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.

tbl\_df(IGN\_data)

## # A tibble: 18,625 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  
## 6 5 Good Total War Battles: Shogun  
## 7 6 Awful Double Dragon: Neon  
## 8 7 Amazing Guild Wars 2  
## 9 8 Awful Double Dragon: Neon  
## 10 9 Good Total War Battles: Shogun  
## # ... with 18,615 more rows, and 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.

# Release year is supposed to be higher than 1995  
  
IGN\_data %>% distinct(release\_year) %>%   
 arrange(release\_year)  
  
IGN\_data[IGN\_data$release\_year == "1970", ]  
  
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))

## release\_year  
## 1 1970  
## 2 1996  
## 3 1997  
## 4 1998  
## 5 1999  
## 6 2000  
## 7 2001  
## 8 2002  
## 9 2003  
## 10 2004  
## 11 2005  
## 12 2006  
## 13 2007  
## 14 2008  
## 15 2009  
## 16 2010  
## 17 2011  
## 18 2012  
## 19 2013  
## 20 2014  
## 21 2015  
## 22 2016  
## 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

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. 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  
## 1 Android  
## 2 Arcade  
## 3 Atari 2600  
## 4 Atari 5200  
## 5 Commodore 64/128  
## 6 Dreamcast  
## 7 Dreamcast VMU  
## 8 DVD / HD Video Game  
## 9 Game Boy  
## 10 Game Boy Advance  
## 11 Game Boy Color  
## 12 Game.Com  
## 13 GameCube  
## 14 Genesis  
## 15 iPad  
## 16 iPhone  
## 17 iPod  
## 18 Linux  
## 19 Lynx  
## 20 Macintosh  
## 21 Master System  
## 22 N-Gage  
## 23 NeoGeo  
## 24 NeoGeo Pocket Color  
## 25 NES  
## 26 New Nintendo 3DS  
## 27 Nintendo 3DS  
## 28 Nintendo 64  
## 29 Nintendo 64DD  
## 30 Nintendo DS  
## 31 Nintendo DSi  
## 32 Ouya  
## 33 PC  
## 34 PlayStation  
## 35 PlayStation 2  
## 36 PlayStation 3  
## 37 PlayStation 4  
## 38 PlayStation Portable  
## 39 PlayStation Vita  
## 40 Pocket PC  
## 41 Saturn  
## 42 Sega 32X  
## 43 Sega CD  
## 44 SteamOS  
## 45 Super NES  
## 46 TurboGrafx-16  
## 47 TurboGrafx-CD  
## 48 Vectrex  
## 49 Web Games  
## 50 Wii  
## 51 Wii U  
## 52 Windows Phone  
## 53 Windows Surface  
## 54 Wireless  
## 55 WonderSwan  
## 56 WonderSwan Color  
## 57 Xbox  
## 58 Xbox 360  
## 59 Xbox One

IGN\_data$platform\_group <- ifelse(grepl("PlayStation", IGN\_data$platform, ignore.case = TRUE), "PlayStation",   
 ifelse(grepl("Game Boy", IGN\_data$platform, ignore.case = TRUE), "Game Boy",  
 ifelse(grepl("Sega", IGN\_data$platform, ignore.case = TRUE) | grepl("Genesis", IGN\_data$platform, ignore.case = TRUE) | grepl("Dreamcast", IGN\_data$platform, ignore.case = TRUE) | grepl("Master System", IGN\_data$platform, ignore.case = TRUE) | grepl("Saturn", IGN\_data$platform, ignore.case = TRUE), "Sega",  
 ifelse(grepl("Nintendo", IGN\_data$platform, ignore.case = TRUE) | grepl("NES", IGN\_data$platform, ignore.case = TRUE) | grepl("Wii", IGN\_data$platform, ignore.case = TRUE) | grepl("GameCube", IGN\_data$platform, ignore.case = TRUE), "Nintendo",  
 ifelse(grepl("Xbox", IGN\_data$platform, ignore.case = TRUE), "Xbox",  
 ifelse(grepl("Iphone", IGN\_data$platform, ignore.case = TRUE) | grepl("Ipod", IGN\_data$platform, ignore.case = TRUE) | grepl("Ipad", IGN\_data$platform, ignore.case = TRUE) | grepl("Macintosh", IGN\_data$platform, ignore.case = TRUE), "Apple",  
 ifelse(grepl("Atari", IGN\_data$platform, ignore.case = TRUE) | grepl("Lynx", IGN\_data$platform, ignore.case = TRUE), "Atari",  
 ifelse(grepl("Windows", IGN\_data$platform, ignore.case = TRUE) | grepl("PC", IGN\_data$platform, ignore.case = TRUE), "Windows",  
 ifelse(grepl("NeoGeo", IGN\_data$platform, ignore.case = TRUE), "NeoGeo",  
 ifelse(grepl("Ouya", IGN\_data$platform, ignore.case = TRUE) | grepl("Android", IGN\_data$platform, ignore.case = TRUE), "Android",  
 ifelse(grepl("SteamOS", IGN\_data$platform, ignore.case = TRUE) | grepl("Linux", IGN\_data$platform, ignore.case = TRUE), "Linux",  
 ifelse(grepl("TurboGrafx", IGN\_data$platform, ignore.case = TRUE), "TurboGrafx",  
 ifelse(grepl("WonderSwan", IGN\_data$platform, ignore.case = TRUE), "WonderSwan", "Other")))))))))))))  
  
IGN\_data %>% group\_by(platform, platform\_group) %>%   
 summarise(n\_distinct(platform))

## # A tibble: 59 x 3  
## # Groups: platform [?]  
## platform platform\_group `n\_distinct(platform)`  
## <fctr> <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 to group genre vectors based on an overall description (e.g., Sports, Cards, Action, etc.) to have a cleaner code visual given the numerous distinct factors. Using those vectors, I created a new variable named genre\_group which brought the number of unique genres from 113 to 21.

# 113 unique factors in genre column   
  
IGN\_data %>% distinct(genre) %>%   
 arrange(genre)

## genre  
## 1   
## 2 Action  
## 3 Action, Adventure  
## 4 Action, Compilation  
## 5 Action, Editor  
## 6 Action, Platformer  
## 7 Action, Puzzle  
## 8 Action, RPG  
## 9 Action, Simulation  
## 10 Action, Strategy  
## 11 Adult, Card  
## 12 Adventure  
## 13 Adventure, Adult  
## 14 Adventure, Adventure  
## 15 Adventure, Compilation  
## 16 Adventure, Episodic  
## 17 Adventure, Platformer  
## 18 Adventure, RPG  
## 19 Baseball  
## 20 Battle  
## 21 Board  
## 22 Board, Compilation  
## 23 Card  
## 24 Card, Battle  
## 25 Card, Compilation  
## 26 Card, RPG  
## 27 Casino  
## 28 Compilation  
## 29 Compilation, Compilation  
## 30 Compilation, RPG  
## 31 Educational  
## 32 Educational, Action  
## 33 Educational, Adventure  
## 34 Educational, Card  
## 35 Educational, Productivity  
## 36 Educational, Puzzle  
## 37 Educational, Simulation  
## 38 Educational, Trivia  
## 39 Fighting  
## 40 Fighting, Action  
## 41 Fighting, Adventure  
## 42 Fighting, Compilation  
## 43 Fighting, RPG  
## 44 Fighting, Simulation  
## 45 Flight  
## 46 Flight, Action  
## 47 Flight, Racing  
## 48 Flight, Simulation  
## 49 Hardware  
## 50 Hunting  
## 51 Hunting, Action  
## 52 Hunting, Simulation  
## 53 Music  
## 54 Music, Action  
## 55 Music, Adventure  
## 56 Music, Compilation  
## 57 Music, Editor  
## 58 Music, RPG  
## 59 Other  
## 60 Other, Action  
## 61 Other, Adventure  
## 62 Party  
## 63 Pinball  
## 64 Pinball, Compilation  
## 65 Platformer  
## 66 Platformer, Action  
## 67 Platformer, Adventure  
## 68 Productivity  
## 69 Productivity, Action  
## 70 Puzzle  
## 71 Puzzle, Action  
## 72 Puzzle, Adventure  
## 73 Puzzle, Compilation  
## 74 Puzzle, Platformer  
## 75 Puzzle, RPG  
## 76 Puzzle, Word Game  
## 77 Racing  
## 78 Racing, Action  
## 79 Racing, Compilation  
## 80 Racing, Editor  
## 81 Racing, Shooter  
## 82 Racing, Simulation  
## 83 RPG  
## 84 RPG, Action  
## 85 RPG, Compilation  
## 86 RPG, Editor  
## 87 RPG, Simulation  
## 88 Shooter  
## 89 Shooter, Adventure  
## 90 Shooter, First-Person  
## 91 Shooter, Platformer  
## 92 Shooter, RPG  
## 93 Simulation  
## 94 Simulation, Adventure  
## 95 Sports  
## 96 Sports, Action  
## 97 Sports, Baseball  
## 98 Sports, Compilation  
## 99 Sports, Editor  
## 100 Sports, Fighting  
## 101 Sports, Golf  
## 102 Sports, Other  
## 103 Sports, Party  
## 104 Sports, Racing  
## 105 Sports, Simulation  
## 106 Strategy  
## 107 Strategy, Compilation  
## 108 Strategy, RPG  
## 109 Strategy, Simulation  
## 110 Trivia  
## 111 Virtual Pet  
## 112 Wrestling  
## 113 Wrestling, Simulation

Action <- c("Action", "Action, Adventure", "Action, Compilation", "Action, Editor", "Action, Platformer", "Action, Puzzle", "Action, RPG", "Action, Simulation", "Action, Strategy")  
  
Adventure <- c("Adventure", "Adventure, Adult", "Adventure, Adventure", "Adventure, Compilation", "Adventure, Episodic", "Adventure, Platformer", "Adventure, RPG")  
  
Sports <- c("Baseball", "Sports", "Sports, Action", "Sports, Baseball", "Sports, Compilation", "Sports, Editor", "Sports, Fighting", "Sports, Golf", "Sports, Other", "Sports, Party", "Sports, Racing", "Sports, Simulation")  
  
Cards <- c("Card", "Card, Battle", "Card, Compilation", "Card, RPG")  
  
Compilation <- c("Compilation", "Compilation, Compilation", "Compilation, RPG")  
  
Education <- c("Educational", "Educational, Action", "Educational, Adventure", "Educational, Card", "Educational, Productivity", "Educational, Puzzle", "Educational, Simulation", "Educational, Trivia")  
  
Fighting <- c("Fighting", "Fighting, Action", "Fighting, Adventure", "Fighting, Compilation", "Fighting, RPG", "Fighting, Simulation")  
Flight <- c("Flight", "Flight, Action", "Flight, Racing", "Flight, Simulation")  
  
Hunting <- c("Hunting", "Hunting, Action", "Hunting, Simulation")  
  
Music <- c("Music", "Music, Action", "Music, Adventure", "Music, Compilation", "Music, Editor", "Music, RPG")  
  
Pinball <- c("Pinball", "Pinball, Compilation")  
  
Platformer <- c("Platformer", "Platformer, Action", "Platformer, Adventure")  
  
Productivity <- c("Productivity, Action", "Productivity")  
  
Puzzle <- c("Puzzle", "Puzzle, Action", "Puzzle, Adventure", "Puzzle, Compilation", "Puzzle, Platformer", "Puzzle, RPG", "Puzzle, Word Game")  
  
Racing <- c("Racing", "Racing, Action", "Racing, Compilation", "Racing, Editor", "Racing, Shooter", "Racing, Simulation")  
  
RPG <- c("RPG", "RPG, Action", "RPG, Compilation", "RPG, Editor", "RPG, Simulation")  
  
Shooter <- c("Shooter", "Shooter, Adventure", "Shooter, First-Person", "Shooter, Platformer", "Shooter, RPG")  
  
Simulation <- c("Simulation", "Simulation, Adventure")  
  
Strategy <- c("Strategy", "Strategy, Compilation", "Strategy, RPG", "Strategy, Simulation")  
  
Wrestling <- c("Wrestling", "Wrestling, Simulation")  
  
IGN\_data$genre\_group <- ifelse(grepl(paste(Action, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Action",  
 ifelse(grepl(paste(Adventure, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Adventure",  
 ifelse(grepl(paste(Sports, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Sports",  
 ifelse(grepl(paste(Cards, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Cards",  
 ifelse(grepl(paste(Compilation, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Compilation",  
 ifelse(grepl(paste(Education, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Education",  
 ifelse(grepl(paste(Fighting, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Fighting",  
 ifelse(grepl(paste(Flight, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Flight",  
 ifelse(grepl(paste(Hunting, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Hunting",  
 ifelse(grepl(paste(Music, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Music",  
 ifelse(grepl(paste(Pinball, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Pinball",  
 ifelse(grepl(paste(Platformer, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Platformer",  
 ifelse(grepl(paste(Productivity, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Productivity",  
 ifelse(grepl(paste(Puzzle, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Puzzle",  
 ifelse(grepl(paste(Racing, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Racing",  
 ifelse(grepl(paste(RPG, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "RPG",  
 ifelse(grepl(paste(Shooter, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Shooter",  
 ifelse(grepl(paste(Simulation, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Simulation",  
 ifelse(grepl(paste(Strategy, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Strategy",  
 ifelse(grepl(paste(Wrestling, collapse = "|"), IGN\_data$genre, ignore.case = FALSE), "Wrestling", "Other"))))))))))))))))))))  
   
unique(IGN\_data$genre\_group)

## [1] "Platformer" "Puzzle" "Sports" "Strategy"   
## [5] "Fighting" "RPG" "Other" "Action"   
## [9] "Adventure" "Shooter" "Music" "Racing"   
## [13] "Simulation" "Education" "Wrestling" "Productivity"  
## [17] "Cards" "Compilation" "Flight" "Pinball"   
## [21] "Hunting"

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")