Twitch Streamer Data Analysis

Rauvina U. Singh

Mitchell Community College

Agriculture and Science Early College

April 30, 2024

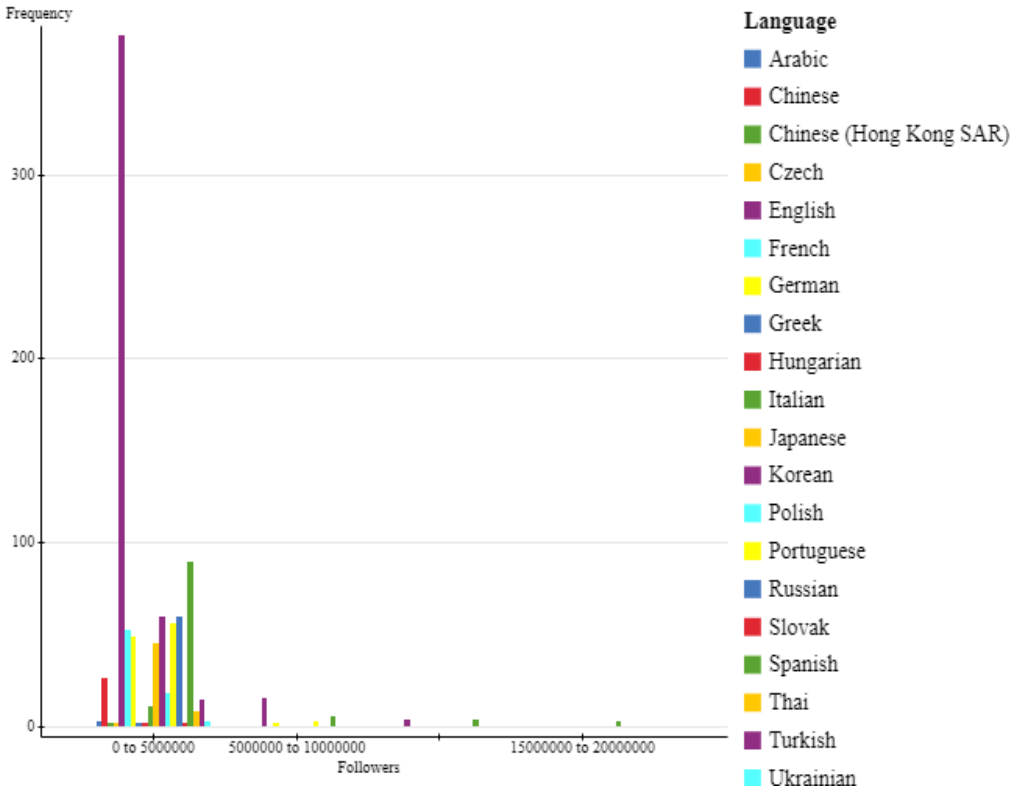
Introduction

Twitch is a streaming platform where people watch different types of content live and interact with the creators. The given data set reveals many factors about the creators and followers on this platform. There is so much we can find concerning the channels, the followers, the languages, the amount of time spent, and the correlation between those variables. There is almost too much information that can be discussed for each category or various pairs of data but I will be discussing some of the more interesting ones in my opinion. There is a lot to assume about the numbers before you see them due to what you may have heard about it, this does not really cloud your judgment but it does make the results of the data much more surprising. The data I will be discussing are the 4 most interesting or surprising ones that stood out to me.

Followers and Language

Considering the fact that the highest proportion of languages was English at 43.8%, I wanted to know if language had anything to do with the number of followers a channel had. By using the Graphing feature on StatCrunch, I was able to find out that surprisingly it did not. Although 394 of the 900 hundred channels that stream are in English, the highest number of followers come from only two streamers that are speaking Spanish. Spanish does come in second in terms of proportion of languages streamed on the platform but due to the huge gap between first and second, I assumed that the highest number of followers would belong to the language that dominated the platform. By testing this out in StatCrunch, I saw that it is a matter of the viewers preference and it made me wonder if it is because there are more viewers using Twitch that speak Spanish. The amount of English speaking channels also makes you wonder if the Spanish speaking channels are streaming something much more entertaining than them to get so much more followers.

Language	**Relative Frequency**	**Language**	**Relative Frequency**
Arabic	0.002	Japanese	0.05
Chinese	0.029	Korean	0.066
Chinese (Hong Kong SAR)	0.001	Polish	0.02
Czech	0.001	Portuguese	0.064
English	0.438	Russian	0.066
French	0.058	Slovak	0.001
German	0.054	Spanish	0.11
Greek	0.001	Thai	0.009
Hungarian	0.001	Turkish	0.016
Italian	0.011	Ukrainian	0.002



Mature Content on Twitch

One of the bigger stereotypes around Twitch streaming is that it has a lot of mature content. Twitch streaming is normally associated with violent game play or swearing due to the fact that it is mostly used by teenagers and young adults. The content created on Twitch is meant to entertain certain audiences and a more mature age group which is why I think that more than 35% of these channels create mature content. To test my claim, I set up a null hypothesis and an alternative hypothesis about the proportion of channels that create mature content.

Null Hypothesis:

$H_0: \rho = 0.35$

Alternative Hypothesis:

$H_A: \rho > 0.35$

Options

One sample proportion summary hypothesis test:

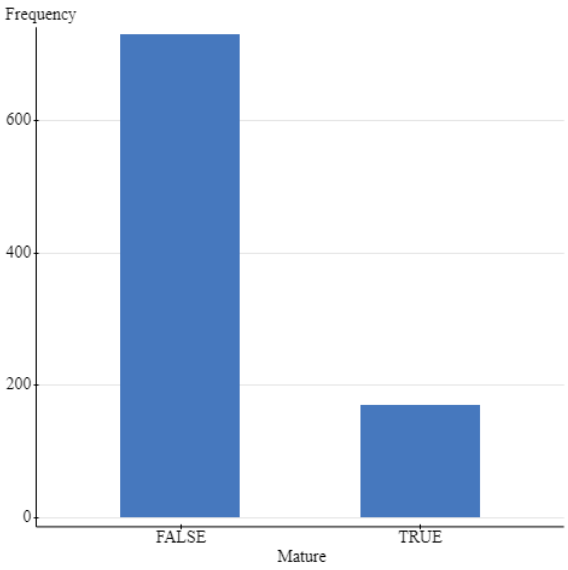
p : Proportion of successes

$H_0 : p = 0.35$

$H_A : p > 0.35$

Hypothesis test results:

Proportion	Count	Total	Sample Prop.	Std. Err.	Z-Stat	P-value
p	170	900	0.18888889	0.015898987	-10.13342	1



Using StatCrunch, I went to Proportion Stats and plugged in the variables required to find out that the p-value is 1. Because the p-value is higher than any significance level you are normally given, we fail to reject the null hypothesis and there was not enough evidence to support the claim that more than 35% of channels on Twitch stream mature content. I am not a user of Twitch but from the clips I have seen and what people normally say about it, I assumed that a large amount of channels would be making mature content. I would have claimed even more than 35% but I considered the fact that people have certain values or even religious ones that prevent that. Looking at the sample proportion being only 18.9% was very shocking considering that I thought Twitch was more mature overall.

Followers and Followers Gained

The followers gained column had very large numbers regardless of whether they were negative or positive. I was able to find the proportion of followers that were made up of the followers gained in 2023. I assumed that the proportion of followers gained that year of followers overall would have a mean greater than 0.15 because although the new followers were

a big impact, the numbers of followers were already very large. To test this claim, I created a null hypothesis and an alternative hypothesis.

Null Hypothesis:

$H_0: \mu = 0.15$

Alternative Hypothesis:

$H_A: \mu > 0.15$

Options

One sample T hypothesis test:

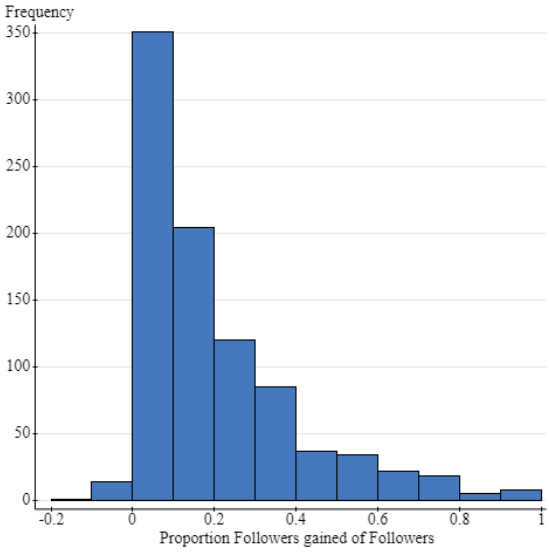
μ : Mean of variable

$H_0 : \mu = 0.15$

$H_A : \mu > 0.15$

Hypothesis test results:

Variable	Sample Mean	Std. Err.	DF	T-Stat	P-value
Proportion Followers gained of Followers	0.20131396	0.006568818	899	7.8117493	<0.0001



Using StatCrunch, I went to T Stats and plugged in the information needed to get results. The results show that the p-value is less than 0.0001 which is less than any normally given significance level so we can reject the null hypothesis. There is enough evidence to support the claim that the mean of the followers gained that make up followers total is greater than 0.15. This relationship between the followers and the followers gained in 2023 is confusing in my opinion. Finding the proportion of the followers gained that year that make up followers total sounds very confusing but it shows how much followers the creators' content from that year make up what they have. It can be useful in the sense that we could see what followers enjoy viewing more of or what they do not by comparing those proportions from previous years as

well. But testing a claim about the mean of a set of proportions does not sound right although it does make sense when trying to find the average growth of followers all the creators in the sample had that year. We can see that the creators managed to put in an average of 20% of the followers they have just over 2023.

Stream Time and Average Viewers

Most people would assume that the more time that a creator spends streaming, the more average viewers they would have from building up their audience. I thought that is what would happen naturally but I decided to look at the correlation between the stream time and average viewers a channel has.

Options (1 of 2)

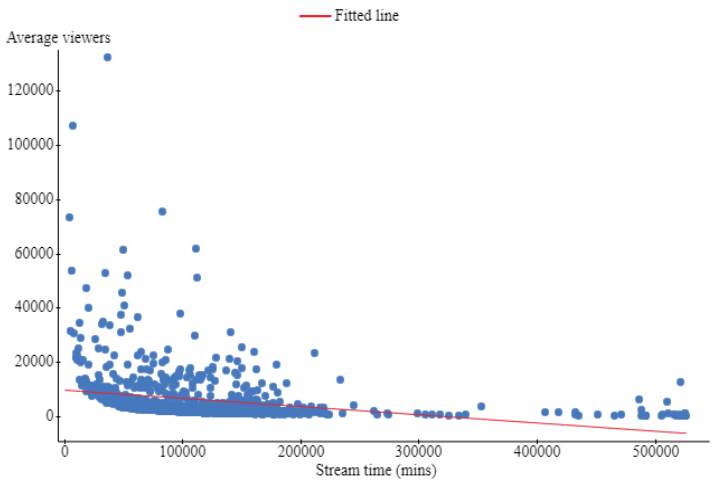
Simple linear regression results:
Dependent Variable: Average viewers
Independent Variable: Stream time (mins)
Average viewers = 10002.647 - 0.030273634 Stream time (mins)
Sample size: 900
R (correlation coefficient) = -0.27503097
R-sq = 0.075642032
Estimate of error standard deviation: 9592.707

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	10002.647	538.35281	≠ 0	898	18.580096	<0.0001
Slope	-0.030273634	0.0035315462	≠ 0	898	-8.5723455	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	6.7621015e9	6.7621015e9	73.485107	<0.0001
Error	898	8.2633984e10	92020027		
Total	899	8.9396086e10			



Surprisingly, the correlation between the two columns of data was a low negative correlation. This means that they barely affect each other and they do so negatively; the correlation coefficient I received from StatCrunch was -0.275. I found this number by going to Regression Stats and plugging in all the information they required. The p-value is less than 0.0001 which is less than any typical given significance level so we can reject the null hypothesis. The stream time and average viewers did not have the positive correlation that I had assumed. This

correlation shows that the time spent streaming does not mean it is entertaining enough for the viewers to watch it often.

Conclusion

There are many more aspects of this data that can be discussed but there is too much to choose from. I would have liked to see more data about the followers and channels so that we could make more connections. If we knew the type of content we would be able to test for a correlation between it and the number of followers. If we knew the language the followers spoke, we would be able to see if the number of followers or views are due to abundance or lack of channels that stream in their language. Unfortunately, that much data would not only be hard to collect but it would be a lot more information to look at than just 900 samples and would have to be in a separate set of data. I know that there are many more things that would make sense to discuss about the data but these relationships caught my attention since the assumptions I made were often completely wrong.