IGN Video Game Reviews

Capstone

Ellen A. Savoye

July 14, 2018

Introduction

Over the last 20 years, a plethora of video games have been released in an ever-growing market of gaming consoles. In tandem with the creation of these games, a website called IGN began publishing reviews and ratings of the games mentioned above. These reviews and resulting ratings can be analyzed to create insights that would be useful to console and video game creators. With analysis, console and video game creators would be able to determine their market/popularity standing in comparison to other consoles and games. Furthermore, they would be able to determine if there is a particular genre that needs more development to increase their rating and market standpoint. Sourced from Kaggle (https://www.kaggle.com/egrinstein/20-years-of-games) and IGN (http://ign.com/games/reviews) via a crawl, the data consists of 20 years worth of video game data.

Caveats

The data does not contain any financial information relating to the volume of games sold or the monetary amount it sold for. Insights are curated based on the number of games released, their rating, the genre, and the console the game was released on. Some games, like 'Gears of War,' have been released on multiple consoles. I have not adjusted the data to constrain these types of games down to one platform.

The Data

The data, 20 years' worth of IGN game reviews, consists of 18,625 records. The raw dataset had ten columns (listed below).

Variable	Description
score_phrase	Phrase given to describe the overall score
title	Game title
url	IGN Game URL
platform	Game Console
genre	Video game genre
score	Overall rating for the video game
editors_choice	Editor Recommended (Y/N)
release_year	Year of game release
release_month	Month of game release
release_day	Day of game release

Data Wrangling

To take a proper look at the data, I loaded the original dataset as a CSV file and the necessary libraries. Of the variables available for use, score_phrase, platform, score, genre, editors_choice, release_year, release_month, and release_day are the ones I used in my analysis. As such, I analyzed them for missing values, outliers, and whether or not the number of distinct factors in each was usable. Editors_choice, score_phrase, and score did not need cleaning. However, when checking release_year, I noticed an outlier titled "The Walking Dead: The Game – Episode 1: A New Day". This record had a release date of 1/1/1970. Given the dataset is spanning 1996 - 2016, I chose to correct the outlier to the correct release date of 4/24/2012.

```
## # A tibble: 5 x 11
##      X1 score_phrase title    url      platform score genre editors_choice
##   <int> <chr>        <chr>    <chr>    <chr>    <dbl> <chr> <chr>
## 1     0 Amazing      LittleB~ /games/~ PlaySta~   9   Plat~ Y
## 2     1 Amazing      LittleB~ /games/~ PlaySta~   9   Plat~ Y
## 3     2 Great        Splice:~ /games/~ iPad       8.5 Puzz~ N
## 4     3 Great        NHL 13   /games/~ Xbox 360   8.5 Spor~ N
## 5     4 Great        NHL 13   /games/~ PlaySta~   8.5 Spor~ N
## # ... with 3 more variables: release_year <int>, release_month <int>,
## #   release_day <int>
```

```
## # A tibble: 5 x 1
##   release_year
##          <int>
## 1         1970
## 2         1996
## 3         1997
## 4         2012
## 5         2013
```

```
## # A tibble: 1 x 11
##      X1 score_phrase title    url      platform score genre editors_choice
##   <int> <chr>        <chr>    <chr>    <chr>    <dbl> <chr> <chr>
## 1   516 Great        The Wal~ /games/~ Xbox 360   8.5 Adve~ N
## # ... with 3 more variables: release_year <int>, release_month <int>,
## #   release_day <int>
```

With the outlier corrected, platform and genre variables remained. The original platform variable consisted of 59 distinct factors. Because platform spanned multiple generations of systems (e.g., PlayStation 1-3) and because not all manufacturers kept system naming consistent, I chose to combine the values into a condensed version based on system name/manufacturer and created a new variable named platform_group. To do so, I loaded a 'platform map' CSV file to merge the new platform_group variable onto the original dataset. After comparing the original platform variable against the new platform_group to ensure no misplaced systems, I moved onto the genre variable.

```
## # A tibble: 59 x 3
## # Groups:   platform [?]
##    platform            platform_group `n_distinct(platform_group)`
##    <chr>               <chr>                                <int>
##  1 Android             Android                                  1
##  2 Arcade              Other                                    1
##  3 Atari 2600          Atari                                    1
##  4 Atari 5200          Atari                                    1
##  5 Commodore 64/128    Other                                    1
##  6 Dreamcast           Sega                                     1
##  7 Dreamcast VMU       Sega                                     1
```

```
##  8 DVD / HD Video Game Other                                    1
##  9 Game Boy            Game Boy                                 1
## 10 Game Boy Advance    Game Boy                                 1
## # ... with 49 more rows
```

Similar to the platform variable, the genre variable has a multitude of factors which makes intelligent analysis a bit difficult. There are 113 unique genres within the field. I chose my grouping based on an overall description (e.g., Sports, Cards, Action, etc.) given the numerous distinct factors. Before cleaning up the column, I checked for any blank cells. Out of 18,625 observations, 36 do not have a genre which is .19%. Due to the unpopulated records being less than 1% of the overall genre column, I chose not to populate them but instead mapped them to 'Other.' To map genre, I loaded a 'genre map' CSV file to merge the new genre_group variable onto the original dataset. In doing so, I brought the number of unique genres from 113 to 21.

```
## # A tibble: 36 x 12
##       X1 score_phrase title    url     platform score genre editors_choice
##    <int> <chr>        <chr>    <chr>   <chr>    <dbl> <chr> <chr>
##  1    12 Good         Wild Bl~ /games~ iPhone     7   <NA>  N
##  2   113 Good         Retro/G~ /games~ PlaySta~   7   <NA>  N
##  3   160 Good         10000000 /games~ iPhone     7.5 <NA>  N
##  4   176 Okay         Colour ~ /games~ PC         6.2 <NA>  N
##  5  9375 Great        Duke Nu~ /games~ Wireless   8   <NA>  Y
##  6  9488 Okay         Rengoku  /games~ Wireless   6.5 <NA>  N
##  7  9767 Good         Super S~ /games~ Wireless   7.5 <NA>  N
##  8  9774 Amazing      Critter~ /games~ Wireless   9   <NA>  Y
##  9 10494 Awful        Clue / ~ /games~ Nintend~   3.5 <NA>  N
## 10 11367 Painful      Jeep Th~ /games~ PlaySta~   2   <NA>  N
## # ... with 26 more rows, and 4 more variables: release_year <int>,
## #   release_month <int>, release_day <int>, platform_group <chr>
```

```
##  [1] Platformer   Puzzle       Sports       Strategy     Fighting
##  [6] RPG          <NA>         Action       Adventure    Shooter
## [11] Music        Other        Racing       Simulation   Education
## [16] Wrestling    Productivity Cards        Compilation  Flight
## [21] Pinball      Hunting
## 21 Levels: Action Adventure Cards Compilation Education ... Wrestling
```

```
## [1] 22
```

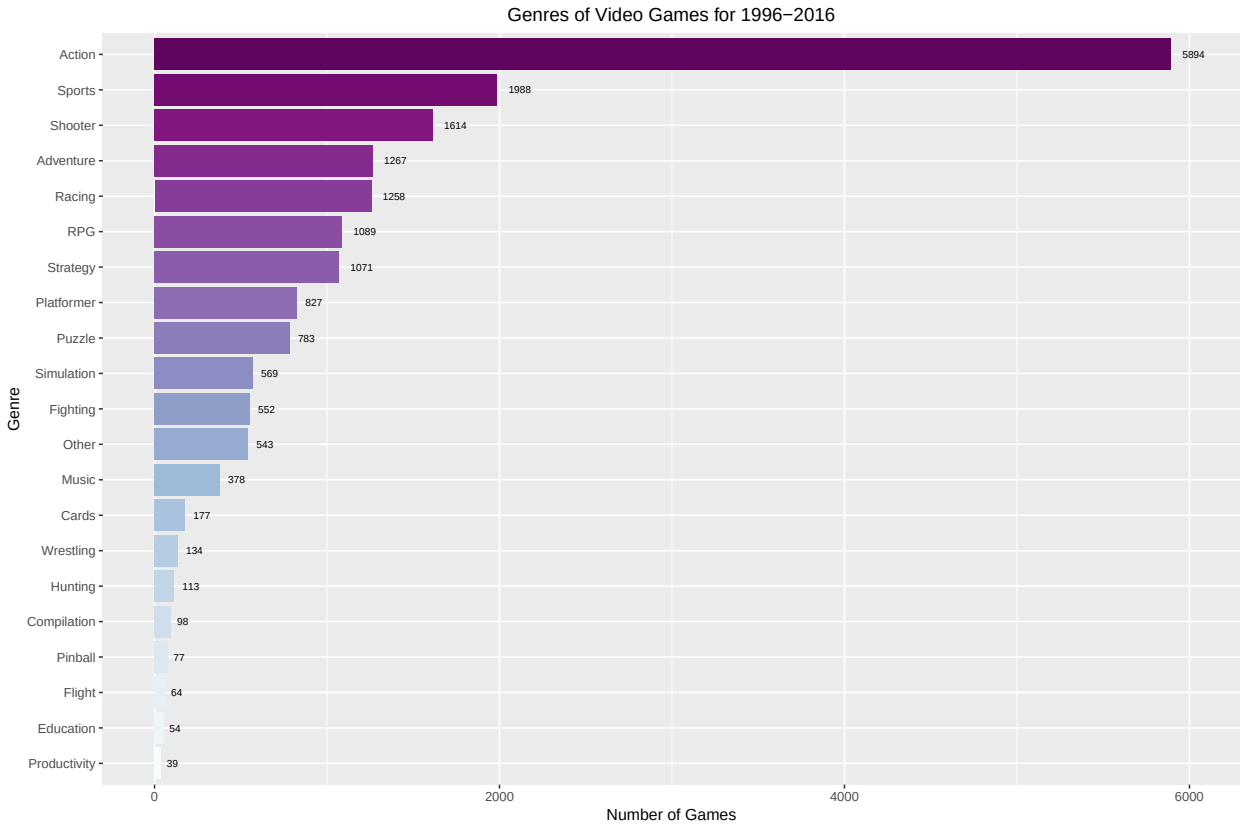
After cleaning up the variables that I will be using in my analysis, I wrote the wrangled data to a new file called "ign_clean.csv" for further use in creating insights.

Exploratory Data Analysis

Genre

Using the condensed genre field, I'm looking to see what the top genres are regarding the number of video game releases and whether or not the scores correspond to the top genres.

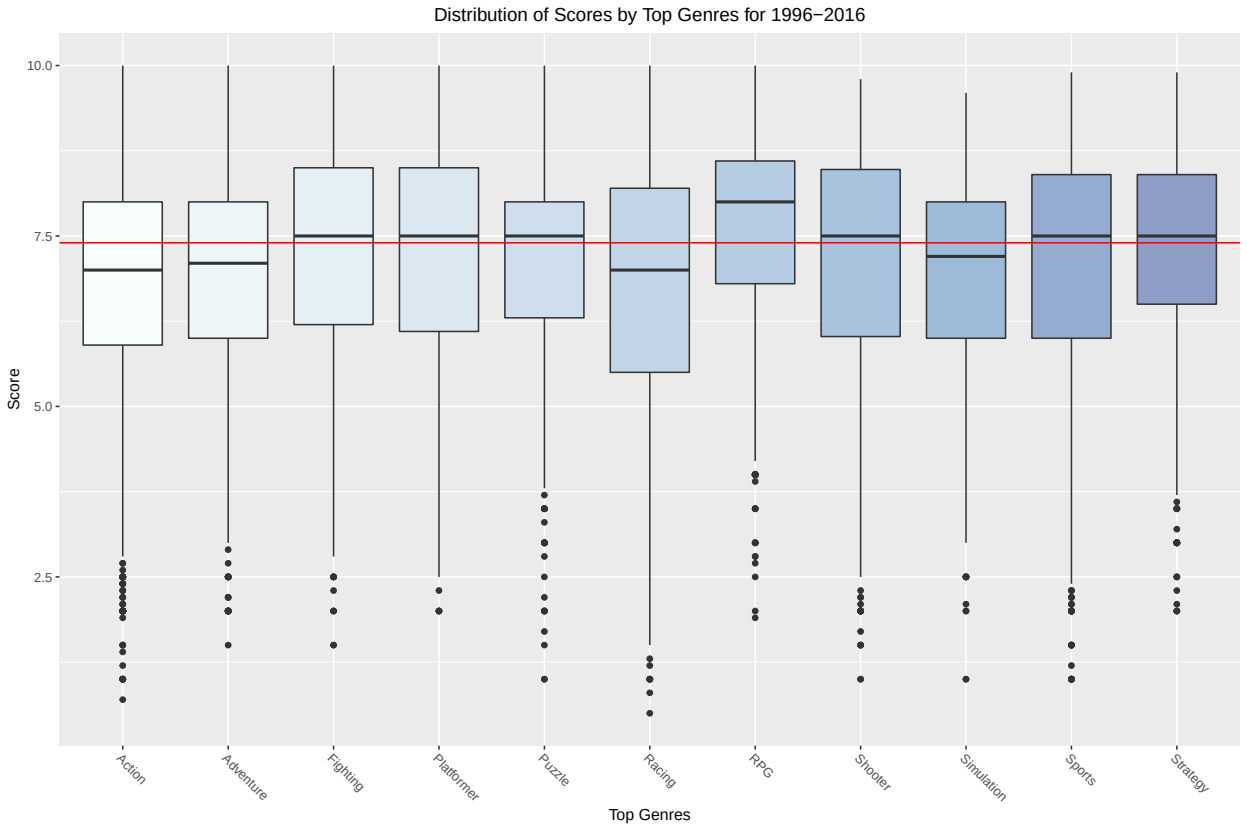
Top Genres



Cumulatively, the top genre is Action followed by Sports, Shooter, Adventure, and Racing. The number of Action games is more than double the next genre, Sports. As I mentioned in the caveats section, this may be due to some games applying across platforms and therefore counted multiple times. Now that we have the top genres, we can see if there are any insights to glean.

Top Genres by Score

Due to a little more than half of the genres having a count of 500+, I'm using 500 as my minimum number of games to filter my data. However, given that 'other' is a catch-all bucket, the minimum will instead be 550 to exclude 'other.'



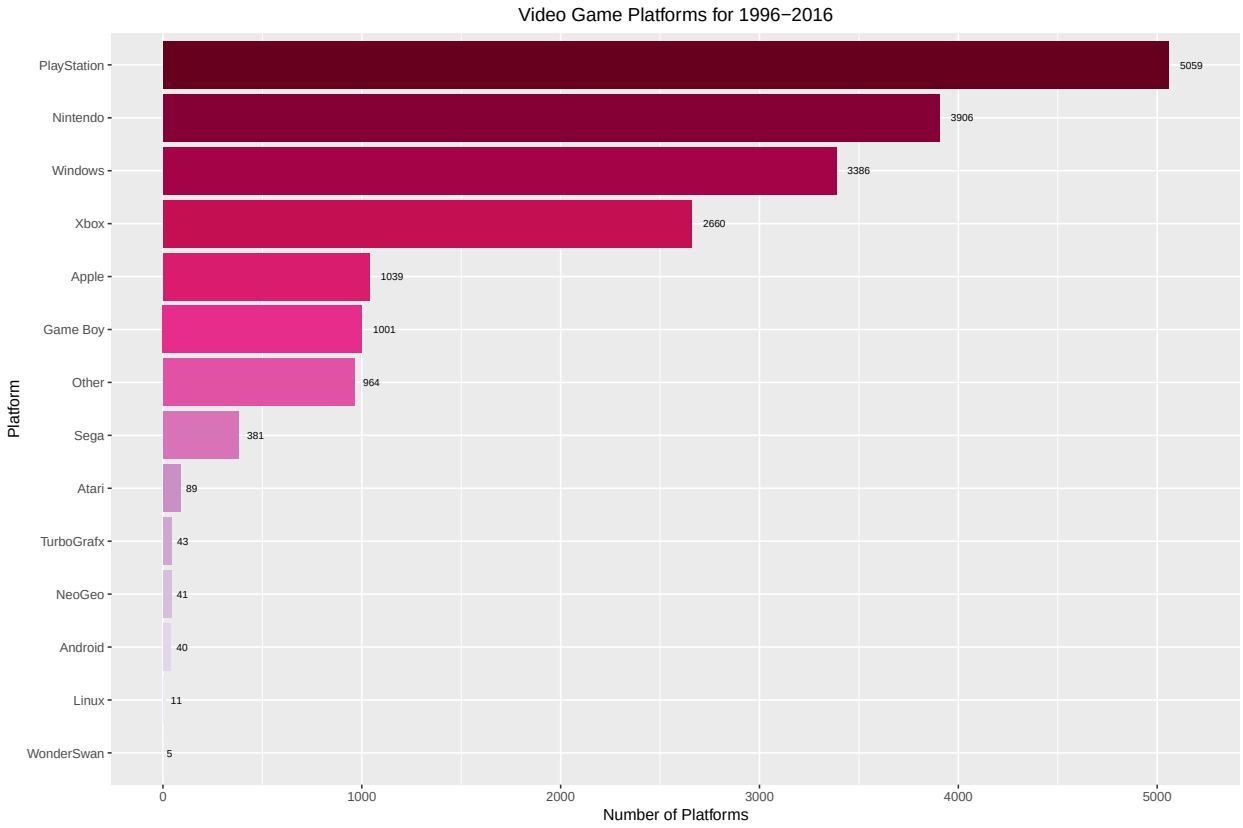

```
##       score_phrase                                    title platform score
## 891       Disaster                    Extreme PaintBrawl       PC   0.7
## 5243      Disaster Looney Tunes: Back in Action: Zany Race Wireless   0.5
## 12514     Disaster                  Action Girlz Racing      Wii   0.8
##        genre editors_choice release_year release_month release_day
## 891   Action              N         1998            10          29
## 5243  Racing              N         2003            10          28
## 12514 Racing              N         2009             2          11
##       platform_group genre_group
## 891          Windows      Action
## 5243           Other      Racing
## 12514       Nintendo      Racing
```

When looking at the distribution of scores for the top 11 genres, we can see that RPG has higher ratings than the other genres even though it sits in sixth place for volume of games released. Given the dearth of games released in the Action genre, I expected it to have the higher scores across the board from a slight volume influenced stand-point. Action, Adventure, and Racing have lower scores with Racing having the largest IQR. Both Action and Racing have games with a rating less than 1. Two of the games are in Racing under while the third is in Action.

Platforms

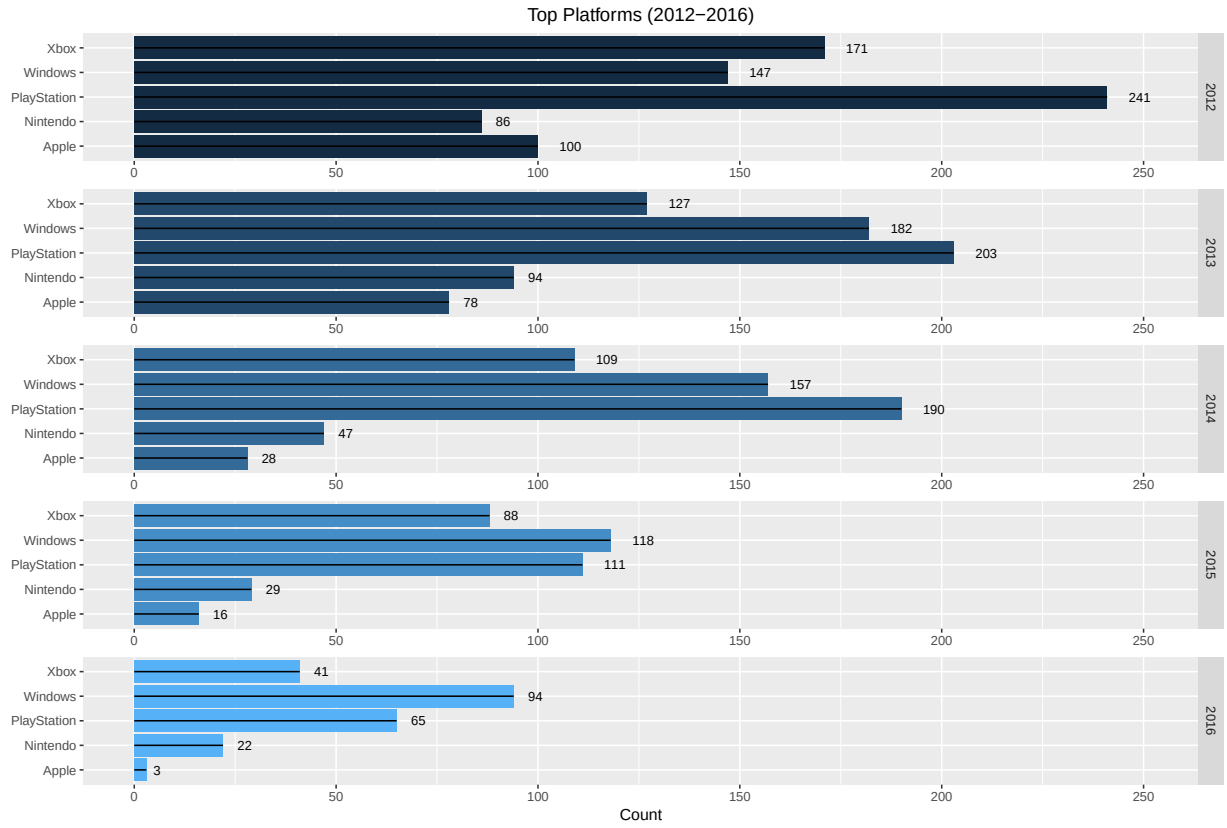
I want to see which platform reigned supreme over the last twenty years and if that holds true when looking at the past five years.

Top Platforms



While some of these platforms have had multiple versions/evolutions over the last 20 years, Nintendo for example, they have been grouped for a more straightforward analysis on the major gaming platforms. Even though PlayStation launched after Nintendo, PlayStation is still the top system followed by Nintendo, Windows, and Xbox. Given the previous graph consisted of all 20 years worth of data, I put together the last five years of data into a chart to see if the top systems remained consistent.

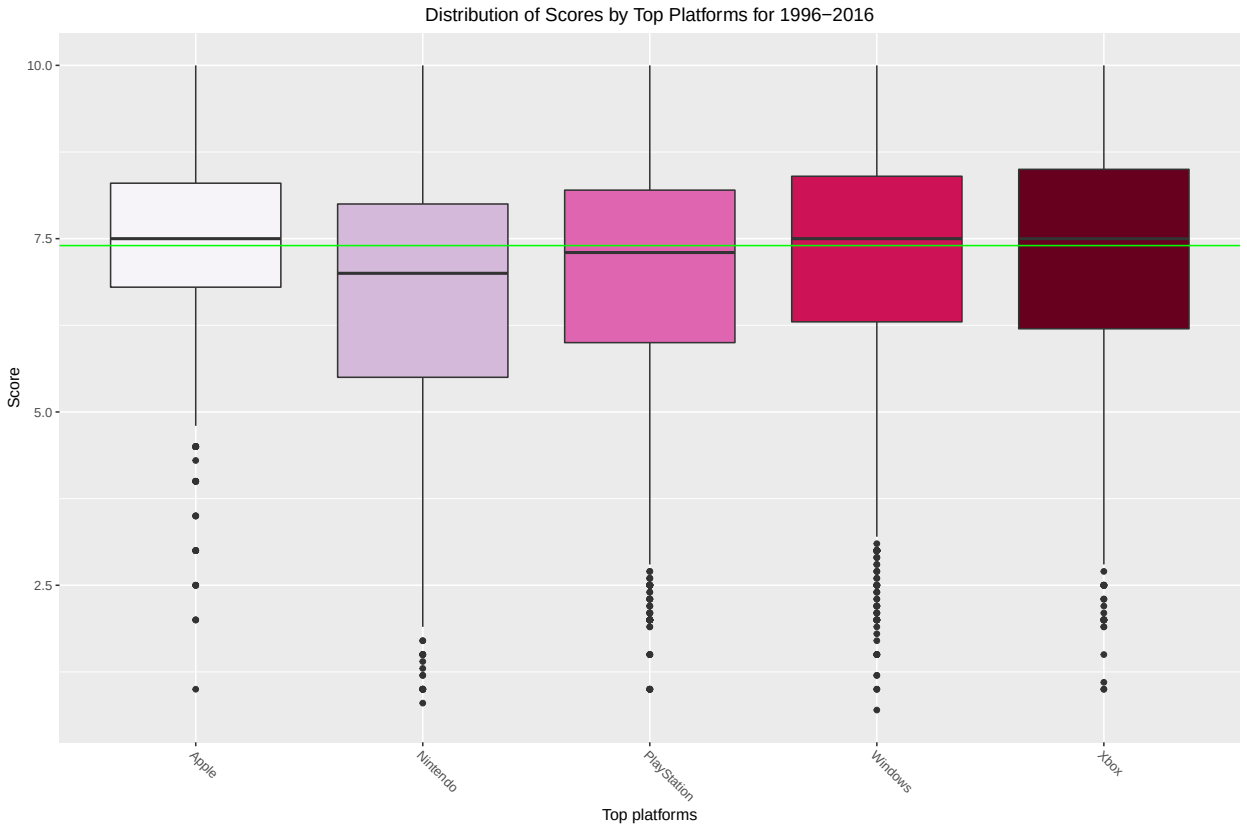
Top Platforms for Past 5 Years



For 2012-2014, PlayStation held down the top spot which is consistent with all years combined. However, 2015-2016, Windows is the top platform. So even though PlayStation holds the top place in the number of games released overall, it doesn't necessarily hold true on a year-by-year basis.

Top Platforms by Score

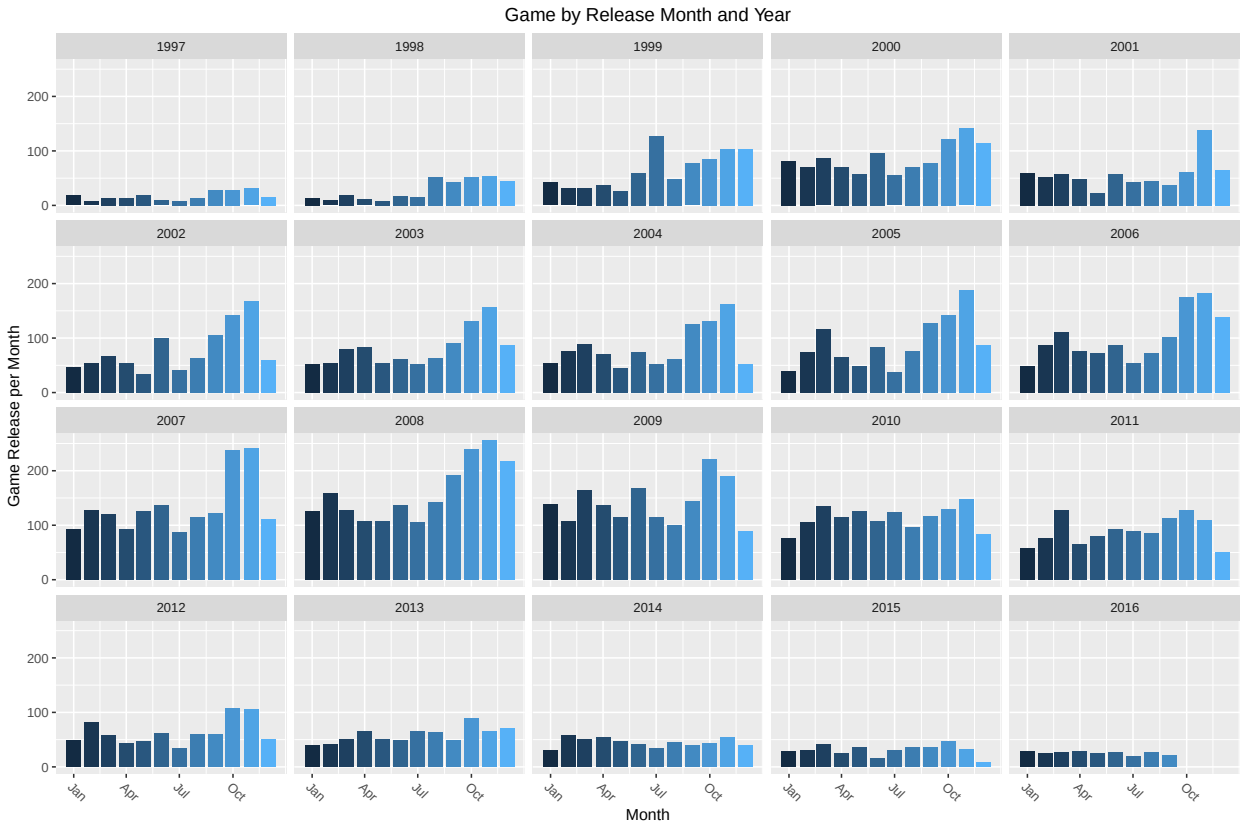
Taking the top 5 platforms found in our "Top Platforms" exhibit above, I want to see how the scores correspond to platform volume.



Given the number of game releases that PlayStation has had, there was some expectation for a better score performance. However, this doesn't appear to be the case. Both Xbox and Windows have higher scores for the 25th, 50th, and 75th IQR values. Windows has the top number of released games for 2015 and 2016, so the higher scores are not wholly a surprise. One thing I've noticed personally is a shift towards computer-based, Windows, games due to the ability to fine-tune graphics cards and speed for better gaming performance.

Important Dates

We want to see if a particular month and year stand out as significant. To do so, I looked at the month and year together in a grid.



From 1997 - 2008, the number of games released continuously increases, with 2008 having the most reviewed and published in video game history. In older years, there is a more significant amount of released video games in October and November over other months. The more recent years have more consistent releases over the course of the year. The decreased number of games released could be due to video games becoming significantly more intricate and graphics intensive. With the capabilities of T.V. and graphics cards now, immersive video game graphics seem like a must-have for a game to be released.

Predictive Models

Subsetting the Data

Before any ventures into modeling, a data set was created to remove any unnecessary features and any blank records. 36 records with a 'blank' genre were removed. Given that some of the kept features are categorical, conversion to binary/dummy variables was needed using a model matrix. I chose not to combine release day, month, and year into a combined date field as I don't believe it would have given any valuable insight as unique a combined value.

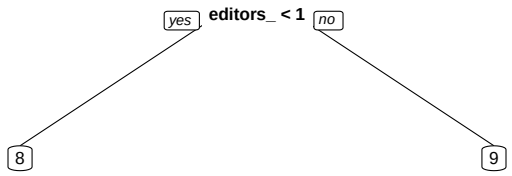
```
##   score_phrase          title       platform          score          genre
##              0              0              0              0             36
## editors_choice   release_year  release_month    release_day platform_group
##              0              0              0              0              0
##    genre_group
##             36
```

Linear Regression

```
##
## Call:
## lm(formula = score ~ ., data = reg_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0058 -0.6557  0.2087  1.0293  3.4429
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.72969    0.03872 173.784  < 2e-16 ***
## editors_choiceY             2.30134    0.02689  85.593  < 2e-16 ***
## platform_groupGame.Boy     -0.67136    0.05717 -11.744  < 2e-16 ***
## platform_groupNintendo     -0.61872    0.04200 -14.732  < 2e-16 ***
## platform_groupOther        -0.36714    0.05788  -6.343 2.31e-10 ***
## platform_groupPlayStation -0.36001    0.04068  -8.850  < 2e-16 ***
## platform_groupWindows      -0.22569    0.04333  -5.209 1.92e-07 ***
## platform_groupXbox         -0.22600    0.04486  -5.038 4.75e-07 ***
## genre_groupRacing          -0.17263    0.04172  -4.138 3.52e-05 ***
## genre_groupRPG              0.40065    0.04496   8.912  < 2e-16 ***
## genre_groupStrategy         0.19513    0.04661   4.186 2.85e-05 ***
## release_month2              0.10024    0.04397   2.280 0.022630 *
## release_month3              0.10592    0.04132   2.563 0.010379 *
## release_month4              0.09026    0.04480   2.015 0.043954 *
## release_month6              0.15790    0.04210   3.751 0.000177 ***
## release_month8              0.20687    0.04384   4.719 2.39e-06 ***
## release_month9              0.28605    0.04005   7.142 9.52e-13 ***
## release_month10             0.20184    0.03598   5.611 2.05e-08 ***
## release_month11             0.12746    0.03437   3.709 0.000209 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.42 on 18570 degrees of freedom
## Multiple R-squared:  0.3127, Adjusted R-squared:  0.312
## F-statistic: 469.3 on 18 and 18570 DF,  p-value: < 2.2e-16
```

The first iteration of the linear regression showed multicollinearity between score and score phrase with a multiple R-squared of 0.9726. After removing score_phrase, R-squared drops to 0.3187. In an attempt to improve R-squared, near zero variance predictors (NZP) were identified and discarded because they are non-informative and tend to occur when breaking categorical variables into dummy variables, as was done above. After removing NZP and checking for high collinearity, a few iterations of the linear model were run to narrow down which features are not significant. A few of the fields removed are platform_groupApple, genre_groupAdventure, and genre_groupSports. After eliminating non-significant fields, the adjusted R-squared is now 0.3127.

CART/Random Forest




```
##
## Call:
##  randomForest(formula = score ~ ., data = Train, ntree = 500)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 6
##
##           Mean of squared residuals: 2.026652
##                     % Var explained: 30.95
```

When looking at the first tree, the only node shown is for editor's choice. If editor's choice is higher than 0.5, it is 'no' or 1. One important note is that a high score can imply an editor's choice designation. However, a high score, say 8 or higher, does not automatically receive an editor's choice designation. Using the complexity parameter (cp), we can force more nodes to appear with cp = .0025. While the complexity parameter forced more nodes, the additional splits did not provide any valuable insights or logical progressions in the decision tree. One thing to note for the second decision tree, editor's choice has remained as the first node. The first tree produces an R-squared of 30.4% which is similar to the linear regression model. When forcing the tree to have multiple nodes, the same R-squared calculation doesn't work as nicely nor is it the most efficient.

The random forest model, built on the training dataset, appears to be the best model based on a comparison of R-squared at 32.69% versus 30.4% (CART), and 31.27% (linear regression).

Conclusion

Between our exploratory data analysis and machine learning models, there are a few takeaways. While the three models don't have high R-squared, this doesn't mean our models are inadequate. The predictor values in each are significant which we can still make important conclusions on. Based on the predictor values, we can see that while specific platforms are significant in their effect on the score, the relationship is negative. Genres RPG and Strategy along with the significant release months have a positive relationship with our score variable. One thing to keep in mind is the human factor. Human opinion influenced the ratings for each game in our source dataset which can make it more challenging to create a precise model.

The human aspect is one limitation to this analysis. Another limitation is the lack of detail in what went into the determination of scores, other than genre and console, for each review: i.e., graphics rating, storyline, etc.

Based on my findings, I would recommend an additional study be performed incorporating more information regarding score determination as well as financial information relating to each game: sales volume, cost of each game, etc. The additional information will, theoretically, provide more detail to aid in accurately predicting video game rating and how they stack up to each other in the market.