Market Exploration for the PS4 games using cluster analysis

Darya Savitskaya

June 23, 2023

Abstract

In this project the datset containing sales performance of various PS4 games over several regions (Europe, Japan, North America and The Rest of the World) was analysed with the help of unsupervised k-means cluster analysis. It was shown that Japan is a separate market with different preferences, some genres are more likely to produce a best-seller, however, are already over-crowded and, lastly, that there needs to be further research into small local markets.

1 Statement of the problem

1.1 Description of the dataset

This dataset is a part of Gregory Smith's web scrape, that concentrates only on the PS4 games' sales. It contains 9 variables: name of the game, year of release, name of the publisher, millions of units sold globally, as well as individually in North America, Europe, Japan and Rest of the World.

1.2 Goals of the analysis

The goal of the analysis is to extract valuable information from the data that can provide insights on market dynamics and patterns based on genre and sales performance in various regions. The unsupervised algorithms can provide information about:

- Market segmentation based on geographic regions and product categories
- Product categorization based on sales patterns
- Sales performance analysis to identify high-performing and low-performing regions, product or attributes.

1.3 Key findings

The data provided insight into game genre popularity based on regions (Europe, Japan, North America and the rest of the world) and into the performance within some popular genres, like action and shooter games. It was shown that Japan is a separate market that presents different preferences to the rest of the world (big popularity of role-playing games and smaller so of sports games), shooters and action genres are more likely to produce a best-seller, however, the genre seems to be over-crowded with average games. Lastly, there are some local markets that are worth exploring such as miscellaneous games and others.

2 Preliminary analysis

2.1 Analysis

The data can be visualised from various angles and it is important to do so to understand better the division proposed by an algorithm.

From Figure 1 it can be seen that the most bought and produced, revenue-bringing game genres are (in descending order): action, shooter, sports, role-playing, action-adventure; genres that have more local audience, that sell and are produced less are (in ascending order): puzzle, party, visual novel, strategy, MMO, Music, simulation miscellaneous, adventure, platform, fighting, racing.

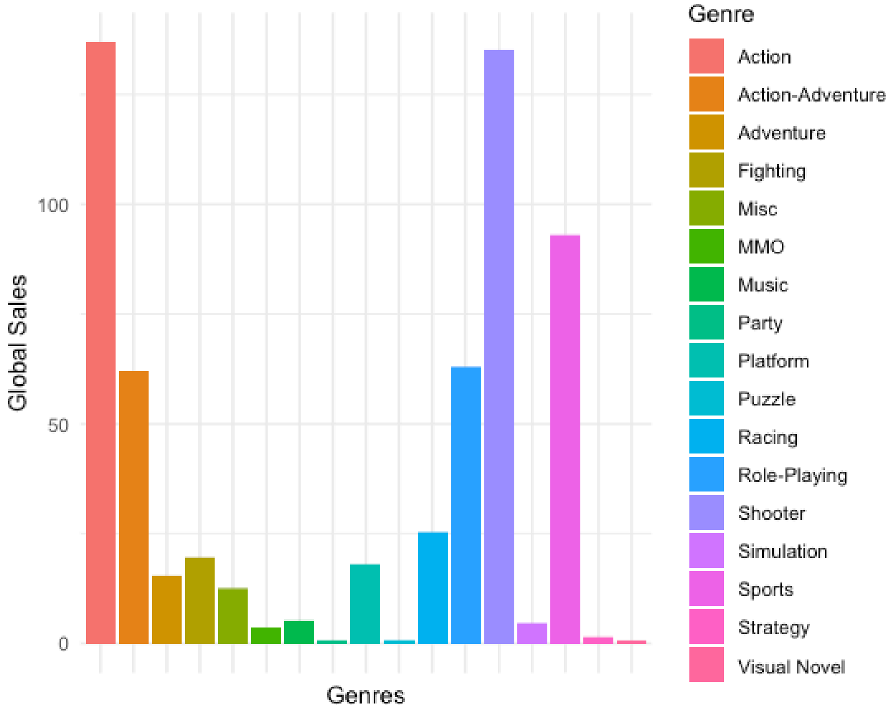
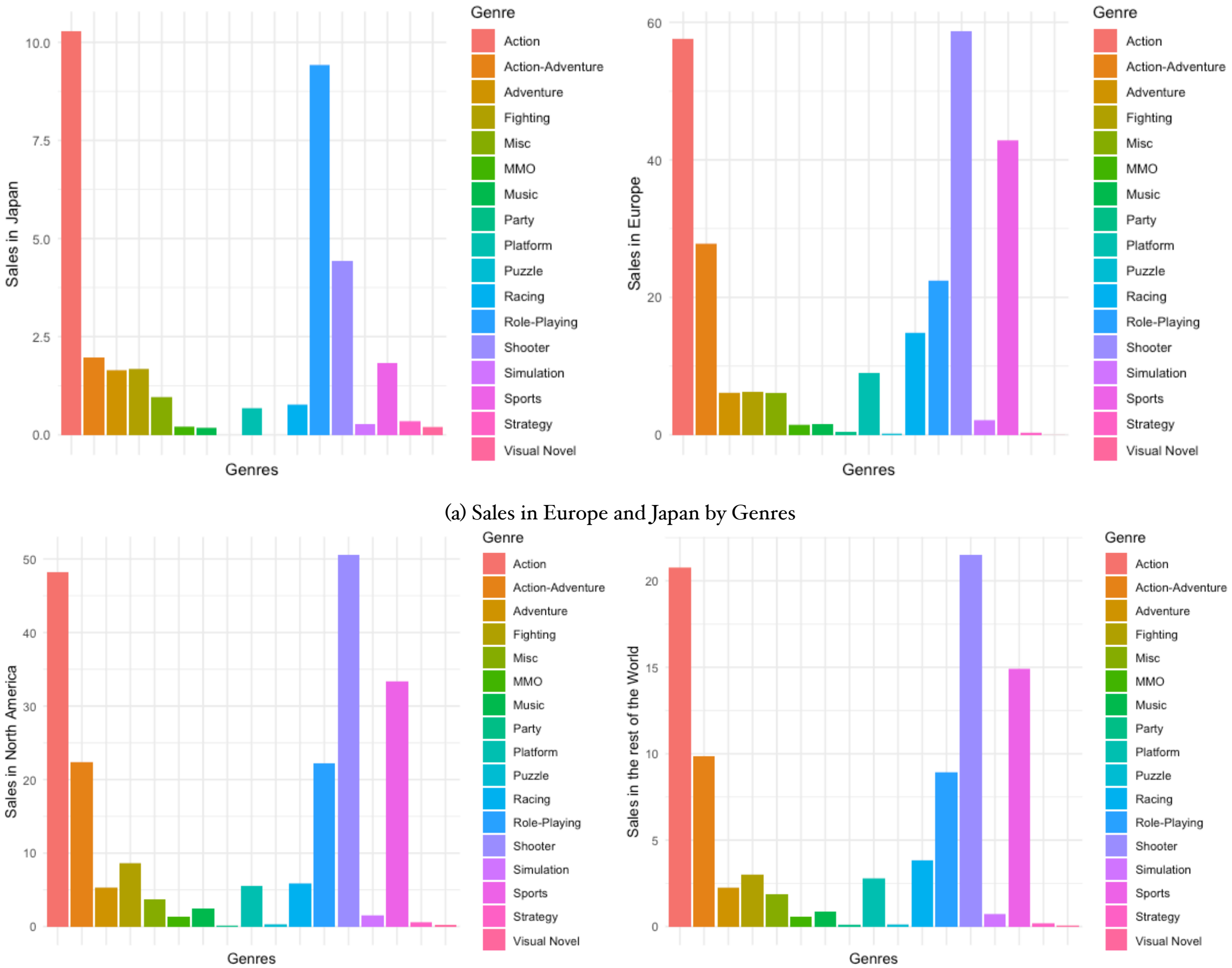


Figure 1: Total sales of games of different genres



(a) Sales in Europe and Japan by Genres

(b) Sales in North America and the rest of the world by Genres

Figure 2: Sales by Genres in different regions

Bar plots in Figure 2 help to understand customer preferences in different regions. It can be seen that Europe, North America and the rest of the world show similar distribution of preferences: the most popular genre is shooter, after the shooters we have action, sports, role-playing. Preferences follow the global distribution precisely and there is minimal difference between regions. Japan, however, already shows itself like a different market, ranking second role-playing games and shooters third, additionally sports games are less popular in Japan than anywhere else, performing on the same level as fighting and adventure games.

2.2 Data Preparation

2.2.1 Outliers

Unsupervised algorithms can be sensitive to outliers, however, when analysing sales data for market segmentation, outliers provide important information. In this case, as seen in figure 3, we have outliers that are over the upper-bounds, i.e. best-sellers.

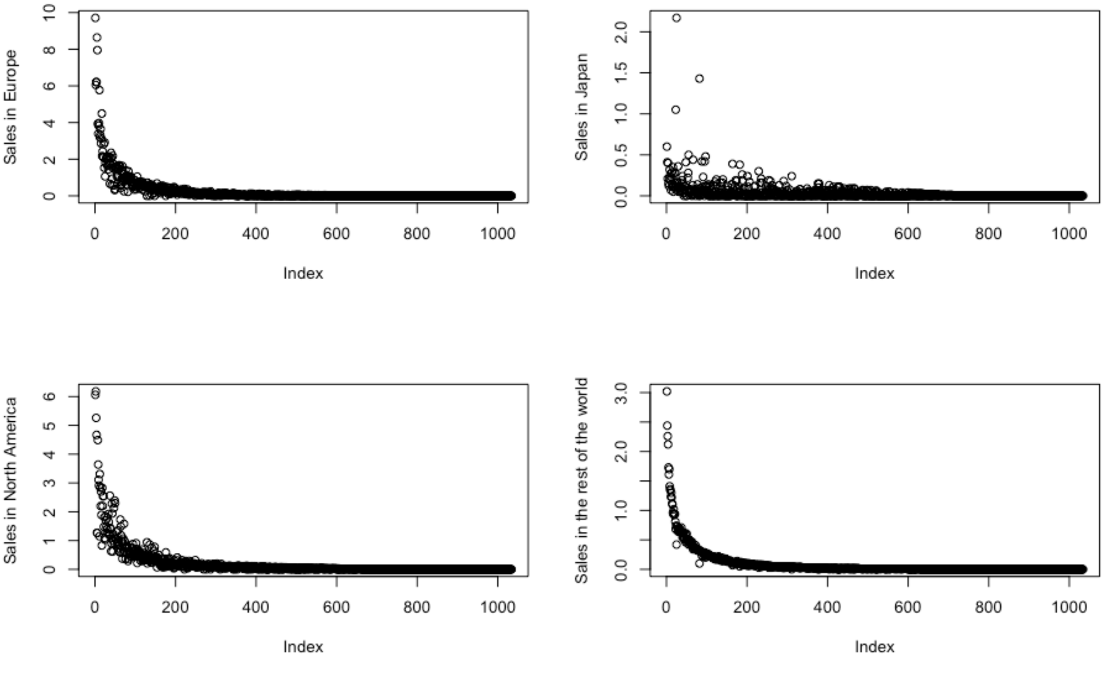


Figure 3: Sales by regions

I hypothesised that there are super bestseller games, likely action or shooter games, that are popular everywhere in the world and heavily pulling weight towards action and shooter genres in general. Chances are, global best-sellers are overall high-quality games that do not represent a genre and are more of a stand alone successful product. To locate them I calculated outliers for every region, using z-scores, with the following results: 19 outliers in Europe, 15 outliers in Japan, 23 outliers in North America and 21 outliers in the rest of the world. All of the regions have three outliers in common: Grand Theft Auto V (action), Call of Duty: Black Ops 3 (shooter) and Call of Duty: WWII (shooter). For the further analysis I will treat them as singular success stories, not related to genre or region, and therefore not helpful in segmentation task. Other outliers show successful games with different success rates in regions, over different genres - it was decided to leave them in the dataset as they carry valuable information.

2.2.2 Data Transformation

Additionally, the dataset had to be transformed for the algorithm to work properly. Firstly, were deleted three duplicated games. Secondly, genre variable was transformed into 17 binary categorical variables, where 1 means that the game belongs to the genre and 0 the opposite. Variables year, publisher and global sales were dropped.

2.2.3 Normalisation

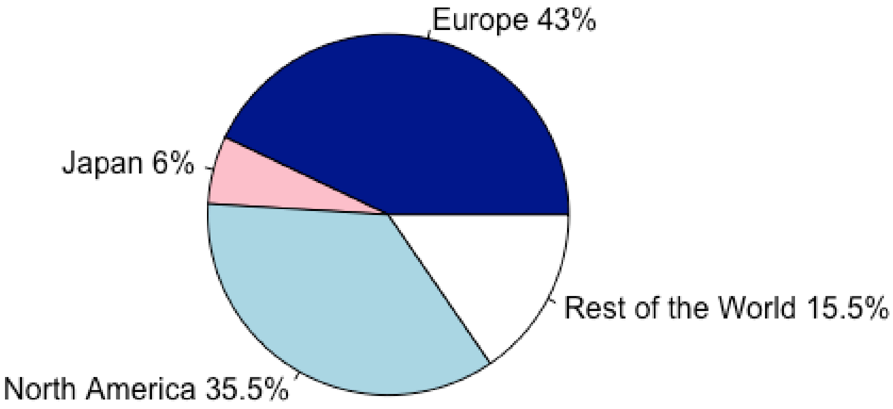


Figure 4: Sales distribution over regions without outliers.

Figure 4 shows that most of the sales happen in Europe or North America. To make sales values directly comparable, we can normalise them over all the regions, confiding them to the same distribution. As a result, the sales variable acts as an index that shows the popularity of the game in a region compared to all other games in other regions, where negative values would mean that the game in this region is more unpopular than on average and positive values would mean that the game is more popular than on average over all regions. This transformation makes the output of the segmentation algorithm more interpretable and makes local outliers have a smaller effect.

3 The analysis

3.1 Model set up

At this point the data set contains both numerical and categorical variables; we need to take it into account when calculating distances and use Gower distance which uses Eucliadian distances for numerical variables and dissimilarity measures for categorical. As data is pretty numerous, hierarchical clustering will be computationally expensive and produce a heavy and hard to interpret dendogram; we will start with k-means clustering. Number of clusters was chosen subjectively based on the interpretability.

3.2 Results

By iteration, the number of clusters was decided to be 7. As soon as number of clusters exceeds 7, clusters that contain only one genre start to present themselves, which do not carry any additional insights.

Table 1: Clusters

	Europe	Japan	N.A.	The world	Genres	Interpretation	Size
1	3.94	1.99	3.4	4.17	Various	Best-sellers	35
2	-0.16	0.3	-0.12	-0.15	Role-playing	Japan favourite	120
3	0.01	-0.11	0.43	0.23	Sports	Sports, unpopular in Japan	66
4	-0.12	-0.05	-0.15	-0.15	Action	Average action	250
5	-0.03	-0.14	-0.03	-0.03	Shooters	Average shooters	74
6	-0.3	-0.25	-0.36	-0.35	Miscellaneous	Miscellaneous	121
7	-0.15	-0.13	-0.19	-0.18	Various	The rest	351

3.2.1 1|Best-sellers

These are the games that performed exceptionally well world-wide everywhere, but less so in Japan. This group includes a number of genres, in particular: 5 action, 13 shooters, 5 sports, 3 role-playing, 1 of miscellaneous, racing, fighting, adventure. This cluster probably includes most of the outliers in the individual regions that were left in the dataset (i.e. are outliers in some regions but not all).

3.2.2 2|Japan Favourite Roleplayers

Already in preliminary analysis Japan showed itself to be a separate market with individual preference. Particularly Japan exhibits a love for role-playing genre that is nit exhibited by other regions.

3.2.3 3|Sport, unpopular in Japan

Unlike role-players, Japan does not show a good demand for sports games, that are popular more than average in North America and the Rest of the world and on average level in Europe.

3.2.4 4|Average Action

As action is the most prevalent genre in the making, analysis showed it benefits to make division inside the genre: we can see that if action game is not a best-seller, it will more likely show a bit smaller than average rate of success. The genre of action then consists of a large number of games - some with exceptional performance, some less than average.

3.2.5 5|Average Shooters

Shooters, shown in preliminary analysis to follow action games in the level of sales, follow the same division as action games. Apart from best-seller shooter games, there is a large number of average or less than average performing shooters.

3.2.6 6|Miscellaneous

Originally, I did not consider singular genres as a good group worth highligting, however, consumers of miscellaneous games show a particular difficulty in targeting. That is because miscellaneous is not a genre and more of a grouping of individual games that do not fit any other genres.That includes indie games, games with innovative elements and etc. Therefore, it is a completely different market segment that is impossible to market to because of its diversity, however, it is still worth investing in as it breeds innovation and industry development.

3.2.7 7|The rest

This group includes various genres with on average unpopular rates, that likely have local small pool of devoted consumers, like puzzles, music games and visual novels.

4 Conclusions

We can make the following conclusions based on the data:

- Japan is a separate market that presents different preferences. In particular, great liking of the role-playing genre and, on the contrary, small liking of the sports games, compared to the average world preferences.

- The biggest best-sellers are likely to be shooters or action games, however, the market for those genres is already over crowded with numerous average and less than average games.

- Miscellaneous games need a different approach as a market segment, because of its heterogeneity, complexity and innovativity.

- There is a number of genres that perform less than average, however, are still being produced and bought, therefore it might be beneficial to study separately local markets for on average poor performing genres.

In the future, it might be a good idea to analyse the distribution of the data in some crowded clusters, such as cluster 7 - the Rest or Miscellaneous, more closely. Otherwise, it is possible to completely eliminate best-sellers to see what constitutes and average performance better.

The Appendix: the R code

```
install.packages(”proxy”, dependencies=TRUE)
library(proxy)
### EDA ###
library(tidyverse)
install.packages(”gridExtra”)
library(gridExtra)
#Total count of games in each genre
ggplot(data=sales)+ geom_bar(mapping=aes(x=Genre,fill=Genre))+
  labs(title=”Total Count of Games in Each Genre”, labels = NULL)+
  scale_x_discrete(labels = NULL)+
  theme_minimal()
layout(matrix(c(1, 2), nrow = 1))
#Most popular genres in Japan
genres_japan <- sales[,c(3,7)]
plot1 <- ggplot(genres_japan, aes(x = Genre, y = Japan, fill = Genre)) +
  geom_bar(stat = ”identity”) +
  labs(x = ”Genres”, y = ”Sales in Japan”) +
  scale_x_discrete(labels = NULL)+
  theme_minimal()

#Most popular genres in Europe
genres_europe <- sales[,c(3,6)]
plot2 <- ggplot(genres_europe, aes(x = Genre, y = Europe, fill = Genre)) +
  geom_bar(stat = ”identity”) +
  labs(x = ”Genres”, y = ”Sales in Europe”) +
  scale_x_discrete(labels = NULL)+
  theme_minimal()

grid.arrange(plot1, plot2, ncol = 2)

#Most popular genres in North America
genres_na <- sales[,c(3,5)]
plot3 <- ggplot(genres_na, aes(x = Genre, y = North.America, fill = Genre)) +
  geom_bar(stat = ”identity”) +
  labs(x = ”Genres”, y = ”Sales in North America”) +
  scale_x_discrete(labels = NULL)+
  theme_minimal()

#Most popular genres in the Rest of the World
genres_world <- sales[,c(3,8)]
plot4 <- ggplot(genres_world, aes(x = Genre, y = Rest.of.World, fill = Genre)) +
  geom_bar(stat = ”identity”) +
  labs(x = ”Genres”, y = ”Sales in the rest of the World”) +
  scale_x_discrete(labels = NULL)+
  theme_minimal()
grid.arrange(plot3, plot4, ncol = 2)
#Sales by genre
global_bygenre <- sales[,c(3,9)]
ggplot(global_bygenre, aes(x = Genre, y = Global, fill = Genre)) +
  geom_bar(stat = ”identity”) +
  labs(x = ”Genres”, y = ”Global Sales”) +
  scale_x_discrete(labels = NULL)+
  theme_minimal()

library(tidyr)
sales_only <- sales[,c(5,6,7,8)]
help(”pivot_longer”)
sales_long <- pivot_longer(sales_only, c(1,2,3,4), names_to = ”Region”, values_to =
                            ”Sales”)

ggplot(sales_long, aes(x = Region, y = Sales, fill = Region)) +
  geom_bar(stat = ”identity”) +
  labs(x = ”Region”, y = ”Sales”) +
  theme_minimal()

###checking for outliers
boxplot(sales$Global)
boxplot(sales$North.America)
boxplot(sales$Japan)
boxplot(sales$Rest.of.World)
boxplot(sales$Europe)

layout(matrix(c(1,2,3,4), nrow = 2))
```

```
plot(sales$Global)
plot(sales$Europe, ylab = ”Sales in Europe”)
plot(sales$North.America, ylab = ”Sales in North America”)
plot(sales$Japan, ylab = ”Sales in Japan”)
plot(sales$Rest.of.World, ylab = ”Sales in the rest of the world”)

layout(1)
##checking outliers in global sales

#Global outliers
z_scores_global <- scale(sales$Global)
outliers_global <- sales[apply(abs(z_scores_global) > 3, 1, any), ]

#Europe outliers
z_scores_europe <- scale(sales$Europe)
outliers_europe <- sales[apply(abs(z_scores_europe) > 3, 1, any), ]

#Japan outliers
z_scores_japan <- scale(sales$Japan)
outliers_japan <- sales[apply(abs(z_scores_japan) > 3, 1, any), ]

#North America outliers
z_scores_na <- scale(sales$North.America)
outliers_na <- sales[apply(abs(z_scores_na) > 3, 1, any), ]

#Rest of the world outliers
z_scores_world <- scale(sales$Rest.of.World)
outliers_world <- sales[apply(abs(z_scores_world) > 3, 1, any), ]

#Common outliers
common_outliers <- Reduce(intersect, list(outliers_europe, outliers_japan, outliers_na,
sales_no_out <- anti_join(sales, common_outliers, by = ”Game”)
dim(sales)- dim(sales_no_out)
### Data preparation ###

sales <- read.csv(”PS4_GamesSales.csv”)
head(sales)
summary(sales)
```

```
unique(sales['Genre'])
unique(sales['Publisher'])
unique(sales['Year'])

### Adding categorical variables for every genre
sales_1<- sales_no_out

#dropping year, publisher and global
sales_1<- sales_no_out[,-c(2,4,9)]

unique(sales['Genre'])
sales_1$Action<- factor(ifelse(sales_1$Genre == 'Action', 1, 0))
sales_1$Shooter<- factor(ifelse(sales_1$Genre == 'Shooter', 1, 0))
sales_1$Action_Adventure<- factor(ifelse(sales_1$Genre == 'Action-Adventure', 1, 0))
sales_1$Sports<- factor(ifelse(sales_1$Genre == 'Sports', 1, 0))
sales_1$Role_Playing<- factor(ifelse(sales_1$Genre == 'Role-Playing', 1, 0))
sales_1$Misc<- factor(ifelse(sales_1$Genre == 'Misc', 1, 0))
sales_1$Platform<- factor(ifelse(sales_1$Genre == 'Platform', 1, 0))
sales_1$Racing<- factor(ifelse(sales_1$Genre == 'Racing', 1, 0))
sales_1$Fighting<- factor(ifelse(sales_1$Genre == 'Fighting', 1, 0))
sales_1$Adventure<- factor(ifelse(sales_1$Genre == 'Adventure', 1, 0))
sales_1$MMO<- factor(ifelse(sales_1$Genre == 'MMO', 1, 0))
sales_1$Simulation<- factor(ifelse(sales_1$Genre == 'Simulation', 1, 0))
sales_1$Music<- factor(ifelse(sales_1$Genre == 'Music', 1, 0))
sales_1$Party<- factor(ifelse(sales_1$Genre == 'Party', 1, 0))
sales_1$Strategy<- factor(ifelse(sales_1$Genre == 'Strategy', 1, 0))
sales_1$Puzzle<- factor(ifelse(sales_1$Genre == 'Puzzle', 1, 0))
sales_1$Visual_Novel<- factor(ifelse(sales_1$Genre == 'Visual Novel', 1, 0))

sales_1 <-sales_1[,-2]
#deleting duplicated games with zero values
duplicated(sales_1$Game)
duplicated_games <- sales_1[duplicated(sales_1$Game), ]
sales_1 <- sales_1[-c(794,961,1023),]

#sales distribution by regions in pie chart
summary(sales_1)
sum(sales_1$Europe)/total
sum(sales_1$Japan)/total
sum(sales_1$North.America)/total
sum(sales_1$Rest.of.World)/total
total = sum(sales_1$Japan) + sum(sales_1$Europe) + sum(sales_1$North.America) + sum(sale
labels <- c(”Europe 43%”, ”Japan 6%”, ”North America 35.5%”, ”Rest of the World 15.5%”)
values <- c(sum(sales_1$Europe), sum(sales_1$Japan), sum(sales_1$North.America), sum(sal
colors <- c(”darkblue”, ”pink”, ”lightblue”, ”white”)
pie(values, labels = labels, col = colors)
help(”pie”)



sales_1 <- data.frame(sales_1, row.names =”Game”)

#majority of sales is in europe, let’s standardize the sales variables together
#also reduce the influence of outliers
#new dataset
sales_stan <- sales_1
sales_stan$North.America <- scale(sales_stan$North.America)
sales_stan$Europe <- scale(sales_stan$Europe)
sales_stan$Japan <- scale(sales_stan$Japan)
sales_stan$Rest.of.World <- scale(sales_stan$Rest.of.World)

sales_scaled <- scale(sales_stan[,c(1,2,3,4)])

sales_scaled <- data.frame(sales_scaled)
summary(sales_scaled)

sales_stan$North.America <- sales_scaled$North.America
sales_stan$Europe <- sales_scaled$Europe
sales_stan$Japan <- sales_scaled$Japan
sales_stan$Rest.of.World <- sales_scaled$Rest.of.World



summary(sales_stan)
### CLUSTERING
### k-means
```

```
library(cluster)
dist_matrix <- daisy(sales_stan, metric = ”gower”)
#5 clusters
set.seed(569)
kmeans_clusters <- kmeans(dist_matrix, centers = 5)
clusterkm_labels <- kmeans_clusters$cluster
sales_stan$cluster5km <- clusterkm_labels
summary(sales_stan[sales_stan$cluster5km==1,])
#shooters with less than average success everywhere 74
summary(sales_stan[sales_stan$cluster5km==2,])
#sports, popular in na and others 62
summary(sales_stan[sales_stan$cluster5km==3,])
#adventure with less than average success everywhere 97
summary(sales_stan[sales_stan$cluster5km==4,])
#not popular games (action, action-adventure, role-playing, racing, fighting, mmo, simul
summary(sales_stan[sales_stan$cluster5km==5,])
#well-received games (action, shooter, a_a, sports, role_playing, ) 57

set.seed(904)
#8 clusters
kmeans_clusters <- kmeans(dist_matrix, centers = 8)
clusterkm_labels <- kmeans_clusters$cluster
sales_stan$cluster5km <- clusterkm_labels

summary(sales_stan[sales_stan$cluster5km==1,])
#best-sellers of different genres
summary(sales_stan[sales_stan$cluster5km==2,])
#slightly unpopular games of different genres
summary(sales_stan[sales_stan$cluster5km==3,])
#action best-sellers
summary(sales_stan[sales_stan$cluster5km==4,])
#role-playing popular in japan
summary(sales_stan[sales_stan$cluster5km==5,])
#slightly un-popular action
summary(sales_stan[sales_stan$cluster5km==6,])
#miscalleneous
summary(sales_stan[sales_stan$cluster5km==7,])
#adventure
```

```
summary(sales_stan[sales_stan$cluster5km==8,])
#sports, unpopular in japan

set.seed(824)
#6 clusters
kmeans_clusters <- kmeans(dist_matrix, centers = 6)
clusterkm_labels <- kmeans_clusters$cluster
sales_stan$cluster5km <- clusterkm_labels

summary(sales_stan[sales_stan$cluster5km==1,])
#334, slightly unpopular games
summary(sales_stan[sales_stan$cluster5km==2,])
#14, bestsellers, action
summary(sales_stan[sales_stan$cluster5km==3,])
#34,best-sellers, different genres
summary(sales_stan[sales_stan$cluster5km==4,])
#74, slightly unpopular shooters
summary(sales_stan[sales_stan$cluster5km==5,])
#240, slightly unpopular action
summary(sales_stan[sales_stan$cluster5km==6,])
#slightly unpopular games that do better in Japan

set.seed(1998)
#7 clusters
kmeans_clusters <- kmeans(dist_matrix, centers = 7)
clusterkm_labels <- kmeans_clusters$cluster
sales_stan$cluster5km <- clusterkm_labels

summary(sales_stan[sales_stan$cluster5km==1,])
#slightly unpopular games 361
summary(sales_stan[sales_stan$cluster5km==2,])
#slightly unpopular shooters 74
summary(sales_stan[sales_stan$cluster5km==3,])
#role-playing popular in japan 120
summary(sales_stan[sales_stan$cluster5km==4,])
#sports popular in the world, na and less europe 66
summary(sales_stan[sales_stan$cluster5km==5,])
#miscalleneous games 121
```

```
summary(sales_stan[sales_stan$cluster5km==6,])
#world best-sellers (shooters, action, action adventure, role-playing) 35
summary(sales_stan[sales_stan$cluster5km==7,])
#slightly unpopular action 250
```