Exploratory Data Analysis on Twitch

Twitch Logo (2019-now)

Twitch is an American (also world-leading) live streaming platform that contains a variety of channels. This analysis examines twitch's data from different dimensions and provides some insights into this platform and hopefully the streaming industry.

Several data processing procedures indicate that the data frame source explores 1000 channels on twitch and evaluates their corresponding watch time (minutes), stream time(minutes), peak viewers, average viewers, followers, followers gained, views gained, partnered, mature, and language. And there is no missing data for the entire 10000(r)*11(c) data frame.

The insights from the data exploration are going to be focused on the following key questions:

1. How well do twitch channels perform? Which are the top ones?
2. Are there any correlations between each performance measurement standard for channels on twitch?
3. In what ways, are language, partnership, and matureness affecting channels' performance on twitch?
4. For newcomers on twitch, how can they estimate and predict their future growth on this platform?

Watch time and average viewers are known to be the dominant factors to evaluate streamers' performance. The four subsequent graphs display the performance on the two bases.

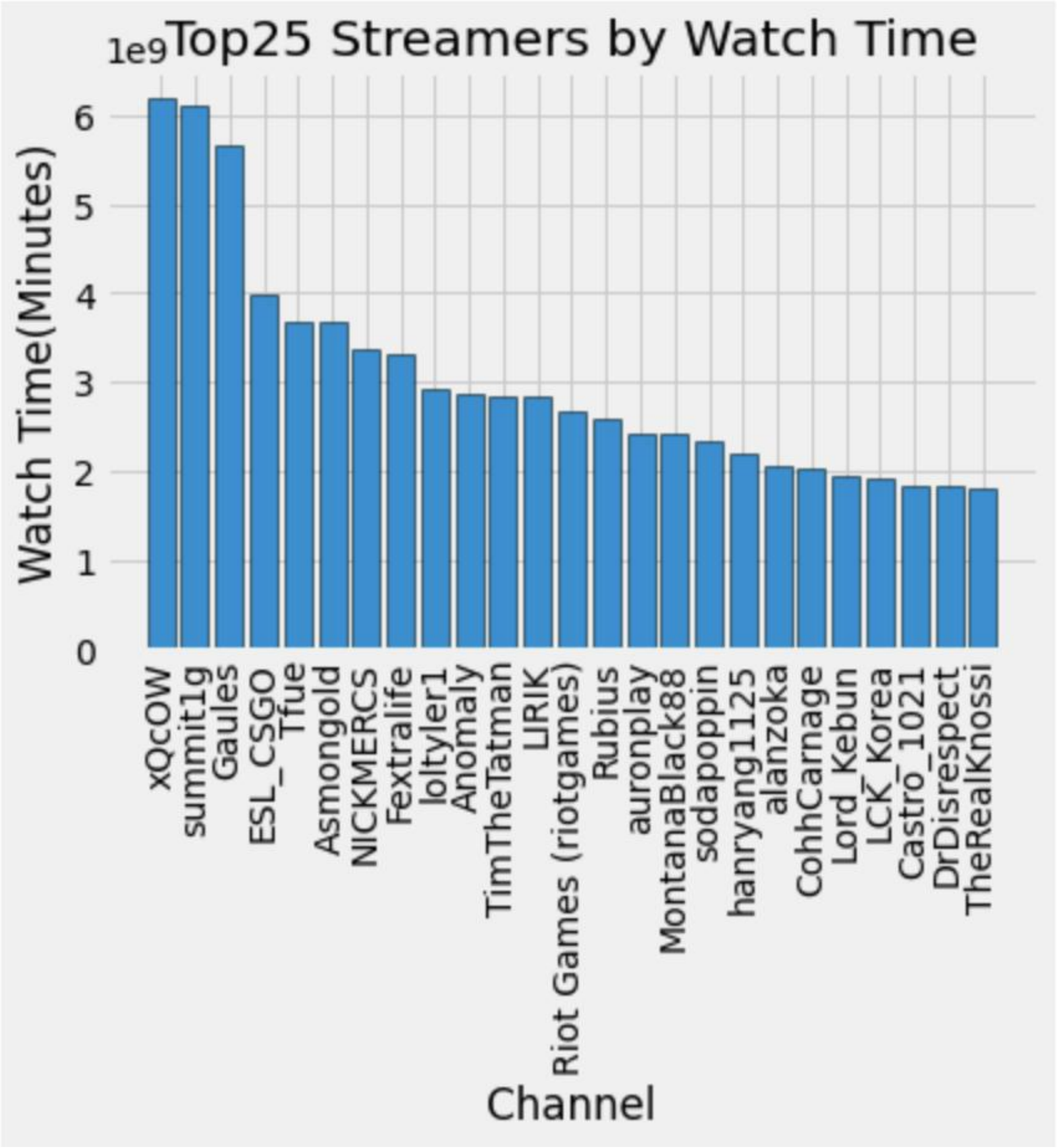


Figure 1 The bar plot shows the top 25 streamers ranked by watch time, with the highest on the very left side. Note that the y-axis is in 10^9. We can observe that xQcOW, summiting, and Gaules stand out by a lot in this scenario.

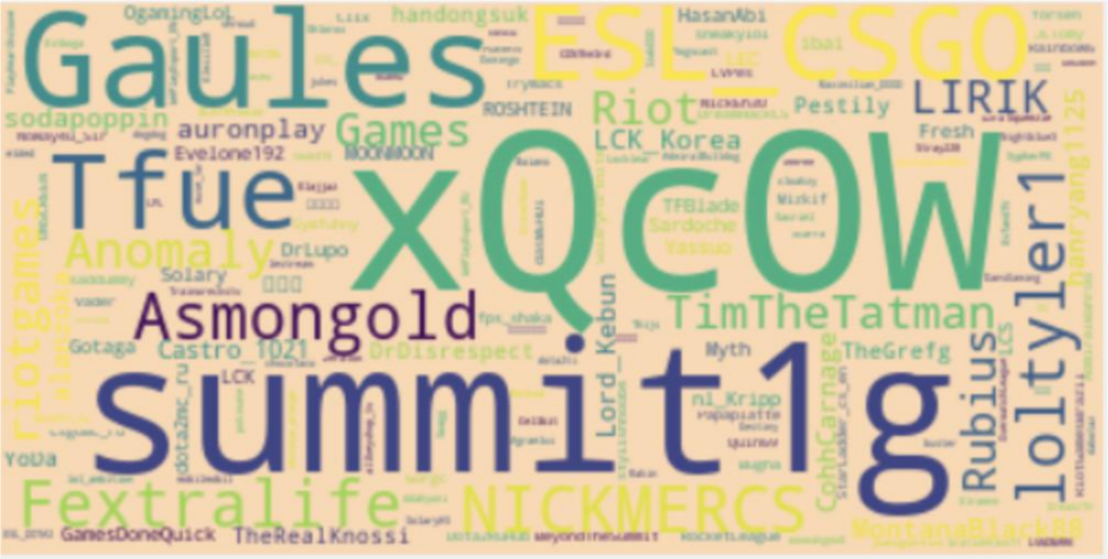


Figure 2 The word cloud displays the channel names ranked by watch time, with the highest having the largest font size. This is a more direct, non-quantitative delivery of Figure 1.

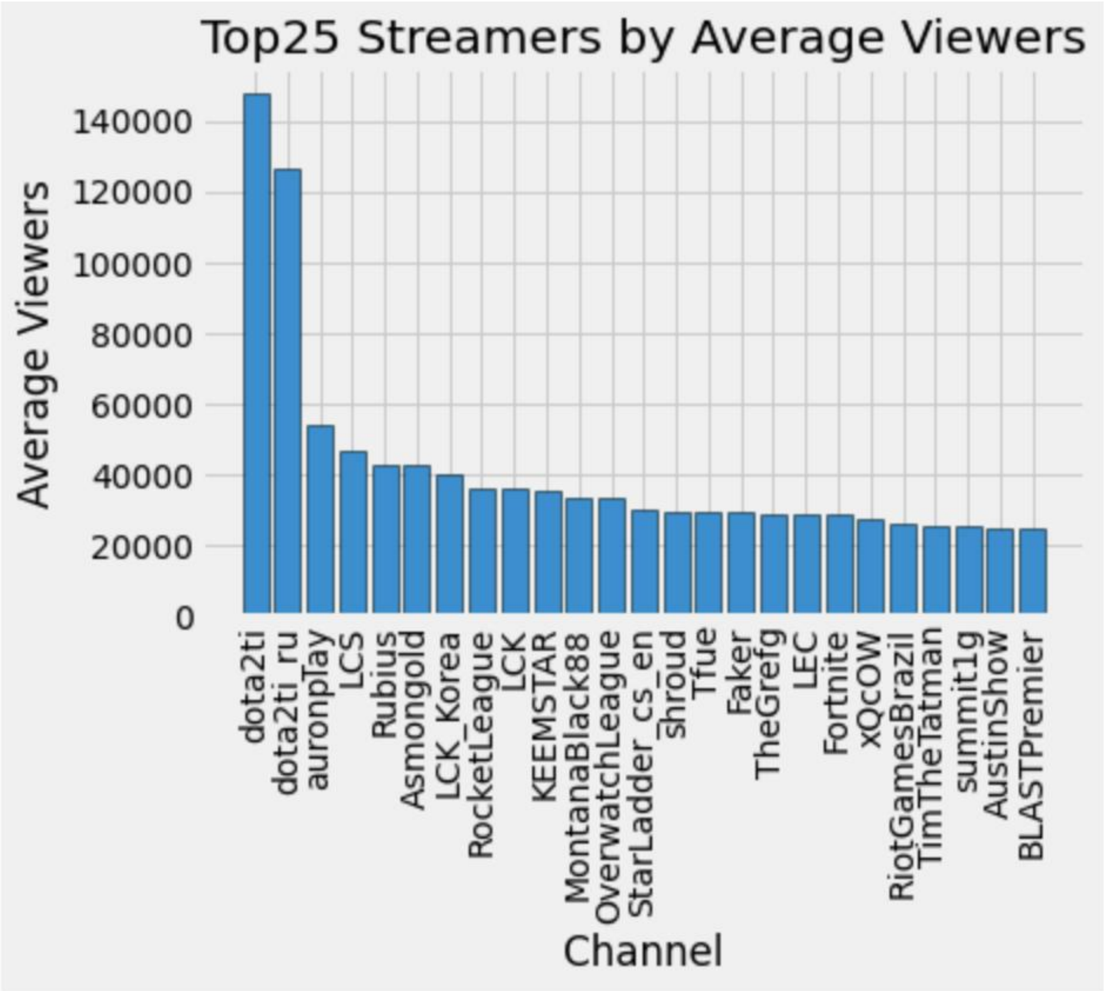


Figure 3 The bar plot shows the top 25 streamers ranked by average viewers, with the highest on the very left side. We can observe that dota2ti and dota2ti_ru stand out by a lot in this scenario. It can be inferred that this video game and related streamings are very popular on twitch.

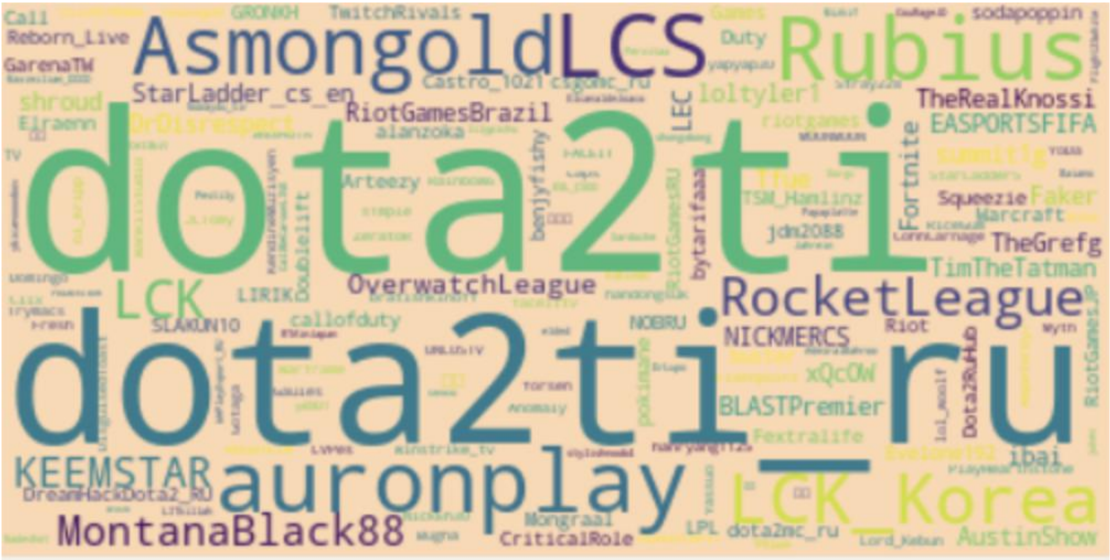


Figure 4 The word cloud displays the channel names ranking by average viewers, with the highest having the biggest font size. This is a more direct, non-quantitative delivery of Figure 3.

Besides the specific insights from each graph, mentioned in corresponding captions, it is also clear that the two performance measurement method (watch time and average viewers) actually give credits to very different streamers. So, after examining the top channels, we are going to see how each dimension is related, and first, let’s take a look at the general correlation map.

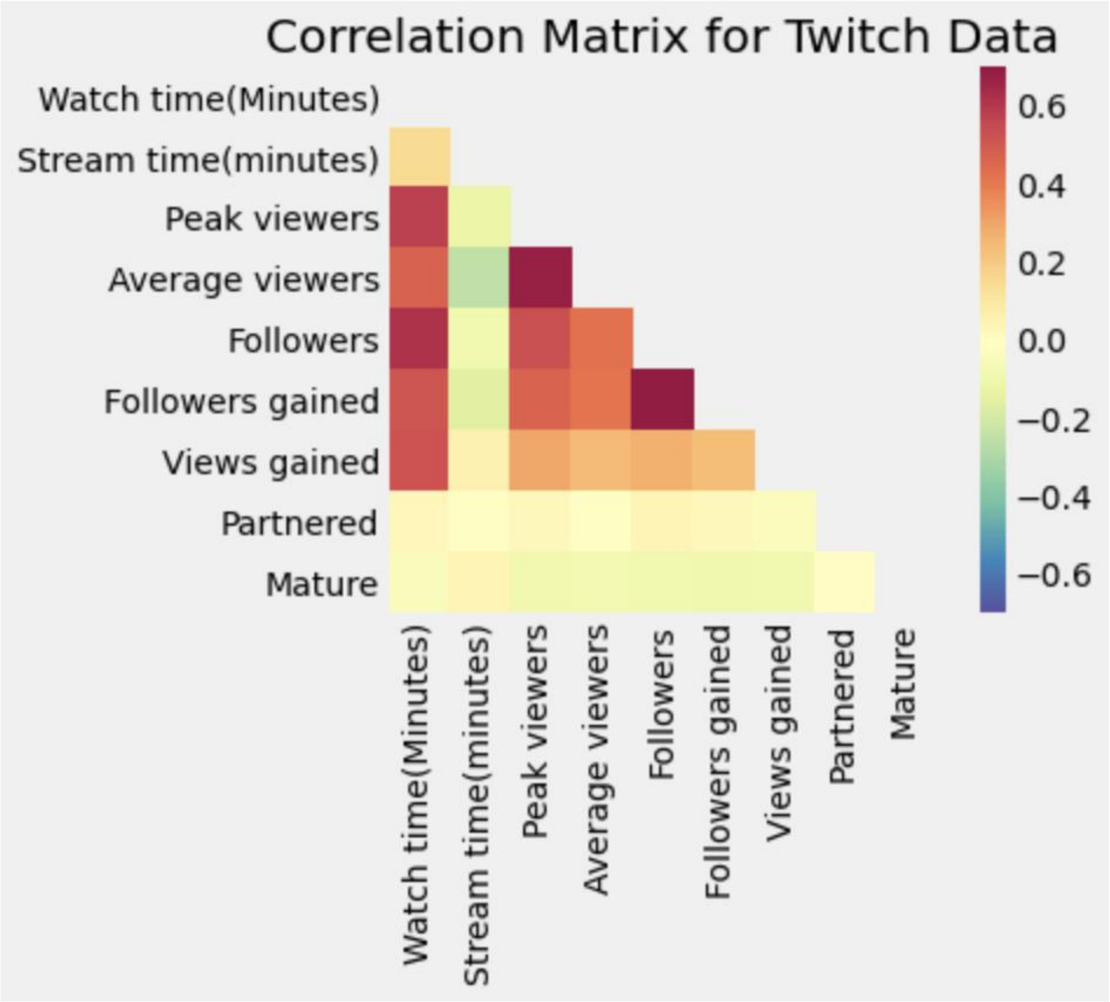


Figure 5 The heatmap shows the correlation matrix for all quantitative twitch data, with the color bar legend on the right side. The warmer colors represent stronger positive correlations, while colder ones represent stronger negative correlations. We can observe that there is no medium/strong negative correlation.

In addition to the apparent observation, mentioned in the above caption, the correlation matrix suggests that the most correlated pairs are average viewers vs. peak viewers, followers vs. followers gained, peak viewers vs. watch time, and followers vs. watch time. So, the below graph offers a closer look at the first two pairs and how each is related to watch time.

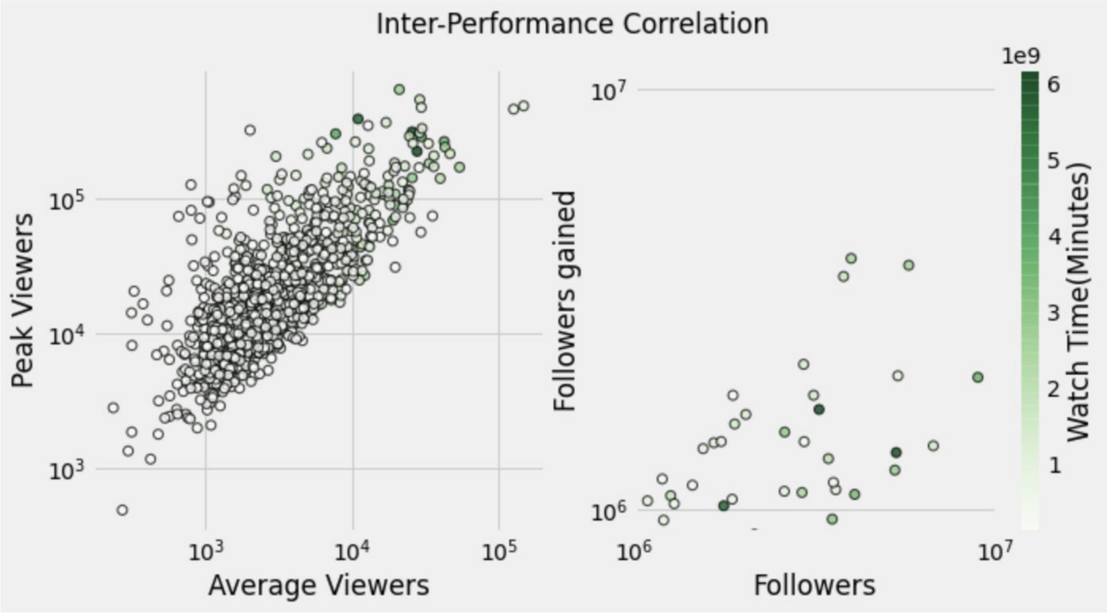


Figure 6 The scatter plots in the subplot show the correlation between different performance measurements. Note that the axis scales are in logs to reduce outliers and that the color-bar on the right represents the value of watch time (the darker the color green is, the longer the watch time it represents). Average viewers vs. peak viewers and followers vs. followers gained display strong positive correlations, and watch time has medium to strong positive correlations with the four.

Besides quantitative variables, the data frame contains categorical ones, including language, partnership, and matureness.

As shown in figure 5, partnership and matureness have very slight correlations with any of the performance measurements. In this case, we are going to find and examine a new one. Subsequent analyses introduce the concept: effective ratio, the quotient of dividing watch time by stream time. It evaluates the effectiveness of streaming, and a higher value means more effective streaming.

In each category's side-by-side bar plot, 'average viewers' are also included next to 'effectiveness ratios.' (Though it is proved to have little correlation with any measurements, it is still an important performance indicator that worth consideration.) We now go for the language category first.

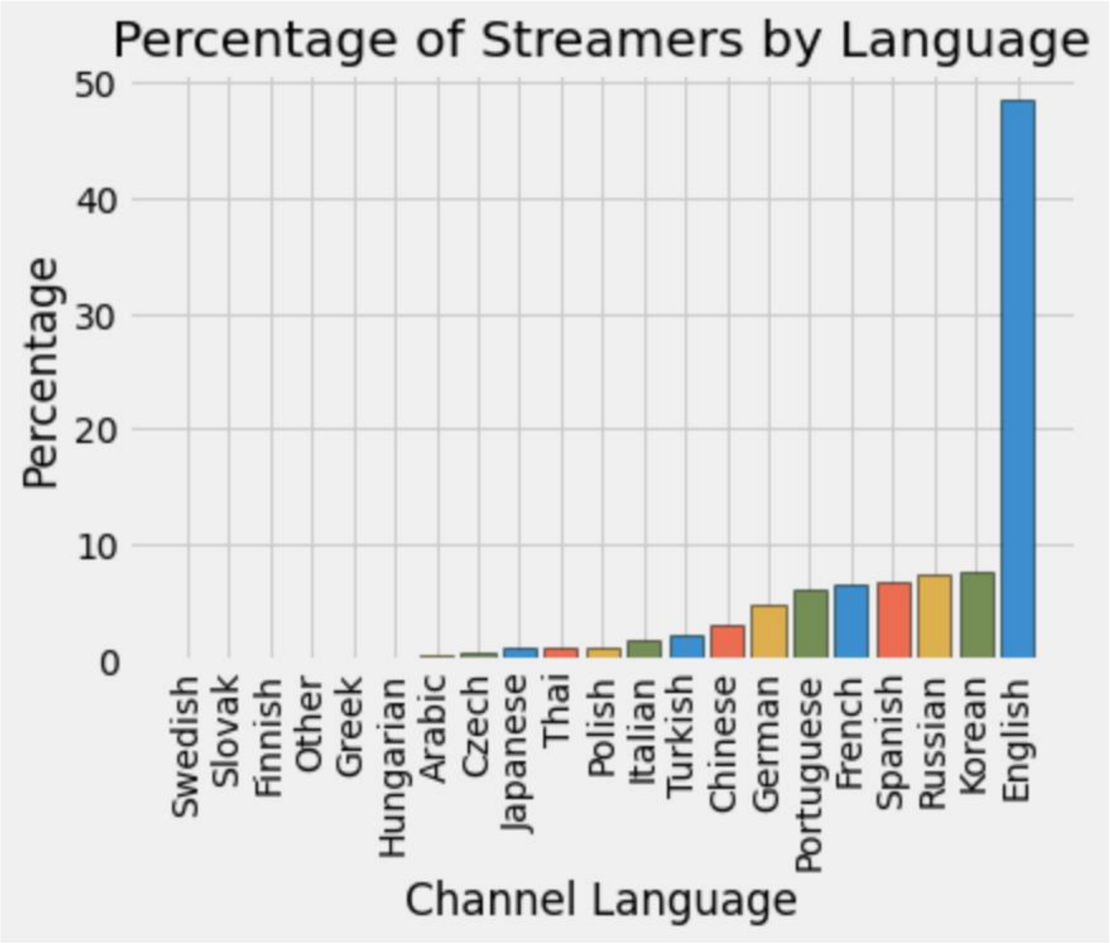


Figure 7 The bar plot shows the percentage of language share of channels on twitch. We can observe that English channels occupy nearly 50% of total channels, while the rest of the languages each has no more than 10% of language share and some have extremely small percentages.

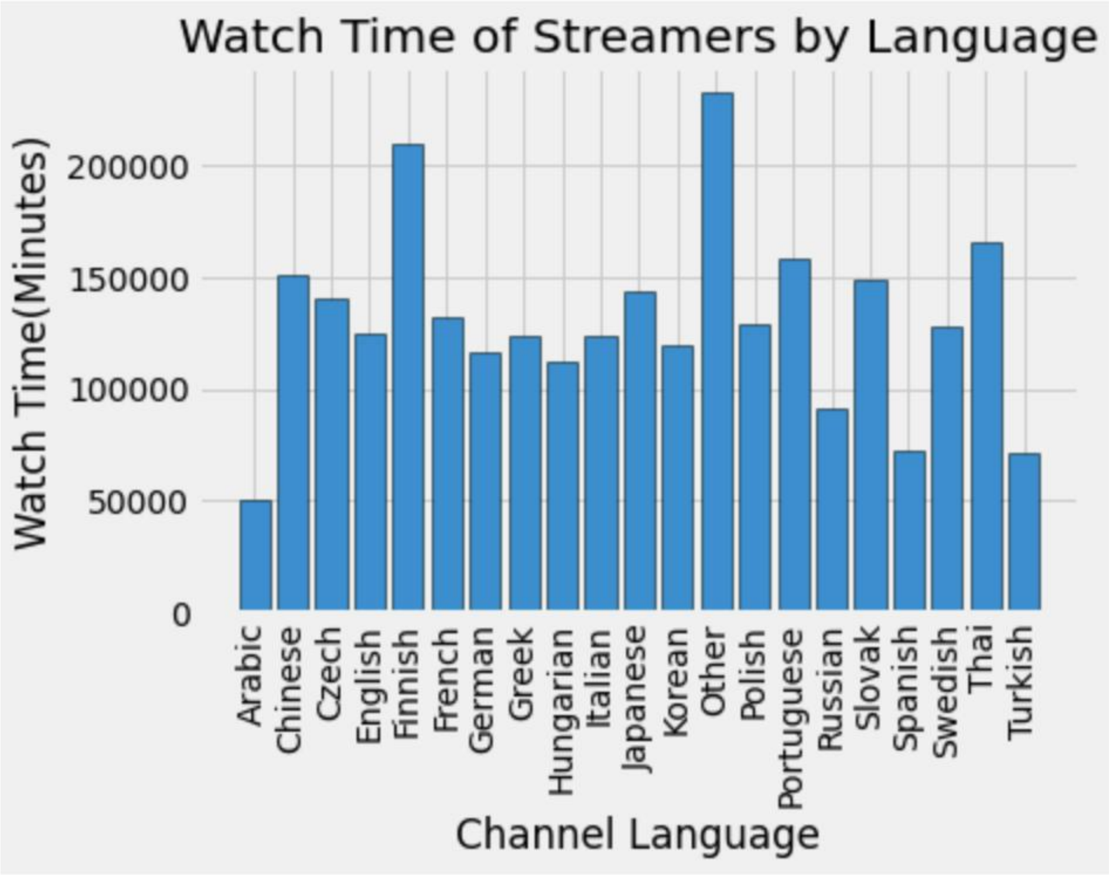


Figure 8 The bar plot shows the watch time (one of the dominant factors of performance) by language. We can observe that some with low streaming language share (Slovak, Finish, and Other) actually have significant amounts of watch time, while English with the most share, does not stand out in watch time. So, twitch or channel streamers might consider re-distribution.

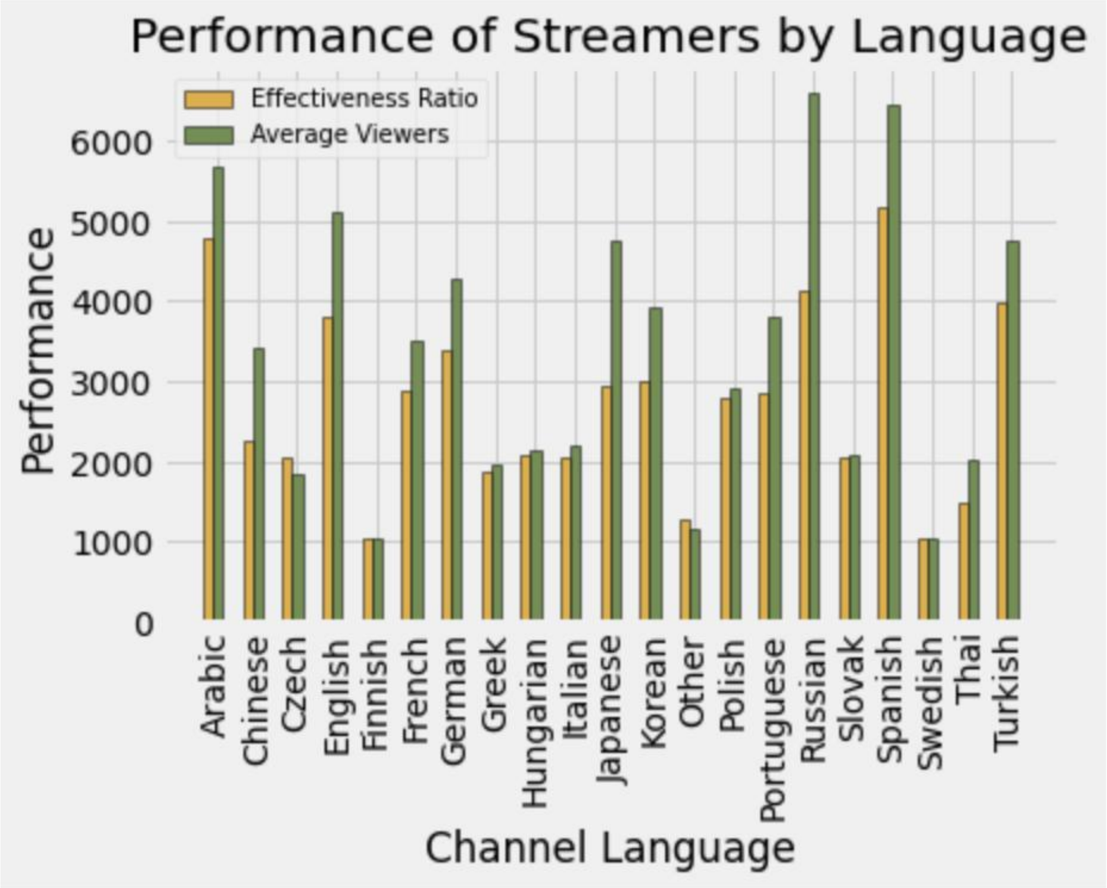


Figure 9 The side-by-side bar plot shows the effectiveness ratio (defined in the previous paragraph) and average viewers by language. The legend is at the upper left. We can observe that Russian, Spanish, and Arabic channels perform well and are effective at turning stream time into watch time.

Therefore, we can learn that channel languages certainly affect performance and that twitch channels need some re-allocations to make streamings more effective. Channels falling into the category of Slovak, Finish, Other, Russian, Spanish, and Arabic probably need more opportunities and are recommended to increase the stream time.

Next, the analysis explores partnership distribution and how partnership influences certain streaming performance.

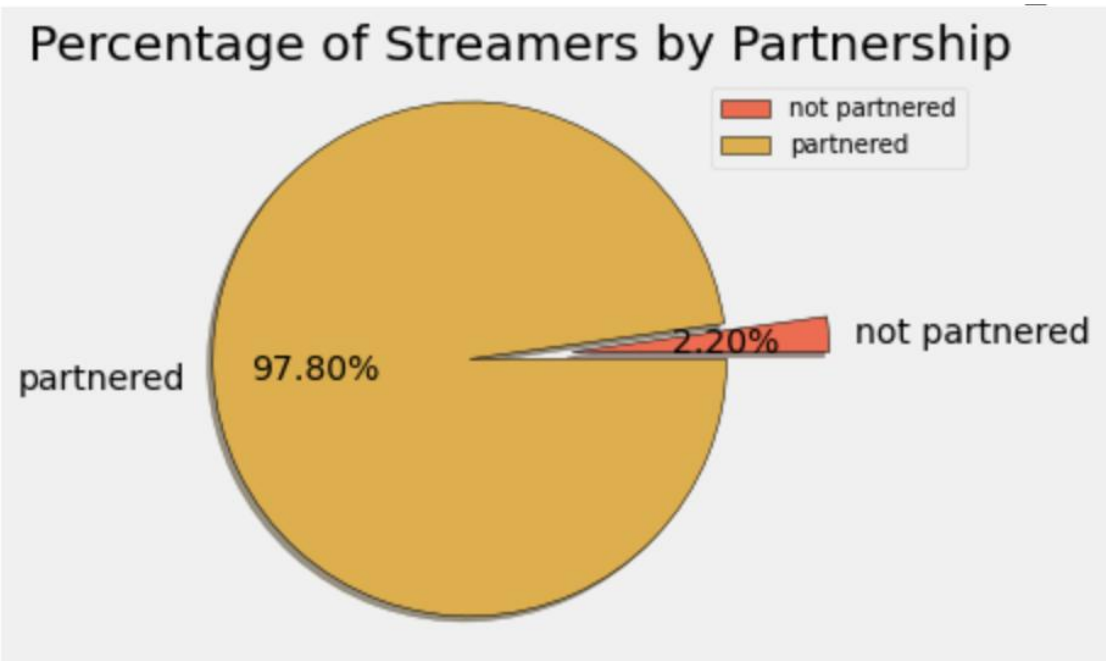


Figure 10 The pie chart shows the percentage of partnership share of channels on twitch. The legend is at the upper right. We can observe that a majority of total channels (97.80%) are partnered, while the rest are not.

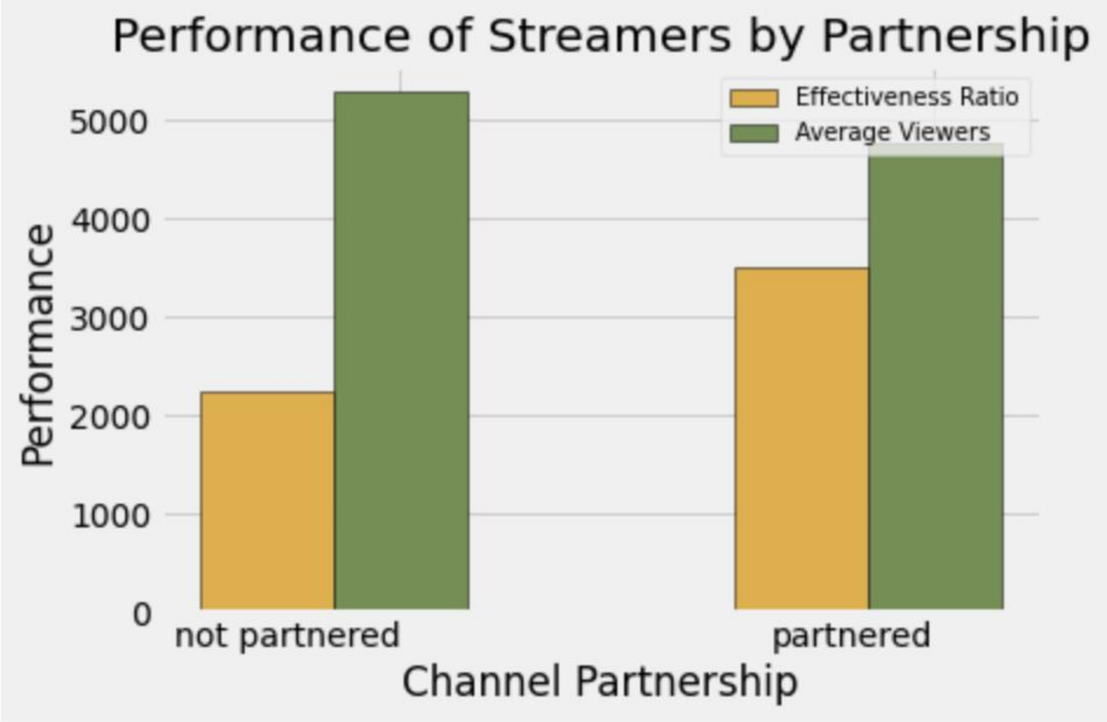


Figure 11 The side-by-side bar plot shows the effectiveness ratio (defined in the ninth paragraph) and average viewers by the partnership. The legend is at the upper left. We can observe that being partnered does not help with increasing average viewers but is effective at turning stream time into watch time.

So, we can learn that a majority of channels pursue partnerships, and streamers have to be clear at the ultimate goal (to have more viewers or to get more watch time per stream time) to decide whether to be partnered or not.

The last categorical variable that is to be evaluated is matureness. Different channels have their own fitting ages, and this difference might raise variety in performance. Let's move on to the matureness of channels.

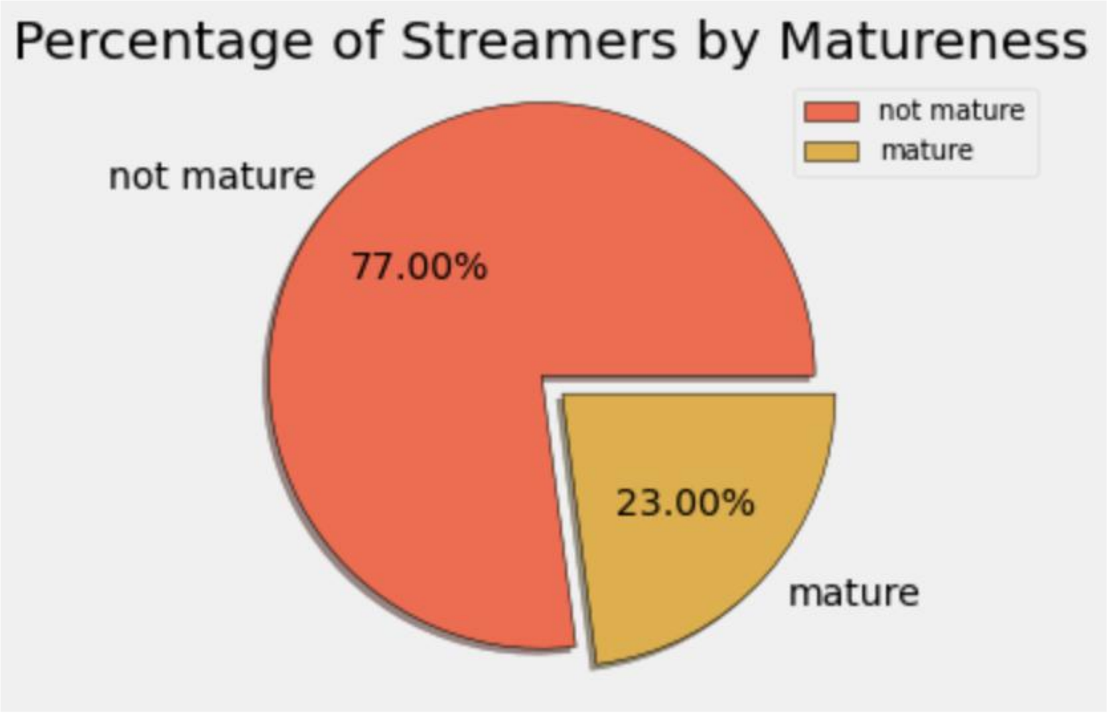


Figure 12 The pie chart shows the percentage of matureness share of channels on twitch. The legend is at the upper right. We can observe that most channels (77.00%) are not mature, while the rest are mature.

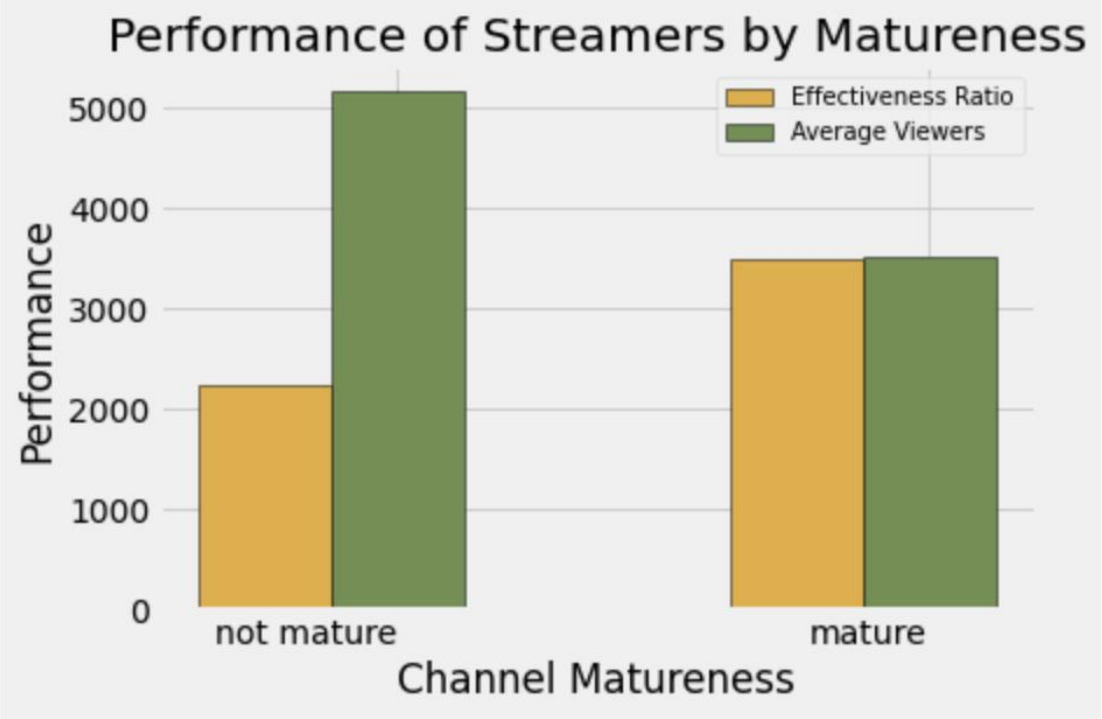
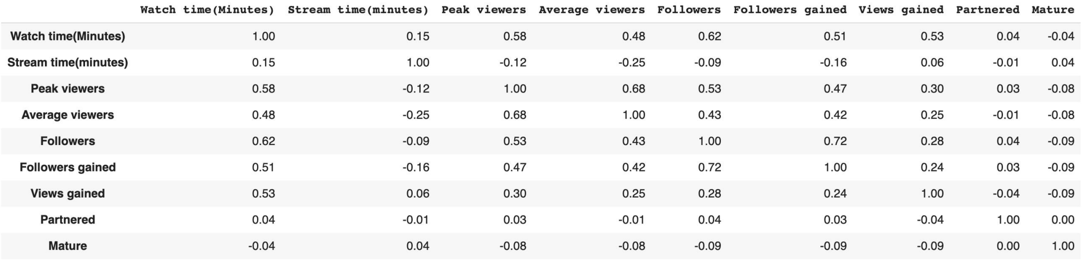


Figure 13 The side-by-side bar plot shows the effectiveness ratio (defined in the ninth paragraph) and average viewers by matureness. The legend is at the upper left. We can observe that not mature channels have more average viewers yet less effective in turning stream time into watch time.

Thus, most channels are not mature. It can be explained by young people are more open to these media and probably are looking for ways to earn money on twitch (older people usually do so in traditional approaches). These channels have more average viewers yet a less watch to stream time ratio. This may be due to young people like to watch young people's channels, but their watch time is often limited by either family rules or study.

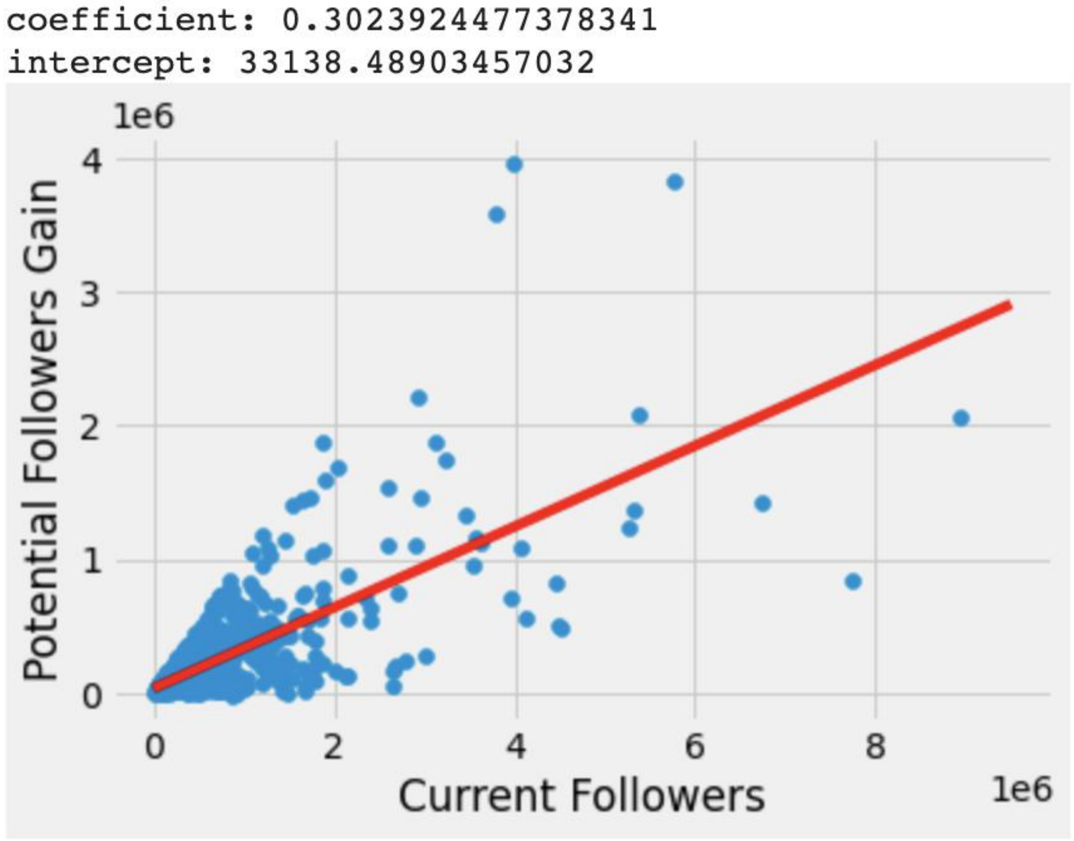
After exploring specific twitch data within categories, we would also want to use current data to do some predictions. Before building the prediction model, it is crucial to find the exact number of variables' correlations.



	Watch time(Minutes)	Stream time(minutes)	Peak viewers	Average viewers	Followers	Followers gained	Views gained	Partnered	Mature
Watch time(Minutes)	1.00	0.15	0.58	0.48	0.62	0.51	0.53	0.04	-0.04
Stream time(minutes)	0.15	1.00	-0.12	-0.25	-0.09	-0.16	0.06	-0.01	0.04
Peak viewers	0.58	-0.12	1.00	0.68	0.53	0.47	0.30	0.03	-0.08
Average viewers	0.48	-0.25	0.68	1.00	0.43	0.42	0.25	-0.01	-0.08
Followers	0.62	-0.09	0.53	0.43	1.00	0.72	0.28	0.04	-0.09
Followers gained	0.51	-0.16	0.47	0.42	0.72	1.00	0.24	0.03	-0.09
Views gained	0.53	0.06	0.30	0.25	0.28	0.24	1.00	-0.04	-0.09
Partnered	0.04	-0.01	0.03	-0.01	0.04	0.03	-0.04	1.00	0.00
Mature	-0.04	0.04	-0.08	-0.08	-0.09	-0.09	-0.09	0.00	1.00

Figure 14 The correlation table shows the exact correlation value between quantitative measurements. We can observe that followers vs. followers gained has the highest correlation.

As we only have 0.72 for the highest correlation value, we are just going to build a model for followers and followers gained (other values are not strong enough to do so).



```
coefficient: 0.3023924477378341
intercept: 33138.48903457032
```

Figure 15 The linear regression shows the relationship between followers and followers gained. Note that both axes are in 10^6. The regression's coefficient is 0.30, and its intercept is 33138.49.

With the relationship and the information about the linear regression line, based on their current followers, twitch's newcomers can quickly estimate their potential follower gain until the next data-collecting time period. They can use these data to determine whether they should stay on twitch or open channels on other platforms.

These are all visuals and insights I obtained from exploring and analyzing twitch data, including the top performance streamers, correlations between each measurement standard, categorical performance differentiation (language, partnership, and matureness), and the prediction model for newcomers. Despite twitch is a large and influential streaming platform, it still has limitations, and its data cannot represent the entire field. Nonetheless, it can at least indicate something. Hope the analysis offers you several thoughts on data analysis and on the streaming industry.