Game Data Analytics using Descriptive and Predictive Mining

SEMINAR REPORT

ON

GAME DATA ANALYSTIC USING DESCRIPTIVE AND PREDICTIVE MINING

BY

MISS PRITI SANTOSH MAURYA

MC21954

GUIDE - PROF.SHAESTA KHAN

POST GRADUATE DEPARTMENT OF COMPUTER SCIENCE,

S.N.D.T. WOMEN’S UNIVERSITY

MASTER OF COMPUTER APPLICATION

SEM VI - (2020 – 2021)

CERTIFICATE

This is to certify that Miss. Priti Santosh Maurya has completed the project on Game Data Analytic Using Descriptive and Predictive Mining satisfactorily as a partial fulfilment of the Post Graduate Degree of Master Of Computer Application (MCA)

Internal Examiner :

Signature :

Name : PROF.SHAESTA KHAN

Date :24/08/21

External Examiner :

Signature :

Name :

Date :

Head of Department :

Signature :

Name :DR.GANESH MAGAR

Date :24/08/21

Acknowledgement

I would like to express my sincere gratitude to my mentor Prof.Shaesta Khan for her valuable guidance and support in completing my research paper

I would also like to express my gratitude towards her for allowing me to do a study on research paper. This study helped me learn many new things. Without the support and suggestions from my mentor, this study would not have been completed.

I take this opportunity to thank SNDT Woman's University for giving me chance to do this project.

Name and signature:

Priti Santosh Maurya

Abstract

This paper aims to predict the top-selling video game sales in North America between 1983 and 2016. The dataset is collected from an internet platform known as Kaggle.com. The dataset was generated by vgchartz.com.

Exploitation the dataset, the RStudio IDE tool and R-programming language are used for data cleaning, analysis, and representation. The machine learning algorithm used in this project is linear regression. Based on the Video Games Sales knowledge, it would be fascinating to know what area unit the required factors that make a game further successfully sold-out than others in North America. So, we'd would like to research what quite video games that area unit further successfully sold-out in North America.

We have a tendency to tend to jointly would like to point the results of this Analysis in Associate in nursing intuitive methodology by visualizing outcome victimization ggplot2 in R.

In this project, we have a tendency to tend to require NA_Sales (North America sales) as response variable and specialize in operative predictions by analysing the rest of variables inside the k video games sales data. The results can facilitate film companies to know the key of generating an advertisement success game.

List of Figures and Tables

Index

CONTENTS **PAGE NO.**

Content

Chapter 1: Introduction

Video game industry needs accurate sales in an exponential market growth. In the last 10 years in the United States the revenue coming from computer and video games increased imposingly. So we have to predict the buying nature of several video game followers by using historical sales data. This study involves extracting the video game sales data and analysing which game has more sales globally when compared to other countries.

With this we used machine learning techniques which predict the sales of video game in the market. This approach is useful to several industries which are interested in predicting the sales data. In this paper, we are concerned with predicting the sales of a video game. For this we have used historical time series sales data. Our dataset consists of 11 variables and 500 samples with a combination of categorical and numeric variables.

We need to perform data pre-processing on dataset to check whether the data is properly loaded or not, is there any missing values or NA values etc.

Out of all these variables few variables are unused so drop those variables. Now find correlation between variables to know the input variable and target variable for applying machine learning algorithms. After applying correlation matrix, we came to know the target variable and input variables. Before applying machine learning algorithms we have to split the dataset into training and testing sets. Finally to obtain better performance, we have to apply possible machine learning algorithms which give us best result.

Machine learning algorithms are classified into three categories: supervised learning, unsupervised learning, reinforcement learning . In supervised learning we have input variables and output variables and we apply machine learning technique to learn mapping function from input to target variable. Supervised learning has two categories: classification and regression. In un-supervised learning we have only input variables and no target variables. It has its own way to discover the structure in the data.

In this project we have used supervised learning algorithms they are linear regression, support vector regression, random forest, and decision tree. We also use performance measures such as root mean square error, r-square, mean absolute error. One of the major objective of this research work is to find the trending sales by using machine learning algorithms. Sales prediction is an essential part of business organizations.

It provides relevant information that can be used to make strategic business decisions . Sales prediction is very important tool for upcoming business ventures etc. Sales and market predictions are two different aspects which determine the client and

market demand respectively. Sales prediction provides relevant information that can be used to make strategic business decisions. In the next sections we formulate the review of the related work, methodologies with detailed descriptions, comparison work, results and discussions. The paper ends with conclusion and future enhancement.

Chapter 2: METHODOLOGY

2.1 System Architecture

There are few steps can be performed for gathering data, analysis and modelling to get best predictions.

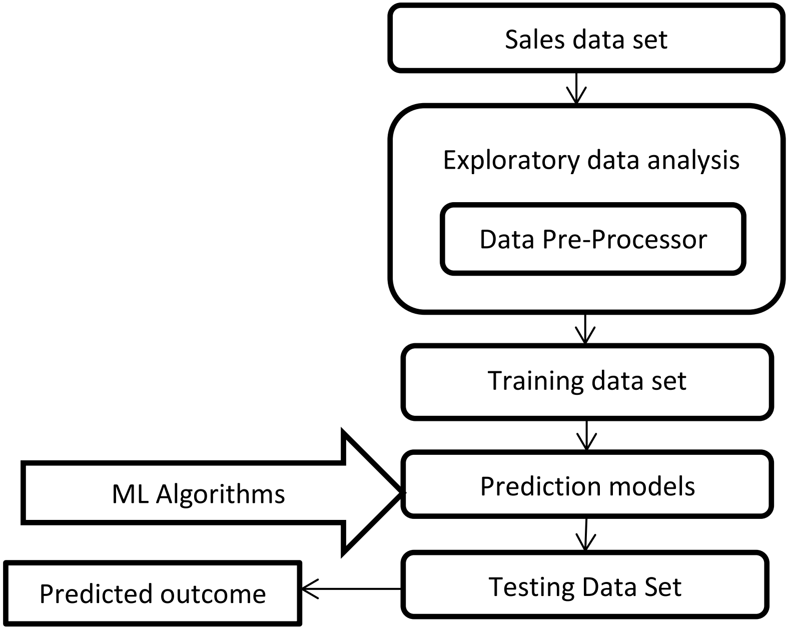


Fig.2.1. System architecture

Descriptive Mining

Descriptive Analytics will help an organization to know what has happened in the past, it would give you the past analytics using the data that are stored.

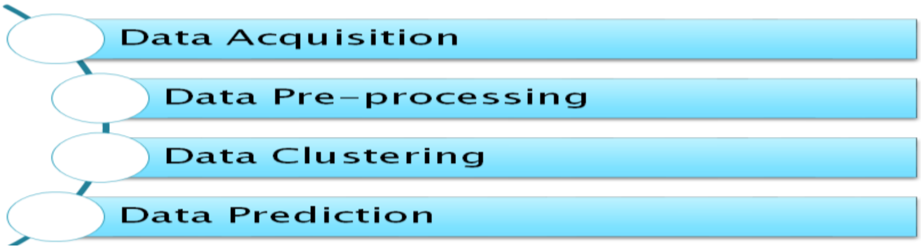
For a company, it is necessary to know the past events that help them to make decisions based on the statistics using historical data.

For example, you might want to know how much money you lost due to fraud and many more.

Experiment and Analysis

The experiment consisted of several stages. The first step is data acquisition to test how data is retrieved from the Steam API.

Then the next step is to preprocess the data that has been obtained. The third step is the phase of testing the clustering of data that has been processed before.



2.2 Data analysis

2.2 . EXPERIMENT

This project/study uses a special video game sale dataset sold in different countries. This dataset is created by VGChartz.com. The RStudio is used to run an experiment. The RStudio is free, open-source and an Integrated Development Environment (IDE) for R-language, used for statistical computing, programming, GUI, and Graphics. RStudio is most famous for Graphical capabilities, but in recent times it gained importance for analysing data.

DATA PREPARATION AND ANALYSIS PROCEDURE:

- **Statistically exploring the data- a check list:**

Check data dimensions

Rows, columns and column names

Data types and unique values per column.

- **Cleaning the data: -**

Look for any missing data.

Identify and convert categorical values to numerical representation or convert numerical to categorical representation using dummy variables if suitable for modelling and check for distinct values in categorical columns.

- **Statistically overview of data: -**

Check head, tails of data to see complete required data loaded. Identify numerical columns and look for insights like median, mean, mode etc. Understand the relationship of columns and how they are effecting each other.

Check correlation and chi-square. Correlation - shows relation of numerical columns

Chi-square - shows relation of categorical columns

- **Graphical representation of data: -**

Perform visualization on dataset attributes.

2.3. METHODOLOGY

A .Machine Learning Algorithm The ML algorithm is a logic that grasp one step ahead when exposed to more information/data. When ML is exposed to training data it produces model. To build a Model, the Machine Learning Algorithm used here Linear Regression (Supervised Learning). It predicts the output values based on the input data fed. This algorithm builds a model based on the training data produced and predicts the new data.

B .Dataset The RStudio is used to import the dataset. Dataset can be in excel or in CSV format. The dataset is reviewed and normalized. Normalization is changing the value of numeric columns of the dataset to common values and fit into a specific range.

This is a video game sales data including game sales of North America, European, Japan and other area, together they make the global sale. The data also give the information about the critic score, user score and the counts of critics or users who gave these two scores. This data was downloaded from https://www.kaggle.com/rush4ratio/video-game-sales-with ratings#Video_Games_Sales_as_at_22_Dec_2016.csv.

This dataset contains a list of video games with sales greater than 100,000 copies.

- Fields include
- Rank - Ranking of overall sales
- Name - The games name
- Platform - Platform of the games release (i.e. PC,PS4, etc.)
- Year - Year of the game's release
- Genre(Category) - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)
- Global_Sales - Total worldwide sales.

The video game industry is the economic sector involved in the development, marketing, and monetization of video games.

It encompasses dozens of job disciplines and its component parts employ thousands of people worldwide. The computer and video game industry has grown from focused markets to mainstream.

They took in about US$9.5 billion in the US in 2007, 11.7 billion in 2008, and 25.1 billion in 2010 (ESA annual report) and 159.3 billon in 2020 or 175.8 billion in 2021(currently counting) in pc market 37 billion and mobile marketing 77 billion in 2021 revenue report up april.

Modern personal computers owe many advancements and innovations to the game industry:

sound cards, graphics cards and 3D graphic accelerators, faster CPUs, and dedicated co-processors like PhysX (Nvidiabased api design) are a few of the more notable improvements.

Analytical study of Video game industry and sales across the world

We study the given data set of sales of video games to find relations between different factors that affect video game sales. Our main objectives are :

- Finding which platform is more popular in which region and affects sale by how much
- Finding which year and period was more popular and how has it affected sales
- Finding which publisher is more popular in which region and affects sale by how much
- Finding which genre is more popular in which region and affects sale by how much
- How are the sales of America, Europe, Japan and other regions of the world correlated and is there a common pattern amongst all, or do they follow different trends
- Realizing and understanding the needs and trends in the video game industry to target the maximum customers in a new release of a game etc.

Chapter 3: Data cleaning

At this stage, by using RStudio we import dataset and remove redundant, missing, duplicate, and unnecessary data for further processing. This stage is the most time-consuming stage in Data Science because to prevent wrongful prediction and get rid of the inconsistencies of data.

Data loading processing

# Loading the database

```
data<-read.csv("C:/Users/priya/Desktop/GameResearch/gamedata/vgsales.csv",
stringsAsFactors =    FALSE)
```

# Removing the Rank column

```
    data$Rank <- NULL
```

# Filtering only the records of interest for this study, removing the records with Year = NaN and records with the year above 2016

```
   data <- data[data$Year != "N/A" & data$Year != "2017" & data$Year != "2020", ]
```

```
  data$Year <- factor(data$Year)(factor = object data wich used as categorize the data and
store is as          level )
```

# Viewing the first 6 DataFrame records

```
    head(data, 6) (head= display the 1st n of row present in data)
```

```
   summary(data) (summary= is generic function producing a summary of the various data )
```

Chapter 4: Data Exploration and Analysis

4.1 Histogram:

A histogram contains rectangular area to display the statistical information which is proportional to the frequency of a variable and its width in successive numerical intervals. A graphical representation that manages a group of data points into different specified ranges. It has a special feature which shows no gaps between the bars and similar to a vertical bar graph. R creates histogram using **hist()** function.

Syntax*:*

Hist(v, main, Xlab, Xlim, Ylim, breaks, col, border)

Parameters:

- **v:** This parameter contains numerical values used in histogram.
- **main:** This parameter main is the title of the chart.
- **col:** This parameter is used to set Color of the bars.
- **xlab:** This parameter is the label for horizontal axis.
- **border:** This parameter is used to set border Color of each bar.
- **xlim**: This parameter is used for plotting values of x-axis.
- **ylim:** This parameter is used for plotting values of y-axis.

Normality Q-Q Plot

As the name suggests, this plot is used to determine the normal distribution of errors. It uses standardized values of residuals. Ideally, this plot should show a straight line. If you find a curved, distorted line, then your residuals have a non-normal distribution (problematic situation).

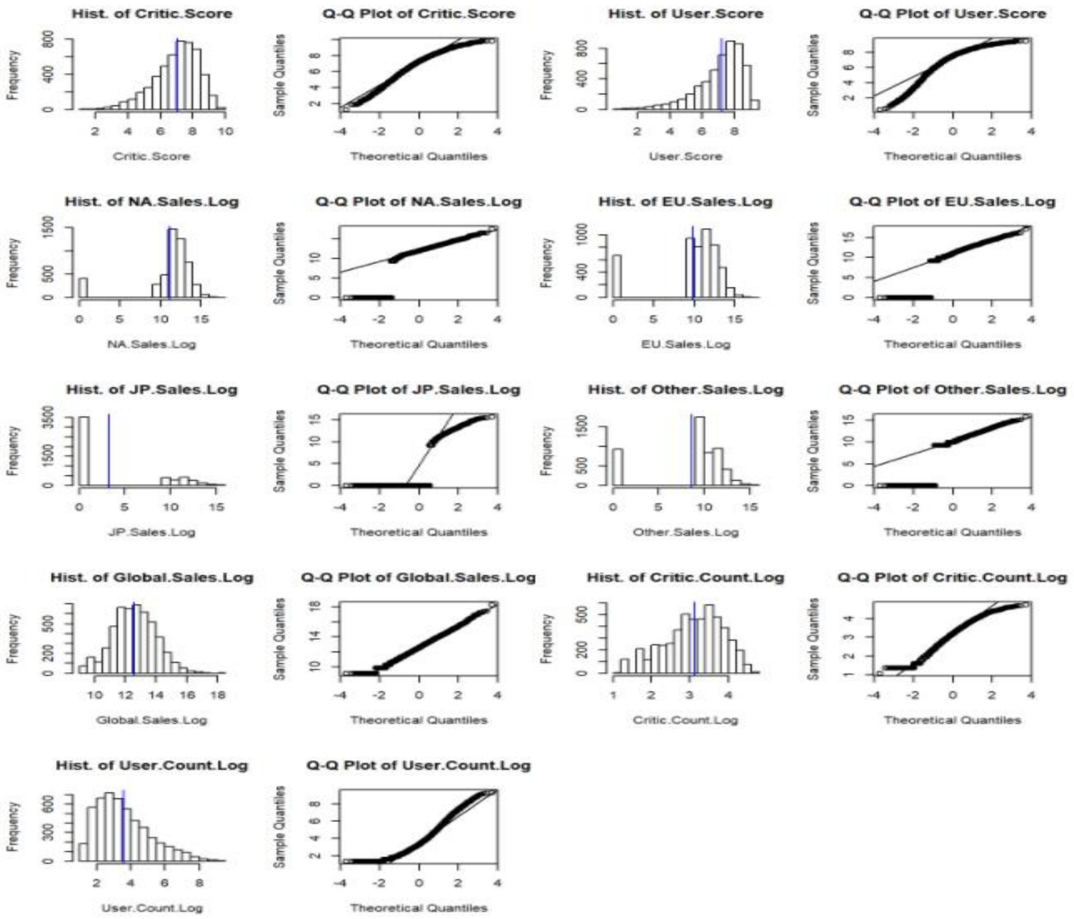
Histogram (QQ-Plot)   Steps:-

- by multiplying 1000000 we get the actual sale,
- adding 1 makes all sales positive which make log possible for all sales later
- Log value's for normal distribution calculation.
- By divide by 10 to make Critic Score the same decimal as User Score
- format column names
- Our pre-analysis shows that these variables are not normally distributed, especially those sales and score counts variables. We take logs to transform these variables.
- we combine the log variables with the original variables.
- the data we use for analysis

Algorithms:

- Normal Distribution and histogram and qq plot :
- Step 1: start add data
- Step 2: Clean data
- Step 3: actual data multiply by 1000000.
- Step 4: Adding +1 to make sales data positive.

  (This is for Log value)
- Step 5: Critical Score /10(Making data vale in decimal)
- Step 6: Log value transformation in variable
- Step 7: Combine Log variable with Original Variable
- Step 8: plot histogram and QQ plot:
- Step 9: Take two columns of Shapiro.test
- ▶ Step 10 : add variable mean



4.1 Histogram & QQ plot

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

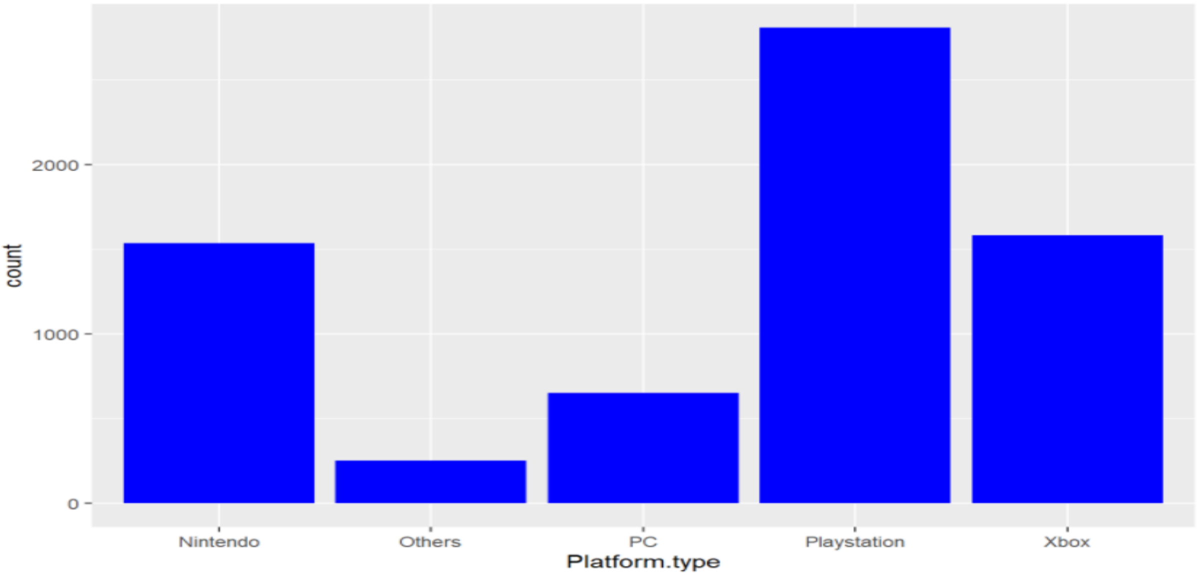
From the histograms and QQ plots we also see that two scores and log values of their counts are close to normal distribution .Though the Shapiro test still deny the normality of these log values. We assume they are normally distributed in our analysis.

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

4.2 Visualization of categorical variables

1. Bar Graph

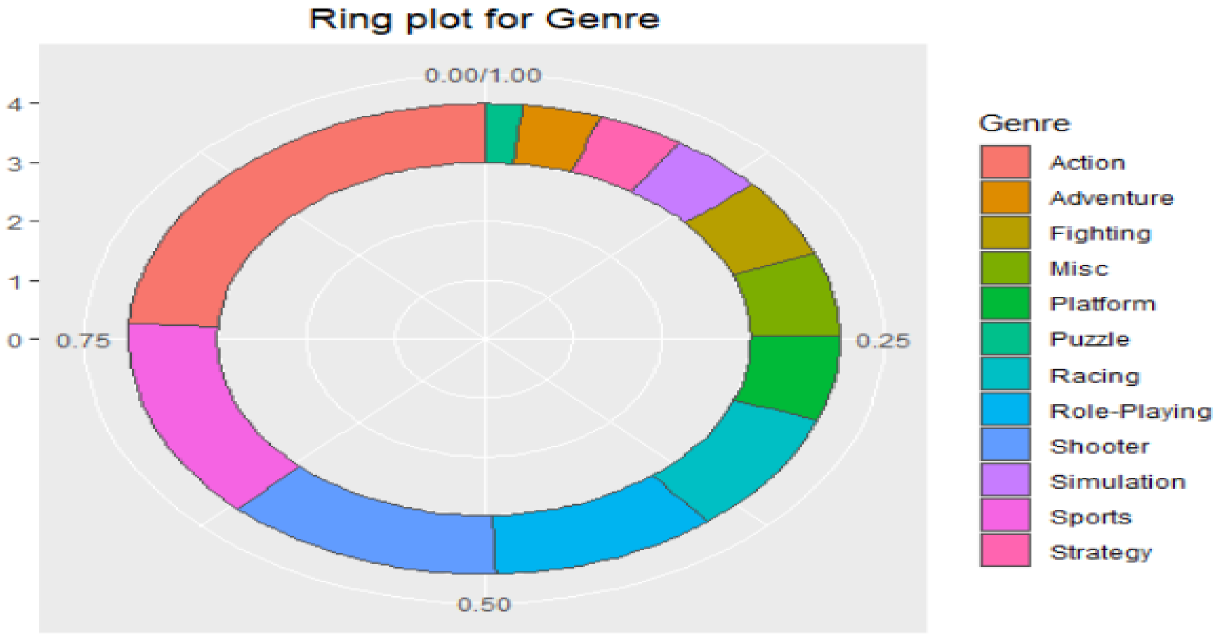
- To simplify platform analysis,We regroup platform as Platform.type.
- Regroup platform as Platform.type
- Bar graph
- regroup Rating as Rating.type
- rename the names of counts for detail information



4.2 Bar Graph

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

2. Ring plot

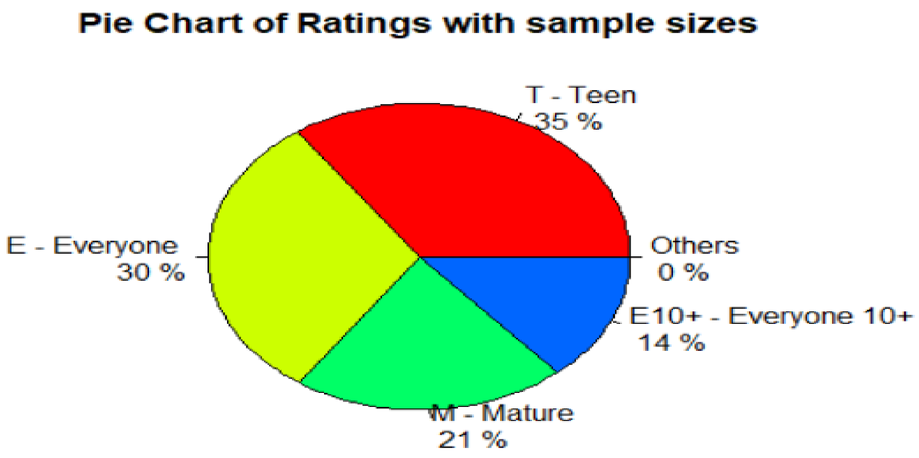


4.3 Ring Plot

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

3. Pie chart

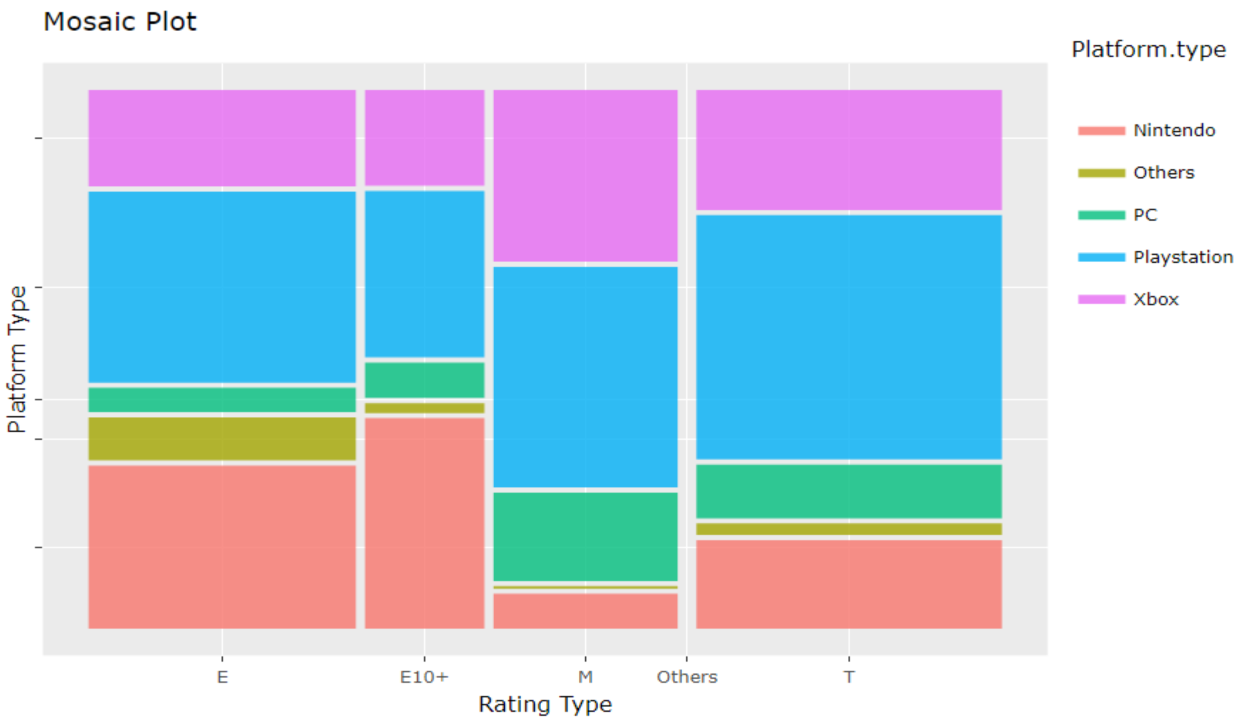
We regroup rating AO, RP and K-A as “Others” because there are only few observations of these ratings.



4.4 Pie Chart

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

4 . Mosaic Plot



4.5 Mosaic Plot

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

4.3. Correlation among numeric variables

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

Key decisions to be made when creating a correlation matrix include: choice of correlation statistic, coding of the variables, treatment of missing data, and presentation.

Applications of a correlation matrix

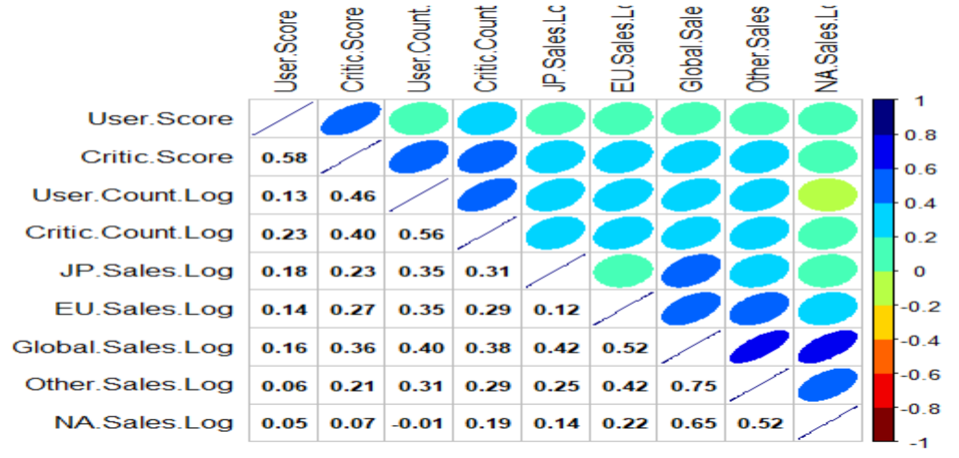
There are three broad reasons for computing a correlation matrix:

To summarize a large amount of data where the goal is to see patterns . In our example above, the observable pattern is that all the variables highly correlate with each other.

To input into other analyses. For example, people commonly use correlation matrixes as inputs for exploratory factor analysis, confirmatory factor analysis, structural equation models, and linear regression when excluding missing values pairwise.

As a diagnostic when checking other analyses . For example, with linear regression, a high amount of correlations suggests that the linear regression estimates will be unreliable.

- Take numeric variables as goal matrix
- Correlation matrix



	User.Score	Critic.Score	User.Count.	Critic.Count	JP.Sales.Lc	EU.Sales.Lc	Global.Sale	Other.Sales	NA.Sales.Lc
User.Score									
Critic.Score	0.58								
User.Count.Log	0.13	0.46							
Critic.Count.Log	0.23	0.40	0.56						
JP.Sales.Log	0.18	0.23	0.35	0.31					
EU.Sales.Log	0.14	0.27	0.35	0.29	0.12				
Global.Sales.Log	0.16	0.36	0.40	0.38	0.42	0.52			
Other.Sales.Log	0.06	0.21	0.31	0.29	0.25	0.42	0.75		
NA.Sales.Log	0.05	0.07	-0.01	0.19	0.14	0.22	0.65	0.52	

4.2.1 Correlation Matrix

There are high r values of 0.75, 0.65, 0.52 and 0.42 between the log value of Global.Sales and regional sales. On the other hand, there are good positive correlation between regional sales too. User Score is positive correlated to Critic Score with r of 0.58. There is little correlation between User Count log value and User Score.

Cluster dendrogram for numeric variables.

As the name itself suggests, Clustering algorithms group a set of data points into subsets or clusters. The algorithms' goal is to create clusters that are coherent internally, but clearly different from each other externally. In other words, entities within a cluster should be as similar as possible and entities in one cluster should be as dissimilar as possible from entities in another.

Broadly speaking there are two ways of clustering data points based on the algorithmic structure and operation, namely agglomerative and divisive.

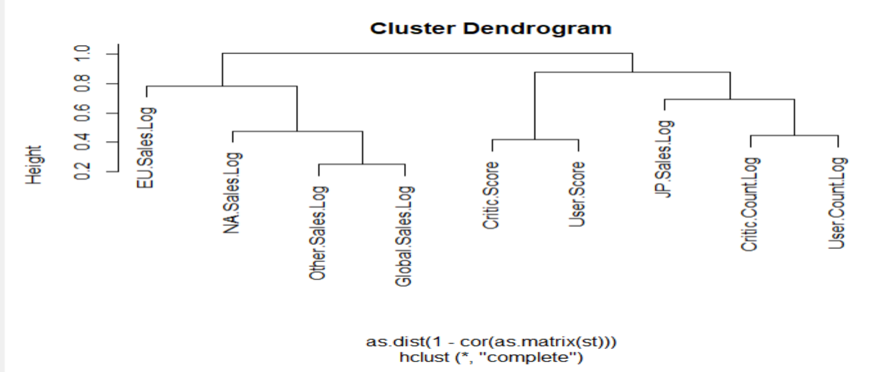
Agglomerative: An agglomerative approach begins with each observation in a distinct (singleton) cluster, and successively merges clusters together until a stopping criterion is satisfied.

Divisive: A divisive method begins with all patterns in a single cluster and performs splitting until a stopping criterion is met.

In this tutorial you are going to focus on the agglomerative or bottom-up approach, where you start with each data point as its own cluster and then combine clusters based on some similarity measure. The idea can be easily adapted for divisive methods as well.

The similarity between the clusters is often calculated from the dissimilarity measures like the euclidean distance between two clusters. So the larger the distance between two clusters, the better it is.

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



4.2.2 Cluster Dedrogram- Hierachical clustring

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

4.4 Analysis of score and count

Pearson Correlation and Linear Regression

A correlation or simple linear regression analysis can determine if two numeric variables are significantly linearly related.

A correlation analysis provides information on the strength and direction of the linear relationship between two variables, while a simple linear regression analysis estimates parameters in a linear equation that can be used to predict values of one variable based on the other.

Correlation

The Pearson correlation coefficient, r, can take on values between -1 and 1. The further away r is from zero, the stronger the linear relationship between the two variables. The sign of r corresponds to the direction of the relationship.

If r is positive, then as one variable increases, the other tends to increase. If r is negative, then as one variable increases, the other tends to decrease.

A perfect linear relationship (r=-1 or r=1) means that one of the variables can be perfectly explained by a linear function of the other.

Linear Regression

A linear regression analysis produces estimates for the slope and intercept of the linear equation predicting an outcome variable, Y, based on values of a predictor variable, X. A general form of this equation is shown below:

$$Y = b_0 + b_1 \cdot X$$

The intercept, b0, is the predicted value of Y when X=0. The slope, b1, is the average change in Y for every one unit increase in X.

Beyond giving you the strength and direction of the linear relationship between X and Y, the slope estimate allows an interpretation for how Y changes when X increases.

This equation can also be used to predict values of Y for a value of X.

Inference

Inferential tests can be run on both the correlation and slope estimates calculated from a random sample from a population. Both analyses are t-tests run on the null hypothesis that the two variables are not linearly related.

If run on the same data, a correlation test and slope test provide the same test statistic and p-value.

Assumptions:

Random samples

Independent observations

The predictor variable and outcome variable are linearly related (assessed by visually checking a scatterplot).

The population of values for the outcome are normally distributed for each value of the predictor (assessed by confirming the normality of the residuals).

The variance of the distribution of the outcome is the same for all values of the predictor (assessed by visually checking a residual plot for a funneling pattern).

Hypotheses:

Ho: The two variables are not linearly related.
Ha: The two variables are linearly related.

Relevant Equations:

Degrees of freedom: df = n-2

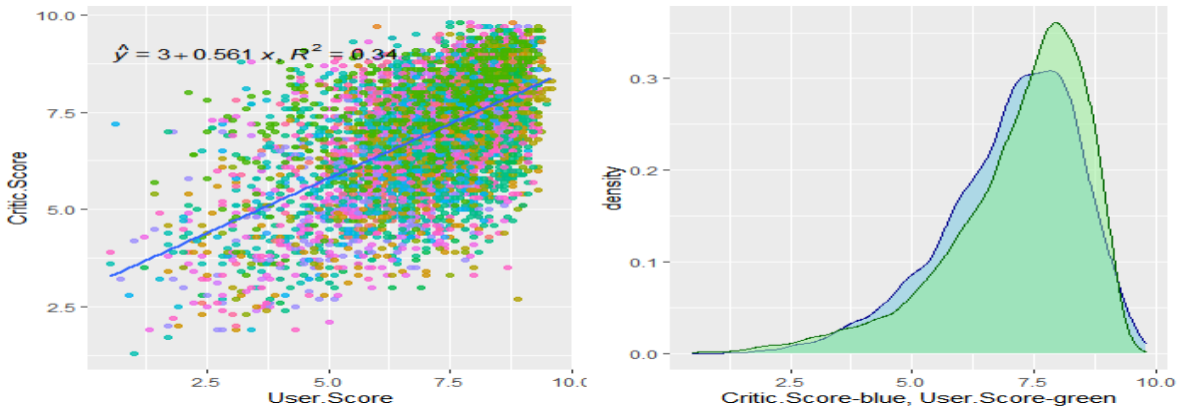
$$r = \frac{\sum z_x z_y}{n-1}$$

$$b_1 = r \cdot \frac{s_y}{s_x}$$

$$b_0 = \bar{Y} - b_1 \cdot \bar{X}$$

Analysis of score and count

- Linear Regration
- formula <- y ~ x
- add regression line
- add regression equation and R square value
- output.type as "expression“



4.3.1 Analysis of Score and Count

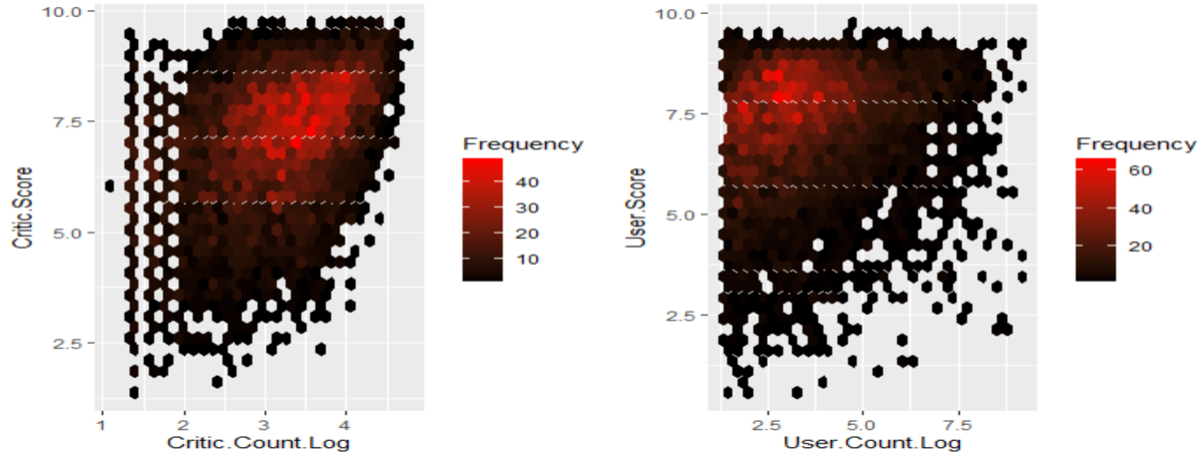
There is positive correlation between Critic.Score and User.Score. In total, Critic score is lower than user score.

Output :

```
T-Test   Welch Two Sample t-test
data:  game$Critic.Score and game$User.Score
 t = -6.5463, df = 13629, p-value = 6.108e-11
 alternative hypothesis: true difference in means is not equal to 0
 95 percent confidence interval:
  -0.2058518 -0.1109834
 sample estimates:
 mean of x mean of y
  7.027209  7.185626
```

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

Correlation Analysis



4.3.2 Correlation Analysis

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

4.5 Analysis of sales :

A. By Year.Release

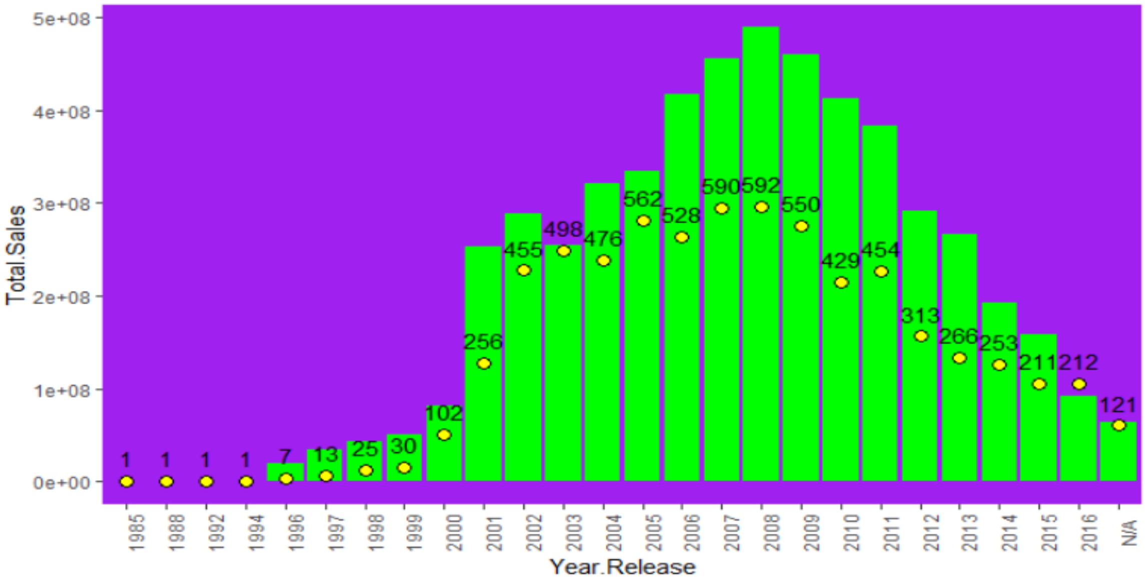


Figure 4.4.1 A : Total global sales and game released by year.

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

B. Analysis of sales - By Year.Release

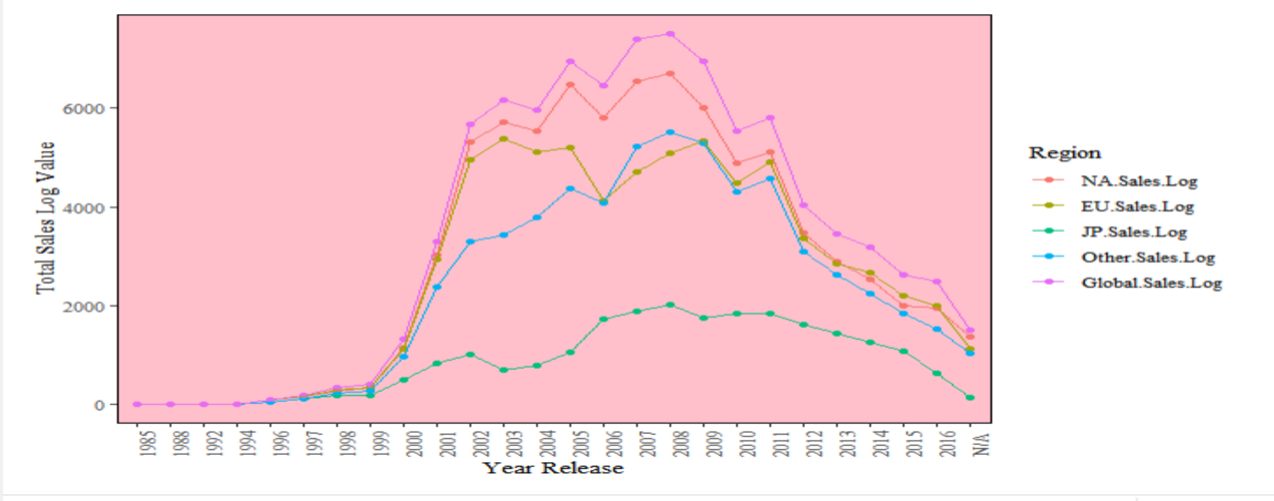


Figure 4.4.2 B: Year wise log global sales by region.

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

C. By Rating

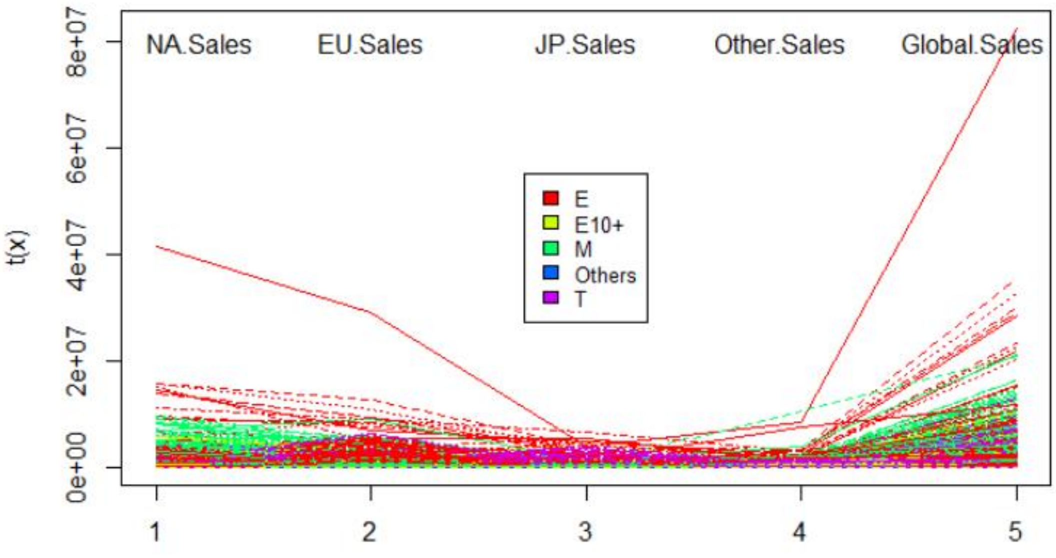


Figure 4.4.3 C: Sales by rating type.

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

C. By Genre

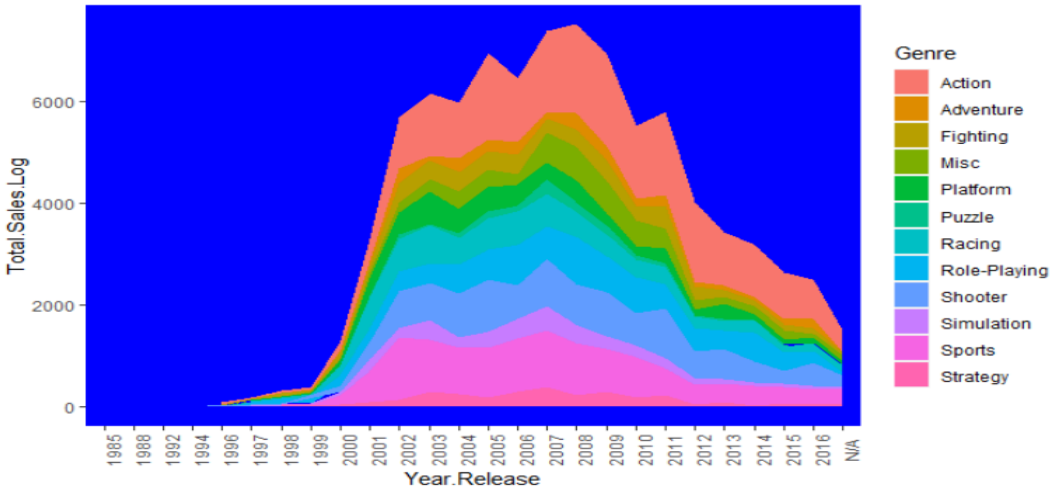


Figure4.4.4 D: Year wise log global sales by Genre.

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

4. 5 By Score

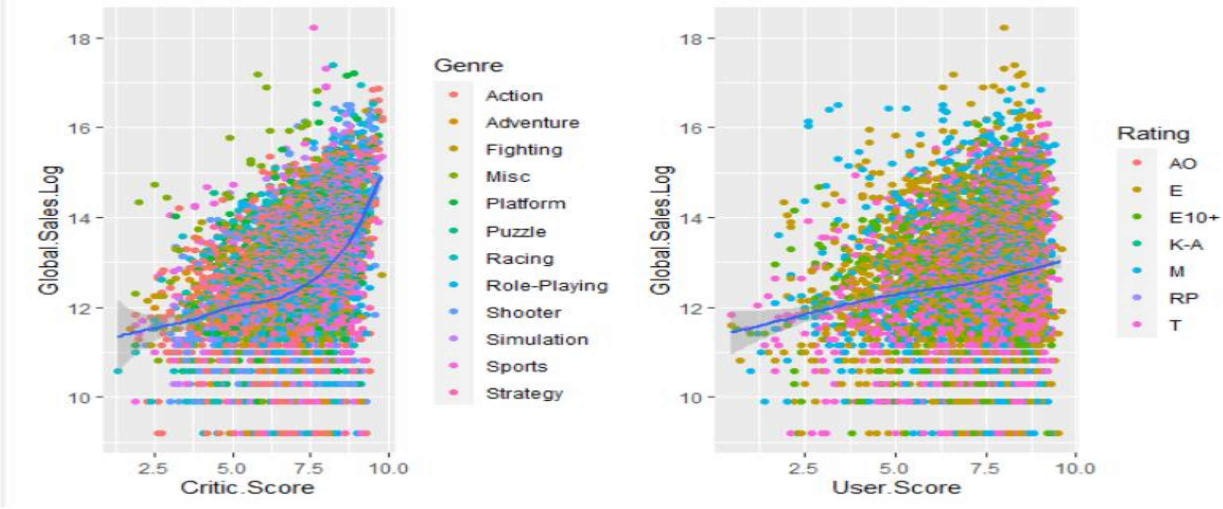
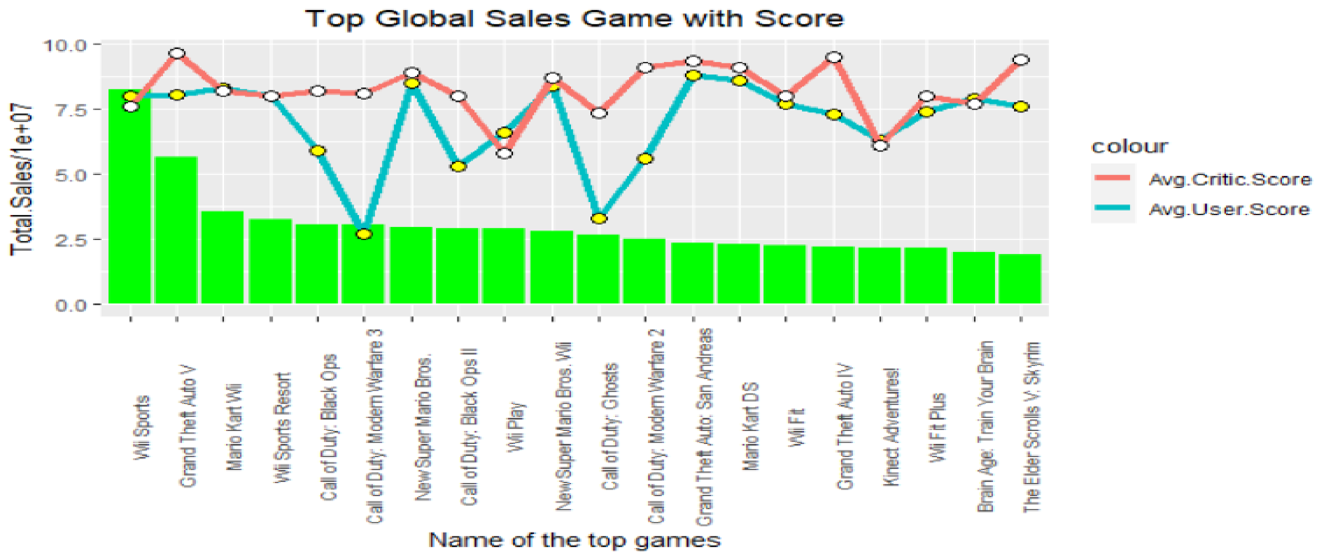


Figure 4.5.1: Global sales by critic and user score.

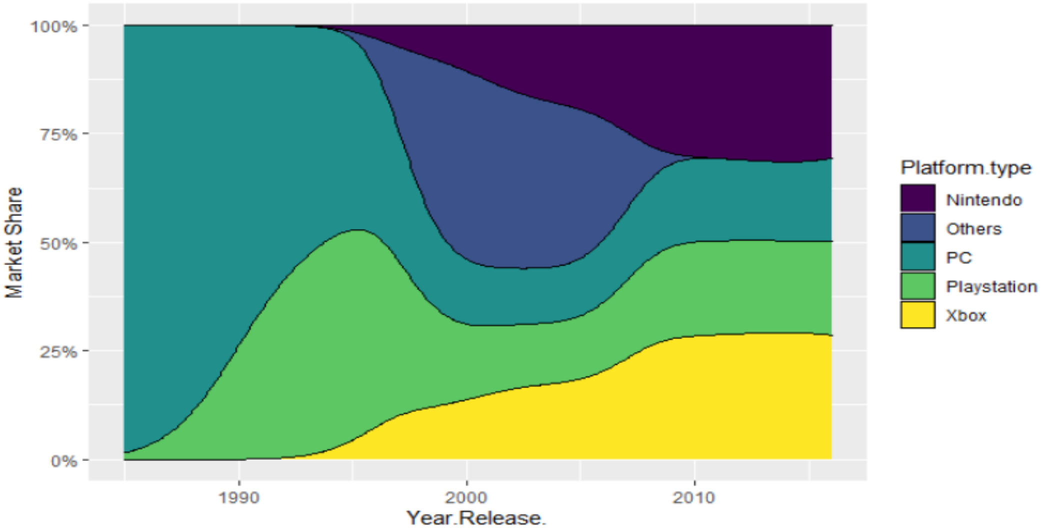
Independent from Genre and Rating, the higher of Critic Score, the better of Global.Sales.Log. Especially for Critic.Score bigger than 9, Global.Sales.Log straightly rises. Global.Sales.Log rises very slowly with User.Score.



4.5.2 Top Global Sales

Among these 20 top sale games, the first two games, Wii Sports and Grand Theft Auto V have much better sales than the others. For most games, average critic score is higher than average user score, which agree with our density plot in Figure 1.5. Two Call of Duty games got really lower average user score comparing with other top sales games.

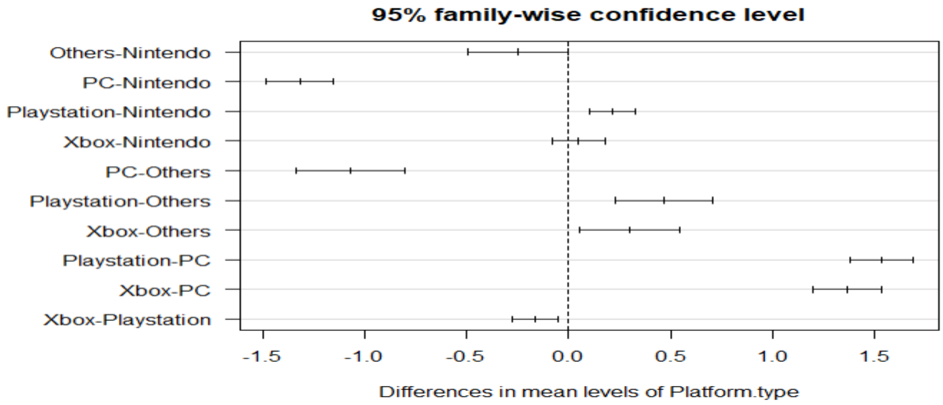
4.6 By Platform



4.6.1 By Platform

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

ANOVA



4.6.2 Compute 1-way ANOVA test for log value of global sales by Platform Type

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

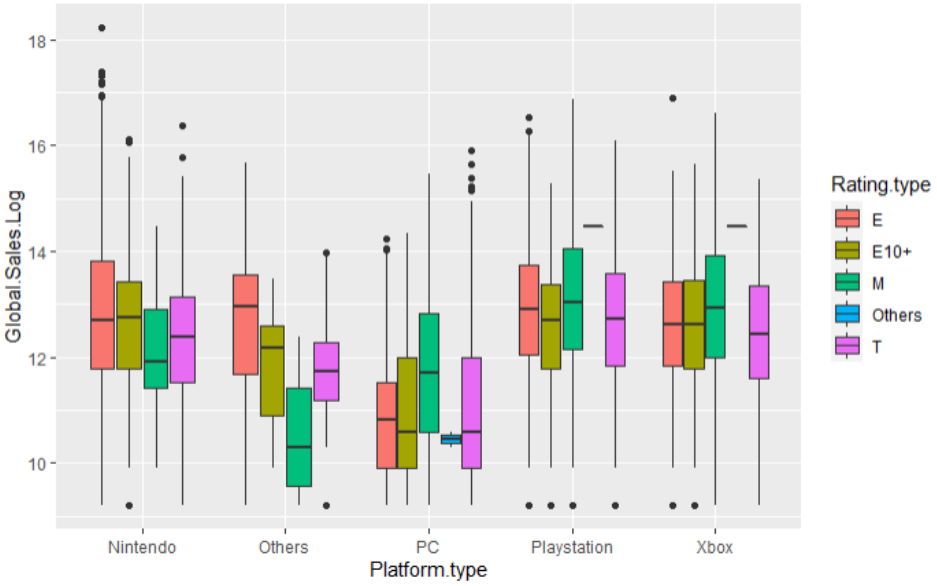


Figure 4.6.3: Global sales log by platform and rating type.

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

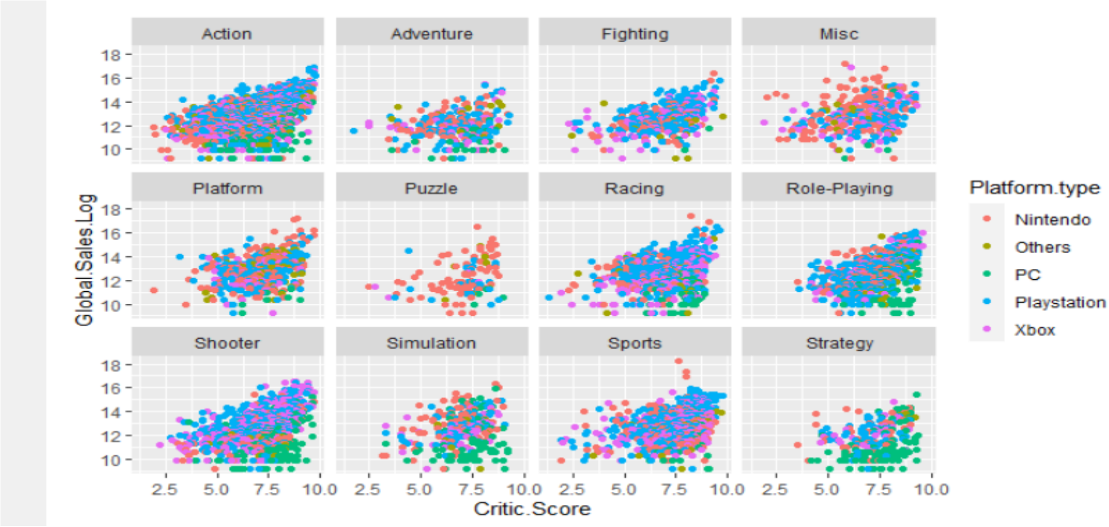


Figure 4.6.4: Global sales log by critic score for different platform type and genre.

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

4.6.2 Effect of platform type to priciple components

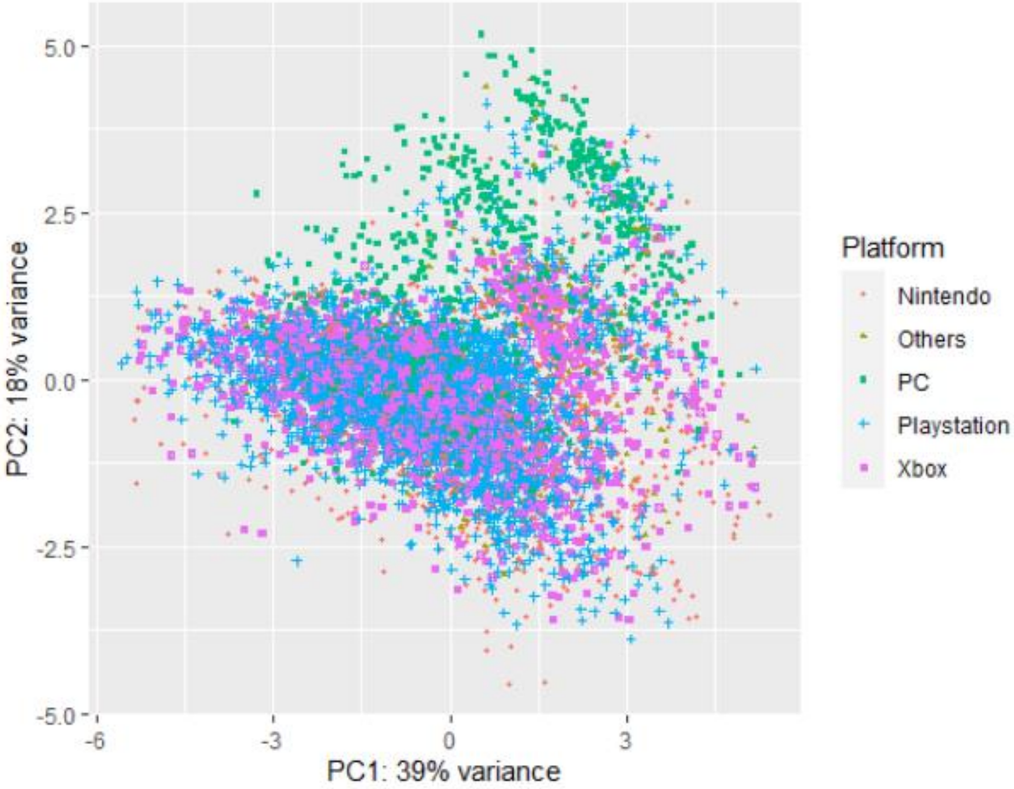


Figure 4.6.2.1 : PCA plot colored with platform type.

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

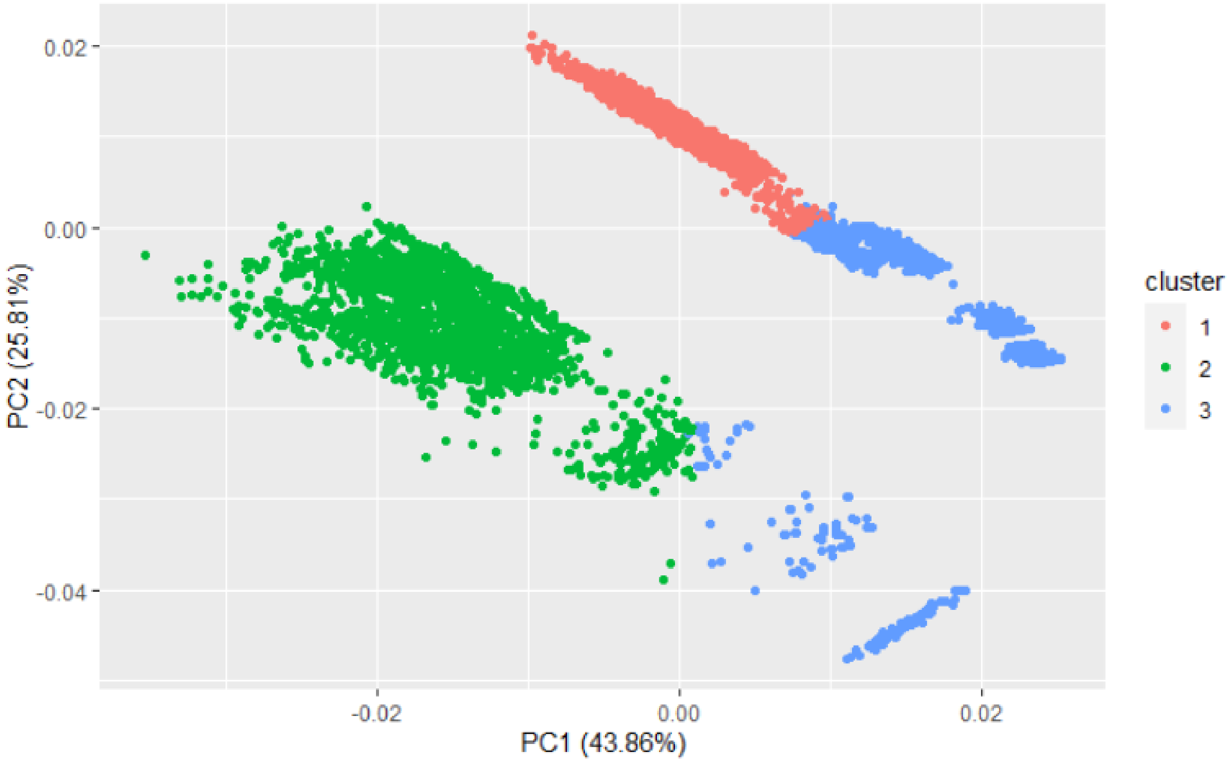


Figure 4.6.2.2: Kmeans PCA figure using ggfortify.

Together with PCA Figure 4.6.2.1, we will find that the first cluster is contributed mainly by PC and Playstation. The second cluster is contributed mainly by Xbox, Nintendo and Playstation. Playstation, Xbox, Nintendo, and PC all together build the third cluster.

4.7 **Models for global sales**

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

```
#re-categorize publisher into 13 groups
```

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

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

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

Algorithms:-

Step - 1: Re-categorize publisher into 13 group.

Step - 2: Sort Data.

Step - 3: Add new variables and preserves existing ones.

Step - 4: Add Linear Regressions.

Step - 5: Create Linear Regression Model using Global sales log values.

Step - 6: The coefficients are statistically significant, the model of two levels of critic score itself will explain the Global.Sales.Log.

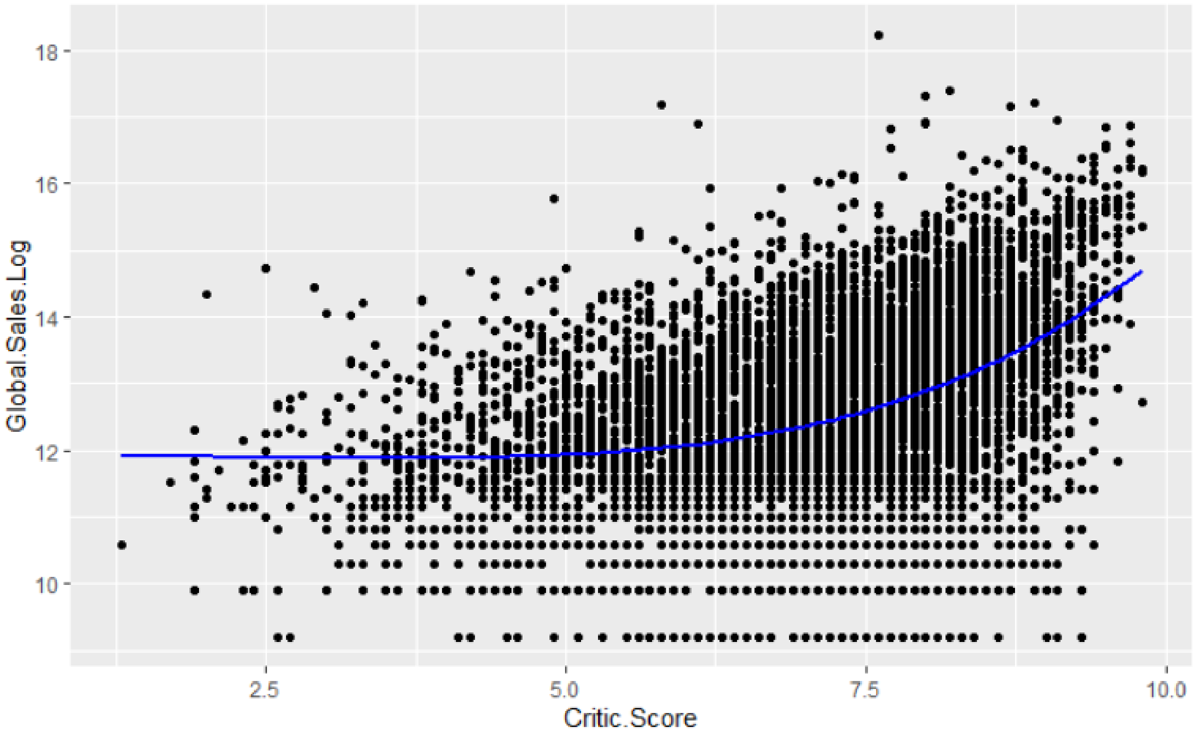
Step - 7: Create Model Function Coefficients.

```
Call:
## lm(formula = Global.Sales.Log ~ Critic.Score + I(Critic.Score^2) +
##     I(Critic.Score^3) + I(Critic.Score^4), data = game.lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0029 -0.7972  0.0916  0.8801  5.6201
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       12.030638   1.468428   8.193 3.02e-16 ***
## Critic.Score      -0.840810   1.095444  -0.768   0.4428
## I(Critic.Score^2)  0.385671   0.292104   1.320   0.1868
## I(Critic.Score^3) -0.060971   0.033147  -1.839   0.0659 .
## I(Critic.Score^4)  0.003434   0.001358   2.528   0.0115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.281 on 6820 degrees of freedom
## Multiple R-squared:  0.1623, Adjusted R-squared:  0.1618
## F-statistic: 330.4 on 4 and 6820 DF,  p-value: < 2.2e-16
```

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

```
summary(model)
##
## Call:
## lm(formula = Global.Sales.Log ~ I(Critic.Score^3) + I(Critic.Score^4),
##     data = game.lm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8634 -0.7892  0.0950  0.8837  5.5807
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.193e+01  7.115e-02 167.722  < 2e-16 ***
## I(Critic.Score^3) -2.989e-03  7.972e-04  -3.749 0.000179 ***
## I(Critic.Score^4)  6.076e-04  8.224e-05   7.387 1.67e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.283 on 6822 degrees of freedom
## Multiple R-squared:  0.1597, Adjusted R-squared:  0.1595
## F-statistic: 648.3 on 2 and 6822 DF,  p-value: < 2.2e-16
```

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



4.7 Model For Global Sales

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

Chapter 5: Predictive Data mining

Predictive Analytics will help an organization to know what might happen next, it predicts future based on present data available. It will analyse the data and provide statements that have not happened yet. It makes all kinds of predictions that you want to know and all predictions are probabilistic in nature.

PREDICTIONS MODELS:

Apply various possible prediction modelling algorithms to see which provides best results. Linear Regression, Decision Tree, Random Forest and Support vector regression algorithms were used on video game sales data.

The summary statistics above tells us a number of things.

One of them is the model's p-Value (in last line) and the p-Value of individual predictor variables (extreme right column under Coefficients). The p-Values are very important.

Because, we can consider a linear model to be statistically significant only when both these p-Values are less than the pre-determined statistical significance level of 0.05.

This can visually interpreted by the significance stars at the end of the row against each X variable. The more the stars beside the variable p-Value, the more significant the variable.

What is the Null and Alternate Hypothesis?

Whenever there is a p-value, there is always a Null and Alternate Hypothesis associated.

So what is the null hypothesis in this case? In Linear Regression, the Null Hypothesis (H0) is that the beta coefficients associated with the variables is equal to zero.

The alternate hypothesis (H1) is that the coefficients are not equal to zero. (i.e. there exists a relationship between the independent variable in question and the dependent variable).

What is t-value?

We can interpret the t-value something like this. A larger t-value indicates that it is less likely that the coefficient is not equal to zero purely by chance. So, higher the t-value, the better.

Pr(>|t|) or p-value is the probability that you get a t-value as high or higher than the observed value when the Null Hypothesis (the ? coefficient is equal to zero or that there is no relationship) is true.

So if the Pr(>|t|) is low, the coefficients are significant (significantly different from zero). If the Pr(>|t|) is high, the coefficients are not significant.

What this means to us?

When p Value is less than significance level (< 0.05), you can safely reject the null hypothesis that the co-efficient ? of the predictor is zero.

In our case, linearMod, both these p-Values are well below the 0.05 threshold.

So, you can reject the null hypothesis and conclude the model is indeed statistically significant.

It is very important for the model to be statistically significant before you can go ahead and use it to predict the dependent variable. Otherwise, the confidence in predicted values from that model reduces and may be construed as an event of chance

5.1 Linear regression (baseline model):

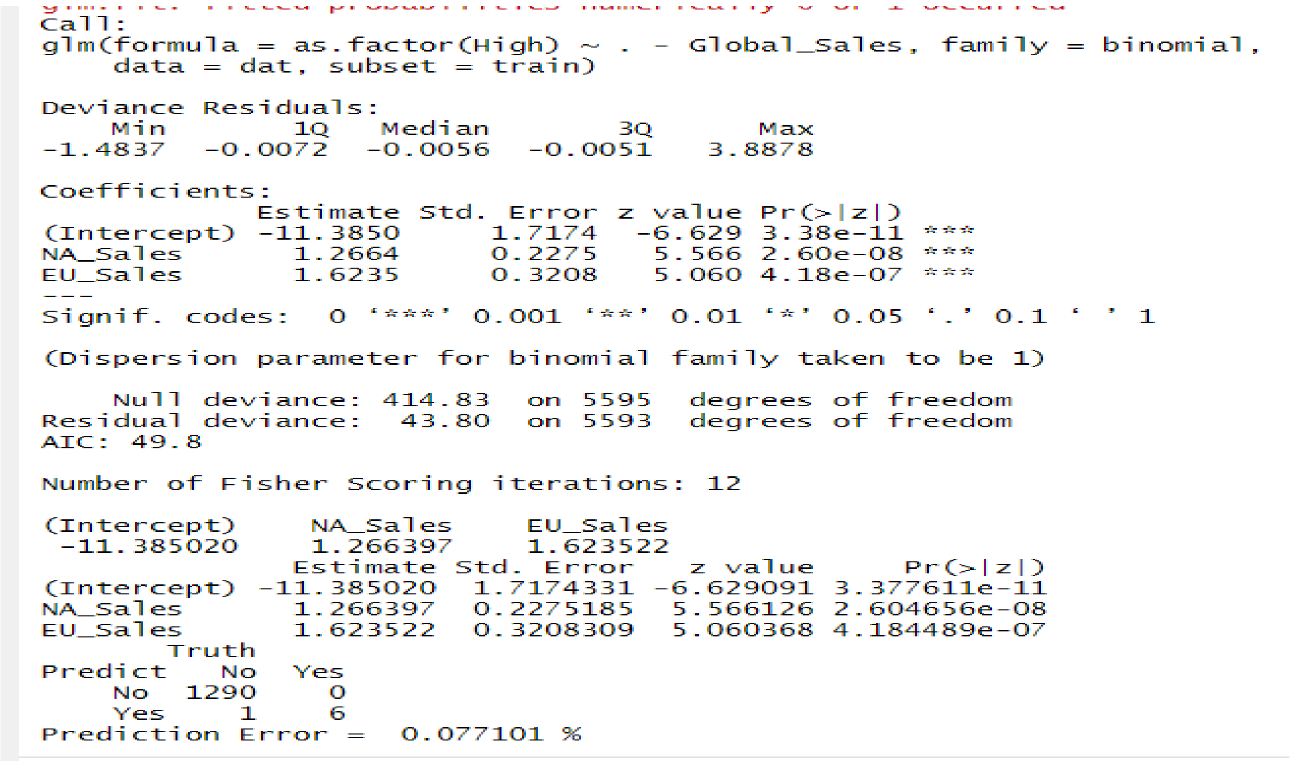
Linear regression is commonly used for predictive modelling techniques. The main theme of this algorithm is to find a mathematical equation for continuous variables Y when we have one or more X variables. This algorithm establishes a relation between two variables one variable is predicted variable and another one is result variable whose value is derived from the predictive variable. Y=aX + b

Function: model = lm (formula, data) Where Y is result variable X is predicted variable a and b are coefficients.

Algorithms:

Prediction through Linear Regression

```
train =( Year_of_Release <= 2011)
num_fact=game[,c("NA_Sales","EU_Sales","Global_Sales")]
High=ifelse(Global_Sales <=10.0,"No","Yes")
dat =data.frame(num_fact,High)
glm.fit=glm(as.factor(High)~.-Global_Sales,dat,subset=train,family=binomial)
summary(glm.fit)
coef(glm.fit)
summary(glm.fit)$coef
dat.test = game[!train,]
High.test=High[!train]
glm.prob=predict(glm.fit,dat.test,type="response")
glm.pred=rep("No",dim(dat.test)[1])
glm.pred[glm.prob >.5]="Yes"
table(Predict=glm.pred ,Truth=High.test)
cat("Prediction Error = ", mean(glm.pred!=High.test)*100,"%")
```



```
glm.fit: fitted probabilities numerically 0 or 1 occurred
Call:
glm(formula = as.factor(High) ~ . - Global_Sales, family = binomial,
    data = dat, subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4837  -0.0072  -0.0056  -0.0051   3.8878

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.3850     1.7174  -6.629 3.38e-11 ***
NA_Sales      1.2664     0.2275   5.566 2.60e-08 ***
EU_Sales      1.6235     0.3208   5.060 4.18e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 414.83  on 5595  degrees of freedom
Residual deviance:  43.80  on 5593  degrees of freedom
AIC: 49.8

Number of Fisher Scoring iterations: 12

(Intercept)    NA_Sales    EU_Sales
 -11.385020    1.266397    1.623522
              Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -11.385020  1.7174331 -6.629091 3.377611e-11
NA_Sales      1.266397  0.2275185  5.566126 2.604656e-08
EU_Sales      1.623522  0.3208309  5.060368 4.184489e-07
       Truth
Predict   No  Yes
    No  1290    0
    Yes    1    6
Prediction Error =  0.077101 %
```

Linear Regression by Generalized Linear Models –

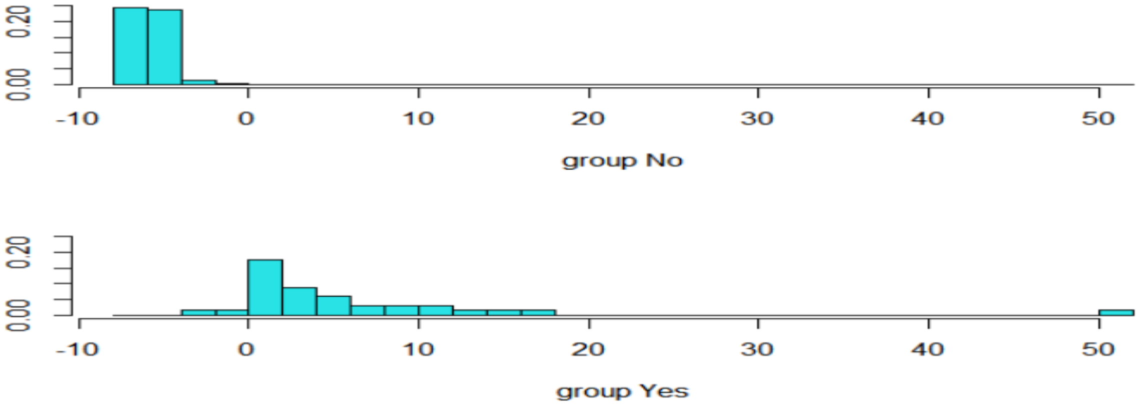
Prediction error of Global Sales will more than 10 million dollar, using North American Sales and European Sales data is 0.19%.

Linear Discriminant Analysis

```
lda.fit=lda(as.factor(High)~.-Global_Sales ,dat,subset=train)
summary(lda.fit)
plot(lda.fit)
dat.test = game[!train,]
High.test=High[!train]
lda.pred=predict(lda.fit,dat.test)
names(lda.pred)
lda.class=lda.pred$class
table(Predict=lda.class ,Truth=High.test)
cat("Prediction Error = ", mean(lda.class!=High.test)*100,"%")
```

```
Length          Class  Mode
prior     2       -none- numeric
counts   2       -none- numeric
means   4       -none- numeric
scaling  2       -none- numeric
lev        2       -none- character
svd       1       -none- numeric
N          1       -none- numeric
call       4       -none- call
terms     3       terms  call
xlevels  0       -none- list
```

```
[1] "class"      "posterior" "x"
       Truth
Predict    No      Yes
    No     1290    0
    Yes    1         6
Prediction Error =  0.077101 %
```



5.1 output of Linear Discriminant Analysis

Linear Regression by Linear Discriminant Analysis - Prediction error of Global Sales will more than 10 million dollar, using North American Sales and European Sales data is 0.14%.

5.2 Support vector regression: Support vector regression uses svm classification algorithm to forecast a continuous variable. But other regression models are used to minimize the error between predicted value and actual value [11]. SVR tries to fit best line among the predefined error value. Svr have few important key words such as kernel, hyper plane, boundary line, support vector. Support vector regressions have two types.

• **Linear SVR :-**

$\sum (\alpha i - \alpha i *). (xi , x) + b\ N\ i$=1 ... Eq.5 • Non linear SVR - $\sum (\alpha i - \alpha i *).K(xi , x) + b\ N$ i=1

• **Kernel function - polynomial:-**

$K(xi , xj)$ = $(xi\ .\ x$) d ... Eq.7 - Gaussian radial basis function:- $K(xi,xj)$ = exp [− ‖$xi-xj$‖ 2 2 σ 2] ...Eq.8

Prediction : Support Vector Machines

Algorithms :

Step - 1: Add data

Step - 2: Divide that data into Training - 70% and Testing - 30%.

Step - 3: Add Linear Classification.

Step - 4: Y. Train in the using if else loop takes sales data.

Step - 5: Add Data Frame.

Step - 6: Add SVMFIT Formula.

Step - 7: Create truth table.

Step - 8: Create Error Model.

Step - 9: Y.Test in using if else loop create and add sales data.

Step - 10 : Create data second model.

Step - 11: Write prediction formula.

Step - 12 : Create Second truth table.

Step – 13: Create Error Second Model

```
train =( Year_of_Release <= 2011)
game.train=game[train,] game.test = game[!train,]
```

Linear classification

```
y.train=ifelse(game.train$Global_Sales>10.0,1,-1)
dat=data.frame(x=game.train$NA_Sales+game.train$EU_Sales, y=as.factor(y.train))
svmfit=svm(y~., data=dat, kernel="linear", cost=10,scale=FALSE)
Summary(svmfit)
table(Model=svmfit$fitted , Truth=dat$y)
cat("Model Error = ", mean(svmfit$fitted!=dat$y)*100,"%")
y.test=ifelse(game.test$Global_Sales>10.0,1,-1)
dat.te=data.frame(x=game.test$NA_Sales+game.test$EU_Sales, y=as.factor(y.test))
pred.te=predict(svmfit, newdata=dat.te)
table(Predict=pred.te, Truth=dat.te$y)
cat("Prediction Error = ", mean(pred.te!=dat.te$y)*100,"%")
```

Output:

Prediction : Support Vector Machines

```
      Truth
Model   -1      1
   -1     5558    2
   1         5      32
Model Error =  0.125067 %


      Truth
Predict   -1    1
    -1    1290   0
     1     1        6
Prediction Error  =  0.077101 %
```

SVM Linear - Prediction error of Global Sales will more than 10 million dollar, using North American Sales and European Sales data is 0.070%

5.3 Random Forest: - Random Forest is a supervised machine learning algorithm creates randomly a forest with several trees . Why we use random forest instead of decision tree, decision trees are easy to implement and work efficiently with training data, but it gives less accuracy this happens due to over fitting . Over fitting occurs when a model trains the data to such an extent that is negatively impacts the performance of the model on new data. For this reason random forest comes into way. Function: train (formula, dataset, method=”rf”,trControl=trcontrol()) [where “rf” is random forest method].

Algorithms:

Step - 1: Normalize the data.

Step - 2: Selecting our predictors and translating status to status.

Step - 3: Create the normalized subset.

Step - 4: Split data into a training set and a set by randomly section.

Step - 5: Set 70% training data or reaming 30% test data.

Step - 6: Translate the values genre into genre as follow.

Step - 7: Summary of data in train and test Sets.

Step - 8: Create a random forest model with default parameter.

Step - 9: Predicting on training set.

Step - 10: Checking classification accuracy.

Step - 11: Predicting on test-set.

Step - 12: Checking Classification accuracy.

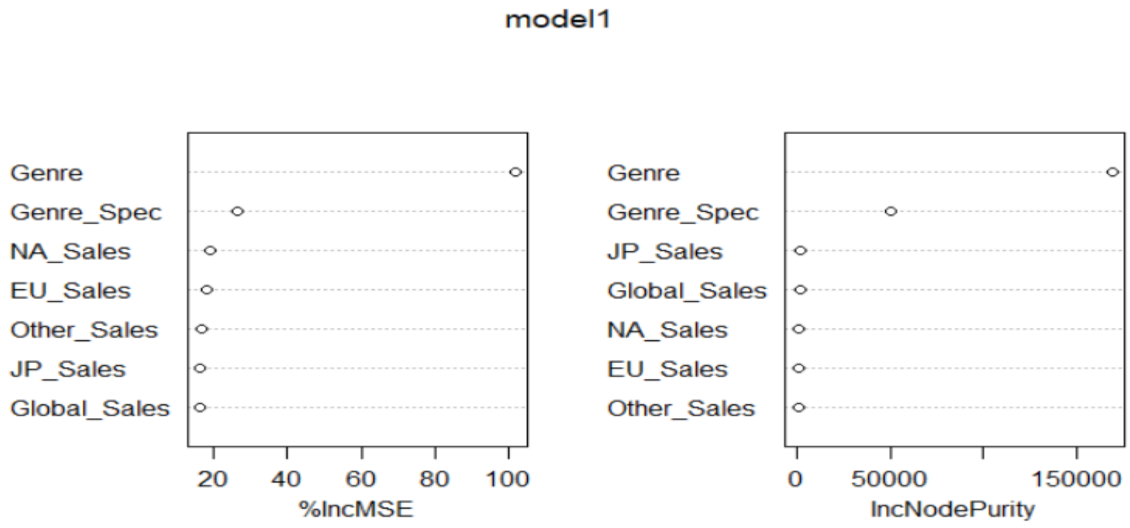
Step - 13: To check important variance.

Model-1:

```
Call:   randomForest(formula = Genre_No ~ ., data = train_set, importance = TRUE,
na.action = na.omit)
  Type of random forest: regression
     Number of trees: 500
  No. of variables tried at each split: 2
  Mean of squared residuals: 0.07436127
   % Var explained: 99.48
```

```
              %IncMSE   IncNodePurity
Genre        101.79744   168951.3354
NA_Sales      19.12285      883.5390
EU_Sales      18.31963      879.5777
JP_Sales      16.44414     1417.1303
Other_Sales   16.85203      504.1532
Global_Sales  16.35002     1343.1926
Genre_Spec    26.67618    50032.3924
```

Plot:



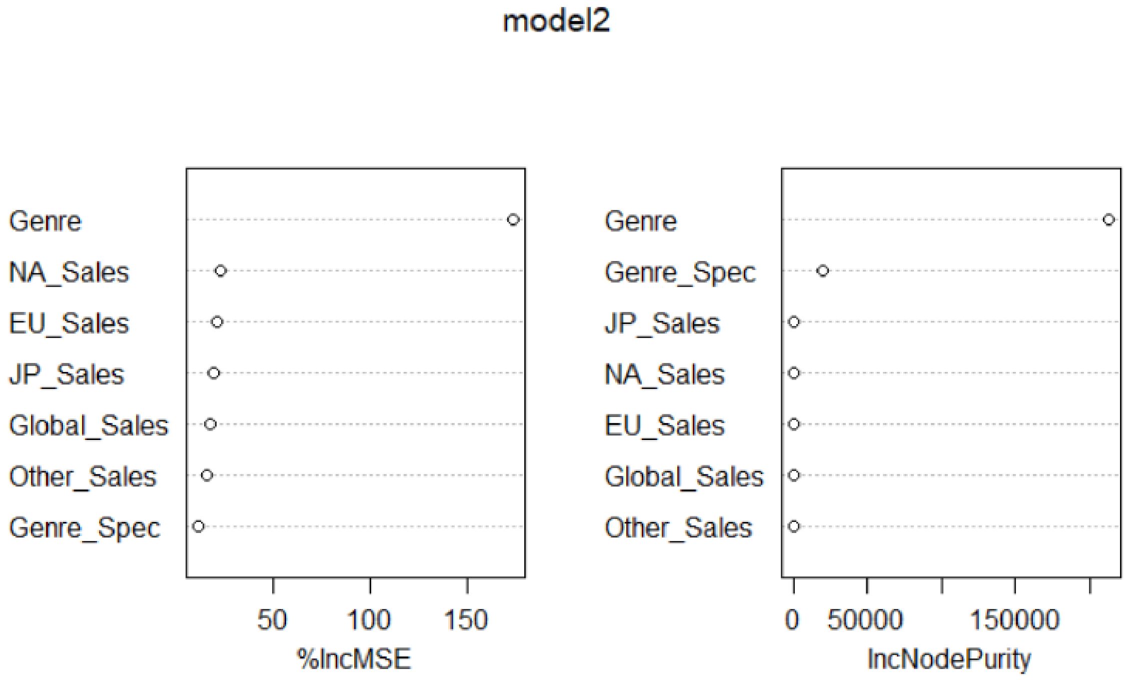
5.3.1 Model -1

Model-2:

```
Call:   randomForest(formula = Genre_No ~ ., data = train_set, ntree = 500,        mtry = 6, importance = TRUE, na.action = na.omit)
    Type of random forest: regression
      Number of trees: 500
   No. of variables tried at each split: 6
   Mean of squared residuals: 0.0002252335
   % Var explained: 100
```

```
                  %IncMSE     IncNodePurity
Genre          173.93003      2.131021e+05
NA_Sales        22.70438     6.594115e+01
EU_Sales        21.36621      1.620178e+01
JP_Sales        19.29366       8.169936e+01
Other_Sales     16.15890      7.722195e+00
Global_Sales    17.80056    1.452238e+01
Genre_Spec      11.56363    1.940848e+04
```

Plot:



5.3.2 Model -2

5.4 KNN Model

What is K-Nearest Neighbors (KNN)?

K-Nearest Neighbors is a machine learning technique and algorithm that can be used for both regression and classification tasks. K-Nearest Neighbors examines the labels of a chosen number of data points surrounding a target data point, in order to make a prediction about the class that the data point falls into. K-Nearest Neighbors (KNN) is a conceptually simple yet very powerful algorithm, and for those reasons, it's one of the most popular machine learning algorithms. Let's take a deep dive into the KNN algorithm and see exactly how it works. Having a good understanding of how KNN operates will let you appreciated the best and worst use cases for KNN.

Algorithm:

Step - 1: Normalize the data.

Step - 2 : Selecting our predictors and translating status to status.

Step - 3: Create the normalized subset.

Step - 4: Split data into a training set and a set by randomly section.

Step - 5: Set 70% training data or reaming 30% test data.

Step - 6: Creating separate data frame for 'Status', feature which is our target.

Step - 7: Find the number of observation.

Step - 8: Square root of 11,403 is 106.78 and so we will create 2 models.

Step - 9: one with 'k' value as 106 and the other model with a 'k' value as 107.

Step -10: Calculate the proportion of correct classification for k=106, 107.

Step - 11: Loop the Calculate the accuracy of the KNN model.

Step - 12: Accuracy plot.

In the KNN Model, we began with selecting the predictors and left out the Genre data (and others) as it's our target study.

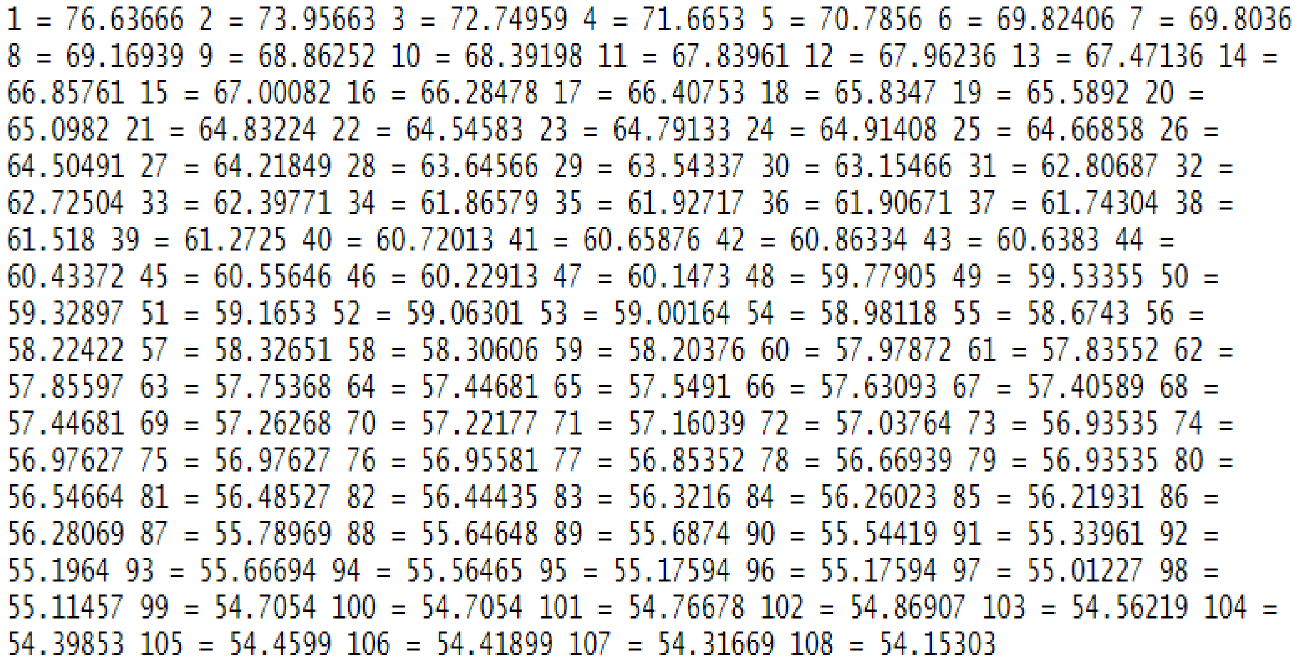
We built the model with the data normalized and used the square root of the number of observations to obtain the K value of 106 and 107. Then we calculated the accuracy of these created models.

## Square root of 11,403 is 106.78 and so we will create 2 models.

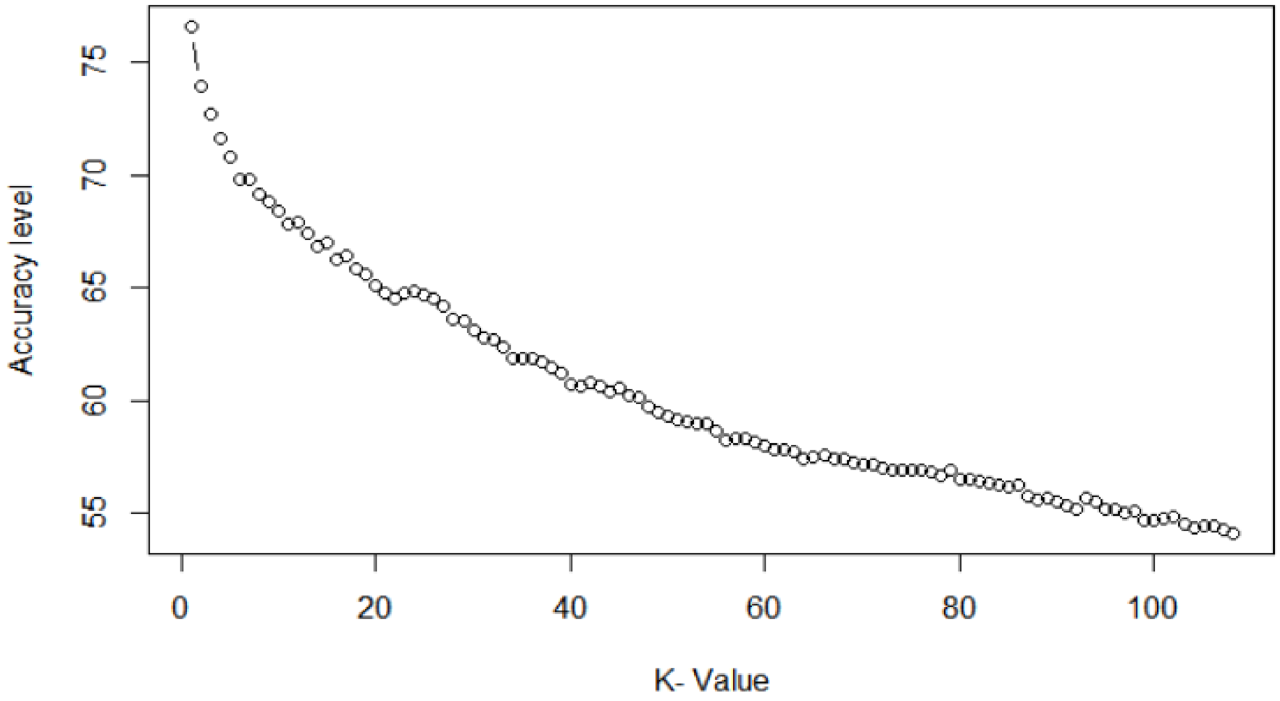
## One with 'K' value as 106 and the other model with a 'K' value as 107.

```
knn_106 <- knn(train=train_set, test=test_set, cl=train_labels$NA_Sales, k=106)

knn_107 <- knn(train=train_set, test=test_set, cl=train_labels$NA_Sales, k=107)

[1] 11403 [1] 54.31669 [1] 54.29624
```

We also created a loop that calculates the accuracy of the KNN model for 'K' values ranging from 1 to 108. This way we can check which 'K' value will result in more accurate model.



```
1 = 76.63666 2 = 73.95663 3 = 72.74959 4 = 71.6653 5 = 70.7856 6 = 69.82406 7 = 69.8036
8 = 69.16939 9 = 68.86252 10 = 68.39198 11 = 67.83961 12 = 67.96236 13 = 67.47136 14 =
66.85761 15 = 67.00082 16 = 66.28478 17 = 66.40753 18 = 65.8347 19 = 65.5892 20 =
65.0982 21 = 64.83224 22 = 64.54583 23 = 64.79133 24 = 64.91408 25 = 64.66858 26 =
64.50491 27 = 64.21849 28 = 63.64566 29 = 63.54337 30 = 63.15466 31 = 62.80687 32 =
62.72504 33 = 62.39771 34 = 61.86579 35 = 61.92717 36 = 61.90671 37 = 61.74304 38 =
61.518 39 = 61.2725 40 = 60.72013 41 = 60.65876 42 = 60.86334 43 = 60.6383 44 =
60.43372 45 = 60.55646 46 = 60.22913 47 = 60.1473 48 = 59.77905 49 = 59.53355 50 =
59.32897 51 = 59.1653 52 = 59.06301 53 = 59.00164 54 = 58.98118 55 = 58.6743 56 =
58.22422 57 = 58.32651 58 = 58.30606 59 = 58.20376 60 = 57.97872 61 = 57.83552 62 =
57.85597 63 = 57.75368 64 = 57.44681 65 = 57.5491 66 = 57.63093 67 = 57.40589 68 =
57.44681 69 = 57.26268 70 = 57.22177 71 = 57.16039 72 = 57.03764 73 = 56.93535 74 =
56.97627 75 = 56.97627 76 = 56.95581 77 = 56.85352 78 = 56.66939 79 = 56.93535 80 =
56.54664 81 = 56.48527 82 = 56.44435 83 = 56.3216 84 = 56.26023 85 = 56.21931 86 =
56.28069 87 = 55.78969 88 = 55.64648 89 = 55.6874 90 = 55.54419 91 = 55.33961 92 =
55.1964 93 = 55.66694 94 = 55.56465 95 = 55.17594 96 = 55.17594 97 = 55.01227 98 =
55.11457 99 = 54.7054 100 = 54.7054 101 = 54.76678 102 = 54.86907 103 = 54.56219 104 =
54.39853 105 = 54.4599 106 = 54.41899 107 = 54.31669 108 = 54.15303
```



5.4.1 KNN Plot

The below graph shows that for ‘K’ value of 1 got the maximum accuracy at 76.636.

5.5 <u>Decision Tree: -</u> A decision tree is a type of supervised machine learning tree which explains about "what the input is and what is the relevant output according to the our data [13]". The main objective of this algorithm is to predict the value of a target variable.

Mostly the decision tree rules are in the form of conditional statements i.e. "if-then-else". Decision trees are used for both classification and regression problems.

Function: train(formula, dataset, method="rpart", trcontrol = trcontrol())

Algorithms :

Step - 1: Create or add sales data columns.

Step - 2: Create High.train(variable where we store the formula) data table or formula.

Step - 3: Create tree high formula using Global sales data and training data.

Step - 4: add summery of tree.

Step - 5: plot the data or add text.

Step - 6: Creating test data help of train data.

Step - 7: include high formula for test data.

Step - 8: Create tree test for and add the classes.

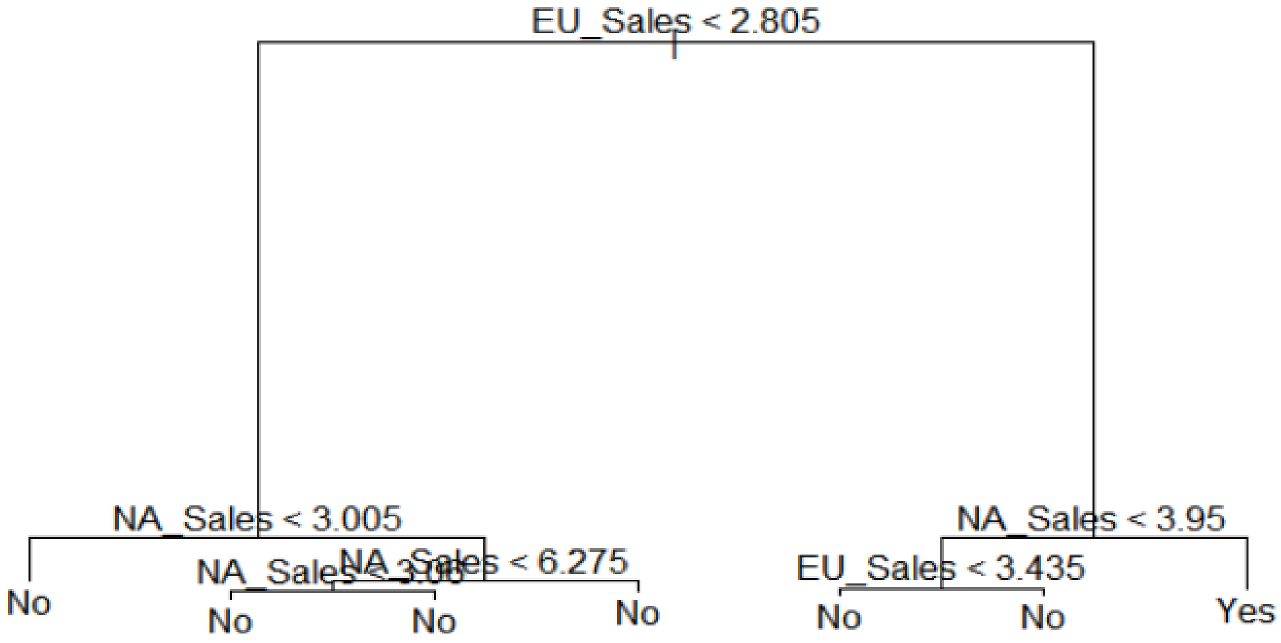
Step - 9: Create table using truth table.

Step - 10: Note Error Prediction table.

Classification tree

```
tree(formula = as.factor(High) ~ . - Global_Sales, data = dat, subset = train)
Number of terminal nodes:  7
Residual mean deviance:  0.005271 = 31.15 / 5910
Misclassification error rate: 0.001183 = 7 / 5917
```

```
   Truth
Predict   No   Yes
    No      971   1
    Yes      1      3
Prediction Error =  0.204918 %
```



5.5.1 Decision Tree

Chapter 6: RESULT AND DISCUSSIONS

For performing the quantitative analysis we have taken few methods, the performance metric value needed to be computed and they are to be compared with the other. Hence, for performing the calculations of the performance metric there are a few formulas which can be utilized for achieving the performance value from the dataset. The formulae for the calculation of the performance metrics are given below in table.

METRIC	FORMULA
Accuracy	(TP+TN)/(TP+TN+FP=FN)
Error Rate	100 - Accuracy
Mean Absolute Error	$1/n \sum \lvert y - y\hat{}\rvert$
Root Mean Square Error	$\sqrt{1/n \sum \lvert y - y\hat{}\rvert}$
Precision	TP/(TP+FP)
F-Value	2/(1/P+1/R)

Table.1. Quantitative Analysis

Sr.No	Algorithm	RMSE VALUE	Accuracy
1	Random Forest	1.4648	0.9605
2	Support vector regression	1.9773	0.8154
3	Decision Tree	2.8762	0.8036
4	Linear Regression	2.4830	0.5734

Table.2. Accuracy and RMSE values

Descriptive mining

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

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

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

Predicting mining

Prediction outcome will gives predicted values of a dataset after applying machine learning algorithms. Among all the algorithms the best algorithm with accuracy can be determined. In order to keep with the results random forest technique offers the best result among alternative algorithms.

MODEL COMPARISON:

In this section, we contrast the predictive effects of linear regression, support vector regression, random forest and decision tree models. We choose linear regression as the baseline model and introduce its prediction result into the result comparison.

The baseline model (linear regression)'s accuracy result, which is 75%, is not ideal. Each of the models we select performs better than the baseline model. Random forest, Decision tree and support vector regression are 96%, 80%, 85% respectively. It is observed that random forest model performs the best among them, with the result of 96%.

Compared with the other prediction models random forest model plays a prominent role in improving video game sales prediction accuracy in the field of large-scale product sales data.

KNN Results:

In the KNN Model, we observed that training set produced starting points for the experiment with K = 106 and K = 107. They yielded 54.50082 and 54.41899 respectively. These values were average and did not perform as well as we had hope.

We believe that data points being further apart contributed to the outcome. K = 1 produced the most accurate result as we saw in the plot prior, which also aligned correctly with the fact that the game ranking 1st was also that one that had the highest sales. This method can be further tuned.

Random Forests Results:

Using Random Forests Model, we extrapolated that as the mean of squared residuals decrease then the % variance is increased. This is a common characteristic in the Random Forests Model. We chose Model 2 to further test and computed the variable importance values to see the effects.

As shown, the Genre variable importance was 191.67, the NA_Sales variable importance was 20.85 and the Genre_Spec variable importance was 10.19. The decreasing variable importance trend seemed appropriate as we gave more weight to the Genre variable in developing this model. The model performed as we suspected but again, can be further refined.

Overall, we thought the Random Forests Model was the most complex of the 3 models we utilized. Transparently, we intended to experiment and learn from the experience. We will take our insights and improve our future analyses.

We began the project with a broad look at the video game sales data and as we visualized and explored more, we begin to understand the trends and implications.

With data-driven knowledge, we developed our algorithm and models to experiment and tuned our prediction practices.

We applied machine learning techniques that went beyond standard linear regression for our video game sales data set. We generated our data set, interpreted the data, performed various algorithms/modelings and presented our insights.

Chapter 7: CONCLUSION

Sales prediction is a crucial part of the strategic planning process. It allows a company to forecast how the company will perform in the future. Predicting sales of a company is not only for planning new opportunities, but also allow knowing the negative trends that appear in the prediction. Finally we conclude that prediction of sales on video games has done and we observed which game has more sales in the market globally. For predicting sales of video games we applied several machine learning algorithms (Linear regression, Random Forest, Decision tree, Support vector regression). Among all these algorithms random forest gave us the best accurate result with minimum error rate.

Chapter 8: LIMITATIONS

We had some limitations in our study. One limitation is that some of our data were missing. We removed the video games without a rating score as it would be hard to properly compare all the data with these data missing. We still had over 7,000 data points. There were a few variables not included that we thought might have been important to consider such as gender, as many males and females play different types of video games; it would be interesting to see if female-targeted games have different global sales numbers than male-targeted games, and to what extent. It is clear that we could have done a better job in terms of handing the dummy variable coding, especially for genre, and we need to re-examine how the coding of platform took place in the analysis process. However, we still believe that what the model revealed was accurate, even allowing for the faulty recoding of genre. We cannot say what the effects of genre and platform are, and this may have contributed to the relatively low, albeit highly significant, value of R2 .

Chapter 9: Future Scope

The main objective of data visualization is the overall idea about the data mining model .In data mining most of the times we are retrieving the data from the repositories which are in the hidden form. This is the difficult task for a user. So this visualization on of the data mining model helps us to provide utmost levels of understanding and trust. The

data mining models are of two types:

Predictive and Descriptive.

The predictive model makes prediction about unknown data values by using the known values. Ex. classification, Regression, Time series analysis, Prediction etc. The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc. Many of the data mining applications are aimed to predict the future state of the data. Prediction is the process of analysing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or classes, this is a supervise learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that map data item to real valued prediction variable. In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis is used for many statistical techniques which will analyze the time-series data such as auto regression methods etc.It is some times used in the two type of modelling (1) ARIMA (II)Long

-memory time -series modelling .The term clustering means analyzes the different data obj

ects without consulting a known class levels. It is also referred to as unsupervised learning or segmentation. It is the partitioning or segmentation of the data in to groups or clusters. The clusters are defined by studying the behaviour of the data by the domain experts. The term segmentation is used in very specific context; it is a process of partitioning of database in to disjoint grouping of similar tuples. Summarization is the technique of presenting the summarize information from the data. The association rule finds the association between the different attributes. Association rule mining is a two-step process:

Finding all frequent item sets, Generating strong association rules from the frequent item sets. Sequence discovery is

a process of finding the sequence patterns in data. This sequence can be used to understand the trend

Details of Computations

1. Correlation Matrix (Hierarchical Agglomerative Clustering)(HAC) :

A correlation matrix is a table showing correlation coefficients between sets of variables. Each random variable (Xi) in the table is correlated with each of the other values in the table (Xj). This allows you to see which pairs have the highest correlation

Step 1: [(x,y)(a,b)] = $\sqrt{(\boldsymbol{x-a})\boldsymbol{2}} + (\boldsymbol{x-b})\boldsymbol{2}$

Form this formula we can find the matrix point through this point we can check the correlation between 2 points

(p1,p3),(p2,p3).

Step 2: plot the correlation matrix. Find the minimum correlation and maximum correlation.

Step 3: to update the distance matrix

MIN[dist(p2,p3),p1]

MIN[dist(p2,p1),(p3,p1)]

Step 4: from this formula we can create clustering dendogram.

Example: Data table

	x	y
P1	8.0	8.3
P2	7.6	8.2
P3	5.7	6.5
P4	3.9	4.2
P5	28.9	29.9
Mean	54.1/5 = 10.82	57.1/5 = 11.42

Pair wise distance matrix

	P1	**P2**	**P3**	**P4**	**P5**
P1	**0**				
P2	**3**	**0**			
P3	**2.18**	**2.40**	**0**		
P4	**5.79**	**5.44**	**7.50**	**0**	
P5	**62.77**	**29.77**	**52.10**	**63.15**	**0**

Euclidean distance between two instances

P1,p2

[(x,y)(a,b)] = $\sqrt{(\boldsymbol{x}-\boldsymbol{a})\mathbf{2}} + (\boldsymbol{y}-\boldsymbol{b})\mathbf{2}$

$= \sqrt{(\mathbf{8.0}-\mathbf{7.6})\mathbf{2}} + (\mathbf{8.3}-\mathbf{8.3})\mathbf{2}$

(P1,p2) = 3

HAC – Linkage criterion

For clustering dendrogram

Distance between p1, to (p2,p3)

Coordinate of the cluster (p2,p3) (7.6 +5.7/2 = 10.45, 8.2+6.5/2 =11.45)

10.45 and 11.45 is the centrode values we will get in hierarchical cluster where 2 cluster find the centre of the cluster

D^2 = p2*p3/p2+p3 = 7.6*5.2/7.6+5.2 = 10.4

We obtain an indexed hierarchy. The merging levels correspond to the measure of dissimilarity between the two groups.

2. Analysis of score count

1. Distance = estimate – mean
2. R2 = actual – mean

Y = b0+b1x (find regression line)

$\boldsymbol{R^2 = \sum(\hat{}y - {}^{-}y)2 \,/\! \sum(y - {}^{-}y)2}$

^y = regression points

t-Test: test significant of an observed sample correlation coefficient.

H0 : Population correlation coefficient µ zero (ϱ =0)

H1: ϱ ! = 0

T = r/ $\boldsymbol{\sqrt{1} - r^2}$ * $\boldsymbol{\sqrt{n} - 2}$

Df , v = n-2

R = sample correlation coefficient.

N = sample size

Example: Data table

	x	y	x-¯x	y-¯y	(x-¯x)^2	(x-¯x)(y-¯y)	^y	^y-y¯	(y^-y¯)^2	R^2
P1	8.0	8.3	0.02	-0.7	0.04	0.014	8.62	2.8	7.84	0.86
P2	7.6	8.2	-3.22	-3.22	10.36	10.36	0	0	0	0
P3	5.7	6.5	-5.12	-4.92	26.21	9.83	9.3	2.11	4.45	0.18
P4	3.9	4.2	-6.92	-7.22	47.88	49.96	7.35	4.07	16.56	0.31
P5	28.9	29.9	18.08	41.3	326.88	746.7	4.89	6.58	43.24	0.025
Mean	54.1/5 = 10.82	57.1/5 = 11.42								

Y¯ = b0 + b1x

B1 = (x-¯x) (y-y¯)/ (x-x¯)2 ,

B1 = -0.014/0.04

= -0.35

11.42 = b0 +(-0.35)(8.0)

11.42 = b0 + (-2.8)

8.62 = b0

Y^ = 8.62+-0.35

R^2 = ∑(y^ - y¯)^2 / ∑(y-y¯)^2

3. Knn : K Nearest Neighbours - Classificationn

Example: Data table

	x	y
P1	8.0	8.3
P2	7.6	8.2
P3	5.7	6.5
P4	3.9	4.2
P5	28.9	29.9
Mean	54.1/5 = 10.82	57.1/5 = 11.42

We can now use training set to classify an unknown cases (x=x10.82 and y = 11.42) using Euclidean distance if k=1 then the nearest neighbour is the last case

In this calculation we are also using the Euclidean distance

$$\mathbf{D} = \sqrt{(\boldsymbol{x1} - \boldsymbol{y1})^{\mathbf{2}}} + (\boldsymbol{x2} - \boldsymbol{y2})\text{\textasciicircum}\mathbf{2}$$

$$= \sqrt{(\mathbf{28.9} - \mathbf{3.9}\ )\mathbf{2}} + (\mathbf{29.9} - \mathbf{4.2})\mathbf{2}$$

d = 35.85

using k =3 find the 3 closest neighbours. The prediction for the unknown cases.

Same Last Step We have to perform again to find the nearest neighbours

References with strictly IEEE format

N. Y. Prathama, R. Asmara and A. R. Barakbah, "Game Data Analytics using Descriptive and Predictive Mining," 2020 International Electronics Symposium (IES), 2020, pp. 398-405, doi: 10.1109/IES50839.2020.9231949. “Research Paper ”.

Chuyachia. (2017, August 21). Video games sales prediction. Kaggle. https://www.kaggle.com/chuyachia/video-games-sales-prediction.

[9]2021.[Online].Available:http://www.stat.columbia.edu/~madigan/DM08/descriptive.ppt.pdf-. [Accessed: 09- Aug- 2021]

[10]2021.[Online].Available:https://www.researchgate.net/publication/267558504_Mining_Enrolment_Data_Using_Predictive_and_Descriptive_Approaches -. [Accessed: 09- Aug- 2021]

Fasih, “Linear regression: Predict using linear regression in r,” Analytics Vidhya, 14-Dec-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/12/predicting-using-linear-regression-in-r/. [Accessed: 10-Aug-2021]

Zev@zevross.com, “Predictive modeling and machine learning in r with the caret package,” Technical Tidbits From Spatial Analysis & Data Science, 02-Oct-2018. [Online]. Available: http://zevross.com/blog/2017/09/19/predictive-modeling-and-machine-learning-in-r-with-the-caret-package/. [Accessed: 10-Aug-2021]

THANK YOU