ASSESSMENT 3 – CASE STUDY AND DATA ANALYSIS

library("tidyverse")

## Warning: package 'tidyverse' was built under R version 4.1.2

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.8  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.2

## Warning: package 'tibble' was built under R version 4.1.2

## Warning: package 'tidyr' was built under R version 4.1.2

## Warning: package 'readr' was built under R version 4.1.2

## Warning: package 'dplyr' was built under R version 4.1.2

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library("dataMeta")

## Warning: package 'dataMeta' was built under R version 4.1.3

library("caret")

## Warning: package 'caret' was built under R version 4.1.3

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library("skimr")

## Warning: package 'skimr' was built under R version 4.1.2

library("ggcorrplot")

## Warning: package 'ggcorrplot' was built under R version 4.1.2

library("tidymodels")

## Warning: package 'tidymodels' was built under R version 4.1.2

## -- Attaching packages -------------------------------------- tidymodels 0.1.4 --

## v broom 0.7.12 v rsample 0.1.1   
## v dials 0.1.0 v tune 0.2.0   
## v infer 1.0.0 v workflows 0.2.6   
## v modeldata 0.1.1 v workflowsets 0.2.1   
## v parsnip 0.2.1 v yardstick 0.0.9   
## v recipes 0.2.0

## Warning: package 'broom' was built under R version 4.1.2

## Warning: package 'dials' was built under R version 4.1.2

## Warning: package 'infer' was built under R version 4.1.2

## Warning: package 'modeldata' was built under R version 4.1.3

## Warning: package 'recipes' was built under R version 4.1.2

## Warning: package 'rsample' was built under R version 4.1.2

## Warning: package 'workflowsets' was built under R version 4.1.3

## Warning: package 'yardstick' was built under R version 4.1.2

## -- Conflicts ----------------------------------------- tidymodels\_conflicts() --  
## x scales::discard() masks purrr::discard()  
## x dplyr::filter() masks stats::filter()  
## x recipes::fixed() masks stringr::fixed()  
## x dplyr::lag() masks stats::lag()  
## x caret::lift() masks purrr::lift()  
## x yardstick::precision() masks caret::precision()  
## x yardstick::recall() masks caret::recall()  
## x yardstick::sensitivity() masks caret::sensitivity()  
## x yardstick::spec() masks readr::spec()  
## x yardstick::specificity() masks caret::specificity()  
## x recipes::step() masks stats::step()  
## \* Use suppressPackageStartupMessages() to eliminate package startup messages

library("vip")

## Warning: package 'vip' was built under R version 4.1.3

##   
## Attaching package: 'vip'

## The following object is masked from 'package:utils':  
##   
## vi

## Summary to CEO

Hello Mr Tuke,

Regarding the predicted sales of ‘The Fatal Empire’. With the information we have about it’s title length, number of platforms released on, specific platform of release, game genre, year of release, and the sales figures of Japan, Europe, and other countries, our random forest model predicts sales of 1.92 million copies.

Taking the data as it is, and assuming the historical data is still representative, we can be over 90% sure in the model’s accuracy. That being said, there are a few areas for improvement, but I’ll cc you in on the manager’s report were that will be discussed further.

Regards, Dylan

## Managers report

### Managers report

1. **Aim and Hypothesis:**

The purpose of this report is to outline the process undertaken to predict the North American sales figure of the PS4 release of ‘The Fatal Empire’ using data available online via kaggle that was originally sourced from www.vgchatz.com. The data provided included information regarding a games:

* Name,
* Year of release
* Platform of release
* Game genre
* Sales figures of North America,
* Sales figures of Japan
* Sales figures of Europe
* Sales figures for the total of all other countries.

From this data, two additional pieces of information were acquired, the title length, as well as the number of platforms a game had been released on as it was hypothesised that human behaviour relating to these two factors would be informative (Title length may influence a potential customers likely hood of picking up a game in store, and more platforms influences word-of-mouth advertising as well as the community available to a game).

Two types of predictive models were trialed, A Lasso regression and a random forest. Their accuracy was measured and function carefully examined.

1. **Data Cleaning**

In this phase:

Summary of what has been accomplished below:

* A duplicate was removed
* An approach to verify the Name and Publisher values was provided
* The Year variable was properly formatted and NA values addressed. 270 NA values were reduced to 133 by setting them equal to the average year for rows with the same title. Five of the remaining titles had a year in their title which made for a good approximation of its release year. The remaining values were given the mean year for rows with the same Platform. This was considered a better approximation on a case by case basis than simply setting them all equal to the global mean year.
* Some of the sales figures were representing weekly sales, these have been converted to annual figures where obvious. Of those, six were for US/American sales which would relate to the Other\_Sales value. As there is no way to easily multiply this column (because it contains multiple countries and not just America) by 52.14, they have been left as is with the knowledge that this will influence the accuracy of the model
* Two incorrect Platform values (as indicated in the Name column) were corrected

Issues Identified but not fixed:

* As mentioned above, the US weekly sales rows remain an issue
* I’ve not changed the names to reflect the corrections made (this remains a source of error for the title length variable)
* There is no simple way to merge data pertaining to the same game because there is no linking variable/ID. And even if there were, it would need to be done individually and manually as some values would be written over in the process.

1. **Exploratory Data Analysis (EDA)**

In this phase, the information most relevant to predicting North American sales was identified. After careful examination of all variables it was found that the title length, Platform, Year, Genre, all Sales figures, and the number of platforms a game had been released on were all considered informative for the prediction of North American sales.

These assertions were identified through simple bivariate analysis, principle component analysis, and parallel coordinate plots.. The parallel coordinate plot found that there was a relationship between platform and region; Nintendo platforms generally selling better in Japan and ‘Sony’ platforms generally selling better outside of Japan. A similar trend could be found for some publishers, but it’s likely just that some publishers produce games for Nintendo platforms/ Japanese audiences. The principle component analysis found that the numerical variables contained the most amount of variance in the data, but that is only saying that no specific category contained much variation on its own. The principle component analysis offered the best evidence at this stage that the additional variables would make good predictors.

Their importances were also confirmed after the fact through variable importance analyses.

1. **Data Processing**

The data was modified as required. In this case, that meant transforming sales figures (by log(x+1)) and normalizing all countable variables. Platform and genre were also treated in a manner that allowed a computer to interpret each category.

The data was then split into training and testing sets. This is done so that the model can be tested on data that was not intrinsically part of its production.

1. **Model Fitting and Model Evaluation**

The model fitting stage involves building four models and measuring their accuracy through out-of-bag error measures (this involves comparing

). A fifth model was created in an attempt to improve/ verify the variable importance measures obtained by the best model, but it provided no benefit in terms of prediction accuracy.

The models created were:

1. LASSO regression model - Initial variables
2. LASSO regression model - initial variables + (Title length, platforms available to game)
3. Random forest model - Initial variables
4. Random forest model - initial variables + (Title length, platforms available to game)
5. Random forest model - initial variables + (Title length, platforms available to game) + preprocessing steps to reduce categories and correlation

There are a few assumptions involved in the LASSO regression model that were verified in this section. However the relevance of the LASSO models was lacking. Of the above list, it was the fourth model that had the best accuracy as it was found to have the highest r-squared value and the lowest route mean squared

## Statisticians report

A detailed analysis of the data intended for a fellow statistician. The analysis should include any code, figures and tables appropriately captioned. This should thoroughly detail your analysis process such that it is independently reproducible by another statistician. [85 marks]

## Data Clean

### Part 1 - Import and skim

# Importing data  
vgsales <- read.csv("vgsales.csv")  
head(vgsales,6)

## Name Platform Year Genre Publisher NA\_Sales  
## 1 Wii Sports Wii 2006 Sports Nintendo 41.49  
## 2 Super Mario Bros. NES 1985 Platform Nintendo 29.08  
## 3 Mario Kart Wii Wii 2008 Racing Nintendo 15.85  
## 4 Wii Sports Resort Wii 2009 Sports Nintendo 15.75  
## 5 Pokemon Red/Pokemon Blue GB 1996 Role-Playing Nintendo 11.27  
## 6 Tetris GB 1989 Puzzle Nintendo 23.20  
## EU\_Sales JP\_Sales Other\_Sales  
## 1 29.02 3.77 8.46  
## 2 3.58 6.81 0.77  
## 3 12.88 3.79 3.31  
## 4 11.01 3.28 2.96  
## 5 8.89 10.22 1.00  
## 6 2.26 4.22 0.58

# Initial skim of Data  
skim(vgsales)

Data summary

|  |  |
| --- | --- |
| Name | vgsales |
| Number of rows | 16598 |
| Number of columns | 9 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 5 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | 0 | 1 | 1 | 132 | 0 | 11493 | 0 |
| Platform | 0 | 1 | 2 | 4 | 0 | 31 | 0 |
| Year | 0 | 1 | 3 | 4 | 0 | 40 | 0 |
| Genre | 0 | 1 | 4 | 12 | 0 | 12 | 0 |
| Publisher | 0 | 1 | 3 | 38 | 0 | 579 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| NA\_Sales | 0 | 1 | 0.26 | 0.82 | 0 | 0 | 0.08 | 0.24 | 41.49 | ▇▁▁▁▁ |
| EU\_Sales | 0 | 1 | 0.15 | 0.51 | 0 | 0 | 0.02 | 0.11 | 29.02 | ▇▁▁▁▁ |
| JP\_Sales | 0 | 1 | 0.08 | 0.31 | 0 | 0 | 0.00 | 0.04 | 10.22 | ▇▁▁▁▁ |
| Other\_Sales | 0 | 1 | 0.05 | 0.19 | 0 | 0 | 0.01 | 0.04 | 10.57 | ▇▁▁▁▁ |

# No missing values here

The data has nine columns, with 16,598 rows. There are currently five character variables and 4 numeric variables. The variables include the name of the game (Name), the platform the game can be played on (Platform), the year the game was released (Year), the genre of the game (Genre), the publisher of the game (Publisher), as well as the number of copies sold in North America, Europe, Japan, and the total for all remaining countries (NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, respectively). No missing values were made apparent in this step.

## Data Clean

### Part 2 - A Good Clean and Tidy

Summary of what has been accomplished below:

* A duplicate was removed
* An approach to verify the Name and Publisher values was provided
* The Year variable was properly formatted and NA values addressed. 270 NA values were reduced to 133 by setting them equal to the average year for rows with the same title. Five of the remaining titles had a year in their title which made for a good approximation of its release year. The remaining values were given the mean year for rows with the same Platform. This was considered a better approximation on a case by case basis than simply setting them all equal to the global mean year.
* A curiosity regarding Wii Sport was examined and validated
* Some of the sales figures were representing weekly sales, these have been converted to annual figures where obvious. Of those, six were for US/American sales which would relate to the Other\_Sales value. As we cannot simply multiply this column (because it contains multiple countries and not just America) by 52, they have been left as is with the knowledge that this will influence the accuracy of the model
* Two incorrect Platform values were corrected

Issues Identified but not fixed:

* As mentioned above, the US weekly sales rows remain an issue
* I’ve not changed the names to reflect the corrections made (because Name is not a predictor)
* I could not find a simple way to merge data pertaining to the same game because there is no linking variable/ID. And even if there were, I would still need to go through each manually as some values would be written over in the process

# Check for Duplicates  
filter(vgsales, duplicated(vgsales) == TRUE)

## Name Platform Year Genre Publisher NA\_Sales EU\_Sales  
## 1 Wii de Asobu: Metroid Prime Wii N/A Shooter Nintendo 0 0  
## JP\_Sales Other\_Sales  
## 1 0.02 0

filter(vgsales, Name == "Wii de Asobu: Metroid Prime")

## Name Platform Year Genre Publisher NA\_Sales EU\_Sales  
## 1 Wii de Asobu: Metroid Prime Wii N/A Shooter Nintendo 0 0  
## 2 Wii de Asobu: Metroid Prime Wii N/A Shooter Nintendo 0 0  
## JP\_Sales Other\_Sales  
## 1 0.02 0  
## 2 0.02 0

clean <- distinct(vgsales)  
# One duplicate removed  
  
# Check for other mistakes/inconsistencies:  
# Name check:  
name\_vgsales <- clean %>%  
 group\_by(Name) %>%  
 tally(sort=TRUE) # currently 11,492 Game titles  
# Need to compare values against a Master list  
  
# Platform check:  
platform\_vgsales <- clean %>%  
 group\_by(Platform) %>%  
 tally(sort=TRUE) # 31 Platforms   
head(platform\_vgsales,5)

## # A tibble: 5 x 2  
## Platform n  
## <chr> <int>  
## 1 DS 2163  
## 2 PS2 2161  
## 3 PS3 1329  
## 4 Wii 1324  
## 5 X360 1265

# Clear  
  
# Publisher check  
publishers\_vgsales <- clean %>%  
 group\_by(Publisher) %>%  
 tally(sort=TRUE) # 579 Publishers  
head(publishers\_vgsales,5)

## # A tibble: 5 x 2  
## Publisher n  
## <chr> <int>  
## 1 Electronic Arts 1351  
## 2 Activision 975  
## 3 Namco Bandai Games 932  
## 4 Ubisoft 921  
## 5 Konami Digital Entertainment 832

# Need to compare values against Master list  
  
 # Check 579 publishers against list found on wiki and at: https://www.kaggle.com/datasets/andreshg/videogamescompaniesregions?resource=download) for quick spell-check I create a Master list to compare Publisher values against:   
dev1 <- read.csv("dev1.csv")  
dev2 <- read.csv("dev2.csv")  
wiki1 <- read\_csv("table-2.csv")

## Warning: One or more parsing issues, see `problems()` for details

## Rows: 865 Columns: 5  
## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (4): Publisher, Location, Notable games published, Notes  
## dbl (1): Est.  
##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

dev1\_publishers <- dev1$Developer  
dev2\_publishers <- dev2$Developer  
wiki\_publishers <- wiki1$Publisher  
publisher\_check <- c(dev1\_publishers, dev2\_publishers, wiki\_publishers)  
publisher\_check <- unique(publisher\_check)  
length(publisher\_check) # 1535 Publishers to check against

## [1] 1535

publisher\_check\_df <- data.frame(matrix(unlist(publisher\_check),  
 nrow=length(publisher\_check),  
 byrow=TRUE))  
 # Comparing lists  
publishers\_vgsales$Publisher %in% publisher\_check\_df$matrix.unlist.publisher\_check...nrow...length.publisher\_check...

## [1] TRUE TRUE FALSE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [13] FALSE TRUE TRUE TRUE FALSE FALSE TRUE TRUE FALSE TRUE FALSE TRUE  
## [25] TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [37] TRUE TRUE TRUE FALSE TRUE TRUE FALSE FALSE TRUE FALSE FALSE TRUE  
## [49] FALSE TRUE TRUE TRUE FALSE FALSE TRUE TRUE FALSE FALSE TRUE TRUE  
## [61] FALSE TRUE FALSE FALSE TRUE TRUE FALSE FALSE FALSE TRUE TRUE FALSE  
## [73] TRUE TRUE FALSE FALSE TRUE FALSE FALSE TRUE TRUE FALSE TRUE FALSE  
## [85] FALSE TRUE TRUE FALSE TRUE FALSE FALSE TRUE TRUE FALSE FALSE FALSE  
## [97] FALSE TRUE FALSE TRUE TRUE FALSE FALSE FALSE FALSE FALSE TRUE TRUE  
## [109] TRUE FALSE TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE TRUE TRUE  
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE  
## [133] TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE TRUE  
## [145] FALSE TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE  
## [157] FALSE TRUE FALSE TRUE FALSE TRUE TRUE FALSE FALSE FALSE FALSE TRUE  
## [169] FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE  
## [181] FALSE TRUE FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE TRUE TRUE  
## [193] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE  
## [205] FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE  
## [217] TRUE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE TRUE  
## [229] FALSE TRUE FALSE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## [241] FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## [253] FALSE TRUE TRUE TRUE FALSE FALSE TRUE FALSE TRUE FALSE FALSE FALSE  
## [265] TRUE FALSE TRUE TRUE TRUE FALSE TRUE FALSE TRUE FALSE FALSE FALSE  
## [277] FALSE TRUE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE TRUE  
## [289] FALSE TRUE FALSE FALSE TRUE TRUE FALSE TRUE FALSE FALSE TRUE FALSE  
## [301] FALSE FALSE TRUE TRUE FALSE FALSE TRUE TRUE FALSE TRUE FALSE FALSE  
## [313] FALSE TRUE FALSE FALSE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE  
## [325] TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE FALSE TRUE TRUE TRUE  
## [337] TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE TRUE FALSE FALSE  
## [349] TRUE FALSE FALSE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## [361] FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE  
## [373] TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE  
## [385] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  
## [397] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE  
## [409] FALSE FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE  
## [421] FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [433] TRUE TRUE TRUE TRUE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE  
## [445] TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE  
## [457] FALSE FALSE FALSE FALSE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE  
## [469] FALSE FALSE FALSE TRUE FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## [481] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE  
## [493] FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE TRUE TRUE FALSE  
## [505] FALSE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE  
## [517] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE  
## [529] FALSE TRUE FALSE FALSE FALSE TRUE FALSE TRUE TRUE FALSE TRUE FALSE  
## [541] FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE TRUE FALSE  
## [553] TRUE FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE FALSE TRUE TRUE  
## [565] FALSE FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE  
## [577] FALSE FALSE FALSE

# With more resources you could verify from a 'master list', mine is incomplete and I have no way of knowing if it has errors itself. But its a proof of concept as to how you could go about checking the Qualitative variables.  
  
# Year Check:  
unique(clean$Year)

## [1] "2006" "1985" "2008" "2009" "1996" "1989" "1984" "2005" "1999" "2007"  
## [11] "2010" "2013" "2004" "1990" "1988" "2002" "2001" "2011" "1998" "2015"  
## [21] "2012" "2014" "1992" "1997" "1993" "1994" "1982" "2003" "1986" "2000"  
## [31] "N/A" "1995" "2016" "1991" "1981" "1987" "1980" "1983" "2020" "2017"

clean <- clean %>%  
 mutate(Year = na\_if(Year, "N/A"), Year = na\_if(Year, ""), Year = na\_if(Year, " ") )  
clean$Year[is.nan(clean$Year)]<-NA  
sum(is.na(clean$Year))

## [1] 270

clean$Year <- as.numeric(clean$Year)  
# Has 270 NA values and is now numeric  
  
# Could just give the average if a game is released on multiple platforms   
na\_rows <- filter(clean, is.na(Year))  
na\_titles <- c(na\_rows$Name)  
# If a game is released on multiple platforms, let NA equal the average of the year   
for(i in na\_titles){  
 if(nrow(filter(clean, grepl(i, Name))) > 1){  
 clean[clean$Name == i, "Year"] <- round(mean(filter(clean, grepl(i, Name))[,3], na.rm = TRUE),0)  
 }  
}  
filter(clean, is.na(Year))

## Name Platform  
## 1 wwe Smackdown vs. Raw 2006 PS2  
## 2 Frogger's Adventures: Temple of the Frog GBA  
## 3 Triple Play 99 PS  
## 4 Legacy of Kain: Soul Reaver PS  
## 5 Donkey Kong Land III GB  
## 6 Air-Sea Battle 2600  
## 7 Suikoden III PS2  
## 8 Yakuza 4 PS3  
## 9 WarioWare: Twisted! GBA  
## 10 Test Drive Unlimited 2 X360  
## 11 Test Drive Unlimited 2 PS3  
## 12 Monster Hunter 2 PS2  
## 13 Advance Wars: Days of Ruin DS  
## 14 Fishing Derby 2600  
## 15 Circus Atari 2600  
## 16 The Chronicles of Riddick: Escape from Butcher Bay XB  
## 17 Maze Craze: A Game of Cops 'n Robbers 2600  
## 18 Silent Hill: Homecoming X360  
## 19 Robert Ludlum's The Bourne Conspiracy X360  
## 20 NHL Slapshot Wii  
## 21 TERA PC  
## 22 NFL GameDay 2003 PS2  
## 23 Silent Hill: Homecoming PS3  
## 24 Harvest Moon: Save the Homeland PS2  
## 25 Robert Ludlum's The Bourne Conspiracy PS3  
## 26 Hangman 2600  
## 27 Cubix Robots for Everyone: Clash 'n' Bash GBA  
## 28 Tomb Raider (2013) PC  
## 29 Dragon Ball Z: Budokai Tenkaichi 2 (JP sales) Wii  
## 30 Dragster 2600  
## 31 All-Star Baseball 2005 PS2  
## 32 Slot Machine 2600  
## 33 The Dukes of Hazzard II: Daisy Dukes It Out PS  
## 34 Disgaea 3: Absence of Detention PSV  
## 35 Haven: Call of the King PS2  
## 36 Unreal Championship 2: The Liandri Conflict XB  
## 37 The Legend of Zelda: The Minish Cap(weekly JP sales) GBA  
## 38 Indy 500 2600  
## 39 Flag Capture 2600  
## 40 DanceDanceRevolution II Wii  
## 41 Big Beach Sports 2 Wii  
## 42 Jet X20 PS2  
## 43 Tribes: Aerial Assault PS2  
## 44 Move Fitness PS3  
## 45 Yu Yu Hakusho: Dark Tournament PS2  
## 46 Ghostbusters II 2600  
## 47 Breakaway IV 2600  
## 48 Valkyria Chronicles III: Unrecorded Chronicles PSP  
## 49 Famista 64 N64  
## 50 Test Drive Unlimited 2 PC  
## 51 Pet Zombies 3DS  
## 52 Trauma Team Wii  
## 53 Backbreaker X360  
## 54 Twisted Metal: Small Brawl PS  
## 55 Otomedius Excellent X360  
## 56 Backbreaker PS3  
## 57 Sword of the Samurai PS2  
## 58 Alone in the Dark: The New Nightmare PS  
## 59 Vegas Party Wii  
## 60 Jurassic Park: The Game X360  
## 61 eJay Clubworld PS2  
## 62 All-Star Baseball 2005 XB  
## 63 Our House Party! Wii  
## 64 Disney's Cinderella: Magical Dreams GBA  
## 65 Nintendo Puzzle Collection GC  
## 66 Charm Girls Club: My Fashion Mall DS  
## 67 Super Robot Wars OG Saga: Masou Kishin II - Revelation of Evil God PSP  
## 68 Saru! Get You! Million Monkeys PS2  
## 69 The Daring Game for Girls DS  
## 70 GiFTPiA GC  
## 71 Atsumare! Power Pro Kun no DS Koushien DS  
## 72 My Healthy Cooking Coach DS  
## 73 Luminous Arc 2 (JP sales) DS  
## 74 The Daring Game for Girls Wii  
## 75 Egg Monster Hero DS  
## 76 Demon Chaos PS2  
## 77 Action Man-Operation Extreme PS  
## 78 Charm Girls Club: My Fashion Show DS  
## 79 Face Racers: Photo Finish 3DS  
## 80 Zero: Tsukihami no Kamen Wii  
## 81 Mega Man Battle Network: Operation Shooting Star DS  
## 82 Dream Trigger 3D 3DS  
## 83 McFarlane's Evil Prophecy PS2  
## 84 Drake of the 99 Dragons XB  
## 85 Port Royale 3 PS3  
## 86 Build-A-Bear Workshop: Friendship Valley Wii  
## 87 Alex Rider: Stormbreaker DS  
## 88 Yoostar on MTV X360  
## 89 Port Royale 3 X360  
## 90 Jewel Link Chronicles: Mountains of Madness DS  
## 91 Chou Soujuu Mecha MG DS  
## 92 Prinny: Can I Really Be The Hero? (US sales) PSP  
## 93 Combat Elite: WWII Paratroopers PS2  
## 94 Flip's Twisted World Wii  
## 95 Mobile Ops: The One Year War X360  
## 96 Tom Clancy's Rainbow Six: Critical Hour XB  
## 97 Reader Rabbit 2nd Grade Wii  
## 98 Mountain Bike Adrenaline PS2  
## 99 Tour de France 2011 X360  
## 100 Drill Dozer GBA  
## 101 B.L.U.E.: Legend of Water PS  
## 102 Sega Rally 2006 PS2  
## 103 Half-Minute Hero 2 PSP  
## 104 Clockwork Empires PC  
## 105 Housekeeping DS  
## 106 Sabre Wulf GBA  
## 107 Beyond the Labyrinth 3DS  
## 108 Bikkuriman Daijiten DS  
## 109 Fullmetal Alchemist: Brotherhood PSP  
## 110 Combat Elite: WWII Paratroopers XB  
## 111 Samurai Spirits: Tenkaichi Kenkakuden PS2  
## 112 Super Duper Sumos GBA  
## 113 Legacy of Ys: Books I & II DS  
## 114 The King of Fighters: Maximum Impact - Maniax XB  
## 115 Combat Wings: The Great Battles of WWII Wii  
## 116 Tube Slider GC  
## 117 Umineko no Naku Koro ni San: Shinjitsu to Gensou no Yasoukyoku PS3  
## 118 Wii de Asobu: Metroid Prime Wii  
## 119 Payout Poker & Casino PSP  
## 120 Steal Princess DS  
## 121 Yu-Gi-Oh! 5D's Wheelie Breakers (JP sales) Wii  
## 122 Writing and Speaking Beautiful Japanese DS DS  
## 123 Virtua Quest PS2  
## 124 Shonen Jump's Yu-Gi-Oh! GX Card Almanac DS  
## 125 Football Manager 2007 X360  
## 126 Ferrari: The Race Experience Wii  
## 127 Jet Impulse DS  
## 128 Dream Dancer DS  
## 129 Homeworld Remastered Collection PC  
## 130 Shorts DS  
## 131 Agarest Senki: Re-appearance PS3  
## 132 Hakuouki: Shinsengumi Kitan PS3  
## 133 Virtua Quest GC  
## Year Genre Publisher NA\_Sales EU\_Sales JP\_Sales  
## 1 NA Fighting N/A 1.57 1.02 0.00  
## 2 NA Adventure Konami Digital Entertainment 2.15 0.18 0.00  
## 3 NA Sports N/A 0.81 0.55 0.00  
## 4 NA Action Eidos Interactive 0.58 0.40 0.00  
## 5 NA Platform Nintendo 0.68 0.31 0.00  
## 6 NA Shooter Atari 0.91 0.06 0.00  
## 7 NA Role-Playing Unknown 0.29 0.23 0.38  
## 8 NA Action Sega 0.15 0.14 0.63  
## 9 NA Puzzle Unknown 0.16 0.06 0.50  
## 10 NaN Racing Atari 0.30 0.32 0.00  
## 11 NaN Racing Atari 0.16 0.35 0.01  
## 12 NA Role-Playing Capcom 0.00 0.00 0.63  
## 13 NA Strategy Nintendo 0.44 0.13 0.00  
## 14 NA Sports Activision 0.48 0.03 0.00  
## 15 NA Action Atari 0.43 0.03 0.00  
## 16 NA Shooter Vivendi Games 0.32 0.11 0.00  
## 17 NA Action Atari 0.42 0.02 0.00  
## 18 NaN Action Konami Digital Entertainment 0.25 0.16 0.00  
## 19 NaN Action Vivendi Games 0.26 0.14 0.00  
## 20 NA Sports Unknown 0.40 0.00 0.00  
## 21 NA Role-Playing Unknown 0.24 0.12 0.00  
## 22 NA Sports Unknown 0.20 0.15 0.00  
## 23 NaN Action Konami Digital Entertainment 0.18 0.15 0.00  
## 24 NA Simulation Unknown 0.19 0.15 0.00  
## 25 NaN Action Vivendi Games 0.18 0.14 0.00  
## 26 NA Puzzle Atari 0.35 0.02 0.00  
## 27 NA Action Unknown 0.26 0.10 0.00  
## 28 NA Action Square Enix 0.06 0.25 0.00  
## 29 NA Action N/A 0.15 0.05 0.14  
## 30 NA Racing Activision 0.30 0.02 0.00  
## 31 NaN Sports Unknown 0.16 0.12 0.00  
## 32 NA Action Atari 0.29 0.02 0.00  
## 33 NA Racing Unknown 0.17 0.11 0.00  
## 34 NA Role-Playing Nippon Ichi Software 0.13 0.06 0.07  
## 35 NA Platform Midway Games 0.14 0.11 0.00  
## 36 NA Shooter Midway Games 0.22 0.05 0.00  
## 37 NA Action N/A 0.00 0.00 0.27  
## 38 NA Racing Atari 0.26 0.01 0.00  
## 39 NA Action Atari 0.25 0.02 0.00  
## 40 NA Misc Konami Digital Entertainment 0.22 0.00 0.00  
## 41 NA Sports THQ 0.09 0.12 0.00  
## 42 NA Racing Unknown 0.11 0.09 0.00  
## 43 NA Shooter Unknown 0.11 0.09 0.00  
## 44 NA Sports Sony Computer Entertainment 0.00 0.17 0.00  
## 45 NA Fighting N/A 0.10 0.08 0.00  
## 46 NA Action Activision 0.20 0.01 0.00  
## 47 NA Puzzle Sears 0.20 0.01 0.00  
## 48 NA Strategy Sega 0.00 0.00 0.21  
## 49 NA Sports Namco Bandai Games 0.00 0.00 0.17  
## 50 NaN Racing Atari 0.05 0.11 0.00  
## 51 NA Simulation Majesco Entertainment 0.18 0.00 0.00  
## 52 NA Simulation Unknown 0.15 0.00 0.02  
## 53 NaN Sports Unknown 0.17 0.00 0.00  
## 54 NA Action Unknown 0.10 0.07 0.00  
## 55 NA Shooter Unknown 0.13 0.00 0.04  
## 56 NaN Sports Unknown 0.16 0.00 0.00  
## 57 NA Fighting Ubisoft 0.00 0.00 0.17  
## 58 NA Adventure Infogrames 0.09 0.06 0.00  
## 59 NA Misc Unknown 0.15 0.00 0.00  
## 60 NA Action Unknown 0.15 0.00 0.00  
## 61 NA Misc Empire Interactive 0.07 0.06 0.00  
## 62 NaN Sports Unknown 0.11 0.03 0.00  
## 63 NA Simulation Unknown 0.13 0.00 0.00  
## 64 NA Platform Disney Interactive Studios 0.10 0.04 0.00  
## 65 NA Puzzle Unknown 0.00 0.00 0.13  
## 66 NA Simulation Unknown 0.12 0.00 0.00  
## 67 NA Strategy N/A 0.00 0.00 0.12  
## 68 NA Platform Unknown 0.00 0.00 0.12  
## 69 NaN Adventure Unknown 0.11 0.00 0.00  
## 70 NA Role-Playing Unknown 0.00 0.00 0.11  
## 71 NA Sports N/A 0.00 0.00 0.10  
## 72 NA Simulation Unknown 0.09 0.00 0.00  
## 73 NA Role-Playing Unknown 0.00 0.00 0.10  
## 74 NaN Adventure Unknown 0.09 0.00 0.00  
## 75 NA Role-Playing Square Enix 0.00 0.00 0.09  
## 76 NA Action Konami Digital Entertainment 0.00 0.00 0.09  
## 77 NA Action N/A 0.05 0.03 0.00  
## 78 NA Simulation Unknown 0.08 0.00 0.00  
## 79 NA Racing Majesco Entertainment 0.08 0.00 0.00  
## 80 NA Action Nintendo 0.00 0.00 0.08  
## 81 NA Role-Playing Capcom 0.00 0.00 0.07  
## 82 NA Shooter D3Publisher 0.04 0.02 0.00  
## 83 NA Action Konami Digital Entertainment 0.03 0.03 0.00  
## 84 NA Shooter Unknown 0.05 0.01 0.00  
## 85 NaN Simulation Kalypso Media 0.06 0.00 0.00  
## 86 NA Misc Unknown 0.06 0.00 0.00  
## 87 NA Action THQ 0.06 0.00 0.00  
## 88 NA Misc Unknown 0.06 0.00 0.00  
## 89 NaN Simulation Kalypso Media 0.06 0.00 0.00  
## 90 NA Puzzle Avanquest 0.00 0.06 0.00  
## 91 NA Simulation N/A 0.00 0.00 0.06  
## 92 NA Action N/A 0.06 0.00 0.00  
## 93 NaN Shooter Unknown 0.03 0.02 0.00  
## 94 NA Platform Unknown 0.05 0.00 0.00  
## 95 NA Simulation Namco Bandai Games 0.00 0.00 0.06  
## 96 NA Shooter Unknown 0.04 0.01 0.00  
## 97 NA Misc Unknown 0.05 0.00 0.00  
## 98 NA Sports Unknown 0.03 0.02 0.00  
## 99 NA Racing Focus Home Interactive 0.00 0.05 0.00  
## 100 NA Platform Unknown 0.04 0.01 0.00  
## 101 NA Adventure N/A 0.00 0.00 0.04  
## 102 NA Racing Unknown 0.00 0.00 0.04  
## 103 NA Role-Playing Rising Star Games 0.00 0.00 0.04  
## 104 NA Strategy Unknown 0.00 0.03 0.00  
## 105 NA Action N/A 0.00 0.00 0.04  
## 106 NA Platform THQ 0.02 0.01 0.00  
## 107 NA Role-Playing Unknown 0.00 0.00 0.03  
## 108 NA Misc N/A 0.00 0.00 0.03  
## 109 NA Action Unknown 0.00 0.00 0.03  
## 110 NaN Shooter Unknown 0.02 0.01 0.00  
## 111 NA Fighting Unknown 0.00 0.00 0.03  
## 112 NA Action Unknown 0.02 0.01 0.00  
## 113 NA Role-Playing Unknown 0.02 0.00 0.00  
## 114 NA Fighting Unknown 0.02 0.01 0.00  
## 115 NA Simulation City Interactive 0.02 0.00 0.00  
## 116 NA Racing Unknown 0.02 0.00 0.00  
## 117 NA Adventure N/A 0.00 0.00 0.02  
## 118 NA Shooter Nintendo 0.00 0.00 0.02  
## 119 NA Misc Unknown 0.02 0.00 0.00  
## 120 NA Platform Unknown 0.02 0.00 0.00  
## 121 NA Racing Unknown 0.00 0.00 0.02  
## 122 NA Misc Unknown 0.00 0.00 0.02  
## 123 NaN Role-Playing Unknown 0.01 0.01 0.00  
## 124 NA Misc Konami Digital Entertainment 0.00 0.00 0.02  
## 125 NA Sports Sega 0.00 0.01 0.00  
## 126 NA Racing System 3 Arcade Software 0.00 0.01 0.00  
## 127 NA Simulation Nintendo 0.00 0.00 0.02  
## 128 NA Misc N/A 0.01 0.00 0.00  
## 129 NA Strategy N/A 0.00 0.01 0.00  
## 130 NA Platform Unknown 0.01 0.00 0.00  
## 131 NA Role-Playing Idea Factory 0.00 0.00 0.01  
## 132 NA Adventure Unknown 0.01 0.00 0.00  
## 133 NaN Role-Playing Unknown 0.01 0.00 0.00  
## Other\_Sales  
## 1 0.41  
## 2 0.07  
## 3 0.10  
## 4 0.07  
## 5 0.04  
## 6 0.01  
## 7 0.08  
## 8 0.05  
## 9 0.02  
## 10 0.07  
## 11 0.12  
## 12 0.00  
## 13 0.06  
## 14 0.01  
## 15 0.00  
## 16 0.02  
## 17 0.00  
## 18 0.04  
## 19 0.04  
## 20 0.02  
## 21 0.05  
## 22 0.05  
## 23 0.06  
## 24 0.05  
## 25 0.06  
## 26 0.00  
## 27 0.01  
## 28 0.05  
## 29 0.01  
## 30 0.00  
## 31 0.04  
## 32 0.00  
## 33 0.02  
## 34 0.04  
## 35 0.04  
## 36 0.01  
## 37 0.01  
## 38 0.00  
## 39 0.00  
## 40 0.02  
## 41 0.03  
## 42 0.03  
## 43 0.03  
## 44 0.05  
## 45 0.03  
## 46 0.00  
## 47 0.00  
## 48 0.00  
## 49 0.03  
## 50 0.03  
## 51 0.01  
## 52 0.01  
## 53 0.01  
## 54 0.01  
## 55 0.01  
## 56 0.01  
## 57 0.00  
## 58 0.01  
## 59 0.01  
## 60 0.01  
## 61 0.02  
## 62 0.01  
## 63 0.01  
## 64 0.00  
## 65 0.00  
## 66 0.01  
## 67 0.00  
## 68 0.00  
## 69 0.01  
## 70 0.00  
## 71 0.00  
## 72 0.01  
## 73 0.00  
## 74 0.01  
## 75 0.00  
## 76 0.00  
## 77 0.01  
## 78 0.01  
## 79 0.00  
## 80 0.00  
## 81 0.00  
## 82 0.01  
## 83 0.01  
## 84 0.00  
## 85 0.01  
## 86 0.00  
## 87 0.00  
## 88 0.00  
## 89 0.00  
## 90 0.01  
## 91 0.00  
## 92 0.00  
## 93 0.01  
## 94 0.00  
## 95 0.00  
## 96 0.00  
## 97 0.00  
## 98 0.01  
## 99 0.01  
## 100 0.00  
## 101 0.00  
## 102 0.00  
## 103 0.00  
## 104 0.00  
## 105 0.00  
## 106 0.00  
## 107 0.00  
## 108 0.00  
## 109 0.00  
## 110 0.00  
## 111 0.00  
## 112 0.00  
## 113 0.00  
## 114 0.00  
## 115 0.00  
## 116 0.00  
## 117 0.00  
## 118 0.00  
## 119 0.00  
## 120 0.00  
## 121 0.00  
## 122 0.00  
## 123 0.00  
## 124 0.00  
## 125 0.00  
## 126 0.00  
## 127 0.00  
## 128 0.00  
## 129 0.00  
## 130 0.00  
## 131 0.00  
## 132 0.00  
## 133 0.00

# Now only 133 NA values  
  
# If the year is in the title, that's a good indication's likely to be within a year of the true value. These 5 titles below only appear in the data once, so we can alter by Name:  
name\_indicates\_year <- filter(clean, grepl(" 20", Name))  
Name\_to\_Year <- filter(name\_indicates\_year, !grepl("", Year))  
clean[clean$Name == "wwe Smackdown vs. Raw 2006", "Year"] <- 2006  
clean[clean$Name == "NFL GameDay 2003", "Year"] <- 2003  
clean[clean$Name == "Tour de France 2011", "Year"] <- 2011  
clean[clean$Name == "Sega Rally 2006", "Year"] <- 2006  
clean[clean$Name == "Football Manager 2007", "Year"] <- 2006  
filter(clean, is.na(Year))

## Name Platform  
## 1 Frogger's Adventures: Temple of the Frog GBA  
## 2 Triple Play 99 PS  
## 3 Legacy of Kain: Soul Reaver PS  
## 4 Donkey Kong Land III GB  
## 5 Air-Sea Battle 2600  
## 6 Suikoden III PS2  
## 7 Yakuza 4 PS3  
## 8 WarioWare: Twisted! GBA  
## 9 Test Drive Unlimited 2 X360  
## 10 Test Drive Unlimited 2 PS3  
## 11 Monster Hunter 2 PS2  
## 12 Advance Wars: Days of Ruin DS  
## 13 Fishing Derby 2600  
## 14 Circus Atari 2600  
## 15 The Chronicles of Riddick: Escape from Butcher Bay XB  
## 16 Maze Craze: A Game of Cops 'n Robbers 2600  
## 17 Silent Hill: Homecoming X360  
## 18 Robert Ludlum's The Bourne Conspiracy X360  
## 19 NHL Slapshot Wii  
## 20 TERA PC  
## 21 Silent Hill: Homecoming PS3  
## 22 Harvest Moon: Save the Homeland PS2  
## 23 Robert Ludlum's The Bourne Conspiracy PS3  
## 24 Hangman 2600  
## 25 Cubix Robots for Everyone: Clash 'n' Bash GBA  
## 26 Tomb Raider (2013) PC  
## 27 Dragon Ball Z: Budokai Tenkaichi 2 (JP sales) Wii  
## 28 Dragster 2600  
## 29 All-Star Baseball 2005 PS2  
## 30 Slot Machine 2600  
## 31 The Dukes of Hazzard II: Daisy Dukes It Out PS  
## 32 Disgaea 3: Absence of Detention PSV  
## 33 Haven: Call of the King PS2  
## 34 Unreal Championship 2: The Liandri Conflict XB  
## 35 The Legend of Zelda: The Minish Cap(weekly JP sales) GBA  
## 36 Indy 500 2600  
## 37 Flag Capture 2600  
## 38 DanceDanceRevolution II Wii  
## 39 Big Beach Sports 2 Wii  
## 40 Jet X20 PS2  
## 41 Tribes: Aerial Assault PS2  
## 42 Move Fitness PS3  
## 43 Yu Yu Hakusho: Dark Tournament PS2  
## 44 Ghostbusters II 2600  
## 45 Breakaway IV 2600  
## 46 Valkyria Chronicles III: Unrecorded Chronicles PSP  
## 47 Famista 64 N64  
## 48 Test Drive Unlimited 2 PC  
## 49 Pet Zombies 3DS  
## 50 Trauma Team Wii  
## 51 Backbreaker X360  
## 52 Twisted Metal: Small Brawl PS  
## 53 Otomedius Excellent X360  
## 54 Backbreaker PS3  
## 55 Sword of the Samurai PS2  
## 56 Alone in the Dark: The New Nightmare PS  
## 57 Vegas Party Wii  
## 58 Jurassic Park: The Game X360  
## 59 eJay Clubworld PS2  
## 60 All-Star Baseball 2005 XB  
## 61 Our House Party! Wii  
## 62 Disney's Cinderella: Magical Dreams GBA  
## 63 Nintendo Puzzle Collection GC  
## 64 Charm Girls Club: My Fashion Mall DS  
## 65 Super Robot Wars OG Saga: Masou Kishin II - Revelation of Evil God PSP  
## 66 Saru! Get You! Million Monkeys PS2  
## 67 The Daring Game for Girls DS  
## 68 GiFTPiA GC  
## 69 Atsumare! Power Pro Kun no DS Koushien DS  
## 70 My Healthy Cooking Coach DS  
## 71 Luminous Arc 2 (JP sales) DS  
## 72 The Daring Game for Girls Wii  
## 73 Egg Monster Hero DS  
## 74 Demon Chaos PS2  
## 75 Action Man-Operation Extreme PS  
## 76 Charm Girls Club: My Fashion Show DS  
## 77 Face Racers: Photo Finish 3DS  
## 78 Zero: Tsukihami no Kamen Wii  
## 79 Mega Man Battle Network: Operation Shooting Star DS  
## 80 Dream Trigger 3D 3DS  
## 81 McFarlane's Evil Prophecy PS2  
## 82 Drake of the 99 Dragons XB  
## 83 Port Royale 3 PS3  
## 84 Build-A-Bear Workshop: Friendship Valley Wii  
## 85 Alex Rider: Stormbreaker DS  
## 86 Yoostar on MTV X360  
## 87 Port Royale 3 X360  
## 88 Jewel Link Chronicles: Mountains of Madness DS  
## 89 Chou Soujuu Mecha MG DS  
## 90 Prinny: Can I Really Be The Hero? (US sales) PSP  
## 91 Combat Elite: WWII Paratroopers PS2  
## 92 Flip's Twisted World Wii  
## 93 Mobile Ops: The One Year War X360  
## 94 Tom Clancy's Rainbow Six: Critical Hour XB  
## 95 Reader Rabbit 2nd Grade Wii  
## 96 Mountain Bike Adrenaline PS2  
## 97 Drill Dozer GBA  
## 98 B.L.U.E.: Legend of Water PS  
## 99 Half-Minute Hero 2 PSP  
## 100 Clockwork Empires PC  
## 101 Housekeeping DS  
## 102 Sabre Wulf GBA  
## 103 Beyond the Labyrinth 3DS  
## 104 Bikkuriman Daijiten DS  
## 105 Fullmetal Alchemist: Brotherhood PSP  
## 106 Combat Elite: WWII Paratroopers XB  
## 107 Samurai Spirits: Tenkaichi Kenkakuden PS2  
## 108 Super Duper Sumos GBA  
## 109 Legacy of Ys: Books I & II DS  
## 110 The King of Fighters: Maximum Impact - Maniax XB  
## 111 Combat Wings: The Great Battles of WWII Wii  
## 112 Tube Slider GC  
## 113 Umineko no Naku Koro ni San: Shinjitsu to Gensou no Yasoukyoku PS3  
## 114 Wii de Asobu: Metroid Prime Wii  
## 115 Payout Poker & Casino PSP  
## 116 Steal Princess DS  
## 117 Yu-Gi-Oh! 5D's Wheelie Breakers (JP sales) Wii  
## 118 Writing and Speaking Beautiful Japanese DS DS  
## 119 Virtua Quest PS2  
## 120 Shonen Jump's Yu-Gi-Oh! GX Card Almanac DS  
## 121 Ferrari: The Race Experience Wii  
## 122 Jet Impulse DS  
## 123 Dream Dancer DS  
## 124 Homeworld Remastered Collection PC  
## 125 Shorts DS  
## 126 Agarest Senki: Re-appearance PS3  
## 127 Hakuouki: Shinsengumi Kitan PS3  
## 128 Virtua Quest GC  
## Year Genre Publisher NA\_Sales EU\_Sales JP\_Sales  
## 1 NA Adventure Konami Digital Entertainment 2.15 0.18 0.00  
## 2 NA Sports N/A 0.81 0.55 0.00  
## 3 NA Action Eidos Interactive 0.58 0.40 0.00  
## 4 NA Platform Nintendo 0.68 0.31 0.00  
## 5 NA Shooter Atari 0.91 0.06 0.00  
## 6 NA Role-Playing Unknown 0.29 0.23 0.38  
## 7 NA Action Sega 0.15 0.14 0.63  
## 8 NA Puzzle Unknown 0.16 0.06 0.50  
## 9 NaN Racing Atari 0.30 0.32 0.00  
## 10 NaN Racing Atari 0.16 0.35 0.01  
## 11 NA Role-Playing Capcom 0.00 0.00 0.63  
## 12 NA Strategy Nintendo 0.44 0.13 0.00  
## 13 NA Sports Activision 0.48 0.03 0.00  
## 14 NA Action Atari 0.43 0.03 0.00  
## 15 NA Shooter Vivendi Games 0.32 0.11 0.00  
## 16 NA Action Atari 0.42 0.02 0.00  
## 17 NaN Action Konami Digital Entertainment 0.25 0.16 0.00  
## 18 NaN Action Vivendi Games 0.26 0.14 0.00  
## 19 NA Sports Unknown 0.40 0.00 0.00  
## 20 NA Role-Playing Unknown 0.24 0.12 0.00  
## 21 NaN Action Konami Digital Entertainment 0.18 0.15 0.00  
## 22 NA Simulation Unknown 0.19 0.15 0.00  
## 23 NaN Action Vivendi Games 0.18 0.14 0.00  
## 24 NA Puzzle Atari 0.35 0.02 0.00  
## 25 NA Action Unknown 0.26 0.10 0.00  
## 26 NA Action Square Enix 0.06 0.25 0.00  
## 27 NA Action N/A 0.15 0.05 0.14  
## 28 NA Racing Activision 0.30 0.02 0.00  
## 29 NaN Sports Unknown 0.16 0.12 0.00  
## 30 NA Action Atari 0.29 0.02 0.00  
## 31 NA Racing Unknown 0.17 0.11 0.00  
## 32 NA Role-Playing Nippon Ichi Software 0.13 0.06 0.07  
## 33 NA Platform Midway Games 0.14 0.11 0.00  
## 34 NA Shooter Midway Games 0.22 0.05 0.00  
## 35 NA Action N/A 0.00 0.00 0.27  
## 36 NA Racing Atari 0.26 0.01 0.00  
## 37 NA Action Atari 0.25 0.02 0.00  
## 38 NA Misc Konami Digital Entertainment 0.22 0.00 0.00  
## 39 NA Sports THQ 0.09 0.12 0.00  
## 40 NA Racing Unknown 0.11 0.09 0.00  
## 41 NA Shooter Unknown 0.11 0.09 0.00  
## 42 NA Sports Sony Computer Entertainment 0.00 0.17 0.00  
## 43 NA Fighting N/A 0.10 0.08 0.00  
## 44 NA Action Activision 0.20 0.01 0.00  
## 45 NA Puzzle Sears 0.20 0.01 0.00  
## 46 NA Strategy Sega 0.00 0.00 0.21  
## 47 NA Sports Namco Bandai Games 0.00 0.00 0.17  
## 48 NaN Racing Atari 0.05 0.11 0.00  
## 49 NA Simulation Majesco Entertainment 0.18 0.00 0.00  
## 50 NA Simulation Unknown 0.15 0.00 0.02  
## 51 NaN Sports Unknown 0.17 0.00 0.00  
## 52 NA Action Unknown 0.10 0.07 0.00  
## 53 NA Shooter Unknown 0.13 0.00 0.04  
## 54 NaN Sports Unknown 0.16 0.00 0.00  
## 55 NA Fighting Ubisoft 0.00 0.00 0.17  
## 56 NA Adventure Infogrames 0.09 0.06 0.00  
## 57 NA Misc Unknown 0.15 0.00 0.00  
## 58 NA Action Unknown 0.15 0.00 0.00  
## 59 NA Misc Empire Interactive 0.07 0.06 0.00  
## 60 NaN Sports Unknown 0.11 0.03 0.00  
## 61 NA Simulation Unknown 0.13 0.00 0.00  
## 62 NA Platform Disney Interactive Studios 0.10 0.04 0.00  
## 63 NA Puzzle Unknown 0.00 0.00 0.13  
## 64 NA Simulation Unknown 0.12 0.00 0.00  
## 65 NA Strategy N/A 0.00 0.00 0.12  
## 66 NA Platform Unknown 0.00 0.00 0.12  
## 67 NaN Adventure Unknown 0.11 0.00 0.00  
## 68 NA Role-Playing Unknown 0.00 0.00 0.11  
## 69 NA Sports N/A 0.00 0.00 0.10  
## 70 NA Simulation Unknown 0.09 0.00 0.00  
## 71 NA Role-Playing Unknown 0.00 0.00 0.10  
## 72 NaN Adventure Unknown 0.09 0.00 0.00  
## 73 NA Role-Playing Square Enix 0.00 0.00 0.09  
## 74 NA Action Konami Digital Entertainment 0.00 0.00 0.09  
## 75 NA Action N/A 0.05 0.03 0.00  
## 76 NA Simulation Unknown 0.08 0.00 0.00  
## 77 NA Racing Majesco Entertainment 0.08 0.00 0.00  
## 78 NA Action Nintendo 0.00 0.00 0.08  
## 79 NA Role-Playing Capcom 0.00 0.00 0.07  
## 80 NA Shooter D3Publisher 0.04 0.02 0.00  
## 81 NA Action Konami Digital Entertainment 0.03 0.03 0.00  
## 82 NA Shooter Unknown 0.05 0.01 0.00  
## 83 NaN Simulation Kalypso Media 0.06 0.00 0.00  
## 84 NA Misc Unknown 0.06 0.00 0.00  
## 85 NA Action THQ 0.06 0.00 0.00  
## 86 NA Misc Unknown 0.06 0.00 0.00  
## 87 NaN Simulation Kalypso Media 0.06 0.00 0.00  
## 88 NA Puzzle Avanquest 0.00 0.06 0.00  
## 89 NA Simulation N/A 0.00 0.00 0.06  
## 90 NA Action N/A 0.06 0.00 0.00  
## 91 NaN Shooter Unknown 0.03 0.02 0.00  
## 92 NA Platform Unknown 0.05 0.00 0.00  
## 93 NA Simulation Namco Bandai Games 0.00 0.00 0.06  
## 94 NA Shooter Unknown 0.04 0.01 0.00  
## 95 NA Misc Unknown 0.05 0.00 0.00  
## 96 NA Sports Unknown 0.03 0.02 0.00  
## 97 NA Platform Unknown 0.04 0.01 0.00  
## 98 NA Adventure N/A 0.00 0.00 0.04  
## 99 NA Role-Playing Rising Star Games 0.00 0.00 0.04  
## 100 NA Strategy Unknown 0.00 0.03 0.00  
## 101 NA Action N/A 0.00 0.00 0.04  
## 102 NA Platform THQ 0.02 0.01 0.00  
## 103 NA Role-Playing Unknown 0.00 0.00 0.03  
## 104 NA Misc N/A 0.00 0.00 0.03  
## 105 NA Action Unknown 0.00 0.00 0.03  
## 106 NaN Shooter Unknown 0.02 0.01 0.00  
## 107 NA Fighting Unknown 0.00 0.00 0.03  
## 108 NA Action Unknown 0.02 0.01 0.00  
## 109 NA Role-Playing Unknown 0.02 0.00 0.00  
## 110 NA Fighting Unknown 0.02 0.01 0.00  
## 111 NA Simulation City Interactive 0.02 0.00 0.00  
## 112 NA Racing Unknown 0.02 0.00 0.00  
## 113 NA Adventure N/A 0.00 0.00 0.02  
## 114 NA Shooter Nintendo 0.00 0.00 0.02  
## 115 NA Misc Unknown 0.02 0.00 0.00  
## 116 NA Platform Unknown 0.02 0.00 0.00  
## 117 NA Racing Unknown 0.00 0.00 0.02  
## 118 NA Misc Unknown 0.00 0.00 0.02  
## 119 NaN Role-Playing Unknown 0.01 0.01 0.00  
## 120 NA Misc Konami Digital Entertainment 0.00 0.00 0.02  
## 121 NA Racing System 3 Arcade Software 0.00 0.01 0.00  
## 122 NA Simulation Nintendo 0.00 0.00 0.02  
## 123 NA Misc N/A 0.01 0.00 0.00  
## 124 NA Strategy N/A 0.00 0.01 0.00  
## 125 NA Platform Unknown 0.01 0.00 0.00  
## 126 NA Role-Playing Idea Factory 0.00 0.00 0.01  
## 127 NA Adventure Unknown 0.01 0.00 0.00  
## 128 NaN Role-Playing Unknown 0.01 0.00 0.00  
## Other\_Sales  
## 1 0.07  
## 2 0.10  
## 3 0.07  
## 4 0.04  
## 5 0.01  
## 6 0.08  
## 7 0.05  
## 8 0.02  
## 9 0.07  
## 10 0.12  
## 11 0.00  
## 12 0.06  
## 13 0.01  
## 14 0.00  
## 15 0.02  
## 16 0.00  
## 17 0.04  
## 18 0.04  
## 19 0.02  
## 20 0.05  
## 21 0.06  
## 22 0.05  
## 23 0.06  
## 24 0.00  
## 25 0.01  
## 26 0.05  
## 27 0.01  
## 28 0.00  
## 29 0.04  
## 30 0.00  
## 31 0.02  
## 32 0.04  
## 33 0.04  
## 34 0.01  
## 35 0.01  
## 36 0.00  
## 37 0.00  
## 38 0.02  
## 39 0.03  
## 40 0.03  
## 41 0.03  
## 42 0.05  
## 43 0.03  
## 44 0.00  
## 45 0.00  
## 46 0.00  
## 47 0.03  
## 48 0.03  
## 49 0.01  
## 50 0.01  
## 51 0.01  
## 52 0.01  
## 53 0.01  
## 54 0.01  
## 55 0.00  
## 56 0.01  
## 57 0.01  
## 58 0.01  
## 59 0.02  
## 60 0.01  
## 61 0.01  
## 62 0.00  
## 63 0.00  
## 64 0.01  
## 65 0.00  
## 66 0.00  
## 67 0.01  
## 68 0.00  
## 69 0.00  
## 70 0.01  
## 71 0.00  
## 72 0.01  
## 73 0.00  
## 74 0.00  
## 75 0.01  
## 76 0.01  
## 77 0.00  
## 78 0.00  
## 79 0.00  
## 80 0.01  
## 81 0.01  
## 82 0.00  
## 83 0.01  
## 84 0.00  
## 85 0.00  
## 86 0.00  
## 87 0.00  
## 88 0.01  
## 89 0.00  
## 90 0.00  
## 91 0.01  
## 92 0.00  
## 93 0.00  
## 94 0.00  
## 95 0.00  
## 96 0.01  
## 97 0.00  
## 98 0.00  
## 99 0.00  
## 100 0.00  
## 101 0.00  
## 102 0.00  
## 103 0.00  
## 104 0.00  
## 105 0.00  
## 106 0.00  
## 107 0.00  
## 108 0.00  
## 109 0.00  
## 110 0.00  
## 111 0.00  
## 112 0.00  
## 113 0.00  
## 114 0.00  
## 115 0.00  
## 116 0.00  
## 117 0.00  
## 118 0.00  
## 119 0.00  
## 120 0.00  
## 121 0.00  
## 122 0.00  
## 123 0.00  
## 124 0.00  
## 125 0.00  
## 126 0.00  
## 127 0.00  
## 128 0.00

# Now only 128 Na values  
  
# The most accurate substitution for the remaining values would be the mean value for the year for games on that platform type... this may take a second  
na\_key <- group\_by(clean, Platform) %>%  
 summarize(m = round(mean(Year, na.rm=TRUE))) # Average year for each platform  
# Make a new row for 'mean year by platform', then back-fill with ifelse statement  
clean <- clean %>%  
 left\_join(na\_key, by = "Platform")  
clean$Year <- ifelse(is.na(clean$Year), clean$m, clean$Year)  
clean <- select(clean, -m)  
head(filter(clean, is.na(Year)))

## [1] Name Platform Year Genre Publisher NA\_Sales   
## [7] EU\_Sales JP\_Sales Other\_Sales  
## <0 rows> (or 0-length row.names)

head(clean)

## Name Platform Year Genre Publisher NA\_Sales  
## 1 Wii Sports Wii 2006 Sports Nintendo 41.49  
## 2 Super Mario Bros. NES 1985 Platform Nintendo 29.08  
## 3 Mario Kart Wii Wii 2008 Racing Nintendo 15.85  
## 4 Wii Sports Resort Wii 2009 Sports Nintendo 15.75  
## 5 Pokemon Red/Pokemon Blue GB 1996 Role-Playing Nintendo 11.27  
## 6 Tetris GB 1989 Puzzle Nintendo 23.20  
## EU\_Sales JP\_Sales Other\_Sales  
## 1 29.02 3.77 8.46  
## 2 3.58 6.81 0.77  
## 3 12.88 3.79 3.31  
## 4 11.01 3.28 2.96  
## 5 8.89 10.22 1.00  
## 6 2.26 4.22 0.58

# Genre Check:  
unique(clean$Genre)

## [1] "Sports" "Platform" "Racing" "Role-Playing" "Puzzle"   
## [6] "Misc" "Shooter" "Simulation" "Action" "Fighting"   
## [11] "Adventure" "Strategy"

genre\_vgsales <- clean %>%  
 group\_by(Genre) %>%  
 tally(sort=TRUE)   
# All clear  
  
# Sales Checks:  
max(clean$NA\_Sales)

## [1] 41.49

max(clean$EU\_Sales)

## [1] 29.02

max(clean$JP\_Sales)

## [1] 10.22

max(clean$Other\_Sales)

## [1] 10.57

head(filter(clean, NA\_Sales <0 | NA\_Sales > 20 ))

## Name Platform Year Genre Publisher NA\_Sales EU\_Sales JP\_Sales  
## 1 Wii Sports Wii 2006 Sports Nintendo 41.49 29.02 3.77  
## 2 Super Mario Bros. NES 1985 Platform Nintendo 29.08 3.58 6.81  
## 3 Tetris GB 1989 Puzzle Nintendo 23.20 2.26 4.22  
## 4 Duck Hunt NES 1984 Shooter Nintendo 26.93 0.63 0.28  
## Other\_Sales  
## 1 8.46  
## 2 0.77  
## 3 0.58  
## 4 0.47

wii\_sport <- filter(clean, NA\_Sales <0 | NA\_Sales > 40 )  
# Wikipedia confirms Wii Sport actually did have 82 million copies sold by 2017  
# Clear  
  
# Is total world wide sales equal to the sum of all sales?  
## filter(clean, (NA\_Sales + EU\_Sales + JP\_Sales + Other\_Sales) != Global\_Sales)  
# Apparently we're not using the Global column seen on kaggle in this assignment...  
  
# Identified a re-occurring issue in sales:  
dplyr::filter(clean, grepl("weekly", Name))

## Name Platform  
## 1 Project Gotham Racing (JP weekly sales) XB  
## 2 Midnight Club 3: DUB Edition (America weekly sales) PS2  
## 3 Pokemon Mystery Dungeon: Red Rescue Team (US weekly sales) GBA  
## 4 The Urbz: Sims In the City (US weekly sales) DS  
## 5 The Legend of Zelda: The Minish Cap(weekly JP sales) GBA  
## 6 Yu-Gi-Oh! The Sacred Cards (JP weekly sales) GBA  
## 7 Ratchet & Clank: Going Commando (JP weekly sales) PS2  
## 8 Tetris 2 (weekly jp sales) GB  
## 9 Chou-Kuukan Night Pro Yakyuu King (weekly JP sales) N64  
## 10 International Superstar Soccer 2000 (JP weekly sales) N64  
## 11 Mickey's Speedway USA (weekly JP sales) N64  
## 12 International Superstar Soccer 64 (weekly JP sales) N64  
## 13 Street Fighter II': Special Champion Edition (JP weekly sales) GEN  
## 14 Medal of Honor: European Assault (weekly JP sales) PS2  
## 15 FIFA Soccer 2003 (weekly jp sales) PS2  
## 16 Project Gotham Racing 2 (JP weekly sales) XB  
## 17 Tom Clancy's Ghost Recon Advanced Warfighter (weekly JP sales) X360  
## Year Genre Publisher NA\_Sales EU\_Sales JP\_Sales  
## 1 2002 Action Microsoft Game Studios 1.54 0.44 0.04  
## 2 2005 Racing Take-Two Interactive 1.22 0.05 0.00  
## 3 2005 Role-Playing Nintendo 0.81 0.30 0.00  
## 4 2004 Simulation Electronic Arts 0.27 0.00 0.00  
## 5 2003 Action N/A 0.00 0.00 0.27  
## 6 2002 Role-Playing Konami Digital Entertainment 0.00 0.00 0.26  
## 7 2003 Action Sony Computer Entertainment 0.00 0.00 0.24  
## 8 1992 Puzzle Nintendo 0.00 0.00 0.23  
## 9 1996 Sports Imagineer 0.00 0.00 0.19  
## 10 1999 Sports Konami Digital Entertainment 0.00 0.00 0.18  
## 11 2000 Racing Nintendo 0.00 0.00 0.10  
## 12 1997 Sports Konami Digital Entertainment 0.00 0.00 0.05  
## 13 1993 Action Capcom 0.00 0.00 0.07  
## 14 2005 Shooter Electronic Arts 0.00 0.00 0.05  
## 15 2002 Sports Electronic Arts 0.00 0.00 0.04  
## 16 2003 Action Microsoft Game Studios 0.00 0.00 0.03  
## 17 2006 Shooter Ubisoft 0.00 0.00 0.01  
## Other\_Sales  
## 1 0.07  
## 2 0.01  
## 3 0.02  
## 4 0.02  
## 5 0.01  
## 6 0.01  
## 7 0.00  
## 8 0.00  
## 9 0.00  
## 10 0.00  
## 11 0.00  
## 12 0.05  
## 13 0.00  
## 14 0.00  
## 15 0.00  
## 16 0.00  
## 17 0.00

dplyr::filter(clean, grepl("Weekly", Name))

## Name Platform Year  
## 1 Tony Hawk's American Wasteland (Weekly american sales) PS2 2005  
## 2 NBA Live 06 (Weekly american sales) PS2 2005  
## 3 Ratchet & Clank: Up Your Arsenal (Weekly american sales) PS2 2004  
## Genre Publisher NA\_Sales EU\_Sales JP\_Sales Other\_Sales  
## 1 Sports Activision 1.38 0.05 0 0.02  
## 2 Sports Electronic Arts 1.35 0.05 0 0.02  
## 3 Platform Sony Computer Entertainment 1.27 0.05 0 0.02

# 20 rows report only weekly sales figures (multiply by 52)  
  
# Altering all weekly Japanese sales figures to be annual  
 # Creating list to sort  
weekly\_to\_annual\_all\_rows <- filter(clean, grepl("weekly", Name) | grepl("Weekly", Name))  
weekly\_to\_annual\_jp\_rows <- filter(weekly\_to\_annual\_all\_rows,  
 grepl("JP", Name) | grepl("jp", Name)| grepl("Jp", Name))  
weekly\_to\_annual\_titles <- c(weekly\_to\_annual\_jp\_rows$Name)  
 # looping through   
for(i in weekly\_to\_annual\_titles){  
 clean[clean$Name == i, "JP\_Sales"] <- clean[clean$Name == i, "JP\_Sales"]\*52.14 # weeks per year   
}  
  
# Investigating these other "weekly sales" figures  
filter(weekly\_to\_annual\_all\_rows, !grepl("JP", Name) & !grepl("jp", Name)& !grepl("Jp", Name))

## Name Platform Year  
## 1 Tony Hawk's American Wasteland (Weekly american sales) PS2 2005  
## 2 NBA Live 06 (Weekly american sales) PS2 2005  
## 3 Ratchet & Clank: Up Your Arsenal (Weekly american sales) PS2 2004  
## 4 Midnight Club 3: DUB Edition (America weekly sales) PS2 2005  
## 5 Pokemon Mystery Dungeon: Red Rescue Team (US weekly sales) GBA 2005  
## 6 The Urbz: Sims In the City (US weekly sales) DS 2004  
## Genre Publisher NA\_Sales EU\_Sales JP\_Sales  
## 1 Sports Activision 1.38 0.05 0  
## 2 Sports Electronic Arts 1.35 0.05 0  
## 3 Platform Sony Computer Entertainment 1.27 0.05 0  
## 4 Racing Take-Two Interactive 1.22 0.05 0  
## 5 Role-Playing Nintendo 0.81 0.30 0  
## 6 Simulation Electronic Arts 0.27 0.00 0  
## Other\_Sales  
## 1 0.02  
## 2 0.02  
## 3 0.02  
## 4 0.01  
## 5 0.02  
## 6 0.02

dplyr::filter(clean, grepl("Tony Hawk's American Wasteland", Name)) # Unclear

## Name Platform Year Genre  
## 1 Tony Hawk's American Wasteland (Old all region sales) PS2 2005 Sports  
## 2 Tony Hawk's American Wasteland (Weekly american sales) PS2 2005 Sports  
## 3 Tony Hawk's American Wasteland XB 2005 Sports  
## 4 Tony Hawk's American Wasteland GC 2005 Sports  
## 5 Tony Hawk's American Wasteland X360 2005 Sports  
## Publisher NA\_Sales EU\_Sales JP\_Sales Other\_Sales  
## 1 Activision 0.80 0.63 0.01 0.19  
## 2 Activision 1.38 0.05 0.00 0.02  
## 3 Activision 0.40 0.20 0.00 0.03  
## 4 Activision 0.33 0.08 0.00 0.01  
## 5 Activision 0.28 0.02 0.00 0.03

dplyr::filter(clean, grepl("NBA Live 06", Name)) # Unclear

## Name Platform Year Genre Publisher  
## 1 NBA Live 06 (All region sales) PS2 2005 Sports Electronic Arts  
## 2 NBA Live 06 (Weekly american sales) PS2 2005 Sports Electronic Arts  
## 3 NBA Live 06 XB 2005 Sports Electronic Arts  
## 4 NBA Live 06 PSP 2005 Sports Electronic Arts  
## 5 NBA Live 06 GC 2005 Sports Electronic Arts  
## 6 NBA Live 06 X360 2005 Sports Electronic Arts  
## NA\_Sales EU\_Sales JP\_Sales Other\_Sales  
## 1 1.44 0.15 0 0.05  
## 2 1.35 0.05 0 0.02  
## 3 0.57 0.04 0 0.03  
## 4 0.39 0.00 0 0.03  
## 5 0.25 0.07 0 0.01  
## 6 0.27 0.00 0 0.02

dplyr::filter(clean, grepl("Ratchet & Clank: Up Your Arsenal", Name)) # Unclear

## Name Platform Year  
## 1 Ratchet & Clank: Up Your Arsenal PS2 2004  
## 2 Ratchet & Clank: Up Your Arsenal (Weekly american sales) PS2 2004  
## Genre Publisher NA\_Sales EU\_Sales JP\_Sales Other\_Sales  
## 1 Platform Sony Computer Entertainment 1.31 0.74 0.31 0.22  
## 2 Platform Sony Computer Entertainment 1.27 0.05 0.00 0.02

dplyr::filter(clean, grepl("Midnight Club 3", Name)) # Unclear

## Name Platform Year Genre  
## 1 Midnight Club 3: DUB Edition PSP 2005 Racing  
## 2 Midnight Club 3: DUB Edition (America weekly sales) PS2 2005 Racing  
## 3 Midnight Club 3: DUB Edition Remix PS2 2006 Racing  
## 4 Midnight Club 3: DUB Edition XB 2005 Racing  
## 5 Midnight Club 3: DUB Edition Remix XB 2006 Racing  
## Publisher NA\_Sales EU\_Sales JP\_Sales Other\_Sales  
## 1 Take-Two Interactive 1.65 1.22 0 0.79  
## 2 Take-Two Interactive 1.22 0.05 0 0.01  
## 3 Take-Two Interactive 0.97 0.04 0 0.16  
## 4 Take-Two Interactive 0.61 0.10 0 0.03  
## 5 Take-Two Interactive 0.15 0.04 0 0.01

dplyr::filter(clean, grepl("Pokemon Mystery Dungeon: Red", Name)) # Unclear

## Name Platform Year  
## 1 Pokemon Mystery Dungeon: Red/Blue Rescue Team GBA 2005  
## 2 Pokemon Mystery Dungeon: Red Rescue Team (US weekly sales) GBA 2005  
## Genre Publisher NA\_Sales EU\_Sales JP\_Sales Other\_Sales  
## 1 Role-Playing Nintendo 0.71 0.52 0.74 0.08  
## 2 Role-Playing Nintendo 0.81 0.30 0.00 0.02

dplyr::filter(clean, grepl("The Urbz: Sims In the City", Name)) # Unclear

## Name Platform Year Genre  
## 1 The Urbz: Sims In the City (US weekly sales) DS 2004 Simulation  
## Publisher NA\_Sales EU\_Sales JP\_Sales Other\_Sales  
## 1 Electronic Arts 0.27 0 0 0.02

# Doesn't appear to be marks for fixing this...

# As the name suggests, PS2 should be PS  
filter(clean, grepl("wrong", Name))

## Name Platform  
## 1 Lunar 2: Eternal Blue(sales, but wrong system) GEN  
## 2 Pachi-Slot Teiou: Golgo 13 Las Vegas (JP sales, but wrong system) PS2  
## Year Genre Publisher NA\_Sales EU\_Sales JP\_Sales Other\_Sales  
## 1 1994 Role-Playing Game Arts 0 0 0.14 0  
## 2 2002 Misc Media Entertainment 0 0 0.01 0

clean[clean$Name=="Pachi-Slot Teiou: Golgo 13 Las Vegas (JP sales, but wrong system)", "Platform"] <- "PS"  
 # Will leave it in, but unsure if this was even sold globally  
clean[clean$Name=="Lunar 2: Eternal Blue(sales, but wrong system)", "Platform"] <- "SCD" # in 1994, the platform should be sega CD  
filter(clean, grepl("wrong", Name))

## Name Platform  
## 1 Lunar 2: Eternal Blue(sales, but wrong system) SCD  
## 2 Pachi-Slot Teiou: Golgo 13 Las Vegas (JP sales, but wrong system) PS  
## Year Genre Publisher NA\_Sales EU\_Sales JP\_Sales Other\_Sales  
## 1 1994 Role-Playing Game Arts 0 0 0.14 0  
## 2 2002 Misc Media Entertainment 0 0 0.01 0

# Identified a re-occurring issue in name:  
multirow <- dplyr::filter(clean, grepl("sales", Name))  
 # Some of the observations have been split into two rows.  
multirow <- filter(multirow, !grepl("all region sales", Name))  
multirow <- filter(multirow, !grepl("All region sales", Name))  
multirow <- filter(multirow, !grepl("All Region sales", Name))  
multirow <- filter(multirow, !grepl("All Region Sales", Name))  
multirow <- filter(multirow, !grepl("all regions sales", Name))  
multirow <- filter(multirow, !grepl("wrong system", Name))  
 # 132 problematic titles that represent at least as many rows but likely more.  
 # They each need to be assessed and bound to one row, or in the case of American/US/us sales, removed. But this is an enormous undertaking, for so few marks so....

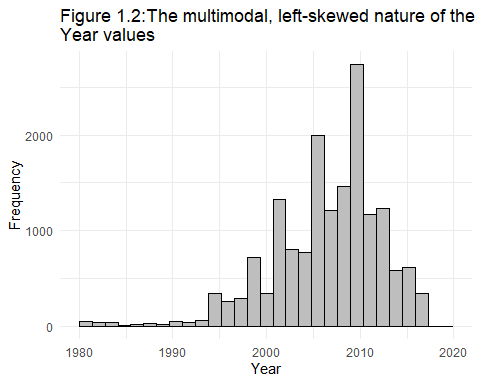
## EDA

### Part 1 - Initial variable inspections

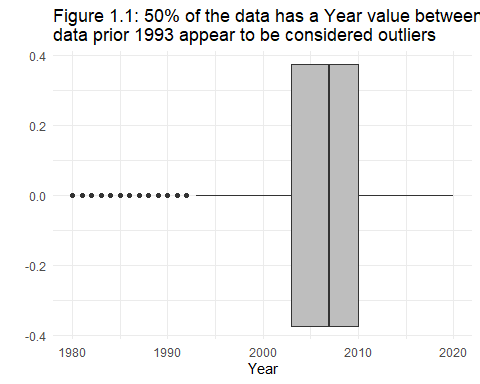
An exploratory data analysis was essential because it provided an opportunity to confirm the validity of our data and ensure it made sense. It would be pointless creating a model from inaccurate or non-nonsensical data. Depending on the model type selected, this phase generally assists in determining what transformations (if any) may be necessary and clearly identifies important characteristics/trends in each variable as well as relationships between variables.

# Univariate analysis - Year  
 # Histogram  
ggplot(clean, aes(x = Year)) +  
 geom\_histogram(fill="grey", colour ="black") +  
 ggtitle(str\_wrap("Figure 1.2:The multimodal, left-skewed nature of the data with known Year values", 70) ) +  
 xlab("Year") +  
 ylab("Frequency") +  
 theme\_minimal()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Boxplot  
ggplot(clean, aes(x = Year)) +  
 geom\_boxplot(fill = "grey") +  
 ggtitle(str\_wrap("Figure 1.1: 50% of the data has a Year value between 2003 and 2010, data prior 1993 appear to be considered outliers", 70 )) +  
 xlab("Year") +  
 ylab("") +  
 theme\_minimal()



# Summary statistics  
summary(clean$Year)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1980 2003 2007 2006 2010 2020

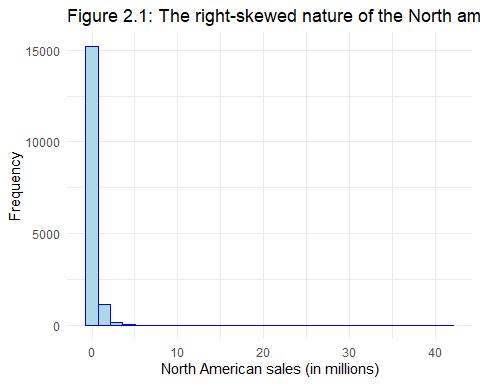
**Year**

The data sits on a domain of 1980 to 2020, with the interquartie range from 2003 to 2010. It is multi-modal with very few games listed prior to the mid-90’s or in the last few years; suggesting that data collection over the entire period has been very inconsistent. The left skew is not what one might expect to see when game development and play has been on the rise over the last few years. This data has been sourced from a public website, perhaps it has seen a decline in users since 2010 resulting in less frequent additions.

The data appears to be a poor representation of historical trends. As copies sold in North America is the outcome variable and not the dollar value, we could ignore the year entirely. However, this leads to a bit of a dangerous assumption. It assumes that wages have grown with inflation and the same proportion of the market can afford to buy a copy no matter the year. It also assumes population sizes (customer population size specifically) have been consistent. Essentially, it removes the ‘timeline’ element and the relationships observed in time (more on this later).

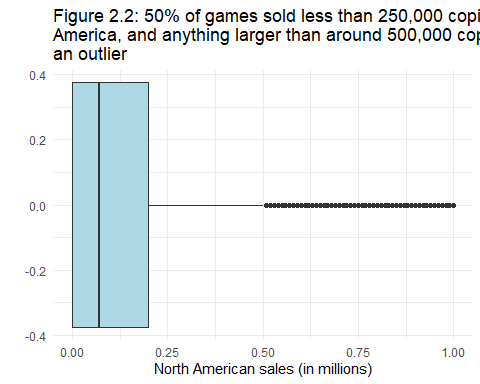
# Univariate analysis - NA\_Sales  
 # Histogram  
ggplot(clean, aes(x = NA\_Sales)) +  
 geom\_histogram(fill="lightblue", colour ="blue") +  
 ggtitle(str\_wrap("Figure 2.1: The right-skewed nature of the North american sales data", 70 )) +  
 xlab("North American sales (in millions)") +  
 ylab("Frequency") +  
 theme\_minimal()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Boxplot  
ggplot(clean, aes(x = NA\_Sales)) +  
 geom\_boxplot(fill = "lightblue") +  
 ggtitle(str\_wrap("Figure 2.2: 50% of games sold less than 250,000 copies in North America, \nand anything larger than around 500,000 copies was considered an outlier",70)) +  
 xlab("North American sales (in millions)") +  
 ylab("") +  
 scale\_x\_continuous(limits = c(0, 1)) +  
 theme\_minimal()

## Warning: Removed 895 rows containing non-finite values (stat\_boxplot).



# Summary statistics  
summary(clean$NA\_Sales)

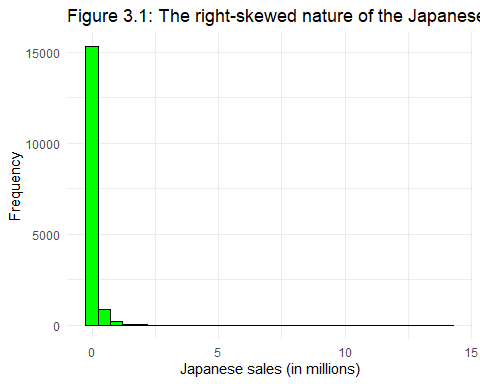
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0800 0.2647 0.2400 41.4900

**NA\_Sales**

The NA\_Sales variable had a domain of 0 to 41.49 representing the number of copies sold in millions. The interquartile range was from 0 to 0.24 million. The mean number of copies sold was 0.2642 million, but the median was only 0.08 million. The data has a uni-modal shape with a clear right skew when looking at the histogram. The boxplot makes it more apparent that there are many small modes resulting from outliers across the domain.

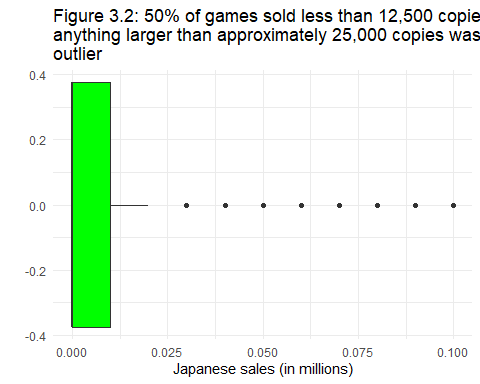
# Univariate analysis - JP\_Sales  
 # Histogram  
ggplot(clean, aes(x = JP\_Sales)) +  
 geom\_histogram(fill="green", colour ="black") +  
 ggtitle(str\_wrap("Figure 3.1: The right-skewed nature of the Japanese sales data", 70) ) +  
 xlab("Japanese sales (in millions)") +  
 ylab("Frequency") +  
 theme\_minimal()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Boxplot  
ggplot(clean, aes(x = JP\_Sales)) +  
 geom\_boxplot(fill = "green") +  
 ggtitle(str\_wrap("Figure 3.2: 50% of games sold less than 12,500 copies in Japan, and anything larger than \napproximately 25,000 copies was considered an outlier", 70)) +  
 xlab("Japanese sales (in millions)") +  
 ylab("") +  
 scale\_x\_continuous(limits = c(0, .1)) +  
 theme\_minimal()

## Warning: Removed 2433 rows containing non-finite values (stat\_boxplot).



# Summary statistics  
summary(clean$JP\_Sales)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00000 0.00000 0.00000 0.08321 0.04000 14.07780

filter(clean, JP\_Sales>108)

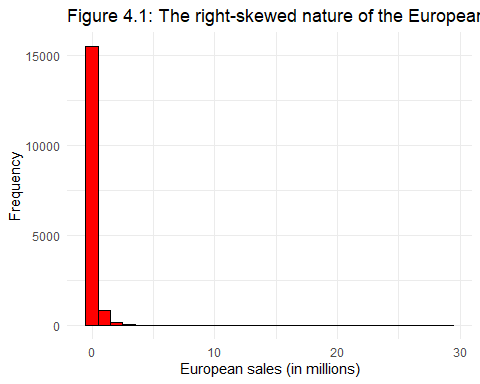
## [1] Name Platform Year Genre Publisher NA\_Sales   
## [7] EU\_Sales JP\_Sales Other\_Sales  
## <0 rows> (or 0-length row.names)

**JP\_Sales**

JP\_Sales had a domain from 0 to 14.08 million copies, (this changed because I converted a weekly sales figure to an annual sales figure in the data cleaning phase). The Median value was 0.00, while the mean value was 0.08 million. Similar to North america, the data has a right skew with a dominant mode with many small modes that represent outliers as observed in the boxplot. The median value of 0 indicates that many of the games were not sold in Japan.

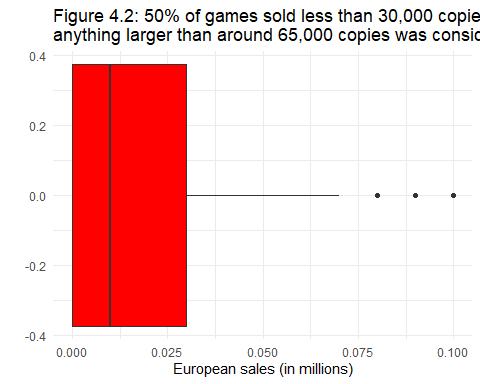
#Univariate analysis - EU\_Sales  
 # Histogram  
ggplot(clean, aes(x = EU\_Sales)) +  
 geom\_histogram(fill = "red", colour ="black") +  
 ggtitle(str\_wrap("Figure 4.1: The right-skewed nature of the European sales data", 70)) +  
 xlab("European sales (in millions)") +  
 ylab("Frequency") +  
 theme\_minimal()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Boxplot  
ggplot(clean, aes(x = EU\_Sales)) +  
 geom\_boxplot(fill = "red") +  
 ggtitle(str\_wrap("Figure 4.2: 50% of games sold less than 30,000 copies in Europe, and anything larger than around 65,000 copies was considered an outlier", 70)) +  
 xlab("European sales (in millions)") +  
 ylab("") +  
 scale\_x\_continuous(limits = c(0, .1)) +  
 theme\_minimal()

## Warning: Removed 4268 rows containing non-finite values (stat\_boxplot).



# Summary Statistics  
 summary(clean$EU\_Sales)

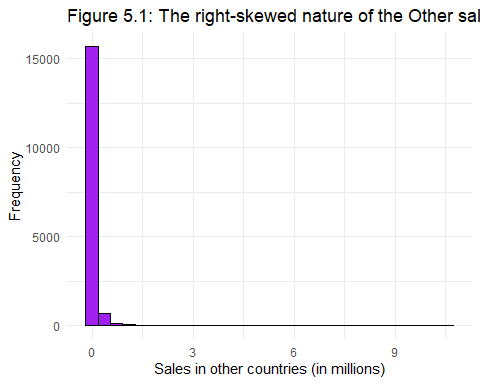
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0200 0.1467 0.1100 29.0200

**EU\_Sales**

The EU\_Sales data had a domain of 0 to 29.02 million. The median value was 0.02 million, while the mean was 0.15 million. Much like the other sales figures, this too has a dominant mode and a right skew with several smaller modes scattered across the domain in positions identified as outliers.

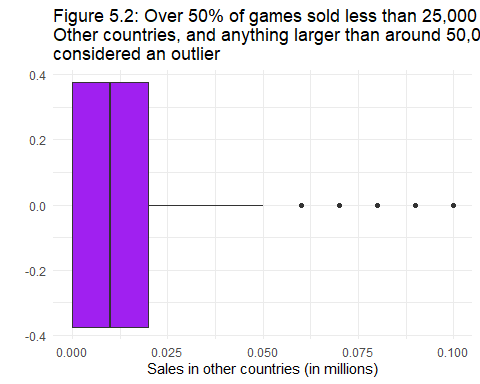
# Univariate analysis - Other\_Sales  
 # Histogram  
ggplot(clean, aes(x = Other\_Sales)) +  
 geom\_histogram(fill = "purple", colour ="black") +  
 ggtitle(str\_wrap("Figure 5.1: The right-skewed nature of the Other sales data", 70 ) ) +  
 xlab("Sales in other countries (in millions)") +  
 ylab("Frequency") +  
 theme\_minimal()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Boxplot  
ggplot(clean, aes(x = Other\_Sales)) +  
 geom\_boxplot(fill = "purple") +  
 ggtitle(str\_wrap("Figure 5.2: Over 50% of games sold less than 25,000 copies in Other countries, and anything larger than around 50,000 copies was considered an outlier", 70 )) +  
 xlab("Sales in other countries (in millions)") +  
 ylab("") +  
 scale\_x\_continuous(limits = c(0, .1)) +  
 theme\_minimal()

## Warning: Removed 1665 rows containing non-finite values (stat\_boxplot).



# Summary statistics  
summary(clean$Other\_Sales)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00000 0.00000 0.01000 0.04807 0.04000 10.57000

**Other\_Sales**

The Other\_Sales data had a domain of 0 to 10.57 million. Its mean value was 0.05 million while it’s median was 0.01 million. And it’s the same story once more, dominant mode, right skew, several outliers.

## EDA

### Part 2 - Improving the Information Exchange between Data and Model

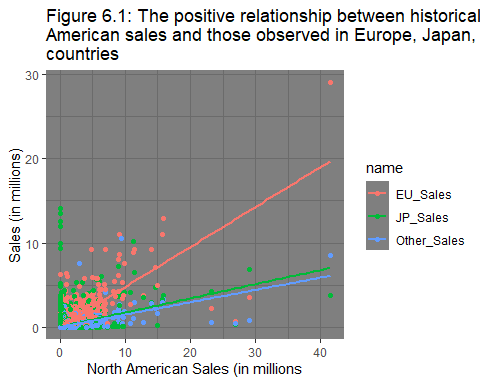
I expect the three most important predictors will be JP\_Sales, EU\_Sales, and Other\_Sales because they represent the response to a game released in a given year on a given Platform of a given Genre; in a sense these variables contain latent information representing the other variables. While cultures, languages and preferences differ around the globe, a game worth buying in one part of the world is most likely worth buying in another:

# Visualizing the relationship between NA\_Sales against all sales variables  
 # Comparison of JP\_Sales, EU\_Sales, and Other\_Sales relationships with NA\_Sales  
sales\_stats <- clean %>%  
 pivot\_longer(cols = c(EU\_Sales, JP\_Sales, Other\_Sales))   
library("ggpubr")

## Warning: package 'ggpubr' was built under R version 4.1.3

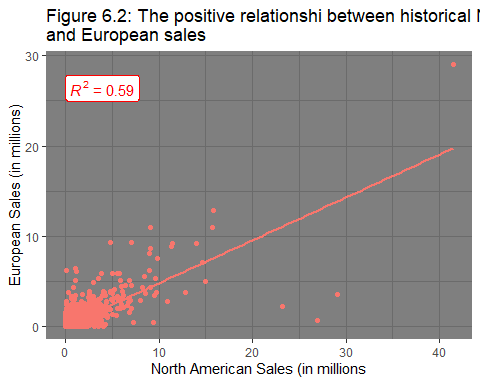
ggplot(sales\_stats, aes(x = NA\_Sales, y = value, colour = name)) +  
 geom\_point() +  
 geom\_smooth(method=lm, se=FALSE) +  
 ggtitle(str\_wrap("Figure 6.1: The positive relationship between historical North American sales and those observed in Europe, Japan, and other countries", 70))+  
 xlab("North American Sales (in millions")+  
 ylab("Sales (in millions)") +  
 theme\_dark()

## `geom\_smooth()` using formula 'y ~ x'



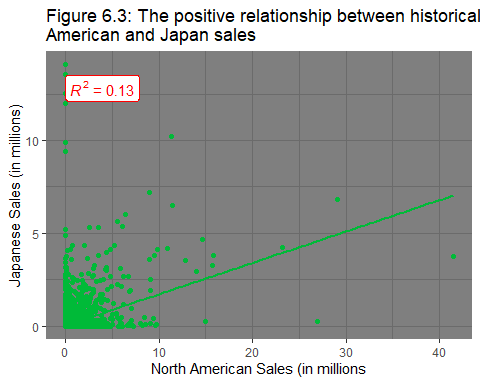
# NA\_sales and EU\_Sales  
ggplot(clean, aes(x = NA\_Sales, y = EU\_Sales)) +  
 geom\_point(colour = '#F8766D', show.legend = FALSE) +  
 geom\_smooth(colour = '#F8766D', method = lm, se=FALSE, show.legend = FALSE) +  
 ggtitle(str\_wrap("Figure 6.2: The positive relationshi between historical North American and European sales", 70))+  
 xlab("North American Sales (in millions")+  
 ylab("European Sales (in millions)") +  
 theme\_dark() +  
 ggpubr::stat\_cor(aes(label = ..rr.label..), color = "red", geom = "label")

## `geom\_smooth()` using formula 'y ~ x'



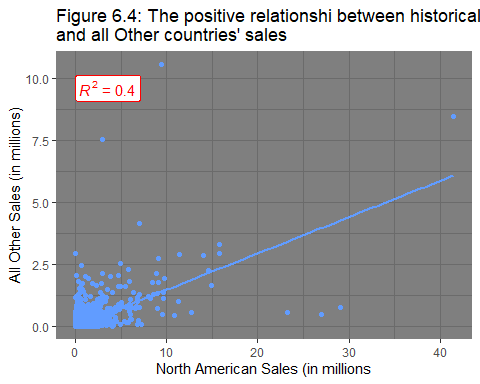
# NA\_sales and JP\_Sales   
ggplot(clean, aes(x = NA\_Sales, y = JP\_Sales)) +  
 geom\_point(colour = '#00BA38', show.legend = FALSE) +  
 geom\_smooth(colour = '#00BA38', method=lm, se=FALSE, show.legend = FALSE) +  
 ggtitle(str\_wrap("Figure 6.3: The positive relationship between historical North American and Japan sales", 70))+  
 xlab("North American Sales (in millions")+  
 ylab("Japanese Sales (in millions)") +  
 theme\_dark() +  
 ggpubr::stat\_cor(aes(label = ..rr.label..), color = "red", geom = "label")

## `geom\_smooth()` using formula 'y ~ x'



# NA\_sales and Other\_Sales  
ggplot(clean, aes(x = NA\_Sales, y = Other\_Sales)) +  
 geom\_point(colour = '#619CFF', show.legend = FALSE) +  
 geom\_smooth(colour = '#619CFF', method=lm, se=FALSE, show.legend = FALSE) +  
 ggtitle(str\_wrap("Figure 6.4: The positive relationshi between historical North American and all Other countries' sales", 70))+  
 xlab("North American Sales (in millions")+  
 ylab("All Other Sales (in millions)") +  
 theme\_dark() +  
 ggpubr::stat\_cor(aes(label = ..rr.label..), color = "red", geom = "label")

## `geom\_smooth()` using formula 'y ~ x'



**Key predictors: EU\_Sales, JP\_Sales, and Other\_Sales**

There does appear to be a positive correlation between these sales figures and NA\_Sales. European\_Sales has a moderate positive relation, perhaps due to cultural similarities as touched on briefly above. The relationship between JP\_Sales is weakly positive. When a game sells well in North America, it doesn’t appear to do as well in Japan, which could be partially related to cultural differences but also to the amount of customers in each country. Perhaps there are just more people to sell to in North America. Other\_Sales follows a similar trend to that of JP\_Sales, but it has a stronger correlation.

The right skew in the data should be addressed as a LASSO model is essentially at it’s heart still a linear model. Linear models make four key assumptions:

1. linearity
2. Normality
3. Homoscedasticity
4. Independence

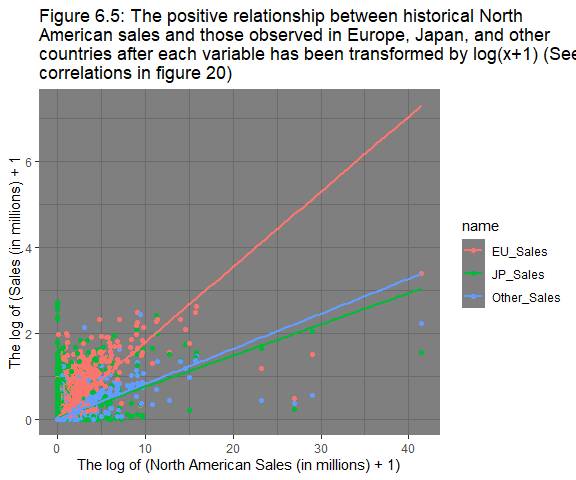
In the case of a LASSO model specifically though, each term in the otherwise linear model equation is given a penalty that relates to it’s importance. Essentially weighing each variable in the equation by importance, with a few being weighed by zero and effectiely removed. When it comes to the assumptions of A LASSO linear model it is far simpler:

1. linearity - a straight line is still the best model
2. Sparcity - only a small number of variables may be relevant
3. The Irrepresentable condition - the important variables are unrelated to the unimportant variables
4. The errors must have a finite variance and a mean of zero, but not necessarily be normally distributed <https://click.endnote.com/viewer?doi=10.9734%2Fbjmcs%2F2016%2F29533&token=WzMzOTQxOTgsIjEwLjk3MzQvYmptY3MvMjAxNi8yOTUzMyJd.NYzJY0PZUBb7vphORpVLYuubCcc>,

The sales figures were transformed by log(x+1) in the modelling phase:

sales\_stats <- clean %>%  
 pivot\_longer(cols = c(EU\_Sales, JP\_Sales, Other\_Sales)) %>%  
 mutate(value=log(value+1))  
  
ggplot(sales\_stats, aes(x = NA\_Sales, y = value, colour = name)) +  
 geom\_point() +  
 geom\_smooth(method=lm, se=FALSE) +  
 ggtitle(str\_wrap("Figure 6.5: The positive relationship between historical North American sales and those observed in Europe, Japan, and other countries after each variable has been transformed by log(x+1) \n(See correlations in figure 20)", 70))+  
 xlab("The log of (North American Sales (in millions) + 1)") +  
 ylab("The log of (Sales (in millions) + 1") +  
 theme\_dark()

## `geom\_smooth()` using formula 'y ~ x'



**Improvement of Information exchange in other variables:**

**Name** - The title length, Title language, or franchise indicators (such as ‘:’, ‘-’, ‘II’,‘2’, reoccurring ‘prefix’ strings, etc.) is probably more informative than the exact String. The language indicates a proportion of the world that can effortlessly play the game and understand the title, and signs of a franchise/sequel indicate a pre-existing fan-base that will help drive sales. Longer titles might be less enticing to consumers; a trend observable in the data. The trend becomes less obvious when you include all the games, but when you take a subset (such as games that sell up to 5 million copies) you effectively remove pre-existing franchises and get a sense of how a new game with no pre-existing fan base sells. Also verified the assumption that no specific platform had guidelines regarding title length.

# Finding Name length and total sales  
name\_length <- clean %>%  
 mutate(total\_sales = (NA\_Sales+EU\_Sales+JP\_Sales+Other\_Sales))%>%  
 group\_by(Name,Platform, total\_sales) %>%  
 tally() %>%  
 mutate(chr\_length = nchar(Name))  
  
# Creating table  
name\_length %>%  
 ungroup() %>%  
 select(c("Name", "chr\_length")) %>%  
 head(26) %>%  
 knitr::kable(caption = "Table \_: The number of characters in each title")

Table \_: The number of characters in each title

| Name | chr\_length |
| --- | --- |
| ’98 Koshien | 11 |
| .hack//G.U. Vol.1//Rebirth | 26 |
| .hack//G.U. Vol.2//Reminisce | 28 |
| .hack//G.U. Vol.2//Reminisce (jp sales) | 39 |
| .hack//G.U. Vol.3//Redemption | 29 |
| .hack//Infection Part 1 | 23 |
| .hack//Link | 11 |
| .hack//Mutation Part 2 | 22 |
| .hack//Outbreak Part 3 | 22 |
| .hack//Quarantine Part 4: The Final Chapter | 43 |
| .hack: Sekai no Mukou ni + Versus | 33 |
| [Prototype 2] | 13 |
| [Prototype 2] | 13 |
| [Prototype 2] | 13 |
| [Prototype] | 11 |
| [Prototype] | 11 |
| 007 Racing | 10 |
| 007: Quantum of Solace | 22 |
| 007: Quantum of Solace | 22 |
| 007: Quantum of Solace | 22 |
| 007: Quantum of Solace | 22 |
| 007: Quantum of Solace | 22 |
| 007: Quantum of Solace | 22 |
| 007: The World is not Enough | 28 |
| 007: The World is not Enough | 28 |
| 007: Tomorrow Never Dies | 24 |

# Revealing any relationship between title length and sales  
ggplot(filter(name\_length, total\_sales < 5), aes(x = chr\_length, y = total\_sales)) +  
 geom\_point(aes(colour = Platform), alpha=0.2, show.legend = FALSE) +  
 geom\_smooth(se=FALSE)+  
 labs(title = str\_wrap("Figure 7: The amount of sales observed compared to the length of the title for games that sold less than 5 million copies", 70),  
 x = "Number of characters in title",  
 y = "Total sales (in millions)") +  
 theme\_minimal()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



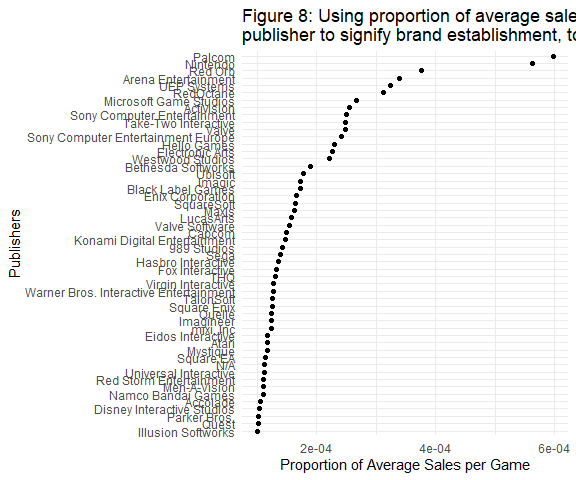
**Publisher** - Although this variable was not a predictor itself, it is worthwhile investigating for trends that might provide some inspiration for the model building phase. For instance, some Publishers (such as ‘Nintendo’) are also the producers of the console. The name alone does not give an indication of how well established the company is nor the fan base they have in terms of market share. We can get an estimate of this if we instead consider the average sales per game for each Publisher. Below I’ve prepared a plot to approximate the average proportion of market share each Publisher has:

# Finding the total sales for each Publisher  
clean$Publisher <- as.factor(clean$Publisher)  
established <- clean %>%  
 mutate(publisher\_sales = (NA\_Sales+EU\_Sales+JP\_Sales+Other\_Sales),  
 releases = 1) %>%  
 group\_by(Publisher, publisher\_sales, releases) %>%  
 tally() %>%  
 group\_by(Publisher) %>%  
 summarise(releases = sum(releases),  
 publisher\_sales = sum(publisher\_sales)) %>%  
 mutate(estimated\_market\_share = ((publisher\_sales/releases)/sum(publisher\_sales)))  
  
# Creating table  
established %>%  
 ungroup() %>%  
 select(c("Publisher", "publisher\_sales")) %>%  
 arrange(desc(publisher\_sales) ) %>%  
 head(100) %>%  
 knitr::kable(caption = "Table \_:The top 100 Publishers in terms of sales in millions")

Table \_:The top 100 Publishers in terms of sales in millions

| Publisher | publisher\_sales |
| --- | --- |
| Nintendo | 1630.6862 |
| Electronic Arts | 675.0526 |
| Activision | 528.0100 |
| Sony Computer Entertainment | 487.6536 |
| Take-Two Interactive | 347.0400 |
| Ubisoft | 325.0714 |
| Microsoft Game Studios | 234.4098 |
| Konami Digital Entertainment | 199.3786 |
| THQ | 193.5200 |
| Sega | 175.1600 |
| Capcom | 166.7698 |
| Warner Bros. Interactive Entertainment | 128.2400 |
| Namco Bandai Games | 124.5300 |
| Square Enix | 118.7400 |
| Atari | 116.0900 |
| Disney Interactive Studios | 96.2800 |
| Eidos Interactive | 87.3700 |
| LucasArts | 82.8800 |
| Bethesda Softworks | 80.3000 |
| SquareSoft | 53.6200 |
| Midway Games | 50.6900 |
| Acclaim Entertainment | 46.4900 |
| 505 Games | 44.9400 |
| Vivendi Games | 42.1700 |
| Virgin Interactive | 41.9500 |
| Codemasters | 36.7700 |
| Enix Corporation | 33.7000 |
| Tecmo Koei | 26.9300 |
| N/A | 26.7378 |
| Unknown | 25.4000 |
| GT Interactive | 24.2200 |
| Sony Computer Entertainment Europe | 23.6100 |
| MTV Games | 20.0100 |
| Deep Silver | 19.6100 |
| Hudson Soft | 18.8900 |
| Universal Interactive | 17.2400 |
| D3Publisher | 15.9900 |
| Hasbro Interactive | 15.2100 |
| Rising Star Games | 14.3900 |
| Banpresto | 13.9900 |
| Infogrames | 13.6000 |
| 989 Studios | 12.9000 |
| Imagineer | 12.1766 |
| Majesco Entertainment | 11.8600 |
| Level 5 | 11.7600 |
| ASCII Entertainment | 10.8200 |
| Empire Interactive | 10.4500 |
| Nippon Ichi Software | 10.4100 |
| Atlus | 9.8800 |
| 3DO | 8.7200 |
| RedOctane | 8.6900 |
| Zoo Digital Publishing | 8.0100 |
| Fox Interactive | 7.3800 |
| TDK Mediactive | 7.3800 |
| Global Star | 6.7100 |
| Crave Entertainment | 6.6800 |
| Activision Value | 6.5200 |
| Mindscape | 6.4100 |
| Interplay | 6.4000 |
| Psygnosis | 6.1000 |
| Focus Home Interactive | 6.0200 |
| Red Orb | 5.2500 |
| Ignition Entertainment | 5.2200 |
| SouthPeak Games | 5.2000 |
| Spike | 5.1500 |
| Parker Bros. | 4.9700 |
| Imagic | 4.8200 |
| Arena Entertainment | 4.7200 |
| Koch Media | 4.7100 |
| Tomy Corporation | 4.5000 |
| City Interactive | 4.4900 |
| Ubisoft Annecy | 4.4800 |
| Palcom | 4.1700 |
| Play It | 4.1500 |
| Zoo Games | 4.0400 |
| ChunSoft | 3.9900 |
| Nordic Games | 3.9900 |
| Gotham Games | 3.7300 |
| Square | 3.6500 |
| GSP | 3.6000 |
| Natsume | 3.5500 |
| BAM! Entertainment | 3.5300 |
| Sony Online Entertainment | 3.5100 |
| Mattel Interactive | 3.4700 |
| Maxis | 3.3900 |
| Ocean | 3.3600 |
| SCi | 3.3400 |
| Destineer | 3.3200 |
| Takara | 3.3200 |
| Valve Software | 3.2200 |
| PQube | 3.1600 |
| GungHo | 3.1500 |
| Oxygen Interactive | 3.1200 |
| Titus | 3.0900 |
| Game Factory | 3.0900 |
| Coleco | 3.0700 |
| Mojang | 2.9100 |
| Scholastic Inc. | 2.8800 |
| Marvelous Interactive | 2.8000 |
| Sammy Corporation | 2.7300 |

# Plotting the Publishers that produced more than 20 million sales  
ggplot(filter(established, estimated\_market\_share > 0.0001),  
 aes( x = estimated\_market\_share, y = fct\_reorder(Publisher, estimated\_market\_share))) +  
 geom\_point() +  
 ggtitle(str\_wrap(" Figure 8: Using proportion of average sales per game for each publisher to signify brand establishment, top 50 Publishers", 70)) +  
 ylab("Publishers") +  
 xlab("Proportion of Average Sales per Game") +  
 theme\_minimal()



# Finding the amount of sales per Platform over time  
meaningful\_platform <- clean %>%  
 mutate(total\_sales = (NA\_Sales+EU\_Sales+JP\_Sales+Other\_Sales))%>%  
 group\_by(Platform, Year, total\_sales) %>%  
 tally() %>%  
 summarise(platform\_sales\_for\_year = sum(total\_sales))

## `summarise()` has grouped output by 'Platform'. You can override using the  
## `.groups` argument.

# Creating table  
clean %>%  
 mutate(total\_sales = (NA\_Sales+EU\_Sales+JP\_Sales+Other\_Sales))%>%  
 group\_by(Platform, total\_sales) %>%  
 tally() %>%  
 summarise(platform\_sales\_for\_year = sum(total\_sales)) %>%  
 arrange(desc(platform\_sales\_for\_year)) %>%  
 knitr::kable(caption = "Table \_: Sales made for each category of platform")

Table \_: Sales made for each category of platform

| Platform | platform\_sales\_for\_year |
| --- | --- |
| PS2 | 720.8162 |
| Wii | 687.7600 |
| X360 | 677.1614 |
| PS3 | 631.3200 |
| DS | 518.6700 |
| PS | 435.2500 |
| GB | 258.9122 |
| PS4 | 249.5000 |
| NES | 236.6900 |
| GBA | 222.5942 |
| 3DS | 202.9300 |
| N64 | 200.8628 |
| PC | 183.1400 |
| SNES | 171.0800 |
| PSP | 160.5100 |
| XB | 146.1998 |
| GC | 130.2400 |
| XOne | 125.2900 |
| 2600 | 75.0700 |
| WiiU | 71.6000 |
| PSV | 36.0700 |
| GEN | 31.5498 |
| SAT | 15.6200 |
| DC | 14.3700 |
| SCD | 1.8600 |
| WS | 1.4200 |
| NG | 1.2400 |
| TG16 | 0.1600 |
| 3DO | 0.0800 |
| GG | 0.0400 |
| PCFX | 0.0300 |