Video Game Industry Analyzer

Kevin Farragher

Dataset and Problem Statement

- The dataset I chose for my project is a video game dataset from Kaggle, which describes information on individual video games (Name, Platform, Release Year, Genre, Publisher, Rating), their physical sales from Vgchartz (regionally and globally) and their scores from users and critics on Metacritic. The dataset covers video games released between 1980 and 2016.
- The problem I tried to solve was finding interesting statistics and relationships between the attributes of my dataset and visualizing them with various graphs. I was also trying to cluster (group) similar video games and see the differences between each cluster. Third, I tried to find a way to predict whether a video game's Global sales were Low or High. The problem I tried to solve was important because by finding interesting statistics and relationships between the attributes of my datasett and visualizing them with various graphs, the findings and results could be used to analyze the video game industry since its beginning.

Methods Explored

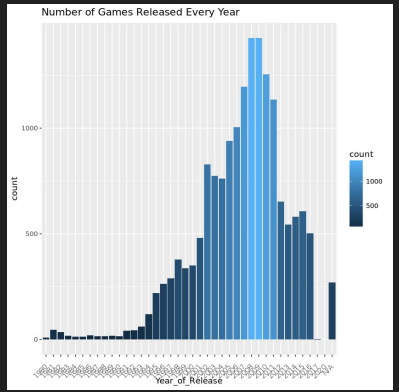
- The tools and methods I used for my project include descriptive statistics (dplyr), visualization (ggplot2), clustering (kmeans), and decision tree analysis (rpart, rpart.plot). Specifically, I used the dplyr, ggplot2 , kmeans, rpart, and rpart.plot libraries.
    - I used descriptive statistics (dplyr) to perform data wrangling and find interesting summary statistics between the attributes of my dataset, displaying them using tables.
    - I used visualization (ggplot2) to visualize the interesting summary statistics between the attributes of my dataset using graphs, plots, etc.
    - I used clustering (kmeans) to cluster (group) similar video games.
    - I used Decision Tree Analysis (rpart, rpart.plot) to predict whether a video game's Global sales is Low or High by building a classification model

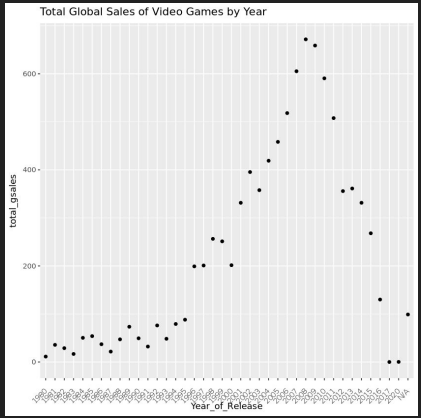
Data Preparation

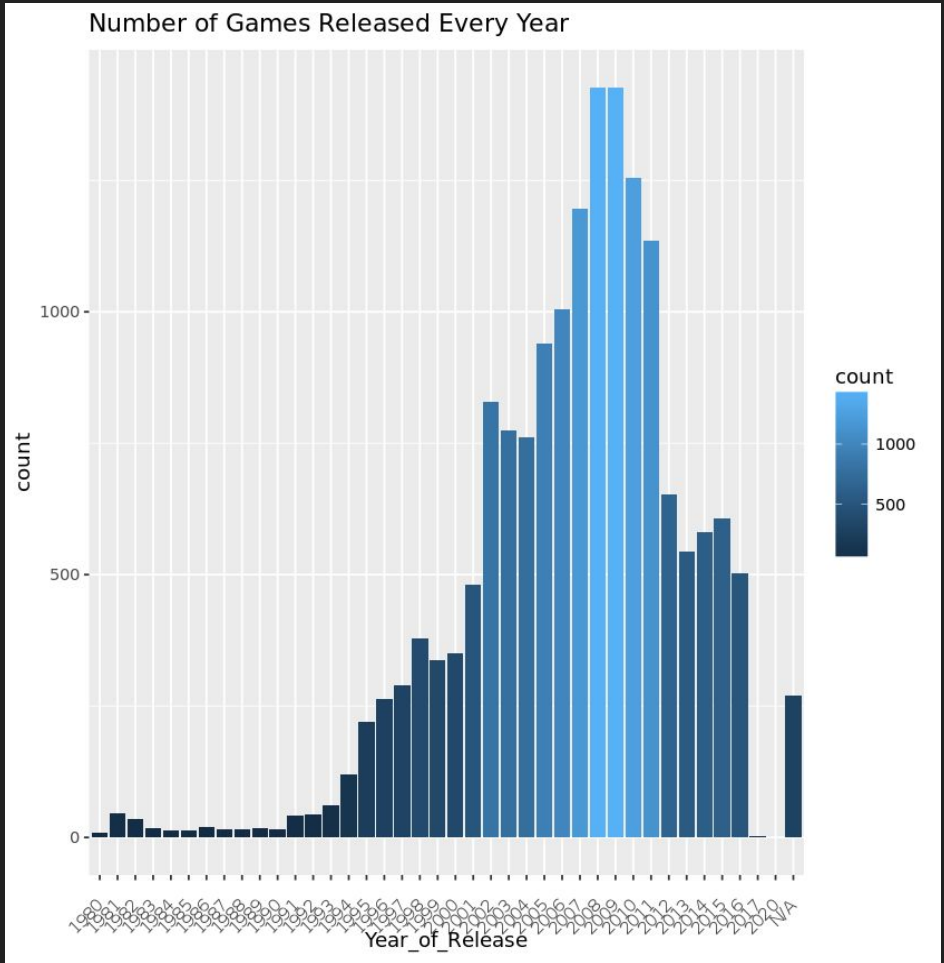
- Changing column data types - When I looked at the structure of my dataset initially, I found out that one of the columns for my dataset, User_Score, was a character field, which shouldn’t have been the case. I changed this User_Score column to a numeric field, an appropriate data type for the column so I could use it for my project.
- Cleaning Consideration - When I initially analyzed my dataset, I found there to be many missing values, especially in the Rating, Critic_Score, Critic_Count, User_Score, User_Count, Developer, and Rating columns of my dataset. There were many missing values in these columns because the data for these columns were brought from Metacritic. Metacritic doesn't cover all the video games and video game platforms used in this dataset, which led to there being many missing values in these columns. For the video games with missing values in these columns, I decided to keep and use them, as removing them would have left out a huge chunk of my dataset, leading to less meaningful and accurate results. I worked around the video games with missing values in the above columns when needed and as best as I can.

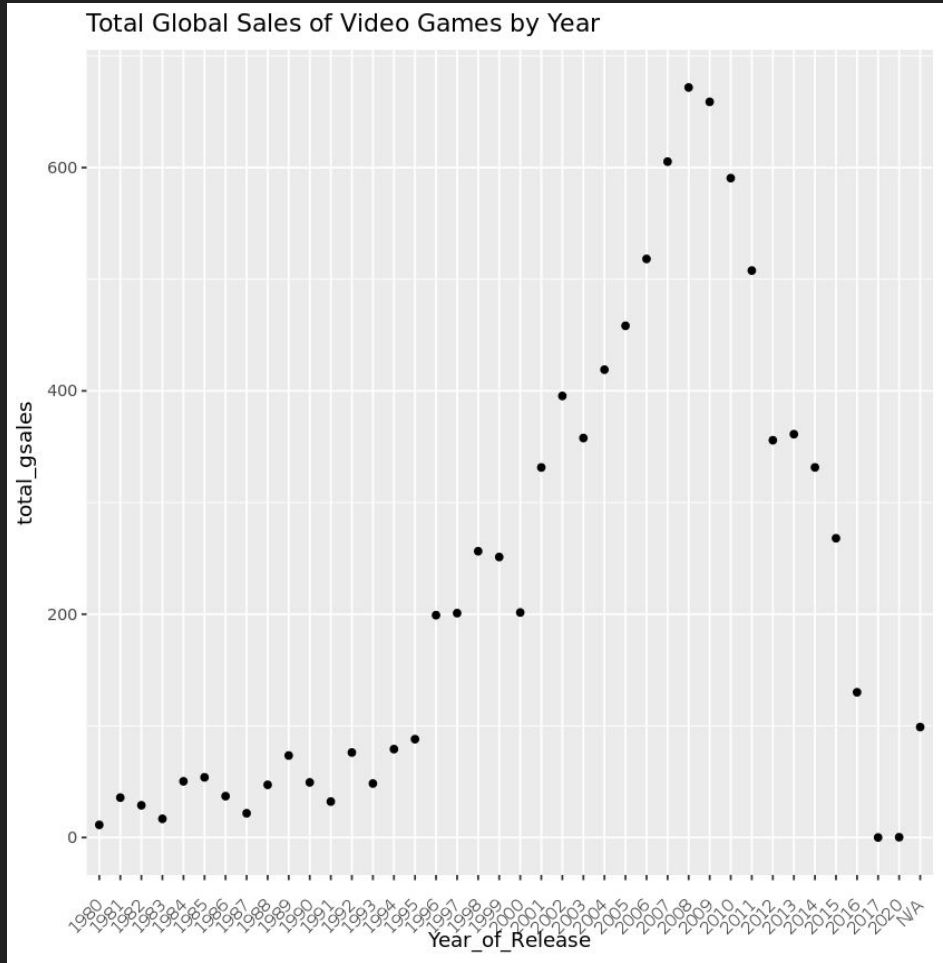
Visualization of Results/Summary Statistics

- Overall, since 1980, the number of video games released per year has increased. However, the number of video games released in a year reached its peak in 2008, and has started to decrease ever since. The graph visualizes the number of video games released each year.
- From 1980 to 2008, the total global sales of video games continued to increase and reached its peak in 2008. Since 2008, the total global sales of video games has been decreasing.The scatterplot visualizes the total global sales for video games in a given year.



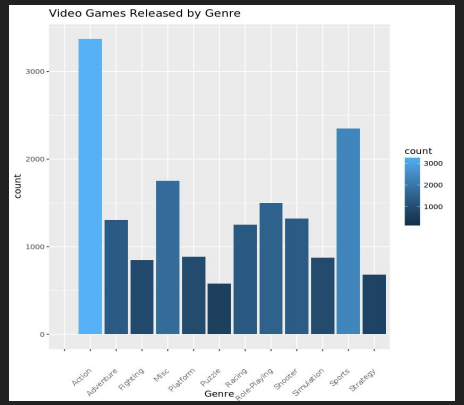


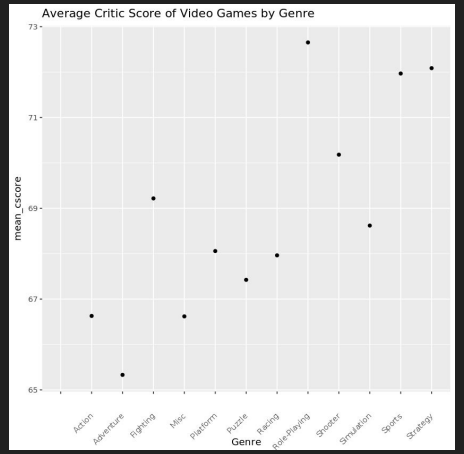

Number of Games Released Every Year
count
1000
500
0
count
1000
500
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2020
NA
Year_of_Release

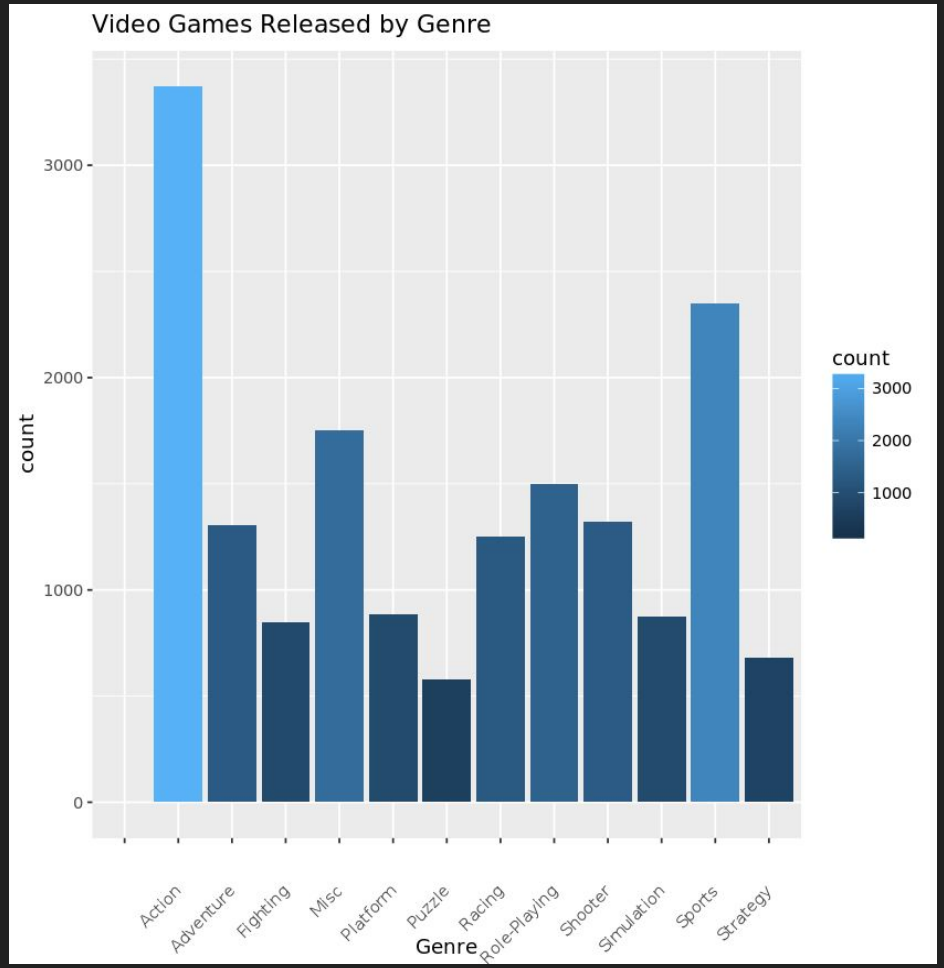

Total Global Sales of Video Games by Year
total_gsales
600
400
200
0
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2020
NA
Year_of_Release

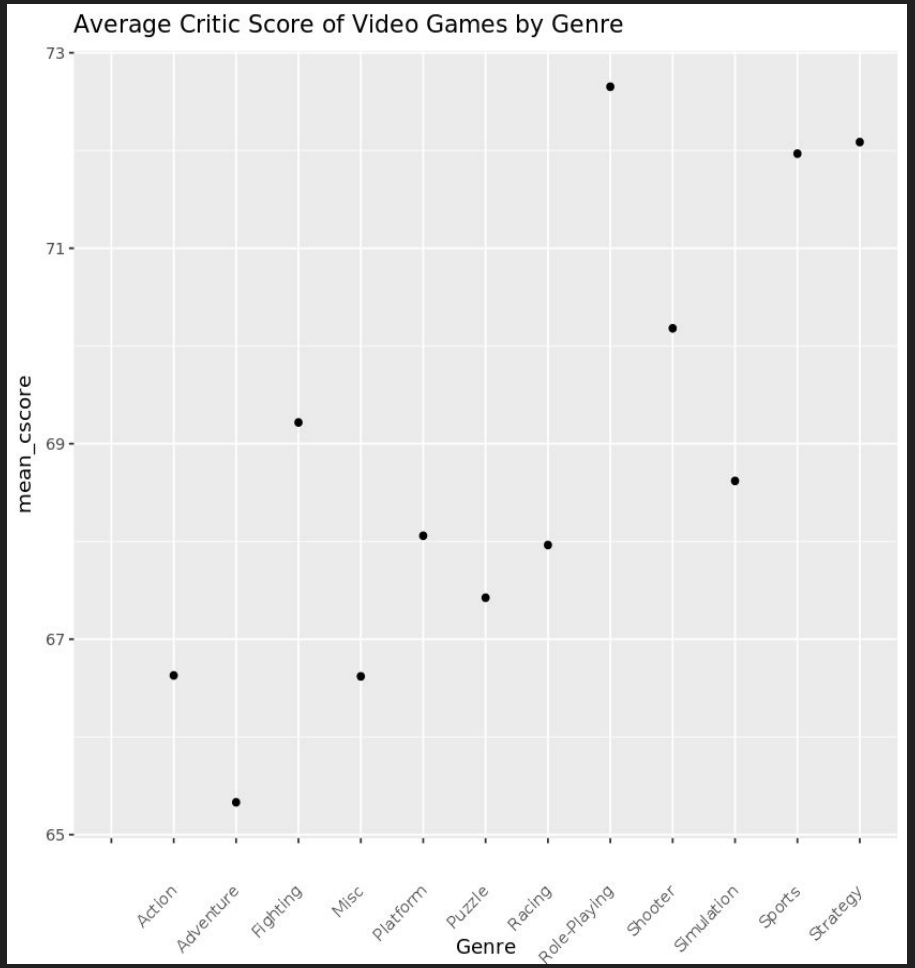
Visualization of Results/Summary Statistics cont.

- Overall, video games are most likely to be part of the action genre or sports genre, while least likely to be part of the puzzle or strategy genre. This implies that people tend to like playing action and sports games the most. The graph visualizes the number of video games that have been released per genre.
- Overall, role-playing and strategy games tend to get the highest scores from critics, while adventure and action games tend to get lowest scores from critics. Role-playing and fighting games to get the highest scores from users, while sports and racing games tend to get the lowest rating from users. Critics and users tend to rate role-playing games the highest, but overall tend to rate games of different genres differently. The scatterplot visualizes the average critic score for video games of each genre.



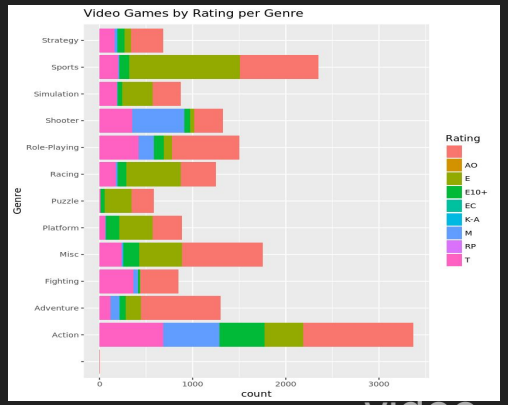



Video Games Released by Genre
count
3000
2000
1000
0
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy
Genre
count
3000
2000
1000

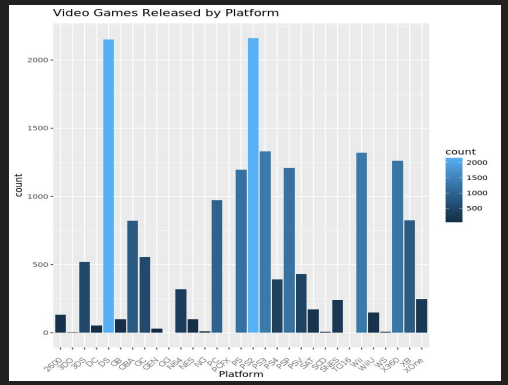

Average Critic Score of Video Games by Genre
mean_cscore
73
71
69
67
65
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy
Genre

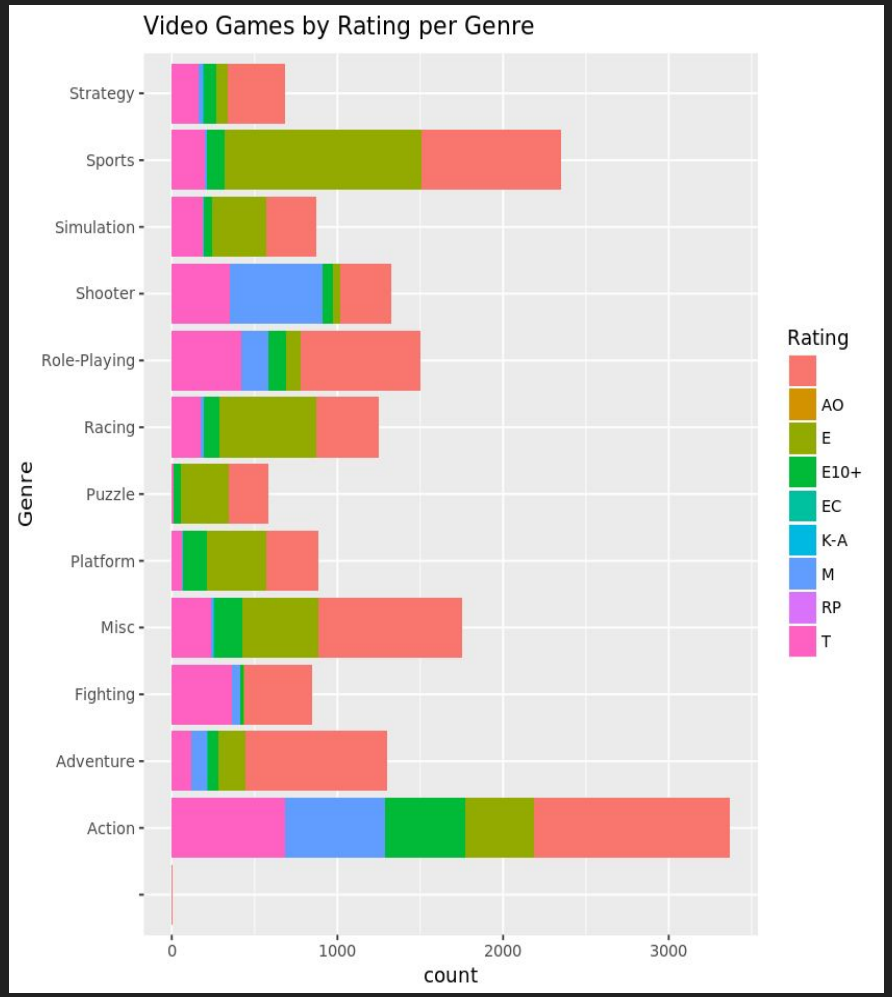
Visualization of Results/Summary Statistics cont.

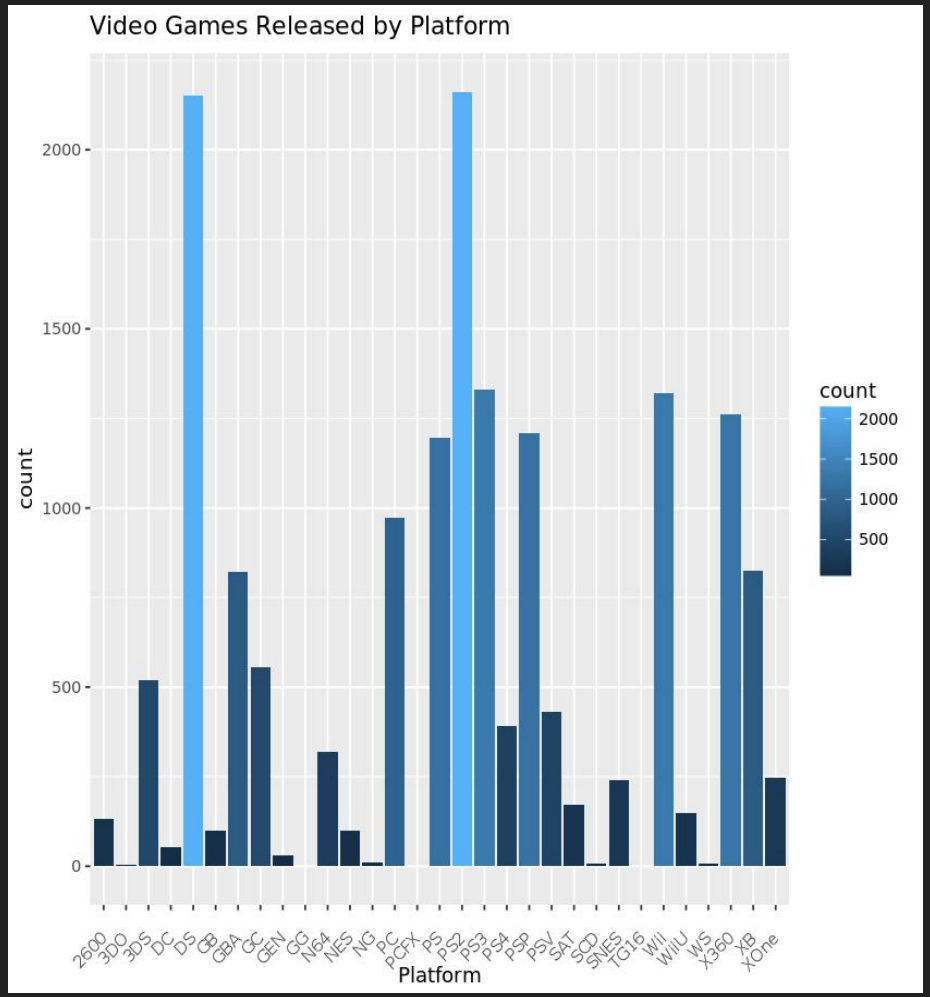
- Overall, Sports, Racing, and Puzzle games tend to be rated Eveyone (E), Fighting games tend to be rated Teen (T), and Shooter games tend to be rated Mature (M). The graph visualizes the proportion of video games by rating per genre.
- Overall, the PS2 and Nintendo DS (DS) have the most games released out of all the platforms. The graph visualizes the number of video games that have been released per platform.



video




Video Games by Rating per Genre
Strategy
Sports
Simulation
Shooter
Role-Playing
Racing
Puzzle
Platform
Misc
Fighting
Adventure
Action
Genre
0
1000
2000
3000
count
Rating
AO
E
E10+
EC
K-A
M
RP
T


Video Games Released by Platform
2000
1500
1000
500
0
count
2600
3DO
3DS
DC
DS
GB
GBA
GC
GEN
GG
N64
NES
NG
PC
PCFX
PS
PS2
PS3
PS4
PSP
PSV
SAT
SCD
SNES
TG16
Wii
WiiU
WS
X360
XB
XOne
Platform
count
2000
1500
1000
500

Other Results

- In 1989, video games had the highest average global sales. Overall, as the years have gone by, individual video games have tended to have less global sales.
- Overall, action games have sold the most globally, while strategy games have sold the least sales globally.
- In Japan, Shooter games tend to have the lowest mean sales, whereas in other regions, shooter games tend to get high mean sales. This shows how Japanese people tend to avoid buying shooter games, probably due to their violence, whereas in other regions, shooter games are very popular.
- Overall, video games released are most likely to have an E (Everyone)and T (Teen) rating, while video games released are least likely to have an AO (Adult Only) rating. Video games tend to have ratings that appeal to a greater audience (age group).
- Electronic Arts, Activision, Namco Bandai Games, Ubisoft, and Konami Digital Entertainment have published the most video games.
- Video games published by Nintendo have sold the most globally, in North America, in Europe, and in Japan, while video games published by Electronic Arts have sold the most in other regions.
- Ubisoft, EA Sports, EA Canada, Konami, and Capcom have developed the most video games.

Learning Algorithms

- Kmeans - I used the kmeans unsupervised learning algorithm to cluster (group) similar video games.
- Decision Tree Analysis - I used Decision Tree Analysis, a supervised learning machine, to predict whether a video game's Global sales is Low or High by building a classification model

Kmeans Clustering/Results

- I decided to cluster the video games using the global sales, NA sales, EU sales, JP sales, and Other Region attributes of my dataset, clustering the video games into four clusters
  - Cluster 1 contained 15,352 video games
  - Cluster 2 contained 17 video games
  - Cluster 3 contained 127 video games
  - Cluster 4 contained 1,223 video games
- Results
  - A video game's Global Sales seems to be the strongest and most important factor in determining which cluster it resides in.
  - Video games in cluster 2 tended to have the highest sales in all regions, while video games in cluster 1 tended to have the lowest sales in all regions.
  - Video games in cluster 4 tended to have the highest critic scores, while video games in cluster 1 tended to have the lowest critic scores.
  - Video games in cluster 2 tended to have the highest amount of critics review them, while video games in cluster 1 tended to have the lowest amount of critics who review them.

Kmeans Clustering/Results continued

- The video games with higher sales (Clusters 2, 3, and 4) tended to have the highest critic scores and highest amount of critics who review them, while the video games with lower sales (Cluster 1) tended to have the lowest critics scores and lowest amount of critics who review them.
    - However, the video games with the highest sales, in Cluster 2, didn’t tend to have the highest critic scores, showing that a video game's high sales doesn't always guarantee a high critic score for the game.
- Video games with the highest sales, in cluster 2, tended to have an Everyone (E) rating. and tended to be on Nintendo platforms.

Decision Tree Analysis/Results

- I used Decision Tree Analysis, a supervised learning machine, to predict whether a video game's Global sales is Low or High by building a classification model.I assigned each video game to having high or low global sales based on whether it had global sales above or equal to 1 (million). If a video game's global sales were less than 1 (million), it was assigned to have low global sales; if a video game's global sales were greater than or equal to 1 (million), it was assigned to have high global sales.
- I built a decision tree that predicts if a video game's global sales are high or low based on Developer and Publisher
- My classification model had an accuracy rate of 89.22% for predicting and classifying whether a video game’s global sales are high or low, being very accurate. This showed that a video game's publisher and developer seem to be important factors when determining whether the video game's global sales are high or low (above or below 1 million sales).

Shiny App

- Allows user to select a platform
  - User can look at a table with all the video games of that platform, and can filter the video games shown for that platform based on its global sales and its critic score
  - User can look at bar graphs for the platform showing the video games released per year for that platform, the video games released per genre for that platform, and the video games released per rating for that platform
  - User can look at boxplots for the platform showing the distribution of global, North America, Europe, Japan, or other Region sales for video games of that platform
- Allows user to select a publisher
  - User can look at a table with all the video games published by that publisher, and can filter the video games shown for that publisher based on its global sales and its critic score
  - User can look at bar graphs for the publisher showing the video games released per year for that publisher, the video games released per genre for that publisher, and the video games released per rating for that publisher
  - User can look at boxplots for the publisher showing the distribution of global, North America, Europe, Japan, or other Region sales for video games of that publisher

Shiny App cont.

- Allows user to select a developer
  - User can look at a table with all the video games developed by that developer, and can filter the video games shown for that developer based on its global sales and its critic score
  - User can look at bar graphs for the developer showing the video games released per year for that developer, the video games released per genre for that developer, and the video games released per rating for that developer
  - User can look at boxplots for the developer showing the distribution of global, North America, Europe, Japan, or other Region sales for video games of that developer
- Allows user to select a genre
  - User can look at a table with all the video games in that genre, and can filter the video games shown for that genre based on its global sales and its critic score
- Allows user to select a release year
  - User can look at a table with all the video games released that year, and can filter the video games shown for that year based on its global sales and its critic score
- Allows user to select a rating
  - User can look at a table with all the video games of that rating, and can filter the video games shown for that rating based on its global sales and its critic score
- Allows user to choose the color for their bar pgraphs and boxplots

Challenges

- The main challenge I faced while working on this project was having to deal with empty or NA values in the dataset. I had to constantly work my way around these values when analyzing the dataset

Lessons Learned/Plans to Mitigate Challenges

- Always find datasets that have no incomplete cases or a very small amount of them. My dataset had many incomplete cases, and it was hard to deal with and work around these incomplete cases throughout the project. Having and choosing a clean, complete dataset would’ve saved me a lot of time and made analysis and mining easier.
- Datasets with many incomplete cases should be avoided, as they make it tougher when mining and analyzing a dataset