The Internet's new Billboard?

COMP 30780 Data Science in Practice

Philip Egan (17489654), Jack Crowley (174844148)

School of Computer Science

University College Dublin

Date

Abstract. This report will discuss how the games marketplace Steam is affected by the active user base of a game and how the live streaming platform Twitch affects the number of players playing a game against how many people are watching and streaming that game . To do this we collected as much data that was available publicly about sales information and streaming statistics. What we wanted to see was how the number of active users of a game affects how much is being spent on these games and if there are trends to this spending. We also wanted to see what the cost of these items that are being purchased are, to see if the items are cheap (<€1), mid range (<€5) or higher cost (>€5), and with that what is the revenue generated from the items in the given bracket, and also the breakdown as a percentage of total cost each bracket is. With the data from Twitch we wanted to look at what engages the audience the most by looking at what games/ categories return the best results in average viewership numbers. The main category we focused on was the "Just Chatting" section and we compared streams that included this category against streams that did not, from the same streamer. We wanted to look at this to see how viewers respond to more focused interaction than just interacting while playing a game and also if there are channels that go against the common trendsWe also wanted to look at changes in the peak playerbase of a game and to see if these changes are related to streamers starting to play the game and are bringing new viewers to the game or it the changes in active player base are encouraging steamers to pick up the game in response and are there games that are influenced more by streamers than others.

Declaration. We (Philip Egan, 1748654 and Jack Crowley, 17484414) declare that this assignment is our own work and that we have correctly acknowledged the work of others. This assignment is in accordance with University and School guidance[1] on good academic conduct in this regard.

[1] See https://www.cs.ucd.ie/sites/default/files/cs-plagiarism-policy_august2017.pdf

1. Introduction

We decided to choose this project due to a mutual interest in games and we both enjoy watching Twitch in our free time. During our initial conversation we started talking about the recent growth of chess on Twitch and both though it was interesting, this evolved to thinking “what makes a game popular on twitch”. From there we started looking at recent games that had spikes in popularity and noticed there were a few of them that also started gaining popularity on Steam. From this we saw that a lot of the games that are trending at the moment have microtransactions built into the games, this then sparked the interest for me in this are in what people are spending their money on in games with microtransactions and how does the value of these items differ in terms of sales from games that are free and games that you have to pay for initially.

after some initial research into the topics some things started to jump out that were interesting to us, like how “just Chatting” had become such a huge part of the Twitch experience and how games that had a large focus on social interactions with other streamers started growing along side the “just Chatting” section. we wanted to look at he differences in viewership between streams that were more focused on the gameplay aspects of livestreaming on twitch and streams that focused more on the interactions with the audience, by reading the live chat and tip messages

to begin with, we needed to get the viewership numbers for games that had “just Chatting” segment and streams that did not, so we decided to use the top streamers on Twitch' past broadcast statistics (collected by sullyGnome) and then pair these statistics with the peak player data over the same time period (collected by SteamCharts.com). the data for the peak players goes back 9 years, so the peak data will start to show if there are trends that could be yearly, like scheduled updates to games that could bring players back to the game, if this is the case it could be the reason for the increase in the twitch views/ peak players. After looking at this we found that games that had consistent peak player base numbers seemed to be rather unaffected by changes in the viewership on twitch, but newer games or games with less established player numbers seemed to be heavily influenced by those changes in the viewership numbers on Twitch.

Then we went back to looking at the comparisons between streams that had dedicated segments for chat interactions and comparing them to streams, by the same streamer, that didn’t have this focused time dedicated to this audience interaction. We could know if a given stream did have this dedicated interaction we had to look for streams that had an included “Just Chatting” section in the data that we collected from SullyGnome, and this data also included the streams average/peak viewership, the followers gained and the viewership per hour. through these metrics we can see if there are changes between the 2 classifications. In most cases we saw that these metrics favored streams that had this streamer interaction, but there were some streamers that contradicted this but they were streams that were more focused on games that had a large social component and this factor is important in the popularity of games in the Twitch space.

When looking at the sale of items on the steam marketplace we wanted to look at different games and different monetization models to see if games that you have to spend money initially to buy the game (pay to play games) have better sales on this marketplace than games without this purchase (free to play games). We expected games that were free to play to perform better than pay to play games, and to an extent we saw this but the pay to play game Counter Strike: Global Offensive was by far the best performed but that may be due to external factors such as a strong relationship with betting and gambling but with that it is also one of the few games that has had items available on the marketplace for 5+ years and has the second largest player base of all games what have items on this marketplace. For most games we could see that when a game has a consistent average player base the amount of money the average player was spending tended to rise but games where this average number fluctuated frequently the amount the average player was spending fluctuated to a trend of the less players the more each player was spending, the exception tho this was the game rust that had a spike

in amount each player was spending just before the spike in playerbase, as the game was starting to gain popularity on Twitch.

when looking at the spending habits on the steam marketplace we gathered to sales history of all the games and took each items average price assigned it a classifier of low, items costing less than 1 euro, medium, items costing less than 5 euro but more than 1, and high, items that cost more than 5 euro. Each of these items had their total revenue per day calculated and then graphed. We expected items that cost less than 1 euro to make up a large amount of sales and high cost items to be less, but it turned out that high cost items were the majority of the sales and bar the first few months of the marketplace existing, medium cost item sales were extremely low. The volume of items in the low price category were the highest and items in the medium cost had the lowest sales volume.

The questions talked about will be explored in more detail in their later sections, following this we will be discussing the motivations, objectives, talking about some of the related work around these topics and how we collected the data used in our research questions

2. Motivations & Objectives

2.1. Background & Motivations

The initial reason for choosing this topic was a mutual interest in the gaming space, as both of us enjoy playing games with a competitive aspect to them, and these games tend to have a strong link to Twitch. the Twitch platform allows anyone to stream their games and in turn anyone to watch others broadcasts. Twitch's broadcasts can be generalised down into two major categories, streamers that are extremely skilled in a game/ games and gain viewership due to this factor primarily.

Shroud (3 most followed streamer) is a retired professional Counter Strike: Global Offensive player that turned to streamer and gained popularity sharing himself playing the game and games like it for others to watch while a streamer like Pokimane (6th most followed streamer) are watched due to their strong and likeable personalities, and streams that have a huge amount of viewer interaction and time dedicated to this factor. The streamers used as examples are both considered giants of the space and bring massive live viewership numbers consistently.

Another space that has been using Twitch to grow is the esport space, most games with few exception broadcast their competitions live on Twitch, this is a relationship that will bring people from games into the Twitch space, and in turn allows Twitches normal audience to find these events and potential want to try the games being played. Some of the most well known examples of this are Rocket League, Fortnight and League of Legends, both these companies host their live broadcasts on Twitch and have follower counts of over 3 million. This also includes companies like ESL that run events for a large number of games, this channel has over 5 million followers. Twitch made itself the hub of esports, and that was something we both enjoy watching, Chess being the common game for us due to recent events held on Twitch.

When it comes to the steam space Philip wanted to know if the number of people steaming games on Twitch related to people actually spending money or playing the games more, so we came up with the idea to look at these factors and look for relationships between them, and then to look at the games themselves to see if there was anything that could be gatherer to show the worth of having people playing the games consistently, and this is where the marketplace came into play. The marketplace allows players to sell items at a price they decide and see fit, this will allow for the players to decide the value of the items which in itself was interesting to Philip.

We decided that we would focus our project on these two intertwined spaces, with Philip focusing on the Steam space and Jack Focusing on the Twitch. This division of labour allowed us to cover more and gave us more space to work free from the constraints of the other as we knew that the data collection would take a long time for the Steam data due to the quantity. This also allowed us more freedom to play with the data and mold it into a way we found most useful while still allowing the other to use it when they wanted.

2.2. Related work

There is a large amount of work done around these topics from a few different angles, the first piece to talk about is “Analysis of the Characteristics and content of Twitch Live-Streaming” by Daniel Farrington, this report focused on what attracted people to watch streams and what people enjoyed watching from a more technical sense. Daiel’s findings also suggested that esport games held popularity better than non-esport games, “ Out of the 4 eSport games, Dota 2 was the only game that moved in popularity by more than 4 places. Minecraft on the other hand, fluctuated between 10 popularity places over the course of our study” (Farrington 38) but he also used a short same time of less than 25 days. This will be broken down more in research question 3

The second report we looked at was “Do Streaming Metrics on Twitch Affect Game Sales” by Nicholas Sherwin. This was the closest paper we could find that tackled a similar perspective as ours in looking at the relationship between game sales and Twitch viewership and this report pointed us in the direction of SullyGnome. This report came to an interesting conclusion that twitch was not a good source of mass advertising but for targeted advertising and used the analogy “watching your favorite band in a massive arena vs. an intimate 500-person club. Advertisers and game developers can similarly take advantage of that tight-knit and niche community to effectively target and communicate at a more personal (and higher-converting) level.” (Sherwin). We believe this to be an accurate description as the viewership of individual streamers may not be at a level that can compete with traditional media yet, it has a very focused audience where targeted advertising could be extremely effective as its going directly to your target market.

The website SteamSpy was created by Sergey Galyonkin and is a huge database of information about the number of people playing games and estimated sales information base of the number of people playing the games on the Steam platform. This website was valuable in figuring out what games we wanted to look at for potential case studies and general inspiration for what we wanted to be able to show by the end of the project.

SullyGnome is a database of Twitch information that was gathered by an individual and again it was interesting to look at then data with on overall view, this allowed us to see if there was anything that started to stand out and pique our interest before diving in, it was also the source for our Twitch data as it was all-inclusive

the streamer Ludwig ahgren was another source of information in how streamers monitise and some of the drawbacks to it he talks about how in exchange for a $5 you can broadcast a message to everyone watching and the streamers themselves, and he also talk about a more unethical advertising method in the video “I MADE money by donating to streamers” (ahgren). In the video “I am not your Friend - a Twitch Documentary” he talks about the parasocial relationship views can have with streams and how the lines between entertainment being made and the having a social relationship with the viewers is being blurred and how harmful to those involved (Atallah)

2.3. Research Questions

2.3.1. RQ1: what is the efficiency of sales per gamer on average

This will give us a chance to look at all games that have items for sales on the SteamCommunity Market, and get daily total sales for each game on a given day and see how much each person is

spending on average on a game. We can also look at games with different monetization options (free to play and pay to play) to see if there are differences and if there are seasonal trends in sales.

2.3.2. RQ2: how are people buying? are they buying cheaper items in greater volume or more expensive items at a lower volume

this was to see that are the favorable price points for buying and selling in game items, again we can look at the games monetization models between games and we will be able to get a percentage breakdown of the total spent at each price point. We decided to look at 3 price points, less than 1 euro, less than 5 euro but greater than 1 euro and lastly more than 5 euro.

2.3.3. RQ3: Do Streamers get more viewers while playing a Game or while just chatting?

The idea behind this question was to see whether or not a streamer gets more viewers when only playing a videogame compared to when there is an aspect of social interaction with the viewers, which is denoted by the phrase “Just Chatting”. We wanted to see if it's beneficial for a streamer to increase their views and therefore have a greater audience on average.

2.3.4. RQ4: Does Playerbase increase before or after a streamer starts playing a game?

This question is based on the two datasets that we created. We looked at the player base statistics from a select few games and also looked at when streamers played those types of games. We wanted to see if the player base is affected by the streamers or is the streamer affected by the increase in playerbase which would also correlate to an increase in the games popularity.

3. Data Wrangling

In this section well look at Data Acquisition and Data Cleaning and Preparation and bring you through the steps in which we took to do both these events.

3.1. Data Acquisition

The data we wanted to collect was divided into 3 parts, the games player data, the Steam community market data and the Twitch data.

The player data had a number of different potential sources but the one that was the best for us was the SteamCharts.com. This site was so good for us as it collected the data in monthly intervals consistently going back 9 years, a similar timeframe to the data that could be collected for the steam community market. The site has a simple system for each game's data, the url followed the same pattern “https://steamcharts.com/app/(SteamAppID)” so the first step was to collect all the app ids from steam, an app id is the game identifier on steam. the app ids were available at “ https://api.steampowered.com/ISteamApps/GetAppList/v0002/ ” and this data was in a json format, some small amounts of cleaning was needed to get the data in the right string format, removal of additional brackets and quotation marks, in total there were 114461 uniques appids, the app id collection was done in the notebook called “AlSteamAppIds”. These appids can then iterate through them to collect the average and peak player data along with the data of each collected datapoint. The web scraper BeautifulSoup was used to do this task and some processing needed to be done to format the data into a manipulatable form. This was done in the “concurrentPlayers' notebook. this data was then further processed in the “compiledPlayerData" to change it into a csv

For the Steam marketplace data the first step was to get a list of all games that had items of sales in the community market, this list was gotten from “https://steamcommunity.com/market/” in the

notebook “SteamComunityAppIds” and looking at all the games that had links on the page, which was a complete list. this data was again scraped from the webpage using BeautifulSoup and collecting all data in the html class “game_button”. From there the next challenge was going to be getting a full list of the items that were for sale for each game that had items for sale, this was done in the notebook “ItemsPerGame”. This was run for each game individually, this was done due to how unreliable the steam api calls were along with inconsistent error messages. there were over 200 different games but the program only ran for a couple seconds for the majority of the games so it was tedious but the result could be guaranteed. After this was done the files were filtered to remove games that had no sales or games we had no player data for. This removed a majority of the games. the market history data was collected using the “marketHistoryPerItemRequest” notebook, and again this had to be done with heavy monitoring for each game, due to errors of unknown origins. This Notebook will run for an extremely long time and was run automated through once for 24 hours +, this missed a large amount of items from games with more than 100 items so that had to be run individually. The format was a json file that was surprisingly difficult to work with and the data call was using the storefronts internal API system, by making a PHP request to the servers it would return the requested data in a raw format that could be used.

In the notebook “marketHistoryDatasheetCreation” due to the datetime format used by the Steam community market there were a large number of issues getting a consistent datetime for all the datasheets, this was corrected in “def csvtoFile” by converting the string initial datetime to a formatted datetime using the datetime packages method “strptime” and then further formatting it into a month day year format, this was the easiest format to use for the datasheets created

The Twitch Data Collection initially started off with using the Twitch API but a problem arose with both the API and the data that was collected so all focus was then placed on another website called SullyGnome. This website gathered data from the Twitch API and aggregated it for us so it was perfect to scrape from. We looked at taking every stream from the top 500 streamers over the last 5 years (2016-2020). This ended up being far too much data due to a time constraint on the project. So we let the scraper run for three days and then used everything we got from it.

We built the scraper using selenium which is an automated web driver. To start off we made a for loop to create the thousands of urls we needed to scrape from. The url would be passed into the function we created to scrape each website and that would be run constantly. An early problem arose when we realised that the games the streamer played wasn't connected to what we were scraping. So all the data we were collecting was pointless without knowing what games the streamer was playing every stream. We had to create another function to be called within our main scraping function called “get_games”, this function would work off the main function so every time a url was accessed it would open the same url and scrape the game data from a different area of the page. Then we would append the data scraped from the url twice together and save it in a csv file labelled whatever the streamers name was. This scraping tool was very slow as we ran into server difficulty with the web page and kept having to restart the scraper.

3.2. Data Cleaning & Preparation

All majority of the data needed basic manipulation in order to make it usable, the code for the Appids needed to have additional brackets and quotation marks but this was done using the .replace() method in the re imported package. The games that had items for sale on the market had to be cleaned as they were gathered by scraping a webpage so by using the re package again the re.findall(r'\d+',x) was used to get the games Appids in a raw form. The market history data was in a format that didn't require much cleaning but it required the item names to be in a specific format. once all the item names were collected from the “ItemsPerGame” notebook. The conversions are shown on the left, this was done using notepad++ as it had an efficient find and replace while also allow me to look through to check if it was a game that we did not want to include (Explicit games and games that were made to be used in item trading scams) none of the game removed had any quantity of sales or a large quantity of

items. Once the data was requested in this format it would allow us to access the full market history for each item but due to the request failing this was a tedious process .

The way we went about cleaning and filtering the Twitch Data started the second we scraped it from the website. Due to the fact we were under some time pressure we thought it would be a good idea to have csv files already filtered and ready to be created into one large and clean dataframe.

We did this by creating a dataframe everytime a url was done scraping and by creating columns for the newly gathered data to be stored in straight away. Also used a lot of small alterations of the csv file to make it as legible as possible and easily appendible to the next csv file that was going to be created. Once all the csv files were created and appended together to create the final large csv file there was still some slight cleaning to do. We had to get rid of any hiccups there with the web driver as sometimes it would time out and not download any data so we had to remove all rows that were empty. Also a lot of extra columns were created by accident so they also had to be removed. We had a small problem with duplicate dates as some streamers would stream more than once a day so that had to be taken care of. All dates that were collected initially from the websites were full dates and times which was unnecessary so we filtered them out while scraping and changed them all into dateTime so that they all are the same data type once in the dataframe. A problem we faced was that the games all came in a list that was already quoted, example being(“ ‘Dota2’ “). This did not lead to any problems as we just quoted the games correctly and split the list up into columns of the dataframe so each game in the list was a new column in the same row.

4. Data Analysis & Results

This is a critical section. In it you will probably focus on each RQ in turn and carefully describe how you went about answering it, and the key results that you found. In addition to presenting the results you need to discuss what they tell you, and their deeper meaning as it relates to your project.

4.1. what is the efficiency of sales per gamer on average

4.1.1. Datasets

The data that was important for this questing was the total sales per day per item information that was collected and stored in a file system where each game had a file for each item. The data initially was quite raw so it was compiled down to a file with all a games daily total sales, this was done in the notebooks “marketHistoryDatasheetCreation” then “manyToOneCsv”. This allowed the data to be easy to read. there was also a data sheet created of all the totals to make it easier to see the data in one place. After all the filtering of games without data / missing game data there were only 33 games that were remaining.

4.1.2. Approach

the way we approached this question was to get a look at the average viewership and look at the sales charts, when looking at these and when talking it out we decided the best way to show that data in a way that could be compared across multiple games, and the way we decided best was to sum the daily figure for sales and change them into monthly intervals, this allows for the data to be far more readable and makes trends more obvious. When this was done we normalised this data with the formula (totalSalesPerMonth) / (AverageplayersPerMonth) and this was done every month. This is a formula that could be applied to all the games without any issues and gives us a number for how much is being spent per month by the average person playing the game. All this was done in the notebook “NormalizeGameDataToSalesData”. The big thing we wanted to look at was how the different monetization models compared, and weather free to play games or pay to play games perform better on average.

4.1.3. Results

when looking at the graphs it became quite obvious that the best performing game was Counter Strike Global Offensive (CSGO) and by a huge margin, when the items were first available in 2012, the average player was spending upwards of 150 euro, and by 2014 they were spending over 300, they average play numbers were quite small at this time but due to having a relatively high number of big spenders (Whales) were able to push this number up greatly. These numbers lowered to its more consistent rate by late 2015 but still this number is hovering around 100 euro per player mark. The average spending lowered in 2019 but this is due to an influx of new players that was caused by the game changing its monetization model to free to play. Comparing CSGO to the second most successful game Team Fortress 2 (TF2), a game that had at peak the average player spending less than 60 euro on average along with having a far lower average player base it is earning massively less while having a free to play monetization model, even though these games have been developed by the same company Valve.

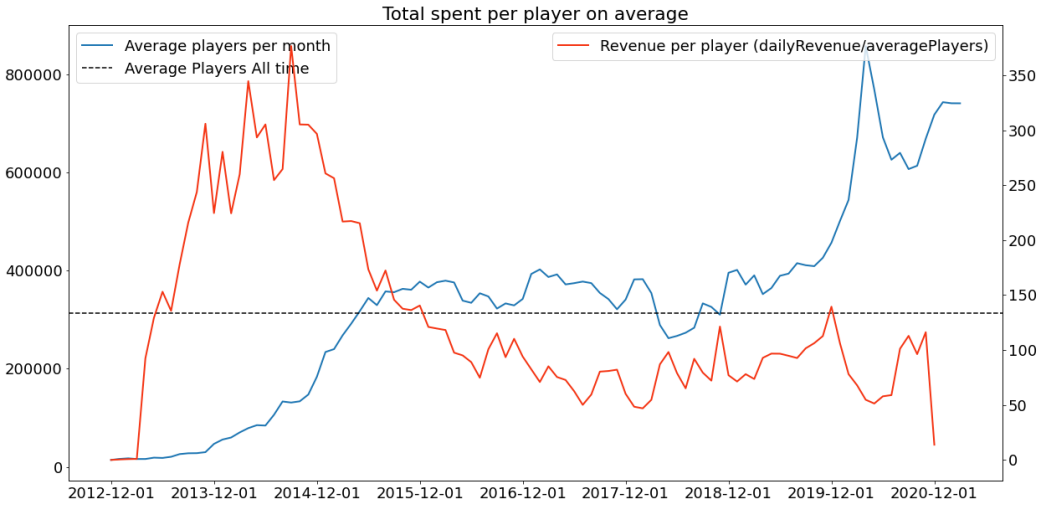


Figure 1 : Counter Strike: Global Offensive average spending per player and average player graph

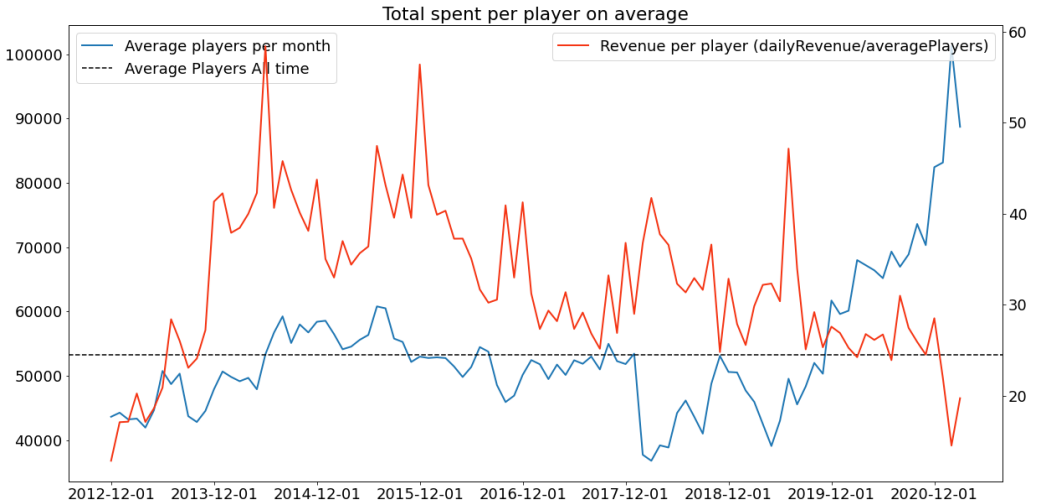


Figure 2 : Team Fortress average spending per player and average player graph

4.1.4. Discussion

We did not get the results we expected at all, our initial thinking was that a game that is free will have more people buying times and spending money than a game you have to pay to play initially. It seems strange to buy a game and then spend more money buying more items that are purely cosmetic, but it seems the general playerbase disagrees. We spent some time looking into this to try find potential reasons for this and there seem to be a few. The first being the items managed to establish value outside of being a cosmetic item in a game, they started to quickly hold high values in 3rd party markets with items selling for prices higher than the Steam community market allows (Max sales price on the community market is 2000 dollars). Another is the scarcity of items, the majority of items are obtained from random drops either after a game or in a "case" which gives a random item from a pool of items but the items have different drop rates, and the lowest rate being 0.31% for a knife skin. With each case costing 2.10 euro to open meaning the average cost to open a random knife skin is around 677 euro so on a market they can cost more depending on the perceived value of the skin, generally based on the appearance. From around 2016 the value of items started to rise again due to the rise in popularity of CSGO items being used as a currency on betting sites and gambling sites and the items have been a staple ever since. From this looks like items that have value outside of the game will demand higher values and be more popular to buy.

4.2. How are people buying? are they buying cheaper items in greater volume or more expensive items at a lower volume

4.2.1. Datasets

for this question we created a new data set from the market history files for each item that was created, this new data set was to reassign the items into 3 different values, low being items that had an average cost less than 1 euro, medium being times that had an average cost less than 5 euro but more than 1 euro and high which contained items that had an average cost more than 5 euro. These data sets were created for the top 5 games, as sales numbers for games below this were too low to matter. The top 5 games were Dota 2, CSGO, TF2, Payday 2 and rust. Once each game's items were grouped into their classifications a new data sheet was created for each classification of items in the game. this was all done in the "SalesAsAPerportionOfItemsAverageCost" notebook

4.2.2. Approach

The approach taken was to create a larger dataframe that would contain all items of a specific item bracket in each of its columns and then have them totaled for each day. This was done to create a clear separation between the categories and not to have several subcategories for each game's items. When this dataframe was created it was then translated into a new dataframe that would give the breakdown as a percentage to give an area plot, this was done because the plot is hard to read in a raw form and it's hard to tell large changes apart. We expected the largest proportion of sales to come from an extremely high volume of low value sales, and then the medium bracket to be the second largest proportion of sales

4.2.3. Results

the results again were not what we expected, the high value items dominated the total spending and when looking into this we found that most of the sales were dominated by CSGO with it having 81% of all sales, and CSGO was also the games that had this highest proportion of items that cost more than 5 euro, together these items attributed for the majority of the sales in this high cost bracket.

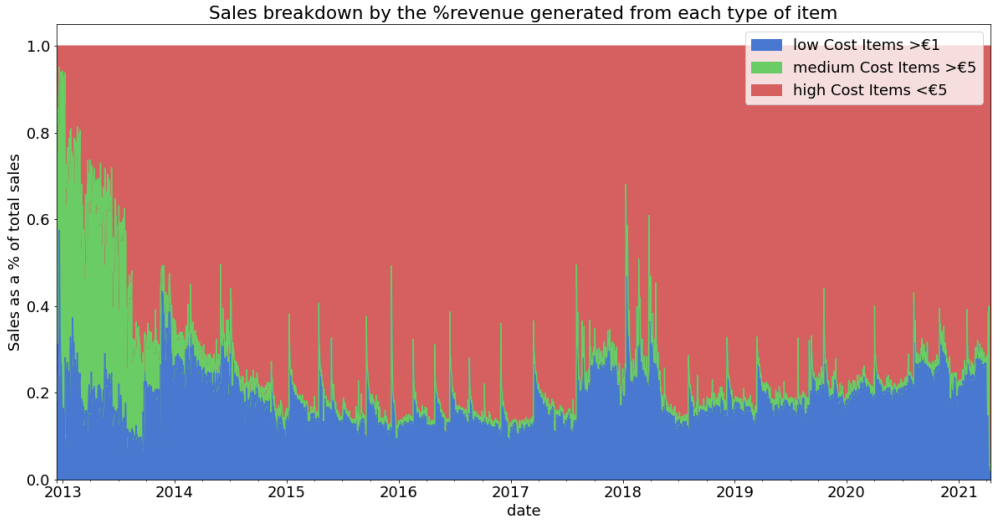


Figure 3: sales as a proportion of the cost of the items for top 5 games showing the massive amount of spending of high cost times

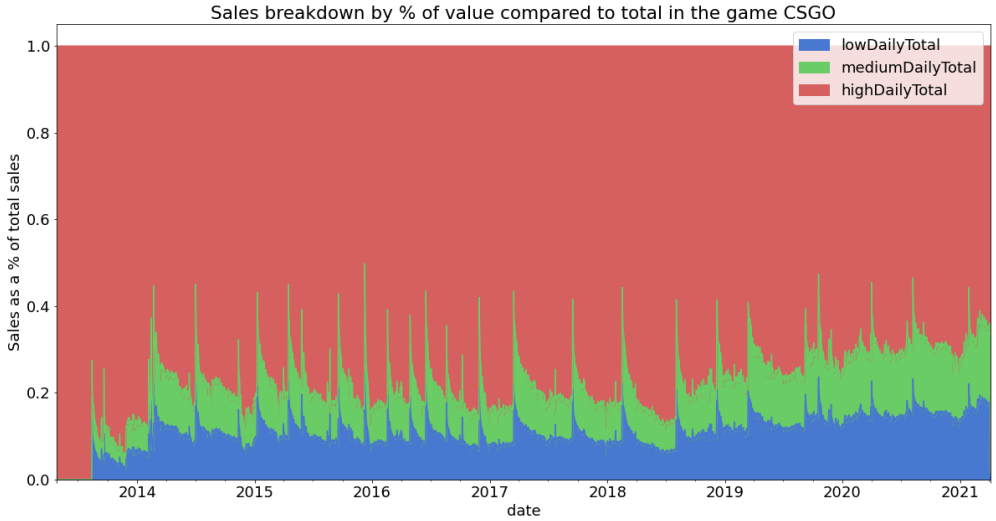


Figure 4 : Sales as a proportion of the cost of the items for CSGO

4.2.4. Discussion

We were quite surprised by the amount of money that was being spent on items over 5 euro. considering that 3.95 billion euro worth of sales have been made on the marketplace, it's amazing to see that people are buying items that cost more than 5 euro enough to make it become the massive majority of this revenue generated. CSGO making up 3.2 billion of these sales shows the value in the items having value outside of the game and marketplace on third party websites can massively influence the sales of a game.

4.3. Do Streamers get more viewers while playing a game or “Just Chatting”

4.3.1. Datasets

The data that was required to answer this question came from the ”Games” and “Avg viewers” columns in the Dataset of Streamers and was grouped through every streamers name. To create this new DataFrame we had to break up the list of Games that the streamers played per stream. we then created a for loop to go through every stream we had and if the stream contained “Just Chatting” to then add it to a separate DataFrame as a separate column to streams that didn’t contain “Just Chatting”, and we compared the total average viewers of each streamer when playing only games or when they also chatted.

4.3.2. Approach

We decided to take every stream of every streamer and create a Dataframe that contained when they just played games vs when they also chatted. This way we could easily tell the average viewers a streamer gets when doing either type of stream. It also allowed for easy calculations to tell the difference between the two types of stream and how well they were doing for a certain streamer. Unfortunately we didn't have access to certain metrics that could have led to interesting findings like whether the person was male or female, this might have shown that males get more views when chatting or vice versa.

4.3.3. Results

The results were very interesting, the average viewers for streamers while chatting was a part of their stream was significantly higher than when just playing games. On average the increase of viewers when chatting was 26% with the highest difference being 68% and only 12 out of 38 streamers on average had more viewers while just playing video games. The average viewers while gaming is about 10000 people, whereas the average while also chatting is 15000 quite a substantial difference especially on a streaming platform.

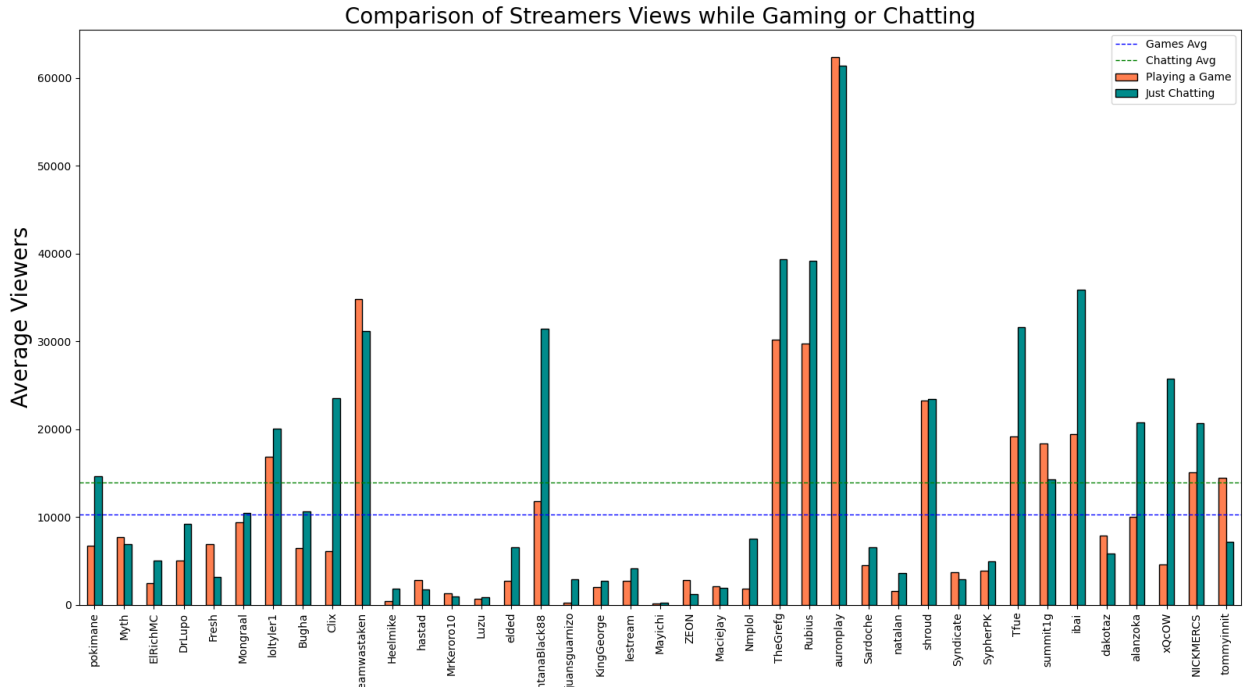


Figure 5: Comparison of Streamers when Gaming vs Chatting

4.3.4. Discussion

The results were not what we expected. With Twitch being a platform for playing video games live we would have assumed that games were the prime interest holder of the viewers, but as clearly shown above majority of the time some form of social interactions with the people watching gains you more viewers. This is helpful knowledge for both the streamer themselves and also any company that wants the streamer to advertise something for them.

4.4. Does Playerbase increase before or after a streamer starts playing a game?

4.4.1. Datasets

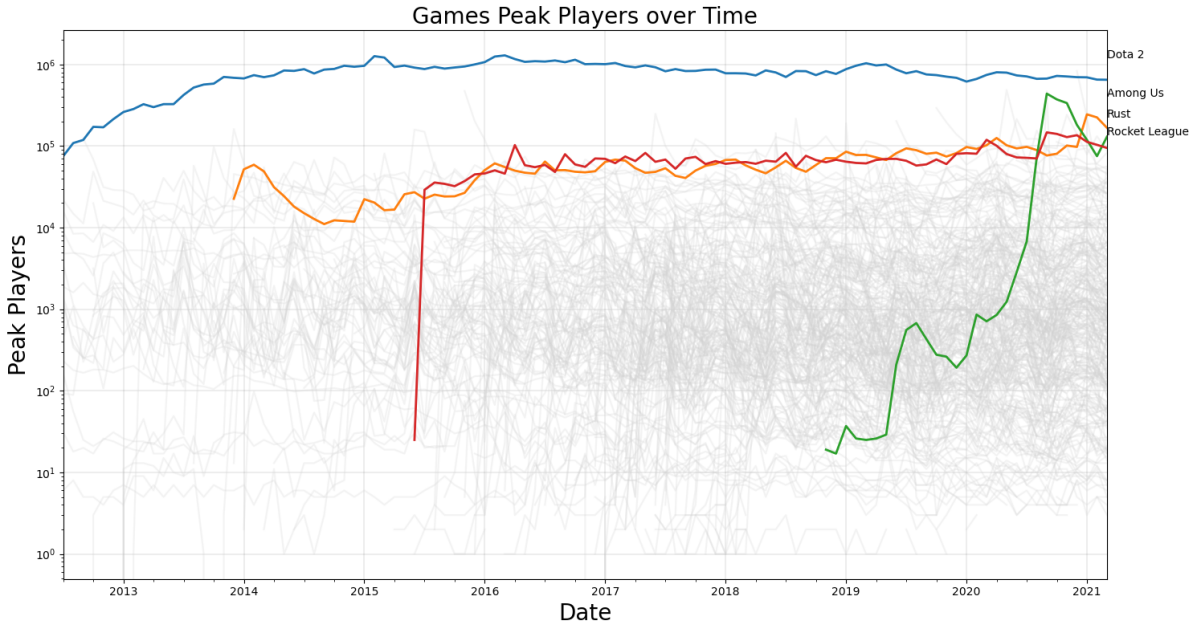
When looking at answering this question we needed to combine both of our datasets to be able to create visualisations that clearly show if there's any influence of streamers on games or if it's vice versa. We used the Steam Data to show how a games peak player base did over time and also took the Twitch Data to show the times when streamers streamed that game and how many viewers they got at a time. We needed this to see if streamers were influenced to play by the games popularity or if the streamer made the game popular.

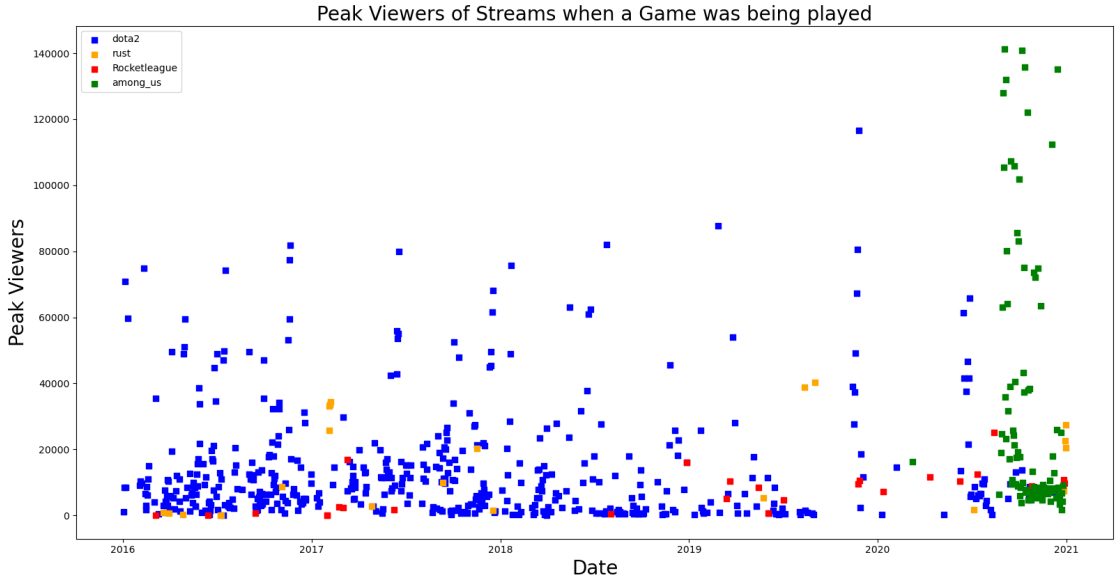
4.4.2. Approach

So initially we had to combine a list of names of the games with the data we had from each game as each game's data came without its name in each row. This was solved by a simple for loop that added a column with that games name to each row of whatever game file it was and then combined all the files into one to create the game dataframe. We then plotted the game frame and selected 4 games that stood out the most and seemed to be the largest, these being (Dota2, Among Us, Rust and Rocket League). We then simply took from our main data frame for twitch streamers any instance of the game being played and took how many peak viewers they got while playing the game. This was so we could see if there was any correlation between the streamers and the games over time.

4.4.3. Results

There were some very interesting results that we found. Dota 2 spikes once a year but not by much as its player base is so huge and dedicated that streamers dont have much influence on whether the game is played or not. Rust spikes when there is an update and also leads to an increase of streamers playing and viewership which would indicate the game having influence on the streamer.Among us blew up in 2020 due to streamers and player base followed the increase. As you can see in the graphs the increase of player base and viewership is very similar. Rocket League would only increase when there was an event but it wasn't influenced by streamers as the viewership for streamers was quite low.





4.4.4. Discussion

Based on the results that we obtained it was clear to us that certain games were influenced more than other games. The Games that were influenced the most by streamers were the ones that never had a playerbase before. Games with a big following and dedicated players were not influenced by streamers unless there was an event for example a tournament or a large update. A good example of this would be PogChamps for chess or when Rust has an update, both the player base and viewership increase significantly.

5. Discussion

From the steam perspective, it is incredible how much is being spent on the games at the top and especially in CSGO. For a game to generate so many sales and in high volume for items that can be

purchased within the game and hold no physical form. There are examples of games holding significant value, for example, magic gathering cards hold an average price of around 6 euro. The other point that stuck out to us was how little items between 1 and 5 euro were being sold and how little of the overall percentage they were contributing to the total sales. We believe this is due to things being under a euro are quite an easy thing to impulse buy and with the more expensive items holding more value in 3rd party websites they have more of a purpose in buying. The items in the middle don't fill into either of these categories. The games CSGO and Dota2 also have the higher cost items on display in the esports game, this again adds value to the items as people want to have the same stuff as the "famous" people. Games outside of the top 3 games are making on average 1153 euro in daily sales, and with them the average rises to 41259 euro, so the difference between the average of all and the average including the top 3 is immense.

When looking at the Twitch Data and what we could see from that it was clear that streamers do have an influence on thousands of people daily. We touched on parasocial relationships and how people are drawn to these streamers as friends. When you have a friendship with someone you usually like the same things and when the viewer is watching the game that is being playing by someone they have a perceived connection with the game appears more fun as they see them playing the game in the same situations and are having fun with the game before actually playing it, so when looking to buy a game will have artificial positive opinion of the game. This is also a reason why "Just Chatting" does so well on Twitch because people want to have that relationship with someone and they feel like that person is their friend. Only the top streamers were considered in this study but they make up 99% of the viewership on Twitch so most viewers watch the same people and the same people play the same games.

5.1. Ethical Considerations

We don't have any real ethical considerations due to the nature of our data

5.2. Reproducibility

Our project is very Reproducible due to the nature of our data collection and code created for that. Our code allows for an individual to press run on the computer and after however many hours they should be able to produce our project. One problem with our Reproducibility is that they would have to stop data collecting at the same points we did in order to obtain our results exactly. If they allowed for more collection of data then they would more than likely obtain a more in-depth version of our project but more than likely the results would be the same.

Our steps taken to collect ,process, and analyze data are quite easy to understand even with a very small knowledge of coding or computers. These primary reasons help make our project reproducible.

5.3. Limitations

The limitations to the steam data is Valve not sharing the number of owners of a game and the in-game sales information, this means we have to focus on the number of people playing the game and the sales on the Steam community marketplace, which is limited to games that have made this a part of the game. If we had full sales information we could make more detailed graphs. There are limitations to the collection of the data, the community market is inconsistent in the number of requests that can be made and it seems to only have limited per game, so small games rarely had issues but games with a huge number of items would fail regularly.

The limitations for the Twitch Data was we didn't have the breakdown of each game that was being played during the stream and how many viewers they had while playing just that game. We only had the bulk data of the entire stream, so it would've been better to know how many viewers someone had

when playing a certain game and we could then even find the popularity of games based on that. While collecting the data from our source we also ran into problems with servers crashing due to the high amounts of requests that were happening every second, and so some of the data would be corrupted or very poorly collected. All that slowed down the process and didn't allow for all the data we wanted to collect to be downloaded in the time frame.

6. Conclusions & Future Work

Our main points that we take away from this project are quite helpful in answering our initial question. We gained knowledge of the steam marketplace and how unless you are a major game where your items have a higher perceived value then you're not going to be very profitable, so if you are a small developer you should start by having your game pay to play, and as player base increases then you can start introducing micro transactions. A benefit we found though for smaller developers who are wishing to increase their player base and revenue is that streamers have a very large impact on the player base and viewing of smaller games. Games like Among Us as an example shot up by player base and views due to Streamers creating groups who play the game together as friends and created a community for people to join.

This leads onto our next point that we found, due to Among Us being a very social game as it's all about communication with your teammates. We noticed that games with a social aspect have a greater increase of players due to streamers than a competitive game.

Interestingly though competitive games are very popular when events are held on Twitch or by streamers. We noticed spikes in games that were quite linear and found that there are three reasons for a spike Updates, Tournaments and Streamer Events. So competitive games in competitive settings lead to an increase of players and viewers.

Streamers who only play video games when they are streaming get on average 26% less views than a streamer who engages with their chat for a certain amount of time per stream. The highest difference is with MontanaBlack88 being 68%, as he gets 68% more viewers when he devotes a section of his time to chatting to his viewers.

So, to conclude and summarise, if a small developer is starting a new game there are a few factors they need to take into consideration. The game should be pay to play until a large enough player base forms, and the way to increase that is by paying a streamer to play their game. It can't be any streamer though it must be one in the top 1% and preferably one that chats to their audience because as we saw a difference of 26% viewers is quite substantial. Finally, they should create a tournament to host the streamers playing the game to allow for a huge audience and therefore increasing the player base and revenue. To answer the main question are streamers a good form of advertising and are they the new billboard, we would say yes.

If we were going to continue working on this project there are a number of additional factors we could have looked at in order to give more informative results. With more time we could have done analysis of the streamers individually and scoring them based on predetermined parameters to see what streamers are most valuable to an advertiser going forward, and in this we would have included an analysis of their chatrooms to give an idea of the interactions between the streamer and the audience. with the Steam analysis if possible getting access to all sales information of all the games on sales from Steam itself would allow for more accurate information and would likely change a large number of charts due to being able to include sales number in the average monthly earnings.

7. Responsibilities

We split the responsibilities into 2 sections. These sections dealt with each of the platforms that our project concerned, Twitch and Steam. Philip took Steam marketplace and playerbase data and Jack took Twitch. During the process of splitting responsibilities we made sure the workload was evened out so one person wasn't getting too difficult a task.

When it came to the Research Questions Philip did the work on RQ1 and RQ2 whereas Jack did the work for RQ3 and RQ4. This includes data collection, filtering and creating of the visualisations.Philip helped with the data collection for the Twitch data due to the error that occurred with the initial Twitch data being corrupted and lost. Due to this issue we both ran the data collection code due to the time constraints

When it came to the creation of the weekly updates and slides we used the previous week's template and typed in the work that we had done during that week, so all that work was split evenly.

For the writing of this Final Report the workload was split by section. Philips Sections were the Introduction, Motivations and Objectives,Related Work and Bibliography. Jacks sections were the Discussion, Reproducibility, Limitations and Conclusion. This splitting of the work is only the typing there was input from both parties on what to write in these sections so the whole project is 50/50 for typing of report. Of course our research questions and data acquisition/preparation is based on our own work so that was written by each other.

8. Bibliography

ahgren, Ludwig, director. *I MADE money by donating to streamers*. 2020. *Youtube*, https://www.youtube.com/watch?v=v7ufQ5Sz-no.

Atallah, Patrick, director. *I am not your Friend - a Twitch Documentary*. *youtube*, https://www.youtube.com/watch?v=wP5ysTqEj7I&t=0s.

Farrington, Daniel. "Analysis of the Characteristics and content of Twitch Live-Streaming." *An Interactive Qualifying Project Report*, vol. 1, no. 1, 2015, p. 63. *web.wpi.edu*, https://web.wpi.edu/Pubs/E-project/Available/E-project-031915-220004/unrestricted/Analysis_of_the_Characteristics_and_Content_of_Twitch.tv_Live-streaming.pdf.

Sherwin, Nicholas. "Do Streaming Metrics on Twitch Affect Game Sales?" *Towards Data Science*, 24 9 2019, https://towardsdatascience.com/do-streaming-metrics-on-twitch-affect-game-sales-cbb4e0ee90e0. Accessed 04 05 2021.