Factors That Affect Top Twitch Streamers Success

Cameron Reggio

Bellevue College

6/12/2021

**Introduction:**

I have always been fascinated by unique non-traditional careers and emerging creative markets that are becoming the forefront especially with the younger generation and social media. It is interesting to me to see the success that can be achieved in these markets and the variables and factors that play a role in that success. The topic I will be analyzing is based on a website known as "Twitch.tv" and this website involves normal everyday people broadcasting live video streams of any category you could think of. While mainly focused on the gaming industry a few examples of other streams you would find could involve cooking, politics, walking around Tokyo, dance competitions and many more. It gives an individual the ability to express themselves in almost any way they want. The specific area of Twitch I have studied is the top streamers and how average viewer count along with other significant variables have an impact on their success and how much total “watch time” they get compared to other streamers that may have lower success. I believe studying the data and doing an analysis may result in some key insights to what makes certain streamers the top of their categories.

**Data Sources:**

I will be pulling and using data from the website "Kaggle" as I believe it is a great source that has a lot of useful and interesting datasets for a wide variety of topics. I managed to find a dataset about the top streamers on Twitch and different variables associated with them such as “Watch time”, “Stream time” and “Followers”. There are 11 variables total (including streamer’s names/channels) and this was exactly the kind of information I was looking for and wanted to analyze. The information in this dataset is extremely relevant to my topic since the variables used are crucial to a streamer’s success. It is especially interesting because the variables within the dataset encompass a streamer’s success in different ways. It is not necessarily just about the highest number of viewers (“peak viewers”) the growth can be significant in other ways such as “followers gained”, “average viewers” and whether the streamers are partnered with the company. Unfortunately this data source does not include a date column but it was estimated the data was collected over the course of one year (2019).

**Models and Methodologies:**

Regarding models and methodologies, I will check the variables for high collinearity in relation to each other. I will create dummy variables for the two variables of “partnered” and “mature” therefore the qualitative nature may become measurable in the analysis if I choose to integrate them. I may also plan on using an interaction between variables of interest if I find a high correlation to study the impact on the dependent variable. I also plan on checking for heteroscedasticity when I run the OLS model and if I find heteroscedasticity I will attempt to resolve it by implementing WLS to the model. Lastly I will have at least four indicator variables but first I believe it is extremely important to start off by doing data exploration to find any possible nulls, outliers or interesting aspects within the dataset.

Since there are no null values as shown in the appendix I won’t need to use “is.na()” to replace any N/A values with zeros. I then noticed that originally the data for “watchtimehours” and “streamtimehours” was in minutes resulting in extremely high values making it significantly difficult to read and interpret. After downloading the dataset I went into the “csv” file and converted the minutes to hours making it a little easier to understand the magnitude of the data. Afterwards I observed the min and max for each of the numeric variables as shown in the appendix.

Observing this output we can see a minimum “streamtimehours” of 57.75 and minimum “followersgained” of -15,772 which indicates they actually lost followers throughout the year. This could be due to a lack of stream time hours and providing new content for their viewers or a different reason entirely. We can also see some very substantial numbers as seen under the maximum “watchtimehours” of 103,269,362.5 which is equivalent to approx. 4,302,890 days. There is also a maximum stream time hours of 8690.75 which is equivalent to approx. 362 days. I realize the magnitude of this output is extreme and a question may arise such as “how is this possible?”. As I mentioned before there are a wide variety of streams found on Twitch.tv and it is not uncommon to have multiple streamers rotate (under the same stream) to provide 24-hour long streams. There are also some celebrities that do occasional sponsored content with streamers which inflates viewer count and watch time. I have checked the numbers of these top streamers to verify them (some of them I occasionally watch too).

Next, I think it would be interesting to see different break downs of variables within the dataset such as the “mature” variable. This variable indicates a stream may not be as family friendly and is intended for a more mature audience (due to language or other factors). A table and a bar plot can be seen in the appendix below but it was found that 77% of the top streamers during this year did not have their stream intended for a more mature audience while 23% did. Currently the dataset is implemented to show whether a stream is intended for a mature audience by resulting in true/false. I will use dummy variables to interchange this data so that it instead shows zero and one (one indicating mature audience and zero indicating family friendly). I will also do the same for the “partnered” variable and the code for this can be seen in the appendix at the bottom. Lastly I would like to just further explore the data by plotting and observing different potential insights about the variables. The histogram in the appendix shows that the majority of content creators streamed around 1,000 to 3,000 hours over the course of 2019 while a few (such as the 24 hour streams I mentioned previously) streamed over 8,000 hours.

**Data Analysis and Interpretation:**

For the data analysis I first want to check the correlations between the variables within the dataset before constructing my model. The correlation between “watchtimehours” and “streamtimehours” as shown in the appendix makes sense because the amount of time a content creator streams does not necessarily guarantee more viewer watch time. This is especially true as Twitch continues to become more popular and therefore more saturated with streamers. Twitch.tv was created in 2011 and has skyrocketed ever since so it was already extremely popular by the time of the creation of this dataset. A plot of this correlation is shown at the bottom along with correlations for the other variables in the dataset. The highest result of 0.62 was between “watchtimehours” and “followers” which is to be expected because the number of followers a streamer has would likely lead to more watch time.

Moving onto the OLS model I would like to explore the impact of specific variables on the dependent variable “watchtimehours”. The independent variables I will be using for the model are “streamtimehours”, “peakviewers”, “averageviewers”, “followers” and “viewsgained” as they showed the most interest during data exploration. I tried to include the categorical variables into my model however after testing it the categorical variables actually slightly lowered my r-squared value and both had p-values over 0.05 which provides no statistical significance. The result might be due to the type of dataset since this dataset includes all the top streamers on Twitch which nearly all are already partnered so the variables provide little unique impact to the dependent variable “watchtime”. The code for my main OLS model can be seen in the appendix and the output gave a relatively significant R-squared value of 0.6283. The p-values for each of the variables are also all under 0.05 indicating statistical significance.

After constructing the model it is important to check it for heteroskedasticity. The BP test shown in the appendix has a low p-value indicating heteroskedasticity therefore I will have to try and resolve it by using WLS to readjust the variances in order to create a homoscedastic model. The steps and process shown at the bottom gave me an updated model and increased my R-squared to a value of 0.7675. All the p-values for my variables are still below 0.05 indicating statistical significance except for the “log(viewsgained)” variable which had a value of 0.5538. I decided to remove this variable since it was insignificant in contributing towards my model and removing it resulted in no change in the r-squared value. All my estimates are positive and have a relatively significant impact on the amount of watch time these top streamers get. I found this analysis and data exploration to be very interesting.

**Conclusion:**

While the data analysis did show the impact of these key variables in relation to a streamers success (“watchtime”) I feel that there are other variables outside of this dataset that could lead to further insights and even more accurate results. An example would be implementing specific categories or types of content a streamer decides to stream on a regular basis. The most popular categories on Twitch rotate frequently as new games are released or people find new unique hobbies to share so the variability in watch time and their viewer base could be substantial. It could be slightly difficult to integrate since more dummy variables may need to be used and there are numerous different categories on Twitch. I still however believe being able to measure other variables and different categories in relation to a streamer’s statistics could provide useful insights on their personal success and in comparison to others.

Appendix



Text

Description automatically generated with low confidence

Text

Description automatically generated with medium confidence



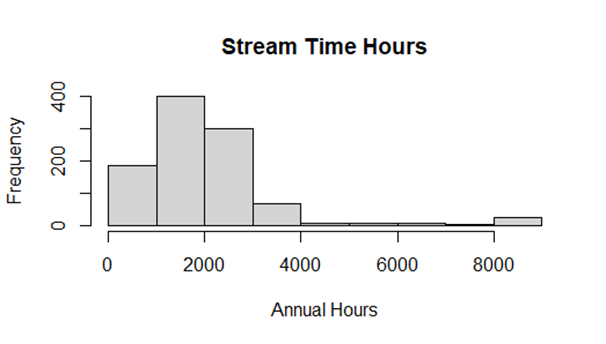
Chart, bar chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated



****





Chart, scatter chart

Description automatically generated

Graphical user interface, application

Description automatically generated

A picture containing logo

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence

A picture containing text, receipt, screenshot

Description automatically generated

**Dataset Source:**

Mishra, A. (2020, August 24). *Top Streamers on Twitch*. Kaggle. https://www.kaggle.com/aayushmishra1512/twitchdata.

**Website collected from:**

*All Categories*. Twitch. (n.d.). https://www.twitch.tv/directory.

(Displays each category found within Twitch.tv and defaults by most popular streams when viewing each section)