MA678-Final Project Report

Jiaheng Li

2020/12/6

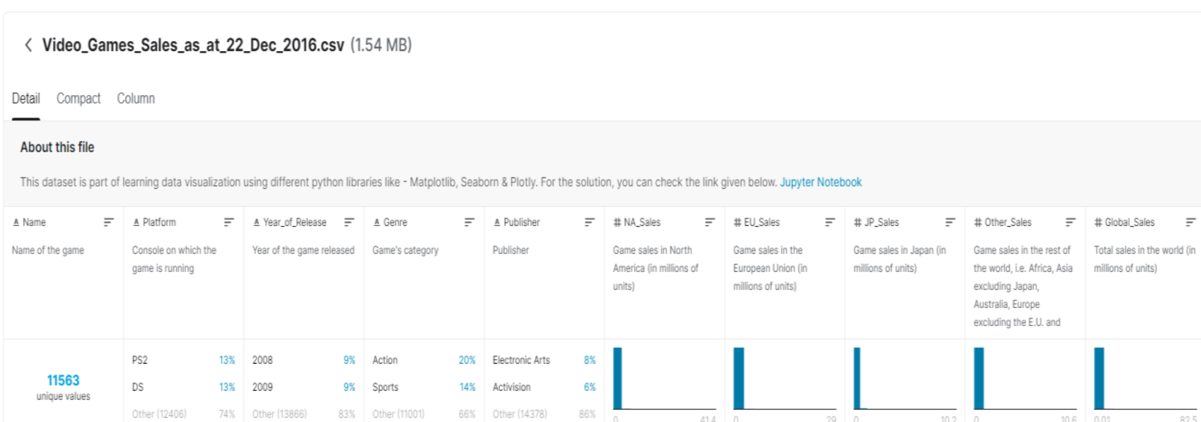
Abstract

Even since the launch of Atari's first home game console in 1977, a whole new revolution has spread in the game industry. Until today, major game manufacturers have successively launched their own game models from generation to generation. Among them, Sony and Nintendo may be the two most familiar mainstream game console manufacturers today. If we take close look of the types of modern game consoles, it is not difficult for us to find that the types of mainstream game consoles today can be divided into two categories, one is Home Video Game Console (hv), and the other is Handheld Game Console (hh). Since it is kind of difficult for us to speculate which game console type will occupy more market value in the future if we only use word-of-mouth among people and friends or read the online comments. Thus, integrating additional analysis may be a necessary step forward for judging the future game market.

Introduction

This report is using the data from Kaggle, (https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings/discussion) It shows us the lots of video game sales with ratings in each mainstream gaming market. And this report will be using this data to analyze the environment of the global game market; how would the gaming ratings affect the game sales and whether there would be a trend of Home Video Game Console (hv) or the Handheld Game Console (hh) in the world game hardware market.

Addition Information



The data is motivated by Gregory Smith's web scrape of VGChartz Video Games Sales, this data set simply extends the number of variables with another web scrape from Metacritic.

Alongside the fields: Name, Platform, YearofRelease, Genre, Publisher, NASales, EUSales, JPSales, OtherSales, Global_Sales, we have:-

Critic_score - Aggregate score compiled by Metacritic staff
Criticcount - The number of critics used in coming up with the Criticscore
User_score - Score by Metacritic's subscribers
Usercount - Number of users who gave the userscore
Developer - Party responsible for creating the game
Rating - The ESRB ratings

Method

Data Cleaning and Selection

Basic on the data set I got from Kaggle, I firstly performed a data categorizing on my data set. I categorized the platform column into two categories, 'hh' stands for the handheld game console, and the 'hv' stands for the home video game console.

Unfortunately, there are missing observations as Metacritic only covers a subset of the platforms. Also, a game may not have all the observations of the additional variables discussed above. In the end, I kept the data of the three major global game markets, which are Europe, North America and Japan. Complete cases are ~ 6,206.

Model selection and transformations

I firstly did a bunch of Exploratory Data Analysis to find whether there is a particular pattern of the game sales for each type of game console in each area. And what I learnt from it is that within the globe area, except for Japan, the sales of the 'hv' games in most situation is higher than the he sales of the 'hh' games. (See Figure 1-1, 1-2, 1-3)

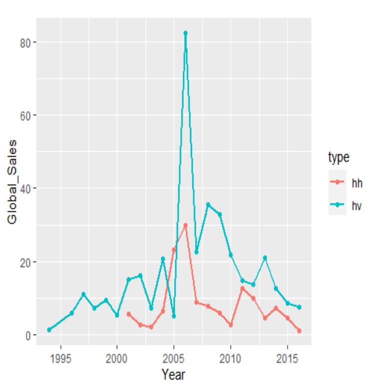


Figure 1-1

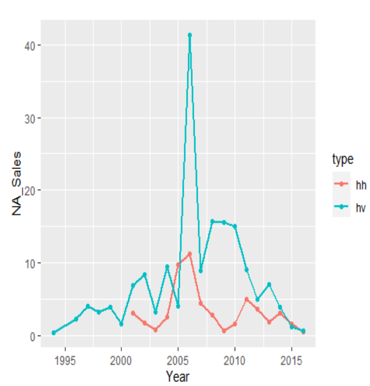


Figure 1-2

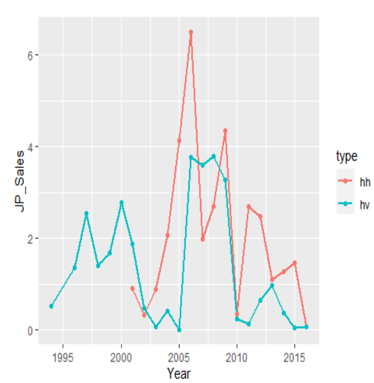


Figure 1-3

Then I tried to find out whether the different levels of the two types of the rating score would have a connection with the number of sales in each area. I finally divide each of the two types rating scores into 4 classes with different levels from low to high, and I fit these levels of ratings and the game sales data in different areas into a multivariable linear regression model to see whether they have a linear relationship.

Result

For global game sales:

1. When other independent variables remain unchanged, sales of 'hv' games are 0.2527 million higher than 'hh' games.

2. When other independent variables remain unchanged, sales of games in the 2nd class of the critic score are 0.19101 million higher than sales of games in the 1st class of the critic score; sales of games in the 3rd class of the critic score are 0.62548 million higher than sales of games in the 1st class of the critic score; sales of games in the 4th class of the critic score are 1.61993 million higher than sales of games in the 1st class of the critic score.

3. When other independent variables remain unchanged, sales of games in the 2nd class of the user score are 0.14136 million higher than sales of games in the 1st class of the user score; sales of games in the 3rd class of the user score are 0.34792 million higher than sales of games in the 1st class of the user score; sales of games in the 4th class of the user score are 0.50788 million higher than sales of games in the 1st class of the user score.

For game sales in Japan:

1. When other independent variables remain unchanged, sales of 'hv' games are 0.076786 million lower than 'hh' games.

2. When other independent variables remain unchanged, sales of games in the 2nd class of the critic score are 0.022503 million higher than sales of games in the 1st class of the critic score; sales of games in the 3rd class of the critic score are 0.063261 million higher than sales of games in the 1st class of the critic score; sales of games in the 4th class of the critic score are 0.161150 million higher than sales of games in the 1st class of the critic score.

3. When other independent variables remain unchanged, sales of games in the 2nd class of the user score are 0.02347 million higher than sales of games in the 1st class of the user score; sales of games in the 3rd class of the user score are 0.04373 million higher than sales of games in the 1st class of the user score; sales of games in the 4th class of the user score are 0.11369 million higher than sales of games in the 1st class of the user score.

For game sales in North America:

1. When other independent variables remain unchanged, sales of 'hv' games are 0.11889 million higher than 'hh' games.

2. When other independent variables remain unchanged, sales of games in the 2nd class of the critic score are 0.08997 million higher than sales of games in the 1st class of the critic score; sales of games in the 3rd class of the critic score are 0.29397 million higher than sales of games in the 1st class of the critic score; sales of games in the 4th class of the critic score are 0.80184 million higher than sales of games in the 1st class of the critic score.

3. When other independent variables remain unchanged, sales of games in the 2nd class of the user score are 0.06661 million higher than sales of games in the 1st class of the user score; sales of games in the 3rd class of the user score are 0.16506 million higher than sales of games in the 1st class of the user score; sales of games in the 4th class of the user score are 0.23886 million higher than sales of games in the 1st class of the user score.

For game sales in Europe:

1. When other independent variables remain unchanged, sales of 'hv' games are 0.07016 million higher than 'hh' games.

2. When other independent variables remain unchanged, sales of games in the 2nd class of the critic score are 0.05730 million higher than sales of games in the 1st class of the critic score; sales of games in the 3rd class of the critic score are 0.20163 million higher than sales of games in the 1st class of the critic score; sales of games in the 4th class of the critic score are 0.47692 million higher than sales of games in the 1st class of the critic score.

3. When other independent variables remain unchanged, sales of games in the 2nd class of the user score are 0.03197 million higher than sales of games in the 1st class of the user score; sales of games in the 3rd class of the user score are 0.10108 million higher than sales of games in the 1st class of the user score; sales of games in the 4th class of the user score are 0.10950 million higher than sales of games in the 1st class of the user score.

Discussion

Base on the result of the above EDAs and the modeling. It can be interpreted that:

1. The home video console is more likely the mainstream game console in the world except in Japan. And the handheld game console is more likely the mainstream game console in Japan.

2. No matter which platforms the game use, the game with a higher rating would generally achieve higher sales. This rule can be applied in both the critic score and the user score.

3. If a game gets a considerable high critic score, it will be more likely to makes much more sales than the games get a low critic score in comparison with the same situation fit in the user score. This may also infer that the critic score is a more reliable rating score to predict the game sales compare with the user score.

Bibliography

Kirubi, R. (2016, December 30). Video Game Sales with Ratings. Retrieved December 09, 2020, from https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

Video game console. (2020, December 06). Retrieved December 09, 2020, from https://en.wikipedia.org/wiki/Video_game_console

A Grammar of Data Manipulation [R package dplyr version 1.0.2]. (2020, August 18). Retrieved December 09, 2020, from https://cran.r-project.org/web/packages/dplyr/index.html

Appendix:

1.Package

```
setwd("C:\\Users\\aaron\\OneDrive\\Desktop\\678 Midterm\\final")

library(dplyr)

library(ggplot2)

library(lubridate)

library(sqldf)
```

2. Data cleaning

2.1 Categorizing

Basic on the data set I got from Kaggle, I firstly performed a data categorizing on my data set. I categorized the platform column into two categories, 'hh' stands for the handheld game console, and the 'hv' stands for the home video game console.

```
data <- read.csv("C:\\Users\\aaron\\OneDrive\\Desktop\\678 Midterm\\fin
al\\game.csv")
gdata <- data[,-c(1,4,5,15,16)]
hv <-  c("Wii","NES","X360","PS3","PS2","SNES","PS4","N64","PS","XB","2
600","XOne","GC","GEN","DC")
hh <- c("GB","DS","GBA","3DS","PSP","WiiU","PSV")
sav <- c("Wii","NES","X360","PS3","PS2","SNES","PS4","N64","PS","XB","2
600","XOne","GC","GEN","DC","GB","DS","GBA","3DS","PSP","WiiU","PSV")
rm <- c("SAT" , "SCD" , "WS" ,  "NG"  , "TG16", "3DO",  "GG"  , "PCFX")
gdata <- filter(gdata,Platform %in% sav)
```

2.2 Dealing with missing values

Unfortunately, there are missing observations as Metacritic only covers a subset of the platforms. Also, a game may not have all the observations of the additional variables discussed above.In the end, I kept the data of the three major global game markets, which are Europe, North America and Japan. Complete cases are ~ 6,206.

```
Nafun <- function(x){
  sum(is.na(x))/length(x)
}
sapply(gdata, Nafun)

##        Platform Year_of_Release         NA_Sales         EU_Sales
   JP_Sales
##       0.0000000       0.0000000       0.0000000       0.0000000
  0.0000000
```

```
##     Other_Sales     Global_Sales     Critic_Score     Critic_Count
 User_Score
##       0.0000000       0.0000000       0.5223631       0.5223631
  0.0000000
##      User_Count
##       0.5611043
```

```
gdata1 <- na.omit(gdata)
gdata1 <- subset(gdata1, gdata1$Year_of_Release!='N/A')
```

2.3 Add type column

I added a column type show whether a game is a 'hh' game or a 'hv' game.

```
gdata1$type <- ifelse(gdata1$Platform %in% hv,'hv','hh')
gdata1$type <- factor(gdata1$type)
```

2.4 Modify data type

Modifying the data type.

```
str(gdata1)
```

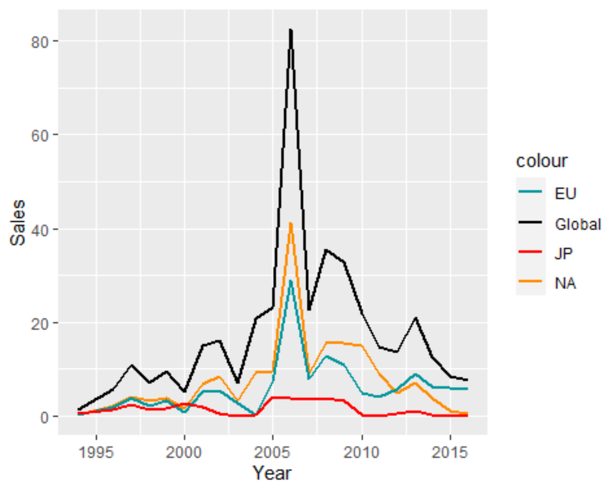
```
## 'data.frame':    6206 obs. of  12 variables:
##  $ Platform       : chr  "Wii" "Wii" "Wii" "DS" ...
##  $ Year_of_Release: chr  "2006" "2008" "2009" "2006" ...
##  $ NA_Sales       : num  41.4 15.7 15.6 11.3 14 ...
##  $ EU_Sales       : num  28.96 12.76 10.93 9.14 9.18 ...
##  $ JP_Sales       : num  3.77 3.79 3.28 6.5 2.93 4.7 4.13 3.6 0.24 2.
53 ...
##  $ Other_Sales    : num  8.45 3.29 2.95 2.88 2.84 2.24 1.9 2.15 1.69
 1.77 ...
##  $ Global_Sales   : num  82.5 35.5 32.8 29.8 28.9 ...
##  $ Critic_Score   : int  76 82 80 89 58 87 91 80 61 80 ...
##  $ Critic_Count   : int  51 73 73 65 41 80 64 63 45 33 ...
##  $ User_Score     : chr  "8" "8.3" "8" "8.5" ...
##  $ User_Count     : int  322 709 192 431 129 594 464 146 106 52 ...
##  $ type           : Factor w/ 2 levels "hh","hv": 2 2 2 1 2 2 1 2 2
2 ...
```

```
gdata1$User_Score <- as.numeric(gdata1$User_Score)
gdata1$User_Score <- gdata1$User_Score *10

gdata1$Year <- year(as.Date(gdata1$Year_of_Release, '%Y'))
```

3. Preliminary EDA

I used the line chart to show a overall game sales stats to get a broad image of the game hardware market from 1995-2016. From the chart we can see that global game sales reached a high peak in year of 2006. It may imply that there are some big events that happened in the gaming industry during that year. An interesting fact is that the PlayStation 3 (a mainstream home video game console at the time) was released in November 2006.

```
y_sales <- sqldf("select Year, Global_Sales, Na_Sales, EU_Sales, JP_Sal
es from gdata1 group by Year")

#Overall situation
p_sales <- ggplot(y_sales)+geom_line(aes(x=Year,y=Global_Sales,col='Glo
bal'),size=0.8)+
  geom_line(aes(Year,NA_Sales,col='NA'),size=0.8)+
  geom_line(aes(Year,EU_Sales,col='EU'),size=0.8)+
  geom_line(aes(Year,JP_Sales,col='JP'),size=0.8)+
   scale_color_manual(values = c('#03969D','black','red','darkorange'))
+
  ylab("Sales")
p_sales
```

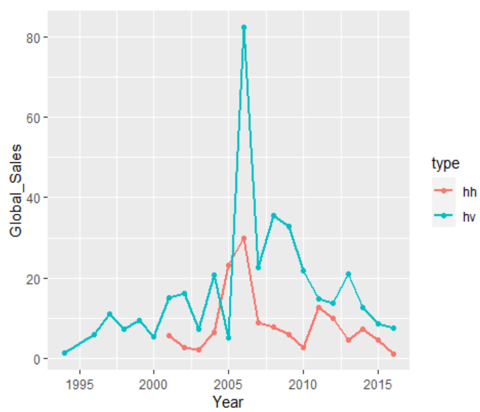


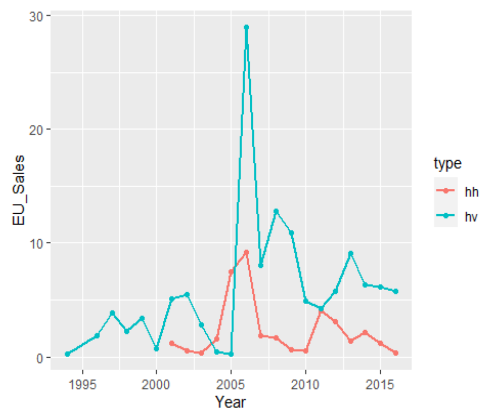
3.2 Total sales for different platform

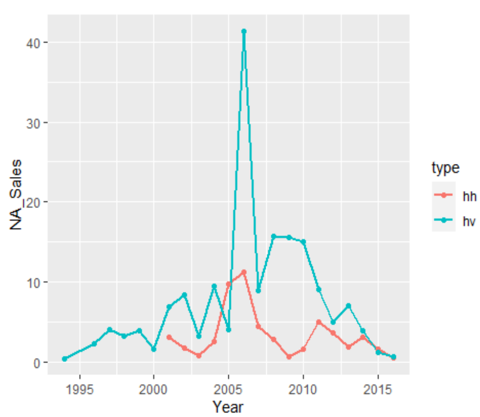
Another series of line charts to show game sales for different platform. We can see that within the globe area, except for Japan, the sales of the 'hv' games in most situation is higher than the he sales of the 'hh' games.

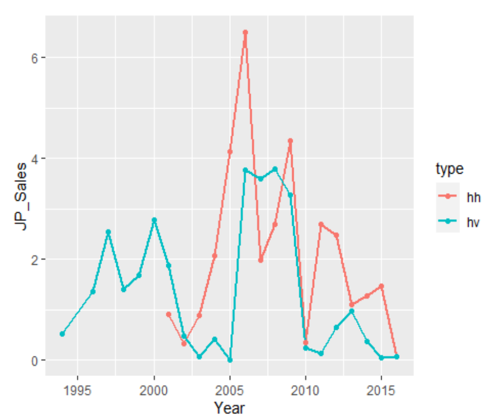
```
y_sales_type <- sqldf("select Year,type, Global_Sales, Na_Sales, EU_Sal
es, JP_Sales from gdata1 group by Year,type")
p_glob <- ggplot(y_sales_type, aes(x=Year,y= Global_Sales, group = type,
 colour=type))+
  geom_line(size=0.8)+geom_point()
p_glob

p_na <- ggplot(y_sales_type, aes(Year, NA_Sales, group = type, colour=t
ype))+
  geom_line(size=0.8)+geom_point()
p_na
```

```
p_eu <- ggplot(y_sales_type, aes(Year, EU_Sales, group = type, colour=t
ype))+
  geom_line(size=0.8)+geom_point()
p_eu

p_jp <- ggplot(y_sales_type, aes(Year, JP_Sales, group = type, colour=t
ype))+
  geom_line(size=0.8)+geom_point()
p_jp
```







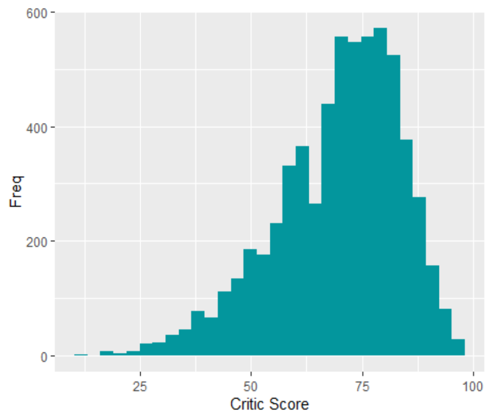


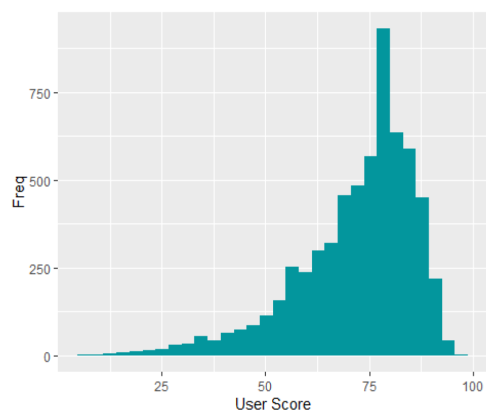
3.3 Rating score histogram

Here are the histograms shows the frequency of the 'critic_Score' and 'User_Score'. We can see that the distribution of the rating score from both sides is kind of similar.

```
ggplot(gdata1, aes(x=Critic_Score))+geom_histogram(bins = 30,fill="#039
69D")+labs(x="Critic Score", y="Freq")
```

```
ggplot(gdata1, aes(x=User_Score))+geom_histogram(bins = 30,fill="#03969
D")+labs(x="User Score", y="Freq")

summary(gdata1$Critic_Score)

summary(gdata1$User_Score)
```





4.Ranking by rating score

I set 4 ranks base on the quartile point of the rating score for both the critic and user score ratings to make it easier to analyze how different levels of the rating scores would affect the game sales.

Class 1 is the group of games with lowest rating score and the Class 4 is group of games with highest rating score.

4.1 Critic Score class

Class 1: Score <= 65 Class 2: Score <=75 Class 3: Score <=82 Class 4: Score >82

```
gdata1$CS_class <- rep(0,length(gdata1$Critic_Score))

for(i in 1:length(gdata1$Critic_Score)){
  if(gdata1$Critic_Score[i] <= 65  ){
    gdata1$CS_class[i] <- '1st class'
  }else if(gdata1$Critic_Score[i]<= 75){
    gdata1$CS_class[i] <- '2st class'
  }else if(gdata1$Critic_Score[i]<= 82){
  gdata1$CS_class[i] <- '3st class'
  }else{
    gdata1$CS_class[i] <- '4st class'
  }
}
```

4.2 User Score class

Class 1: Score <= 61 Class 2: Score <=72 Class 3: Score <=80 Class 4: Score >80

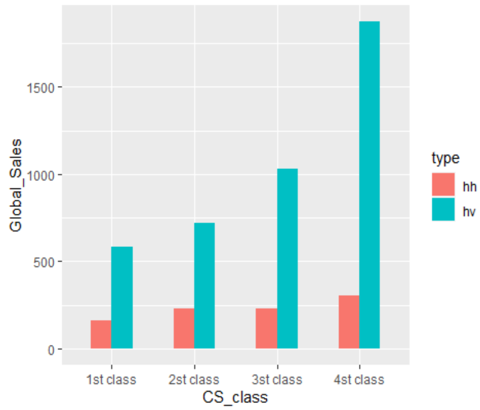
```r
gdata1$US_class <- rep(0,length(gdata1$User_Score))

for(i in 1:length(gdata1$User_Score)){
  if(gdata1$User_Score[i] <= 61  ){
    gdata1$US_class[i] <- '1st class'
  }else if(gdata1$User_Score[i]<=72.00){
    gdata1$US_class[i] <- '2st class'
  }else if(gdata1$User_Score[i]<=80.00){
  gdata1$US_class[i] <- '3st class'
  }else{
    gdata1$US_class[i] <- '4st class'
  }
}
```

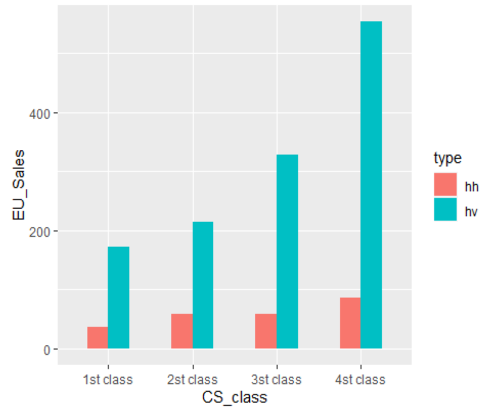
5. Ranked rating score bar charts

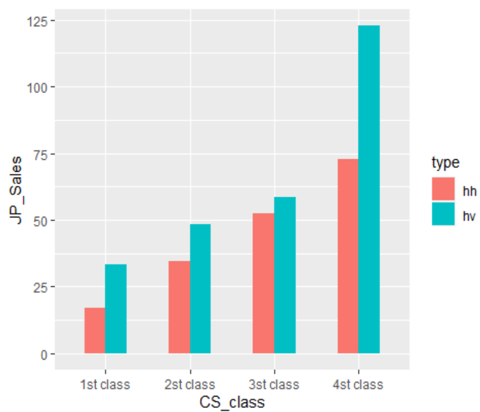
5.1 critic score and sales

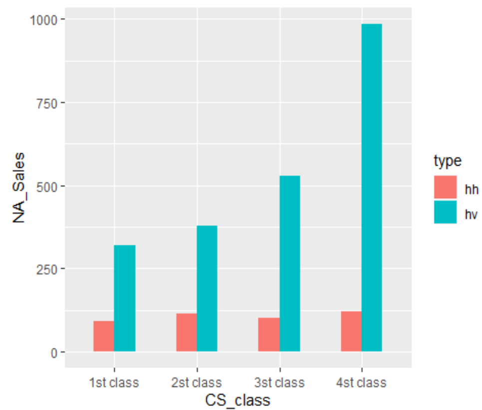
Here is a series of bar charts of the games sales fit into the all 4 ranks of critic score in different area. We can clearly see from those charts that for the 'hv' game, in all areas the higher the critic score the game has the higher sales the game would have. However, for the 'hh' game this kind of pattern only happens in Japan. Thus, It can be informed that compared with the popularity of 'hv' games in globalization, 'hh' games only have a higher market position in Japan.

```r
csco_sal <- sqldf("select type, CS_class, sum(Global_Sales) as Global_S
ales, sum(NA_Sales) as NA_Sales, sum(EU_Sales) as EU_Sales, sum(JP_Sale
s) as JP_Sales from gdata1 group by type, CS_class")

user_sal <- sqldf("select type, US_class, sum(Global_Sales) as Global_S
ales, sum(NA_Sales) as NA_Sales, sum(EU_Sales) as EU_Sales, sum(JP_Sale
s) as JP_Sales from gdata1 group by type, US_class")

ggplot(csco_sal, aes(x=CS_class, y=Global_Sales, fill=type)) + geom_bar
(stat = "identity", position = "dodge", width = 0.5)

ggplot(csco_sal, aes(x=CS_class, y=NA_Sales, fill=type)) + geom_bar(sta
t = "identity", position = "dodge", width = 0.5)

ggplot(csco_sal, aes(x=CS_class, y=EU_Sales, fill=type)) + geom_bar(sta
t = "identity", position = "dodge", width = 0.5)

ggplot(csco_sal, aes(x=CS_class, y=JP_Sales, fill=type)) + geom_bar(sta
t = "identity", position = "dodge", width = 0.5)
```







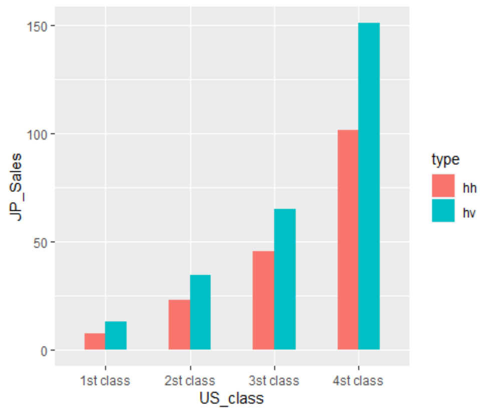


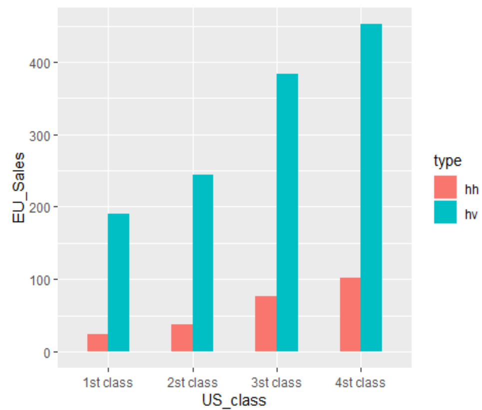
5.2 User score and sales

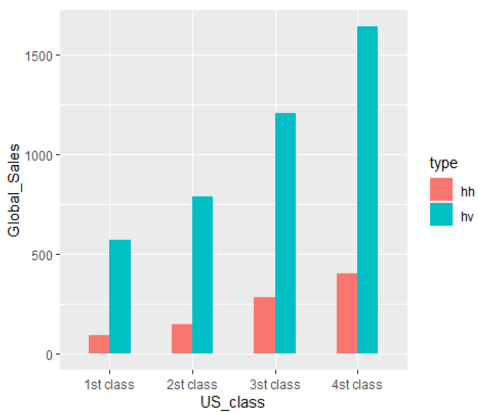
Here is a series of bar charts of the games sales fit into the all 4 ranks of user score in different area. The results is similar as the 5.1, and one thing is worth mentioning is that the user score may have a higher impact for the game sales in the US and Europe compare with the effect that the critic score brings to the game sales in these area because the fluctuation here under the 'hh' part of these area is larger than the one in 5.1.

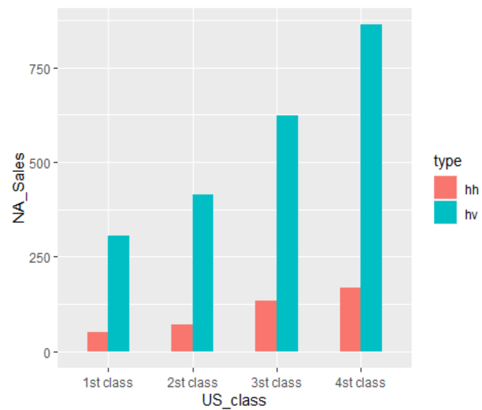
```
ggplot(user_sal, aes(x=US_class, y=Global_Sales, fill=type)) + geom_bar
(stat = "identity", position = "dodge", width = 0.5)

ggplot(user_sal, aes(x=US_class, y=NA_Sales, fill=type)) + geom_bar(sta
t = "identity", position = "dodge", width = 0.5)

ggplot(user_sal, aes(x=US_class, y=EU_Sales, fill=type)) + geom_bar(sta
t = "identity", position = "dodge", width = 0.5)
```

```
ggplot(user_sal, aes(x=US_class, y=JP_Sales, fill=type)) + geom_bar(sta
t = "identity", position = "dodge", width = 0.5)
```









6. Modeling

I use the multivariable linear regression model to analyze the relationship between the number of game sales and all 4 classes game ratings of both the critic and user score. The result I got met my expectations.

```
fit1 <- lm(Global_Sales~ type  + CS_class , data = gdata1)
summary(fit1)
fit2 <- lm(Global_Sales~ type  + US_class , data = gdata1)
summary(fit2)
fit_JP <- lm(JP_Sales~ type  + CS_class , data = gdata1)
summary(fit_JP)
fit_JP <- lm(JP_Sales~ type  + US_class , data = gdata1)
summary(fit_JP)
fit_NA <- lm(NA_Sales~ type  + CS_class , data = gdata1)
summary(fit_NA)
fit_NA <- lm(NA_Sales~ type  + US_class , data = gdata1)
summary(fit_NA)
fit_EU <- lm(EU_Sales~ type  + CS_class , data = gdata1)
```

```r
summary(fit_EU)
fit_EU <- lm(EU_Sales~ type  + US_class , data = gdata1)
summary(fit_EU)
```