VideoGames

Eduardo Guiliani

6/30/2021

Overview

The Video Games project is part of the HarvardX: PH125.9x Data Science: Capstone course. The aim of the project is to develop and train recommendation machine learning algorithms to predict North American video game sales from a set of video games spanning the years 1985 to 2016 in the data set. The Residual Mean Square Error (RMSE) will be used to evaluate the accuracy of the algorithms. This report will present methods used in exploratory data analysis and visualization, results for the RMSE model and a conclusion based on results of the model. The objective is to demonstrate knowledge acquired in the 9 courses of the professional certificate.

The Video Games Sales data set was downloaded from Kaggle (https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings) in excel csv format. The code separated the data into two subsets for training (train_set) and testing (test_set). The algorithms used to train and test the model were Naive Bayes, Generalised Linear Model (GLM), K-nearest neighbor (Knn), Random Forest, and Classification Trees, The data set was loaded as a data frame, its dimensions and a sample of its features are provided below:

Methods

Data Set Video games

```
## The video games data set has 16719 rows and 16 columns.
```

```
## There are 11563 different video games and 582 different publishers in the video games data se
t.
```

```
## All sales are in millions of units sold.
```

To prepare the data set for further analysis it's rows were evaluated in order to identify NA and Blank values in it's 16 columns. From the following table we can observe that Critic Score, Critic Count, and User Count have a considerable amount of NA values, and will reduce the data set significantly.

```
##            Name        Platform Year_of_Release           Genre       Publisher 
##               0               0               0               0               0 
##        NA_Sales        EU_Sales        JP_Sales     Other_Sales    Global_Sales 
##               0               0               0               0               0 
##    Critic_Score    Critic_Count      User_Score      User_Count       Developer 
##            8582            8582               0            9129               0 
##          Rating 
##               0
```

```
## After removing NA's the video games data set now has 7017 rows.
```

Upon further investigation we noticed that the “Year of release” column included some NA values that were not identified in the first removal and were subsequently removed. Rows with blank values in the “Developer” and “Ratings” column were also removed from the data set.

```
##            Name        Platform Year_of_Release           Genre       Publisher 
##               0               0               0               0               0 
##        NA_Sales        EU_Sales        JP_Sales     Other_Sales    Global_Sales 
##               0               0               0               0               0 
##    Critic_Score    Critic_Count      User_Score      User_Count       Developer 
##               0               0               0               0               4 
##          Rating 
##              68
```

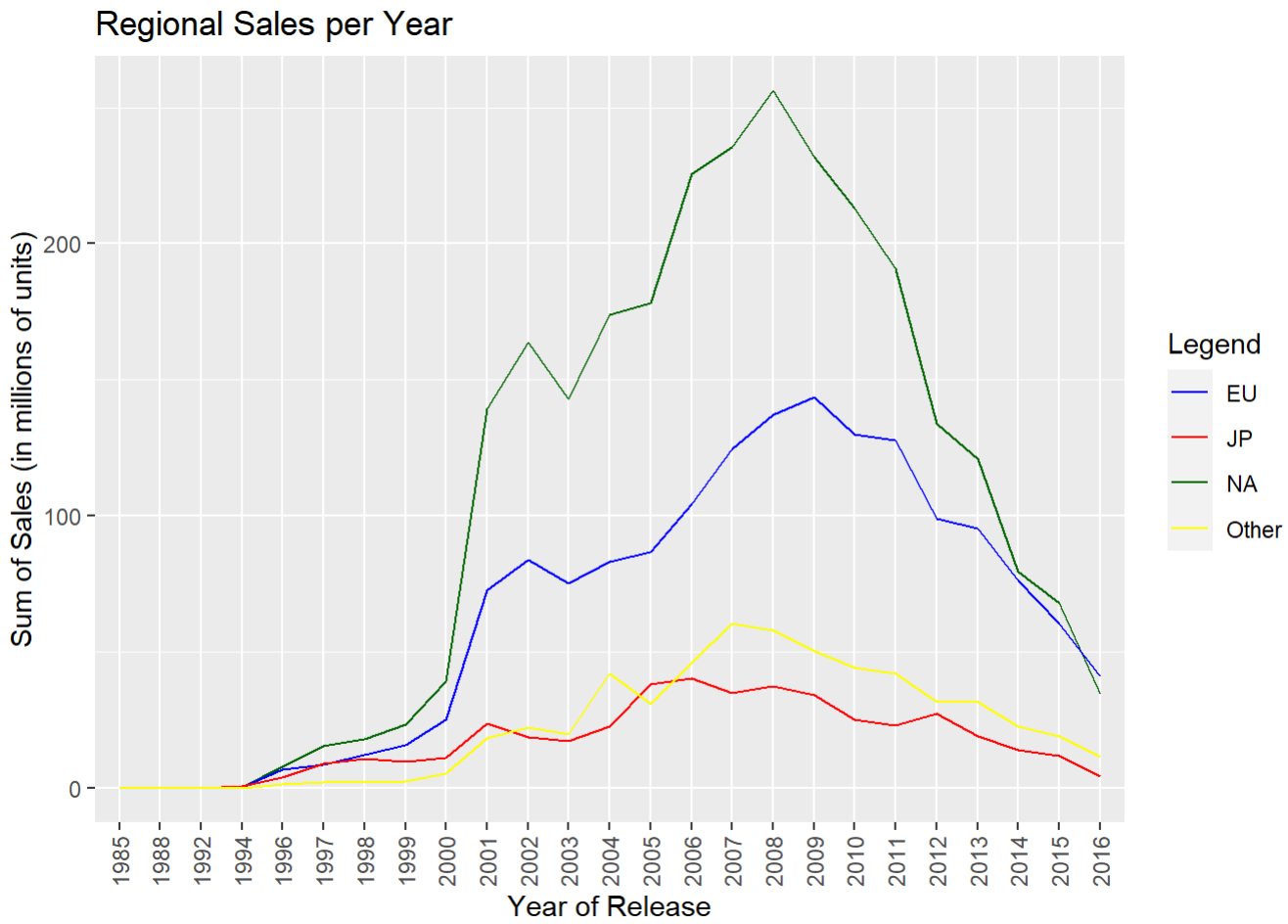
```
## After removing all NA's and blanks, the video games data set now has 6826 rows.
```

The following summary presents a description of the data set in it’s final form previous to exploratory data analysis :

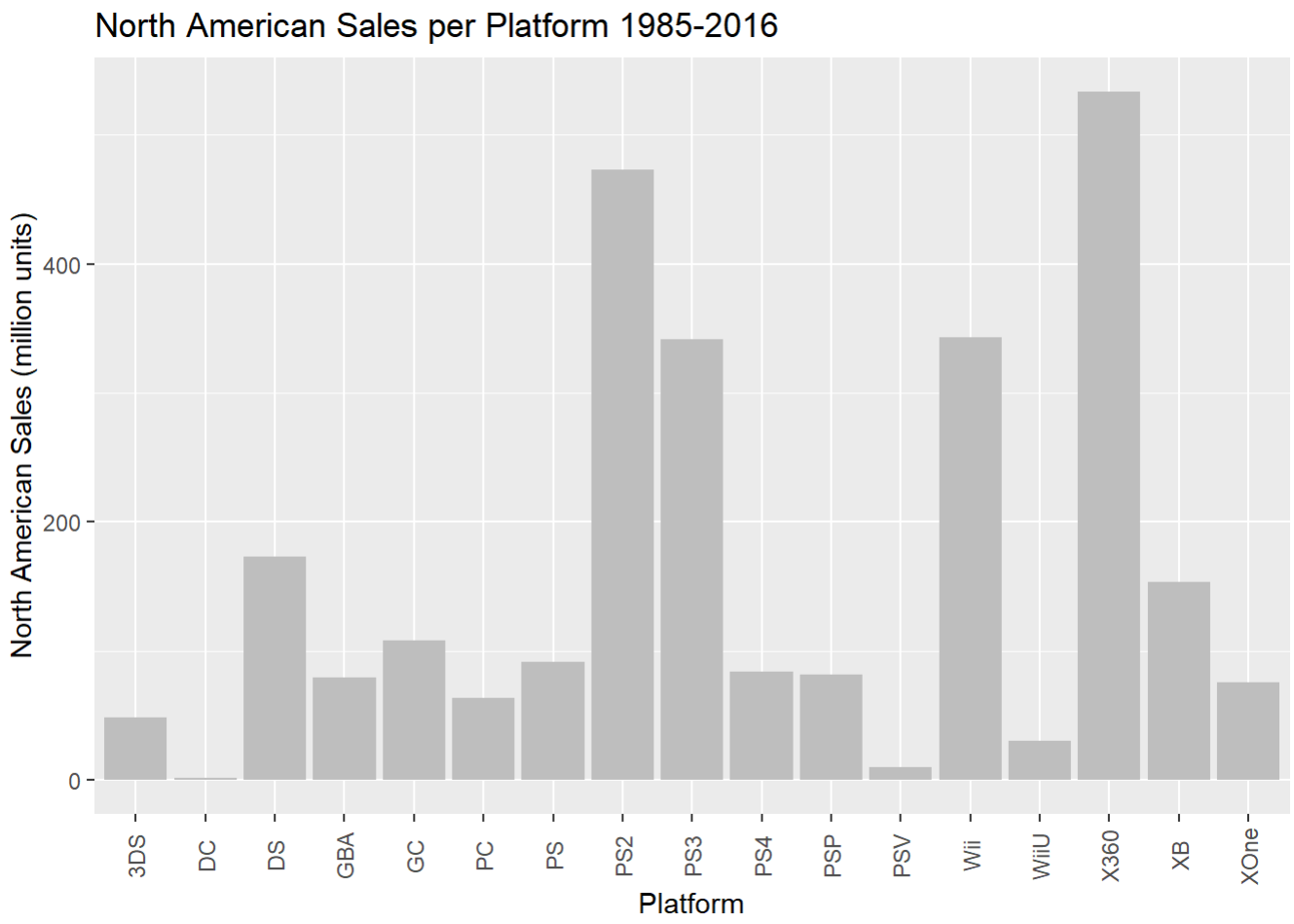
```
##                                       Name         Platform
##  LEGO Star Wars II: The Original Trilogy  :   8   PS2    :1140
##  Madden NFL 07                            :   8   X360   : 858
##  Need for Speed: Most Wanted              :   8   PS3    : 769
##  Harry Potter and the Order of the Phoenix:   7   PC     : 652
##  Madden NFL 08                            :   7   XB     : 565
##  Need for Speed Carbon                    :   7   Wii    : 479
##  (Other)                                  :6781   (Other):2363
##  Year_of_Release          Genre                          Publisher
##  2008   : 592    Action      :1630   Electronic Arts            : 944
##  2007   : 590    Sports      : 943   Ubisoft                    : 496
##  2005   : 562    Shooter     : 864   Activision                 : 492
##  2009   : 550    Role-Playing: 712   Sony Computer Entertainment: 316
##  2006   : 528    Racing      : 581   THQ                        : 307
##  2003   : 498    Platform    : 403   Nintendo                   : 291
##  (Other):3506    (Other)     :1693   (Other)                    :3980
##     NA_Sales          EU_Sales          JP_Sales        Other_Sales
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :0.00000   Min.   : 0.00000
##  1st Qu.: 0.0600   1st Qu.: 0.0200   1st Qu.:0.00000   1st Qu.: 0.01000
##  Median : 0.1500   Median : 0.0600   Median :0.00000   Median : 0.02000
##  Mean   : 0.3944   Mean   : 0.2361   Mean   :0.06415   Mean   : 0.08267
##  3rd Qu.: 0.3900   3rd Qu.: 0.2100   3rd Qu.:0.01000   3rd Qu.: 0.07000
##  Max.   :41.3600   Max.   :28.9600   Max.   :6.50000   Max.   :10.57000
##
##   Global_Sales      Critic_Score    Critic_Count      User_Score
##  Min.   : 0.0100   Min.   :13.00   Min.   :  3.00   Min.   : 5.00
##  1st Qu.: 0.1100   1st Qu.:62.00   1st Qu.: 14.00   1st Qu.:64.00
##  Median : 0.2900   Median :72.00   Median : 25.00   Median :74.00
##  Mean   : 0.7775   Mean   :70.27   Mean   : 28.93   Mean   :70.85
##  3rd Qu.: 0.7500   3rd Qu.:80.00   3rd Qu.: 39.00   3rd Qu.:81.00
##  Max.   :82.5300   Max.   :98.00   Max.   :113.00   Max.   :95.00
##
##    User_Count                Developer        Rating
##  Min.   :    4.0   EA Canada       : 149   T      :2378
##  1st Qu.:   11.0   EA Sports       : 142   E      :2082
##  Median :   27.0   Capcom          : 126   M      :1433
##  Mean   :  174.7   Ubisoft         : 103   E10+   : 930
##  3rd Qu.:   89.0   Konami          :  95   AO     :   1
##  Max.   :10665.0   Ubisoft Montreal:  87   K-A    :   1
##                    (Other)         :6124   (Other):   1
```

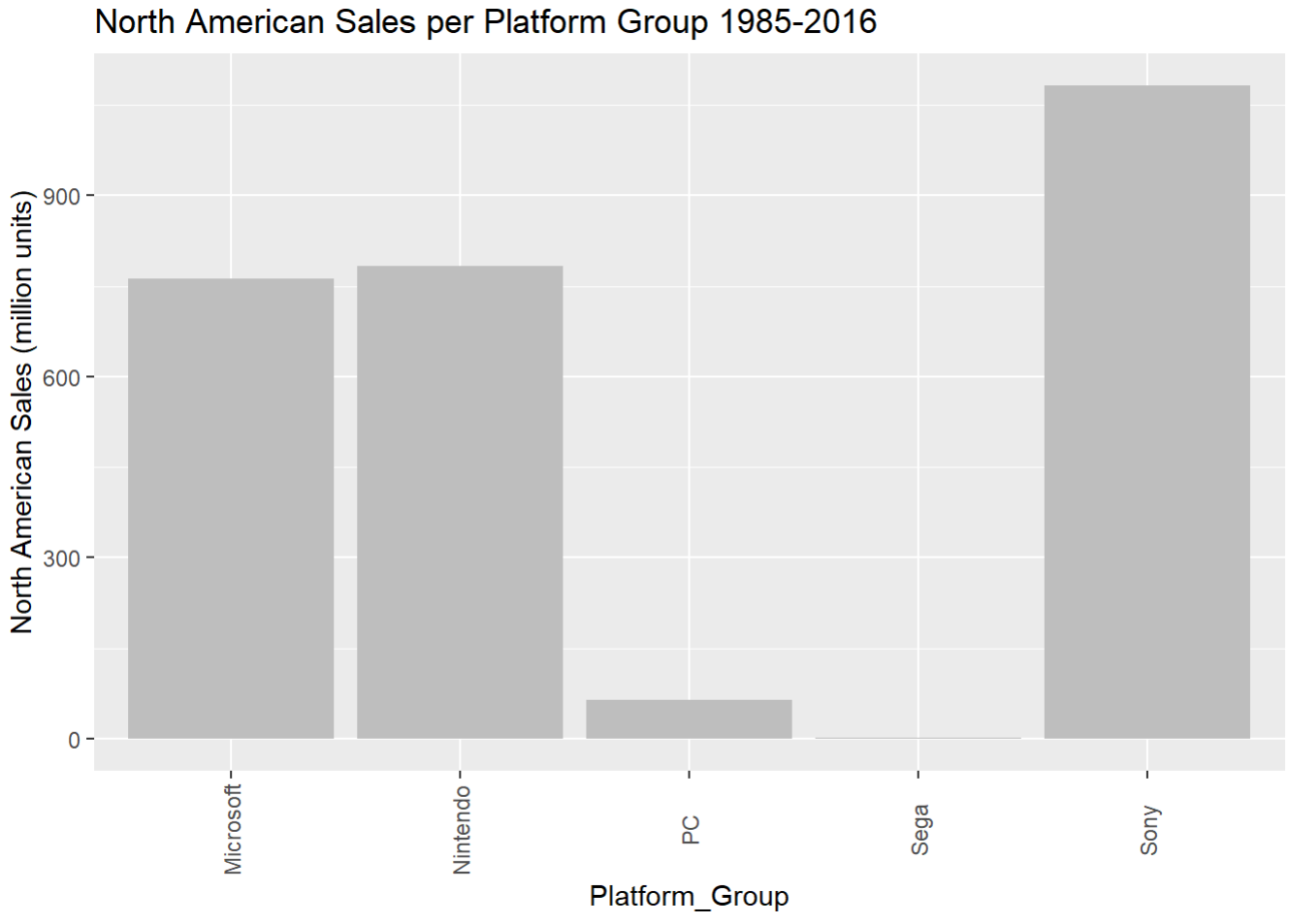
Exploratory Data Analysis

In our initial exploratory analysis we observe that North America is by far the most important market for the period observed (1985-2016), although there is some convergence in sales by the EU and North America in the latter years. The importance of the North American market to global sales will shape our approach in our further exploration of the data.



Subsequently, by aggregating North American sales by platforms we can observe that the X360 is the most used platform in the North American Market. However, by observing the x axis we notice that some of the platforms are in fact different versions of the consoles used to play video games. Thus, we aggregated platforms by platform groups to gain a better understanding of the popularity of the different consoles in the North American market. Sony is the clear leader in this respect, followed by Nintendo and then Microsoft.


North American Sales per Platform 1985-2016
North American Sales (million units)
0
200
400
3DS
DC
DS
GBA
GC
PC
PS
PS2
PS3
PS4
PSP
PSV
Wii
WiiU
X360
XB
XOne
Platform


North American Sales per Platform Group 1985-2016
North American Sales (million units)
0
300
600
900
Microsoft
Nintendo
PC
Sega
Sony
Platform_Group

In the following tables we can observe that the developer with most sales is Nintendo. However, if we look at publisher sales we see that the largest video game publisher is Electronic Arts (EA). In terms of releases, EA is the most active video game publisher well ahead of Ubisoft.

```
## # A tibble: 5 x 2
##   Developer  NA_Sales
##   <fct>         <dbl>
## 1 Nintendo      231.
## 2 EA Sports      83.9
## 3 EA Tiburon     65.7
## 4 EA Canada      60.9
## 5 Treyarch       56.2
```

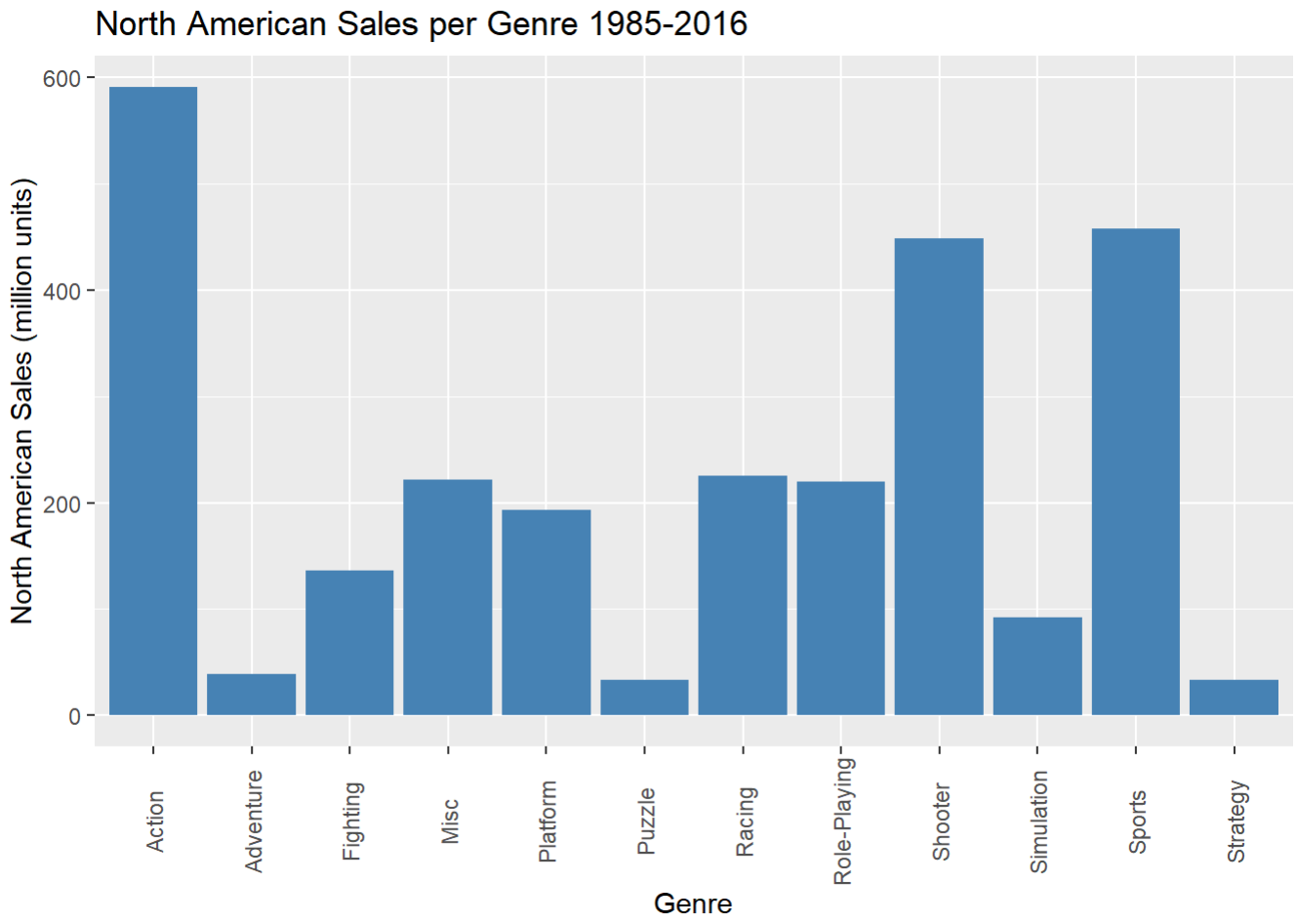
```
## # A tibble: 5 x 2
##   Publisher                    NA_Sales
##   <fct>                           <dbl>
## 1 Electronic Arts                  465.
## 2 Nintendo                         371.
## 3 Activision                       307.
## 4 Take-Two Interactive             188.
## 5 Sony Computer Entertainment      177.
```

```
## # A tibble: 5 x 2
##   Publisher                    Releases
##   <fct>                           <int>
## 1 Electronic Arts                   944
## 2 Ubisoft                           496
## 3 Activision                        492
## 4 Sony Computer Entertainment       316
## 5 THQ                               307
```

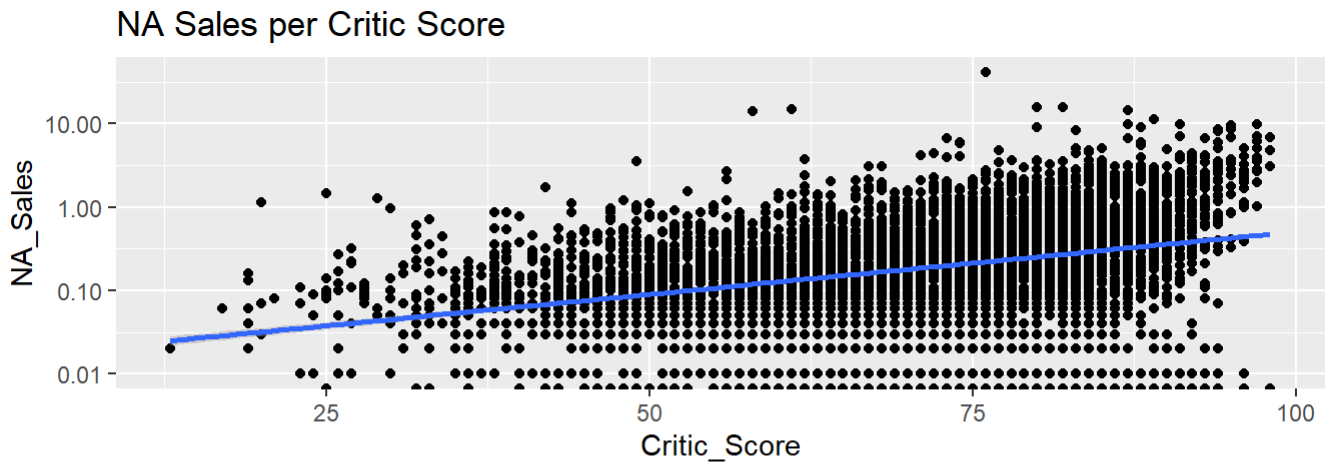
In terms of the most popular game in North America, Wii Sports has a significant lead in this regard.

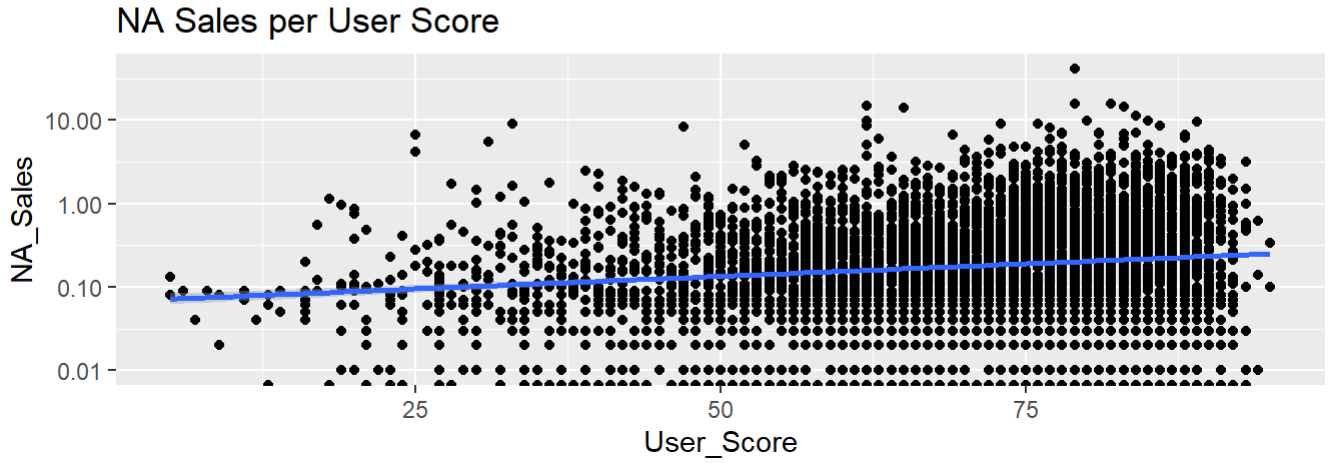
```
## # A tibble: 5 x 2
##   Name                     NA_Sales
##   <fct>                       <dbl>
## 1 Wii Sports                   41.4
## 2 Grand Theft Auto V           23.8
## 3 Call of Duty: Black Ops      17.0
## 4 Mario Kart Wii               15.7
## 5 Wii Sports Resort            15.6
```

Furthermore, when grouping by genre we see that the action genre is the most important to North American sales. An action game is a video game genre that emphasizes physical challenges, including hand–eye coordination and reaction-time. For example, Enemy attacks and obstacles deplete the player character’s health and lives, and the player receives a game over when they run out of lives.

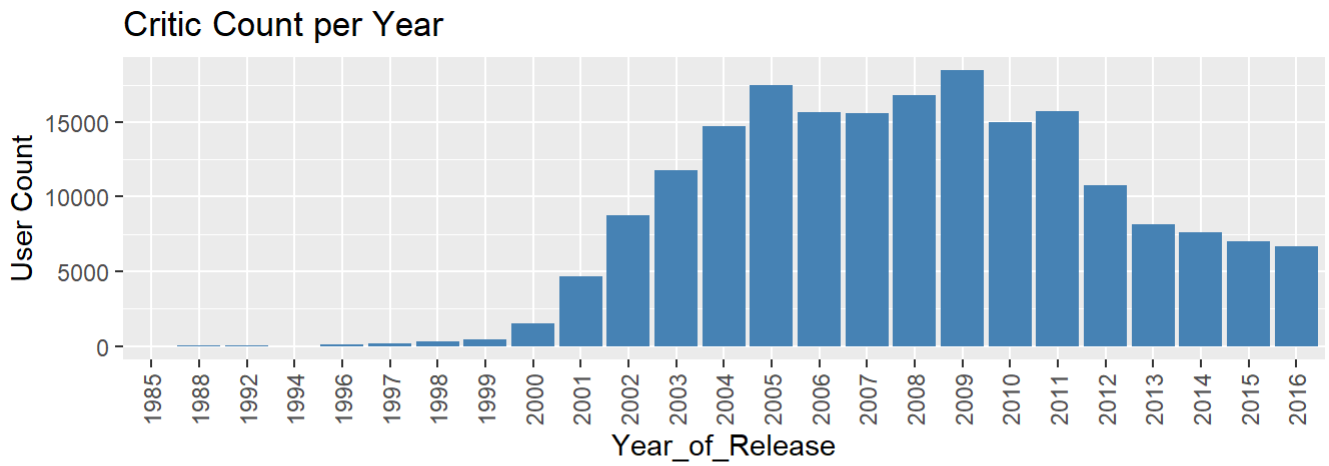


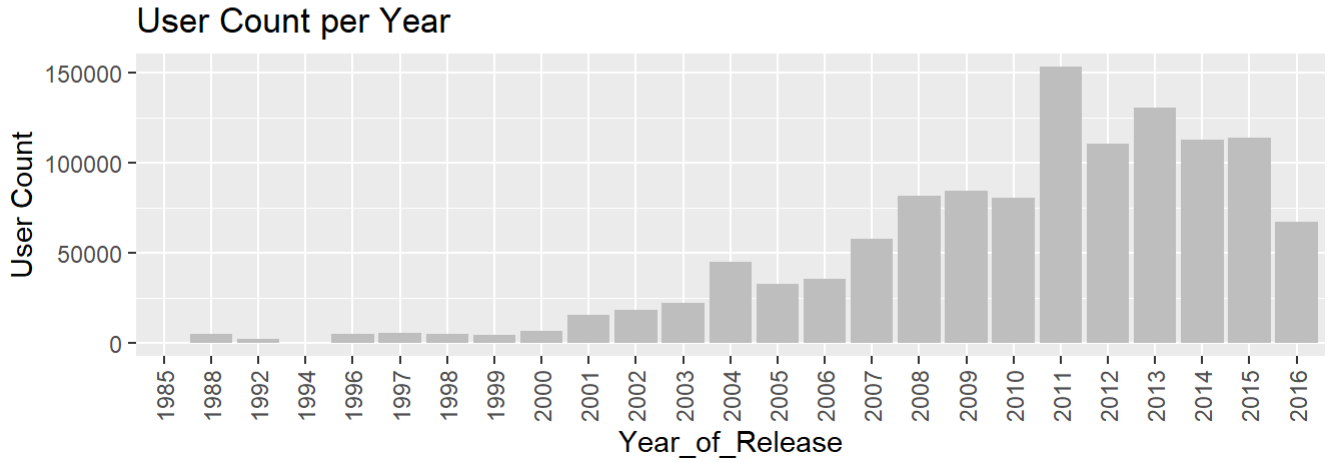
The data set also provides a critic score and user score as features that rate the quality of the game as rated by critics or users. The following graphs display the fact that higher critic and user scores translate to better North American sales, with a slight preference to critic score considering the steeper trend line of the two graphs.



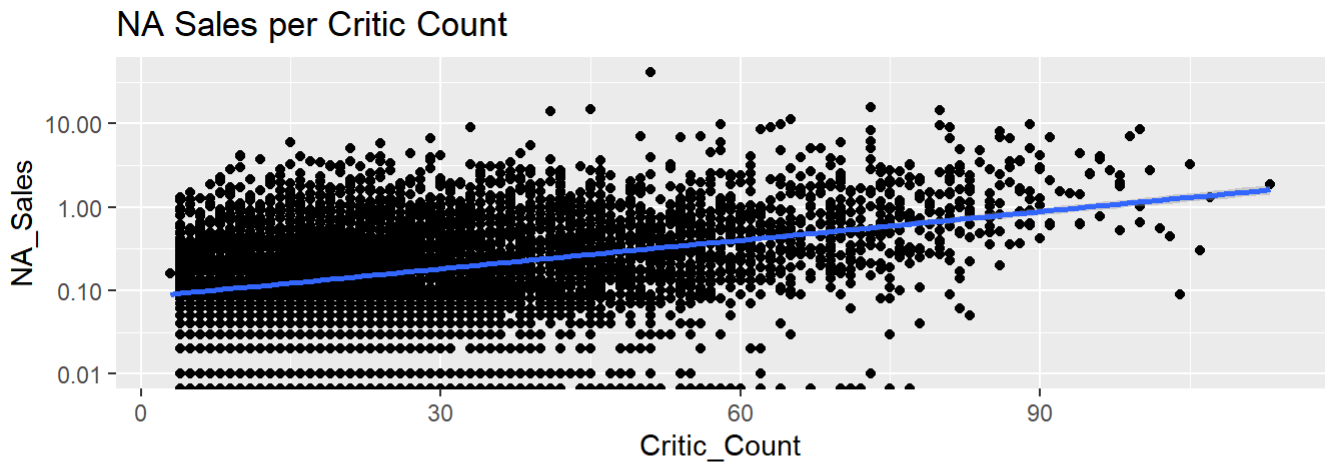


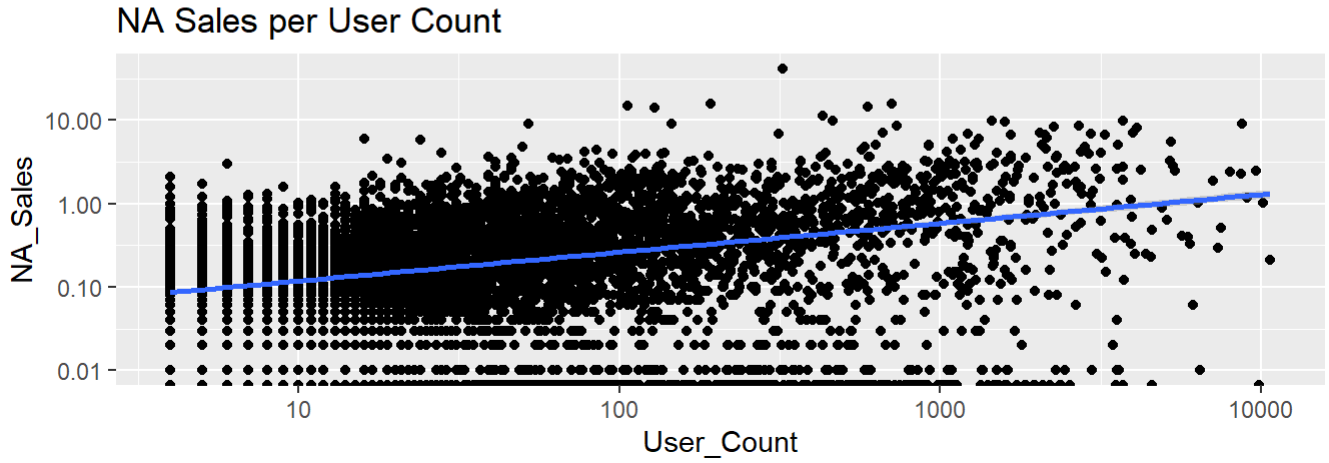
Furthermore, by analyzing the amount of critic and user reviews across the time series we observe that they trend upwards peaking in 2009 and 2011 respectively, and then trend downwards.





In terms of North American sales we observe that the critics and user counts positively affect sales numbers.





Model

The data set was partitioned into train_set and test_set in order to train the model and then test the results. The split was .5 due to the fact that the data set was greatly reduced when removing NA and blank values. The seed was set to 527 in order to get the same results when running the algorithms. The features selected to train the model are critic score, critic count, user score, user count, platform, and genre. Platform and Genre are categorical features, thus the algorithms selected (GLM, KNN, random forest, and classification trees) are able to handle categorical data.

The RMSE Model

The evaluation of the predictions were to be executed via an RMSE loss function, defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{u,i}(y_{u,i} - \hat{y}_{u,i})^2}$$

with $\hat{y}_{u,i}$ and $y_{u,i}$ being the predicted and actual ratings, and *N*, the number of possible combinations between user *u* and movie *i*.This function evaluates the square root of the mean of the differences between true and predicted ratings.

Algorithms

Naive Bayes: is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points.

General Linear Model:used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution.

KNN:is a non-parametric classification method used for classification and regression. In both cases, the input consists of the k closest training examples in data set.

Random Forest:is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.

Classification Trees:is an algorithm used to create a model that predicts the value of a target variable based on several input variables.

Results

Naive Bayes

```
mu <- mean(train_set$NA_Sales)
mu
```

```
## [1] 0.3991176
```

```
NB_Model <- RMSE(test_set$NA_Sales, mu)
```

```
## The resulting RMSE is 0.8071075
```

Linear Regression Model

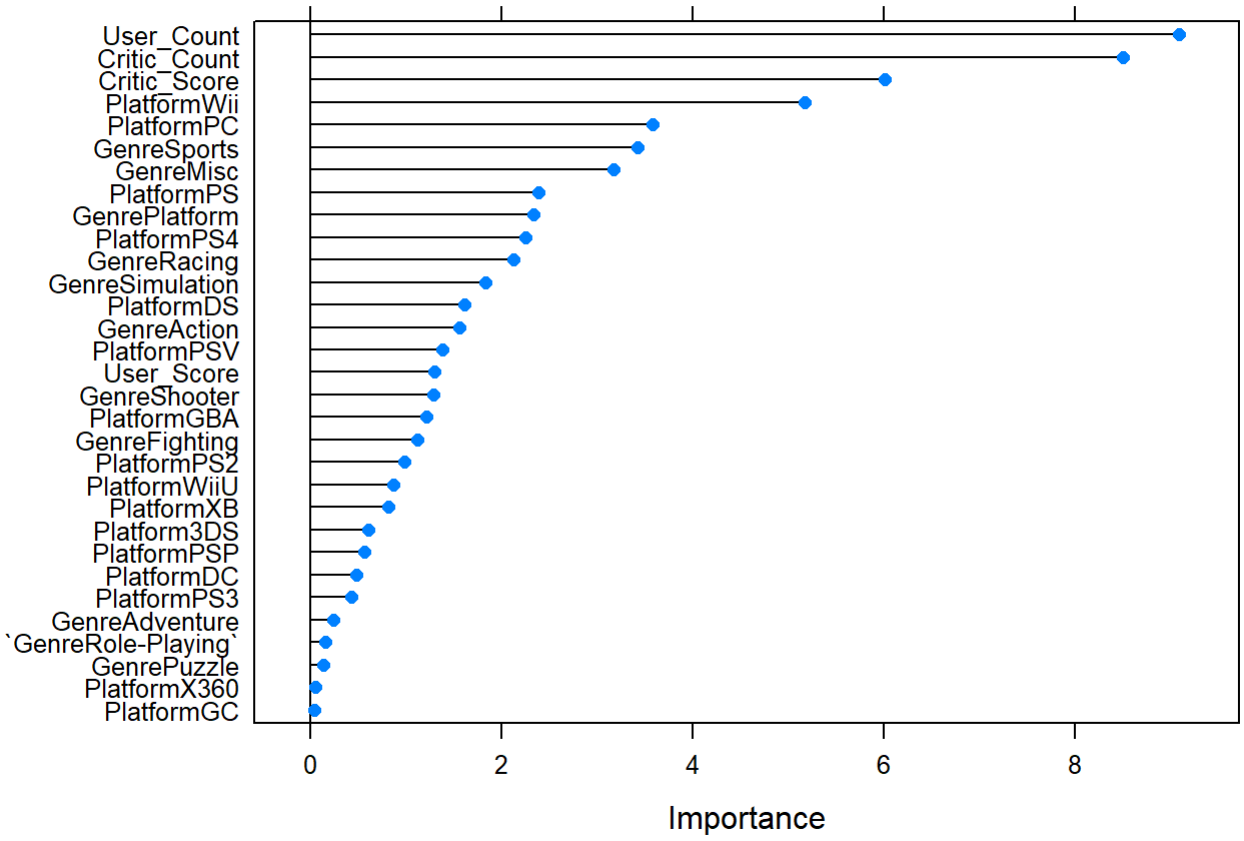
```
fit_GLM <- train(NA_Sales ~ Critic_Score + User_Score + Platform + Genre + Critic_Count + User_C
ount,
       data=train_set,
       method="lm")

Results_2 <- predict(fit_GLM, test_set)

GLM_Model <- RMSE(Results_2, test_set$NA_Sales)
```

```
## The resulting RMSE is 0.70951
```

From the following graph we can observe that user and critic count are the most important features, followed by critic score and platform Wii.



KNN

```
fit_knn <- knn3(NA_Sales ~ Critic_Score + User_Score + Platform + Genre + Critic_Count + User_Co
unt,
                data=train_set, k=10)
Results_3 <- predict(fit_knn, test_set)
Knn_Model <- RMSE(Results_3, test_set$NA_Sales)
```

```
## The resulting RMSE is 0.8948646
```

Note: the varImp function does not provide support for the knn algorithm on feature importance.

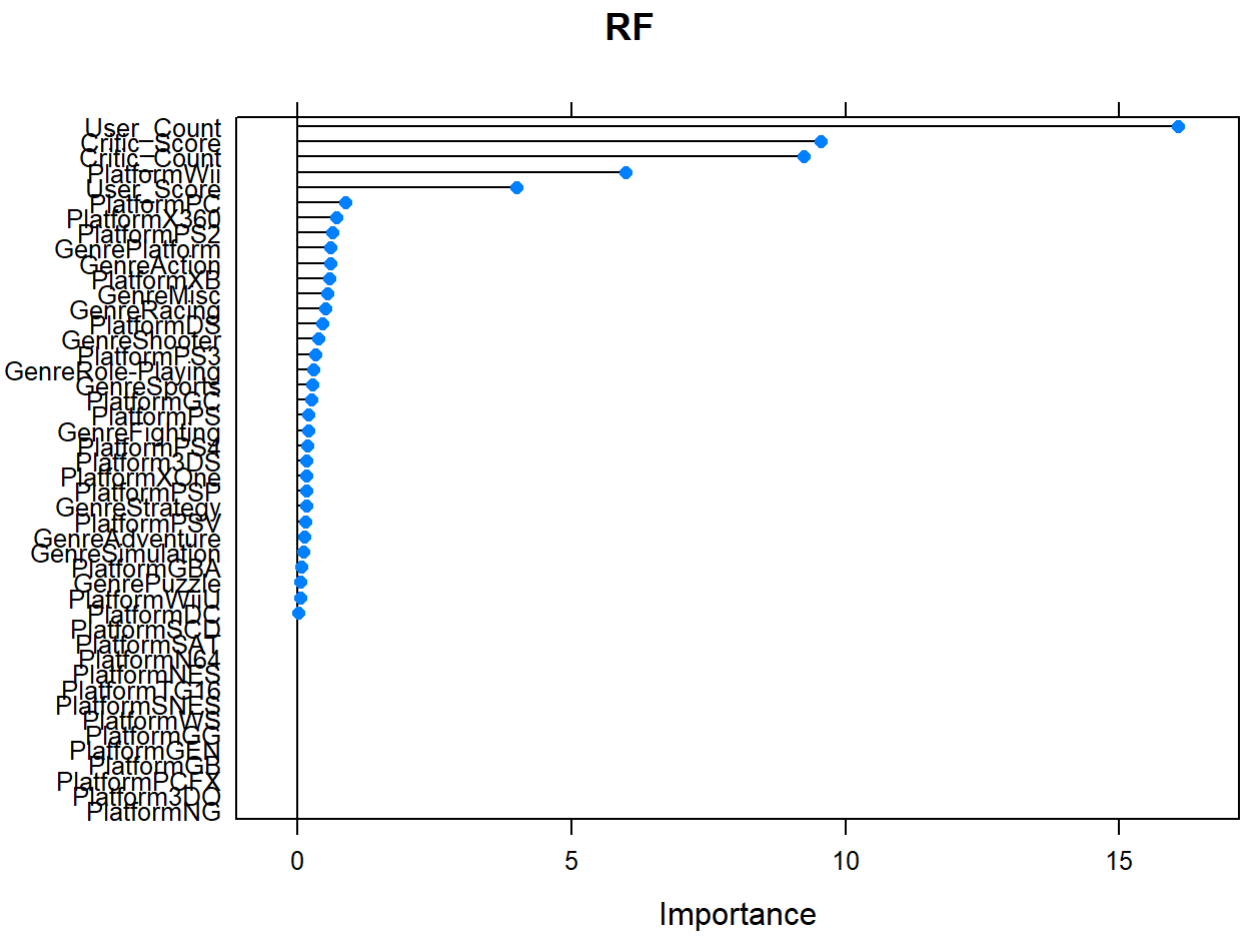
Random Forest

```
fit_rf <- train(NA_Sales ~ Critic_Score + User_Score + Platform + Genre + Critic_Count + User_Co
unt,
                  data=train_set, method="Rborist", tuneGrid = data.frame(predFixed = 2, minNode
= c(3, 50)),)

 Results_4<- predict(fit_rf, test_set)
 RF_Model<- RMSE(Results_4, test_set$NA_Sales)
```

```
## The resulting RMSE is 0.6860307
```

We can observe from the following graph that the most important features are user count and critic score, followed by critic count and platform Wii.



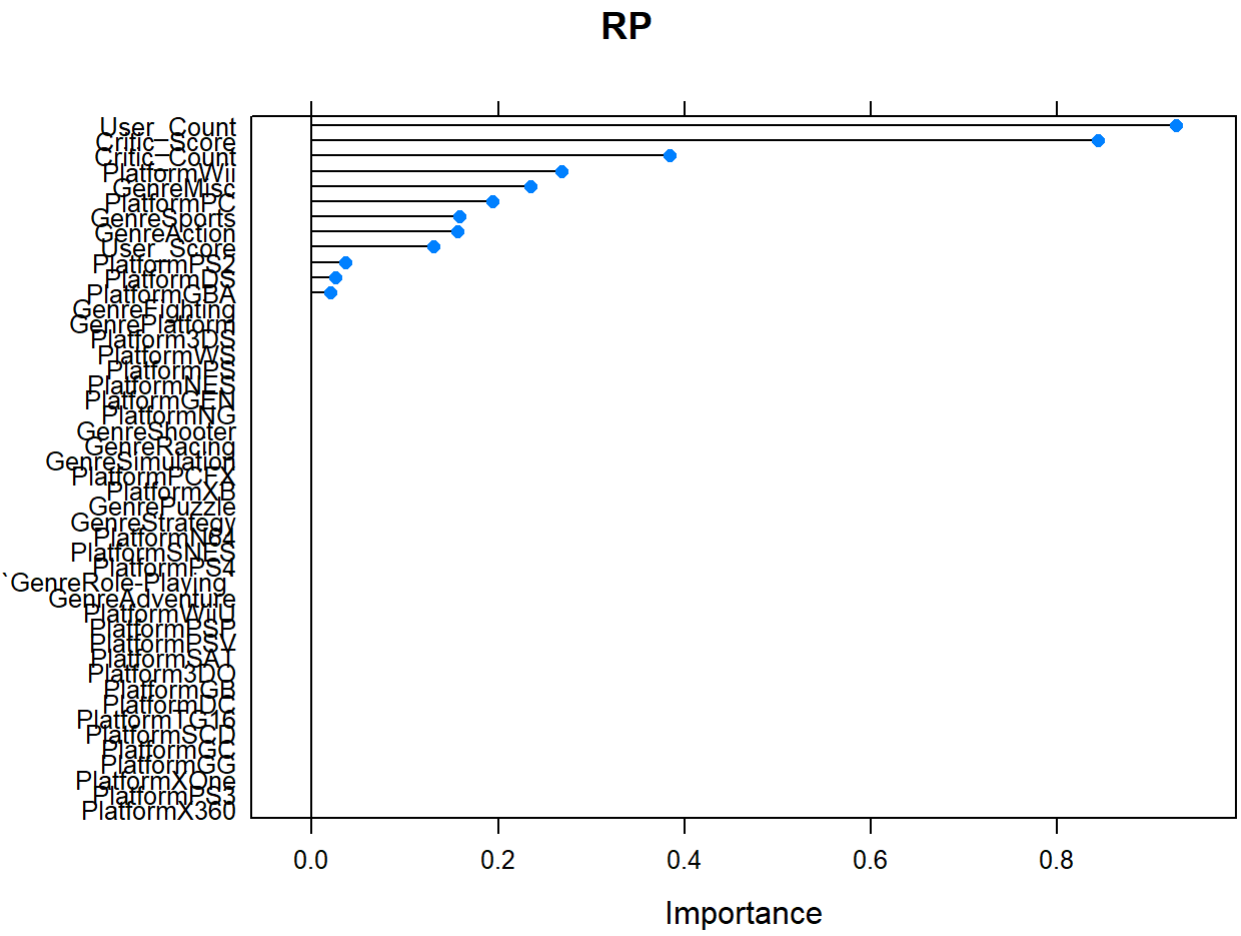
Classification Trees

```
fit_rp <- train(NA_Sales ~ Critic_Score + User_Score + Platform + Genre + Critic_Count + User_Co
unt,
                data=train_set, method="rpart", tuneGrid = data.frame(cp = seq(0.0, 0.1, len =
25)),)

 Results_5<- predict(fit_rp, test_set)
 RP_Model<- RMSE(Results_5, test_set$NA_Sales)
```

```
## The resulting RMSE is 0.8023646
```

We can observe from the following graph that the most important features are user count and critic score, followed by critic score.



In summary, from the following table we can observe that the method that minimizes the RMSE is the Random Forest Model. However, the GLM algorithm was a close second.

```
##                        Model      RMSE
## 4       Random Forest Model 0.6860307
## 2                 GLM Model 0.7095100
## 5      Classification Trees 0.8023646
## 1 Naive Mean-Baseline Model 0.8071075
## 3                 knn Model 0.8948646
```

Conclusion

From the results presented we concluded that the Random Forest model was the algorithm, among the ones selected, that best performed. The data set selected for the project was greatly reduced after the data cleaning process, and despite this issue, the Random Forest model achieved an RMSE of less than 0.7. The most important features in general, were critic and user counts and scores. Herein lies the impact of the report, the ability to predict sales in your most important region by understanding the features that drive sales. It can be used in terms of the production of the video game itself, and how to market it to customers by making sure it receives a decent volume of user and critic scores. The model is limited by the size of data set, as previously mentioned. Future work could be centered around incorporating other regional sales figures into the model.