Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

A Case Study of Evaluating sales of video game with Data Mining Tools and CRISP-DM Methodology

Stage1

Yingzhang Liu 440517742

Business Understanding

o Determine Business Objectives

🕮 Background

GameShop is a retailer that sales video games. It is a small and new company which is located in Sydney and was established in 2016 by Jack Jackson. Jack was studying Bachelor of Engineering in the Empire State University. Jack loves video games so much, he was always dreaming about selling video games in the future. After his graduation, he got a huge investment from his father and then he started his company.

In this company, Jack hires four people as salesman, and their job is sell video games in the physical store. He also hires a manager who should in charge of employees' daily work performance and evaluation, customers' complaint and he can make decision in some urgent situation. And also, there are two accountants who record daily income, daily purchase. For Jack, he purchases video games.

Jack is new in this industry and he has no experience doing this. In his company, there is no professional analysis group and he doesn't have analysis skills to help him do his job. He doesn't know how many number for each game he should purchase. So for every hot video games, he purchases 200 for each of them. But he didn't really do the research for these hot games. He thought those hot games are popular because he saw their post and advertisement everywhere. He doesn't spend time on the sale statistics of each game. For other games, which are not "hot", he purchases 100 for each of them every time.

After several months, when Jack check the monthly record about the income. He found that the income is much less than expense. And also, there are a lot of "hot" games that are still stay in the warehouse.

To save his own business, Jack started to look for the solution for his problem. He thought the problem might be the poor management of the manager and the poor skills of the salesmen. He fired one of the salesmen and asked the manager to put more advertisements and sale promotion. It became a little bit better, but when Jack purchased new games again, the problem came out again.

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

The current solution could attractive more customer to come, but at the meantime it also decreases the profit that Jack could get. It helps sell more games, but every time after Jack purchase new game, a lot of games pile up in the warehouse and it is not easy to sell all of them.

🖃 Business Objectives

Jack wants to make the profit reach the maximum, he doesn't think sale promotion would be a long-term strategy and at the same time, Jack would like to know how to purchase game every time, in other words, he wants to find out which game he should purchase more, which game is easy to sell, so that later there wouldn't be much game left in the warehouse.

🖃 Business Success Criteria

Objective:

• Get more profit from selling the games.

• Purchase should be more reasonable. In other words, the amount of purchase for each game should be different according to its popular level.

Subjective:

• In the next season's evaluation report, the number of profit should grow up instead of getting lower.

• The number of remaining games should be in a reasonable range.

• The number of less popular game should not be more than the number of popular game.

o Assess Situation

This is the first time that Jack attempt at data mining, he decided to consult a data mining specialist to help him get started and try to analyze data to know different sale statistic for each game in genre, published year, published area and so on.

🖃 Inventory of Resources

Hardware: A computer that can access the sales data of every video game using excel.

Data: The data is about the number of sales of video games. The number of sales include sales in north America, Europe, Japanese and other places. The data is public and can be accessed by a website called Kaggle. People have to sign up to access the data.

Personnel: Since Jack just started his small business and he has no experience doing data mining before, there is no any data analysis group or department in

his company. But Jack hired a data mining specialist. If this data mining technique is helpful, Jack will consider to have specialized data group.

🗁 Requirements, Assumptions, and Constraints

Requirements:

• The data is generated by vgchartz.com, and the research result of the data should be securely kept in company's computer and only be accessed by Jack and the data mining specialist.

• The data mining specialist should submit the final report about the data within 14 days.

• The report should show the top genre by revenue in current year.

• The report should show the best-selling games in current year.

• The report should show the top publisher by revenue in current year

• The report should show the top platform by revenue in current year

Assumptions:

• Jack couldn't afford the price to hire the best data mining specialist, the one he hired might not provide the correct result that he wants.

• The data is found on website, the quality of data might not be 100% accurate.

• Jack want to simply view the results, and he expects the specialist could explain any diagram, which is generated by the data, to him.

Verify Constraints:

• The dataset is public resource, it is free to access, and could be used for private purpose, but need sign up in the website first.

• The budget just cover the salary for the specialist. There is no budget for purchasing data mining tools. They are just using excel. Other tools might be needed and purchased in the future.

🗁 Risks and Contingencies

Risk Contingency Plan The specialist might be absent or the work might get delayed.

Hire one more specialist.

The data quality could not be guaranteed then the result could not be completely accurate.

Check the data before getting started.

Even the result would come out, there is still some chance that result would not solve Jack's problem.

Have to find problems in other aspect for example, marketing strategy.

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

The process could be time consuming.

Pay extra money for overtime work or hire more workers.

🗁 Terminology

Data Mining: The practice of examining large pre-existing databases in order to generate new information.

CRISP-DM: A freely available model that has become the leading methodology in data mining. It provides guidelines for organized and transparent execution of any project.

🗁 Costs and Benefits

The data is free to collected and the cost would be salary for the specialist. Also, some other tools might be needed, that could be considered as potential cost. After the data mining done, Jack would know how to purchase video games each time. And that would increase the profit for his company and he would not deal with a large amount of video games each time.

o Determine Data Mining Goals

🗁 Data Mining Goals

Use historical data to generate diagram for:

• The top genre by revenue in current year.

• The best-selling games in current year.

• The top publisher by revenue in current year

• The top platform by revenue in current year

🗁 Data Mining Success Criteria

• The diagram should be easy to understand by Jack.

• The data mining would provide specific number.

• The different between each category should be clearly to see.

o Produce Project Plan

🗁 Project Plan

Phase Time Resources Risks Business understanding

1 day Data mining

specialist

Economic change

Data understanding

3 days Data mining

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

Specialist

Data quality problems, technology problems

Data preparation 5 days Data mining

specialist

Data quality problems, technology problems Modeling 2 days Data mining

specialist

Technology problems, inability to find adequate model Evaluation 2 day Data mining

specialist

Economic change, inability to implement results Deployment 1 day Data mining

specialist

Economic change, inability to implement results

🖃 Initial Assessment of Tools and Technique

Use excel to understand and analyze the data at the beginning step. Data Understanding

o Collect Initial Data

🖃 Initial Data Collection Report

The data they use is a public resource on Kaggle website. Once sign up on the website, the dataset is free to download and use. The dataset contains sale statics for more than 16,500 games. It has detail information like rank, game name, genre, publish year, platform, publisher and number of sale in different area, these information could be really helpful because it could compare the number of sale in many ways. At the current situation, Jack has no plan to purchase extra datasets, he only provides the data that was found on Kaggle website, and ask the specialist to use it. However, for further research, if higher requirement is needed they might consider to purchase other valuable datasets which would have more details. For example, different age group must have different interest in different type of game. If Jack could have datasets that cover every area, it would be more accurate to help Jack.

o Describe Data

🖃 Data Description Report

Data Quantity:

• The data has number and string.

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

• Rank, year, NA_sales, EU_sales, Jp_sales, Other_sales, Global_sales are numbers.

• Name, platform, genre, and publisher are strings.

• The dataset is downloaded from Kaggle website.

• The dataset has 16,599 rows and 11 columns (with labels).

Data Quality:

• The data has the most important thing we want, which is the number of sales.

• Two data types: numeric and categorical (string).

• The global_sales for each game could tell whether the game is popular or not.

• The name and sales would be the highest priority attribute.

• The platform, genre, publisher would be lower priority attribute

• The year would be the least priority attribute

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

440517742 Yingzhang Liu

A Case Study of Evaluating sales of video game with Data Mining Tools and CRISP-DM Methodology Stage2

Data Understanding

o Explore Data

From the data set, we hope to determine the poplar game, in other words, find those poplar games, and it would be better to predict which game would be better for sale, and finally, it would bring more profit to the company. From the data set, every attribute seems promising for further analysis, especially “platform”, “year”, “Genre”, “publisher”, “global sale”. Form theses attribute, we could figure out poplar platform, genre, publisher through different year, and from different year, we could see the trend in different attribute, therefore, we could get more accurate result in the prediction part. From the exploration we find following characteristic:

• Through the years, 2008 and 2009 has the biggest number of new release game.

• 2008 has the highest video game revenue. (Although the number of released game in 2008 is a little bit lower than that in 2009)

• For the recent 4 years, 2013, 2014, 2015, 2016, the top publisher by revenue for each year are Electronic Arts($211.68millions), Nintendo($194.6millions), Electronic Arts($181.68millions), Electronic Arts($49millions).

• Action game has been the most popular genre for 10 years(2007-2016)

• For the recent four years, 2013, 2014, 2015, 2016, the most popular games are Grand Theft Auto V($151.12million), Call of Duty: Advanced Warfare($87.6million), Call of Duty: Black Ops 3($113.24million), FIFA17($27.64million)

• For the recent four years, 2013, 2014, 2015, 2016, the most popular platforms are PS3($469.56million), PS4($395.04million), PS4($461.2million), PS4($157million)

• Through all the data in the dataset, Action game contribute almost 20% of the game released. The second is Sports game and the third is Misc game.

• Through all the data in the dataset, Action game contribute almost 20% percent of the revenue, Sports game is the second, and the third is the Shooter game.

We cannot just get the “popular” game by looking at the sales number, there are more attribute need to think about. For example, there are more games released in 2009 than the game released in 2008, but game released in 2008

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

made more revenue than the game in 2009. The top publisher, top platform, top game sales, top genre of game for the recent four years would be considered for later use. After this exploration, it is more clear to the business goal and the data mining goal doesn't change.

o Verify Data Quality

⏏ Data Quality Report

In our dataset, there are some data that has 0 values, and there are N/A values. The 0 values appear in the sales area. It happens due to the following reason: (1) Did not release in the current area. (2) The number of sales of that game is too low and there is no data for that game.

The N/A values appears in Year and Publisher, it happened due to the following reason: (1) the game is canceled. (2) the data for the game is not complete.

Some games are missing data in Year and Publisher, but they are actually have the data of sales. Which is quite confused that Year and Publisher are much more simple than Sales to collect. Otherwise, there is no spelling inconsistencies. And also, there is one game that come out in year 2020. That is incorrect.

In our dataset, there are some games that are released a long time ago. Actually, we need focus recent games only, and other data would be no valuable for our further study and has no impact on our hypotheses

The data is stored in the csv file and could be open in excel. Every attribute is clearly shown to the user.

Data Preparation

o Select Data

⏏ Selecting Items (rows): There are 16,599 data in our dataset, we cannot use all of them. We have to delete most of them, since there are some games that are too old and they will not be useful to help us achieve our business goal. So we need choose the data in recent years, that would be 2014, 2015, and 2016. And also there are some games with N/A value, we would not choose them as well.

⏏ Selecting attributes: Since we want to purchase popular games, from those attributes, we need Name for sure, and Platform, this will help us to see the popular console in the current year. And Year, this is an very important attribute. And Genre, it would help us to decide which type of game should be

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

purchased more. And so does Publisher, we could see which publisher is the most popular one. And the last, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, these attributes are necessary.

o Clean Data

Missing data: For some games, their data about number of sale is 0. That usually happened on those old games which are released a long time ago. Or it means the game is not released in that area. Because Jack's store is located in Australia, So we would consider just delete the rows with 0 value in Other_Sales attribute. Some of games have N/A value in Year or Publisher and we will delete those rows with N/A values.

Data errors: The data in the row 5959, has the publish year of 2020, that is in the future, but it also has sales data, which has contradiction since it is 2017 now. It is easy to find its correct publish year in Google and fix it. There might be some games with wrong publish year, unless it is the same type of error as the one we mentioned, otherwise we could not check every games' information among 16,599 rows of data.

o Construct Data

In my dataset, we have attribute of name, year, publisher, genre, platform. That include all the necessary attributes for a game and we don't need add any more attribute. Also, there are data about global sales for each game, which is the most important thing that we focus on. All of them could not be separate to derive new attribute.

For the modeling algorithm, there is particular type of data, it would be better to keep the data as original type

o Integrate Data

🚕 **Adding current dataset with customer attribute**

It is a good option to merge the dataset with the Jack's store data which is recorded by the accountant. For each customer ID, game purchased, purchased time are correctly merged with the related game data. So an event would correctly associated with right customer ID and the right game details.

o Format Data

We are planning to use the predictive analytics. We could use Naïve Bayes to do the predictive modeling and use data mining at the same time to analyze current and historical data to make prediction about future or otherwise unknown event. There is no particular format for this.

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

A Case Study of Evaluating sales of video game with Data Mining Tools and CRISP-DM Methodology Stage3

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

Modelling

o Select Modeling Techniques

♣ Modeling Technique

The cleaned dataset Jack and his team have right now is based on 6 most significant attributes which are name, platform, year of release, genre and sales(in millions). In sales, this attribute is divided into 5 different branches which are NA_sales, EU_sales, JP_sales, OTHER_sales and GLOBAL_sales. Mainly targeting those continents and regions with the greatest number of players or producers. Different gamers from different regions and areas certainly have a different taste on games, like Japanese gamers may always be huge fans of Nintendo and they prefer to play Japanese homemade games than English games. In NA and EU, gamers may use X-box more often rather than using PS4, however, Asian gamers will choose the opposite gaming platform. Jack need to classify the dataset into different small train sets to do further analysis.

The next step for Jack to do is to select and build a model which is based on the cleaned dataset, and do some deeper data mining and design some tests to test his data. After doing so, he supposes that he will find a rightful way of selling his games more successfully. What exactly is a rightful way of selling games in Jack's opinion? After building the model, Jack will be able to classify what is the most significant factor that affects players decisions of buying games. Is it the genre, publisher or platform?

After the implementation of the model, test and training dataset are needed to test how good the model is and how accurate and precise this model can be. For this point, the data analyst will randomly choose 150 samples from the cleaned

version of our dataset and input them to our model and run the test. Currently, there are over 800 observations in our dataset. Even though the dataset is relatively small, it will be definitely enough for our models.

And for clustering, there is no particular type of data, and certain level of data quality, since we have cleaned our dataset already, those missing data and error are already removed and our dataset will be ready to implement the model.

♣ **Modeling Assumptions**

Jack is facing a lot of choices of models and he and his team get to narrow down the modelling tools of choice and choose the one which is the most appropriate model that help them reach their goal. The only numerical attribute Jack has is the sales while the others are all categorical, and sometimes the value of sales can be arbitrary. The popularity in different regions is also different, therefore, this not only affects the value of sales but also affects game selections. For example, if a game sales well in JP area does not mean it can also earn a lot of money in other regions such as NA or EU. Therefore, Jack makes some following modelling assumptions that can help him analyze.

Decision Tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. Most importantly, it works with both numerical and categorical data.

Naive Bayes is a classifier that applies the Bayes theorem with independence assumptions between the features. In this case, Jack has to assume that these classes or attributes are independent from each other.

K-means clustering is one method of building the clustering model. Jack and his analysis group are going to try this method on the current dataset. In K-means clustering, it does not require any data type before execution. So they would just put cleaned version of dataset into the model.

o Generate Test Design

♣ **Test Design**

Jack and his team will use a new feature which is called ***Sell*** to help them measure the goodness of a model. By using an IF-statement formula on excel

$$Sell = IF(Li > mean, 1, 0)$$

Jack has calculated the mean global_sales in his dataset and by comparing it with the actual global_sales of each game in the dataset, he defined that if the Sell = 1, this game is worthy and good to sell. Jack's goal is to split up all the

possible Sell = 1 samples in a new training set, and do further analysis with this training sets.

After building a model, the analysis team would randomly pick 150 data from the dataset and put it into the model to test how good the model is. The analysis team would put test data into different model. With different model, will generate different type of value, we could measure the success with these type of value. And also, some models, like decision tree, we could use the test data to run on decision tree and see if it fit the one that is generated by the training data. For models like clustering, the team might run three times to find the most appropriate one since every time when rerun the clustering, the result might change.

The criteria of goodness of the model will be considered as something that is related to data mining goal, the good model should help Jack to analyse the current video games market. It could help Jack identify which game is popular so that it could bring more profit to the company.

o Build the Model

♣ Model Description

Decision Tree:

This model failed since there are too many classes in each class. But we cannot modified it. For example, in class Platform, there are PS4, XBOX, PC, Nintendo and so on. For Genre class, there are action, sports, shooting, role playing and so on. We considered that to decrease the number of types for each class, but this is something that we cannot modify.

K-means Cluster:

This is a popular method for cluster analysis in data mining. K-means clustering aims to partition n items into K clusters. We try to run the code in R to create K-mean Cluster

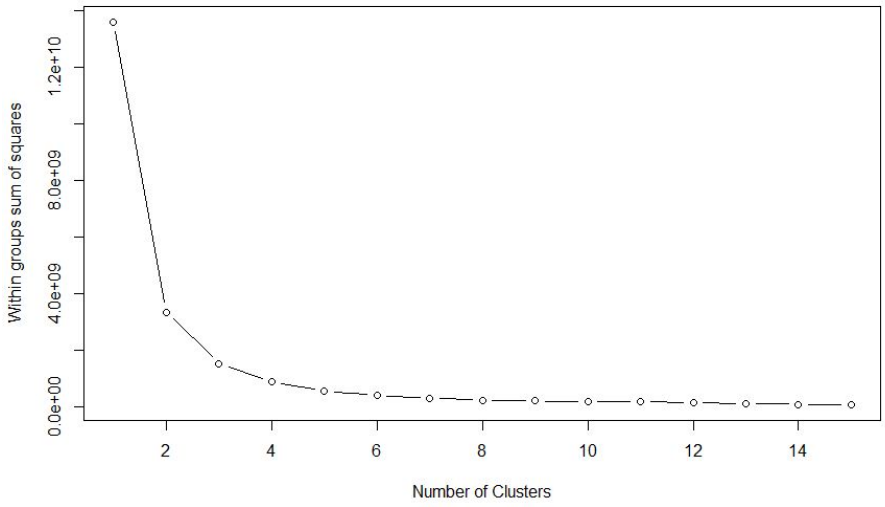
First of all, Jack needs to select an appropriate value of k in terms of the following diagram. In this case, we get a diagram below and using elbow method, which

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

point out that K = 4 would be good options.

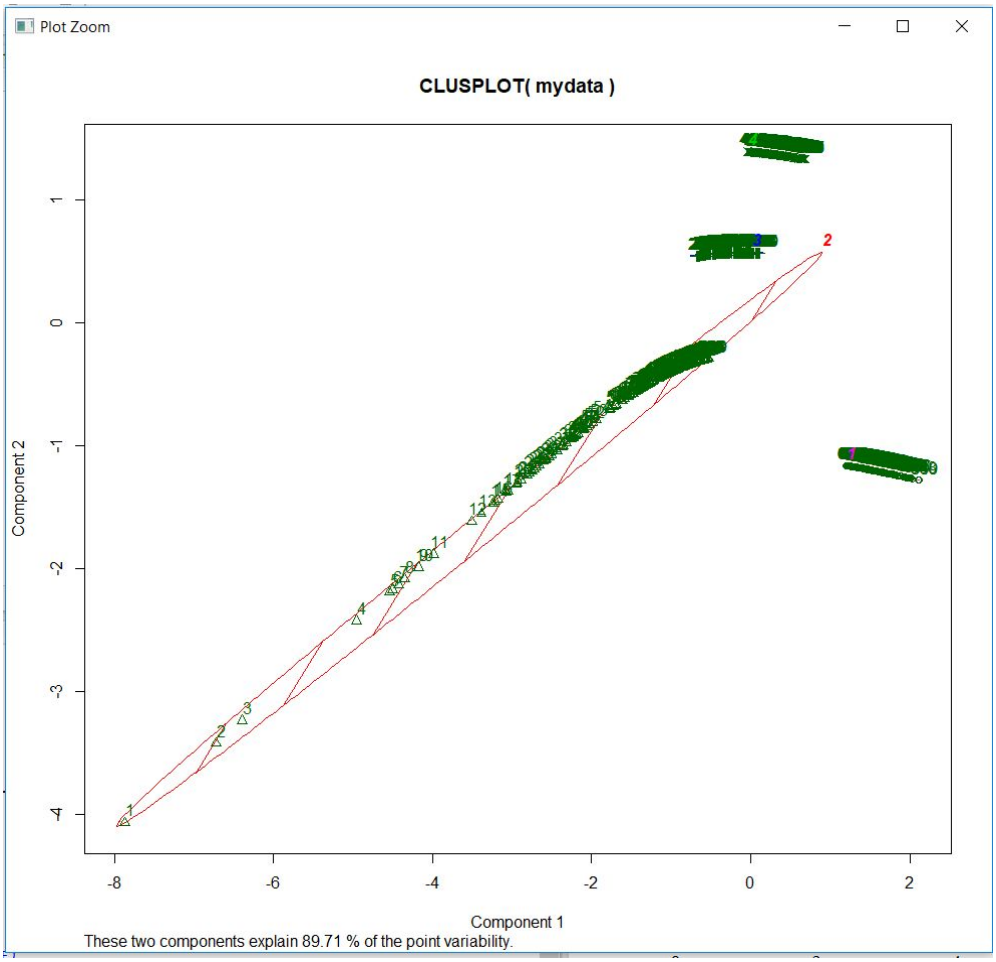


Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

After Jack find the appropriate value K, we get the corresponding K-means model:



In this model, we could see the more similar objects that are grouped together inside the same cluster. We can only find out the objects that are grouped together, besides that, we could not find any meaningful information from this model. Also, the problem for this model is, because there are too many objects, we can not really see clearly in our model that which objects are grouped together.

Naive Bayes:

Jack considers Genre of games will be the main factor that determines gamers choice of game, so he decides to use Naive Bayes model to verify his assumption. Naive Bayes is a very simple technique to classify data. Since there are around 800 samples left in the current data set, Jack reckons it would be more convenient and accurate to use 67% of the data samples as training data set, while the other 33% will be considered as testing set. By applying the Bayes theorem, the first model is built successfully. Jack can apply this theorem to different attributes to gain more results and conclude a final result from them.

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

With the help of the prediction result, Jack can focus on the specific platform and genre of games they conclude from the result in the future.

o Assess the Model

♣ Model Assessment

```
Conditional probabilities:
             Sell
Y                 [,1]      [,2]
  Action       0.2020725 0.4025904
  Adventure    0.0750000 0.2667468
  Fighting     0.2258065 0.4250237
  Misc         0.2727273 0.4558423
  Platform     0.3684211 0.4955946
  Puzzle       1.0000000        NA
  Racing       0.2068966 0.4122508
  Role-Playing 0.1857143 0.3916837
  Shooter      0.5147059 0.5034996
  Simulation   0.1111111 0.3333333
  Sports       0.2769231 0.4509605
  Strategy     0.0000000 0.0000000
```

This is the conditional probability we got from the training set and testing set by using the attribute Genre.

```
Conditional probabilities:
      Sell
Y            [,1]      [,2]
  3DS  0.17948718 0.3887764
  PC   0.07894737 0.2732763
  PS3  0.19480519 0.3986477
  PS4  0.34193548 0.4758957
  PSV  0.04347826 0.2061846
  Wii  0.60000000 0.5477226
  WiiU 0.31428571 0.4710082
  X360 0.25000000 0.4364358
  XOne 0.27083333 0.4467230
```

This is the conditional probability we got from the training set and testing set by using the attribute Platform.

According to the result Jack got so far, in overall, PS4,Wii and XOne are the top 3 most selling gaming platforms, as a very high proportion of games on these platforms can make their sales values more than the average, which is the 0.74 millions approximately.

On the other hand, Action, Shooter and Sports are the 3 most popular genres of all the samples we got. Considering there is no record shown in the testing set for Puzzle game, Jack does not want to take it into account for his future plan.

Group: 42

Name: *Yuanxi ZENG* & *Yingzhang Liu*

SID: 450105465 & 440517742

```
> table(training$Genre)

      Action    Adventure     Fighting         Misc     Platform       Puzzle       Racing Role-Playing
         287           51           39           44           25            3           45          103
     Shooter   Simulation       Sports     Strategy
          99           17          116           11
> table(training$Platform)

 3DS   PC  PS3  PS4  PSV  Wii WiiU X360 XOne
  60   65  117  241   67    6   46   91  147


> pred
       Action  Adventure   Fighting       Misc   Platform      Puzzle     Racing Role-Playing   Shooter
[1,] 0.2727273 0.01069519 0.04278075 0.05347594 0.04812834 0.005347594 0.04278075    0.1176471 0.2139037
     Simulation    Sports
[1,] 0.01604278 0.1764706
> pred <- predict(e1071model2,newdata = "Sell",type = "raw")
> pred
            3DS          PC         PS3         PS4         PSV         Wii        WiiU        X360        XOne
[1,] 0.05347594 0.02673797 0.1229947 0.3850267 0.02139037 0.02139037 0.07486631 0.0855615 0.2085561
```

It seems like there is not any huge difference in gaming market, apart from Role-playing will be replaced by the Shooter games while Action and Sports are still the leader of the most popular genres. What catches Jack's attention is that Wii and Wiiu have incredibly low distribution but they do have some best selling games than the other platforms. Maybe this is caused by players from different regions. Japanese players seem like to be very keen on their national made games.

In next stage, Jack will purchase and stock more PS4,PS3 and XOne games which mostly are sports, shooter and action games in these platforms.

Evaluation

o Evaluation Results

♣ Assessment of Data Mining Results:

Multiple model diagrams and numerical data results were generated by using different models and help analyze for Jack's future business goal. Overall, the whole analyzing process produced useful and constructive results. Besides, the dataset itself has already very straightforward data for Jack to take into consideration. Choosing the best selling genres and combine them with the best selling platforms, Jack will guarantee himself a bright future and great success. One problem needs to be taken care of is Jack has to figure out what market is he targeting. The global sales could be a very promising standard, however, people different culture and different background of certain areas can sometimes be very stubborn in game choosing. As we all know that USA is a country that allows its citizens to own guns, this explains why the best selling games in US is Shooter type games. On the contrary, the highest sales in EU region are Sports games with no doubt. The most popular sports in the world, not only the world, is the football, and EU people seem like they have the craziest fanaticism which beyonds your imagination. No wonder the FIFA games can be so hot during the

past 4 years. So this is a problem Jack needs to figure out in the future. But there is nothing wrong about selling some games which are globally popular.

New Questions: Since the model is constructed based on current data from 2014 to 2016. Jack would not know if the current model could accurately predict the situation about video games in the near future.

o Review Process

♣ Review of process

The process of review will lead Jack to understand:

- A return to Business Understanding phase is necessary if there is any error happened.
- In Model and Evaluation phase, it requires extra patience and attention. Since it takes time and a small error could lead the project to a wrong way.
- Our dataset does not suit Decision Tree at all since there are too many classes under each data attribute. It requires understanding to every model's requirement of variables before executing the model.
- Clustering is not better than Naive Bayes for the dataset, which we put a lot of hope on it during the modeling phase
- The whole process requires a lot of data mining skills, we have to learn new knowledge while we are analyzing and modeling at all time
- Gaming industry is developing in an incredible speed, the analyze must keep up with the trend

o Determine Next Steps

♣ List of Possible Actions Decision

The next stage is, of course, the deployment stage which is also the last stage of Jack's project. Jack is pretty satisfied with current process and result. And at the same time, the project team is ready to go back and hope they could find some model that is possible to predict the future video games market.