GyanMatrix Challange

Saurav Kaushik

October 8, 2016

Firstly, importing the required libraries.

```
library('ggplot2')
library('randomForest')
library('ggthemes')
library('dplyr')
```

Then, importing Dataset into R IDE.

```
ign<-read.csv("ign.csv",stringsAsFactors = T)
```

1. Analyze and list the platforms with the most "Editor's Choice" awards?

The platforms with most "Editor's Choice" awards are :

```
sort(table(ign$platform[ign$editors_choice=='Y']),decreasing = T)
```

```
##
##                  PC            Xbox 360        PlayStation 3
##                 679                 366                  342
##        PlayStation 2            Wireless                 Xbox
##                 314                 268                  212
##                 Wii              iPhone          PlayStation
##                 179                 152                  124
##            GameCube         Nintendo DS     Game Boy Advance
##                 110                 100                   99
## PlayStation Portable      PlayStation 4            Dreamcast
##                  77                  62                   59
##         Nintendo 3DS            Xbox One          Nintendo 64
##                  43                  43                   42
##           Macintosh    PlayStation Vita       Game Boy Color
##                  40                  35                   32
##                iPad               Wii U          Nintendo DSi
##                  30                  24                   23
##                 NES            Super NES               N-Gage
##                   9                   9                    8
##             Genesis             Android   NeoGeo Pocket Color
##                   5                   4                    4
##                iPod               Linux        Master System
##                   3                   3                    3
##            Game Boy              Saturn         TurboGrafx-16
##                   2                   2                    2
##              Arcade              NeoGeo         Nintendo 64DD
##                   1                   1                    1
##           Pocket PC             SteamOS         TurboGrafx-CD
##                   1                   1                    1
##       Windows Phone     WonderSwan Color           Atari 2600
##                   1                   1                    0
##          Atari 5200     Commodore 64/128         Dreamcast VMU
##                   0                   0                    0
```

```
##  DVD / HD Video Game              Game.Com                Lynx
##                  0                     0                   0
##     New Nintendo 3DS                  Ouya             Sega 32X
##                  0                     0                   0
##            Sega CD               Vectrex            Web Games
##                  0                     0                   0
##    Windows Surface             WonderSwan
##                  0                     0
```
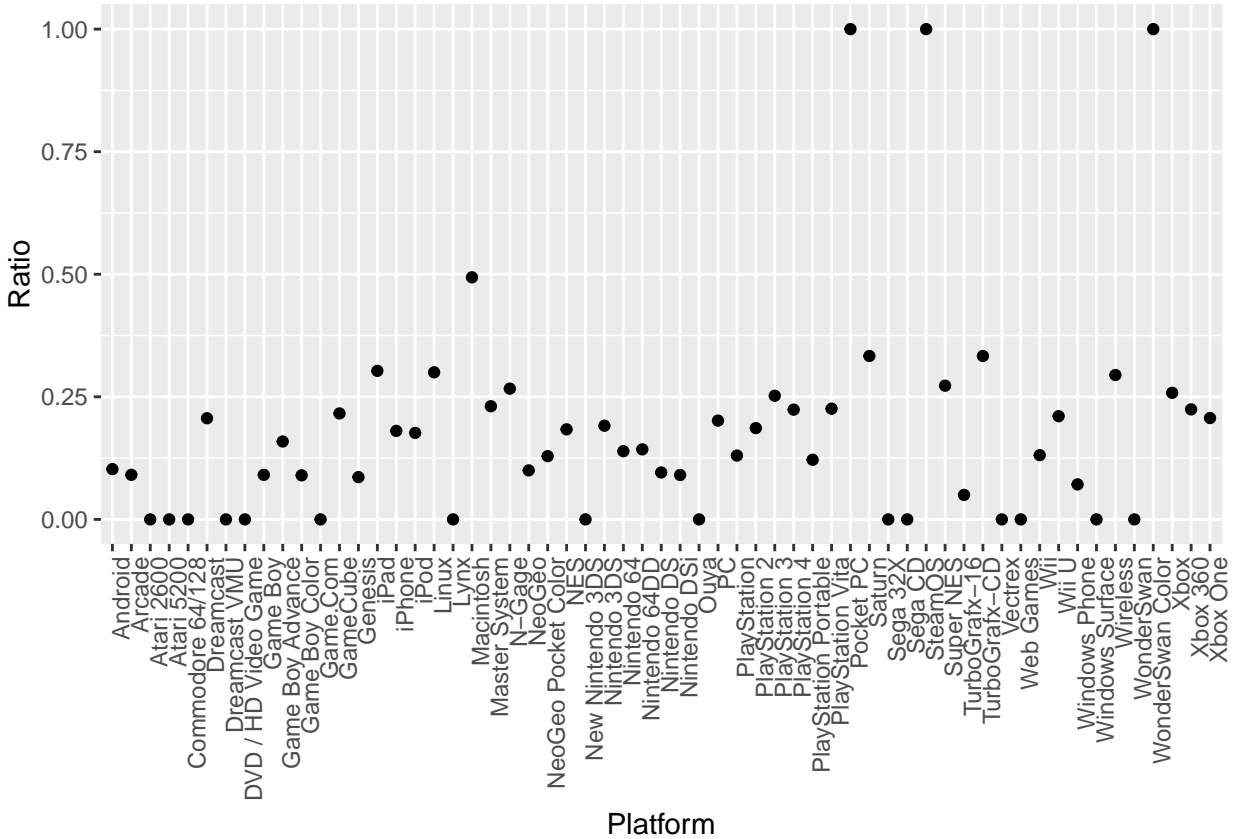
PC is the platform with the most "Editor's Choice" awards followed by Xbox 360 and PlayStation 3.

But there is a catch, if we look at the faliures in receiving Editor's Choice award, based on platforms :

```r
sort(table(ign$platform[ign$editors_choice=='N']),decreasing = T)
```

```
##
##                 PC        PlayStation 2               Xbox 360
##               2691                 1372                   1265
##                Wii        PlayStation 3            Nintendo DS
##               1187                 1014                    945
##        PlayStation               iPhone               Wireless
##                828                  690                    642
##               Xbox  PlayStation Portable        Game Boy Advance
##                609                  556                    524
##           GameCube       Game Boy Color            Nintendo 64
##                399                  324                    260
##       Nintendo DSi            Dreamcast           PlayStation 4
##                231                  227                    215
##       Nintendo 3DS             Xbox One         PlayStation Vita
##                182                  165                    120
##              Wii U                 Lynx                   iPad
##                 90                   82                     69
##            Genesis            Macintosh                    NES
##                 53                   41                     40
##       TurboGrafx-16              Android    NeoGeo Pocket Color
##                 38                   35                     27
##           Super NES               N-Gage               Game Boy
##                 24                   22                     20
##           Sega 32X                 iPod           Windows Phone
##                 16                   14                     13
##             Arcade        Master System                 NeoGeo
##                 10                   10                      9
##              Linux       Commodore 64/128          Nintendo 64DD
##                  7                    6                      6
##         Atari 2600               Saturn             WonderSwan
##                  5                    4                      4
##           Game.Com           Atari 5200        New Nintendo 3DS
##                  3                    2                      2
##        TurboGrafx-CD             Vectrex           Dreamcast VMU
##                  2                    2                      1
##  DVD / HD Video Game                Ouya                Sega CD
##                  1                    1                      1
##          Web Games      Windows Surface              Pocket PC
##                  1                    1                      0
##             SteamOS     WonderSwan Color
##                  0                    0
```

Here, its clear that PC as a platform has failed the maximum no. of times to get Editor's Choice Award as well.

This can be better understood by the following visual :

```
q<-qplot(platform,data = ign,fill=editors_choice,xlab = 'Platform',ylab='Count')
q + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```


Now, its clear that PC is by far the most preffered platform for games and therefore, we need to shift our prespective from 'maximum Edotor's Choice awards' to 'maximum Editor's Choice awards per release' for a platorm :

```
ec_yes<-as.data.frame(table(ign$platform[ign$editors_choice=='Y']))
ec_no<-as.data.frame(table(ign$platform[ign$editors_choice=='N']))
ec<-merge(ec_yes,ec_no,by ='Var1')
ec$total<-ec$Freq.x+ec$Freq.y
ec$ratio<-ec$Freq.x/ec$total

q<-qplot(Var1,ratio,data = ec,xlab = 'Platform',ylab='Ratio')
q + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Hence, there were three platforms which had the perfect ratio of Editor's Choice awards per release namely :
Pocket PC, SteamOS and WonderSwan Color, reflecting best chance to get an Editor's choice award, although
another fator to be taken into account is that there was just one release on these three platforms.

2. Does number of games by a platform in a given year have any effect on these awards?

```
ec_year_yes<-as.data.frame(table(ign$release_year[ign$editors_choice=='Y']))
ec_year_no<-as.data.frame(table(ign$release_year[ign$editors_choice=='N']))
ec_year<-merge(ec_year_yes,ec_year_no,by ='Var1')
ec_year$ratio<-ec_year$Freq.x/(ec_year$Freq.x+ec_year$Freq.y)
q<-qplot(Var1,ratio,data=ec_year,xlab = 'Year',ylab='Ratio')
q + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```
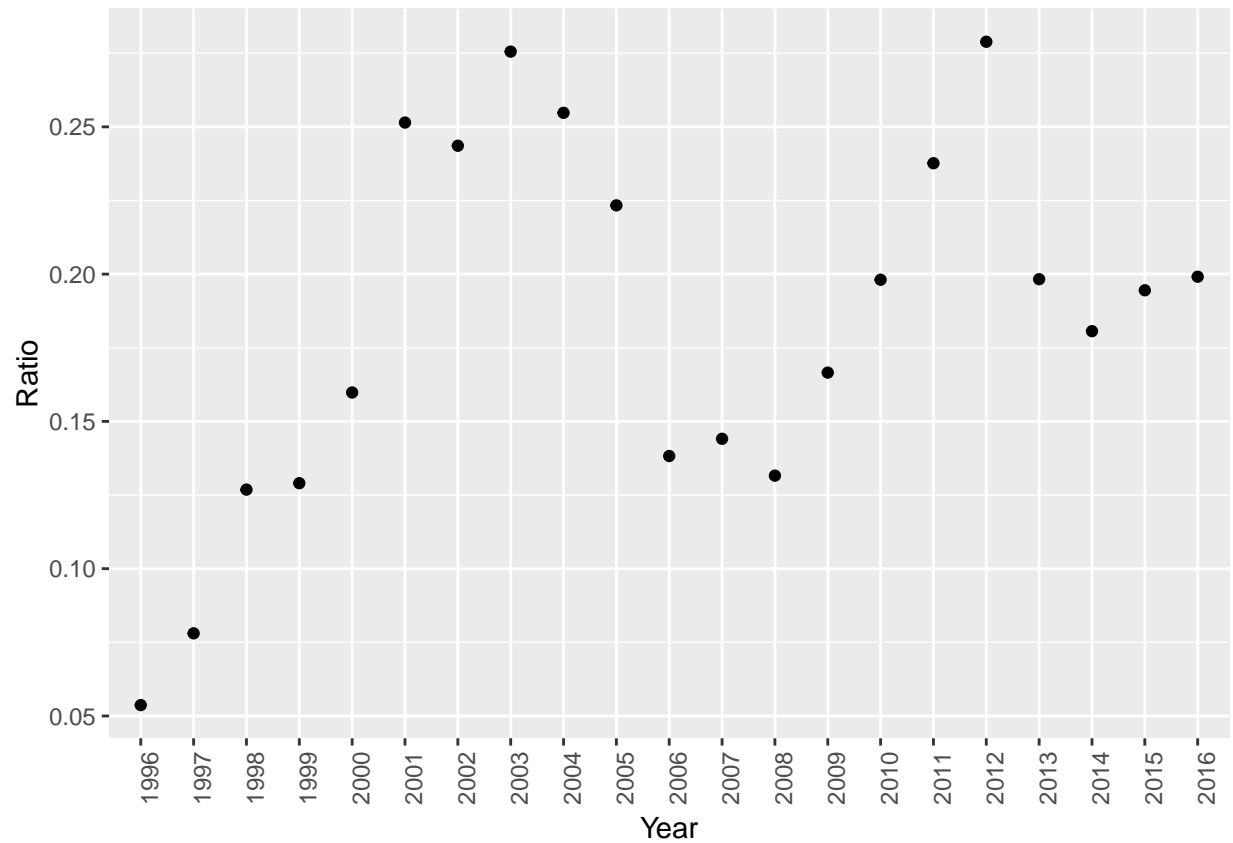
As clearly visible, the yearly ratios of total games awarded with editor's award and total games released are very fluctuating. The best ratio was found in year 2012 where 27.8% of the total releases received Editor's award followed by year 2003 with 27.5%.

Also the total no. of Editor's award given per year are represented as :

```
q<-qplot(Var1,Freq.x,data=ec_year,xlab = 'Year',ylab='Total Editors Award')
q + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```
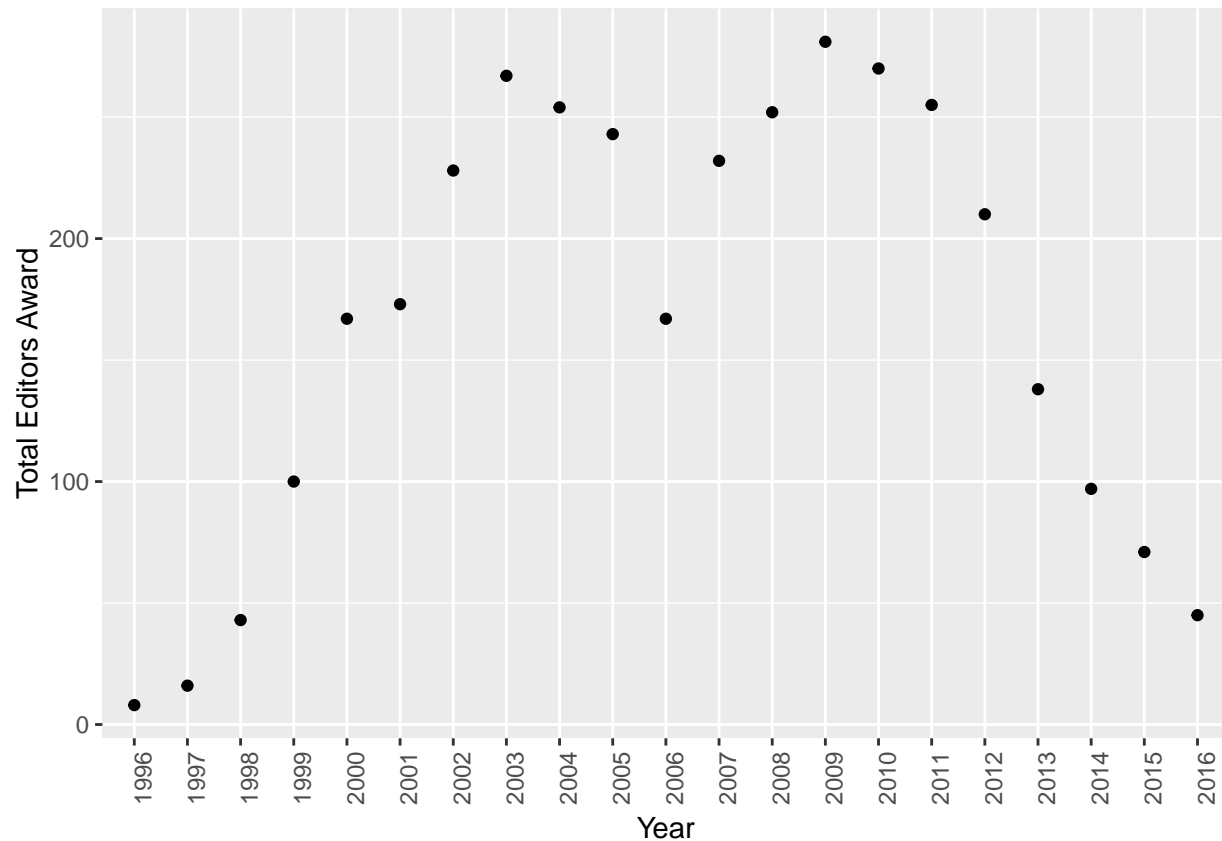
This shows that there is variation in the no. of Editor's choice awards distributed per year with the most Editor's choice awards being distributed in year 2009 followed by 2010 and 2003.

3. What is Macintosh's average award count?

Over the years, the no. of Games released for Macintosh platform are :

```
table(ign$release_year[ign$platform=='Macintosh'])
```

```
##
## 2002 2003 2012 2013 2014 2015 2016
##   37    3   13   18    6    3    1
```

Out of these releases in respective years, the no. of games released for Macintosh platform that bagged Editor's choice award are :

```
table(ign$release_year[ign$editors_choice=='Y'&ign$platform=='Macintosh'])
```

```
##
## 2002 2003 2012 2013 2014 2016
##   20    2    7    7    3    1
```

Hence, whenever a game is released for Macintosh platform (7 years out of 22 in given data), the Macintosh's average award count is :

```
(20+2+7+7+3+1)/7
```

```
## [1] 5.714286
```

For general, Macintosh's average award count considering all years (22 years) is :

```
(20+2+7+7+3+1)/22
```

```
## [1] 1.818182
```

4. What is the optimal month for releasing a game?

Finding optimal month for releasing a game based on Editor's Choice award :

```
ec_month_yes<-as.data.frame(table(ign$release_month[ign$editors_choice=='Y']))
ec_month_no<-as.data.frame(table(ign$release_month[ign$editors_choice=='N']))
ec_month<-merge(ec_month_yes,ec_month_no,by ='Var1')
ec_month$total<-ec_month$Freq.x+ec_month$Freq.y
ec_month$ratio<-ec_month$Freq.x/ec_month$total

qplot(Var1,ratio,data = ec_month,xlab = 'Month',ylab='Editors Choice Vs Total Ratio')
```
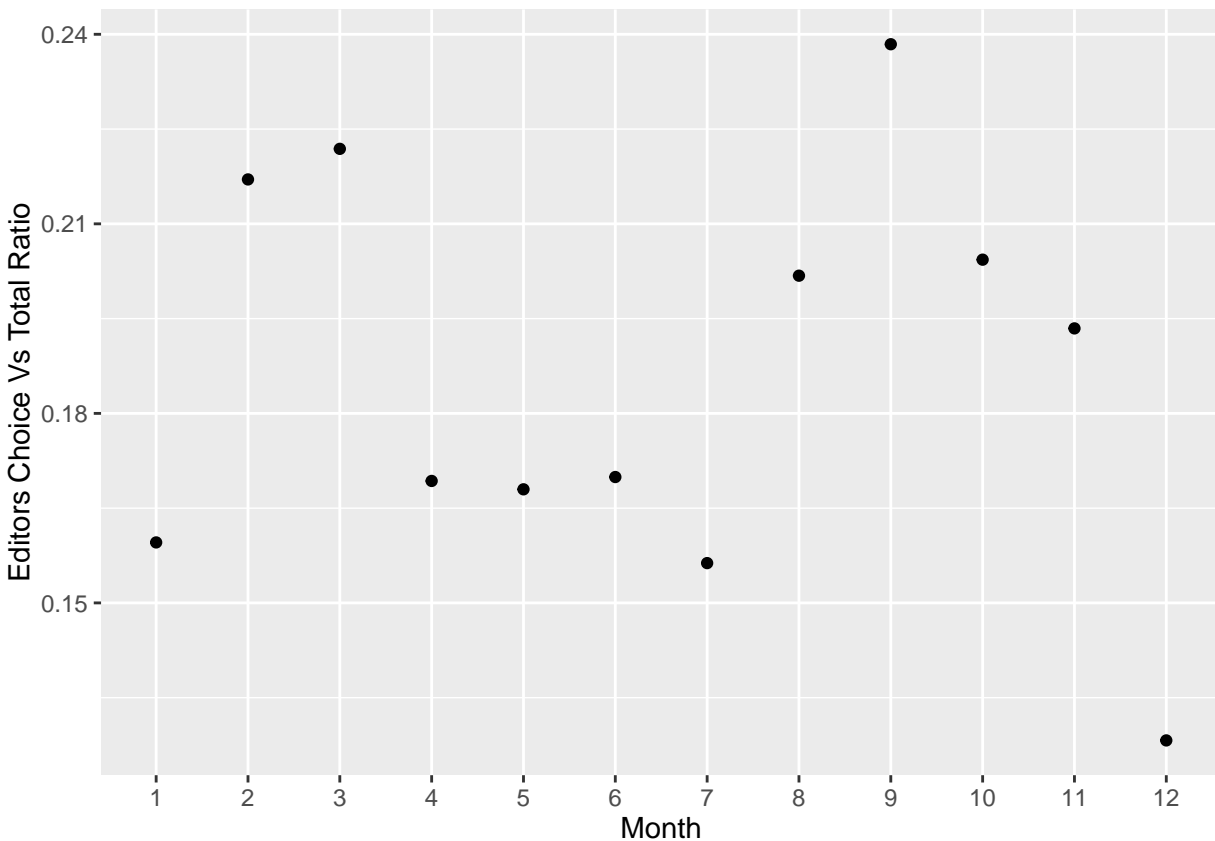

Clearly, September(9) has historically been the optimal month for releasing a game with close to 24% of games released getting Editor's choice award follwed by March(3) and Febraury(2).

Finding optimal month for releasing a game based on Scores :

```
scores<-data.frame(1:12)
for(i in 1:12){
  scores$X2[scores$X1.12==i]<- mean(ign$score[ign$release_month==i])
}

scores$X1.12<-as.factor(scores$X1.12)

qplot(X1.12,X2,data = scores,xlab = 'Month',ylab='Average Scores')
```
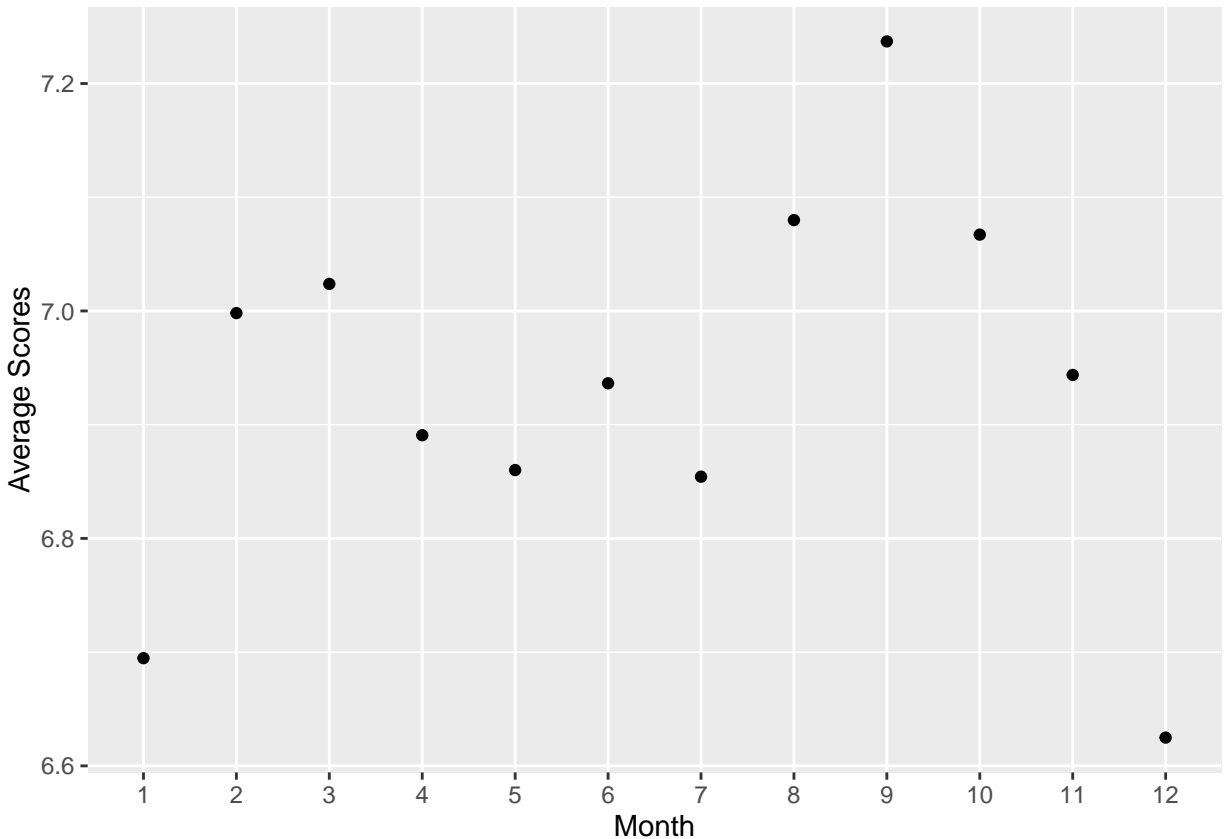
Based on monthly mean scores, September(9) is the optimal month for releasing a game with mean score of 7.23 follwed by August(8) and October(10).

5. Analyze the percentage growth in the gaming industry over the years.

Over the years, there has been a steady growth in gaming industry till 2008 with an exception of 2001 but since 2008, it has been continously on decline :

```
yearly<-as.data.frame(table(ign$release_year))
q<-qplot(Var1,Freq,data=yearly,xlab = 'Year',ylab='Total Games Released')
q + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```
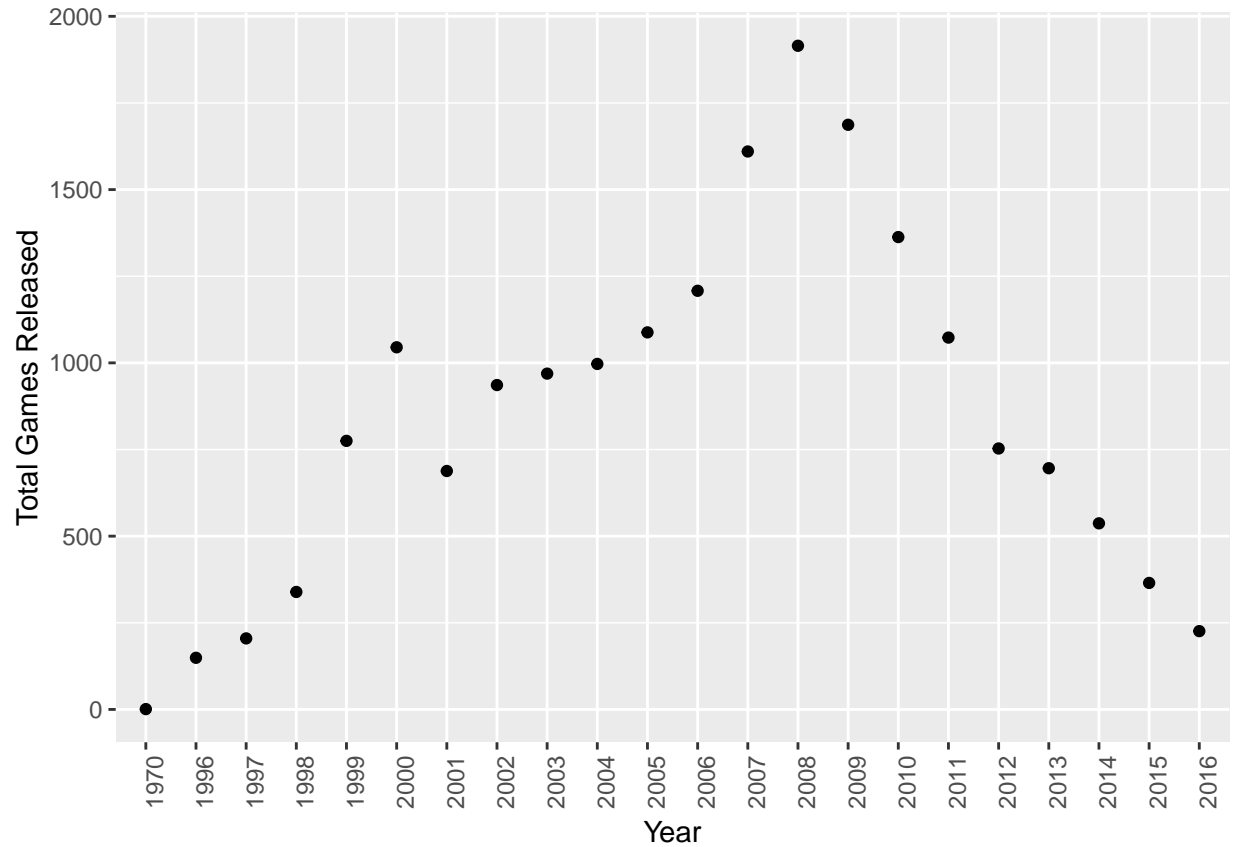
For, yearly percentage growth (i.e. comparing percentage diff. between two consecutive years. with an asumption of games released in 1995 to be same as of 1970 i.e. 1):

```r
perc<-as.data.frame(table(ign$release_year))

a<-(diff(perc$Freq))
perc<-perc[-1,]
perc$diff<-a

perc$pg<-(perc$diff/perc$Freq)*100

q<-qplot(Var1,pg,data=perc,xlab = 'Year',ylab='%age diff. from Last Year')
q + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```
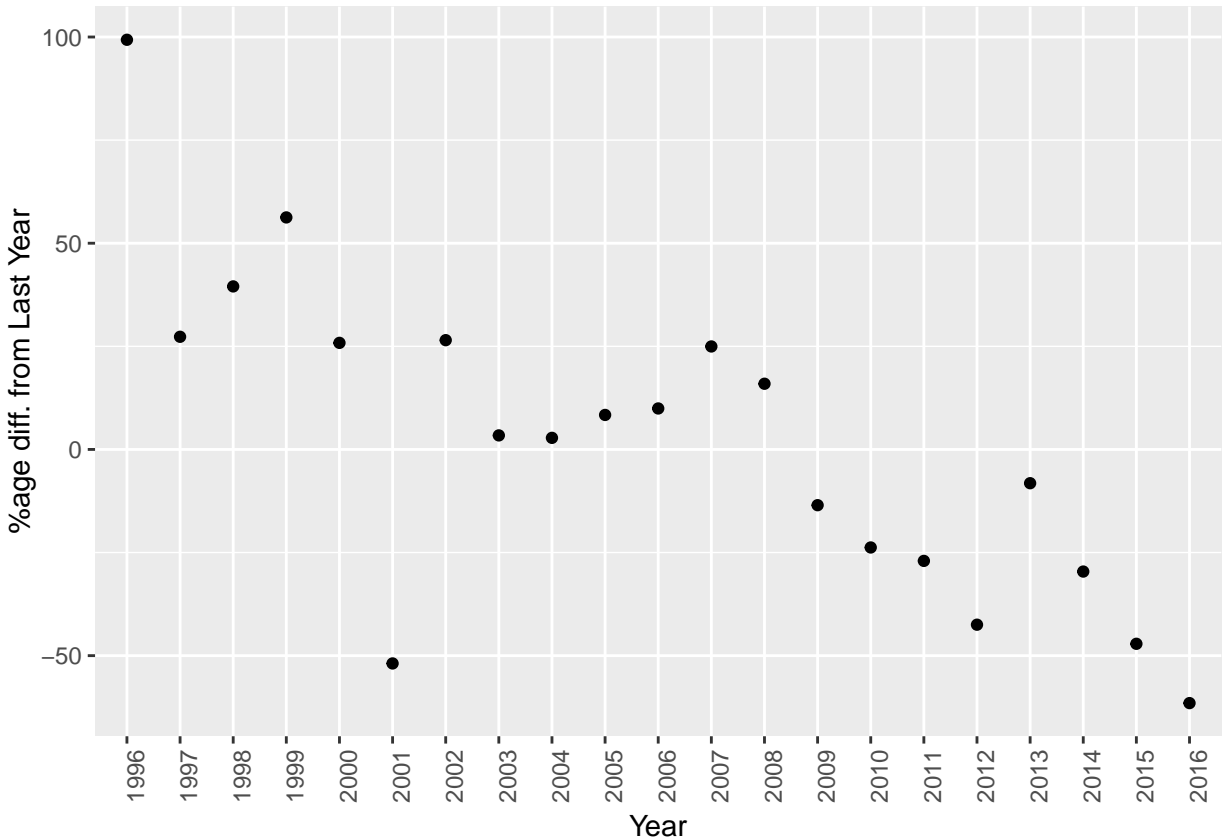
There has been a negative trend in percentage growth of gaming industry according to the given data after 2008. Also,there is severe decline in 2001 visible in the graph.

6. Use the data to build a predictive model to predict which games will get "Editor's Choice" awards in a given year?

The given problem is a classification problem. We'll be using randomForest (ensemble of decesion trees) :

```
ign$editors_choice<-ifelse(ign$editors_choice=='Y',1,0)

rf_model<-randomForest(factor(editors_choice)~score_phrase+
                    score+release_year+release_month+
                       release_day,data=ign)
```

The Out of Bag error (OOB) for the above model is given as :

```
plot(rf_model)
legend('topright', colnames(rf_model$err.rate), col=1:3, fill=1:3)
```
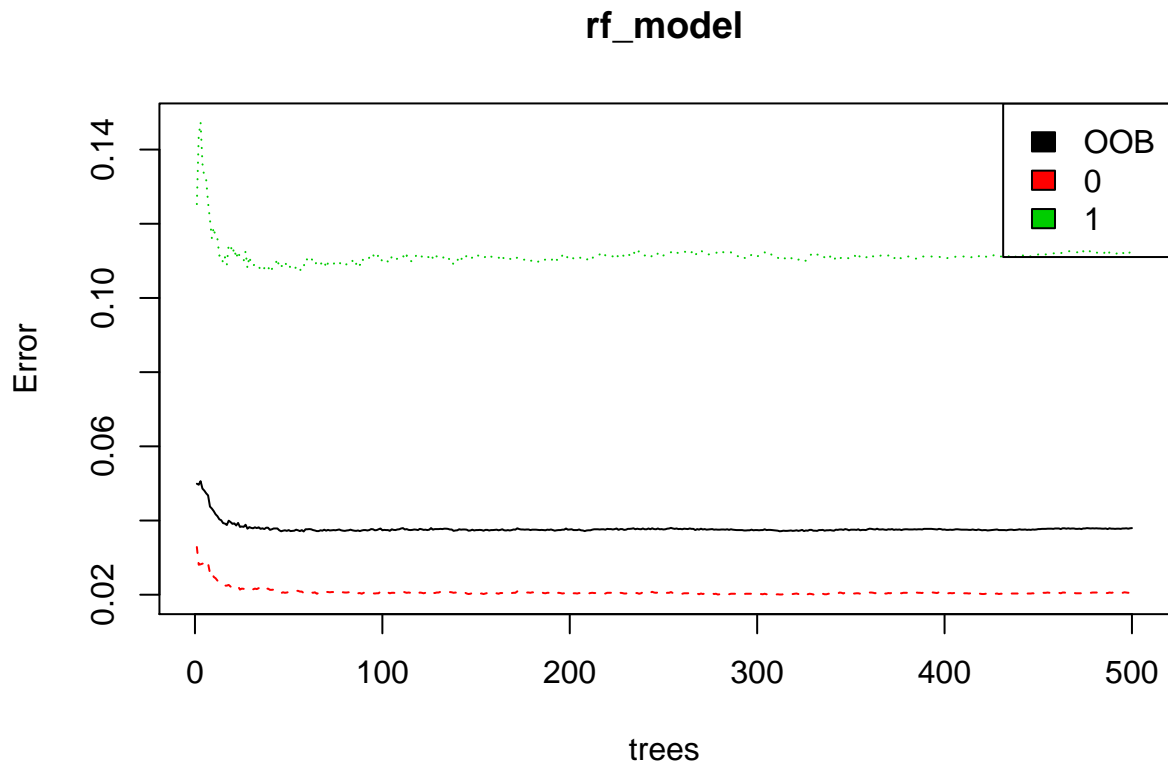
rf_model

7. What attributes are the strongest predictors of whether a game will get an award?

```
importance    <- importance(rf_model)
varImportance <- data.frame(Variables = row.names(importance),
                            Importance = round(importance[ ,'MeanDecreaseGini'],2))

# Create a rank variable based on importance
rankImportance <- varImportance %>%
  mutate(Rank = paste0('#',dense_rank(desc(Importance))))

# Use ggplot2 to visualize the relative importance of variables
ggplot(rankImportance, aes(x = reorder(Variables, Importance),
                           y = Importance, fill = Importance)) +
  geom_bar(stat='identity') +
  geom_text(aes(x = Variables, y = 0.5, label = Rank),
            hjust=0, vjust=0.55, size = 4, colour = 'red') +
  labs(x = 'Variables') +
  coord_flip() +
  theme_few()
```
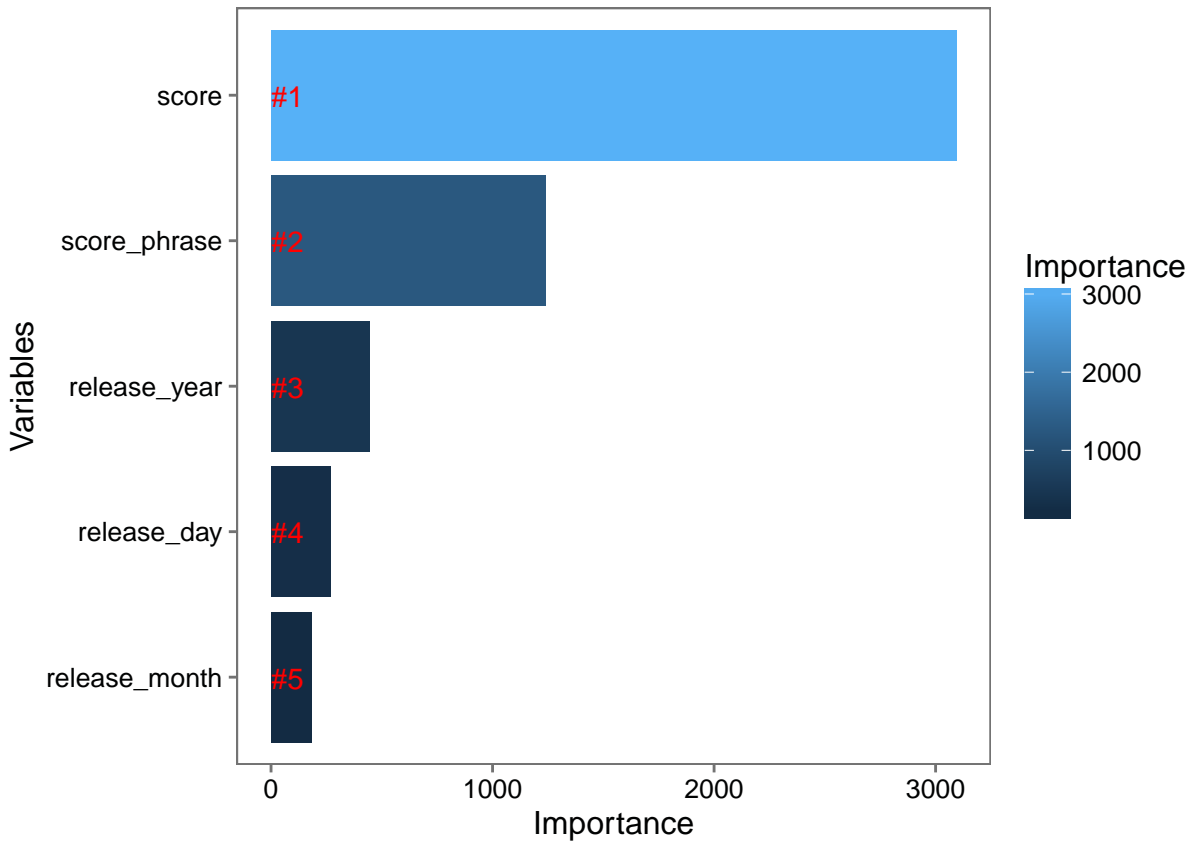
Therefore, the attributes that are the strongest predictors of whether a game will get an award are Score and Score phrase. Also, there is significant importance observed on the release year,relsease day and release month.

Intrestingly, relsease day exhibits stronger relationship with whether a game will get an award than release month attribute.