action	T	681	0.2020771513
Adventure		857	0.6577129701
Adventure	E	162	0.1243284728
Adventure	E10+	68	0.0521872602
Adventure	EC	2	0.0015349194
Adventure	M	99	0.0759785111
Adventure	T	115	0.0882578665
Fighting		411	0.4840989399
Fighting	E	8	0.0094228504
Fighting	E10+	19	0.0223792697
Fighting	M	49	0.0577149588
Fighting	T	362	0.4263839812
Misc		868	0.4960000000
Misc	E	457	0.2611428571
Misc	E10+	167	0.0954285714
Misc	EC	5	0.0028571429
Misc	K-A	1	0.0005714286
Misc	M	13	0.0074285714
Misc	T	239	0.1365714286
Platform		319	0.3592342342
Platform	E	358	0.4031531532
Platform	E10+	144	0.1621621622
Platform	M	3	0.0033783784
⋮	⋮	⋮	⋮
Racing	M	18	0.0144115292

Genre	Rating	n	proportion
Racing	RP	1	0.0008006405
Racing	T	172	0.1377101681
Role-Playing		723	0.4820000000
Role-Playing	E	84	0.0560000000
Role-Playing	E10+	111	0.0740000000
Role-Playing	M	162	0.1080000000
Role-Playing	T	420	0.2800000000
Shooter		304	0.2297808012
Shooter	E	48	0.0362811791
Shooter	E10+	58	0.0438397581
Shooter	M	565	0.4270597128
Shooter	T	348	0.2630385488
Simulation		305	0.3489702517
Simulation	E	326	0.3729977117
Simulation	E10+	48	0.0549199085
Simulation	M	5	0.0057208238
Simulation	T	190	0.2173913043
Sports		839	0.3573253833
Sports	E	1188	0.5059625213
Sports	E10+	107	0.0455706985
Sports	M	16	0.0068143101
Sports	T	198	0.0843270869
Strategy		344	0.5036603221
Strategy	E	70	0.1024890190
Strategy	E10+	78	0.1142020498
Strategy	K-A	2	0.0029282577
Strategy	M	25	0.0366032211
Strategy	RP	2	0.0029282577
Strategy	T	162	0.2371888726



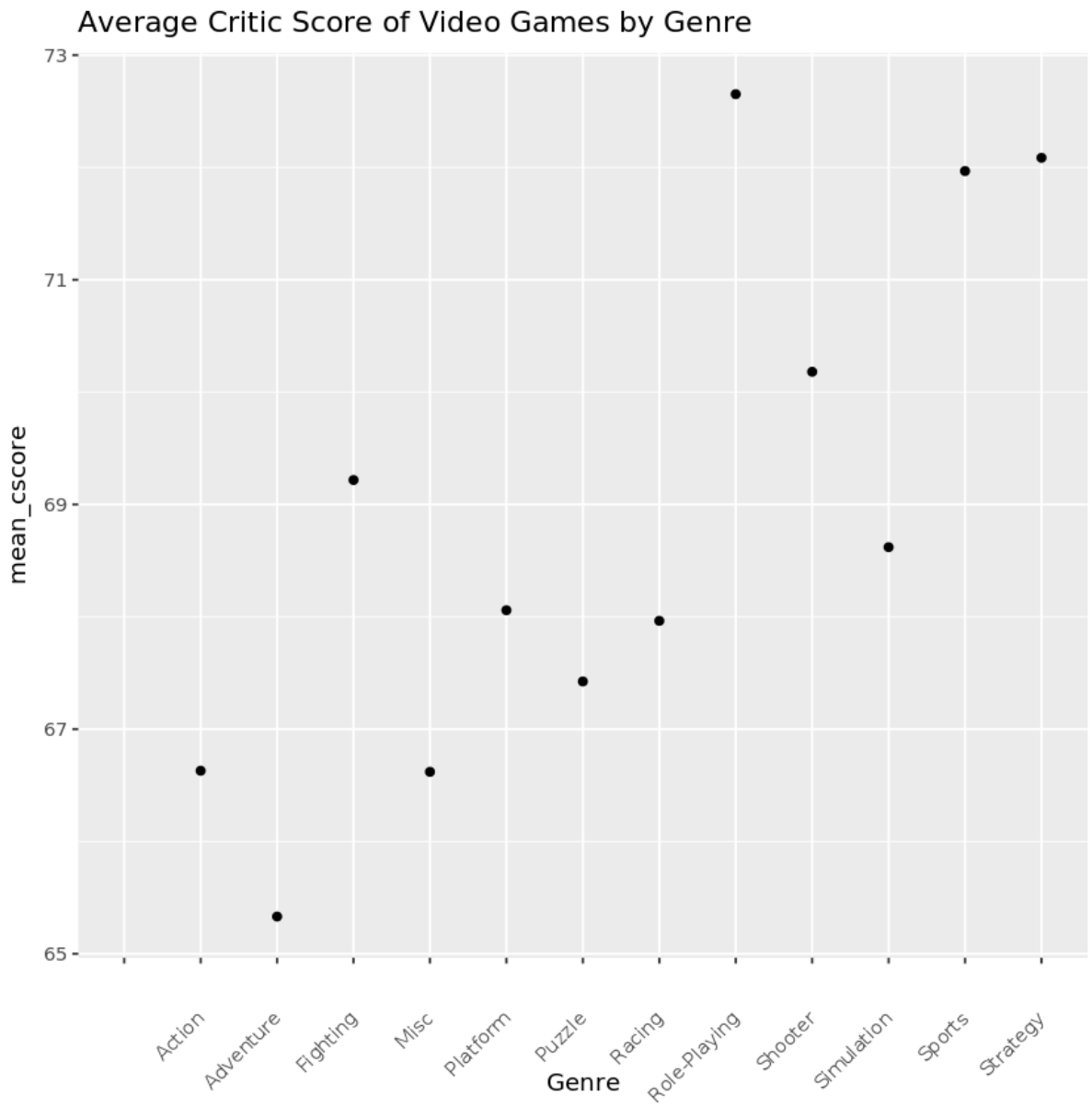
Average Critic Score, User Score of Video Games by Genre

The table below summarizes the average critic score and user score for games of each genre, with N/A critic score and user score values being removed. The scatterplot visualizes each genre and the average critic score of video games of that genre. Overall, role-playing and strategy games tend to get the highest scores from critics, while adventure and action games tend to get the lowest rating from critics. Role-playing and fighting games tend to get the highest scores from users, while sports and racing games tend to get the lowest rating from users. Critics and users tend to rate role-playing games the highest, but overall tend to rate games of different genres differently.

In [12]:
```
critic_score_genre<-vg %>% group_by(Genre) %>% summarize(mean_cscore=mean(Crit
ic_Score,na.rm=TRUE)) %>% arrange(desc(mean_cscore))
user_score_genre<-vg %>% group_by(Genre) %>% summarize(mean_uscore=mean(User_S
core,na.rm=TRUE)) %>% arrange(desc(mean_uscore))
cbind(critic_score_genre,user_score_genre)
ggplot(critic_score_genre, aes(x=Genre,y=mean_cscore))+
geom_point()+
labs(title="Average Critic Score of Video Games by Genre")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Genre	mean_cscore	Genre	mean_uscore
Role-Playing	72.65265	Role-Playing	75.19515
Strategy	72.08609	Fighting	72.02506
Sports	71.96817	Platform	72.01869
Shooter	70.18114	Strategy	71.95498
Fighting	69.21760	Puzzle	70.75000
Simulation	68.61932	Simulation	70.34593
Platform	68.05835	Adventure	70.34333
Racing	67.96361	Action	69.54153
Puzzle	67.42411	Shooter	69.41883
Action	66.62910	Racing	69.36661
Misc	66.61950	Sports	68.61559
Adventure	65.33127	Misc	67.19590
	NaN		NaN

```
Warning message:
“Removed 1 rows containing missing values (geom_point).”
```

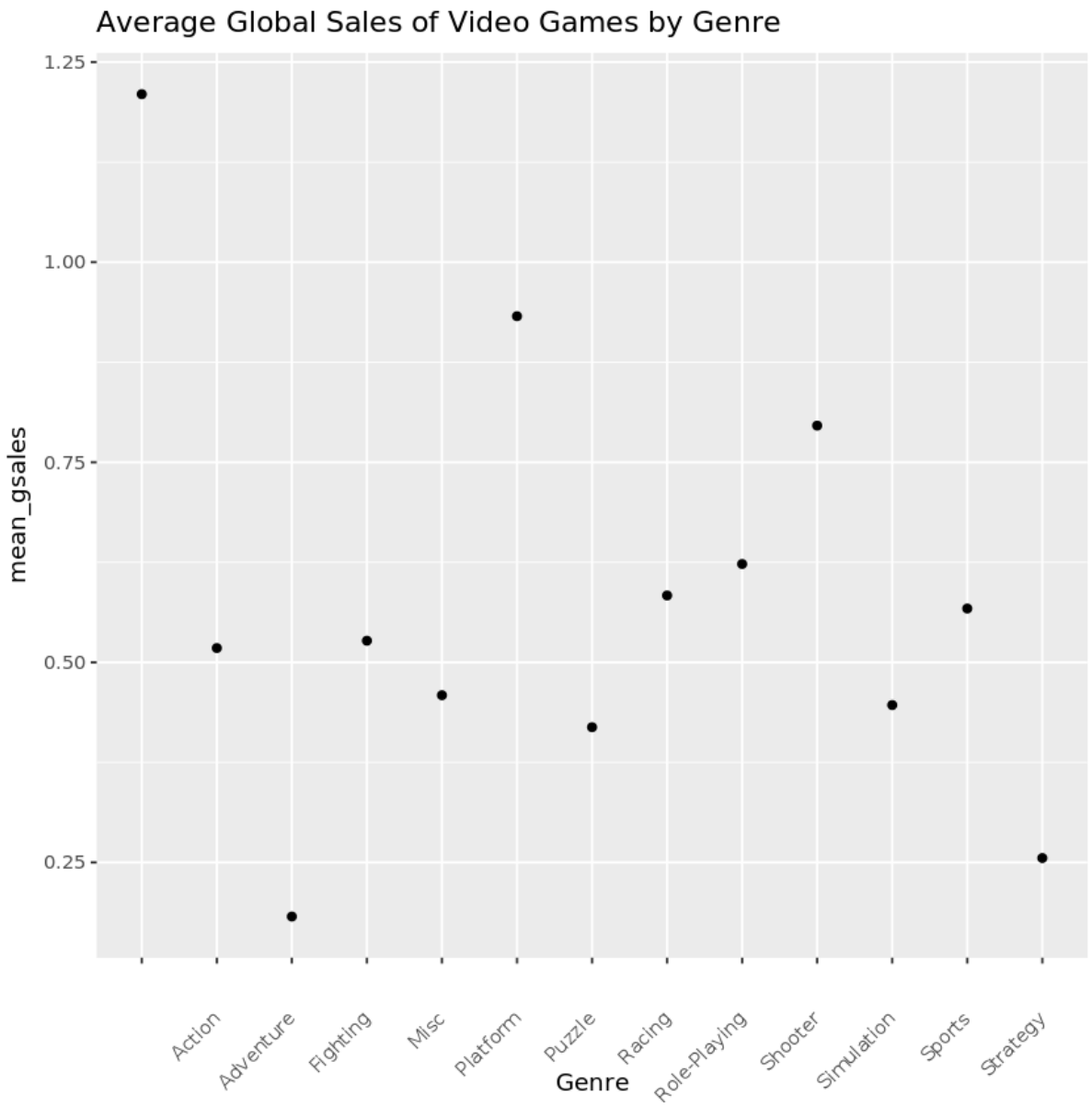


Average Global, NA, EU, JP, Other Region Sales of Video Games by Genre

The table below summarizes the average global, NA, EU, JP, and other region sales for games of each genre. The scatterplot visualizes each genre and the average global sales of video games of that genre. Overall, platform and shooter games tend to have the highest mean global sales, while adventure and strategy games tend to have the lowest mean global sales. Platform and shooter games tend to have the highest mean sales in North America, while Adventure and strategy games tend to have the lowest mean sales in North America. Shooter and platform games tend to have the highest mean sales in Europe, while Adventure and strategy games tend to have the lowest mean sales in Europe. Role-Playing and platform games tend to have the highest mean sales in Japan, while Shooter and adventure games tend to have the lowest mean sales in Japan. Shooter and racing games tend to have the highest sales in other regions, while Adventure and strategy games tend to have the lowest sales in other regions. In all regions besides Japan, Adventure games tend to get the lowest mean sales. In Japan, Shooter games tend to have the lowest mean sales, whereas in other regions, shooter games tend to get high mean sales. This shows how Japanese people tend to avoid buying shooter games, probably due to their violence, whereas in other regions, shooter games are very popular.

In [13]:
```
gsales_genre<-vg %>% group_by(Genre) %>% summarize(mean_gsales=mean(Global_Sal
es)) %>% arrange(desc(mean_gsales))
NAsales_genre<-vg %>% group_by(Genre) %>% summarize(mean_NAsales=mean(NA_Sales
)) %>% arrange(desc(mean_NAsales))
EUsales_genre<-vg %>% group_by(Genre) %>% summarize(mean_EUsales=mean(EU_Sales
)) %>% arrange(desc(mean_EUsales))
JPsales_genre<-vg %>% group_by(Genre) %>% summarize(mean_JPsales=mean(JP_Sales
)) %>% arrange(desc(mean_JPsales))
osales_genre<-vg %>% group_by(Genre) %>% summarize(mean_osales=mean(Other_Sale
s)) %>% arrange(desc(mean_osales))
sales_genre<-cbind(gsales_genre,NAsales_genre,EUsales_genre,JPsales_genre,osal
es_genre)
sales_genre
ggplot(gsales_genre, aes(x=Genre,y=mean_gsales))+
geom_point()+
labs(title="Average Global Sales of Video Games by Genre")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Genre	mean_gsales	Genre	mean_NAsales	Genre	mean_EUsales	Genre
	1.2100000		0.89000000		0.26500000	Role-Playing
Platform	0.9325225	Platform	0.50168919	Shooter	0.23986395	Platform
Shooter	0.7958730	Shooter	0.44764928	Platform	0.22561937	Fighting
Role-Playing	0.6229333	Sports	0.29149489	Racing	0.18935949	Puzzle
Racing	0.5835869	Racing	0.28771017	Sports	0.16047274	Simulation
Sports	0.5672913	Fighting	0.26308598	Action	0.15404451	Strategy
Fighting	0.5270671	Action	0.26083383	Simulation	0.12988558	Misc
Action	0.5178843	Misc	0.23272571	Role-Playing	0.12580667	Sports
Misc	0.4589600	Role-Playing	0.22054000	Misc	0.12156571	Action
Simulation	0.4467048	Puzzle	0.21184483	Fighting	0.11817432	Racing
Puzzle	0.4190000	Simulation	0.20845538	Puzzle	0.08622414	Adventure
Strategy	0.2554905	Strategy	0.10036603	Strategy	0.06613470	Shooter
Adventure	0.1824175	Adventure	0.08078281	Adventure	0.04876439	

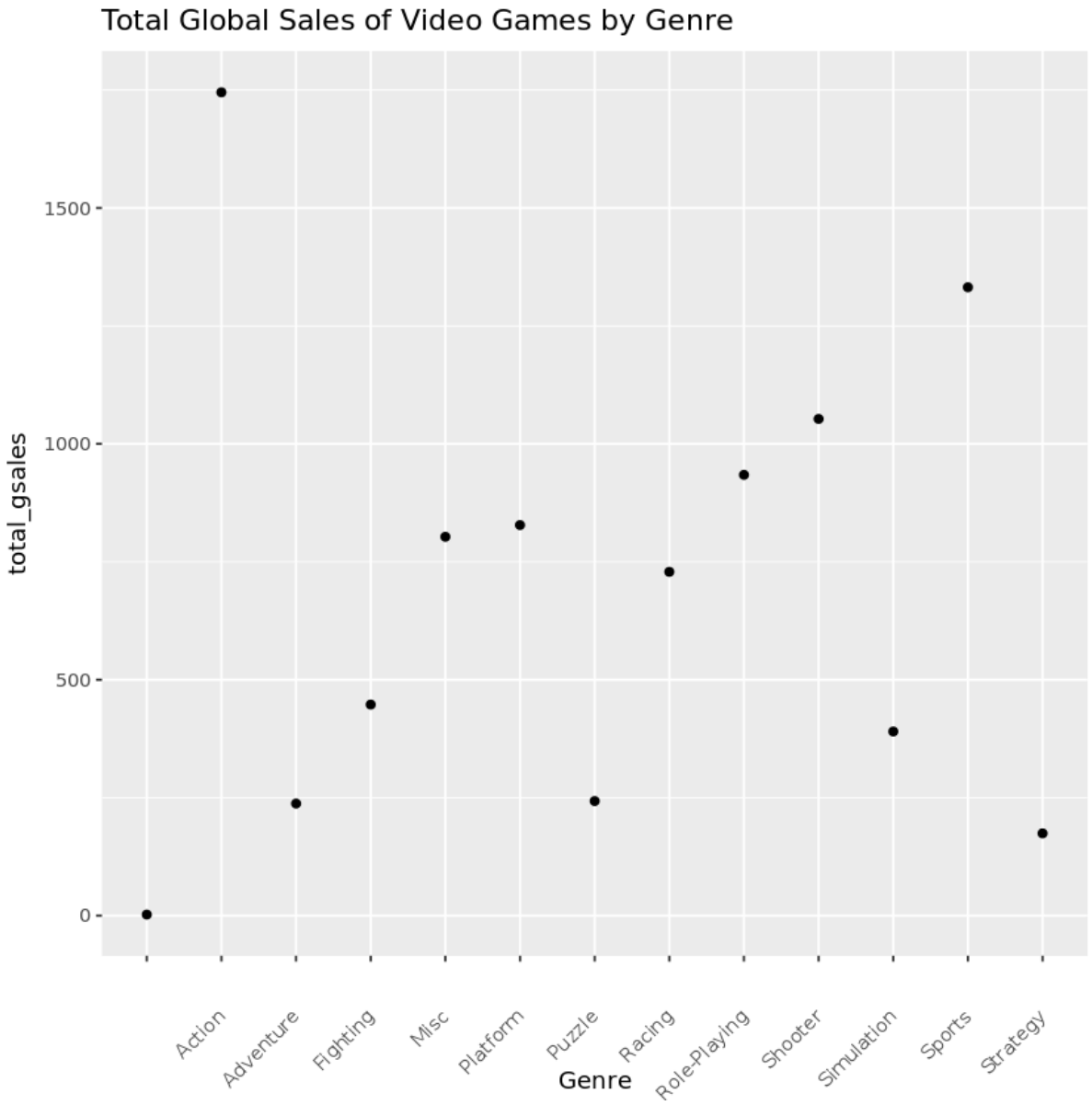


Total Global, NA, EU, JP, Other Region Sales of Video Games by Genre

The table below summarizes the total global, NA, EU, JP, and other region sales for games of each genre. The scatterplot visualizes each genre and the total global sales of video games of that genre. Overall, action and sports games have sold the most globally, in North America, in Europe, and in other regions, while role-playing games have sold the most in Japan. Strategy games have sold the least sales globally, in North America, in Europe, and in other regions, while shooter games have sold the least in Japan. Japanese people tend to buy less-violent video games, whereas in all other regions, people tend to buy more violent games.

In [14]:
```
gsales_genre<-vg %>% group_by(Genre) %>% summarize(total_gsales=sum(Global_Sal
es)) %>% arrange(desc(total_gsales))
NAsales_genre<-vg %>% group_by(Genre) %>% summarize(total_NAsales=sum(NA_Sales
)) %>% arrange(desc(total_NAsales))
EUsales_genre<-vg %>% group_by(Genre) %>% summarize(total_EUsales=sum(EU_Sales
)) %>% arrange(desc(total_EUsales))
JPsales_genre<-vg %>% group_by(Genre) %>% summarize(total_JPsales=sum(JP_Sales
)) %>% arrange(desc(total_JPsales))
osales_genre<-vg %>% group_by(Genre) %>% summarize(total_osales=sum(Other_Sale
s)) %>% arrange(desc(total_osales))
sales_genre<-cbind(gsales_genre,NAsales_genre,EUsales_genre,JPsales_genre,osal
es_genre)
sales_genre
ggplot(gsales_genre, aes(x=Genre,y=total_gsales))+
geom_point()+
labs(title="Total Global Sales of Video Games by Genre")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Genre	total_gsales	Genre	total_NAsales	Genre	total_EUsales	Genre	t
Action	1745.27	Action	879.01	Action	519.13	Role-Playing	3
Sports	1332.00	Sports	684.43	Sports	376.79	Action	1
Shooter	1052.94	Shooter	592.24	Shooter	317.34	Sports	1
Role-Playing	934.40	Platform	445.50	Racing	236.51	Platform	1
Platform	828.08	Misc	407.27	Misc	212.74	Misc	1
Misc	803.18	Racing	359.35	Platform	200.35	Fighting	8
Racing	728.90	Role-Playing	330.81	Role-Playing	188.71	Simulation	6
Fighting	447.48	Fighting	223.36	Simulation	113.52	Puzzle	5
Simulation	390.42	Simulation	182.19	Fighting	100.33	Racing	5
Puzzle	243.02	Puzzle	122.87	Adventure	63.54	Adventure	5
Adventure	237.69	Adventure	105.26	Puzzle	50.01	Strategy	4
Strategy	174.50	Strategy	68.55	Strategy	45.17	Shooter	3
	2.42		1.78		0.53		0

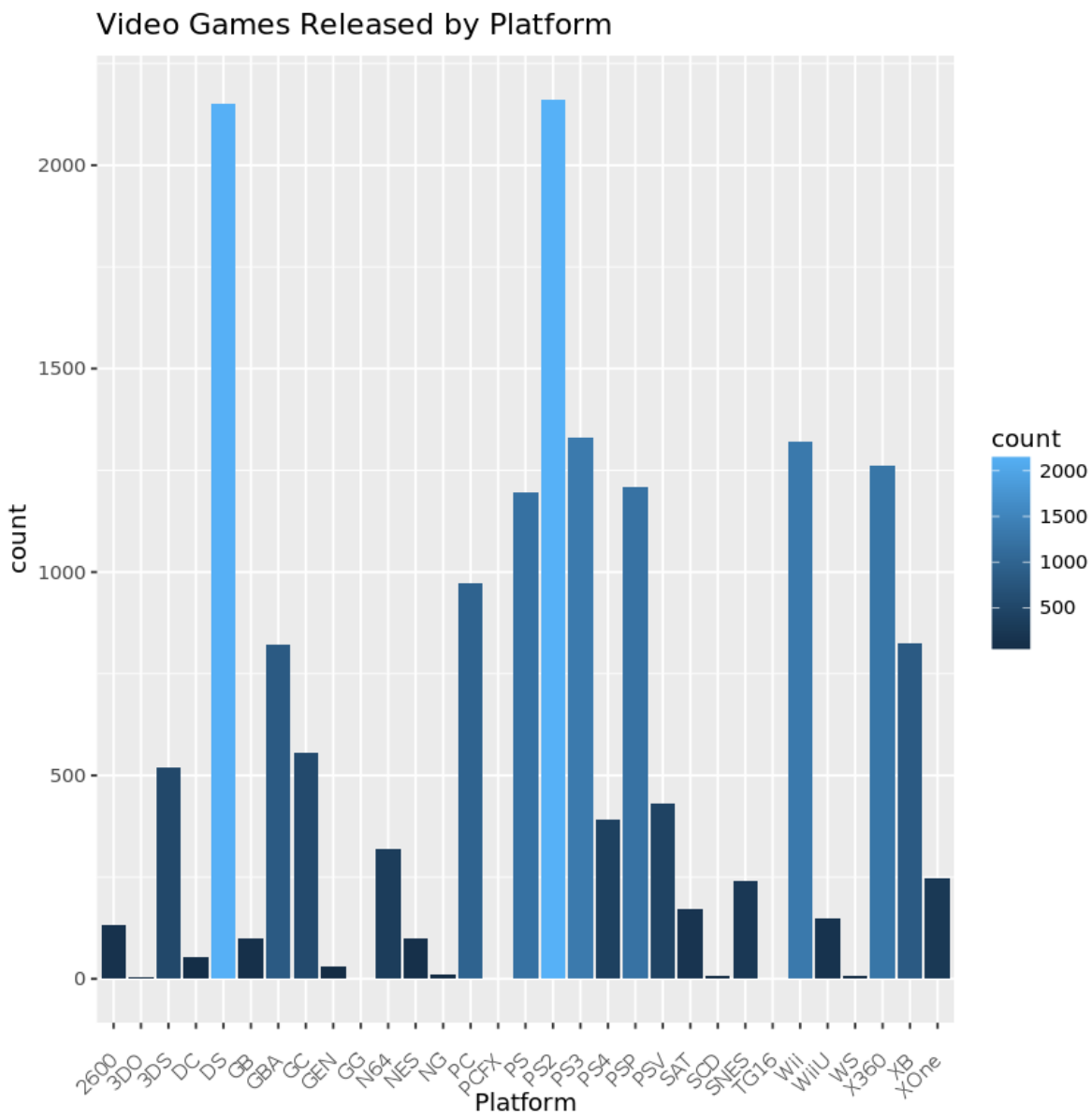


Number of Video Games Released By Platform

The table below summarizes the number of video games that have been released per platform. The graph visualizes the number of video games that have been released per platform. Overall, the PS2 and NIntendo DS (DS) have the most video games released out of all the platforms, while the PCFX and GG have the least video games released out of all the platforms.

In [15]:
```
vg_platform<-vg %>% group_by(Platform) %>% tally() %>% arrange(desc(n))
vg_platform
ggplot(vg)+
geom_bar(aes(x=Platform,fill=..count..))+
labs(title="Video Games Released by Platform")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Platform	n
PS2	2161
DS	2152
PS3	1331
Wii	1320
X360	1262
PSP	1209
PS	1197
PC	974
XB	824
GBA	822
GC	556
3DS	520
PSV	432
PS4	393
N64	319
XOne	247
SNES	239
SAT	173
WiiU	147
2600	133
GB	98
NES	98
DC	52
GEN	29
NG	12
SCD	6
WS	6
3DO	3
TG16	2
GG	1
PCFX	1

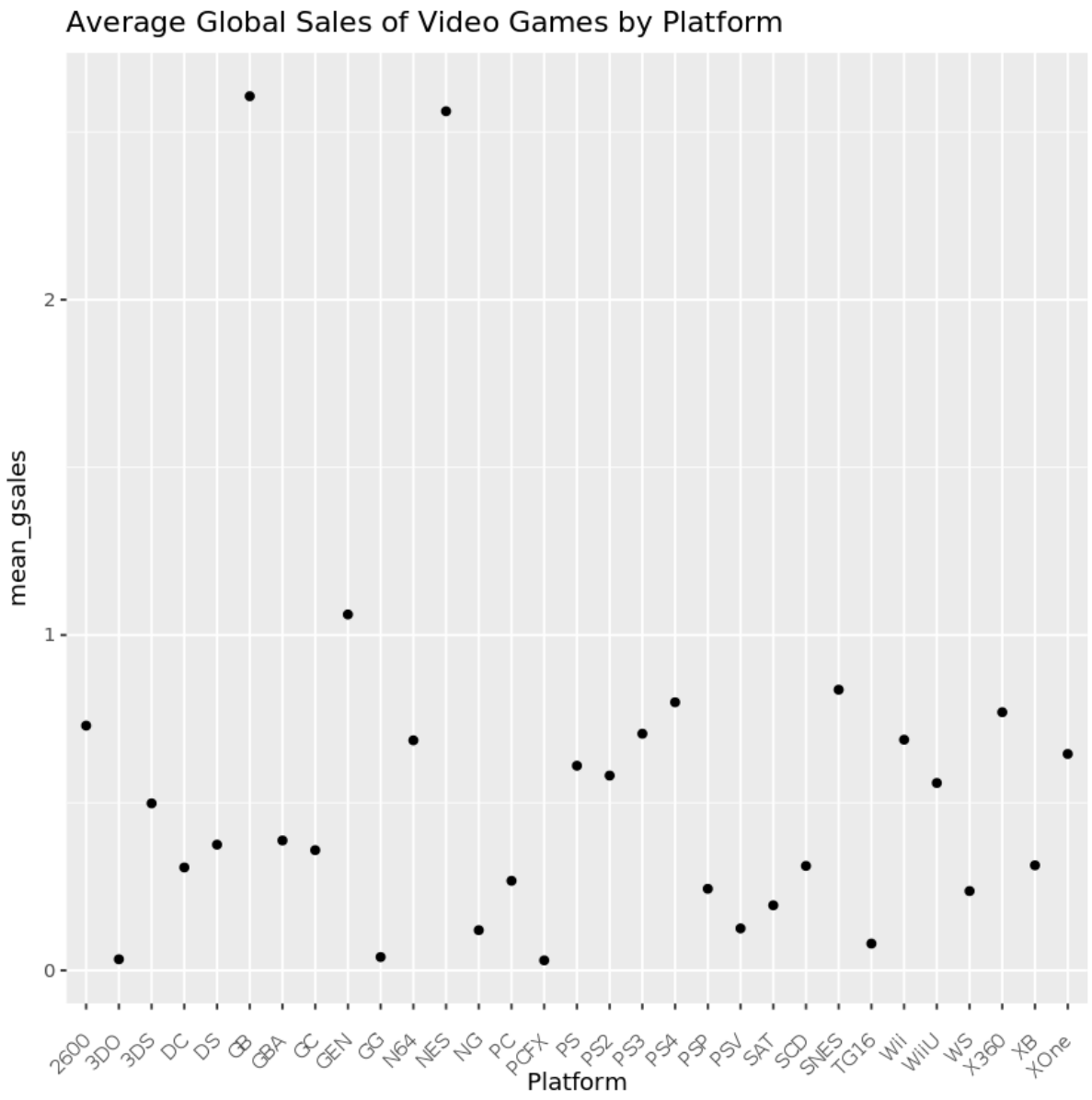


Average Global, NA, EU, JP, Other Region Sales by Platform

The table below summarizes the average global, NA, EU, JP, and other region sales for games of each platform. The scatterplot visualizes each platform and the average global sales of video games of that platform. Overall, Gameboy (GB) and Nintendo Entertainment (NES) games tend to have the highest mean global sales. Nintendo Entertainment System and Gameboy games tend to have the highest mean sales in North America. Gameboy and Playstation 4 games tend to have the highest mean sales in Europe. Nintendo Entertainment System and Gameboy games tend to have the highest mean sales in Japan. Playstation 4 (PS4) and Playstation 3 (PS3) games tend to have the highest mean sales in other regions. In North America, Europe, and Japan, games for Nintendo consoles tend to have the highest mean sales, while in other regions, games for Sony (Playstation) consoles tend to have the highest mean sales.

In [16]:
```
gsales_platform<-vg %>% group_by(Platform) %>% summarize(mean_gsales=mean(Glob
al_Sales)) %>% arrange(desc(mean_gsales))
NAsales_platform<-vg %>% group_by(Platform) %>% summarize(mean_NAsales=mean(NA
_Sales)) %>% arrange(desc(mean_NAsales))
EUsales_platform<-vg %>% group_by(Platform) %>% summarize(mean_EUsales=mean(EU
_Sales)) %>% arrange(desc(mean_EUsales))
JPsales_platform<-vg %>% group_by(Platform) %>% summarize(mean_JPsales=mean(JP
_Sales)) %>% arrange(desc(mean_JPsales))
osales_platform<-vg %>% group_by(Platform) %>% summarize(mean_osales=mean(Othe
r_Sales)) %>% arrange(desc(mean_osales))
sales_platform<-cbind(gsales_platform,NAsales_platform,EUsales_platform,JPsale
s_platform,osales_platform)
sales_platform
ggplot(gsales_platform, aes(x=Platform,y=mean_gsales))+
geom_point()+
labs(title="Average Global Sales of Video Games by Platform")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Platform	mean_gsales	Platform	mean_NAsales	Platform	mean_EUsales	Platform	mea
GB	2.60663265	NES	1.28510204	GB	0.487959184	NES	1.0
NES	2.56193878	GB	1.16653061	PS4	0.359007634	GB	0.8
GEN	1.06137931	GEN	0.72586207	PS3	0.248151766	SNES	0.4
SNES	0.83702929	2600	0.68120301	NES	0.215816327	WS	0.2
PS4	0.79956743	X360	0.47739303	X360	0.214548336	3DS	0.1
X360	0.76991284	N64	0.43579937	XOne	0.208866397	SAT	0.1
2600	0.72992481	XOne	0.37700405	GEN	0.208620690	DC	0.1
PS3	0.70580766	Wii	0.37643939	Wii	0.198643939	NG	0.1
Wii	0.68797727	PS3	0.29563486	PS	0.178454470	PS	0.1
N64	0.68614420	PS	0.28113617	WiiU	0.170952381	N64	0.1
XOne	0.64550607	PS4	0.27669211	PS2	0.157006016	GEN	0.0
PS	0.61042607	PS2	0.27017122	PC	0.146242300	WiiU	0.0
PS2	0.58104581	WiiU	0.25918367	N64	0.128714734	DS	0.0
WiiU	0.55891156	SNES	0.25619247	3DS	0.118230769	TG16	0.0
3DS	0.49825000	GC	0.24003597	GBA	0.091545012	SCD	0.0
GBA	0.38746959	GBA	0.22815085	DS	0.087774164	PS2	0.0
DS	0.37504647	XB	0.22656553	SNES	0.079665272	PSP	0.0
GC	0.35856115	DS	0.17782063	XB	0.073968447	PS3	0.0
XB	0.31342233	SCD	0.16666667	GC	0.069622302	GBA	0.0
SCD	0.31166667	3DS	0.16055769	SCD	0.060000000	Wii	0.0
DC	0.30711538	DC	0.10442308	PSP	0.055153019	PSV	0.0
PC	0.26724846	PC	0.09705339	2600	0.041127820	PS4	0.0
PSP	0.24342432	PSP	0.09029777	DC	0.032500000	GG	0.0
WS	0.23666667	PSV	0.02912037	PSV	0.030370370	GC	0.0
SAT	0.19416185	SAT	0.00416185	SAT	0.003121387	3DO	0.0
PSV	0.12527778	3DO	0.00000000	3DO	0.000000000	PCFX	0.0
NG	0.12000000	GG	0.00000000	GG	0.000000000	X360	0.0
TG16	0.08000000	NG	0.00000000	NG	0.000000000	XB	0.0
GG	0.04000000	PCFX	0.00000000	PCFX	0.000000000	XOne	0.0
3DO	0.03333333	TG16	0.00000000	TG16	0.000000000	PC	0.0
PCFX	0.03000000	WS	0.00000000	WS	0.000000000	2600	0.0

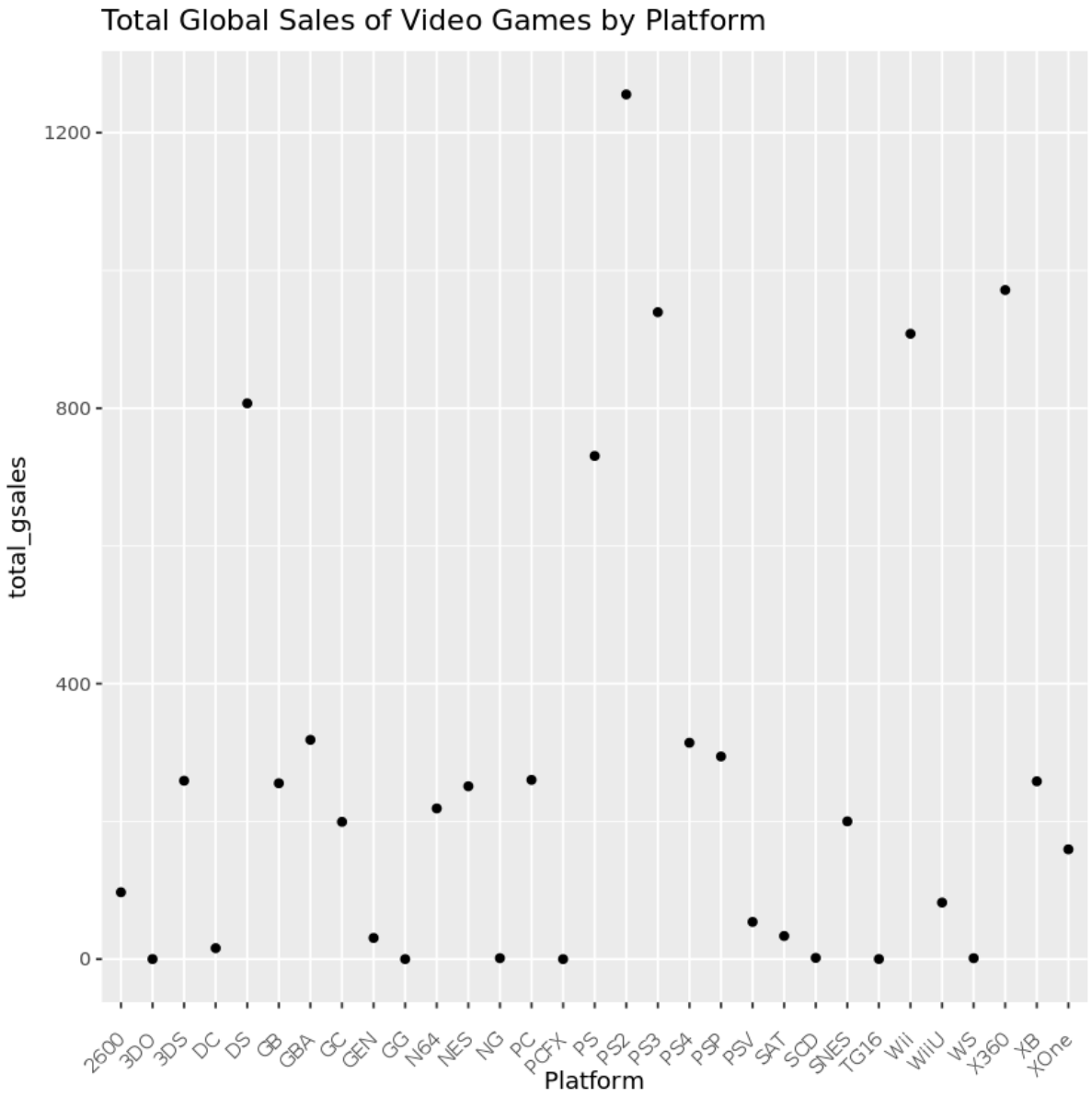


Total Global, NA, EU, JP, Other Region Sales of Video Games by Platform

The table below summarizes the total global, NA, EU, JP, and other region sales for games of each platform. The scatterplot visualizes each platform and the total global sales of video games of that platform. Overall, PS2 games have sold the most globally, in Europe, and in other regions, games for the Xbox 360 have sold the most in North America, and games for the Nintendo DS (DS) have sold the most in Japan.

In [17]:
```
gsales_platform<-vg %>% group_by(Platform) %>% summarize(total_gsales=sum(Glob
al_Sales)) %>% arrange(desc(total_gsales))
NAsales_platform<-vg %>% group_by(Platform) %>% summarize(total_NAsales=sum(NA
_Sales)) %>% arrange(desc(total_NAsales))
EUsales_platform<-vg %>% group_by(Platform) %>% summarize(total_EUsales=sum(EU
_Sales)) %>% arrange(desc(total_EUsales))
JPsales_platform<-vg %>% group_by(Platform) %>% summarize(total_JPsales=sum(JP
_Sales)) %>% arrange(desc(total_JPsales))
osales_platform<-vg %>% group_by(Platform) %>% summarize(total_osales=sum(Othe
r_Sales)) %>% arrange(desc(total_osales))
sales_platform<-cbind(gsales_platform,NAsales_platform,EUsales_platform,JPsale
s_platform,osales_platform)
sales_platform
ggplot(gsales_platform, aes(x=Platform,y=total_gsales))+
geom_point()+
labs(title="Total Global Sales of Video Games by Platform")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Platform	total_gsales	Platform	total_NAsales	Platform	total_EUsales	Platform	total_J
PS2	1255.64	X360	602.47	PS2	339.29	DS	175.57
X360	971.63	PS2	583.84	PS3	330.29	PS	139.82
PS3	939.43	Wii	496.90	X360	270.76	PS2	139.20
Wii	908.13	PS3	393.49	Wii	262.21	SNES	116.55
DS	807.10	DS	382.67	PS	213.61	3DS	100.67
PS	730.68	PS	336.52	DS	188.89	NES	98.65
GBA	318.50	GBA	187.54	PC	142.44	GB	85.12
PS4	314.23	XB	186.69	PS4	141.09	PS3	80.19
PSP	294.30	N64	139.02	GBA	75.25	PSP	76.78
PC	260.30	GC	133.46	PSP	66.68	Wii	69.33
3DS	259.09	NES	125.94	3DS	61.48	GBA	47.33
XB	258.26	GB	114.32	XB	60.95	N64	34.22
GB	255.45	PSP	109.17	XOne	51.59	SAT	32.26
NES	251.07	PS4	108.74	GB	47.82	PSV	21.93
N64	218.88	PC	94.53	N64	41.06	GC	21.58
SNES	200.05	XOne	93.12	GC	38.71	PS4	16.00
GC	199.36	2600	90.60	WiiU	25.13	WiiU	13.01
XOne	159.44	3DS	83.49	NES	21.15	X360	12.43
2600	97.08	SNES	61.23	SNES	19.04	DC	8.56
WiiU	82.16	WiiU	38.10	PSV	13.12	GEN	2.70
PSV	54.12	GEN	21.05	GEN	6.05	NG	1.44
SAT	33.59	PSV	12.58	2600	5.47	WS	1.42
GEN	30.78	DC	5.43	DC	1.69	XB	1.38
DC	15.97	SCD	1.00	SAT	0.54	SCD	0.45
SCD	1.87	SAT	0.72	SCD	0.36	XOne	0.34
NG	1.44	3DO	0.00	3DO	0.00	PC	0.17
WS	1.42	GG	0.00	GG	0.00	TG16	0.16
TG16	0.16	NG	0.00	NG	0.00	3DO	0.10
3DO	0.10	PCFX	0.00	PCFX	0.00	GG	0.04
GG	0.04	TG16	0.00	TG16	0.00	PCFX	0.03
PCFX	0.03	WS	0.00	WS	0.00	2600	0.00



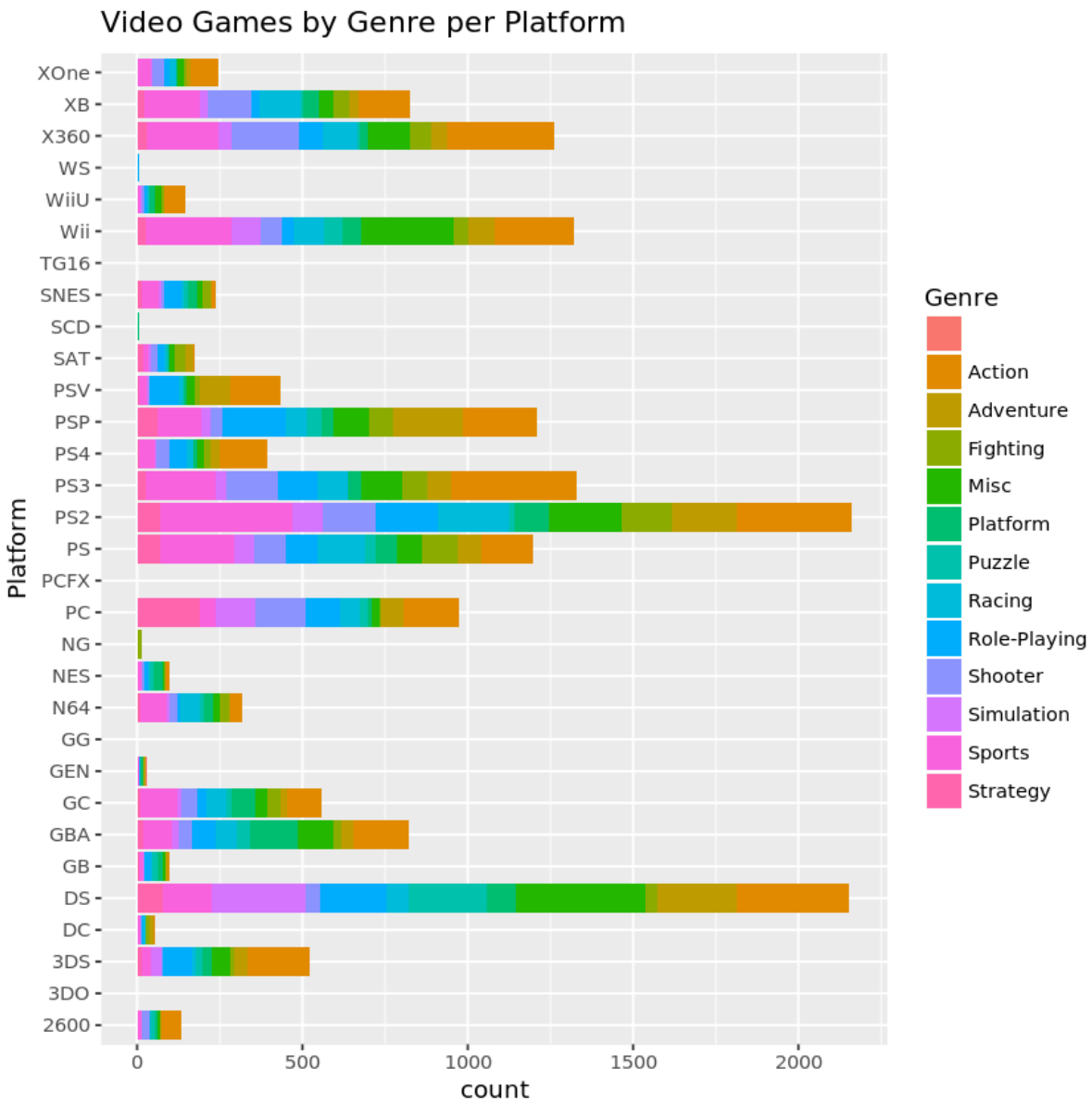
Video Games by Genre per Platform

The table below summarizes the number and proportion of video games by genre per platform. The graph visualizes the proportion of video games by genre per platform.

In [18]:
```
genre_platform<-vg %>% group_by(Platform,Genre) %>% tally() %>% mutate(proport
ion=n/sum(n))
genre_platform
ggplot(vg)+
geom_bar(aes(x=Platform,fill=Genre))+
labs(title="Video Games by Genre per Platform")+
coord_flip()
```

Platform	Genre	n	proportion
2600	Action	61	0.458646617
2600	Adventure	2	0.015037594
2600	Fighting	2	0.015037594
2600	Misc	5	0.037593985
2600	Platform	9	0.067669173
2600	Puzzle	11	0.082706767
2600	Racing	6	0.045112782
2600	Shooter	24	0.180451128
2600	Simulation	1	0.007518797
2600	Sports	12	0.090225564
3DO	Adventure	1	0.333333333
3DO	Puzzle	1	0.333333333
3DO	Simulation	1	0.333333333
3DS	Action	188	0.361538462
3DS	Adventure	38	0.073076923
3DS	Fighting	14	0.026923077
3DS	Misc	54	0.103846154
3DS	Platform	28	0.053846154
3DS	Puzzle	20	0.038461538
3DS	Racing	11	0.021153846
3DS	Role-Playing	88	0.169230769
3DS	Shooter	7	0.013461538
3DS	Simulation	31	0.059615385
3DS	Sports	26	0.050000000
3DS	Strategy	15	0.028846154
DC	Action	3	0.057692308
DC	Adventure	11	0.211538462
DC	Fighting	12	0.230769231
DC	Platform	2	0.038461538
DC	Racing	6	0.115384615
⋮	⋮	⋮	⋮
X360	Puzzle	7	0.005546751

Platform	Genre	n	proportion
X360	Racing	104	0.082408875
X360	Role-Playing	75	0.059429477
X360	Shooter	203	0.160855784
X360	Simulation	40	0.031695721
X360	Sports	217	0.171949287
X360	Strategy	28	0.022187005
XB	Action	155	0.188106796
XB	Adventure	26	0.031553398
XB	Fighting	48	0.058252427
XB	Misc	46	0.055825243
XB	Platform	49	0.059466019
XB	Puzzle	7	0.008495146
XB	Racing	123	0.149271845
XB	Role-Playing	23	0.027912621
XB	Shooter	132	0.160194175
XB	Simulation	24	0.029126214
XB	Sports	170	0.206310680
XB	Strategy	21	0.025485437
XOne	Action	85	0.344129555
XOne	Adventure	14	0.056680162
XOne	Fighting	7	0.028340081
XOne	Misc	19	0.076923077
XOne	Platform	5	0.020242915
XOne	Racing	20	0.080971660
XOne	Role-Playing	14	0.056680162
XOne	Shooter	38	0.153846154
XOne	Simulation	4	0.016194332
XOne	Sports	38	0.153846154
XOne	Strategy	3	0.012145749



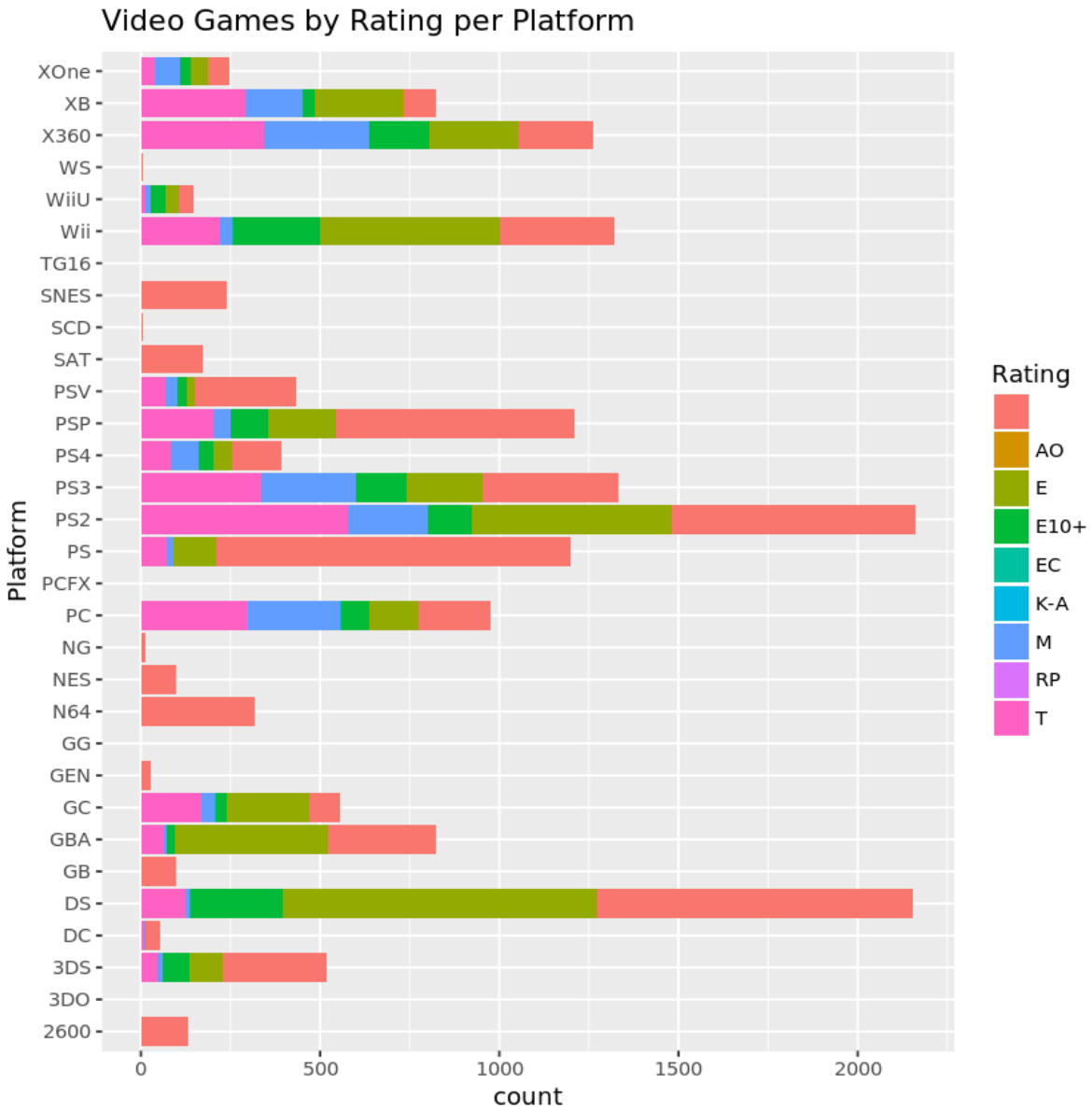
Video Games by Rating per Platform

The table below summarizes the number and proportion of video games by rating per platform. The graph visualizes the proportion of video games by rating per platform. Overall, Nintendo DS (DS),Gamboy Advance (GBA), Gamecube (GB), and Wii video games tend to be rated E, which are all platforms created by Nintendo.

In [19]:
```
rating_platform<-vg %>% group_by(Platform,Rating) %>% tally() %>% mutate(propo
rtion=n/sum(n))
rating_platform
ggplot(vg)+
geom_bar(aes(x=Platform,fill=Rating))+
labs(title="Video Games by Rating per Platform")+
coord_flip()
```

Platform	Rating	n	proportion
2600		133	1.000000000
3DO		3	1.000000000
3DS		292	0.561538462
3DS	E	92	0.176923077
3DS	E10+	76	0.146153846
3DS	M	12	0.023076923
3DS	T	48	0.092307692
DC		38	0.730769231
DC	E	3	0.057692308
DC	M	1	0.019230769
DC	T	10	0.192307692
DS		878	0.407992565
DS	E	877	0.407527881
DS	E10+	259	0.120353160
DS	EC	2	0.000929368
DS	M	10	0.004646840
DS	T	126	0.058550186
GB		98	1.000000000
GBA		300	0.364963504
GBA	E	426	0.518248175
GBA	E10+	25	0.030413625
GBA	M	7	0.008515815
GBA	T	64	0.077858881
GC		87	0.156474820
GC	E	230	0.413669065
GC	E10+	31	0.055755396
GC	EC	1	0.001798561
GC	M	38	0.068345324
GC	T	169	0.303956835
GEN		29	1.000000000
⋮	⋮	⋮	⋮
SNES		239	1.000000000

Platform	Rating	n	proportion
TG16		2	1.000000000
Wii		318	0.240909091
Wii	E	503	0.381060606
Wii	E10+	242	0.183333333
Wii	EC	2	0.001515152
Wii	M	33	0.025000000
Wii	T	222	0.168181818
WiiU		42	0.285714286
WiiU	E	35	0.238095238
WiiU	E10+	42	0.285714286
WiiU	M	15	0.102040816
WiiU	T	13	0.088435374
WS		6	1.000000000
X360		210	0.166402536
X360	E	247	0.195721078
X360	E10+	168	0.133122029
X360	M	294	0.232963550
X360	T	343	0.271790808
XB		91	0.110436893
XB	AO	1	0.001213592
XB	E	248	0.300970874
XB	E10+	31	0.037621359
XB	M	161	0.195388350
XB	T	292	0.354368932
XOne		61	0.246963563
XOne	E	45	0.182186235
XOne	E10+	31	0.125506073
XOne	M	70	0.283400810
XOne	T	40	0.161943320

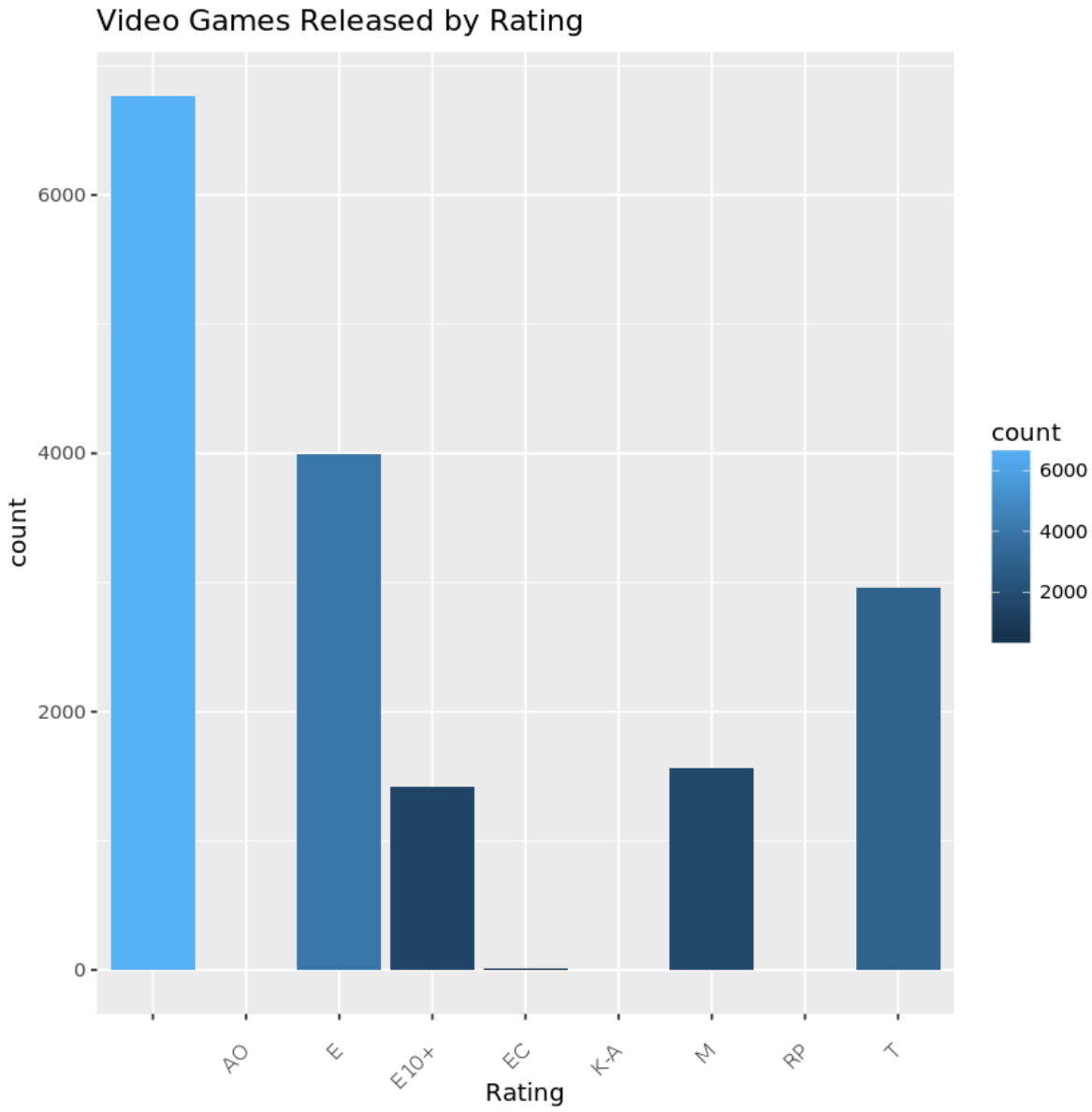


Video Games Released by Rating

The table below summarizes the number and proportion of video games that have been released per rating. The graph visualizes the number of video games that have been released by each rating. Overall, video games released are most likely to have an E (Everyone)and T (Teen) rating, while video games released are least likely to have an AO (Adult Only) rating. Video games tend to have ratings that appeal to a greater audience (age group).

In [20]:
```
vg_rating<-vg %>% group_by(Rating) %>% tally() %>% arrange(desc(n)) %>% mutate
(proportion=n/sum(n))
vg_rating
ggplot(vg)+
geom_bar(aes(x=Rating,fill=..count..))+
labs(title="Video Games Released by Rating")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Rating	n	proportion
	6769	4.048687e-01
E	3991	2.387104e-01
T	2961	1.771039e-01
M	1563	9.348645e-02
E10+	1420	8.493331e-02
EC	8	4.784975e-04
K-A	3	1.794366e-04
RP	3	1.794366e-04
AO	1	5.981219e-05

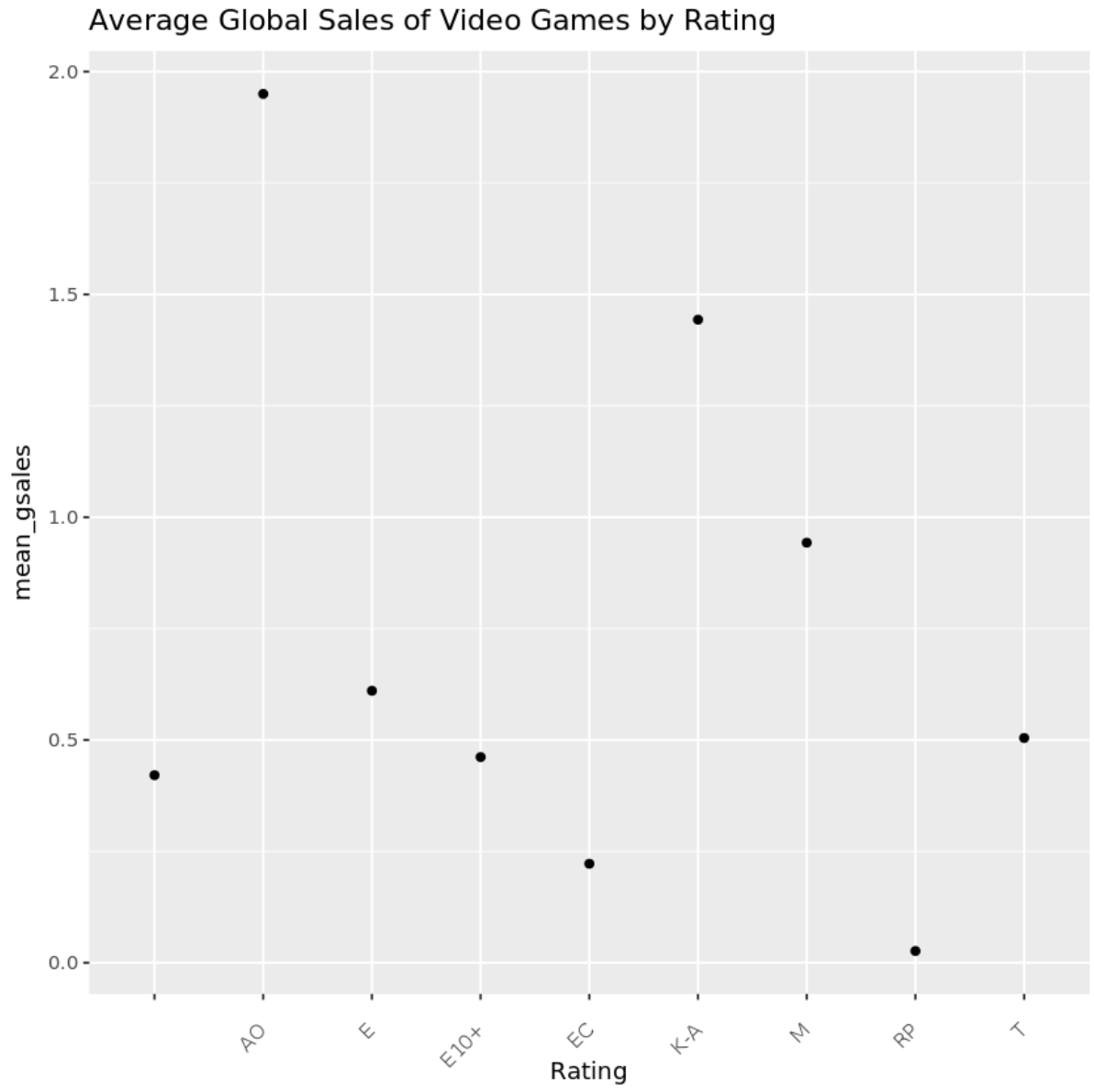


Average Global, NA, EU, JP, Other Region Sales by Rating

The table below summarizes the average global, NA, EU, JP, and other region sales for games of each rating. The scatterplot visualizes each rating and the average global sales of video games of that rating. Overall, Adult Only (AO) games tend to have the highest mean sales globally, in North America, and in Europe. K-A rated games tend to have the highest mean sales in Japan, and Mature (M) games tend to have the highest mean sales in other regions. In all regions, E10+ games tend to have the lowest sales.

In [21]:
```
gsales_rating<-vg %>% group_by(Rating) %>% summarize(mean_gsales=mean(Global_S
ales)) %>% arrange(desc(mean_gsales))
NAsales_rating<-vg %>% group_by(Rating) %>% summarize(mean_NAsales=mean(NA_Sal
es)) %>% arrange(desc(mean_NAsales))
EUsales_rating<-vg %>% group_by(Rating) %>% summarize(mean_EUsales=mean(EU_Sal
es)) %>% arrange(desc(mean_EUsales))
JPsales_rating<-vg %>% group_by(Rating) %>% summarize(mean_JPsales=mean(JP_Sal
es)) %>% arrange(desc(mean_JPsales))
osales_rating<-vg %>% group_by(Rating) %>% summarize(mean_osales=mean(Other_Sa
les)) %>% arrange(desc(mean_osales))
sales_rating<-cbind(gsales_rating,NAsales_rating,EUsales_rating,JPsales_rating
,osales_rating)
sales_rating
ggplot(gsales_rating, aes(x=Rating,y=mean_gsales))+
geom_point()+
labs(title="Average Global Sales of Video Games by Rating")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Rating	mean_gsales	Rating	mean_NAsales	Rating	mean_EUsales	Rating	mean_JPsa
AO	1.95000000	AO	1.2600000	AO	0.61000000	K-A	0.48666667
K-A	1.44333333	K-A	0.8533333	M	0.30964171		0.12439356
M	0.94295585	M	0.4788740	E	0.17796292	T	0.05113137
E	0.61059885	E	0.3240441	T	0.14421817	E	0.04963919
T	0.50469436	T	0.2565856	E10+	0.13276056	M	0.04110045
E10+	0.46183803	E10+	0.2488169		0.09068252	E10+	0.02830986
	0.42121584	EC	0.1912500	K-A	0.09000000	AO	0.00000000
EC	0.22250000		0.1835515	RP	0.02666667	EC	0.00000000
RP	0.02666667	RP	0.0000000	EC	0.01375000	RP	0.00000000

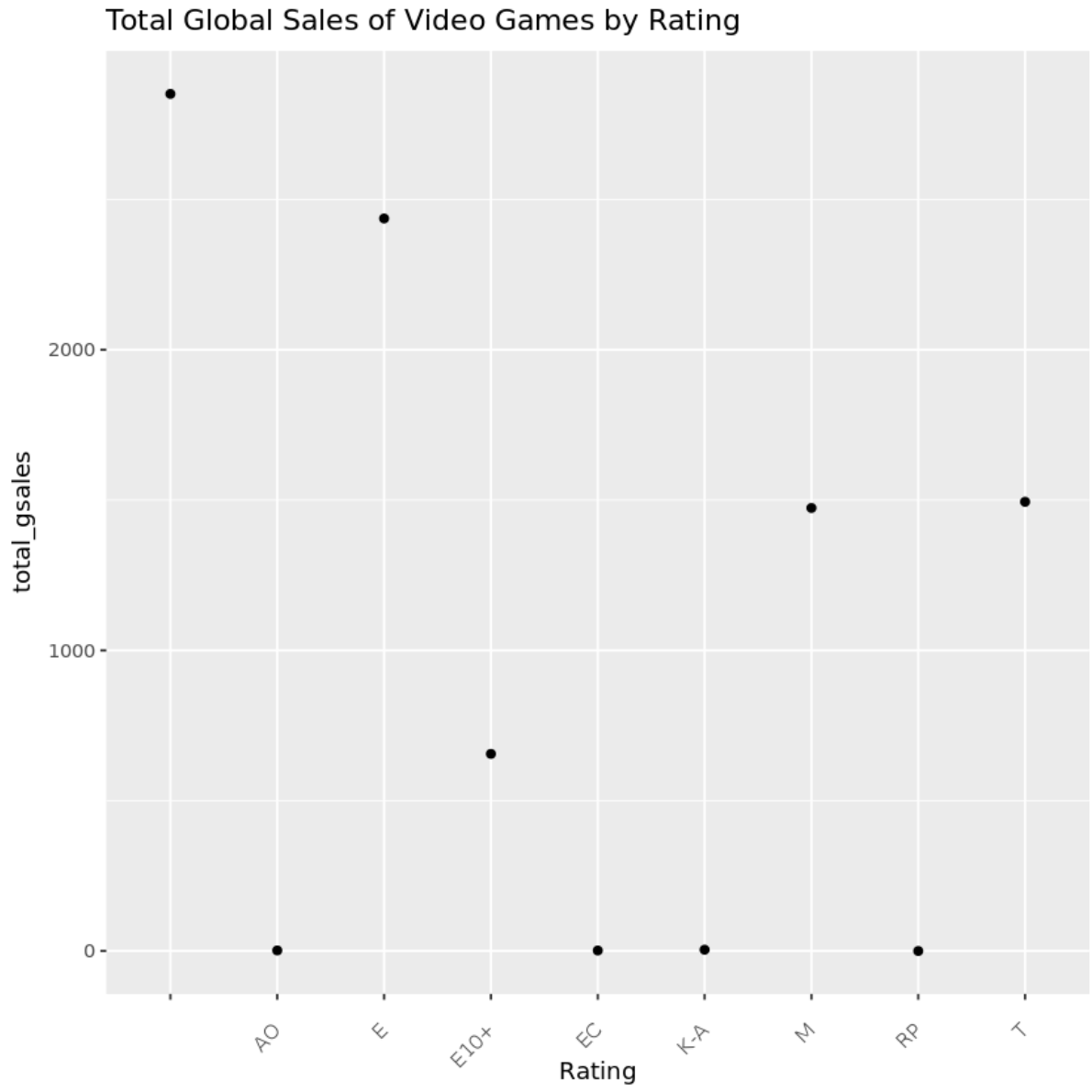


Total Global, NA, EU, JP, Other Region Sales of Video Games by Rating

The table below summarizes the total global, NA, EU, JP, and other region sales for games of each rating. The scatterplot visualizes each rating and the total global sales of video games of that rating. Overall, games rated Everyone (E) have sold the most in all regions.

In [22]:
```
gsales_rating<-vg %>% group_by(Rating) %>% summarize(total_gsales=sum(Global_S
ales)) %>% arrange(desc(total_gsales))
NAsales_rating<-vg %>% group_by(Rating) %>% summarize(total_NAsales=sum(NA_Sal
es)) %>% arrange(desc(total_NAsales))
EUsales_rating<-vg %>% group_by(Rating) %>% summarize(total_EUsales=sum(EU_Sal
es)) %>% arrange(desc(total_EUsales))
JPsales_rating<-vg %>% group_by(Rating) %>% summarize(total_JPsales=sum(JP_Sal
es)) %>% arrange(desc(total_JPsales))
osales_rating<-vg %>% group_by(Rating) %>% summarize(total_osales=sum(Other_Sa
les)) %>% arrange(desc(total_osales))
sales_rating<-cbind(gsales_rating,NAsales_rating,EUsales_rating,JPsales_rating
,osales_rating)
sales_rating
ggplot(gsales_rating, aes(x=Rating,y=total_gsales))+
geom_point()+
labs(title="Total Global Sales of Video Games by Rating")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Rating	total_gsales	Rating	total_NAsales	Rating	total_EUsales	Rating	total_JPsales
	2851.21	E	1293.26	E	710.25		842.02
E	2436.90		1242.46		613.83	E	198.11
T	1494.40	T	759.75	M	483.97	T	151.40
M	1473.84	M	748.48	T	427.03	M	64.24
E10+	655.81	E10+	353.32	E10+	188.52	E10+	40.20
K-A	4.33	K-A	2.56	AO	0.61	K-A	1.46
AO	1.95	EC	1.53	K-A	0.27	AO	0.00
EC	1.78	AO	1.26	EC	0.11	EC	0.00
RP	0.08	RP	0.00	RP	0.08	RP	0.00



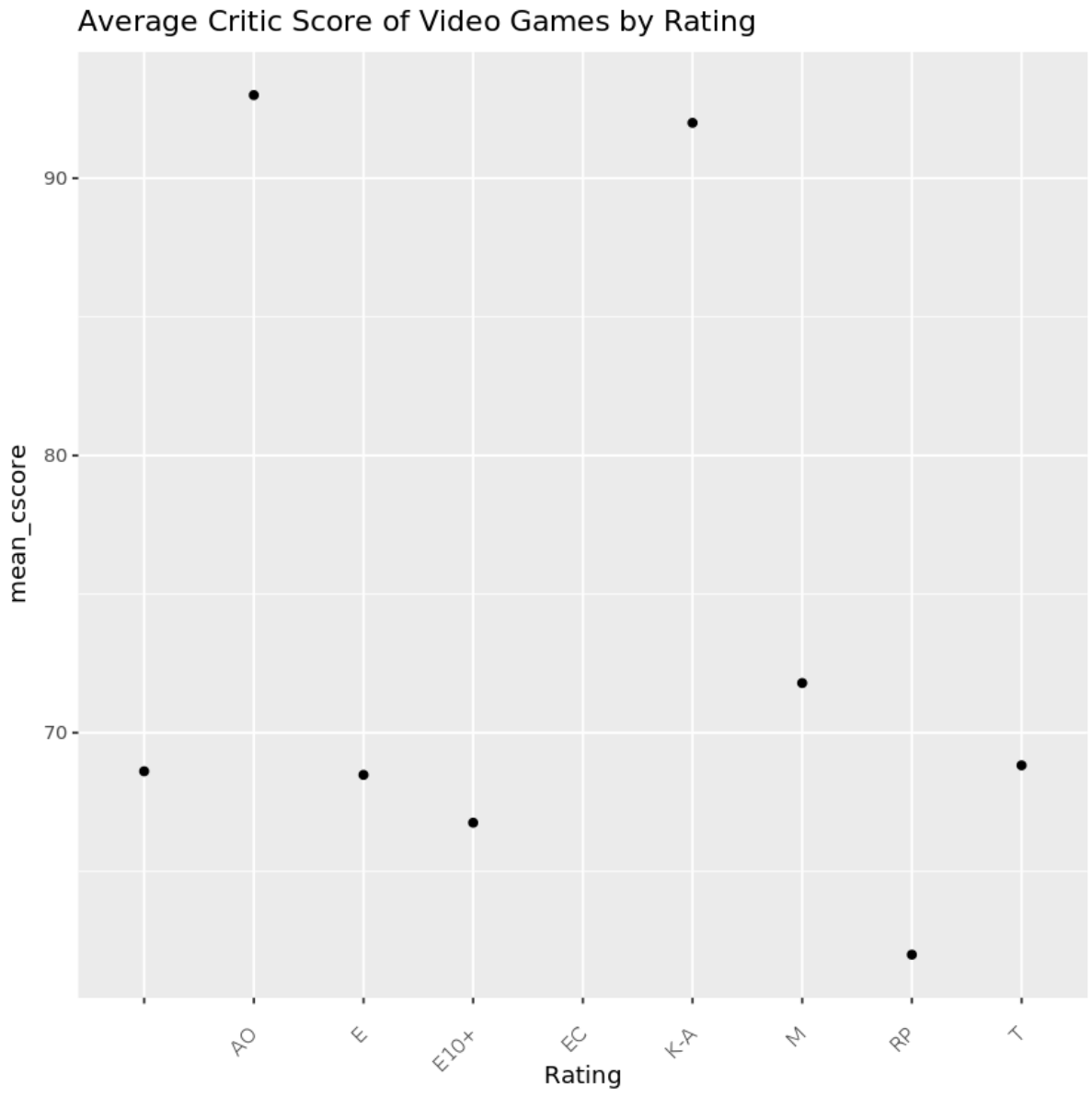
Average Critic Score, User Score of Video Games by Rating

The table below summarizes the average critic score and user score for games of each rating, with N/A critic score and user score values being removed. The scatterplot visualizes each rating and the average critic score of video games of that rating. Overall, Adult Only (AO) rated games tend to get the highest scores from critics and users, while Everyone (E) Everyone 10+ (E10) games tend to get the lowest rating from critics and users.

In [23]:
```
critic_score_rating<-vg %>% group_by(Rating) %>% summarize(mean_cscore=mean(Cr
itic_Score,na.rm=TRUE)) %>% arrange(desc(mean_cscore))
user_score_rating<-vg %>% group_by(Rating) %>% summarize(mean_uscore=mean(User
_Score,na.rm=TRUE)) %>% arrange(desc(mean_uscore))
cbind(critic_score_rating,user_score_rating)
ggplot(critic_score_rating, aes(x=Rating,y=mean_cscore))+
geom_point()+
labs(title="Average Critic Score of Video Games by Rating")+
theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

Rating	mean_cscore	Rating	mean_uscore
AO	93.00000	AO	85.00000
K-A	92.00000	K-A	80.66667
M	71.79703	T	71.86763
T	68.82841	M	70.59881
	68.61446	E	69.53011
E	68.48469	EC	69.00000
E10+	66.75939		68.65116
RP	62.00000	E10+	67.55321
EC	NaN	RP	52.00000

```
Warning message:
“Removed 1 rows containing missing values (geom_point).”
```



Top 10 Publishers Who Have Published the Most Games

The table below shows the top 10 publishers' who've published the most games. The graph visualizes the number of video games published by each publisher in the top 10. Electronic Arts, Activision, Namco Bandai Games, Ubisoft, and Konami Digital Entertainment have published the most video games.

In [24]:
```
vg_publisher<-vg %>% group_by(Publisher) %>% tally %>% arrange(desc(n))
vg_publisher[1:10,]
```

Publisher	n
Electronic Arts	1356
Activision	985
Namco Bandai Games	939
Ubisoft	933
Konami Digital Entertainment	834
THQ	715
Nintendo	706
Sony Computer Entertainment	687
Sega	638
Take-Two Interactive	422

Top 10 Publishers Who Have the Highest Average Global, NA, EU, JP, and Other Region Sales

The table below shows the top 10 publishers' who have the highest mean Global, EU, JP, and Other Region sales. The graph visualizes the top 10 publishers' who have the highest mean Global sales. Video games published by Palcom tend to have the highest mean sales globally and in North America. Video games published by Red Orb tend to have the highest mean sales in Europe, video games published by Enix Corporation tend to have the highest mean sales in Japan, and video games published by Hello Games tend to have the highest mean sales in other regions.

In [25]:
```
gsales_publisher<-vg %>% group_by(Publisher) %>% summarize(mean_gsales=mean(Gl
obal_Sales)) %>% arrange(desc(mean_gsales))
NAsales_publisher<-vg %>% group_by(Publisher) %>% summarize(mean_NAsales=mean(
NA_Sales)) %>% arrange(desc(mean_NAsales))
EUsales_publisher<-vg %>% group_by(Publisher) %>% summarize(mean_EUsales=mean(
EU_Sales)) %>% arrange(desc(mean_EUsales))
JPsales_publisher<-vg %>% group_by(Publisher) %>% summarize(mean_JPsales=mean(
JP_Sales)) %>% arrange(desc(mean_JPsales))
osales_publisher<-vg %>% group_by(Publisher) %>% summarize(mean_osales=mean(Ot
her_Sales)) %>% arrange(desc(mean_osales))
sales_publisher<-cbind(gsales_publisher,NAsales_publisher,EUsales_publisher,JP
sales_publisher,osales_publisher)
sales_publisher[1:10,]
```

Publisher	mean_gsales	Publisher	mean_NAsales	Publisher	mean_EUsales
Palcom	4.170000	Palcom	3.3800000	Red Orb	1.8050000
Red Orb	2.620000	Arena Entertainment	1.8500000	Hello Games	0.7600000
Nintendo	2.533725	RedOctane	1.6100000	Valve	0.6000000
Arena Entertainment	2.360000	Westwood Studios	1.5500000	Nintendo	0.5934986
UEP Systems	2.250000	UEP Systems	1.5200000	Sony Computer Entertainment Europe	0.5786667
RedOctane	2.170000	Nintendo	1.1571813	Black Label Games	0.4600000
Hello Games	1.700000	Imagic	1.1275000	UEP Systems	0.4600000
Valve	1.700000	Valve	0.8300000	Arena Entertainment	0.4400000
Sony Computer Entertainment Europe	1.558000	Microsoft Game Studios	0.8242408	Maxis	0.4400000
Westwood Studios	1.550000	Quelle	0.8100000	Palcom	0.4400000

Top 10 Publishers Who Have the Highest Total Global, NA, EU, JP, and Other Region Sales

The table below shows the top 10 publishers' who have the highest total Global, EU, JP, and Other Region sales. The graph visualizes the top 10 publishers' who have the highest total Global sales. Video games published by Nintendo have sold the most globally, in North America, in Europe, and in Japan, while video games published by Electronic Arts have sold the most in other regions.

In [26]:
```
gsales_publisher<-vg %>% group_by(Publisher) %>% summarize(total_gsales=sum(Gl
obal_Sales)) %>% arrange(desc(total_gsales))
NAsales_publisher<-vg %>% group_by(Publisher) %>% summarize(total_NAsales=sum(
NA_Sales)) %>% arrange(desc(total_NAsales))
EUsales_publisher<-vg %>% group_by(Publisher) %>% summarize(total_EUsales=sum(
EU_Sales)) %>% arrange(desc(total_EUsales))
JPsales_publisher<-vg %>% group_by(Publisher) %>% summarize(total_JPsales=sum(
JP_Sales)) %>% arrange(desc(total_JPsales))
osales_publisher<-vg %>% group_by(Publisher) %>% summarize(total_osales=sum(Ot
her_Sales)) %>% arrange(desc(total_osales))
sales_publisher<-cbind(gsales_publisher,NAsales_publisher,EUsales_publisher,JP
sales_publisher,osales_publisher)
sales_publisher[1:10,]
```

Publisher	total_gsales	Publisher	total_NAsales	Publisher	total_EUsales	Pu
Nintendo	1788.81	Nintendo	816.97	Nintendo	419.01	Nir
Electronic Arts	1116.96	Electronic Arts	599.50	Electronic Arts	373.91	Na Ba Ga
Activision	731.16	Activision	432.59	Activision	215.90	Ko Dig En
Sony Computer Entertainment	606.48	Sony Computer Entertainment	266.17	Sony Computer Entertainment	186.56	So Co En
Ubisoft	471.61	Ubisoft	252.74	Ubisoft	161.99	Ca
Take-Two Interactive	403.82	Take-Two Interactive	222.94	Take-Two Interactive	119.25	Se
THQ	338.44	THQ	207.72	THQ	93.78	Sq
Konami Digital Entertainment	282.39	Microsoft Game Studios	157.43	Sega	80.66	Sq
Sega	270.35	Atari	109.84	Konami Digital Entertainment	68.98	En Co
Namco Bandai Games	254.62	Sega	108.61	Microsoft Game Studios	68.64	Te

Top 10 Publishers Who Have the Highest Critic Scores, User Scores

The table below summarizes the top 10 publishers who have the highest mean critic scores, with N/A critic score and user score values being removed. The scatterplot visualizes the top 10 publishers who have the highest mean critic scores. Overall, video games published by Valve, Blue Byte, 2D Boy, Graphism Entertainment, and Number None tend to get the highest scores from critics. Video games published by Square, Sunflowers, ValuSoft, Kadokawa Shoten, and SNK tend to get the highest scores from users.

In [27]:
```
critic_score_publisher<-vg %>% group_by(Publisher) %>% summarize(mean_cscore=m
ean(Critic_Score,na.rm=TRUE)) %>% arrange(desc(mean_cscore))
t10_cscore<-critic_score_publisher[1:10,]
user_score_publisher<-vg %>% group_by(Publisher) %>% summarize(mean_uscore=mea
n(User_Score,na.rm=TRUE)) %>% arrange(desc(mean_uscore))
t10_uscore<-user_score_publisher[1:10,]
cbind(t10_cscore,t10_uscore)
```

Publisher	mean_cscore	Publisher	mean_uscore
Valve	95.00	Square	90.0
Valve Software	93.00	Sunflowers	90.0
Blue Byte	91.00	ValuSoft	90.0
2D Boy	90.00	Kadokawa Shoten	89.0
Graphsim Entertainment	90.00	SNK	87.5
Number None	90.00	Blue Byte	87.0
Psygnosis	89.75	Havas Interactive	87.0
Square	89.00	Kool Kizz	87.0
Devolver Digital	88.00	Nihon Falcom Corporation	87.0
Havas Interactive	88.00	Telegames	86.5

Top 10 Publishers Who Have the Highest Critic Count, User Count

The table below summarizes the top 10 publishers who have the highest mean amount of critics and users who review their games, with N/A critic count and user count values being removed. The scatterplot visualizes the top 10 publishers who have the highest mean amount of critics who review their games. Overall, video games published by Hello Games, Maxis, Sony Computer Entertainment Europe, inXile Entertainment, and Valve Software tend to have the highest amount of critics who review them. Video games published by Hello Games, Maxis, Valve Software, id Software, and Wargaming.net tend to have the highest amount of users who review them.

In [28]:
```
critic_count_publisher<-vg %>% group_by(Publisher) %>% summarize(mean_ccount
=mean(Critic_Count,na.rm=TRUE)) %>% arrange(desc(mean_ccount))
t10_ccount<-critic_count_publisher[1:10,]
user_count_publisher<-vg %>% group_by(Publisher) %>% summarize(mean_ucount=m
ean(User_Count,na.rm=TRUE)) %>% arrange(desc(mean_ucount))
t10_ucount<-user_count_publisher[1:10,]
cbind(t10_ccount,t10_ucount)
```

Publisher	mean_ccount	Publisher	mean_ucount
Hello Games	94.00000	Hello Games	5096.000
Maxis	75.00000	Maxis	4572.000
Sony Computer Entertainment Europe	65.58333	Valve Software	3175.333
inXile Entertainment	65.00000	id Software	1796.000
Valve Software	57.66667	Wargaming.net	1466.000
Nihon Falcom Corporation	56.00000	Sony Computer Entertainment Europe	1203.667
GOA	54.00000	inXile Entertainment	1170.000
Valve	54.00000	Activision Blizzard	1154.000
Microsoft Game Studios	50.74830	Havas Interactive	1056.000
RedOctane	49.50000	SCS Software	992.000

Top 10 Developers Who Have Developed the Most Games

The table below shows the top 10 developers' who've developed the most games. The graph visualizes the number of video games developed by each developer in the top 10. Ubisoft, EA Sports, EA Canada, Konami, and Capcom have developed the most video games.

In [29]:
```
vg_developer<-vg %>% group_by(Developer) %>% tally %>% arrange(desc(n))
vg_developer[2:11,]
```

Developer	n
Ubisoft	204
EA Sports	172
EA Canada	167
Konami	162
Capcom	139
EA Tiburon	108
Electronic Arts	104
Ubisoft Montreal	101
Visual Concepts	99
Omega Force	80

Top 10 Developers Who Have the Highest Average Global, NA, EU, JP, and Other Region Sales

The table below shows the top 10 developers' who have the highest mean Global, EU, JP, and Other Region sales. The graph visualizes the top 10 developers' who have the highest Global sales. Video games developed by Good Science Studio tend to have the highest mean sales globally, in North America, and in Europe. Video games developed by Retro Studios, Entertainment Analysis & Development Division tend to have the highest mean sales in Japan, and video games developed by Polyphony Digital tend to have the highest mean sales in other regions.

In [30]:
```
gsales_developer<-vg %>% group_by(Developer) %>% summarize(mean_gsales=mean(Gl
obal_Sales)) %>% arrange(desc(mean_gsales))
NAsales_developer<-vg %>% group_by(Developer) %>% summarize(mean_NAsales=mean(
NA_Sales)) %>% arrange(desc(mean_NAsales))
EUsales_developer<-vg %>% group_by(Developer) %>% summarize(mean_EUsales=mean(
EU_Sales)) %>% arrange(desc(mean_EUsales))
JPsales_developer<-vg %>% group_by(Developer) %>% summarize(mean_JPsales=mean(
JP_Sales)) %>% arrange(desc(mean_JPsales))
osales_developer<-vg %>% group_by(Developer) %>% summarize(mean_osales=mean(Ot
her_Sales)) %>% arrange(desc(mean_osales))
sales_developer<-cbind(gsales_developer,NAsales_developer,EUsales_developer,JP
sales_developer,osales_developer)
sales_developer[1:10,]
```

Developer	mean_gsales	Developer	mean_NAsales	Developer	mean_EUsal
Good Science Studio	21.810000	Good Science Studio	15.000000	Good Science Studio	4.890000
Retro Studios, Entertainment Analysis & Development Division	12.660000	Retro Studios, Entertainment Analysis & Development Division	5.030000	Retro Studios, Entertainment Analysis & Development Division	4.020000
Infinity Ward, Sledgehammer Games	9.923333	Infinity Ward, Sledgehammer Games	4.996667	Infinity Ward, Sledgehammer Games	3.650000
Polyphony Digital	9.314286	Bungie Software	4.113333	DMA Design	3.290000
Rockstar North	8.533571	DMA Design	4.060000	Polyphony Digital	2.957143
DMA Design	8.260000	Rockstar North	3.945000	Rockstar North	2.726429
Nintendo	7.384861	Neversoft Entertainment, BudCat	3.490000	Nintendo	2.324583
Bungie Software	5.203333	Polyphony Digital	3.245714	SCEE, Zoe Mode	2.300000
Neversoft Entertainment, BudCat	4.980000	Nintendo	3.213056	Agenda	1.790000
343 Industries	4.955000	343 Industries	3.195000	Team Soho	1.770000

Top 10 Developers Who Have the Highest Total Global, NA, EU, JP, and Other Region Sales

The table below shows the top 10 developers' who have the highest total Global, EU, JP, and Other Region sales. The graph visualizes the top 10 developers' who have the highest total Global sales. Video games developed by Nintendo have sold the most in all regions.

In [31]:
```
gsales_developer<-vg %>% group_by(Developer) %>% summarize(total_gsales=sum(Gl
obal_Sales)) %>% arrange(desc(total_gsales))
NAsales_developer<-vg %>% group_by(Developer) %>% summarize(total_NAsales=sum(
NA_Sales)) %>% arrange(desc(total_NAsales))
EUsales_developer<-vg %>% group_by(Developer) %>% summarize(total_EUsales=sum(
EU_Sales)) %>% arrange(desc(total_EUsales))
JPsales_developer<-vg %>% group_by(Developer) %>% summarize(total_JPsales=sum(
JP_Sales)) %>% arrange(desc(total_JPsales))
osales_developer<-vg %>% group_by(Developer) %>% summarize(total_osales=sum(Ot
her_Sales)) %>% arrange(desc(total_osales))
sales_developer<-cbind(gsales_developer,NAsales_developer,EUsales_developer,JP
sales_developer,osales_developer)
sales_developer[1:10,]
```

Developer	total_gsales	Developer	total_NAsales	Developer	total_EUsales	Developer
	2829.46		1238.31		605.07	
Nintendo	531.71	Nintendo	231.34	Nintendo	167.37	Nintendo
EA Sports	175.38	EA Sports	96.84	EA Sports	55.51	Capcom
EA Canada	142.32	EA Tiburon	78.66	EA Canada	54.32	SquareSoft
Ubisoft	132.54	Ubisoft	76.91	Ubisoft Montreal	41.88	Square Enix
Rockstar North	119.47	EA Canada	66.20	Ubisoft	41.64	Omega Force
Capcom	115.71	Treyarch	56.87	Rockstar North	38.17	Level 5
Ubisoft Montreal	108.31	Visual Concepts	55.96	Treyarch	33.45	Namco
Treyarch	103.16	Rockstar North	55.23	Capcom	26.68	Konami
EA Tiburon	96.12	Ubisoft Montreal	51.62	Konami	26.27	Polyphony Digital

Top 10 Developers Who Have the Highest Critic Scores, User Scores

The table below summarizes the top 10 developers who have the highest mean critic and user scores for their games with N/A critic score and user score values being removed. The scatterplot visualizes the top 10 developers who have the highest critic scores. Overall, video games developed by Irrational Games, Digital Extremes, Kojima Productions, Bungie Software, and DMA Design, Rockstar North tend to get the highest scores from critics. Video games developed by Inferno Games, Infinite Dreams, Tecmo, Pax Softonica, and Telenet tend to get the highest scores from users.

In [32]:
```
critic_score_developer<-vg %>% group_by(Developer) %>% summarize(mean_cscore=m
ean(Critic_Score,na.rm=TRUE)) %>% arrange(desc(mean_cscore))
t10_cscore<-critic_score_developer[1:10,]
user_score_developer<-vg %>% group_by(Developer) %>% summarize(mean_uscore=mea
n(User_Score,na.rm=TRUE)) %>% arrange(desc(mean_uscore))
t10_uscore<-user_score_developer[1:10,]
cbind(t10_cscore,t10_uscore)
```

Developer	mean_cscore	Developer	mean_uscore
Irrational Games, 2K Marin	96.00000	Inferno Games	94.0
Digital Extremes, 2K Marin	94.00000	Infinite Dreams, Paragon 5	94.0
Kojima Productions, Moby Dick Studio	94.00000	Tecmo, Graphic Research	93.0
Bungie Software	93.66667	Pax Softonica	92.0
DMA Design, Rockstar North	93.00000	Telenet	92.0
Rockstar North	92.85714	Handheld Games	91.5
ThatGameCompany	92.00000	Activision Value	91.0
Irrational Games	91.50000	Housemarque	91.0
Valve Software	91.38462	Telegames	91.0
Headgate	91.00000	Codo Games	90.0

Top 10 Developers Who Have the Highest Critic Count, User Count

The table below summarizes the top 10 developers who have the highest mean amount of critics and who review their games, with N/A critic count and user count values being removed. The scatterplot visualizes the top 10 developers who have the highest amount of critics who review their games. Overall, video games developed by The Coalition, Hello Games, Ready at Dawn, Ruffian Games, and Bungie tend to have the highest amount of critics who review them. Video games developed by Infinity Ward, Sledgehammer Games, Hello Games, CD Projekt Red Studio, Bethesda tend to have the highest amount of users who review them.

In [33]:
```
critic_count_developer<-vg %>% group_by(Developer) %>% summarize(mean_ccount
=mean(Critic_Count,na.rm=TRUE)) %>% arrange(desc(mean_ccount))
t10_ccount<-critic_count_developer[1:10,]
user_count_developer<-vg %>% group_by(Developer) %>% summarize(mean_ucount=m
ean(User_Count,na.rm=TRUE)) %>% arrange(desc(mean_ucount))
t10_ucount<-user_count_developer[1:10,]
cbind(t10_ccount,t10_ucount)
```

Developer	mean_ccount	Developer	mean_ucount
The Coalition	96.0	Infinity Ward, Sledgehammer Games	6537.000
Hello Games	94.0	Hello Games	5096.000
Ready at Dawn, SCE Santa Monica	94.0	CD Projekt Red Studio	4682.714
Ruffian Games	91.0	Bethesda Game Studios	3854.778
Bungie	90.0	Valve Software	2999.692
Playground Games	87.5	Irrational Games, 2K Marin	2927.500
Big Blue Box	87.0	Irrational Games	2764.000
Good-Feel	83.0	LucasArts, BioWare	2748.000
PLAYGROUND, Playground Games	83.0	SCE/WWS, Media Molecule	2704.500
343 Industries	82.5	Crystal Dynamics, Nixxes Software	2673.000

K-means Clustering

Below, I am using the k-means clustering algorithm, a type of unsupervised machine learning, to cluster (group) similar video games.

In [34]:
```
head(vg)
```

Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_S
Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81
Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79
Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28
Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22
Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22

Below, I am clustering the video games based on global sales, NA sales, EU sales, JP sales, and Other Region sales.

In [35]:
```
sel<-vg[,c(-1:-5,-11:-16)]
head(sel)
```

NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
41.36	28.96	3.77	8.45	82.53
29.08	3.58	6.81	0.77	40.24
15.68	12.76	3.79	3.29	35.52
15.61	10.93	3.28	2.95	32.77
11.27	8.89	10.22	1.00	31.37
23.20	2.26	4.22	0.58	30.26

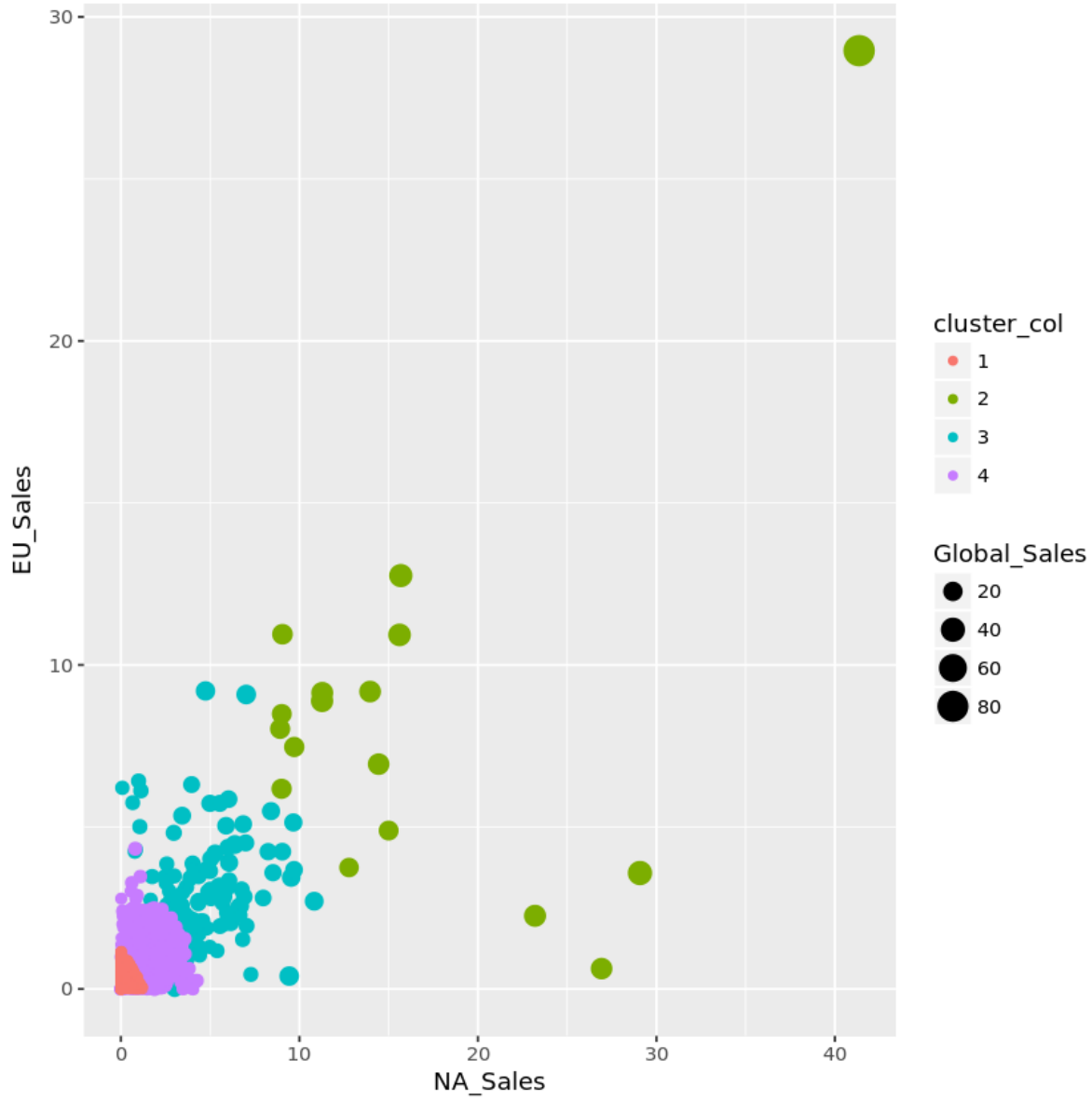
Below, I am clustering the video games using four clusters.

In [36]:
```
set.seed(4)
kclust<-kmeans(sel,4)
```

In [37]:
```
cluster_col<-kclust$cluster
cluster_col<-factor(cluster_col)
```

The scatterplot below visualizes the North America sales and Europe sales of video games in each cluster, with the size of each data point being based on the video game's global sales. Overall, the video games in cluster 2 have the highest NA, EU, and global sales, while the video games in cluster 1 have the lowest NA, EU, and global sales.

In [38]:
```
ggplot(vg, aes(x=NA_Sales,y=EU_Sales, label=row.names(vg)))+
geom_point(aes(size=Global_Sales, color=cluster_col))
```



Below, I am assigning each video game to its appropriate cluster.

In [39]:
```
vg$Cluster<-cluster_col
head(vg)
```

Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_S
Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81
Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79
Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28
Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22
Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22

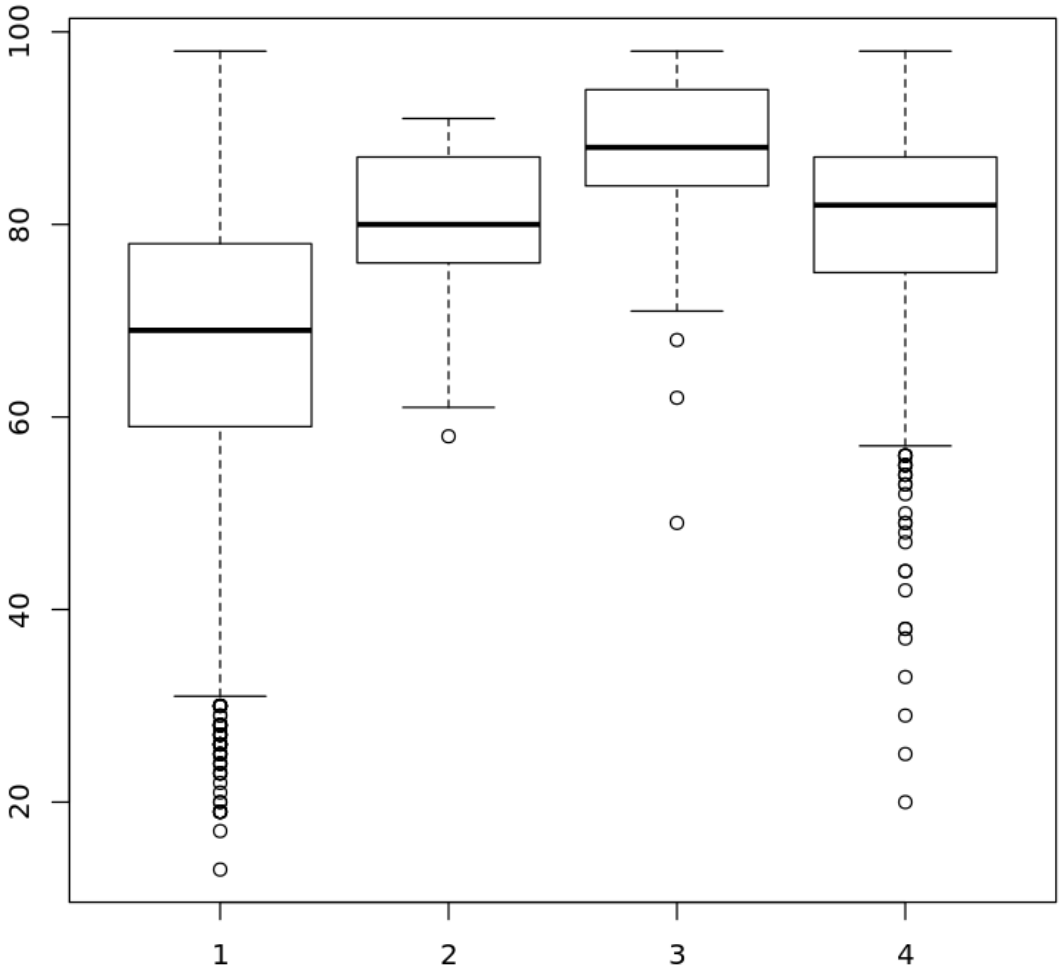
The table below summarizes the number of video games in each cluster. Cluster 1 has the most amount of video games within it, while Cluster 2 has the least amount of video games within it.

In [40]:
```
games_cluster<-vg %>% group_by(Cluster) %>% tally()
games_cluster
```

Cluster	n
1	15352
2	17
3	127
4	1223

The boxplot below shows the distribution of critic scores for the video games of each cluster. Overall, video games in cluster 3 tend to have higher critic scores, while video games in cluster 1 tend to have lower critic scores.

In [41]:
```
boxplot(vg$Critic_Score~vg$Cluster)
```



The tables below summarizes the average global sales, North America sales, Europe sales, Japan sales, Other region sales, critic score, and critic count for the video games of each cluster, with the NA values in the critic score and critic count categories being removed. Overall, video games in cluster 2 tend to have the highest sales in all regions, while video games in cluster 1 tend to have the lowest sales in all regions. Video games in cluster 4 tend to have the highest critic scores, while video games in cluster 1 tend to have the lowest critic scores. Video games in cluster 2 tend to have the highest amount of critics review them, while video games in cluster 1 tend to have the lowest amount of critics who review them. The video games with higher sales (Clusters 2, 3, and 4) tend to have the highest critic scores and highest amount of critics who review them, while the video games with lower sales (Cluster 1) tend to have the lowest critics scores and lowest amount of critics who review them. However, the video games with the highest sales, in Cluster 2, don't tend to have the highest critic scores, which shows that a video game's high sales doesn't always guarantee a high critic score for the game.

In [42]:
```
vg %>% group_by(Cluster) %>% summarize(mean(Global_Sales),mean(NA_Sales),mean(
EU_Sales),mean(JP_Sales),mean(Other_Sales),mean(Critic_Score,na.rm=TRUE),mean(
Critic_Count,na.rm=TRUE))
```

Cluster	mean(Global_Sales)	mean(NA_Sales)	mean(EU_Sales)	mean(JP_Sales)	mean(
1	0.2639168	0.1293232	0.06543838	0.04658481	0.0222
2	30.9370588	16.2517647	8.41352941	4.09823529	2.1752
3	9.6588189	4.4625197	2.98409449	1.25314961	0.9588
4	2.5478741	1.2871872	0.73430090	0.28899428	0.2374

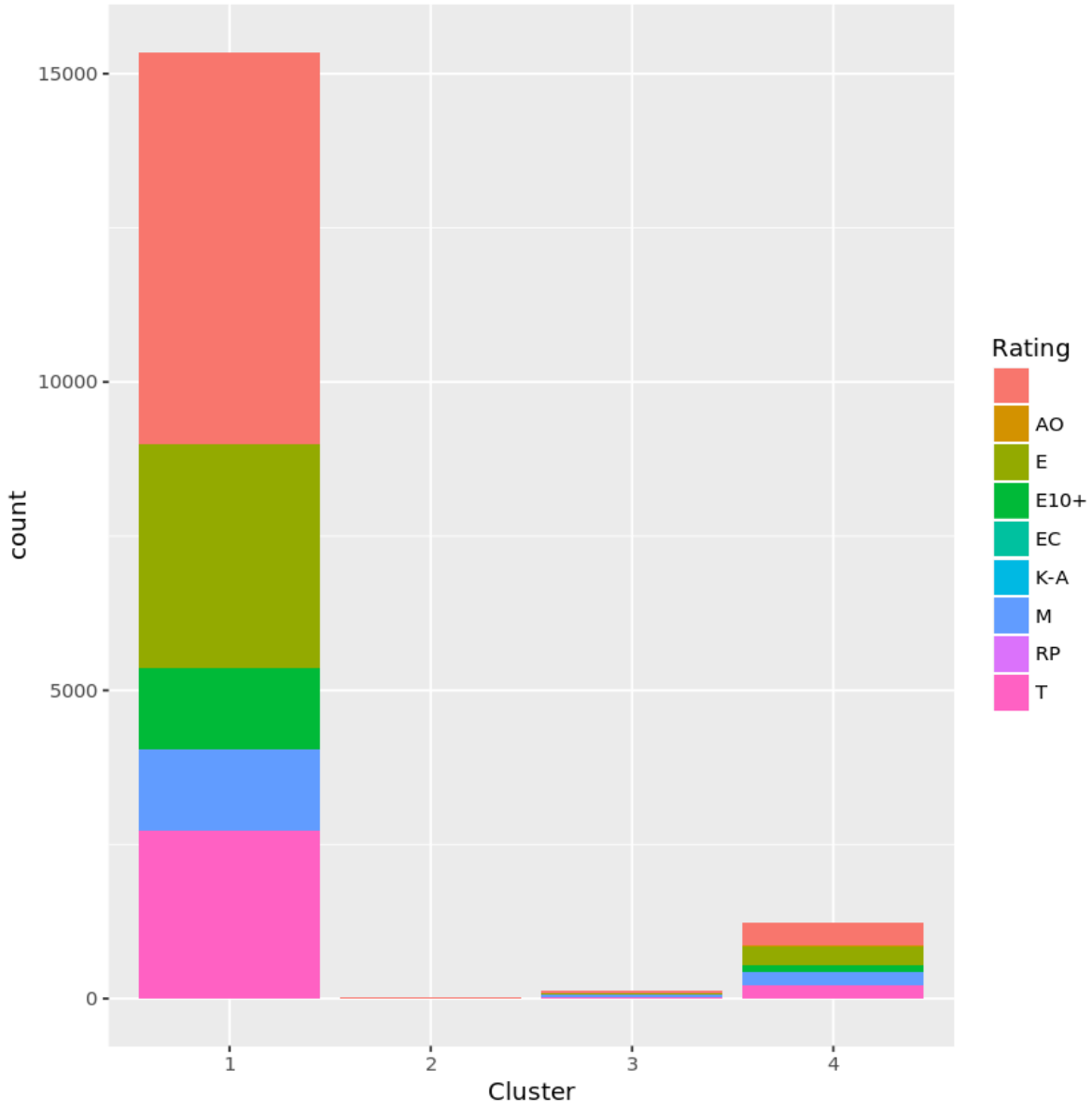
A video game's Global Sales seems to be the strongest and most important factor in determining which cluster it resides in.

Video Games by Rating per Cluster

The table below summarizes the number and proportion of video games by rating per cluster. The graph visualizes the proportion of video games by rating per cluster. Video games with the highest sales, in cluster 2, tend to have an Everyone (E) Rating.

In [43]:
```
genre_cluster<-vg %>% group_by(Cluster,Rating) %>% tally() %>% filter(Rating!=
"") %>% mutate(proportion=n/sum(n))
genre_cluster
ggplot(vg)+
geom_bar(aes(x=Cluster,fill=Rating))
```

Cluster	Rating	n	proportion
1	E	3628	0.4033351862
1	E10+	1323	0.1470817121
1	EC	8	0.0008893830
1	K-A	1	0.0001111729
1	M	1305	0.1450806003
1	RP	3	0.0003335186
1	T	2727	0.3031684269
2	E	10	1.0000000000
3	E	31	0.3406593407
3	E10+	5	0.0549450549
3	M	41	0.4505494505
3	T	14	0.1538461538
4	AO	1	0.0011709602
4	E	322	0.3770491803
4	E10+	92	0.1077283372
4	K-A	2	0.0023419204
4	M	217	0.2540983607
4	T	220	0.2576112412



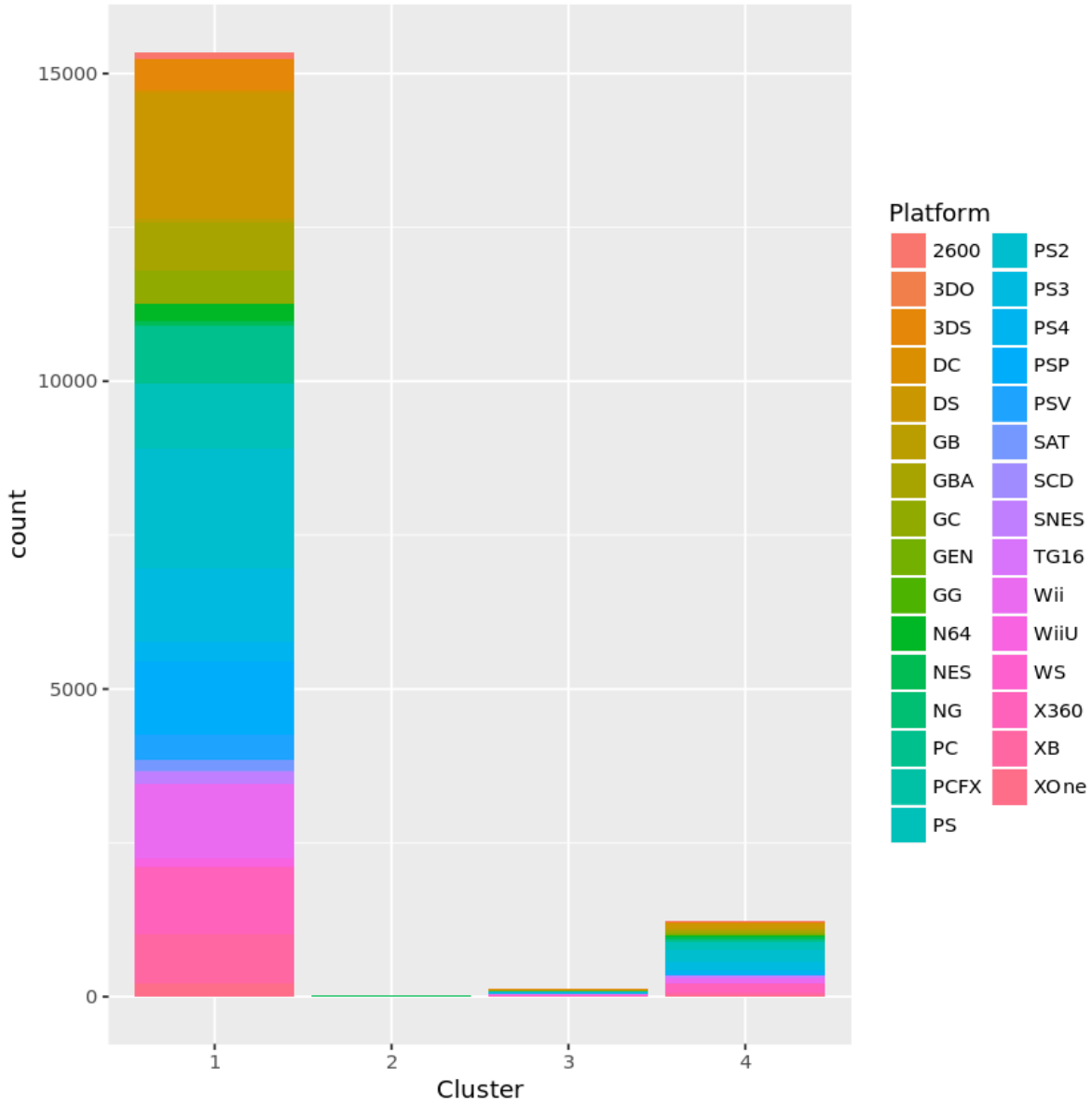
Video Games by Platform per Cluster

The table below summarizes the number and proportion of video games by platform per cluster. The graph visualizes the proportion of video games by platform per cluster. Video games with the highest sales, in cluster 2, tend to be on Nintendo platforms.

In [44]:
```
platform_cluster<-vg %>% group_by(Cluster,Platform) %>% tally() %>% filter(Pla
tform!="") %>% mutate(proportion=n/sum(n))
platform_cluster
ggplot(vg)+
geom_bar(aes(x=Cluster,fill=Platform))
```

Cluster	Platform	n	proportion
1	2600	117	7.621157e-03
1	3DO	3	1.954143e-04
1	3DS	487	3.172225e-02
1	DC	50	3.256905e-03
1	DS	2060	1.341845e-01
1	GB	61	3.973424e-03
1	GBA	779	5.074257e-02
1	GC	526	3.426264e-02
1	GEN	18	1.172486e-03
1	GG	1	6.513809e-05
1	N64	281	1.830380e-02
1	NES	57	3.712871e-03
1	NG	12	7.816571e-04
1	PC	928	6.044815e-02
1	PCFX	1	6.513809e-05
1	PS	1072	6.982804e-02
1	PS2	1936	1.261073e-01
1	PS3	1183	7.705836e-02
1	PS4	328	2.136529e-02
1	PSP	1185	7.718864e-02
1	PSV	429	2.794424e-02
1	SAT	172	1.120375e-02
1	SCD	5	3.256905e-04
1	SNES	208	1.354872e-02
1	TG16	2	1.302762e-04
1	Wii	1208	7.868682e-02
1	WiiU	135	8.793643e-03
1	WS	6	3.908286e-04
1	X360	1091	7.106566e-02
1	XB	797	5.191506e-02
⋮	⋮	⋮	⋮
3	Wii	13	0.1023622047

Cluster	Platform	n	proportion
3	WiiU	1	0.0078740157
3	X360	21	0.1653543307
3	XB	2	0.0157480315
3	XOne	1	0.0078740157
4	2600	15	0.0122649223
4	3DS	25	0.0204415372
4	DC	2	0.0016353230
4	DS	78	0.0637775961
4	GB	30	0.0245298446
4	GBA	40	0.0327064595
4	GC	27	0.0220768602
4	GEN	10	0.0081766149
4	N64	34	0.0278004906
4	NES	36	0.0294358136
4	PC	44	0.0359771055
4	PS	115	0.0940310711
4	PS2	215	0.1757972200
4	PS3	130	0.1062959935
4	PS4	58	0.0474243663
4	PSP	23	0.0188062142
4	PSV	3	0.0024529845
4	SAT	1	0.0008176615
4	SCD	1	0.0008176615
4	SNES	27	0.0220768602
4	Wii	92	0.0752248569
4	WiiU	11	0.0089942764
4	X360	149	0.1218315617
4	XB	25	0.0204415372
4	XOne	32	0.0261651676



Decision Tree Analysis

I am using Decision Tree Analysis, a supervised learning machine, to predict whether a video game's Global sales is Low or High by building a classification model using Decision Tree Analysis.

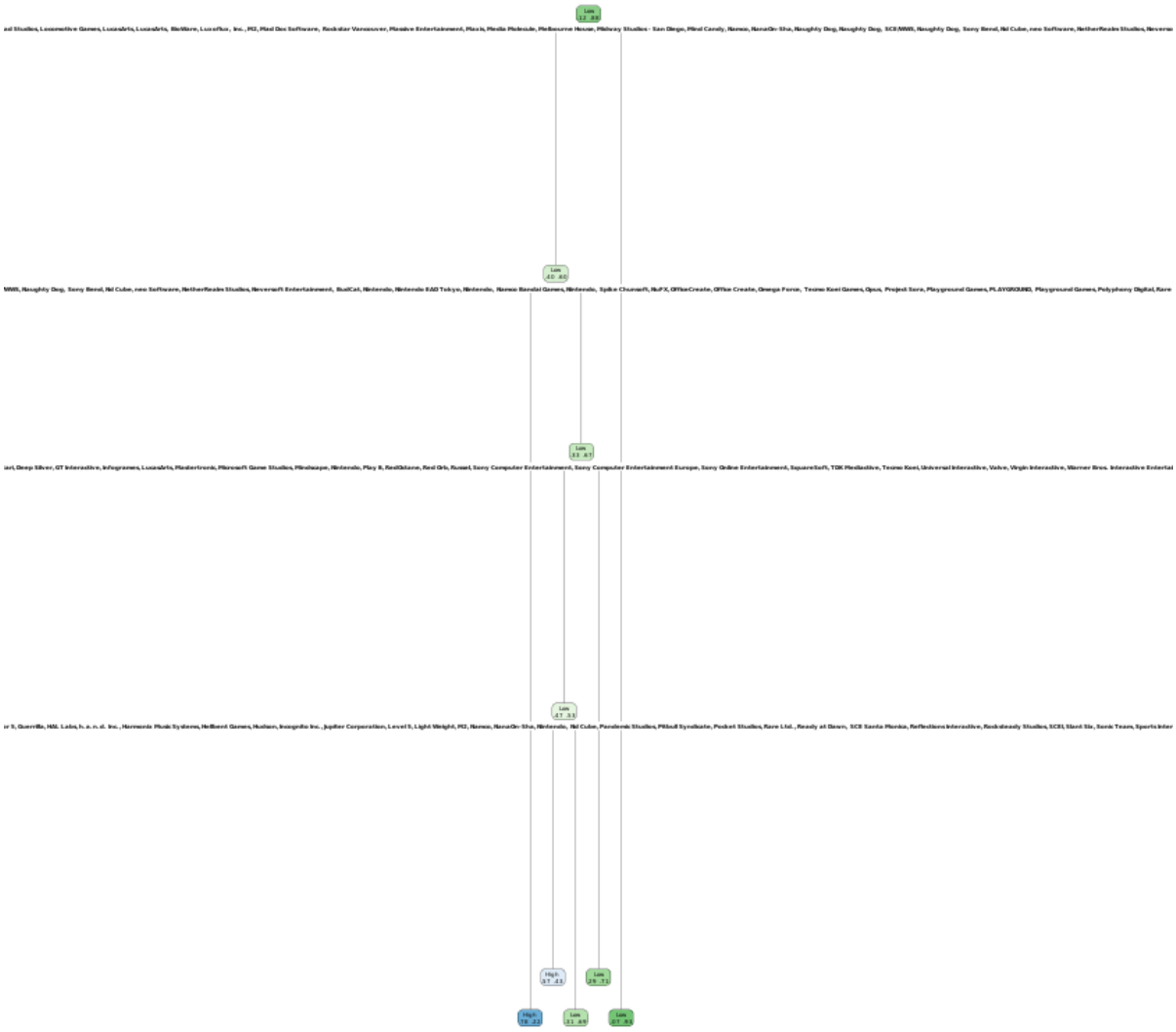
I am assigning each video game to having high or low global sales based on whether it had global sales above or equal to 1 (million). If a video game's global sales are less than 1, it has low global sales; if a video game's global sales are greater than or equal to 1, it has high global sales.

In [45]:
```
Global_Sales_Category<-ifelse(vg$Global_Sales>=1,"High","Low")
vg_t<-data.frame(vg,Global_Sales_Category)
head(vg_t)
```

Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_S
Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81
Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79
Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28
Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22
Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22

Below, I am building a decision tree that predicts if a video game's global sales are high or low based on Developer and Publisher.

In [46]:
```
decision_tree_1<-rpart(Global_Sales_Category ~ Developer+Publisher,data=vg_t)
rpart.plot(decision_tree_1,extra=4)
```



Below, I am calculating the accuracy of my model. My model has an accuracy of 89.22%. Overall, my model is very accurate and is likely to predict and classify correctly whether a video game's global sales are high or low. A video game's publisher and developer seem to be important factors in determining whether the video game's global sales are high or low (above or below 1 million sales).

In [47]:
```
pred_val<-predict(decision_tree_1, vg_t, type="class")
conf_m<-table(pred_val, vg_t$Global_Sales_Category)
conf_m
accr<-((conf_m[1,1])+(conf_m[2,2]))/dim(vg_t)[1]
round(accr*100,2)
```

```
pred_val  High   Low
    High   486   212
    Low   1590 14431
```

89.22

Results

In this section give a high-level summary of your results. If the reader only reads one section of the report, this one should be it, and it should be self-contained. In particular emphasize any results that were surprising, and if so, what your exploration of them yielded.

- Wii Sports, Grand Theft Auto 5, Super Mario Bros, Tetris, and Mario Kart Wii are the best-selling games globally. Wii Sports, Super Mairo Bros, Cuck Hunt, Tetris, and Grand Theft Auto 5 are the best-selling games in North America. Wii Sports, Grand Theft Auto 5, Mario Kart Wii, FIFA 15, and Call of Duty: Modern Warfare 3 are the best-selling games in Europe. Pokemon Red/Blue, Pokemon Gold/Silver, Super Mario Bros, New Super Mario Bros, and Pokemon Diamond/Pearl are the best-selling games in Japan. Grand Theft Auto: San Andreas, Wii Sports, Grand Theft Auto 5, Gran Turismo 4, and Call of Duty: Black Ops 2 are the best-selling games in other regions. Pokemon and Super Mario games are very popular in Japan.
- SoulCalibur, Metroid Prime, NFL 2K1, Super Mario Galaxy, and Grand Theft Auto 5 are the highest rated video games by critics. Boktai:The Sun is in Your Hand, Harvest Moon: Friends of Mineral Town, Cory in the House, Golden Sun: The Lost Age, and Karnaaj Rally are the highest rated games by users. From this, there appears to be a huge difference between the scores critics and users give toward video games.
- Overall, since 1980, the number of video games released per year has increased. However, the number of video released in a year reached its peak in 2008, and has started to decrease ever since, which I found to be surprising.
- In 1989, video games had the highest average global sales. Overall, as the years have gone by, individual video games have tended to have less global sales.
- From 1980 to 2008, the total global sales of video games continued to increase and reached its peak in 2008. Since 2008, the total global sales of video games has been decreasing, which I found to be surprising
- Overall, video games are most likely to be part of the action genre or sports genre, while least likely to be part of the puzzle or strategy genre. This implies that people tend to like playing action and sports games the most.
- Overall, role-playing and strategy games tend to get the highest scores from critics, while adventure and action games tend to get the lowest rating from critics. Role-playing and fighting games tend to get the highest scores from users, while sports and racing games tend to get the lowest rating from users, which I found to be surprising. Critics and users tend to rate role-playing games the highest, but overall tend to rate games of different genres differently.
- Overall, platform and shooter games tend to have the highest mean global sales, while adventure and strategy games tend to have the lowest mean global sales. Platform and shooter games tend to have the highest mean sales in North America, while Adventure and strategy games tend to have the lowest mean sales in North America. Shooter and platform games tend to have the highest mean sales in Europe, while Adventure and strategy games tend to have the lowest mean sales in Europe. Role-Playing and platform games tend to have the highest mean sales in Japan, while Shooter and adventure games tend to have the lowest mean sales in Japan. Shooter and racing games tend to have the highest sales in other regions, while Adventure and strategy games tend to have the lowest sales in other regions. In all regions besides Japan, Adventure games tend to get the lowest mean sales. In Japan, Shooter games tend to have the lowest mean sales, whereas in other regions, shooter games tend to get high mean sales. This shows how Japanese people tend to avoid buying shooter games, probably due to their violence, whereas in other regions, shooter games are very popular, which I found really surprising.
- Overall, action and sports games have sold the most globally, in North America, in Europe, and in other regions, while role-playing games have sold the most in Japan. Strategy games have sold the least sales globally, in North America, in Europe, and in other regions, while shooter games have sold the least in Japan. Japanese people tend to buy less-violent video games, whereas in all other regions, people tend to buy more violent games.

- Overall, the PS2 and Nintendo DS (DS) have the most video games released out of all the platforms, while the PCFX and GG have the least video games released out of all the platforms.
- Overall, Gameboy (GB) and Nintendo Entertainment (NES) games tend to have the highest mean global sales. Nintendo Entertainment System and Gameboy games tend to have the highest mean sales in North America. Gameboy and Playstation 4 games tend to have the highest mean sales in Europe. Nintendo Entertainment System and Gameboy games tend to have the highest mean sales in Japan. Playstation 4 (PS4) and Playstation 3 (PS3) games tend to have the highest mean sales in other regions. In North America, Europe, and Japan, games for Nintendo consoles tend to have the highest mean sales, while in other regions, games for Sony (Playstation) consoles tend to have the highest mean sales, which I found surprising.
- Overall, PS2 games have sold the most globally, in Europe, and in other regions, games for the Xbox 360 have sold the most in North America, and games for the Nintendo DS (DS) have sold the most in Japan.
- Overall, Sports, Racing, and Puzzle games tend to be rated Eveyone (E), Fighting games tend to be rated Teen (T), and Shooter games tend to be rated Mature (M).
- Overall, Nintendo DS (DS),Gamboy Advance (GBA), Gamecube (GB), and Wii video games tend to be rated E, which are all platforms created by Nintendo.
- Overall, video games released are most likely to have an E (Everyone)and T (Teen) rating, while video games released are least likely to have an AO (Adult Only) rating. Video games tend to have ratings that appeal to a greater audience (age group).
- Overall, Adult Only (AO) games tend to have the highest mean sales globally, in North America, and in Europe. K-A rated games tend to have the highest mean sales in Japan, and Mature (M) games tend to have the highest mean sales in other regions. In all regions, E10 games tend to have the lowest sales, which I found surprising.
- Overall, games rated Everyone (E) have sold the most in all regions.
- Overall, Adult Only (AO) rated games tend to get the highest scores from critics and users, while Everyone (E) and Everyone 10+ (E10) games tend to get the lowest rating from critics and users.
- Electronic Arts, Activision, Namco Bandai Games, Ubisoft, and Konami Digital Entertainment have published the most video games.
- Video games published by Palcom tend to have the highest sales globally and in North America. Video games published by Red Orb tend to have the highest sales in Europe, video games published by Enix Corporation tend to have the highest sales in Japan, and video games published by Hello Games tend to have the highest sales in other regions.
- Video games published by Nintendo have sold the most globally, in North America, in Europe, and in Japan, while video games published by Electronic Arts have sold the most in other regions.
- Overall, video games published by Valve, Blue Byte, 2D Boy, Graphism Entertainment, and Number None tend to get the highest scores from critics. Video games published by Square, Sunflowers, ValuSoft, Kadokawa Shoten, and SNK tend to get the highest scores from users.
- Overall, video games published by Hello Games, Maxis, Sony Computer Entertainment Europe, inXile Entertainment, and Valve Software tend to have the highest amount of critics who review them. Video games published by Hello Games, Maxis, Valve Software, id Software, and Wargaming.net tend to have the highest amount of users who review them.
- Ubisoft, EA Sports, EA Canada, Konami, and Capcom have developed the most video games.
- Video games developed by Good Science Studio tend to have the highest sales globally, in North America, and in Europe. Video games developed by Retro Studios, Entertainment Analysis & Development Division tend to have the highest sales in Japan, and video games developed by Polyphony Digital tend to have the highest sales in other regions.
- Video games developed by Nintendo have sold the most in all regions, which I found surprising.

- Overall, video games developed by Irrational Games, Digital Extremes, Kojima Productions, Bungie Software, and DMA Design, Rockstar North tend to get the highest scores from critics. Video games developed by Inferno Games, Infinite Dreams, Tecmo, Pax Softonica, and Telenet tend to get the highest scores from users.
- Overall, video games developed by The Coalition, Hello Games, Ready at Dawn, Ruffian Games, and Bungie tend to have the highest amount of critics who review them. Video games developed by Infinity Ward, Sledgehammer Games, Hello Games, CD Projekt Red Studio, Bethesda tend to have the highest amount of users who review them. #### K-means Results I clustered the video games using the Global sales, NA sales, EU sales, JP sales, and Other Region attributes of my dataset, clustering the video games into four clusters. Cluster 1 contained 15,352 video games, Cluster 2 contained 17 video games, Cluster 3 contained 127 video games, and Cluster 4 contained 1,223 video games.
- A video game's Global Sales seems to be the strongest and most important factor in determining which cluster it resides in.
- Video games in cluster 2 tended to have the highest sales in all regions, while video games in cluster 1 tended to have the lowest sales in all regions.
- Video games in cluster 4 tended to have the highest critic scores, while video games in cluster 1 tended to have the lowest critic scores.
- Video games in cluster 2 tended to have the highest amount of critics review them, while video games in cluster 1 tended to have the lowest amount of critics review them.
- The video games with higher sales (Clusters 2, 3, and 4) tended to have the highest critic scores and highest amount of critics who review them, while the video games with lower sales (Cluster 1) tended to have the lowest critics scores and lowest amount of critics who review them. However, the video games with the highest sales, in Cluster 2, didn’t tend to have the highest critic scores, showing that a video game's high sales doesn't always guarantee a high critic score for the game.
- Video games with the highest sales, in cluster 2, tended to have an Everyone (E) rating and tended to be on Nintendo platforms, which I found surprising. #### Decision Tree Analysis Results I used Decision Tree Analysis, a supervised learning machine, to predict whether a video game's Global sales is Low or High by building a classification model.I assigned each video game to having high or low global sales based on whether it had global sales above or equal to 1 (million). If a video game's global sales were less than 1 (million), it was assigned to have low global sales; if a video game's global sales were greater than or equal to 1 (million), it was assigned to have high global sales.
- I built a decision tree that predicts if a video game's global sales are high or low based on Developer and Publisher
- My classification model had an accuracy rate of 89.22% for predicting and classifying whether a video game’s global sales are high or low, being very accurate. This showed that a video game's publisher and developer seem to be important factors when determining whether a video game's global sales are high or low (above or below 1 million sales).

In [ ]: