WHAT FACTORS INFLUENCE VIDEO GAME SUCCESS?

6DATA004W - Data Visualisation and Dashboarding

Coursework Report

Tomasz Wasowski – w1684891

Table of Contents

1. Research Question

The purpose of this report is to formulate, analyse and answer a research question related to a personal area of interest. Following a passion for the development and consumption of video games, I am interested in understanding '**What factors influence video game success?**'. In order to answer this question, it is important to first define what is meant by 'factors' and what is meant by 'success'.

The 'success' of a video game can be measured in many different ways, usually depending on the context of the inquiring party. It can be measured by comparing total revenue or profit made from the video game for the companies that developed and published it. It could also mean the number of sales made, or the number of concurrent users. However, for the purpose of this report, 'success' will be measured from the perspective of the end user; defined by the ratio of positive user reviews attributed to the video game in question across its lifespan.

The 'factors' that may influence the success of a video game refer to measurable statistics or features pertaining to that specific video game. For example, the average playtime, the genre or even the price of the video game can all indicate how it has been received by the consumer. Many of these factors can be plotted directly against the user review ratio in order to inspect visual trends and correlations.

2. Data Acquisition

The data used to answer the question at hand will be originally sourced from Steam, the biggest digital video game distribution service and storefront run by the Valve company. The reason for this choice is that steam has been holding around 75% of market share when it comes to the digital distribution of PC video games, offering over 34,000 games for sale and maintaining more than 95 million monthly active users (Wikipedia, 2022). As a result, data gathered from Steam is likely to show the most complete picture when it comes to answering the research question.

The data in question is publicly available and can be acquired directly from the Steam API, however, as this process can be quite time consuming, an already existing dataset sourced from the Steam API can be acquired from the Kaggle website instead. Kaggle is a community website for data scientists where datasets for projects such as this one can often be found. The datasets used were uploaded approximately 3 years ago by a Kaggle user named Nik Davis, who sourced it by polling the Steam and SteamSpy APIs.

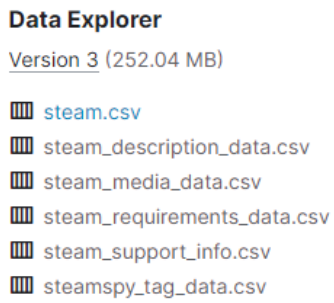

Figure 2-1 - Steam data .csv files from Kaggle

The data set consists of multiple .csv files containing information about individual video game releases on the digital Steam store. These files include information such as the release date, developer, and price, but also statistics regarding user reviews, average and median play times, as well as an estimated number of players that own that specific game. All files have a common 'AppID' column by which they can be joint together for analysis. This column contains the unique identifier attributed to each video game released on the digital Steam store.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	appid	name	release_d	english	developer	publisher	platforms	required_	categories	genres	steamspy_	achievem	positive_r	negative_	average_p	median_p	owners	price
2	10	Counter-S	########	1	Valve	Valve	windows;	0	Multi-play	Action	Action;FP	0	124534	3339	17612	317	10000000-	7.19
3	20	Team Fort	########	1	Valve	Valve	windows;	0	Multi-play	Action	Action;FP	0	3318	633	277	62	5000000-1	3.99
4	30	Day of De	########	1	Valve	Valve	windows;	0	Multi-play	Action	FPS;Worlc	0	3416	398	187	34	5000000-1	3.99
5	40	Deathmat	########	1	Valve	Valve	windows;	0	Multi-play	Action	Action;FP	0	1273	267	258	184	5000000-1	3.99
6	50	Half-Life:	########	1	Gearbox S	Valve	windows;	0	Single-pla	Action	FPS;Actior	0	5250	288	624	415	5000000-1	3.99

Figure 2-2 Snippet of the main steam.csv file containing a portion of the data that will be used for this project

This dataset will be suitable for analysis because it includes information about all factors that could potentially influence the success of a video game by our established definitions. Additionally, it holds just over 27,000 records, which is a significant amount of Steam's entire video game catalogue; certainly, enough to analyse and answer our question without limitation or bias.

3. Preparation

Preparing data for exploratory analysis consists of multiple steps to ensure data integrity, join separate tables, and make any necessary aggregations. This process begins in excel, where it is necessary to check the data for any erroneous or missing values. The dataset acquired has already been cleaned, so there should not be any issues, however, it is imperative that integrity is confirmed before analysis begins.

3.1. Ensuring Data Integrity

Several excel formulas, filters and conditional formatting will be used to ensure that the dataset acquired is clean and ready to be used for analysis. The first formula used will be to check for erroneous values, which include any cells that contain either '#VALUE' or '#N/A'. These values are present if an error occurs and the value cannot be displayed or calculated, and a record with this value would not be usable for analysis.

To check for this issue, we can use the formula =IF(ISERROR(A2),1,0), where A2 is the cell reference of the cell we want to check. This will result in a value of 1 if the value is an error, or 0 if the value is correct. Once the formula is typed, it can be dragged across horizontally and vertically to match the dimension of the dataset, creating a matrix which can then be filtered to highlight only rows which resulted in a 1.

	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	W	X	Y
	develo	publish	platfor	require	categor	genres	steams	achieve	positiv	negativ	averag	mediar	owners	price			
1	#NAME?	#NAME?	windows	0	Multi-play	Action;Inc	Early Acce	18	74	16	0	0	0-20000	5.19	1	1	0
1	#NAME?	#NAME?	windows;	0	Single-pla	Indie	Indie;Puz	0	14	0	0	0	0-20000	3.99	1	1	0

Figure 3-1 - Two erroneous records found while filtering through the data

As can be seen in Figure 2-1 above, this method has located a few records with '#NAME?' errors in them. These errors originated from the developer name starting with the '=' symbol, which in excel translated to the beginning of a formula. These rows were fixed by using an escape characters to treat the '=' symbol as a string and display the names properly. All the other columns filtered using this method didn't show signs of errors across the dataset.

This process was repeated using another formula which checks specifically for null values within the cell. The formula used for this was =IF(ISBLANK(A2),1,0), replacing the ISERROR function with the ISBLANK function, which instead checks specifically if the cell contains a null value. With each column filtered, it seems that no null values were found and that all cells contain some data within them.

Additionally, a final check can be used to ensure that the data does not have any anomalous results. For this, the formula is replaced with a conditional checking whether the value of the cell is less than or greater than 0. This formula highlights any cells which contain a negative value, which would not make sense in this data set, as you cannot have a negative value for playtime, price, reviews, etc. As with null values, this check came back empty, meaning that the data does not have any negative values.

3.2. Data Joining and Aggregation

In order for the entire set to be analysed, the tables from the individual .csv files have to be joint together by their common 'AppID' columns. To do this, the files are imported into Tableau and an inner join is created to link each additional table onto the main steam.csv file. The result of this can be seen in Figure 3-2 shown below.

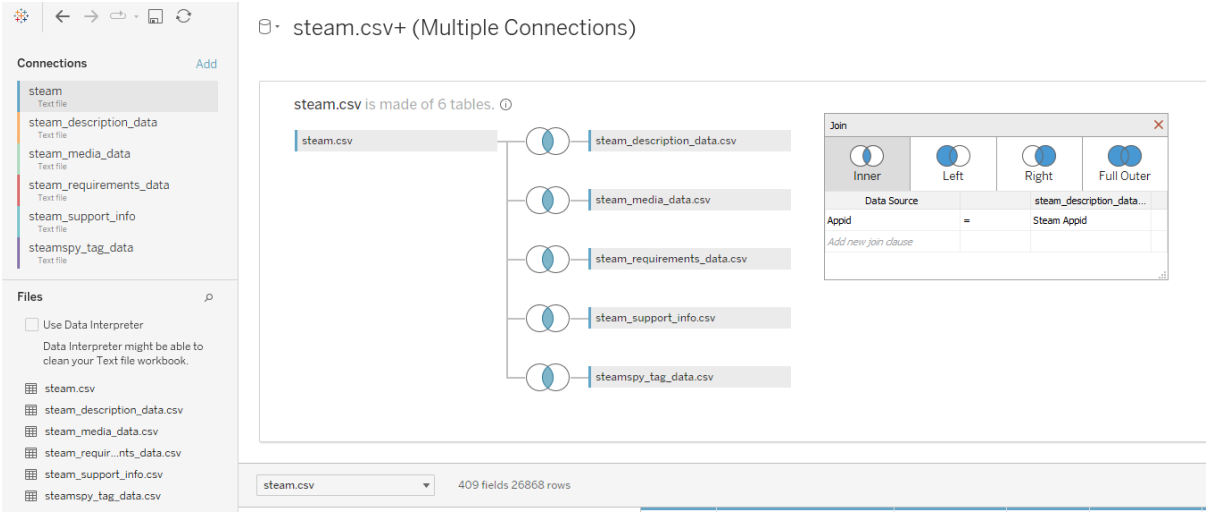

Figure 3-2 - Joining dataset tables by their 'AppID' column

Now that the data is joined the last thing that has to be prepared is for the positive review percentage, used to determine the success of the games, to be calculated. This is done by creating a new calculated field in tableau and using the formula: Positive Reviews / Total Reviews * 100. This creates a field that shows us the percentage of positive reviews for each video game in the data set. This operation can be seen in Figure 3-3 shown below.

Figure 3-3 - Creating a field to calculate positive review percentage for each game in the data set

With these preparations in place, the data is ready to be analysed in order to answer the research question.

4. Exploratory Data Analysis

Exploratory data analysis (Vigni et al., 2013) performed on this data set has followed a specific methodology (Vanawat, 2021) thus far, and will consist primarily of bivariate data analysis (Sims,

2000), as the question specifically asks about which factors have influence on the success of a video game. For this, key factors will be identified from the data set and measured against the calculated positive review percentage in order to identify any trend patterns and measure the factor's impact on the game's success. Where appropriate, additional variables will be taken into consideration depending on the context of the data.

Once all the factors are identified, they will be categorised based on relevance and explored within the context of their groupings. Finally, results will be presented across multiple visualisations displayed on a dashboard for each grouping.

4.1. Factor Identification and Categorisation

Factors which influence the success of a video game can be statistics and metrics collected about its user base, or intricacies and details related to its development and release. From the data set utilised, the following factors have been identified:

- ***Estimated number of owners*** – This is the number of users that have purchased the video game in question. The larger the number, the more people own the video game.
- ***Number of user reviews made*** – This is the total number of reviews that have been created. A large number may indicate a more popular or controversial video game.
- ***Average and median playtime*** – This is the amount of time that users spend playing a specific video game. A larger number indicates a longer play time amongst its player base.
- ***Number of achievements available*** – Achievements are awards used to track progress within a video game. A larger number means that the video game has more awards available for the player to collect.
- ***Language support*** – This refers to whether a game supports the English language. Games that support it will be written or voice acted in English.
- ***Video game genre*** – This refers to the style of gameplay the user can expect from a video game. An example of a video game genre would be 'action' or 'sports'.
- ***Video game category*** – This refers to the type of game that each record is. An example of a category could be 'single-player' or 'co-operative'.
- ***Release date*** – This refers to the date when the game has officially been released, determining how old the game is.
- ***Age restrictions*** – This number specifies the age the user has to be to play the video game.
- ***Publishing company*** – This refers to the company that helped the developers publish the video game.
- ***Video game price*** – This refers to how much the game costs on the Steam digital marketplace.

These factors all have the potential to impact the success of a video game, as defined in the first chapter of this report. In order to keep analysis focused and concise, these factors will be split into four key categories revolving around video games' release, type, contents, and popularity. Within these groupings, the factors will be analysed, and the results will be discussed.

4.2. Video Game Release Analysis

The first aspect of analysis revolves around the release of a video game. The factors discussed in this category will be the release date, game price, publishing company and age restrictions. The results of this analysis can be seen across four visualisations in the dashboard shown in Figure 4-1 below.

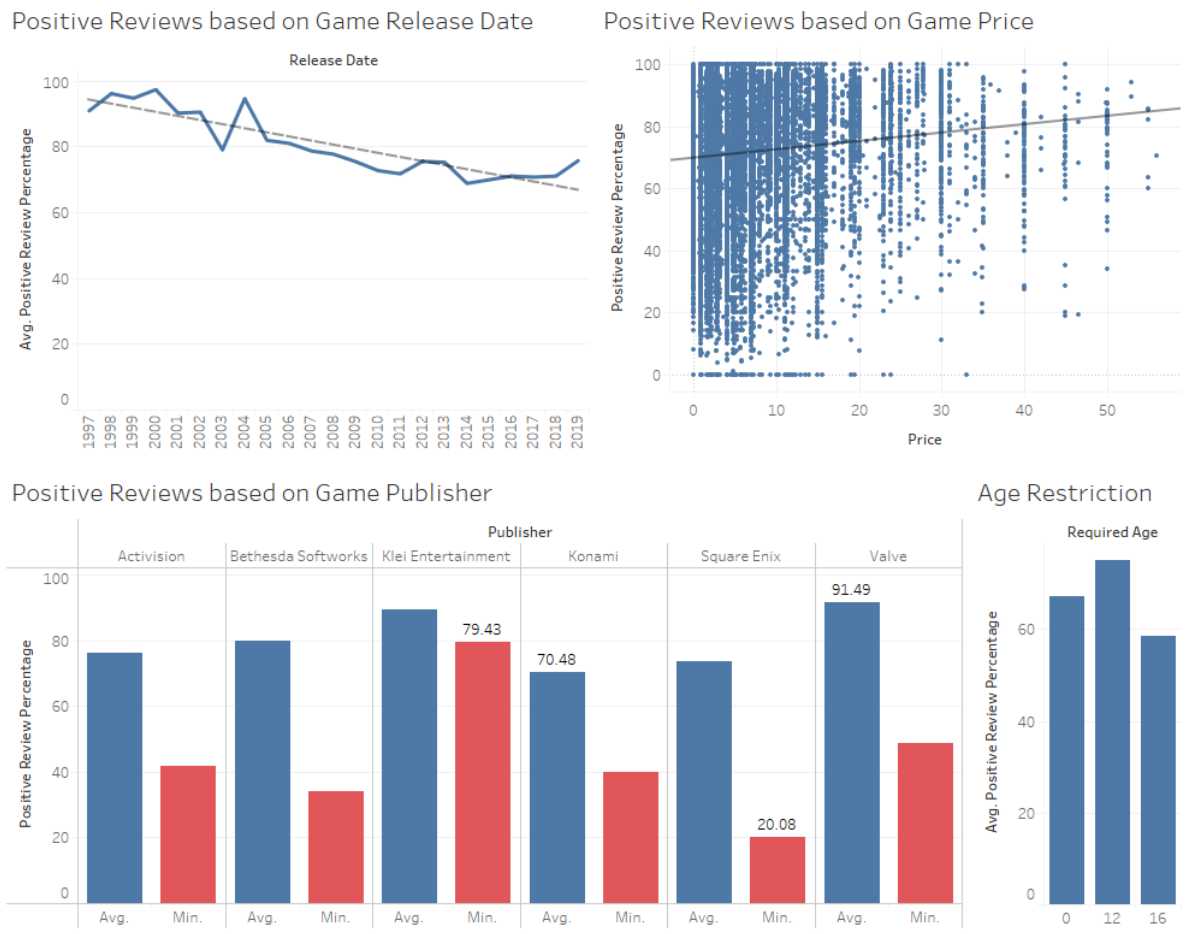

Figure 4-1 - Release analysis dashboard displaying influence of release date, game price, publishing company and age restrictions

As can be seen, the factors revolving around the release of a video game have a large correlation with the ratio of positive user reviews made. The influence of each factor will be discussed in a list below:

- ***Release date*** – The chart shows a clear negative correlation between the video game's release date and the average positive review percentage. This means that older games receive higher reviews on average than newer ones do. This could be due to a nostalgic factor, with players leaving higher reviews on games they are already familiar with, while newer games receive much more realistic levels of scrutiny.
- ***Age restrictions*** – We can see that the positive review ratio does appear to be impacted by different age restrictions, however, no clear pattern can be identified from the graphic.
- ***Publishing company*** – It appears that the publishing company has a big influence on the success of a video game, with the average difference in positive review percentage noticeable between different companies. However, the largest differences can be seen when looking at the lowest positive review percentage for each company. This could be because some companies publish many more video games or aim for much larger audiences.
- ***Video game price*** – In this visualisation we can see a slight positive correlation of video game price and average positive review percentage. This could be because games that cost more are generally produced to a higher quality, or that users that spend a larger amount of money ensure that they will enjoy a game before they purchase it.

Overall, it can be seen that the details of a game's release play a large role in its determining its success, by the definitions determined for this project.

4.3. Video Game Type Analysis

The next aspect of analysis refers to how the type of video game affects its success. The factors discussed in this category are the language support, categories, and genres of a video game. The results of this analysis can be seen across three visualisations in the dashboard shown in Figure 4-2 below.

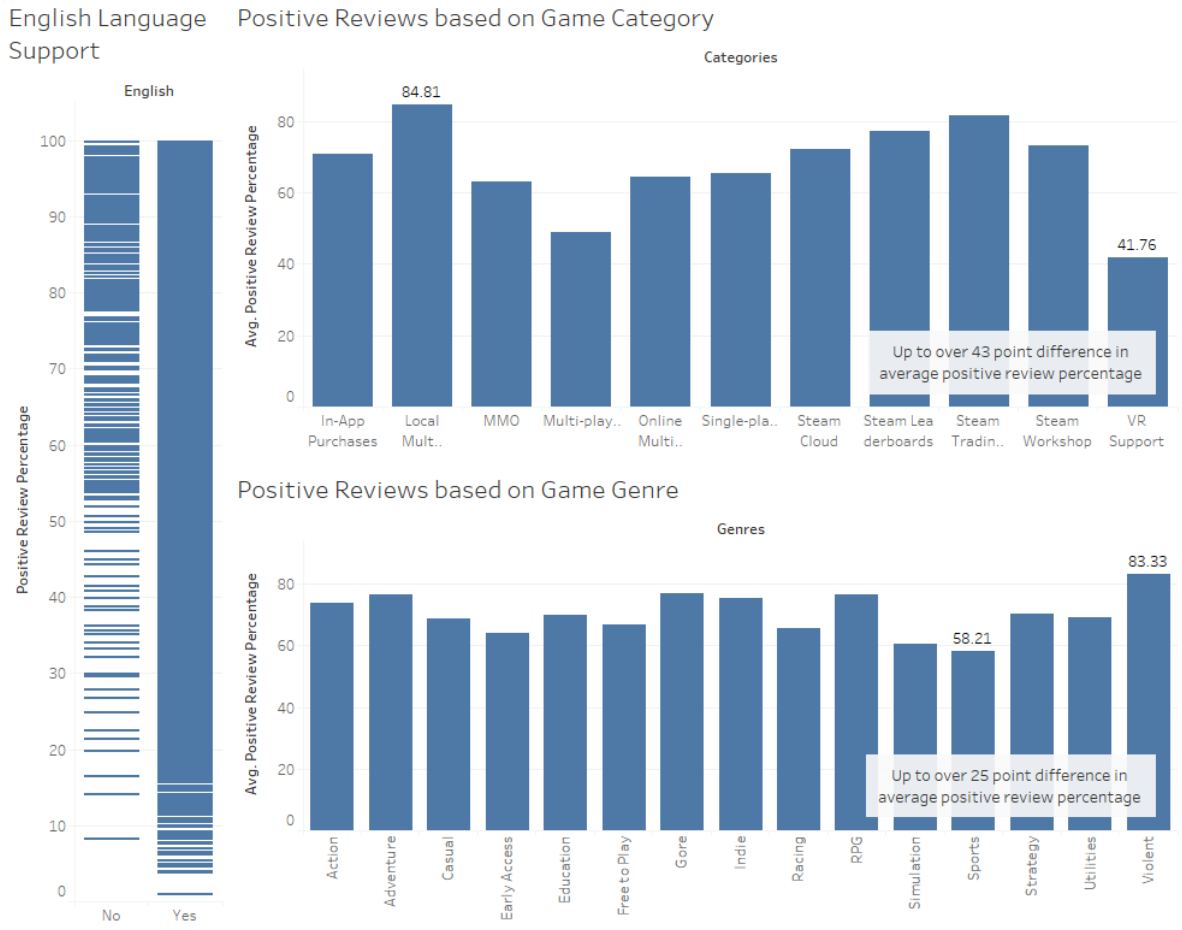

Figure 4-2 - Type analysis dashboard displaying influence of language support, categories, and genres

As can be seen, the factors revolving around the type of a video game have some correlation with the ratio of positive user reviews made but are not as significant. The influence of each factor will be discussed in a list below:

- ***Language support*** – It can be seen that there are many more reviews on made on English-supporting video games, and these reviews have a much larger range of positive review ratios. By contrast, games that do not support the English language seem to have a much higher concentration of positive review percentages, possibly due to bias with reviews made solely by the community that can understand the original language of the video game.
- ***Video game genre*** – We can see that the video game gerne seems to have a mild impact on the average positive review ratio of the video game, which could be due to varied expectations or scrutiny from the community interested in that particular game.

- ***Video game category*** – A significantly larger correlation can be seen with video game categories, with games that offer local multiplayer having a significantly larger average positive review percentage than games which offer VR support. This could be due to the quality of implementation.

From this analysis, it can be seen that the type of video game in question incurs variation in its success, however, the results do not highlight any conclusive trends that cannot be explained with community bias.

4.4. Video Game Contents Analysis

The third category of analysis refers to how the contents of video game affects its success. The discussion within this category will revolve around the average and median playtime, as well as the number of achievements available. The results of this analysis can be seen across two visualisations in the dashboard shown in Figure 4-3 below.

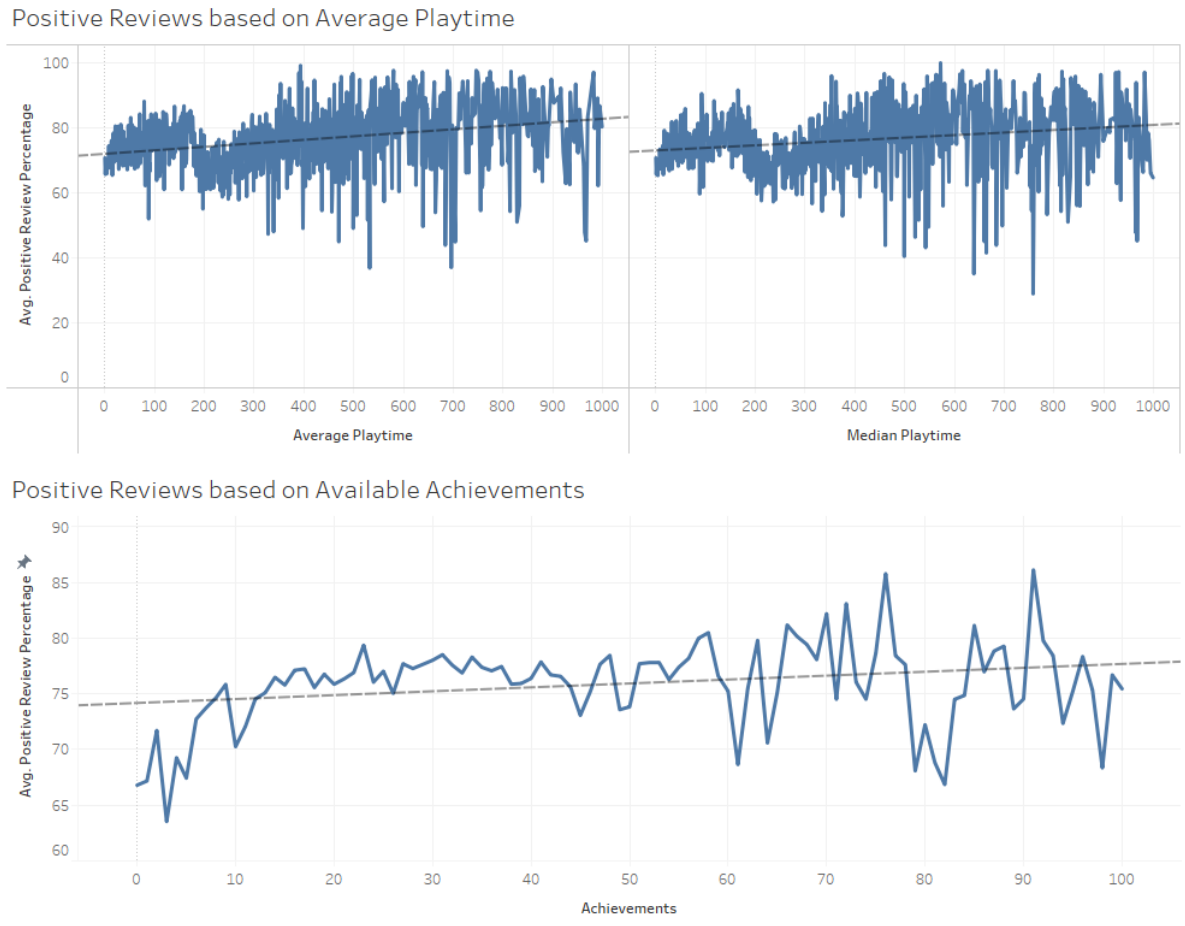

Figure 4-3 - Contents analysis dashboard displaying influence of average and median playtime, as well as available achievements

As can be seen, the factors involved in the contents of a video game have very minor correlation with the ratio of positive user reviews made. The influence of each factor will be discussed in a list below:

- ***Average and median playtime*** – As can be seen from the graphics, both average and median playtime shows a very small positive corelation with the average positive review ratio. This could be by an assumption that a player who plays a game more is more likely to enjoy it.

- ***Number of achievements available*** – Again, a very minor positive correlation can be seen in the number of achievements available and the average positive review ratio. It would make sense that more awards/goals within a game, the more likely a player is to stay engaged with the game.

From this analysis, it can be seen that the contents of a video game do not seem to have a significant influence on its success criteria.

4.5. Video Game Popularity Analysis

The final category of analysis refers to the popularity of a video game, and how that affects its success. The discussion will focus on the factors of ownership and total review count. The results of this analysis can be seen across three visualisations in the dashboard shown in Figure 4-4 below.

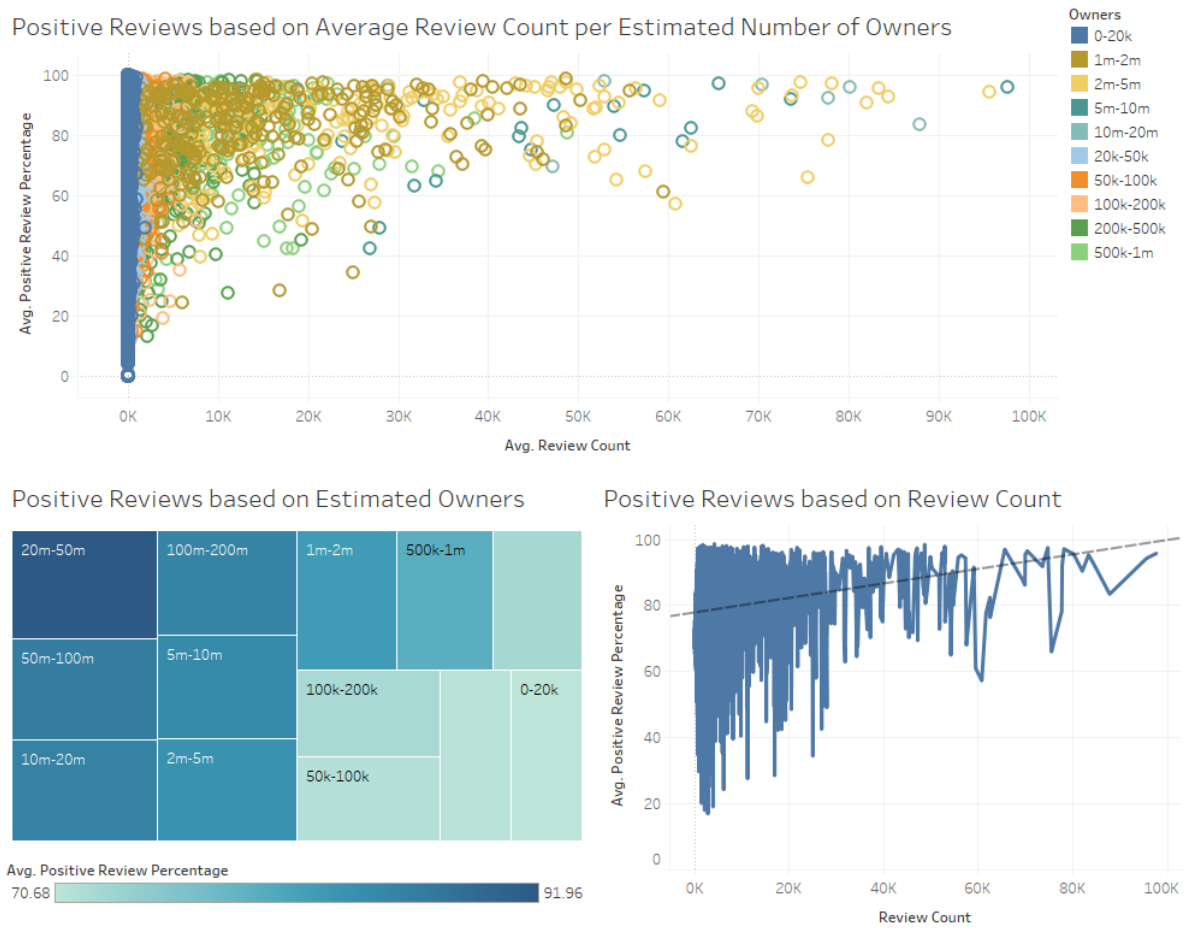

Figure 4-4 - Popularity analysis dashboard displaying influence of owner and review count

As shown in the visualisations, the number of owners of a game and user reviews left on it seem to have a high correlation with its success. The influence of each factor will be discussed below:

- ***Estimated number of owners*** – As can be seen, the more people own a game, the larger the number of reviews that are made, but also the higher the ratio of positive reviews left. This could be due to the fact that, if a game garnishes a large player base, it is likely because it has the qualities to interest a large number of players.

- ***Number of user reviews made*** – Likewise, the more reviews left the better they are on average, it seems. This could be due to the fact that trending titles are likely to be trending for a positive reason, garnering positive feedback from a large percentage of the reviews.

From this analysis, it can be seen that the popularity of a game appears to have a large influence on its success.

4.6. Critical Evaluation and Summary

From the analysis conducted it would appear that the key factors which appear to influence the success of a video game the most are related to its **release specification and its popularity**. These factors such as the release date and price, but also the number of owners and overall total of user reviews made. To conclude, the methodology and analysis steps performed in this chapter will be critically evaluated.

It is important to highlight that most of the analysis performed revolved around the comparison of only two variables. It could be argued that multivariate analysis could provide clearer, more detailed results. Additionally, it is significant to note that many of the visualisations displayed used filters to highlight a specific portion of the data. This was mostly done as a means to remove outliers and improve the granularity of the charts, however, analysis which revolved around genres and categories only used a small subset of the data and did not consider entries with combinations of factors.

Finally, many of the results gathered were based on trend lines and averages. It could be argued that different aggregation of data would provide different results, and that all bounds of the positive review percentage value should be evaluated for each factor.

5. Visualisations

The creation of each visualisation started with the concept of the data that needed to be explored. As highlighted in the previous chapter, each factor had to be explored to determine its influence on the success criteria established at the beginning of the report. As such, it made sense to use one visualisation to represent the influence of one factor on the positive review ratio.

5.1. Visualisation Choice and Justification

The type of charts used was determined on a case-by-case basis, with the majority of visualisations picked specifically to match their data type. For example, when reviewing the effect of average playtime on the positive review percentage, a line chart was used as both features were displayed on continuous axis, so it made sense to map them with a line chart. An exception to this was used for the price chart, as it was beneficial to show not only the trend of the data, but also the concentration of reviews amongst differently price points.

Likewise, charts which utilised discrete values, such as the genre or category of the game, or the publishing company, were displayed as bard charts, so that individual elements can be directly compared between each other. Once again, an exception to this was made to display the influence of the number of owners on the positive review ratio, which was displayed as a tree map. The reason for this is partially aesthetic, to break up the use of bar charts and to show the data in a varied-yet-visually-appealing way, but also to visually highlight the pattern across a colour spectrum.

Finally, the visualisations were placed within relevant dashboards to help contextualise the results. Though the factor charts could be evaluated and analysed individually, in many instances it makes sense to look at them together for added context. An example of this can be seen in the popularity dashboard, where a clear distinction can be seen between the number of owners of the video game

and the number of reviews that have been made. In this instance, a multivariate analysis visualisation was used to highlight this relationship, however, across other dashboards these relationships were left noticeable across the different bivariate chart.

5.2. Colour Scheme and General Aesthetics

The colour scheme used for the visualisations largely follows a default schema suggested by the Tableau software. The reason behind this is that the colour scheme used is a good option for colour-blind viewers and provides a great alternative to the commonly used red and green colours. In visualisations with multiple colours, contrasting and easily identifiable colours were chosen, and in the tree map a suitable colour-blind-friendly scheme was chosen for the colour spectrum used to visualise the differences.

As for the general aesthetic of the visualisations, it was aimed to maximise the data-ink ratio and ensure clarity, while still providing the user with relevant explanatory labels and annotations. Some visualisations may appear cluttered due to the amount of data displayed, and a potential improvement would be to aggregate this data in order to display the trend without necessarily showing all the data to the user.

5.3. Feedback and Improvement

After the initial analysis of the data and the creation of the visualisations, they results were shown to some family members and friends that also share an interest in video game development and or consumption. The aim with this was to gather feedback to improve the way that results are visualised.

The key pieces of feedback gathered was that too much data was being shown in individual visualisations, and that some visualisations are complicated and could benefit from annotations or labels to help contextualise them. An example of this feedback being implemented can be seen in Figure 5-1 shown below.

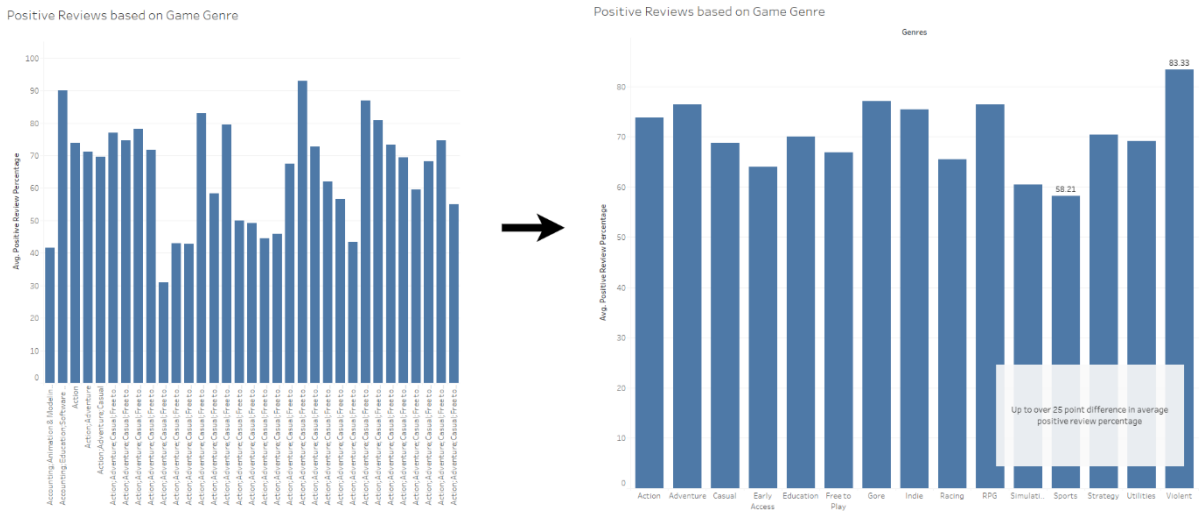

Figure 5-1 - Using filters to limit the amount of data displayed and adding labels and annotations to contextualise result

This feedback was implemented to varying degrees across all visualisations. Another notable example of this can be seen on the contents dashboard, in Figure 4-3, where both the y and x axis are filtered, with the y axis starting at the value 60 to display more data on the screen, and the x axis being limited to maximum value of 100, to limit outliers far beyond this range from distorting the data displayed.

6. References

Sims, R.L. (2000). *Bivariate data analysis: A practical guide.* Nova Publishers.

Vanawat, N. (2021). *How To Perform Exploratory Data Analysis -A Guide for Beginners. How To Perform Exploratory Data Analysis -A Guide for Beginners.* Available from: https://www.analyticsvidhya.com/blog/2021/08/how-to-perform-exploratory-data-analysis-a-guide-for-beginners/#:~:text=optimal%20factor%20settings-,Why%20EDA%20is%20important%3F,better%20before%20making%20any%20assumptions. [Accessed 15/05/2022].

Vigni, M.L. et al. (2013). Exploratory data analysis. *Data handling in science and technology.* Elsevier, 55-126.

Wikipedia (2022). *Steam (service). Steam (service).* Available from: https://en.wikipedia.org/wiki/Steam_(service) [Accessed 14/05/2022].