Daniel Saul
March 14, 2021

I. Insights and Future Analysis

Business Insights

The focal point of my analysis revolves around which genres of games from different platforms receive the best ratings. I was able to achieve the goal of creating influential visuals to help explain my recommendations. First, the platforms that generate the most sales are Xbox, PlayStation, and Nintendo, with the majority of the sales coming from North America and Europe (see Figure 1). I recommend that GameStop focuses on selling these platforms and games, compatible with each platform, in North America and Europe with the items as the center of attention in stores. This would help boost sales for the company. Second, the top three video game genres with the best ratings are role-playing, strategy, and sports (see Figure 2). I believe GameStop should include games that adhere to these genres in their commercials and advertisements. Since the genres have the highest ratings, we can assume customers and users play or enjoy these kinds of video games the most. This would help GameStop draw attention to the most popular and highest-selling games in their stores, advancing their business goals. Third, GameStop should put their attention towards partnering with the best video game publishers such as Nintendo, Sony, Activision, and Electronic Arts (see Figure 3). The publishers mentioned put out some of the highest rated games such as Sony's "God of War" and Nintendo's "Super Mario Odyssey" (see Figure 4). In my opinion, GameStop would not have success without strategically targeting popular games from high-rated publishers. Selling well-known games would benefit both GameStop and the publisher's popularity in the gaming industry.

What's next?

After completing this case study, I have come to the realization that there are many more recommendations and analysis I could perform for GameStop. If more time was available, GameStop should shift their focus from purely ratings to understanding and counting new variables such as total active users per region, count of video games purchased per region, and a grand total of users from when the game was released to present. Figuring out how many players use these games could be quite valuable for future insights. Furthermore, subcategorizing players into a hierarchy of age groups, genders, and other demographic groups would benefit videogame analysis. Lastly, GameStop should look into analyzing video games that are popularly streamed on platforms such as Twitch, YouTube, and Mixer. This could provide insights to the younger generation of gamers utilizing streaming services more and gaming advancements in the industry.

II. Reproducibility and Documentation

This report on video game analysis was created with Microsoft Word and converted to PDF format for submission. The R programming language, along with the RStudio software, were utilized for performing web scraping of data from the Open Critic website, data wrangling, and some aspects of data transformation. However, RMarkdown, a file formatter, was not used for the generation of this report. I used the Google Chrome web browser to access the websites for scraping. Along with Google Chrome, I employed SelectorGadget, a Chrome extension, to select HTML nodes to target valuable variables for data scraping. The reproducible R code used is available in Appendix A. The Tableau software was used for data visualization of the key insights of video game data. The Tableau documentation and tools used are available in

Appendix B. The R code applied in this report was supported by a web scraping lecture from MIST 4630 with Dr. Salge.

III. Web Scraping and Data Import

Web Scraping

I successfully web scraped the data from the Open Critic websites provided for me—utilizing RStudio, R code, and SelectorGadget. First, I used the SelectorGadget to pick the information and variables I wanted to extract from the website. I then used each website link provided by my supervisor to manipulate my code and download the data into tibbles in RStudio. The tibbles were then convert to CSV files and exported out of RStudio to my computer. The overall data included 6 variables and 6351 observations, with the variables and descriptions included below. In addition, the Open Critic dataset consists of many null or N/A values in the critic score variable. I assume the reasoning for this void of data is because certain games were not scored by critics yet. The release year and website source were manually entered in with R code because scraping these variables off of the website were not possible. If I were to web scrape this type of data again, I recommend collecting more variable types such as active players, total number of game users, genre, sales regions, sales, and other variables similar to the dataset from Metacritic—making analysis easier and more thorough. The data scraped off of Open Critic can be found in the file submitted along with this report called "Open_Critic_Data.csv." For more detailed information on the working R code, see Appendix A.1.

Variable	Description
page_no	Page number of the website where the observation was found
Name	Name of the video game
Platform	Platform(s) that the video game is played on
Year_of_Release	Year that the video game was released to the public
Critic_Score	Score given on a video game by the Open Critic staff out of 100
Source	Place where the observation was pulled from

Data Import

I successfully imported the data provided for me from Metacritic and Vgchartz. The file imported is called "videogame.csv." The overall data included 16 variables and 16719 observations, with the variables and descriptions included below. There are many null or N/A values in the critic score, critic count, user score, user count, developer, and rating. This discrepancy is due to the fact that some observations are spread over multiple rows and the data fails to be repeated. Also, some games may not have been scored yet by users and critics, or they do not have an ESRB rating.

Variable	Description
Name	Name of the video game
Platform	Platform that the video game is played on
Year_of_Release	Year that the video game was released to the public
Genre	Type of video game
Publisher	Party responsible for publishing and selling the game
NA_Sales	North America game sales in millions of USD
EU_Sales	Europe game sales in millions of USD
JP_Sales	Japan game sales in millions of USD

Other_Sales	Other region game sales in millions of USD
Global_Sales	Total sales summed from the previous 4 sales variables in millions of USD
Critic_Score	Aggregate score compiled by Metacritic staff out of 100
Critic_Count	Number of critics used in coming up with Critic_Score
User_Score	Score by Metacritic's subscribers out of 10
User_Count	Number of users who gave the User_Score
Developer	Party responsible for creating the game
Rating	ESRB ratings (everyone, teen, adults only, etc)

IV. Wrangling and Transformation

Wrangling

The data from both Open Critic and Metacritic/Vgchartz were untidy in their own ways. The Open Critic data extracted via web scraping was untidy because each cell in the platform variable was not a single measurement. For example, each game had many platforms labeled as such "XB1, XBXS, PS3" in just one column. I should have separated the cells into new variables based on the different platforms; however, I could not find an effective way to do so. An insight that came to mind while performing these actions was the realization of how many ways there are to play video games. Video games can be played and compatible with countless platforms. When scraping the data, I forged my code to assign the data taken into four different tibbles for the four release years scraped (see Appendix A.1). This is also where I added the year of release to the tibbles, since the dates on the website were not in a useful format. I then performed three full joins on the tables scraped to create a final tibble. The working code for the joins can be found in Appendix A.1.

The Metacritic/Vgcharts data provided in "videoGame.csv" was also untidy because there were multiple observations spread across many rows. This is a result of video games being compatible on different platforms, generating different amounts of sales on each platform. A typical analyst would spread the table to make it narrower and longer to account for the multiple observations across many rows, but I believed that I must leave this data the way it is to perform my visual analysis in Tableau. This was necessary to be able to analyze sales for each platform.

Transformation

Both datasets stated above required transformations. The Open Critic tibble needed a new variable called "Source," instructed by my supervisor. I did this by mutating the tibble with R code to create the new variable to be put last in the order of the variables (see Appendix A.3). Once this was done, the final tibble was transformed into a CSV file with RStudio which the code for can also be found in Appendix A.3. The rest of the data transformation performed was in Tableau and can be found in Appendix B. In Tableau, variables were renamed in the worksheets, preventing the generated variable names from being an eye sore in the visuals and were easier to understand. I also used grouped summaries such as averages of critic scores and totals of sales to enhance my analysis on the video game data. Also, variables and data were filtered to sharpen the breadth of data in my hands which was used for visuals (see Appendix B.2, B.3, B.4, B.5).

V. Visualizations

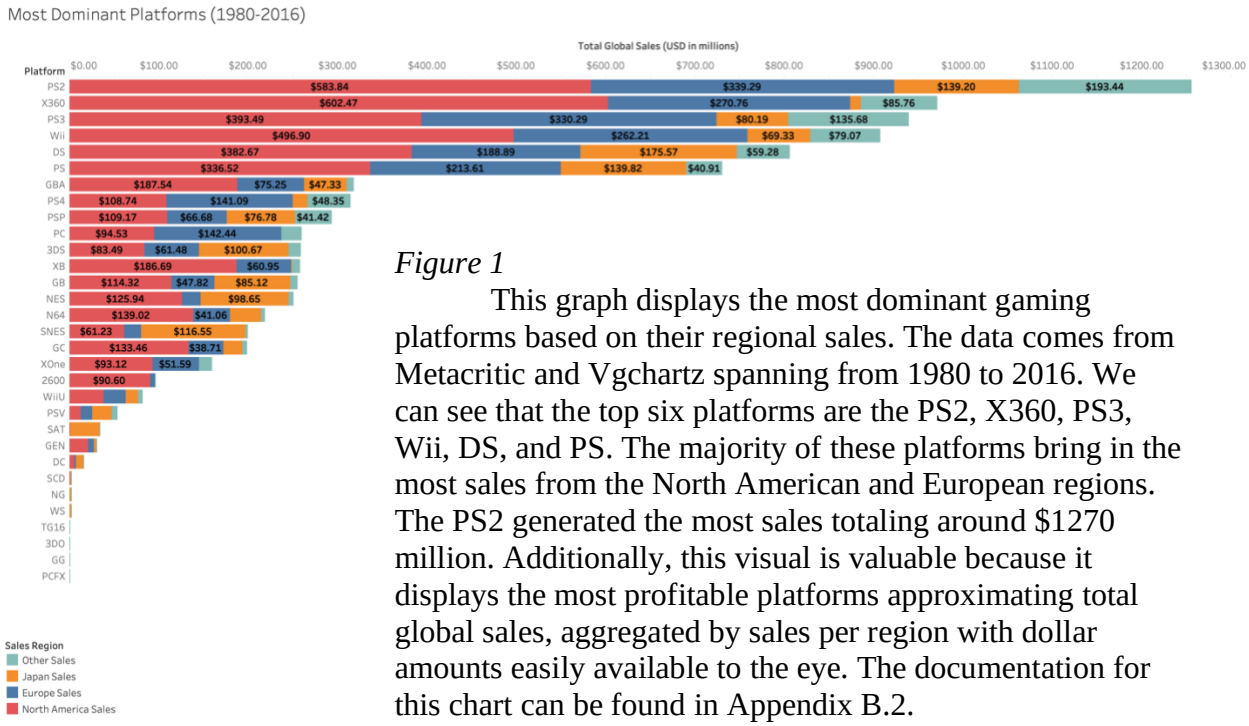
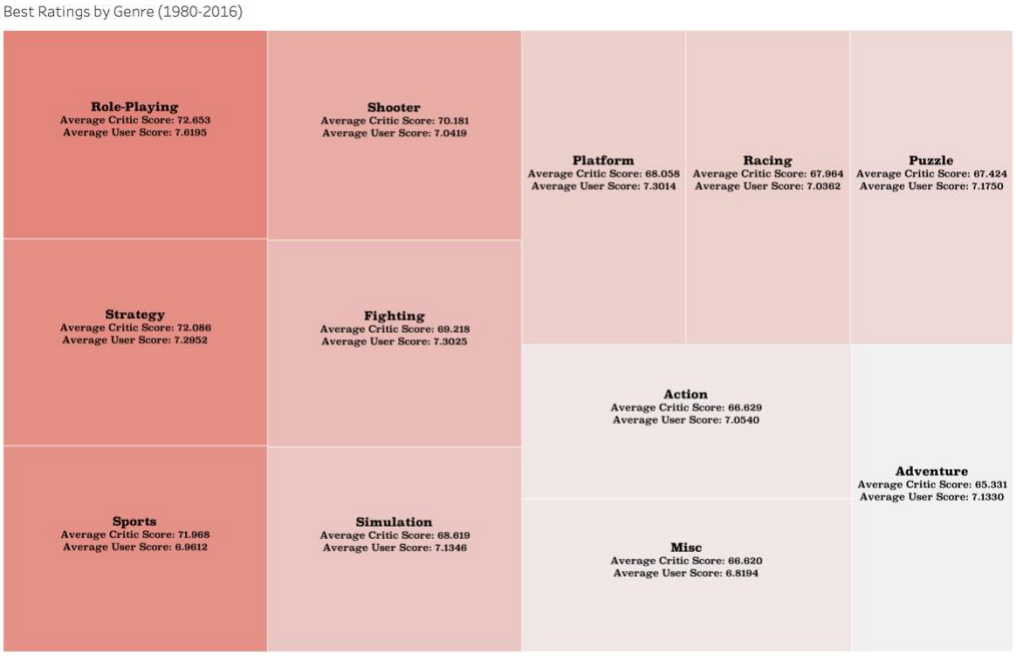


Figure 1

This graph displays the most dominant gaming platforms based on their regional sales. The data comes from Metacritic and Vgchartz spanning from 1980 to 2016. We can see that the top six platforms are the PS2, X360, PS3, Wii, DS, and PS. The majority of these platforms bring in the most sales from the North American and European regions. The PS2 generated the most sales totaling around $1270 million. Additionally, this visual is valuable because it displays the most profitable platforms approximating total global sales, aggregated by sales per region with dollar amounts easily available to the eye. The documentation for this chart can be found in Appendix B.2.

Figure 2

This visual shows the video game genres with the highest average critic and user score from Metacritic spanning from 1980 to 2016. The data is summarized by average score for each genre because summing the scores would not be useful, and it is filtered by the size of the critic score due to game critics having more experience than users. We can see that role-playing, strategy, and sport video games carry the best ratings. Role-playing holds the highest average with a 72.653 critic score. The documentation for this visual can be found in Appendix B.3.



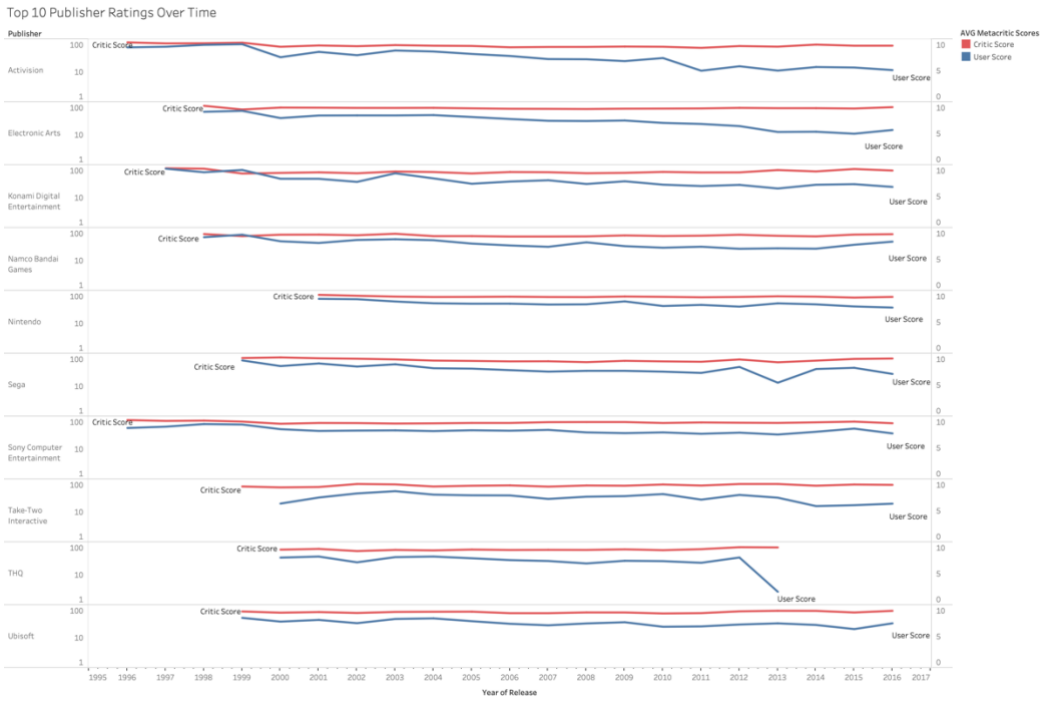
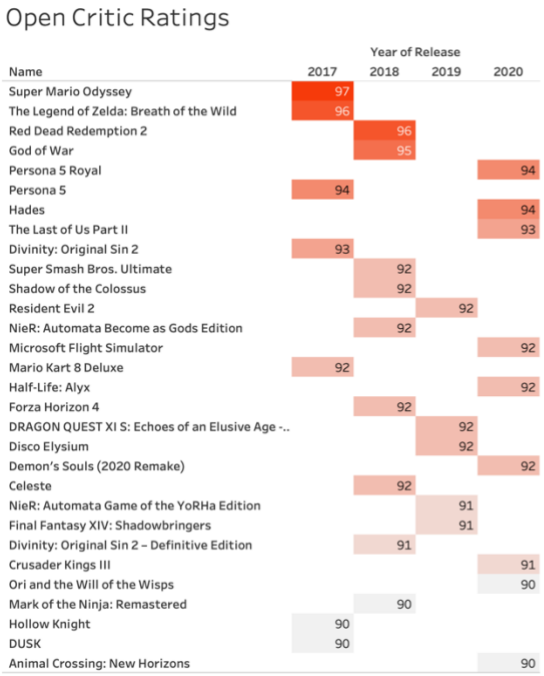


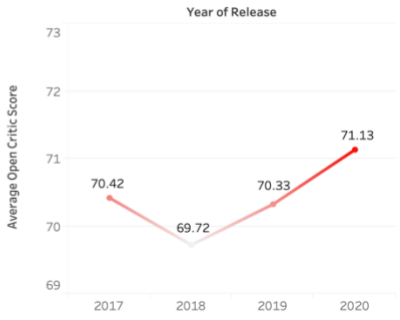
Figure 3

This graph displays the changes in average user and critic score from Metacritic for the top ten highest rated video game publishers over time. Once again, the data is summarized by average score for each genre because summing the scores would be of little help with understanding ratings. We can see that two of the highest rated platforms, Activision and Electronic Arts, have either increased or kept their critic scores constant, while user score has slightly declined over time. This graph is effective because it allows us to understand which game publishers have distributed the best games over time. For instance, Electronic Arts develops mostly sports games such as "FIFA," one of the highest rated genres in the previous visual (see Figure 2). The documentation for this visual can be found in Appendix B.4.

Figure 4

This optic combines two types of graphs based off of Open Critic ratings: one shows the top thirty highest-rated video games from the years 2017 to 2020, and the second displays the change in average critic score from 2017 to 2020. We can see that "Super Mario Odyssey" and "The Legend of Zelda: Breath of the Wild" are the two highest rated games in the past four years. This visual is important because it allows us to see types of games that have high ratings in the past when critic score is slowly increasing. We can use this data to help predict the rating and popularity of future video game releases. The documentation for this visual can be found in Appendix B.5.





OpenCritic

Appendix A
R Code

A.1 *Web Scraping Code*

The source of code comes from a web scraping lecture from MIST 5620. Each year (2017, 2018, 2019, 2020), from the data that was web scraped, used the same code format with the only difference being the website links and code where the year need to be inserted. The code within the "sprintf()" function is a wrapper allowing us to combine text with the variable values. The "map()" function loops through the links of each page and the "read_html()" function reads each link. The next part of code retrieves the data for Name, Platform, Year_of_Release, and Critic_Score by using the SelectorGadget Google Chrome Extension to select HTML nodes of the data we needed. The "tibble()" function creates a table with the data scraped, and the "bind_rows()" function merges the data collected with the page ID. In addition, it creates the new variable page_no. Directly below and to the right are the R packages utilized to run this code.

```
library(rvest)
library(tidyverse)
library(xml2)
library(tibble)
library(dplyr)
```

```
#OpenCritic 2017
opencritic2017 <- sprintf("https://opencritic.com/browse/all/2017/date?page=%d",
1:84)

oc2017 <- map(opencritic2017, ~ {

  doc <- read_html(.x)

  name <-
    doc %>%
    html_nodes(".col a") %>%
    html_text()

  platform <-
    doc %>%
    html_nodes(".platforms") %>%
    html_text()

  year_of_release <- "2017"

  critic_score <-
    doc %>%
    html_nodes(".score") %>%
    html_text()

  tibble(Name = name,
         Platform = platform,
         Year_of_Release = year_of_release,
         Critic_Score = critic_score)

}
) %>%

  bind_rows(.id = 'page_no')
```

```r
#OpenCritic 2018
opencritic2018 <- sprintf("https://opencritic.com/browse/all/2018/date?page=%d", 1:84)

oc2018 <- map(opencritic2018, ~ {

  doc <- read_html(.x)

  name <-
    doc %>%
    html_nodes(".col a") %>%
    html_text()

  platform <-
    doc %>%
    html_nodes(".platforms") %>%
    html_text()

  year_of_release <- "2018"

  critic_score <-
    doc %>%
    html_nodes(".score") %>%
    html_text()

  tibble(Name = name,
         Platform = platform,
         Year_of_Release = year_of_release,
         Critic_Score = critic_score)

}
) %>%

  bind_rows(.id = 'page_no')
```

```r
#OpenCritic 2019
opencritic2019 <- sprintf("https://opencritic.com/browse/all/2019/date?page=%d", 1:76)

oc2019 <- map(opencritic2019, ~ {

  doc <- read_html(.x)

  name <-
    doc %>%
    html_nodes(".col a") %>%
    html_text()

  platform <-
    doc %>%
    html_nodes(".platforms") %>%
    html_text()

  year_of_release <- "2019"
```

```
  critic_score <-
    doc %>%
    html_nodes(".score") %>%
    html_text()

  tibble(Name = name,
         Platform = platform,
         Year_of_Release = year_of_release,
         Critic_Score = critic_score)

}
) %>%

  bind_rows(.id = 'page_no')
```

```
#OpenCritic 2020
opencritic2020 <- sprintf("https://opencritic.com/browse/all/2020/date?page=%d", 1:75)

oc2020 <- map(opencritic2020, ~ {

  doc <- read_html(.x)

  name <-
    doc %>%
    html_nodes(".col a") %>%
    html_text()

  platform <-
    doc %>%
    html_nodes(".platforms") %>%
    html_text()

  year_of_release <- "2020"

  critic_score <-
    doc %>%
    html_nodes(".score") %>%
    html_text()

  tibble(Name = name,
         Platform = platform,
         Year_of_Release = year_of_release,
         Critic_Score = critic_score)

}
) %>%

  bind_rows(.id = 'page_no')
```

A.2 *Wrangling*

After scraping the data from the website into four separate tibbles, the data needed to be wrangled. I performed three full joins using the "full_join()" function to combine all rows and all columns from each tibble. Each tibble was joined with the previous until a final, larger tibble was processed and completed.

```
open_critic <- full_join(oc2017, oc2018)

## Joining, by = c("page_no", "Name", "Platform", "Year_of_Release", "Critic_Score")

open_critic1 <- full_join(open_critic, oc2019)

## Joining, by = c("page_no", "Name", "Platform", "Year_of_Release", "Critic_Score")

open_critic2 <- full_join(open_critic1, oc2020)

## Joining, by = c("page_no", "Name", "Platform", "Year_of_Release", "Critic_Score")
```

A.3 *Transformation*

The final tibble required a new variable to display where the data came from. This new variable was called "Source" and repeated "Open Critic" in each row, because that is where the web-scraped data originated from. To do this, I used the "mutate()" function to add a new column to the tibble. The data was then ready for analysis, therefore I used the "write_csv()" function to create a text file of the data to be stored as a csv file on my computer.

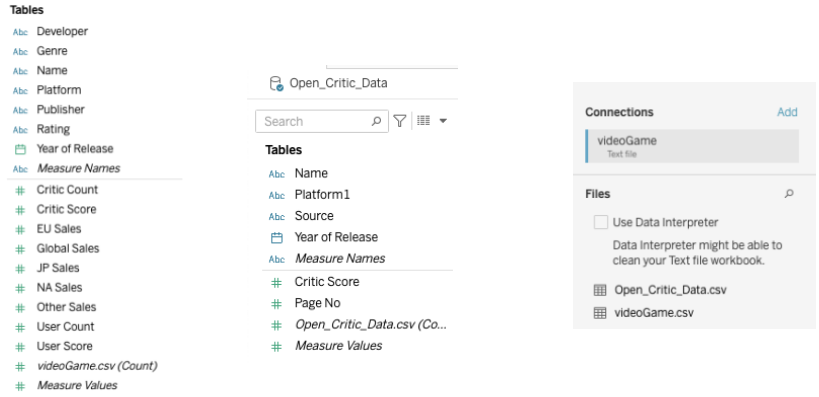
```
open_critic3 <- open_critic2 %>%
                  mutate(., Source = "Open Critic")

write_csv(open_critic3, "Open_Critic_Data.csv")
```

Appendix B
Tableau Visualization

Tableau was utilized for data visualization and certain aspects of transformation. In the following sections, the pictures and text explain how the visuals were created for analysis. It is important to mention that "Open_Critic_Data.csv" uses ratings from Open Critic while "videoGame.csv" uses ratings from Metacritic. Both have similar variable names, so please keep the data source in mind. Pictures of the Tableau tools are included to enhance understanding of figure creation.

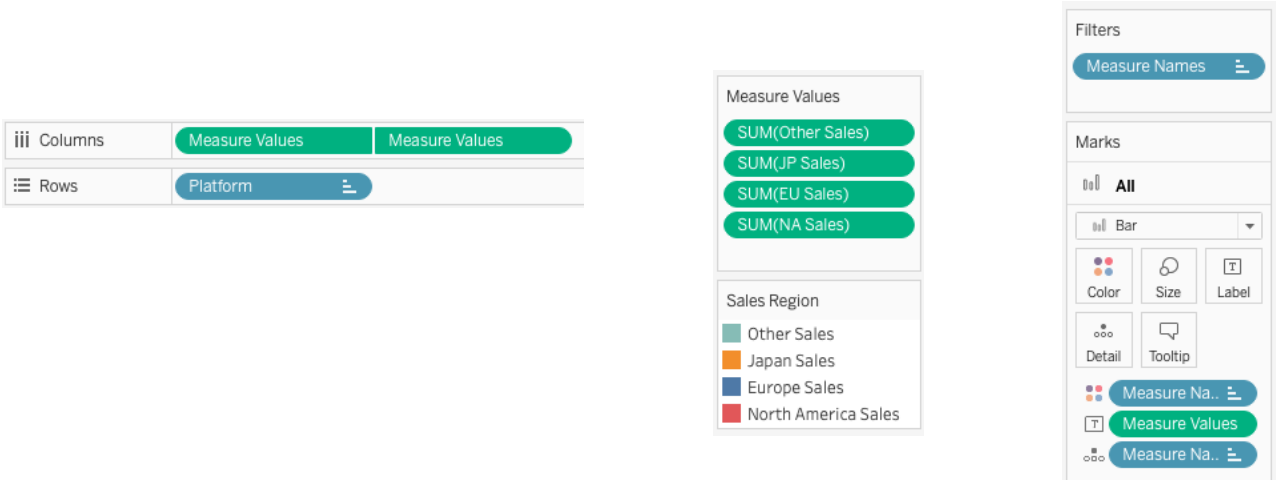
B.1 *Data Connection*

Data connection is the first step to using Tableau. Here, you can see where the two text files "Open_Critic_Data.csv" and "videoGame.csv" data are imported and connected to the software. On the left are the measures and dimensions for "videoGame.csv". In the middle are the measures and dimensions of "Open_Critic_Data.csv."



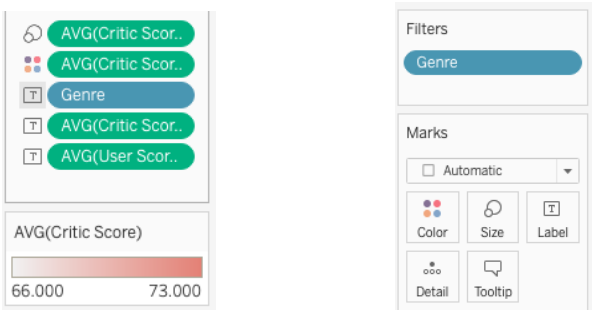
B.2 *Figure 1*

Figure 1 is a horizontal and stacked bar chart. It uses the "videoGame.csv" dataset. The "Platform" variable is on the y-axis and "Global_Sales" variable in millions of USD on the x-axis. The bars of the chart include the label of regional sales for each region. In addition, they are colored (grey, orange, blue, red) by regions. The sales in each region are summed and displayed for each platform. "Other_Sales," "JP_Sales," "EU_Sales," and "NA_Sales" are the regions that were summed to make up global sales. The detail of the chart is made up of the region variables above.



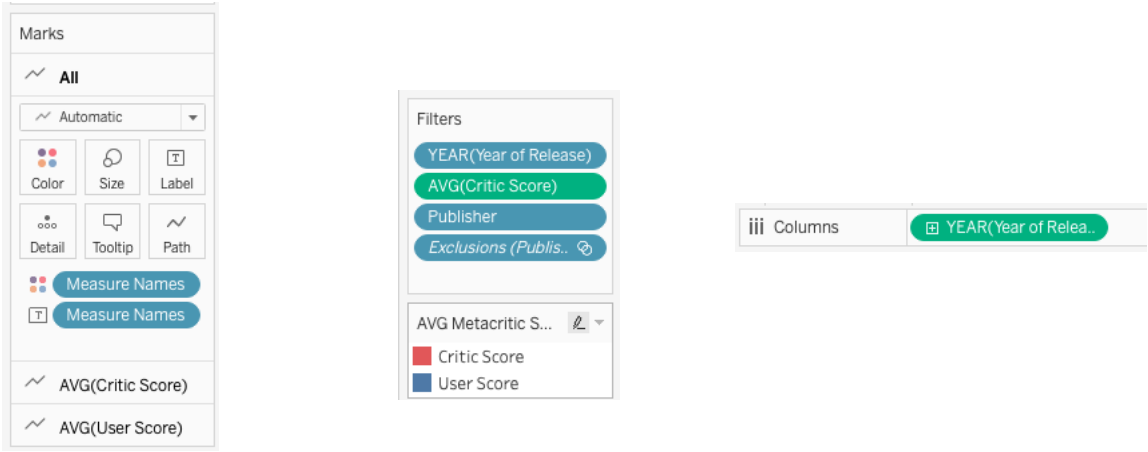
B.3 *Figure 2*

Figure 2 is a treemap. It uses the "videoGame.csv" dataset. The variables used in this visual are "Genre," "Critic_Score," and "User_Score." The branches of the tree are colored (red) and sized by the average "Critic_Score." This variable is deemed more useful than "User_Score" because game critics typically have more experience with a variety of games, making their opinion superior. The labels consist of all three variables mentioned above.



B.4 *Figure 3*

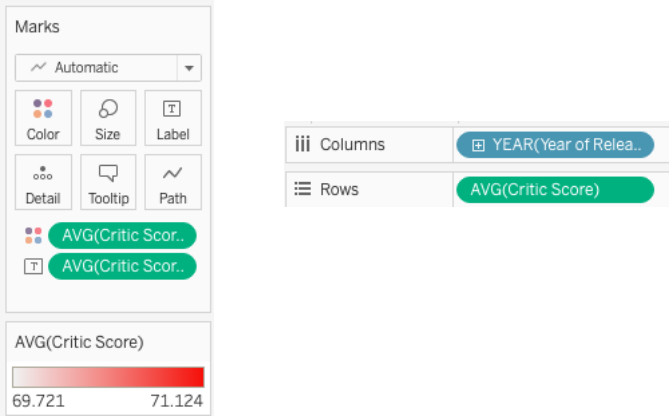
Figure 3 is a dual-axis line graph. It uses the "videoGame.csv" dataset. The variables used are "Publisher," "Year_of_Release," "Critic_Score," and "User_Score." The x-axis is the years of releases (1995-2017) and the y-axis are the top 10 publishers. Null values are filtered out. The top 10 publishers with games who earned the highest ratings are also filtered. Critic score and user score are averaged and displayed over time for each publisher displayed. The left side of the chart includes the publisher and axis of the critic score while the right side displays just the axis of the user score. The axis are logarithmic to show more detail of the rating differences. The lines are colored by critic score in red and the user score in blue.



B.5 *Figure 4*

Figure 4 is a dashboard comprised of the Open Critic logo and two worksheets: a line chart and a highlight table. The logo was added for the aesthetic. It uses the "Open_Critic_Data.csv" dataset. The line chart uses the variables "Year_of_Release" and "Critic_Score." The x-axis is the years of release (2017-2020) and the y-axis is the average Open Critic score on a logarithmic scale to show differences in averages by year. The average critic score was calculated to be used in the graph. It is colored by the average critic score. The highlight table uses the variables "Critic_Score," "Year_of_Release," and "Name." The names of games are on the y-axis and the year of release is on the x-axis. The names of the games were filtered to show the top 30 games with the highest critic score in the past 4 years. It is colored (red) by critic score. The darker the red shading, the higher the rating the game had.

Line Chart



Highlight Table