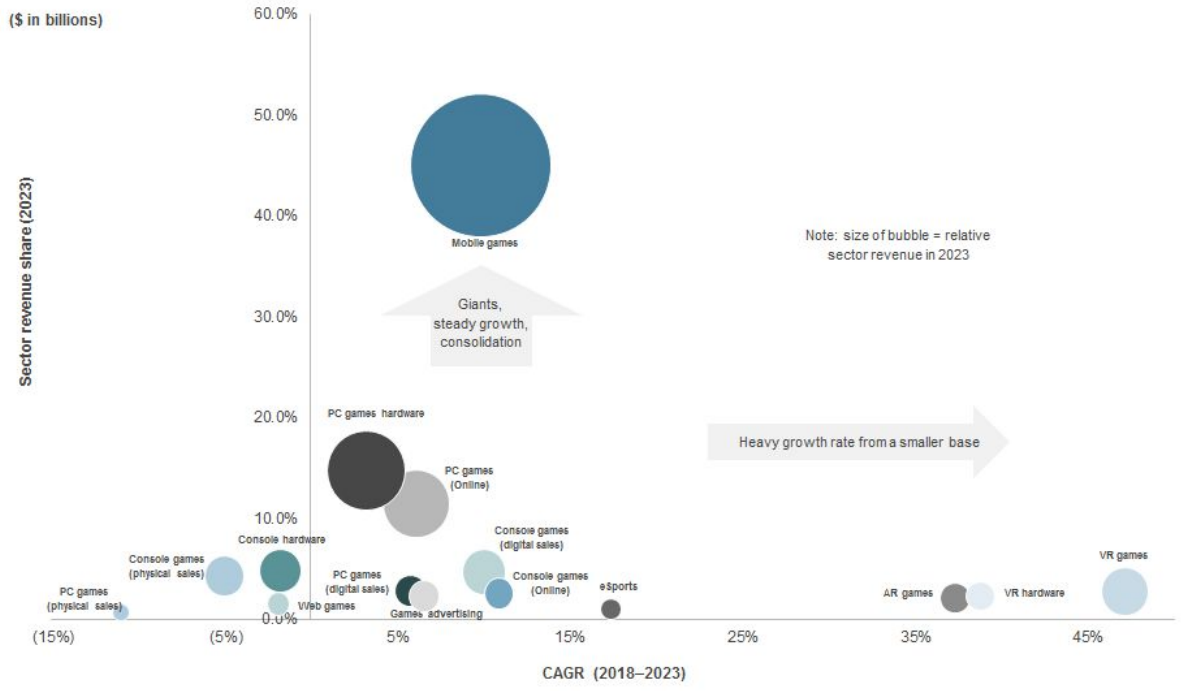
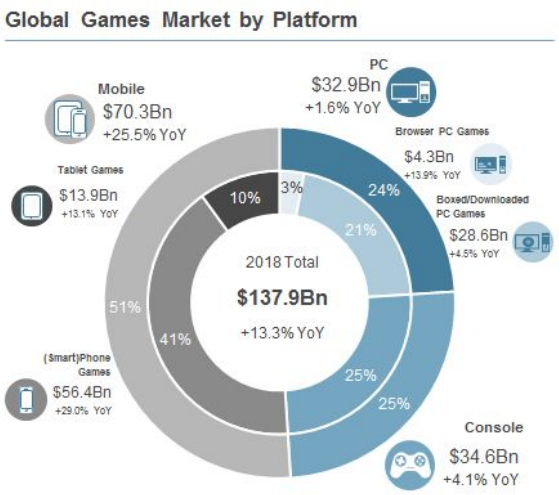
ANLY 545-51-B
Matt Hubley

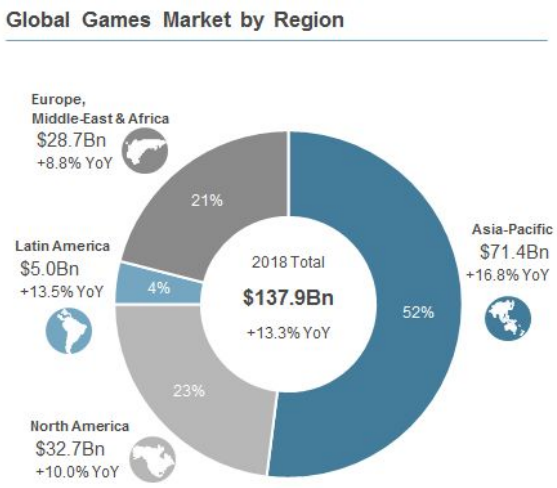
ANALYSIS OF HISTORICAL PHYSICAL VIDEO GAME SALES TO HELP UNDERSTAND FUTURE TRENDS

Avi Kumar, Ting-I Lin, Wei-Hsin Liu, Jatinder Midha, Ha Nul Ryou, Jinqi Zhang

Growth in mobile is far outpacing console and PC platforms; mobile games spending increased to $70Bn in 2018 (26% YoY growth). Mobile is the largest games platform, currently accounts for 51% of total global gaming spend and China, Japan, and South Korea account for nearly half of all global games revenue in 2018. These are high numbers and since the 1980s every child and adult has played or has been interested in video games. There is a lot of datasets to analyze and we believe that given massive shifts from consoles / physical game copies to mobile/electronic subscriptions and eSports we can analyze historical data to forecast major trends which would allow us to predict future areas to invest in the game industry.







From this study, first of all, we are hoping to suggest which publisher could be used as a benchmark for other video game publishers, which they may earn a takeaway to generate a profit. Second of all, we are hoping to identify which genre of a videogame is most popular to customers and attempt to suggest the target genre to generate a profit. Lastly, we are hoping to provide a sales volume of each geographic area (by continents) to suggest which area has a lack of sales and what actions must be taken to enhance a sales volume. On the other hand, where the area has the highest sales volume, in order to maintain its high sales volume, we can suggest what sales strategy has been used and what customers demand are to find the cause of its highest sales volume. Not only maintaining the high sales volume, but we would also like to suggest any predictive analysis to take action for a future to have a growth in profit.

Since 1980, videogames were out in the public for customers to enjoy arcade gaming. Initial publishers were Atari (United States) and Activision (United States). In 1985, the new company named Nintendo was introduced to the world, which was established from Japan. Nintendo can be referred to as a game-changer since most of the games that the company

launched, it was selling millions of games worldwide. One of the most popular games was Super Mario Brothers, and it was released in the same year as Nintendo established its business in the United States. It was the most popular game that has recorded the best-selling game from Nintendo all the time. Since then other publishers such as Electronic Arts (EA), Namco Bandai Games, Sony Computer Entertainment, Sega and many others started to jump into the videogame industry.

The video game industry had steady growth from the 1980s until early 2010. Video games used to have completed stories in its game, which brought interest to customers. Nowadays, if you want to have a complete story of the game, you would need to make a payment, so-called Micro-transactions. Micro-transactions are additional transaction need to be made to continue your game and to purchase items to easily and broadly enjoy the game. Many customers may feel that it is unfair to pay additional payments to enjoy the entire game even though they have purchased a game either electronically or physically. Per personal experience, right after purchasing PlayStation 4, there were numbers of additional payments required and things to be installed to actually enjoy the game. Therefore, the purchased PlayStation 4 was immediately returned, due to a lack of patient and realizing its unfairness. Also, these days many games are primarily made to be played with other people, which is known as multiplayer mode. It is not wrong to have such a mode, but giving options to people to enjoy their game on their own must be given.

In order to overcome the above issues that have been addressed to video game publishers, they may need to take a look at which genre of the game has the highest demand and how to effectively have customers enter into a video gaming experience. First of all, video game

publishers must consider what kind of innovation they can bring to attract new customers. For example, Nintendo has launched its new device “Nintendo Wii”, which enabled customers to interact with video games more actively by recognizing a person’s movement to play the game. In addition, PlayStation launched the Virtual Reality (VR) machine, which enabled customers to more actively interact with video games than any other video gaming systems than before.

The data was discovered on Kaggle database. The data was generated by a data scrape of vghcartz.com which is a video game sales tracking website that provides weekly sales figures of console software and hardware by region. The data contains a list of video games with sales greater than 100,000 copies.

The fields of the data include the following:

Rank - Ranking of overall sales

Name - The games name

Platform - Platform of the games release (i.e. PC,PS4, etc.)

Year - Year of the game's release

Genre - Genre of the game

Publisher - Publisher of the game

NA_Sales - Sales in North America (in millions)

EU_Sales - Sales in Europe (in millions)

JP_Sales - Sales in Japan (in millions)

Other_Sales - Sales in the rest of the world (in millions)

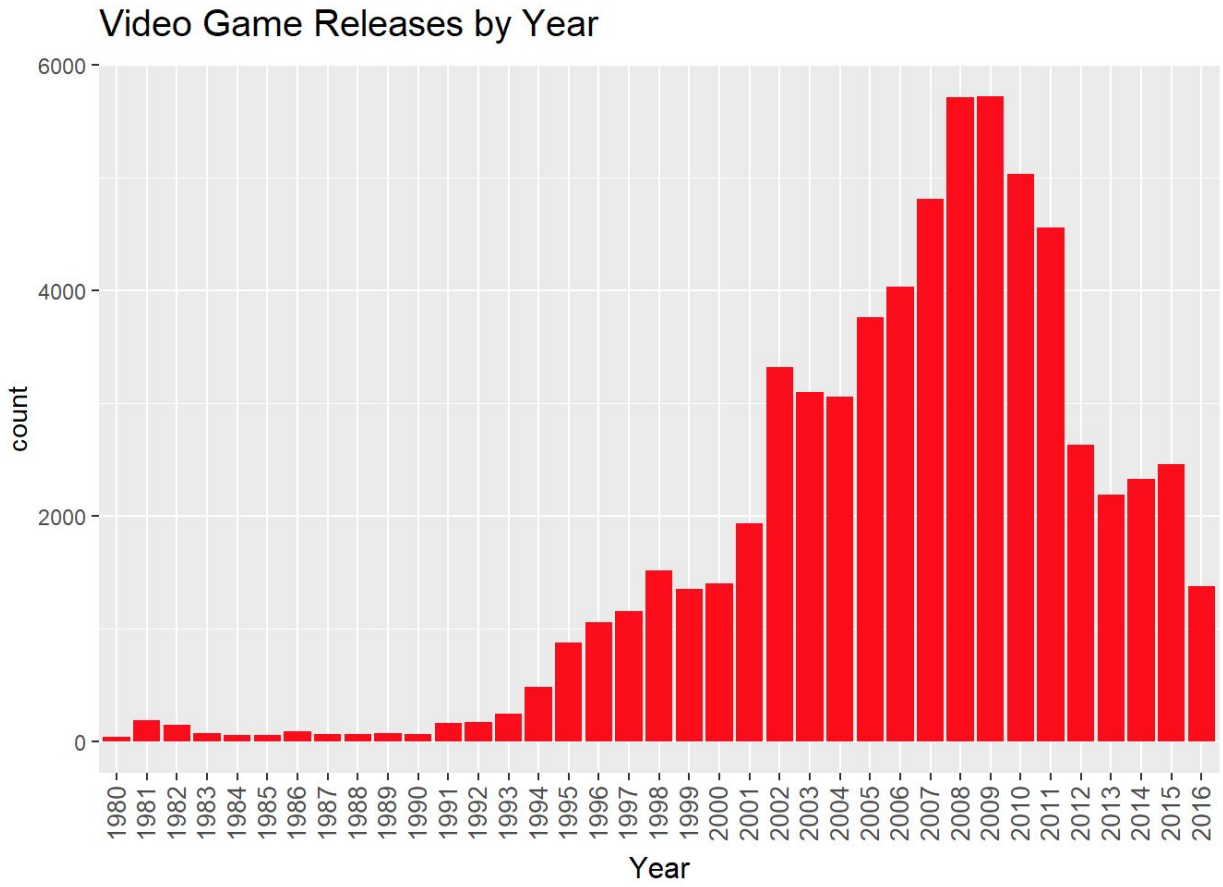
Global_Sales - Total worldwide sales.

Before we start analyzing data, exploratory data analysis is necessary for understanding dataset variables. It is also helpful to analyze information that is hidden or missing in rows or column format by summarizing and interpreting the data before extensive data analysis. Once the exploratory analysis is conducted, we can gain insights about the games and the publishers with sales in primary viewpoint.

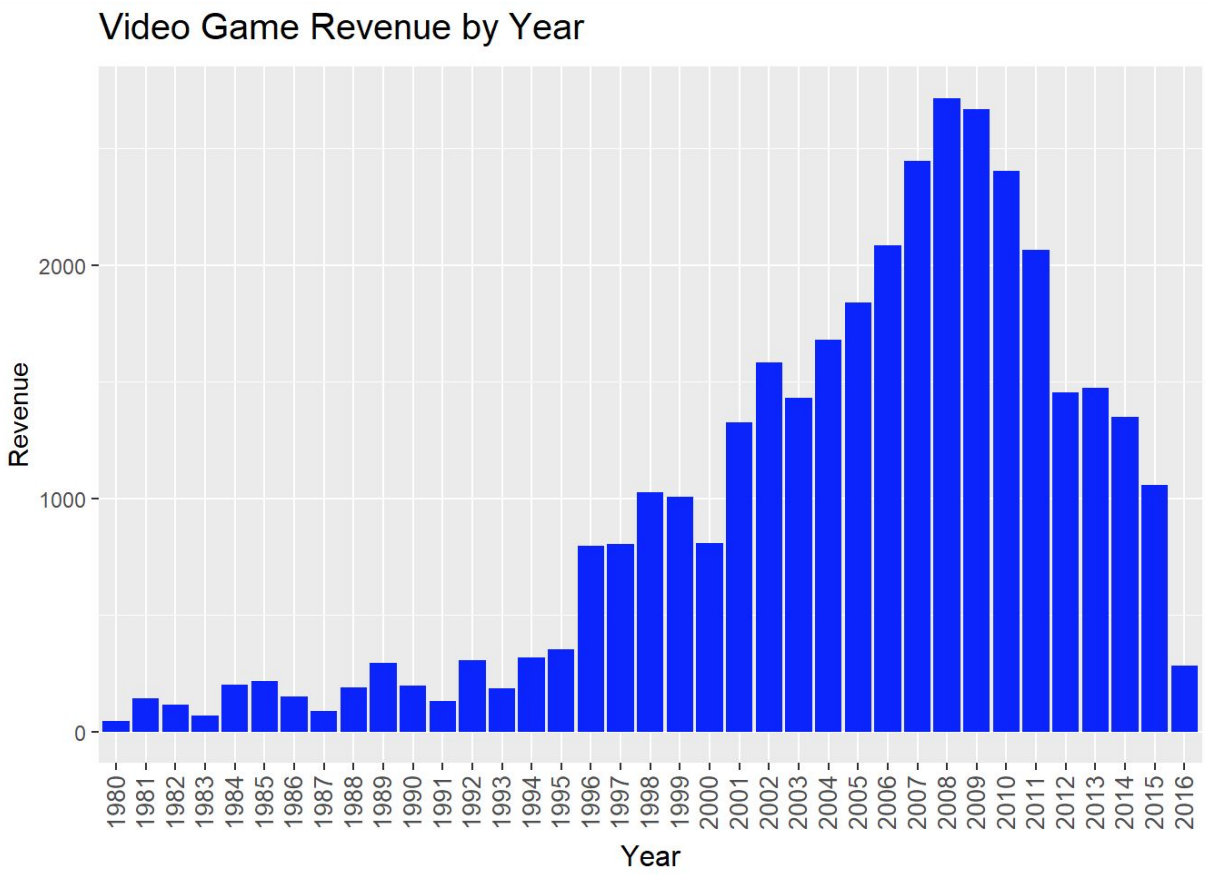
By using the str() and summary() functions, we understood the type variable types in the dataset. We transformed factor variables into character variables and converted and year of release and user score to a numeric value. Also, we discovered that the year data is not consistent for 2016 and onwards. Therefore N/A, 2017 and 2020 was taken out from the dataset.

We explored the data and created charts that can conclude most of the objectives:

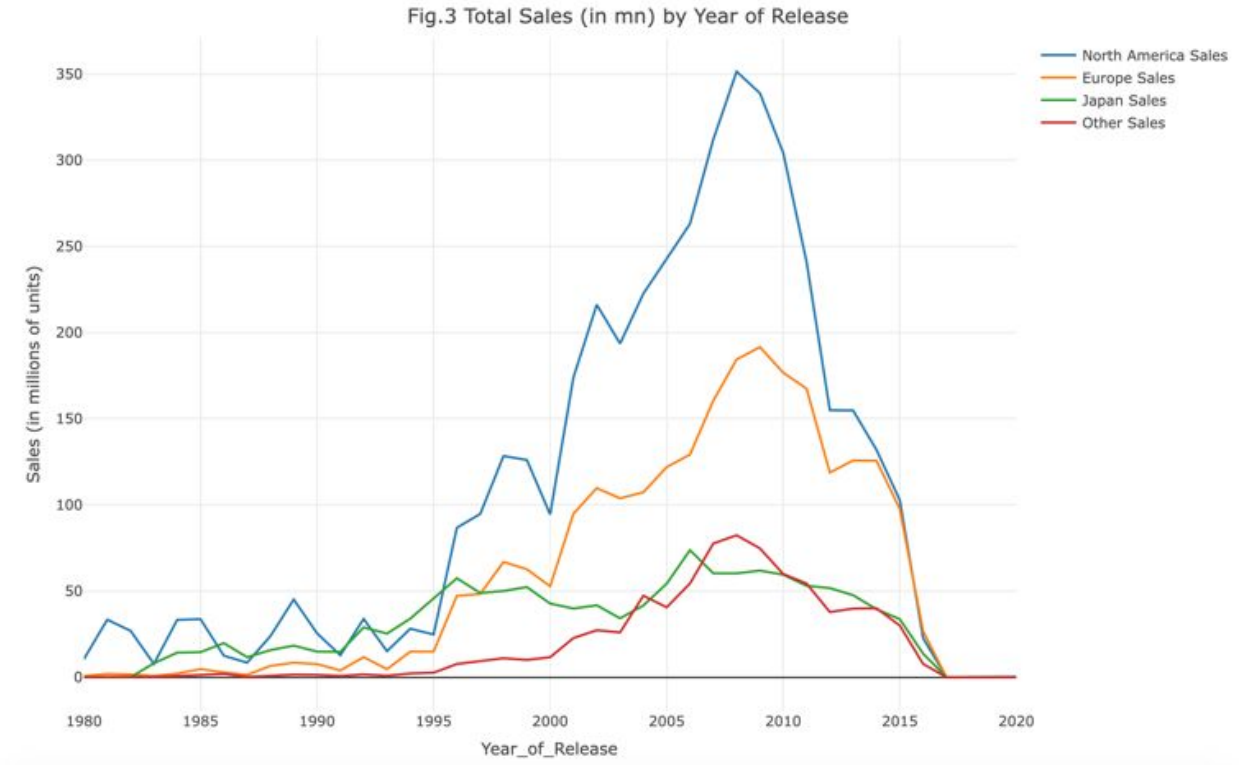
1.1 Number of Releases by Year – The releases are distributed normally but left-skewed from the year 1980 to 2016, while 2008 and 2009 are the peak of the count of video game releases. From 1980 to 1992, there were constant low releases, and from 1993, the releases in each year kept increasing until 2008. After the peak in 2009, the releases started to decline until now. There was a significant drop from 2011 to 2012, which may cause by a change in the gaming trend, such as more demand in mobile games.



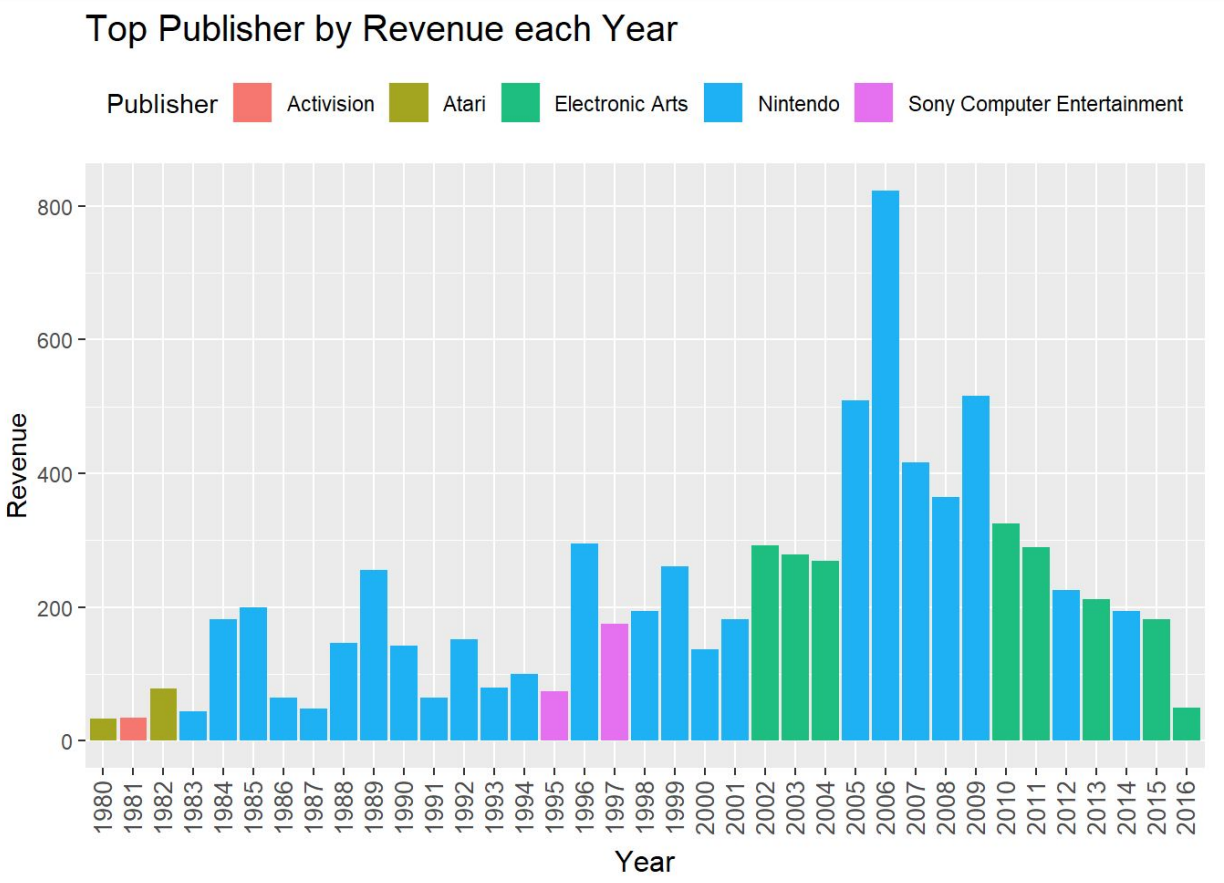
1.2 Revenue by Year – The revenue is also distributed normally but left-skewed. 2008 and 2009 are also the peak of the revenue, just like a number of releases. The revenue also has a drop from 2011 to 2012. Overall, Revenue by Year and Number of Releases by Year have an identical shape so we can possibly conclude that Number of Releases by Year and Revenue by Year are correlated.



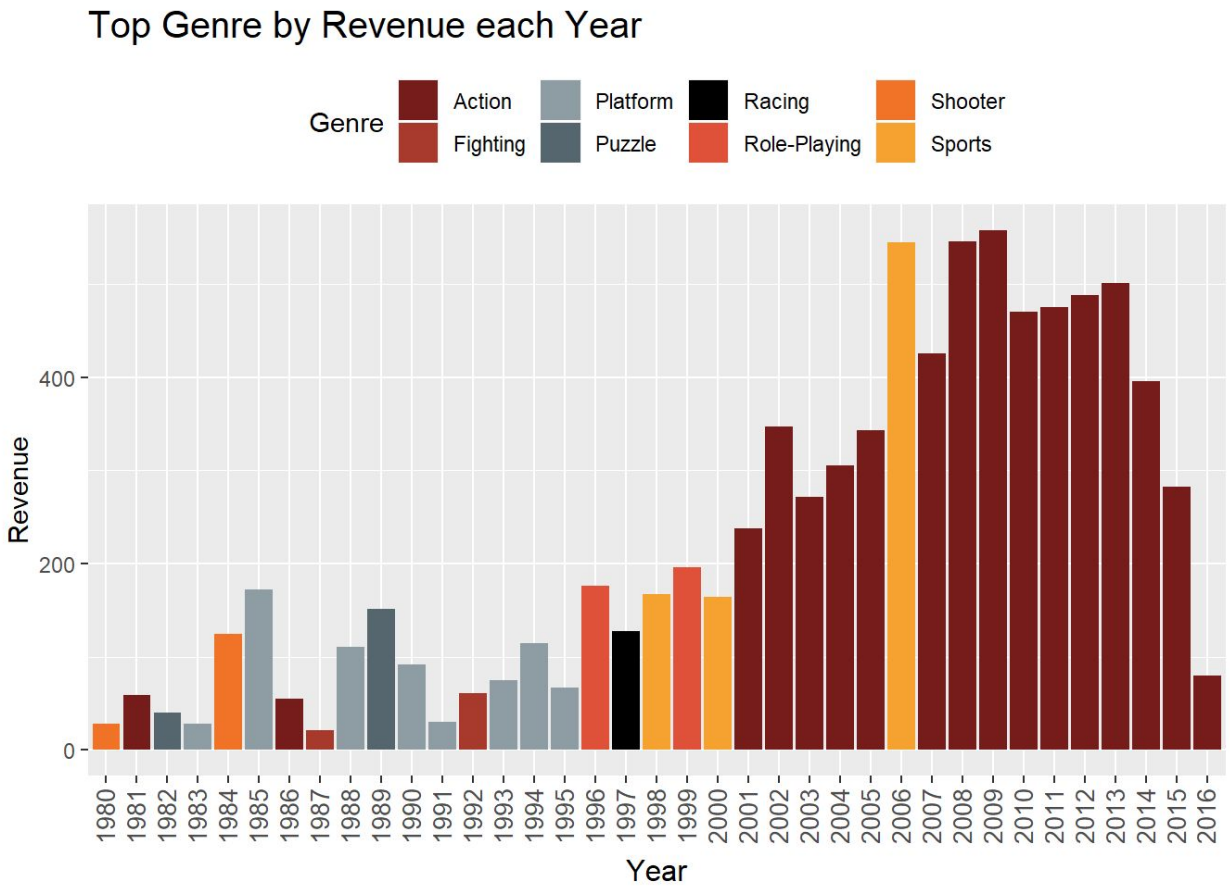
2. Sales by Year in each Region – first, we can see that the shape of Sales by Year is also similar to Revenue by Year and Number of Releases by Year. Japan has a slightly different trend; comparing to other regions, Japan as a more constant sales from 1995 to 2015. Japan's sales have almost reached its peak in 1995, whereas other regions' sales were just started to rise. It indicates an early matured market in Japan and a good explanation is that Japan has had a more developed gaming industry than other regions. With a greater population, North America has the greatest sales, which is even more than half of Japan and Other region's sales.



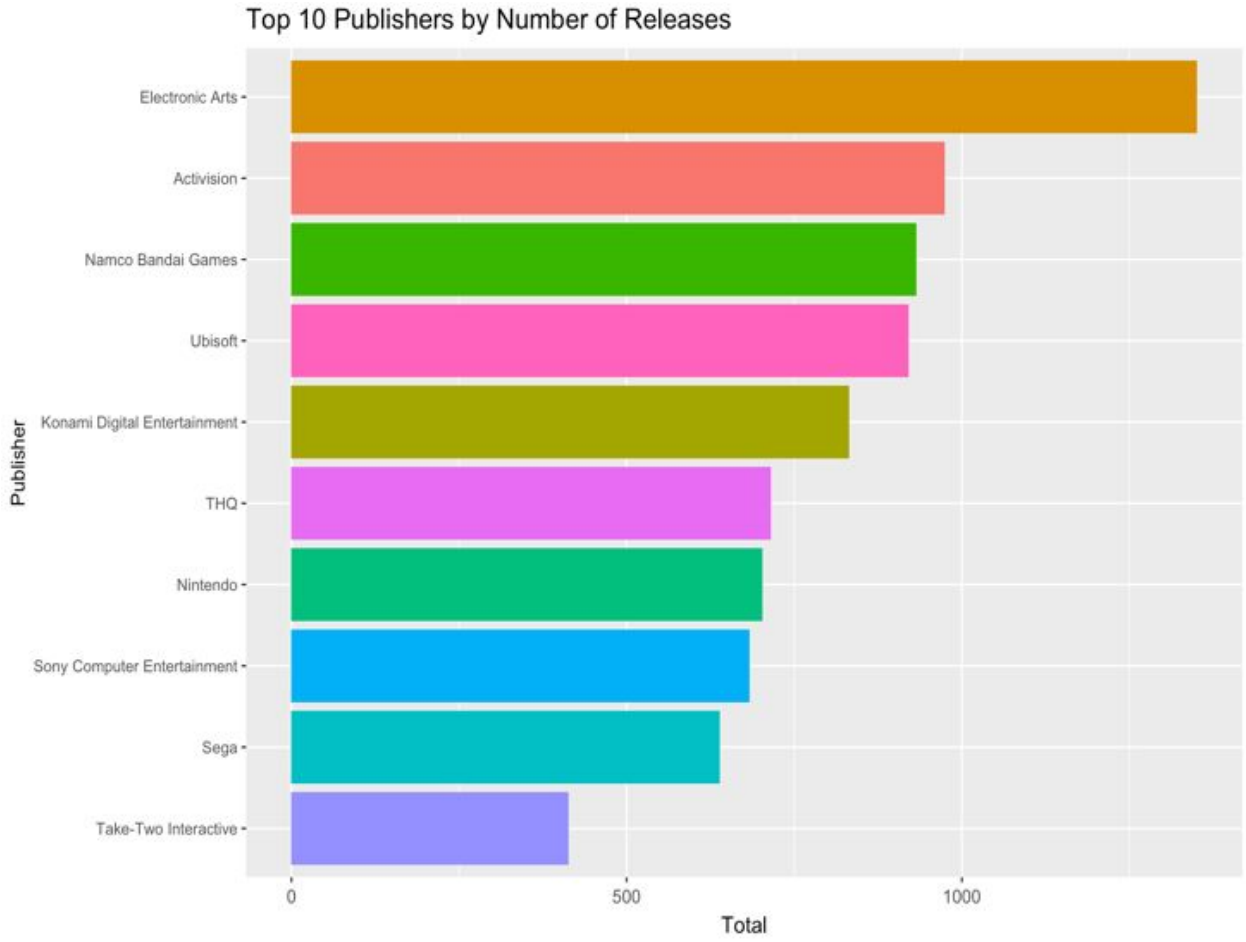
3. Top Publisher by Revenue per Year – Nintendo is the publisher with the highest number of highest revenues, followed by Electronic Arts. Nintendo has the highest revenue in 2006, the year that it published Wii.



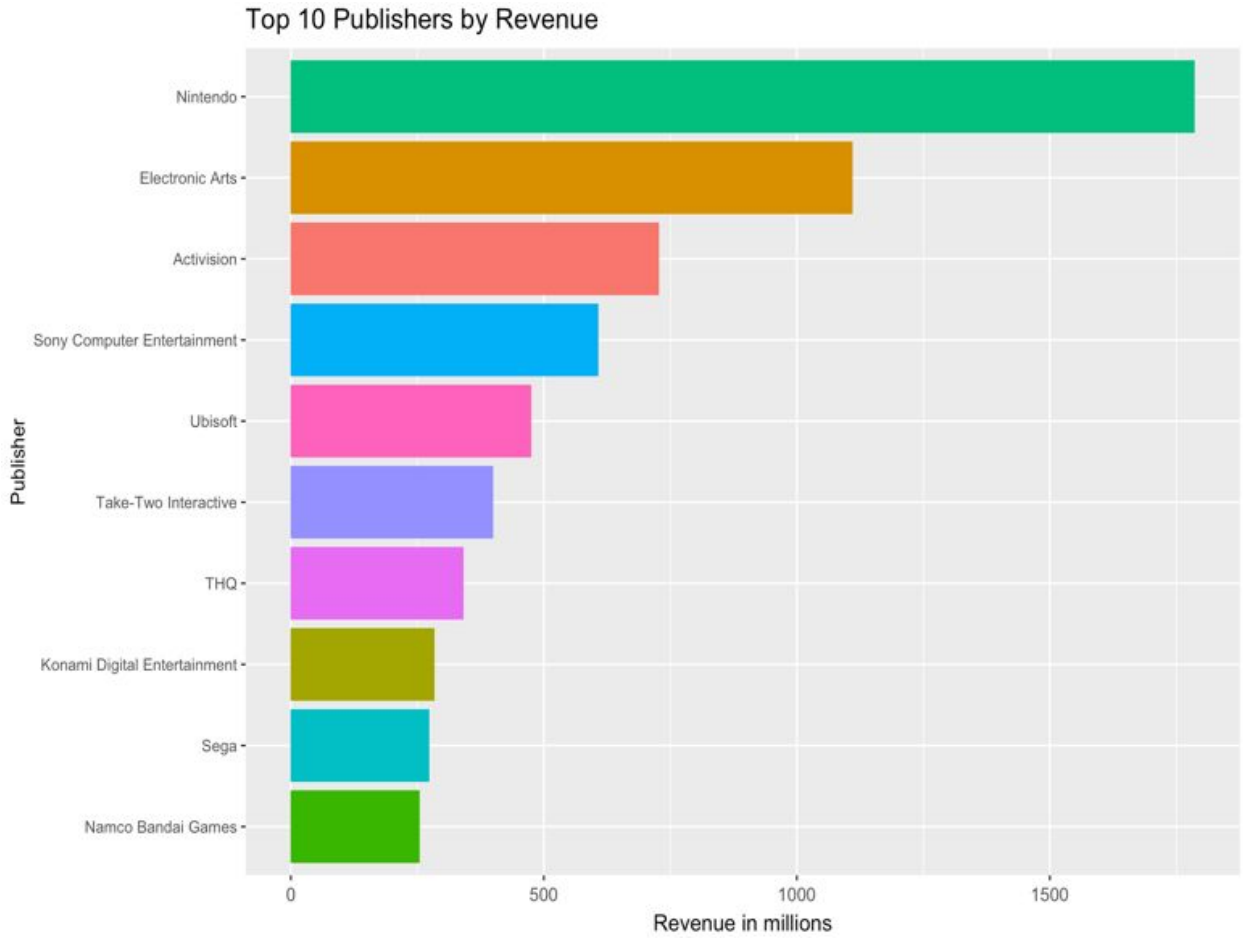
4. Top Genre by Revenue per Year – Before 2001, the top genres by revenue were random among the eight genres, but starting from 2001, Action is the top genre except for 2006. In 2006, Sports is the top genre, which may also due to the release of Wii Sports that year.



5.1 Top 10 Publishers by Number of Releases – Electronics Art is the top publisher. It published 500 more than the second publisher, Activision.

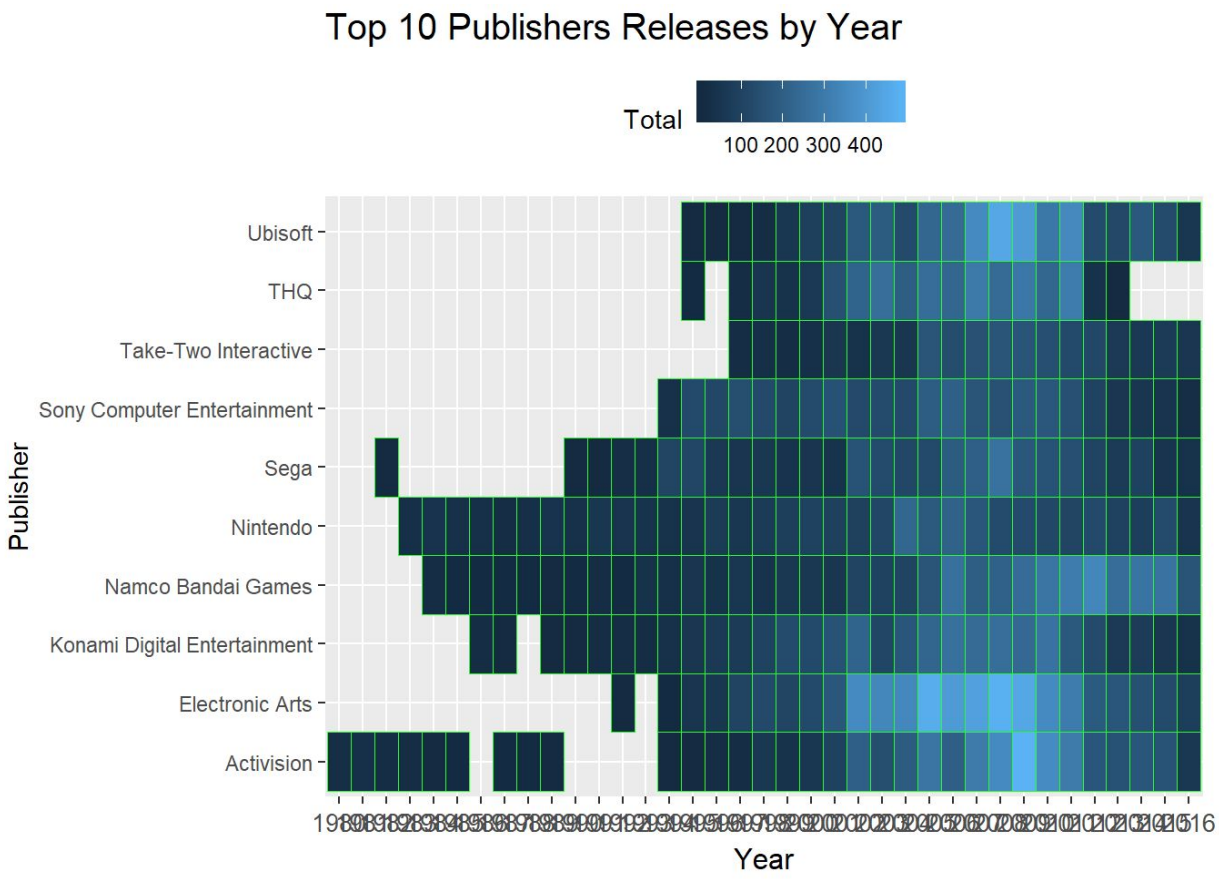


5.2 Top 10 Publishers by Revenue – Nintendo ranks 1 instead of Electronics Art, which becomes second. Nintendo exceeds Electronic Arts by 2 billion and accounts for 20% of all publishers' revenue.

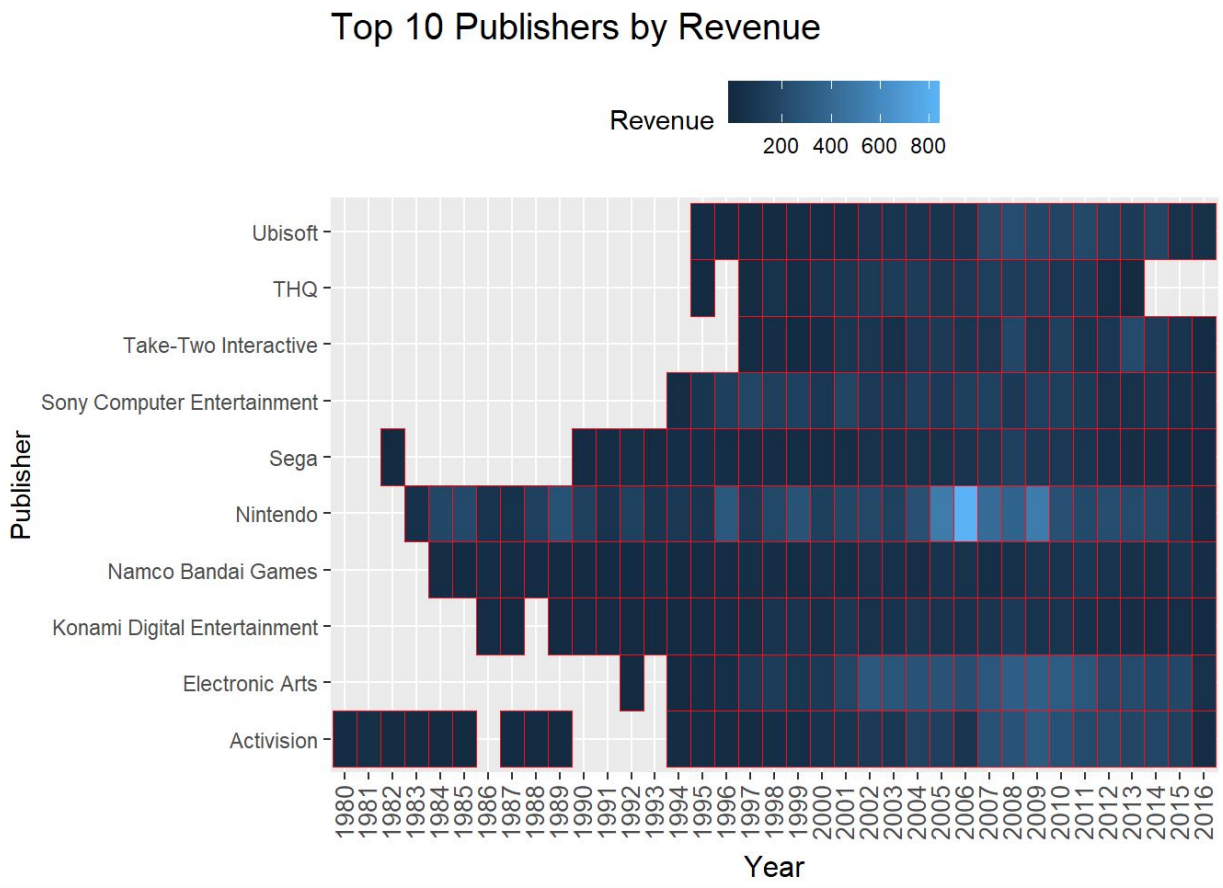


There is a change in the position from the list by number by releases to number by revenue. Nintendo is top 1 with almost 21% of the overall revenue, whereas Electronics Arts being top 2 with nearly half of the revenue of the Nintendo. Nearly 70% of the overall revenue is generated by the top 10 publishers.

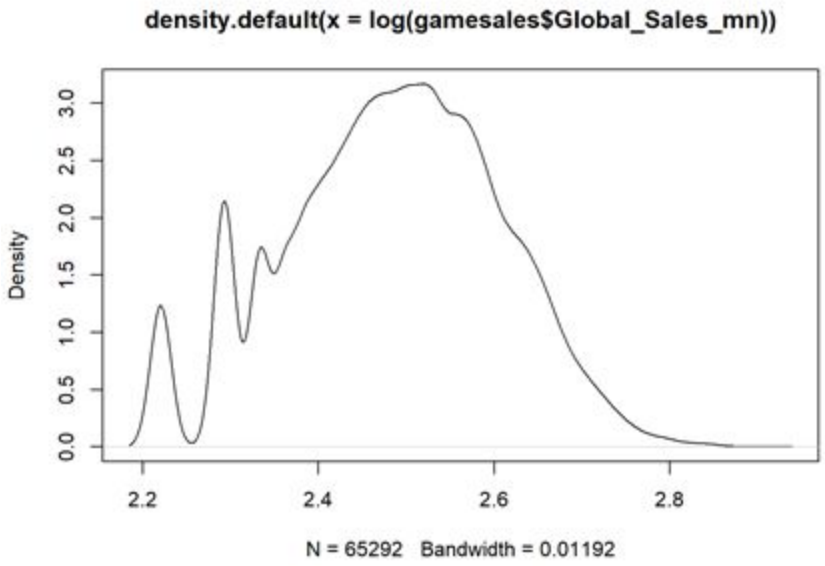
6. Top 10 Publishers' Number of Releases by Year – Activision is the oldest in the market and is considered as a veteran. EA is a latecomer and has the highest number of releases between 2002 and 2011.

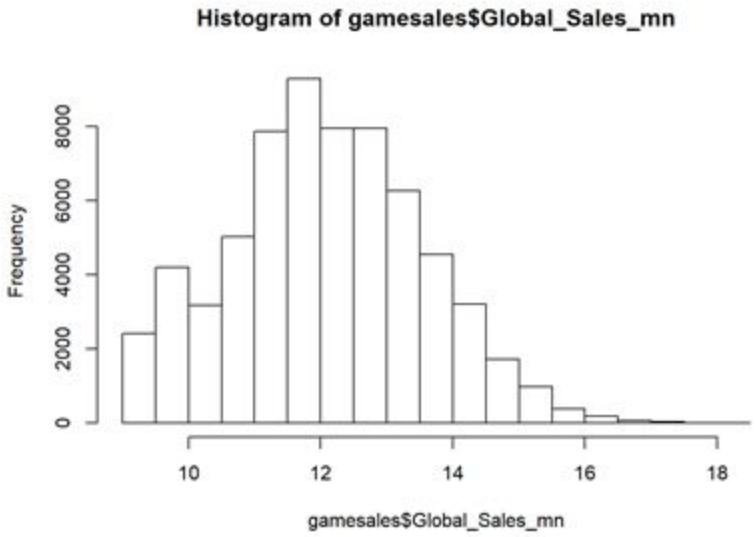


7. Top 10 Publishers' Revenue by Year – Nintendo has an outstanding year in 2006, which could be an outlier. Nintendo, Electronic Arts, and Activision had relatively good revenues in the last 10 years.



In order to understand what factors influence global sales of video games, we ran a regression model. Based on the distribution plot and histogram below, we could see that the distribution is a little rightly skewed, but almost normal distributed. Since the data for global sales are normally distributed, we could directly use this dataset to do analysis without transformation. The mean global sales of the video game was $12.08 million, the minimum global sales was $9.21 million, and the maximum global sales was $ 18.23 million.





In our regression model, the dependent variable was global sales of the video game. Independent variables were video game platform, genre, publisher, region, and ranking. Based on the result below, we could conclude that platform, genre, publisher, and ranking were all significant at the .05 level. Therefore, we could say that video game platform, genre, publisher, and ranking had a significant impact on video game global sales.

```
                      Df Sum Sq Mean Sq   F value Pr(>F)
gamesales$Platform     30  16491     550    8395.1 <2e-16 ***
gamesales$Genre        11   5184     471    7197.1 <2e-16 ***
gamesales$Publisher   576  34463      60     913.7 <2e-16 ***
gamesales$Region         3      0       0       0.0      1
gamesales$Rank           1  79279   79279 1210747.6 <2e-16 ***
Residuals            64670   4235       0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In addition to the regression model, we used the function “tapply” to analyze the means of several factors. In “Genre” category, the mean global sales among different genres was actually very close, hovering around 12 million. Turned out that there is not a lot of variation in our data in the genre variable.

```
tapply(gamesales$Global_Sales_mn , gamesales$Genre, mean)
```

```
##       Action    Adventure     Fighting         Misc     Platform
##     12.16132     11.08525     12.26776     12.00187     12.54508
##       Puzzle       Racing Role-Playing      Shooter   Simulation
##     11.68019     12.18790     12.17086     12.39884     11.98102
##       Sports     Strategy
##     12.32016     11.53598
```

Secondly, we turned our attention to the variable, Region. We observed a similar pattern as before: there was not a lot of variation in terms of regions. Every region had nearly around $12 million, sale.

```
tapply(gamesales$Global_Sales_mn , gamesales$Region, mean)
```

```
##    EU_Sales    JP_Sales    NA_Sales Other_Sales
##    12.07855    12.07855    12.07855    12.07855
```

Thirdly, we did see a lot of variation among the variable, publishers. Some publisher earned around $9 million global sales, while others could earn as high as to $12 million global sales. A difference of 3 million sales! For example, Nintendo, a well-known publisher, had around 12 million global sales in our dataset. We could that Nintendo is a preferred publisher in terms of global sales of video games.

In summary, we learned: Nintendo was the leading company in the video game market, which can be used as a benchmark, action was the top genre video game that generated most revenue out of other genre and North America had the highest video game sales volume.

We look to the future and see there are a lot of data points providing information that can be useful for us as video game data scientists and maybe working for publishers. We expect we would do more work on eSports given that may be a field which is driving new growth of sales as well as trying to get hold of mobile data sales and microtransactions sales since they may a difference between physical copy sales. We would also look to bifurcate games now in types of players such as casual players, midcore players and hardcore players - games now are more focused on different demographics and no longer are targeting just people interested in action, but action hardcore where they can extract more money for longers.

References

(n.d.). Retrieved from https://www.nintendo.com/corp/history.jsp
Is the Video Game Industry Dying? (2018, November 17). Retrieved from https://www.darkspawngaming.com/is-the-video-game-industry-dying/
8 Reasons The Video Game Industry Is Going To Crash Again. (n.d.). Retrieved from https://www.fraghero.com/8-reasons-the-video-game-industry-is-going-to-crash-again/

R SYNTAX

```
' Let us start with cleaning our data. The data is not consistent for
2016 onwards. We have taken out 2017, 2020 and N/A
This was followed by converting the Year of Release and User score
to a numeric value
'

library(ggplot2)
library(dplyr)
library(DT)
library(tidyr)
library(wesanderson)
library(plotly)

gamesales <- vgsales
gamesales <- gamesales[!(gamesales$Year %in% c("N/A", "2017", "2020")),]
gamesales$Year_of_Release <- as.numeric(as.character(gamesales$Year))
gamesales$Rank <- as.numeric(as.character(gamesales$Rank))
gamesales$Global_Sales <- as.numeric(as.character(gamesales$Global_Sales))
'gamesales$User_Score <- as.numeric(as.character(gamesales$Rank))'
gamesales$Genre <- as.character(gamesales$Genre)
gamesales <- gamesales %>% gather(Region, Revenue, 7:10)
gamesales$Region <- factor(gamesales$Region)

## STEP 1 : DATA EXPLORATION (Descriptive Analysis)

# Let us define some colours and themes for our charts
mycolors <- c("#771C19", "#AA3929", "#8E9CA3", "#556670", "#000000", "#E25033",
"#F27314", "#F8A31B", "#E2C59F", "#B6C5CC")

theme_1 <- function() {

  return(theme(axis.text.x = element_text(angle = 90, size = 10, vjust = 0.4), plot.title =
element_text(size = 15, vjust = 2),axis.title.x = element_text(size = 12, vjust = -0.35)))

}
```

```
theme_2 <- function() {

  return(theme(axis.text.x = element_text(size = 10, vjust = 0.4), plot.title = element_text(size =
15, vjust = 2),axis.title.x = element_text(size = 12, vjust = -0.35)))

}



# 1.1 Breaking down the number of releases by year and revenue by year

ggplot(gamesales, aes(Year)) +
  geom_bar(fill = "Red") +
  theme_1() +
  ggtitle("Video Game Releases by Year")

revenue_by_year <- gamesales %>%
  group_by(Year) %>%
  summarize(Revenue = sum(Global_Sales))

ggplot(revenue_by_year, aes(Year, Revenue)) +
  geom_bar(fill = "Blue", stat = "identity") +
  theme_1() +
  ggtitle("Video Game Revenue by Year")



# 1.2 Let us look at the sales broken down by Regions
'install.packages("plotly")
library("plotly")'
gamesales_2 <- gamesales
gamesales_2$Year_of_Release <- as.numeric(as.character(gamesales_2$Year))
Sales_NA <- gamesales_2%>% select(Year_of_Release,NA_Sales) %>%
  group_by(Year_of_Release)%>%
  summarise(Sum_NA_Sales=sum(NA_Sales))
Sales_EU <- gamesales_2%>% select(Year_of_Release,EU_Sales) %>%
  group_by(Year_of_Release)%>%
  summarise(Sum_EU_Sales=sum(EU_Sales))
Sales_JP <- gamesales_2%>% select(Year_of_Release,JP_Sales) %>%
  group_by(Year_of_Release)%>%
```

```
  summarise(Sum_JP_Sales=sum(JP_Sales))
Sales_OH <- gamesales_2%>% select(Year_of_Release,Other_Sales) %>%
  group_by(Year_of_Release)%>%
  summarise(Sum_OH_Sales=sum(Other_Sales))

Sales_evo <- Reduce(function(x,y)
merge(x,y,all=TRUE,by="Year_of_Release"),list(Sales_NA,Sales_EU,Sales_JP,Sales_OH))

plot_ly(data=Sales_evo,x=~Year_of_Release)%>%
  add_trace(y=~Sum_NA_Sales,name="North America Sales",mode="lines",type = 'scatter')
%>%
  add_trace(y=~Sum_EU_Sales,name="Europe Sales",mode="lines",type = 'scatter') %>%
  add_trace(y=~Sum_JP_Sales,name="Japan Sales",mode="lines",type = 'scatter') %>%
  add_trace(y=~Sum_OH_Sales,name="Other Sales",mode="lines",type = 'scatter') %>%
  layout(title = "Fig.3 Total Sales (in mn) by Year of Release",
         yaxis = list(title="Sales (in millions of units)"))

# 1.3 Top Publisher by Revenue each Year

top_publisher_year <- gamesales %>%
  group_by(Year, Publisher) %>%
  summarize(Revenue = sum(Global_Sales)) %>%
  top_n(1)

datatable(top_publisher_year)

ggplot(top_publisher_year, aes(Year, Revenue, fill = Publisher)) +
  geom_bar(stat = "identity") +
  ggtitle("Top Publisher by Revenue each Year") +
  theme_1() +
  theme(legend.position = "top")

# 1.4 Top Genre by Revenue each year
top_genre <- gamesales %>%
  group_by(Year, Genre) %>%
  summarize(Revenue = sum(Global_Sales)) %>%
  top_n(1)

datatable(top_genre)
```

```
ggplot(top_genre, aes(Year, Revenue, fill = Genre)) +
  geom_bar(stat = "identity") +
  ggtitle("Top Genre by Revenue each Year") +
  theme_1() +
  theme(legend.position = "top") +
  scale_fill_manual(values = mycolors)

#1.5 Top 10 publishers by


length(unique(gamesales$Publisher))
#there are 577 Pubishers
by_publishers <- gamesales %>% group_by(Publisher) %>% summarize(Total = n()) %>%
arrange(desc(Total)) %>% head(10)
by_publishers$Percentage <- by_publishers$Total/dim(gamesales)[1] * 100
by_publishers$Publisher <- factor(by_publishers$Publisher)

datatable(by_publishers, filter = "none")

ggplot(by_publishers, aes(reorder(Publisher, Total), Total, fill = Publisher)) +
  geom_bar(stat = "identity") +
  ggtitle("Top 10 Publishers by Number of Releases") +
  theme(legend.position = "none") +
  xlab("Publisher") +
  theme_2() +
  coord_flip()

#EA IS THE TOP PUBLISHER


top_publishers <-gamesales %>% group_by(Publisher) %>% summarize(Revenue =
sum(Global_Sales),
            Percentage = Revenue/sum(gamesales$Global_Sales) * 100) %>%
arrange(desc(Revenue)) %>% head(10)

top_publishers$Publisher <- factor(top_publishers$Publisher)

datatable(top_publishers)
```

```
ggplot(top_publishers, aes(reorder(Publisher, Revenue), Revenue, fill = Publisher)) +
  geom_bar(stat = "identity") +
  ggtitle("Top 10 Publishers by Revenue") +
  theme(legend.position = "none") +
  xlab("Publisher") +
  ylab("Revenue in millions") +
  theme_2() +
  coord_flip()

# There is change in the positions from the list by number of releases.
# Nintendo is Top 1 with almost 21% of the overall revenue
# EA being Top 2 with nearly half the revenue of the Nintendo.
# Nearly staggering 70% of the overall revenue is generated by the Top 10 publishers

# 1.6 Let us see how the top 10 publishers have grown bt year in terms of releases


top_publishers <- gamesales[gamesales$Publisher %in% by_publishers$Publisher,] %>%
group_by(Publisher, Year) %>% summarize(Total= n())

top_publishers$Publisher <- factor(top_publishers$Publisher)

ggplot(top_publishers, aes(Year, Publisher, fill = Total)) +
  geom_tile(color = "green") +
  theme_2() +
  ggtitle("Top 10 Publishers Releases by Year") +
  xlab("Year") +
  theme(legend.position = "top")


# EA is one of the late comers !!
# Activision is in the market since 1980 and is considered a Veteran
# EA has highest number of releases between 2002 and 2011.
# THQ has not released any games from 2014


# 1.7 Lets check how top 10 publishers have grown over the years in terms of revenue
```

```
top_publishers <- gamesales[gamesales$Publisher %in% by_publishers$Publisher,] %>%
  group_by(Publisher, Year) %>%
  summarize(Revenue = sum(Global_Sales))

top_publishers$Publisher <- factor(top_publishers$Publisher)

ggplot(top_publishers, aes(Year, Publisher, fill = Revenue)) +
  geom_tile(color = "red") +
  theme_1() +
  ggtitle("Top 10 Publishers by Revenue") +
  xlab("Year") +
  theme(legend.position = "top")


# Nintendo had great year in 2006, it is a pretty clear outlier
# Nintendo, EA and Activision had good run in last 10-15 years in terms of revenue.




# lET run some regression
#install.packages('moments')
#library('moments')
gamesales$Global_Sales_mn <-  log(gamesales$Global_Sales*1000000)
plot(density(log(gamesales$Global_Sales_mn)))
plot(hist(gamesales$Global_Sales_mn))
summary(gamesales$Global_Sales_mn)
model_1 <- aov(gamesales$Global_Sales_mn ~ gamesales$Platform + gamesales$Genre +
gamesales$Publisher + gamesales$Region + gamesales$Rank)
summary(model_1)
qqnorm(model_1$residuals)
qqline(model_1$residuals)
tapply(gamesales$Global_Sales_mn , gamesales$Genre, mean)
# This tells us that the Genre, even though is significant, doesn't actually impact the Global
Sales. All the genres almost have an equal chance
# of success
tapply(gamesales$Global_Sales_mn , gamesales$Region, mean)
tapply(gamesales$Global_Sales_mn , gamesales$Publisher, mean)
```

#On the Other hand, we can see how Nintendo is a preferred publisher wiht a strong relationship with the Global Sales

'model_2 <- aov(gamesales$Revenue ~ gamesales$Platform + gamesales$Genre + gamesales$Publisher + gamesales$Region + gamesales$Rank + gamesales$Global_Sales)'
'shapiro.test(model$residuals)'

scatterplot(gamesales$Global)