## Samray Estifanos

6/17/2021

# Final Project

**Describe the data set and why you selected it for this project.**

I chose the 'Video Game sales' data set for my final project. I chose this data set because I enjoy playing video games and it would be interesting to analyze which games were popular and for what reasons.

Video Game Sales Data: This document has more that 16,500 Observations.

```
vgsales.csv <- read.csv('vgsales.csv')
View(vgsales.csv)
summary(vgsales.csv)
```

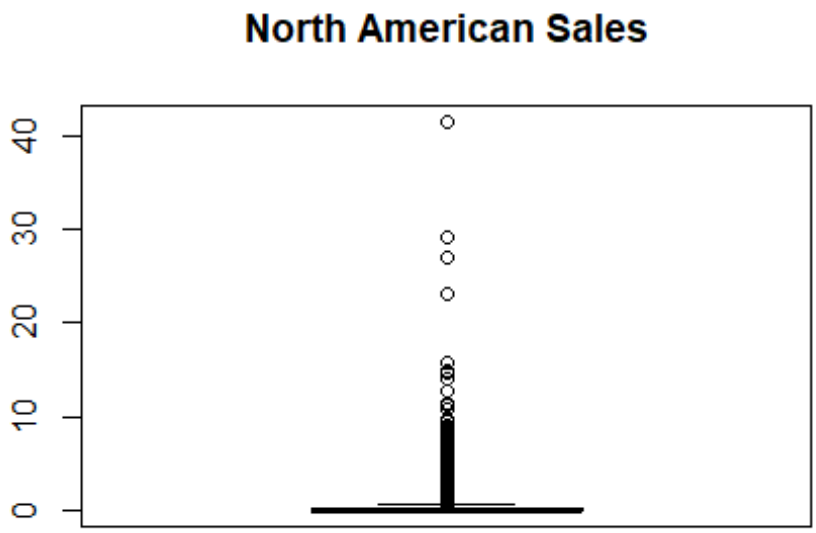
```
##       Rank           Name              Platform            Year          
##  Min.   :    1   Length:16598       Length:16598       Length:16598      
##  1st Qu.: 4151   Class :character   Class :character   Class :character  
##  Median : 8300   Mode  :character   Mode  :character   Mode  :character  
##  Mean   : 8301                                                           
##  3rd Qu.:12450                                                           
##  Max.   :16600                                                           
##     Genre            Publisher            NA_Sales          EU_Sales       
##  Length:16598       Length:16598       Min.   : 0.0000   Min.   : 0.0000  
##  Class :character   Class :character   1st Qu.: 0.0000   1st Qu.: 0.0000  
##  Mode  :character   Mode  :character   Median : 0.0800   Median : 0.0200  
##                                        Mean   : 0.2647   Mean   : 0.1467  
##                                        3rd Qu.: 0.2400   3rd Qu.: 0.1100  
##                                        Max.   :41.4900   Max.   :29.0200  
##     JP_Sales         Other_Sales        Global_Sales      
##  Min.   : 0.00000   Min.   : 0.00000   Min.   : 0.0100  
##  1st Qu.: 0.00000   1st Qu.: 0.00000   1st Qu.: 0.0600  
##  Median : 0.00000   Median : 0.01000   Median : 0.1700  
##  Mean   : 0.07778   Mean   : 0.04806   Mean   : 0.5374  
##  3rd Qu.: 0.04000   3rd Qu.: 0.04000   3rd Qu.: 0.4700  
##  Max.   :10.22000   Max.   :10.57000   Max.   :82.7400
```

**Describe any processing problems you identified with the data and how you overcame those issues.**

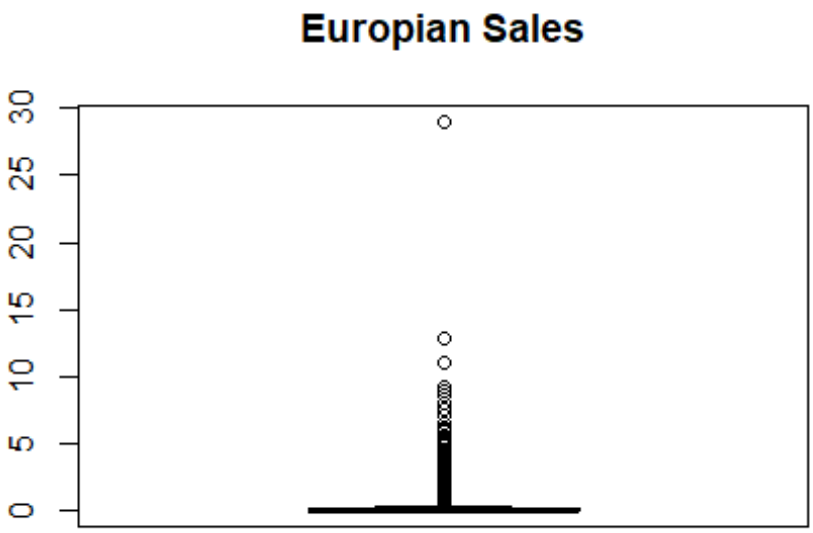
There were a lot of problems that I came across.

- The first being that I never used R before and learning it quickly was overwhelming. I had a hard time with cleaning the data set and pulling information that I wanted to analyze. I was able to narrow down what I wanted to analyze and create something.
- The second problem was trying to figure out which direction I wanted to go with this project. I found it easier to pick out some variables and create box plots that showed the sales in different parts of the globe.

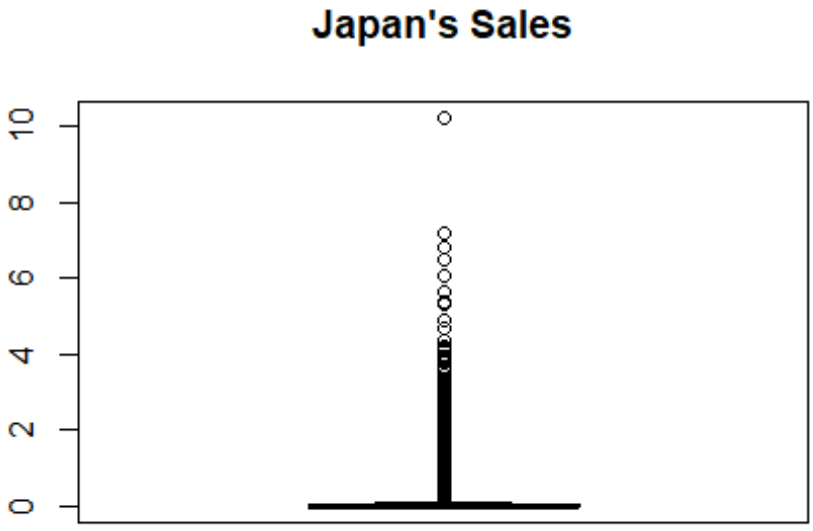
```
title=c("North American Sales")
boxplot(vgsales.csv$NA_Sales,main=title)
```



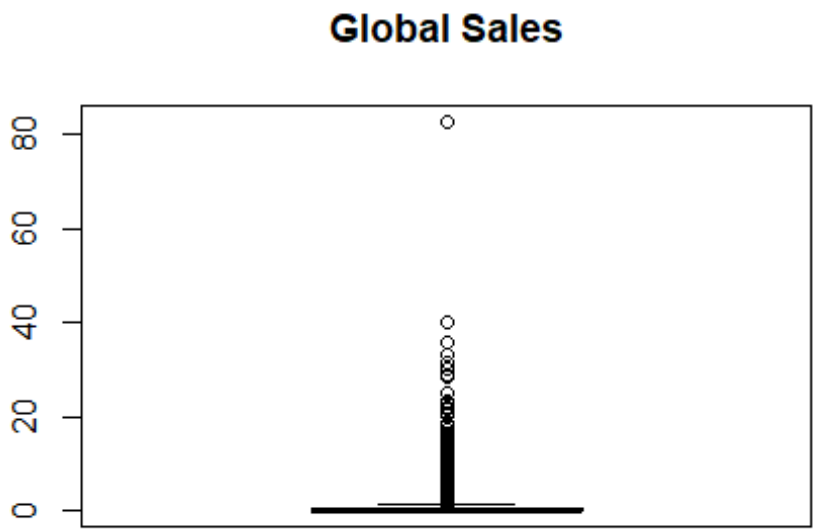

```
title=c("Europian Sales")
boxplot(vgsales.csv$EU_Sales,main=title)
```




```
title=c("Japan's Sales")
boxplot(vgsales.csv$JP_Sales,main=title)
```



```
title=c("Global Sales")
boxplot(vgsales.csv$Global_Sales,main=title)
```

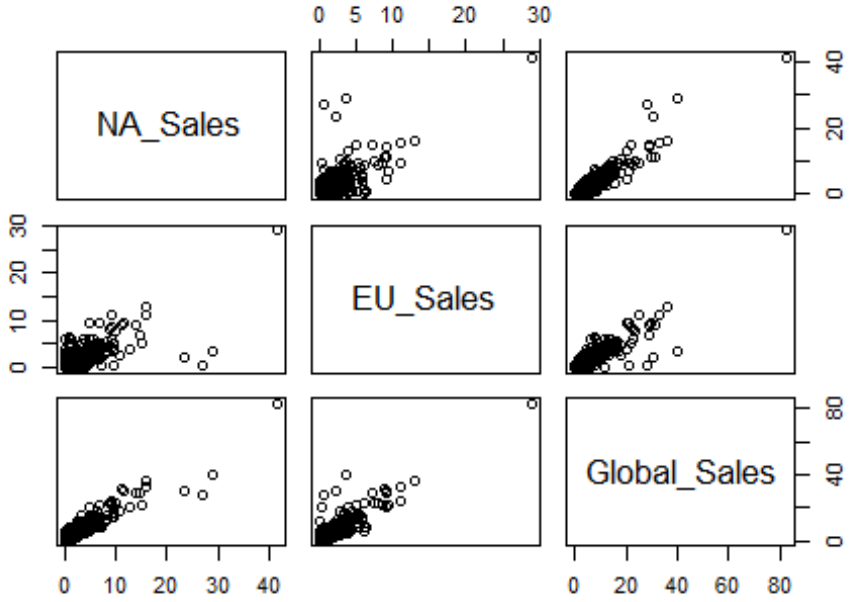



**Describe the results of your exploratory analysis and what preliminary conclusions you were able to draw based on this analysis.**

For my exploratory analysis I wanted to pull out North American sales, European Sales and Global sales to compare and contrast which sets of data were better in which parts of the world. I created a scatter plot matrix that shows the data. I found this very helpful because I can see how North American sales compares to Global sales. I can also see the best fit line in each plot. This showed me the average sale price across the world as well as the different outliers. By analyzing these plots I was able to come to the conclusion that Nintendo had the best sales in North America and the Wii out performed every other system. I have a theory as to why that is. I think that based on the data, each game that was listed for the Nintendo Wii are multiplier games. Whereas on the other platforms are single player games. The Nintendo Wii has a larger target audience, which is why its sales were so successful.

```
pairs(~NA_Sales+EU_Sales+Global_Sales,data=vgsales.csv,
   main="Scatterplot Matrix about Video Game Sales Globally")
```

## Scatterplot Matrix about Video Game Sales Globally




***Describe how you selected the methodology for your analysis of the big question and the pros and cons of that method and any alternative methods you considered.***

I chose a data set that I was comfortable with. Out of all of the options I chose Video Game Sales because I am into video games and I know/played some of the games that were mentioned in the data set. Once I went through the data I saw the different games, systems, manufacturers and the different sales. So, I was interested in the sales portion of the data set and wanted to find out 'What makes the top games so popular?'.

There were a lot of pros and cons because, in the data set there is a ranking category. Based on that ranking category it looked obvious which games were the most popular and the corresponding systems. So, I wanted to visualize that. It was really difficult for me to find an initial direction for me to go for this project, because the data was overwhelming. I decided to think small and look at the first 10 columns of data and analyze that to find my answer. The only con with that was trying to find a way to incorporate all of the data.

After analyzing the first 10 column I wanted to compare it from a global perspective so I created a scatter plot matrix to show all the data. By doing this I hope to answer my big question.

***Describe your final conclusions based on your analysis and support them with analytics on your data set.***

In conclusion I chose the video game sales data to find out which games were the most popular based on their sales. I found out that Nintendo was the best publisher and the Nintendo Wii was the best publisher. When examining the data and having prior knowledge of video game history. I noticed that the Nintendo Wii had more multiplayer games than single player games, which increased sales while others only had single player games.

***Describe any additional analyses that you would have liked to carry out and any additional data that would have been needed in order to extend your analysis.***

I would have liked to find out what were the least performing platforms. I feel like if I had more time to learn more about R I could have created a better scatter plot that showed some sort of correlation between the popular platforms and the least popular platforms.