Project Title: An Analysis of the Top 1000 Twitch Streamers

Group Members: Logan Kreisher

Abstract:

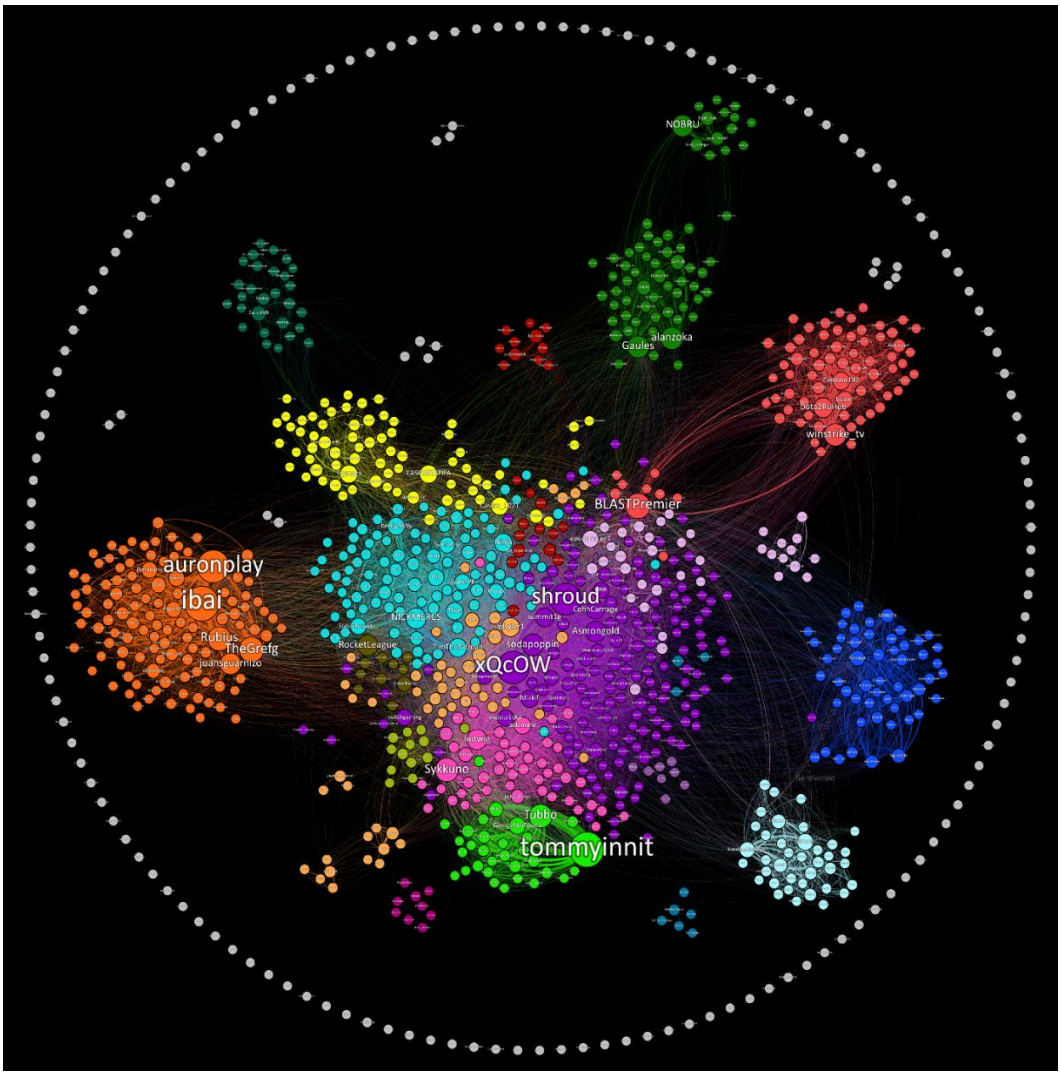
Live-streaming video games has become more and more popular over the last few years and is becoming a larger sphere of online entertainment. As streamers become larger, it is important to see what factors contribute to the success of large streamers. Looking at data of the top 1000 streamers might show us important properties of what helps make a streamer become big. It is also important to view other aspects of the platform such as viewer distribution among the top 1000 channels to see if the platform is growing as a whole or if certain channels are leading the pack with a considerable gap.

Introduction:

In the last few years, live streaming platforms continue to grow, both with users, as well as streamers. Of these live streaming platforms, Twitch seems has been one of the largest. While other platforms such as YouTube Gaming are growing, a large majority of streamers still use Twitch as their platform of choice. By taking a deep dive into the top thousand of streamers on Twitch, it might be possible to see what helps make a streamer successful in terms of watch hours. Looking deeper into top Twitch creators is interesting because video game live streaming is becoming increasingly popular and is now a legitimate form of entertainment for many people.

Because of Twitch still being a relatively new and rising form of entertainment, there has not been a ton of prior work in terms of visualization directly with Twitch. However, there are a couple of good relevant visualizations that relate to Twitch in some sort of way. This first

visualization below is a visualization of Twitch's community as a whole. The visualization shows how tight-knit certain communities are with each other. Many communities are tight-knit while others are not as close. One critique that I have about this visualization is that most of the communities that appear to be far away from others are communities of other languages. Focusing on just one language of content creators might yield more interesting results. Another critique of this visualization is that it uses many colors, far more than our brains can interpret at one given time. Lastly, the communities are categorized currently in very broad communities, many of which are just their language such as French, Spanish, and Turkish.



Color	Community Name	Top Streamers in Community
	Variety (English)	xQcOW, Shroud
	Spanish	Ibai, auronplay
	First Person Shooter (English)	NICKMERCS, TimTheTatman
	Russian/English CS:GO	BLASTPremier, Dota2RuHub
	German/English FIFA	Trymacs, MontanaBlack
	Portuguese	Gaules, alanzoka
	League of Legends/Polish	loltyler1, Jankos
	OfflineTV and Friends (English)	Sykkuno, Pokimane
	French	Gotaga, LeBouseuh
	Turkish	KendineMuzisyen, Elraenn
	English Dota2/Thai	epicenter_en1, AdmiralBulldog
	Minecraft (English)	tommyinnit, Tubbo
	Italian	iMasseo, Tumblurr
	English Apex/Japanese	fps_shaka, TSM_ImperialHal
	GTA V	Ramee, RatedEpicz
	Rocket League/Brawlhalla (Eng.)	RocketLeague, SqushyMuffinz

Below is a visualization for a similar platform, YouTube. While YouTube does have a live streaming part of their website called YouTube Gaming, the main focus of YouTube has been to play back prerecorded videos in the format of on demand. While Twitch is nowhere near as a large as YouTube, live streaming content is starting to see growth similar to that of the early days of YouTube. It could be possible that we see live streamers that reach follower numbers of the YouTube channels below in this visualization. One critique that I have about this visualization is that having profile pictures to the right of each bar is distracting and unnecessary. It could also accidentally manipulate our perception of how large a channel is due to the profile picture seemingly increasing the length of each bar.

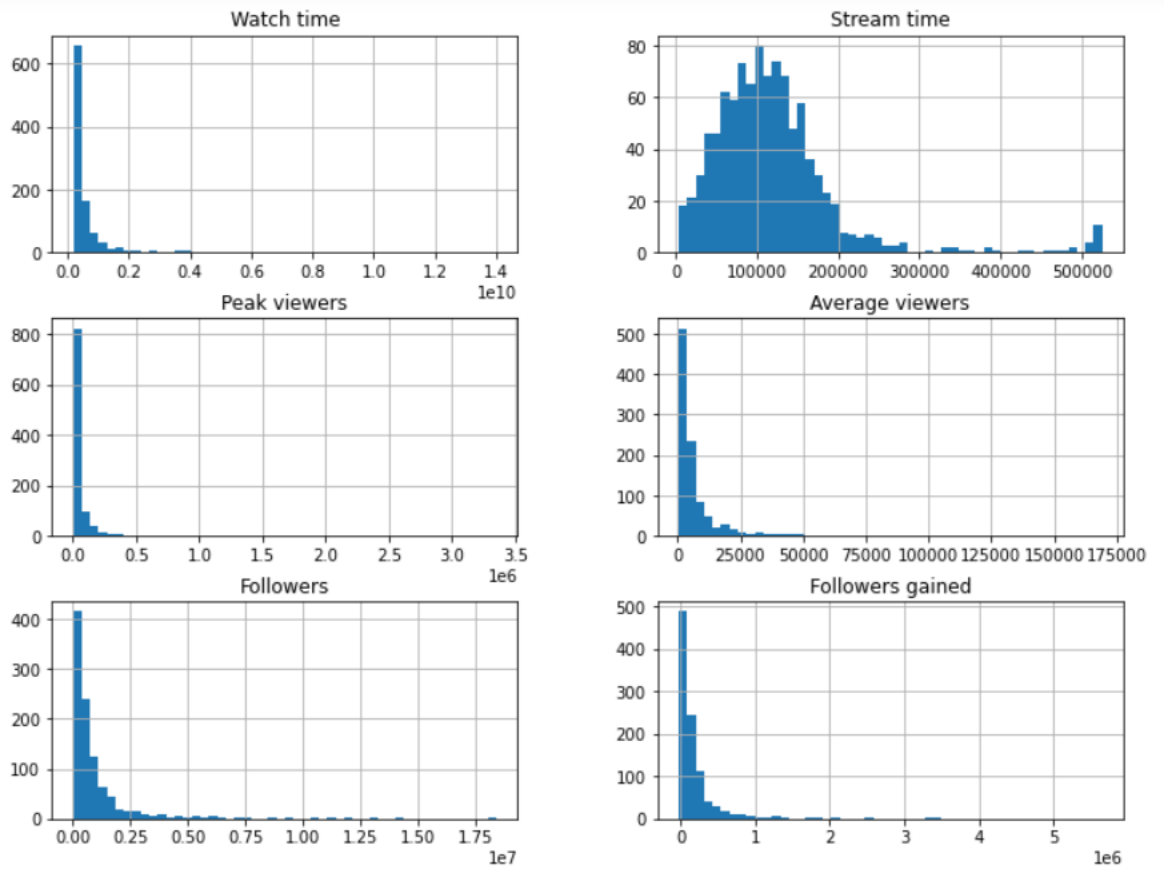
At the end of this project, I wanted to know what helps a channel gain more viewers and if there is a viewer oligopoly among the top creators. I expected to see some sort of indication related to games and watch time. In my visualizations, I focused primarily on watch time and factored in the success of each game or category each streamer tried. This differs from the visualizations above which focused primarily on viewers and follower counts. The first reference visualization focused on the network of viewers throughout the platform. The second reference visualization represents nothing more than a subscriber count. My visualizations provide much more depth into the top streamers and what helps them achieve their watch time.

Data and Methods (Process):

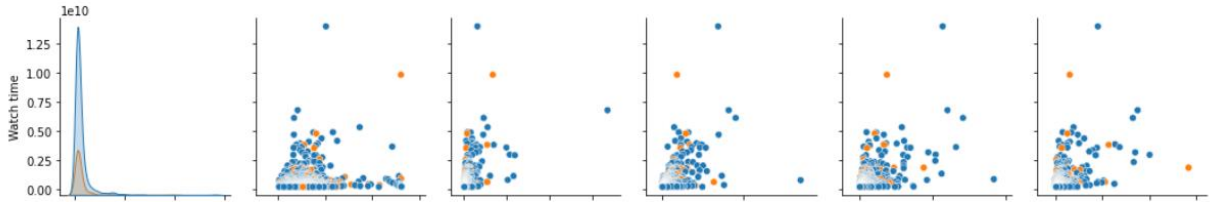
The first data set I have is of the top 1000 twitch streamers based on watch time. This is calculated by multiplying the average viewership of a channel by the number of minutes streamed to get the final result of watch time in minutes. This essentially counts 1 minute for each viewer watching at any given time during a stream. This data set includes information such as: channel name, watch time, stream time, peak viewership, average viewership, total followers, followers gained, partner status, mature channel status, and language of stream. The second data set is the top 200 games based on watch time. This includes data such as: game name, watch time, stream time, peak viewership, peak channels, streamers, average viewers, average channels, and average viewer ration. The data I am using is easily harvested from sullygnome.com/365 by downloading the pre-available csv files and combining them.

In the initial review of the data, it became very apparent that the data was heavily spread out. In these histograms below, there isn't a whole lot of meaningful insight coming from them

due to the fact most of the data falls in the first couple of bars for each histogram. This shows that there is a large gap between where most of the streamers are and the very top streamers.



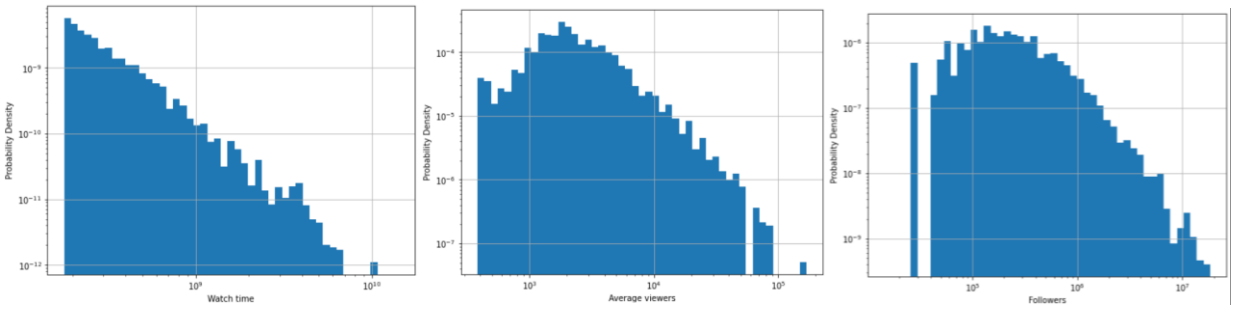
Initially, I was also hoping to find some sort of correlation in the data and go from there. However, as seen in the initial pair plot, in regard to watch time, there really aren't any pair plots that heavily correlate. Most of these have a large blob in the bottom left of the scatterplots, and no clear line of best fit.



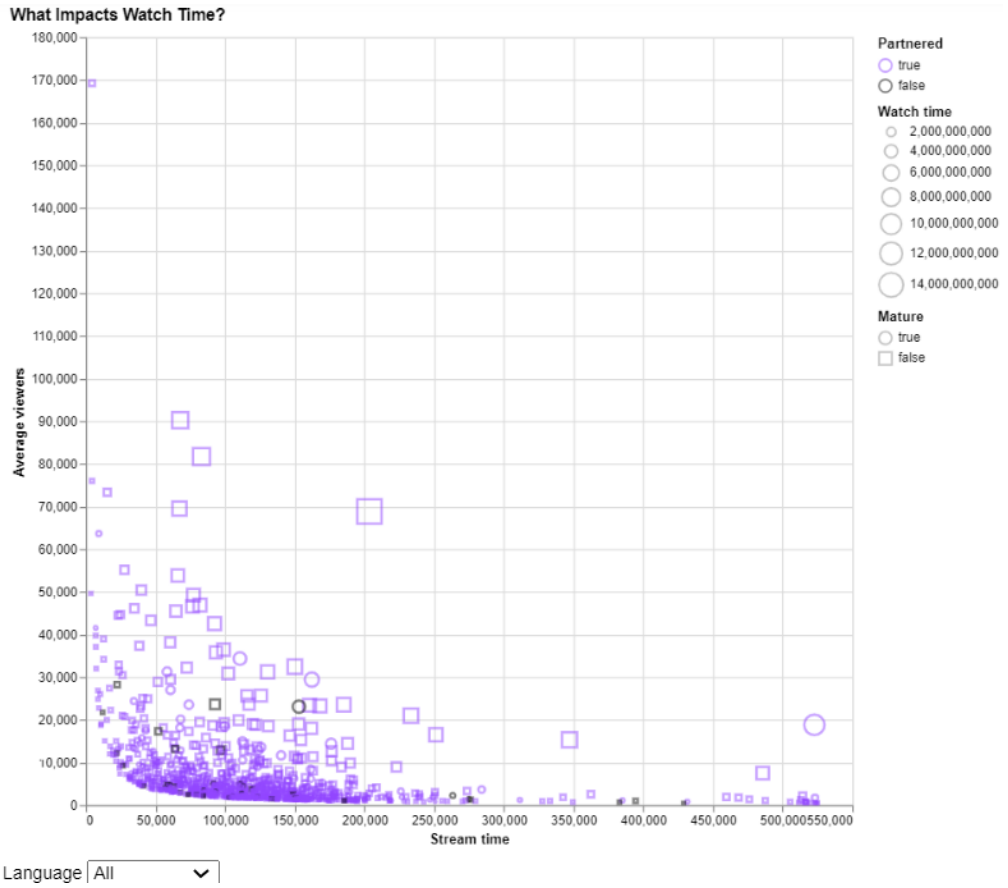
After seeing these initial failed experiments from the start, the main idea that came to me was to create a more detailed version of these scatterplots. I also decided that I wanted to factor in game data, as it became apparent that there was more to the story than the initial dataset provided. I wanted to create an interactive scatter plot that laid out many of the details from the data set and allowed the user to look around and come to some of their own conclusions, as well as look at where some of their favorite streamers might have fallen. The pro to this is that it offers a lot of information in a small visualization, while also not making the area too convoluted. The major con to this however is that there wouldn't necessarily be a clear answer to the initial questions.

Results:

After a few of these initial failed visualizations, I reformatted the histograms in terms of their log scale and probability densities. As seen below, these newer histograms show a little more information due to being in log scale. We can see that the data is very spread out in these categories, where the very top streamers in these categories are very far above the rest.



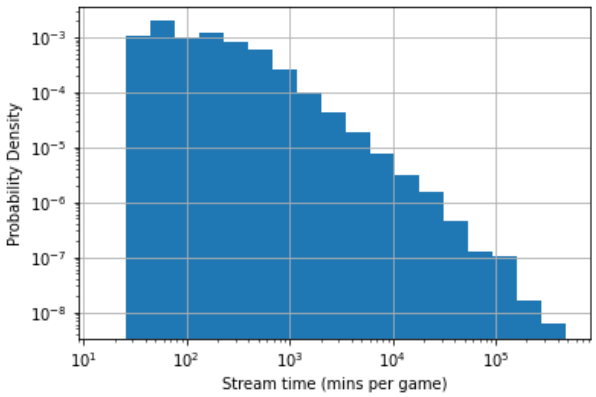
Then, after initially considering the pair plots, I decided a basic, yet effective approach would be to create an interactive scatterplot. This would allow the user to zoom in, look around, and make potential connections on their own through the data. Below is the result of the initial dataset.



While the above image is static, the actual visualization is interactive including the ability to change the language, view more information with the tooltip, and it can be zoomed in on. The Stream time is on the x axis and the average viewers is on the y axis. The watch time can be seen in terms of the size of the point. The watch time can also be seen as the area of the square between the 2 axes and the point. The partner status is indicated with the color, and the mature tag is indicated with the shape. While this visualization doesn't paint the best picture outright, it is still very useful where users can look around, use the tooltip, and try to make some conclusions from the data.

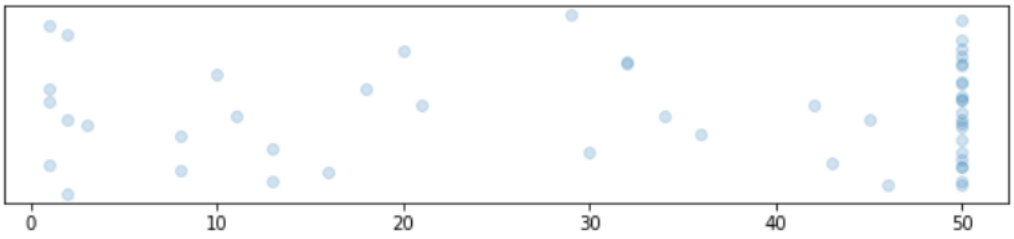
After seeing this data from the main dataset, it became apparent that including game data could be very useful. I decided to take a case study of the top 50 streamers from the initial

dataset and look at the top 50 games or categories in which they streamed. The reason for stopping at the top 50 is because once you get past the 50th ranked game for each streamer, it is only played for a couple hours or less. Including even more games could lead to an added clutter in the visualization. The reason for investigating the games played is that throughout the year, there are often many different trends in the gaming sphere. For instance, some games quickly rise to the top and become super popular for a period of time. Other popular games have frequent major updates, and many times, these streamers want to try out the newest version of some of these games. After taking a look at the initial distribution of the game data, it became apparent that the game data was quite similar to the stream data in terms of watch time. This shows that many times, there are a lot of games that are streamed for a pretty short time; however, sometimes streamers also have a couple of games that might be their specific preference, so they get played a lot.

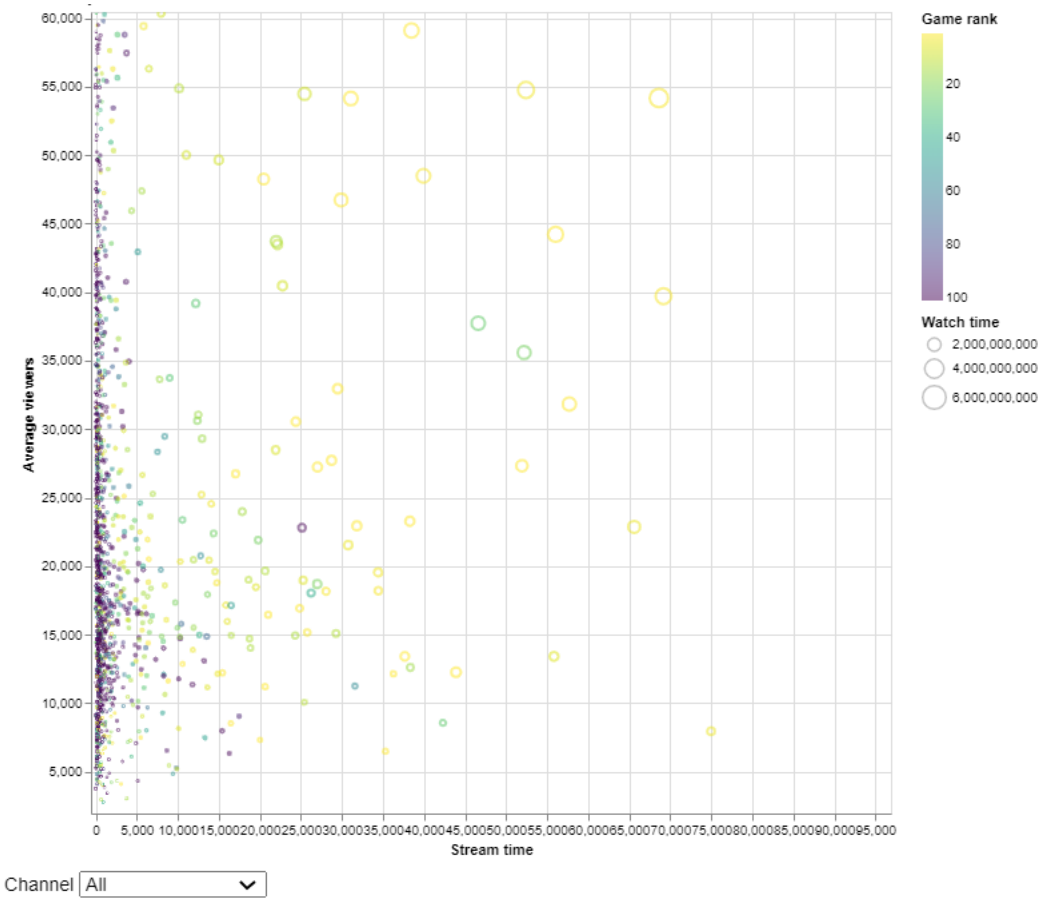


Another interesting thing is in this 1D scatterplot below, which shows the number of games each streamer played over the course of the last year. Most streamers seem to play a lot of games, with most of them playing 30 or more games in the year. One interesting note is that a few don't seem to play many games, however, many of those are game specific channels. For instance, we see in the bottom area a Valorant channel, two league of legends channels, a CS:GO

channel, and a Rocket League channel. These channels aren't individual people, but rather a channel dedicated to the stream of these games, in which they often host large events for their respective games. Since so many streamers are playing a variety of games, it could be interesting to see how often certain games are played in their channels and what seems to be the most successful. This also seems to indicate that maybe the variety of many of these top channels leads to success through the means of fresh content.



After seeing the similarities in terms of data distribution, I decided to do something similar with this game data. As a result, I get another interactive scatterplot with the top 50 streamers and their top 50 games. As similar with the last visualization, the x axis is stream time and the y axis is average viewers. Note that each data point is a separate game with their corresponding streamer, so the x axis is in terms of stream time of that specific game. The watch time is represented as the size of the data mark, as well as the area as explained above. Lastly, the game rank is represented in a color scale, where the number 1 game is in yellow, and the bottom games are in purple. The user of this visualization can also select a streamer from the top 50 in the dropdown menu. Another important note is that the rank of the game is determined from the platform as a whole. The dataset factors in the watch time of the whole platform, not just the top 1000 streamers, and that is how the game ranks are derived.



One interesting thing that immediately can be noticed by zooming in on the bulk of the data is that there does seem to be some correlation between the game ranks and the watch time. Ideally, we would see a trend of worst to best going from the bottom left to the top rate. While it isn't a perfect trend, we do see this somewhat in the data. In the bottom left, almost all of the points are in purple. As we move to the top right, we start to see some blue and green, and then eventually towards the top right, we mostly see yellow and light green data points.

While these visualizations are great at looking at a broader picture, it is important to also talk about the limitations of these visualizations. First and foremost, the data from this project is simply a snapshot of the past year. A future project might consider the possibility of creating a way for the user to look at the data over the years. While this data could be harvested, managing

this much data over this much time would have been far too complex for the scope of this project. Similarly, I also think it might be interesting to look at streamers outside of the top 100 streamers. Twitch has far more than 1000 partnered streamers, and many are not nearly as successful as the top 1000. As we saw in these visualizations, there is quite a disparity within even the top 1000, and I imagine the disparity somewhat continues as we look at the next 1000 and so on.

In summary, at the end of this project, there does not appear to be any single variable that necessarily is an indicator for success. One of the more promising variables is the popularity of the game (game rank) in which the streamer is playing. While this does show some promise, it is far from being the only thing that contributes to success. As we saw with the histograms, the climb to the very top seems to be exponential rather than linear. The top 1% of even this data set seems to be miles ahead of the rest of the pack.

References

Bellan, R. (2020, March 23). *Is 2020 the year of Twitch?* Forbes. Retrieved October 14, 2022, from https://www.forbes.com/sites/rebeccabellan/2020/03/23/is-2020-the-year-of-twitch/?sh=593811475564

Gershenfeld, K. (2020, December 28). *Insights from Visualizing Public Data on twitch*. Medium. Retrieved October 14, 2022, from https://towardsdatascience.com/insights-from-visualizing-public-data-on-twitch-a73304a1b3eb

Iqbal, M. (2022, September 6). *YouTube revenue and Usage Statistics (2022)*. Business of Apps. Retrieved October 14, 2022, from https://www.businessofapps.com/data/youtube-statistics/

MaxData. (2021). *Top 10 Most Subscribed Youtube Channels | Subscriber Count History (2006-2021)*. *MaxData*. Retrieved October 14, 2022, from https://www.youtube.com/watch?v=1c-qrKCAK7c&t=544s.

Sheng, J. T., & Kairam, S. R. (2020). From virtual strangers to irl friends. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–34. https://doi.org/10.1145/3415165

Twitch statistics and analytics. Hi I'm SullyGnome! (n.d.). Retrieved October 14, 2022, from https://sullygnome.com/365