The Name of the Game

By Daniel Reiter

TABLE OF CONTENTS

1. Proposal

To all video game developers, thank you for your time in discussing my proposal to hopefully help further your sales in this booming industry. While the games may not have started off as hot as their current state, the video game industry is one of the most successful in the world but it does come with some issues. While I doubt a crash as bad as the one that occurred in 1983 would happen again, there is certainly a fear of oversaturation in the market. Sure, many companies have successful franchises as well as systems that keep them in the black but more is not always better and smaller companies can struggle to survive. Releasing a similar game each year can grow stale and gamers will always latch on to a new game eventually. My proposal is to look into the sales of video games, especially the all-time best sellers, to see what kinds of games sell the best.

Looking into a dataset from kaggle, I can dig into all of the best-selling games from 1985 to 2017 with much more information than just the title and the amount sold. I can look at the genre, release year, sales by regions, review scores, ESRB rating and more. I could find out what types of games sell the best in different regions. Should all of the focus be just on those areas or can certain games that are huge in one region become big in another? And while first-person shooters may be big now, how long will they stay on top? When looking at the years where some of the all-time greats came out, we may get an idea of how long certain genres stayed big. Plus, how important are review scores? Should companies focus more on user scores and sales of a franchise instead? There is a lot to be gleaned from this dataset just on the surface alone. However, this project looks to go a little deeper than that.

As helpful as this kaggle dataset will be, I also want to compare it with another set from the University of Portsmouth that contains sales info from 2004 to 2010. The point of looking at this group is to better compare the inferences made from the best-sellers dataset and whether they are truly applicable. As useful as that data is, we are talking about some of the best franchises ever and that name recognition can have a big factor on their ability to sell well. To compare it with a different and somewhat more contemporary list will lead to better analysis as well as predictions for the future of the industry.

And that really is the goal of this project, to analyze all of the trends in the selling of video games and create a predictive model that will help decide what a company should do moving forward. To do that, I will need to use some in-depth machine learning techniques to create a strong model. But to take this to another level, I will also use some natural language processing practices to help look further into what games sell best. These techniques would look at a kaggle dataset of video game reviews from Metacritic and see which words stick out. While there will always be mixed opinions on games, these reviews will help find what words are used most often when people discuss a game they like as well as dislike. This can give a gaming company more insight into what gamers prefer for either in a specific game or an entire franchise.

With everything put together, I hope this project can show deeper strategies for video game sales. Not only will I have the code uploaded to my GitHub but also write a report that breaks down all of the major steps I take as well as construct a PowerPoint presentation to help explain the biggest takeaways from this endeavor. While there will always be fads in the industry as well as new consoles that come out, gamers will persist. This industry has stayed around and grown because of its fans but that does not mean a recession cannot happen. Companies need to stay diligent and focused in order to stay afloat in this competitive field. Hopefully this project will give them all the insight needed to be successful and also maintain the love that gamers have for their craft.

2. Data Wrangling Report

This project entails three different datasets that need to be inspected and cleaned before going into further analysis. The first set is the all-time video game sales list from kaggle that includes not only the names of the games but also the sales for different areas, the consoles they were on, critic as well as user scores and even more. Following that dataset is the video game Metacritic user reviews, also from kaggle, that contains hundreds of thousands of reviews to cypher through with natural language processing. Finally, the last dataset is from the University of Portsmouth, which has the U.S. sales info of video games from 2004 to 2010. Most of this will be used in the machine learning aspect as a comparison to the models made with the other two datasets. Although none of these sets of data were overly messy, there were several necessary steps for cleaning before heading into exploratory data analysis.

When it came to the all-time sales list, most of the columns and rows were clean. The dataset was pretty tidy with each row being an observation (game) and each column was a feature. However, there was one issue that needed to be resolved. For many of the older games, there was missing data especially nonexistent review info (critic or user). This is mainly because they were around before the Internet was created as well as before reviews became common. And unfortunately, with reviews especially, there is no simple statistical measurement that would be helpful for the data because the scores can range wildly in either direction. Because of this, the only solution was to drop any row that contained an “NaN.” While many classics were lost, this was necessary in order to have clean data.

Also, to help make this dataset better, the features needed to be further enhanced. First off, the release years were changed from floats to integers so they would look cleaner and hopefully avoid any upcoming confusion. Secondly, each specific genre was made into its own binary column, which will be helpful when it comes to plotting. Lastly, the different systems that video games came out for were also made into individual binary columns for the same reason as the genre category. With these new columns, the dataset was now ready to be saved into a newer and cleaner CSV that would be used for exploratory data analysis.

For the second dataset, the few “NaN”s were dropped. Then, for later merging, many of the platform names were changed to match with those of the first dataset. Because games could come out on multiple consoles, these needed to be matched correctly. The main cleaning took place with the natural language processing. First, all of the reviews that contained symbols or just numbers needed to be removed because they would cause errors with tokenization. To do that, a function was created to remove any non-letters and replace them with spaces followed by minimizing any extra white space. Then, only the reviews that had words were kept. Finally, the reviews were whittled down more so that only those in English remained.

After the cleaning, this DataFrame was merged with the first one on the name of the game as well as the platform and together there was no missing data. After renaming a couple columns and dropping unnecessary ones, it was time for the actual natural language processing. Because of how many reviews and words were available, the matrix would become

too large if everything was run together and my computer could not handle it. To deal with that, a sample of 10,000 reviews were taken and put through a function to help find the most common words. The function made sure all of the words were lowercase, tokenized each word, removed the stop words, broke them down to the stems and lastly counted each word in each review. This made a dictionary that contained dictionaries for each review and the count of each word.

Now this data had to be transformed into a DataFrame with each row being a review and each column being an individual word that could be in the review. The values would then be how many times that word showed up in the review. In order to do that, the dictionary of dictionaries was turned into a list before then being transformed into a DataFrame. Because each column is an individual word, there were many “NaN”s. These were changed into zeroes, and then the DataFrame was put through a TF-IDF Matrix, which is better at weighting words by importance. With this new DataFrame, the most important words were chosen as the ones that had a mean value greater than 0.001. This created series of a little over 1,000 words that were considered the most important and would help with tokenizing the entire dataset.

The same process of tokenization was used on all of the reviews except a change in the stemming portion where only the important words from the earlier sample were stemmed and kept. From there, all of the DataFrame and TF-IDF transformations were the same as they created a new DataFrame that contained all of the reviews and how often the most common words occurred in each review. This was merged with the earlier DataFrame and then modified to only have the mean for each word for all reviews, grouped by each game and platform along with all of the other important features from the first dataset. Lastly, the DataFrame was saved into a CSV that would be used for the modeling section in the machine learning portion of the project.

Finally, the last dataset, which contained U.S. sales info from 2004-2010, was very clean and did not need any work. Each row was a different game and there were many different features as columns including the different genres for a game similar to what was done with the first set of data. Also, there was no missing data so no cleaning was needed. Now there were three clean DataFrames that would be used throughout the rest of the data process to find the deeper trends inside the sale of video games.

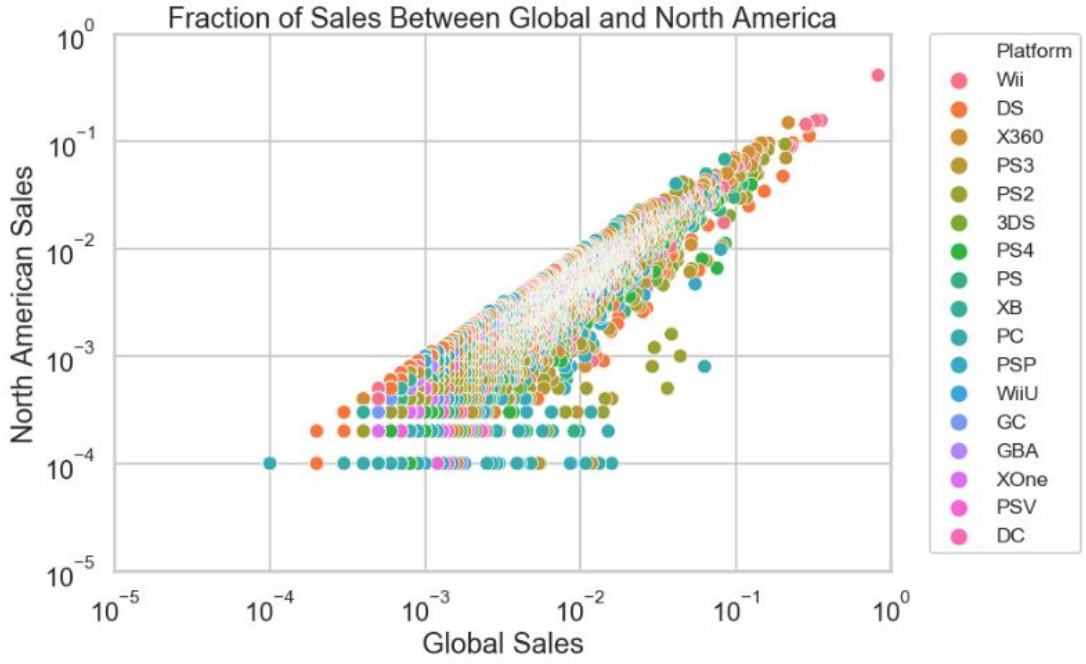
3. Data Story

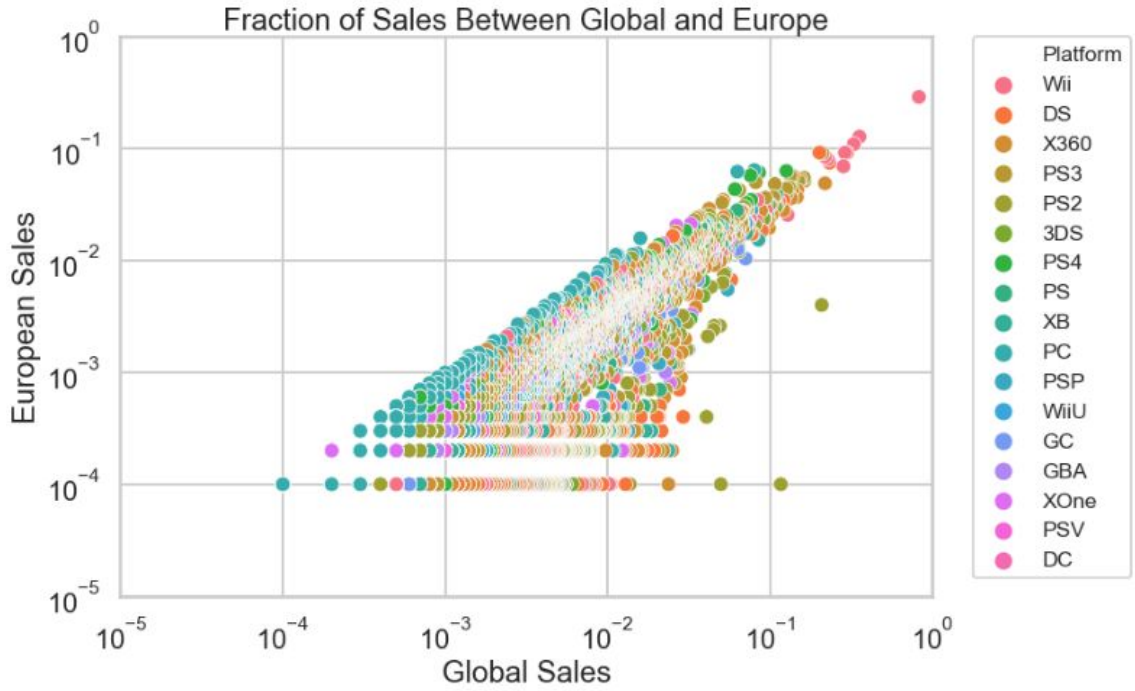
For this project, one of the key aspects I wanted to focus on was how video game sales could differ in the three main regions: North America, Europe and Japan. While there is still plenty to see globally, the goal is to see trends in each region and what factors are more specific for one compared to another. Of course, just because a plot shows that certain types of games do not do well in some areas does not mean that all will struggle there. This is more of a general idea of how video game sales are driven in these different regions.

To do this, I tried to keep plots grouped together so that it would be easier to see the difference in each. The main types of plots are Scatter Plots (including some with a logarithmic scale) as well as Bar Plots. There are plenty more visualizations I looked into but these ones stood out to me because of the trends and patterns they illustrate. After looking at all of these plots, I now have a clearer picture of what to delve into for my Inferential Statistical Analysis.

Plot Group #1

Before we look into each region specifically, I think it is good to see how they compare with Global Sales. In the end, what matters most for a video game company is how much total money they make on a game, not just what a certain region brings in. So I looked at how the fraction of sales in each area is related to the fraction of Global Sales. These were used on a logarithmic scale so it would be easier see all of the points as well as any possible correlation. When it all came together, one can see which regions have the strongest relationship with Global Sales.





There is a strong positive correlation between Global and North American sales as well as European Sales. The slopes have a steep upward direction with Nintendo Wii games sticking out near the top. This might signify how sales in these areas could be predictive of Global Sales overall. Japan, however, sees less of a positive correlation. That is not to say that this

correlation is not good but what can be inferred is that there are games that sell really well worldwide but do not sell well in Japan.

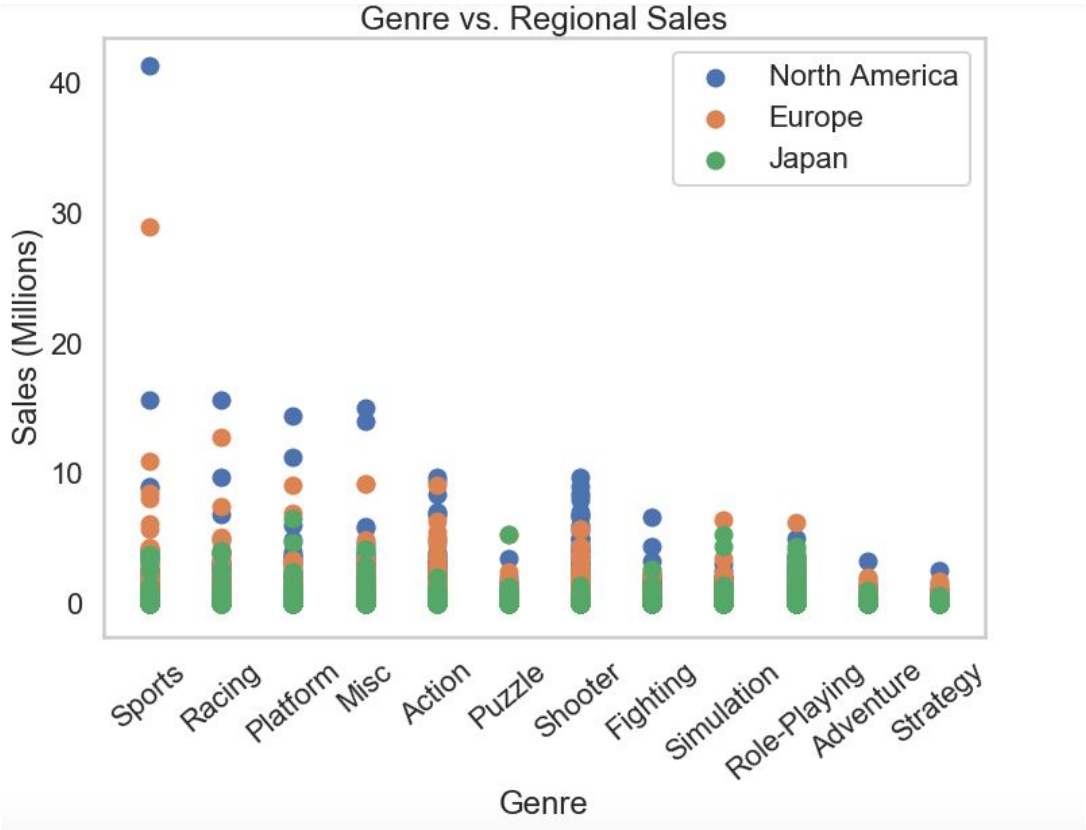
Plot Group #2

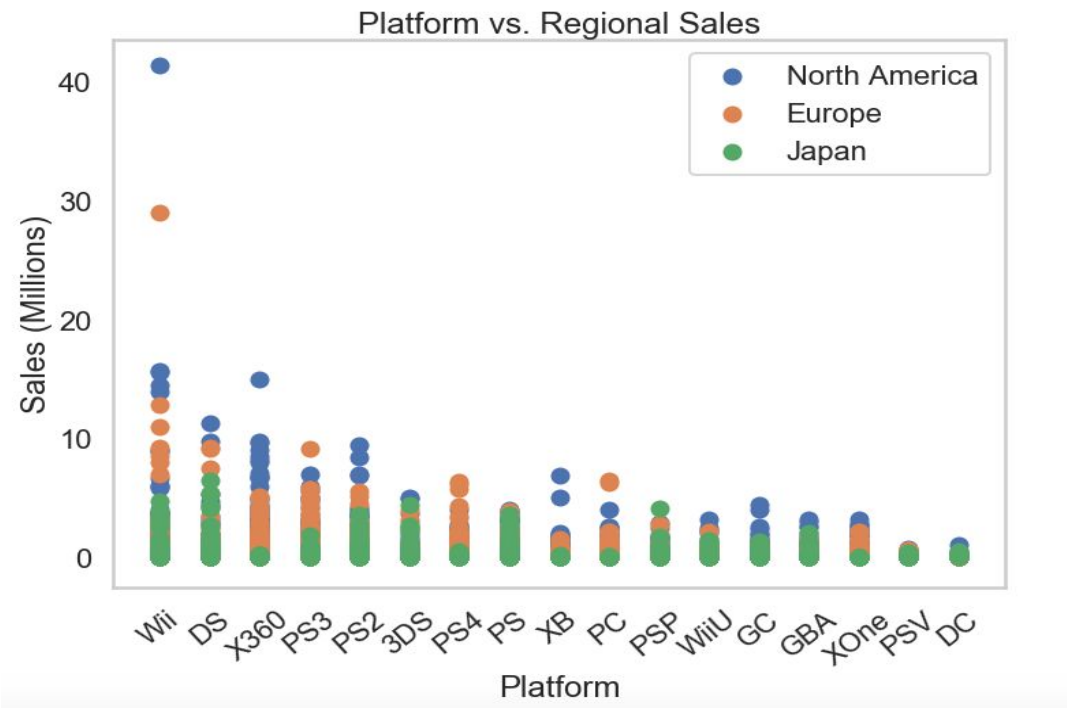
One thing to keep in mind when comparing these regions is the difference in population. First off, North America and Europe represent multiple countries including the U.S. so there are a lot more people in those areas compared to Japan, which is just one country. While the population is still large, it is still much smaller than the other combined areas. To that end, I looked at the fraction of sales between each region to see if there were any deeper connections between them.

Much like before, there is a strong positive correlation between North American and European sales while Japan's relationship with both is weaker. As mentioned earlier, part of this is the different population sizes but another may be opposite preferences. Some of the bigger, more mainstream types of games might be more popular in North America and Europe compared to Japan. While this group of plots offers plenty of insight, it is time to dig deeper into what types of games are bigger in these regions.

Plot Group #3

For looking at types of games, there were two specific factors I wanted to delve into: Genre and Platform. The genre was always important to me because there are so many different types of games out there and most gamers play more than one type of game. The platforms of the games are also important because they tell a lot about the consoles as well as companies that players prefer. One thing to keep in mind is that there are plenty of games that come out on multiple consoles so not every game is unique in that light. By looking at the below plots, one gets a better idea of the preferred genres and consoles in these three regions.





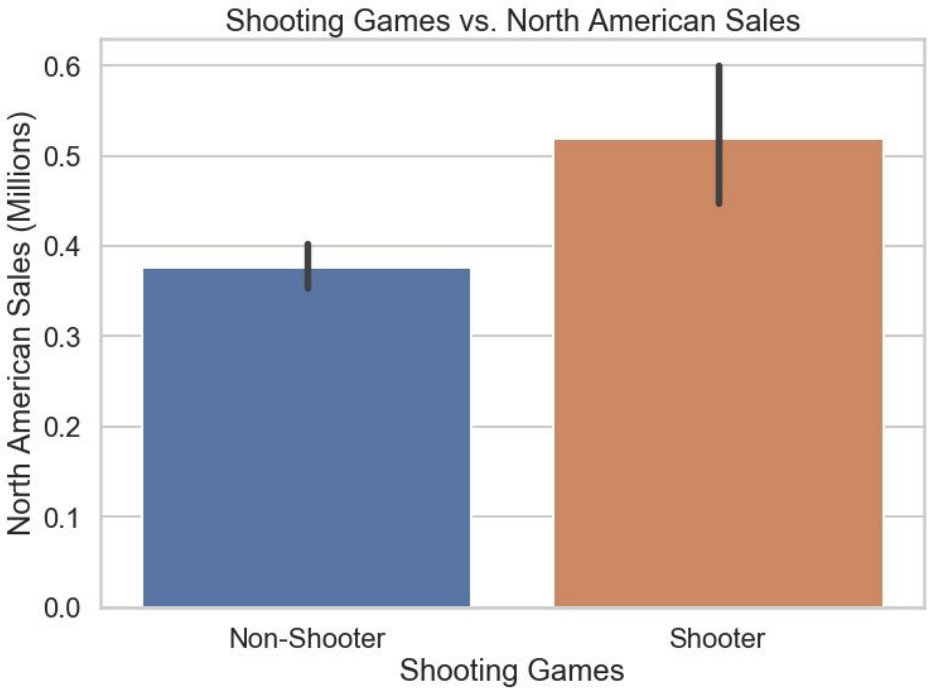
For Genre, Sports games stand out the most followed by Racing and Platform games. Sports is not surprising to see at the top because of games like “Madden” as well as “FIFA” that come out every year and are always hot sellers. The outlier at the top is “Wii Sports,” which is a little misleading because, especially early on, the game came with the Wii console. One of the

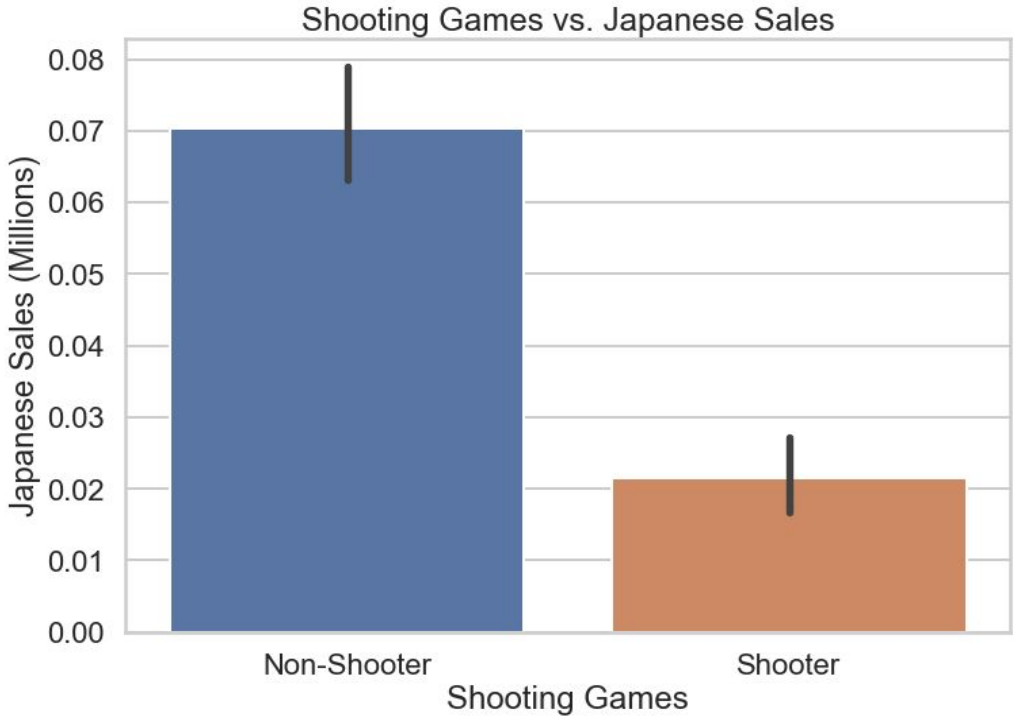
more surprising plot locations was that of Shooter games, which was more in the middle. That style has become big over the last decade or so, which may explain why it is the middle as of now. The genres near the end (Role-Playing, Adventure and Strategy) are not too surprising to see there as they are thought to be more for hardcore players.

In the Platform plot, “Wii Sports” is again the outlier at the top but Nintendo leads the way as Wii and DS games saw the most sales followed by Xbox 360 as well as PlayStation 2 and 3. This says a lot about Nintendo's hardcore fans as these systems are from previous generations of consoles and while the Wii U may have struggled, it is apparent how big Nintendo's fan base is. The order for most of these consoles by games sold is not too surprising except for possibly Xbox One. But one thing to keep in mind is that newer systems are still growing by the time this data was available so in a few years, they could be much higher. And if you are wondering why more classic consoles are not on this (like NES and SNES) that is because those games had a lot of missing data so they were dropped. Overall, this plot shows what many would expect when it comes to games sold by console.

Plot Group #4

After looking at the previous plots, the next step is looking at where certain types of genres are big. While there are many different genres to break down, one of the biggest ones is Shooters. Because of games like “Call of Duty,” that genre has become huge but yet it was only in the middle of the plot. Why is that? A look at those games sold in each region tells part of the story.



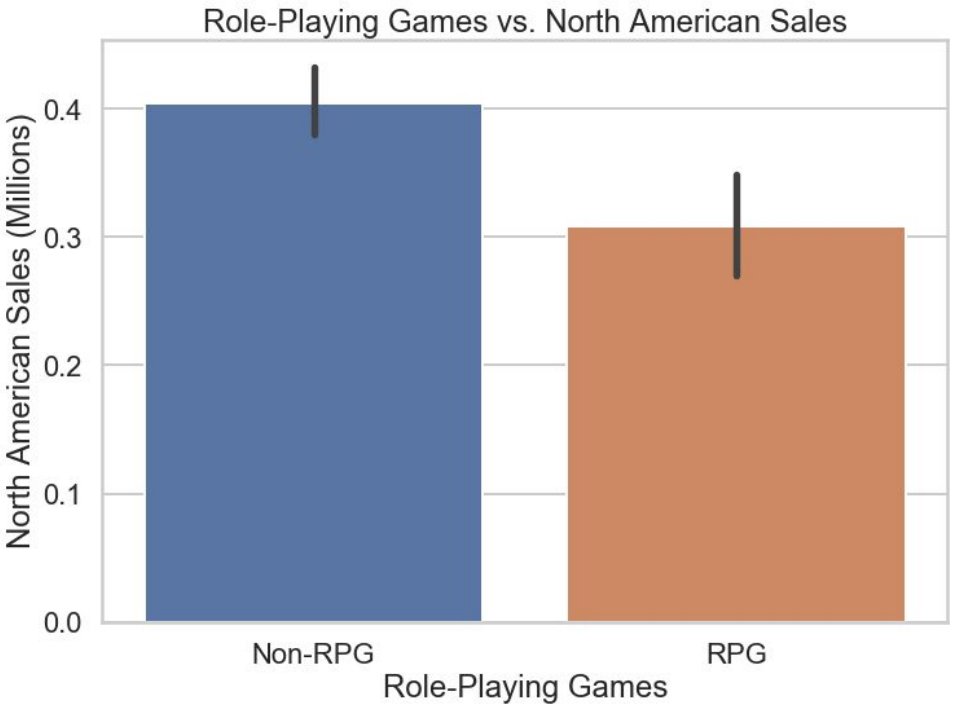


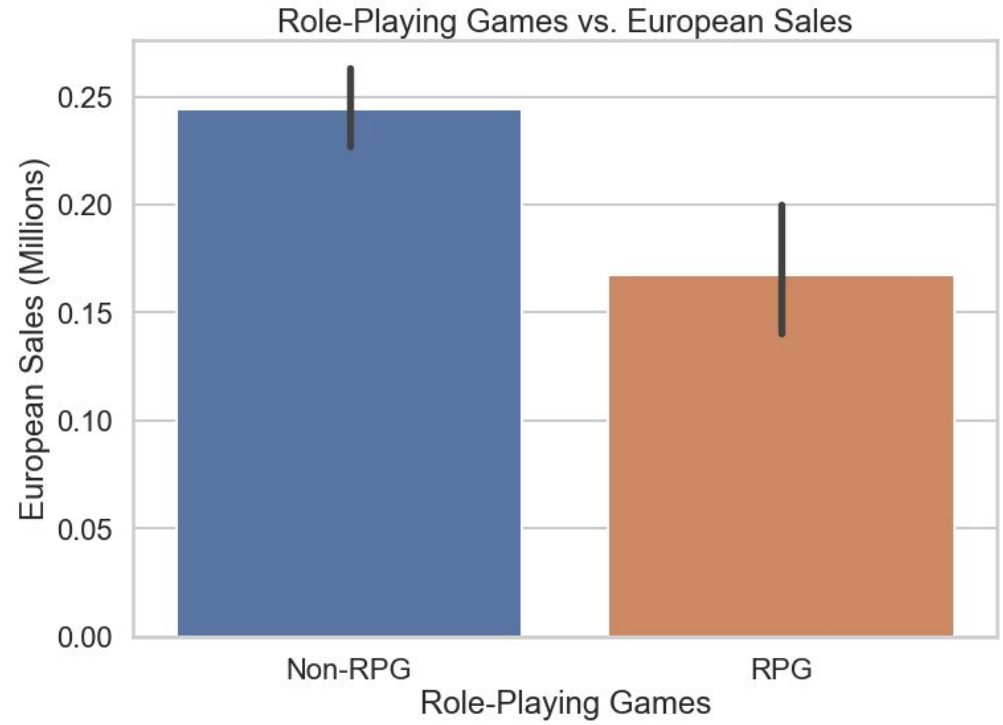
While there is a definite preference for Shooters in North America and Europe, there is a steep drop when it come to Japan. Only about 0.02 million in sales came from Shooting games

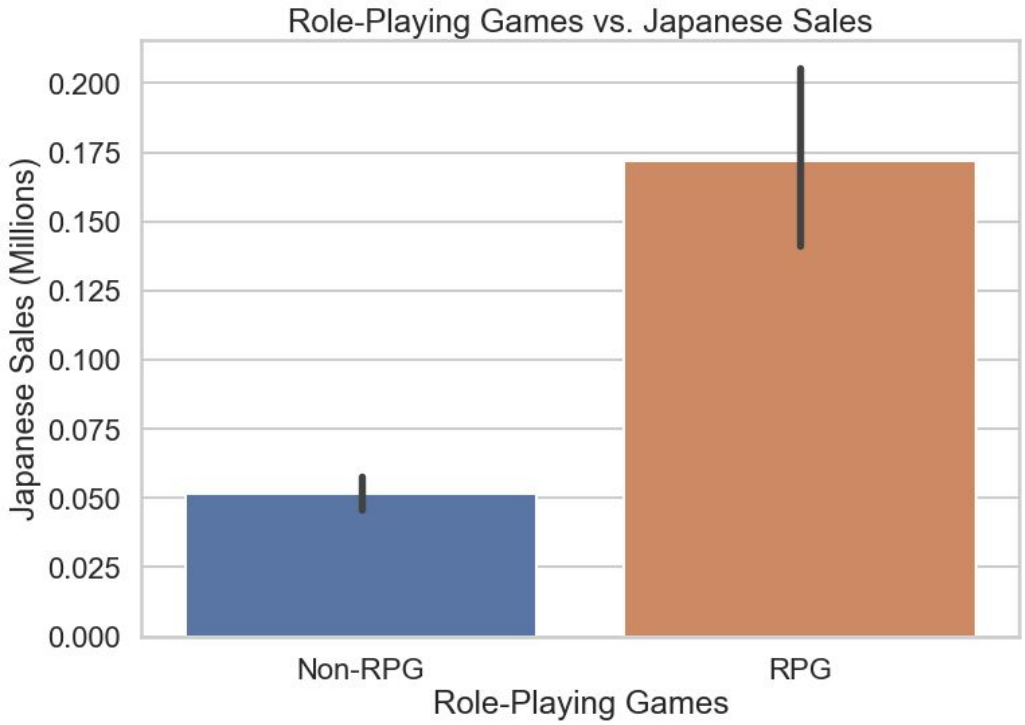
while North America and Europe had much higher amounts. Although Japan may have the smallest population of these regions, it still says a lot about why Shooters have not reached the level of Sports and Racing games yet. As big as they are in the U.S., Shooting games still have a ways to go in Japan.

Plot Group #5

Continuing to look at genres, I think it is important to consider the less popular categories and get an idea of why they may not be selling that well. Role-Playing games (or RPGs as they are known by) are near the end of the spectrum. Yet, major franchises like “Pokémon” and “Final Fantasy” are RPGs so clearly there is a market for them. The question is, where is it?





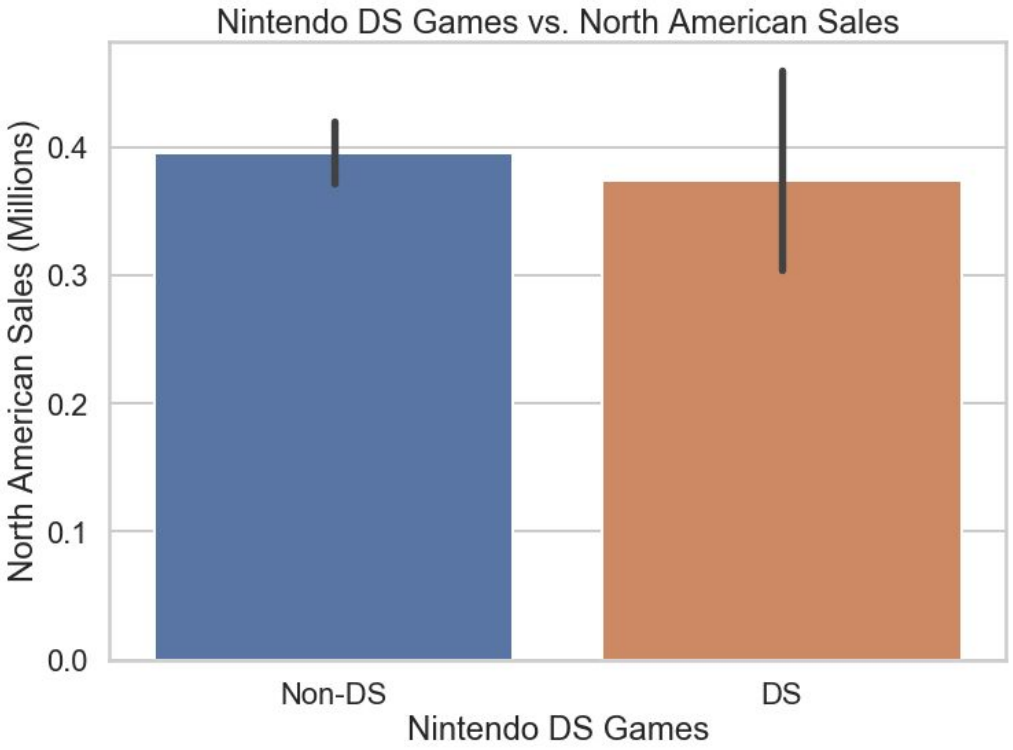


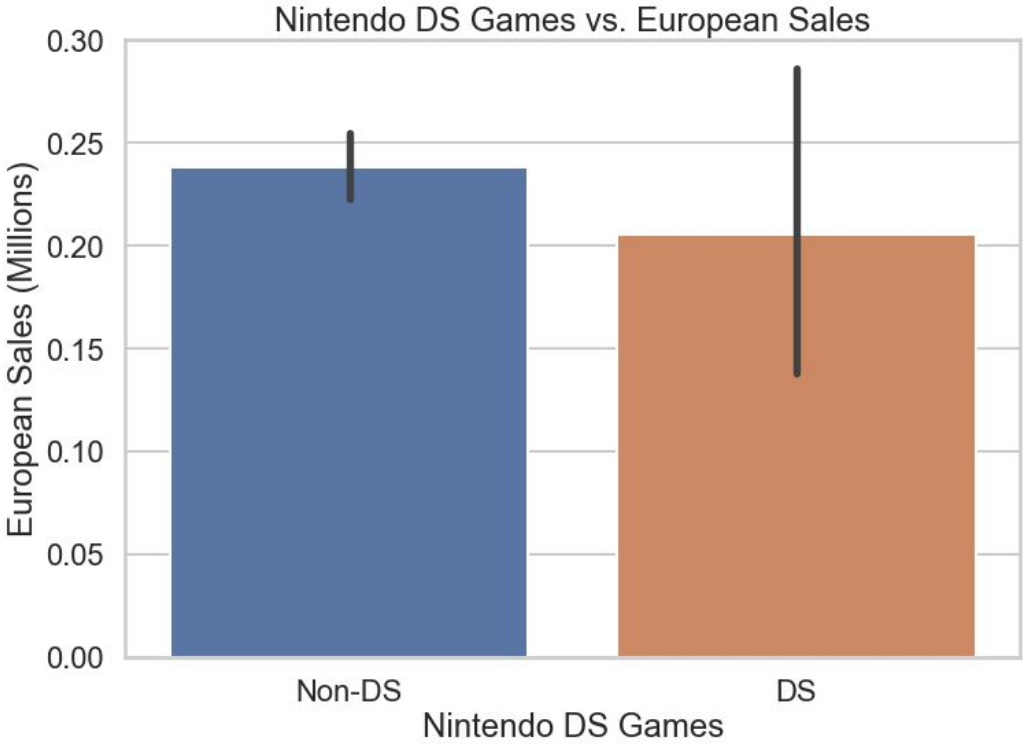
Unlike Shooters, RPGs are huge in Japan but not very big in North America and Europe. While that is not saying they are unpopular, it makes a little more sense for why it is lower on the list of popular genres. This speaks more to the point of the earlier plots that showed how much

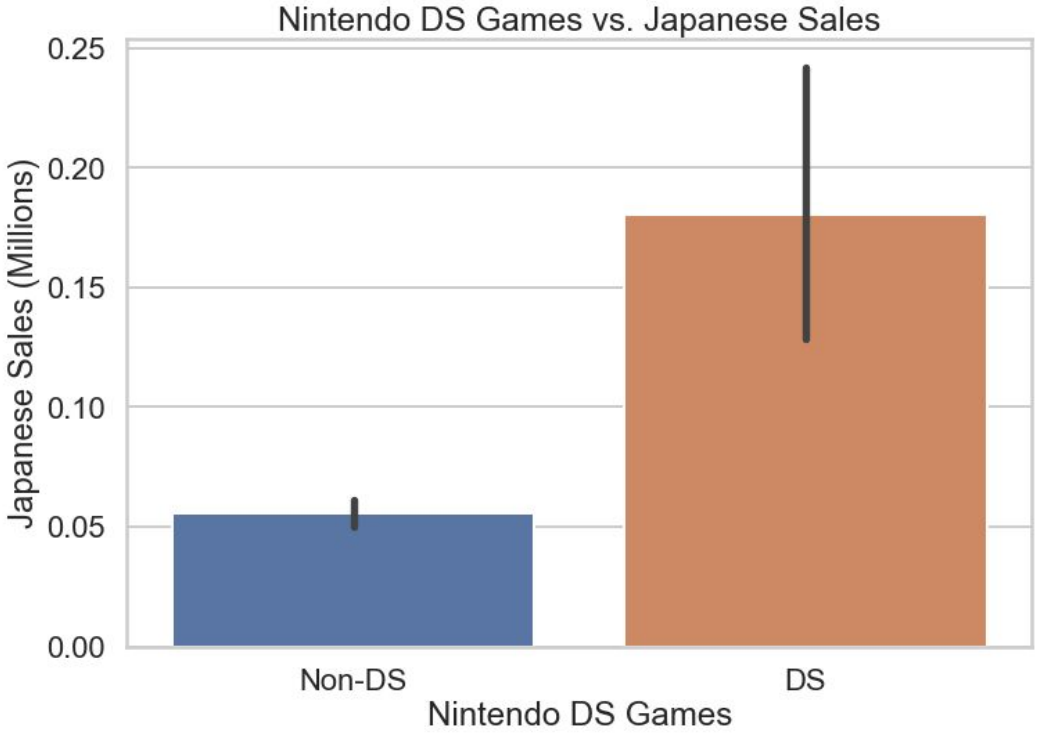
correlation there is between Global Sales with North American as well as European sales. Despite how popular RPGs are in Japan, that is not enough to make it one of the most popular genres worldwide.

Plot Group #6

Looking at the major consoles from before, one that may have surprised some is how high up Nintendo DS games were. While the Nintendo Wii made sense because of its mainstream appeal with motion controls, the DS is just a handheld system, which, while popular, normally do not do as well compared to home consoles. Also, the Nintendo DS was a generation before some of the other top platforms, while its follow-up, the Nintendo 3DS, was more in the middle. So what made Nintendo DS games place so highly on the plot?



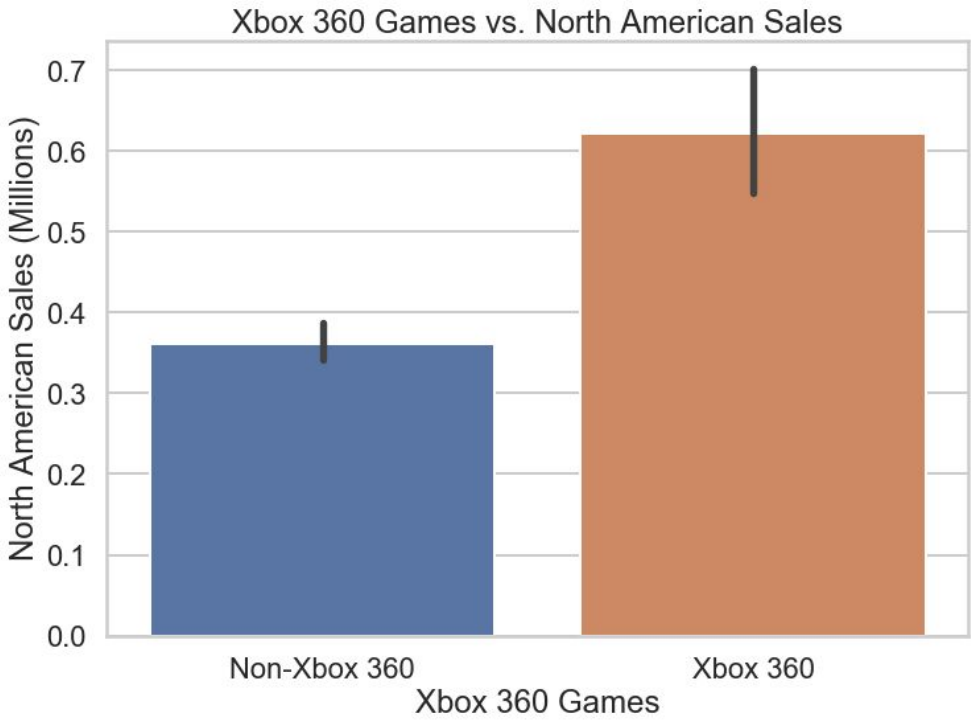

Nintendo DS Games vs. European Sales
European Sales (Millions)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Non-DS
DS
Nintendo DS Games

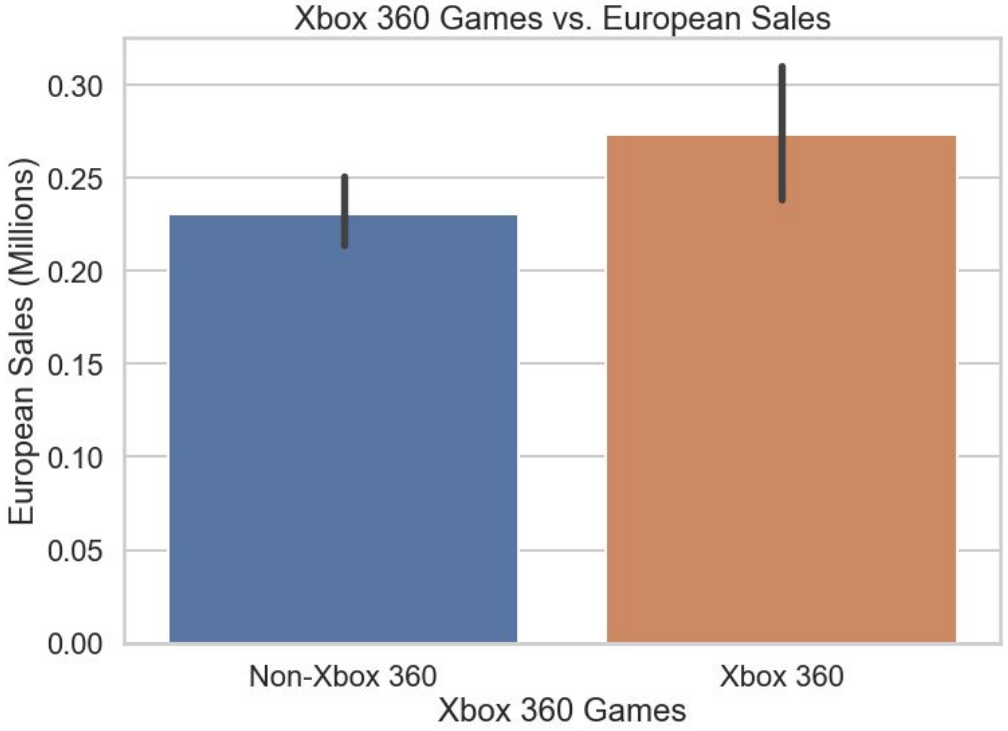

Nintendo DS Games vs. Japanese Sales
Japanese Sales (Millions)
0.25
0.20
0.15
0.10
0.05
0.00
Non-DS
DS
Nintendo DS Games

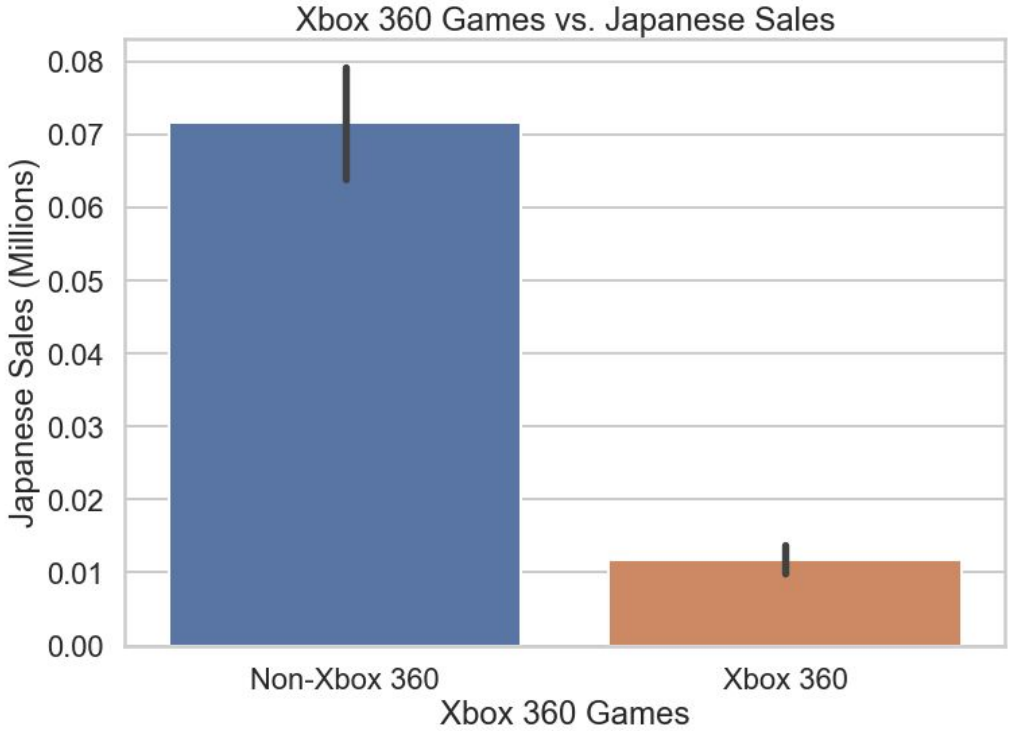
Despite many of the previous plots, the impact of Japan is finally seen with the sale of DS games. While the games would not be called unpopular in North America and Europe, they did not sell better than non-DS games yet Japan saw a pretty sizeable increase. Although its population may be the smallest of these three regions, it can still have a big impact on sales and the Nintendo DS is a great example why. Nintendo may be popular worldwide but Japan will always be one of their staunchest supporters.

Plot Group #7

Soon after the Nintendo DS in the console games spectrum is the Microsoft Xbox 360. Doing far better than its predecessor, the Xbox, this system rose in the last decade or so similar to the Shooting genre. And with Microsoft being an American company, it would make sense that it did well in North America. But while those games may have been popular here, it does not mean they were big everywhere.





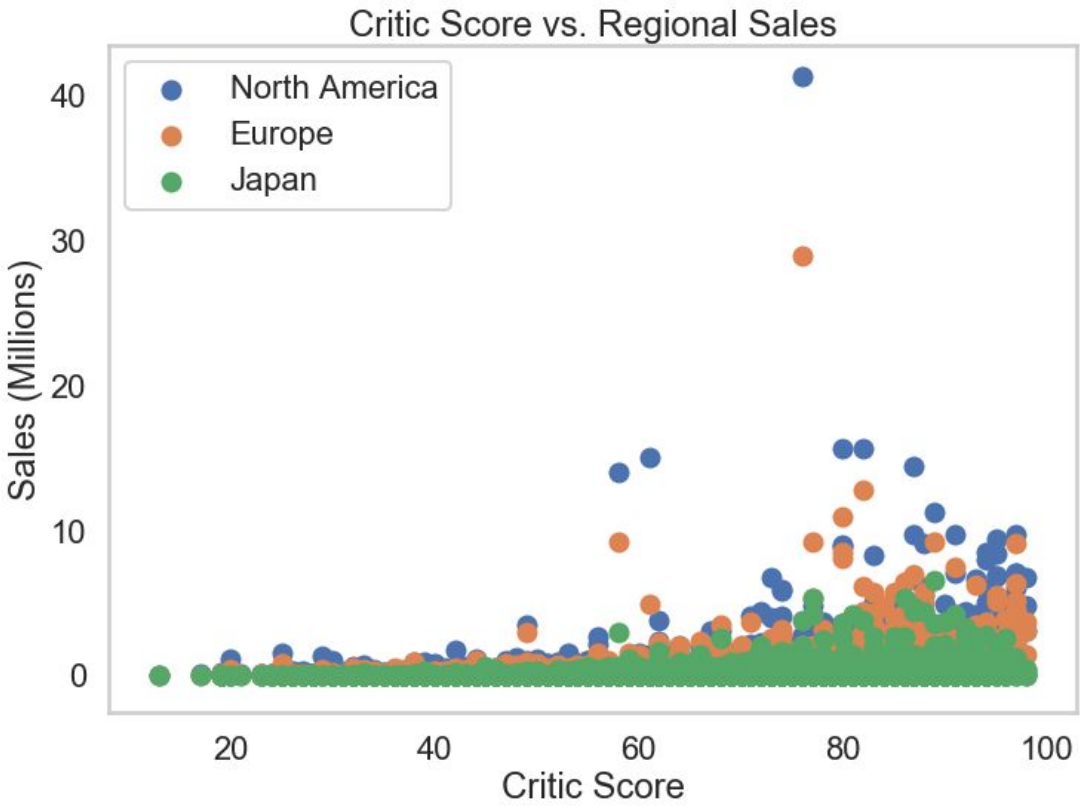


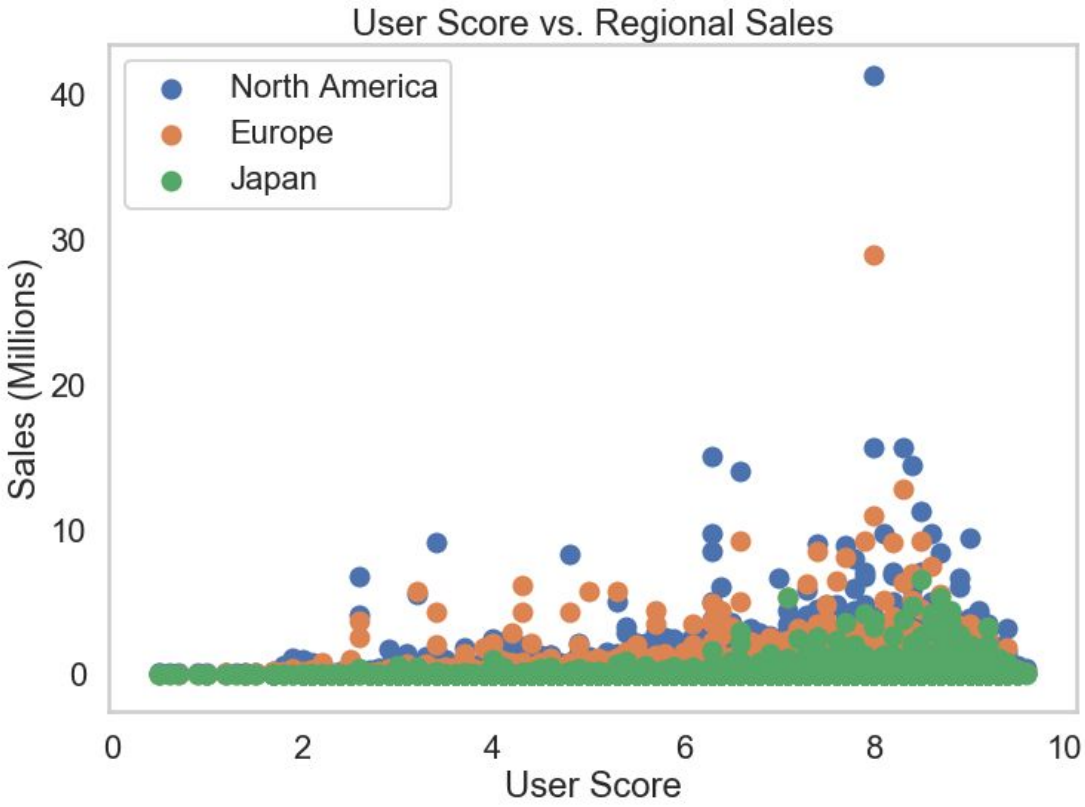
It is surprising to see just how unpopular Xbox 360 games are in Japan as there is not just a decline but instead more of a massive jump down. North America and Europe were similar

to what I would have expected but nothing prepared me for how poorly these games did in the East. Couple this with the previous plots about the Nintendo DS games and there is certainly pattern that describes the preferences of Japanese gamers. This speaks a lot to what Microsoft should do with the Xbox One and any upcoming consoles when it comes to Japan.

Plot Group #8

One of the final aspects I wanted to look into was how important Critic and User scores are to gamers. Personally, I do not focus too much on what other people say about a game unless I am already on the fence. If it is a game I know I want, it is unlikely that a review would shy me away from buying it. But some people do take these scores seriously as well as write user reviews so it does have an impact. The question is, how much?





For both Critic and User score, the trend is as expected with sales tending to be stronger for higher-rated games. There are still plenty of games sold with lesser scores, which may be games with brand names or pop culture ties. However, the games with the highest sales do not tend to be ones with perfect scores. In fact scores around 80% seem to have the best sales. Also, it is important to consider an inverse relationship as the more popular games may receive more total reviews, which leads to more varied scores. But, overall, these plots explain what is expected when it comes to Critic and User scores vs. sales.

Conclusion

What stands out the most from these plots is how important North America and Europe are to sales overall. They not only have a positive correlation with each other but also globally. Yet Japan should not be an afterthought. While its population is smaller than the other regions, it had a big impact on the sale of Nintendo DS games and could be critical for how much bigger Shooting games become in the future. And when it came to reviews, 80% looks to be the ideal score for any video game developer.

While there were plenty more plots I looked at, these stood out because of the important trends as described above. These regions have varied preferences especially when looking at Japan, but that does not mean these are set in stone. Throughout the history of video games, different genres always rise much like how Platformers were big in the '90s. The next step is

digging deeper into these plots with inferential statistics to find out which really are important factors in video game sales.

4. Inferential Statistical Analysis

For the inferential statistics in this project, I decided to break down essentially all of the important relationships in the data. To start, I ran t-tests for both the Genre and Platform of the video games against the Global as well as regional sales. Then, for the categorical data like Genre, Platform, Rating and Developer, I used a chi-squared test to determine their relationships with each other and also dug deeper in certain situations. Lastly, Pearson correlation tests were used to see the relationship between the different sales of games as well as their relationships with Critic and User scores. For all tests, the alpha was set to 0.05 but the null hypothesis was somewhat different for each. In t-tests, it was that the factor had no impact on sales. The null hypothesis for the chi-squared tests was that there was no association between the factors, which is similar to the correlation tests that stated there was no relationship between each factor. All together, there were many p-values that showed statistical significance and explained a lot about the sale of video games.

When it came to the t-tests for Genre vs. sales, most of them showed statistical significance for at least one type of sale. Racing and Platform games were the only ones that did not, which means they do not sell better or worse in any of the measured areas. One thing that stood out though was that while most of the significant areas had a positive relationship, Japan showed a negative relationship for most types of games. For example, Sports games sold better in North America but sold worse in Japan. Also, Shooters sold better in Global, North American and European sales but sold worse in Japanese Sales. Even Action games only showed significance in Japanese Sales, where they sold worse. The lone exception was Role-Playing games, which sold better in Japan while selling worse in North America and Europe.

For the other genres, Miscellaneous games sold better globally and in North America while Fighting games struggled in Global and European sales. Puzzle and Simulation games only showed statistical significance in North America, where they sold worse. Finally, Adventure and Strategy games had negative t-statistics globally and regionally, which explains why they are the weakest selling genres in the Data Story plot. These tests add credibility to the plots in explaining which genres do better in each area and how important they are to overall sales. Sports games do the best because they sell better in North America while Shooter games are stuck in the middle since they struggle in Japan. The bottom three (Role-Playing, Adventure and Strategy games) do worse in most areas so it makes sense why they are at the end of the spectrum.

The t-tests for Platform vs. sales had many results that showed statistical significance in all types of sales. Games for consoles like Xbox, PC and GameCube sold worse in all areas while the Wii and original PlayStation games sold better in each. The Xbox 360 games sold better in all areas except Japan, where it sold worse, while the Dreamcast games only did better in Japan and struggled in the other areas. Also, Japan's passion for Nintendo was exemplified as the DS and 3DS games only showed statistical significance there and each sold better.

Otherwise, PlayStation 3 games sold better globally as well as in North America and Europe while PSP and PlayStation Vita games sold worse; PlayStation 4 games did well in Global and European sales but struggled in Japanese Sales. Lastly, Xbox One games only showed statistical significance in Japan, where they sold worse, and Game Boy Advance games struggled in both areas they showed significance, Global and European sales.

Altogether these t-tests further illustrate what the Data Story plots displayed earlier. Nintendo does very well in Japan but can be hit-or-miss in other areas. Microsoft, on the other hand, consistently struggled in Japan with a lot of success elsewhere when it came to Xbox 360 games. Games for Sony's home consoles did well overall especially in Global and European sales. One thing that stood out to me, however, was how both the PlayStation 2 and Wii U games did not show statistical significance in any area. For the PS2, that makes some sense, as it was so successful overall that the games did not sell better or worse in any area. However, the Wii U is considered a disappointment for Nintendo, implying it sold worse but none of the areas showed any significance when it came to games sold. While this does not necessarily make the console a success, it does go to show that there is more to sales than just the overall numbers.

Moving to the chi-squared tests, all showed statistical significance. The associations between Genre, Platform, Rating and Developer were not random. This means that there is some statistical relationship between any two. To delve further, I compared some of the top genres, platforms and ratings with each other to see just how significant the relationships are. For Genre, I looked at Action, Shooter and Sports games while PlayStation 2, Nintendo Wii and Xbox 360 games were the focuses in Platform. Lastly, the key ratings I delved into were E (Everyone), T (Teen) and M (Mature), which are the standard types of ratings games receive. When these chi-squared tests were run again and only focused on the main three factors, they all still showed statistical significance. This means that even the top genres, platforms and ratings all have non-random associations with each other. Although sales may not have been a focus in these tests, they still explain the importance of other factors in the field.

Similar to the chi-squared tests, all of the Pearson correlation tests showed statistical significance with some showing stronger positive relationships than others. The relationships between Global Sales as well as North American and European sales were each over 0.9, which is very high. Its relationship with Japanese Sales was a little weaker, around 0.61. Between North America and Europe, the relationship was also very strong, close to 0.84 while their relationships with Japan were also weaker, around 0.47 and 0.52 respectively. These tests help support the earlier plots that showed the same strong relationships when it came to fraction of sales. North America and Europe have a strong connection with each other as well as with Global Sales.

For Critic and User scores, all were statistically significant but the positive relationships were much weaker. For Global Sales, the relationship with Critic Score was around 0.24 and User Score around 0.09. In North American Sales, they were around 0.23 and 0.09 while European Sales had relationships close to 0.21 and 0.06 respectively. Japanese Sales had similarly low relationships but had a lower one with Critic Score (around 0.15) and a higher one with User Score (around 0.13). While these scores do not have much of a predictive relationship between the factors, it does show that User scores are more important in Japan while Critic

scores are less critical. Although the positive relationships are not strong, there is still some significance for Critic and User scores when it comes to sales.

All of these statistical tests continued to stress the important trends from the Data Story. The driving force for sales in Genre tended to be North America while Japan saw many sell worse including Shooters. When it came to consoles, it became more clear how much Microsoft's systems struggled in Japanese game sales unlike Nintendo while Sony probably did the best overall. All of the categorical data showed statistical association between each other including when focused on the biggest types of genres, platforms and ratings. Lastly, the relationships between the different types of sales further exemplified what the plots from the Data Story illustrated with North American Sales having a strong relationship with Global and European sales while Japanese Sales had the weakest relationships. And while Critic and User scores did have statistically significant relationships with sales, they were not strongly correlated. When you put all of these tests together, they show just how deep the statistical relationships are when it comes to video game sales.

5. Machine Learning Analysis

For the machine learning portion of this project, I attempted multiple different models until I found one that worked best on the DataFrame that went through natural language processing. After trying out Lasso Regression, Support Vector Machines and more, Random Forest turned out to be the best overall. Since I was going to use that same model on each of the major types of sales (Global, North American, European and Japanese), one of the important preprocessing steps was removing the other types of sales when setting up the variables because they would have too big of an impact and take away from the importance of variables whose impact I wanted to observe. To dig deeper into this model, I also split each type of sales by their median and ran the model over the values equal to or above it. This would help when dealing with the different proportion of sales in each area. Lastly, new dataset from the University of Portsmouth with U.S. Sales info from 2004 to 2010 would be put through the same model (including sales equal to or above the median) to further test its accuracy as well as find out what other features could be important.

With each use of the model, the “X” value was set to most of the features except for things like types of sale, name, year of release and the categorical variables that either had been, or would be, turned into binary columns. The “y” value was set to whichever type of sale the model was meant to predict on. When it came to testing above the median, the data was filtered before the variables were created. Also, the data was split into training and testing sets with the test size being 20%. Each model used GridSearchCV with the parameters always having n_estimators from 10 to 100. When it came to parameters for the Random Forest Regressor, the random state was always set to 42 however, in certain situations, extra parameters were added to improve R^2 including min_samples_leaf of 3 and max_leaf_nodes of 1,000. For metrics in each, I got R^2 from the testing data for “X” and “y” and the Root Mean Squared Error by the way of the square root of the Mean Squared Error of the testing data of “y” as well as GridSearchCV predicting on the testing data of “X.”

When it came to Global Sales, I received an R^2 of about 0.43 and a Root Mean Squared Error around 2.72. While not too high, this would prove to be one of the better uses of his model especially since it used the extra parameters. For features, the most important ones were Critic and User scores plus User Count as well as words like “super,” “mario” and “mw” (likely referring to “Modern Warfare” from “Call of Duty”). The lesser features tended to be different developers and also certain consoles. Yet when it came to the biggest difference between test and prediction, User Count appeared to have a big impact with one score under 40 and another over 800. After the data was split by the median, R^2 actually went up to around 0.45 and the RMSE raised to around 2.85. Because this is a smaller subset of the data, the model actually did better without the extra parameters. The feature importance was very similar with User Count and Critic Score leading the way but also included the word “kart,” like in “Mario Kart,” higher than before. And while developers were still lower, certain ones like RockStar

North and Polyphony Digital were fairly high. Also, once again, User Count had a sizable impact on the extreme differences between test and prediction.

For North American Sales, R^2 was also around 0.43 but yielded a lower RMSE, around 1.33 and this was using the extra parameters like for overall Global Sales. The important features were essentially the same as for Global Sales except for the rise in the word “madden.” This helps supplement all of the exploratory data analysis from earlier that showed the correlation between Global and North American sales. And yet again, the extremes in User Count helped explain the differences between test and prediction. Yet, when North American Sales were split above the median, R^2 dropped drastically to around 0.16 with an RMSE around 1.51. Done without the extra parameters, the reason R^2 may have decreased so much is because of how small the subset became. Much like the previous models though, the same features were near the top and bottom of the spectrum though the word “halo” was higher this time around. Extreme values in User Count also had a big impact on the difference between test and prediction especially “Grand Theft Auto V,” which had over 3,000 users. Between the first two types of sales, it is becoming apparent how important Critic Score and User Count are to sales as well as the words of certain major game franchises like “Mario” and “Call of Duty.”

Things started to change when the model was used on European Sales. First off, the model for overall sales did better without the extra parameters with an R^2 around 0.38 and a Root Mean Squared Error around 0.93. This makes sense as it did have a very strong positive correlation with Global Sales but not quite as strong as North American Sales, which may be why R^2 was lower. Feature importance also differed a little bit with PC games as well as the words “fifa” and “engin” being higher though User Count led the way with a noticeable impact on the difference between test and prediction. After subsetting European Sales by the median, R^2 and the RMSE both dipped to around 0.33 and 0.85 respectively as well as did better with extra parameters. The most important feature ended up being word “kart” with the standard group of highly impactful features soon after. The reason “kart” may have become so important is because of how big Racing games are in the area including “Mario Kart.” And unlike the previous models, there was only one game in the most extreme difference between test and prediction, “Blur.” The reason may be because of how frequently the words “mario” and “kart” were used in its reviews.

To wrap up this dataset, Japanese Sales saw the biggest difference when a Random Forest model was used on it. Without the extra parameters, it received an R^2 around 0.37 and a RMSE around 0.46. But while the word “super” was the most important feature, Japanese Sales was the only one that had publishers or developers very high with Nintendo and SquareSoft having big impacts. For the other models, both types tended to be near the bottom of the spectrum. Much like what was shown in the plots from the exploratory data analysis, Nintendo is very popular in Japan. And this was further exemplified in the difference between test and prediction as a common trait for those extreme games was that Nintendo or SquareSoft did not make them. But this model gets even stranger when split above the median; everything stayed the exact same. And not just R^2 and the Root Mean Squared Error but feature importance and the difference between test and prediction did not change at all. This may be due to Japanese Sales being lower than the other types so the median could be close to, if not

exactly zero so nothing really changed with the model. Either way, Japanese Sales definitely tells a different story when run through the Random Forest model.

After using this model on the four major types of sales, it is interesting how R^2 overall got weaker in the different regions, signifying weaker models, while the RMSE also got lower, signifying stronger models. Also, despite what I may have thought in the exploratory data analysis, User Count and Critic Score had big impacts on these models. For the natural language processing, some of the words of major franchises tended to have big impacts including “Mario,” “Call of Duty,” “FIFA” as well as “Halo.” And except for Japan, the companies that made the games were not that important, nor were the genres or consoles. While this data was not used in the exploratory data analysis, it does say a lot about what features are truly important in the sales of video games when used with a predictive model.

To finish off this section, I used the exact same model and process on the data from University of Portsmouth on U.S. Sales from 2004 to 2010. Although this data does not have any natural language processing, it does include binary columns of when in the year a game was released. When the Random Forest model was used (with the extra parameters), it received an R^2 around 0.86 with a Root Mean Squared Error around 0.41; this is a very strong model especially compared to the previous ones. The features that were most important were the time of year the games came out especially Block 4, the holidays. This was further evident in the difference between test and prediction when none of the extreme games came out in any of the major blocks of the year. When the data was filtered around the median, both R^2 and RMSE increased to around 0.92 and 0.42 respectively with a very similar spectrum of feature importance. Although Block 4 was still the most important (by a lot), some of the others changed position including Block 2 and Block 0.5, which dropped, while Block 1 increased. Along with the drop in the importance of the Nintendo Wii console and its possible that many of the games below the median came out in Blocks 2 and 0.5 as well as were played on the Wii. And once again, the blocks of the year had a noticeable impact on the difference between test and prediction.

What this new data shows is how important the time of the year is for the sale of video games. Holiday time is as important as expected and, in general, these blocks of the year have a bigger impact than any genre or console. Couple this with the earlier dataset and it becomes more apparent about what is critical for video game sales. Name recognition for major series is important along with the amount of users who rate a game. Things like ESRB rating as well as publishers and developers are not as key (except in Japan). And while some games are hindered by rushed deadlines, when a game comes out during the year is very important when it comes to successful sales. Instead of focusing on the hot type of game, like an online Shooter, or a specific console, companies should look more at the most popular brands in the medium.

6. Conclusion

After all of the different steps throughout this project, I now have a clearer picture of what impacts video game sales. When looking at the different regions, it is apparent how much North America and Europe affect sales globally, but that does not mean Japan is unimportant. In fact, when it came to major genres and platforms, Japanese Sales had a major impact on where they fell in the spectrum. All one has to do is look at the placement of the Shooter genre as well as Nintendo consoles to see how important Japan is to video game sales.

But as I dug deeper into the machine learning portion of this venture, I got a better idea of what factors had the biggest impact on video game sales. Brand names for major franchises like “Mario,” “FIFA,” “Call of Duty,” “Halo” and more were some of the most critical features. While this may seem obvious to some, I think it illustrates how important it is to keep the name of the series in the title. Without that, less-informed gamers may not realize a certain game is part of one of their favorite franchises. While PC was a higher factor in certain models, in general genres and platforms were not that important for the predictive models.

What was most surprising to me, though, was how important Critic Score and User Count were. As mentioned earlier, I personally do not put too much stake into people’s thoughts on games unless it is one I am already on the fence about. Most games I want, I would buy regardless of what others thought. Yet, the opinions of professional reviewers tended to have a noticeable impact on most models signifying that it is an important factor. And when it came to the biggest differences between test and prediction, the amount of users reviewing the specific game on Metacritic appeared to have a sizable effect. This implies that it is important to have many gamers rate a game. While this does not mean they necessarily liked the game, this factor is illustrating that the more it is discussed the better.

Lastly, when using the University of Portsmouth data, not only was the model stronger but I also found how important it is to pick the right time of the year to release a game. The holidays are always expected to be key but that became much clearer when using a Random Forest model. While not every game will be guaranteed success based on when it comes out, this explains why companies want their games to be released at a certain point. Rushed deadlines can be a hindrance but there is a reason why games need to come out by a specific date. If they come out during the wrong block of the year, they can sell a lot less than if they came out during a stronger block. Just like with movies, there is a right and wrong time to release a video game.

Now there are ways to help improve this project, to go even deeper into this medium. For one, a more complete dataset may be helpful. With the natural language processed DataFrame, none of the models were that strong as R^2 never reached 0.5, which may be due to the amount of features. And while the University of Portsmouth dataset yielded very strong models, the features were different including the blocks of the year and lack of natural language processing. Also, because of the scarcity of information from games on older consoles, they were not used for this project. While it was the right decision to drop them, it could be interesting

to see what impacted their success especially with the rise of retro consoles being released recently. And although this project looked at the impact of consoles, it was always about the games released for the system. Although Xbox consoles may not be big in Japan, it would be useful to also have the overall platform sales as another factor to consider. Maybe the systems sold better than the games. Finally, while the natural language processing had a strong impact, it would be useful to have more reviews especially those used by critics. This project only used user reviews so there could be different keywords that reviewers use that have an impact on the market.

In the end, this project added great insight into the science of video game sales. The impacts and relationships with certain regions were explored as well as the preferences for each. Now a company may have a better idea for which type of game or console sell better in each of the major areas. And, when going even deeper, it became clear how important brand names and user reviews are along with when a game is released. As always, no matter how important any feature or aspect could be, there are plenty of factors that are hard to measure like quality of the game. The love people have for a franchise or the stories they cherish are also indefinable qualities. But, after going through all of the data and analysis, there is now a much greater understanding for what truly impacts the sale of video games.