Gaming_Data_Markdown

Omar Faruk

9 January 2019

Video game sales

For the Hackathon I wanted to analyse the different behaviors over the years of gaming companies, genres of games and the consoles they are played on. The relationship in sales between different countries are explored and my thoguhts on how to expand this dataset are given at the end of this document. The data found goes up to the year 2016, so may be limited in its scope, however I am sure interesting findings can be found regardless.

Firstly, I will load the appropriate libraries.

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.5.3
```

```
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 3.5.3
```

```
##
## Attaching package: 'skimr'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(ggplot2)
```

Next, we must import video game dataset

```
Video_Game_Data <- read_excel("C:/Users/omarf/OneDrive/Desktop/Hackathon
Assignment/Book1.xlsx")
```

View the dataset

```
View(Video_Game_Data)
```

Observe the structure of the dataset to find the variables of interest.

```
str(Video_Game_Data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    16598 obs. of  11 variables:
##  $ Rank         : num  1 2 3 4 5 6 7 8 9 10 ...
```

```
##  $ Name        : chr  "Wii Sports" "Super Mario Bros." "Mario Kart Wii"
"Wii Sports Resort" ...
##  $ Platform    : chr  "Wii" "NES" "Wii" "Wii" ...
##  $ Year        : chr  "2006" "1985" "2008" "2009" ...
##  $ Genre       : chr  "Sports" "Platform" "Racing" "Sports" ...
##  $ Publisher   : chr  "Nintendo" "Nintendo" "Nintendo" "Nintendo" ...
##  $ NA_Sales    : num  41.5 29.1 15.8 15.8 11.3 ...
##  $ EU_Sales    : num  29.02 3.58 12.88 11.01 8.89 ...
##  $ JP_Sales    : num  3.77 6.81 3.79 3.28 10.22 ...
##  $ Other_Sales : num  8.46 0.77 3.31 2.96 1 0.58 2.9 2.85 2.26 0.47 ...
##  $ Global_Sales: num  82.7 40.2 35.8 33 31.4 ...
```

Summary statistics for the sale figures of interest. This includes sales in North America, Japan, Europe and Globally.

```
Video_Game_Data %>%
  select('NA_Sales',
         'JP_Sales',
         'EU_Sales','Global_Sales')%>%
  skim()
```

```
## Skim summary statistics
##  n obs: 16598
##  n variables: 4
##
## -- Variable type:numeric ------------------------------------------------
-------------------------------------------------------
##      variable missing complete     n  mean   sd   p0  p25  p50  p75  p100
##      EU_Sales       0    16598 16598 0.15  0.51 0    0    0.02 0.11 29.02
##  Global_Sales       0    16598 16598 0.54  1.56 0.01 0.06 0.17 0.47 82.74
##      JP_Sales       0    16598 16598 0.078 0.31 0    0    0    0.04 10.22
##      NA_Sales       0    16598 16598 0.26  0.82 0    0    0.08 0.24 41.49
##      hist
```

Convert condition colums to a factor.

```
Video_Game_Data$Publisher <- as.factor(Video_Game_Data$Publisher)

Video_Game_Data$Genre <- as.factor(Video_Game_Data$Genre)

Video_Game_Data$Platform <- as.factor(Video_Game_Data$Platform)
```

Place newly factorised data into a new data frame. Whilst selecting other variables of interest.

```
VGdata1 <- Video_Game_Data %>% select(`Publisher`, `Genre`, `Platform`,
                                      `Global_Sales`, `NA_Sales`,
                                      `JP_Sales`, 'EU_Sales', 'Year')
```

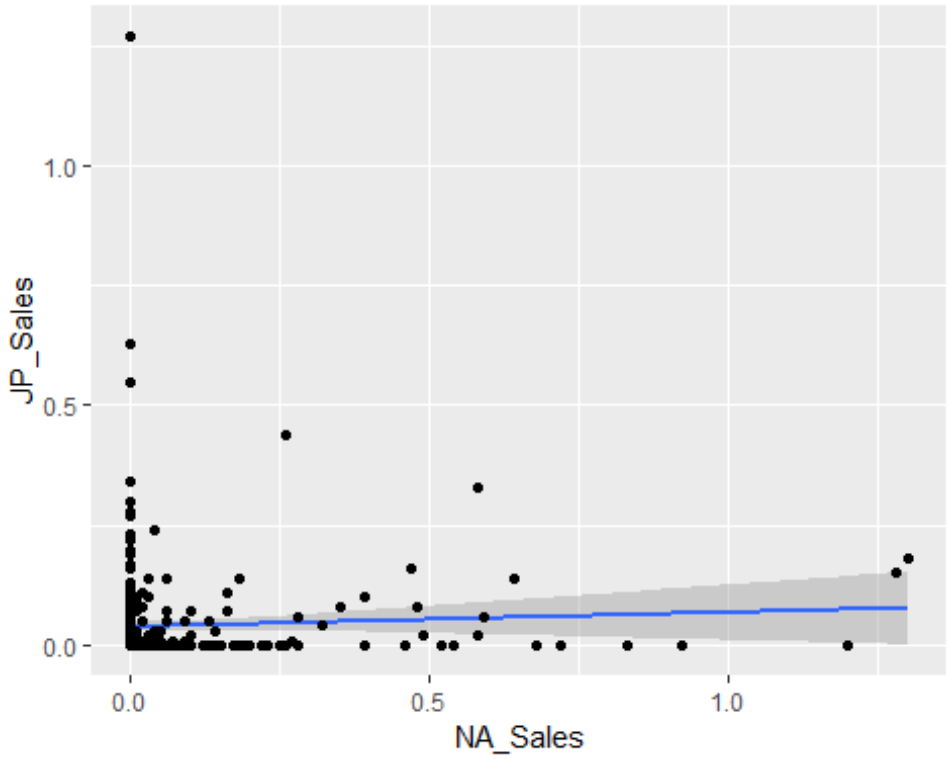
The first question posed was is there a correlation between the sale of video games and the region they are sold in the year 2016?

Firstly, we need to filter the dataset to include only data that was gathered in the year 2016.

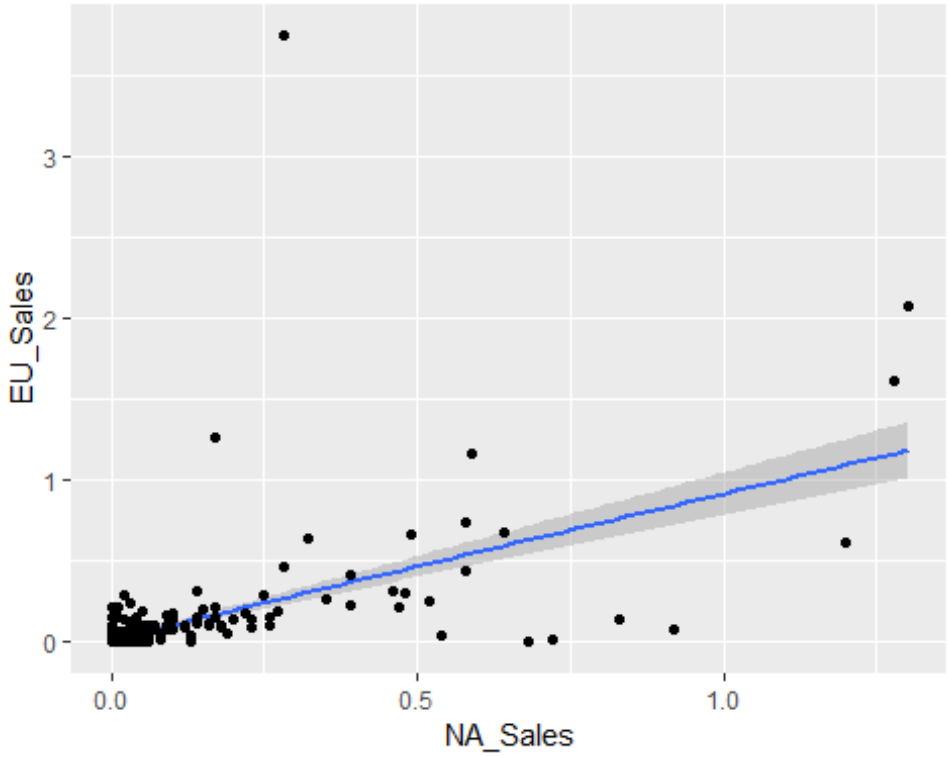
```
VG_Filter <- filter(VGdata1, Year ==2016)
```

Scatterplots are then created to help visualse the relationship between North American, Japanese and European sales.

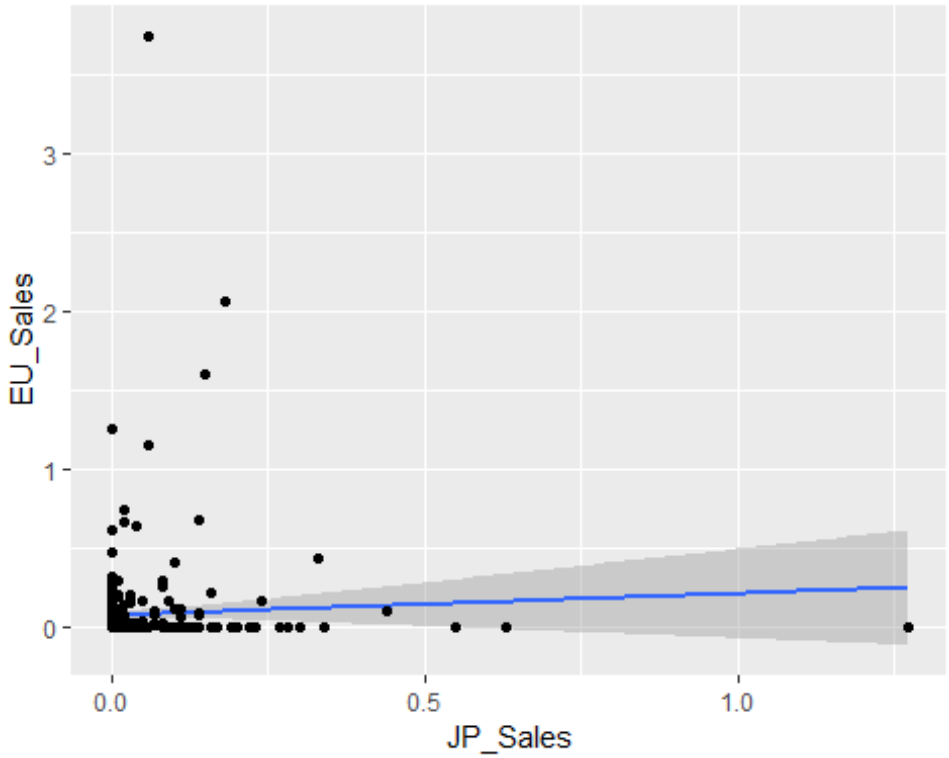
```
ggplot(VG_Filter, aes(x=NA_Sales, y= JP_Sales)) + geom_smooth(method = "lm")
+geom_point()
```




```
ggplot(VG_Filter, aes(x=NA_Sales, y= EU_Sales)) + geom_smooth(method = "lm")
+geom_point()
```




```
ggplot(VG_Filter, aes(x=JP_Sales, y= EU_Sales)) + geom_smooth(method = "lm")
+geom_point()
```



The correlation coefficients for each of the correlations.

```
cor(VG_Filter$NA_Sales, VG_Filter$JP_Sales)
```

```
## [1] 0.05251218
```

```
cor(VG_Filter$NA_Sales, VG_Filter$EU_Sales)
```

```
## [1] 0.5646121
```

```
cor(VG_Filter$EU_Sales, VG_Filter$JP_Sales)
```

```
## [1] 0.04985794
```

All three correlations display a moderate correlation between each of the sales figues.

As North American sales take up most of the Global sales it would be intersting to conduct a linear regression between North American sales with Japanese and EU sales. This would allow us to predict future sales in the US based on Japanese and EU sales.

The following makes a linear regression model that will use the Japanese sales as the predictor variables and the North American sales as the outcome variable.

```
Lin_NA_JP <- lm(NA_Sales ~ JP_Sales, data=VG_Filter)
```

The following code allows us to see the outcome of the model.

```
summary(Lin_NA_JP)
```

```
##
## Call:
## lm(formula = NA_Sales ~ JP_Sales, data = VG_Filter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17864 -0.06405 -0.06222 -0.03222  1.22128
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06222    0.01013   6.142 2.26e-09 ***
## JP_Sales     0.09167    0.09427   0.972    0.332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1745 on 342 degrees of freedom
## Multiple R-squared:  0.002758,   Adjusted R-squared:  -0.0001584
## F-statistic: 0.9457 on 1 and 342 DF,  p-value: 0.3315
```

The same is applied to the European sales.

```
Lin_NA_EU <- lm(NA_Sales ~ EU_Sales, data=VG_Filter)
```

Output

```
summary(Lin_NA_EU)
```

```
##
## Call:
## lm(formula = NA_Sales ~ EU_Sales, data = VG_Filter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08985 -0.03825 -0.03825 -0.01294  0.94159
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.038249   0.008077   4.736 3.21e-06 ***
## EU_Sales    0.355092   0.028069  12.651  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1442 on 342 degrees of freedom
## Multiple R-squared:  0.3188, Adjusted R-squared:  0.3168
## F-statistic:   160 on 1 and 342 DF,  p-value: < 2.2e-16
```

Next we need to compare both models with a model with just the intercept (so the mean of our outcome) predicting the outcome (NA sales). Firstly, we need to create such a model.

```
model0 <- lm (NA_Sales ~ 1, data = VG_Filter)
```

We will then run an ANOVA to see if the models using the Japanese and European sales are a better fit than just using the intercept (mean).

```
anova(model0, Lin_NA_EU)
```

```
## Analysis of Variance Table
##
## Model 1: NA_Sales ~ 1
## Model 2: NA_Sales ~ EU_Sales
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    343 10.4435
## 2    342  7.1143  1    3.3293 160.05 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model0, Lin_NA_JP)
```

```
## Analysis of Variance Table
##
## Model 1: NA_Sales ~ 1
## Model 2: NA_Sales ~ JP_Sales
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    343 10.444
## 2    342 10.415  1  0.028798 0.9457 0.3315
```

The F ratio for the European model is signficant indicating it is a better fit in comparison to the null model that simply used the intercept. The Japanese model on the other hand did not show a significant F value and therefore would not be a good fit for predicting the North American sales.

It was shown that for every 0.038 million units of games sold in the EU there will be a million units sold in North America. Therefore, using sales figures in the EU would be a good indicator to predict sales in North America.

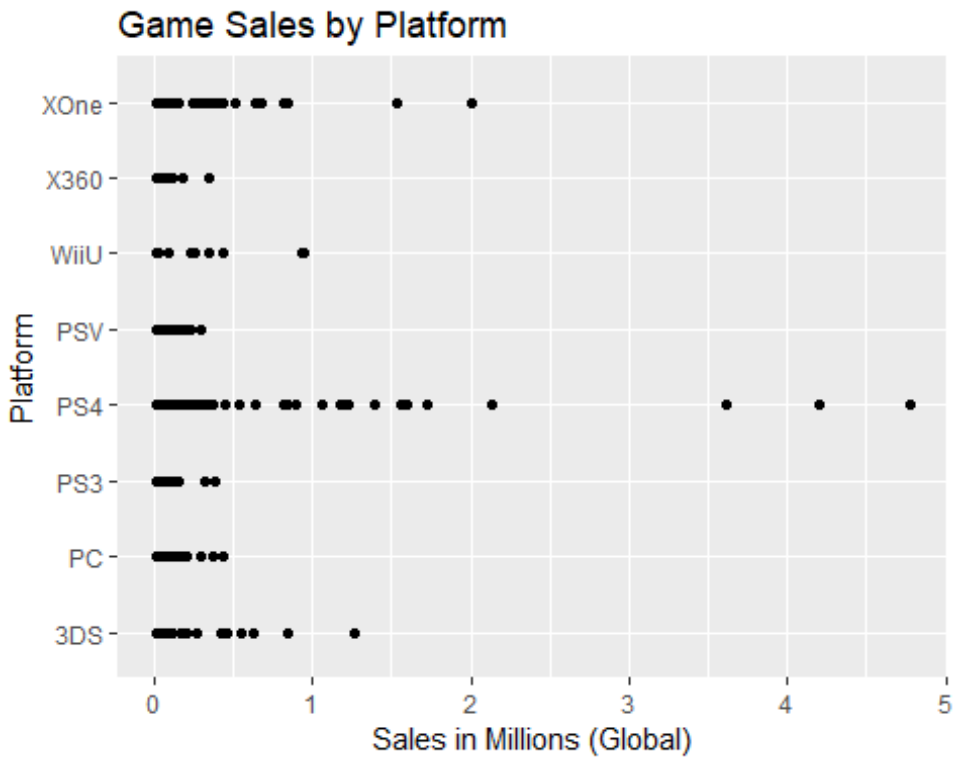
It would also be interesting to know the popularity of Platform, Publisher and Genre in which these games were sold in the year 2016.

Firstly, the mean score for each category were compiled.

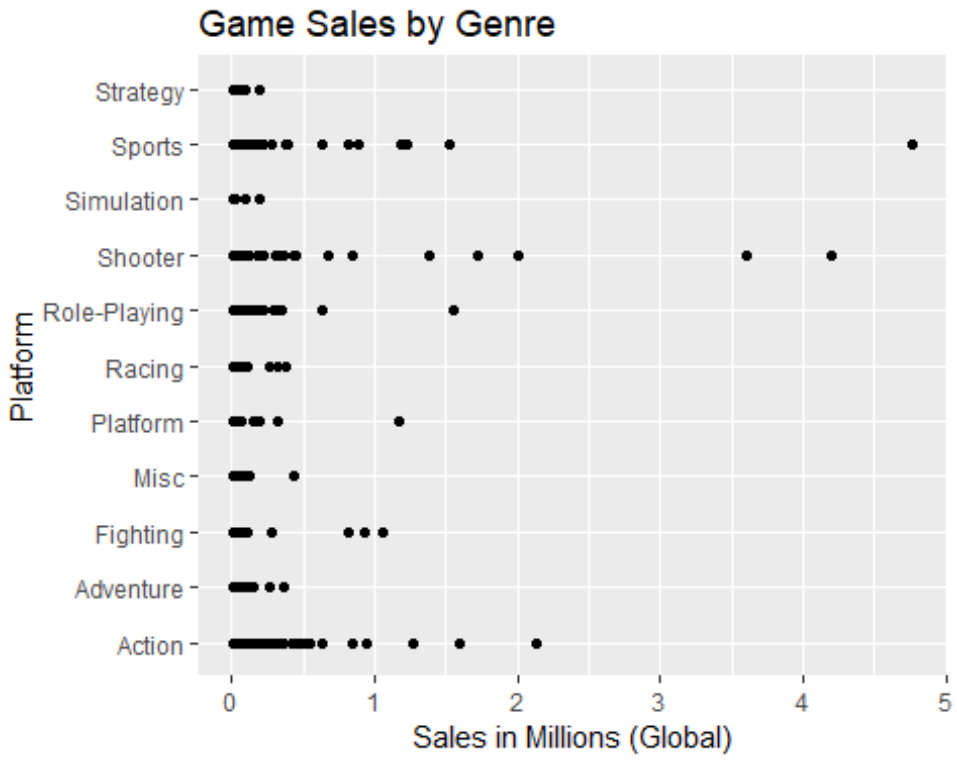
```
attach(VG_Filter)
names(VG_Filter)

## [1] "Publisher"    "Genre"        "Platform"     "Global_Sales"
## [5] "NA_Sales"     "JP_Sales"     "EU_Sales"     "Year"

mean(Global_Sales)

## [1] 0.2061919

tapply(Global_Sales, Publisher, mean)

tapply(Global_Sales, Genre, mean)

##       Action    Adventure     Fighting         Misc     Platform
##   0.16731092   0.05323529   0.27571429   0.06500000   0.20700000
##       Puzzle       Racing Role-Playing      Shooter   Simulation
##           NA   0.08200000   0.16900000   0.56937500   0.04333333
##       Sports     Strategy
##   0.38421053   0.05000000

tapply(Global_Sales, Platform, mean)

##       2600        3DO        3DS         DC         DS         GB
##         NA         NA 0.18857143         NA         NA         NA
##        GBA         GC        GEN         GG        N64        NES
##         NA         NA         NA         NA         NA         NA
##         NG         PC       PCFX         PS        PS2        PS3
##         NA 0.06842105         NA         NA         NA 0.08093750
##        PS4        PSP        PSV        SAT        SCD       SNES
## 0.36682243         NA 0.05666667         NA         NA         NA
##       TG16        Wii       WiiU         WS       X360         XB
##         NA         NA 0.32900000         NA 0.10375000         NA
##       XOne
## 0.22907407
```

This data is then visualised.

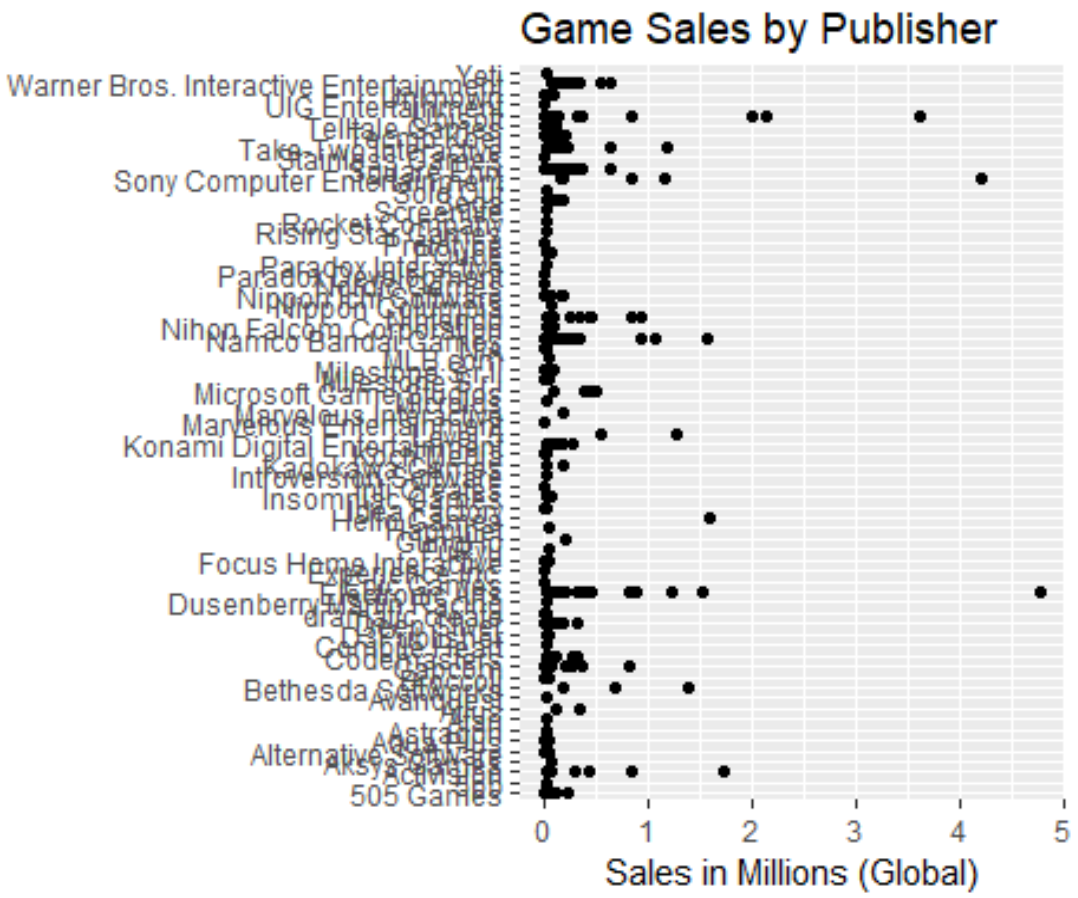
```
VGdata1 %>%
  group_by(Platform) %>%
  filter(!is.na(`Global_Sales`) & Year==2016)%>%
  ggplot(aes(x = Global_Sales, y = Platform)) +
  geom_point() +
  labs(x = "Sales in Millions (Global)", y = "Platform", title = "Game Sales
by Platform")
```




```
VGdata1 %>%
  group_by(Genre) %>%
  filter(!is.na(`Global_Sales`) & Year ==2016)%>%
  ggplot(aes(x = Global_Sales, y = Genre)) +
  geom_point() +
  labs(x = "Sales in Millions (Global)", y = "Platform", title = "Game Sales
by Genre")
```




```
VGdata1 %>%
  group_by(Publisher) %>%
  filter(!is.na(`Global_Sales`) & Year ==2016) %>%
  ggplot(aes(x = Global_Sales, y = Publisher)) +
  geom_point() +
  labs(x = "Sales in Millions (Global)", y = "Platform", title = "Game Sales
by Publisher")
```



In terms of the plaform that has sold the most games it appears that the Xbox One, the Playstation 4 and the 3DS have sold the most globally. The highest selling genre appear to be sports, shooting and action games. Lastly, in terms of the highest selling publishers Sony, Electronic Arts and Ubisoft appear to have sold the most games, with sales surpassing 3 million globally.

This topic of interest was expanded upon to understand the change in sales over the past ten years.

Firstly, we need to understand the mean score of Global Sales for Platform, Genre and Publisher in the past ten years. The following filter specifies games released after 2006.

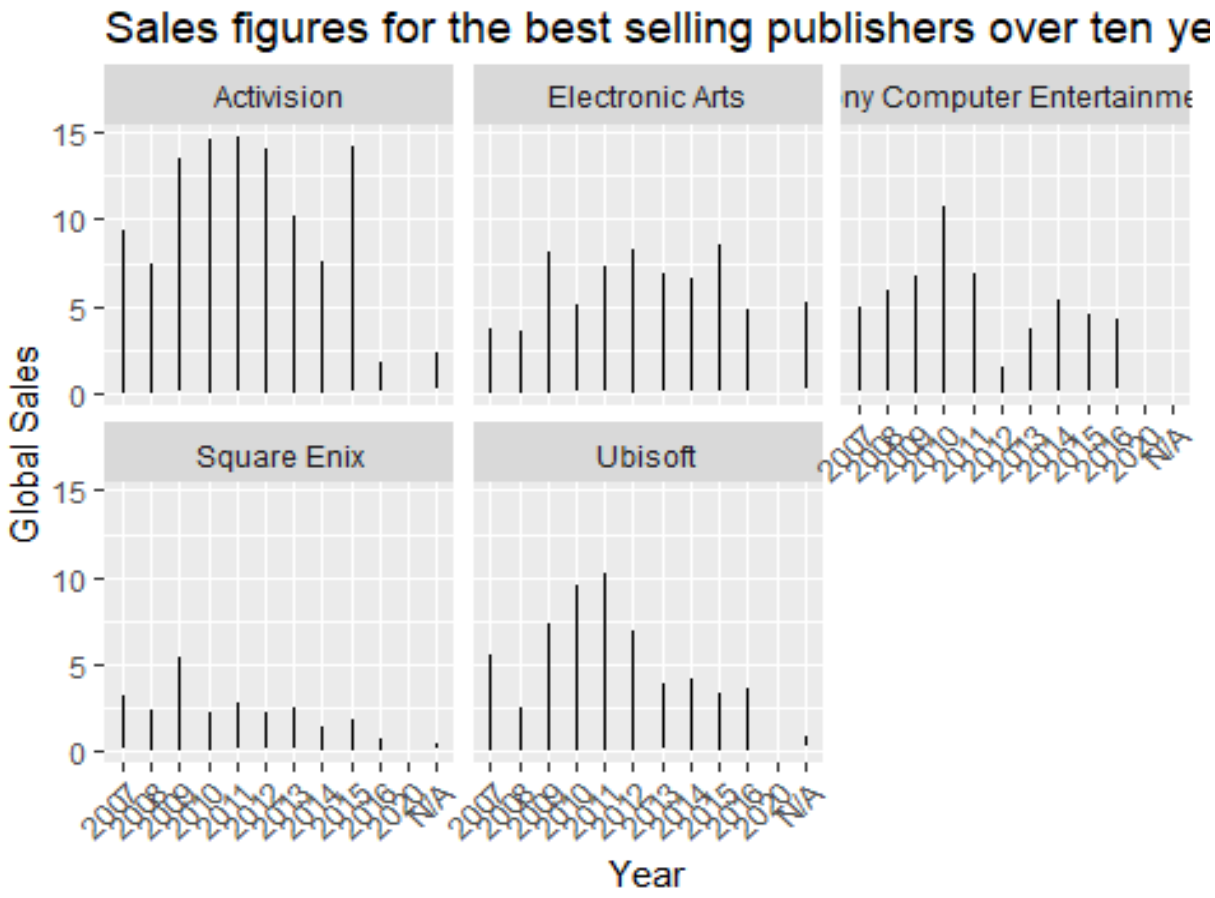
```
VG_Filter1 <- filter(VGdata1, Year > 2006)
```

Next we need to specify the highest selling category of each factor into a data frame.

```
Top_Genre <- c("Sports", "Shooter", "Role-Playing", "Platform", "Racing")

Top_Pub <- c("Sony Computer Entertainment", "Ubisoft", "Electronic Arts",
"Activision", "Square Enix")

Top_Platform <- c("PS4", "XOne", "Wii", "X360", "PS3", "WiiU")
```

The following graphs display the change in sales for each factor over time for publishers.

```
VG_Filter1 %>%
  filter(Publisher %in% Top_Pub) %>%
  group_by(Year) %>%
  filter(!is.na(`Global_Sales`)) %>%
  ggplot(aes(x = Year, y = `Global_Sales`)) +
  geom_line() +
  facet_wrap(~ Publisher) +
  labs(x = "Year", y = "Global Sales", title = "Sales figures for the best
selling publishers over ten years") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

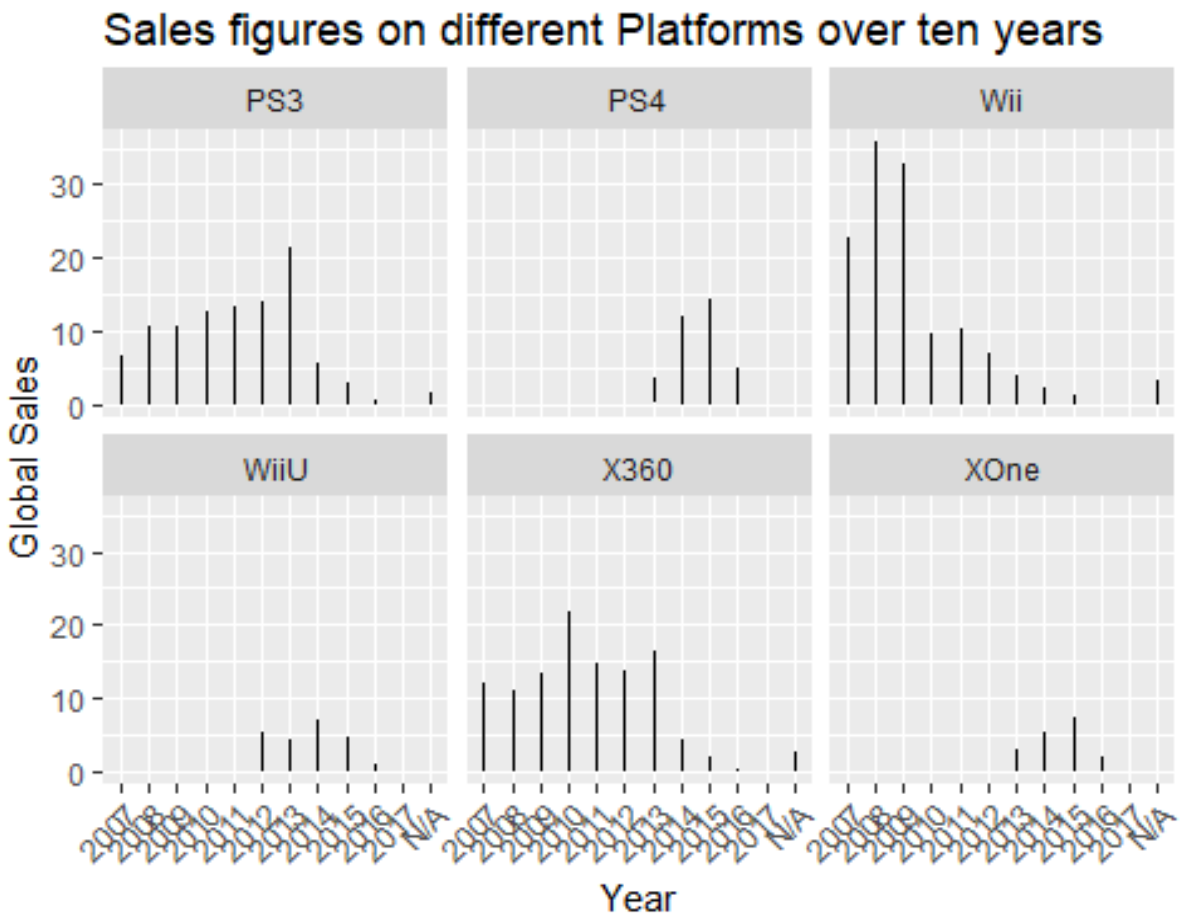


It is interesting Electronic Arts, Ubisoft and Sony have all reached a 5 million sale plateu in 2016. Wheras Activision and Square Enix seem to have had a all time low in sales in comaparison to their previous years in 2016.

The next graph will display change in sales over time in accordance to what platform the games are released under.

```
VG_Filter1 %>%
  filter(Platform %in% Top_Platform) %>%
  group_by(Year) %>%
```

```
  filter(!is.na(`Global_Sales`)) %>%
  ggplot(aes(x = Year, y = `Global_Sales`)) +
  geom_line() +
  facet_wrap(~ Platform) +
  labs(x = "Year", y = "Global Sales", title = "Sales figures on different
Platforms over ten years") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



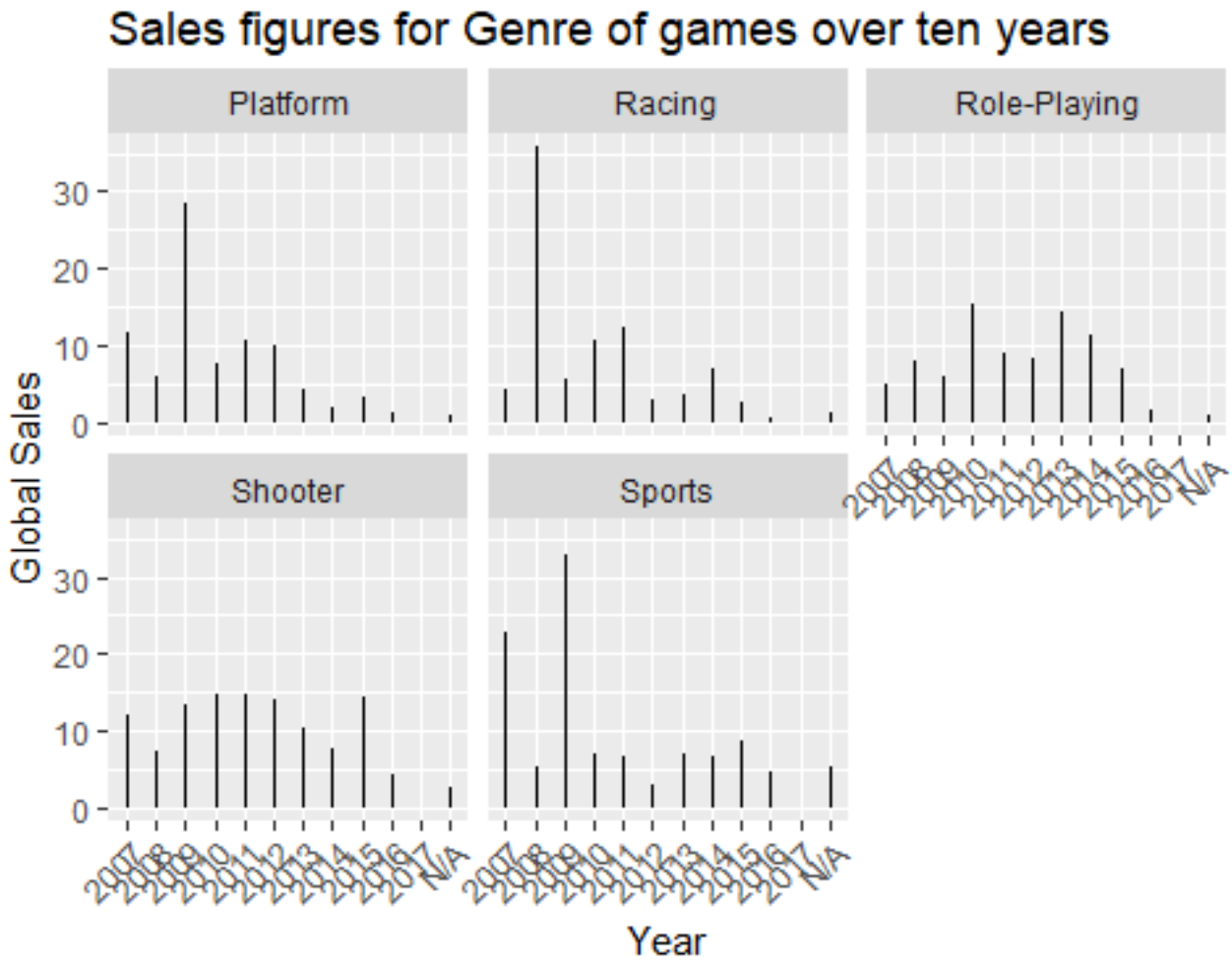
The dropoff in sales for the Xbox 360, PS3 and wII are expected as newer versions of such consoles (the PS4, Xbox One and Wii U) would intice consumers to buy games on the next generation of systems. This would explain the rise in sales for the newer consoles and the dropoff in sales for the older consoles.

Lastly, the last graph display global sales over time in accordance to genre.

```
VG_Filter1 %>%
  filter(Genre %in% Top_Genre) %>%
  group_by(Year) %>%
  filter(!is.na(`Global_Sales`)) %>%
  ggplot(aes(x = Year, y = `Global_Sales`)) +
  geom_line() +
  facet_wrap(~ Genre) +
```

```
  labs(x = "Year", y = "Global Sales", title ="Sales figures for Genre of
games over ten years") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Platform, racing and sports games seem to be dropping sales in comparison to



the years before 2010 where each genre was surpassing more than 20 million units. Role-playing and shooting games in comparison have a relatively steady global sales figure.

It would be interesting for this dataset in the future to include games which are part of a series. The exposure to series could affect the sales of a game immensely. Another thing that could be added is a region code for certain developers. As Japanese publishers are more well known to the Japanese market this could have affected the sales of certain games in Japan and the West and understanding this relationship could have been crucial to explaining the lack of relationship when modelling predictions for the North American sales with the Japanese sales