Michael Hart
Information Visualization
Final Project

This is a dataset of all video games sold on the popular digital game distribution platform Steam. The dataset was put together from information made publicly available by Valve Software, and contains data on game release dates, text descriptions of the game content, genres, reviews, developers, publishers, user-generated reviews, average playtimes, and number of owners.

This dataset is limited to video game activity on Windows, Mac, and Linux systems, but provides much more fine-grain detail due to Steam's in-built analytics. The visualizations created from this dataset would be useful for sales and marketing teams in the video game industry, investors who want to know which game developers/publishers are the most successful, and video game shoppers who are looking for suggestions for their next purchase. I used Voyant for my text analysis and Tableau for my graphs.

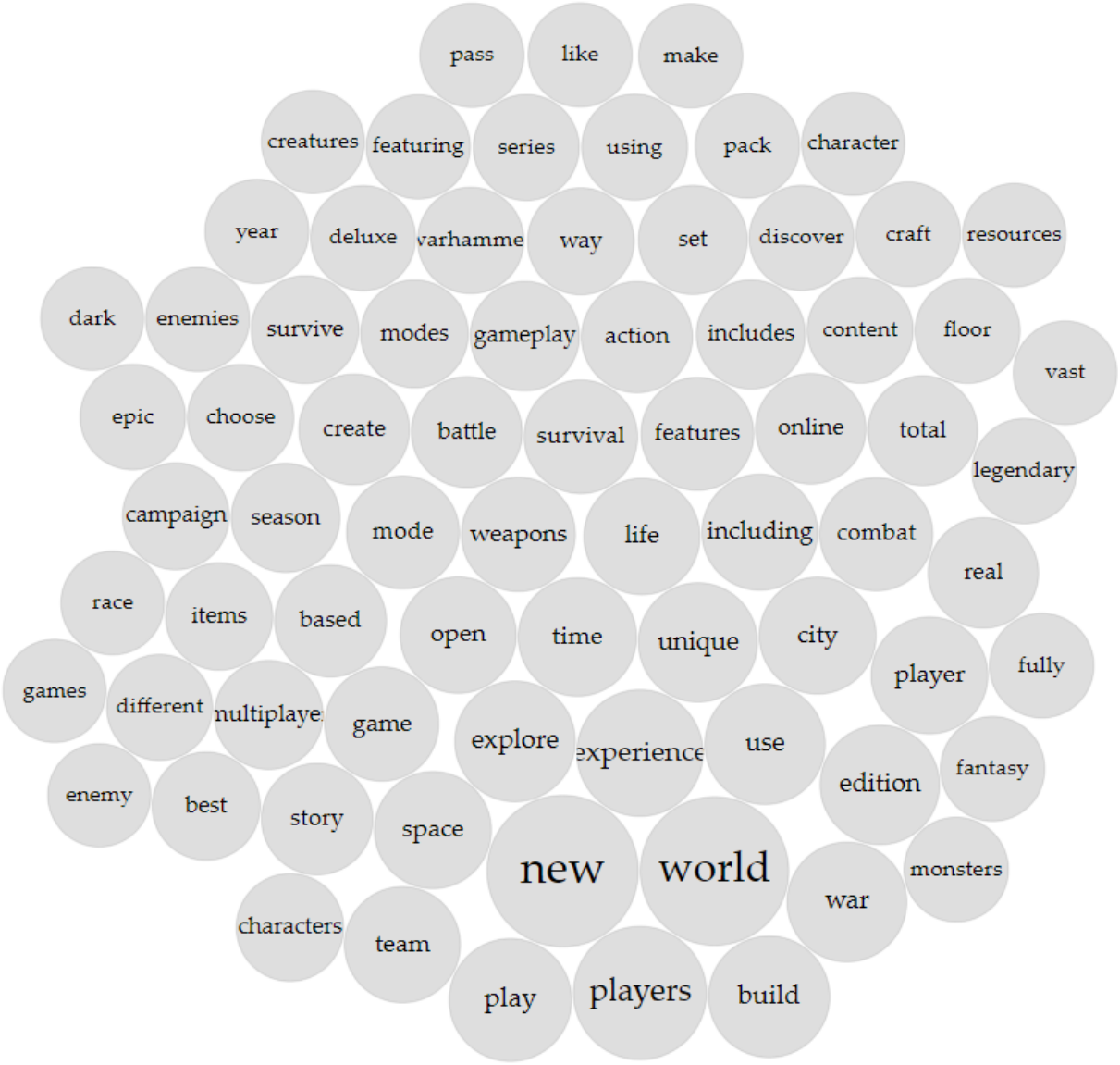
Questions

- Of Interest to Game Companies and Investors:
    - Which game genres are most profitable?
    - Is there a correlation between the percentage of positive/negative ratings and the number of purchases?
    - Do the most-purchased games generally have good reputations?
    - Which game developers and publishers have the best reputations?
    - Which game developers and publishers are the most successful?
    - Text Analysis:
        - What words are most commonly used the marketing descriptions of the top 50 games from the past four years?
        - Which terms are associated with each other in the text?
- Of Interest to Video Game Shoppers:
    - What games of a particular genre are popular and have mostly positive reviews?
    - Which games are the most addictive and time-consuming?
    - Is the average playtime for games distorted by a small number of obsessive gamers?

I will open my presentation with a simple word cloud of the marketing description for the top 50 most profitable games from the past four years. This will give my audience a broad sense of what words appeal to game shoppers.

The larger a word is, the more common it is in the text. Viewers should read the largest words first. I removed the word “game” since it is irrelevant and redundant.

The most common seven words are “new”, “world”, “players”, “experience”, “play”, “build”, and “explore”. This suggests that video game marketers want their games to be viewed as an exciting new adventure for game players.

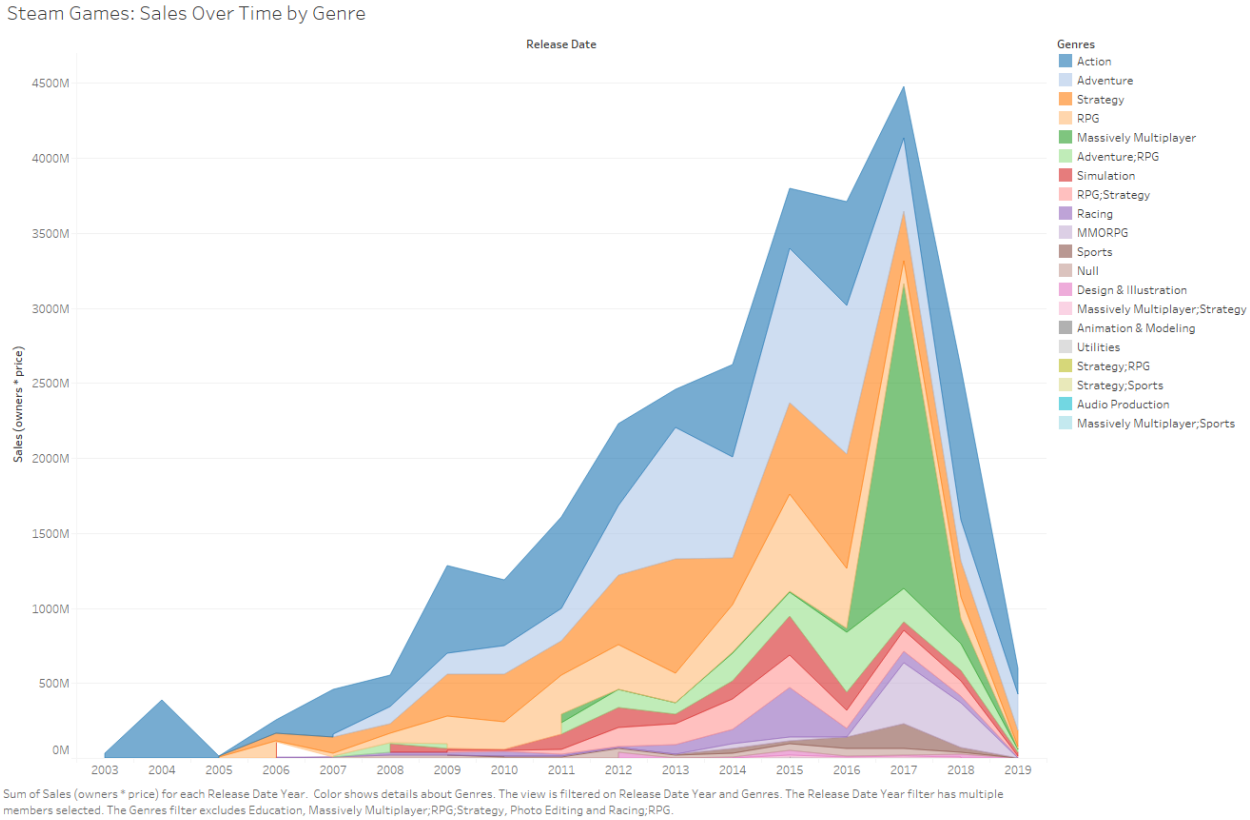


I chose to use a Voyant TermsBerry rather than a word network or word web because the interactive version (which I will show during my presentation) is easier to read.

The highest frequency terms appear in the middle of the visualization. The TermsBerry words display how many times the word appears in the text when the user hovers the mouse pointer over it. Other words that appear within two words to the right or left of the highlighted word will also appear, with that frequency listed. Viewers can thus figure out which words are often paired together in the video game descriptions.

When hovering the mouse over a given word, we can see trends in what kind of video games have been popular in the past four years. For example, hovering over the word “open” reveals that it appears before the word “world” 45 times. This suggests that “open world” games have been very popular recently. Also, the word “new” frequently appears before a wide variety of words, which shows that game marketing teams are striving to convince gamers that they are offering a fresh experience. The top associations for the word “new” are “new” and “world” (18 times), “new” and “features” (10 times), “new” and “including” (8 times), and “new” and “way”

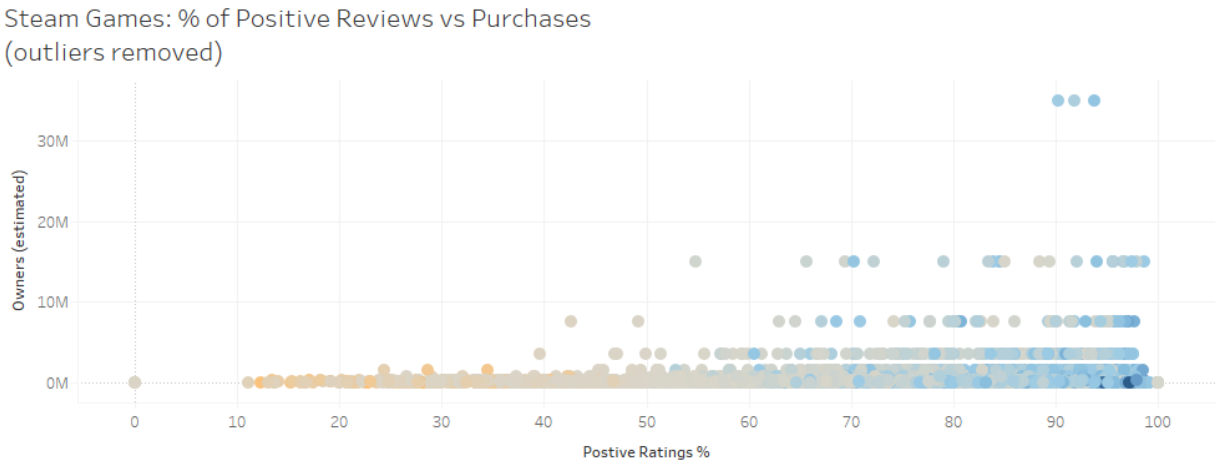
(7 times)". This TermsBerry would be extremely useful to consult while writing the marketing materials of a new game.

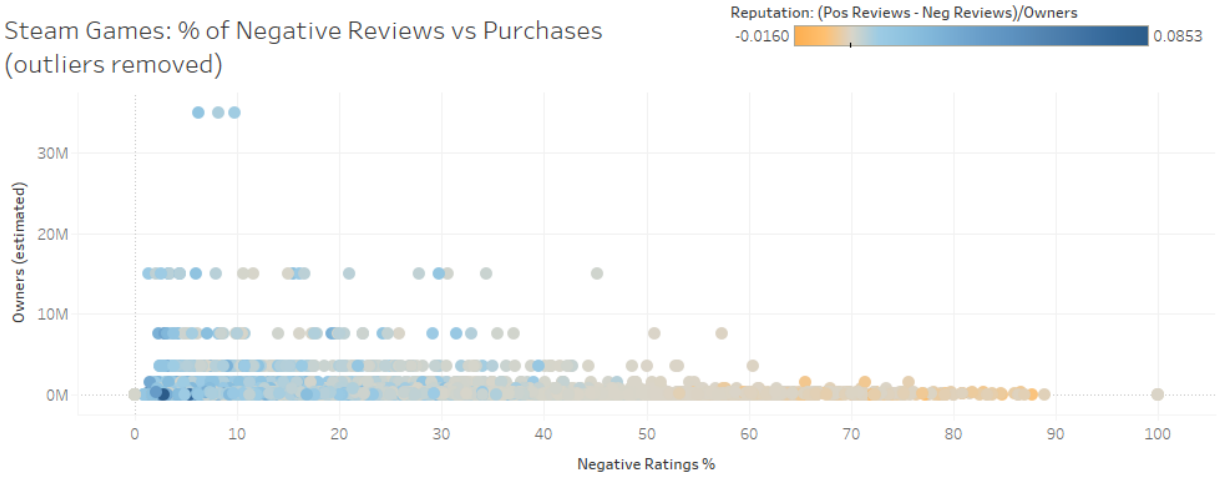


I chose a stacked area chart for the time series of game sales to show the trends in profitability for different genres of games. This image also serves to demonstrate that video game sales are steadily increasing every year (the sales only went down in 2019 because this dataset was created in mid-2019). Video games are an extremely profitable industry and could surpass movies in a few years.

Viewers can use the color legend on the right to see the relative sales of each genre from the years 2003 (when Steam launched) to 2019. The x-axis is the release year of the games, not the sales year of all games, which explains the downturn in 2018-2019.

Action, Adventure, Strategy, and RPG games have consistently been profitable game genres. Massively Multiplayer games released in 2016-2018 saw a spike in popularity, so investors and game creators should be aware of that trend. The amount of game sales will probably increase even more in 2020 due to coronavirus-related lockdowns worldwide.
This visualization also reveals that games released about 2 years prior to 2019 have the highest sales numbers, highlighting the need for companies to constantly produce new games. Game consumers, for the most part, are not buying classic games in large numbers.

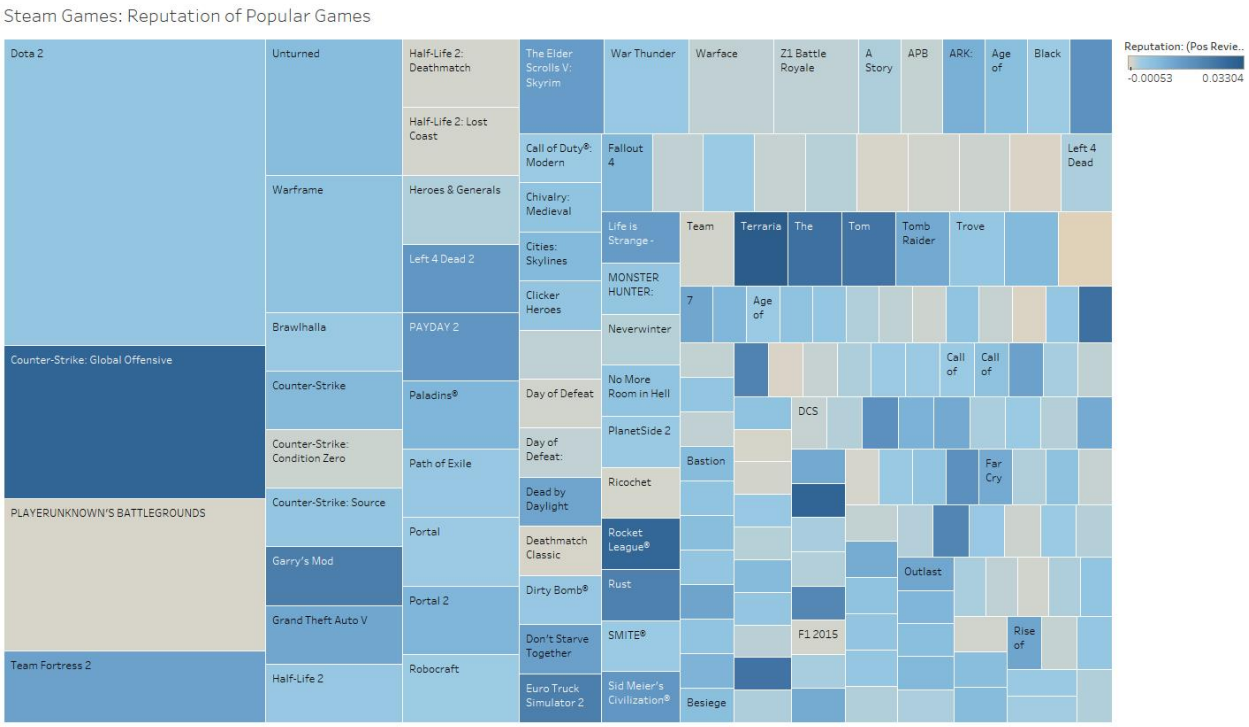




I made a quick scatterplot to see if the percentage of positive/negative ratings was correlated with the number of people who own games.

Read these like a typical scatterplot, and determine if there is positive or negative correlation between the X and Y variables. I removed four or five extreme outliers to make the graph easier to read. The outliers followed the general positive/negative correlation of the two graphs, but stretched the scatterplot out too much.

There is a positive correlation between positive reviews and ownership numbers, and a negative correlation between negative reviews and ownership numbers. It is difficult to tell from this data if negative reviews drive sales down, or if positive reviews increase sales. But these graphs should increase our trust in the Steam review system, as popular games generally have better reviews.

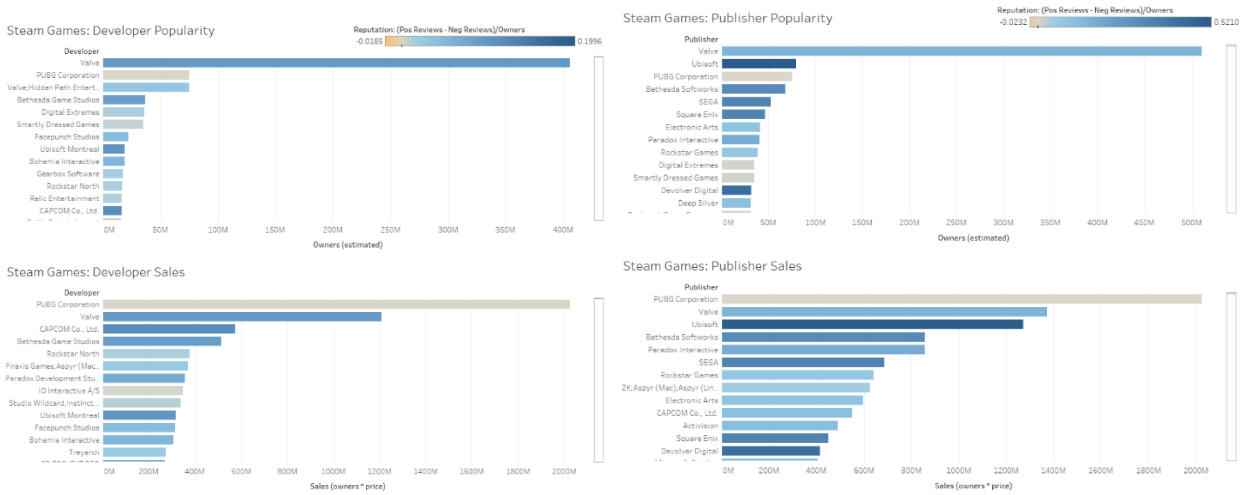


Game Name. Color shows sum of Reputation: (Pos Reviews - Neg Reviews)/Owners. Size shows sum of Owners (estimated). The marks are labeled by Game Name. The view is filtered on Game Name, which keeps 174 of 6,170 members.

I created a new metric "Reputation" calculated as ([Positive Ratings]-[Negative Ratings])/[Owners (estimated)] to get a clearer visualization of which games have mostly positive reviews.

This treemap has the rectangle size reflect the total number of owners of games, and the orange-to-blue gradient tracks the games' negative-to-positive reputation. If the viewer wants to see what games have sold many copies and have a good reputation, they should look for large blue rectangles. If they want to see what games have sold many copies but have a bad reputation, they should loop for the large beige or orange rectangles. If used interactively, the user could hover their mouse over the unlabeled rectangles for the game's name.

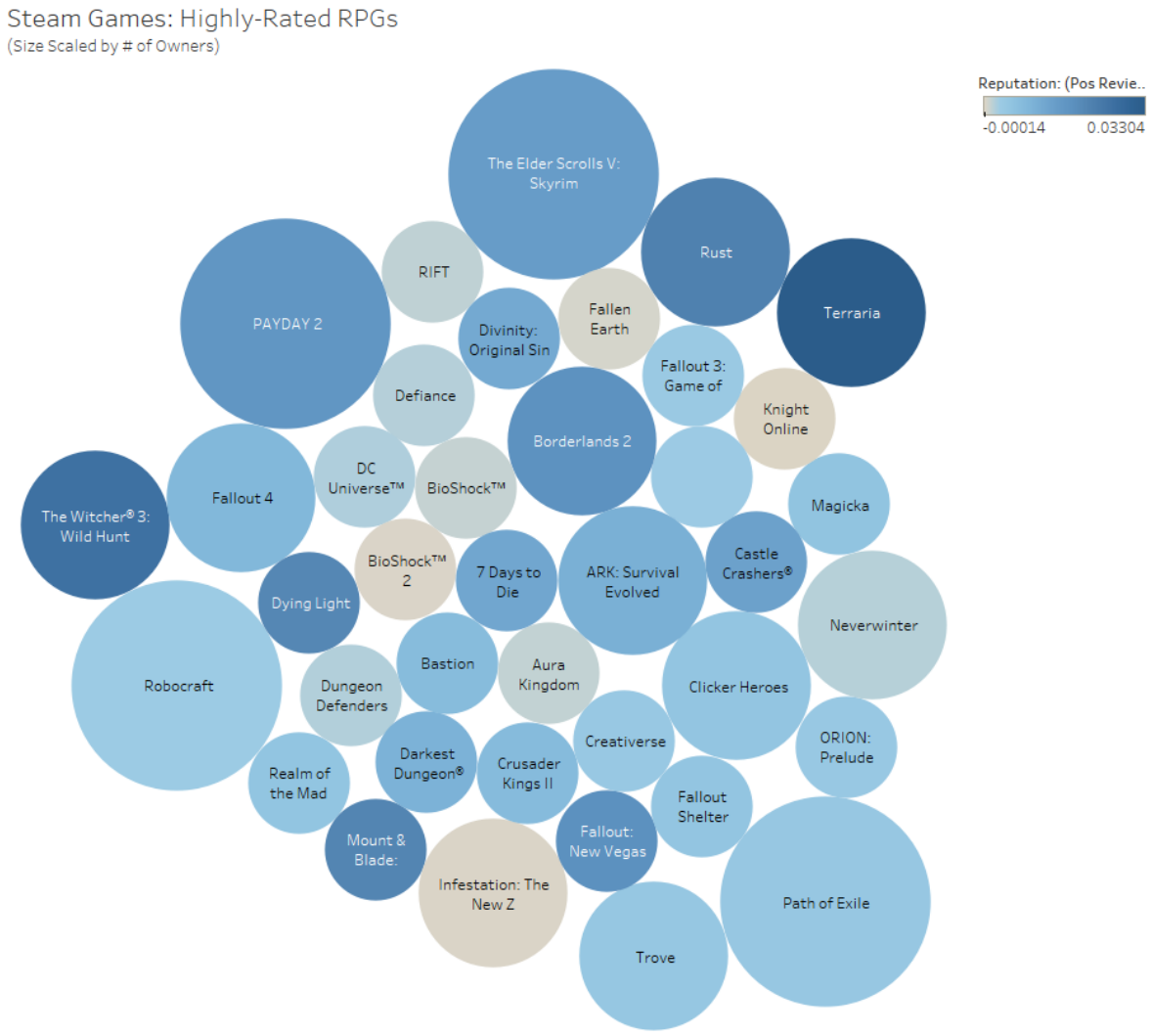
The treemap demonstrates that there are many games that sold a lot of copies but have negative or mixed reviews, such as PLAYERUNKOWN'S BATTLEGROUNDS, Counter-Strike: Condition Zero, and Half-Life 2: Deathmatch. In my opinion, this demonstrates that the Steam reviews are trustworthy, because customers are willing to give negative reviews regardless of the game's popularity.



I chose horizontal bar graphs so the names of the developers and publishers would be easier to read. I stacked the popularity graph on top of the sales graph to account for the fact that many games these days are “free-to-play”, allowing customers to download the game for free and optionally pay for additional features later on. Therefore, the “sales” graph might be a truer representation of a company’s success (although it doesn’t account for in-game purchases using real money). I also included my Reputation metric as the orange-blue color scale.

These can be read like a typical horizontal bar graph. The interactive Tableau version would allow the viewer to scroll down to see the more company names and bars.

The orange-blue Reputation metric reviews that financial success doesn’t guarantee a good reputation; PUBG Corporation has by far the highest sales figures, but has the worst reputation as both a developer and a publisher among the top 13 companies. If an investor were trying to decide which company has the best prospects, they might use this information to invest in companies with lower sales but a better reputation, since many customers will abandon a company after feeling betrayed by one bad experience with a game. Ubisoft, Valve, Capcom, Square Enix, Sega, and Bethesda Softworks all have very good sales numbers and solid blue Reputations, so a wise investor would probably favor them over PUBG Corporation, Rockstar North, or IO Interactive.

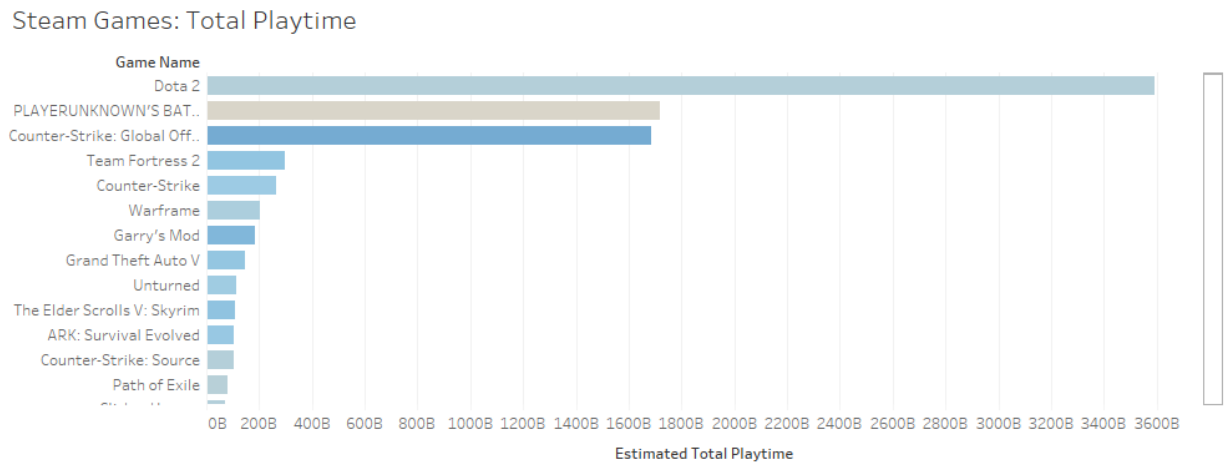


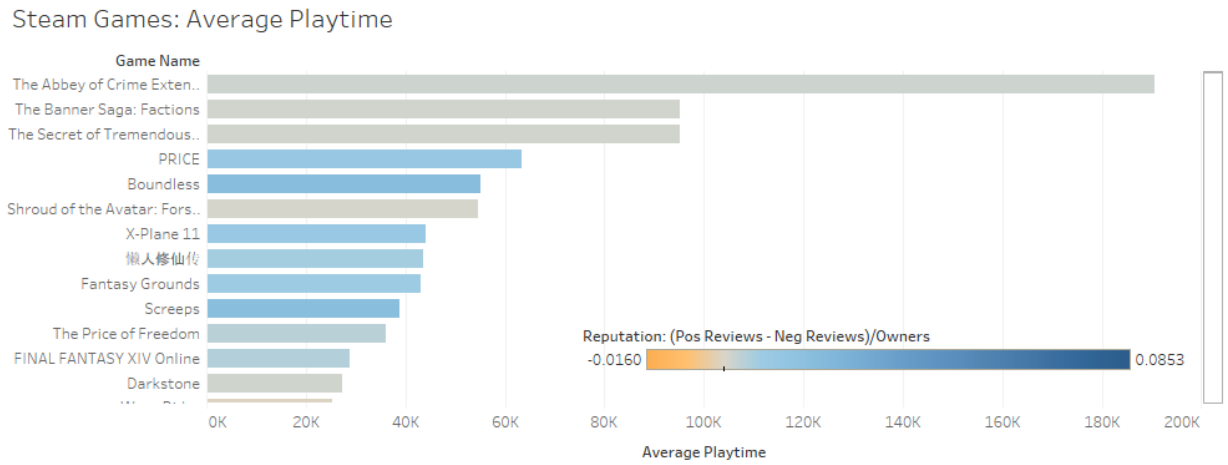
Game Name. Color shows sum of Reputation: (Pos Reviews - Neg Reviews)/Owners. Size shows sum of Owners (estimated). The marks are labeled by Game Name. Details are shown for Game Name. The data is filtered on Genres, which has multiple members selected. The view is filtered on Game Name, which has multiple members selected.

I chose this packed bubble visualization as an example of how this dataset could be used to help a video game shopper find a new game to play in their favorite genre. For me, that is RPGs, so I filtered for RPGs with high purchase numbers and shaded them for Reputation.

The larger the circle, the more copies have been sold. The bluer the circle, the better the reputation. Beige or orange circles have bad reputations.

This visualization provides suggestions for which RPG games a viewer should try, and which ones they should avoid. The size difference between the bubbles and the Reputation coloring also allows for quick comparisons of popularity and quality of the games.





Looking at the raw data of the average playtime and median playtime in Excel, I realized that the average playtime was probably being inflated by a small number of people playing the games for a very long time. I made these two horizontal bar graphs to see if there was a significant difference between which games have the highest average playtime vs the highest total playtime ([Average Playtime]*[Owners (estimated)]).

This should be read like a normal horizontal bar graph, with particular attention given to the difference in game names on the left. I again included my orange-blue Reputation metric.

The two lists are almost completely different. The games with the highest total playtime have been purchased by a large number of people and are played for many hours by those people. The games with the highest average playtime need not have been purchased by a large number of people; a small group of fans could pump up the average playtime. Therefore, if a game shopper is looking to purchase a game that the average person could enjoy playing for many hours, they should look at the games with the highest total playtime. The games with high average play time might only be enjoyable for a small subset of the population, so should be avoided unless the game player's preferences are well known.

The text analysis of the top 50 games from 2015 to 2019 revealed some trends that game companies should be aware of. The most common seven words are "new", "world", "players",

“experience”, “play”, “build”, and “explore”. The word “open” appears before the word “world” 45 times. This suggests that “open world” games have been very popular recently. Also, the word “new” frequently appears before a wide variety of words, which shows that game marketing teams are striving to convince gamers that they are offering a fresh experience. The top associations for the word “new” are “new” and “world” (18 times), “new” and “features” (10 times), “new” and “including” (8 times), and “new” and “way” (7 times)”.

My time series analysis of game sales revealed that Action, Adventure, Strategy, and RPG games have consistently been profitable genres. Massively Multiplayer games released in 2016-2018 saw a spike in popularity, so investors and game creators should also be aware of that trend. In general, the amount of online game sales will probably increase even more in 2020 due to coronavirus-related lockdowns worldwide, as people seek to maintain connections with friends. The visualization also revealed that games released about 2 years prior to 2019 have the highest sales numbers, highlighting the need for companies to constantly produce new games. Game consumers, for the most part, are not buying older games in large numbers.

We also found that high sales numbers don’t guarantee a good reputation or popularity for a company’s games. There are many games that sold a lot of copies but have negative or mixed reviews, such as PLAYERUNKOWN’S BATTLEGROUNDS, Counter-Strike: Condition Zero, and Half-Life 2: Deathmatch. In my opinion, this demonstrates that the Steam reviews are trustworthy, because customers are willing to give negative reviews despite games being successful or popular. Ubisoft, Valve, Capcom, Square Enix, Sega, and Bethesda Softworks all have very good sales numbers and solid blue Reputations, so a wise investor would probably favor them over PUBG Corporation, Rockstar North, or IO Interactive.

My visualizations would also be useful for game shoppers. If a gamer is interested in buying a new game in a particular genre, they could view somethings similar to my packed bubbles visualization to find a popular game that has a good reputation. Game shoppers who want to maximize their hours of game enjoyment on their next purchase can consult my visualization of total playtime vs average playtime for popular games; the games with higher total playtime among all players are more likely to be a safe purchase.

This analysis is limited in that it only handles data from games bought and played on the Steam app on Windows, Mac, or Linux. Game data played on handheld consoles or traditional TV consoles is therefore neglected, which is unfortunate because console video games are just as popular as computer games. Additionally, my visualizations using the playtime metrics are probably not totally reliable because many players leave them on in the background while doing other things, which would inflate the total playtime numbers. The data related to sales is also uncertain, because Steam regularly offers sales that range from 25% to 80% off, so my “owners multiplied by price” calculation probably does not reflect the actual amount of money that companies made from their game sales. Game production costs also vary drastically, so a more thorough exploration of game profitability would have to include how much money was spent creating and marketing each game.

If I were to expand upon the work with this dataset, I would like to take a look how the “Steamspy Tags” (player-generated labels for game genres) are correlated with sales and positive/negative reviews. It would be interesting to see if the user-generated genre tags significantly differed from how companies label their own games. Considering how knowledge of Google AdWords has become essential when doing any business online, I can imagine game companies will put even more effort into labeling and branding their games as the market

becomes ever-more saturated and competitive. Companies may even alter how they label their games to more closely match the user-generated labels.

It should also be possible to use this dataset in a machine learning project to predict which video game companies will be successful in the coming years based on their past performance. Such predictive models are already the norm in the world of stocks and finance. When gathering this data, I did come across one article where a data scientist attempted to predict which upcoming games would be hits or not based on past and current trends. In the music industry, popular songs have been fed into machine learning algorithms in an attempt to produce new hit songs automatically (although to my knowledge, no top 10 hits have been product yet); the same might be eventually possible for video games. A good starting point would be to take the text descriptions of the top 50 video games from the past four years, and use a machine learning algorithm to generate the text describing the next hit game. A game development studio could then use that text description as inspiration for their next game.

Sources:

Chavarria, I. (2017, August 2). Predicting video game hits with Machine Learning. Retrieved from https://towardsdatascience.com/predicting-hit-video-games-with-ml-1341bd9b86b0

Davis, N. (2019, May). Steam Stores Games (Clean dataset). Retrieved from https://www.kaggle.com/nikdavis/steam-store-games#steam_description_data.csv

Sinclair, S. & G. Rockwell. (2020). *Voyant Tools*. Retrieved May 3, 2020, from https://voyant-tools.org/?panels=collocatesgraph%2Creader%2Ctrends%2Cphrases%2Cbubblelines&corpus=c420fbb6646f6c425088cb4466ff2933