StatCrunch Competition
Twitch Dataset

Matthew Carson
University of California, Los Angeles

April 29, 2024

Contents

List of Tables

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 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

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: Summary Statistics.

represents a median increase of approximately 16 percent. Because of the heavy skewness of the distributions, easy-to-interpret visualizations were difficult to make (Fig. 4).

2 Correlation

To assess the relationships between numeric variables, Spearman's correlation coefficients were calculated (Fig. 1). Because of the non-linearity of the relationships (Fig. 5), typical Pearson's R correlation coefficients would be inappropriate. Spearman's correlation coefficients are preferred for assessing the strength of non-linear relationships.

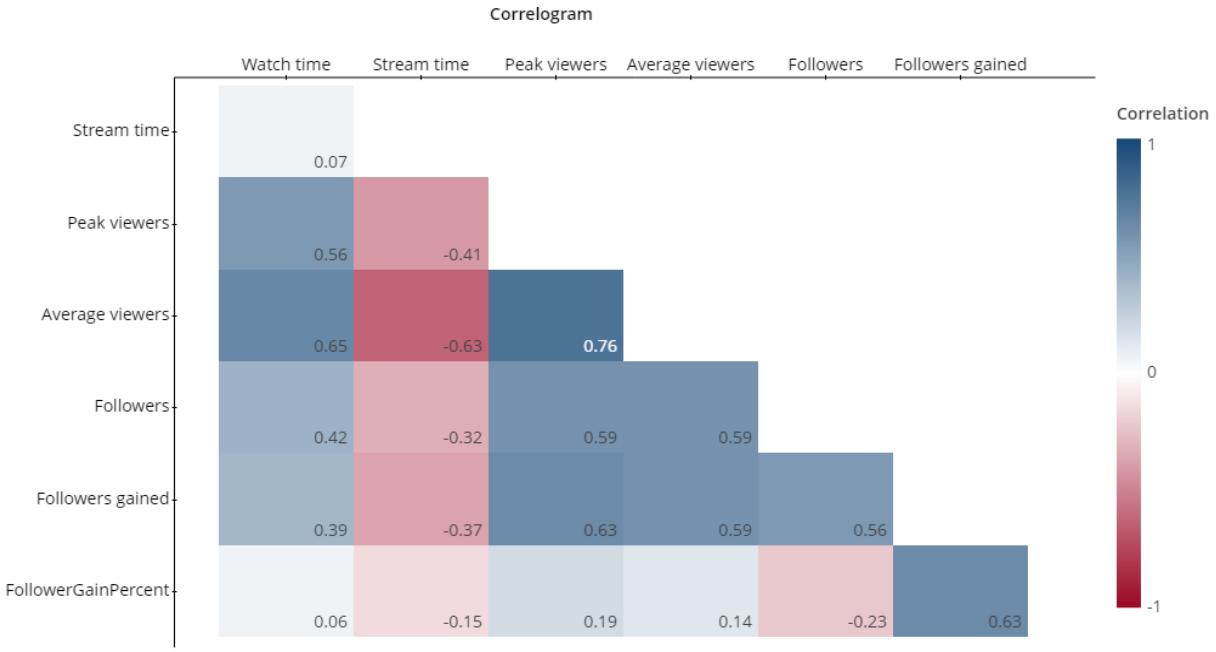

Figure 1: Spearman Correlogram

The relationships between numeric variables are surprising, especially the absence of some correlations where one would think they they would exist (Fig. 1). `'Stream time'` has a moderately strong negative Spearman correlation (ρ) with `'Average viewers'` ($\rho = -0.63$), which is counterintuitive 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, accounts that streamed more hours did *not* gain more followers; indeed, the accounts that were in the top decile of weekly stream time gained less than one-fifth of the followers that the accounts in the lowest stream time decile gained (Table 2; Fig. 2). With respect to surprising absences of relationships, `'Stream time'` has practically no relationship with `'Watch time'` ($\rho = 0.07$;

Fig. 1). 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.

Summary statistics for Followers gained:
Group by: Decile(Mean weekly stream hours)

Decile(Mean weekly stream hours)	n	Min	Max	Q1	Median	Q3	IQR	Skewness	Sum
1	90	-50044	2357734	76157	161864	334780	258623	3.208376	25113566
2	90	1794	2908884	53652	118075	223059	169407	3.8221382	24067622
3	90	-73927	3333126	43905	111803	259317	215412	4.8287635	21967095
4	90	5060	3188636	32716	70134.5	177347	144631	4.9468279	18171458
5	90	-6884	1077925	24736	61008.5	145078	120342	3.1782676	10812529
6	90	-6992	4942097	22146	56464.5	138969	116823	7.6579575	15845831
7	89	-18066	3020501	15108	34484	77542	62434	6.5304191	11092581
8	91	619	550107	15513	52120	112217	96704	2.0740535	8868969
9	90	-25062	593969	16319	44798.5	113868	97549	2.5940324	7838226
10	90	-3416	379434	9776	28283	62415	52639	2.5942559	4884724

Table 2: Followers Gained by Stream Time Deciles. (The lowest decile streamed the least.)

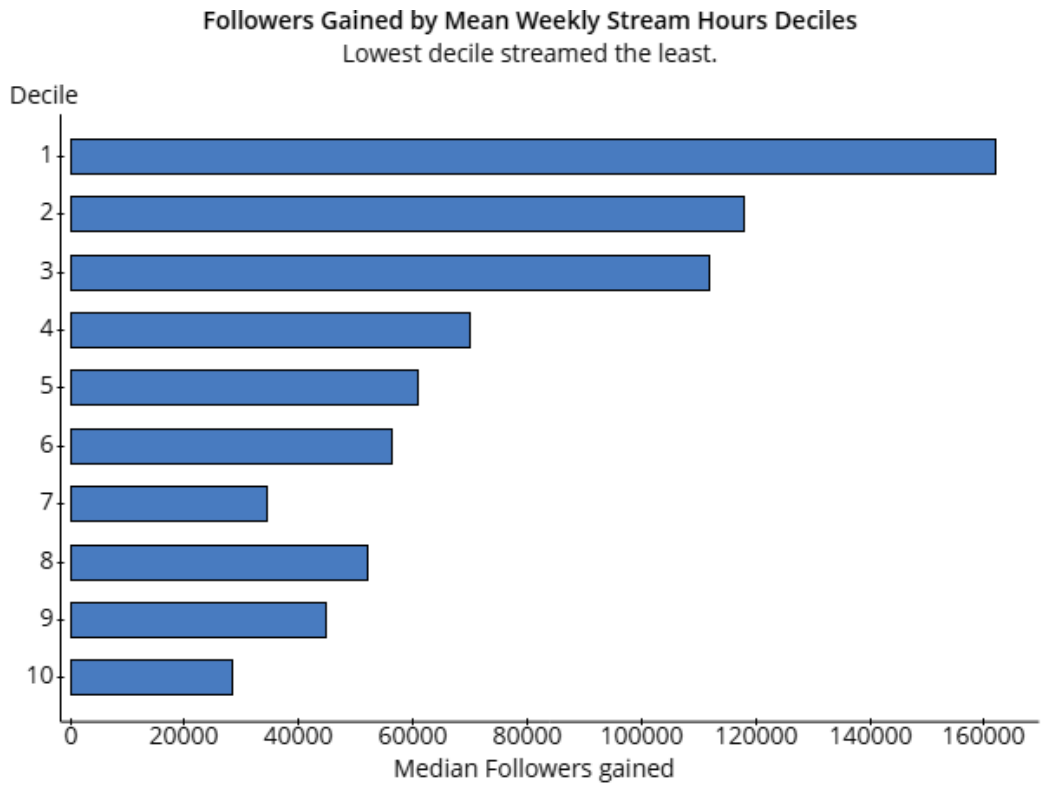

Figure 2: Followers Gained by Stream Time Deciles. (The lowest decile streamed the least.)

3 Simple Linear Regressions

Because the relationships between the variables were not linear, it was difficult to fit models using simple linear regression. However, using transformations, I was able to fit some models.

3.1 Stream Hours to Predict Average Viewers

A simple linear regression was run to assess the strength between `'Mean weekly stream hours'` (independent variable) and `'average viewers'` (dependent variable). Because of the non-linear relationship between the variables, the residuals were highly non-normal. A inverse (reciprocal) transformation of `'Average viewers'` was performed to correct for non-normality of the residuals. The transformation greatly improved the distribution of the residuals, making them nearly normal. The transformation could not correct for heteroskedasticity, but this is not an issues since inference is not being conducted.

3.1.1 Model Specification

The regression model is as follows:

$$\frac{1}{Average\ viewers_i} = \beta_0 + \beta_1 * Mean\ weekly\ stream\ hours_i \tag{2}$$

where i is a Twitch account. *'Average viewers'* is the average number of viewers that watched the respective Twitch account; and *'Mean weekly stream hours'* is the number of hours that the respective Twitch account streamed over the year divided by 52.

3.1.2 Results

Table 3: Simple linear regression model showing a moderate relationship between average viewers and mean weekly stream hours.

R-squared is moderate, suggesting that the mean weekly stream hours can explain 56 percent of the variation in the average number of viewers. Because of the inverse transformation of the dependent variable, the signs of the intercept and mean weekly stream hours coefficient are reversed. This makes sense though, since an increase in the denominator of the equation (when the right hand side of the equation is back transform; Eq. 3) will diminish the predicted number of average viewers.

$$Average\ viewers_i = \frac{1}{4.04e^{-5} + 9.15e^{-6} * Mean\ weekly\ stream\ hours_i} \tag{3}$$

Figure 3 shows the relationship between the mean weekly stream hours and the inverse of average viewers. The other scatter plots show the residuals. The histogram and Q-Q plots show that the residuals are now nearly normal, although excess kurtosis is still high, making the distribution of the residuals leptokurtic (skewness = -0.57036091; excess kurtosis = 6.0727871). Removing extreme dependent variable observations would help correct for this, but it is unlikely to be helpful since inference is not being conducted. Heteroskedasticity also was not corrected, but it also is not an issue as no hypothesis testing is being conducted, nor are confidence intervals being calculated.

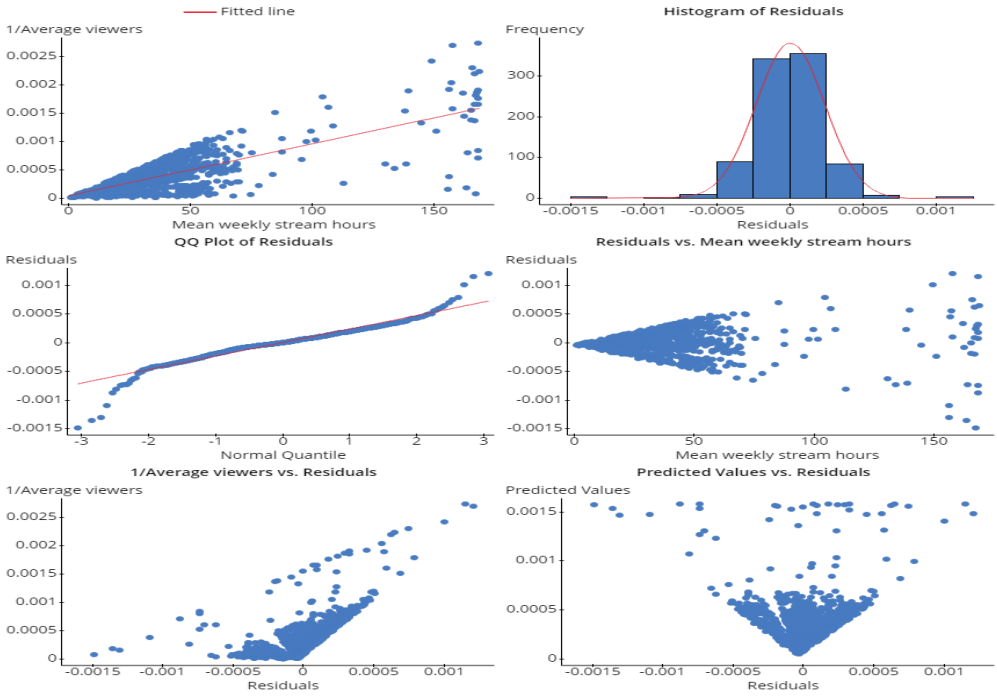

Figure 3: Residual Plots of the model. There residuals are symmetric, making the model a decent fit, nothwithstanding the non-constant variance.

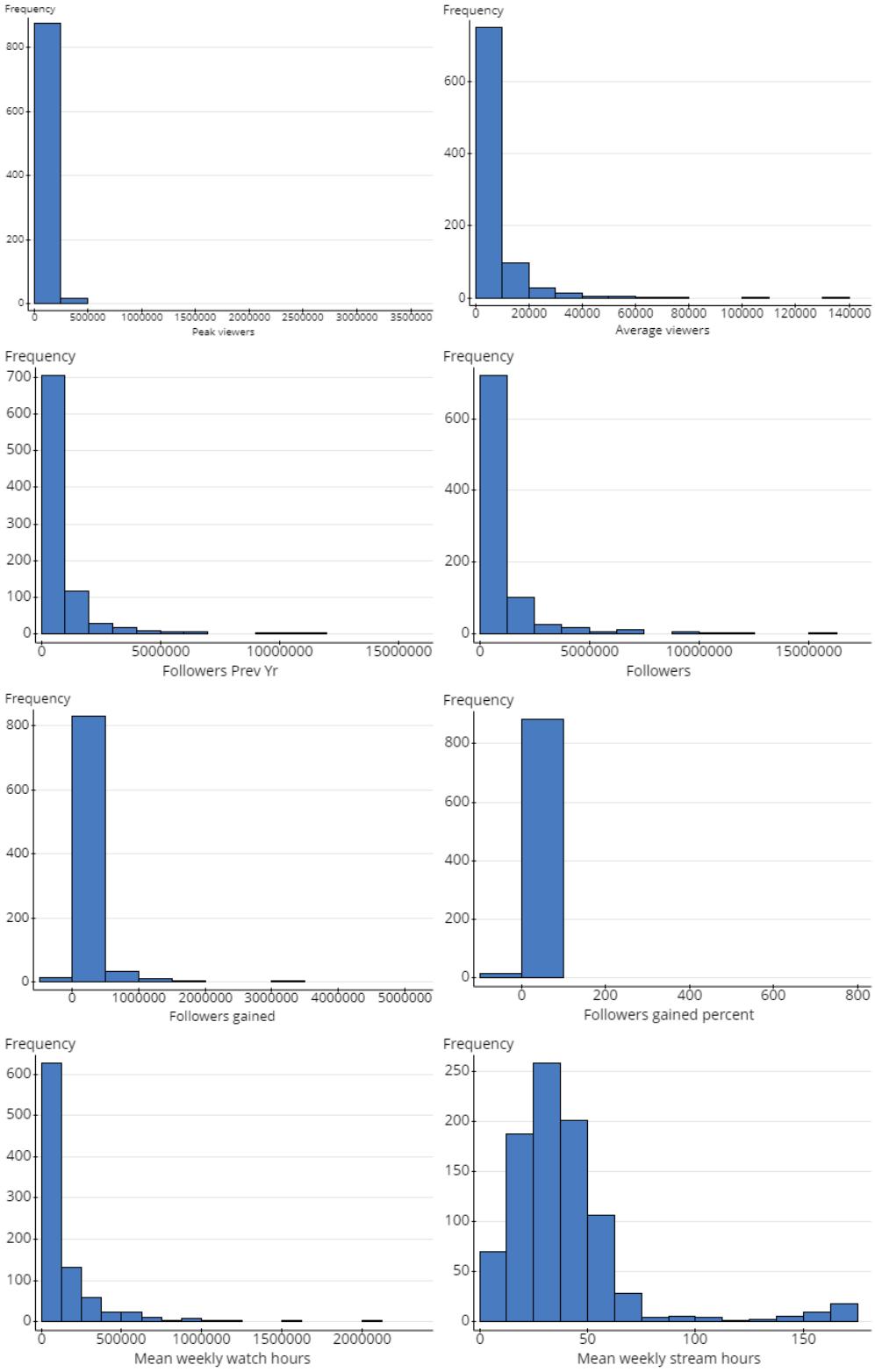

Figure 4: All distributions are heavily skewed and non-normal.

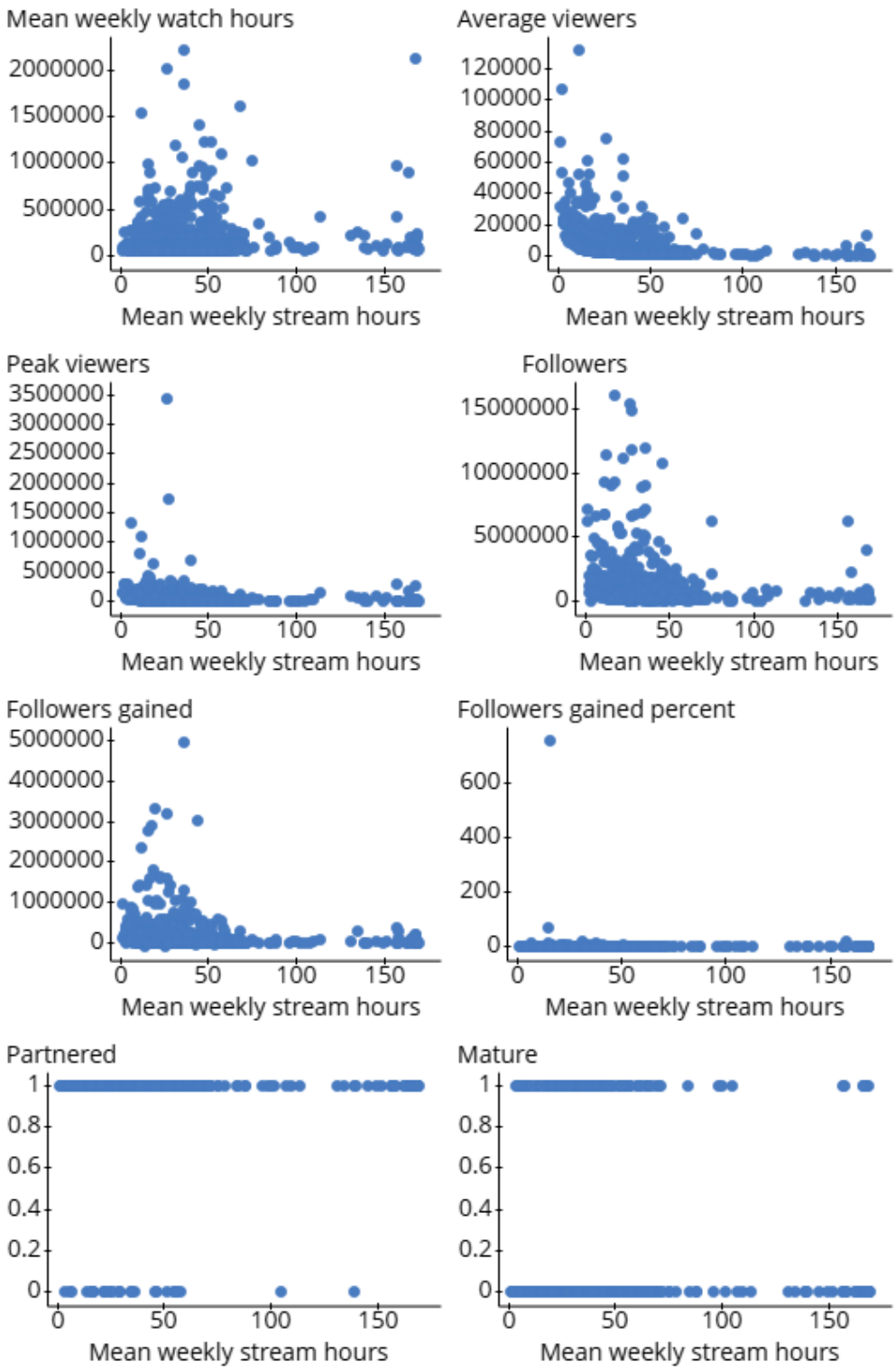

Figure 5: Relationships between `'Mean weekly stream hours'` and other variables.

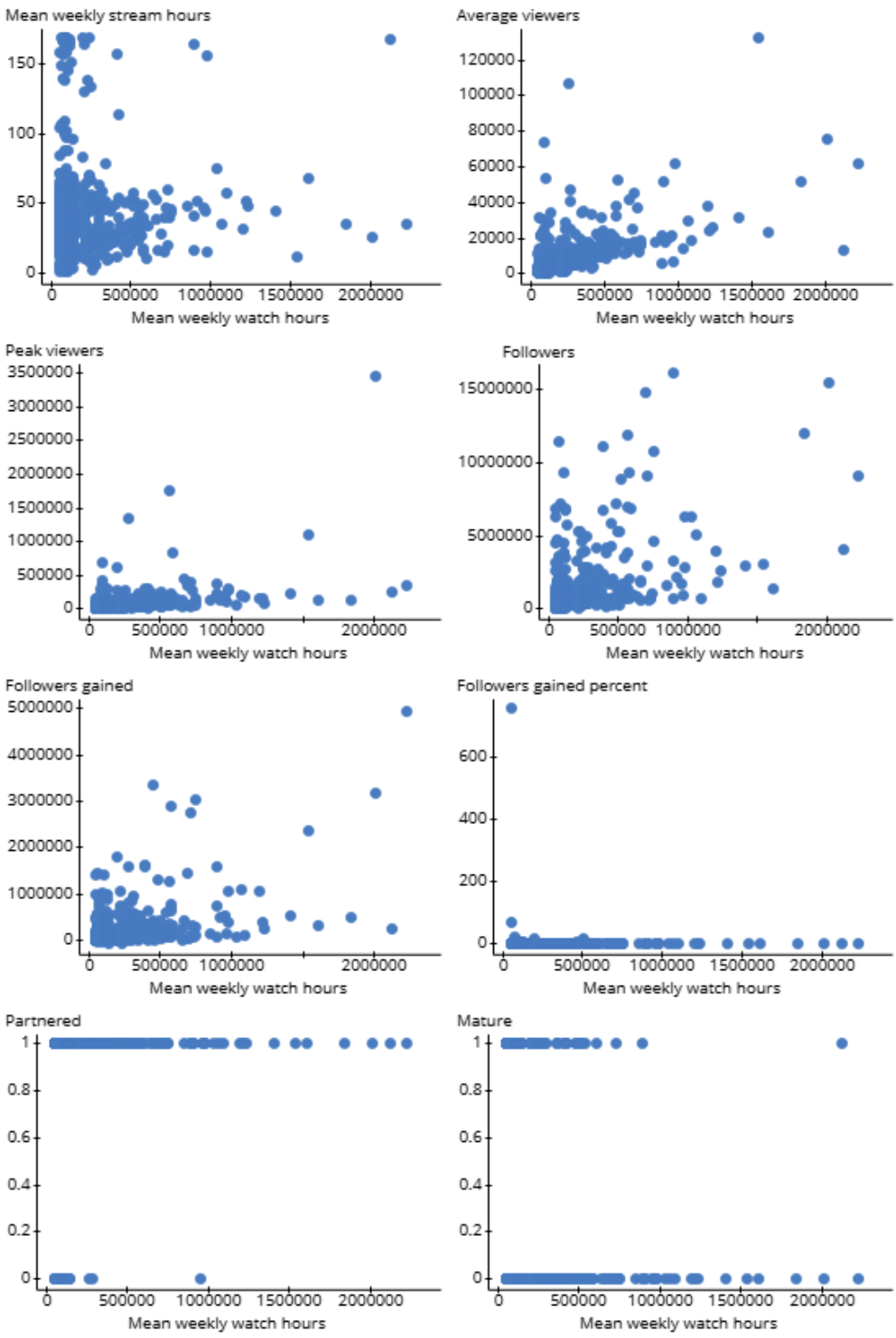

Figure 6: Relationships between `'Mean weekly watch hours'` and other variables.

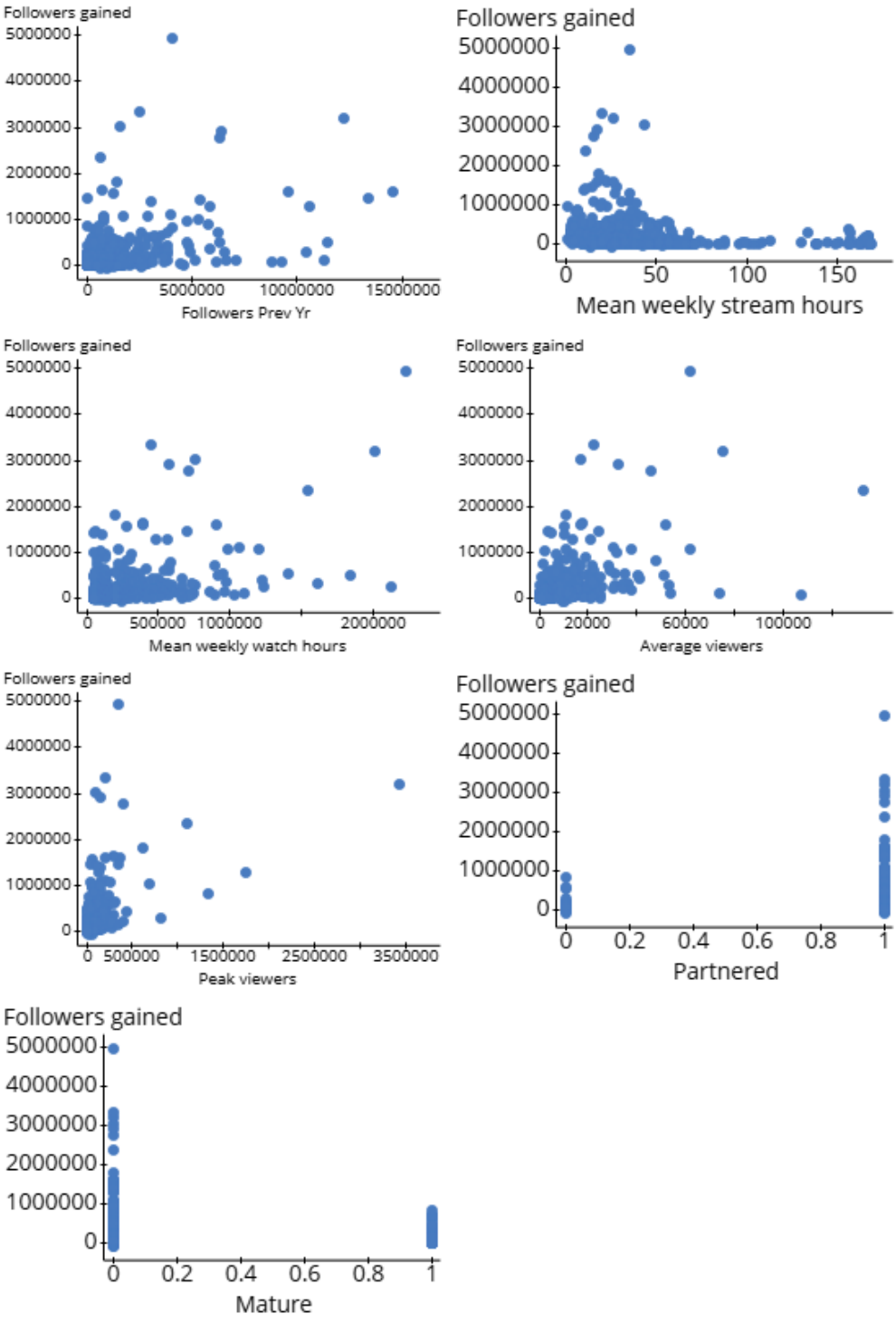

Figure 7: Relationships between `'Followers gained'` and other variables.