IGN Video Game Reviews

Data Story

by Ellen A Savoye

May 8, 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 releasing reviews and ratings of the aforementioned games. These ratings and reviews could potentially be compiled and 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 in order to potentially increase their rating and market standpoint. To do such an analysis, data was sourced from Kaggle (https://www.kaggle.com/egrinstein/20-years-of-games) and IGN (http://ign.com/games/reviews), via a crawl, consisting 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 was sold for. Any potential insights are solely related to the number of games released, their rating, and the console the game was released on. Some games are released across multiple consoles. I have not adjusted the data to constrain games from multiple platforms down to one.

The Data

The data, 20 years' worth of IGN game reviews, consists of 18,625 records. The raw dataset had 10 columns (listed below).

Variable	Description
score_phrase	Phrase given to describe overall score
title	Game title
url	IGN Game URL
platform	Game Console
genre	Video game genre
score	Overall rating for 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

Table 2: IGN Data (continued below)

X1	score_phrase	title
0	Amazing	LittleBigPlanet PS Vita
1	Amazing	LittleBigPlanet PS Vita – Marvel Super Hero Edition
2	Great	Splice: Tree of Life
3	Great	NHL 13
4	Great	NHL 13

Table 3: Table continues below

url
/games/littlebigplanet-vita/vita-98907
/games/littlebigplanet-ps-vita-marvel-super-hero-edition/vita-20027059
/games/splice/ipad-141070
/games/nhl-13/xbox-360-128182
/games/nhl-13/ps3-128181

Table 4: Table continues below

platform	score	genre	editors_choice	release_year
PlayStation Vita	9	Platformer	Y	2012
PlayStation Vita	9	Platformer	Y	2012
iPad	8.5	Puzzle	N	2012
Xbox 360	8.5	Sports	N	2012
PlayStation 3	8.5	Sports	N	2012

release_month	release_day
9	12
9	12
9	12
9	11
9	11

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.

```
head(tbl_df(IGN_data), 5)
```

```
## # 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~  9.00 Plat~ Y
## 2     1 Amazing      LittleB~ /games/~ PlaySta~  9.00 Plat~ Y
## 3     2 Great        Splice:~ /games/~ iPad      8.50 Puzz~ N
## 4     3 Great        NHL 13   /games/~ Xbox 360  8.50 Spor~ N
## 5     4 Great        NHL 13   /games/~ PlaySta~  8.50 Spor~ N
## # ... with 3 more variables: release_year <int>, release_month <int>,
## #   release_day <int>
```

```
# Release year is supposed to be higher than 1995

head(IGN_data %>% distinct(release_year), 5) %>% arrange(release_year)
```

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

```
IGN_data[IGN_data$release_year == "1970", ]
```

```
## # 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  8.50 Adve~ N
## # ... with 3 more variables: release_year <int>, release_month <int>,
## #   release_day <int>
```

```
IGN_data <- IGN_data %>% mutate(release_year = if_else(title ==
    "The Walking Dead: The Game -- Episode 1: A New Day", as.integer(2012),
    release_year)) %>% mutate(release_month = if_else(title ==
    "The Walking Dead: The Game -- Episode 1: A New Day", as.integer(4),
    release_month)) %>% mutate(release_day = if_else(title ==
    "The Walking Dead: The Game -- Episode 1: A New Day", as.integer(24),
    release_day))
```

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.

```
# 59 variables in original platform column

IGN_data %>% distinct(platform) %>% arrange(platform)

Platform_Map <- read.csv("platform_map.csv")

IGN_data_v2 <- IGN_data %>% mutate_at(vars(platform), funs(as.factor(as.character(.))))

IGN_data_v2 <- IGN_data_v2 %>% left_join(Platform_Map, by = c(platform = "platform"))
```

```
IGN_data_v2 %>% group_by(platform, platform_group) %>% summarise(n_distinct(platform_group))
```

```
## # A tibble: 59 x 3
## # Groups:   platform [?]
##    platform            platform_group `n_distinct(platform_group)`
##    <fct>               <fct>                                 <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 blank 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.

```
# Check for blanks in genre column

IGN_data_v2 %>% distinct(genre) %>% arrange(genre)
```

```
group_by(IGN_data_v2[is.na(IGN_data_v2$genre), ])
```

```
## # A tibble: 36 x 12
##       X1 score_phrase title    url     platform score genre editors_choice
##    <int> <chr>        <chr>    <chr>   <fct>    <dbl> <chr> <chr>
##  1    12 Good         Wild Bl~ /games~ iPhone    7.00 <NA>  N
##  2   113 Good         Retro/G~ /games~ PlaySta~  7.00 <NA>  N
##  3   160 Good         10000000 /games~ iPhone    7.50 <NA>  N
##  4   176 Okay         Colour ~ /games~ PC        6.20 <NA>  N
##  5  9375 Great        Duke Nu~ /games~ Wireless  8.00 <NA>  Y
##  6  9488 Okay         Rengoku  /games~ Wireless  6.50 <NA>  N
##  7  9767 Good         Super S~ /games~ Wireless  7.50 <NA>  N
##  8  9774 Amazing      Critter~ /games~ Wireless  9.00 <NA>  Y
##  9 10494 Awful        Clue / ~ /games~ Nintend~  3.50 <NA>  N
## 10 11367 Painful      Jeep Th~ /games~ PlaySta~  2.00 <NA>  N
## # ... with 26 more rows, and 4 more variables: release_year <int>,
## #   release_month <int>, release_day <int>, platform_group <fct>
```

```
# 113 unique factors in genre column

IGN_data_v2 %>% distinct(genre) %>% arrange(genre)

Genre_Map <- read.csv("genre_map.csv")

Genre_Map <- Genre_Map %>% mutate_at(vars(genre), funs(as.character(as.factor(.))))
```

```
IGN_data_v2 <- IGN_data_v2 %>% left_join(Genre_Map, by = c(genre = "genre"))
```

```
unique(IGN_data_v2$genre_group)
```

```
##  [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
```

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.

```
write_csv(IGN_data_v2, "ign_clean.csv")
IGN_data_cleaned <- read.csv("ign_clean.csv")
```

Exploratory Data Analysis

Genre

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

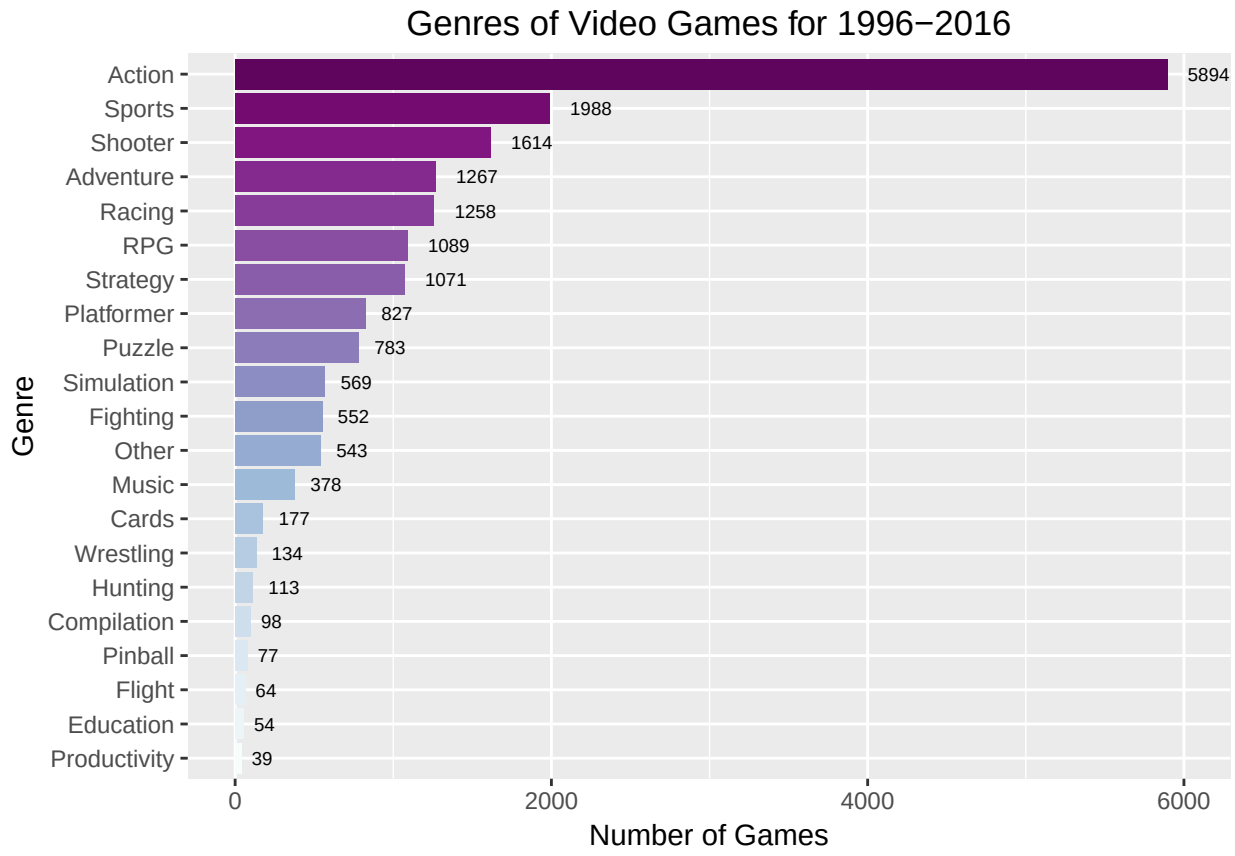
Top Genres

```
top_genres <- IGN_data_cleaned %>% group_by(genre_group) %>% summarize(genres_count = n()) %>%
    arrange(desc(genres_count))

top_genres$genre_group <- factor(top_genres$genre_group, levels = top_genres$genre_group[order(top_genr

colourCount = length(unique(top_genres$genre_group))
fill_purple <- colorRampPalette(brewer.pal(9, "BuPu"))

top_genres %>% filter(genre_group != "NA") %>% ggplot(aes(x = genre_group, y = genres_count,
    fill = genre_group)) + geom_bar(stat = "identity") + coord_flip() + geom_text(aes(label = genres_co
    size = 2.5, color = "black", hjust = -0.5) + labs(x = "Genre", y = "Number of Games",
    title = "Genres of Video Games for 1996-2016") + theme(legend.position = "none",
    plot.title = element_text(hjust = 0.5)) + ylim(0, max(top_genres$genres_count +
    100)) + scale_fill_manual(values = fill_purple(colourCount))
```



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 are therefore being counted multiple times. Now that we have the top genres, we can see if there are any insights to glean.

Top Genres by Score

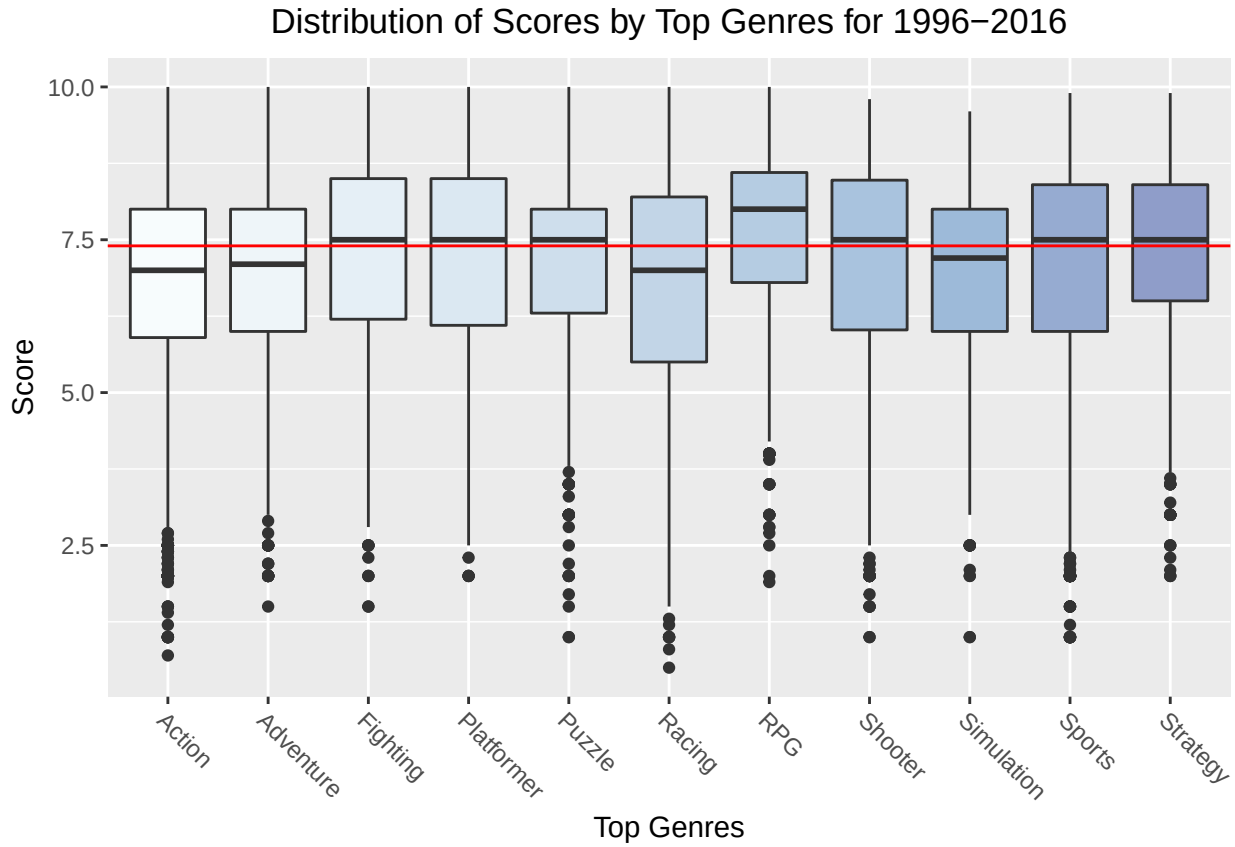
Due to a little more than half of the genres having a count of 500+, I'm using 500 as my minimum number of games to filter my data. However, given that 'other' is a catch-all bucket, the miminum will instead be 550 to exclude 'other'.

```
top_genre_scores <- IGN_data_cleaned %>% group_by(genre_group) %>% summarize(sum_genres_count = n()) %>%
    arrange(desc(sum_genres_count)) %>% filter(sum_genres_count > 550)
top_genre_scores <- top_genre_scores[, 1]

genre_scores_df <- IGN_data_cleaned[IGN_data_cleaned$genre_group %in% top_genre_scores$genre_group,
    ]

colourCount = length(unique(top_genres$genre_group))
fill_purple <- colorRampPalette(brewer.pal(9, "BuPu"))

genre_scores_df %>% ggplot(aes(x = genre_group, y = score, fill = genre_group)) +
    geom_boxplot(alpha = 1) + labs(x = "Top Genres", y = "Score", title = "Distribution of Scores by To
    theme(legend.position = "none", plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = -45, hjust = 0)) + geom_hline(aes(yintercept = median(genre_
    color = "red") + scale_fill_manual(values = fill_purple(colourCount))
```




```
subset(genre_scores_df, genre_scores_df$score < 1)
```

```
##          X1 score_phrase                                    title
## 891     890     Disaster                      Extreme PaintBrawl
## 5243   5242     Disaster Looney Tunes: Back in Action: Zany Race
## 12514 12513     Disaster                     Action Girlz Racing
##                                                          url platform
## 891                        /games/extreme-paintbrawl/pc-10455       PC
## 5243  /games/looney-tunes-back-in-action-zany-race/cell-611813 Wireless
## 12514                    /games/action-girlz-racing/wii-889218      Wii
##       score  genre editors_choice release_year release_month release_day
## 891     0.7 Action              N         1998            10          29
## 5243    0.5 Racing              N         2003            10          28
## 12514   0.8 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 scores than the other genres eve though it sits in sixth place for volumne of games released. Given the darth of games released in the Action genre, I expected it to have the higher scores across the board from a slightly 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 score 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.

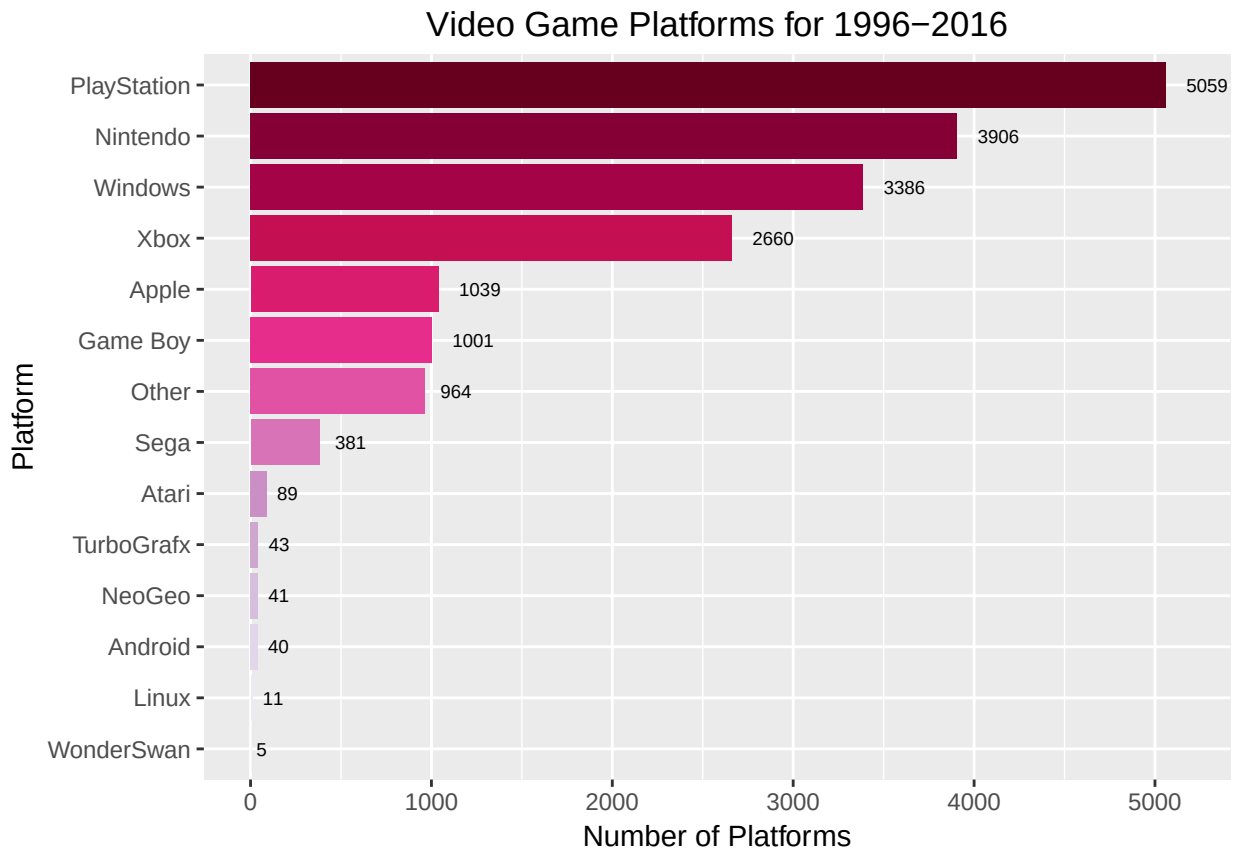
Top Platforms

```
top_platform <- IGN_data_cleaned %>% group_by(platform_group) %>% summarize(platform_count = n()) %>%
    arrange(desc(platform_count))

top_platform$platform_group <- factor(top_platform$platform_group, levels = top_platform$platform_group

colourCount = length(unique(top_platform$platform_group))
fill_purple <- colorRampPalette(brewer.pal(9, "PuRd"))

top_platform %>% filter(platform_group != "NA") %>% ggplot(aes(x = platform_group,
    y = platform_count, fill = platform_group)) + geom_bar(stat = "identity") +
    coord_flip() + geom_text(aes(label = platform_count), size = 2.5, color = "black",
    hjust = -0.5) + labs(x = "Platform", y = "Number of Platforms", title = "Video Game Platforms for 19
    theme(legend.position = "none", plot.title = element_text(hjust = 0.5)) +
    ylim(0, max(top_platform$platform_count + 100)) + scale_fill_manual(values = fill_purple(colourCount
```



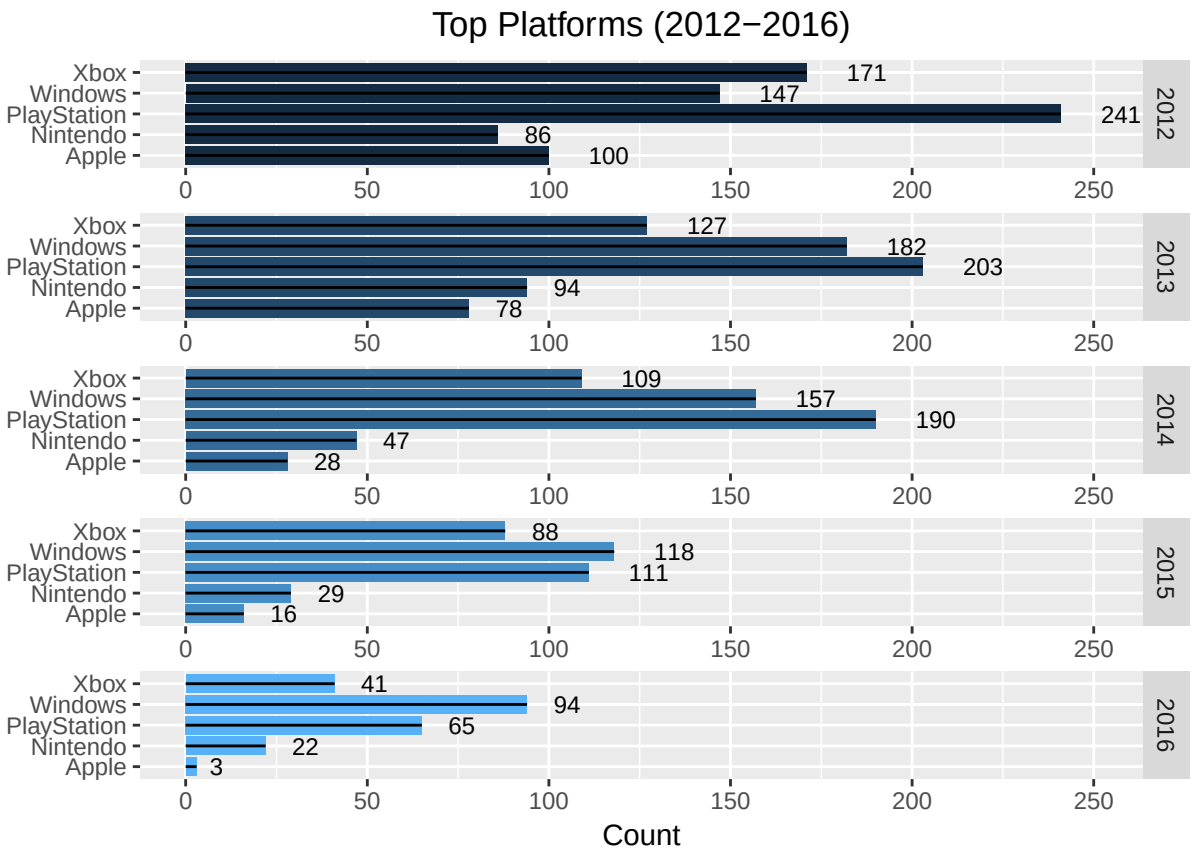
While some of these platforms have had multiple versions/evolutions over the last 20 years, Nintendo for example, they have been grouped together for a simpler analysis on the major gaming platforms. Even though Playstation was released after Nintendo, Playstation is still the top system followed by Nintendo, Windows, and Xbox. Given the previous graph consisting of all 20 years worth of data, I put together a

graph showing the last 5 years of data to see if the top systems held consistent.

Top Platforms for Past 5 Years

```
platform_years <- IGN_data_cleaned %>% filter(release_year >= 2012) %>% group_by(release_year,
    platform_group) %>% summarize(sum_platform_count = n()) %>% arrange(desc(release_year),
    desc(sum_platform_count)) %>% top_n(5, sum_platform_count)

platform_years %>% ggplot(aes(x = platform_group, y = sum_platform_count, fill = release_year)) +
    geom_bar(stat = "identity") + facet_wrap(~release_year, nrow = 5) + coord_flip() +
    geom_segment(aes(x = platform_group, xend = platform_group, y = 0, yend = sum_platform_count)) +
    labs(title = "Top Platforms (2012-2016)", x = "", y = "Count") + geom_text(aes(label = sum_platform_
    hjust = -1, size = 3) + theme(legend.position = "none", plot.title = element_text(hjust = 0.5)) +
    facet_wrap(~release_year, nrow = 5, scales = "free", strip.position = "right") +
    ylim(0, max(platform_years$sum_platform_count + 10))
```



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 spot in 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.

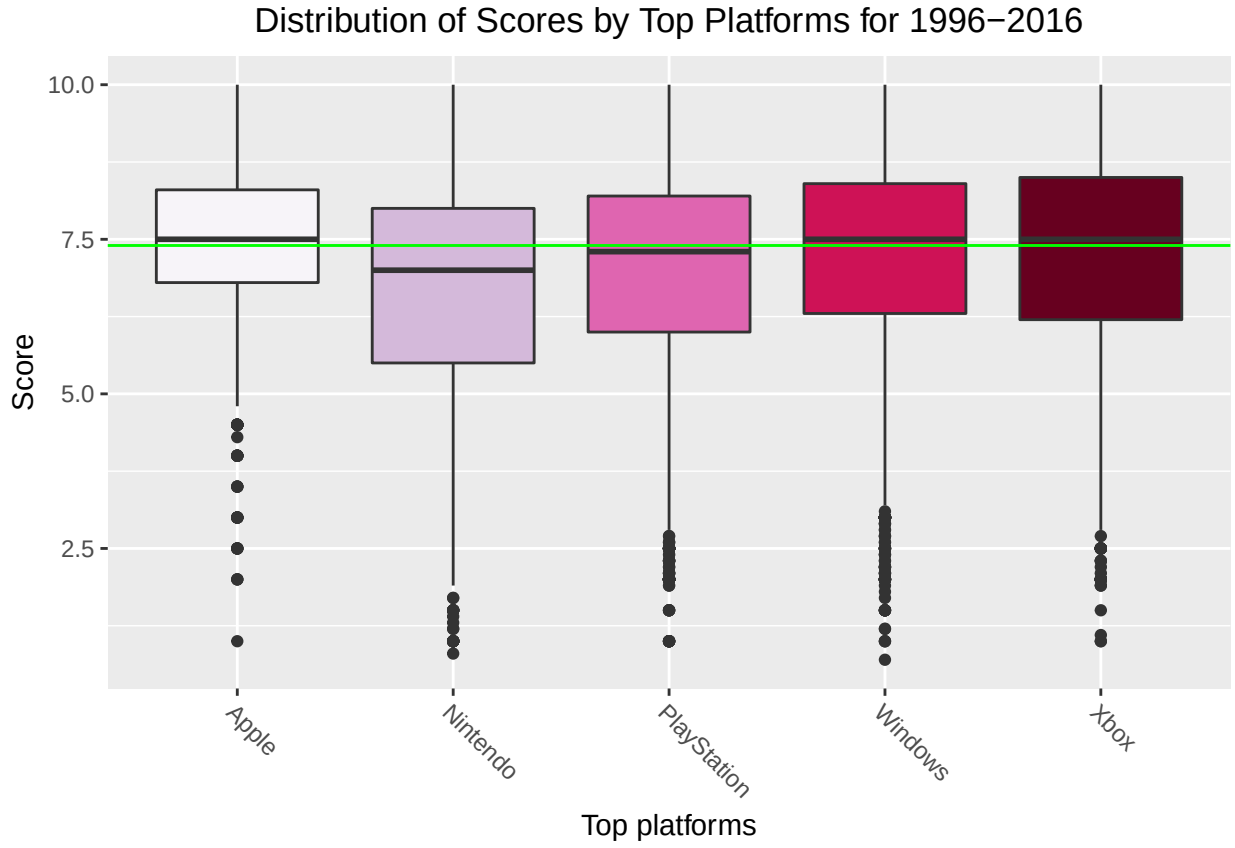
```
top_platform_scores <- IGN_data_cleaned %>% group_by(platform_group) %>% summarize(sum_platforms_count =
    arrange(desc(sum_platforms_count)) %>% top_n(5, sum_platforms_count)
```

```
top_platform_scores <- top_platform_scores[, 1]

platform_scores_df <- IGN_data_cleaned[IGN_data_cleaned$platform_group %in%
    top_platform_scores$platform_group, ]

colourCount = length(unique(top_platform_scores$platform_group))
fill_purple <- colorRampPalette(brewer.pal(9, "PuRd"))

platform_scores_df %>% ggplot(aes(x = platform_group, y = score, fill = platform_group)) +
    geom_boxplot(alpha = 1) + labs(x = "Top platforms", y = "Score", title = "Distribution of Scores by
    theme(legend.position = "none", plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = -45, hjust = 0)) + geom_hline(aes(yintercept = median(genre_
    color = "green") + scale_fill_manual(values = fill_purple(colourCount))
```



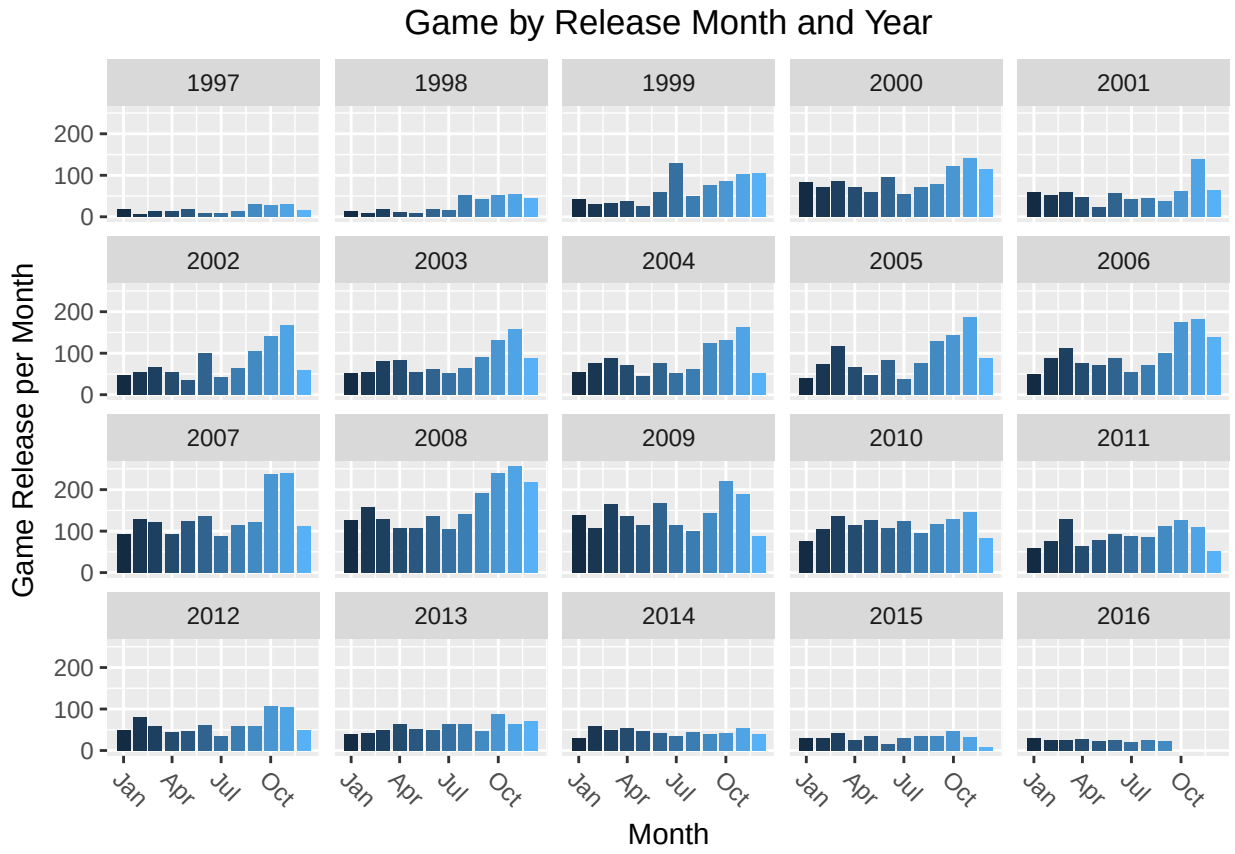
Given the number of game releases that Playstation has had, there was some expectation for a better score performance. However, this doesn't appear to be the case. Both Xbox and Windows have higher scores for the 25th, 50th, and 75th IQR values. Windows has the top number of released games for 2015 and 2016 so the higher scores are not completely a surprise.

Important Dates

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

```
Mon_Yr_Ct <- IGN_data_cleaned %>% filter(release_year > 1996) %>% group_by(release_month,
    release_year) %>% summarize(games_per_mon = n()) %>% arrange(desc(games_per_mon))
```

```
Mon_Yr_Ct %>% mutate(release_month_2 = as.Date(paste0(release_month, "-01"),
    "%m-%d")) %>% ggplot(aes(x = release_month_2, y = games_per_mon, fill = release_month_2)) +
    geom_bar(stat = "identity") + facet_wrap(~release_year, ncol = 5) + labs(title = "Game by Release M
    y = "Game Release per Month", x = "Month") + theme(legend.position = "none",
    plot.title = element_text(hjust = 0.5), axis.text.x = element_text(angle = -45,
        hjust = 0)) + scale_x_date(date_labels = "%b")
```



From 1997- 2008, the number of games released largely increases with 2008 having the most reviewed and released in video game history. Not so much in the more recent years with more consistent releases over the year, in the prior years, there is the larger amount of released video games in Octover and November over other months. The decreased number of games released could be due to video games becoming significantly more intricate and graphics intensive. With the capailities of T.V. and graphics cards now, immerisive video game graphics seem like a must-have for a game to be released.

Conclusion

After looking at platform, genre, and release date in a myriad of ways, Action can be conclusively named the top pick in Genre for types of video games released. When looking at Genre by score, RPG is the favorite when it comes to review score. When looking at the data by Platform, the answer is not as clear. While Playstation dominated for all years combined, even though it was released years after Nintendo or Windows, the exhibit by score does not show Playstation to have the highest rated games. This would imply that while a platform can release a large amount of games, it does not necessarily mean that those games will be favorably reviewed. To truly determine what platform reigns supreme, if there is one, I would look into tying video game sale data with the video game reviews to look at sale volume versus platform, genre, date, and

score.