ERASMUS UNIVERSITEIT ROTTERDAM

The effects of partnering with Superstars on Video Games' Quality and Video Games' Sales: A Social Network Perspective

Author:

Victor Raadsheer (619175)

Supervisor:

Dr. Nuno Almeida Camacho

Co-Reader:

Prof. Dr. Dennis Fok

08-08-2023

Abstract

Research on *superstar* phenomena using Social Network Analyses is currently insufficiently researched in the academic field. *Superstars* are currently defined in different ways by different authors. In this study, a new definition is provided to analyse whether video game developers that work together with *superstar*s (other video game developers with high connections) are releasing video games of higher quality and higher sales. The data (N = 18,893) was partially collected from Kaggle and partially scraped from the website PlayTracker. One Quasi-Binomial model and two general multivariate regression models were set up to find the influence of *superstar* presence on video game quality and video game sales. The results showed that the presence of a *superstar* is statistically significant and has a positive effect on video game quality as well as video game sales.

1 Introduction

The *small world* hypothesis is a principle that states that people are connected in a certain number of steps. This concept was originally coined by Stanley Milgram (1967). His research on this concept consisted of a series of experiments exploring how well-connected social networks of people are within the United States. He found that the average length of a social network path is around five to six steps. Because of his findings, his research became closely related to another similar concept, *six degrees of separation*. This concept states that all people are connected on six or fewer connections.

The way to measure the length of a social network is by giving the connection steps a certain numeric value. For example, Goyal et al. (2004) operationalized this distance by taking two persons. When person A and person B know each other, they are at a distance of 1, and if they do not know each other, they are at a distance of 2. Goyal et al. (2004) applied this method to perform an in-depth study on exploring the small world hypothesis by analysing the interconnectedness of academic economists. They modelled the co-authorship links of economists from the period of 1970 to 1999, split into three ten-year periods. They found that over the years, the distance between economists had shrunk considerably, even though the clustering remained high.

More specifically, for analysing the co-authorship Goyal et al. (2004) used a *Social Network Analysis* method, which seemed incredibly useful to explore the concept of a small world hypothesis. At present, using Social Network Analysis to study the small world hypothesis has found quite a bit of coverage in academic literature. Prior research has examined social networks in a variety of contexts. Some examples include co-authorship networks (Goyal et al., 2004), networks of ideators (Stephen, Zubcsek, & Goldenberg, 2016), networks of companies and their performance (Ye & Li, 2022), and inter-organizational networks of employees (Cross, Borgatti, Parker, 2002). Currently, most literature seems to use this method on an individual level. Although some examples exist of this method being used on higher levels, such as a group or organizational level, literature is relatively scarce. For instance, research on interfirm cooperation has found that cooperation can be both a driver for new product development, efficiency, and other benefits (Wuyts, Dutta, & Stremersch, 2004; Gulati, Lavie, & Singh, 2009), as well as oppose product innovation if cooperative trust is too excessive (Molina-Morales & Martínez-Fernández, 2009). Overall, research on an interfirm level does exist, but it is fewer than on an individual level and at times contradicts each other. Therefore, in this paper, a way to expand the Social Network Analysis to a company perspective is presented.

To operationalize this further, I will apply this method to the gaming industry. The reason for choosing this specific field is twofold. First, some video game developers frequently work together to produce new video games. This leads to a relatively easy connectivity measurement between developers. Second, even though game developers work together relatively often, they are still direct competitors (Zackariasson & Wilson, 2010). This gives an interesting avenue to analyse cooperation between firms.

Another seemingly closely related concept is the concept of *superstars*. Academic research covers this concept often, though authors define and apply this concept differently. For example, Binken & Stremersch (2008) apply it to video game titles. They name video games of exceptionally high-quality *superstars* (Rosen, 1981; Binken & Stremersch, 2008). As these games are of significantly higher quality, they are characterized by a significant disproportionate pay-off in comparison to other video games. Nonetheless, in this paper, a different approach that is more akin to the approach of Goyal et al. (2004) will be used. They classify authors of economics papers as *stars,* which are economists who write with many other economists. Of these other economists, they have few co-authors and generally do not write with each other. This is then operationalized by analysing the degree centrality of specific authors. In this paper, this concept will be applied to video game developers instead.

This research commences by proposing and answering three “exploratory” research questions. The reason for formulating these without a hypothesis is that previous literature is scarce on these topics. Therefore, these questions will be answered in a descriptive analysis fashion.

> Exploratory Question 1: *“How much are video game developers working together?”*
>
> Exploratory Question 2: *“How does the video game developer cooperation change over time?”*
>
> Exploratory Question 3: *“Are video game developers cooperating and what is the nature of cooperation? (I.e., are superstars connected to non-superstars; Are non-superstars connected to non-superstars?)*

After answering these questions, the main research questions will be answered. Three main Research Questions have been composed, which are as follows:

> Research Question 1: *“Do superstar firms publish higher quality video games?”*
>
> Research Question 2: *“To what degree does cooperating with superstars pay off in terms of video game sales?”*
>
> Research Question 3: *“Does video game quality pay off in terms of sales?”*

By utilising a Social Network Analysis on video game developers, a potential avenue is created to see how well-connected the video game industry is. On top of this, further inferences about the interconnectedness of firms and whether it is interesting for firms to work together on a more horizontal level could be made.

2 Superstars: Literature Review

2.1 Construct Origins and Definition

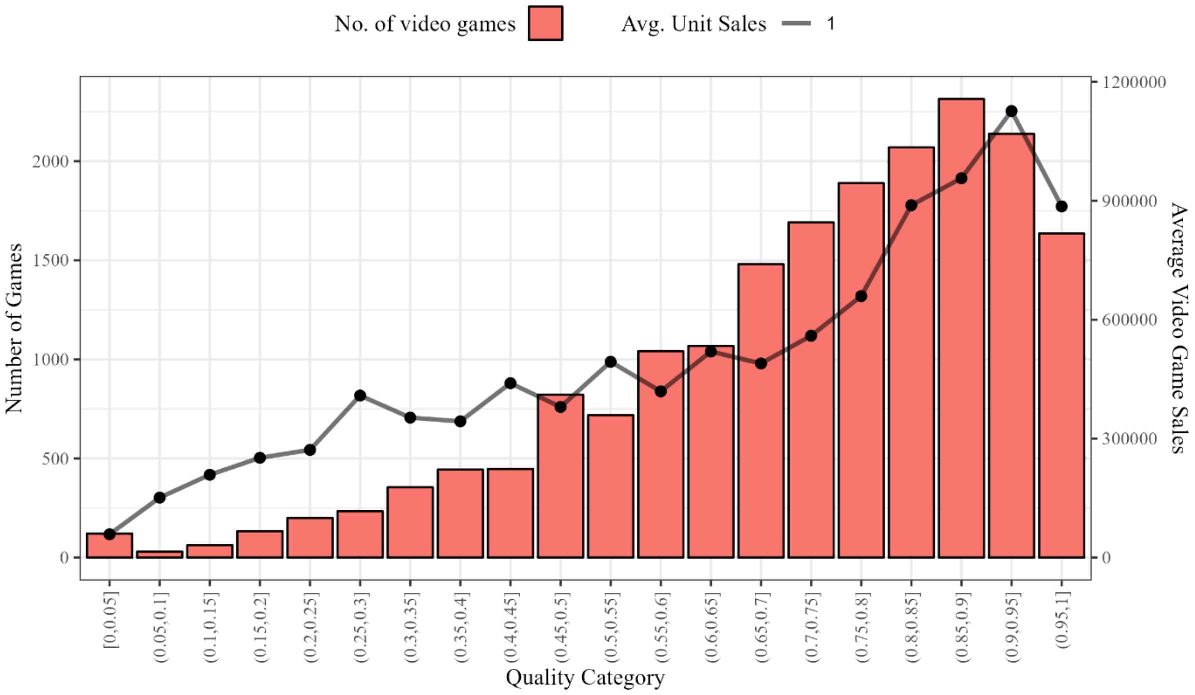
In this section, the existing academic literature on the *superstar*-concepts will be discussed. Over the years, the concepts of *superstars* have been applied quite differently and given different names by different authors. The earliest mentions of *superstars* reach back to Rosen (1981), which he deems to be a phenomenon where a relatively small number of people earn enormous amounts of money and dominate the activities in which they engage. In this paper, a different definition will be proposed, in which a *superstar*-video game developer is one that develops games with many other developers. In the following paragraphs, it will be elaborated upon as to why this concept was chosen.

Blinken & Stremersch (2008) used this *superstar* concept and applied it to the video game industry. Instead of discerning *superstars* on a firm level, they discerned them on both a product- and industry-level, giving so-called *superstar*-products three important characteristics. Overall, *superstars* have a disproportionally large payoff, generally because a *superstar* has unique attributes that drive this payoff. Furthermore, in a *superstar*-industry, there is a small number of *superstars* that dominate their industry. Lastly, due to the scarce nature of high-quality, a *superstar*-industry shows increasing returns to quality (Rosen, 1981; Binken & Stremersch, 2008). Here, it is clear that the *superstar* term is used to characterize a product and industry, and not so much a firm. Based on these characteristics, it is found that among other industries, such as the music and movie industry, the video game industry adheres to these characteristics as well (Binken & Stremersch, 2008; Gretz et al., 2019; Chung & Cox, 1994; De Vany & Walls, 1996; McAndrew & Everett, 2015).

Other research affirms the idea that there is a heterogeneity in the quality of games, making only a small number of video game titles actual *superstar*-products (Corts & Ledermand, 2009).

Looking at Figure 1, this trend is seen visualized, where higher quality games also demand more sales. One thing to note is that this figure is slightly different than observed in previous literature (Binken & Stremersch, 2008). Nowadays, there are many more video games belonging to a high-quality category, demanding a relatively smaller number of sales. Nonetheless, based on the three characteristics of Binken & Stremersch (2008), it is possible to consider the video game industry a *superstar*-industry.

Figure 1: Number of Video Games Per Quality Category



Previous literature also covers how software and hardware sales in a *superstar* industry interact with each other. For example, it mainly delves into whether software drives hardware sales, or vice versa. Papers find that higher quality software, i.e. *superstar software*, drives the adaptation of video game platforms, such as consoles.

Even though some authors use the *superstar* concept to describe a product or industry, others again use it to describe a certain type of firm (Tambe et al., 2020; Autor et al., 2020; Gutiérrez & Philippon, 2019; Ayyagari et al., 2019). These papers generally refer to *superstar*-firms, which are firms that are unique in their ability to scale up innovations (Tambe et al, 2020; Autor et al, 2020). Autor et al. (2020) further expand this concept by proposing that *superstar*-firms are the most productive firms in a sector. They have above-average markups and below-average labour shares, leading to a "winner take most" mechanism in such industries. They further claim that the rise of these *superstar* firms could happen due to increased consumer sensitivity to quality-adjusted prices from greater competition.

Generally speaking, it is clear that different authors understand and apply the concept of *superstars* differently. However, for the purpose of this paper, its own definition of *superstar*-game developers will be offered. This paper departs from a network perspective, rather than an output or outcome perspective. The approach will be based on the one that is used by Goyal et al. (2004). They define *superstar*-economists as those who write with many other co-authors. Of these co-authors, most do not write with each other. Similarly, *superstar*-game developers will be defined as those who develop games with many other game developers. Those game developers that are not a *superstar* are then defined as having a much smaller number of game-developing partners than the *superstar*.

2.2 Construct Operationalization

Now a broad definition of *superstar*-game developers has been made, the next step is to find a more concrete operationalization on how to appropriately measure them. In their paper, Goyal et al. (2004) define a *superstar*-author as a person who has a large number of co-authors, specifically 25 times the average number of co-authors. The issue with their specification is that it is somewhat unclear as to why this specific number was chosen as it is not elaborated upon any further. Therefore, it seems to be somewhat arbitrary how they concretely operationalize *superstar*-authors.

When looking at other research for a *superstar* specification, it is found that most papers simply acknowledge the fact that certain industries have *superstar* characteristics, providing numerical examples of the phenomena. For example, Chung & Cox (1994) found that 10.8% of musical performers obtain more than 43.1% of the total gold-records in the industry, making them *superstar* players. Similarly, McAndrew & Everett (2015) define a group of 13% as higher-performing musicians as *superstars* and cluster them in a similar group, and 87% of lower-performing musicians as *non-superstars*. De Vany & Walls (1996) found that 20% of the films earned 80% of the box office revenues, finding that a small group of movies dominate the industry. However, none of these examples provide a clear cut-off and generally classify a *superstar* group to be between 10 to 20 percent.

The only two researches found providing a concrete specification for their *superstar* definition are from Binken & Stremersch (2008) and Gretz et al. (2019). They both use the same specification based on a quality rating of 90 or higher out of 100. Binken & Stremersch (2008) identify that software in this category typically sell more than 1 million units and show an increasing return to quality at this point. Perhaps the most noteworthy observation is that out of their 5800 software titles, only 89 can be considered *superstars* by their definition. This is only about 1.53% of their entire sample. For Gretz et al. (2019), 259 out of 7424 games can be classified as *superstars*, which is about 3.49%.

As the *superstar* definition in this paper is based on the number of developer connections, the direct quality specification that Binken & Stremersch (2008) and Gretz et al. (2019) use cannot directly be applied to this paper. Instead, the specification in this paper is based on the relative size of their *superstar* groups. After running a Social Network Analysis the degree centrality measures were retrieved, which are displayed in Table 1. Here, the total number of links and how often they occur within the data can be seen. One thing that is immediately noticed is that most of the developers have zero links, which comprises about 83.9% of the total developers. Furthermore, the number of developers having more links quickly falls off. About 1,668 (13.5%) developers have 1 link, 202 (1.6%) developers have 2 links, and after that it dives under the 1% margin. To make the group of star developers more comprisable, but still keep it in line with prior literary categorizations, a *superstar*-game developer will be specified as one that has 3 or more links with other developers. This will define a total of 117 (0.945%) developers as *superstars*, which encompasses just under one percent of all developers.

Table 1 – Number of Times Each Linkage Occurs

Number of Links	Occurrence	% Occurrence
0	10,388	83.943
1	1,668	13.479
2	202	1.632
3	64	0.517
4	13	0.105
5	12	0.097
6	9	0.073
7	8	0.065
8	2	0.016
9	1	0.008
10	3	0.024
11	2	0.016
12	0	0.000
13	0	0.000
14	0	0.000
15	0	0.000
16	2	0.016
17	1	0.008

Note. N = 12,375

Concluding, a *superstar* is defined by how often a developer works together with another developer. A *superstar* works together with other developers a lot, of which the other developers are not working together nearly as often as the *superstar*. As no prior literature uses a clear specification to determine what a *superstar* is, this paper determines a *superstar*-developer to be one with 3 or more links with other developers, based on degree centrality. This comprises about 1% of the total developers in the data. In Table 2 it is possible to review 9 published papers on *superstar* concepts and their main takeaways for this paper.

Table 2 – Literature Table with top 9 papers about *superstars*

Paper	Journal	Google Cites	Key Theme	Takeaways
Binken & Stremersch (2008)	*Journal of Marketing*	199	Definition & Operationalization	• Higher quality video games demand higher sales. • Video game industry is *superstar* industry. • *Superstar* specification when quality is 90 or higher out of 100. • 1.53% of video games are *superstars*
Goyal et al. (2004)	*Journal of Political Economy*	638	Definition & Operationalization	• *Superstars* are authors who write with many other authors. • Those who are not *superstar* have smaller number of co-authors. • *Superstars* generate higher output
Rosen (1981)	*The American Economic Review*	4,279	Definition	• Original concept of *superstars* • *Superstars* are those who earn a lot of monetary rewards and dominate in activities in which they engage.
Gretz. et al. (2019)	*Journal of the Academy of Marketing Science*	15	Operationalization	• *Superstar* specification when quality is 90 or higher out of 100. • 3.49% of video games are *superstars*
Chung & Cox (1994)	*Review of Economics and Statistics*	225	Definition	• Small amount of musical performers demand most of the rewards.
McAndrew & Everett (2015)	*Cultural Sociology*	85	Definition	• Small amount of higher-performing musical artists
De Vany & Walls (1996)	*Economic Journal*	536	Definition	• Small amount of films demand most of the revenues.
Tambe et al. (2020)	*National Bureau of Economic Research*	73	Definition	• Refers to *superstar* firms with unique characteristics
Autor et al. (2020)	*Quarterly Journal of Economics*	2,294	Definition	• *Superstar* firms are most productive firms in a sector • Above-average markup & below-average labour share

2.3 Superstar Performance

Lastly, it will be discussed what the existing literature has to say about the ability of *superstars* to generate more output and/or higher quality output. Overall, the results seem to vary slightly. When considering the network statistics of Goyal et al. (2004), we find that *superstars* do publish more papers. This indicates that *superstars* do have a higher output than *non-superstars*. This is further reinforced by looking at the top 100 authors in their dataset, who publish way more than the overall average. Though, this is not fully consistent with the findings of Newman (2001), who considered the author collaboration statistics of scientists publishing biomedical, physics, and computer science papers. He finds that papers with more authors and papers with high collaboration are less common overall. Even though a *superstar* concept is not explicitly mentioned in his research, his findings do suggest that *superstars* do not publish more papers.

When considering whether *superstars* generate higher quality, social science networks are considered. It is found that authors who publish in complete or priority journals are more likely to be at the core of the networks (Moody, 2004). This could suggest that *superstars* are publishing higher quality, due to them publishing in journals that are generally more renowned. This is reinforced by Björk & Magnusson (2009) who find that when groups show higher network connectivity, the quality of ideas produced within a firm increase. Additionally, Tsai (2001) finds evidence that both innovation and performance of a company increase when central network positions are occupied by said company. Both Björk & Magnusson (2009) and Tsai (2001) suggest that this is due to the exposure, acquisition and sharing & transferring of knowledge of new information within a network. This lends further credibility that *superstars* potentially generate higher quality due to their higher connectivity.

Even though Moody (2004), Björk & Magnusson (2009), and Tsai (2001) find potential evidence for *superstars* generating higher quality, none of them specifically focused on *superstars* in the same way this paper does. In general, literature seems to be lacking on whether *superstars* generate higher quality, though some evidence exists. Furthermore, the lack of literature when considering the quantity generated by *superstars* seems to hold as well. Even though Goyal et al. (2004) provide the most concrete evidence for higher quantity generated, it does seem to be somewhat contradictory when considering the research of Newman (2001).

3 Hypotheses

Due to the lack of literature and sometimes contradictory results of the few existing papers, an interesting avenue is opened to research how *superstars* could affect video game quality and video game sales.

As discussed in the literature review, Autor et al. (2020) find *superstar*-firms to be firms with above-average mark-ups and below-average labour shares which reap disproportionate monetary rewards and find substantiated proof for these mark-ups and below-average labour shares. Even though they do not find any evidence for the quality of their products, they do hypothesize about the rise in *superstar* firms could be partially explained by customers becoming more sensitive to quality-adjusted prices.

Furthermore, some authors find evidence that higher connectivity does increase quality, suggesting *superstars* could indeed increase quality (Moody, 2004; Björk & Magnusson, 2009; Tsai, 2001). This higher connectivity could lead to a higher exchange of information and knowledge, which in turn leads to better ideas, and ultimately a potential increase in quality. This leads to the belief that *superstar* firms do create higher quality products, thus leading us to the hypothesis of Research Question 1:

H_{0A}: Superstar video game developers do not publish higher quality video games as compared to non-superstar video game developers.

H_{1A}: Superstar video game developers do publish higher quality video games as compared to non-superstar video game developers.

Regarding Research Question 2, Goyal et al. (2004) found the most concrete evidence that *superstars* generate higher output, leading to believe that *superstars* therefore also generate higher sales. Though, this might be contradicted by Newman (2001). Therefore, another paper written by Klimas & Czakon (2018) is considered. They focus specifically on coopetition[1] between video game developers. They found that within the Polish video game industry, coopetition is a popular strategy. About 68% of the 506 video game developer firms surveyed participate in coopetition. They further found that *organizational innovativeness* and its corresponding dimensions are positively related to coopetition. More specifically, video game developers engage in coopetition due to 1) an openness to innovate, 2) strategic innovative focus, and 3) extrinsic monetary motivation. While this does not clarify whether working together with *superstars* leads to higher sales in terms of video games, it does state that monetary motivation is a driver for coopetition. Therefore, based on this, the hypothesis for Research Question 2 would be as follows:

H_{0B}: *Video games co-developed with a superstar do not have higher sales than video games developed without a superstar involved.*

H_{1B}: *Video games co-developed with a superstar have higher sales than video games developed without a superstar involved.*

A third important question is whether higher video game quality would result in higher sales. There is a plethora of literature discussing this topic. Generally speaking, spending resources on quality is found to increase costs. Though, it is also found that increased quality can be traced back to the profits of a firm (Rust et al., 1995; Narasimhan, Ghosh & Mendez, 1993). This effect could be explained through, for example, retaining customer loyalty (Purwati et al., 2020), which in turn could drive long-term sales for a company. It is further found that improved

[1] 'Coopetition' is a term used to indicate a collaborative relationship of a firm with a competitor, likely originating as a mixture between 'cooperation' and 'competition' (Klimas & Czakon, 2018).

general product quality drives initial sales, and can even lead to repeat sales (De Langhe et al., 2016).

Even though quality is found to increase sales, sales are also influenced heavily by other variables, such as price and advertising campaigns (Köcher & Köcher, 2018; Yang, Cao, Wang, Lu, 2022; Thiesing, Middelberg, Vornberger, 1995). Depending on the product, other factors could also play a role, such as healthiness regarding food products (Morano et al., 2018).

Since the rise of the Internet, it seems that online user ratings have become an increasingly important tool to assess the quality of a product. It is found that the average user rating of a product has become a highly significant driver of sales for many products across several industries. When a user rating is favourable, it effectively reduces the quality uncertainty, converting people from potentially not buying the product to buying the product (Hu, Lui & Zhang, 2008). This effectively means that people are influenced by the reviews of others, influencing their buying decision, suggesting that network effects are at play here as well (Binken & Stremersch, 2008). Due to this, it is important to examine whether the quality of a video game influences video game sales as well.

This leads to the following hypothesis regarding Research Question 3:

H_{0C}: Video games with a higher quality do not demand more sales than video games with a lower quality.

H_{1C}: Video games with a higher quality do demand more sales than video games with a lower quality.

4 Data

In this chapter, the data collection, data cleaning, and descriptive statistics will be discussed[2]. The data is based on aggregated statistics from the online gaming platform Steam, consisting of a combination of several sources. The main dataset is obtained from Kaggle and contains data from 71.171 different games on the Steam platform (Roman, 2022).

The dataset contains variables on many video game statistics, such as the number of positive reviews, negative reviews, total reviews, developers, publishers, popular tags, and more. The data ranges from 1997 to 2023, from which the bulk was released between 2017 and 2022. Even though the data does not contain an indication of where the developer studio is from, it does seem that it covers studios from all over the world, as video game studios sometimes have names in different languages, such as Mandarin, Japanese, or Russian.

The dataset is collected from several sources, using both the Steam API and additional data from Steam Spy. It is updated monthly, and by the time of writing, the dataset was last updated on 4-4-2023. The data further contains an estimate of the number of people owning the game, however, the error margins are extremely large for specific games. Therefore, the website PlayTracker was scraped for a more accurate estimation of game ownership statistics.

PlayTracker[3] is a project that gathers public data from video game players around the world in a random sample and then extrapolates that random sample to estimate total values for video game statistics. As its statistics are based on samples of its user accounts, they are estimates, and can therefore be off due to various factors. For most of the estimations, PlayTracker is 90% certain that the estimates are accurate within a 10% confidence interval (Marijan, N.D.).

[2] Uncleaned data, cleaned data, code for scraping & cleaning, and code for models can be found on the following Github link: https://github.com/MetaKingDedede/MThesis-Code.git

[3] More can be found on the Playtracker website: https://playtracker.net/

Overall, PlayTracker offers information about several video game platforms, such as Steam, PlayStation, XBOX, and Origin. However, only the information for Steam is given for free, whereas more detailed statistics are locked behind a subscriber service offered through Patreon. On top of this, to keep the data mergeable with the dataset obtained from Kaggle, only the scraped observations of the Steam platform are considered.

PlayTracker was chosen to scrape over other websites that offer similar services, as the creator of this project allows the Steam data to be used for free. Similar projects such as SteamDB[4], SteamSpy[5], or VGInsights[6] either only give free samples, or lock the data behind Patreon paywalls. Furthermore, the way PlayTracker calculates estimated ownership is similar to other projects, making it somewhat arbitrary which one is chosen.

4.1 Cleaning

After the two datasets were obtained, they were cleaned and merged. Overall, a substantial number of observations from the original dataset were dropped. This was done for several reasons.

1. PlayTracker only tracks the games that are owned by its users, thus a smaller number of games had accurate ownership statistics. This is not deemed an issue, however, as a quick analysis shows that from the 71.716 original observations, about 57.360 games had an average playtime of zero, implying they had never been bought or played.
2. Some games did not have any developer values. These are games that are still on the Steam store and can still be bought, but do not have an official developer displayed anymore. One such example is the 2013 video game remake of the

[4] More about SteamDB can be found on: https://steamdb.info/
[5] More about SteamSpy can be found on: https://steamspy.com/
[6] More about VGInsights can be found on: https://vginsights.com/

Flashback title, which was originally released in 1992. This remake was developed by the studio *VectorCell* and published by *Ubisoft*. Due to the lack of success of the remake, the company suffered bankruptcy and ceased to exist (Chopard, 2004). This is presumably why their developer studio was taken off the product page on Steam[7], though it can still be bought as the video game license is likely owned by publisher *Ubisoft*.

3. Some observations were duplicated. The duplicates were promptly removed.
4. Some video game genres did not correspond to video games. The Steam store offers some additional products besides video games, which are still categorized under their genre system. Steam offers the additional genres of Animation & Modelling, Audio Production, Design & Illustration, Game Development, Photo Editing, Software Training, Utilities and Video Production. These genres do not contain any actual video game titles in the traditional sense, but rather other kinds of software such as video and audio editors.
5. Lastly, when creating and transforming some variables, some NA or extreme values were created, which were removed.

Overall, after all the cleaning and transforming of data, a total of 18,893 video game observations across all genres, from 12,375 different developer studios, released between June 1997 and February 2023 remained.

[7] Studio closure/discontinuation seems to be a common theme. Another example is the video game *Shallow Space*, presumably released by developer *Special Circumstances* (PCGamingWiki, 2022). The game was an early access title, meaning it was not yet a finished product at the time it was released. At some point during development, the developers abandoned the project. Even though there is no concrete evidence pointing towards this, it is possible that the developer studio disbanded shortly after.

4.2 Data Structure

In this part, brief comments on the descriptive statistics of the dependent and independent variables that are going to be used in the regression analyses will be made as found in Table 3, as well as explain how some of the variables were created.

Table 3 – Descriptives

	M	SD	Min	Max
Dependent Variables				
Video Game Sales ^{PT}	696,969	3,589,720	0	331,000,000
Log Video Game Sales* ^{PT}	12.271	1.022	6.909	19.618
Video Game Quality ^{S}	0.811	0.202	0	1
Independent Variables				
Superstar (Dummy) ^{S}	0.036	0.185	0	1
No. Developers ^{S}	1.090	0.420	1	15
No. Previous Games ^{S}	2.044	6.295	0	101
DLC Count ^{S}	0.811	5.521	0	461
Price ^{S}	7.912	8.981	0	99.99
Log Price ^{S}	1.719	1.022	0	4.615
Degree Centrality ^{S}	0.337	1.293	0	17
Genre: Action (Dummy) ^{S}	0.438	0.496	0	1
Genre: Adventure (Dummy) ^{S}	0.387	0.487	0	1
Genre: Casual (Dummy) ^{S}	0.375	0.484	0	1
Genre: Early Access (Dummy) ^{S}	0.066	0.248	0	1
Genre: Education (Dummy) ^{S}	0.000	0.016	0	1
Genre: Free To Play (Dummy) ^{S}	0.069	0.254	0	1
Genre: Gore (Dummy) ^{S}	0.010	0.099	0	1
Genre: Indie (Dummy) ^{S}	0.718	0.450	0	1
Genre: Massively Multiplayer (Dummy) ^{S}	0.026	0.158	0	1
Genre: Nudity (Dummy) ^{S}	0.004	0.063	0	1
Genre: RPG (Dummy) ^{S}	0.038	0.191	0	1
Genre: Racing (Dummy) ^{S}	0.165	0.372	0	1
Genre: Sexual Content (Dummy) ^{S}	0.004	0.061	0	1
Genre: Simulation (Dummy) ^{S}	0.186	0.389	0	1
Genre: Sports (Dummy) ^{S}	0.044	0.206	0	1
Genre: Strategy (Dummy) ^{S}	0.201	0.401	0	1
Genre: Violent (Dummy) ^{S}	0.016	0.125	0	1

Note. *M, SD, Min* and *Max* are used to represent the mean, standard deviation, minimum value and maximum value respectively. All variables have an N = 18,893, except for * which has N = 18,066. Video games can belong to more than one genre. ^{S} indicates origin from Steam, ^{PT} indicates origin from PlayTracker. For dummies, the 'mean' should be interpreted as percentage of sample taking the value 1. Please note that each game can be classified in multiple genres.

Video Game Sales

The first variable covered will be Video Game Sales. As stated in the data collection section, the sales statistics were scraped from the PlayTracker website. The sales are at least rounded to a thousand. This means that the lowest value after 0 sales, is 1000 sales. This means that the sales are not accurate to the exact sale, but they provide a rough but accurate indication.

Looking at Table 3, video game sales seem to have a very large standard deviation (*M = 696,969, SD = 3,589,720)*. The reason for this is most likely due to outliers. There are a few games that have attracted such a large player base over the years that they have seen incredible sales numbers. As can be seen by the maximum value, the largest game has sold over 331,000,000 units. This feat belongs to the video game *Counter-Strike: Global Offensive* from the developer *Valve*. To account for these outliers, the variable was transformed using a logarithm. This massively decreases the standard deviation (*M = 12.271, SD = 1.022)*, and provides it with a more normal distribution.

Video Game Quality

The Video Game Quality variable was created by taking the proportion of positive reviews divided by total reviews of a video game. This makes it a proportion between 0 and 1. Research on hotel reviews has shown that people place reviews to signal quality to other consumers and managers alike (Chevalier, Dover, Mayzlin, 2017). Furthermore, people can only leave a review if they paid and own the game on their Steam account, which reduces the possibility of competitors trying to negatively affect review scores (Mayzlin, Dover, Chevalier, 2012). This makes the proportion of reviews a useful proxy to determine the quality of a video game. When considering descriptives, it is seen that it is slightly skewed towards the higher review scores as it averages around 81.1% (*M = 0.811*) and a standard deviation of 20.2% (*SD* = 0.202*)*.

Superstar Game Developers

The *Superstar* Game Developers variable is a simple dummy variable. Using a Social Network Analysis, the degree centrality of each developer in the social network of all game developers in the Steam database was computed to identify superstar developers (the approach will be made precise within the methodology section). A dummy was created per video game, indicating whether or not a *superstar* developer was present during the development of the corresponding video game. As discussed in the literature review, a *superstar* is a game developer having relatively many connections in comparison to *non-superstar* developers. This was further operationalized as a developer having 3 or more links with other developers. Thus, a video game developer obtained the *superstar* tag if they had a degree centrality of 3 or more. In total, the mean shows that 3.6% (*M = 0.036)* of all video games in the dataset had at least one *superstar* developer present.

As was covered in the literature review section, *superstars* are relatively small groups within their corresponding context. The fact that about 3.6% of the video game titles in this dataset had at least one *superstar* developer involved seems to be in line with previous literary observations (Gretz et al., 2019, Binken & Stremersch, 2008).

Degree Centrality

The Degree Centrality variable is the degree centrality measures for the main video game developer per video game observations. It essentially measures how many neighbours a developer has, i.e. the number of edges connected to a node. This concept will be expanded upon in the Methodology section. Even though the degree centrality ranges from 0 to 17 links, it is skewed towards 0, as can be seen by the mean (*M = 0.337)*. This indicates that many developers have a lower number of links.

Number of Developers, Number of Previous Games, DLC Count

The Number of Developers, Number of Previous Games, and DLC (downloadable content) Count variables are all simple numeric count variables. The Number of Developers tracks how many developers were present in the development process of a video game. The Number of Previous Games tracks, per observation, how many games the first developer has developed before the development of the current video game. Lastly, the DLC Count tracks how many DLCs have been published for the game.

Price

The Price of a video game is measured in U.S. dollars and varies from \$0 to \$99,99. As with the Video Game Sales variable, the standard deviation of the price seems to be a bit larger than the mean (*M = 7.912, SD = 8.981).* When looking a little deeper into why the average price of a game seems to be only \$7.91, it is found that a relatively large number of video games have a very low price. Of the 18.893 video games in the dataset, 10.198 are under \$5.00, of which 4,662 have a price of \$1 or less. This skewness towards lower prices was brought down considerably by taking a logarithmic transformation (*M = 1.719, SD = 1.022).*

Genres

The Genre variables were created by checking which genre tags a game observation had associated with it, and making a binary variable whether that game had the corresponding tag or not. A video game could have several tags, i.e. a game could belong to both the action genre and casual genre at the same time. It is further important to note that most of the genre variables seem to be skewed towards 0, indicating that most video game observations do not contain that tag. The only observation going against this trend is the indie genre, which has a mean of 0.718, indicating that about 71.8% of the games published on Steam were Indie[8] games.

Year

Lastly, the Year variables will be considered. The descriptives for the year variables can be found in Table 4. When considering this table, it is immediately clear that the bulk of the games

[8] Short for *independent video game*. Generally, indie games are developed by individuals or smaller development teams. Indie games have smaller budgets, more unique mechanics, shorter stories, and more stylized arts in comparison to AAA games made by bigger companies (Khomych, 2022).

were released within the period 2014-2019. Especially the years 1997-2005 and 2023 lack observations. Nonetheless, for completeness' sake, all the years will be considered in the upcoming analyses.

Table 4 – Games Released on Steam Per Year

Years	Games Released
1997	2
1998	1
1999	3
2000	1
2001	3
2002	1
2003	3
2004	5
2005	6
2006	38
2007	58
2008	86
2009	160
2010	137
2011	179
2012	240
2013	359
2014	1330
2015	2194
2016	3434
2017	4420
2018	3786
2019	1287
2020	397
2021	355
2022	406
2023	12

Note. N = 18,893. Year variable originates from Steam data

5 Methodology

In this chapter, the methodology of this paper will be covered. This will be heavily based on the paper of Goyal et al. (2004). In this paper, it is crucial to properly operationalize the *superstar* status of each developer, as this variable is the main one used to study the impact of *superstar* status on video game sales and video game quality. Therefore, it is important to carefully describe how the Social Network Analysis will take shape.

5.1 Operationalizing Superstar Developers using Social Network Analysis

It is important to cover some basic notation of the features of a Social Network Analysis. First, we have $N = \{1,2,\dots,n\}$, which will be the set of nodes in a network, where *n* is the number of nodes in a set. In this paper, binary undirected links will be covered. The reason binary links will be covered is that even though there are cases where video game developers work together more than once, these are very sporadic and therefore do not result in many additional insights. For two different nodes $i, j \in N$, we will define $g_{i,j} \in \{0,1\}$ as a link between them, where $g_{i,j} = 1$ signifying a link, and $g_{i,j} = 0$ signifying no link. In this paper, game developers have a link when they have developed a game together, and game developers do no have a link if they have not developed a game together.

Game developers can belong to the same network if and only if there exists a path between them. A path between i and j can exist either if $g_{i,j} = 1$, or if there is a set of distinct intermediate co-game developers such that $g_{i,j_1} = g_{j_1,j_2} = \cdots = g_{j_n,j} = 1$. The collection of links will be denoted by g. A network will then be the set of nodes and the links between these sets of nodes, denoted by $G(n, g)$. The set of game developers who have a link with i will be defined as $N_i(G) = \{j \in N : g_{i,j} = 1\}$ in network G.

To see how important entities are in a network, different node centrality measures exist. From the different forms, only one will be used. This used measurement is *degree centrality*.

The simplest way to measure entity importance is *degree centrality*, which examines to which extent a specific node is connected with the other nodes in a network. As binary undirected networks are used in this paper, the degree centrality will be formulated as follows:

$$C_D(N_i) = \sum_{j=1}^{g} x_{ij}(i \neq j) \qquad (1)$$

As described by Knoke & Yang (2020), $C_D(N_i)$ will be the degree centrality of node *i*, where $\sum_{j=1}^{g} x_{ij}$ counts the number of direct ties that node *i* has to the *g – 1* other nodes *j*. This degree centrality measure will be used as a control variable. It is also used as an indicator to construct *superstar* involvement.

For performing the Social Network Analysis, the *igraph* package (Csárdi et al, 2023) in R was used. This package provides a variety of tools to build social networks and analyse their corresponding statistics. First, the Social Network Analysis was performed containing all the developers. This was 12,858 developers in total. After that, the degree centrality was found per developer by running the *degree* function. Then, the degree centrality was matched to each developer in each video game observation by imputing additional columns corresponding to which position the developer had (i.e., the first developer got a new column containing their degree centrality, the second developer got a new column containing their degree centrality and so on). Then, all these columns were checked to see if they contained degree centralities higher than 3. This led to the creation of the *superstar* variable. If one of the columns had a degree centrality of 3 or higher, then a *superstar* was present and the variable obtained a ‘true’ value, otherwise it obtained a ‘false’ value. Regarding the descriptives of the degree centrality, the mean of the degree centrality across all developers is 0.211, and the standard deviation is 0.653, with a minimum of 0 and a maximum of 17. This shows that the overall degree centrality of all developers skewed towards the lower band.

As explained in the literature review, the degree centrality cut-off for determining *superstars* was chosen to be 3 as it is similar to what previous research on the topic of *superstars* has shown (Gretz et al., 2019; Binken & Stremersch, 2008).

5.2 Empirical Regressions

The last step in this research is to regress the *superstar* indicator against the video game sales variable and the video game quality variable separately. The different regression formulas for each of these analyses will be covered. In the next chapter, the coefficients for these regression models will be estimated.

The different regression formulas include several control variables based on existing research and available data. All three regression formulas contain all possible genres, which correspond to coefficient 7 to coefficient 23. As a video game can belong to more than one genre, it is not necessary to leave one out as a baseline category as they are not mutually exclusive (Hanck et al., 2023). On the other hand, the year variables are mutually exclusive and exhaustive, and therefore the year 1997 variable is left out as a baseline category.

The regression formula to answer H_{1A} is as follows:

$$\begin{aligned} E[Y_i(\%\,Video\ Game\ Quality)|x] \\ &= e(\alpha + \beta_1 D(Superstar\ Involved)_{1i} + \beta_2(No.Developers)_{2i} \\ &+ \beta_3(No.Previous\ Games)_{3i} + \beta_4(DLC\ Count)_{4i} + \beta_5(Log\ Video\ Game\ Price\ in\ \$)_{5i} \\ &+ \beta_6(Node\ Degree)_{6i} + \beta_7 D(Genre\ Action)_{7i} + \beta_8 D(Genre\ Adventure)_{8i} + \cdots \\ &+ \beta_{23} D(Genre\ Violent)_{23i} + \beta_{24}(Year\ 1998)_{24i} + \beta_{25}(Year\ 1999)_{25i} + \cdots \\ &+ \beta_{49}(Year\ 2023)_{49i} + \varepsilon_i) \end{aligned} \tag{2}$$

For H_{1B}, the regression formula is as follows:

$$
\begin{aligned}
&Y_i(Log\ Video\ Game\ Sales) \qquad (3)\\
&= \alpha + \beta_1 D(Superstar\ Involved)_{1i} + \beta_2(No. Developers)_{2i} + \beta_3(No. Previous\ Games)_{3i}\\
&+ \beta_4(DLC\ Count)_{4i} + \beta_5(Log\ Video\ Game\ Price\ in\ \$)_{5i} + \beta_6(Node\ Degree)_{6i}\\
&+ \beta_7 D(Genre\ Action)_{7i} + \beta_8 D(Genre\ Adventure)_{8i} + \cdots + \beta_{23} D(Genre\ Violent)_{23i}\\
&+ \beta_{24}(Year\ 1998)_{24i} + \beta_{25}(Year\ 1999)_{25i} + \cdots + \beta_{49}(Year\ 2023)_{49i} + \varepsilon_i
\end{aligned}
$$

Lastly, for H_{1C}, the regression formula is as follows:

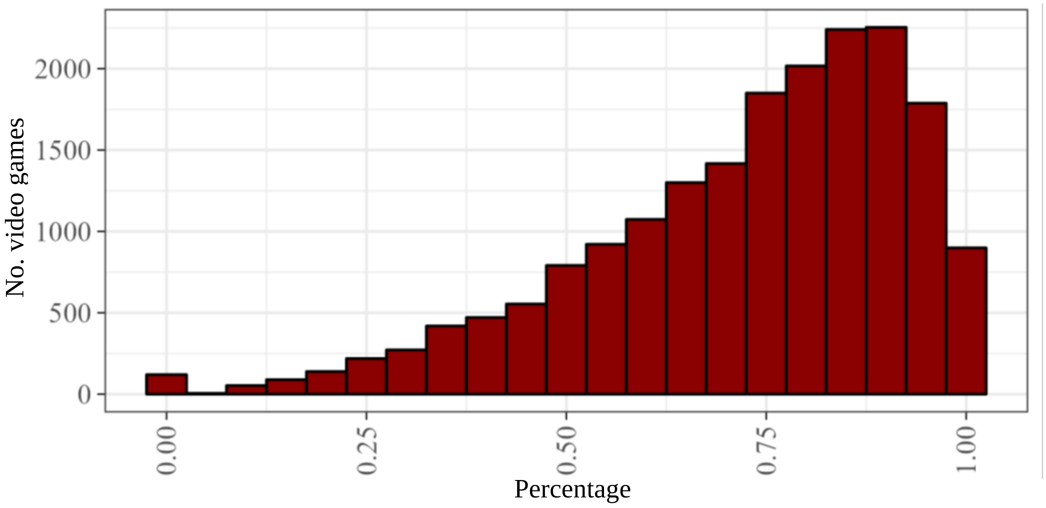
$$
\begin{aligned}
&Y_i(Log\ Video\ Game\ Sales) \qquad (4)\\
&= \alpha + \beta_1(\%\ Video\ Game\ Quality)_{1i} + \beta_2(No. Developers)_{2i}\\
&+ \beta_3(No. Previous\ Games)_{3i} + \beta_4(DLC\ Count)_{4i} + \beta_5(Log\ Video\ Game\ Price\ in\ \$)_{5i}\\
&+ \beta_6(Node\ Degree)_{6i} + \beta_7 D(Genre\ Action)_{7i} + \beta_8 D(Genre\ Adventure)_{8i} + \cdots\\
&+ \beta_{23} D(Genre\ Violent)_{23i} + \beta_{24}(Year\ 1998)_{24i} + \beta_{25}(Year\ 1999)_{25i} + \cdots\\
&+ \beta_{49}(Year\ 2023)_{49i} + \varepsilon_i
\end{aligned}
$$

The control variables in these regression formulas are variables that could theoretically explain variation in either unit sales or product quality, such as video game genre, year of release, price, and other video game characteristics (Cox, 2013; Collins et al., 2002; Köcher & Köcher, 2018; Yang, Cao, Wang, Lu, 2022). As this research also has access to the price and the count of DLC released for a game, these will also be included as video game characteristics. Lastly, in his paper on box office movies, Moretti (2011) suggests that consumers receive a signal on quality based on other observable characteristics, such as directors. This leads to the belief that the previous games released by a specific developer could provide such a signal on quality as well. Therefore, aside from the previous control variables mentioned, another control variable will be used stating how many games a developer has released before this game.

Superstars and Video Game Quality (H_{1A})

To test hypothesis H_{1A}, a Quasi-Binomial multivariate regression model with a logistic link will be estimated on the video game quality (Dunteman & Ho, 2006; Cox, 1996; Papke & Wooldridge, 1996). The reason for choosing a Quasi-Binomial model has several reasons. First, as the dependent variable video game quality is measured as a proportion of positive reviews against total reviews, models that are able to handle a dependent variable that is a proportion, such as a binomial model or beta model, were considered (Papke & Wooldridge, 1996; Baum, 2008; Cox, 1996; Cook, Kieschnick & McCullough, 2008; Ferraria & Bribari-Neto, 2004). Second, the distributional assumption of the dependent variable was checked, which can be seen in Figure 2.

Figure 2: Histogram of Video Game Quality



As is shown clearly, the data is left-skewed. This would initially suggest a beta-regression, as such a model can deal with left-skewed distributions. However, the main drawback of using a beta-regression is that it can only analyse data that is bound between [0, 1], and not values that are 0 and 1 (Ferraria & Bribari-Neto, 2004). As the video game quality variable contains such values, a Binomial model was definitely chosen.

Second, it is a common technique to model a proportion dependent variable using a binomial model with a logistic link function, as using a proportion in a linear model could yield nonsensical predictions (Baum, 2008; Papke & Wooldridge, 1996; Cox, 1996).

However, it is common for data analysed using a binomial family to exhibit dispersion (Dean, 1992). For example, when we take the following formula which is the distribution that can be used for any exponential family, under which the Binomial family falls:

$$f(y\ |\theta,\phi) = e^{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)} \qquad (5)$$

Where θ is a function of the mean μ of the distribution; the dispersion parameter ϕ which plays a role in defining the variance of y, and $c(y,\phi)$ which is a function of the observation and dispersion parameters. The important part here is that the dispersion parameter ϕ in a binomial model is normally assumed to be 1 (Dunteman & Ho, 2006).

Nevertheless, it is most likely not the case that the dispersion parameter ϕ in this paper's data can realistically be assumed to be 1. The data as shown in Figure 2 visualizes it as heavily left-skewed. This skewness generally indicates *underdispersion*. Underdispersion happens when the observed observations are more clumped around the mean μ, which is something that could be inferred from Figure 2. This indicates that the variance the model estimates is larger than the variance of the observed mean μ. A potential explanation for this happening is when the upper end of the distribution of the dependent variable Y has a shortened upper end, which can be seen in Figure 2 (Gardner, Mulvey & Shaw, 1995; Dunteman & Ho, 2006).

This led to the belief that underdispersion could be present if a regular Binomial model would be applied to the data in this research. After running a dispersion test on a regular Binomial model, it was confirmed there was heavy underdispersion in the data, as the dispersion parameter was estimated to be 0.193. Overall, underdispersion is found to be less of an issue than its opposite, overdispersion. This is because underdispersion makes the standard errors in the predicted model larger, and therefore the statistical inference more conservative. Consequently, the H_0 would be rejected less often than if there was no underdispersion. On the other hand, it would also mean that if underdispersion is too high, the statistical inference would become too conservative (Gardner, Mulvey & Shaw, 1995). Therefore, it is still an issue that should be dealt with.

Luckily, an easy way to adjust for underdispersion is by taking a ‘quasi-model’, which is incidentally the third reason why a Quasi-Binomial regression was chosen. When taking a Quasi-Binomial model, the dispersion parameter is estimated from the data. The coefficients of the regression remain the same as in a regular Binomial model, but standard errors and significance values are adjusted for over- or underdispersion (Zeileis, Kleiber & Jackman, 2008; Cox, 1996). Quasi-Poisson and Negative-Binomial models were also considered to deal with the underdispersion, but were not chosen as those are exclusively for count data (Dunteman & Ho, 2006; Cameron & Trivedi, 2013).

Even though the remaining assumptions for this regression will be covered later in the paper, it is important to note that no influential observations are found, nor that any multicollinearity is present. The final regression equation for the Quasi-Binomial model is as seen in formula (2) and will be estimated through maximum likelihood estimation.

Superstars and Video Game Sales (H_{1B})

To test hypothesis H_{1B}, a general multivariate regression model with a dependent variable of the logarithm of sales will be made. As was mentioned in the data chapter, the video game sales variable was transformed using a logarithmic scale to account for outliers and give it a more normal distribution. Looking at Figure 3 and Figure 4, it is possible to see the distribution of sales before and after the transformation. As is clear, a more normal distribution is obtained after the logarithmic transformation, making it adhere to the distributional assumptions of a general multivariate regression model. As will be covered later in this paper, the model displays heteroskedasticity, though no multicollinearity is present.

Figure 3: Histogram of Sales

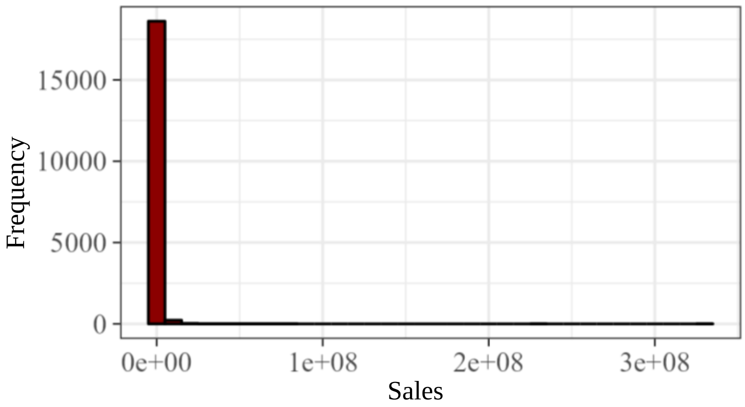
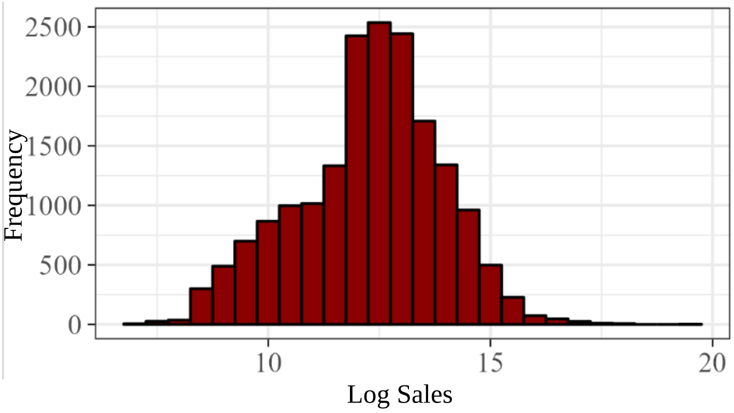


Figure 4: Histogram of Log Sales



The regression equation for the general multivariate regression model is as seen in formula (3) and will be estimated through ordinary least squares estimation.

Video Game Quality and Video Game Sales (H_{1C})

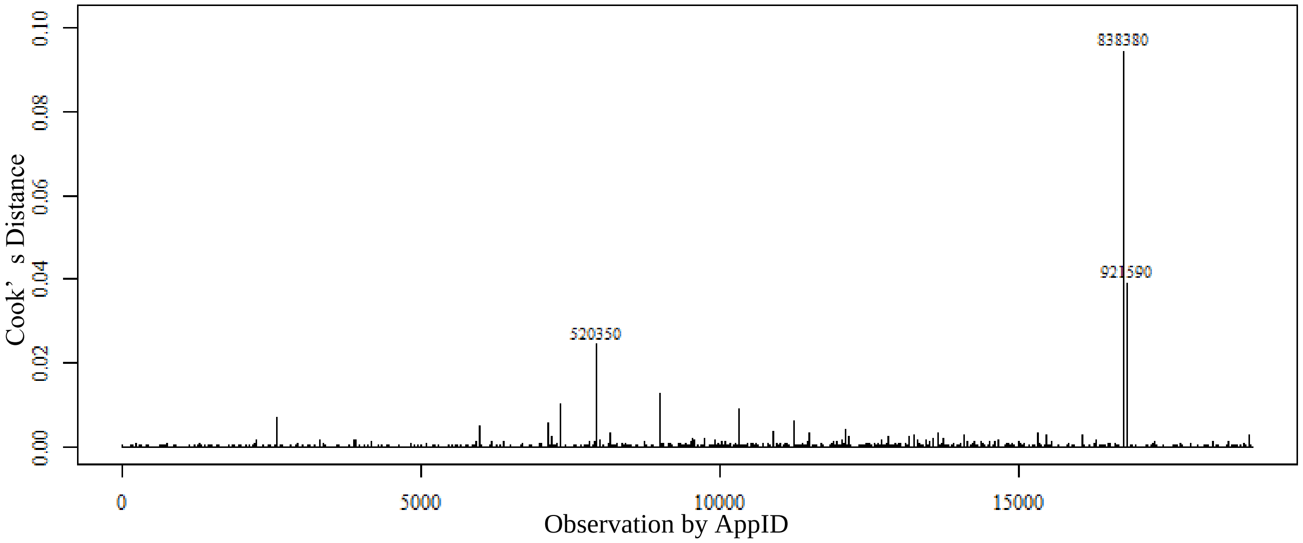
Lastly, to test hypothesis H_{1C}, another general multivariate regression model with the dependent variable of the logarithm of sales will be made, this time using the video game quality as an independent variable. As this model uses the same dependent variable as the previous model, the assumption of a normal distribution also holds here. However, as will be discussed later as well, heteroskedasticity is present, though no multicollinearity is found.

The regression equation for the model is as seen in formula (4) and will be estimated through ordinary least squares estimation.

5.3 Testing Regression Assumptions

Before covering the results, the assumptions for the multivariate regression analyses on video game quality and the log of video game sales were tested. As the first model on video game quality is a Quasi-Binomial model, traditional linear regression assumptions such as normality and homoskedasticity are not relevant, and instead other assumptions should be checked (Osborne, 2015). The first assumption of whether the data follows a Binomial distribution was already covered in the *Empirical Regressions* chapter.

Figure 5: Cook's Distance Plot for Quasi-Binomial Model



The second assumption checked is whether there are any significant outliers or high leverage points in the model. When looking at the Cook's Distance plot in Figure 5, three observations stand out. Observations 838380, 925190, and 934710 seem to have the largest distance. Some apply the rule of thumb that if Cook's distance is greater than 0.5, it may be influential. In Figure 3, however, it is found that none of the observations reach this threshold. Therefore, we can assume that there are no significant outliers or high leverage points in our model (Kutner, Nachtsheim, Neter & Li, 2005).

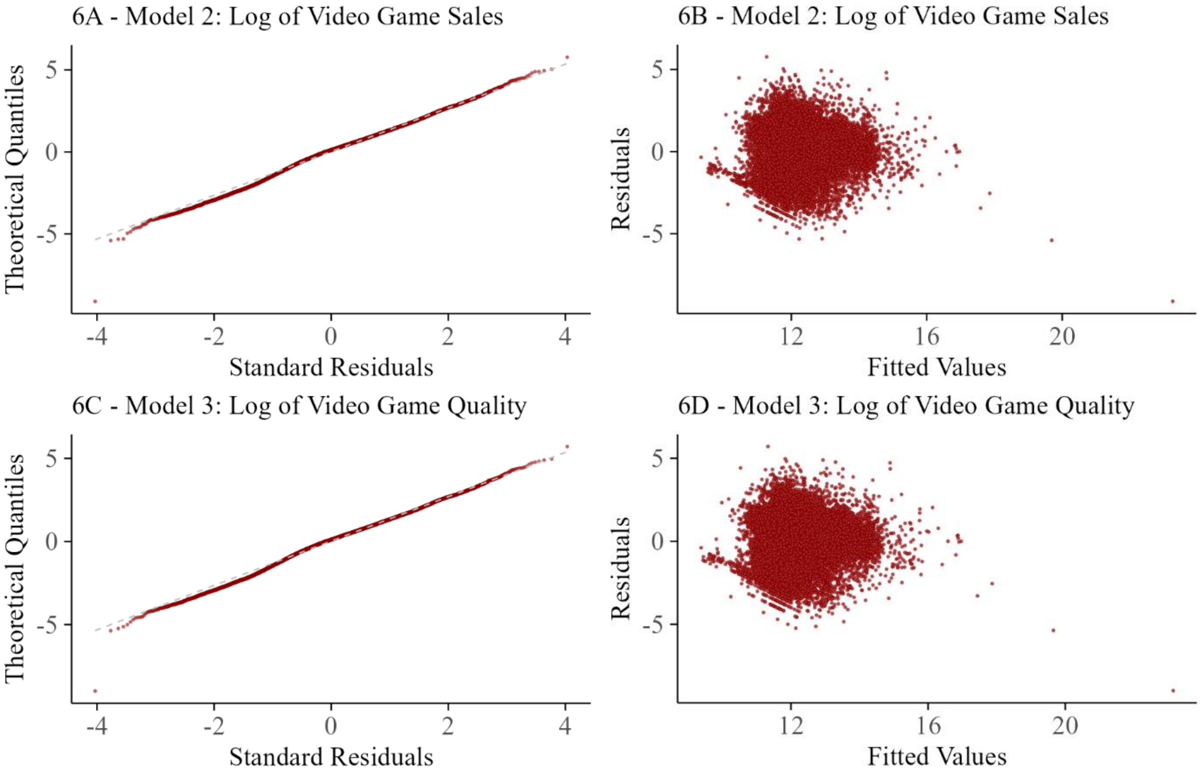
The third and last assumption checked is whether any multicollinearity is present in the model. When considering the VIF values of this model, there seems to be very little multicollinearity[9]. The *Superstar* dummy variable and the *Degree Centrality* variable are standing out. They have

[9] *Superstar, VIF 2.35| No. Dev., VIF = 1.12| Prev. Games, VIF = 1.10| DLC Count, VIF = 1.04| Log Price, VIF = 1.43| Degree Centrality, VIF = 2.41| G. Action, VIF = 1.14| G. Adventure, VIF = 1.11| G. Casual, VIF = 1.13| G. Early Access, VIF = 1.06| G. Education, VIF = 1.00| G. Free To Play, VIF = 1.41| G. Gore, VIF = 1.88| G. Indie, VIF = 1.14| G. Massively Multiplayer, VIF = 1.17| G. Nudity, VIF = 1.35| G. RPG, VIF = 1.11| G. Racing, VIF = 1.11| G. Sexual Content, VIF = 1.32| G. Simulation, VIF = 1.12| G. Sports, VIF = 1.13| G. Strategy, VIF = 1.11| G. Violent, VIF = 1.93| Year, GVIF = 1.28*

a VIF of 2.35 and 2.41 respectively. However, this should pose no problems regarding multicollinearity in the model.

For model 2 and model 3, the regular linear regression assumptions are tested. For models 2 and 3, the QQ-Plot of standardised residuals showed that the data for the log of video game sales contained approximately normally distributed errors (*see Figures 6A & 6C)*. Therefore, we can assume multivariate normality. However, the fitted against residuals scatterplot for the logarithm of video game sales shows that the assumption of homogeneity of variance and linearity might not be met entirely (*see Figures 6B & 6D*). For models 2 and 3, one could argue that there is a cone shape due to the small tail on the left, which gradually moves downward. This shows a pattern, as well as creating a potential cone shape in the residuals. Therefore, an amount of heteroskedasticity might be present in the data. To account for this issue, robust standard errors based on sandwich covariance matrix estimators will be presented next to the normal standard errors for both models (Zeileis, 2006).

Figure 6 – Scatterplots & QQ-Plots for model 2 and model 3



Lastly, there seems to be very little to no multicollinearity amongst the variables in model 2[10] and model 3[11]. The only two outliers in model 2 are the *Superstar* dummy variable and the *Degree Centrality* variable which have a VIF of 2.38 and 2.44 respectively. This is almost more than double the size of most other variables. However, it is still low enough to not pose any serious multicollinearity problems.

[10] *Superstar, VIF 2.39| No. Dev., VIF = 1.12| Prev. Games, VIF = 1.00| DLC Count, VIF = 1.03| Log Price, VIF = 1.54| Degree Centrality, VIF = 2.44| G. Action, VIF = 1.14| G. Adventure, VIF = 1.12| G. Casual, VIF = 1.15| G. Early Access, VIF = 1.05| G. Education, VIF = 1.01| G. Free To Play, VIF = 1.42| G. Gore, VIF = 1.80| G. Indie, VIF = 1.16| G. Massively Multiplayer, VIF = 1.14| G. Nudity, VIF = 1.36| G. RPG, VIF = 1.11| G. Racing, VIF = 1.10| G. Sexual Content, VIF = 1.34| G. Simulation, VIF = 1.11| G. Sports, VIF = 1.12| G. Strategy, VIF = 1.12| G. Violent, VIF = 1.84| Year, GVIF = 1.32*

[11] *Video Game Quality, VIF = 1.11| No. Dev., VIF = 1.11| Prev. Games, VIF = 1.10| DLC Count, VIF = 1.03| Log Price, VIF = 1.55| Degree Centrality, VIF = 1.17| G. Action, VIF = 1.14| G. Adventure, VIF = 1.13| G. Casual, VIF = 1.13| G. Early Access, VIF = 1.06| G. Education, VIF = 1.01| G. Free To Play, VIF = 1.42| G. Gore, VIF = 1.83| G. Indie, VIF = 1.16| G. Massively Multiplayer, VIF = 1.16| G. Nudity, VIF = 1.33| G. RPG, VIF = 1.11| G. Racing, VIF = 1.11| G. Sexual Content, VIF = 1.31| G. Simulation, VIF = 1.13| G. Sports, VIF = 1.12| G. Strategy, VIF = 1.12| G. Violent, VIF = 1.89| Year, GVIF = 1.35*

6 Results

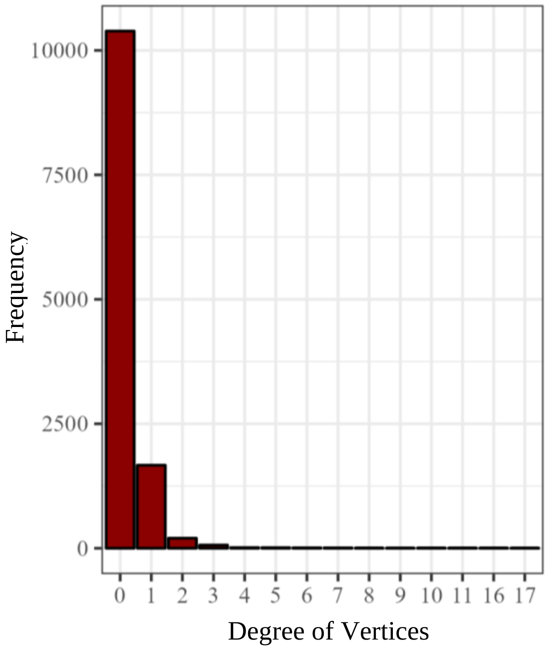
6.1 Exploratory Research Questions

In this section, some additional statistics will be covered to answer the exploratory research questions.

How much are video game developers working together?

The first set of statistics covered will correspond to the first exploratory research question: *"How much are video game developers working together?"*.

Figure 7: Barplot of Degree of Nodes SNA



First, Figure 7 will be considered, which is a visualization of how often each degree centrality for each developer in the Social Network Analysis occurs. One glaring observation is that it is heavily skewed towards zero. That indicates that more than 10,000 developers simply have no connection to another developer. After that, it is observed that about 1,500 developers have one connection to another developer, and about 200 have two connections. Only about 100 developers have three or more connections. The majority of video game developers are relatively unconnected to one another. These findings are further reinforced by Table 5, which shows how many games are released per developer combination.

Table 5 – Games Released per Developer Combination

	Games Released	% Games Released
One Developer	17,588	93.088
Two Developers	1,085	5.743
Three Developers	145	0.767
Four Developers	37	0.196
Five Developers	16	0.085
Six Developers	8	0.042
Seven Developers	7	0.037
Eight Developers	2	0.011
Nine Developers	1	0.005
Ten Developers	1	0.005
Eleven Developers	1	0.005
Twelve Developers	2	0.011
Thirteen Developers	0	0.000
Fourteen Developers	0	0.000
Fifteen Developers	1	0.005

Note. N = 18,894

It shows that the majority of all video games released on the Steam platform were developed by just one developer. Only 1,306 (6,912%) of video games released had more than one developer involved in the creation process. This shows that overall, video game developers are not working together very often.

In conclusion, to answer the question *"How much are video game developers working together?"*, it is found that the majority of video game developers do not work together and have not made any links to other video game developers.

How does the video game developer cooperation change over time?

The second set of statistics covered will correspond to the second exploratory research question: *"How does the video game developer cooperation change over time?"*. For this, Figure 7 and Figure 8 will be covered. These figures show the percentage of games released by individual developers and the percentage of games released by other developer combinations respectively. Starting with Figure 7, it is observed that from 1997 to 2005 all the games released were developed by one studio each. After 2005, there is a drop in video games developed by single studios. There is a downward trend to 2023, going from 100% to about 91%. From Figure 8 it is evident that this is compensated by a steady increase in games developed by, mostly, two developers. From 2005 to 2023, an upward trend is observed for two developer pairs, going from 0% to more than 8%. More than two developer pairs are observed as well.

Games developed by three developers see an interesting trend as well. After 2008 it takes off, remaining relatively stable at around 1% of total games developed. However, it drops down again after 2020. Anything more than four developer pairs remains low over the years. They are mostly found after 2012, but barely even reach the 1% mark of total games developed. In conclusion, these observations show that there is a slow and relatively small decrease in games released by just one developer over the years. This is mostly compensated by two-developer releases. This indicates that even though video game developer cooperation is rare, it is becoming more common over time.

Figure 7: Percentage of Games Released by Individual Developers

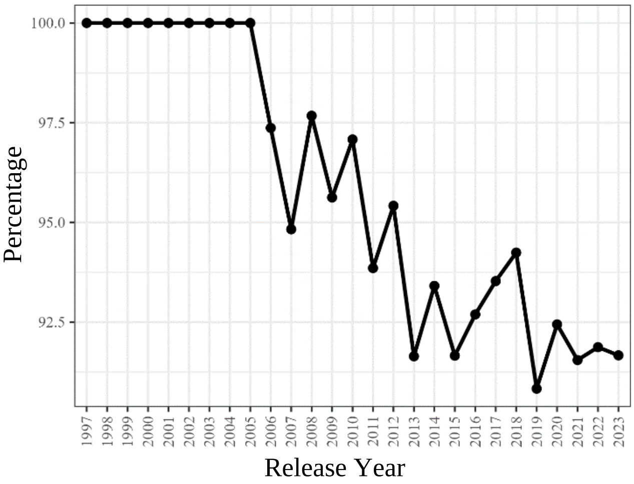
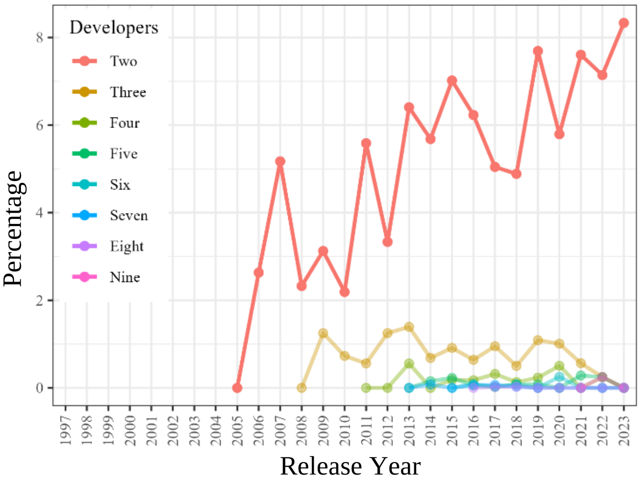


Figure 8: Percentage of Games Released per other Developer Combinations

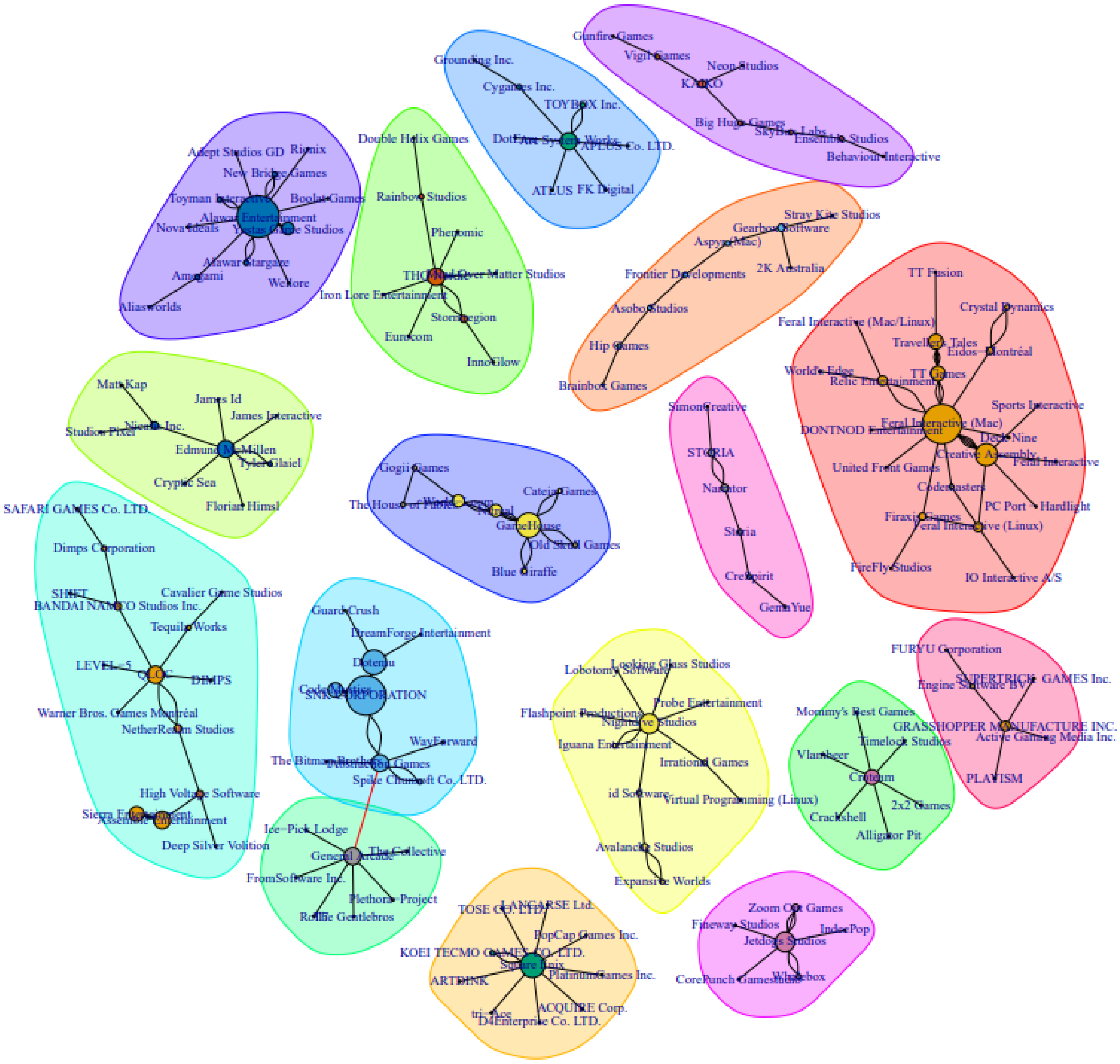


In conclusion, to answer the question *"How does the video game developer cooperation change over time?"*, it is found that after 2005 the cooperation of video game developers increases.

Are video game developers cooperating and what is the nature of cooperation?

The third and last set of statistics covered correspond to the last exploratory research question: *“Are video game developers cooperating and what is the nature of cooperation? (I.e., are superstars connected to non-superstars; Are non-superstars connected to non-superstars?)”*. As could already be seen in Figure 7 and Table 5, and noted earlier in this paragraph, there is cooperation present between developer studios. Overall though, there are very few collaborations. Most of the cooperation is between two studios, and that is the extent of the collaboration. However, there are a few larger pockets of developer collaborations as well. If we move to Figure 9, all clusters in the dataset that contain six or more nodes can be seen.

Figure 9: Clusters with at least 6 nodes



A filter was applied on the number of node connections for two reasons. First, to sift out the lower connections and make the graph more readable. Second, showing clusters of just two or three nodes does not give the insight to comment on the nature of the cooperation.

From this graph, it is shown that there are two types of relationships in larger collaboration networks. The first type shows a typical *superstar – non-superstar* relationship pattern as defined in the literature review, with one developer in the middle, being connected to several *non-superstar* studios. Take for example the light-orange cluster at the bottom of the graph, with *Square Enix* in the middle. *Square Enix* seems to be the *superstar* developer here, working together with several other studios that would be classified as *non-superstar* in this paper's definition.

Square Enix can also be found in Table 6, indicating this studio to be the sixth-highest *superstar* developer. Overall, they have 10 links with other developers and have published 32 games, of which about one-third were co-developed with at least one other studio.

When considering the other top *superstar* developers in Table 6, two thing stands out. Most *superstar* developers, on average, release high-quality titles. The exception to this is the developer Tero Lunkka, which has an average quality of 47.187%, of which its best-selling game is even lower, at 22%. This shows that *superstar* developers do not publish high-quality titles per definition.

Returning to Figure 9, it is found that the second type of network visualizes a more 'chain' like network, without a clear *superstar* player being present. Such networks could be seen in the pink cluster in the middle of the graph, or the purple cluster at the top of the graph. A possible explanation for these 'chain' clusters seems to indicate *non-superstar* developers working together with other *non-superstar* studios per game release. For example, when considering the purple cluster, it seems that the studio *Behavioural Interactive* made a connection to *Ensemble*

Studios. *Ensemble Studios* then developed a game with *SkyBox Labs*. *SkyBox Labs* then developed a game with *Big Huge* Games, and so on. However, it seems that such clusters are less common. In conclusion, it seems that typical *superstar – non-superstar* relationships are the most common.

In conclusion, to answer the question *"Are video game developers cooperating and what is the nature of cooperation?",* it is found that there are two network types. The first visualizes a relationship with the *superstar* player in the middle, whereas the second is more of a 'chain' network. High superstar linkage do not always seem to correlate with higher quality.

Table 6 – Top Superstar Developers

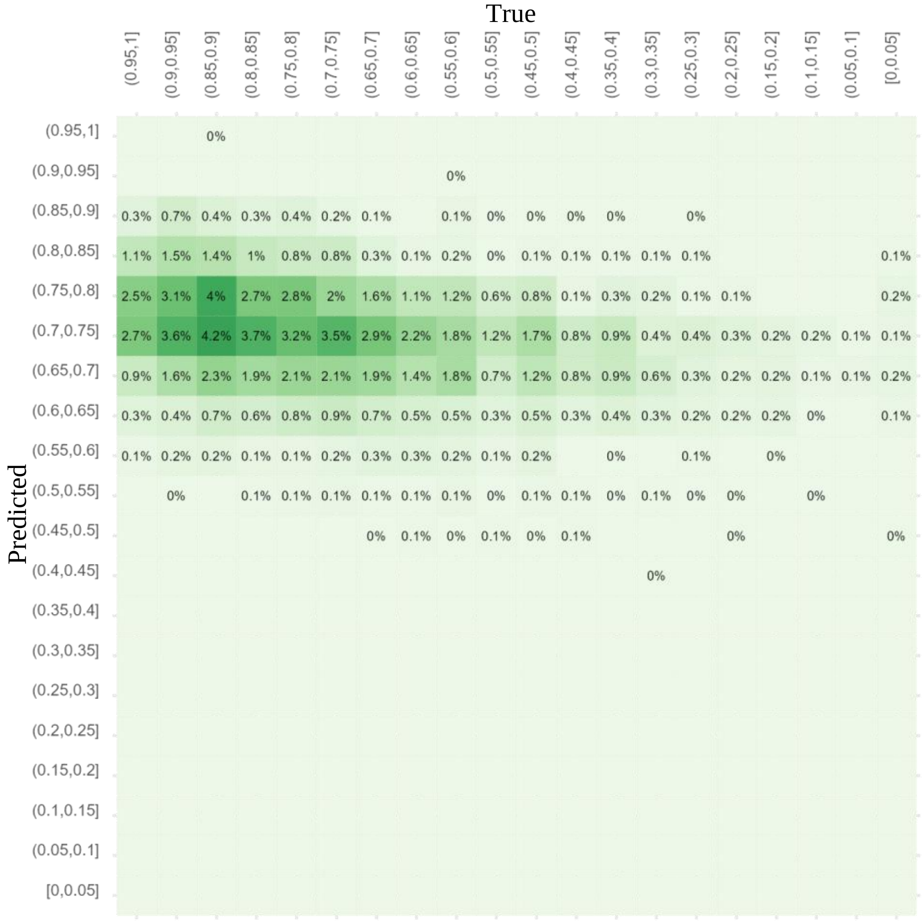
Developer	Links	No. Games	% Co-Developed	% Avg. Quality	Total Unit Sales	Best Selling Title	% Quality Best Selling Title	Unit Sales Best Selling Title
Alawar Entertainment	17	32	53.125	80.136	8,468,000	Sacra Terra: Angelic Night	86.032	856,000
SNK CORPORATION	16	20	80	79.335	13,998,000	METAL SLUG 3	89.630	2,900,000
Feral Interactive	16	22	100	82.568	159,167,000	Tomb Raider	96.186	37,000,000
Idea Factory	11	18	64.706	82.377	8,571,000	Hyperdimension Neptunia Re;Birth3 V Generation	95.063	1,400,000
Tero Lunkka	11	24	50	47.187	10,949,000	Tales of Destruction	22.000	1,200,000
Square Enix	10	32	33.333	73.728	25,939,000	NieR: Automata^{TM}	84.805	3,800,000
Dotemu	10	14	71.429	79.236	12,577,000	METAL SLUG 3	89.630	2,900,000
GameHouse	10	20	50	85.193	706,000	Heart's Medicine – Time to Heal	93.880	175,000
Creative Assembly	9	10	90	78.398	44,905,000	Total War: SHOGUN 2	90.654	14,800,000
All Superstars	4.615*	7.222*	59.584*	77.702	947,664,000	Borderlands 2[1]	93.725	38,400,000
All Developers	0.211*	1.626*	6.912*	72.465	13,169,042,000	Counter-Strike: Global Offensive[2]	88.261	331,000,000

Note. * Indicates average. [1] has developers Gearbox Software, Aspyr (Mac) and Aspyr (Linux). [2] has developer Valve and Hidden Path Entertainment.

6.2 Regression Results: Video Game Quality, H_{1A}

Do *superstar* video game developers publish higher quality video games as compared to *non-superstar* video game developers? The answer will be discussed using the results of the Quasi-Binomial regression analysis. As was described in the empirical regression chapter, a Quasi-Binomial was chosen over a normal Binomial model due to the underdispersion present in the model. Starting, an indication of model quality will be examined. For this, the dataset was split up into a training set and a test set, making a percentage split of 80/20 respectively. This means that the train set obtained 15,071 observations, while the test set obtained 3,820 observations. First, the Quasi-Binomial model was trained on the training set. This model was then promptly saved and used to predict the dependent variable video game quality. The Root Mean Squared Error (RMSE) and a confusion matrix will be presented from the obtained predictions to analyse the model fit. First, the RMSE found was 0.189. Considering how the video game quality

Figure 10: Confusion Matrix True/Predicted



	(0.95,1]	(0.9,0.95]	(0.85,0.9]	(0.8,0.85]	(0.75,0.8]	(0.7,0.75]	(0.65,0.7]	(0.6,0.65]	(0.55,0.6]	(0.5,0.55]	(0.45,0.5]	(0.4,0.45]	(0.35,0.4]	(0.3,0.35]	(0.25,0.3]	(0.2,0.25]	(0.15,0.2]	(0.1,0.15]	(0.05,0.1]	[0,0.05]
(0.95,1]			0%																	
(0.9,0.95]									0%											
(0.85,0.9]	0.3%	0.7%	0.4%	0.3%	0.4%	0.2%	0.1%		0.1%	0%	0%	0%	0%		0%					
(0.8,0.85]	1.1%	1.5%	1.4%	1%	0.8%	0.8%	0.3%	0.1%	0.2%	0%	0.1%	0.1%	0.1%	0.1%	0.1%					0.1%
(0.75,0.8]	2.5%	3.1%	4%	2.7%	2.8%	2%	1.6%	1.1%	1.2%	0.6%	0.8%	0.1%	0.3%	0.2%	0.1%	0.1%				0.2%
(0.7,0.75]	2.7%	3.6%	4.2%	3.7%	3.2%	3.5%	2.9%	2.2%	1.8%	1.2%	1.7%	0.8%	0.9%	0.4%	0.4%	0.3%	0.2%	0.2%	0.1%	0.1%
(0.65,0.7]	0.9%	1.6%	2.3%	1.9%	2.1%	2.1%	1.9%	1.4%	1.8%	0.7%	1.2%	0.8%	0.9%	0.6%	0.3%	0.2%	0.2%	0.1%	0.1%	0.2%
(0.6,0.65]	0.3%	0.4%	0.7%	0.6%	0.8%	0.9%	0.7%	0.5%	0.5%	0.3%	0.5%	0.3%	0.4%	0.3%	0.2%	0.2%	0.2%	0%		0.1%
(0.55,0.6]	0.1%	0.2%	0.2%	0.1%	0.1%	0.2%	0.3%	0.3%	0.2%	0.1%	0.2%		0%		0.1%		0%			
(0.5,0.55]		0%		0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0%	0.1%	0.1%	0%	0.1%	0%	0%		0%		
(0.45,0.5]							0%	0.1%	0%	0.1%	0%	0.1%				0%				0%
(0.4,0.45]														0%						
(0.35,0.4]																				
(0.3,0.35]																				
(0.25,0.3]																				
(0.2,0.25]																				
(0.15,0.2]																				
(0.1,0.15]																				
(0.05,0.1]																				
[0,0.05]																				

variable was distributed between 0 and 1, an RMSE of 0.189 seems to be relatively high. It is nearly a 20% deviation.

Directing to Figure 10, the confusion matrix is found. Both the true and predicted values of video game quality were organized into 20 brackets of quality categories. Each quality category spans 5%, i.e. [0% to 5%], (5% to 10%), and so on. The glaring result of this confusion matrix is that the predicted values seem to be quite wrong when compared to the true values. First off, the predictions never predict a quality lower than 40%. This indicates that the model tends to overestimate quality in video games. Second, even though it mostly predicts the quality to be too high, it still does not do that accurately. For example, consider the greenest square. About 4.2% of all observations are predicted to fall into the (70% to 75%] category. However, the true value for these 4.2% of the observations belong to the (85% to 90%] bracket. All-in-all, the accuracy of the model (no. correct predictions divided by total predictions) is 10.5%. That indicates that nearly 90% of all predictions are misclassified.

Lastly, the goodness-of-fit Pearson's Chi-Square as shown in Table 7 is considered. It is found that the Pearson Chi-Square is 3568.306. Performing a Chi-Square test shows that, with 49 degrees of freedom, the Chi-Square statistic is highly significant (*p = 0.000*), indicating that the data does decrease the deviance in the model quite well (Cox, 1996; Osborne, 2015; Sonderegger, 2020). This is remarkable, as even though the data seems to fit the model well, the previous findings indicate that its prediction capabilities are poor.

Moving on to the results of the Quasi-Binomial regression presented in Table 7, hypothesis H_{1A} will be answered. In this model, a statistically significant positive association between video game quality and *superstar* involvement was found (*Superstar*: *B = 0.235, p = 0.000)*. When estimating the effect, it shows that when all else is held equal, the presence of a *superstar* results in *exp(0.235)* ≈ 1.265 = 26.5% higher video game quality.

Table 7 – Results of the Quasi-Binomial Regression Analysis for Video Game Quality

Independent variables	Estimates	SD	Sig.	Independent variables	Estimates	SD	Sig.
Superstar (Dummy)	0.235***	0.063	0.000	**Year 2000**	1.479	2.918	0.959
No. Developers	0.042*	0.019	0.029	**Year 2001**	-0.157	1.229	0.590
No. Previous Games	-0.005***	0.001	0.000	**Year 2002**	-0.088	1.712	0.987
DLC Count	0.012***	0.002	0.000	**Year 2003**	-0.621	1.154	0.752
Log Price	0.214***	0.008	0.000	**Year 2004**	-0.019	1.131	0.540
Degree Centrality	-0.012	0.008	0.187	**Year 2005**	-0.337	1.067	0.418
Genre: Action	-0.135***	0.015	0.000	**Year 2006**	-0.585	0.954	0.265
Genre: Adventure	-0.022	0.016	0.152	**Year 2007**	-0.766	0.947	0.211
Genre: Casual	0.032*	0.016	0.043	**Year 2008**	-1.051	0.942	0.215
Genre: Early Access	-0.236***	0.028	0.000	**Year 2009**	-1.175	0.939	0.207
Genre: Education	0.365	0.465	0.433	**Year 2010**	-1.166	0.940	0.274
Genre: Free To Play	0.426***	0.033	0.000	**Year 2011**	-1.184	0.939	0.247
Genre: Gore	-0.001	0.081	0.989	**Year 2012**	-1.028	0.938	0.147
Genre: Indie	0.115***	0.017	0.000	**Year 2013**	-1.086	0.938	0.159
Genre: M.M.	-0.290***	0.046	0.000	**Year 2014**	-1.357	0.936	0.208
Genre: Nudity	-0.060	0.126	0.632	**Year 2015**	-1.319	0.936	0.216
Genre: RPG	-0.033	0.020	0.107	**Year 2016**	-1.180	0.936	0.217
Genre: Racing	-0.176***	0.038	0.000	**Year 2017**	-1.157	0.936	0.335
Genre: Sexual Content	0.047	0.132	0.720	**Year 2018**	-1.157	0.936	0.429
Genre: Simulation	-0.226***	0.019	0.000	**Year 2019**	-0.903	0.937	0.466
Genre: Sports	-0.066	0.036	0.065	**Year 2020**	-0.741	0.938	0.516
Genre: Strategy	-0.159***	0.019	0.000	**Year 2021**	-0.684	0.938	0.259
Genre: Violent	-0.478***	0.072	0.000	**Year 2022**	-0.609	0.938	0.959
Year 1998	1.135	2.560	0.658	**Year 2023**	-1.110	0.984	0.590
Year 1999	-0.235	1.172	0.841				
Null deviance	4060.5 on 18892 DF						
Residual deviance	3690.8 on 18843 DF						
Pearson’s Chi^2	3568.306***		0.000				

Note. SD is used to represent the standard deviation. Sig. is used to represent the significance value. * indicates p<0,05. ** indicates p<0,01. *** indicates p<0,001. N = 18.893

Besides this, there are quite a few other coefficients that are statistically significant. First, it is found that the number of previous games (*B = -0.005, p = 0.000*), action genre (*B = -0.135, p = 000*), early access genre (*B = -0.236, p = 0.000*), massively multiplayer genre (*B = -0.290, p = 0.000*), racing genre (*B = -0.176, p = 0.000*), simulation genre (*B = -0.226, p = 0.000*), strategy genre (*B = -0.159, p = 0.000*) and violent genre (*B = -0.478, p = 0.000*) all have a mediating effect on the video game quality. On the other hand, it is found that the number of developers (*B = 0.042, p = 0.029)*, DLC count (*B = 0.012, p = 0.000)*, log price (*B = 0.214, p = 0.000)*, casual genre (*B = 0.032, p = 0.043*), free-to-play genre (*B = 0.426, p = 0.000)*, and indie genre (*B = 0.115, p = 0.000*) all have an empowering effect on the video game quality.

Lastly, it is interesting to note that almost all years have a negative effect on video game quality, however, none of them are found to be statistically significant.

In general, the most important observation is that the presence of a *superstar* significantly influences the quality of a video game release. Therefore, the null-hypothesis H_{0A}: '*Superstar video game developers do not publish higher quality video games'* can be rejected and evidence is found that collaborating with *superstars* does pay off in terms of video game quality.

6.3 Regression Results: Video Game Sales, H_{1B}

Do video games that are developed while cooperating with a *superstar* lead to a higher pay off in terms of sales than video games developed without a *superstar* involved? The answer to this will be discussed using the results of the first general multivariate regression analysis. As in the previous section, a start is made with an indication of model quality. The dataset used the same split of training and test, making a percentage split of 80/20 respectively. After training the regression model and using it to predict video game sales, the RMSE was once again calculated. The RMSE for this model is 2.861. When considering the scale on which the variable was taken, which is the logarithm of sales on a scale from 1.627 to 12.267, it seems relatively low.

Figure 11: True Log Sales Values

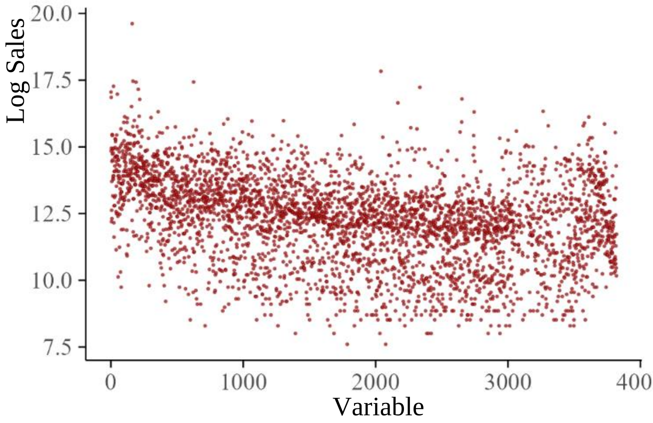
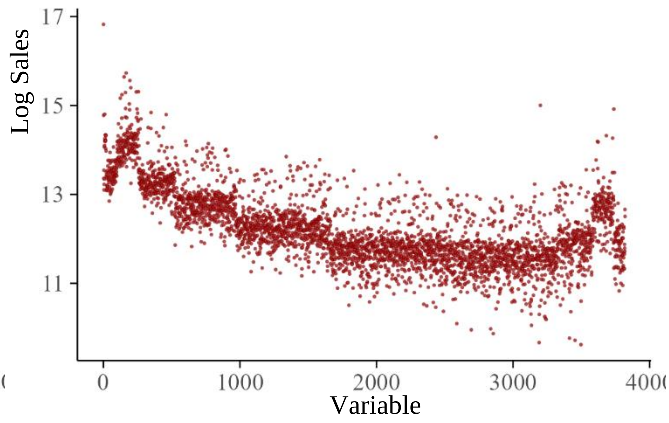


Figure 12: Predicted Log Sales Values



However, when considering Figure 11 and Figure 12, the true and predicted values can be compared with each other. It is found that even though the predicted values follow a similar distribution as the true values, they are distributed much narrower than the true values, which indicates that this model is not very useful for prediction purposes. Lastly, the adjusted R^2 of this model as seen in Table 8 will be considered. The adjusted R^2 is 0.260, which means that this model explains about 26.0% of the variance in the data.

Moving on to the results for this model, which can be found in Table 8. With these results, an answer for hypothesis H_{1B} will be formed. In this model, a statistically significant positive association between log sales and *superstar* involvement was found (*Superstar*: *B = 0.349, p = 0.000)*. Interpreting this coefficient further shows that, when all else is held equal, the presence of a *superstar* results in *exp(0.349)* ≈ 1.418 = 41.8% higher sales. This means that when compared to the previous model, *superstar* involvement leads to a higher increase in sales than in video game quality.

There are a few other control variables that are statistically significant as well. It is found that the number of previous games (*B = -0.024, p = 0.000*), log price (*B = -0.080, p = 0.000*), casual genre (*B = -0.177, p = 0.000*), early access genre (*B = -0.732, p = 0.000*), education genre (*B = -1.282, p = 0.000*), indie genre (*B = -0.086, p = 0.000*), sexual content genre (*B = -0.394, p = 0.043*), sports genre (*B = -0.289, p = 0.000*) and violent genre (*B = -0.371, p = 0.001*) all have a mediating effect on the log sales. On the other hand, it is also interesting to note that the number of developers (*B = 0.055, p = 0.038)*, DLC count (B = 0.026, p = 0.000), Degree Centrality (*B = 0.067, p = 0.000)*, the action genre (*B = 0.267, p = 0.000)*, the adventure genre (*B = 0.070, p = 0.002)*, free to play genre (*B = 0.886, p = 0.000)*, indie genre (*B = 0.088, p = 0.000)*, massively multiplayer genre (*B = 0.287, p = 0.000)*, RPG genre (*B = 0.098, p = 0.000)*, massively multiplayer genre (*B = 0.265, p = 0.000)*, RPG genre (*B = 0.078, p = 0.008)* and strategy genre (*B = 0.073, p = 0.007)* all have an empowering effect on the log sales. Furthermore, it seems that a few year variables are showing a small statistical significance at the 5% level. These are the years 2004, 2017, 2018, 2019, and 2022, which could indicate that some years affect game sales.

Table 8 – Results of the Logistic Regression Analysis for Log Sales

Independent variables	Estimates	SD	RSD	Sig.	Independent variables	Estimates	SD	RSD	Sig.
Superstar (Dummy)	0.349***	0.085	0.084	0.000	**Year 2000**	2.822	1.714	0.741	0.099
No. Developers	0.055*	0.027	0.028	0.038	**Year 2001**	0.984	1.278	0.924	0.441
No. Previous Games	-0.024***	0.002	0.002	0.000	**Year 2002**	0.929	1.714	0.741	0.588
DLC Count	0.026***	0.002	0.008	0.000	**Year 2003**	0.520	1.278	0.917	0.684
Log Price	-0.080***	0.013	0.013	0.000	**Year 2004**	2.824*	1.171	0.770	0.016
Degree Centrality	0.067***	0.012	0.012	0.000	**Year 2005**	0.321	1.143	0.991	0.779
Genre: Action	0.267***	0.022	0.022	0.000	**Year 2006**	-0.506	1.016	0.781	0.618
Genre: Adventure	0.070**	0.023	0.023	0.002	**Year 2007**	0.200	1.007	0.754	0.843
Genre: Casual	-0.177***	0.023	0.023	0.000	**Year 2008**	-0.619	1.001	0.756	0.537
Genre: Early Access	-0.732***	0.043	0.047	0.000	**Year 2009**	-0.601	0.996	0.749	0.546
Genre: Education	-0.734	0.627	0.545	0.242	**Year 2010**	-0.548	0.997	0.751	0.583
Genre: Free To Play	0.886***	0.050	0.042	0.000	**Year 2011**	-0.124	0.995	0.746	0.901
Genre: Gore	0.140	0.140	0.139	0.319	**Year 2012**	-0.037	0.994	0.744	0.970
Genre: Indie	0.098***	0.024	0.026	0.000	**Year 2013**	-0.045	0.993	0.743	0.964
Genre: M.M.	0.265***	0.071	0.067	0.000	**Year 2014**	-0.836	0.991	0.741	0.399
Genre: Nudity	0.066	0.189	0.178	0.729	**Year 2015**	-1.387	0.990	0.741	0.161
Genre: RPG	0.078**	0.029	0.029	0.008	**Year 2016**	-1.849	0.990	0.741	0.062
Genre: Racing	0.063	0.057	0.057	0.271	**Year 2017**	-2.340*	0.990	0.741	0.018
Genre: Sexual Content	-0.394*	0.195	0.190	0.043	**Year 2018**	-2.475*	0.990	0.741	0.012
Genre: Simulation	0.027	0.028	0.029	0.340	**Year 2019**	-2.167 *	0.991	0.742	0.029
Genre: Sports	-0.289***	0.053	0.055	0.000	**Year 2020**	-1.380	0.992	0.745	0.164
Genre: Strategy	0.073**	0.027	0.027	0.007	**Year 2021**	-1.407	0.993	0.744	0.156
Genre: Violent	-0.371***	0.112	0.116	0.001	**Year 2022**	-2.114*	0.993	0.743	0.033
Year 1998	2.361	1.714	0.741	0.168	**Year 2023**	-1.973	1.070	0.837	0.065
Year 1999	1.208	1.278	0.879	0.344					
R^2/Adj. R^2	0.262/0.260								

Note. SD is used to represent the standard deviation. Sig. is used to represent the significance value. * indicates p<0,05. ** indicates p<0,01. *** indicates p<0,001. N = 18.066

On top of this, it is found that some variables are showing different effects in comparison to the previous model on video game quality. For example, in the model on log sales, it is found that there is no statistically significant effect, whereas it is the case in the model on video game quality. The log price of a game will show a decrease in sales, whereas it shows an increase in quality. An increase in the Degree Centrality will result in a statistically significant increase in log sales, whereas it shows no apparent statistically significant effect on video game quality. Lastly, the genres show to have very different effects overall. Where some estimates show positive results on video game quality, they show negative results on video game sales (for example the casual & indie genres), or vice versa.

The robust standard deviations made to account for the heteroskedasticity barely show any difference from the normal standard deviations. Only a few variables see a small increase in their standard deviation, for example, DLC Count (*SD = 0.002, RSD = 0.008*), the early access genre (*SD = 0.043, RSD = 0.0047*) and the indie genre (*SD = 0.024, RSD = 0.026*). Most other variables remain roughly the same or gain an even lower standard deviation. The year variables seem to be impacted the largest, having a relatively large change in RSD. Nonetheless, the overall small changes in RSD lend credibility to the idea that the estimators are correctly estimated.

Perhaps the most important observation is that the presence of a *superstar* has a positive statistically significant effect on the log of sales. This, therefore, means the null-hypothesis H_{0B}: '*Collaborating with stars does not pay off in terms of video game sales'* can be rejected, and find evidence that collaborating with *superstars* does pay off in terms of video game sales.

6.4 Regression Results: Does quality pay off in terms of sales, H_{1C}

Figure 13: True Log Sales Values

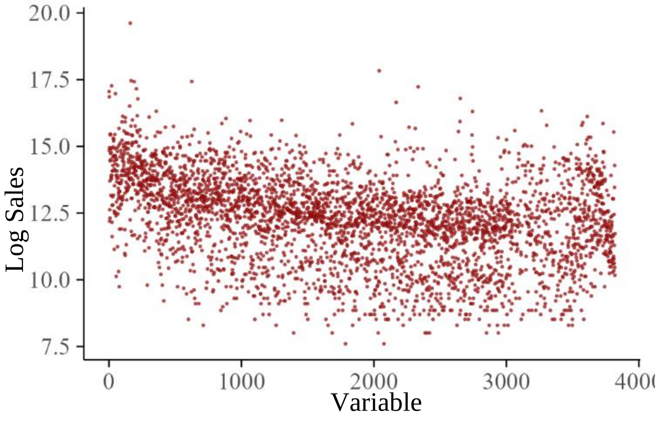
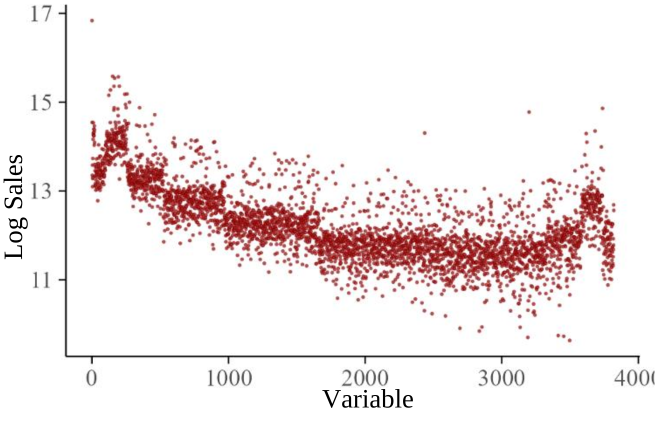


Figure 14: Predicted Log Sales Values



Do video games with a higher quality demand more sales than video games with a lower quality? This will be answered using the results of the second general multivariate regression analysis. As before, a start is made with an indication of model quality with the same training/test specifications. After training the regression model and using it to predict video game sales, the RMSE was calculated. The RMSE for this model is 2.859, which seems to be ever so slightly lower than the previous general multivariate regression model, suggesting that this model would do slightly better at predicting, though the difference is probably barely noticeable.

In Figure 13 and Figure 14, the true and predicted values are once again compared with each other. The predicted values are almost identical to the ones discussed in the previous model, suggesting that this model also is not suited well for prediction purposes. Lastly, the adjusted R^2 of this model as seen in Table 9 will be considered. The adjusted R^2 is 0.261, which means that this model explains about 26.1% of the variance in the data.

Table 9 – Results of the Logistic Regression Analysis of Quality on Sales

Independent variables	Estimates	SD	RSD	Sig.	Independent variables	Estimates	SD	RSD	Sig.
Video Game Quality	0.346***	0.054	0.059	0.000	**Year 2000**	2.764	1.713	0.735	0.107
No. Developers	0.065*	0.026	0.029	0.014	**Year 2001**	1.060	1.277	0.935	0.406
No. Previous Games	-0.024***	0.002	0.002	0.000	**Year 2002**	0.936	1.713	0.736	0.585
DLC Count	0.025***	0.002	0.008	0.000	**Year 2003**	0.540	1.277	0.900	0.672
Log Price	-0.094***	0.013	0.013	0.000	**Year 2004**	2.792*	1.170	0.760	0.017
Degree Centrality	0.103***	0.009	0.009	0.000	**Year 2005**	0.366	1.142	0.978	0.749
Genre: Action	0.278***	0.022	0.022	0.000	**Year 2006**	-0.486	1.1025	0.775	0.632
Genre: Adventure	0.073**	0.023	0.023	0.001	**Year 2007**	0.246	1.006	0.748	0.807
Genre: Casual	-0.180***	0.023	0.023	0.000	**Year 2008**	-0.469	1.001	0.750	0.570
Genre: Early Access	-0.716***	0.043	0.047	0.000	**Year 2009**	-0.539	0.996	0.743	0.588
Genre: Education	-0.779	0.627	0.514	0.214	**Year 2010**	-0.484	0.996	0.745	0.627
Genre: Free To Play	0.853***	0.051	0.041	0.000	**Year 2011**	-0.057	0.995	0.740	0.955
Genre: Gore	0.138	0.140	0.139	0.324	**Year 2012**	0.012	0.993	0.738	0.990
Genre: Indie	0.088***	0.025	0.026	0.000	**Year 2013**	0.016	0.992	0.737	0.987
Genre: M.M.	0.287***	0.071	0.067	0.000	**Year 2014**	-0.762	0.990	0.736	0.441
Genre: Nudity	0.068	0.189	0.178	0.721	**Year 2015**	-1.315	0.990	0.736	0.184
Genre: RPG	0.082**	0.029	0.029	0.005	**Year 2016**	-1.786	0.990	0.736	0.071
Genre: Racing	0.075	0.057	0.057	0.189	**Year 2017**	-2.277*	0.990	0.735	0.021
Genre: Sexual Content	-0.397*	0.195	0.187	0.041	**Year 2018**	-2.415*	0.990	0.736	0.015
Genre: Simulation	0.043	0.028	0.029	0.131	**Year 2019**	-2.121*	0.990	0.737	0.032
Genre: Sports	-0.281***	0.053	0.055	0.000	**Year 2020**	-1.341	0.992	0.739	0.176
Genre: Strategy	0.085**	0.027	0.027	0.002	**Year 2021**	-1.371	0.992	0.738	0.167
Genre: Violent	-0.335**	0.112	0.116	0.003	**Year 2022**	-2.088*	0.992	0.738	0.035
Year 1998	2.308	1.713	0.735	0.178	**Year 2023**	-1.882	1.069	0.833	0.078
Year 1999	1.257	1.277	0.890	0.325					
R^2/Adj. R^2	0.263/0.261								

Note. SD is used to represent the standard deviation. Sig. is used to represent the significance value. * indicates p<0,05. ** indicates p<0,01. *** indicates p<0,001. N = 18.066

Furthermore, there are quite a few control variables that are statistically significant as well. It is interesting to note that the number of previous games (*B = -0.024, p = 0.000*), log price (*B = -0.094, p = 0.000*), casual genre (*B = -0.180, p = 0.000*), early access genre (*B = -0.716, p = 0.000*), education genre (*B = -1.282, p = 0.000*), indie genre (*B = -0.086, p = 0.000*), sexual content genre (*B = -0.397, p = 0.041*), sports genre (*B = -0.281, p = 0.000*) and violent genre (*B = -0.335, p = 0.003*) all have a mediating effect on the log sales. On the other hand, it is also interesting to note that the number of developers (*B = 0.065, p = 0.014),* DLC count (B = 0.025, p = 0.000), Degree Centrality (*B = 0.103, p = 0.000),* the action genre (*B = 0.278, p = 0.000),* the adventure genre (*B = 0.073, p = 0.001),* free to play genre (*B = 0.853, p = 0.000),* indie genre (*B = 0.088, p = 0.000),* massively multiplayer genre (*B = 0.287, p = 0.000),* RPG genre (*B = 0.082, p = 0.005),* and strategy genre (*B = 0.085, p = 0.002)* all have an empowering effect on the log sales. Furthermore, it seems that a few year variables are showing a small statistical significance at the 5% level. These are the years 2004, 2017, 2018, 2019, and 2022, which could indicate that some years affect the sales of games. Overall, all these findings seem to be very similar to the results found for the previous model.

To answer hypothesis H_C, null-hypothesis H_{0C}: '*Video game quality does not pay off in terms of sales'* can be rejected, and evidence is found that video game quality does pay off in terms of video game sales.

7 Conclusion and Discussion

This study attempted to measure to which extent *superstar* video game developers are associated with video game quality and video game sales. Based on existing literature on this topic, the expectation was that *superstar* presence would lead to higher video game quality and higher video game sales. A short overview of the main findings, implications, and restrictions of this study will be given.

7.1 Findings & Implications

Due to a lack of previous existing literature on the topic of video game developers working together, a few exploratory research questions were proposed and answered. First off, most of the video games developed were made by just one studio, and connections between video game developers are low. Of the games in this dataset, 17,588 were made by just one developer, and relatively speaking, video games developed by 2 or more studios are low. Furthermore, more than 10,000 developers had no connections to others, and only a small number of developers had 1 or more connections. When considering developer cooperation over time, it is found that cooperation only started after 2005, mostly between two developer pairs. Pairs of 3 or more developers remained rare. Lastly, the nature of the collaborations is mostly between *superstars* and *non-superstars*, though some networks of *non-superstars* only do exist.

Even though none of the models presented in this research have strong predictive capabilities[12], all the regression models presented in this study have found evidence for statistically significant associations between *superstar* presence and video game sales, *superstar* presence and video game quality, and video game sales and video game quality. Therefore, the main implication of this study is that *superstars* are associated with higher sales and create higher quality products.

[12] The most likely reason for the low predictive capabilities could be due to the skewness of some variables. For example, video game quality was skewed towards larger values. Therefore, the predictions could be more skewed towards those larger values as well.

Keeping this in mind, answers to the three research questions have been formulated. The answer to *“Do superstar firms publish higher quality video games?”* would be a simple yes. The presence of a *superstar* developer leads to an increase in the review score of the video game, indicating higher quality. Similarly, the answer to “*To what degree does cooperating with superstars pay off in terms of video game sales?”* would be just as affirmative. The presence of a *superstar* in development does lead to a higher demand for sales than when no *superstar* is present. Lastly, answering “*Does video game quality pay off in terms of sales”* would be an additional positive answer. It seems that higher quality leads to higher sales.

The results presented in this paper are in line with previous literature on the topic of *superstars*. Chung & Cox (1994), McAndrew & Everett (2015), and De Vany & Walls (1996) all found that *superstar* entities demanded a large portion of their respective industry’s rewards. In this research, it is found that a group of 117 developers, which is 0.945% of the total sample, have an impact on the quality of the products produced, as well as demanding more rewards from these products.

Overall, with the results presented in this paper, insightful knowledge about *superstars* has been added to the literature existing on this topic. Namely, 1) the video game industry is not yet highly connected, 2) *superstars* publish higher quality video games and their video game titles demand higher sales, and 3) video game quality pays off in terms of sales.

7.2 Limitations & Suggestions for future research

A few limitations of this research should be discussed. Most of the limitations had to do with the data and data collection. As the data was collected from two different sources, eventual issues regarding merging the two datasets popped up. This mostly had to do with the dataset obtained from Kaggle and the data scraped from the website PlayTracker having different names for different games, different names for different developers, or sometimes having older versions of games that were long removed from the Steam game store. For example, the developer *Creative Assembly,* famous for their real-time strategy war-simulating series *Total War*, was recorded in different ways. It existed as both 'Creative Assembly' and 'CREATIVE ASSEMBLY' in the data. This example was cleaned, but due to the size and scope of the data, not all developers may have been covered.

This is in close relation to a second limitation regarding the cleaning process, which had to do with grouping certain developer studios based on their relationship with a 'main' studio. For example, the developer studio *Ubisoft*, famous for games such as *Rayman* and *Assassin's Creed*, has several sub-studios in a variety of different countries. For example, *Assassin's Creed Valhalla* was developed solely by their studio *Ubisoft Montreal*. One could argue that it would be logical to group all the Ubisoft studios together under one denominator called 'Ubisoft', as it is the same company after all. However, it was chosen not to do this. The reasoning behind this decision had to do with the fact that maybe not all developer studios would have such a clear-cut case as Ubisoft, and considering the size and scope of the data would have taken too long to analyse on a case-by-case basis.

The third limitation is on the video game quality variable. Prior literature (De Langhe et al., 2016) shows that online user ratings might be trusted too much by consumers. Their research found that there is a disconnect between the objective quality information of user reviews and the extent to which consumers trust them as indicators of objective quality. Therefore, the quality measurement in this paper might not be the most accurate, and future research should perhaps find other ways to measure quality.

The fourth and last limitation has to do with the heteroskedasticity found in the models. Even though an attempt has been made to cover this by including robust standard deviations, it does beg the question of whether a different model or even more rigorous data cleaning could solve these issues entirely.

To tackle these issues, a few suggestions are proposed. If time permits for a larger study, it would be interesting to spend more time on the data-cleaning process. For example, spending time on more accurately grouping developers could yield more exact results. Even though some outliers and other suspicious observations were removed, another suggestion could be to look into outliers even stricter. An example of two cases spring to mind: *Counter-Strike: Global Offensive* and *Dota 2,* which were both developed by *Valve*. These games have an estimated number of owners of 331,000,000 and 231,700,000 respectively. Compare this to the third largest game, *PAYDAY 2,* which has an estimated number of owners of 77,200,000, which is significantly less. In the end, it was not chosen to remove these specific observations, as they seemed like observations containing useful information. On top of that, these specific cases did not seem to influence the regression diagnostics. Despite this, taking another look at some suspicious observations might improve the results.

In conclusion, in this research, a positive statistically significant link was found between *superstar* presence and video game sales, as well as *superstar* presence and video game quality. These findings are an interesting addition to the current field of research on *superstars*.

8 Bibliography

Autor, D., Dorn, D., Patterson, C., Katz, L. F., & Reenen, J. V. (2020). The fall of the labor share and the rise of superstar firms. *Quarterly Journal of Economics*, 135(2), 645–709. https://doi.org/10.1093/qje/qjaa004

Ayyagari, M., Demirguc-Kunt, A., & Maksimovic, V. (2019). The rise of star firms: intangible capital and competition (Ser. Policy research working paper, no. 8832). Retrieved 2023, from http://elibrary.worldbank.org/doi/book/10.1596/1813-9450-8832.

Baum, C. (2008). *Stata Tip 63: Modeling proportions.* The Stata Journal, 8, Number 2, pp. 299-303.

Björk Jennie, & Magnusson, M. (2009). Where do good innovation ideas come from? exploring the influence of network connectivity on innovation idea quality. *Journal of Product Innovation Management*, *26*(6), 662–670. https://doi.org/10.1111/j.1540-5885.2009.00691.x

Binken, J.L.G., J., & Stremersch, S. (2008). The effect of superstar software on hardware sales in system markets. *Erim Report Series Research in Management Erasmus Research Institute of Management*. Retrieved 2023, from http://repub.eur.nl/pub/12339.

Cameron, A. C., & Trivedi, P. K. (2013). Regression analysis of count data (Second, Ser. Econometric society monographs, no. 53). Cambridge University Press.

Chevalier, J. A., Dover, Y., Mayzlin, D., & National Bureau of Economic Research. (2017). *Channels of impact : user reviews when quality is dynamic and managers respond* (Ser. Nber working paper series, no. w23299). National Bureau of Economic Research.

Chung, K. H. and R. A. K. Cox (1994), “A Stochastic Model of Superstardom: An Application of the Yule Distribution,” Review of Economics and Statistics, 76 (4), 771-775

Chopard, B. (2004). Enchères, redressement ou liquidation judiciaire. *L'Actualité économique*, *80*(4), 655–669. https://doi.org/10.7202/012131ar

Collins A., Hand C., Snell M.C. (2002). What makes a blockbuster? Economic analysis of film success in the United Kingdom. Managerial and Decision Economics 23: 343–354

Cook, D. O., Kieschnick, R., & McCullough, B. D. (2008). Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance*, *15*(5), 860–867. https://doi.org/10.1016/j.jempfin.2008.02.001

Corts, K. S., & Lederman, M. (2009). Software exclusivity and the scope of indirect network effects in the U.S. home video game market. *International Journal of Industrial Organization*, *27*(2), 121–136. https://doi.org/10.1016/j.ijindorg.2008.08.002

Cox, C. (1996). Nonlinear quasi-likelihood models: applications to continuous proportions. *Computational Statistics and Data Analysis*, *21*(4), 449–461. https://doi.org/10.1016/0167-9473(95)00024-0

Cox, J. (2013). What makes a blockbuster video game? an empirical analysis of us sales data. Managerial and Decision Economics, (april 2013). https://doi.org/10.1002/mde.2608

Cross, R., Borgatti, S. P., & Parker, A. (2002). Making invisible work visible: using social network analysis to support strategic collaboration. *California Management Review*, 44(2), 25–46. Retrieved from https://doi.org/10.2307/41166121.

Csárdi G, Nepusz T, Traag V, Horvát S, Zanini F, Noom D, Müller K (2023). *igraph: Network Analysis and Visualization in R*. doi:10.5281/zenodo.7682609, R package version 1.5.0, https://CRAN.R-project.org/package=igraph.

Dean, C. B. (1992). Testing for overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association*, *87*(418), 451–457.

De Langhe, B., Fernbach, P. M., & Lichtenstein, D. R. (2016). Navigating by the stars: investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research*, *42*(6), 817–833. Retrieved from https://doi.org/10.1093/jcr/ucv047.

De Vany, A. and W. D. Walls (1996), "Bose-Einstein Dynamics and Adaptive Contracting in the Motion Picture Industry," Economic Journal, 106 (439), 1493-1514.

Dunteman, G. H., & Ho, M.-H. R. (2006). *An introduction to generalized linear models* (Ser. Quantitative applications in the social sciences, 145). Sage Publications.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: poisson, overdispersed poisson, and negative binomial models. *Psychological Bulletin*, *118*(3), 392–404.

Gretz, R. T., Malshe, A., Bauer, C., & Basuroy, S. (2019). The impact of superstar and non-superstar software on hardware sales: the moderating role of hardware lifecycle. *Journal of the Academy of Marketing Science*, *47*(3), 394–416. https://doi.org/10.1007/s11747-019-00631-3

Gulati, R., Lavie, D., & Singh, H. (2009). The nature of partnering experience and the gains from alliances. *Strategic Management Journal*, 30(11), 1213–1233. Retrieved from https://doi.org/10.1002/smj.786.

Gutiérrez, G., & Philippon, T. (2019). Fading stars. *Working Paper Series*, 25529.

Goyal, S, van der Leij, M.J, & Moraga-Gonzalez, J.L. (2004). Economics: An Emerging Small World? *Tinbergen Institute Discussion Paper Series*. Retrieved from http://hdl.handle.net/1765/6696.

Hanck, C., Arnold, M., Gerber, A., Schmelzer, M. (2023). Introduction to Econometrics with R. *6.4 OLS Assumptions in Multiple Regression*. Retrieved on 25-6-2023, from https://www.econometrics-with-r.org/1-introduction.html.

Hu, N., Liu, L., & Zhang, J. J. (2008). Do online reviews affect product sales? the role of reviewer characteristics and temporal effects. *Information Technology and Management*, *9*(3), 201–214. Retrieved from https://doi.org/10.1007/s10799-008-0041-2

Khomych, A. (2022). What is an Indie Game and Why is It So Popular. Retrieved on 20-6-2023, from https://blog.getsocial.im/what-is-an-indie-game-and-why-is-it-so-popular/

Knoke, D., & Yang, S. (2020). Social network analysis. *SAGE*. Retrieved 2023.

Köcher, S. & Köcher, S. (2018). Should we Reach For The Stars? Examining the Convergence Between Online Product Ratings and Objective Product Quality and Their Impacts on Sales Performance. *Journal of Marketing Behaviour*, 3(2), 167-183. Retrieved from DOI:10.1561/107.00000050.

Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W. (2005). *Applied Linear Statistical Models*. (Vol 5). New York: McGraw-Hill Irwin.

Mayzlin, D., Dover, Y., Chevalier, J. A., & National Bureau of Economic Research. (2012). *Promotional reviews : an empirical investigation of online review manipulation* (Ser. Nber working paper series, no. 18340). National Bureau of Economic Research.

McAndrew, S., & Everett, M. (2015). Music as collective invention: a social network analysis of composers. Cultural Sociology, 9(1), 56–80. https://doi.org/10.1177/1749975514542486

Marijan. (N.D.). Insight. Retrieved on 7-5-2023, from: https://web.archive.org/web/20220528094616/https://playtracker.net/insight/about/

Milgram, S. (1967) The Small World Problem. *Psychology Today*, 2, 60-67.

Molina-Morales, F.X., & Martínez-Fernández M.T. (2009). Too much love in the neighborhood can hurt: how an excess of intensity and trust in relationships may produce negative effects on firms. Strategic Management Journal, 30(9), 1013–1023.

Moody, J. (2004). The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999. *American Sociological Review*, *69*(2), 213–238.

Morano, R. S., Barrichello, A., Jacomossi, R. R., & D'Acosta-Rivera, J. R. (2018). Street food: factors influencing perception of product quality. *Rausp Management Journal*, *53*(4), 535–554. Retrieved from https://doi.org/10.1108/RAUSP-06-2018-0032

Moretti, E. (2011). Social learning and peer effects in consumption: evidence from movie sales. Review of Economic Studies, 78(1), 356–393. https://doi.org/10.1093/restud/rdq014

Narasimhan, R., Ghosh, S., & Mendez, D. (1993). A dynamic model of product quality and pricing decisions on sales response. *Decision Sciences*, *24*(5), 893–908. Retrieved from https://doi.org/10.1111/j.1540-5915.1993.tb00495.x.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 404–409.

Osborne, J. W. (2015). Best practices in logistic regression. SAGE. Retrieved 2023.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics*, *11*(6), 619–632.

Roman, M.B. (2022). Steam Games Dataset. *Kaggle*. Retrieved on 5-4-2023, from: https://doi.org/10.34740/KAGGLE/DS/2109585

Rosen, S. (1981). The economics of superstars. *The American Economic Review*, *71*(5), 845–858.

Sonderegger, D.L. (2020). *Statistical Methods II*. Retrieved on 6-7-2023, Retrieved from https://bookdown.org/dereksonderegger/571/

Tambe, P., Hitt, L. M., Rock, D., Brynjolfsson, E. (2020). *Digital capital and superstar firms* (No. w28285). National Bureau of Economic Research.

Thiesing, F.M., Middelberg, U., Vornberger, O. (1995). Short term prediction of sales in supermarkets. *Proceedings of ICNN'95 – International Conference on Neural Networks*. 1028-1031 vol.2, doi: 10.1109/ICNN.1995.487562.

Tsai, W. (2001). Knowledge transfer in intraorganizational networks: effects of network position and absorptive capacity on business unit innovation and performance. *The Academy of Management Journal*, *44*(5), 996–1004.

Wuyts, S., Dutta, S., & Stremersch, S. (2004). Portfolios of interfirm agreements in technology-intensive markets: consequences for innovation and profitability. *Journal of Marketing*, 68(2), 88.

Yang, Z., Cao, X., Wang, F., & Lu, C. (2022). Fortune or prestige? the effects of content price on sales and customer satisfaction. *Journal of Business Research*, *146*, 426–435. https://doi.org/10.1016/j.jbusres.2022.03.075

Zackariasson, P., & Wilson, T.L. (2010). Paradigm shifts in the video game industry. *Competitiveness Review: An International Business Journal*, 20(2), 139–151. Retrieved from https://doi.org/10.1108/10595421011029857.

Zeileis, A. (2006). Object-Oriented Computation of Sandwich Estimators. *Journal of Statistical Software*, 16(9), 1–16. doi:10.18637/jss.v016.i09

Zeileis, A., Kleiber, C., Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software*, *27(8)*. 1–25. https://doi.org/10.18637/jss.v027.i08