IGN Video Game Reviews

Data Story

by Ellen A. Savoye

April 21, 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.

The Data

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

Table 1: IGN Data

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

X1	score_phrase	title	url
0	Amazing	LittleBigPlanet PS Vita	/games/littlebigplanet-vita/vita-98907
1	Amazing	LittleBigPlanet PS Vita – Marvel Super Hero Edition	/games/littlebigplanet-ps-vita-marvel-super-h
2	Great	Splice: Tree of Life	/games/splice/ipad-141070
3	Great	NHL 13	/games/nhl-13/xbox-360-128182
4	Great	NHL 13	/games/nhl-13/ps3-128181

The Approach

In brief, outline your approach to solving this problem (knowing that this might change later)

My approach can be broken out into the following sections:

1. Data wrangling and cleaning
   - Take into account any missing values or outliers
   - Are there any discontinued consoles? How should they be accounted for?
2. Exploratory Data Analysis
   - Use a combination of inferential statistics and data visualization to identify trends between game consoles and video game ratings
     - Determine potentially significant variables
     - Identify trends and correlations between variables
3. Machine Learning
   - Dig deeper into the correlation between consoles and game ratings using classification techniques and/or regression models
4. Data visualization and report out
   - Compile all relevant information into deliverables (listed below)

Deliverables

What are your deliverables?

My deliverables will consist of a report on my findings, a slide deck, and the corresponding R code used in analyzing the data. In addition, the aforementioned deliverables will be submitted and published on GitHub.

To take a proper look at the data, I loaded the original dataset as a CSV file and the necessary libraries.

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

```
## # A tibble: 5 x 11
##       X score_phrase title    url      platform score genre editors_choice
##   <int> <fct>        <fct>    <fct>    <fct>    <dbl> <fct> <fct>
## 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>
```

Of the variables available for use, score_phrase, platform, score, genre, editors_choice, release_year, release_month, and release_day are the ones I am using 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.

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

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

```
##   release_year
## 1         1970
## 2         1996
## 3         1997
## 4         2012
## 5         2013
```

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

```
##       X score_phrase                                              title
## 517 516        Great The Walking Dead: The Game -- Episode 1: A New Day
##                                                                  url platform
## 517 /games/the-walking-dead-season-1-episode-1/xbox-360-135866 Xbox 360
##     score     genre editors_choice release_year release_month release_day
## 517   8.5 Adventure              N         1970             1           1
```

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

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

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

```
group_by(IGN_data[IGN_data$genre == "", ])
```

```
## # A tibble: 36 x 12
##        X score_phrase title    url     platform score genre editors_choice
##  * <int> <fct>        <fct>    <fct>   <fct>    <dbl> <fct> <fct>
##  1    12 Good         Wild Bl~ /games~ iPhone    7.00 ""    N
##  2   113 Good         Retro/G~ /games~ PlaySta~  7.00 ""    N
##  3   160 Good         10000000 /games~ iPhone    7.50 ""    N
##  4   176 Okay         Colour ~ /games~ PC        6.20 ""    N
##  5  9375 Great        Duke Nu~ /games~ Wireless  8.00 ""    Y
##  6  9488 Okay         Rengoku  /games~ Wireless  6.50 ""    N
##  7  9767 Good         Super S~ /games~ Wireless  7.50 ""    N
##  8  9774 Amazing      Critter~ /games~ Wireless  9.00 ""    Y
##  9 10494 Awful        Clue / ~ /games~ Nintend~  3.50 ""    N
## 10 11367 Painful      Jeep Th~ /games~ PlaySta~  2.00 ""    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 %>% distinct(genre) %>%
  arrange(genre)

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

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

```
unique(IGN_data$genre_group)
```

```
##  [1] Platformer   Puzzle       Sports       Strategy     Fighting
##  [6] RPG          Other        Action       Adventure    Shooter
## [11] Music        Racing       Simulation   Education    Wrestling
## [16] Productivity Cards        Compilation  Flight       Pinball
## [21] 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 later in the course.

```
write.csv(IGN_data, "ign_clean.csv")
```

After the original dataset was compiled, I corrected the singular outlier. In addition, platorm and genre variables were pared down into a more manageable number of vari

Given the diversity, I would like to determine if, by year and over the years, there is a particular console that is the most popular by use and by average game rating. In addition, I would like to determine if, by year and

over the years within reason, there is a particular video game genre that is heralded as the favorite by rating and amount of games within that genre with a positive rating.