ANALYSIS OF
VIDEO GAMES SALES

Sijin Zhou & Junfan Chen
IST 707 Data Analytics
Professor: Jesse Cases
szhou41@syr.edu | jchen269@syr.edu

Abstract

As a cultural product, games have become an important part of people's cultural life. They have greatly enriched people's lives and brought huge profits to the company. In order to further understand the development of the game industry, in this paper we conducted Data Exploration on industry data and used Association Rules and Classification Analysis. We found that Play Station is the most popular gaming platform; North America, Europe, and Japan are the main markets for current games; teenagers are the main consumers of games, and their favorite is action games; for companies, they can use Random Forest to assist market expectation analysis

I. Introduction

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. 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 below. Complete cases are 6,900. We first perform Data Exploration and Association Rules on the data set, finding out the combination of attributes that account for high satisfaction. Then we perform classification analysis on the factors in the data set, observing groups of similar games and find out the suitable classifier for this dataset. According to the previous analysis results, we summarize the business advice for game companies.

II. Methods

2.1 Data Preparing

Our project is going to explore the Video Games Sales Dataset from Kaggle (https://www.kaggle.com/sidtwr/videogames-sales-dataset). There are 3 csv files under this dataset, and we finally decide to use 'Video_Games_Sales_as_at_22_Dec_2016.csv', which contains data of other two files, to be our object. This csv file is consist of 11563 lines of sales information and 16 columns: Name, Platform, Year_of_Release, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, Critic_Score, Critic_Count, User_Score, User_Count, Developer, and Rating after data processing.

2.2 Data Exploration

In this step, we mainly deleted NAs, Spaces and check outliers. After preparing our data, we use correlation matrix to explore correlations among each numeric variable. And visualize the relationship of sales and Year_of_Release, sales and genre, sales and platforms.

2.3 Association Rules

In this part, we define any games that score is above 8.2 is satisfied and others are not satisfied. The Left Hand Side variables are Platform, Genre, Publisher, Developer, Rating, NA_Sales_Class, EU_Sales_Class, JP_Sales_Class, Other_Sales_Class. The parameter of support is 0.01 and the confidence is 0.5.

2.4 Classification Analysis

In this step, we want to explore further on the relationship between customer safisfaction and other factors. Based on the results of previous steps, we decided to slice the original data and focus on these columns: Platform, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Critic_Score, Developer, Rating, and Customer_Satisfaction. After setting Customer_Satisfaction as the independent variable, we will perform Naïve Bayes, Bagged Cart, Random Forest, and Decision Tree on this dataset, finding out which of them is the most suitable classifier for this task. For the training step, we randomly sample 80% of data to train each classifier, and use the train() function for Naïve Bayes, Bagged Cart, Random Forest, and Decision Tree. In order to generate a reliable classifier, we set the value of ‘control’, a parameter of train() function, as trainControl(method = 'repeatedcv', number = 10), which means we will use repeated random sub-sampling validation with 10 folds during the training. As for the Random Forest, we set ‘control’ equals to trainControl(method = "repeatedcv", number = 5, repeats = 5), because it will come out a tree with reasonable branches. Then use plot() in ‘rpart’ package to draw the tree.

III. Results

3.1 Data Exploration

Data preprocessing:
In order to give the appropriate outcome of str( ) and summary ( ) function, there are some data types formatting steps to take.

Names

1. Check replicate, all the values are unique

2. There arw two rows (660th and 14247th) have null value in this column, we deleted them.
3. We deleted spaces before the games' name

Platform:

1. Check replications, all the values are unique
2. Check Null values, there are no Null values in the column
3. Deleted spaces before the platform name

Year_of_release

1. There is no NA in this column
2. Converted the format to date format

Genre

1. There is no NA in this column
2. The genre includes: Action, Adventure, Fighting, Misc, Platform, Puzzle, Racing, Role-Playing, Shooter, Simulation, Sports and Strategy

Publisher

1. There is no NA in this column.
2. There are 531 publishers in total in the dataset.

NA_Sales/EU_Sales/JP_Sales/Other Sales and Global Sales have no NA

Critic_Score/Critic_Count/User_Score/User_Count have a lot of NA values. Because in the associate rule, we need to consider those indexes, therefore we filter all the null values in this column and subset a new one, which there is no NAs in those columns.

Developer and Rating have spaces and NA values, we used gusb( ) function deleted spaces.

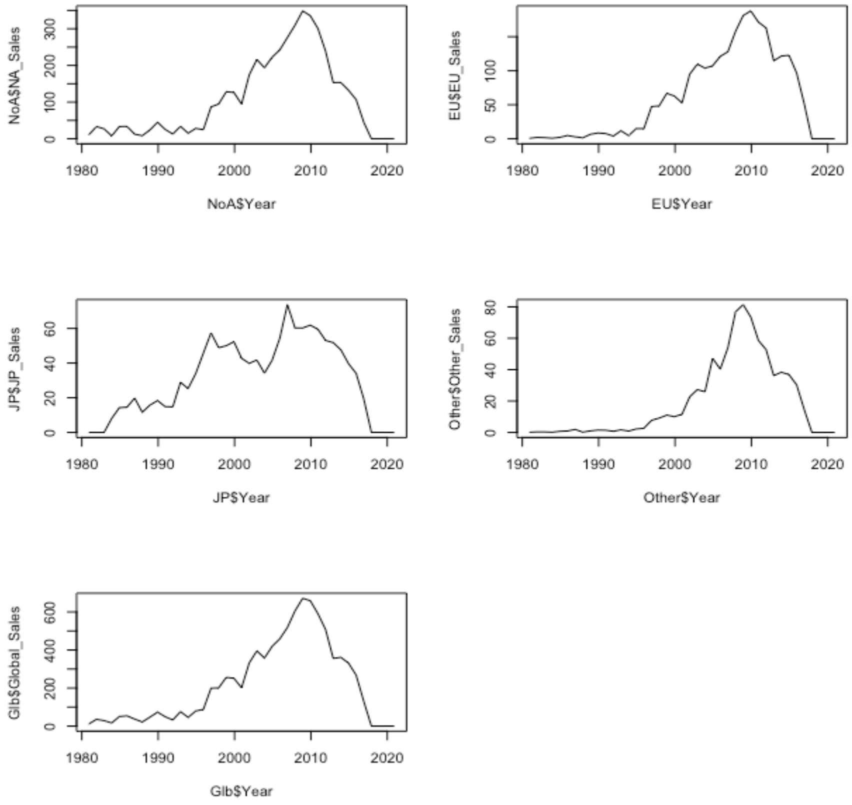
Rating:
ESRB ratings provide information about what's in a game or app so parents and consumers can make informed choices about which games are right for their family. Ratings have 3 parts: Rating Categories, Content Descriptors, and Interactive Elements.

Please see more information using this link: https://www.esrb.org/ratings-guide

3.2 Analysis:

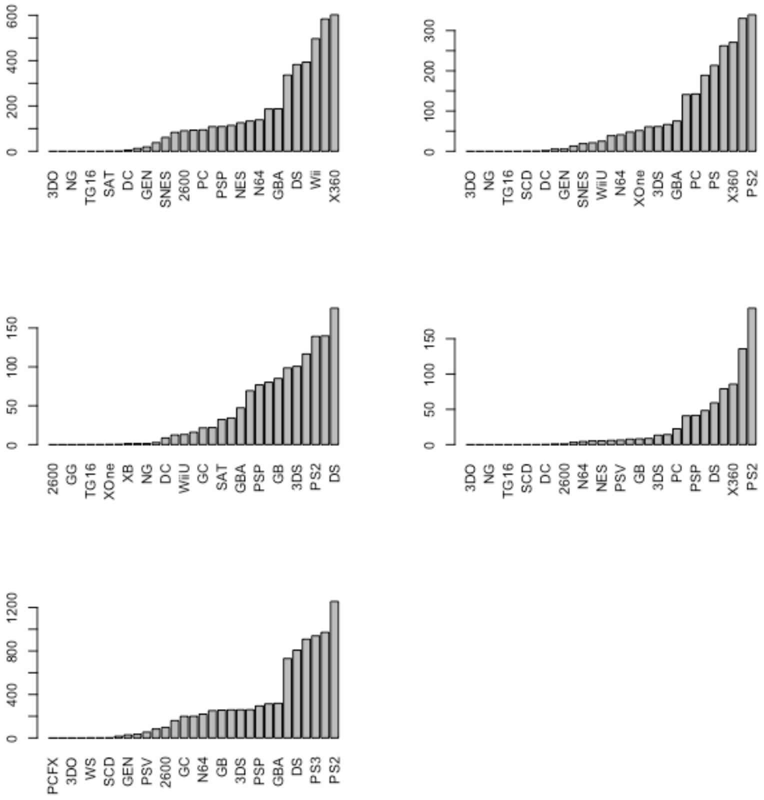
3.2.1. Sales of Year_of_Release

- We could see the sales trend in each region had the similiarity pattern. The 2010 is the year that sales reached the peek ineach region.
- Compared with other region, games sales in North American were much higher
- Because there was only 1 record of 2020 in the original dataset, the sales in 2020 was preety low.
- We can ignore the time effect to the sales according to the plot.



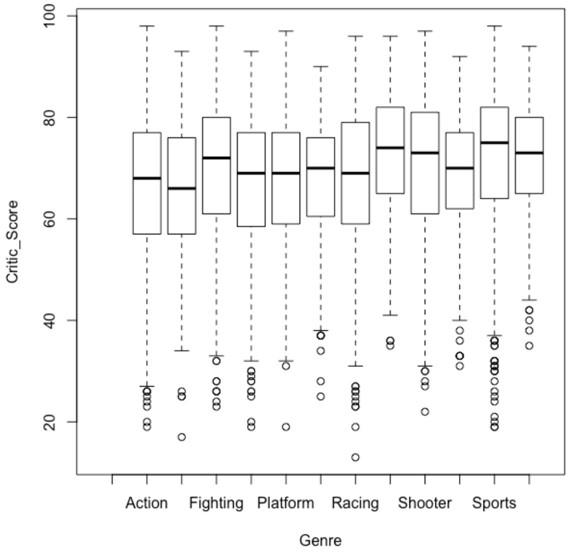
3.2.2. Sales of platform

- For each different region, the most popular platform was different. PS2 stays the most popular platform in the galbal games market.



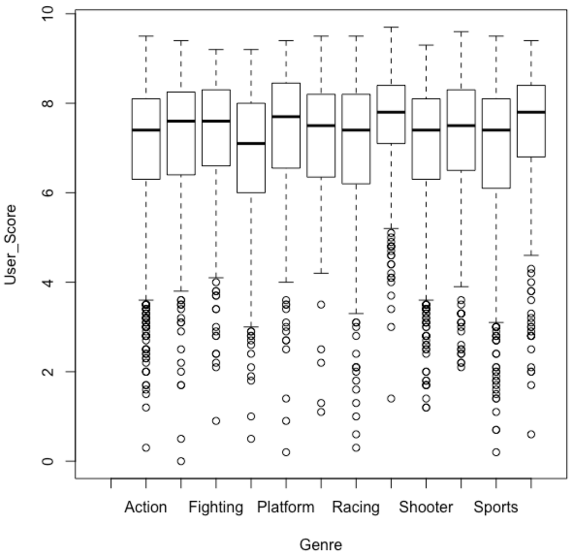
3.2.3. Critic score

- Aggregate score compiled by Metacritic staff
- Critic score range from 1 to 100
- The distribution of Critic score has similarity pattern for each genre



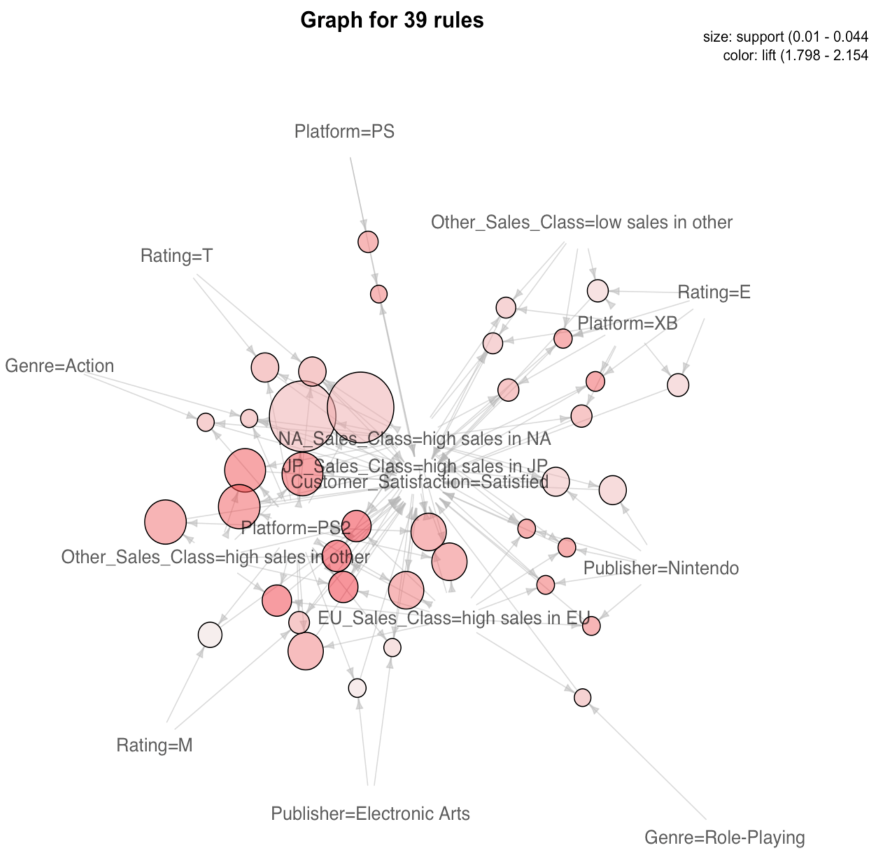
3.2.4. User_Score

- Scale from 1 to 10
- The distribution of User score has similarity pattern for each genre



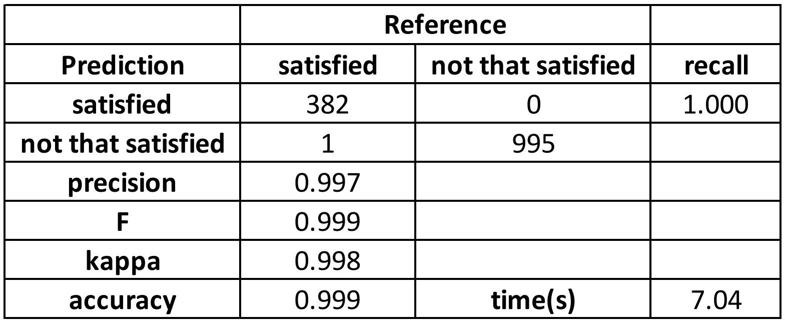
3.2.5 Association Rules

1. For the 39 rules, most rules' support concentrate in the range from 0.01 to 0.03. And the confidence range from 0.5 to 0.6
2. From graph for 39 rules below, for higher customer satisfaction of this game, the North American sales and Japan sales are very important because those indexes are in the center of graph. And also, the platform is PS, rating is T(which is game made for teenagers), and the game genre is action,these are the most important features of popular and higher satisfaction game.
3. For game publisher, Nintendo has high reputation for popular games
4. Role playing game and Action game have potential market


Graph for 39 rules
size: support (0.01 - 0.044
color: lift (1.798 - 2.154
Platform=PS
Other_Sales_Class=low sales in other
Rating=T
Rating=E
Platform=XB
Genre=Action
NA_Sales_Class=high sales in NA
JP_Sales_Class=high sales in JP
Customer_Satisfaction=Satisfied
Platform=PS2
Other_Sales_Class=high sales in other
Publisher=Nintendo
EU_Sales_Class=high sales in EU
Rating=M
Publisher=Electronic Arts
Genre=Role-Playing

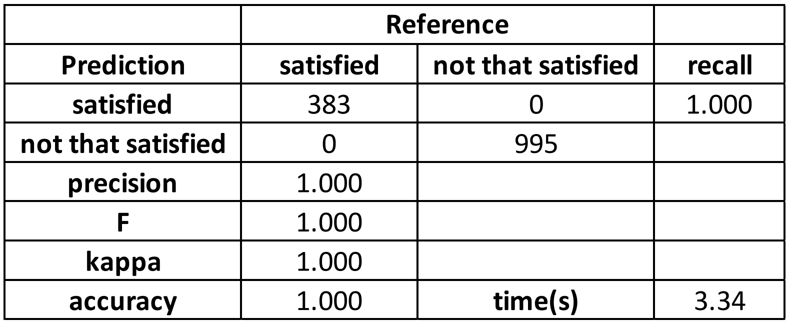
3.3. Classification Analysis

The results of each classifier are shown below.



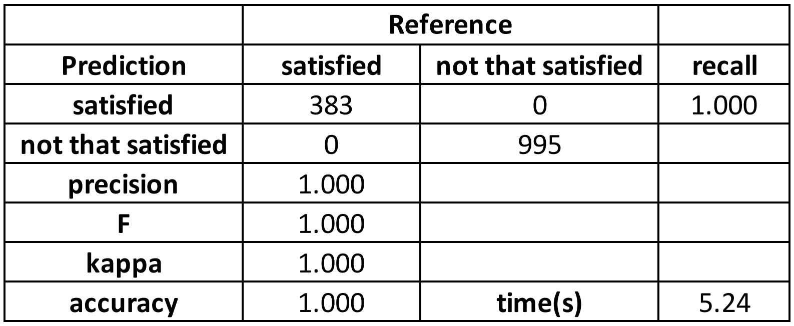
	Reference		
Prediction	**satisfied**	**not that satisfied**	**recall**
satisfied	382	0	1.000
not that satisfied	1	995	
precision	0.997		
F	0.999		
kappa	0.998		
accuracy	0.999	**time(s)**	7.04

Figure 3.3.1. The result of Naïve Bayes classifier



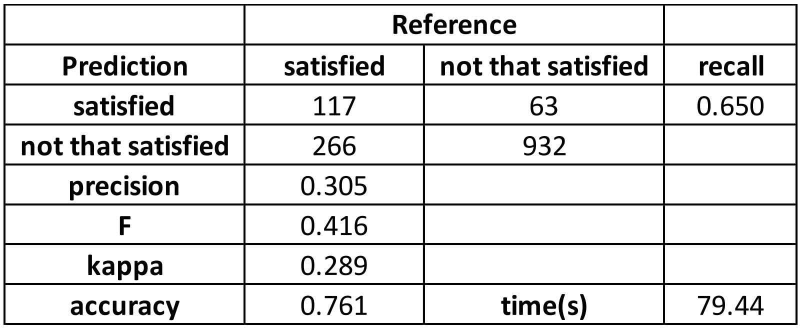
	Reference		
Prediction	satisfied	not that satisfied	recall
satisfied	383	0	1.000
not that satisfied	0	995	
precision	1.000		
F	1.000		
kappa	1.000		
accuracy	1.000	time(s)	3.34

Figure 3.3.2. The result of Bagged Cart classifier



	Reference		
Prediction	satisfied	not that satisfied	recall
satisfied	383	0	1.000
not that satisfied	0	995	
precision	1.000		
F	1.000		
kappa	1.000		
accuracy	1.000	time(s)	5.24

Figure 3.3.3. The result of Random Forest classifier



	Reference		
Prediction	satisfied	not that satisfied	recall
satisfied	117	63	0.650
not that satisfied	266	932	
precision	0.305		
F	0.416		
kappa	0.289		
accuracy	0.761	time(s)	79.44

Figure 3.3.4. The result of Decision Tree classifier

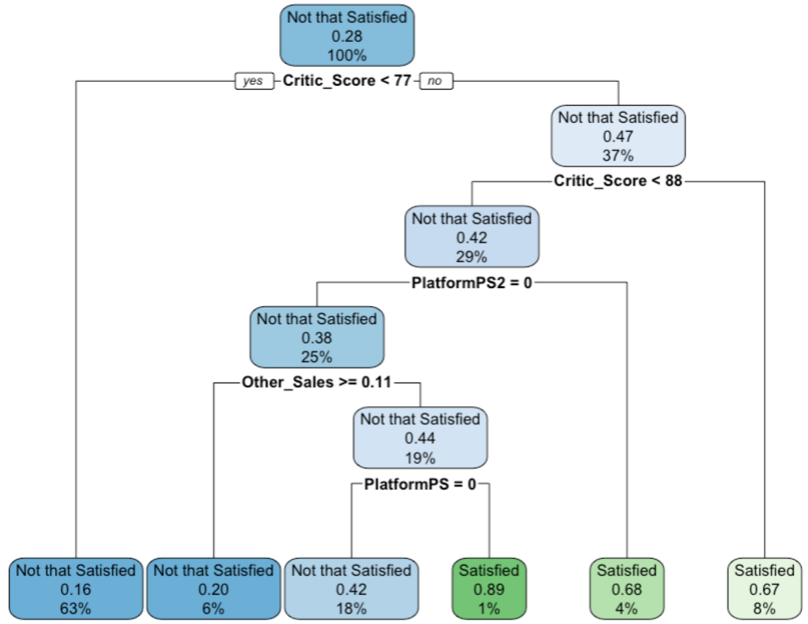


Figure 3.3.5. The tree generated with Decision Tree classifier

We decide to use kappa and F-Measure to evaluate the prediction. Because F-Measure combines the Precision and Recall, which can evaluate the model more objectively, and kappa show the agreement level of model's prediction. From the pictures above we can say the Bagged Cart classifier and Random Forest classifier do a better job in this step, because they both have the highest accuracy, which is equal to 1. And the values of their kappa and F are also the highest amount these classifiers. Also, these two classifiers took a very short time to run, which shows the efficiency of them.

IV. Conclusion

From the results of Data Exploration and Association Rules, we find the most popular platform for companies to develop games is PS, which full name is PlayStation. Therefore, if companies want to increase the popularity of their games, they can consider adding a version suitable for the PlayStation platform when the game is developed and released. The other fact we can know from the results is that teenagers are the main audience for video games, and the genre they like is Action game. In the light of this statement, developers can add action elements to the game according to the preferences of teenagers, thereby increasing the popularity and sales of the game. Furthermore, we find North America, Europe, and Japan are currently major market for video games sales, which means that, in the development process, companies should add the languages of these regions into the game. Then as the results of classification shown, companies can choose Bagged Cart classifier and Random Forest classifier to do a market forecast analysis, assessing the popularity of the released game.

References

[1] ENTERTAINMENT SOFTWARE RATING BOARD https://www.esrb.org/ratings-guide/

Aappendix

[1] List of association rules

lhs	support	confidence	coverage	lift	count
[1] {Platform=PS}	0.01261967	0.5686275	0.02219321	2.044923	87
[2] {Platform=PS,					
JP_Sales_Class=high sales in JP}	0.01029881	0.568	0.01813171	2.042667	71
[3] {Publisher=Nintendo,					
EU_Sales_Class=high sales in EU}	0.01102408	0.5714286	0.01929214	2.054997	76
[4] {Publisher=Nintendo,					
NA_Sales_Class=high sales in NA}	0.01740644	0.5263158	0.03307224	1.892760	120
[5] {Platform=XB,					
Rating=E}	0.01348999	0.5224719	0.02581955	1.878937	93
[6] {Platform=XB,					
NA_Sales_Class=high sales in NA}	0.01319988	0.554878	0.0237888	1.995477	91
[7] {Genre=Role-Playing,					
EU_Sales_Class=high sales in EU}	0.01029881	0.5419847	0.01900203	1.949109	71
[8] {Platform=PS2,					
Publisher=Electronic Arts}	0.01087903	0.5033557	0.021613	1.810190	75
[9] {Platform=PS2,					
Rating=M}	0.01508558	0.5	0.03017116	1.798122	104
[10] {Platform=PS2,					
Other_Sales_Class=high sales in other	0.02697998	0.5723077	0.04714244	2.058158	186

[11] {Platform=PS2,					
EU_Sales_Class=high sales in EU}	0.02277343	0.562724	0.04046997	2.023693	157
[12] {Platform=PS2,					
NA_Sales_Class=high sales in NA}	0.04366115	0.5384615	0.081085	1.936439	301
[13] {Publisher=Nintendo,					
NA_Sales_Class=high sales in NA,					
EU_Sales_Class=high sales in EU}	0.01102408	0.5757576	0.01914708	2.070565	76
[14] {Publisher=Nintendo,					
EU_Sales_Class=high sales in EU,					
JP_Sales_Class=high sales in JP}	0.01102408	0.5757576	0.01914708	2.070565	76
[15] {Publisher=Nintendo,					
NA_Sales_Class=high sales in NA,					
JP_Sales_Class=high sales in JP}	0.01740644	0.5263158	0.03307224	1.892760	120
[16] {Platform=XB,					
Rating=E,					
JP_Sales_Class=high sales in JP}	0.01160429	0.5797101	0.02001741	2.084779	80
[17] {Platform=XB,					
Rating=E,					
Other_Sales_Class=low sales in other	0.01319988	0.5170455	0.02552945	1.859422	91
[18] {Platform=XB,					
NA_Sales_Class=high sales in NA,					
JP_Sales_Class=high sales in JP}	0.01319988	0.554878	0.0237888	1.995477	91
[19] {Platform=XB,					

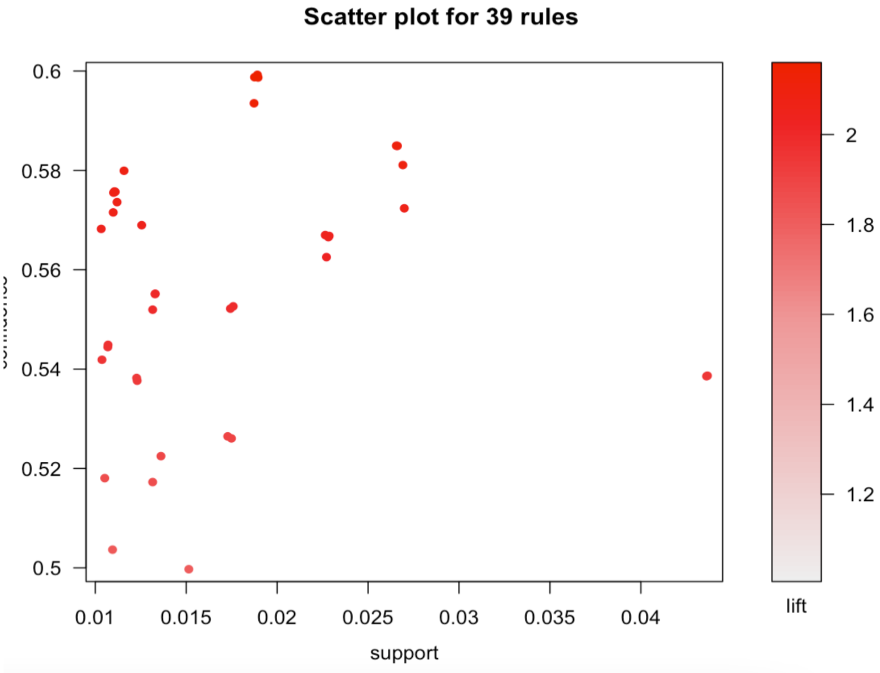
NA_Sales_Class=high sales in NA,					
Other_Sales_Class=low sales in other	0.01232956	0.5379747	0.02291848	1.934688	85
[20] {Platform=PS2,					
Publisher=Electronic Arts,					
JP_Sales_Class=high sales in JP}	0.01058892	0.5177305	0.02045257	1.861885	73
[21] {Platform=PS2,					
Rating=M,					
JP_Sales_Class=high sales in JP}	0.01305483	0.5521472	0.02364375	1.985656	90
[22] {Platform=PS2,					
EU_Sales_Class=high sales in EU,					
Other_Sales_Class=high sales in other	0.01885698	0.5936073	0.03176675	2.134757	130
[23] {Platform=PS2,					
NA_Sales_Class=high sales in NA,					
Other_Sales_Class=high sales in ot	0.02654482	0.5846645	0.0454018	2.102596	183
[24] {Platform=PS2,					
JP_Sales_Class=high sales in JP,					
Other_Sales_Class=high sales in ot	0.02697998	0.58125	0.04641717	2.090317	186
[25] {Platform=PS2,					
Genre=Action,					
NA_Sales_Class=high sales in NA}	0.01058892	0.5447761	0.01943719	1.959148	73
[26] {Platform=PS2,					
NA_Sales_Class=high sales in NA,					
EU_Sales_Class=high sales in EU}	0.02277343	0.566787	0.04017987	2.038304	157

[27] {Platform=PS2,					
EU_Sales_Class=high sales in EU,					
JP_Sales_Class=high sales in JP}	0.02277343	0.566787	0.04017987	2.038304	157
[28] {Platform=PS2,					
Rating=T,					
NA_Sales_Class=high sales in NA}	0.01755149	0.5525114	0.03176675	1.986966	121
[29] {Platform=PS2,					
NA_Sales_Class=high sales in NA,					
JP_Sales_Class=high sales in JP}	0.04366115	0.5384615	0.081085	1.936439	301
[30] {Publisher=Nintendo,					
NA_Sales_Class=high sales in NA,					
EU_Sales_Class=high sales in EU,					
JP_Sales_Class=high sales in JP}	0.01102408	0.5757576	0.01914708	2.070565	76
[31] {Platform=XB,					
Rating=E,					
JP_Sales_Class=high sales in JP,					
Other_Sales_Class=low sales in other	0.01131419	0.5735294	0.0197273	2.062552	78
[32] {Platform=XB,					
NA_Sales_Class=high sales in NA,					
JP_Sales_Class=high sales in JP,					
Other_Sales_Class=low sales in other	0.01232956	0.5379747	0.02291848	1.934688	85
[33] {Platform=PS2,					
NA_Sales_Class=high sales in NA,					

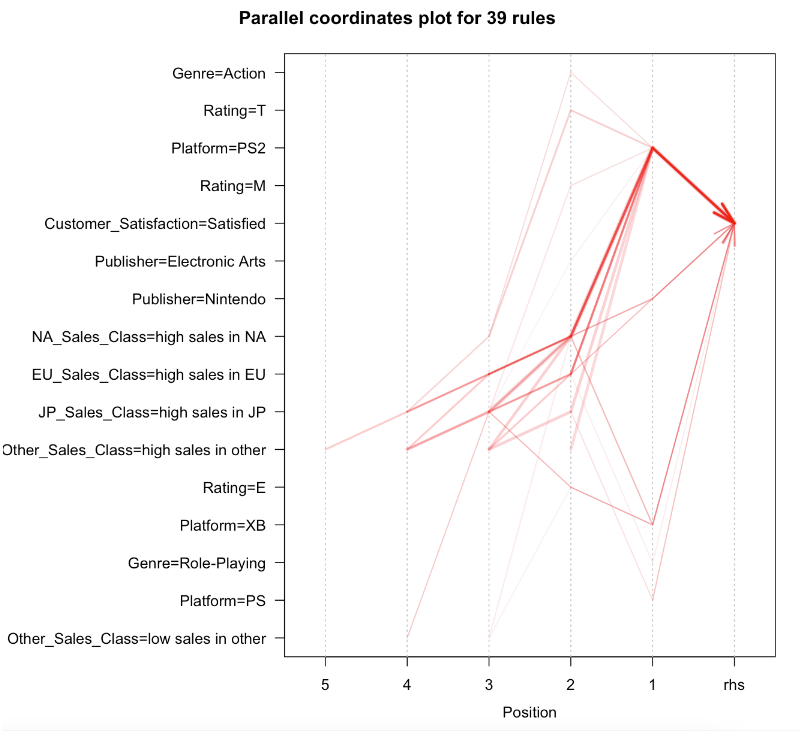
EU_Sales_Class=high sales in EU,					
Other_Sales_Class=high sales in other	0.01885698	0.5990783	0.03147665	2.154432	130
[34] {Platform=PS2,					
EU_Sales_Class=high sales in EU,					
JP_Sales_Class=high sales in JP,					
Other_Sales_Class=high sales in other	0.01885698	0.5990783	0.03147665	2.154432	130
[35] {Platform=PS2,					
NA_Sales_Class=high sales in NA,					
JP_Sales_Class=high sales in JP,					
Other_Sales_Class=high sales in other	0.02654482	0.5846645	0.0454018	2.102596	183
[36] {Platform=PS2,					
Genre=Action,					
NA_Sales_Class=high sales in NA,					
JP_Sales_Class=high sales in JP}	0.01058892	0.5447761	0.01943719	1.959148	73
[37] {Platform=PS2,					
NA_Sales_Class=high sales in NA,					
EU_Sales_Class=high sales in EU,					
JP_Sales_Class=high sales in JP}	0.02277343	0.566787	0.04017987	2.038304	157
[38] {Platform=PS2,					
Rating=T,					
NA_Sales_Class=high sales in NA,					
JP_Sales_Class=high sales in JP}	0.01755149	0.5525114	0.03176675	1.986966	121
[39] {Platform=PS2,					

NA_Sales_Class=high sales in NA,					
EU_Sales_Class=high sales in EU,					
JP_Sales_Class=high sales in JP,					
Other_Sales_Class=high sales in other	0.01885698	0.5990783	0.03147665	2.154432	130

[2] Scatter plot of 39 rules



[3] Parallel coordinates plot 39 rules


Parallel coordinates plot for 39 rules
Genre=Action
Rating=T
Platform=PS2
Rating=M
Customer_Satisfaction=Satisfied
Publisher=Electronic Arts
Publisher=Nintendo
NA_Sales_Class=high sales in NA
EU_Sales_Class=high sales in EU
JP_Sales_Class=high sales in JP
Other_Sales_Class=high sales in other
Rating=E
Platform=XB
Genre=Role-Playing
Platform=PS
Other_Sales_Class=low sales in other
5
4
3
2
1
rhs
Position