North American Video Game Market: Understanding Which Games Win

Sam Rosenberg

Introduction

In this project, we will be analyzing games that sold over a million copies from the dataset provided by Kaggle (https://www.kaggle.com/gregorut/videogamesales). This data included a total of 16598 games from the year 1977 to 2018. This data isn't necessarily a random sample, however, because it does not reflect the entirety of all games published. Thus, it will be considered a subpopulation of some super-population of all potentially published games in the dataset's timeframe. The question we are seeking to answer is what variables are important in predicting North American sales of video games.

The history of game production has been full of different consoles and different companies. This report looks to understand the ways in which these different factors work together to predict the success of some of the most successful games and to see if certain companies have better luck selling copies of the games in the North American market. Companies that make their games come from all across the world: Nintendo from Japan, Ubisoft from France, and Activision in the United States, so it remains a question as to whether or not famous companies that originate from outside a region have more difficulty selling in a foreign market. Further, does the American market buy certain genres of games more than others? These are some of the questions that are intended to be answered with this analysis given the variables made available to us in this dataset.

We will also be looking at some numerical variables. Our data includes the global sales of a game, so it is possible to see the total non-American sales a game has had. We will be exploring if the international success of a game is a good predictor of how well the game sells in

the North American region. We will also be analyzing age as a factor of game sales, to see if it has any impact on the number of games sold. While older games have had more time to sell, it's possible newer games could have been able to cash in on a growing market.

While not many public studies have been conducted on the factors that make a game sell well in the North American market, there have been a handful of publications on the growth of the industry. As time has gone on, the industry of video games has expanded and become more mainstream. With a larger audience, newer games may have found themselves with the ability to sell to a wider customer base, leading to an increase in sales.

Methods

The game dataset includes so many flops (games that didn't sell more than a couple thousand copies) that the skew in the sales make it such that we have to consider only games that at least got some traction. In this case, we looked at games that sold over one million copies globally in the dataset. Thus, all conclusions are drawn in relation to games that have over a million copies sold, meaning coefficients are interpreted in terms of how they relate to already successful games. We will be analyzing what factors explain an increase in the number of game copies sold based on the age, genre, producer, and platform of the game. The data was scraped from VGChartz, a website that compiled information about game sales up through the year 2018.

The games we are looking at are ones that were bought over 1 million times across the globe (or at least recorded regions) and more than 100 thousand times in the North American region. This second bound for North American sales is meant to prevent the handful of wildly popular global games that simply weren't put on sale in the US.

To prevent overfitting we chose to take into account the nine most successful publishers globally: Nintendo, Microsoft Game Studios, Take-Two Interactive, Sony Computer Entertainment, Activision, Ubisoft, Bethesda Softworks, Electronic Arts, Sega. All other producers were considered the same for the sake of this analysis. While these aren't necessarily the largest publishers in the dataset, they were the producers of the most popular games.

Further, the large number of game platforms included in the dataset had a handful of oddities that were removed before creating our model. First, all platforms of the same family were consolidated: PlayStations 1-4 were combined to simply act as a single console, as were the Xbox consoles and several others. Any system with less than 100 total games in the set was combined into an "other" category to represent less supported hardware.

The greatest number of null values was in the data the year was produced in. To solve this issue, we scraped the database of idgb, a popular video game website, for the games that were lacking a release date. We first searched for each of the 270 or so games on the list that lacked a date, then recorded the first release date available in the search results (for the game that best matched the search). This had the potential to cause some errors in the data with two games with similar names, but generally, this seems to be about as accurate as it can be. This data acquisition was done through the python package selenium (Appendix C).

Results

Initial data exploration indicated that the North American sales numbers of the individual games were heavily right-skewed, as seen in the following figure. A log transformation was conducted in order to help alleviate this skew, which led to the data being approximately normal (see Appendix A, 1).

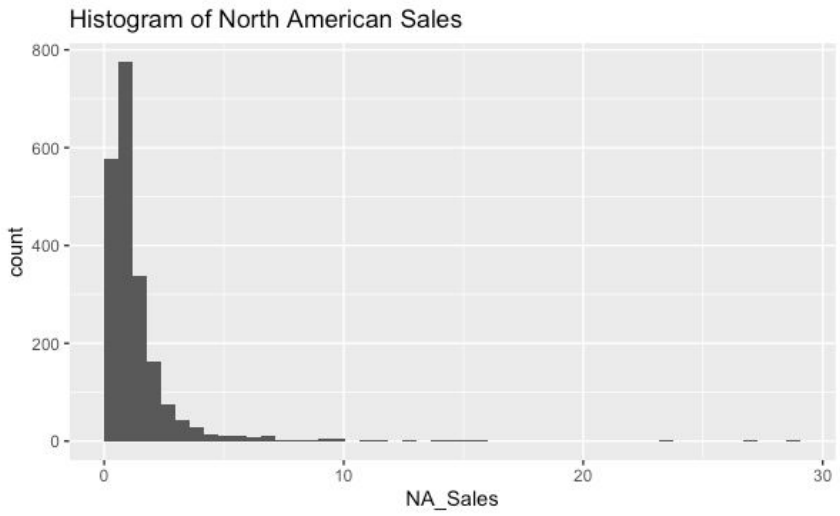

Figure 1: Histogram of North American Game Sales

Before these changes were made to the data, the summary statistics of such data can be seen in the table below:

	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std. Deviation
age	0.0	10.0	13.0	13.6	17	43	5.88
NA_Sales	0.00	0.00	0.0800	0.2647	0.2400	41.4900	1.56
Global_Sales	0.01	0.06	0.17	0.54	0.47	82.74	1.56

Table 1: Table of Numerical Data in All Game Sales

Publisher	**Electronic Arts**	**Activision**	**Ubisoft**	**Nintendo**	**Sony**	**Sega**	**Take-Two**	**Microsoft**	**Bethesda**
Count	1351	975	921	703	683	639	413	189	71

Table 2: Publishers of All Games in Data

Genre	**Action**	**Sports**	**Misc**	**Role-Playing**	**Shooter**	**Adventure**	**Racing**	**Platform**	**Simulation**	**Fighting**	**Strategy**	**Puzzle**
Count	3316	2346	1739	1488	1310	1286	1249	886	867	848	681	582

Table 3: Genre of All Games in Data

After selecting the games which sold more than one million total copies and removing an outlier that will be addressed later, the following is the summary of the data:

	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std. Deviation
age	4.00	10.00	14.00	15.09	19.00	43.00	7.10
Global_Sales	1.000	1.290	1.750	2.765	2.870	40.240	3.28
NA_Sales	0.100	0.610	0.940	1.393	1.540	29.080	1.78

Table 4: Table of Numerical Data in 1M+ Game Sales

Publisher	**Electronic Arts**	**Activision**	**Ubisoft**	**Nintendo**	**Sony**	**Sega**	**Take-Two**	**Microsoft**	**Bethesda**
Count	342	161	117	338	149	74	91	54	22

Table 5: Publishers of All Games in 1M+ Game Sales

Genre	**Action**	**Sports**	**Misc**	**Role-Playing**	**Shooter**	**Adventure**	**Racing**	**Platform**	**Simulation**	**Fighting**	**Strategy**	**Puzzle**
Count	426	302	173	203	252	42	182	195	93	124	32	56

Table 6: Genre of All Games in 1M+ Game Sales

We have four explanatory variables planned in our model, along with two numerical variables and two categorical variables. The numerical variables include the age of a game and the number of sales from non-NA regions. The categorical variables include the genre of game (with twelve levels shown above) and the nine popular publishers shown above.

Because our intention is to fit a regression with the variables that explain the NA sales of Games, we will be trying to build the best fit model given all variables at our disposal. During the initial build process of the model, several issues became clear they needed addressing.

The popular title Wii Sports has been excluded from this analysis. With over 80 million copies sold, it was clearly the most popular game in the dataset. The reason for its exclusion is that it came pre-bundled with every Wii sold throughout the lifespan of the console. Because it was the only such game to be intrinsically tied to the sale of an incredibly popular console, it was

not consistent with the rest of the dataset, had issues with independence and was removed from consideration by the model.

The untransformed data is very heteroskedastic in the model (Appendix B, 1), the residuals don't have a trend up or down, but have a clearly defined cone shape. Taking the log of the NA sales helps the problem, but it should be noted that there may continue to be an issue with the randomness of variance that would need to be solved with advanced techniques (Appendix B, 2). For now, however, it is the closest we can get to having an evenly spread distribution of errors.

The model assumes that all points of game sales are independent of each other. This assumption could be flawed, as was true with the Wii Sports example (if someone bought any Wii game they also bought Wii sports). Games on the same console are likely going to be sold in proportion to the amount the console itself was sold, however, in an attempt to standardize that, the factor of the game's platform was added in the model.

The log of North American sales and other sales have a relationship that at least trends positively, however no transformation we attempted made the relationship fully linear. We will be adding the other sales to the model to see if it has significant predicting power in determining the number of NA sales. This linear relationship becomes shaky at best as it approaches the low end of both values, however, the general trend of data does appear to be positive.

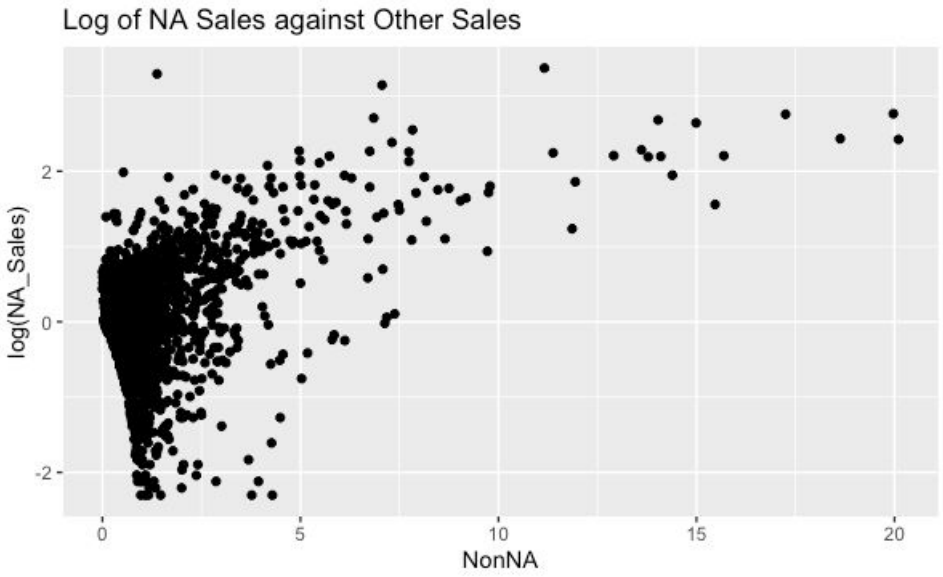

Figure 2: Log of NA Sales against Other Sales

Upon observing initial models, the only major game companies that had a significant impact ($p < 0.05$) on predicting the number of copies of a game sold in the North American region is Microsoft, Take-Two, Sony, Activision, Bethesda, and EA. Nintendo, Sega, and Ubisoft don't have a significant impact when compared to the rest of the gaming companies.

Our final theoretical model ends up being:

$$E[log(NA\_Sales)| age + Genre + Micro + TakeTwo + Sony + Activision + Bethesda + EA + Platform + NonNA] = \beta_0 + \beta_1 age + \beta_2 Micro + \beta_3 TakeTwo + \beta_4 Sony + \beta_5 Activision + \beta_6 Bethesda + \beta_7 EA + \beta_8 NonNA + \beta_{9\text{-}19} Genres + \ + \beta_{20\text{-}34} Platforms$$

In the model above, age represents the number of years since the release of the game, with 2020 games having an age of 0. The names of companies (Micro, TakeTwo, Sony, Activision, Bethesda and EA) each take on the value of either 1 or 0 depending on whether a game is produced by that specific company. NonNA is the total game sales internationally (in millions). The number of different genres are marked as the coefficients for Genres to save space, the

genres are Adventure, Fighting, Misc, Platform, Puzzle, Racing, Role-Playing, Shooter, Simulation, Sports and Strategy. These are all binary variables based on the genre of the game, and games cannot have multiple genres. Like Genre, Platform represents a number of different platforms. These are 3DS, DS, Gboy, GC, N64, NESes, PC, PSconsole, PSP, PSV, SAT, Wiiconsole, XBconsole, and XOne.

The coefficients and their standard errors are shown in the table below.

Coefficient	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-0.627117	0.176708	-3.549	0.000396
age	0.025269	0.003255	7.763	1.33e-14
GenreAdventure	-0.063511	0.101953	-0.623	0.533397
GenreFighting	0.112936	0.064004	1.765	0.077802
GenreMisc	0.018813	0.059149	0.318	0.750474
GenrePlatform	0.079791	0.056674	1.408	0.159321
GenrePuzzle	-0.231239	0.094045	-2.459	0.**014027**
GenreRacing	-0.150902	0.057164	-2.640	0.008363
GenreRole-Playing	-0.196237	0.057853	-3.392	0.000708
GenreShooter	0.068948	0.050427	1.367	0.171698
GenreSimulation	-0.203538	0.076363	-2.665	0.007754
GenreSports	-0.156301	0.050388	-3.102	0.001950
GenreStrategy	-0.016544	0.131171	-0.126	0.899648
Micro	0.386732	0.093994	4.114	4.04e-05
TakeTwo	0.348589	0.069813	4.993	6.47e-07
Sony	0.162657	0.059770	2.721	0.006559
Activision	0.247708	0.054943	4.508	6.92e-06
Bethesda	0.276191	0.135934	2.032	0.042311
EA	0.122706	0.043995	2.789	0.005338
PlatformOther	0.035610	0.193864	0.184	0.854279
Platform3DS	-0.216117	0.187068	-1.155	0.248117
PlatformDS	-0.177266	0.159327	-1.113	0.266022
PlatformGboy	-0.054850	0.147584	-0.372	0.710193
PlatformGC	0.066690	0.167175	0.399	0.689994
PlatformN64	0.149458	0.159675	0.936	0.349384
PlatformNESes	-0.699973	0.139618	-5.013	5.83e-07
PlatformPC	0.079209	0.170490	0.465	0.642274
PlatformPScon	-0.243888	0.147202	-1.657	0.097717 .
PlatformPSP	-0.661601	0.170577	-3.879	0.000109
PlatformPSV	-0.719062	0.267121	-2.692	0.007166

PlatformSAT	-1.781208	0.376664	-4.729	2.42e-06
PlatformWiicon	0.075986	0.159435	0.477	0.633706
PlatformXBcon	0.195030	0.155454	1.255	0.209781
PlatformXOne	0.213801	0.187853	1.138	0.255207
NonNA	0.237823	0.009090	26.163	< 2e-16

Table 7: Final Model of Game Sales

While there are too many coefficients to walkthrough in this results section, we can take a look at the interpretation of a categorical and numerical variable to understand the way in which these coefficients predict the North American game sales. Looking at the Non-North American coefficient (representing the total sales from outside the NA region, we can see a coefficient of 0.237823. This means that for each million sales the game received outside of NA, the game sales in North American increase by 1.27 times. This has a 95% confidence interval of 1.25 to 1.29 times the sales for each additional million games sold.

For each genre, the model predicts the increase in sales over if the game was an action title instead. This would mean that to compare one genre to another you could compare the change between the two genres in comparison to action. Racing has a coefficient of -0.150902, meaning the model expects a racing game with greater than one million global sales and 100 thousand North American sales to get .86 times the number of North American sales as an action game with 95% confidence between .77 and .96 times the sales holding all else constant. This same analysis can be done for the publishers of the game, with the comparison is to the same game made by any other publisher than itself.

It's worth noting that this data set doesn't have a lot of games in the last couple of years, and the ones it does contain are not exceptionally well selling. We can know from a cursory google search and generalized knowledge that there are recent games that sold incredibly well in

North America but were not represented in the model. This brings us to the major foundational problems a lot of this analysis is stuck with.

Our data is not a random sample of all video games across the existence of the medium. We don't have representative quantities across the years from a solid mix of publishers. The data we do have is often quite robust, especially towards the early and mid-2010s, but isn't fully randomized. It's likely this data should be qualified further as to be most accurate at explaining games before the year 2016, as that appears to be the last year of high volume recordings.

Discussion

The data modeled included 1962 games from the year 1977 to 2018, with information on the year the game was published, the platform it was released on, the company that published the game, and the regional sales of the game. Our final model with the constraints on the level of success needed included We used a linear model to determine which of the aforementioned factors were significant predictors of success games had in the North American region, and in what way those predictors impacted the sales of the games.

We were able to conclude that many of these different factors were significant in understanding how well games would do. The number of copies the game sold (in millions) was a significant positive predictor of the number of NA copies sold ($p\text{-value} < 2e\text{-}16$), and shows that as a whole, games with greater success internationally translates to greater predicted success in North America.

We also concluded that the age of a game did predict a significant increase in sales with each year the game had been out ($p\text{-value} = 1.33e\text{-}14$). While there are several reasons to be skeptical about this conclusion given potential issues with the assumption of linearity, as well as

a potentially unrepresentative dataset, the model did find that if this data were taken to be a randomized sample of some larger population of games then newer games are not predicted to have sold more than ones released in years past (p-value = 1.33e-14). For each additional year the game is out, it is predicted that its total predicted sales with increase.

While this study has found a number of potentially interesting significant relationships between some of the tested variables, it's important to understand the limitations of this analysis. First, the assumptions made about linearity between the numerical data is very shaky. There is a positive trend, and the model finds a significant relationship, however, it needs to be understood that this comes in potential violation of another assumption. The data set comes from a time before the modern era of video games, which has seen the rise of free to play games built on monetization models other than traditional sales, making this analysis built for what may not be representative of the current video game environment and thus not at all applicable to the games of the modern-day.

For another study, we would recommend looking at a smaller more specific window of time, as well as constructing at least a reasonably random sample of the gaming population. As games have moved to online distributors such as Steam, data continues to get easier to gather, and information about those games has never been more accessible. Looking at a more concentrated dataset than simply games this bot managed to record from a certain website would allow for much more sound statistical conclusions than what could have been drawn in this study. The analysis in this project makes it clear that there is likely merit to examining the modern genres and games to see how well those sell and discovering current leading publishers in the game design space.

Appendix A

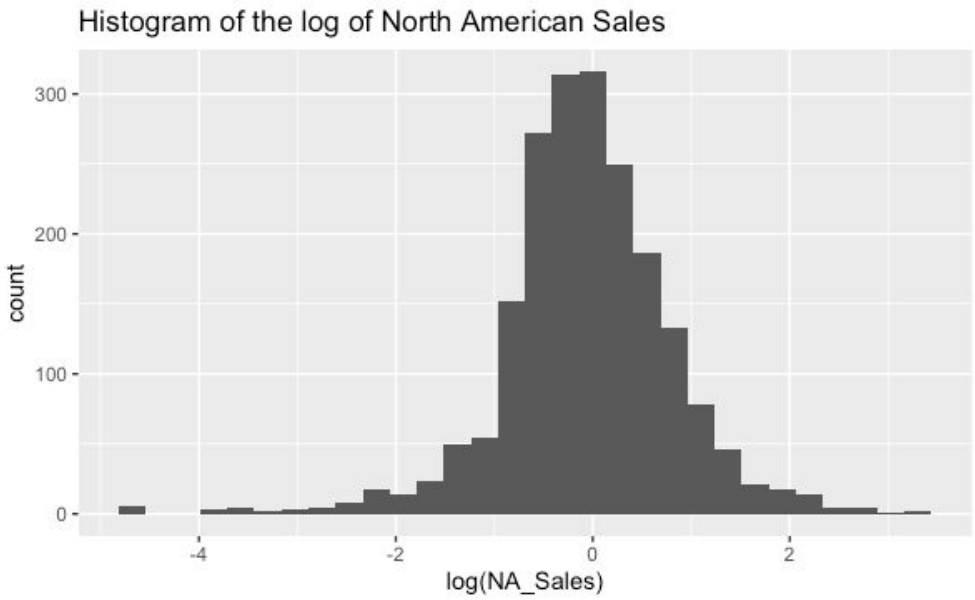

Figure 1: Histogram of the log of North American Sales

Appendix B

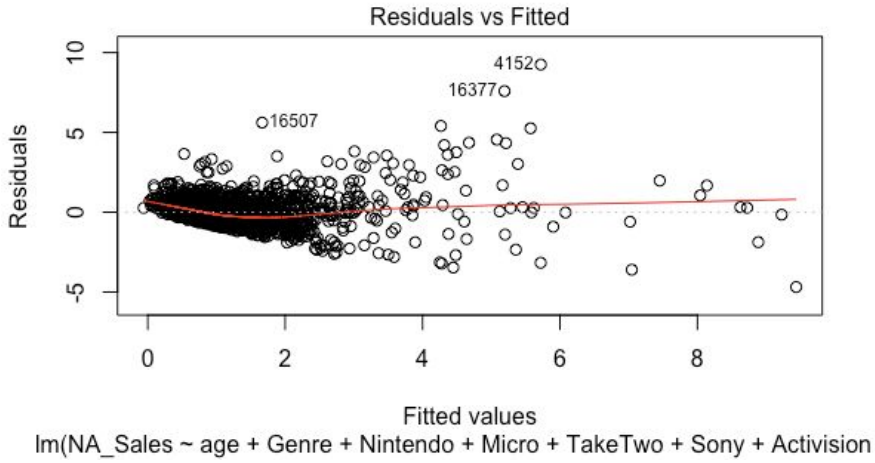

Figure 1: Residuals when NA_Sales untransformed

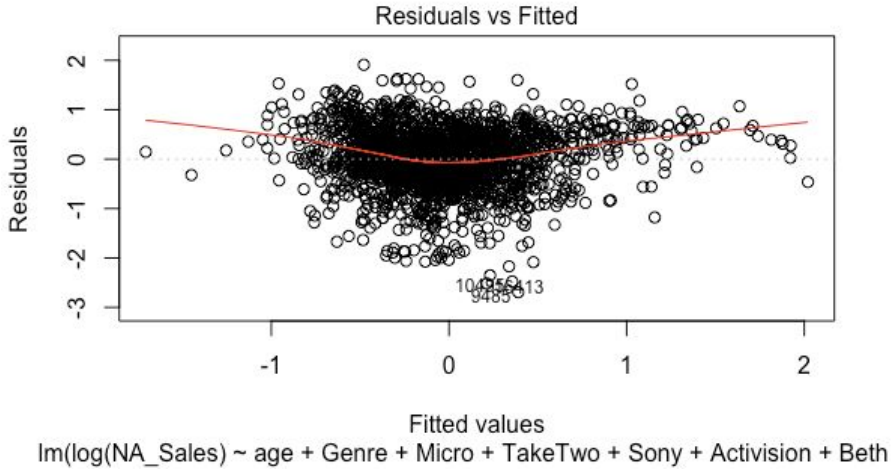

Figure 2: Residuals of Final Linear Regression on Log of North American Sales

Appendix C - Python for Scraping of Dates from IGDB

```
# list of games missing dates, 1 per line
games = """

"""

#https://stackoverflow.com/questions/13169725/how-to-convert-a-string-that-
has-newline-characters-in-it-into-a-list-in-python
out = []
buff = []
for c in games:
   if c == '\n':
       out.append(''.join(buff))
       buff = []
   else:
       buff.append(c)
else:
   if buff:
      out.append(''.join(buff))

print out

import pygal
from selenium import webdriver
import time
import requests
import csv
result_array = []
final_games = []

for each in out:
   final_games.append(each.replace(" ", "-"))

#This has to be the location of chrome driver on the computer
driver = webdriver.Chrome('/Users/samrosenberg/Desktop/chromedriver')  #
Optional argument, if not specified will search path.
with open('years.csv', mode='w') as csv_file:
   fieldnames = ['Game', 'year']
   writer = csv.DictWriter(csv_file, fieldnames=fieldnames)
   writer.writeheader()
   # go to the first game page
```

```
#this uses Selenium to properly search the Javascript/HTML.
for each in out:
    driver.get("https://www.igdb.com/games/" + each)
    time.sleep(2)
    try:
        date = driver.find_elements_by_class_name("game-shortdate")
        datefin = (date[0].text)
        print(each + " : " + datefin)
        writer.writerow({'Game': each, 'year': datefin})
    except:
        continue

driver.quit()

csv_file.close()
```

Appendix D - R Code

```
# Library imports
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(plyr)
library(tidyr)
library(car)

# upload data #
vgsales <- read.csv('/Users/samrosenberg/Documents/vgsales.csv')

# head the csv #
head(vgsales)
vgsales$age <- 2020 - vgsales$Year

# sumarize the data #
summary(vgsales)
sd(vgsales$age)
sd(vgsales$NA_Sales)
table(vgsales$Publisher)
table(vgsales$Genre)

# remove wii sports #
vgsales.subsport <- vgsales[-9474,]

# Plot North American Sales vs Age #
ggplot(vgsales.subsport, aes(y = NA_Sales, x = age)) + geom_point()
#boxplots
ggplot(vgsales.subsport, aes(x = as.factor(Genre), y = NA_Sales)) +
geom_boxplot() + xlab("Genre")
#plot against age
ggplot(vgsales.subsport, aes(y = NA_Sales, x = age)) + geom_point()
```
```
# Data with more than 1M copies sold #
vgsalesmil <- vgsales.subsport
vgsalesmil <- vgsales.subsport[which(vgsales.subsport$Global_Sales >= 1),]
ggplot(vgsalesmil, aes(x = as.factor(Genre), y = log(NA_Sales))) +
geom_boxplot() + xlab("Genre")

# Create Non North American Sales Column #
vgsalesmil$NonNA <- vgsalesmil$Global_Sales - vgsalesmil$NA_Sales
```

```
# plot log NA sales against age and NonNA #
ggplot(vgsalesmil, aes(y = log(NA_Sales), x = age)) + geom_point()
ggplot(vgsalesmil, aes(y = log(NA_Sales), x = NonNA)) + geom_point()

vgsales.new <- vgsalesmil

# qqplot of sales #
ggplot(vgsales.new, aes(sample = NA_Sales)) + geom_qq() + geom_qq_line()
# histogram of sales #
ggplot(vgsales.new, aes(x = NA_Sales)) + geom_histogram(boundary = 0, bins
= 50) + ggtitle("Histogram of North American Sales")

# Making variables for top 10 publishers #
vgsales.new$Nintendo <- ifelse(vgsales.new$Publisher == "Nintendo", 1, 0)
vgsales.new$Micro <- ifelse(vgsales.new$Publisher == "Microsoft Game
Studios", 1, 0)
vgsales.new$TakeTwo <- ifelse(vgsales.new$Publisher == "Take-Two
Interactive", 1, 0)
vgsales.new$Sony <- ifelse(vgsales.new$Publisher == "Sony Computer
Entertainment", 1, 0)
vgsales.new$Activision <- ifelse(vgsales.new$Publisher == "Activision", 1,
0)
vgsales.new$Ubisoft <- ifelse(vgsales.new$Publisher == "Ubisoft", 1, 0)
vgsales.new$Bethesda <- ifelse(vgsales.new$Publisher == "Bethesda
Softworks", 1, 0)
vgsales.new$EA <- ifelse(vgsales.new$Publisher == "Electronic Arts", 1, 0)
vgsales.new$Sega <- ifelse(vgsales.new$Publisher == "Sega", 1, 0)

# boxplot for Nintendo #
ggplot(vgsales.new, aes(x = as.factor(Nintendo), y = NA_Sales)) +
geom_boxplot() + xlab("Nintendo")

# logged histogram and QQ plot #
ggplot(vgsales.new, aes(sample = log(NA_Sales))) + geom_qq() +
geom_qq_line()
ggplot(vgsales.new, aes(x = log(NA_Sales))) + geom_histogram() +
ggtitle("Histogram of the log of North American Sales")

# combine the consoles #
table(vgsales.new$Platform)
vgsales.sim <- vgsales.new
```

```
# Makes all playstation consoles the same #
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("PS"="PScon"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("PS2"="PScon"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("PS3"="PScon"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("PS4"="PScon"))
#Makes all Xboxes the same
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("X360"="XBcon"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("XB"="XBcon"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("Xone"="XBcon"))
#Gameboy
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("GB"="Gboy"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("GBA"="Gboy"))
#NES
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("NES"="NESes"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("SNES"="NESes"))
#Wii
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("Wii"="Wiicon"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("WiiU"="Wiicon"))
#Other fo consoles with less than 100 games
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("3DO"="Other"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("DC"="Other"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("GEN"="Other"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("GG"="Other"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("NG"="Other"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("PCFX"="Other"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("SCD"="Other"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("TG16"="Other"))
vgsales.sim$Platform <- revalue(vgsales.sim$Platform, c("WS"="Other"))
table(vgsales.sim$Platform)

# Greater than 100k units sold in North America
vgsalesmodel <- vgsales.sim[which(vgsales.sim$NA_Sales >= 0.1),]

# summarize statistics
summary(vgsalesmodel)
sd(vgsalesmodel$age)
sd(vgsalesmodel$NA_Sales)
sd(vgsalesmodel$Global_Sales)
sort(table(vgsalesmodel$Publisher))
sort(table(vgsalesmodel$Genre))
```

```
# replot some of the earlier relationships #
ggplot(vgsalesmodel, aes(y = log(NA_Sales), x = sqrt(NonNA))) +
geom_point() + ggtitle("Log of NA Sales against Square Root of Other
Sales")
ggplot(vgsalesmodel, aes(y = log(NA_Sales), x = NonNA)) + geom_point() +
ggtitle("Log of NA Sales against Other Sales")

# remove the highly highly successful global games #
final <- vgsalesmodel[-which(vgsalesmodel$Rank <= 10),]

# histogram of age #
ggplot(final, aes(x = age)) + geom_histogram()
# log of NA sales against Non North American sales
ggplot(vgsalesmil, aes(y = log(NA_Sales), x = NonNA)) + geom_point()

# model untransformed data #
vgsales.lm2 <- lm(NA_Sales ~ age + Genre + Nintendo + Micro + TakeTwo +
Sony + Activision + Ubisoft + Bethesda + EA + Sega + Platform + NonNA, data
= final)
summary(vgsales.lm2)
plot(vgsales.lm2, which = 1)
#Cone shaped residuals, problems with a constance in variance

# model log with all factors #
vgsales.lm3 <- lm(log(NA_Sales) ~ age + Genre + Nintendo + Micro + TakeTwo
+ Sony + Activision + Ubisoft + Bethesda + EA + Sega + Platform + NonNA,
data = final)
summary(vgsales.lm3)
anova(vgsales.lm3)

# check diagnostic plots #
plot(vgsales.lm3, which = 1)
plot(vgsales.lm3, which = 2)
plot(vgsales.lm3, which = 4)

# test to remove Sega and Ubisoft from list of included "major publishers"?
vgsales.lm4 <- lm(log(NA_Sales) ~ age + Genre + Nintendo + Micro + TakeTwo
+ Sony + Activision + Bethesda + EA + Platform + NonNA, data = final)
summary(vgsales.lm4)
anova(vgsales.lm3, vgsales.lm4)

#Fail to reject reduced, no longer contains Sega and Ubisoft.
```

```
# test removal of genre
vgsales.lm5 <- lm(log(NA_Sales) ~ age + Nintendo + Micro + TakeTwo + Sony +
Activision + Bethesda + EA + Platform + NonNA, data = final)
summary(vgsales.lm5)
anova(vgsales.lm5, vgsales.lm4)

#Reject reduced, keep Genre

# remove Nintendo
vgsales.lm6 <- lm(log(NA_Sales) ~ age + Genre + Micro + TakeTwo + Sony +
Activision + Bethesda + EA + Platform + NonNA, data = final)
summary(vgsales.lm6)
anova(vgsales.lm6, vgsales.lm4)

#Fail to reject reduced, drop Nintendo

# check diagnostics #
plot(vgsales.lm6, which = 1)
plot(vgsales.lm6, which = 2)
plot(vgsales.lm6, which = 4)
influencePlot(vgsales.lm6)
```