**Video Games Sales Analysis-Factors which affect sales**

**Project Abstract:**

Video games are a type of electronic game or software that can be run and played on a computing device, such as a personal computer, game console or mobile device. Different platforms exist besides the personal computer such as the Playstation 4, XBox One and Nintendo Switch which can run different video game software. Recently there has also been an emergence of certain video games which can only be run on mobile devices thus creating a new subcategory of video games. Since its conception in the 1970s, current video games have advanced significantly, offering photorealistic graphics and realism to extraordinary degrees.

Video game design and development is currently a billion-dollar industry spanning multitudes of studios and independent developers across many countries. The video game market is expected to be worth over 20 billion U.S dollars by 2020 and the industry is expected to continue to grow. While it has high penetration rates among children, increasingly more adults have been found to be playing video games, in part due to their higher spending power. Video games are playable across multiple platforms, with mobile phones currently seeing the greatest number of users followed by PCs and then various consoles and nearly every household having at least one gaming device.

The purpose of this project is to analyze video game sales across platforms, taking into account various factors and investigate how these factors affect sales. The dataset used is by GregorySmith and is obtained from Kaggle.com. Link to the datasets will be placed in the appendix. The dataset contains a list of video games with global sales greater than 100,000 copies and contains a list of over 16,500 games. Sales volumes are listed in millions. The dataset only contains games released between the years 1980 to 2016.

Values contained include:

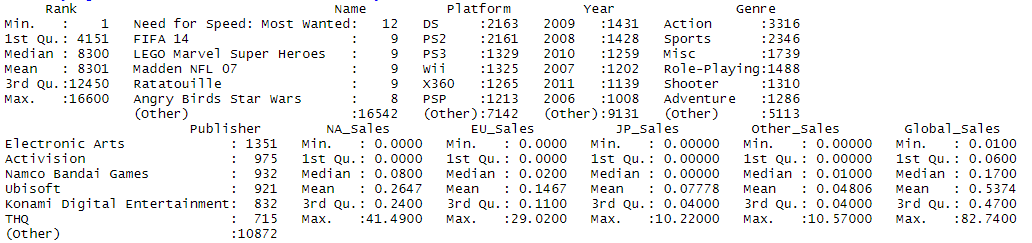
* Rank - Ranking of overall sales
* Name - The game’s name
* Platform - Platform of the games release (i.e. PC, PS4, etc.)
* Year - Year of the game's release
* Genre - Genre of the game
* Publisher - Publisher of the game
* NA\_Sales - Sales in North America (in millions)
* EU\_Sales - Sales in Europe (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.

Furthermore, a journal article titled “Value Creation in the Video Game Industry: Industry Economics, Consumer Benefits, and Research Opportunities” will be used for further study. The website vgchartz.com contains a database of all video games released to date and can be used to create further datasets. R scripting was used for visualization and organizing the data for observation and study in RStudio. The code used is included in the appendix.

The data will be used to obtain insight on video game sales volumes and how different categories such as platform, genre and region might affect sales The data will also be used to study specific trends in video games over the years, for example certain genres may display increase in sales in later years. Further study can be conducted by obtaining relevant data on sales of video game platforms. Graphs and charts have been used to display trends in video game sales throughout the years.

**Summary Statistics**

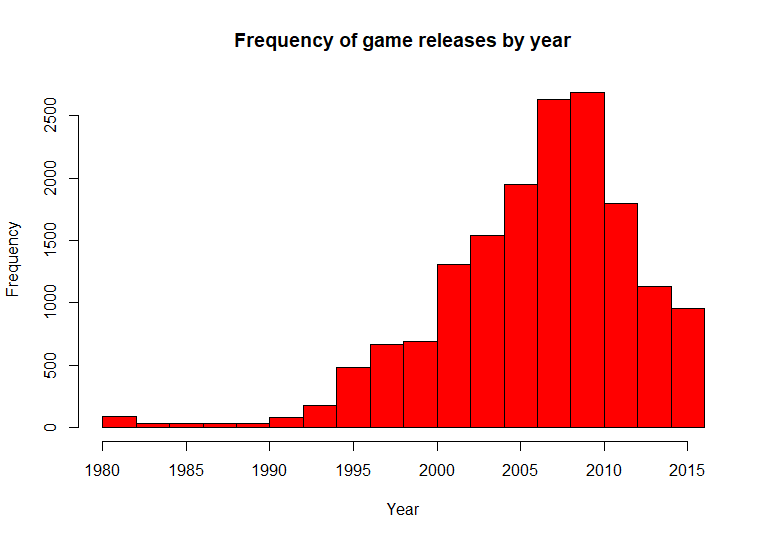
Here are the summary statistics of the dataset:



We can see from the above statistics that the total number of games listed, as seen from the rank are 16600. Need for Speed: Most Wanted was listed 12 times in the dataset, this is likely due to the game being released in multiple platforms in different years. The same can inferred for FIFA 14 which is the second most listed game in the dataset. The Nintendo DS has 2163 entries in the dataset, followed by the PS2 with 2161 entries. From this it can be inferred that these two are likely the greatest selling platforms and would account for a good portion of the sales volumes. The PS3 is the third highest selling platform of choice, but with the number of entries being 1329, there is certainly a big dip compared to the DS and PS2. 2009 has 1431 entries meaning this is likely when video game sales were at its peak. Action has 3316 entries; hence it can be assumed that it is most popular genre for video games.

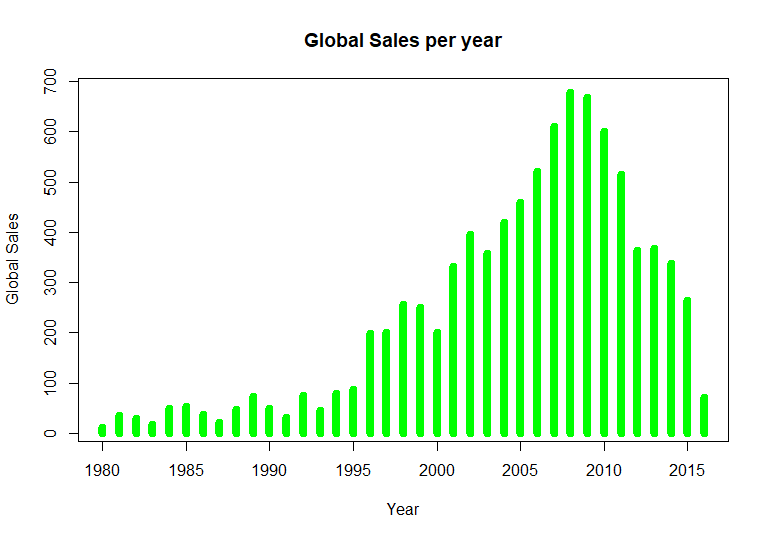
Data cleaning involved removing missing values for game titles which contained entries with N/A. There were also some values for 2017 and 2020 for games not yet released resulting in outliers. These values were also removed.

**Frequency of game releases by year**



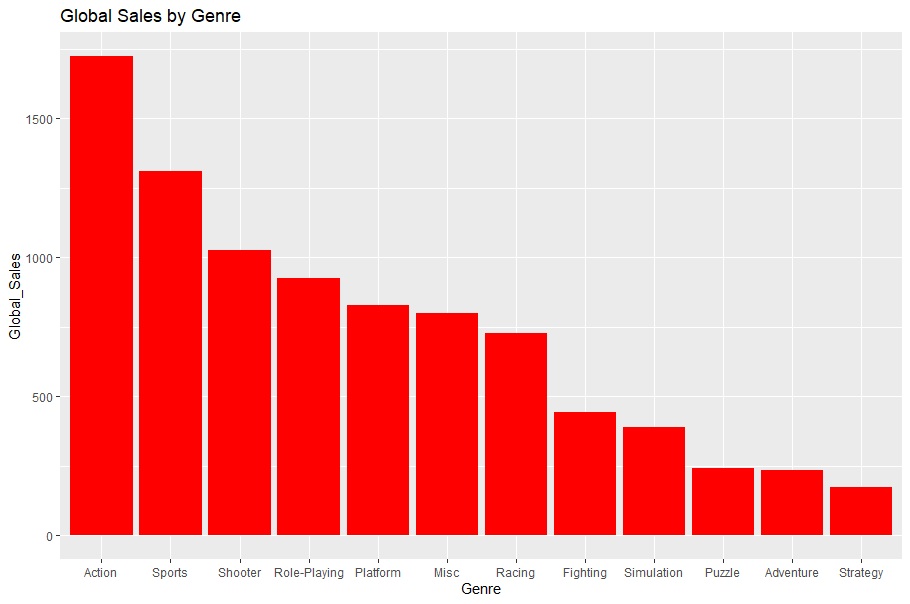
The bar chart shows the frequency of games released each year between 1980 to 2016. An increasing trend in the number of video games released over time can be observed, with a rapid increase from 2000 (the PS2 was released in this year) and the peak number of releases being between 2009 to 2010. Although there seems to be a downward trend following that year. This is somewhat surprising as with improvements in technology it can be expected that more games would be released, or at least the frequency of releases would remain fairly consisted. One explanation for this trend may be that with improvements in graphics, the focus has shifted to releasing better quality games as opposed to greater quantities. Another explanation could be that with newer generation of consoles being released after 2010 (the PS4 and Xbox one was released in 2013), development was shifted towards making games for the newer platforms and older platforms (such as the PS3 and Xbox 360) became obsolete. Over time, older consoles become obsolete and video game development tend to halt for such older platforms.

**Global sales per year**



While there has been an overall increase in sales over time, following 2010 the sales volume can be seen to be decreasing, this is perhaps in direct relation to how fewer games have been released following 2010. This closely matches the previous chart of frequency of game releases per year. While the dataset contains sales volumes of individual titles, it does not take into consideration the release of DLC (Downloadable content) which is often sold as an add-on to already released games. It also does not take into consideration additional costs such as online services and microtransactions. Consoles released after 2010, such as the PS3 and Xbox 360 make greater use of internet connectivity. Hence game publishers are able to make use of online services where they can charge extra money from customers and even provide updates and improvements to already released games. This might explain the overall downward trend of sales as more revenue can be generated from online services and updating currently released games rather than developing and releasing new ones.

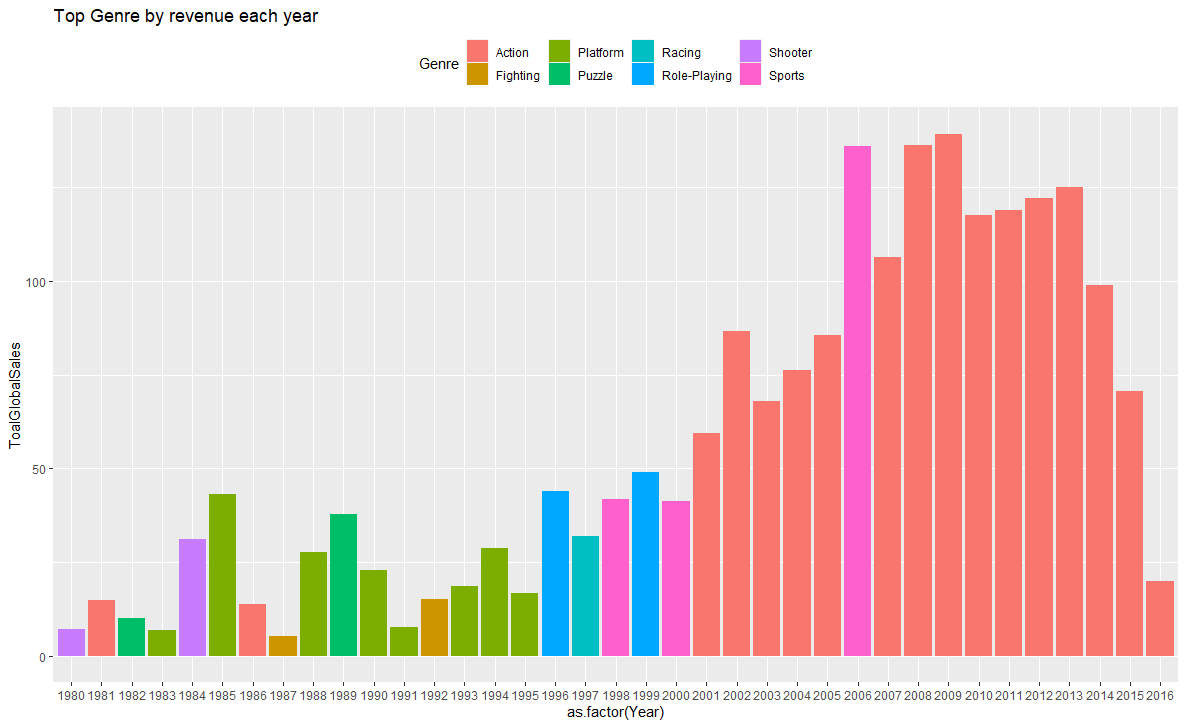
**Global sales by genre**



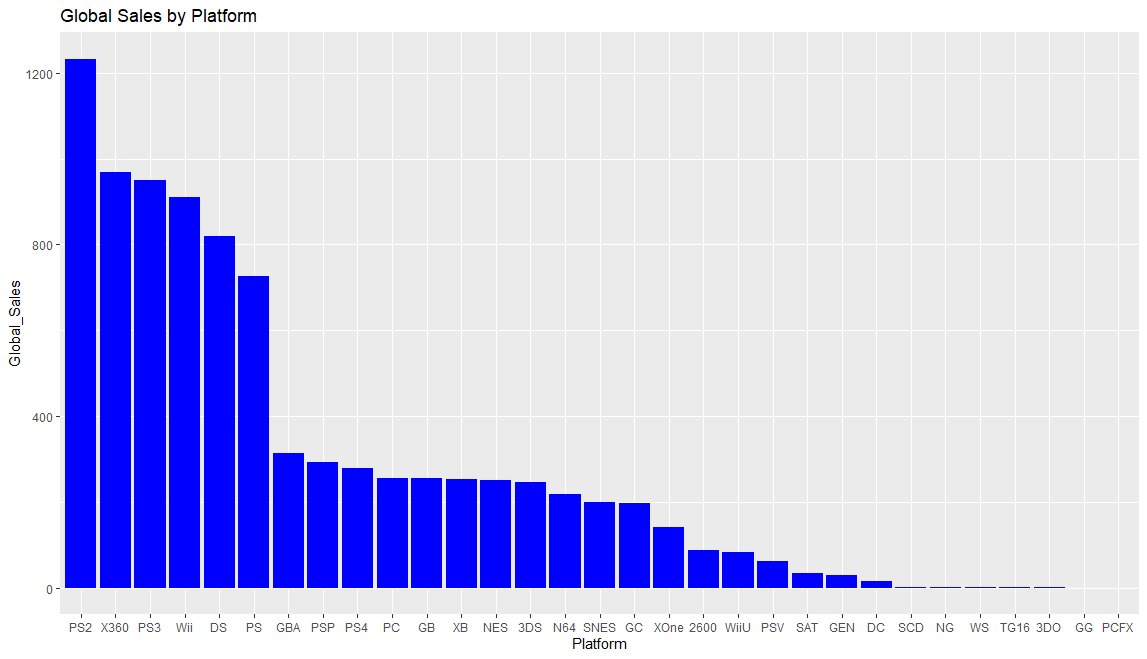
Action games seem to have the greatest number of sales globally whereas strategy has the least number of sales. Although Shooters, role-playing games and fighting games are ranked as separate genres, even though these genres are often interchangeable with action. Hence the real question is perhaps where the distinction lies between action games and other sub genres of games.

**Top genre by global sales per year**

Action games hold the greatest sales volumes throughout multiple years. More action games have been sold since 2001 only being disrupted by sports games in 2006. More games were sold in other genres prior to 2001. With advancements in graphics and processing power, action games have become easier to make in following years which would explain its popularity. Given its popularity, new developers will benefit more from developing action games if they wish to increase sales volumes.

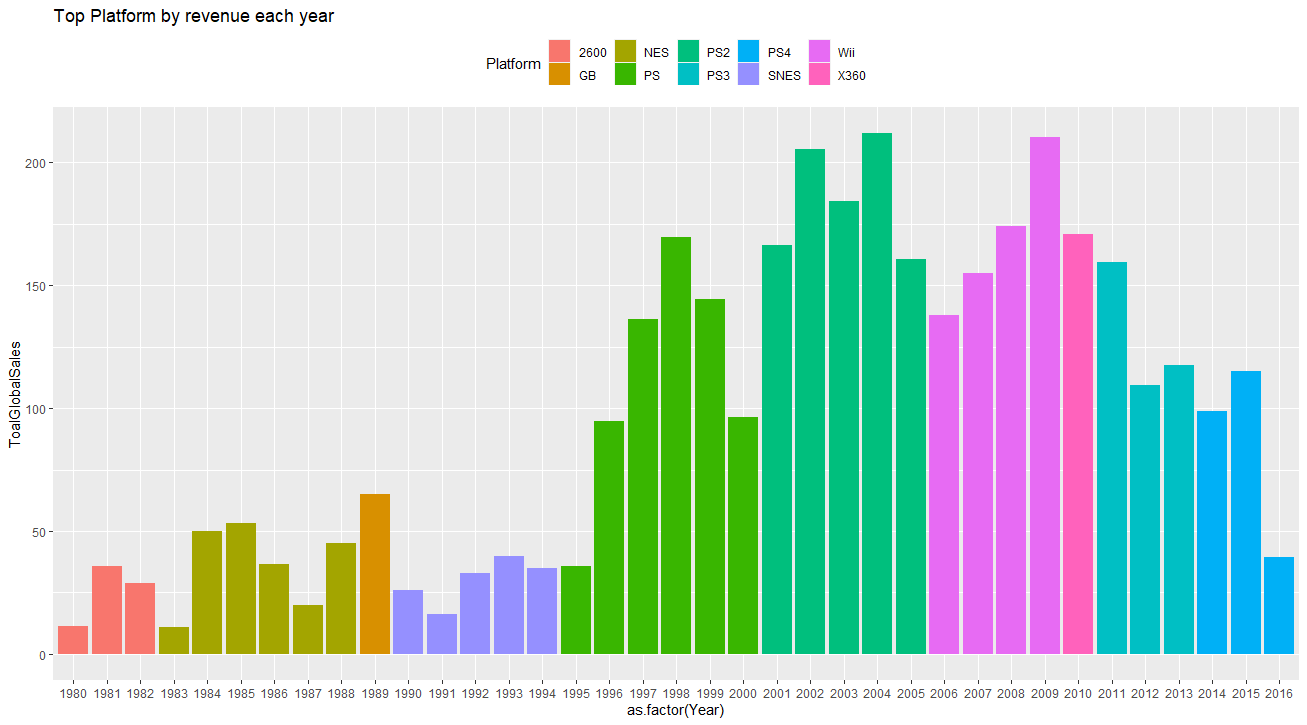


**Global sales by platform**



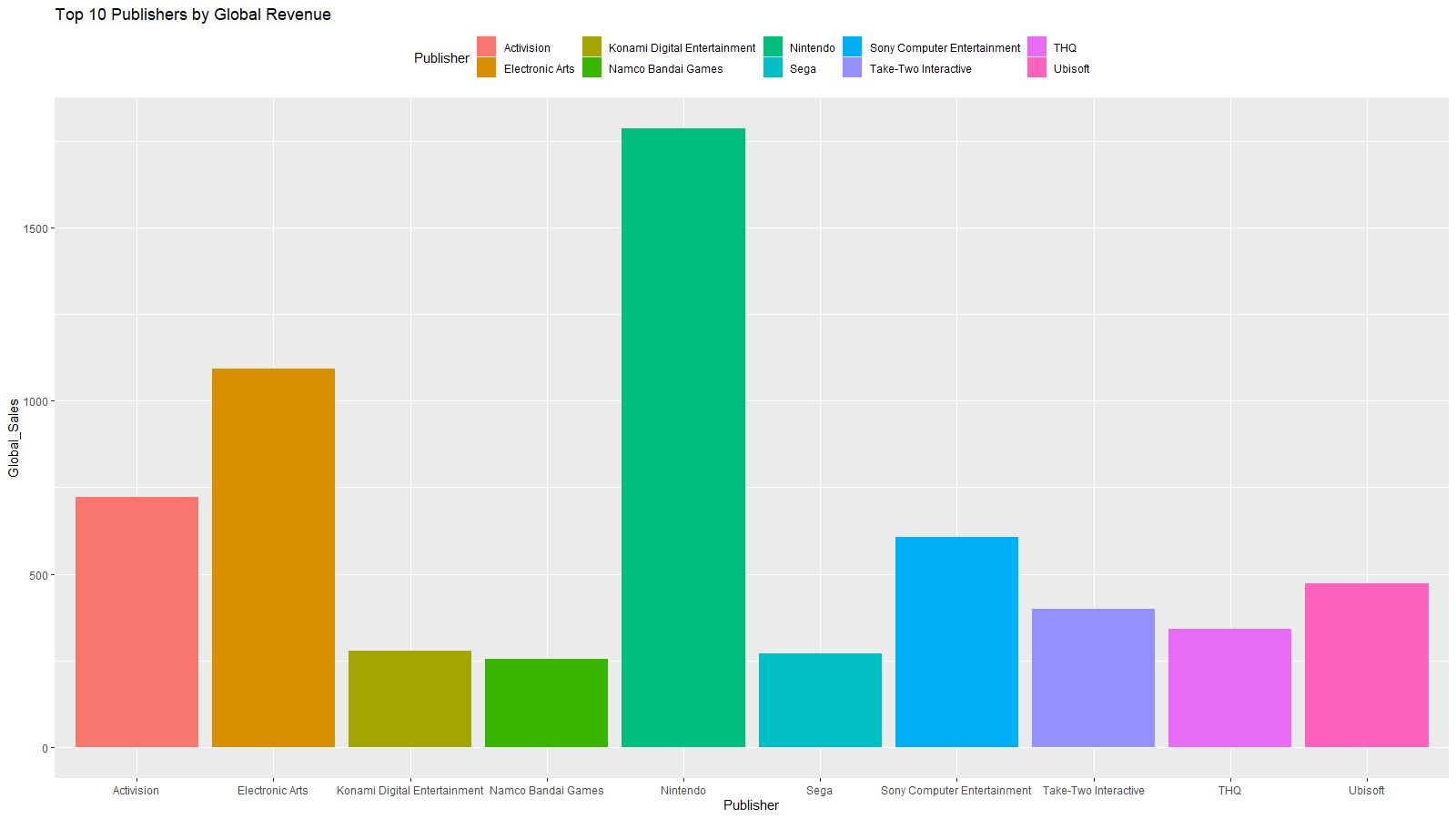
PS2 games by far have the greatest sales volumes compared to other platforms. Sony can be seen to be the leading competitor when it comes to games released for its systems with the PS, PS2, PS3 and the PS4 all having significant market share in sales volume. The PS was released in 2000 and the PS2 was released in 2006, which is when the largest increase in sales volumes can be seen in previous graphs. Newer generation consoles such as the PS4 and the Xbox One have fewer sales, however. As with previous charts, this could be explained by the idea of generating revenue through the sales of additional content and online services.

**Top platform by global sales per year**



The Wii was the most popular platform between 2006 to 2009 which would perhaps explain Nintendo’s hold as the top publisher of games. The PS2 was the most popular platform between 2001 and 2005, the PS being the most popular between 1995 and 2000, the PS3 between 2011 and 2013 and the PS4 from 2014 onwards. With Sony being the top platform in multiple years, particularly during the peak years of game release, it makes sense it dominated as the platform of choice. This might indicate that more people own one of these platforms for video games as most games are released for multiple platforms. Nintendo and Sony platforms would be ideal choices for developing new games for, as they remain consistent in terms of game releases.

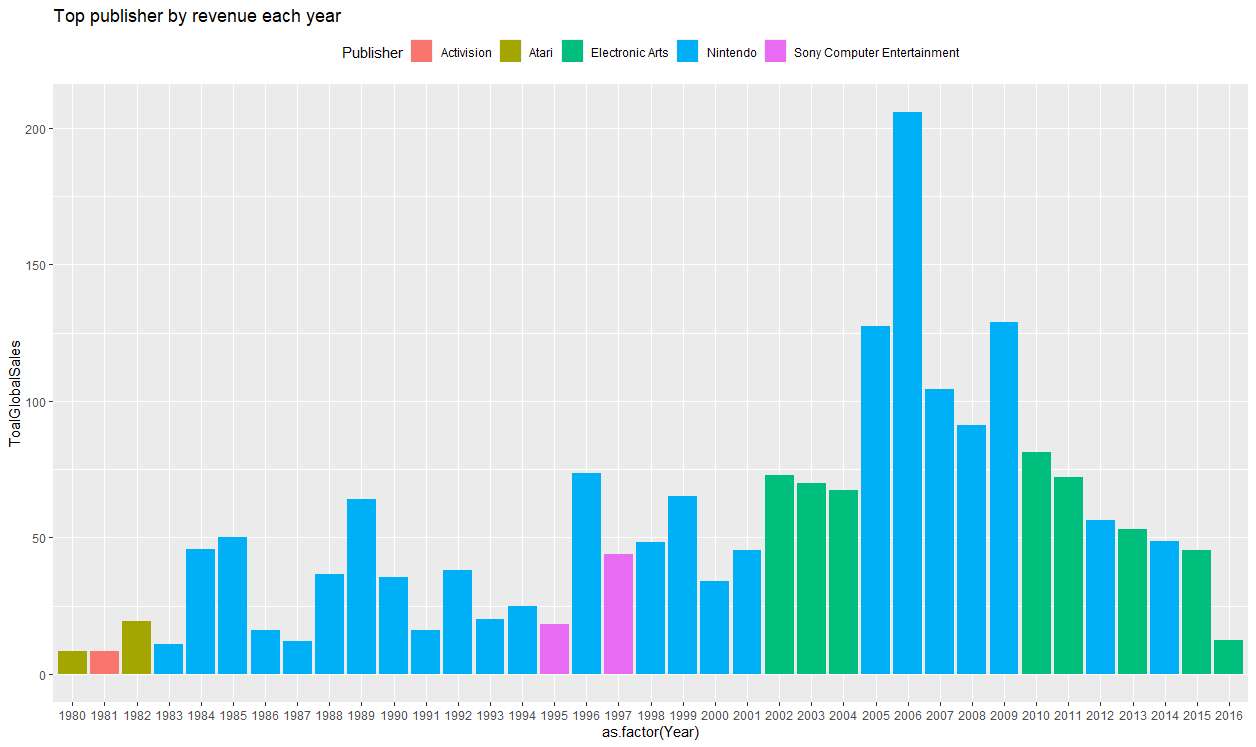
**Top 10 publishers by global sales**



For this chart I only used the top 10 publishers due to the data being too large otherwise. Nintendo appears to be the top publisher of video games. This is interesting as from previous graphs it can be seen that Sony’s platforms generally have greater sales volumes. An explanation for this may be since Nintendo games are never published for platforms other than its own. Other publishers release their games for multiple platforms thus if we were to combine sales from all other publishers and compare that to Nintendo, it would likely dwarf Nintendo’s sales volumes. Overall Nintendo has a decent market share of platforms and it is the biggest publishers of games as well.

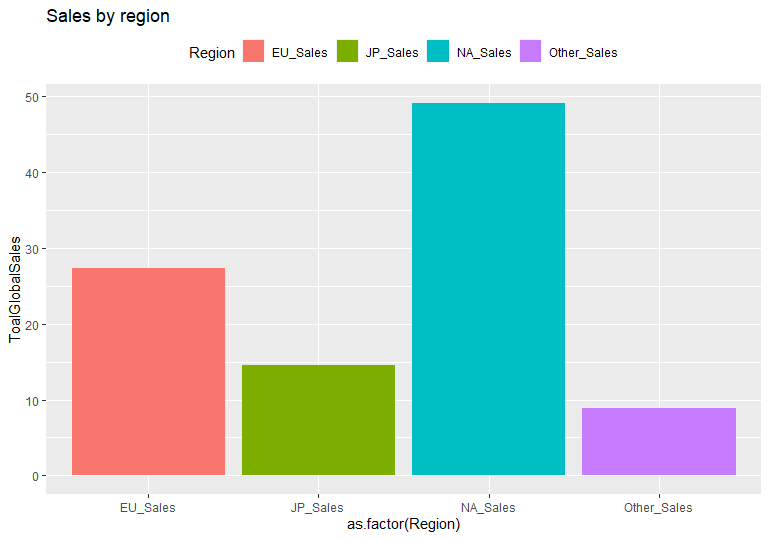
**Top publisher by sales per year**

Nintendo as a publisher for games clearly owns the greatest market share in terms of overall sales volumes throughout, which is in line with the previous graph. It has also been the top publisher for games in multiple years. It would be interesting to see how other publishers fared in those same years as Nintendo

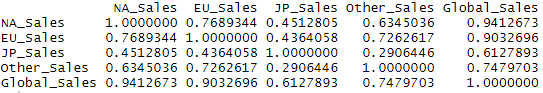


**Sales by Region**

This chart compares the sales in individual regions as a factor of the global sales. North America has the greatest sales volumes of video games overall, being almost twice the number for the EU. This would make North America an ideal location for initial release of new games.

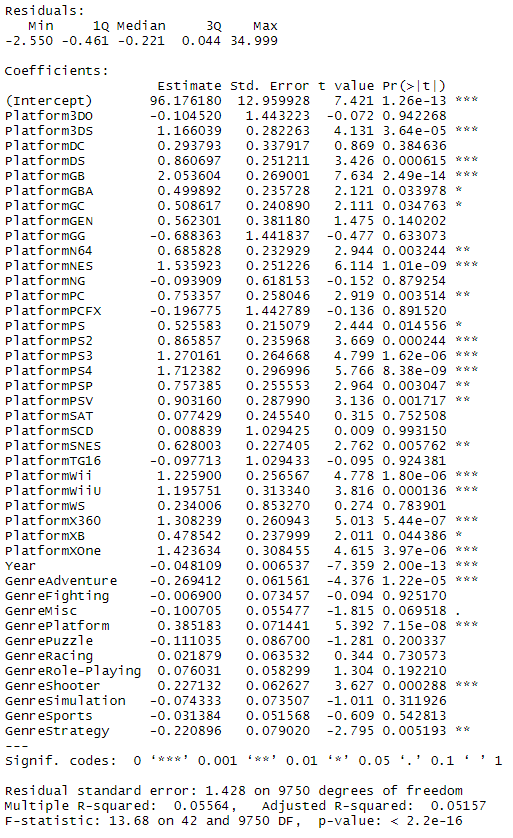


**Correlation of regional sales with global sales**



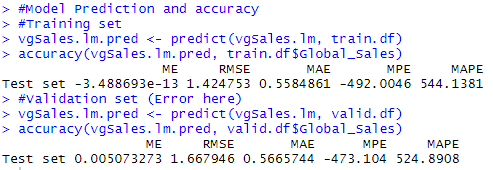
While there is mostly positive correlation between sales in different regions, only the correlation with global sales can be considered of note. Higher sales in certain regions would affect global sales as opposed to between certain regions. Correlation between other factors with global sales such as platform, genre and publisher might give more insight into sales trends.

**Multiple Linear Regression Model**



I ran a multiple linear regression model to see how certain factors explain global sales. The variables used were platform, genre, and year. Publisher was excluded as it would introduce too many variables into the model. The model was ran on a set of training data containing 60% of the variables. The model shows an R-squared value of 0.05564 and an adjusted R-squared value of 0.05157. Both values are quite low indicating that the independent variables are not explaining much of the variation of global sales.

**Model Prediction and Accuracy**



The RMSE values for both the training set and validation set are above 1. The MAE values are roughly 0.5 for both sets. This indicates a poor fit for the model.

**Issues with the dataset**

The dataset itself only contains game releases up to the year 2016 and hence might contain incomplete data for that year. The games listed and their sales amounts are only based on its year of release and its sales during release. Some games may have sold more in the following years and the data does not take that into account. Some games, particularly those released in the late 2010s are also released such that more copies are sold in following years as the games are constantly updated and more players buy those games. Many games also have DLC released for its which players must purchase separately, accounting for a greater portion of the revenue that is earned by publishers. With improvements in broadband connection, publishers are earning more revenue through online services and subscription models. Hence a game might be released for long term playability through in-game purchases and subscription models. The dataset also does not account for mobile games which has seen more growth in recent years.

**Conclusion**

This study can be used as a guideline for obtaining an idea of trends in video game sales and how different factors can impact sales. It can also be used to gain insight into steps that can be taken to improve sales. Further research needs to be conducted before drawing final conclusions. Sales in individual regions alongside global sales against each factor can be studied to gain better insight. As the dataset only contains values up to the year 2016, more recent data needs to be used to for obtaining more accurate and up to date results.

**Appendix**

Code used:

library(dplyr)

library(ggplot2)

library(e1071)

library(MASS)

library(tidyr)

library(caret)

library(forecast)

vgSales.df <- read.csv("vgsales.csv")

#View(vgSales.df)

#plot(vgSales.df)

summary(vgSales.df) # Find summary stats for each column

#vgSales.df[vgSales.df == "N/A"] <- NA

#vgSales.df <- na.omit(vgSales.df)

#attach(vgSales.df)

vgSales.df <- vgSales.df[!(vgSales.df$Year %in% c("N/A", "2017", "2020")),]

vgSales.df$Name <- as.factor(as.character(vgSales.df$Name))

vgSales.df$Platform <- as.factor(as.character(vgSales.df$Platform))

vgSales.df$Year <- as.numeric(as.character(vgSales.df$Year))

vgSales.df$Genre <- as.factor(as.character(vgSales.df$Genre))

vgSales.df$Publisher <- as.factor(as.character(vgSales.df$Publisher))

max(vgSales.df$Year, na.rm=TRUE)

#Histogram of frequency of game releases by year

hist(vgSales.df$Year, col = "red", xlab = "Year", ylab = "Frequency",

main = "Frequency of game releases by year")

#Bar chart of Global sales by year

aggregateRevenue <- aggregate(Global\_Sales~Year, vgSales.df, sum)

plot(aggregateRevenue, type = 'h', xlab="Year", ylab = "Global Sales",

col = "green", lwd = 8, main = "Global Sales per year")

#Bar chart of Global sales of games by genre

revenueByGenre <- aggregate(Global\_Sales~Genre, vgSales.df, sum)

arrangeByGenre <- arrange(revenueByGenre, desc(Global\_Sales))

arrangeByGenre$Genre <- factor(arrangeByGenre$Genre, levels = arrangeByGenre$Genre)

ggplot(arrangeByGenre, aes(Genre, Global\_Sales)) + geom\_bar(fill="red",stat = "identity")

+ ggtitle("Global Sales by Genre")

#Bar chart of Global sales of games by platform

revenueByPlatform <- aggregate(Global\_Sales~Platform, vgSales.df, sum)

arrangeByPlatform <- arrange(revenueByPlatform, desc(Global\_Sales))

arrangeByPlatform$Platform <- factor(arrangeByPlatform$Platform, levels = arrangeByPlatform$Platform)

ggplot(arrangeByPlatform, aes(Platform, Global\_Sales)) + geom\_bar(fill="blue",stat = "identity")

+ ggtitle("Global Sales by Platform")

#Bar chart of Global sales of games by top 10 Publishers

revenueByPublisher <- aggregate(Global\_Sales~Publisher, vgSales.df, sum)

arrangeRevenueByPublisherByGlobalSales <- arrange(revenueByPublisher,desc(Global\_Sales))

top10=arrangeRevenueByPublisherByGlobalSales[1:10,]

ggplot(top10,aes(Publisher,Global\_Sales, fill=Publisher))+ geom\_bar(stat = "identity")

+ ggtitle("Top 10 Publishers by Global Revenue") + theme(legend.position = "top")

#Top publisher by revenue each year

arrangeByYearAndPublisher <- vgSales.df %>% group\_by(Year, Publisher) %>%

summarize(ToalGlobalSales = sum(Global\_Sales)) %>% arrange(desc(ToalGlobalSales)) %>% top\_n(1)

ggplot(arrangeByYearAndPublisher, aes(x=as.factor(Year), ToalGlobalSales, fill=Publisher))

+ geom\_bar(stat = "identity") + ggtitle("Top publisher by revenue each year") + theme(legend.position = "top")

#Top genre by revenue each year

arrangeByYearAndGenre <- vgSales.df %>% group\_by(Year, Genre) %>%

summarize(ToalGlobalSales = sum(Global\_Sales)) %>% arrange(desc(ToalGlobalSales)) %>% top\_n(1)

ggplot(arrangeByYearAndGenre, aes(x=as.factor(Year), ToalGlobalSales, fill=Genre))

+ geom\_bar(stat = "identity") + ggtitle("Top Genre by revenue each year") + theme(legend.position = "top")

#Top platform by revenue each year

arrangeByYearAndPlatform <- vgSales.df %>% group\_by(Year, Platform) %>%

summarize(ToalGlobalSales = sum(Global\_Sales)) %>% arrange(desc(ToalGlobalSales)) %>% top\_n(1)

ggplot(arrangeByYearAndPlatform, aes(x=as.factor(Year), ToalGlobalSales, fill=Platform))

+ geom\_bar(stat = "identity") + ggtitle("Top Platform by revenue each year") + theme(legend.position = "top")

#Sales by region

vgSalesRegion <- vgSales.df %>% gather(Region, Revenue, 7:10)

vgSalesRegion$Region <- factor(vgSalesRegion$Region)

by\_regions <- vgSalesRegion %>%

group\_by(Region) %>%

summarize(ToalGlobalSales = (sum(Revenue)/sum(Global\_Sales))\*100.0) %>%

arrange(desc(ToalGlobalSales))

ggplot(by\_regions, aes(x=as.factor(Region), ToalGlobalSales, fill = Region)) +

geom\_bar(stat = "identity") +

ggtitle("Sales by region") +

theme(legend.position = "top")

#Correlation of sales among regions

regionSales <- vgSales.df[,c("NA\_Sales","EU\_Sales","JP\_Sales","Other\_Sales","Global\_Sales")]

cor(regionSales)

#Linear regression of Genre, Platform and Publisher against Global\_Sales

#Partition data

selected.var <- c(3, 4, 5, 11)

size <- dim(vgSales.df)[1]

size

set.seed(1)

train.index <- sample(size, size\*0.6)

train.df <- vgSales.df[train.index, selected.var]

valid.df <- vgSales.df[-train.index, selected.var]

vgSales.lm <- lm(Global\_Sales ~ ., data=train.df)

options(scipen = 999)

summary(vgSales.lm)

#Model Prediction and accuracy

#Training set

vgSales.lm.pred <- predict(vgSales.lm, train.df)

accuracy(vgSales.lm.pred, train.df$Global\_Sales)

#Validation set (Error here)

vgSales.lm.pred <- predict(vgSales.lm, valid.df)

accuracy(vgSales.lm.pred, valid.df$Global\_Sales)

**References**

Marchand, A., & Hennig-Thurau, T. (2013). Value Creation in the Video Game Industry: Industry Economics, Consumer Benefits, and Research Opportunities. *Journal of Interactive Marketing*, *27*(3), 141–157. doi: 10.1016/j.intmar.2013.05.001

Babb, J., Terry, N., & Dana, K. (2013). The Impact Of Platform On Global Video Game Sales. *International Business & Economics Research Journal (IBER),* *12*(10), 1273. doi:10.19030/iber.v12i10.8136

<https://www.kaggle.com/gregorut/videogamesales>