Predicting Video Game Success

Drew Blik & Drake Watson

Research Questions:

1. What determines a video games' success?

   There were definitely some consistencies amongst the higher selling games, but there was also more diversity amidst certain characteristics that we did not expect. A majority of total games sales have the ESRB rating of E, the publisher popularity was dominated by Nintendo, but the genre popularity was quite diverse with an even distribution.

2. How do critic scores influence sales?

   Even though critic scores are a polarizing hot button issue in today's gaming community, we found almost no correlation at all between total sales and critic score even after removing the single most egregious outlier.

3. Do different countries have different tastes in video games?

   The European and North American sales were quite similar with sports, action, and shooter titles coming in as the top 3, but after those genres you can see the differences in taste as Europe consumes more racing titles whereas North America has a higher tendency of playing platformers compared to Europe. Japan's taste had stronger differences, with role-playing games coming in at the #1 spot and fighting games coming in at #6, both respective genres being much higher ranked in Japan's pie chart than Europe or North America.

4. Can we predict where a video game will be successful?

   Using a classifier machine learning model we were able to predict which country a video game will be most successful about 78% of the time given the features of genre, platform, developer, critic score, and ESRB rating. The attempt to create a regressor model that predicted sales was not successful.

Motivation and background:

The video game industry generates hundreds of millions of dollars a year and is growing exponentially as a cultural phenomenon. No longer are video games seen as obscure hobbies for nerds and outcasts, but rather as central forms of entertainment. With the growing popularity of streaming it seems to only be trending up, so it would be interesting to see what it is that makes a game popular (both from the consumer side and the sales side). Finding what characteristics are correlated most strongly with success would both give us insight into gamers psychology and might tip us off to what the next big successes might look like (and where they will be most successful).

Datasets:

https://www.kaggle.com/gregorut/videogamesales/code

https://www.kaggle.com/ashaheedq/video-games-sales-2019

Methodology (algorithm or analysis):

The very first task that was done was the joining of the two datasets. One dataset has ~16k observations through the year 2016 with a much more complete scrape including sales data, whereas the other dataset has ~30k observations through the year 2019 without complete sales data but it does include ESRB ratings and metacritic scores, which are features that are not included in the first data set.

After merging the datasets, we answered our challenge questions by filtering for the observations that provide the relevant data and generating visualizations and machine learning models.

- The first analysis question was to seek out the characteristics that determine a successful video game, which we did by isolating certain features (genre, esrb, and publisher) and then creating visualizations that displayed the distribution of these features when joined with global sales. The ESRB and genre comparisons were completed using pie charts due to the number of unique values, but the publisher comparison was done using a bar chart that isolated the top 10 due to the sheer amount of publishers in the data set (it was in the thousands).
- To analyze these features in respect to the regions in the data set we were able to isolate the sales for an individual region before creating similar visualizations to the first analysis question. The regional publisher comparison was a much more difficult task as it required some new pandas techniques but we were able to create a grouped bar chart that split the top 10 publishers into separate bars to display how they performed in each region.
- Using only the games that have both critic score and sales data we generated a scatter plot that allowed us to create a best fit line and analyze whether or not there is a correlation between those two things or not. We considered that the extreme outlier(s) may be influencing the correlation, but the sheer size of the data set seemed to make it immune to such an influence.
- After learning about correlations of various characteristics of video games, we attempted to predict a video game's lifetime success (sales) based on those features. Using a machine learning library, we trained a machine learning model to predict which region it would be most successful in based on attributes such as critic score, ESRB rating,

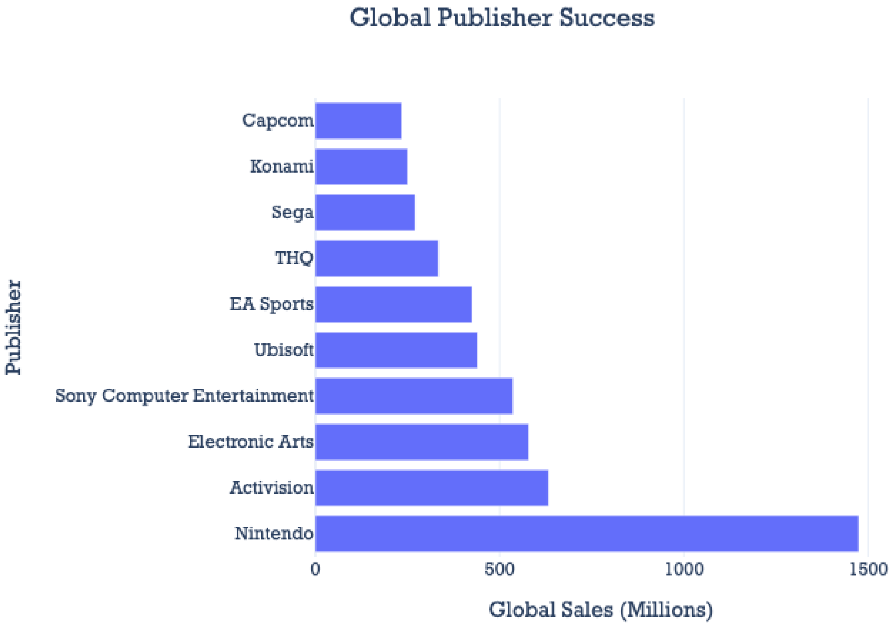
publisher and genre. After training the model, we tested the model on a random sample of the data to determine its accuracy.

- We also attempted to create a machine learning model that was able to estimate the total sales that a game generated, but we were unable to do so given the nuanced nature of sales (marketing, accessibility, release timing, etc.) and the sort of rudimentary information we had from the data set.

Results:

1. What determines a video games' success?

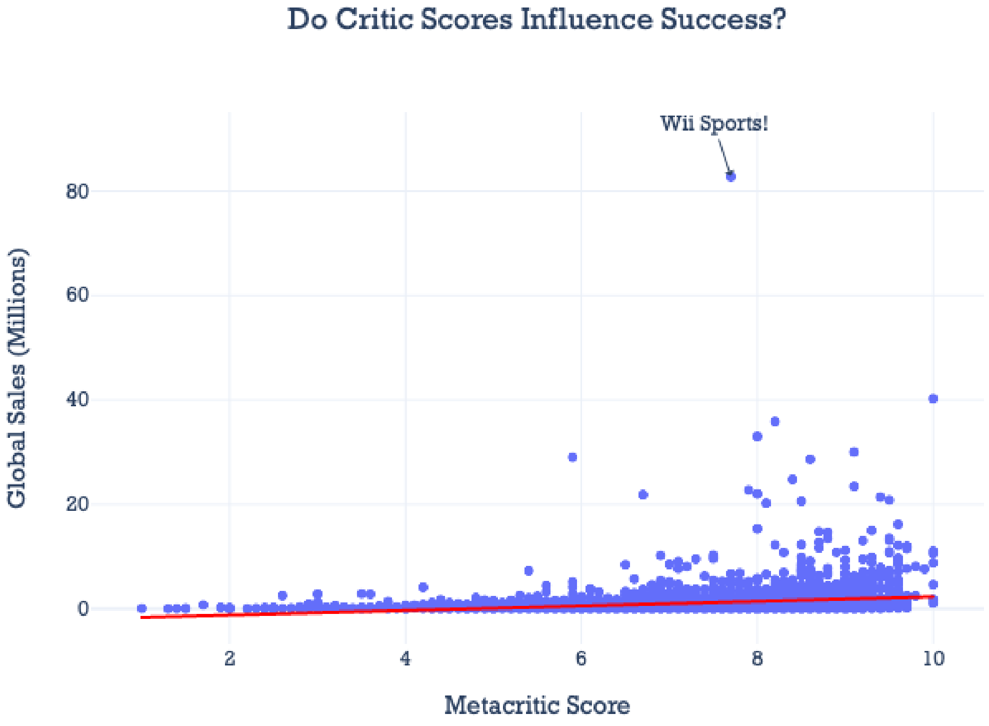
   We actually found that there is a very diverse set of characteristics that may determine whether or not a game is successful. Genre popularity is very evenly distributed (which can be seen in the following results section covering regional genre popularity), and the other two features that we investigated truly only had one dominant unique value (Nintendo as a publisher and the E for everyone ESRB rating). We actually believe that these two majority stakeholders may be related as every Nintendo game is rated E, but sports games are also making a large contribution to the E rating games sales. The publisher bar chart is only displaying the top 10 due to the thousands of publishers that were included in the data set, but the chart followed a very similar path for the first ~30 publishers until leveling out around the same level in the tens of millions range. We felt it was most interesting and digestible to focus on the top publishers in the world who generate the vast majority of sales in the gaming market.



2. How do critic scores influence sales?

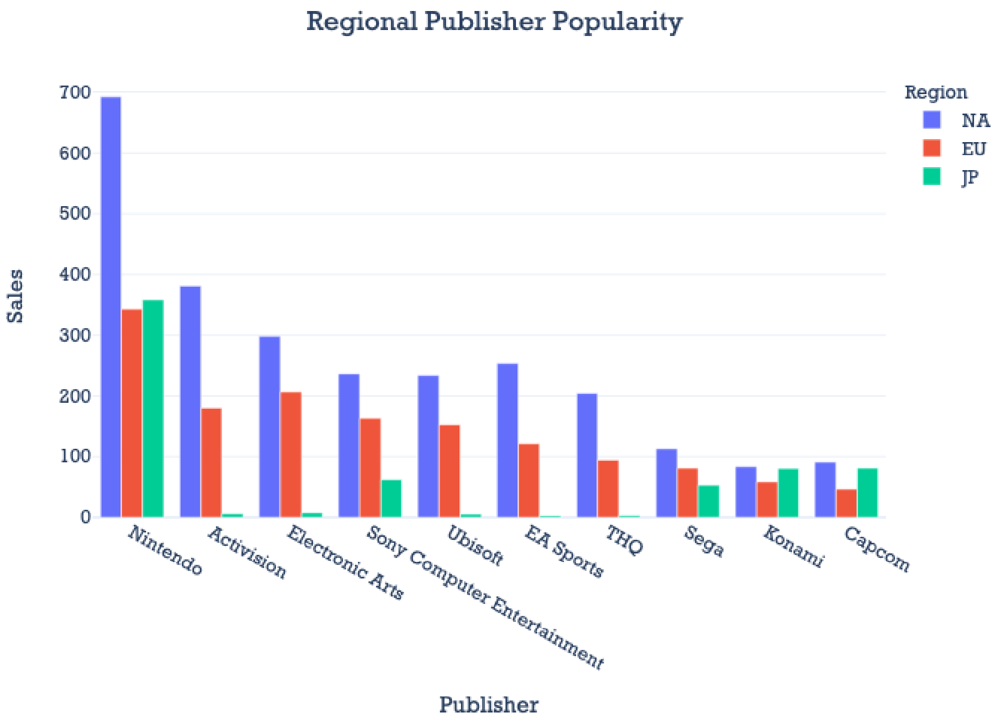
   Somewhat surprisingly we found almost no correlation between critic sales and success (measured in global sales). Why this is surprising is because there is a lot of focus on reviews and ratings in the gaming community. Lots of consumers complain every year about the quality of some of the most popular games (Call of Duty and Madden immediately come to mind) but in essence this shows that the negative public perception doesn't actually seem to result in a loss of profit. People will still buy Call of Duty every year even if they *think* it's a hot pile of garbage, and I believe this speaks to the fear of missing out and interactivity that the gaming community creates. If a game is being talked about, for better or worse, people want to check out what all the fuss is about.

   Something of note was that we did explore what the correlation was if we removed the Wii Sports outlier that is denoted in the visualization, but due to the mass amount of samples in the data it did not move the line at all, so we decided not to include that in the final report.

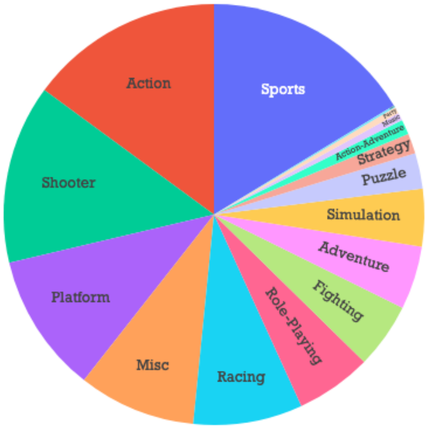


3. Do different countries have different tastes in video games?

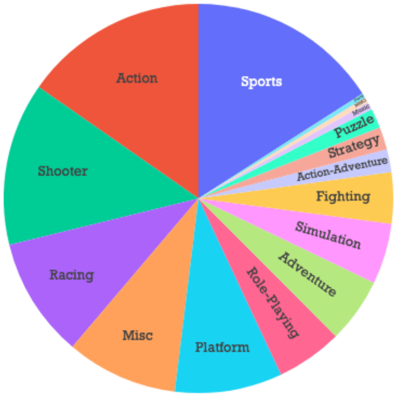
   The results we found in this section of the analysis were not surprising, as we uncovered some differences in preferred genres and some considerable variance in publisher sales. Nintendo was the dominant publisher for all three of these regions, but after that there is quite a bit of variance particularly with Japan. Japan is a considerably smaller market compared to Europe and North America, but it's interesting to note that Japan actually has a higher total sales for Nintendo than Europe. It is also notable that Japan has higher rates for all of the Japanese publishing companies (Nintendo, SCE, Sega, Konami, Capcom) and seems to have very little interest in many of the other top 10 publishers. Europe and North America seem to have a consistent proportional relationship with only some slight variance in their sales totals compared to one another. These differences in taste are a bit more pronounced in the following pie charts that display what genres each market favors (seen on the next page). North America and Europe have the same top 3 but then you can see the differences in preference as Europe is more inclined to play racing games, which meets our expectations since racing culture is much more prominent in Europe. It can be seen that there are much greater differences in the Japanese market as role-playing comes in at #1, platforming at #3, and fighting games come in at #6, which are all respectively higher than in Europe or North America. These preferences in tastes were nearly exactly what our expectations were going into the project and they mirror the commonly held public perception of cultural video game tastes.



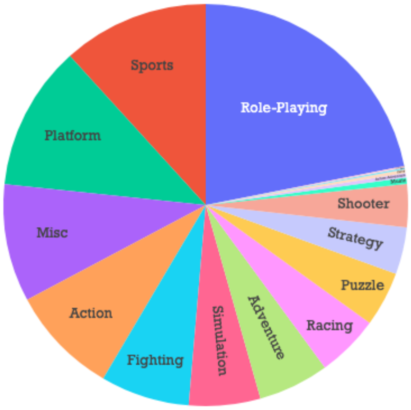
Genre Popularity in North America



Genre Popularity in Europe

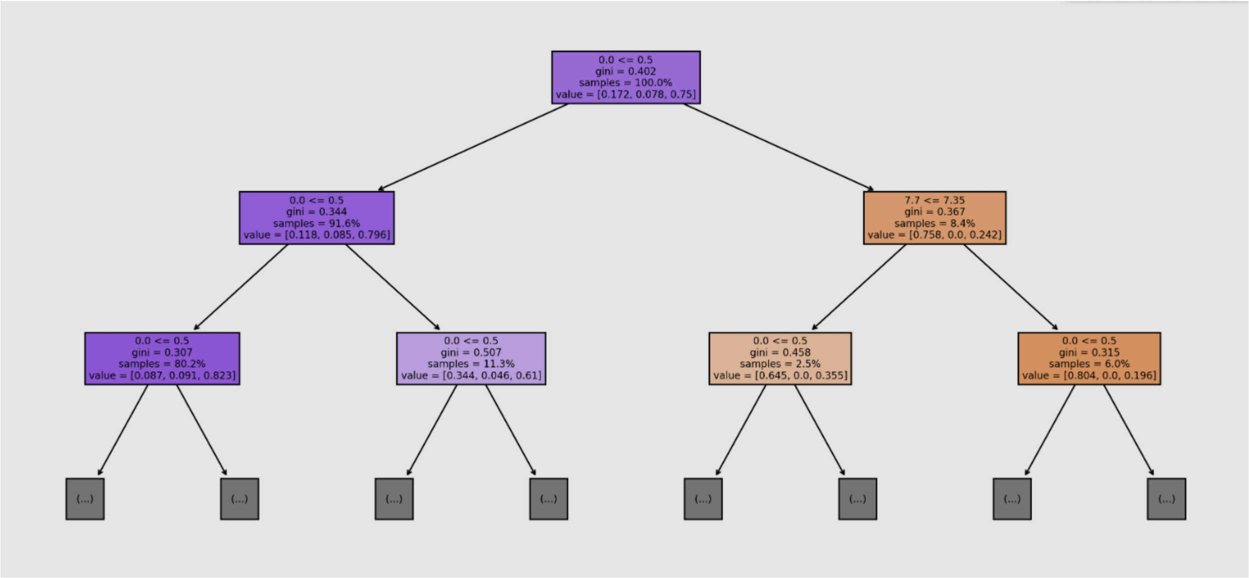


Genre Popularity in Japan



4. Can we predict where a video game will be successful?

   Our machine learning model for predicting what region a game was most successful in was a relative success. After focusing on the features of genre, platform, developer, critic score, and ESRB rating and creating a new column that identified which region a game sold the most in, we were able to create a model with a test accuracy of 78%. We attempted to take it a step further and see how well a regressor model would do at predicting the actual amount of money that a game would generate, but we did not succeed with this because the information we had in the data set was not enough to accurately predict sales totals (which makes sense since there are many things that go into total sales such as marketing, budget, release date, etc.). We attempted to show a visualization of the decision tree for the classifier model but it seems to be using some strange notation and a multitude of steps that made it basically illegible, nevertheless we included the first few decision steps below.



Challenge Goals:

1. Machine Learning

   Our project used machine learning to predict which region will a video game be the most successful. We used features (genre, publisher, etc.) to train a model that attempts to predict what region of the world it was most successful in. We also used machine learning to predict the global sales of a game based on the features mentioned above. While this prediction was not accurate at all, this is most likely due to the fact that our dataset did not include variables like video game marketing and development budgets.

2. New Library / Advanced Visualizations

   We made use of a new library called plotly to visualize our data analysis in a more advanced way than we had been shown in the course so far. We used our data to create scatter plots + trend lines, histograms, and pie charts detailing the different features in the data set. All of these charts are interactive because the user can use their mouse to scroll over different parts of the chart to learn more about that particular element. We also had to learn new data wrangling techniques such as the melt() function that allowed us to pivot the data into a representable version that could be used in a grouped bar chart.

3. Multiple Datasets

   The two datasets that this project utilized contained video game sales, ESRB ratings, and critic and user scores. Merging these two datasets allowed us to analyze how these different factors affect sales.

Work Plan Evaluation:

- Merge Data sets: This task should take < 1 hour, and should hopefully be a relatively smooth process as it is the precursor to the rest of the analysis in our report.

  This estimation was a bit optimistic, we ran into some issues such as having duplicate entries for a single game due to there being several releases across multiple platforms. This actually ended up taking roughly 2 hours as we had to figure out how we could merge the datasets for analysis while maintaining all of the important information that was needed to answer our analysis questions.

- Wrangling data: We need to filter and organize our merged dataframe into several sub data frames that will answer specific questions. This process should take no longer than half an hour, seeing as we will only need to decide on how many smaller data frames to create and which columns to filter for.

  This estimation was fairly accurate, instead of creating multiple smaller dataframes we instead filtered at the beginning of every function for the relevant information for that particular function. It ended up being a step that was taken several times throughout the process, but each time only took a few minutes to figure out which columns/rows were relevant.

- Various visualizations: Each visualization should take ~2 hours to complete depending on the complexity of plotly and having to learn new functions, but this process could become much quicker as we become more familiar with the package. In total we’re hoping to have ~10 visualizations, and hopefully the entirety of them will take less than 8 hours.

This estimation was accurate, learning the plotly library was rather difficult at times due to the customization that we wanted to do and led to learning several new pandas commands that allowed us to plot the way that we ideally wanted to. A lot of the work time was spent on the initial understanding of the plotting functions and arguments, but once we had created a few visualizations it became easier to recreate with new variations.

- Machine Learning model: developing a model should take ~2 hours, we will have to decide on parameters that we may need to set and also how to handle certain features, but this time frame is probably the most variable out of all of the pieces. The model could work out and give good results in an hour or we could run into issues and not be able to provide anything at all.

  This estimation was also fairly accurate, most of the work came into theorizing which columns should serve as the features and how to represent the sales columns as a label. Once we figured out how to create a new column that served as the highest selling country and zeroed in on the columns we wanted as features it became a fairly simple process that was very similar to the machine learning examples we'd done in class earlier this quarter.

Testing:

- The majority of the analysis was done on a massive data set of real world data from 1985-2016 and the only way we could conceive of testing our results is to check if our outcomes were reflective of other public data. Yet this would also be awkward to test because of large outliers (such as Wii Sports in our data set) and because of the date of the public data that we could compare our data to. Still, we wanted to see what the closest public information looks like in comparison to our analysis at a high level.

  Statista is a website with mass data analysis about many global markets including video game sales. To see the full analysis you need a premium membership, but a quick glance at the Video Game & Esports sections gives us enough info to get a rough comparison to our own analysis. Under the “Industry Leaders” section it is denoted that “The biggest publishers by market cap are Activision Blizzard, Nintendo, and Electronic Arts” which is accurately portrayed within our analysis. On the Video Games Topics page we can see that the most recognized gaming company in the U.S. is Nintendo, which is also reflected in our analysis by Nintendo’s dominance as a publisher. Statista also has a genre breakdown from 2019 that shows a bit of a different result than our analysis, with a slightly different order of genres in the top 5 but this is most likely due to outliers and difference in time period (our data set does not account for a few games that have exploded in popularity over the last few years such as GTA V and Minecraft, which are now the top 2 selling games of all time to date). Overall we get the sense that our data analysis is accurate for the time that it was taken from.

Statista Links:

Video Games Overview:
https://www.statista.com/markets/417/topic/478/video-gaming-esports/#overview

Video Games Topics:
https://www.statista.com/topics/868/video-games/

US Genre Sales:
https://www.statista.com/statistics/189592/breakdown-of-us-video-game-sales-2009-by-genre/

- The testing of the machine learning process is virtually built into the process itself when we run the training and testing accuracy of the model. We would've liked to test this model on more examples that were not included in the data set (such as newer games that have been released in the last few years) but we would need those observations (games) to have the same columns (features) so that they could be fit into the model. Ideally this would come from the source of our data, the vgcharts website, but their most recent data is very incomplete and does not have the information we would need to create an even newer test set.

Collaboration:

- We did not collaborate with anyone else on this project but we did make use of several of the HW assignments for the machine learning portion of the project and we regularly referred to the plotly documentation online to learn more about how to customize the visualizations. We also made use of sites such as stack overflow when we ran into wrangling issues, particularly when we were attempting to create the grouped bar chart.