Daniel Chang

BAIS:3250

12/09/22

Project Report

Steam Game Ratings

1. Introduction

Steam is currently the most beloved video game platform in the world. When the purchase and storage of digital games were complicated in the past, Steam was the first game platform to be created and absorbed many customers with many game reserves. Steam provided convenient and interesting services to customers compared to other competitors: convenient UI & periodic discounts and events. Since many new games are coming onto Steam, I believe the features of current games in the Steam store will be a great indicator to pick valuable games and will lead to better business decision-making.

In this project, I analyze the steam game dataset from Kaggle and examine how to distribute and sort out better games from several features.

2. Data

This project uses two primary sources of data: Steam games complete dataset from Kaggle that contains games from Steam shop with detailed data, and number of 24-hour peak players data that are scrapped from SteamDB website.

- Kaggle dataset: https://www.kaggle.com/datasets/trolukovich/steam-games-complete-dataset
- SteamDB(24h_peak): https://steamdb.info/graph/

2.1 Steam games complete dataset

I downloaded “Steam games complete dataset” from Kaggle. It is in csv file format and includes 40,834 apps and bundles with 20 columns: developer, genre, price, etc. Since it is a huge dataset, data reduction was needed.

Columns that are not related to this project were dropped(url, des_snippet, developer, publisher, recent_reviews, game_description, mature_content, minimum_requirements, recommended_requirements, discount_price, languages). Any rows with NA value were also dropped. Steam contains lot of contents such as games, games bundles, DLC, etc. and I only left games. Some of names are crashed so I also removed it. There were a lot of genres, and game details. I only remained very first one since it represents the game. For the comparison and analysis, I replace text rating to points like below:

- Overwhelmingly Positive → 90 pts
- Very Positive → 80 pts
- Positive → 70 pts
- Mostly Positive → 60 pts
- Mixed → 50 pts
- Mostly Negative → 40 pts
- Negative → 30 pts
- very Negative → 20 pts
- Overwhelmingly Negative → 10 pts

Lastly, I dropped games that were made in 19XX because I would like this analysis to apply to current gaming product trends. Total process of reduction reduces dataset from 40k to 6,908. Data reduction and cleaning code for "Steam games complete dataset" is also attached on ICON

2.2 SteamDB: Number of 24 hours peak players by game

Since Steam blocks its link for web scrapping, instead, I used "power automate" app, recommended by Professor Colbert during project check-in. I extracted the names of games, and number of 24 hours peak players. This dataset has games, dlcs, and bundles but since I only left games for first dataset,

2.3 Integrate Steam Kaggle dataset and SteamDB dataset

I integrated two datasets with names of games. Some data that doesn't match and has NA in any columns are dropped. With this process, data is reduced and strengthened. The final dataset has 8 columns and 1,293 rows with no NA value.

Table 1 Description of Data

Column	Type	Description
name	text	Name of the game
rating	numeric	Rating of game in points
year	text	year the game was released
game_type	factor	Type of game
achievements	numeric	Number of achievements users can get

genre	factor	Genre of the game
price	numeric	Price of the game
24h_peak	numeric	Peak number of players of the game in 24 hours

3. Analysis Questions

The goal of this project is to examine the relationship between various features and game success(ratings). This will be helpful for decision-making when Steam receives a new game on their platform. Analysis questions include:

- Is there a significant relationship between game type and game rating?
- Does the release year have a significant impact on the ratings of the game? (New released games are always better than old games?)
- How does each feature (achievements, genre, price) impact the ratings of games?
- What type of games are receiving the most attention right now? (Filtering by 24-Hour Peak & compare with game rating)

<u>3.1 What I predict before analyzing with my own experience</u>

- Is there a significant relationship between game type and game rating?
    - Multi-player games will have higher ratings than single-player games because I personally think playing games with people or friends is much more fun than playing by myself.
- Does the release year have a significant impact on the ratings of the game? (New released games are always better than old games?)
    - New released games will have higher ratings than game that released in the past since new released games have better performances and gaming graphics which is important features for gamers
- How does each feature (achievements, genre, and price) impact the ratings of games?
    - Games that have action or survival tags will have higher ratings
    - More achievements mean more contents in the game so games that have large numbers of achievements will have higher ratings
    - I personally like playing action genre games so action genre games will have higher ratings
    - About the price, I can’t make any prediction. Lots of people might think expensive games will have higher ratings because higher price means more content and better graphics but I personally experienced great games with lower price.
- What type of games are receiving the most attention right now? (Filtering by 24-Hour Peak)

- Regardless of rating, the game that is receiving the most attention now will have a direct impact on the current game trend. So, it is important to analyze

3.2 Game rating by game type

Is there a significant relationship between game type and game rating? I created a summary table by using dplyr package that shows the game type, number, and the average rating of games that has same game type. Table 2 displays the resulting summary table.

Table 2 Game type summary

Rank	Game_type	Number	Average_Rating
1	Local Multi-Player	5	76
2	Single-Player	1,245	68.4
3	Online Multi-Player	43	66

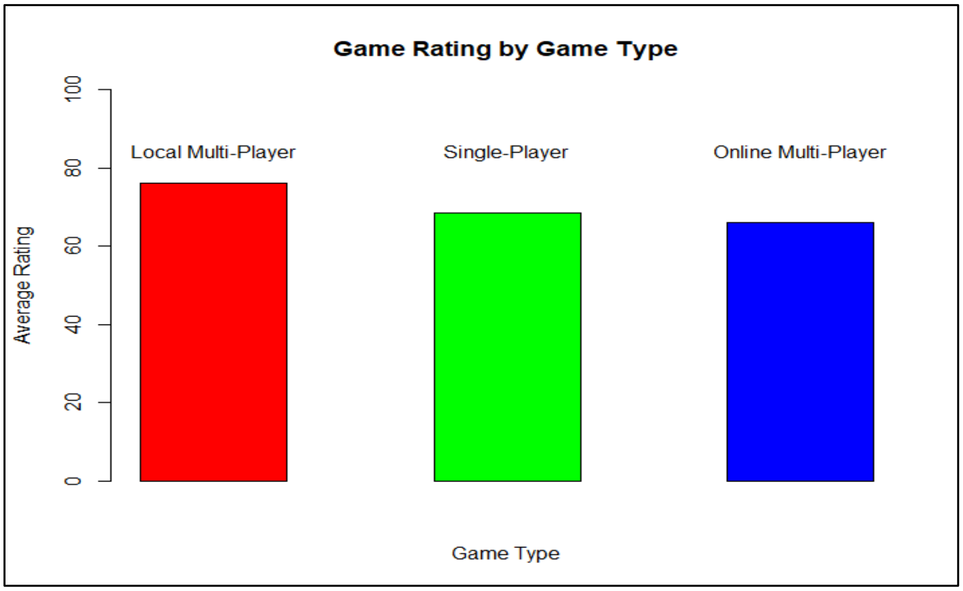

Figure 1 Bar graph of game rating by game type

The summary table and bar chart show that local multi-player games have the highest average rating. However, the dataset only contains 5 local multi-player games, and this is probably a result of small data size. I would like to focus on the single-player and online multi-player. Before I analyzed this dataset, I thought online multi-player games would have the highest average rating. There are still some differences in data number of two game type. However, as the table indicates, single-player games have a little bit higher rating than online multi-player. I still need to analyze games with more variables but from this analysis, I can say that Steam users give higher ratings to single player games than online multi-player games. I cannot guarantee the reason but, in my opinion, as one of steam users, one of the major reasons why players give low rating to game is the server problem and the single-player game doesn't need server for many people, and single-player games can be enjoyed without being bound by the number of players or other reasons such as hacker cheaters who cheat in game.

3.3 Game rating by release year

Does the release year have a significant impact on the ratings of the game? (New released games are always better than old games?). As one of Steam users, I always wonder if release year has a significant relationship with game rating. New games probably have better performance and better graphics, but old games might have their own contents and people might want to play original games. I created a summary table that shows the release year and the game rating below.

Table 3 Release year summary

Rank	*Release Year*	*Number*	*Average Rating*
1	*2004*	*2*	*90*
2	*2003*	*2*	*85*
3	*2005*	*2*	*85*
4	*2006*	*3*	*83.3*
5	*2009*	*11*	*78.2*
6	*2010*	*23*	*77.4*
7	*2011*	*37*	*77.3*
8	*2007*	*3*	*76.7*
9	*2012*	*48*	*75.4*
10	*2013*	*79*	*74.2*
11	*2008*	*7*	*72.9*
12	*2014*	*133*	*71.1*
13	*2019*	*74*	*70.4*
14	*2018*	*147*	*67.7*
15	*2016*	*293*	*66.7*
16	*2015*	*184*	*64.7*
17	*2017*	*245*	*64.4*

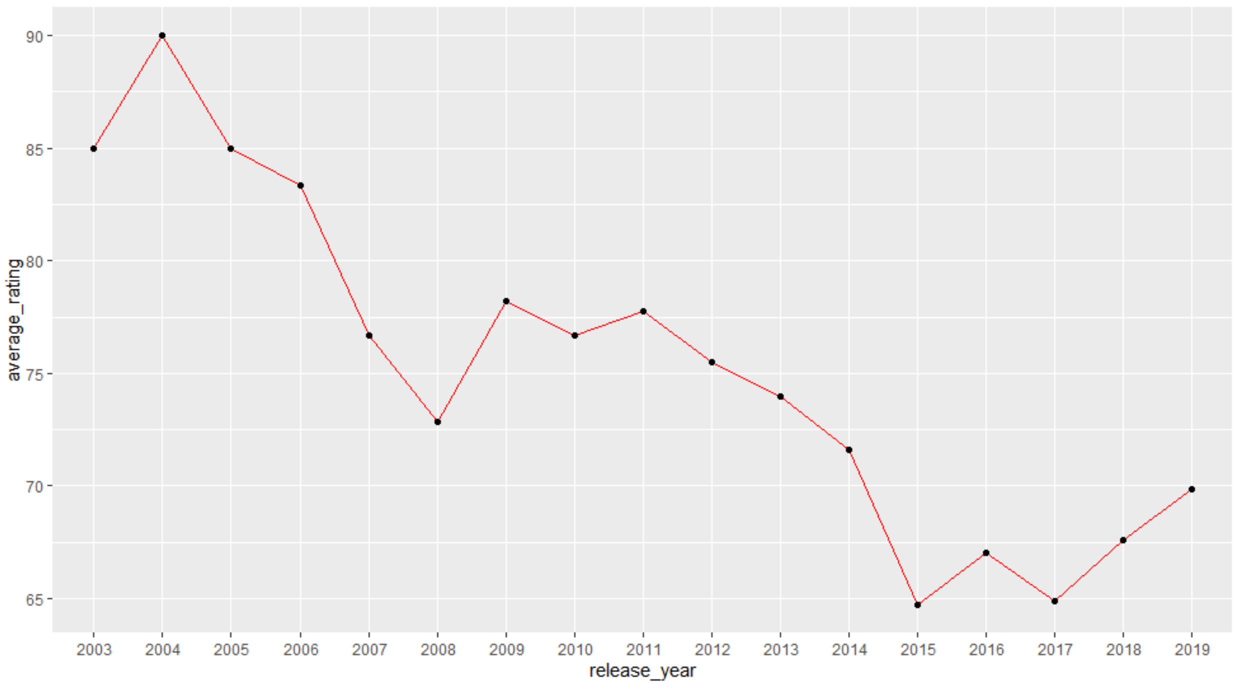

Figure 2 line graph of game rating by release year

According to the table 3 release year summary, the years 2003, 2004, and 2005 got into top 3 for highest average game ratings. However, just like the previous summary table, the amount of data for those three years is small. All of them have only 2 games concluded in dataset. Even though most of the games released after 2015 are ranked low but is still difficult to draw a conclusion that old games are better only with the summary table.

Figure 2 line graph helps a lot with drawing conclusions in this analysis. I created a line graph with ggplot2 package by average rating and release year. As you can see from the graph. The trend-line is decreasing from 2003 to 2014 which refers to those old games getting higher ratings compared to games that released in 2010 ~ 2014 and shows big drop from 2014 to 2015. However, From the 2015, the average ratings start to increase from 64.7 points to 69.9. In conclusion, old games from steam store got higher average ratings and the games released in 2010 ~ 2015 experienced a bad period. But from 2015, the game ratings have been increased rapidly and looking at the trend, and consideration of COVID19 period(more players at home playing games), it is possible that it has increased by 2022.

3.4 Game Rating by each feature (achievements, genre, and price)

3.4(a) Achievements

As I mentioned in previous, I believe more achievements means more content in the game and games with lots of achievements will have higher ratings. For this feature, the scatter plot can be helpful to draw a conclusion.

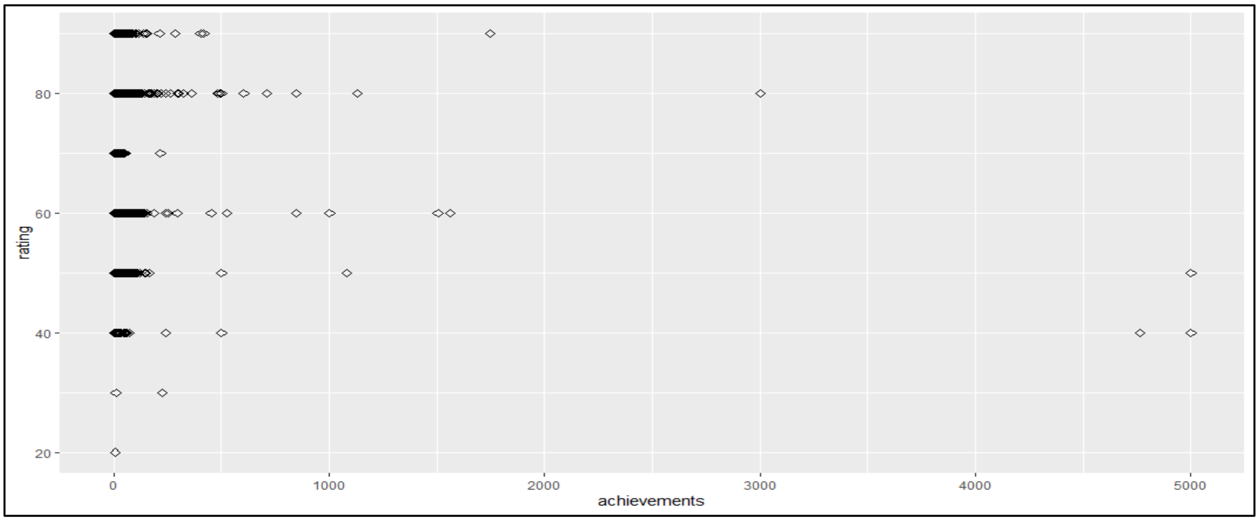

Figure 3 Scatter plot of game rating by number of achievements

As you see in Figure 3 scatter plot, lots of games has 0 ~ 500 achievements. This plot shows that number of achievements doesn't have significant relationship with game rating. Rather too many achievements decrease the rating(games with more than 4,000 achievements have 50 or lower points of ratings).

3.4(b) Genre

Genre might be the major feature when people look up games at the Steam store. In this dataset, there are 9 genres and below table 4 shows the average rating by genre.

Table 4 Genre summary

Rank	Genre	Number	Average rating
1	RPG	50	71.8
2	Indie	141	71.4
3	Adventure	199	70.4
4	Action	639	68.0
5	Casual	143	67.1
6	Strategy	34	62.9
7	Racing	15	62.7
8	Simulation	64	62.5

9	Sports	4	57.5

According to Table 4 Genre summary, RPG genre games got first place with 71.8 average rating and Indie, and adventure follow respectively. I can infer that Indie and RPG genre games have high ratings because most of those games are single-player games and we found out that people give higher ratings to single-player in previous analysis. Strategy, racing, simulation, and sports got low rank in ratings, and it shows that people are likely to play games that they can control the characters and likely to play more dynamic games.

3.4(c) Price

Price is also one of the major features when people decide to purchase a game. I used summary() function to see the statistics summary of game prices.

Table 5 Price summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.99	4.99	14.99	20.48	24.99	624.74

Since we dropped all free games, the lowest game price we can see is $0.99. The average game price is $20.48 which seems reasonable price, and the most expensive game in this dataset is $624.74. I checked the linear relationship between the price and rating, and I got -0.0411 which is negative relationship. This refers to the fact that increasing the game price is associated with lower game rating.

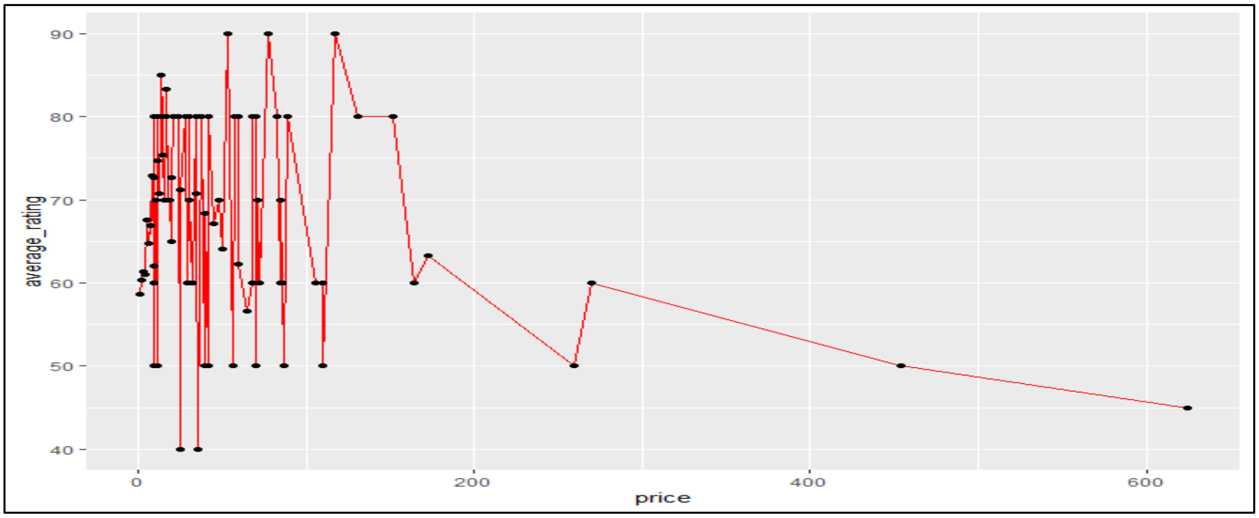

Figure 4 line graph of game rating by price

Figure 4 above supports the statement of rating and price having negative relationship. Between $0 and $100, the rating of games goes up and down repeatedly. However, it shows that after $100, the game rating drops rapidly while the price goes up.

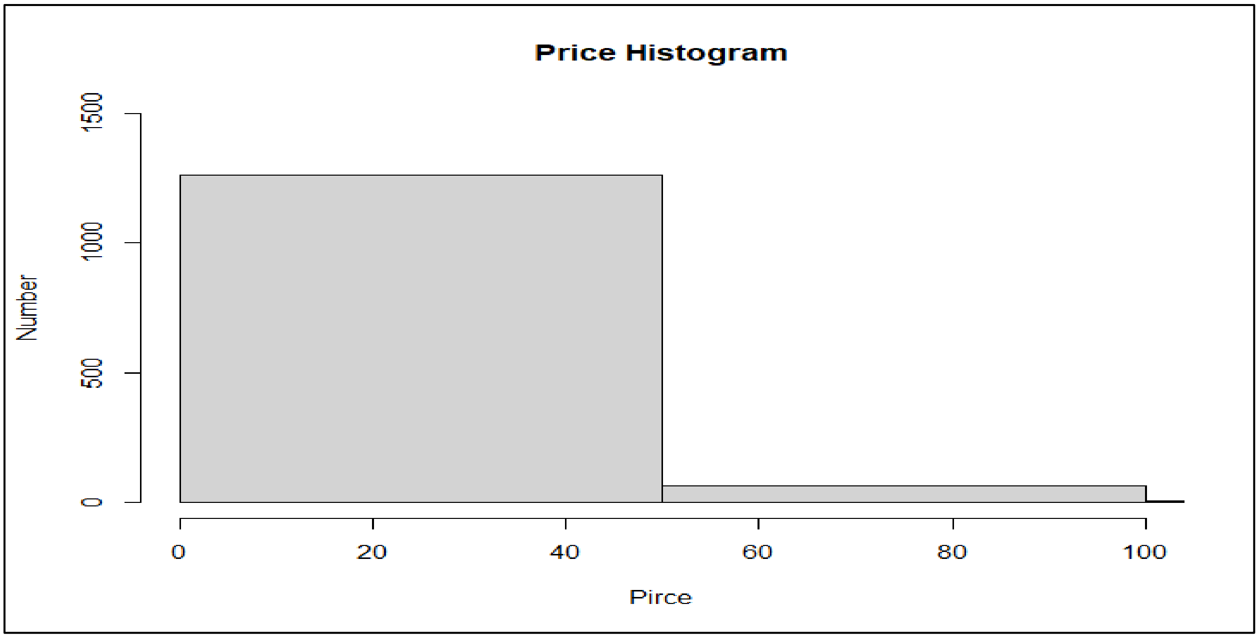

Figure 5 Histogram of number of games by price

With relationship test and line graph visualization, we can guess that many games might set prices between 0 and 100, or between 0 and 50. The steam dataset also shows that most games are in the range 0 to 100. And you can check the distribution of games by price with Figure 5 Histogram of number of games by price(dropped outliers).

3.5 24-hour peak

24-hour peak means the greatest number of players in 24 hours. Rating can be the feature to indicate game but also, 24-hour peak can be an indicator of the game trends. In this section, I will find out what type of games are played a lot by filtering games with top 10 24-hour peak. Some of the games that were released too long ago are forgotten and not played a lot, so they have low 24-hour peak, but I believe that it also gives us more information about current game trends. Before filtering games, I created a summary table for 24-hour peak.

Table 6 24-hour peak summary

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	4	34	950	174	124,996

As I just mentioned, there were some games that are not played these days because they were released too long time ago so the minimum shows 0. There are big differences between median and mean which are 34 and 950 respectively. This statistic shows that game users are more likely to be focused on only a few limited games, and lots of games are not played much or getting less attention.

Table 7 Top 10 Games by 24-hour peak

Rank	Name	Rating	Year	Game Type	Achi	Genre	Price	24-Hour
1	Grand Theft Auto V	60	2015	Single	77	Action	$29.99	124,996
2	Rust	80	2018	Multi	48	Action	$34.99	100,203
3	ARK: Survival Evolved	50	2017	Single	32	Action	$49.99	68,378
4	Warframe	80	2013	Single	187	Action	$19.99	53,161
5	Terraria	90	2011	Single	88	Action	$9.99	45,465
6	PAYDAY 2	80	2013	Single	1130	Action	$9.99	44,312
7	Left 4 Dead 2	90	2009	Single	70	Action	$9.99	40,462
8	Dead by Daylight	60	2016	Multi	128	Action	$19.99	36,836
9	Euro Truck Simulator 2	90	2012	Single	67	Indie	$19.99	34,644
10	Hearts of Iron IV	80	2016	Single	92	Simulation	$39.99	31,036

Table 7 shows the top 10 games by 24-hour peak, and their average rating is 77.8 which is between “Positive” and “Very Positive”. 8 out 10 games are distributed as single player, the top 8 games are action genre, and all the games are cheaper than $50. This table represents the results analyzed above and also the linear relationship between 24-hour peak and rating is positive(0.0646) which means more 24-hour peak number, rating goes up.

4. Conclusion

I faced several limitations while I was working on this project. First, Original steam dataset from Kaggle had more than 40k data, However, lots of them had NA value and some of texts were crashed so the data reduction was needed. In the process of integrating data from Kaggle and data from SteamDB, most games didn’t have a 24-hour peak. In the process of data reduction, many data were excluded from the original dataset and that leads to 1,293 instances of final data from 40k. Second, all games have more than one game type and genre, but I decided to keep the very first one which represents the game, and I think this might affect the analysis. For further analysis, I could possibly obtain data from other gaming platforms such as EA play, Epic games, or Battle.net so I can combine ratings and make it overall average ratings which can give me more accuracy ratings.

This project integrates Steam data from Kaggle and 24-hour peak data from SteamDB website to analyze the effect of each feature affecting Steam game ratings. I used three analysis questions: relationship between game type and game rating, relationship between the release year and the game rating, and how do other features(achievements, genre, price) impact the ratings. And lastly, regardless of game ratings, I used 24-hour peak data to evaluate what type of games are playing a lot and getting attention right now. With visualization, summary table, linear regression test, I was able to find out what type in each features impact the good ratings. About the game-type, single-player games got higher ratings than multi-player games, release year didn't really have significant relationship with game ratings, but we found out that newly released games are not always better. Rather, games released in the past showed higher ratings, but from 2015, ratings are gradually showing higher ratings and show positive trend line. About result of achievement feature, most of games are focused on between 0 and 1,000 achievements, but several games over 4,000 achievements got lower ratings which means a reasonable number of achievements is fine, but too many may result to low ratings. Several genres positively impacted the game ratings: RPG, Indie, Adventure, Action, and Casual, while simulation and sports got lower ratings. Lastly, about the price, it seems like there is no significant relationship between price and game ratings, because a lot of games are distributed between price $1 ~ $100 and ratings goes up and down repeatedly between price range $1 ~ 100, but from $150, I was able to find out that the game rating decreased. Based on this analysis, I can say that games with single-player type, number of achievements between 0 and 1000, positively impact genre(RPG, Indie, Adventure, Action, and Casual), and price range between $0 ~ $100 are likely to get higher ratings. In addition, comparing my final analysis and the 24-hour peak analysis, I was able to double-check my analysis and it can be seen that they have similar direction and result.