615 Final Project

Video Game Sales

Xinyi Wang

12/9/2018

Introduction

Video gaming has a long history as far back as the early 1950s but it did not reach mainstream popularity until the 1970s and 1980s. Video games are a billion-dollar business and have been for many years. In 2016, the video game market in the United States was valued at 17.68 billion U.S. dollars.

The purpose of this project is to demonstrate what I have learned and my ability to extend knowledge as apply R in new situations - which is explore the development of the video game industry from 1983-2016.

Read and Clean Data

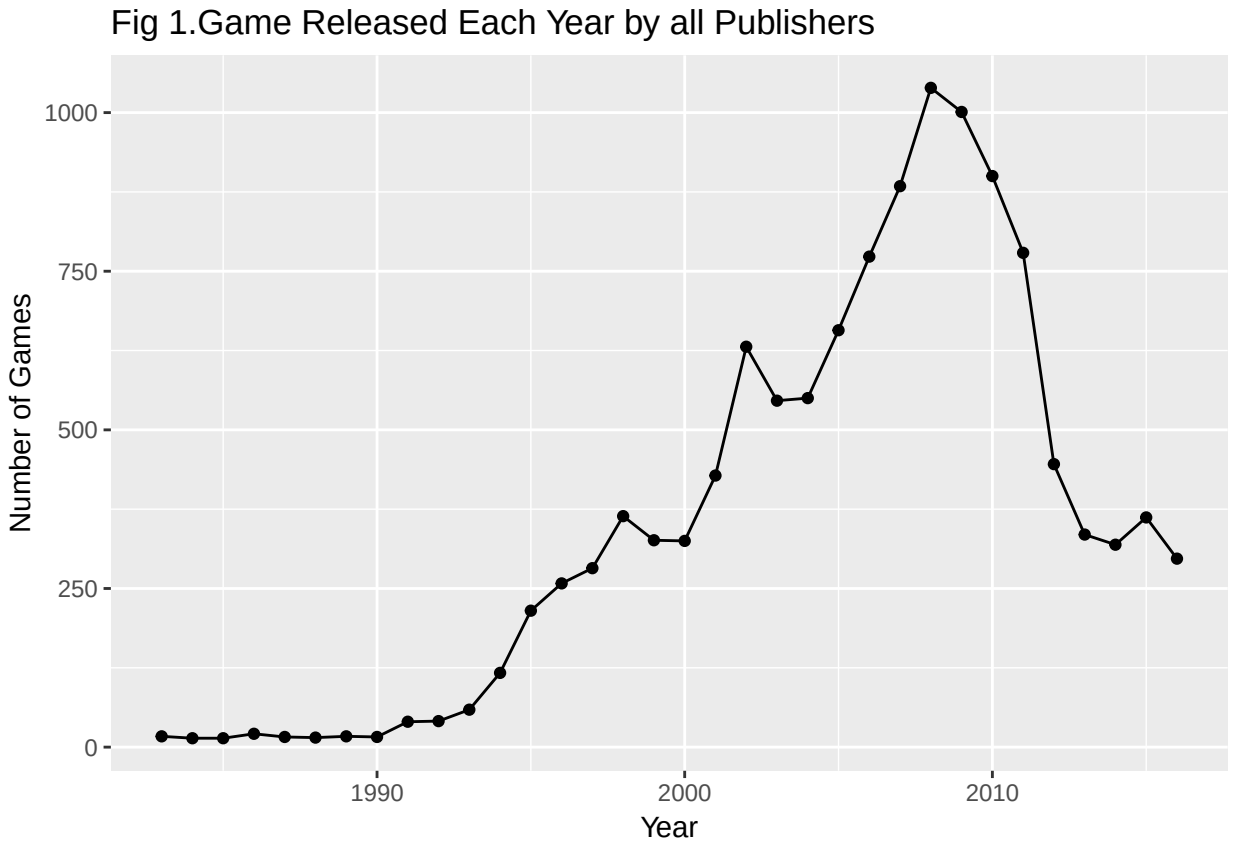
```
## Observations: 16,353
## Variables: 16
## $ Name            <chr> "Wii Sports", "Super Mario Bros.", "Mario Kart...
## $ Platform        <fct> Wii, NES, Wii, Wii, GB, GB, DS, Wii, Wii, NES,...
## $ Year_of_Release <dbl> 2006, 1985, 2008, 2009, 1996, 1989, 2006, 2006...
## $ Genre           <fct> Sports, Platform, Racing, Sports, Role-Playing...
## $ Publisher       <fct> Nintendo, Nintendo, Nintendo, Nintendo, Ninten...
## $ NA_Sales        <dbl> 41.36, 29.08, 15.68, 15.61, 11.27, 23.20, 11.2...
## $ EU_Sales        <dbl> 28.96, 3.58, 12.76, 10.93, 8.89, 2.26, 9.14, 9...
## $ JP_Sales        <dbl> 3.77, 6.81, 3.79, 3.28, 10.22, 4.22, 6.50, 2.9...
## $ Other_Sales     <dbl> 8.45, 0.77, 3.29, 2.95, 1.00, 0.58, 2.88, 2.84...
## $ Global_Sales    <dbl> 82.53, 40.24, 35.52, 32.77, 31.37, 30.26, 29.8...
## $ Critic_Score    <int> 76, NA, 82, 80, NA, NA, 89, 58, 87, NA, NA, 91...
## $ Critic_Count    <int> 51, NA, 73, 73, NA, NA, 65, 41, 80, NA, NA, 64...
## $ User_Score      <fct> 8, , 8.3, 8, , , 8.5, 6.6, 8.4, , , 8.6, , 7.7...
## $ User_Count      <int> 322, NA, 709, 192, NA, NA, 431, 129, 594, NA, ...
## $ Developer       <fct> Nintendo, , Nintendo, Nintendo, , , Nintendo, ...
## $ Rating          <fct> E, , E, E, , , E, E, E, , , E, , E, E, E, M, M...
```

First step is to read and clean data. I removed all observations that do not have game names and released after 2016 since the dataset is from web scrape of VGChartz in late 2016.

EDA

The next important part is data visualization. In this project, I will focus on publishers, game sales by region, platform, and genre.

1.Games Released Each Year by All Publishers



We can see from Fig 1 that there are a lot of video games released during 2005-2008, which means the competition became intense in the late 2000s.

2.top 10 publishers with higher revenue

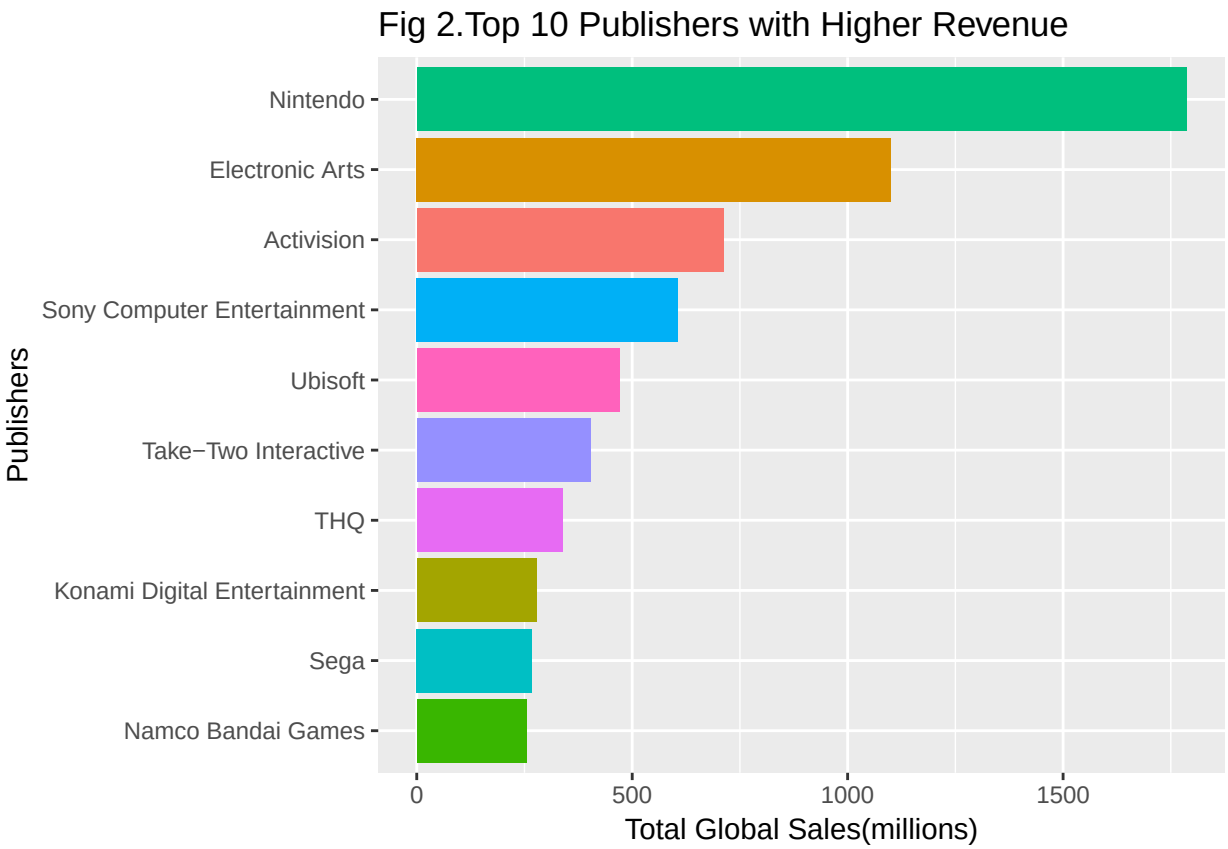


Fig 2 shows that form 1983 to 2016, Nintendo is the publisher who get highest revenue and follows by Electronic Arts and Activision. Surprisingly that Sony Computer Entertainment not reached half of Nintendo's total revenue, by looking into the whole publishers name I found that is more likely because Sony Company has a lot of branches such as "Sony Online Entertainment","Sony Music Entertainment",etc.

3.Sales per region

Fig 3.Sales per region

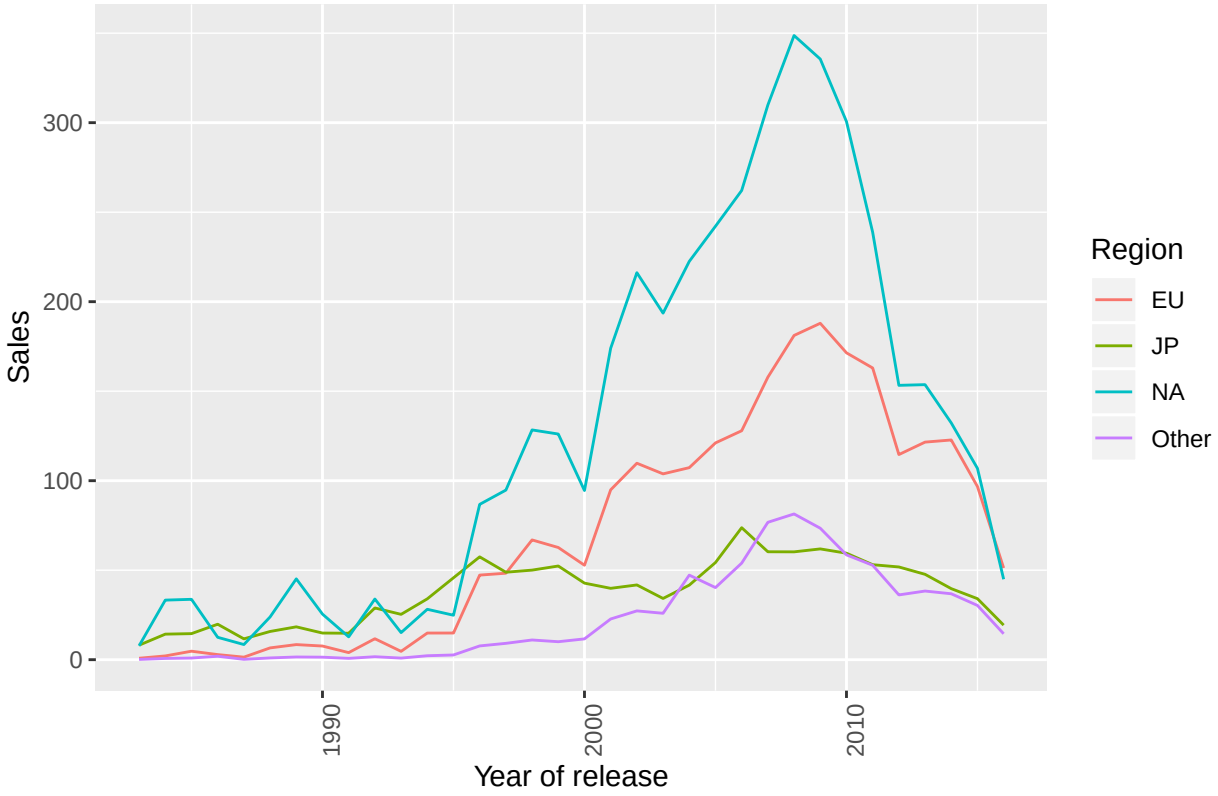


Fig 3 shows the sales trend for different regions from 1983 to 2016. It's clear that North America had the most drastic spike in the late 2000s.

4.global sales map by platform and year

Fig 4.global sales map by platform and year

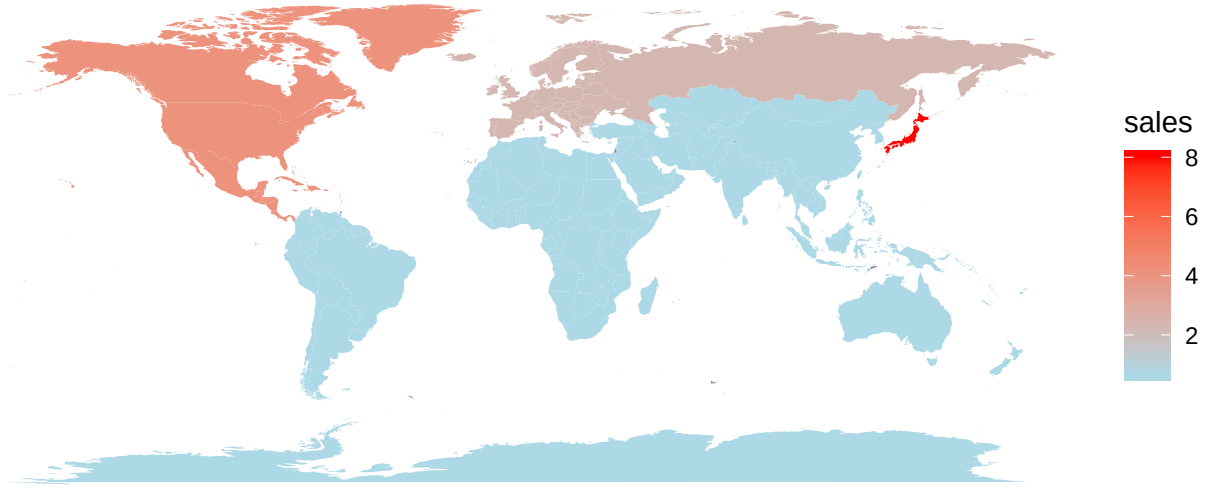


Fig 4 is a global heat map to visualize sales revenue by platform and year. I take Platform="3DS" and

Year_of_Release="2016" as example. As we can see, Japan had highest sales revenue for 3DS video games in 2016. Users could also adjust Platform and Year of Release in Shiny App.

5.game name text analysis (word cloud)

Fig 5. Word Cloud

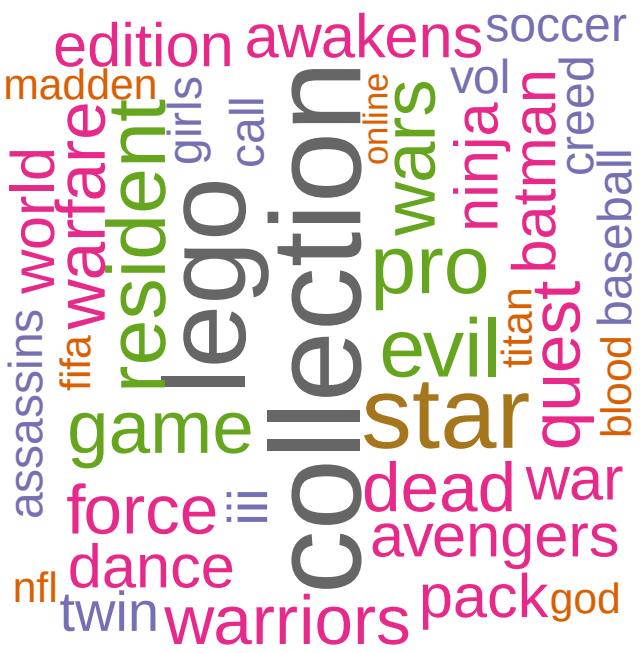


Fig 5 is a word cloud allows us to highlight the most frequently used keywords in texts. The above word cloud clearly shows that "collection", "lego", "wars", "pro" and "evil" are the five most important words in game names which released 2016.

6.Sales by Genre

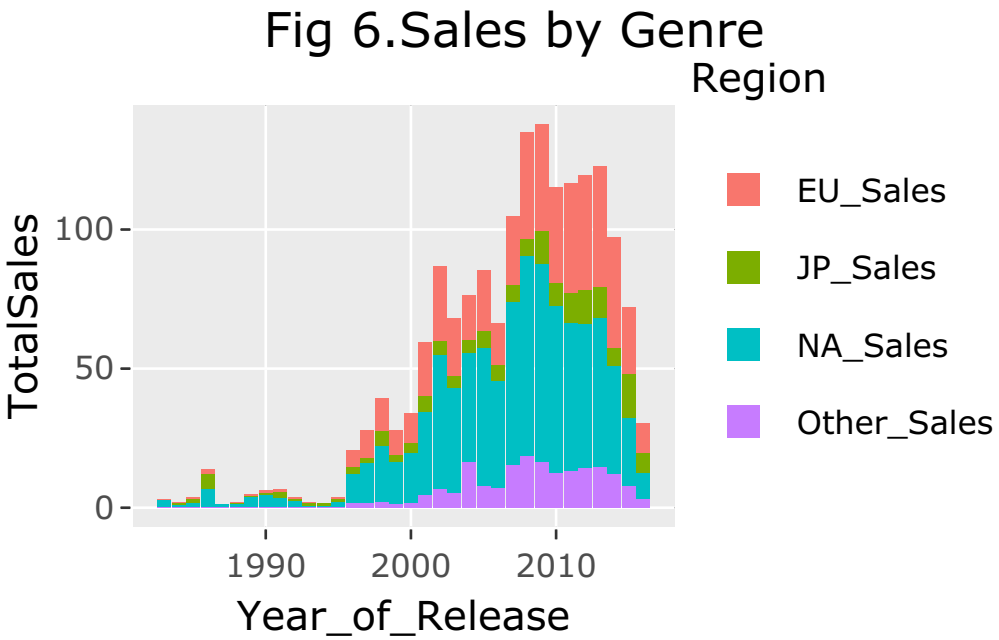


Fig 6 shows the distribution of sales by genre. It is clear that North America always contributes most of sales, however in 2015 and 2016, it seems all four regions had equal sales amount.

7.Pie chart of genres in certain year and platform

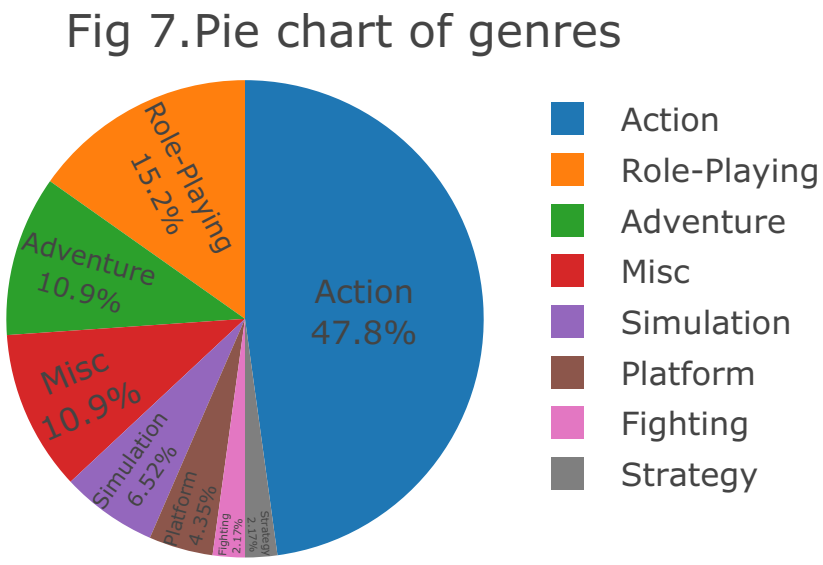


Fig 7 is pie chart of genres in certain year and platform. I took Platform="3DS" and Year_of_Release="2016" as example, it shows us Action is the largest portion. Same as the global sales map above(Fig 1), Platform and Year of Release are also adjustable in Shiny App.

Top 10 Games

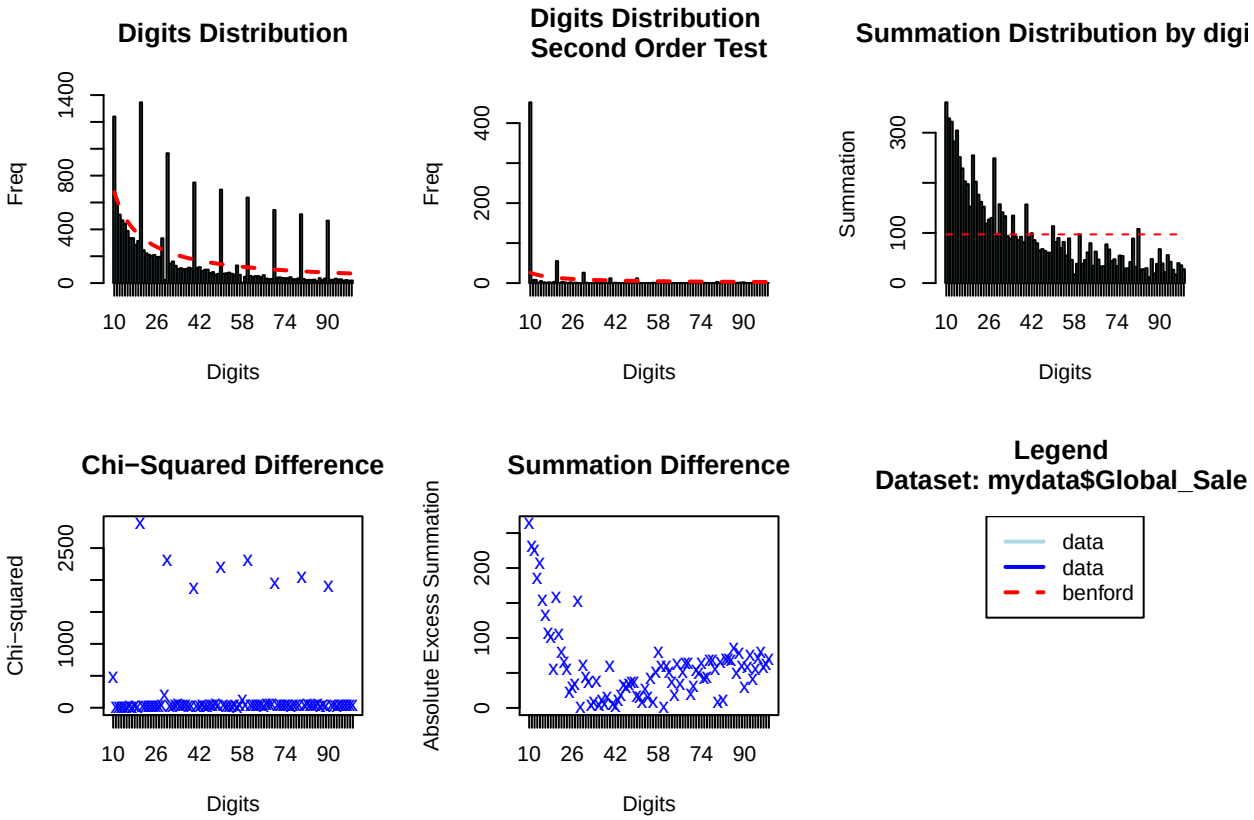
```
#Take 2016 as example
toptable <- mydata %>%
  select(Name,Global_Sales,Year_of_Release,Platform) %>%
  filter(Year_of_Release==2016) %>%
  arrange(desc(Global_Sales))%>%
  select(Name,Platform,Global_Sales)%>%
  head(10)
toptable
```

```
##                             Name Platform Global_Sales
## 1                       FIFA 17      PS4         7.59
## 2                Pokemon Sun/Moon      3DS         7.14
## 3      Uncharted 4: A Thief's End      PS4         5.38
## 4  Call of Duty: Infinite Warfare      PS4         4.46
## 5                   Battlefield 1      PS4         4.08
## 6       Tom Clancy's The Division      PS4         3.80
## 7                       FIFA 17     XOne         2.65
## 8  Call of Duty: Infinite Warfare     XOne         2.42
## 9                 Far Cry: Primal      PS4         2.26
## 10                  Battlefield 1     XOne         2.25
```

I also made a table to show users that most popular games in a certain year, here is an example when "2016" is selected.

Benford Law

```
library(benford.analysis)
bfd <- benford(mydata$Global_Sales)
plot(bfd)
```




```
library(BenfordTests)
# Euclidean Distance Test for Benford's Law
edist.benftest(mydata$Global_Sales)
```

```
##
##  Euclidean Distance Test for Benford Distribution
##
## data:  mydata$Global_Sales
## d_star = 2.6476, p-value < 2.2e-16
```

The p-value is smaller than 0.05 so that we reject the null hypothesis. Therefore, the goodness-of-fit test based on the Euclidean distance between the first digits' distribution and Benford's distribution shows the data does not conform to Benford's law very well.