Key aspects for a popular video game: using online databases to evaluate the success of a video game.

Juan Montenegro M2020586 Data Science Student

Abstract

As the video game industry grows, extensive amounts of data are generated daily, and video game developers are increasingly becoming more dependent of the data created by their players. This data can be of help for the analysis of new games, mechanics, ideas and sales. The present research has the objective of identifying which aspects make a video game popular and therefore, successful. This is done through a qualitative analysis using the database from the two biggest video games APIs: RAWG and Steam.

Introduction

According to the Cambridge Dictionary a game is “an entertaining activity or sport, especially one played by children, or the equipment needed for such an activity” (Cambridge-a, n.d.) Through the evolution of technology, games have changed and developed into a different means of portrayal. This is the case of videogames. Video games are a type of game in which the player controls moving pictures on a screen by pressing buttons. (Cambridge-b, n.d.). This type of games have gained an incredible amount of popularity in recent years, to the extent of generating a worldwide 159.3-billion-dollar market in 2019 and is expected to surpass the 200-billion-dollar mark by 2023. (Gough-a, 2020)

The immense demand of video games has given rise to its own industry. The video game industry is the economic sector involved in the development, marketing, and monetization of video games. (Zackariasson, P. and Wilson, T.L. eds. 2012). By the end of 2020, this industry alone is expected to count for 2.7 billion gamers worldwide, with 2.5 billion of those users playing on mobile phones, 1.3 billion playing on PC, and 0.8 billion on a console. The number of users is expected to increment by 2023 by more than 3.07 billion video game players. This is due to the increasing growth of markets in regions like Asia-Pacific, the Middle East, Africa, and Latin America. (Wijman, 2020)

Is specially the Asia Pacific region where the main point of sales is for the video game industry. Over 1.5 billion video gamers were counted in 2020 for this region alone. The Asia Pacific region has generated a combined revenue of 78.3 billion U.S. dollars. Gamers worldwide spend an average of over 123 U.S. dollars on video games over a three-month period in 2018. (Gough-b, 2020).

With all this information and daily generation of new data, video game developers are increasingly becoming more dependent of the data created by their players. The data is used for a variety of reasons: to determine how to increase aspects like engagement, calculate the average number of sessions, each session length and average revenue per user. By a quantified understanding of which elements are popular and those that are not, where and at what level players lapse, and what features players enjoy more, developers can increase player engagement and therefore revenue. (OC&C, 2020).

Nonetheless, despite this increasing amount of data, there is not yet a complete understanding of the factors that make a video game successful[1]. For example, in 2019, high budget video games like "Call of duty" required a budget of 50 million dollars to make but earning more than 1 billion dollars (Activision, 2019). Meanwhile, micro budget games like "Among us" cost around 100 thousand dollars to produce. Nonetheless, this specific game has earned around 3.2 million dollars in 2020 alone. (Handrahan, 2020). When talking about popularity levels, in an specific example, "Call of duty: Warzone" had more than 75 million downloads by September 2020 (Takahashi, 2020) 12 but "Among us" had 100 million downloads by the same period. (Fenlon, 2020)

The present investigation intends to search and determine which are the key aspects that make a video game popular and therefore, being successful. In specific, an effort is made to understand these key aspects and offer a superficial guide to aid future video games developers at the moment of creating a video game by demonstrating a Exploratory Data Analysis (EDA) of the features that players enjoy the most based on more than 35,000 games. Additionally, a regression model will be presented as a tool to input different characteristics of an upcoming game and prognosticate the popularity and (implicitly) success of a particular video game.

Methodology

For the development of this investigation, web scraping tools will be used to extract information of the RAWG and Steam APIs (Application Programming Interface)[2]. Both of

[1] For the purpose of this investigation, the word "successful" is referenced as the number of sales and playability a video game has through a period of time. In this case, it will be counted as a year.
[2] API stands for application programming interface. An API is a way to programmatically interact with a separate software component or resource. (Freeman, 2019)

these sites contain information for more than 350,00 video games and different aspect of them than can be used on this analysis (ex: reviews scores, prices, genres, popularity level, etc.).

Web Scraping is a technique used to extract large amounts of data from websites where the extracted data is saved to a local file or to a database in table (spreadsheet) format for further analysis (Web harvy, n.d.). The built-in functions from "Beautiful Soup" will be used. "Beautiful Soup" is a free open-source Python library with the aim of extracting data from HTML files.

Is important to note that the predictor variable will be the popularity of the game. This popularity number starts from 1, which indicates that the game positioned in this place is the most popular game at the moment of the collection of the data. The bigger the popularity number the less popular the game is in the database.

The regression algorithm is created by collecting the data and performing feature engineering. This process is required to preparing a proper input dataset that is compatible with the machine learning algorithm that is planned to use. Additionally, feature engineering aims to improve the performance of the machine learning models. This includes cleaning the data by getting rid of unnecessary, missing or problematic data points that can impair the analysis. Punctuation marks, symbols and additional white spaces from the data are eliminated. Additionally, three feature scaling methods will be implemented to the data: Normalization, Standardization and Powerful Transformation. This is done to compare which of these scaling methods provide the best results.

Once the feature engineering is culminated, the data is then separated as train and test data. The train data is used to form and train the model in order to predict the test data. Different regression models will be used to make the prediction of the dependent variable. In this investigation, linear regression, Elastic Net, Ridge, Lasso, XG Boost Regressor and Decision Tree Regressor will be used. Afterwards, the models will be evaluated and compared according to how well they scored against the test data.

Research framework

The present research will make use of an exploratory analysis and machine learning techniques to find out the key aspects that make a game successful and to predict if a new game can be or not successful according to a variety of different aspects that games share. It uses quantitative analysis of the obtained data to find relationships between these features and game success. The exploration of the data has been made with descriptive statistics and graphical analysis.

Hypothesis development

It has been studied that there are certain aspects that successful video game share between them. There is the existence of heuristic analysis which suggest that the most important elements of good game design are cohesion, variety, good user interaction and some form of good social interaction. (Bond, 2009). Other studies propose that the aspects that truly make a video game successful are due to games being released by a popular video game publisher for a popular hardware platforms and higher quality video games are significantly more likely to sell a greater number of units than those of a lower quality (Cox, 2013)

Nonetheless, one of the purposes behind this study is due to the insufficient data taken into account by previous studies on the subject to formulize conclusions about the aspects that make a video game successful. In previous investigations, no more than 5,000 videogame titles were analyzed, and in some studies, only samples of fewer than 100 titles were investigated. The present investigation makes use of 23,579 title games which represents a more representative sample for an investigation. Therefore, this study proposes that:

Hypothesis 1a (H1a): There are common aspects shared between different successful videogames. The aspects found in the present investigation that make a video game successful are different from the ones mentioned in previous investigations done on the subject.

Hypothesis 1b (H1a): There are common aspects shared between different successful videogames. The aspects found in the present investigation that make a video game successful are the same as the ones mentioned in previous investigations done on the subject.

Hypothesis 2 (H2): There are no common aspects shared between different successful videogames. Each game success is independent from the aspects that other games possess and any game is defined by its own concept and creativity.

Statistical Analysis

Exploratory data analysis was performed on the database to find different results that describe the success of videogame due to its popularity in two of the biggest game web pages in the world. Categorical variables like game genre were encoded as dummy variables by the generate dummies tools from the SK-Learn library. The R2 score was calculated between the main dependent variable: popularity score and the rest of the independent variables.

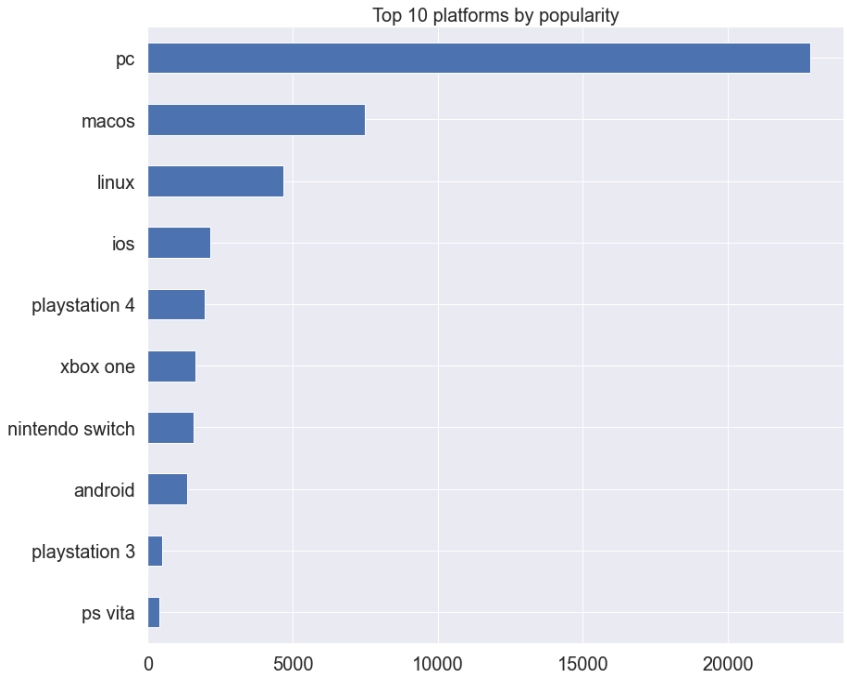
Results

No.	Aspect	Description
1	Presence	How many articles on social media sights relating to the game
2	Platform	Which platforms a certain video games will run on. Ex: PC, macOS, PlayStation, etc.
3	Ratings Breakdown	Distribution of game reviews on RAWG
4	Franchise	Whether or not this game is a part of a franchise
5	Original Cost	How much the game cost without being on sale
6	Controller	Does the game use a controller or not
7	Achievements	Indicates how many achievements are available to earn in the game
8	Languages	A list of supported languages. Not all languages are supported on the same level, though if they are present then they have some support.
9	Storage	Minimum storage requirements for running the game.
10	Memory	Required system memory to run the game
11	Genres	A list of genres assigned to the game
12	Tags	Player assigned tags, usually designating the genre, number of players.
13	Id/ popularity of a videogame	Database primary key. It represents the order the games were scraped, which was sorted by popularity.

Source: Self-made with RAWG and Steam data

The present investigation makes use of Exploratory Data Analysis (EDA) to find different type of relationships between different variables against the popularity variable. Here, the most important aspects are presented. In total, 13 relevant aspects were extracted from the RAWG and Steam databases. From this 13, twelve were used to predict the dependent variable: id, which is renamed as “popularity of a videogame”.

Due to space constraints, the most relevant findings will be detailed in the present study. The first relevant analysis comes from the examination between platform and popularity. Video games are released and played in different platforms. For this dataset, the combination between platforms is dominated by the popularity of video games played on PC (or video games played on desktop computers).

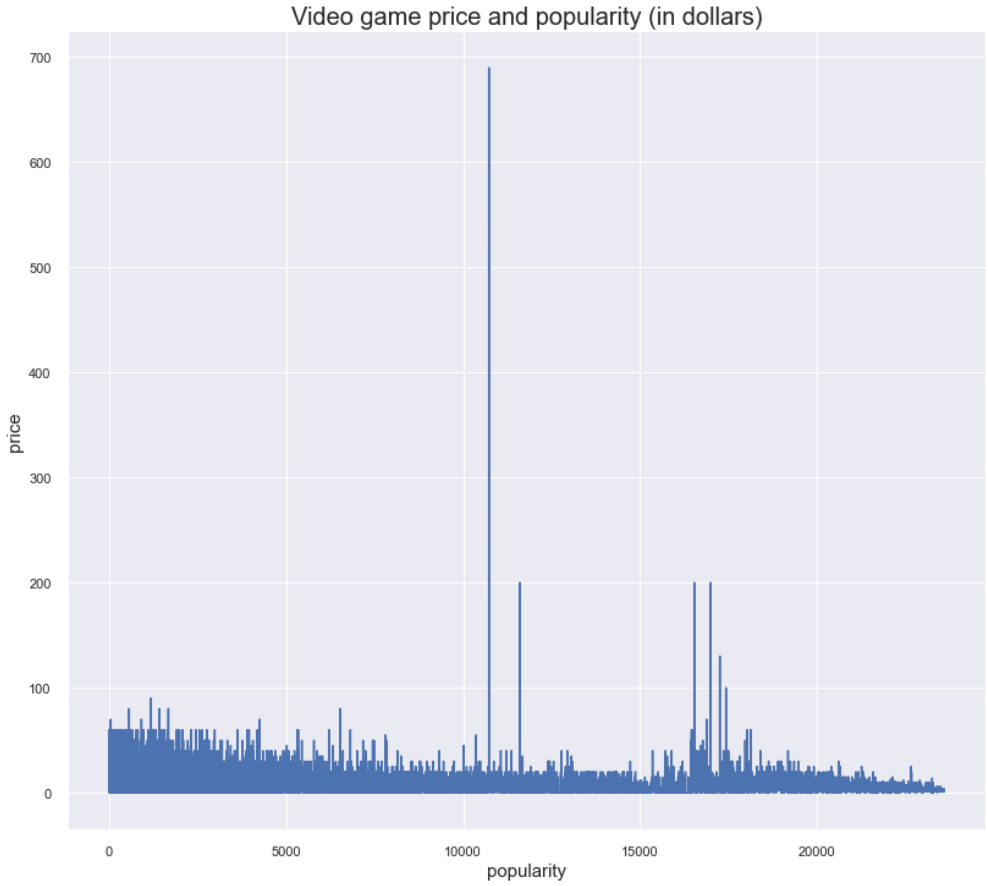


Source: Self-made with RAWG and Steam data

The three most popular platforms for a video game, are represented in the use of a desktop computer. Is important to note that PC is referred for the Windows operating system.

As a first finding, it can be concluded that the popularity of a videogame is affected by the platform to which is released. The preferred platform would be a desktop computer, especially for the Windows operating system. This is followed by the PlayStation 4, Xbox one and Nintendo switch video game consoles.

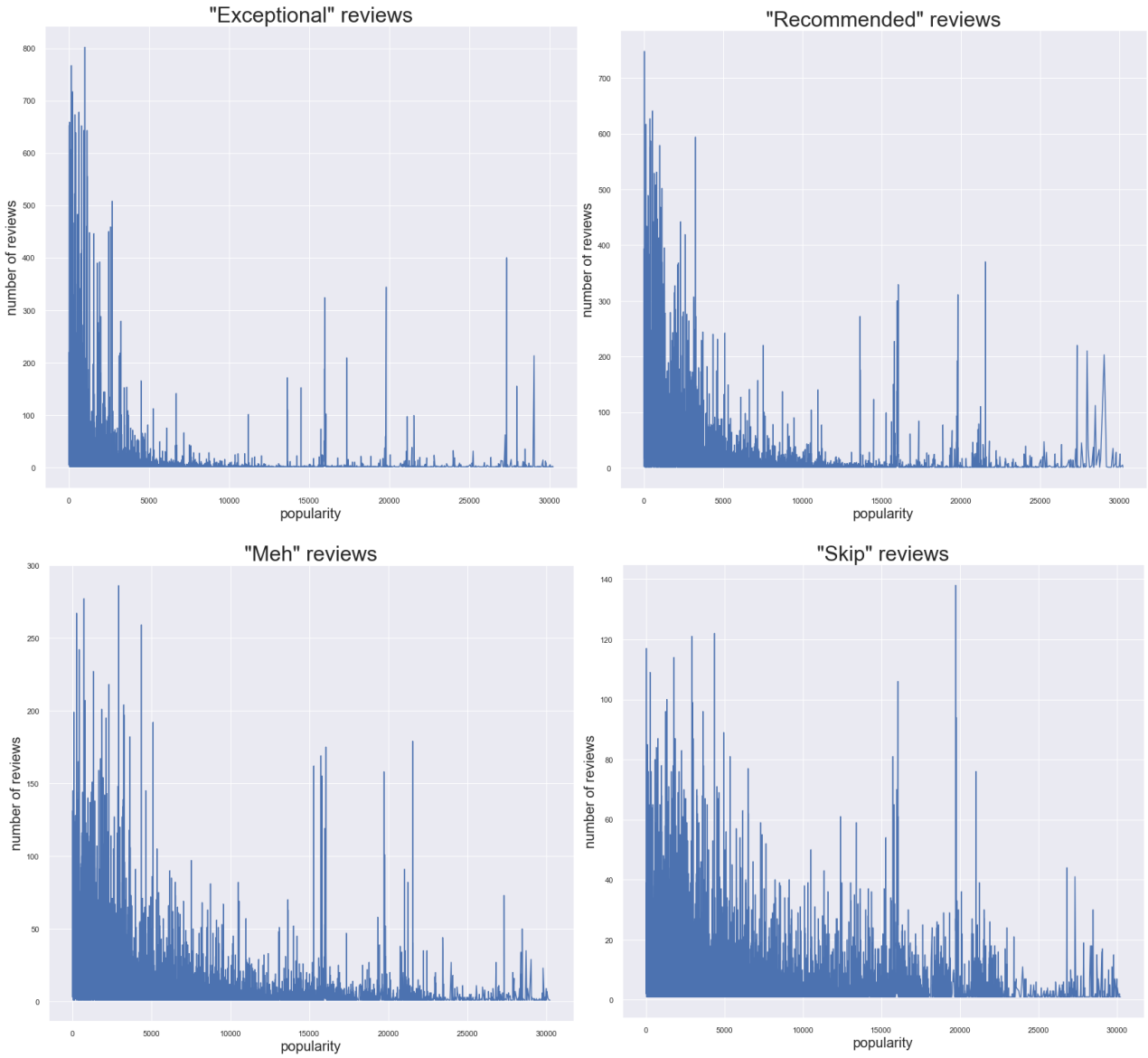
For the next analysis, the original cost was compared to a video game popularity. In the "Video game price and popularity (in dollars)" graph, some interesting facts can be observed. First, the most popular video games tend to be around the mark of 80 dollars. Therefore, a popular video game is not affected by "bad pricing" as stated by the 2009 study by Matthew Bond were one of the negative aspects for a good game was expensiveness. Even so, the most expensive game on the list, "X-Plane 10 Global", with a price of around 678 dollars is situated on the middle of the popularity scale.



As a second finding, it can be stated that a moderate high price (around 80 dollars) is a positive factor for the popularity of a video game. This can be explained by the relation with users that a high price conveys high quality.

Source: Self-made with RAWG and Steam data

As a next point for analysis, the reviews for the RAWG database will be taken into consideration. These reviews are formatted in a scale of four categorical variables. In other words, as an ordinal variable: Exceptional, Recommended, Meh and Skip. Exceptional is intended for games that are, as its name indicates, exceptionally good. Recommended is assigned to those games that can be shared as a preference within users. Meh is a category used for games that have been played but didn't provide enough enjoyment. Finally, Skip is for those games that are not even considered worth to even play.



Source: Self-made with RAWG and Steam data

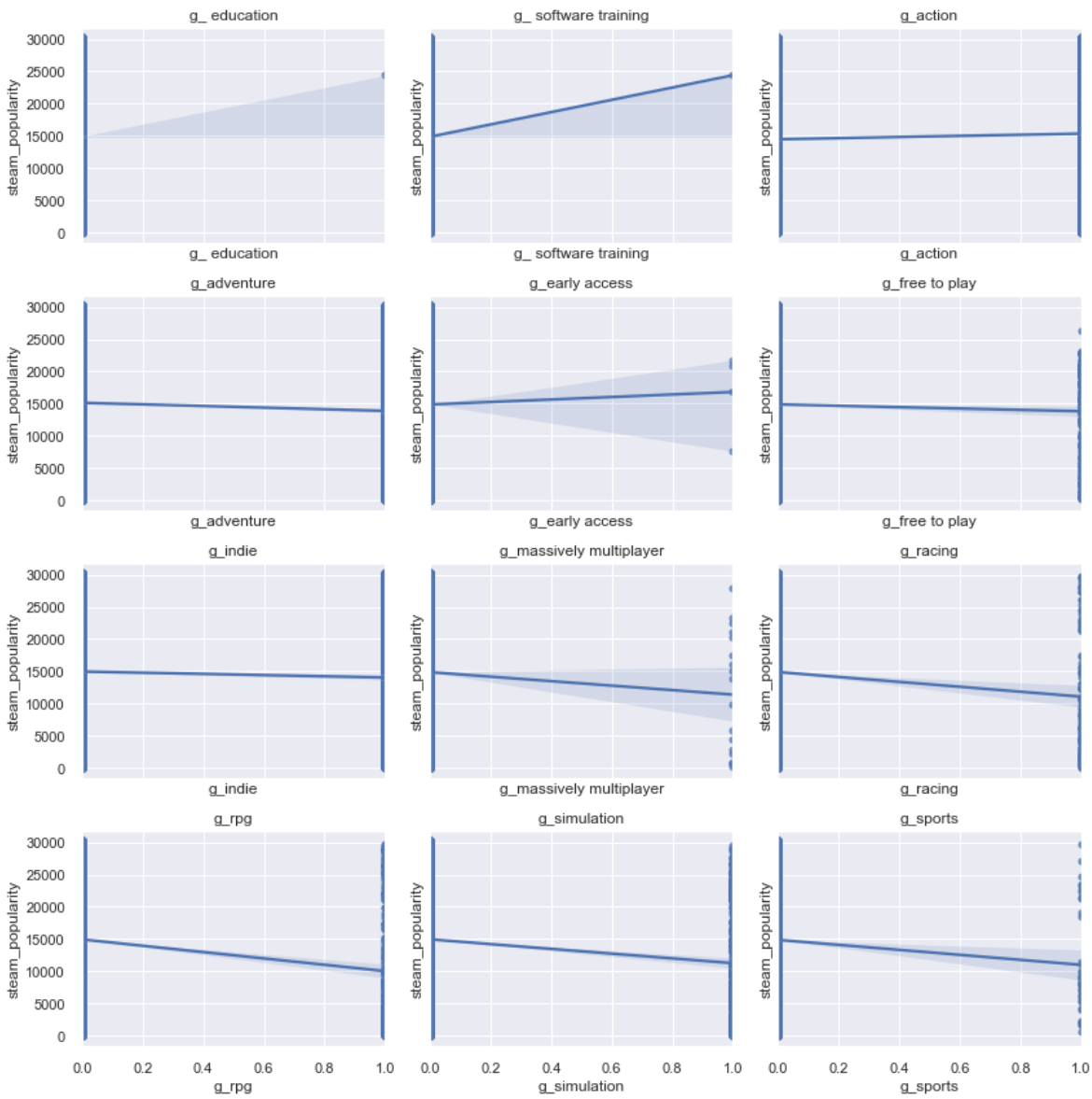
From these graphs, it can be understood that games with higher popularity (the ones that are close to cero) possess more exceptional reviews. Nonetheless, these types of games also manage to have reviews for the recommended, meh and skip categories. Even more for this last category. Therefore, it can be concluded that taking into account reviews for an analysis of what makes a video game good, although can be good as a form of feedback, is not fundamental to take into account when analyzing the key aspects of a good video game.

This analysis can be furthered explained with the use of the variable presence and video game popularity. The "presence" variable dictates how many articles appear on social media related to a game. This aspect is useful due to the direct relationship between the publicity a game obtains when displayed in a social or online news platform and consequentially, the amount of people that can get interested on it and therefore, playing the game which ultimately contributes to its popularity.

As it can be seen, the points of the scatter plot are concentrated in the majority around the most popular games indicating that a major presence is equal to more popularity.

Source: Self-made with RAWG and Steam data

For both sets of graphs it can be stated that having numerous reviews is a better indicator of popularity than having just some "exceptional" reviews as described by RAWG. The difference between good and bad reviews is not significative when it comes to describing the popularity of a games. The importance comes from having those reviews.



Source: Self-made with RAWG and Steam data

The above graph shows a connection between video games' genres and their popularity. It also contains the tags that users have added to the game which is represented on the right line. Nonetheless, the importance of the graph is to demonstrate which game genre is more popular. As the slope of the line declines, the popularity of the game increases. Surprisingly, traditional game genres like action and adventure are not as popular as some genres that have surfaced or had different changes on their core mechanics in recent years:

RPG[3](Role Playing Game), simulations, racing, sports and massive multiplayers dominate when it comes to popularity.

Additionally, the least preferable genres to be popular are the ones related to education (which doesn't present many titles) and Software training (classified as a game because it can be targeted to both professionals and casual players. An example of this kind of game would be the X-Plane 10 Global game discussed in the video game price section).

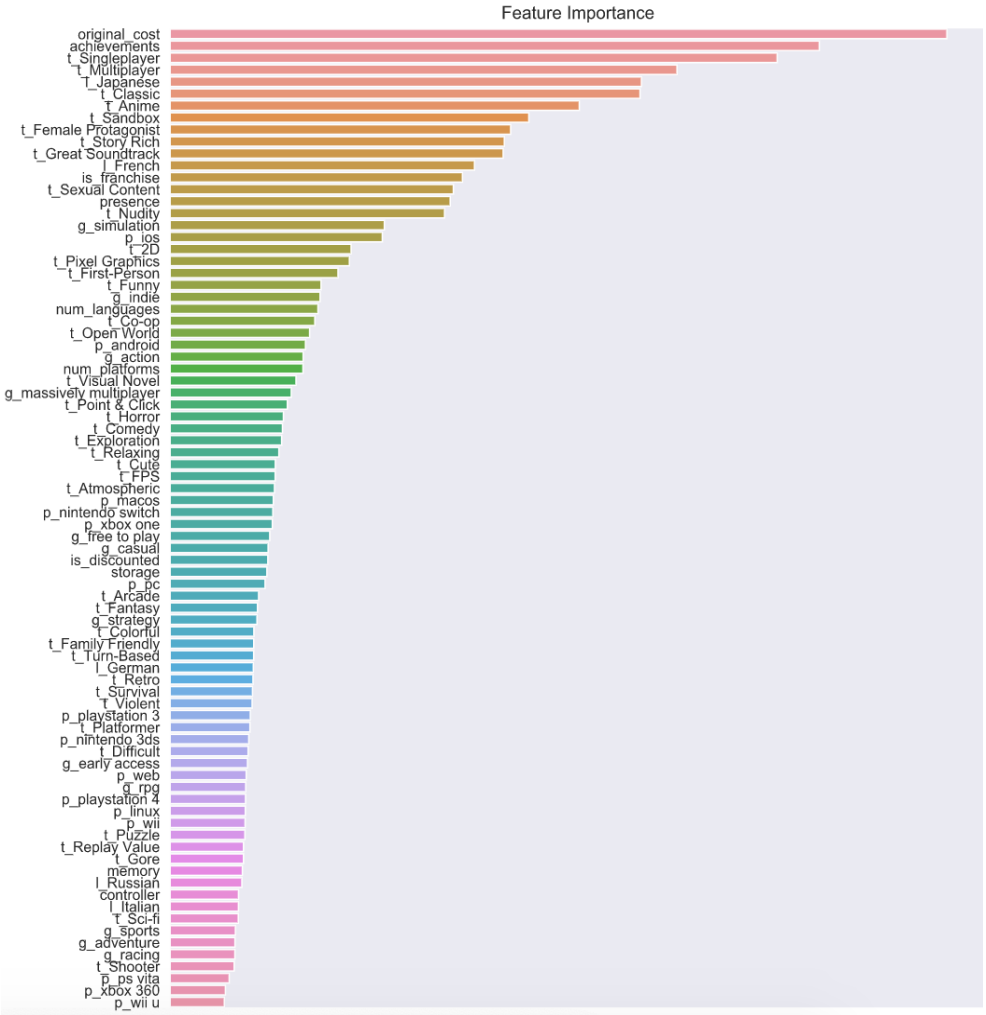
Another fact that can be taken from this analysis is that games made by independent small studios or even made by one individual are not popular. Therefore, respecting this analysis, the success of a video game can also be interpreted in the form that a video game will have more probabilities of success when it comes from a renowned publisher or developer with previous records of making other video game titles. Further investigation should be done on this aspect.

Machine Learning section

This section presents an application of different machine learning models for predicting the popularity of upcoming games and therefore, their success. This is done with the intention of helping game developers to, not only be aid with the previous exploratory analysis, but also possess a quantifiable tool to measure popularity according to the different aspects taken into account in this study and the aspects of an upcoming video game.

[3] An RPG genre is a game in which players advance through a story quest, and often many side quests, for which their character or party of characters gain experience that improves various attributes and abilities. (Hosch, n.d.)

For the analysis of the model, feature engineering was performed in first place. After all, the model results will be as good as the features used in it. For this, certain categorical variables like platform or languages needed to be split and then processed with the "get_dummies" function from the pandas package. With this, all the features are presented as quantitative.



Source: Self-made with RAWG and Steam data

For the creation of the different models, a small Decision Tree Classifier was used to evaluate the features importance and so, contribution to the regression task for the models. The most prominent feature was original cost.

As it was stated in the methodology, the data was scaled in three ways: Standardization, Normalization and Power Mean. This is a necessary step to improve the score of the models. The SkLearn preprocessing library was used for this task. For normalization the "Min Max Scaler" package was used. Respecting Standardization and

Powerful Transformation the “Standard Scaler” and “Power Transformer” were used respectively.

Is important to note that these two last scaling methods were not applied to the dummy generated variables due to the difference in values these would have generated. The models were tested with each scaling method and the results were compared.

For each model a set of different parameters were given and a Grid Search was done to find the best parameters from the set. This is done to automate the task of finding the best hyperparameters for the model. For this task, a cross validation of 3 was used. The results of the models are presented in the following table:

Scaling	Model	Score on test set	R2
Min Max scaling	Linear Regression	0.401	0.383
	XG Boost Regressor *	0.513	0.491
	Decision Tree Regressor*	0.174	0.175
	Lasso	0.383	0.391
	Ridge	0.384	0.391
	Elastic net	0.383	0.391
Standard Scaler	Linear Regression	0.364	0.380
	XG Boost Regressor*	0.513	0.491
	Decision Tree Regressor*	0.174	0.175
	Lasso	0.384	0.392
	Ridge	0.384	0.391
	Elastic net	0.384	0.392
Power Transformer	Linear Regression	0.344	0.430
	XG Boost Regressor*	0.513	0.491
	Decision Tree Regressor*	0.174	0.175
	Lasso	0.431	0.442
	Ridge	0.431	0.442

Source: Self-made with RAWG and Steam data

The model that had the best performance was the XGBoost Regressor with a test score of 0.513 and and R2 of 0.491. This was done with the following set of hyperparameters: objective='reg:squarederror', learning_rate= .1, max_depth=7, min_child_weight=4, subsample=.7, colsample_bytree=.7, n_estimators=300. The rest of the regressor models didn't performed as well giving low results in both the test score and the R2 score. The least recommended model for this task is the Decision Tree Regressor with a test score of 0.174 and an R2 of 0.175.

Is important to note that the XGBoost Regressor also didn't performed as well as a proper machine learning model should. This is an opportunity for future investigations to realize a finer tunning process for this machine learning task. Future investigations could prepare more complex models that give even better results, like neural networks. These models could also be ensembled to create an even better score than just comparing each model individually. As another recommendation, a more in dept analysis could be made to find the best hyperparameters using other search tools like Bayesian optimization.

Study Limitations

The use of an API comes with certain limitations. For this study the extraction of the data using web scraping to collect the information presented some issues. In the first place, the process took hours to complete. Around 11 hours were employed to get 30,249 video games titles and their respective information from the APIs (23,579 final games were

*Decision trees and decision trees-based models do not require feature scaling to be performed as they are not sensitive to the variance in the data.

analyzed after drooping repeated games and feature engineering). This is due to the possibility of IP blocking by the server hosting the API.

IP blocking is a method used by webpages to stop scraping techniques from accessing data of a website. It usually occurs when a website detects a high number of requests from the same IP address (Octoparse, 2019). Due to the high traffic generated by this type of demand of information, a website could confuse the web scraping with an informatic attack and the website would procced to ban the IP or restrict its access by breaking down the scraping process. For this, the Steam and RWAG APIs require between 1 and 5 second of delay between the scrapping of information for each game.

This time constraint makes it difficult to obtain all the information of all the games the website host. This represents a constraint to a full-scale analysis of the information. Another limitation presented in the study is the errors presented while scrapping the data. In some cases, around 1000 games presented issues were the information belonging to one game was repeated in the other 1000 games slots. Due to time constraints, these games had to be dropped from the analysis.

As a final limitation, there is the ever-present missing data aspect when obtaining information. For this investigation, 6 features had to be dropped from the analysis due to having more than half of missing values. Dropping these features represent a loss of information and makes the analysis less informative.

Conclusions

The results obtained from this research should be considered in light of their limitations. The EDA performed in the first section of the investigation demonstrates some

interesting results that defy the findings of different previous investigations on the subject. This can be attributed to the use of a more extensive database and the fact that it comes from two different renowned sources in the video game community.

From the EDA we can stablish that the aspects that make a video game successful are: games that are oriented to Desktop computers, especially for the Windows operating system. Video games that cost around 80 dollars also have much more popularity within costumers due to its price being related to high quality. This doesn't mean that a game done by a less renowned publisher or doesn't follow the levels of quality that these 80 dollars games have, should have this type of price. The video game price should go accordingly to the complexity technical value it has.

Video games that possess great amounts of reviews have a higher level of popularity, and therefore success. The important key from this aspect is that the nature of the reviews is not significant. Good or bad reviews are helpful to boost the popularity of the game and make other potential customers decide to try the video game.

As a final aspect to take from this investigation, videogames that belong to the RPG, simulations, racing, sports or massive multiplayers genre posse's higher popularity within the video game community. This goes accordingly with previous works done on the subject, like the Newzoo Global Esports Market Report from the present year. Globally, the total esports audience will grow to 495.0 million people in 2020, a year-on-year growth of +11.7% (Newzoo, 2020). The most popular game played in this type of competitions is "League of legends" which belongs to the massive multiplayer category. The "Among Us" mentioned in the beginning also belongs to this genre, which explains its popularity rise.

References

Activision. (2019, December 18). *Investor Activision*. Retrieved from https://investor.activision.com/news-releases/news-release-details/call-duty-modern-warfare-1-most-played-call-duty-multiplayer

Bond, M. (2009). What makes a good game? Using reviews to inform design. (p. 6). Research Gate.

Cambridge-a. (n.d.). *Cambridge Dictionary*. Retrieved from https://dictionary.cambridge.org/dictionary/english/game

Cambridge-b. (n.d.). *Cambridge Dictionary*. Retrieved from https://dictionary.cambridge.org/dictionary/english/video-game

Cox, J. (2013). *What Makes a Blockbuster Video Game?An Empirical Analysis of US Sales Data.* Hampshire: Wiley Online Library.

Debra J. Brody, M. L. (218). *Prevalence of Depression Among Adults Aged 20 and Over.* NCHS Data Brief .

Fenlon, W. (2020, September 24). *PC Gamer*. Retrieved from https://www.pcgamer.com/how-among-us-became-so-popular/

Freeman, J. (2019, August 8). *Info World*. Retrieved from https://www.infoworld.com/article/3269878/what-is-an-api-application-programming-interfaces-explained.html

Gough-a, C. (2020). *Gaming market value worldwide 2012-2023.* Statista.

Gough-b, C. (2020). *Number of video gamers worldwide 2015-2023.* Statista.

Handrahan, M. (2020, September 16). *Games Industry*. Retrieved from https://www.gamesindustry.biz/articles/2020-09-16-among-us-got-70-percent-of-lifetime-downloads-in-45-days#:~:text=In%20August%202020%20alone%2C%20Among,its%20revenue%20is%20more%20modes

Hosch, W. (n.d.). *Britannica*. Retrieved from https://www.britannica.com/topic/role-playing-video-game

Newzoo. (2020). *Newzoo*. Retrieved from https://newzoo.com/insights/trend-reports/newzoo-global-esports-market-report-2020-light-version/#:~:text=Global%20esports%20revenues%20will%20grow,from%20%24950.6%20million%20in%202019.&text=Globally%2C%20the%20total%20esports%20audience,year%20growth%20

OC&C. (2020). *Growth in the Video Gaming Market: the changing state of play.* OC&C Strategy Consultants.

Octoparse. (2019, November 19). *Octoparse*. Retrieved from https://www.octoparse.com/blog/9-web-scraping-challenges

Takahashi, D. (2020, August 4). *Venture beat*. Retrieved from https://venturebeat.com/2020/08/04/call-of-duty-warzone-hits-75-million-downloads-in-less-than-5-months/#:~:text=Log%20Out-,Call%20of%20Duty%3A%20Warzone%20hits%2075%20million,in%20less%20than%205%20months&text=Call%20of%20Duty%3A%20Warzone%20%20grew,popu

Web Harvy. (n.d.). *Web Harvy*. Retrieved from https://www.webharvy.com/articles/what-is-web-scraping.html#:~:text=Web%20Scraping%20(also%20termed%20Screen,in%

Wijman, T. (2020). *Three Billion Players by 2023: Engagement and Revenues Continue to Thrive Across the Global Games Market.* Newzoo.

Zackariasson, P. and Wilson, T.L. eds. (2012). *The Video Game Industry: Formation, Present State, and Future*. New York: Routledge