StatCrunch Competition
Twitch Dataset

Matthew Carson
University of California, Los Angeles

April 28, 2024

Contents

List of Figures

1 Summary Statistics

Since these data were not randomly sampled, it would be inappropriate to conduct inference (i.e., construct confidence intervals or conduct hypothesis tests). Because of this, it is not possible to estimate population parameters; that is, make claims or generalizations about the broader population of Twitch users. However, since these data represent the top 900 Twitch users, statistics can be calculated and relationships can be discovered about that population.

Initial calculations were made before presenting the summary statistics. Since values for `'Watch time (mins)'` were typically very large, requiring scientific notation to express, values were rescaled to `'Mean weekly watch hours'` by dividing `'Watch time (mins)'` by the product of 60 times 52 (number of weeks in a year) to make the numbers more manageable:

$$Mean\ weekly\ watch\ hours = \frac{Watch\ time\ (mins)}{60 * 52} \tag{1}$$

Additional statistics were calculated as well:

- `'Followers Prev Yr'` = `'Followers'` - `'Followers gained'`.
- `'Followers gained percent'` = `'Followers gained'` / `'Followers Prev Yr'`.

Because all distributions are heavily right skewed (skewness ≥ 2.6), medians, represented with Greek letter eta (η), are reported instead of means (all values reported are from Table 1). The majority of the top nine hundred accounts stream content at least 30 hours per week ($\eta \approx 34.23$) and are watched more than 90 thousand hours per week ($\eta \approx$ 91,422). Most accounts gained a substantial number of followers from the previous year ($\eta \approx$ 66,003), which represents a median increase of approximately 16 percent. Because of the heavy skewness of the distributions, easy-to-interpret visualizations were difficult to make (see appendix for histograms).

To assess the relationships between numeric variables, Spearman's correlation coefficients were calculated. Because of the non-linearity of the relationships, typical Pearson's R correlation coefficients would be inappropriate. Spearman's correlation coefficients are preferred for assessing the strength of non-linear relationships.

The relationships between numeric variables are surprising, especially the absence of some correlations where one would think they they would exist. 'Stream time' has a moderately strong negative correlations with 'Average viewers', which is counterintuitive

Summary statistics:

Column	n	Range	Min	Q1	Median	Q3	Max	IQR	Skewness	Kurtosis
Mean weekly watch hours	900	2162429	54741.952	68682.611	91421.861	137050.45	2217170.9	68367.839	4.5932318	27.933122
Mean weekly stream hours	900	167.11538	1.2451923	22.632212	34.230769	46.65625	168.36058	24.024038	2.6807611	8.8522208
Peak viewers	900	3441783	962	10632.5	19709.5	42840.5	3442745	32208	14.536547	278.83939
Average viewers	900	131990	366	1850	3113.5	6044	132356	4194	5.66002	48.595566
Followers Prev Yr	900	14546214	135	186861.5	396875	866859	14546349	679997.5	4.7607089	28.621801
Followers	900	16119419	18437	249737	489640	1043609	16137856	793872	4.6234316	27.164653
Followers gained	900	5016024	-73927	23587	66002.5	164270.5	4942097	140683.5	6.6702471	60.633802
Followers gained percent	900	755.04229	-0.11636715	0.063403004	0.15944607	0.38231599	754.92593	0.31891299	29.50479	879.30102

Table 1: These are summary statistics for the dataset.

given that one might expect more frequent streaming to result in more viewers, but that is not the case. In terms of change over time, the more one streams With respect to surprising absences of relationships, 'Stream time' has practically no relationship with 'Watch time'. Together, these findings suggest that a strategy of merely increasing one's streaming time does not "pay off" in terms of the number of followers or viewers.