**1. Background Reading Introduction**

In this project I will try to answer some questions based on video game data collected from Steam (digital distribution platform) in 2016. Going forward, I decided to concentrate on those three questions:

1. “*Are there any associations that can be observed between games that received a Metacritic score and those that didn’t?*”
2. “*Is there an association between number of screenshots in game description and the sales of the game?*”
3. “*Is there an association between how many copies of the game were sold and the price of the game?*”

Also, one of my initial questions that I realized is basically impossible to assess:

“*Does relative success of a game means that most likely it’s localized in most popular*

*languages and are there any exceptions for this assumption?*”

I won’t be trying to answer it because success of a game might be subjective thing to different people. And even if I wanted to evaluate game’s success from a financial stand point, since I don’t have data on how much was spent by each developer and how many people worked on it to develop a game and for time etc. - I can’t calculate the profit that developer made. In addition to that, the data required for this is most likely confidential and can’t be found for many games. Thus, I decided to withhold from trying to answer this question.

**2. Topics of research:**

1. ***Metacritic score***

Since one of my questions deals with Metacritic score, I had to research how it works and basically dig any information that might be useful in my understanding of it, since it will aid in answering the question(s) later.

Official “About” and “FAQ” pages on Metacritic website turned out to be very useful in my research and cleared out most uncertainties I had about it [1]. Here are my findings:

First, Metacritic website covers virtually all new game releases in United States and other English-speaking territories, as long as they are reviewed by multiple publications (at least four).

There are two main scores on the website: Metacritic score (based on weighted calculation of scores from different publications/critics) and a User score (based on user reviews). User scores are not used in Metacritic score calculation and in contrast to publication/critic reviews, user reviews may be changed down the line (Metacritic only accepts first review from a publication to avoid situations when publication’s score changes down the line by means of potential external pressure on the publication). Also, some of publication/critic reviews might not provide a discrete value for their score, so it’s up to Metacritic staff to assign a numeric value to such reviews. Usually they work in increments of 10 (0-100), but sometimes might fall somewhere in between (for instance: 75).

Speaking of Metacritic score, weightings for it are kept in secret (User score is not weighted). And lastly, low score for Metacritic score doesn’t necessarily mean that game is that bad in terms of experience (but it certainly might); it means that most of publication/critic reviews were generally negative.

1. ***Steam and Steam Spy***

Steam is a digital distribution platform developed by Valve

Corporation for purchasing and playing video games [7]. The Steam

platform is the largest digital distribution platform for PC gaming,

estimated in 2013 to have 75% of the market space [8]. By 2017, users

purchasing games through Steam totaled roughly $4.3 billion,

representing at least 18% of global PC games sales [9]. In terms of

number of games, in March 2017, Steam had about 80-90% of all

games on PC platform (based on number of games in leading

distribution services on the time: around 14K games out of 16K

games on the platform were on Steam) [10].

Steam Spy is a website that uses an application programming interface

(API) to the Steam to estimate the number of sales of software titles

offered by the service. Estimates are made based on the API polling

using profiles from Steam to determine what software titles (primarily

video games) they own and using statistics to estimate overall sales.

Estimates for number of owners and players for specific titles (games)

were taken from this resource and added to the dataset that I’ll be

using.

1. ***Work of the original authors for the dataset I’ll be using***

Authors of the dataset I’ll be using did a data analysis of their own.

During their research, they proved that free games do receive more

recommendations on average, compared to non-free games. But free

games do score lower on average (Metacritic score) than non-free

games. Another prediction that they made was that the most

recommended and the highest rated genre is action, but they found

out that the most recommended genre was free to play and, for highest

rated, it was sports instead of action. Final prediction of theirs was that

Metacritic scores are an inverse bell curve when sorted by

recommendation, i.e. lower and higher scoring games would have

more recommendations that games with a middle score. While not

exactly related to their original prediction, they found that pricing

compared to Metacritic scores is mostly uniform and that pricing

compared to user recommendations is also nearly uniform. Those two

facts don’t support the idea of an inverse bell curve, thus most likely

it’s not. In my research I would like to investigate Metacritic score

from perspective of which games received this score and which didn’t

to see if there are any commonalities/associations that could be

observed.

1. ***Finding associations/correlations in game data***

Most sources I found didn’t state any clear questions that they wanted to be answered and were more like articles than scientific papers. Only one paper [2] used regression and some other Data Mining techniques to find correlation between certain game characteristics (genre, tags, price, etc.) and game sales. Author determined that having some successful games behind your belt does help you to succeed with your new game. Also, author determined a set of parameters that influence the sales of games the most (some of them: genres, length of description, launch price, age requirements, presence of multi-player and others). Other sources used more generic analysis techniques, such as comparing averages of different categories [3] [4]. One of the articles I stumbled upon [5], while didn’t really do much in terms of analysis of video games data, provided and interesting outlook on why someone might be interested in doing their own research on video games if they are planning to get into game development or are already one (basically to have a higher chance of success).

**3. Background Reading Conclusion**

The most useful and comprehensive information was found for Metacritic score, which will certainly help me with answering my question. Also, while I found out that a game receives a Metacritic score when it’s been reviewed by at least four critics/publishers it keeps my question of trying to find out if there are some associations to be found in terms of how different game parameters affect game having a Metacritic score or not.

In terms of research in this area, there is quite a bit of research more based on implications of gaming in general and more sophisticated studies on specific aspects of the games. Couldn’t find much in terms of good studies on how different parameters of the video games correlate/associate with each other. Most of articles that I found did a more rudimentary analysis – comparing averages of certain categories or just analyzing distributions of certain parameters. One of the articles used regression for finding correlation between variables, it might be a good fit for some of my questions.

Finally, for original analysis on this data set, authors did have a similar style of question to my question (2) and (3). It was about association between whether a game is free or not and number of recommendations. Since authors didn’t asses the question I want to answer and I haven’t seen my questions answered from other works on this subject, I’ll keep my questions unchanged.

**4. Methodology**

***Data Collection:***

For this project I’m using a preexisting dataset [6]. Most of the data was collected from Steam via Steam API and some extra data (for instance owner count and players estimate) was acquired from Steam Spy (service that collects the kind of data mentioned before from games on Steam). Also, Metacritic score for games was included in that dataset. For Steam data, it consists of all listings (mostly games, but also some non-game software) from Steam dating December 12 of 2016.

This dataset most likely is a rather big (around 70-80% of population) non-random sample of games on PC platform (excluding games that can be emulated and games that don’t run any more on recent at the time versions of operating systems). This estimate was made from comparison of number of games on Steam and some other popular game-distribution services that usually don’t have many games in common (for instance GOG.com “remakes” many old games to make them compatible with modern systems, or Uplay releases only Ubisoft (game publisher) games) [10]. In this comparison around 14,000 games in 2017 was on Steam and around 2,000 on GOG and Uplay (numbers were taken from [10]). Finding specific numbers proved to be almost an impossible task, thus I’ll have to do with these results.

While Steam games most likely represent most games in the population, there are distribution services with more unique games (like old games from GOG.com and fan-made or not officially published games) that sway me to think that I should not generalize my result to all PC games. Because of this I’ll generalize any results I find only to games on Steam, which will still be enough in my opinion, since Steam is the most popular digital game-distribution platform and probably, while can’t represent all games on the platform, most likely represents the platform that most PC gamers use, which makes it relevant.

Also, speaking of data, original authors [6] indicated that there are quite a few games that nobody bought and thus never played on Steam – this might result in big number of outliers or other obstacles and I’ll need to take care of it. If that will be the case, I’ll try to linearize the relationships between the variable by either taking the log of the data or taking the square root of it and see which works best. Another thing that I already noticed – the dataset contains some non-game titles, they will need to be removed before I move onto the data analysis.

***Data Analysis:***

- “*Are there any associations that can be observed between games that received a Metacritic score and those that didn’t?*” For this question I’ll probably compare the averages for different features (those that are numeric) in two different groups – those that received a Metacritic score and those that didn’t. Also, might look at the distribution of features for each of the groups, will most likely provide some more insight into this question.

For my next questions, I decided to focus on descriptive and visual measure (game screenshots) of the games since those are usually are a part of advertising for the game and thus will probably have some effect on game sales.

- “*Is there an association between number of screenshots in game description and the sales of the game?”* For this one, most likely I’ll try to use some sort of regression analysis, to determine if there is any association.

- “*Is there an association between how many copies of the game were sold and the price of the game?*”, for this one, also, regression seems like the most appropriate choice from what we learned so far in DwD1.

***Data Visualization:*** All the relevant visualizations will be done in R.

**Works Cited**

[1] About Metacritic score, <https://www.metacritic.com/about-metascores>

FAQ, <https://www.metacritic.com/faq#item18>

[2] “Using Steam data to tell if your game will sink or swim”, <https://venturebeat.com/2017/06/28/using-steam-data-to-tell-you-if-your-game-will-sink-or-swim/>

[3] “Steam – What’s your Game?”,<https://nycdatascience.com/blog/student-works/web-scraping/steam-whats-game/>

[4] “What’s in the Name? Data analysis of 5,820 Steam Games”, <https://gamedevelopment.tutsplus.com/articles/whats-in-a-name-data-analysis-of-5820-steam-games--cms-30101>

[5] “Understanding your game through data”, <https://galyonk.in/understanding-your-game-through-data-8b09ca93ec11>

[6] Steam data, <https://github.com/CraigKelly/steam-data>

[7] Steam (software), <https://en.wikipedia.org/wiki/Steam_(software)>

[8] Edwards, Cliff (November 4, 2013), “Valve Lines Up Console Partners in Challenge to Microsoft, Sony”

[9] Bailey, Dustin (March 22, 2018), “With $4.3 billion in sales, 2017 was Steam’s biggest year yet”

[10] Answer to “How many video games exist?” on Quora by John Mcmillagan, <https://www.quora.com/How-many-video-games-exist>