

Video Game Analysis

By the Super Mario Bros.

Nathan Shepherd and Mike Mahoney

2018

**Background**

Video games as a medium have been around less than 50 years. Although pong is widely recognized as being the first video game, near a decade before “SpaceWar!” was created by Massachusetts Institute of Technology. It was a space computer game released on the “ PDP-1 (Programmed Data Processor-1), then a cutting-edge computer mostly found at universities. “(1962, 1). In 1972, Pong was introduced as an arcade game, “Selling at just $1095”, marketed and distributed at social gatherings around cities (1973, 2). Pong was essentially tennis, with player limitations of only controlling the Y-Axis. In 1975, Atari created the console market and marketed their first system with ironically Sears even distributing 150,000 copies as a Sears brand. From here the video game market was born, many variations of console systems were created over the next decade, but few were actually successful.

In 1985 Nintendo changed the small but growing gaming world by releasing the NES or Nintendo Entertainment System. Widely considered to be first truly popular gaming system, it would end up “selling over 60 million units” (3). Nintendo would continue to make cutting edge products, such as the Game Boy, SNES, Nintendo-Switch and the highly successful Wii; and set industry standards with innovations such as, the button configuration and layouts, On the go gaming, and the now famous motion control. The PlayStation would be created over a contract dispute in which Sony was assisting Nintendo with integrating the CD-ROM in their console, on the Super Disc (4). Despite this dispute, Sony would release the PlayStation X in 1994, and it would be one of the first consoles to actually use a CD-ROM to play games.

In 2001 Microsoft debuted the Xbox, a new more powerful gaming system that used windows 2000. The Xbox was released during a saturated point of video game consoles, where PS2, the Sega Dreamcast and Nintendo GameCube dominated the market. It was however when Microsoft acquired Bungie (creator of Halo) for $30 million, that the Xbox began to disrupt the console market (5). Microsoft now had exclusive rights to what will be one of the largest video game franchises ever Halo. Microsoft would become known for their exclusive games, only being offered on the Xbox system, starting what would become the growing console race.

In 2006, Nintendo revolutionized the gaming industry again with the introduction of the Wii, which for the first time allowed the player’s motions to control avatars in a gaming system. Players were no longer bound to their sofas for hours on end playing games, there were instead bound to their carpets in front of their televisions, learning to master the controls to the motion system. The Wii would usher Microsoft and Sony’s own adaptations to their gaming systems, indeed fueling the console race. However, few games would be able to compete with Wii’s initially released games, like Wii Sports.

Fast forward to 2018, the three gaming giants are still entrenched in a console war mentality. The current buzzwords around the videogame industry are virtual reality and augmented reality. Virtual Reality completely immerses a gamer in a simulation, like games do now except often with a headset to create the perception of being in the simulation. Augmented Reality takes our current reality, and adds filters to create a modified version of reality. The most famous example is Pokémon Go, in which players traversed entire cities to find rare collectibles often in the most unusual places such a random stranger’s house or even the Holocaust Museum; and the game has even inspired legislation, like Pidgey’s Law (6).

Despite the relative infancy of the video game era, the video game industry is worth a phenomenal $100 billion estimate for 2017. In September 2013, Grand Theft Auto was released, earning a staggering $800 million in 24 hours, reaching $1 billion after 3 days (7). Video games can no longer be viewed as a fad, with these sales figures. What once was a solitary sedimentary hobby, has evolved into a behemoth that is allowing players to interact and control their worlds.

**Motivation**

Why video games? Why analyze the customer trends and buying patterns? Why attempt to analyze the next video game success? Besides being video game fanatics, the video game industry is evolving in to a behemoth, producing fantastic simulations and epic failures. If one could figure out the trends of customer buying habits or what makes a great game, it could be invaluable to game publishers. The current video game market is estimated at 75 billion dollars, so an accurate prediction and depiction of the market would be invaluable to both video game publishers and the companies producing the console.

As the video game market is still in infancy (really teen-age), our proposed research could help these companies optimize their markets such as; if racing games are not popular in Europe the company might forgo marketing and release in that market. The company might save resources by not imputing those languages in their game and save research and development time, improving their release dates. Understanding genre success, would be invaluable to a company if the right information and trends are obtained.

From our research there is not a standard measure of success for a video game. Grand Theft Auto 5 is most certainly an outlier, but it is present never less. If one uses the Pareto principle (The 80/20 rule) then 80% of our gamers are playing 20% of the games, then Grand Theft Auto would most certainly be at the top of the 20%. This is precisely why our research is important, to find not only what is an important game but where in the world are they being bought.

**Analysis Goals**

Our analysis goals include the ability to determine if a game will be a global success, is it possible to predict a successful game? Can we accurately predict the next juggernaut game? We are looking to see what makes a particular game successful for our training set and using our test set to gauge our prediction.

Our other main goal in the analysis is market segmentation, by observing and comparing sales in each of the corresponding geographic areas. We looked at genre, platform, Critic Score, User Score as well as console, to determine if we could accurately cluster our geographic areas. If we can accurately understand the patterns by geographic area, the information could be invaluable for marketing in these markets.

**Data Sources**

Our data is from Kaggle, an open source website for finding data sets. It is a web scrape of the website VGCharz.com, appended with a user score and critic score from Metacritic. Our data was scraped in January of 2017, and does not the 4th quarter sales from publishers which is thee most important quarter for these publishers. Our data was scrapped by Gregory Smith and can be found at <https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>. The original data set contained 17416 observations with 17 variables. These observations include games from year 1976 like Super Breakout for Atari to popular games of 2016 like Madden 2017 (publishers often release sports games early). We also attempted to web scrape VGCharz with python, but were unable to access VGCharz’s API to gain access to the server. Some of our variables include the name of the game, the publisher and the year, so let’s dive into them a little deeper.

The name of the game and the year are self-explanatory. The parent game publisher is included but not the child publisher. For example, the very popular (and awesome) Assassin’s Creed franchise is produced by Ubisoft, but each sub publisher gets to produce their own independent franchise game and that information is omitted.

There was also a Rating variable which broke down each game. These ratings correspond to the Entertainment Software Rating Board (ESRB), which is a third party non-profit organization mainly serving the United States, which gives ratings (based on factors like drug use, profanity and violence) to media to help parents understand the content they are giving their children. Our Ratings variable is broken down into: RP, E, E+10, T and M.

One variable includes the genre of the game, which can be divided up into categories such as: Action, Adventure, Fighting, Miscellaneous, Platform, Puzzle, Racing, Role Playing, Shooter, Simulation, Sports and Strategy. This helps to cluster games into certain groups, despite many games being combinations of these categories.

The platform each game is released on is included, which is what video game console it was released on. The list of platforms includes the early systems such as Atari to the widely popular systems of today including: Xbox One and Ps4. One important note is games can be released on multiple platforms and often have different sales on each platform, but we did not combine the platforms so each game is also divided by the platforms of release

We also have the video game sales for four distinct geographic areas that include: North America (NA\_Sales), Japan (JP\_Sales), Europe (EU\_Sales) and Other (Other Sales). One important note is China roughly represents 25% of the global market and it the dominant force in Other sales. Together these four regions encompass Global Sales (Global\_Sales), which is going compose to be the dependent variable.

Other variables include the Metacritic score such as: User Score (User\_Score) which is what gamers give as the rating and Critic Score (Critic\_Score) which is the rating the professional give the games. It is important to note we cleaned our data to include both Critic and User score, which shortened our data to 7192 observations and included data up to the mid 1990’s.

Our dependent variable is success, a binary variable we created, where 1 indicates success and failure. We defined success as global sales over 5 million. We understand this number is subjective, but without a measure of success we had to try different metrics. After changing the range of our definition of success, we decided to return to games sold over 5 million globally as successful.

**Data Analysis**

To accurately predict success using the critic score and user score, we removed any observation missing these two variables. As a result, we were only using about 40% of the original data. Despite the data truncation, we agreed it was best because user score and critic score were essential.

User Score and Critic Score were not on the same scale, as User Score was on a 10-point scale and Critic Score was on a 100-point scale. We agreed it was best to have both on the 100-point scale so we multiplied each observation in User Score by 10. Otherwise, the data was very clean and required little data preparation.

Our entire project was done with the R program, including cleaning and analysis. To accurately predict success, we focused on three methods: Logistic Regression, Tree based models and Support Vector Machines. We planned on using more methods, but as you will learn in the results, it would not have likely changed the outcome.

**Predictive Model Results**

Logistic Regression was the first model we ran, attempting to predict success with the NA\_Sales variable. Our first logistic model was built with NA\_Sales, Genre and Critic\_Score. Despite our best efforts we were unable to accurately produce a model with logistic regression. In fact, we were unable to produce a model using logistic regression for any of our distinct four regions. So, lets move on to our next model, the tree-based methods. One important note, using a summary function on the logistic model we did learn that Critic Score (2.2e-16) was much more significant than User Score (1.4e-5), as a result we focused on only Critic Score.

Our first tree model was built using NA\_Sales, Critic\_Score and Genre, predicting success. We pruned the tree at the optimal value of 7, and found that the most influential genre predicting success was Sports. We obtained an accuracy of 98.78%, a sensitivity of 0.65 and specificity of 0.99.

Our next tree model was EU\_Sales, Critic\_Score and Genre to predict success. We pruned the tree to the optimal value of 8 and found the most influential genre predictors were strategy and sports. We obtained an accuracy of 98.87%, a sensitivity of 0.6 and specificity of 0.997.

The next tree model was JP\_Sales, Critic\_Score and Genre predicting our success of above 5 million in sales. We found the optimal tree size of 11, and found Critic Score to be the most important factor splitting the nodes at above and below a Critic Score of 83.5. The accuracy was found to be 98.6% with a sensitivity 0.49 and specificity of 0.997.

The final tree model was Other\_Sales, Critic\_Score and Genre to find success. The optimal tree size was found to be 6, and there were not any clear splits in genre, just further cut points of Other\_Sales. The accuracy was 98.87%, sensitivity of 0.6 and specificity of 0.997.

Our tree models were able to predict with very high accuracies and high specificity rates, but we were producing low sensitivity rates (between 0.49 and 0.65). In an attempt to find the best model for success we then switched our approach to the Support Vector Machines (SVM’s). We discovered the best kernel for the SVM was the default kernel radial.

The first SVM consisted of NA\_Sales, Critic\_Score and Genre, to predicting success metric. After tuning the model, the best parameters were gamma of 0.1 and a cost of 1, the accuracy obtained after tuning was 96.4%. This model yielded an accuracy of 99.12%, a sensitivity of 0.7 and specificity of 0.997, which was better than our tree-based model.

|  |  |  |
| --- | --- | --- |
| JP Sales | 0 | 1 |
| 0 | 2109 | 35 |
| 1 | 1 | 12 |

Our Europe SVM model used EU\_Sales, Critic\_Score and Genre. The tuned model also gave the best parameters of the best parameters were gamma of 0.1 and a cost of 1 and a tuned accuracy of 99.12%. It yielded an accuracy of 99.12% with a sensitivity of 0.60 and specificity of 1, only better than our tree model, with similar sensitivities and specificities.

|  |  |  |
| --- | --- | --- |
| EU Sales | 0 | 1 |
| 0 | 2110 | 19 |
| 1 | 0 | 28 |

Our Japan SVM model used JP\_Sales, Critic\_Score and Genre, predicting success above 5 million in global sales. Again, the best parameters were a gamma of 0.1 and a cost of 1, with a tuned accuracy of 98.33%. The final SVM model gave an accuracy 98.33% with a sensitivity of 0.25 and specificity of 0.999. This model yields substantially lower sensitivity than our Japan tree model with only a slight dip in accuracy.

|  |  |  |
| --- | --- | --- |
| Other Sales | 0 | 1 |
| 0 | 2107 | 16 |
| 1 | 3 | 31 |

The Other SVM model used Other\_Sales, Critic\_Score and Genre. The tuned model also gave the best parameters of the best parameters were gamma of 0.1 and a cost of 1 and a tuned accuracy of 96.4%. It yielded an accuracy of 99.12% with a sensitivity of 0.65 and specificity of 0.999, still better than our tree model.

|  |  |  |
| --- | --- | --- |
| NA Sales | 0 | 1 |
| 0 | 2017 | 46 |
| 1 | 3 | 1 |

Three of our SVM models preformed only slightly better than their corresponding tree-based models, giving modest improvements to sensitivity but only slightly to specificity and accuracy. However, the Japan SVM model gave substantially lower sensitivity and a minimal decrease in accuracy. Despite the shortcomings of the Japan SVM model, we still felt SVM gave us a better prediction than just the tree-based models alone.

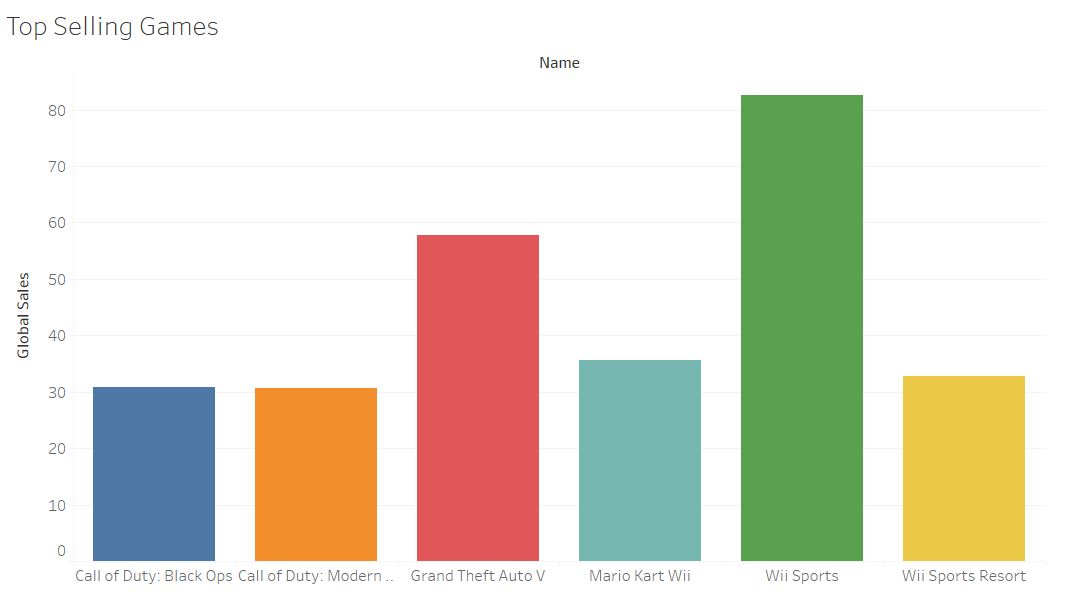
In an attempt to further understand success, we focused on what makes the individual markets successful. To understand what makes each geographic market successful and to better understand the global market, we started with North American Sales because it is the driving force in Global Sales. We again used Critic\_Score and genre to predict NA\_Sales, but moved our metric for NA\_Sales. We tried many different values for our North American success in the range of (0.5 to 4 million) and found each gave us the same result. We were unable to predict any success in NA\_Sales, despite trying different variables like User\_Score, Platform or even Rating.

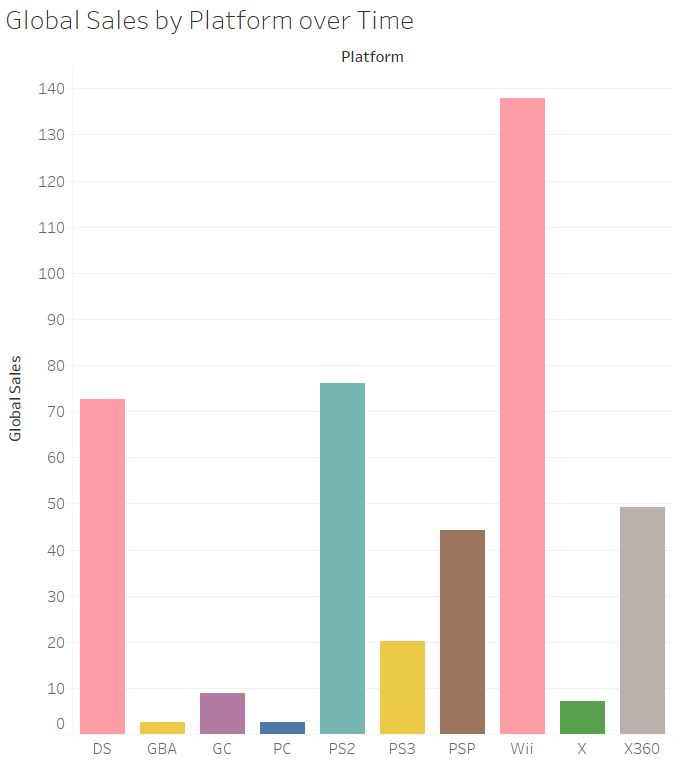
We also attempted to predict sales of our remaining three markets, modifying the metrics and predictors but reached the same conclusion. We were unable to predict and successful games, in any market. We believe that because we only had categorical variables and two interval variables (User\_Score and Critic\_Score), we were unable to predict success in each geographic area. In fact, we also believe our predictions of global success utilizing each geographic region did not provide substantial models.

The Global\_Sales is an aggregate of each market for that particular game, as a result our Global\_Sales was highly correlated with each geographic area. NA\_Sales and Global\_ Sales had a correlation of 0.95, knowing that the United States market completely dominates every other market, if the game is successful in North America it will be successful across the world. However, this correlation suggests the necessity for more variables that break down NA\_Sales to try and get a better prediction. We also found that each remaining market was also highly correlated with Global\_Sales, the breakdown is EU\_Sales (0.94), JP\_Sales (0.61), and Other\_Sales (0.81). The only region not extremely correlated was Japan, which shows that it is most likely the smallest market considering the country only has a population of 127 million versus the 320 million in the United States, the 740 million in Europe and over a billion in the rest of the world. Despite the extremely high accuracy rates achieved in our models, we these models do not currently produce the best predictions for success using the variables given in the data set.

**Customer Segmentation Results**

With our perplexing accuracy, we decided to visualize our variables to detect any peculiar trends. Comparing Global Sales by Critic Score one variable stands out from the rest. The game Wii sports is rated low, but it has the highest sales out of any game. To investigate further a chart of games earning greater than 30 million in sales was created from the depths of Tableau. From that chart Wii Sports is still the top seller, followed by Grand Theft Auto V, Wii Sports Resort, and a couple games from the Call of Duty franchise. This still doesn’t explain why a low rated game sold so well. A graph was churned in the Tableau cauldron and out burst a graph tracking global sales of consoles over time. When you reach the release of the Wii, 2006, it had the highest sales out of the other consoles. Here’s what they look like:

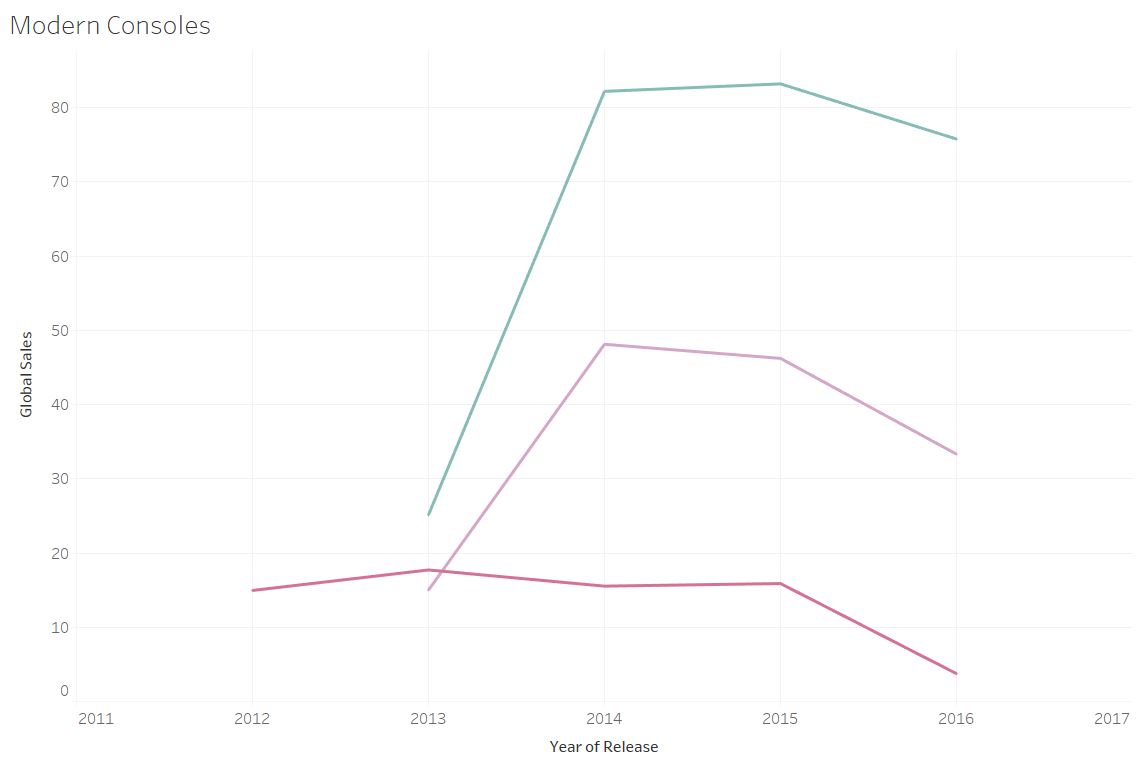




With further research it was discovered that when the Wii initially released, a game came bundled with it. With every Wii this game came packaged with it. That game is Wii Sports, DEDUCED. Wii sports sales are artificially inflated, which would explain how a low rated game sold so well. It’s interesting to note that the same case occurred for Wii Sports Resort, but it wasn’t bundled with the console. It was bundled with a Wii controller that contained hardware call the Wii Motion Plus, which improved the Wii’s motion controls. This occurred in 2009. In 2010 there was a later bundle that included the Wii Motion Plus and Wii Sports Resort with newer Wii consoles at the same price as earlier Wii consoles. A good deal with great sales but didn’t have the same impact as the initial Wii with Wii Sports bundle. That satisfied the Wii Sports question, but now there was a question about bundles. How much of an affect do bundles have? Sadly, our data does not contain any data about bundles but with some extra research we found a couple more examples.

In 2010 Xbox 360 sells were dominant. A highly anticipated franchise game was released that year, Halo Reach. A Halo Reach Bundle was advertised, which included a Halo Reach themed Xbox, two corresponding themed controllers, and a copy of the game. The game sold well, but it wasn’t the highest selling game for Xbox that year. Kinect Adventures was the highest selling game. The Kinect is Microsoft’s version of Wii motion controls. Instead of motion controls there is an infrared camera that tracks your movement. If you guessed that there was a bundle that contained the Kinect and Kinect Adventures, you are correct and win a solid high five. In 2011 a highly anticipated PlayStation exclusive released, Uncharted 3: Drake’s Deception. A PS3 bundle was released that contained Uncharted 3. PS3 sales beat out the other competitors for that year and this bundle may have played a big part. Our analysis could have been better if our dataset had sales for game and whether those sales were individual or part of a bundle.

Those were metrics about the previous console generation. How are today’s modern consoles doing?



PS4 and Xbox One have been doing well. The declines in their curves are to be expected since the consoles have been out for a couple years. The sharp increase is correlated with the initial release of a console. What happened to the Wii U? It’s initial release sales pale to the other consoles, and the Wii U never really recovered. This a perfect time to bring up a variable that would have been helpful in at least explaining why the Wii U did so poorly. The variable is Launch Titles. Launch titles are crucial in getting consumers to purchase your product. Nintendo is well known for their Mario and Zelda titles, but there were no such launch titles for the Wii U. There were no popular franchise games, just games like ZombiU and Just Dance 4. When introducing new consoles, a good first impression is key. A console may be the next cutting-edge piece of technological genius, but if there are no good games to play why bother? It would have been nice if our dataset had the sales for Nintendo’s new console, The Switch. The Switch had great launch titles and it’s selling just as much if not more than the Wii did. Launch titles would not be the sole reason a console fails, but it would help add more detail than what our dataset has.

When looking at global sales by publisher majority of the sales are going to three publishers. The top publishers in order of greatest to least is Electronic Arts, Nintendo, and Activision. Over the years Activision and Electronic Arts have done numerous mergers and acquisitions. They may have not had a hand in making the actual game, but they do publish it. Nintendo is interesting since they haven’t really merged with other companies. They have worked with other companies but never outright bought them.

We wanted to see if certain genres were popular to specific regions. There are quite a few genres so here is a visualization to expedite the explanation.



North American and European sales per genre are almost identical. Action genre reigns supreme with shooter and sports taking second and third place respectively. Japan sales tells us a different story. Role-playing games are the most popular in Japan. Shooter and sports games don’t even come close. An interesting thing to note though is when comparing the sales between North America and Japan for the role-playing genre. Even though role-playing games are not as popular in North America, sales for that genre are still higher in North America. This reinforces our point that if your game is successful in North America then in can be considered a successful game.

Here is a snapshot of the top ten games with the highest concurrent player count on Steam. Steam is solely on the PC but it has easy access to their statistics. Acquiring statistics for player count on consoles is much harder to obtain. This snapshot reinforces the previous bar chart in showing what genre of game most people are playing:



Eight out of the ten games are shooters. There is only one shooter out of those eight that has single player as well. Dota 2 is a strategy game but it’s also free to play. Someone can download Steam at this exact moment and install Dota 2. Dota 2 does have microtransactions for cosmetics and other items that do not affect the core gameplay. Microtransactions would have been some excellent data to have, especially in ranking popular free to play games. The other non-shooter is Subnautica which is more of an adventure submarine game. Shooters and action games are indeed popular, but action and shooter are broad genre categories. Shooter could even be considered an action genre. This dataset would benefit the inclusion of more detailed genres or a category of sub-genres. A good example is the Player Unknown’s Battlegrounds (PUBG). With roughly 700,000 more concurrent players than Dota 2, what makes this shooter any different? PUBG is a different kind of shooter. It’s multiplayer only like other shooters but it’s technically called a battle royal game. Battle royal means that 100 players start the match in a plane with nothing but the clothes on their back, or no clothes if they feel so inclined. Don’t worry, underwear is mandatory. The plane flies over the map or play area, people parachute out, and land at different parts of the map. The whole purpose of the game is to kill everyone else and be the last person standing. Once you die, that’s it. You get sent back to the title screen and must start the process from the beginning. It’s a very high stake, adrenaline fueled survival/arena shooter. There are also few games that use this type of genre because the genre itself is still relatively new.

**Recommendations**

Through our data exploration, we discovered several problems with our data set. One problem is that video game sales data is such a highly guarded secret, we are only being given the data these publishers want us to have. One example is the complete omission of the popular N64 video game console, which includes the extremely popular Golden Eye game which set the current standard for multiplayer first person shooters. One would wonder why Nintendo would not volunteer sales for such a successful system. There is also a lack of mobile application games, as its not really a secret that Angry Birds is one of the highest grossing games of all time.

Another issue we found was the lack of variables that could actually predict success such as Multiplayer, the Rate of Change (over months), Franchise and Bundle. Through domain knowledge we know that if a game has multiplayer, regardless of its actually a good game, will drastically improve how long the gamer actually plays that game. The presence of multiplayer also gives the publisher more profits with microtransactions, in which the gamer could uses actually currency for non-transferable in game currency to buy anything from more lives, to better in-game accessories or even to customize avatars. The presence of multiplayer allows the publishers to literally print money and its absence is almost foolish. Still we believe a binary variable indicating the presence of multiplayer.

We believe if there was a variable showing the rate of change over months, we would have a better understanding not only of sales but how quickly a game is selling. Do certain game sell extremely well during their release or the holiday season, or do the they fizzle out and completely stop being sold? Do certain genres have better rates of change? We felt that another presence of time, might be independent of sales in of the geographic region and could be an excellent indicator of how well a game is really selling, and could be less correlated with just regional sales.

If a game is a franchise drastically impacts a game’s success. In our top selling game chart, two franchises are present: Call of Duty and Wii Sports. In fact, of the top 50 selling games several franchises are present, which makes sense if a game sells well the first time, the sequel or reboot will have gamers lining up to buy it on release day. A binary variable indicating the game as a franchise, could possibly show us how much being a franchise really matters to global sales and each individual market.

The final variable that could aid in our predictions is if the game is a bundle. Often times when a gamer purchases a new console (sometimes even used) a game is included. Take Wii Sports, it has terrible reviews bother User and Critic, but still it is one of the highest selling games of all time. If Wii Sports was not a bundle it probably would have abysmal sales. Still, being able to accurately understand which games are bundled could provide insight into why certain games might have artificially inflated sales.

After reading the data set and the comments associated with the data set, it is unknown whether or not these sales are new or used. Often publishers will release Downloadable Content (DLC) several months after a game has been released, which spurs gamers to buy a used copy after they completed and traded these games. For this information GameStop would be a tremendous resource because it would have that data and it is the largest used video game store in the United States.

**Future Analysis**

As a team we agree that more information is needed to predict success. If we potentially had data with variables we mentioned and the sacred data video games publishers keep in their Fort Knox like vaults, we might be able to give a more accurate depiction of success. Furthermore, our results are not a deterrent but perhaps gave a deeper understanding of what information is needed to provide an accurate but actual prediction of a truly successful video game.

We believe that the video game market is due for it’s next evolution. With the growing technologies of Virtual Reality and Augmented Reality, that which ever publisher and console can capitalize on these new technologies by producing quality games and consoles will dominate. We already know that Nintendo is the company that not only adapts the new technology but executes quality in a way both critics and users agree. It would be a costly mistake for any console or game publisher to ignore this futuristic trend.

References

(<http://www.history.com/topics/history-of-video-games>), 1

(<https://www.atari.com/history/1972-1984-0>), 2

(<https://www.nintendo.com/corp/history.jsp>), 3

(<http://www.computinghistory.org.uk/det/4032/Sony-Playstation/>). 4

(<https://www.digitaltrends.com/gaming/the-history-of-the-xbox/>), 5

(<http://www.businessinsider.com/pidgeys-law-bill-aims-to-force-removal-of-pokestops-in-pokemon-go-2016-8>), 6

(<http://www.guinnessworldrecords.com/news/2013/10/confirmed-grand-theft-auto-breaks-six-sales-world-records-51900>), 7