---

title: "video\_game\_clustring"

author: "RStudio"

Refrence: https://gexijin.github.io/learnR/the-game-sales-dataset.html,https://github.com/gexijin/learnR/blob/master/09-Exploring-game-sale-data-set.Rmd

date: "28/06/2021"

output: html\_document

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

################################################################################

# Generate the vgsales (video game sales) data

################################################################################

# Install libraries if they have not been installed yet

if(!require(tidyverse)) install.packages("tidyverse",

repos = "http://cran.us.r-project.org")

if(!require(caret)) install.packages("caret",

repos = "http://cran.us.r-project.org")

if(!require(data.table)) install.packages("data.table",

repos = "http://cran.us.r-project.org")

if(!require(dslabs)) install.packages("dslabs",

repos = "http://cran.us.r-project.org")

if(!require(dslabs)) install.packages("dslabs",

repos = "http://cran.us.r-project.org")

if(!require(class)) install.packages("class",

repos = "http://cran.us.r-project.org")

if(!require(knitr)) install.packages("knitr",

repos = "http://cran.us.r-project.org")

if(!require(kableExtra)) install.packages("kableExtra",

repos = "http://cran.us.r-project.org")

# Load libraries

library(tidyverse)

library(caret)

library(data.table)

library(dslabs)

library(class)

library(randomForest)

library(knitr)

library(kableExtra)

library(ggmosaic)

library(plotly)

library(reshape2)

library(viridis)

library(scales)

library(dplyr)

```{r}

tem <- read.csv("C:/Users/priya/Desktop/Game Research/Video\_Games\_Sales\_as\_at\_22\_Dec\_2016.csv")

tem <- na.omit(tem) #remove NA

```

```{r}

game <- tem %>% filter(Rating != "") %>% droplevels() #remove empty rating observations

```

The detail about the data is listed as follow:

Name: Name of the game

Platform: Console on which the game is running

Year\_of\_Release: Year of the game released

Genre: Game’s category

Publisher: Publisher

NA\_Sales: Game sales in North America (in millions of units)

EU\_Sales: Game sales in the European Union (in millions of units)

JP\_Sales: Game sales in Japan (in millions of units)

Other\_Sales: Game sales in the rest of the world, i.e. Africa, Asia excluding Japan, Australia, Europe excluding the E.U. and South America (in millions of units)

Global\_Sales: Total sales in the world (in millions of units)

Critic\_Score: Aggregate score compiled by Meta critic staff

Critic\_Count: The number of critics used in coming up with the Critic\_score

User\_Score: Score by Metacritic’s subscribers

User\_Count: Number of users who gave the user\_score

Developer: Party responsible for creating the game

Rating: The ESRB ratings: E for “Everyone”; E10+ for “Everyone 10+”; T for “Teen”; M for “Mature”; AO for “Adults Only”; RP for “Rating Pending”; K-A for kids to adults.

After downloading the data, We replaced N/A with NA first in excel and saved it as csv file and read in. We removed these observations with empty string of Rating and 6825 observations was left in our data set “game”. We noticed that there were lots of sales value with zero. To prepare these variables ready for being taken log value to make them normal or close to normal distribution, we transformed the sales into its basic units and plus 1. We also divided critic score by 10 to march the scale unit of user score.

```{r}

#by multiplying 1000000 we get the actual sale,

#adding 1 makes all sales positive which make log possible for all sales later

game$Year\_of\_Release <- as.factor(as.character(game$Year\_of\_Release))

game$NA\_Sales <- game$NA\_Sales \* 1000000 + 1

game$EU\_Sales <- game$EU\_Sales \* 1000000 + 1

game$JP\_Sales <- game$JP\_Sales \* 1000000 + 1

game$Other\_Sales <- game$Other\_Sales \* 1000000 + 1

game$Global\_Sales <- game$Global\_Sales \* 1000000 + 1

```

```{r}

# By divide by 10 to make Critic Score the same decimal as User Score

game$Critic\_Score <- as.numeric(as.character(game$Critic\_Score)) / 10

game$User\_Score <- as.numeric(as.character(game$User\_Score))

game$Critic\_Count <- as.numeric(game$Critic\_Count)

game$User\_Count <- as.numeric(game$User\_Count)

```

```{r}

# format column names

colnames(game) <- c("Name", "Platform", "Year.Release", "Genre", "Publisher", "NA.Sales", "EU.Sales", "JP.Sales", "Other.Sales", "Global.Sales", "Critic.Score", "Critic.Count", "User.Score", "User.Count", "Developer", "Rating")

head(game)

```

```{r}

str(game)

```

```{r}

summary(game)

```

Summary of these variables tells us that some of games were published in the same name; PS2 is the most popular platform; Action is the most popular Genre; Electronic Arts has the most high frequency among the publishers; Rating T and E are the two most released ratings; For these sales, though the minimums, several quantiles and medians are small, but the maximums are high, which means there are real good sale games among them; Extreme big maximum User count hints so many users scored some special games.

Our pre-analysis shows that these variables are not normally distributed, especially those sales and score counts variables. We take logs to transform these variables.

```{r}

NA.Sales.Log <- log(game$NA.Sales)

EU.Sales.Log <- log(game$EU.Sales)

JP.Sales.Log <- log(game$JP.Sales)

Other.Sales.Log <- log(game$Other.Sales)

Global.Sales.Log <- log(game$Global.Sales)

Critic.Count.Log <- log(game$Critic.Count)

User.Count.Log <- log(game$User.Count)

```

Then we combine the log variables with the original variables.

```{r}

game.log <- cbind.data.frame(NA.Sales.Log, EU.Sales.Log, JP.Sales.Log, Other.Sales.Log,

Global.Sales.Log, Critic.Count.Log, User.Count.Log)

game <- cbind.data.frame(game, game.log) # the data we use for analysis

head(game)

```

```{r}

str(game)

```

Now we plot histogram and QQ plot for the transformed data set.

```{r}

name <- colnames(game)[c(11, 13, 17:23)] # pick up the numeric columns according to the names

par(mfrow = c(5, 4)) # layout in 5 rows and 4 columns

for (i in 1:length(name)){

sub <- sample(game[name[i]][, 1], 5000)

submean <- mean(sub)

hist(sub, main = paste("Hist. of", name[i], sep = " "), xlab = name[i])

abline(v = submean, col = "blue", lwd = 1)

qqnorm(sub, main = paste("Q-Q Plot of", name[i], sep = " "))

qqline(sub)

if (i == 1) {s.t <- shapiro.test(sub)

} else {s.t <- rbind(s.t, shapiro.test(sub))

}

}

s.t <- s.t[, 1:2] # take first two columns of shapiro.test result

s.t <- cbind(name, s.t) # add variable name for the result

s.t

```

Histograms and qq plots of these original sales show no normal distribution, but the log value of these sales are much close to normally distributed, especially the log value of global sales. Though the Shapiro test with p value lass than 0.05 deny its normality, it’s much better than the other sales or other log values of sales. Maybe the missing value of sale is the reason of abnormality. We will pay more attention to the log value of global sales later.

From the histograms and qq plots we also see that two scores and log values of their counts are close to normal distribution. Though the Shapiro test still deny the normality of these log values. We assume they are normally distributed in our analysis.

There are lots of interest points in this data set such as the distribution of global and regional sales, their relationship; the correlation of critic score and user score, and their counts; whether these scores are the main effect for sales, or the effect of other factors matter to sales such as genre, rating, platform, publisher, and so on. First let’s do visualization.

```{r}

#Visualization of categorical variables

#To simplify platform analysis, We regroup platform as Platform.type.

#regroup platform as Platform.type

pc <- c("PC")

xbox <- c("X360", "XB", "XOne")

nintendo <- c("Wii", "WiiU", "N64", "GC", "NES", "3DS", "DS")

playstation <- c("PS", "PS2", "PS3", "PS4", "PSP", "PSV")

game <- game %>%

mutate(Platform.type = ifelse(Platform %in% pc, "PC",

ifelse(Platform %in% xbox, "Xbox",

ifelse(Platform %in% nintendo, "Nintendo",

ifelse(Platform %in% playstation, "Playstation", "Others")))))

ggplot(game, aes(x = Platform.type)) + geom\_bar(fill = "blue")

```

```{r}

#As the bar plot shown here, Playstation is the biggest group, then xbox and nintendo. While “Others” is the smallest type.

dat <- data.frame(table(game$Genre))

dat$fraction = dat$Freq / sum(dat$Freq)

dat = dat[order(dat$fraction), ]

dat$ymax = cumsum(dat$fraction)

dat$ymin = c(0, head(dat$ymax, n = -1))

names(dat)[1] <- "Genre"

library(ggplot2)

ggplot(dat, aes(fill = dat$Genre, ymax = ymax, ymin = ymin, xmax = 4, xmin = 3)) +

geom\_rect(colour = "grey30") + # background color

coord\_polar(theta = "y") + # coordinate system to polar

xlim(c(0, 4)) +

labs(title = "Ring plot for Genre", fill = "Genre") +

theme(plot.title = element\_text(hjust = 0.5))

#Action, Sports and Shooter are the first three biggest genre. Action occupies almost 25% genre. Three of them together contribute over half of genre count. Puzzle, Adventure and Stratage have relatively less count.

#We regroup rating AO, RP and K-A as “Others” because there are only few observations of these ratings.

```

```{r}

#regroup Rating as Rating.type

rating <- c("E", "T", "M", "E10+")

game <- game %>% mutate(Rating.type = ifelse(Rating %in% rating, as.character(Rating), "Others"))

counts <- sort(table(game$Rating.type), decreasing = TRUE)

# rename the names of counts for detail information

names(counts) <- c("T - Teen", "E - Everyone", "M - Mature", "E10+ - Everyone 10+", "Others")

pct <- paste(round(counts/sum(counts) \* 100), "%", sep = " ")

lbls <- paste(names(counts), "\n", pct, sep = " ") # Rating information and percentage

pie(counts, labels = lbls, col = rainbow(length(lbls)),

main="Pie Chart of Ratings with sample sizes")

#According to the order, the most popular ratings are T, E, M and E10+. “Others” rating only occupy very little portion in the all games.

```

```{r}

p <- ggplot(game) +

geom\_mosaic(aes(x = product(Rating.type), fill = Platform.type), na.rm = TRUE) +

labs(x = "Rating Type", y = "Platform Type", title="Mosaic Plot") +

theme(axis.text.y = element\_blank())

ggplotly(p)

#As we noticed previously, Rating Type of “Others” cannot be seen here in plot because of its small amount. For all platform and rating combination, Playstation games occupy the most portion in all other three different rating types except Everyone 10 age plus. Nintendo is the most popular game for Everyone 10+, it’s the second popular platform for rating Everyone. Xbox is the second popular platform for rating Mature and Teenage, and it’s the third favorite platform for rating Everyone and Everyone 10+. Most “Others” platform games are rated as Everyone

```

```{r}

#9.2 Correlation among numeric variables

st <- game[, c(11, 13, 17:23)] # take numeric variables as goal matrix

st <- na.omit(st)

library(ellipse)

library(corrplot)

corMatrix <- cor(as.matrix(st)) # correlation matrix

col <- colorRampPalette(c("#7F0000", "red", "#FF7F00", "yellow", "#7FFF7F",

"cyan", "#007FFF", "blue", "#00007F"))

corrplot.mixed(corMatrix, order = "AOE", lower = "number", lower.col = "black",

number.cex = .8, upper = "ellipse", upper.col = col(10),

diag = "u", tl.pos = "lt", tl.col = "black")

#There are high r values of 0.75, 0.65, 0.52 and 0.42 between the log value of Global.Sales and regional sales, we will consider to use Global.Sales.Log as our target sales to analyze the relationship of sales with other variables later. On the other hand, there are good positive correlation between regional sales too. User Score is positive correlated to Critic Score with r of 0.58. There is little correlation between User Count log value and User Score.

```

```{r}

plot(hclust(as.dist(1 - cor(as.matrix(st))))) # hierarchical clustering

#All sales’ log value except JP.Sales.Log build one cluster; Scores, log value of counts and JP.Sales build the second cluster. In first cluster, Other.Sales.Log is the closest to Global.Sales.Log, then NA.Sales.Log, and EU.Sales.Log is the next.

```

#9.3 Analysis of score and count

```{r}

library(ggpmisc) #package for function stat\_poly\_eq

formula <- y ~ x

p1 <- ggplot(game, aes(x = User.Score, y = Critic.Score)) +

geom\_point(aes(color = Platform), alpha = .8) +

geom\_smooth(method = 'lm', se = FALSE, formula = formula) + #add regression line

theme(legend.position = "none") +

stat\_poly\_eq(formula = formula, #add regression equation and R square value

eq.with.lhs = "italic(hat(y))~`=`~", # add ^ on y

aes(label = paste(..eq.label.., ..rr.label.., sep = "\*plain(\",\")~")),

label.x.npc = "left", label.y.npc = 0.9, # position of the equation label

parse = TRUE) # output.type as "expression"

p2 <- ggplot() +

geom\_density(data = game, aes(x = Critic.Score), color = "darkblue", fill = "lightblue") +

geom\_density(data = game, aes(x = User.Score), color = "darkgreen", fill = "lightgreen", alpha=.5) +

labs(x = "Critic.Score-blue, User.Score-green")

library(gridExtra)

grid.arrange(p1, p2, nrow = 1, ncol = 2)

#There is positive correlation between Critic.Score and User.Score. In total, Critic score is lower than user score.

```

```{r}

t.test(game$Critic.Score, game$User.Score)

#T-test with p value of much less than 0.05 let us accept the alternative hypothesis with 95% confidence that there is significant difference between the means of critic score and user score. The mean of critic score is 7.03, and mean of user score is 7.19.

````

```{r}

p1 <- ggplot(game, aes(x = Critic.Count.Log, y = Critic.Score)) +

stat\_binhex() + # Bin 2d plane into hexagons

scale\_fill\_gradientn(colours = c("black", "red"),

name = "Frequency") # Adding a custom continuous color palette

p2 <- ggplot(game, aes(x = User.Count.Log, y = User.Score)) +

stat\_binhex() +

scale\_fill\_gradientn(colours = c("black", "red"), name = "Frequency") # color legend

grid.arrange(p1, p2, nrow = 1, ncol = 2)

#Critic.Score has a pretty good correlation to Critic.Count.Log, with an r value of 0.41 in the correlation analysis above, though Critic.Count.Log doesn’t have impact over Critic.Score. While User.Score looks like independent on User.Count.Log

````

9.4 Analysis of sales

9.4.1 By Year.Release

```{r}

Year.Release <- game$Year.Release

counts <- data.frame(table(Year.Release))

p <- game %>%

select(Year.Release, Global.Sales) %>%

group\_by(Year.Release) %>%

summarise(Total.Sales = sum(Global.Sales))

q <- cbind.data.frame(p, counts[2]) #add counts to data frame

names(q)[3] <- "count"

q$count <- as.numeric(q$count)

ggplot(q, aes(x = Year.Release, y = Total.Sales, label = q$count)) +

geom\_col(fill = "green") +

geom\_point(y = q$count \* 500000, size = 3, shape = 21, fill = "Yellow" ) +

geom\_text(y = (q$count + 50) \* 500000) + # position of the text: count of games each year

theme(axis.text.x = element\_text(angle = 90),

panel.background = element\_rect(fill = "purple"),

panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank()) +

scale\_x\_discrete("Year.Release", labels = as.character(Year.Release), breaks = Year.Release)

#We can see from the histogram of total sales that there is very little sales before 1996, only one game was released for each year. For several years between 1996 and 2000 the sales increased slowly. The count of games too. After that there is a big rise in total sales and the number of released games. The top sales happened in 2008, and the most count games was released in that year too. After that both total sales and count of games went downhill.

````

```{r}

game %>%

select(Year.Release, NA.Sales.Log, EU.Sales.Log, JP.Sales.Log,

Other.Sales.Log, Global.Sales.Log) %>%

melt(id.vars = "Year.Release") %>% #stacks other columns into "Year.Release"

group\_by(Year.Release, variable) %>%

summarise(total.sales = sum(value)) %>%

ggplot(aes(x = Year.Release, y = total.sales, color = variable, group = variable)) +

geom\_point() + geom\_line() +

labs(x = "Year Release", y = "Total Sales Log Value", color = "Region") +

theme(axis.text.x = element\_text(angle = 90),

panel.background = element\_rect(fill="pink"),

panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank())

#The pattern of log value for these regional sales in those years are similar for Global, North America, Europe, and Others. Japan is much different from them, which matches the previous cluster analysis.

````

9.4.3 By Rating

```{r}

game$Rating.type <- as.factor(game$Rating.type)

x <- game[, c(6:10)]

matplot(t(x), type = "l", col = rainbow(5)[game$Rating.type])

legend("center", levels(game$Rating.type), fill = rainbow(5), cex = 0.8, pt.cex = 1)

text(c(1.2, 2, 3, 3.9, 4.8), 80000000, colnames(x))

#The figure shows one E game(for everyone) which was sold mainly in North America and Europe produced a sale tale of over 80 millions’ global sale, while North America contributed half of the global sales. We can check the game data and know it’s Wii Sports released in 2006. We also noticed that Mature game is popular in North America(green), which contributed a lot to global sales, Everyone games(red) have good sale in Europe, while Japanese like Teen(purple) and Everyone(red) games. It’s balance in rating for “other” region.

````

9.4.4 By Genre

```{r}

game %>%

select(Year.Release, Global.Sales.Log, Genre) %>%

group\_by(Year.Release, Genre) %>%

summarise(Total.Sales.Log = sum(Global.Sales.Log)) %>%

ggplot(aes(x = Year.Release, y = Total.Sales.Log, group = Genre, fill = Genre)) +

geom\_area() +

theme(legend.position = "right", axis.text.x = element\_text(angle = 90),

panel.background = element\_rect(fill = "blue"),

panel.grid.major = element\_blank(),

panel.grid.minor=element\_blank())

#The figure shows the golden year for games are from 2007 to 2009, these games together produced above 7000 total.sales.log in each of those years. Action and sports keeps on the top sale for almost all of those 20 years, occupying biggest portion of the total global sales log. Adventure, Puzzle and Strategy are on the bottom of the sale log list.

````

9.4.5 by Score

```{r}

p1 <- ggplot(game, aes(x = Critic.Score, y = Global.Sales.Log)) +

geom\_point(aes(color = Genre)) +

geom\_smooth() # Regression line

p2 <- ggplot(game, aes(x = User.Score, y = Global.Sales.Log)) +

geom\_point(aes(color = Rating)) +

geom\_smooth()

grid.arrange(p1, p2, nrow = 1, ncol = 2)

#Independent from Genre and Rating, the higher of Critic Score, the better of Global.Sales.Log. Especially for Critic.Score bigger than 9, Global.Sales.Log straightly rises. Global.Sales.Log rises very slowly with User.Score

```

```{r}

game$Name <- gsub("Brain Age: Train Your Brain in Minutes a Day", #shorten the game name

"Brain Age: Train Your Brain", game$Name)

p1 <- game %>%

select(Name, User.Score, Critic.Score, Global.Sales) %>%

group\_by(Name) %>%

summarise(Total.Sales = sum(Global.Sales), Avg.User.Score = mean(User.Score),

Avg.Critic.Score = mean(Critic.Score)) %>%

arrange(desc(Total.Sales)) %>%

head(20)

ggplot(p1, aes(x = factor(Name, levels = Name))) +

geom\_bar(aes(y = Total.Sales/10000000), stat = "identity", fill = "green") +

geom\_line(aes(y = Avg.User.Score, group = 1, colour = "Avg.User.Score"), size = 1.5) +

geom\_point( aes(y = Avg.User.Score), size = 3, shape = 21, fill = "Yellow" ) +

geom\_line(aes(y = Avg.Critic.Score, group = 1, colour = "Avg.Critic.Score"), size = 1.5) +

geom\_point(aes(y = Avg.Critic.Score), size = 3, shape = 21, fill = "white") +

theme(axis.text.x = element\_text(angle = 90, size = 8)) +

labs(title = "Top Global Sales Game with Score", x = "Name of the top games" ) +

theme(plot.title = element\_text(hjust = 0.5))

#Among these 20 top sale games, the first two games, Wii Sports and Grand Theft Auto V have much better sales than the others. For most games, average critic score is higher than average user score, which agree with our density plot in Figure 9.5. Two Call of Duty games got really lower average user score comparing with other top sales games.

```

9.4.6 By Rating & Genre & Critic score

```{r}

p1 <- game %>%

select(Rating.type, Global.Sales, Genre, Critic.Score) %>%

group\_by(Rating.type, Genre) %>%

summarise(Total.Sales = sum(Global.Sales)/10^8, Avg.Score = mean(Critic.Score))

p2 <- p1 %>% group\_by(Genre) %>% summarise(Avg.Critic.Score = mean(Avg.Score))

ggplot() +

geom\_bar(data = p1, aes(x = Genre, y = Total.Sales, fill = Rating.type),

stat = "Identity", position = "dodge") +

geom\_line(data = p2,

aes(x = Genre, y = Avg.Critic.Score, group = 1, color = "Avg.Critic.Score"),

size = 2) +

geom\_point(data = p2, aes(x = Genre, y = Avg.Critic.Score, shape = "Avg.Critic.Score"),

size = 3, color = "Blue") +

scale\_colour\_manual("Score", breaks = "Avg.Critic.Score", values = "yellow") +

scale\_shape\_manual("Score", values = 19) + #add scale for line and point plot

theme(axis.text.x = element\_text(angle = 90),

legend.position="bottom",

panel.background = element\_rect(fill = "black"),

panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank())

#For genre & rating & Global.sale combination, Everyone\*sports game is so popular that it occupy the first in global sales among all combinations. Rating Mature contributes big portions in both Action and Shooter global sales. On average these three top sales genres show relatively higher critic score. Fighting, adventure and racing games get relatively lower average critic score. We can also see from the figure that adventure, puzzle and stratage do sell less comparing with other genres.

```

9.4.7 By Platform

```{r}

p <- game %>%

group\_by(Platform.type, Year.Release) %>%

summarise(total = sum(Global.Sales))

p$Year.Release. <- as.numeric(as.character(p$Year.Release))

ggplot(p, aes(x = Year.Release., fill = Platform.type)) +

geom\_density(position = "fill") +

labs(y = "Market Share") +

scale\_fill\_viridis(discrete = TRUE) +

scale\_y\_continuous(labels = percent\_format())

#Nintendo and Xbox came after 1990. Before that PC and Playstation occupied the game market, PC are the main platform at that time. After 1995, the portion of PC and Playstation shrinked, while Nintendo and Xbox grew fast and took over more portion than Playstation and PC in the market. Together with Nintendo and Xbox, there were other game platform sprouting out in early 1990s, but they last for 20 years and disappeared. From around 2010, the portions of Nintendo, PC, Playstation, and Xbox, these 4 platforms keep relatively evenly and stably.

```

```{r}

#compute 1-way ANOVA test for log value of global sales by Platform Type

model <- aov(Global.Sales.Log ~ Platform.type, data = game)

summary(model)

```

```{r}

tukey <- TukeyHSD(model)

par(mar = c(4, 10, 2, 1))

plot(tukey, las = 1)

#ANOVA test shows that there is at lease one of the mean values of Global.Sales.Log for those platform types is significant different from the others. In detail, the plot of Turkey tests tells us that there is significant difference between all other pairs of platform types but between Xbox and Nintendo, others and Nintendo.

```

```{r}

game$Platform.type <- as.factor(game$Platform.type)

ggplot(game, aes(x = Platform.type, y = Global.Sales.Log, fill = Rating.type)) +

geom\_boxplot()

#In total, PC has lower Global sales log comparing with other platform type, while Playstation and Xbox have higher sale mediums for different rating types. Rating of Everyone sold pretty well in all platform type, while rating Mature sold better in PC, Playstation and Xbox.

```

```{r}

ggplot(game, aes(Critic.Score, Global.Sales.Log, color = Platform.type)) +

geom\_point() +

facet\_wrap(~ Genre)

#Most genre plots in Figure 9.15 illustrate that there are positive correlation between Global.Sales.Log and Critic Score, the higher the critic score, the better the global sales log value. Most puzzle games were from Nintendo, while lots of stratage games are PC. For other genres, all platforms shared the portion relatively evenly. Lots of PC(green) shared lower market portion in different genres, while some of Nintendo(red) games in sports, racing, platform, and misc were sold really well. At the same time, Playstation with genre of action, fighting, and racing games, Xbox with genre of misc, action, and shooter games show higher global sales log too

```

9.5 Effect of platform type to priciple components

```{r}

st <- game[, c(11, 13, 17:23)]

pca = prcomp(st, scale = T) #scale = T to normalize the data

percentVar <- round(100 \* summary(pca)$importance[2, 1:9], 0) # compute % variances

pcaData <- as.data.frame(pca$x[, 1:2]) #First and Second principal component value

pcaData <- cbind(pcaData, game$Platform.type) #add platform type as third col. for cluster purpose

colnames(pcaData) <- c("PC1", "PC2", "Platform")

ggplot(pcaData, aes(PC1, PC2, color = Platform, shape = Platform)) +

geom\_point(size = 0.8) +

xlab(paste0("PC1: ", percentVar[1], "% variance")) + # x label

ylab(paste0("PC2: ", percentVar[2], "% variance")) + # y label

theme(aspect.ratio = 1) # width and height ratio

#PC, Xbox, Playstation and Nintendo occupy in their own positions in the PCA figure, which illustrate that they play different important role in components of the variance of PC1 and PC2.

```

```{r}

library(ggfortify)

set.seed(1)

autoplot(kmeans(st, 3), data = st, label = FALSE, label.size = 0.1)

```

9.6 Models for global sales

Because there are too many of levels in Publisher and Developer, and there is apparent correlation between them, we use only top 12 levels of Publisher and classified the other publishers as “Others”; Because of the good correlation between Critic.Score and User.Score, we use only critic score; Also we use only log value of user score count because of it’s closer correlation to global sales log. We will not put other sales log variables in our model because their apparent correlation with global sales log.

```{r}

#re-categorize publisher into 13 groups

Publisher. <- head(names(sort(table(game$Publisher), decreasing = TRUE)), 12)

game <- game %>%

mutate(Publisher.type = ifelse(Publisher %in% Publisher., as.character(Publisher), "Others"))

game.lm <- game[, c(3:4, 11, 21, 23:26)]

model <- lm(Global.Sales.Log ~ ., data = game.lm)

summary(model)

```

```{r}

model <- aov(Global.Sales.Log ~ ., data = game.lm)

summary(model)

```

Global sales log is mostly effected by factors of critic score, user count log, platform type, Publisher type and genre in glm analysis. ANOVA shows every factor is significant in the contribution to global sales log. Critic score and user count log are the most important factors.

Critic score and User.Count.Log positively affect the global sales log, while other factors like Platform type and Genre either lift up or pull down the global sales according to their types. This model will explain the global sales log with R-Square of 0.57.

Because of the curve smooth lin2e at global sale ~ critic score plot in our previous analysis(Figure 9.11) and critic score’s big contribution in linear model analysis, We try a polynomial fit of critic score only

```{r}

model <- lm(Global.Sales.Log ~

Critic.Score + I(Critic.Score^2) + I(Critic.Score^3) + I(Critic.Score^4), data = game.lm)

summary(model)

```

The first two levels are not statistically significant according to our pre-analysis, so here we use the third and fourth levels only.

```{r}

model <- lm(Global.Sales.Log ~ I(Critic.Score^3) + I(Critic.Score^4), data = game.lm)

summary(model)

```

In total, the coefficients are statistically significant, the model of two levels of critic score itself will explain the Global.Sales.Log with R square 0.16.

```{r}

ModelFunc <- function(x) {model$coefficients[1] + x^3\*model$coefficients[2] + x^4\*model$coefficients[3]}

ggplot(data = game.lm, aes(x = Critic.Score, y = Global.Sales.Log)) +

geom\_point() + stat\_function(fun = ModelFunc, color = 'blue', size = 1)

#Here is the scatter plot of Global.Sales.Log ~ Critic Score and the model line which predict the global sales log with critic score.

```

9.7 Conclusion

Global and regional sales are not distributed normally, while their log values are close to normal distribution. Most regional sales have the similar pattern as global sales.

There is positive correlation between critic score and user score. In total, Critic score is lower than user score. No apparent correlation was found between scores and their counts.

Critic score, user score count log, genre, rating, platform, and publisher together affect the global sales log. Critic score is the most important contributor.