CSCI 385 - First Deliverable

Brandon Weaver

9/26/2020

Video Games Sales Analysis

Introduction

The domain for this portfolio is Video Games. My reason for picking this topic is that Video Games are one of my favorite hobbies and I think it would be interesting if it would be possible to make predictions of future trends based on the success and failures of the past. Studying Video Games is important because it is one of the fastest growing entertainment industries and I believe that they will only grow in relevance in the future. The ultimate goal of this research would be to create a predictive model where an algorithm could predict the sales of a video game before they are released.

Data Set

```r
VGSales <- read_csv("Video_Games_Sales.csv")
```

```
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   Platform = col_character(),
##   Year_of_Release = col_double(),
##   Genre = col_character(),
##   Publisher = col_character(),
##   NA_Sales = col_double(),
##   EU_Sales = col_double(),
##   JP_Sales = col_double(),
##   Other_Sales = col_double(),
##   Global_Sales = col_double(),
##   Critic_Score = col_double(),
##   Critic_Count = col_double(),
##   User_Score = col_double(),
##   User_Count = col_double(),
##   Developer = col_character(),
##   Rating = col_character()
## )
```

This data set comes from the Kaggle users Rush Kirubi and Gregory Smith. This data is the results of a web scrape of the game review data from the website Metacritic and a web scrape of the sales records of video games from the website VGChartz.

One current limitation with this data set is that games older than 2002 have sales data from VGChartz but do not have critic scores from Metacritic. One way to fix this limitation would be to explore another critic score aggregate website and see if they include an average score for older games.

Another limitation of this data is that it does not include any data on games that are hosted on the mobile phone platform. This is important because some of the best selling games are located on the phone. Therefore this introduces a bias to the data because certain genres that are more popular on phones, like puzzle games (Ex. Candy Crush), are not represented. While genres that are more traditionally popular on consoles, like action and shooter games, receive more representation.

One form of Tidying that I did to the data was that some of the user scores were labeled as "tbd," which means to be determined. I altered those to be equal to NA and then I was then able to make that column a double rather than characters.
This was done by using the replace feature by pressing "command f" when inside of the csv and then I swapped all of the instances of "tbd" with a blank space.

I had to do a similar process with the Year_of_Release column and replace several rows that stated "N/A" with a blank space.

Another Tidying that I had to do to the data was that I multiplied all of the values in User_Score by 10 so that both the `User_Score` and the `Critic_Score` were out of 1 to 100 rather than the `User_Score` being out of 0.01 to 10. This will allow for the two scores to be properly compared.

```
VGSales$User_Score <- VGSales$User_Score * 10
```

Here is a showcase of the first 6 rows in the data set

```
head(VGSales)
```

```
## # A tibble: 6 x 16
##   Name  Platform Year_of_Release Genre Publisher NA_Sales EU_Sales JP_Sales
##   <chr> <chr>              <dbl> <chr> <chr>        <dbl>    <dbl>    <dbl>
## 1 Wii ~ Wii                 2006 Spor~ Nintendo      41.4    29.0      3.77
## 2 Supe~ NES                 1985 Plat~ Nintendo      29.1     3.58     6.81
## 3 Mari~ Wii                 2008 Raci~ Nintendo      15.7    12.8      3.79
## 4 Wii ~ Wii                 2009 Spor~ Nintendo      15.6    10.9      3.28
## 5 Poke~ GB                  1996 Role~ Nintendo      11.3     8.89    10.2
## 6 Tetr~ GB                  1989 Puzz~ Nintendo      23.2     2.26     4.22
## # ... with 8 more variables: Other_Sales <dbl>, Global_Sales <dbl>,
## #   Critic_Score <dbl>, Critic_Count <dbl>, User_Score <dbl>, User_Count <dbl>,
## #   Developer <chr>, Rating <chr>
```

- `Name`- `chr` - The games name
- `Platform` - `chr` - Platform of the games release (i.e. PC, PS4, etc.)
- `Year_of_Release` - `dbl` - Year of the game's release, ranges from 1980 to 2016
- `Genre` - `chr` - Genre of the game (i.e Action, Puzzle, Sports, etc.)
- `Publisher` - `chr` - Publisher of the game
- `NA_Sales` - `dbl` - Sales in North America (in million copies sold)
- `EU_Sales` - `dbl` - Sales in Europe (in million copies sold)
- `JP_Sales` - `dbl` - Sales in Japan (in million copies sold)
- `Other_Sales` - `dbl` - Sales in the rest of the world (in million copies sold)
- `Global_Sales` - `dbl` - Total worldwide sales. (in million copies sold)
- `Critic_Score` - `dbl` - Aggregate score compiled by Metacritic staff, ranges from 1 to 100
- `Critic_Count` - `dbl` - The number of critics used in coming up with the Critic score
- `User_score` - `dbl` - Score by Metacritic's subscribers, ranges from 1 to 100

- `User_Count` - `dbl` - Number of users who gave the user score
- `Developer` - `chr` - Party responsible for creating the game
- `Rating` - `chr` - The ESRB ratings from E (Everyone) to AO (Adults Only). This is the rating system used in North America.

Exploratory Data Analysis

Considering each of the numeric variables in the `VGSales` data set individually, we find the following summary statistics.

```
dblInVGSales <- select(VGSales, Year_of_Release, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sale
summary(dblInVGSales)
```

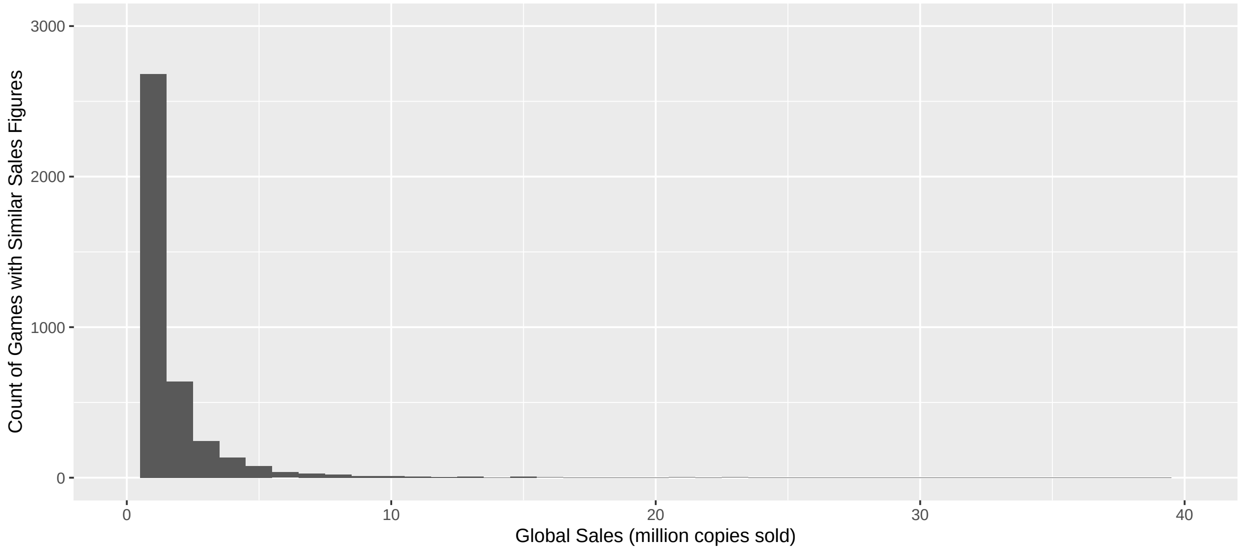
```
##  Year_of_Release    NA_Sales          EU_Sales         JP_Sales
##  Min.   :1980    Min.   : 0.0000   Min.   : 0.000   Min.   : 0.0000
##  1st Qu.:2003    1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.0000
##  Median :2007    Median : 0.0800   Median : 0.020   Median : 0.0000
##  Mean   :2006    Mean   : 0.2633   Mean   : 0.145   Mean   : 0.0776
##  3rd Qu.:2010    3rd Qu.: 0.2400   3rd Qu.: 0.110   3rd Qu.: 0.0400
##  Max.   :2020    Max.   :41.3600   Max.   :28.960   Max.   :10.2200
##  NA's   :269
##   Other_Sales        Global_Sales      Critic_Score     Critic_Count
##  Min.   : 0.00000   Min.   : 0.0100   Min.   :13.00   Min.   :  3.00
##  1st Qu.: 0.00000   1st Qu.: 0.0600   1st Qu.:60.00   1st Qu.: 12.00
##  Median : 0.01000   Median : 0.1700   Median :71.00   Median : 21.00
##  Mean   : 0.04733   Mean   : 0.5335   Mean   :68.97   Mean   : 26.36
##  3rd Qu.: 0.03000   3rd Qu.: 0.4700   3rd Qu.:79.00   3rd Qu.: 36.00
##  Max.   :10.57000   Max.   :82.5300   Max.   :98.00   Max.   :113.00
##                                       NA's   :8582    NA's   :8582
##    User_Score      User_Count
##  Min.   : 0.00   Min.   :    4.0
##  1st Qu.:64.00   1st Qu.:   10.0
##  Median :75.00   Median :   24.0
##  Mean   :71.25   Mean   :  162.2
##  3rd Qu.:82.00   3rd Qu.:   81.0
##  Max.   :97.00   Max.   :10665.0
##  NA's   :9129    NA's   :9129
```

One take away from this summary is a possible error in the data collection. This data set was supposed to only included games that were released before 2017. However the max year that is given is 2020. upon looking at those rows, it seems that it was possibly due to a typo on one of the websites that was scrapped because all of the games that had a 2020 value were actually released on an earlier year.

```
ggplot(data = VGSales) + geom_histogram(mapping = aes(x = Global_Sales), binwidth = 1) + xlim(0,40) + y
  labs(x = "Global Sales (million copies sold)", y = "Count of Games with Similar Sales Figures")
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



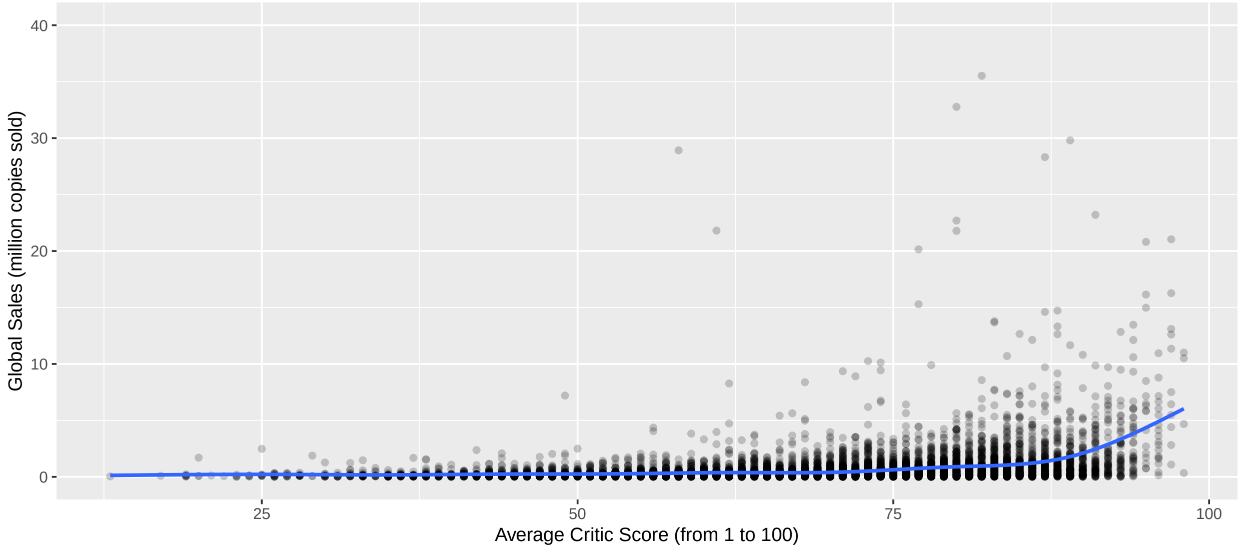
This graph displays how most video games struggle to exceed 5 million copies sold. This means that the distribution of games sales is Right Skewed. This is important because it demonstrates how only a select few of games are actually able to succeed in the current market and how it is important to do market research in order to give a game a better chance of selling well.

```r
ggplot(data = VGSales, aes(x = Critic_Score, y = Global_Sales)) + geom_point(alpha = 0.2) + geom_smooth
  labs(x = "Average Critic Score (from 1 to 100)", y = "Global Sales (million copies sold)")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 8583 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 8583 rows containing missing values (geom_point).
```



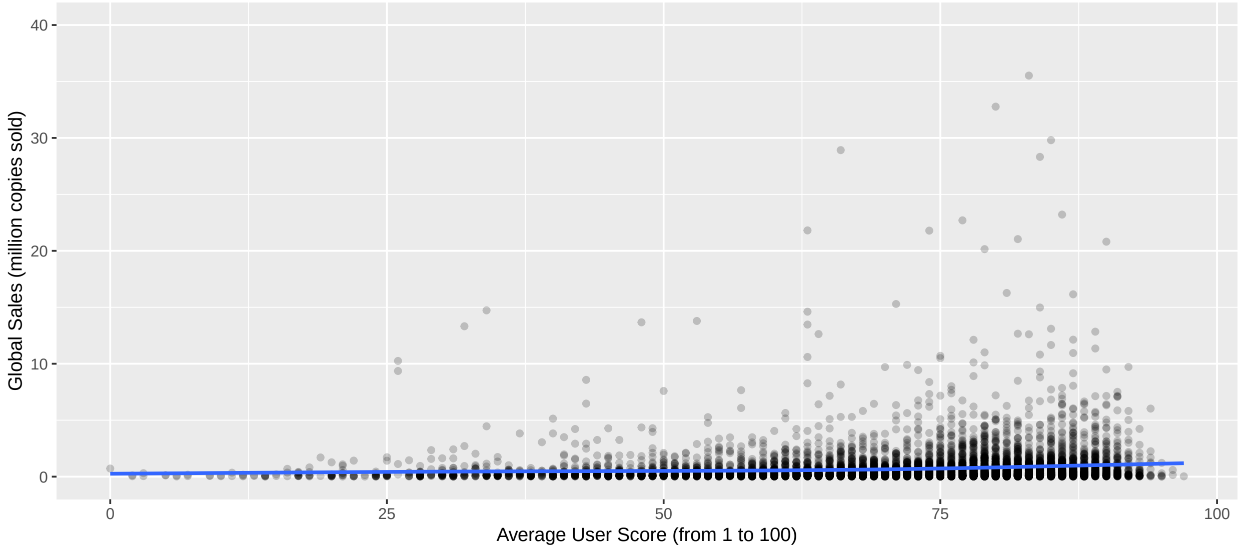
This graph leads to the conclusion that the Critic Score has little impact on a game's sales unless the game has an average critic score that is greater than 75. In which case there seems to be an increasing trend. This shows how critics still play a vital role in helping to promote game sales even in the age of internet where information about games is more accessible.

```
ggplot(data = VGSales, aes(x = User_Score, y = Global_Sales)) + geom_point(alpha = 0.2) + geom_smooth(s
  labs(x = "Average User Score (from 1 to 100)", y = "Global Sales (million copies sold)")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 9130 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9130 rows containing missing values (geom_point).
```



In contrast, when looking at the average User Scores, there is q small increase in sales when the average User Score is higher. However, there is no pronounced increase in the slope of the line. In comparison, in the previous graph there was a dramatic increase after the 75 score threshold was breached. This is interesting because one would think that what the average user thinks about a game would be a valid way to tell if a game sells well.

```
salesByGenre <- VGSales %>%
  select(Genre, NA_Sales, JP_Sales, EU_Sales, Other_Sales) %>%
  group_by(Genre) %>%
  summarise(NorthAmerica = sum(NA_Sales), Japan = sum(JP_Sales), Europe = sum(EU_Sales), Other = sum(Ot
  gather("Region", "Games_Sold", -Genre)
```

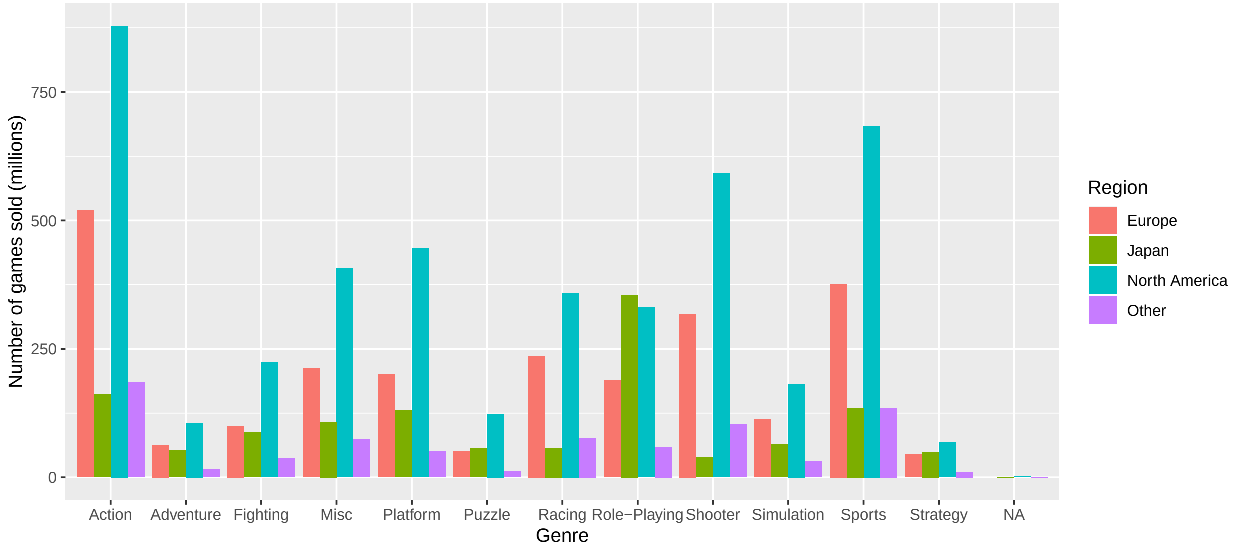
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(salesByGenre)
```

```
## # A tibble: 6 x 3
##   Genre     Region       Games_Sold
##   <chr>     <chr>             <dbl>
## 1 Action    NorthAmerica       879.
## 2 Adventure NorthAmerica       105.
## 3 Fighting  NorthAmerica       223.
## 4 Misc      NorthAmerica       407.
## 5 Platform  NorthAmerica       446.
## 6 Puzzle    NorthAmerica       123.
```

This new table groups all the games by their Genre and then adds together all of their sales in each region that is explored in this data set. Then it is reconstructed so that each all of the regions are in the same column. This was done so that they can all be compared on the same histogram.

```
ggplot(data = salesByGenre) + geom_bar(aes(x=Genre, y=Games_Sold, fill = Region),stat = "identity", posi
  labs(y = "Number of games sold (millions)") + scale_fill_discrete(labels=c("Europe", "Japan", "North A
```



One interesting take away from this graph is how the sales trends in Europe compared to North America are very similar, the only difference being that the sales in Europe are on a smaller scale. This is in contrast to Japan where the sales trends are completely different. For example, Role Playing Games are by far the most popular genre in Japan with other genres like action and sports not being popular at all. Where as in North America and Europe, Action and Sports games are the two most popular genres. This shows hows if a publisher is trying to maximize there sales in a certain region that the genre of the game will have a major impact based on the region.

Another interesting static to explore in the future would be to take this graph but also factor in population size. For example, the population of North America and Europe is much greater than Japan. Therefore the sales of games should be relatively higher. So it would be important to see if the populations were normalized if the comparisons between genre sales and the region would be similar.

Data Science Questions

1. Do Critic scores correlate with higher game sales?
   This data science question is important because in the past couple of years review outlets such as IGN and GameSpot have been criticized because they have been know to generally only give high review scores. Therefore it is important to explore a possible reason in the shift of giving higher scores.

2. How much of an impact does the genre and the platform affect game sales?
   This data science question is important because the role that a platform fills for its console generation has a definite impact on how well a game based on its genre will sell. For example, In the 1990s the two main consoles were the Super Nintendo and the Sega Genesis, with the Super Nintendo targeting the family audience while the Sega Genesis targeted the more mature audience. Therefore if an algorithm was made to predict game sales. Then looking into what current platform it is being developed for and what role in the market that this platform provides, family or mature audience, would be an important factor to consider.

3. Does the User score have a bigger impact on sales compared to the Critic score? For example, does a game that has a critic score of 70 and a user score of 90 sell less than a game that has the inverse. This question could be used to explore the importance of public opinion verses the opinion of the video game critics.

4. Does the rating of the game have an impact on sales based on the region? For example, are games that are rated mature sell better in Japan compared to America?
   This question would help rate the importance of the maturity of a game when creating an algorithm to determine the sales of a game before it releases.

5. Can we predict a games success (in terms of sales) based on the given factors, Region, Genre, Platform, Critic_Score, and the User_Score.
   This data science question is important because it encapsulates the entire goal of this research. Which is to create a predictive model that could be used to find what the next trend in popular video games is in order to capitalize on the market.

One social implication that could arise from this type of research is that if trends are clear as to what games sell the best then games could start to become less original and more homogeneous. For example, based on the data every company should make an action game that is targeted to a North American audience. Then it could lead to a world where every game is relatively similar and lead to less creativity in the medium.

One important note to bring up is that this affect has already been seen in the mid 2000s to early 2010's with the rise of the generic military shooter (Ex. Call of Duty). However in the past couple of years the market has started to correct itself with more experiential games being released that are still in similar genres (Ex. Destiny).