Project Report

Video games Sales Prediction

TEAM MEMBERS:

Harsha Vardhan Elluru—---->HXE210013
Pooja Ananthula—-------->PXA220007
Afsaruddin Mohammed—------->AXM210415
Chaitanya Kommineni—------->CXK220004
Sai Manikanta Anumalasetty—-->AXM220023

ABSTRACT:

Playing video games for many years has led to a large volume of gaming data that consist of gamer's likings and their playing behavior. Such data can be used by game creators to extract knowledge for enhancing games. Most of the video gaming business organizations highly depend on a knowledge base and demand prediction of sales trends. However, no studies are conducted to work out the variables that inspire industrial sales predict involvement in and contribution to the sales prediction method. This project relates to the sales of these video games based on different regions and analyzes the sales. Also, we have analyzed which genre, platform or publisher is the most popular and has the maximum number of sales. The idea was to visualize the sales for different genres, publishers, and platforms. This would give the basic idea about the most popular genres, publishers, and platforms amongst all. Also analyzing the effect of genres on sales in different regions. Predictive modeling has long been the goal of many individuals and organizations. This science has many techniques, with simulation and machine learning at its heart. Machine learning techniques are very effective tools in extracting hidden knowledge from an enormous dataset to enhance accuracy and efficiency in predictions. In this project, we briefly analyzed the concept of gaming sales data and sales predictions. The various machine learning techniques and measures used for this sales prediction. Based on a performance evaluation, best suited predictive models like linear regression, support vector regression, random forest and decision trees etc. are used for the sales trend predictions.

PROPOSED WORK:

This project uses a special video game sale dataset sold in different countries.We collected our data from the data website, Kaggle; the dataset was titled Video Game Sales, and was released in 2016 and used data from the website, vgchartz. We cleaned up the data by removing certain variables that we thought were insignificant to our research such as the year the video game was published and the specific name of the developer who released the game. We then ran a stepwise regression analysis, to identify key factors that contribute to the final model. Data cleaning involves Removing rows with missing values, Transforming non-numeric values, Refining categorical variables, Transforming time variables, Removing variables based on business rules, Removing dependent variables.As a result, the new dataset with 11 columns will be used in analysis.The project is trying to explore what would be key factors to determine the sales of a game and to predict how many percentages of newly developed games could be successful. We Evaluated all columns to find out one or several the ultimate determinant(s) of game sales. The conclusion gives publishers and developers actionable strategies to achieve higher sales when developing new games. We also analyzed when games have been released and tried to predict potential success rate. In this project we performed Exploratory data analysis on video games sales data for creating visualizations in order to answer different business problems.After completion of exploratory data analysis we have to split the dataset into training and testing with split ratio 70 and 30. Training set contains 70% of data and the test set contains 30% of data. For predicting sales of video games we applied several machine learning algorithms (Linear regression, Random Forest, Decision tree, Support vector regression, Logistic Regression, KNN,etc.). After running these all algorithms, accuracy was calculated to check which algorithm best fits the model. Decision trees and KNN proves to be the best to fit the given dataset. (Both training and testing data set).

DATA SUMMARY:

In this project we choose video game sale data, our dataset consists of 16 variables with a combination of categorical and numeric variables. They are Rank, Name of video game, Platform, Year, Genre, Publisher, North American Sales, Europe Sales, Japan Sales, Other Sales and Global Sales.

Feature	Explanation	Data type	Count of Data
Name	Name of the game	Factor	16717
Platform	Game console	Factor	16719
Year_of_Release	Year of the Game's release date	int	16450
Genre	Game type (action, sports, etc.)	Factor	16717
Publisher	Game studio	Factor	16665
NA_Sales	Sales in north america	num	16719
EU_Sales	Sales in europe	num	16719
JP_Sales	JP_Sales	num	16719
Other_Sales	Sales in other regions	num	16719
Global_Sales	Sales around the globe	Num	16719
Critic_score	Aggregate score compiled by Metacritic staff	int	8137
Critic_Count	The number of critics used in coming up with the Critic Score	int	8137
User_Score	Score by Metacritic's subscribers	Factor	10015

User_Count	Number of users who gave the user score	int	7590
Developer	Party responsible for creating the game	Factor	10096
Rating	The ESRB ratings	Factor	9950

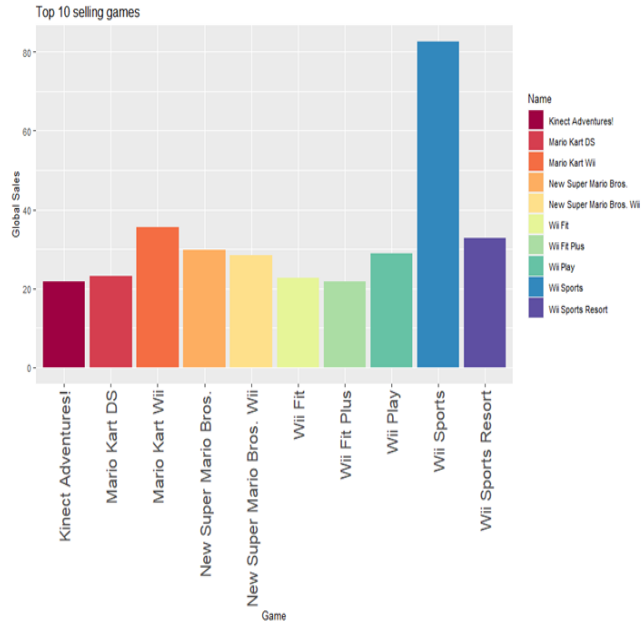
DESCRIPTIVE STATISTICS:

```
> summary(vgsales)
                                          Name         Platform    Year_of_Release          Genre
 LEGO Star Wars II: The Original Trilogy  :   8   PS2     :1140   Min.   :1985    Action      :1630
 Madden NFL 07                            :   8   X360    : 858   1st Qu.:2004    Sports      : 943
 Need for Speed: Most Wanted              :   8   PS3     : 769   Median :2007    Shooter     : 864
 Harry Potter and the Order of the Phoenix:   7   PC      : 651   Mean   :2007    Role-Playing: 712
 Madden NFL 08                            :   7   XB      : 565   3rd Qu.:2011    Racing      : 581
 Need for Speed Carbon                    :   7   Wii     : 479   Max.   :2016    Platform    : 403
 (Other)                                  :6780   (Other):2363                    (Other)     :1692
                  Publisher       NA_Sales          EU_Sales          JP_Sales         Other_Sales
 Electronic Arts            : 944   Min.   : 0.0000   Min.   : 0.0000   Min.   :0.00000   Min.   : 0.00000
 Ubisoft                    : 496   1st Qu.: 0.0600   1st Qu.: 0.0200   1st Qu.:0.00000   1st Qu.: 0.01000
 Activision                 : 492   Median : 0.1500   Median : 0.0600   Median :0.00000   Median : 0.02000
 Sony Computer Entertainment: 316   Mean   : 0.3945   Mean   : 0.2361   Mean   :0.06416   Mean   : 0.08268
 THQ                        : 307   3rd Qu.: 0.3900   3rd Qu.: 0.2100   3rd Qu.:0.01000   3rd Qu.: 0.07000
 Nintendo                   : 291   Max.   :41.3600   Max.   :28.9600   Max.   :6.50000   Max.   :10.57000
 (Other)                    :3979
  Global_Sales      Critic_Score    Critic_Count      User_Score     User_Count                  Developer
 Min.   : 0.0100   Min.   :13.00   Min.   :  3.00   7.8     : 294   Min.   :    4.0   EA Canada       : 149
 1st Qu.: 0.1100   1st Qu.:62.00   1st Qu.: 14.00   8       : 259   1st Qu.:   11.0   EA Sports       : 142
 Median : 0.2900   Median :72.00   Median : 25.00   8.2     : 258   Median :   27.0   Capcom          : 126
 Mean   : 0.7776   Mean   :70.27   Mean   : 28.93   8.5     : 238   Mean   :  174.7   Ubisoft         : 103
 3rd Qu.: 0.7500   3rd Qu.:80.00   3rd Qu.: 39.00   7.9     : 235   3rd Qu.:   89.0   Konami          :  95
 Max.   :82.5300   Max.   :98.00   Max.   :113.00   7.5     : 234   Max.   :10665.0   Ubisoft Montreal:  87
                                                    (Other):5307                     (Other)         :6123
     Rating
 T       :2377
 E       :2082
 M       :1433
 E10+    : 930
 AO      :   1
 K-A     :   1
 (Other):   1
```

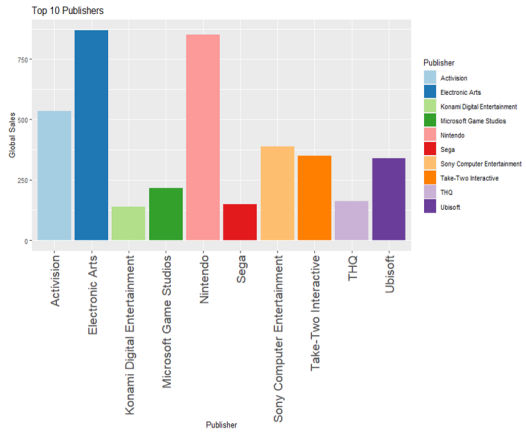
```
> skim(vgsales)
── Data Summary ────────────────────────
                           Values
Name                       vgsales
Number of rows             16719
Number of columns          16
_________________________
Column type frequency:
  character                8
  numeric                  8
_________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────────────────────
  skim_variable   n_missing complete_rate min max empty n_unique whitespace
1 Name                    0             1   0 132     2    11563          0
2 Platform                0             1   2   4     0       31          0
3 Year_of_Release         0             1   3   4     0       40          0
4 Genre                   0             1   0  12     2       13          0
5 Publisher               0             1   3  38     0      582          0
6 User_Score              0             1   0   3  6704       97          0
7 Developer               0             1   0  80  6623     1697          0
8 Rating                  0             1   0   4  6769        9          0

── Variable type: numeric ────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate     mean      sd    p0   p25   p50   p75    p100 hist
1 NA_Sales              0             1    0.263    0.814  0     0     0.08  0.24   41.4 ▇▁▁▁▁
2 EU_Sales              0             1    0.145    0.503  0     0     0.02  0.11   29.0 ▇▁▁▁▁
3 JP_Sales              0             1    0.0776   0.309  0     0     0     0.04   10.2 ▇▁▁▁▁
4 Other_Sales           0             1    0.0473   0.187  0     0     0.01  0.03   10.6 ▇▁▁▁▁
5 Global_Sales          0             1    0.534    1.55   0.01  0.06  0.17  0.47   82.5 ▇▁▁▁▁
6 Critic_Score       8582             0.487 69.0     13.9  13    60    71    79     98  ▁▁▃▅▇
7 Critic_Count       8582             0.487 26.4     19.0   3    12    21    36    113  ▇▃▂▁▁
8 User_Count         9129             0.454 162.     561.   4    10    24    81  10665  ▇▁▁▁▁
>
```

DATA VISUALIZATIONS:

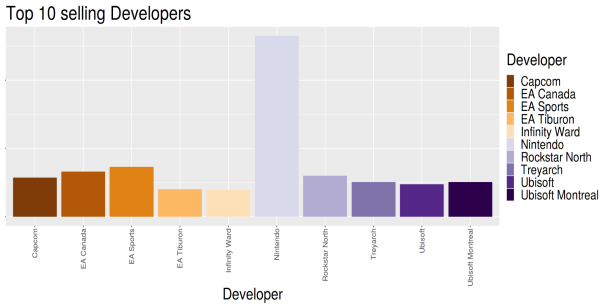
A)gamesales %>% select(Name,Global_Sales) %>% arrange(desc(Global_Sales))%>% head(10)%>%
ggplot(aes(x=Name,y=Global_Sales,fill=Name))+geom_bar(stat="identity")+
theme(text = element_text(size=10),legend.position="right",axis.text.x=element_text(angle = 90,vjust = 0.5,hjust = 1,size=15))+labs(x="Developer",y="Total Sales",title="Top 10 selling Developers")+labs(x="Game",y="Global Sales",title="Top 10 selling games")+scale_fill_brewer(palette="Spectral")



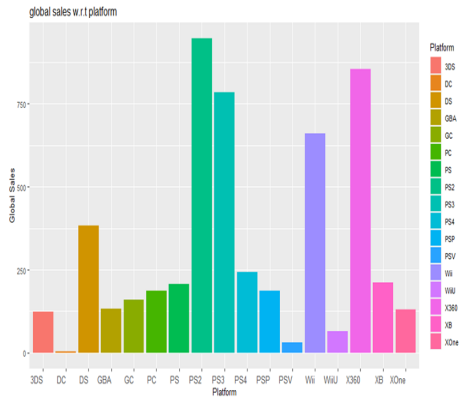
b)gamesales%>% select(Publisher,Global_Sales)%>%group_by(Publisher)%>%
summarise(Total_sales=sum(Global_Sales))%>%arrange(desc(Total_sales))%>% head(10)%>%
ggplot(aes(x=Publisher,y=Total_sales,fill=Publisher))+geom_bar(stat="identity")+
theme(text = element_text(size=10),legend.position="right",axis.text.x=element_text(angle = 90,vjust = 0.5,hjust = 1,size=15))+labs(x="Publisher",y="Global Sales",title="Top 10 Publishers")+scale_fill_brewer(palette="Paired")



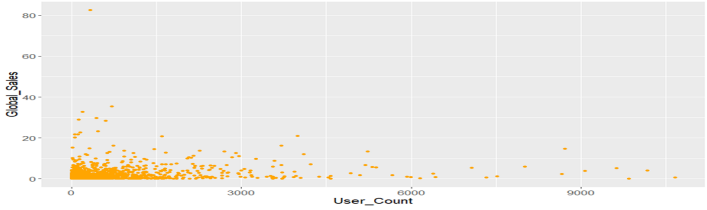
c)gamesales %>% select(Name,Global_Sales) %>% arrange(desc(Global_Sales))%>% head(10)%>%
ggplot(aes(x=Name,y=Global_Sales,fill=Name))+geom_bar(stat="identity")+
theme(text = element_text(size=10),legend.position="right",axis.text.x=element_text(angle = 90,vjust = 0.5,hjust = 1,size=10))+labs(x="Developer",y="Total Sales",title="Top 10 selling Developers")+labs(x="Game",y="Global Sales",title="Top 10 selling games")+scale_fill_brewer(palette="Spectral")



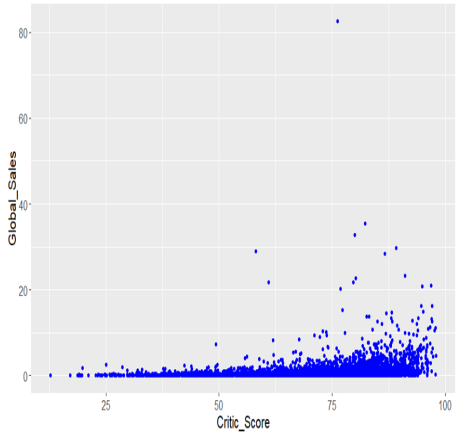
d)salesbyplatform <- ggplot(gamesales, aes(Platform,Global_Sales,fill =Platform)) +geom_bar(stat = "identity") +
theme(text = element_text(size=10),legend.position="right",axis.text.x=element_text(vjust = 0.5,hjust = 1,size=10))+labs(x="Platform",y="Global Sales",title="global sales w.r.t platform")
salesbyplatform



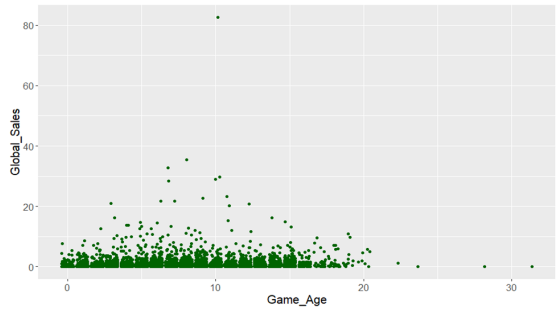
e)Usercount <- ggplot(gamesales, aes(User_Count,Global_Sales))+geom_jitter(color = "orange") + theme(text = element_text(size = 15))
Usercount



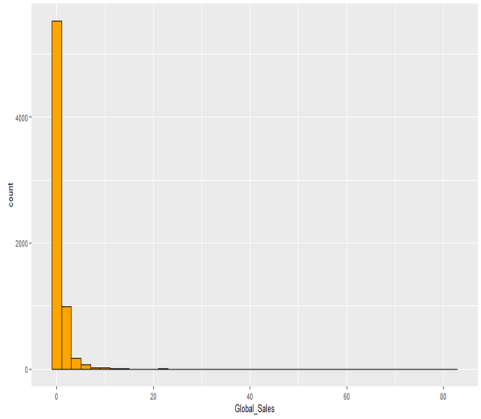
f)Criticscore <- ggplot(gamesales, aes(Critic_Score,Global_Sales))+geom_jitter(color = "blue") + theme(text = element_text(size = 15))
Criticscore



g)Globalsales <- ggplot(gamesales, aes(Game_Age,Global_Sales)) + geom_jitter(color = "darkgreen") + theme(text = element_text(size = 15))
Globalsales

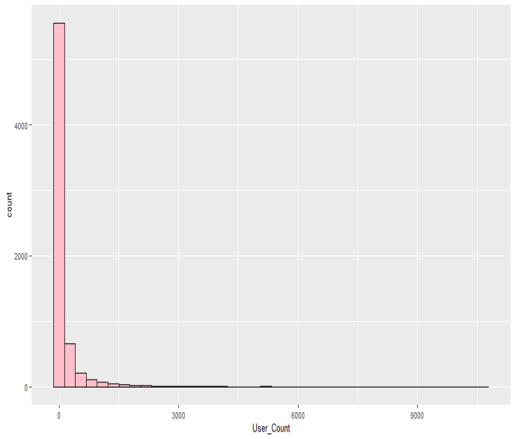


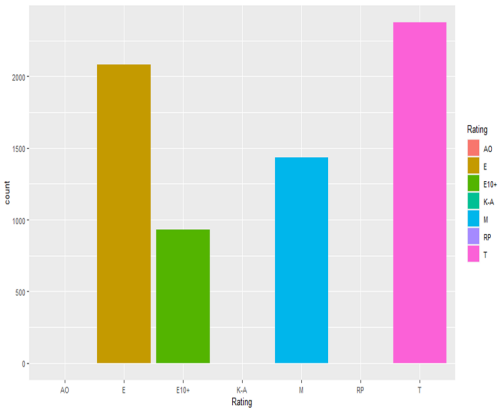
H)Usercount <- ggplot(gamesales, aes(User_Count,Global_Sales))+geom_jitter(color = "orange") + theme(text = element_text(size = 15))
Usercount

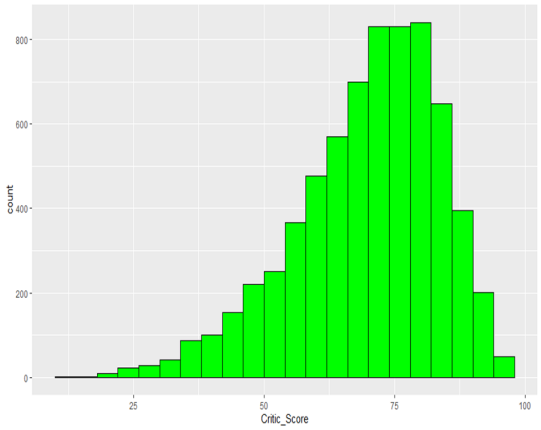


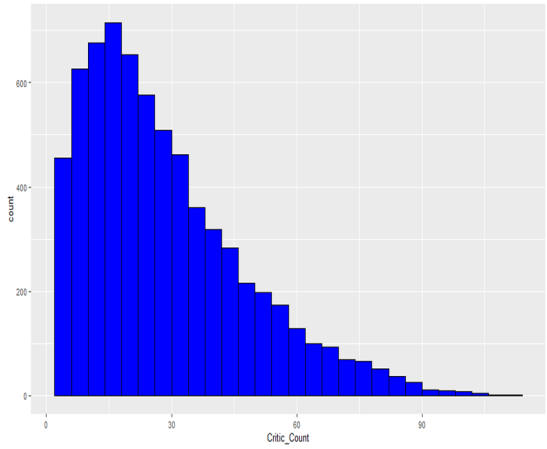
I)GlobalSalesplot <- ggplot(gamesales, aes(Global_Sales)) + geom_histogram(binwidth = 2, color = "black",fill = "orange") + theme(text = element_text(size=10))
GlobalSalesplot

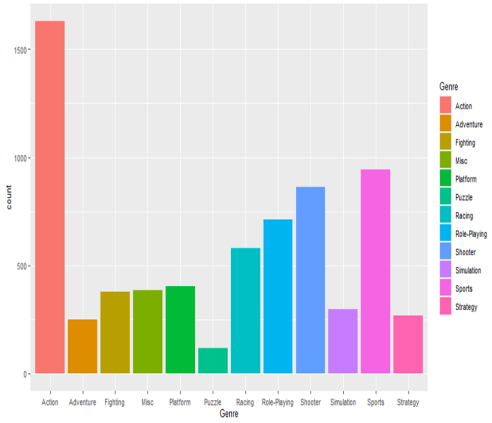
G)UserCountplot<-ggplot(gamesales, aes(User_Count)) + geom_histogram(color = "black",fill = "pink",bins = 40) + theme(text = element_text(size=10))
UserCountplot

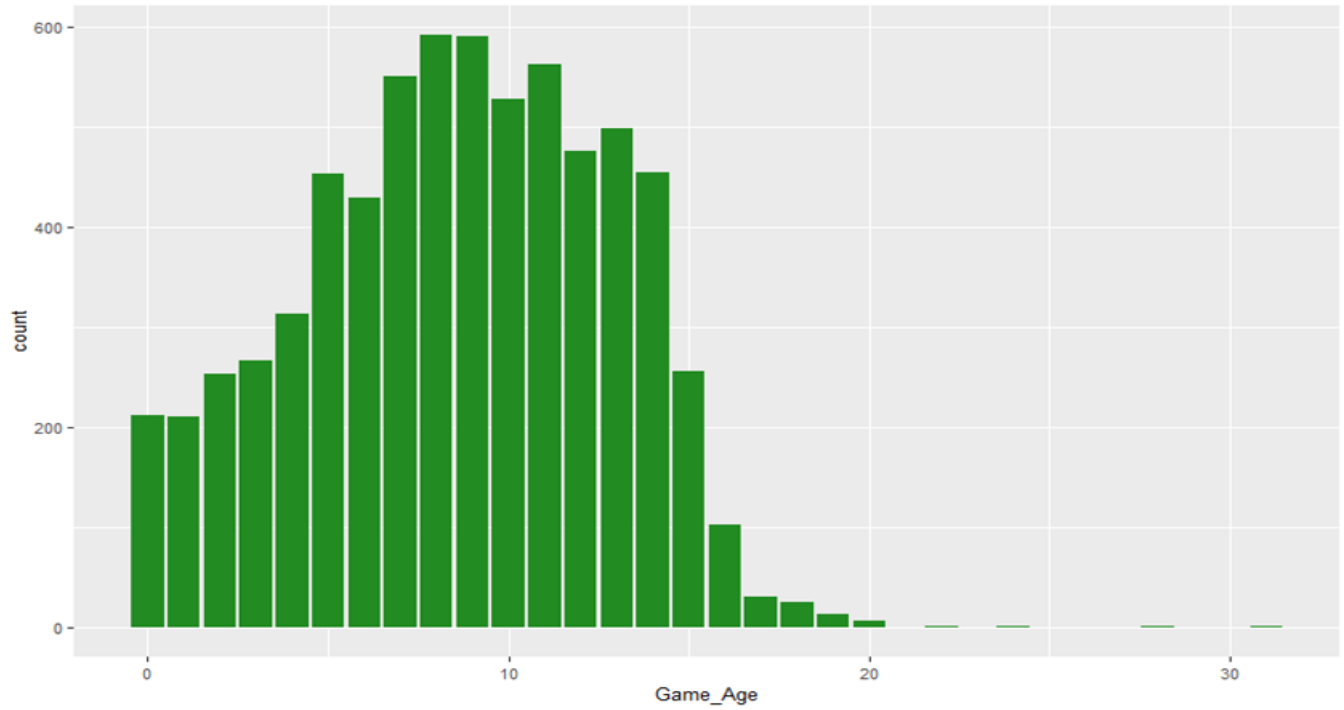


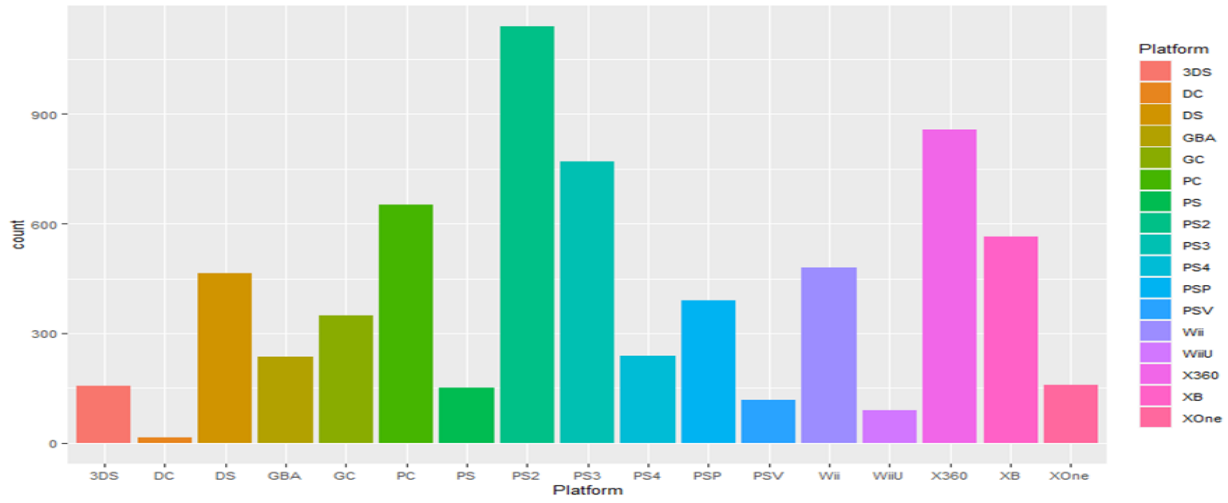



count
800
600
400
200
0
25
50
75
100
Critic_Score


count
600
400
200
0
0
30
60
90
Critic_Count


count
1500
1000
500
0
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy
Genre
Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy





DATA CLEANING:

	na_count
Name	2
Platform	0
Year_of_Release	269
Genre	2
Publisher	54
NA_Sales	0
EU_Sales	0
JP_Sales	0
Other_Sales	0
Global_Sales	0
Critic_Score	8582
Critic_Count	8582
User_Score	6704
User_Count	9129
Developer	6623
Rating	6769

	na_count
Name	0
Platform	0
Year_of_Release	0
Genre	0
Publisher	0
NA_Sales	0
EU_Sales	0
JP_Sales	0
Other_Sales	0
Global_Sales	0
Critic_Score	0
Critic_Count	0
User_Score	0
User_Count	0
Developer	0
Rating	0

Feature Selection:

This section is used to explore the first question proposed in the objective: what would be key factors to determine the sales of a game? All columns will be evaluated to find out one or several the ultimate determinant(s) of game sales. The conclusion gives publishers and developers actionable strategies to achieve higher sales when developing new games.

LINEAR REGRESSION:

This algorithm establishes a relation between two variable one variable is predicted variable and another one is result variable whose value is derived from the predictive variable.

```
lm1<-lm(formula = Global_Sales ~ User_Score, data = Videogame)
summary(lm1)

lm2<-lm(formula = Global_Sales ~ Critic_Score, data = Videogame)
summary(lm2)

lm3<-lm(formula = Global_Sales ~ User_Score + Critic_Score, data = Videogame)
summary(lm3)

lm4<-lm(formula = Global_Sales ~ Rating, data = Videogame)
summary(lm4)

lm5<-lm(formula = Global_Sales ~ User_Score + Critic_Score, data = Videogame_ms)
summary(lm5)

lm6<-lm(formula = Global_Sales ~ Critic_Score, data = Videogame_ms)
summary(lm6)
```

In the linear regression method, significant predictors are: **Genre** (Adventure, Puzzle, Role-Playing, Strategy), **Publishers** (Electronic Arts, Nintendo), **Critic Score**, **User Score**, and **Platform** (DS, PS, PS2, PS3, PS4, Wii, X360). Through further selection, **User Score** and **Critic Score** are the ultimate determinants.

Logistic regression is **an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring**.

```
> logit <- glm(formula = aboveave ~ Genre + Publisher + Critic_Score + User_Score +
+        Rating + year2 + Platform, data =Videogame_aboveave_train  )
```

```
glm.pred
        0         1
0.8933144 0.1066856
```

```
0.6799431
```

In the logistic regression method, the percentage of video games will be successful is 10.67%, while the accuracy is only 67.99%. The result needs further validation.

KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

When K=1, the accuracy of the model and the successful rate should be:

```
knn.pred.avgsales
        0         1
0.8008535 0.1991465
```

When K=5, the accuracy of the model and the successful rate should be:

```
knn.pred.avgsales5
        0         1
0.8421053 0.1578947
```

When K=10, the accuracy of the model and the successful rate should be:

```
knn.pred.avgsales10
         0          1
0.8556188 0.1443812
```

KNN CONCLUSION:

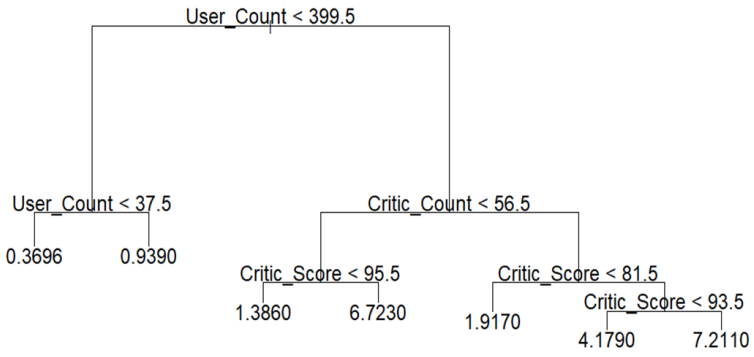
In the KNN method, when K=1, the successful rate is 19.91% and the accuracy is 88.62%. When K=5, the successful rate is 15.79%, the accuracy is 89.47%. When K=10, the successful rate is 14.44%, and the accuracy is 89.12%.

Comparing the two models, they all predict that the successful rate should be between 10% and 20% while KNN have a higher accuracy rate. The percentage result demonstrated by KNN implies a typical business principle: **Pareto principle**, or **80/20 principle** which indicates that roughly 80% effects come from 20% of contents/contributors. Although games might be thought to have long-tail markets, the result suggests that it might be another mass market and investors could use this principle in the game industry to support investment decisions.

DECISION TREE:

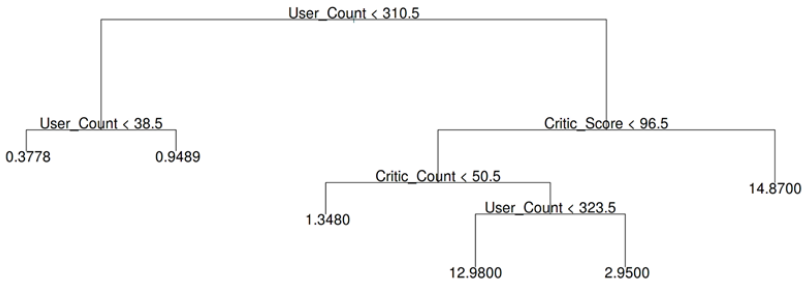
Decision trees are an approach used in supervised machine learning, a technique which uses labelled input and output datasets to train models. The approach is used mainly to solve classification problems, which is the use of a model to categorise or classify an object.

```
Regression tree:
tree(formula = Global_Sales ~ Critic_Score + Critic_Count + User_Count,
    data = dt.df, subset = train.index)
Number of terminal nodes:  7
Residual mean deviance:  2.505 = 11950 / 4770
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-6.3130 -0.3496 -0.2190  0.0000  0.1004 31.8300
```



PRUNING:

Pruning reduces the size of decision trees by removing parts of the tree that do not provide power to classify instances. Decision trees are the most susceptible out of all the machine learning algorithms to overfitting and effective pruning can reduce this likelihood




```
> t8 = mean((yhat2-vgs.test)^2)
> t6 = mean((yhat-vgs.test)^2)
> dt.mae <- mean(abs(yhat-vgs.test))
> cat("full tree mse = ", t8,"\n")
full tree mse =  4.857161
> cat("6 leaf tree mse = ", t6)
6 leaf tree mse =  4.936105
```

RESULTS:

In the logistic regression method, the percentage of video games will be successful is 10.67%, while the accuracy is only 67.99%. The result needs further validation.

Following the application of all models to the data, decision tree models, linear regression models, and KNN models are the best fixes for the data with understating the mean square error and mean absolute error as

Models	MAE	MSE
K - Nearest Neighbours	0.77	1.46
Linear Regression	0.79	1.66
Decision Tree	0.72	3.58

CONCLUSION:

The examination of worldwide sales of video games is something that interests you as a team. 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 allows knowing the negative trends that appear in the prediction. Finally we conclude that prediction of sales on video games has been 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, Logistic Regression, Decision tree, KNN). Among all these algorithms KNN and decision Trees gave us the best accurate result with minimum error rate.