Game Sales Data Analysis

Jussi Pylkkänen

2025-01-20

Abstract

This project aims to analyze the factors influencing video game sales across the globe. The analysis focuses on identifying trends and patterns of games in different regions of the world. Through data analysis and visualization, we find out the varying sales dynamics across regions, platforms, genres and publishers. Key findings include the significant difference in sales between Japan and global markets, the top-performing genres for Nintendo, Activision and Electronic Arts, and lastly, the observation that, apart from sales in different regions, publisher and platform play the biggest roles in predicting global sales.

Introduction

The dataset used is from https://www.kaggle.com/datasets/gregorut/videogamesales made by GREGORY-SMITH. The data is originally scraped from vgchartz.com which tracks and analyzes global video game and console sales data.

In this dataset we have console game sales from across the globe. We have 10 variables: Games name, Platform, Release year, Genre, Publisher, North America Sales, Europe sales, Japan Sales, Other Sales and Global Sales. Sales are in millions. We have 16598 observations before preprocessing.

In this project I will show some graphs which showcase different data trends and patterns, see which genres are most profitable for certain publishers and see which variables are the most important for global sales.

Data trends and Patterns

First let's bring the data in and preprocess it slightly to make a more usable format for analysis and visualization.

```r
set.seed(123) #For reproducibility

sales_data <- read_csv("vgsales.csv", show_col_types = FALSE)
#Check to see if we have any empty values
print(colSums(is.na(sales_data) | sales_data == "N/A", na.rm = TRUE))
```

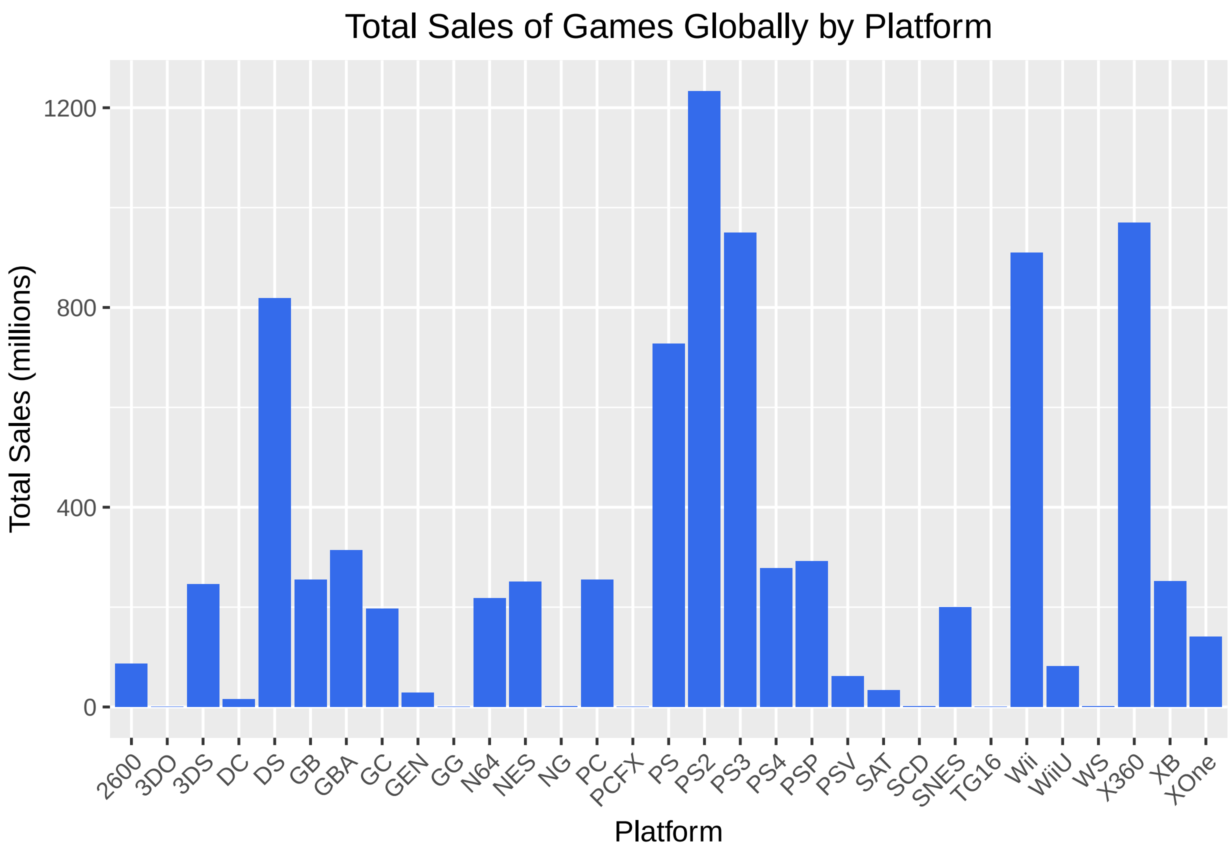
```
##         Rank         Name     Platform         Year        Genre    Publisher 
##            0            0            0          271            0           58 
##     NA_Sales     EU_Sales     JP_Sales  Other_Sales Global_Sales 
##            0            0            0            0            0
```

```r
factor_list <- c('Platform', 'Genre', 'Publisher') #Factorize these variables
for (i in 1:length(factor_list)) {
  sales_data[[factor_list[i]]] <- as.factor(sales_data[[factor_list[i]]])
}

sales_data <- sales_data %>% #Filter out N/A from Year variable
  filter(Year != "N/A") %>%
  mutate(Year = as.numeric(Year))

platform_sales_Global <- sales_data %>%  #Group all Global sales of games for a platform together
  group_by(Platform) %>%
  summarize(Total_Sales_Global = sum(Global_Sales, na.rm = TRUE))

platform_sales_JP <- sales_data %>%  #Group all Japan sales of games for a platform together
  group_by(Platform) %>%
  summarize(Total_Sales_JP = sum(JP_Sales, na.rm = TRUE))
```

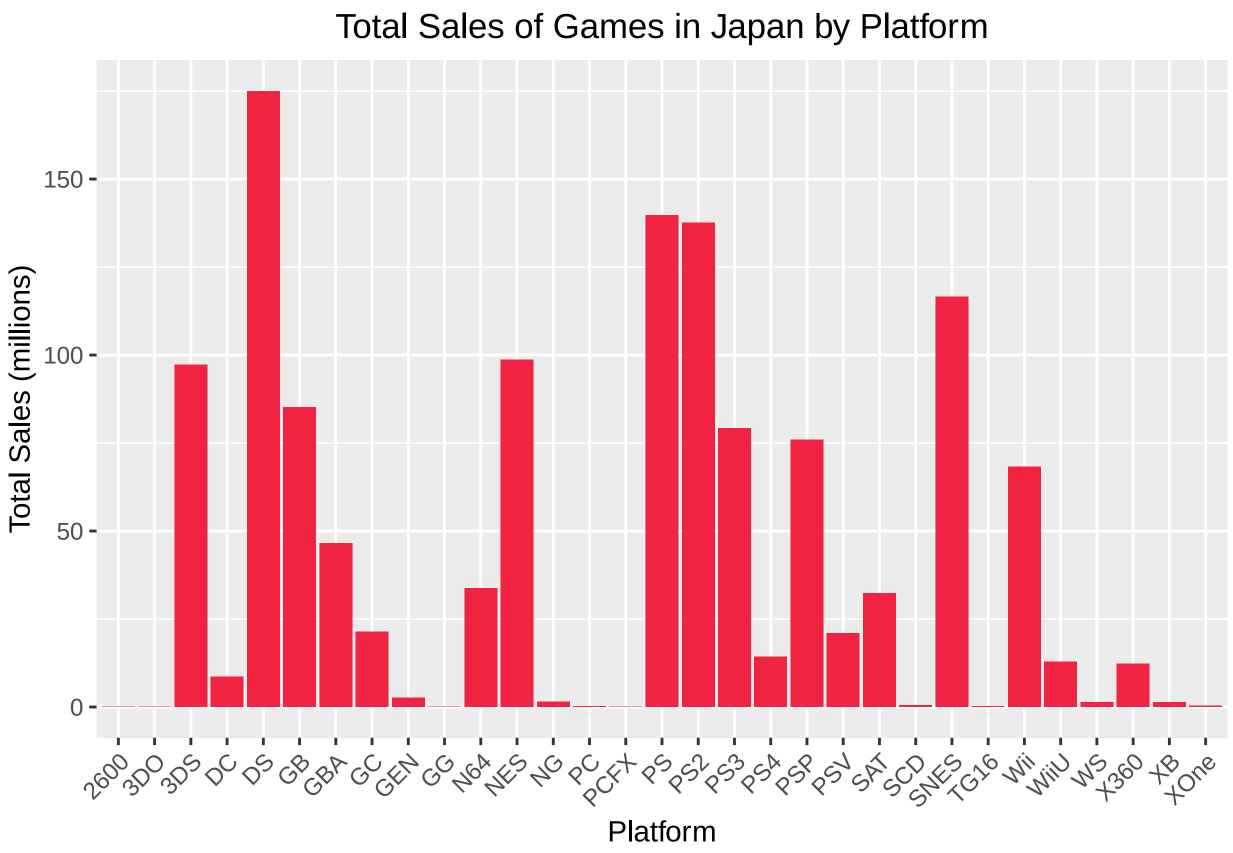
Next let's produce some graphs to showcase some trends. We can see that the global sales and sales in Japan differ quite a lot. For example Playstations 1, 2, 3 and Nintendo DS have done well in both groups, but we can see that consoles by Nintendo like 3DS, Gameboy (GB), NES and SNES have sold proportionally better in Japan than globally. And on the other hand, the Xbox consoles (X360, XB, XOne) didn't really sell much in Japan. This is likely due to Sony (Playstation) and Nintendo being originally Japanese companies and Xbox is an American company.

Japan has an unique market when compared to western society. Japanese companies can better market their products to Japanese people than western companies can. This same trend is seen in other places aswell, for example Toyota being the leading car manufacturer in Japan. We can see this difference between western market and Japanese market in the correlation table as well. We can see that all the other sales correlate much better between each other than with sales in Japan.

```r
#Global sales by Platform
ggplot(platform_sales_Global, aes(x = Platform, y = Total_Sales_Global)) +
  geom_bar(stat = 'identity', fill = '#346beb') +
  labs(x = "Platform", y = "Total Sales (millions)", title = "Total Sales of Games Globally by Platform
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1)) #Rotate labels for readability
```



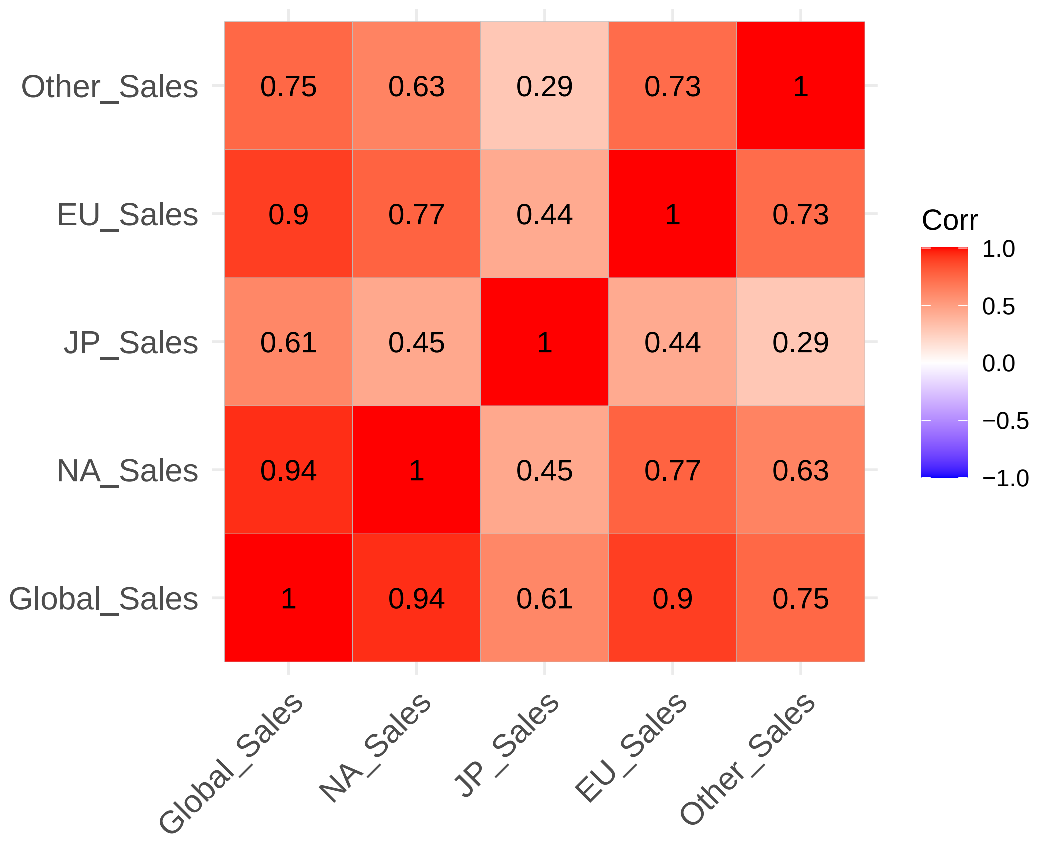
```r
#Japan sales by Platform
ggplot(platform_sales_JP, aes(x = Platform, y = Total_Sales_JP)) +
  geom_bar(stat = "identity", fill = '#f22241') +
  labs(x = "Platform", y = "Total Sales (millions)", title = "Total Sales of Games in Japan by Platform
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
#Correlations
library(ggcorrplot)
```

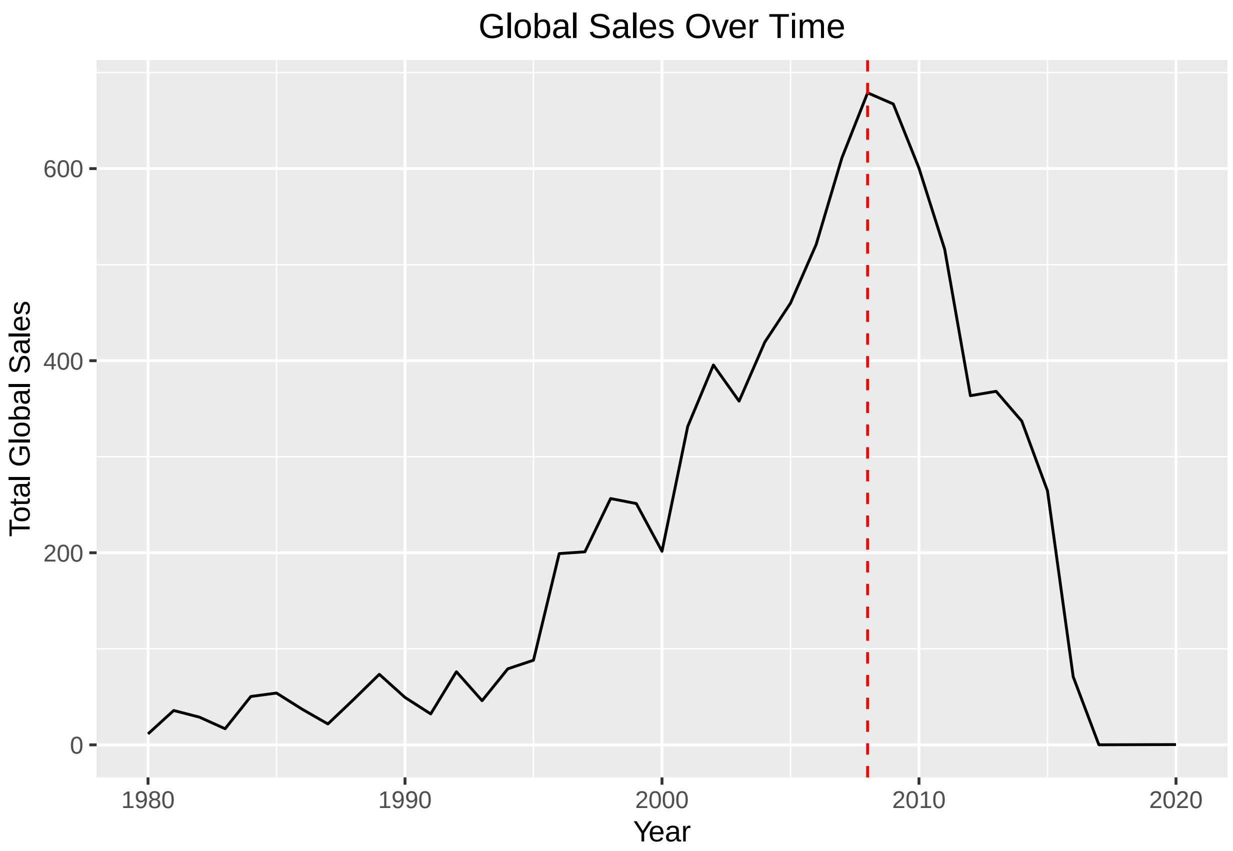
```
## Warning: package 'ggcorrplot' was built under R version 4.4.2
```

```r
numeric_data <- sales_data %>%
  select(Global_Sales, NA_Sales, JP_Sales, EU_Sales, Other_Sales)

cor_matrix <- cor(numeric_data, use = "complete.obs")
ggcorrplot(cor_matrix, lab = TRUE)
```



Next we can see the trend of console game sales through the years. 2008 seemed to be the best year for console game sales. After that the numbers have gone significantly down, and in the year 2018 the site VGChartz.com stopped producing estimates for software sales because the digital market makes it more difficult to produce reliable retail estimates (https://www.vgchartz.com/methodology.php). Other than that this could be caused by many things. PS3, Xbox360, Nintendo Wii and Nintendo DS all came out withing a few years of 2008. Them being some of the most popular consoles, you could expect more games to be sold during that time. But my best guess would also be the increased popularity of PC gaming, which is not taken into account in this dataset.

```
#Global sales by Year
sales_data_by_year <- sales_data %>%
  filter(Year != "N/A") %>%
  mutate(Year = as.numeric(Year)) %>%
  group_by(Year) %>%
  summarize(Total_Sales_Global = sum(Global_Sales, na.rm = TRUE))

max_year <- sales_data_by_year$Year[which.max(sales_data_by_year$Total_Sales_Global)]
ggplot(sales_data_by_year, aes(x = Year, y = Total_Sales_Global, group = 1)) +
  geom_line() +
  geom_vline(xintercept = max_year, linetype = "dashed", color = "red") +
  labs(title = "Global Sales Over Time", x = "Year", y = "Total Global Sales") +
  theme(plot.title = element_text(hjust = 0.5))
```



Next a small discussion about the importance of genre for different publishers. We will use linear regression to see which genres have sold the best for Nintendo, Activision and Electronic Arts. The variables serving as the respective reference categories will be “Sports”, “Shooter” and “Racing”. My guess was that those were the best performing genres for each of these three publishers.

In the linear regression output we can simply see that if the estimate is negative, that level performs worse than the reference, and if it’s positive it performs better. And the Pr(>|t|) (p-value) gives us an indicator of how strongly the data gives evidence to support the estimate. Generally p-value of 0.05 and lower is held as being “statistically significant” but it is not a straight yes or no answer. We can just see here that a lower p-value means that the estimate is more believable. In the output also marked by the amount of stars.

For Nintendo we can see that Sports games seem to be the best sellers. For Activision Shooters seem to also be the best sellers. Lastly for Electronic Arts, racing games seem to be better sellers than most. Although there is some indication that shooter games might perform better, but the p-value suggest that there is not enough evidence to draw a definite conclusion.

```r
compare_publisher_genres <- function(publisher_name, genre_name) {
  sales_data_filtered <- sales_data %>%
    filter(Publisher == publisher_name)

  sales_data_filtered$Genre <- relevel(sales_data_filtered$Genre, ref = genre_name)

  lm_model <- lm(Global_Sales ~ Genre, data = sales_data_filtered)
  summary(lm_model)
}

compare_publisher_genres("Nintendo", "Sports")
```

```
##
## Call:
## lm(formula = Global_Sales ~ Genre, data = sales_data_filtered)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.989 -2.202 -1.302 -0.029 78.776
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.9638     0.7587   5.224 2.32e-07 ***
## GenreAction        -2.3215     0.9907  -2.343   0.0194 *
## GenreAdventure     -2.9435     1.2166  -2.419   0.0158 *
## GenreFighting      -0.9999     1.5279  -0.654   0.5131
## GenreMisc          -2.1571     0.9446  -2.284   0.0227 *
## GenrePlatform      -0.1244     0.9278  -0.134   0.8934
## GenrePuzzle        -2.2763     1.0017  -2.272   0.0234 *
## GenreRacing         0.1254     1.1964   0.105   0.9166
## GenreRole-Playing  -1.2536     0.9366  -1.339   0.1812
## GenreShooter       -1.0601     1.3765  -0.770   0.4415
## GenreSimulation    -0.9192     1.3063  -0.704   0.4819
## GenreStrategy      -3.1019     1.2637  -2.455   0.0144 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.627 on 684 degrees of freedom
## Multiple R-squared:  0.03315,    Adjusted R-squared:  0.0176
## F-statistic: 2.132 on 11 and 684 DF,  p-value: 0.01652
```

```
compare_publisher_genres("Activision", "Shooter")
```

```
##
## Call:
## lm(formula = Global_Sales ~ Genre, data = sales_data_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8958 -0.4405 -0.2205  0.0668 12.8542
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.9058     0.1251  15.237  < 2e-16 ***
## GenreAction        -1.4454     0.1534  -9.425  < 2e-16 ***
## GenreAdventure     -1.6890     0.3356  -5.033 5.78e-07 ***
## GenreFighting      -1.4972     0.6017  -2.488  0.01301 *
## GenreMisc          -1.1626     0.1980  -5.873 5.91e-09 ***
## GenrePlatform      -1.3491     0.2368  -5.698 1.61e-08 ***
## GenrePuzzle        -1.7544     0.6017  -2.916  0.00363 **
## GenreRacing        -1.6731     0.2210  -7.569 8.86e-14 ***
## GenreRole-Playing  -0.7646     0.2735  -2.796  0.00528 **
## GenreSimulation    -1.5467     0.3480  -4.445 9.82e-06 ***
## GenreSports        -1.3765     0.1809  -7.610 6.58e-14 ***
## GenreStrategy      -1.1013     0.3548  -3.104  0.00196 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.557 on 954 degrees of freedom
## Multiple R-squared:  0.1086, Adjusted R-squared:  0.09833
## F-statistic: 10.57 on 11 and 954 DF,  p-value: < 2.2e-16
```

```
compare_publisher_genres("Electronic Arts", "Racing")
```

```
##
## Call:
## lm(formula = Global_Sales ~ Genre, data = sales_data_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1286 -0.5877 -0.3060  0.1790  7.6440
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.91679    0.08305  11.039   <2e-16 ***
## GenreAction       -0.28306    0.11367  -2.490   0.0129 *
## GenreAdventure    -0.55141    0.30208  -1.825   0.0682 .
## GenreFighting     -0.10495    0.18909  -0.555   0.5790
## GenreMisc         -0.45772    0.18000  -2.543   0.0111 *
## GenrePlatform     -0.50867    0.27465  -1.852   0.0642 .
## GenrePuzzle       -0.26679    0.40442  -0.660   0.5096
## GenreRole-Playing  0.09178    0.19552   0.469   0.6389
## GenreShooter       0.22177    0.12160   1.824   0.0684 .
## GenreSimulation   -0.14498    0.12787  -1.134   0.2571
## GenreSports       -0.07078    0.09421  -0.751   0.4526
## GenreStrategy     -0.53625    0.19114  -2.806   0.0051 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.047 on 1327 degrees of freedom
## Multiple R-squared:  0.02814,    Adjusted R-squared:  0.02009
## F-statistic: 3.493 on 11 and 1327 DF,  p-value: 7.802e-05
```

Here we use random forests to see which of the variables have the highest predictive power to the global sales. In summary, higher %IncMSE (Mean Squared Error) and IncNodePurity indicate that the variable is a better predictor for the target variable.

We can see that sales in Europe and North America are the most important variables in explaining global sales. Which are higher than sales in Japan and other parts of the world. This is the same trend we saw on the graphs. Other than those, we can see that the platform and publisher are the most important factors in game sales. This is fairly straightforward logic: More popular consoles and publishers lead to better global sales. And finally we can see that the genre and release year of the game have had the least amount of impact on global sales.

```
publisher_freq <- table(sales_data$Publisher)
threshold <- 50
sales_data$Publisher <- ifelse(sales_data$Publisher %in% names(publisher_freq[publisher_freq < threshol
                         "Other",
                         sales_data$Publisher)
```

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

rf_model <- randomForest(Global_Sales ~ Platform + Year + Genre + Publisher +
                           JP_Sales + EU_Sales + NA_Sales + Other_Sales, data = sales_data, importance =
importance(rf_model)
```

```
##               %IncMSE IncNodePurity
## Platform    12.672141     1050.6981
## Year         6.573761      471.8115
## Genre        4.666340      724.5230
## Publisher   14.157199     1362.8403
## JP_Sales    19.596483     4365.0831
## EU_Sales    27.853691    10057.9435
## NA_Sales    33.887358    13988.4700
## Other_Sales 20.366490     7564.5641
```