Annalaura Cranford
Destinee Sprinkle
Noelle Warner
CSC 272
12/4/2017

Is This Game Lame?: An Analysis of User Ratings and Sales of Video Games

Introduction:

In our analysis, we found that our initial ideas of how user and critic scores were mostly inconclusive; it seems as though a lower score doesn't necessarily affect sales after all. We did find that sales across the world are very related to each other. We also learned that we could take all of the other attributes and predict sales pretty accurately, if we had more data.

Background:

For our project, we chose to look at data about video game sales and their respective user and critic review scores scraped from a website called Metacritic. This website compiles critic reviews from various other games journalism websites, such as IGN, GameSpot, and Polygon, and gives each game a score based on the average sentiments from all of these sources. It also provides an average user score from reviews that users can input right on each game's page.

Our goal going into this project was to see if we could find any correlation between user or critic scores and sales for each game. This is a simple problem, but it speaks to something that is actually becoming a large problem within the games industry. Certain game studios, such as Bethesda or EA, are now refusing to send out review copies, or copies of the game distributed to reviewers and games journalists about two weeks before the public release so that they can write their reviews of the game. This is a problem because these initial reviews allow the consumer to get a sense of what this game is like, and whether or not they will want to buy it. Without review copies, it is impossible to know if a game is riddled with glitches or microtransactions.

While we thought this is an interesting problem that says something about human behavior (i.e. whether reviews have any real bearing on whether or not gamers will buy a certain title), we understood that this is a bit of a narrow topic. Our goal then became looking for other associations or correlations within the data that could tell us anything interesting, as much of data mining is about learning from what the data has to say that we never could have thought of or seen ourselves beforehand.

Dataset description:

https://www.kaggle.com/kendallgillies/video-game-sales-and-ratings

Our dataset was obtained from Kaggle, a website where people can post datasets that they create, which is a platform for "the world's largest community of data scientist and machine learning engineers." They house competitions for people to find the create the most accurate predictive model for various datasets. We simply searched for a topic of interest to us, video

games, and found a dataset that we thought would provide an interesting insight into how video games sell in accordance to their ratings.

The data includes:

Attribute	Description	Type(Value)
Name	The name of the game (no pun intended).	Nominal (over 11,000 unqiue entries)
Platform	The console(s) on which the game is available.	Nominal (Wii, Nintendo Entertainment System (NES), Gameboy (GB), Nintendo DS, Xbox 360(X360), Playstation 4 (PS4), Playstation (PS3), Playstation 2 (PS2), Super Nintendo Entertainment System (SNES), Gameboy Advance (GBA), 3DS, Nintendo 64 (N64), Playstation (PS), Xbox, Personal Computer (PC), Xbox One (XOne), WiiU, GameCube (GC), Genesis (GEN), Playstation Portable (PSP), Atari 2600 (2600), Dreamcast (DC), Saturn (SAT), Playstation Vita (PSV), Sega Mega-CD (SCD), TurboGrafx-16 (TG16), Panasonic 3DO (3DO), Game Gear (GG), Neo Geo Pocket Color (NG), NEC PC-FX (PCFX), WonderSwan (WS))
Year_of _Release	The year that the game was initially released. Lowest Year = 1980 Highest Year = 2020 Mean = 2006.489 Standard Deviation = 5.878	Numeric
Genre	The genre of the game.	Nominal (Adventure, Action, Fighting, Misc, Platform, Puzzle, Racing, Role-Playing, Shooter, Simulation, Sports, Strategy)
Publisher	The publisher of the game. (The developer is the studio that actually creates the	Nominal (There are 583 publishers, so that's probably too many to name here.)

	game, while the publisher produces the game and is in charge of distribution. Sometimes the developer and the publisher are on in the same.)	Here are a few: Atari, Sega, Nintendo, Warner Bros., Activision, THQ, Lucas Arts, Bethesda, Disney Interactive Studios, EA, Sony, Ubisoft, Universal, etc.
NA _Sales	Sales in the North American region in millions from release date to present. Lowest Sale = 0 Max = 41.36 Mean = .263286 Mean W/O 0's = .3605177 Standard Dev = .813518	Numeric
EU_Sales	Sales in the European region in millions from release date to present. Lowest Sale = 0 Highest Sale = 28.96 Mean = .1450278 Mean W/O 0's = .22354867 Standard Dev = .503331	Numeric
JP_Sales	Sales in Japan in millions from release date to present. Lowest Sale = 0 Highest Sale = 10.22 Mean = 0.0776 Mean W/O 0's = .2964896 Standard Dev = .308853	Numeric
Other_Sales	Sales in other regions in millions from release date to present. Lowest Sale = 0 Highest Sale = 10.57 Mean = 0.04733 Standard Dev = .186731	Numeric
Global_Sales	Global sales in millions from release date to present. Lowest Sale = 0.1 Highest Sale = 82.53 Mean = .5335214 Standard Dev = 1.5480	Numeric
Critic_Score	Score that critics gave the game on Metacritic.	Numeric

	Lowest Score = 13 Highest Score = 98 Mean = 68.96768 Standard Dev = 13.93816	
Critic_Count	Number of critics that reviewed the game. Lowest Count = 3 Highest Count = 113 Mean = 26. 36082 Standard Dev = 18.98049	Numeric
User_Score	Score that users gave the game on Metacritic. Lowest Score = 0 Highest Score = 9.7 Mean = 7.125 Stand Dev = 1.5	Numeric
User_Count	Number of users that reviewed the game. Lowest Count = 4 Highest Count = 10665 Mean = 162.2299 Stand Dev = 561.2823	Numeric
Developer	The name of the studio that developed the game.	Nominal (There are 1,698 developers in the dataset.) Here are a few: 2K, Activision, Amaze Entertainment, Atari, Bandai, Besthesda Softworks, Bioware, Nintendo, Destineer, Disney Interactive Studios, Dreamcatcher, EA Games, EA Canada, Microsoft Game Studios, Naughty Dog, etc.
Rating	The age rating for a game, similar to the rating system that movies have. Describes what is appropriate for different age groups.	Nominal (E, E10+, T, M, C Company, RP, EC, AO)

3. In order to get our data to work in Weka, we had to do quite a bit of preprocessing, which did end up taking a bit longer than we expected. First, we tried to open the file in Weka as a CSV, but we were given an error saying that there was a problem encountered on line 2. This was extremely vague, and line 2 in the CSV file seemed fine. This sent us on a journey of trying to

figure out what was wrong with this data. We opened our data in Excel as a CSV file, but our first problem was that if we tried to use a text editor to reformat the data as an ARFF file by hand, there were just too many possible attributes (over 11,000 unique entries) for game titles for this to be a viable option. Instead, we went back to Excel, copied all of the names of the games, and then transposed and pasted them in a new Excel sheet so that they were going horizontally across the page as opposed to vertically, so that they would then be in a comma delimited format. We then copied that into the text editor, and we had all of the attributes we needed.

The next problem was that we kept getting weird formatting errors, such as Weka was expecting 16 Tokens and received 19, when we tried to load it into Weka, even though it seemed like our formatting was fine. Our first idea was that there was text in a cell somewhere in the file that we couldn't see. We copied and pasted just the video game data into a new Excel file, but we still had the same error as we loaded the new Excel file into Weka. In the end, it turned out that our problem was that all of the attributes that had an apostrophe such as "Disney's Interactive Studios", were being seen as quotation marks, which Weka was looking for as a set. We were able to take out all of the apostrophes with find and replace and the data set.

Once we loaded it into Weka, we realized the Year of Release and User Score attributes were not being recognized as either nominal or numeric, which we didn't understand because we expected them to be numerical. We converted our CSV file to an ARFF file using Weka's arff viewer. We edited the file to change the word string to numeric in Year_of_Release and User_Score, but Weka still could not recognize the file as ARFF format. After some investigation, we noticed that there were some "tbd"s and "n/a"s hiding amongst the numerical values, which was causing the problem. We promptly removed them and the was working with the CSV and the Year_of_Release and User_Score attributes were numeric. After we got it all working, we had to take out the name attribute, because that could lead to overfitting as it is a unique identifier. We discretized the numeric attributes using unsupervised attribute discretize, with settings attribute indices first to last, bins 5, and use equal frequency true. Everything else was left the same. We also left Global_Sales as an attribute later once we realized this may be overweighting the sales attributes since European, North American, Japan, and other sales were also included in the dataset. All of the numeric attributes were discretized using an equal distribution scheme.We chose to use this scheme because our data was so skewed to one side that other discretization schemes would have laced all the instances in one bin.

Data set Analysis

To classify this data set we used the 10-fold cross validation scheme to make our test and training data. This uses 10 different folds of the data as the test data, meaning that the first time the model runs, it uses the first n rows of the data set as the training data. The second time it runs, it uses the next n rows of the data set as the training data. This means that ultimately every instance was at some time used as part of the test data and as the training data.

Table 1) Dataset With Missing Values in all attributes and 0's in each Sale attribute included. This data is the original dataset after we pre-processed it.

Algorithm	**Classifying Attribute**	**Accuracy**

ZeroR	NA_Sales	26.9698%
ZeroR	EU_Sales	35.1241%
ZeroR	JP_Sales	62.9016%
OneR	NA_Sales	58.1693%
OneR	EU_Sales	43.7272%
OneR	JP_Sales	69.1235%
Naive-Bayes	NA_Sales	66.8023%
Naive-Bayes	EU_Sales	62.7221%
Naive-Bayes	JP_Sales	70.673%
J48	NA_Sales	67.0894%
J48	EU_Sales	65.7015%
J48	JP_Sales	71.5525%
IBK k(5)	NA_Sales	69.297%
IBK k(5)	EU_Sales	66.1382%
IBK k(5)	JP_Sales	71.8456%

Table 2) with missing values, without 0's in Sales attributes

Algorithm	**Classifying Attribute**	**Accuracy**
ZeroR	NA_Sales	26.9698%
ZeroR	EU_Sales	35.1241%
ZeroR	JP_Sales	62.9016%
OneR	NA_Sales	46.4014%
OneR	EU_Sales	45.4981%
OneR	JP_Sales	70.9423%
Naive-Bayes	NA_Sales	52.9106%
Naive-Bayes	EU_Sales	54.837%
Naive-Bayes	JP_Sales	73.5208%

J48	NA_Sales	49.6081%
J48	EU_Sales	66.0503%
J48	JP_Sales	73.437%
IBK	NA_Sales	52.0251%
IBK	EU_Sales	67.2462%
IBK	JP_Sales	74.0233%

Table 3) Sales 0 Replaced with Mean and Missing Values deleted. This is the data set after we removed the missing values from the attributes Critic_Score, Critic_Count, User_Score, User_Count, Developer, and Rating and replaced the instances where the Sales equaled zero with the mean sales for that region.

Algorithm	**Classifying Attribute**	**Accuracy**
ZeroR	NA_Sales	34.3458%
ZeroR	EU_Sales	32.3881%
ZeroR	JP_Sales	70.8363%
OneR	NA_Sales	42.0901%
OneR	EU_Sales	38.8945%
OneR	JP_Sales	66.3308%
Naive Bayes	NA_Sales	51.9793%
Naive Bayes	EU_Sales	49.3019%
Naive Bayes	JP_Sales	75.6145%
J48	NA_Sales	42.5363%
J48	EU_Sales	43.5872%
J48	JP_Sales	72.0023%
IBK k(5)	NA_Sales	49.6329%
IBK k(5)	EU_Sales	49.9928%
IBK k(5)	JP_Sales	75.7305%

To start, we weren't quite sure what we were going to do, as there are so many different

directions we could have taken this in. We began where data analysis usually begins, by running the ZeroR algorithm on our data. This algorithm is a baseline test; it looks at the data to see which of the class attributes was most common and chooses that. As you can see from the table, the Japanese Sales is by far the most accurate with a 62.9016% accuracy. It's unusual that ZeroR for Japan is so high. This is because there are many zeros in the JP_Sales attribute, therefore it can easily classify it as 0. This was gathered with the missing values intact. However, when the missing values were removed, the accuracy went up to 70%. After this, we decided the remove all the missing values from every instance our dataset because they do not necessarily mean something, just that the review study was not done for that game. We still produced similar results. This is when we saw that our data set had a lot of zeros in the place of the value for the sales. We decided it would be worth while to see how our numbers would change if we removed them. We hypothesised that these zeros could be missing values and not that the game had no sales in the region, although we cannot prove if this is true. As you can see in Table 2, this negatively affected our accuracies because our data set decreases from around 16,000 instances to around 2,000. Since we did not have enough data to accurately predict the sales, the next step took was to replace the zeros with the mean sales for the region. Combining both these tactics gave us higher accuracies because it is now giving us a better representation of the real world if our hypothesis is true. We then took these data sets and applied OneR, Naive Bayes, J48, and IBK with a K of 5. For each of these, we saw that the accuracies are around the same.

Our next algorithm we used was OneR. OneR finds the attribute that predicts the class attribute with the highest accuracy. The average accuracies were much higher compared to ZeroR, and the Japanese Sales still came out the highest accuracy. This is because it is using one attribute versus absolutely no attributes. One attribute used is better than none.

The accuracies were even higher compared to OneR using the Naive-Bayes algorithm, and were again highest with the Japanese sales. Naive-Bayes multiplies the probability of each attribute by the a priori probability. The a priori probability is the probability that the entire instance will happen. This is probably not the best algorithm for our dataset since our data in sales is severely skewed to the right, and this particular data set needs a equally distributed curve to work properly.

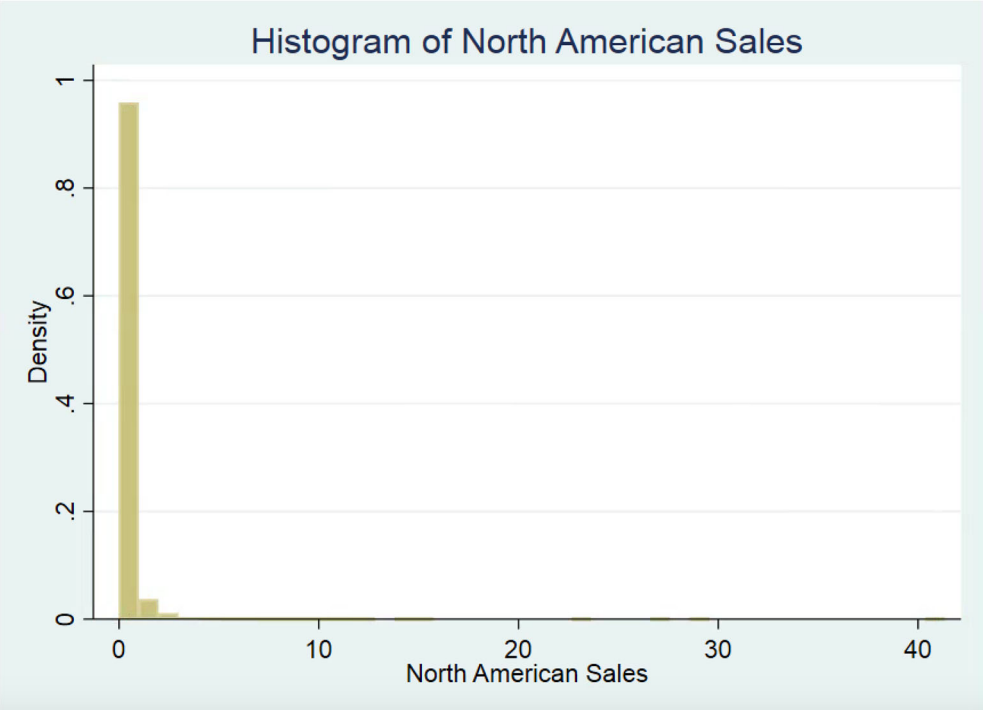
The J48 algorithm had particularly high accuracies, which we were quite excited about. However, after we visualized the tree, we realized that we were giving too much weight to the Global Sales attribute, making the tree a bit redundant. This wouldn't be as helpful for someone who was trying to use our algorithm to predict sales because they would need to know Global Sales before predicting NA Sales, which doesn't make sense because they would have had to have already sold the game. Once we removed the Global Sales attribute, we saw that this was a problem with all of the sales. We then conducted J48, and the other algorithms again, but this time with only the sales for the region we were trying to predict.

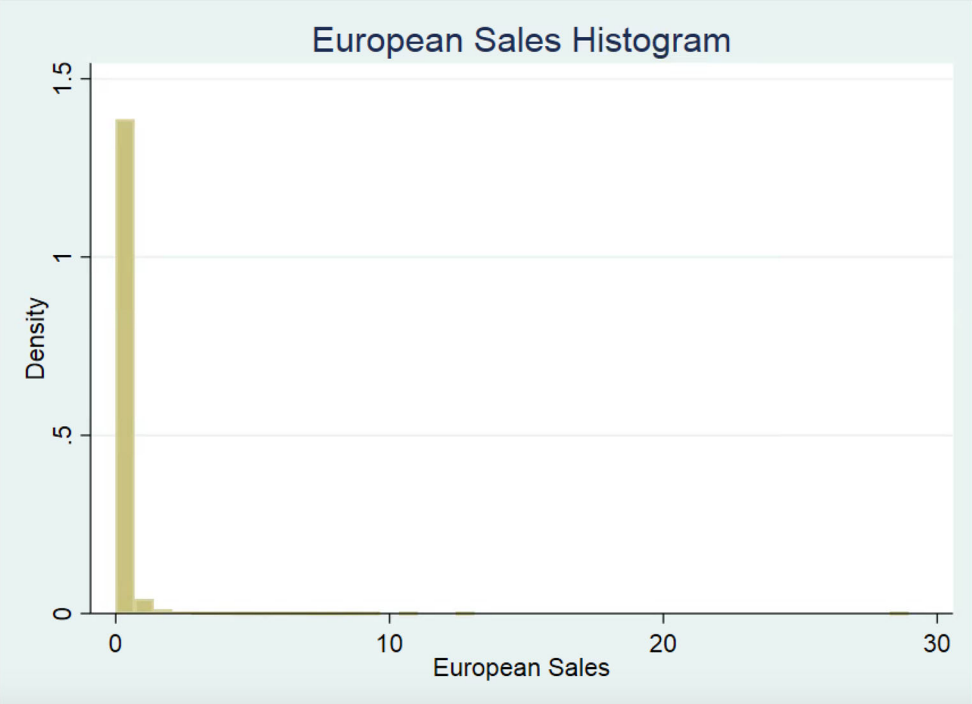
The next algorithm we used was IBk, which is sometimes known as the nearest neighbor algorithm. We used a K of 5, which finds 5 most similar instances and then averages them based off of the euclidean distance formula to classify the new instance. We thought this would be useful if the video game studios wanted to know how a new video game they had developed, but not yet released would do on the market, and had the data for user and critic ratings, developer, platform, etc.

Both the User and Critic Scores are lower than the sales percentages for each algorithm; this suggests that the scores are not nearly as predictive as we may have originally believed. For some reason, Japanese Sales always seem to have a much higher accuracy than any of the other attributes.

Results:

We first created some visualizations to get to know our data a little better, shown below.


Histogram of North American Sales
Density
0
.2
.4
.6
.8
1
0
10
20
30
40
North American Sales


European Sales Histogram
Density
0
.5
1
1.5
0
10
20
30
European Sales

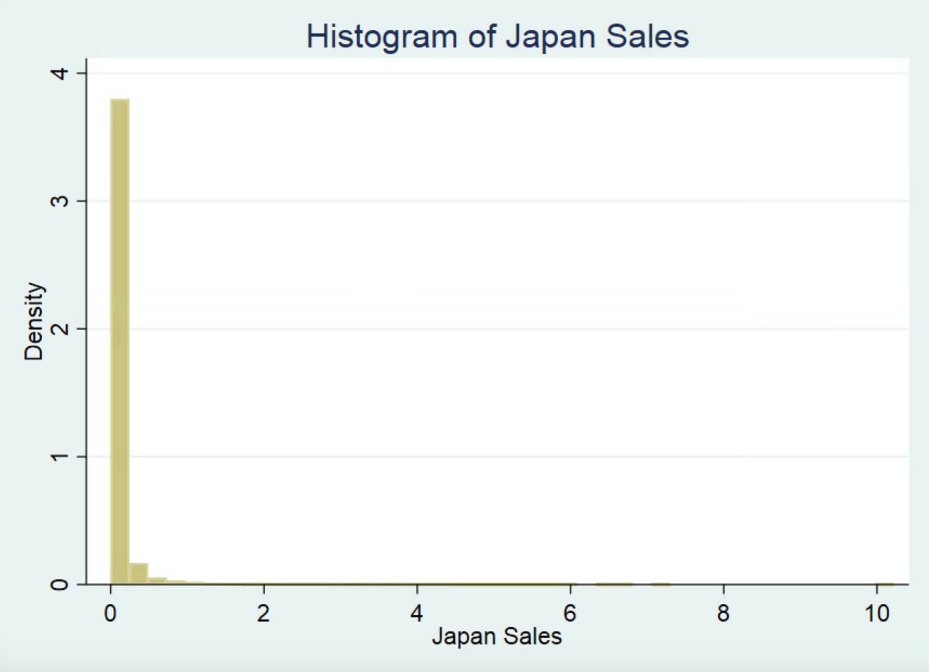
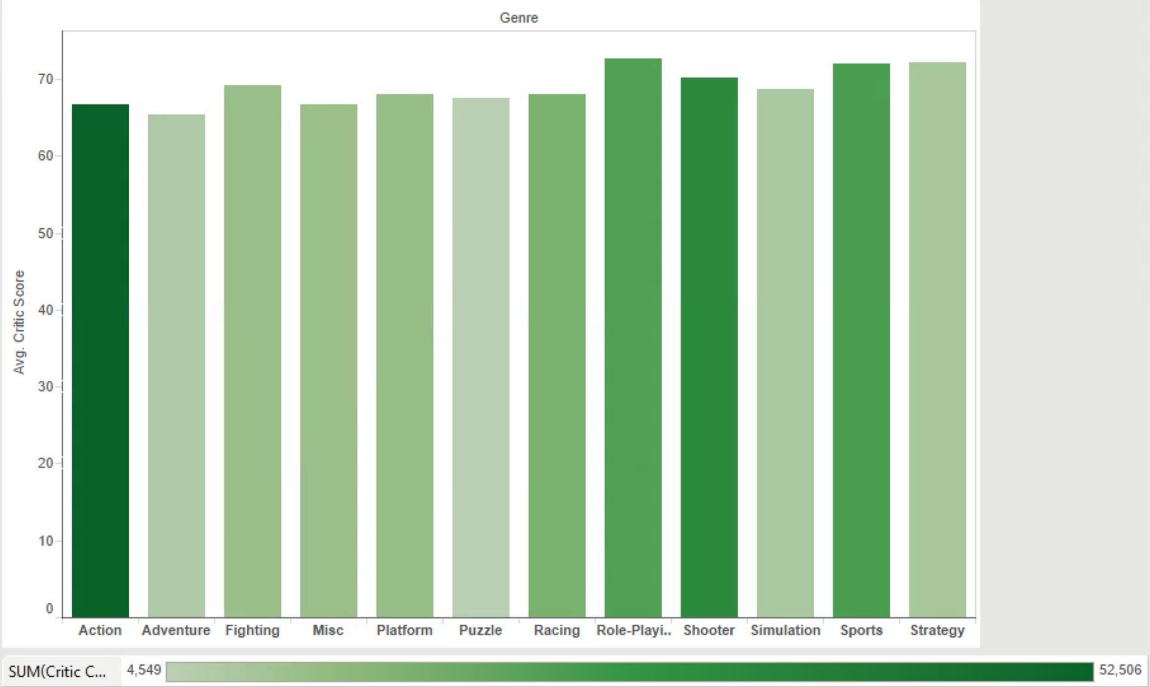


Figure 1) The graphs above display frequency of the amount of times a certain number occurs in North American Sales, European Sales, and Japan Sales. As we can see, the graphs are severely skewed to the right with 0 being the most common number in all cases. Our predictions would have better with a more normally distributed curve because we wouldn't have huge making the mean of the data bigger than it actually is.



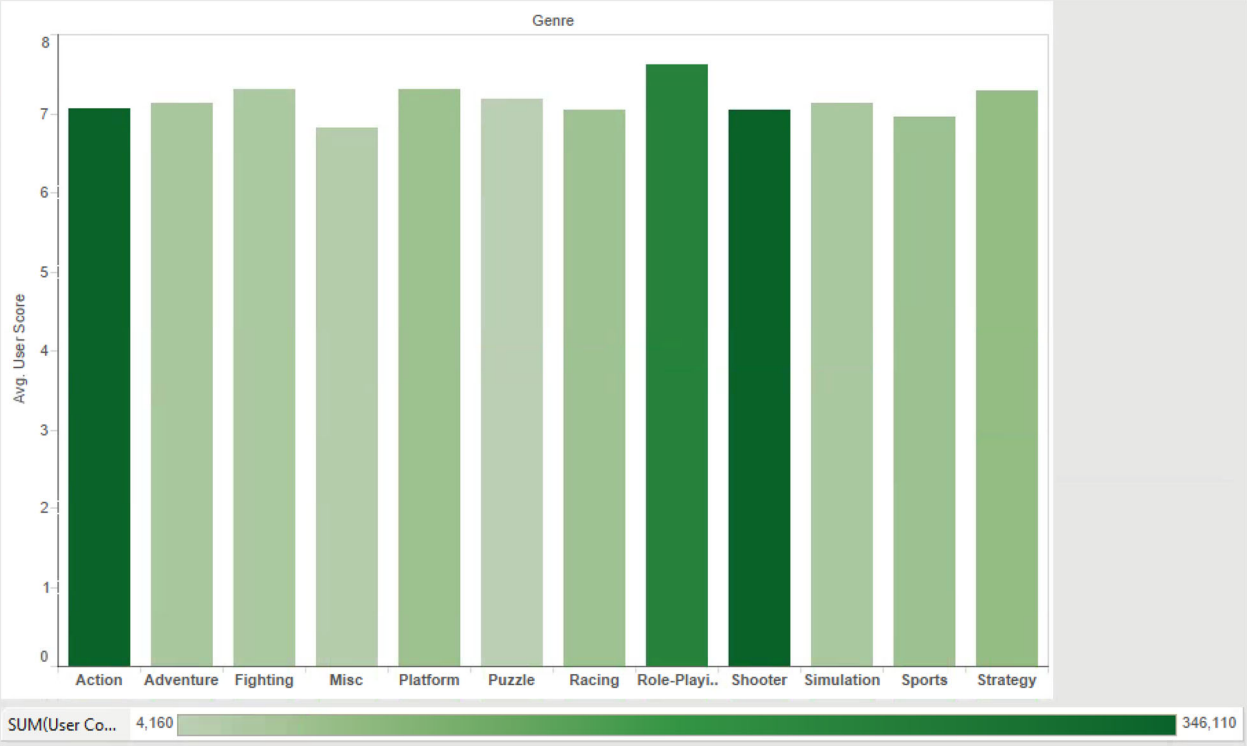


Figure 2) This bar graphs above demonstrate that both critics and users rate action, shooter, and role-playing games more than others. All game are rated relatively the same for both critics and users, but role-playing games have the highest ratings overall. It would be interesting to see if all users rated each genre of games. This shift the scores in a better way that reflects what the true population would approve. For instance, people who enjoy playing puzzle games are rating puzzle games, but would people who play action games rate puzzle games as high?

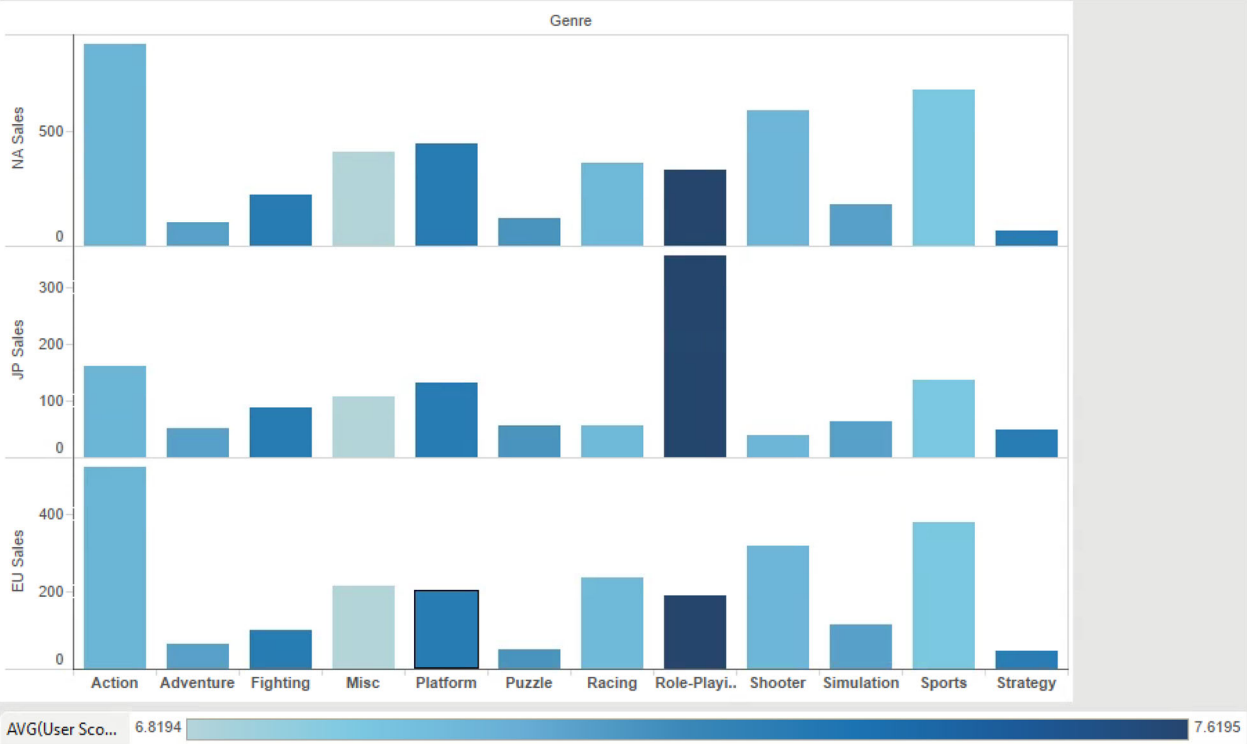


Figure 3)

The graph above displays sales in North America, Japan, and Europe in each genre. The color displays the average user score with deeper blue the higher the score. (As we can see in the description of user score below, the spread is very narrow from 6.8194 and 7.6195. This makes sense from Figure 2., which shows that most scores are similar across genres). Role-playing is sold much more in Japan compared to Europe and North America. In Japan role-playing has the highest sales, whereas in both Europe and North America, Action video games have the highest sales. The graph also shows that the role-playing genre is particularly popular in Japan, as it actually originated in that country. Japan produces more role-playing games than any other genre, and their games are popular in other regions of the world as well.

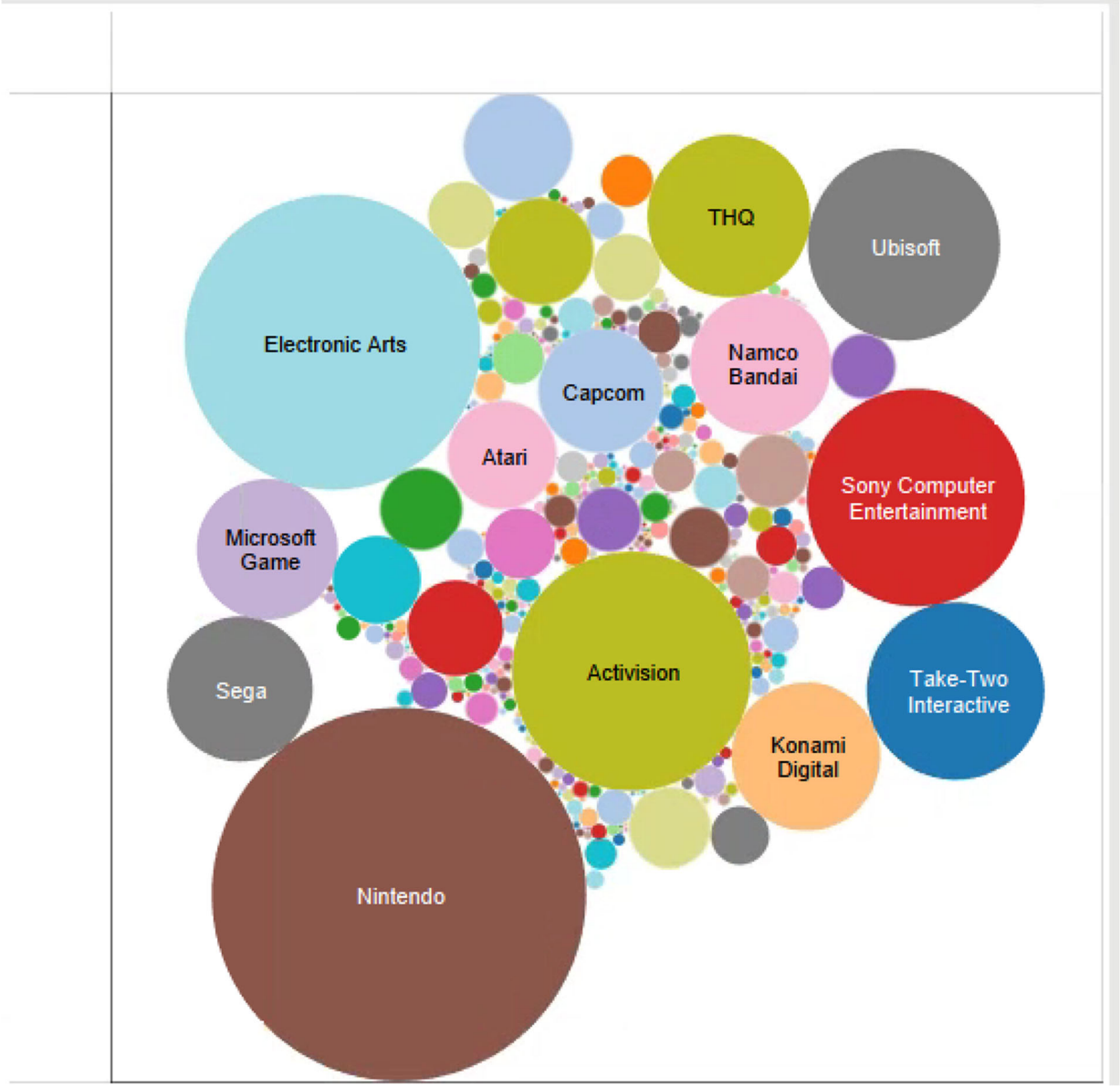


Figure 4)
The chart above shows different publishers as bubbles and color with a bigger bubble representing higher sales. Nintendo is the best publisher with the highest sales.

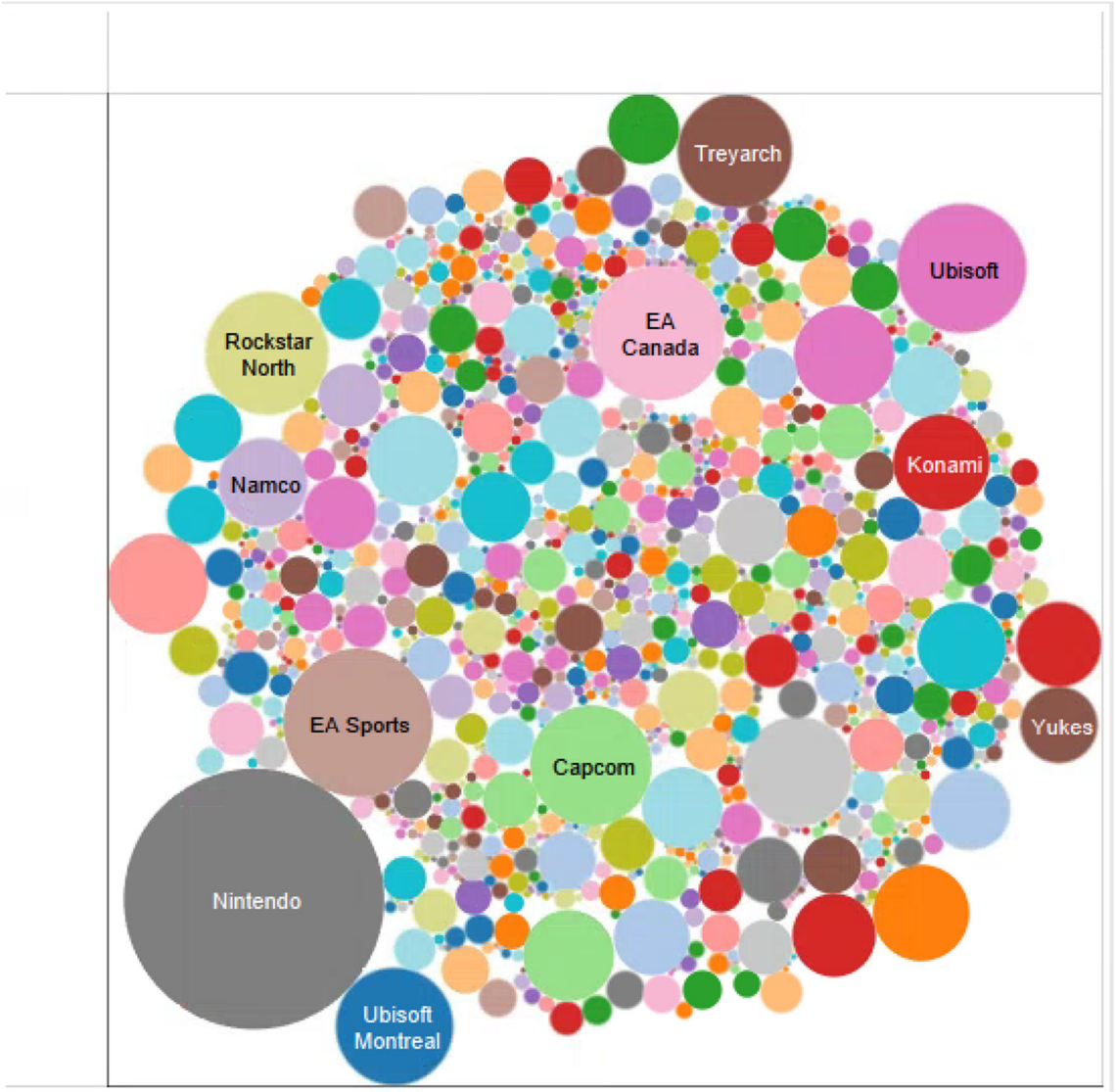


Figure 5)

The picture above explains with developer produces the highest global sales. In this graph we all removed the null values since many developers were left as null in the dataset. Nintendo is also the best developer with the most amount of global sales.

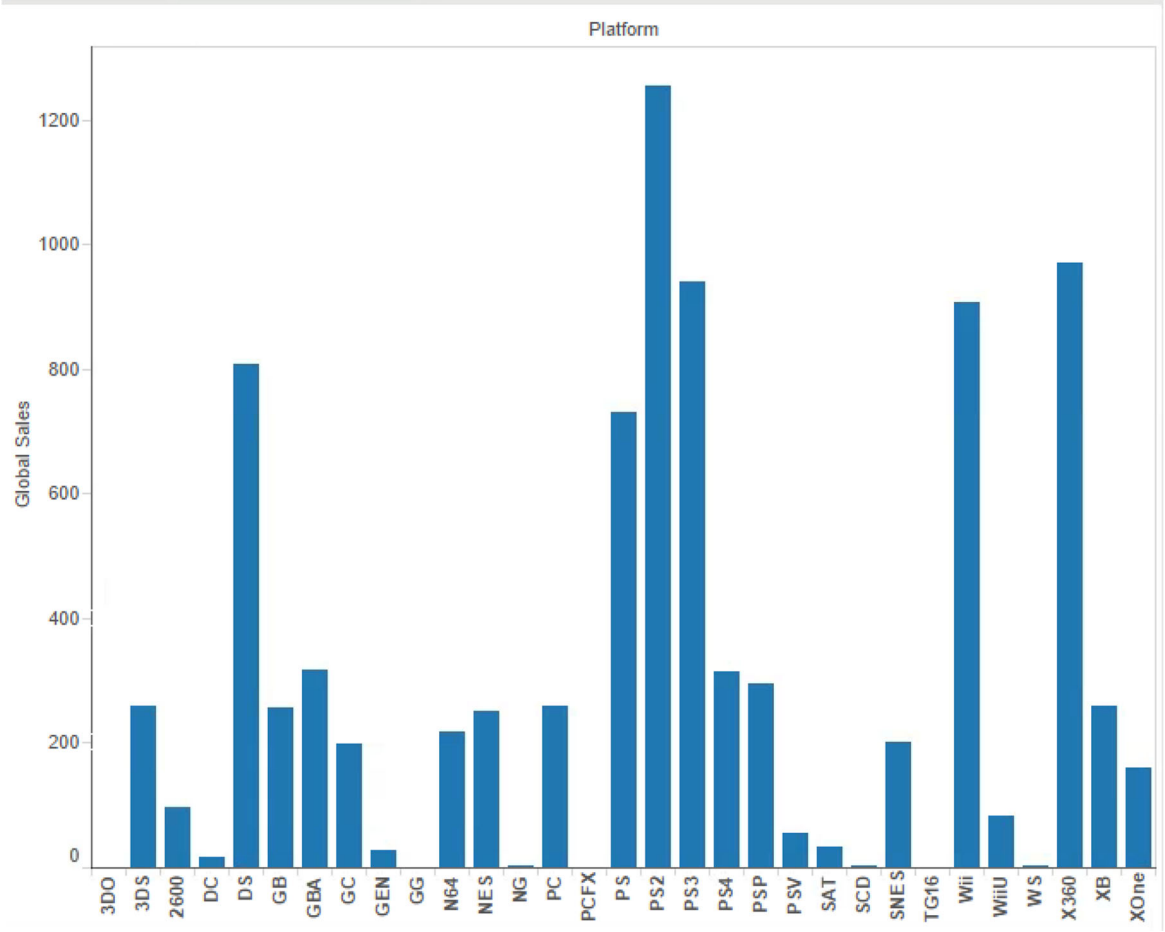


Figure 6)

The graph displays which platform for video games produces the highest amount of global sales. PS2 in our dataset has the highest amount of global sales.

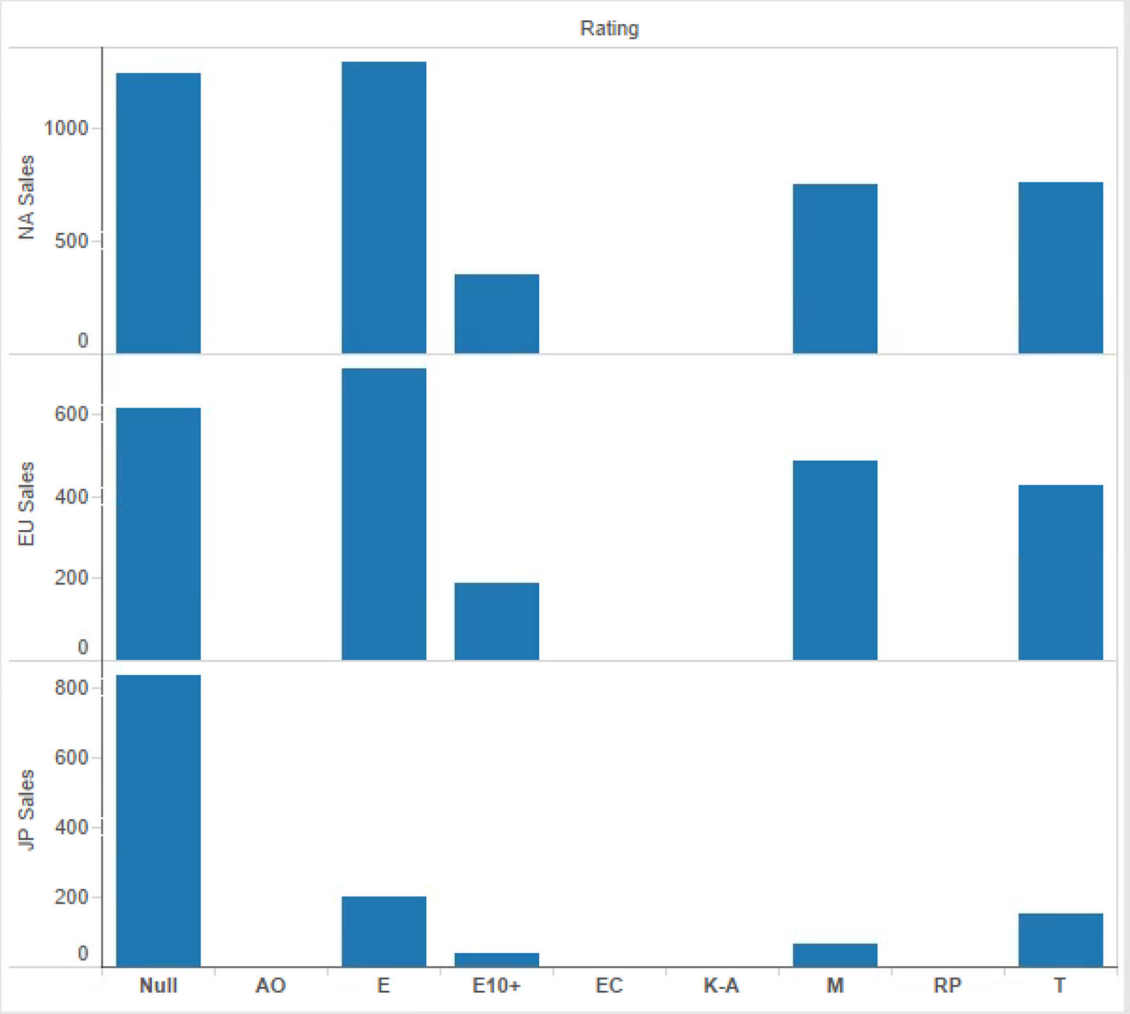


Figure 7)

The graph above displays rating against North American, European, and Japan Sales. Excluding the null values, E rated games have the highest sales in most all areas. AO, EC, K-A, and RP rated games are played the least.

Our original hypothesis when we first started this project was that reviews would be a predictor of sales, but after we ran the data through Weka, we found that the percent accuracy for this relationship was actually pretty low, around 40%. Though this is opposite than what we thought, it may actually show that the way reviews are handled in the world of video games isn't as much of a big deal. This result could have also stemmed from the many missing values we had in the reviews section of our data.

This suggests that reviews are not taken into very much account when consumers are looking for what to buy, which could perhaps the withholding of review copies is pointless for developers because the reviews don't have much to do with sales anyway, except for a few extreme cases. In the same breath, this would also mean that gamers have no reason to be upset about not getting to read reviews for a game in regards to hearing about whether it is

worth buying or not, although they may be upset about not getting a teaser about what's in the game.

One of our most inconclusive results is that the attributes we had were not that predictive of sales in a particular region. Though we felt this was not the best result, we understand that this may not be due to a lack of connection, but due to the skewed distribution of our data and to the sparseness of our data set.

The second this we found from our data is that the regional sales are all correlated to one another. Below are the association rules generated by Weka:

1. eu_sales='(-inf-0.035]' other_sales='(0.043669-0.048669]' 783 ==> jp_sales='(0.0738-0.0788]' 746 <conf:(0.95)> lift:(1.34) lev:(0.03) [191] conv:(6.01)
2. na_sales='(0.266643-inf)' other_sales='(0.105-inf)' 1116 ==> eu_sales='(0.147514-inf)' 1013 <conf:(0.91)> lift:(2.8) lev:(0.09) [651] conv:(7.26)
3. eu_sales='(0.147514-inf)' other_sales='(0.105-inf)' 1117 ==> na_sales='(0.266643-inf)' 1013 <conf:(0.91)> lift:(2.64) lev:(0.09) [629] conv:(6.98)
4. other_sales='(0.105-inf)' 1237 ==> eu_sales='(0.147514-inf)' 1117 <conf:(0.9)> lift:(2.79) lev:(0.1) [716] conv:(6.91)
5. other_sales='(0.105-inf)' 1237 ==> na_sales='(0.266643-inf)' 1116 <conf:(0.9)> lift:(2.63) lev:(0.1) [691] conv:(6.66)

Each one has something to do with sales. This proved tricky for us when trying to predict sales, without having another region's sales be included.

Conclusion:

From our work, we found that reviews are not as big of a deal in sales than the industry makes them out to be. We also discovered that having more data, as well as the data we have being more complete, would have been quite helpful (although this seems to be a theme for most data analysts). Through association learning, we also discovered that the global sales have a strong correlation, meaning that if a game sells well in one region they will sell well in another.