IGN Data: Data Wrangling

by Ellen A. Savoye

The IGN dataset sourced from Kaggle (https://www.kaggle.com/egrinstein/20-years-of-games) and IGN (http://ign.com/games/reviews), via a crawl, consists of 20 years worth of video-game data. To take a proper look at the data, I loaded the original dataset as a CSV file and the necessary libraries.

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

```
## # A tibble: 5 x 11
##       X score_phrase                                                title
##   <int>       <fctr>                                               <fctr>
## 1     0      Amazing                              LittleBigPlanet PS Vita
## 2     1      Amazing LittleBigPlanet PS Vita -- Marvel Super Hero Edition
## 3     2        Great                                 Splice: Tree of Life
## 4     3        Great                                               NHL 13
## 5     4        Great                                               NHL 13
## # ... with 8 more variables: url <fctr>, platform <fctr>, score <dbl>,
## #   genre <fctr>, editors_choice <fctr>, 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.

```r
# 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
```

```r
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
```

```r
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 <- 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)`
##                 <fctr>         <fctr>                        <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
##  * <int>       <fctr>                                     <fctr>
##  1    12         Good                                 Wild Blood
##  2   113         Good                                Retro/Grade
##  3   160         Good                                        10000000
##  4   176         Okay                                     Colour Bind
##  5  9375        Great                                Duke Nukem Arena
##  6  9488         Okay                                         Rengoku
##  7  9767         Good                                  Super Sketcher
##  8  9774      Amazing                                  Critter Crunch
##  9 10494        Awful Clue / Mouse Trap / Perfection / Aggravation
## 10 11367      Painful                                    Jeep Thrills
## # ... with 26 more rows, and 9 more variables: url <fctr>,
## #   platform <fctr>, score <dbl>, genre <fctr>, editors_choice <fctr>,
## #   release_year <int>, release_month <int>, release_day <int>,
## #   platform_group <fctr>
```

```
# 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")
```