The attached submission won 1st place in the StatCrunch statistics contest:

https://www.pearson.com/en-us/higher-education/campaigns/statcrunch/statcrunch-contest.html

StatCrunch Competition
Twitch Dataset

Matthew Carson
University of California, Los Angeles

April 30, 2024

Contents

List of Tables

List of Figures

1 Note on Absence of Inference

Since these data were not randomly sampled, it would be inappropriate to conduct inference (e.g., 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.

2 Summary Statistics

Initial calculations were made before presenting the summary statistics. Since values for `‘Watch time (mins)’` and `‘Stream time (mins)’` were typically very large, requiring scientific notation to express, values were rescaled by dividing the values in both columns by the product of 60 (60 minutes in an hour) times 52 (number of weeks in a year) to make the numbers more manageable:

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

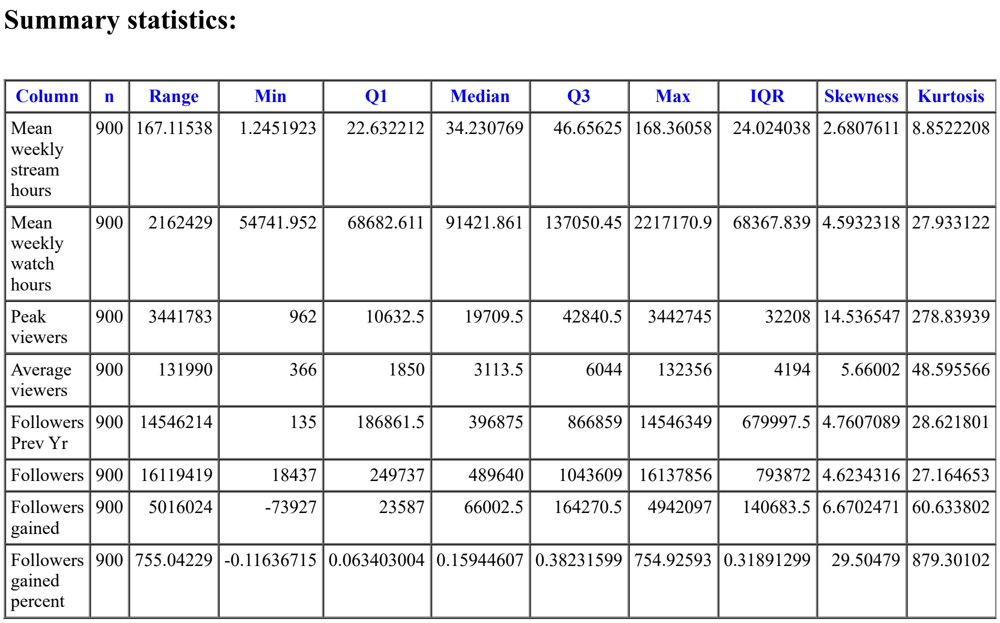
and

$$Mean\ weekly\ stream\ hours = \frac{Stream\ time\ (mins)}{60 * 52} \tag{2}$$

Additional statistics were calculated as well.[1]

- `‘Followers Prev Yr’` = `‘Followers’` - `‘Followers gained’`.
- `‘Followers gained percent’` = `‘Followers gained’` / `‘Followers Prev Yr’`.

[1] Entire dataset available: (click here)



Summary statistics:

Column	n	Range	Min	Q1	Median	Q3	Max	IQR	Skewness	Kurtosis
Mean weekly stream hours	900	167.11538	1.2451923	22.632212	34.230769	46.65625	168.36058	24.024038	2.6807611	8.8522208
Mean weekly watch hours	900	2162429	54741.952	68682.611	91421.861	137050.45	2217170.9	68367.839	4.5932318	27.933122
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.

Because all distributions are heavily right-skewed (skewness ≥ 2.6), medians, represented with the 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 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. 13).

3 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. 10, Fig. 11, Fig. 12), typical Pearson's R correlation coefficients would be inappropriate. Spearman's correlation coefficients are preferred for assessing the strength of non-linear relationships.

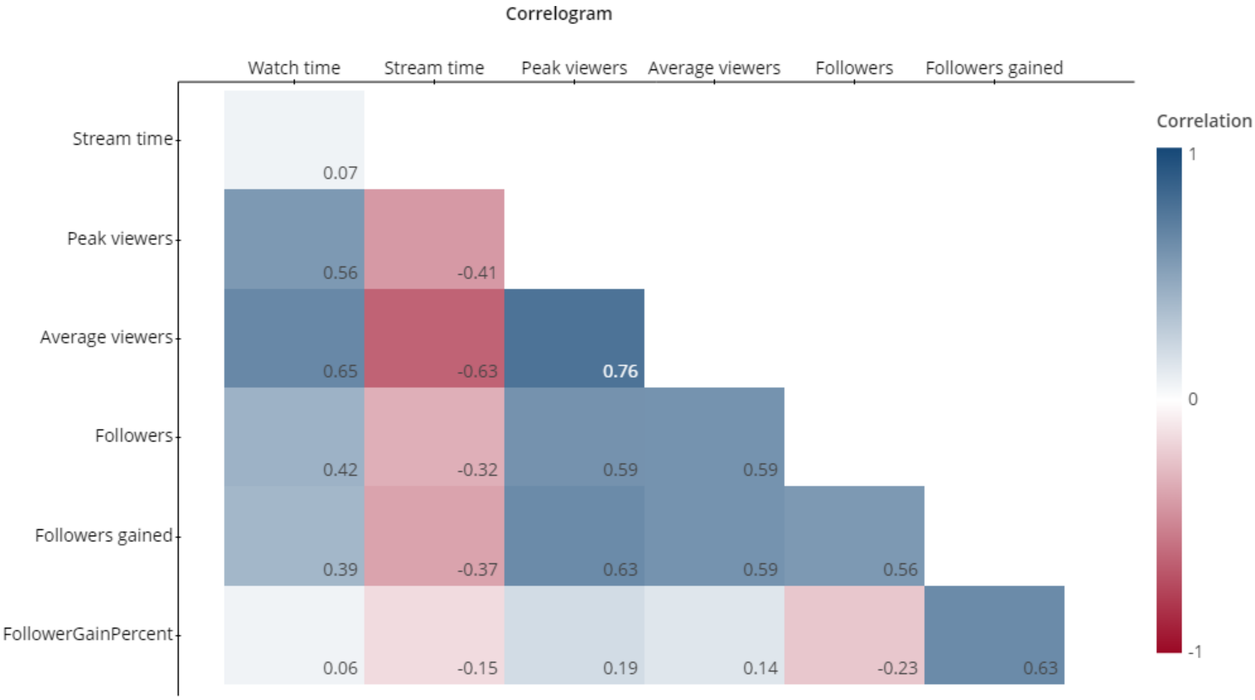
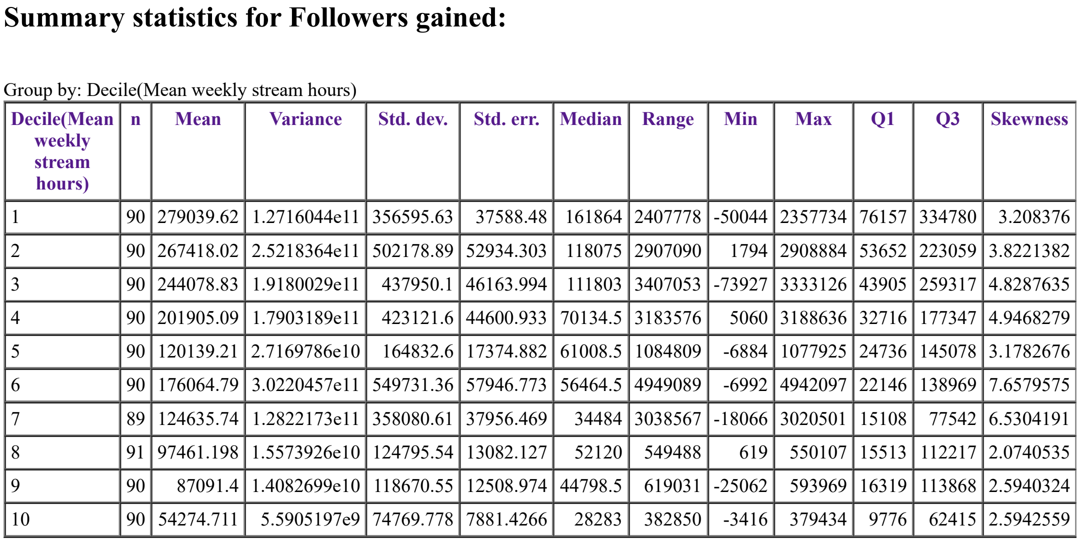


Figure 1: Spearman Correlogram

The relationships between numeric variables are surprising, especially the absence of some correlations where one would think they 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	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3	Skewness
1	90	279039.62	1.2716044e11	356595.63	37588.48	161864	2407778	-50044	2357734	76157	334780	3.208376
2	90	267418.02	2.5218364e11	502178.89	52934.303	118075	2907090	1794	2908884	53652	223059	3.8221382
3	90	244078.83	1.9180029e11	437950.1	46163.994	111803	3407053	-73927	3333126	43905	259317	4.8287635
4	90	201905.09	1.7903189e11	423121.6	44600.933	70134.5	3183576	5060	3188636	32716	177347	4.9468279
5	90	120139.21	2.7169786e10	164832.6	17374.882	61008.5	1084809	-6884	1077925	24736	145078	3.1782676
6	90	176064.79	3.0220457e11	549731.36	57946.773	56464.5	4949089	-6992	4942097	22146	138969	7.6579575
7	89	124635.74	1.2822173e11	358080.61	37956.469	34484	3038567	-18066	3020501	15108	77542	6.5304191
8	91	97461.198	1.5573926e10	124795.54	13082.127	52120	549488	619	550107	15513	112217	2.0740535
9	90	87091.4	1.4082699e10	118670.55	12508.974	44798.5	619031	-25062	593969	16319	113868	2.5940324
10	90	54274.711	5.5905197e9	74769.778	7881.4266	28283	382850	-3416	379434	9776	62415	2.5942559

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

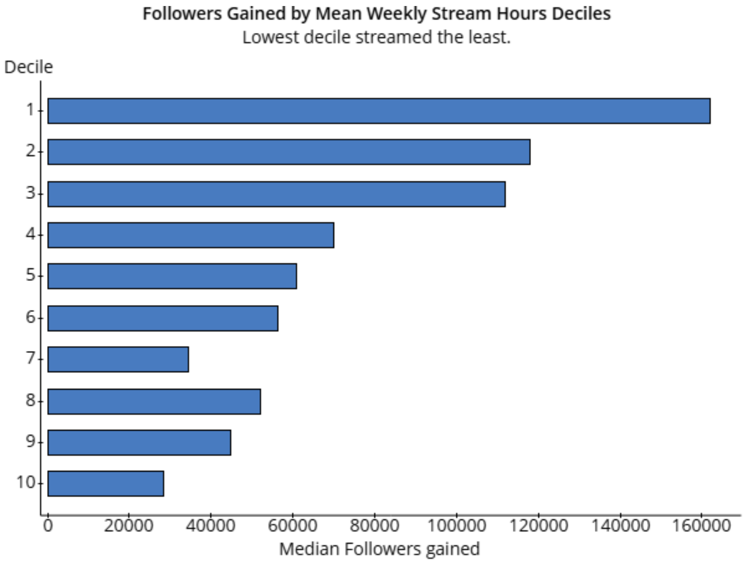


Figure 2: Followers Gained by Stream Time Deciles. (The lowest decile streamed the least.)

4 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.

4.1 Stream Hours to Predict Average Viewers

Do accounts that stream more hours have more viewers? Intuition would suggest yes, but the question deserves empirical examination. 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. An inverse (reciprocal) transformation of `‘Average viewers’` was performed to correct for the 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 issue since inference is not being conducted.

4.1.1 Model Specification

The regression model is as follows:

$$\frac{1}{Average\ viewers_i} = \beta_0 + \beta_1 * Mean\ weekly\ stream\ hours_i \tag{3}$$

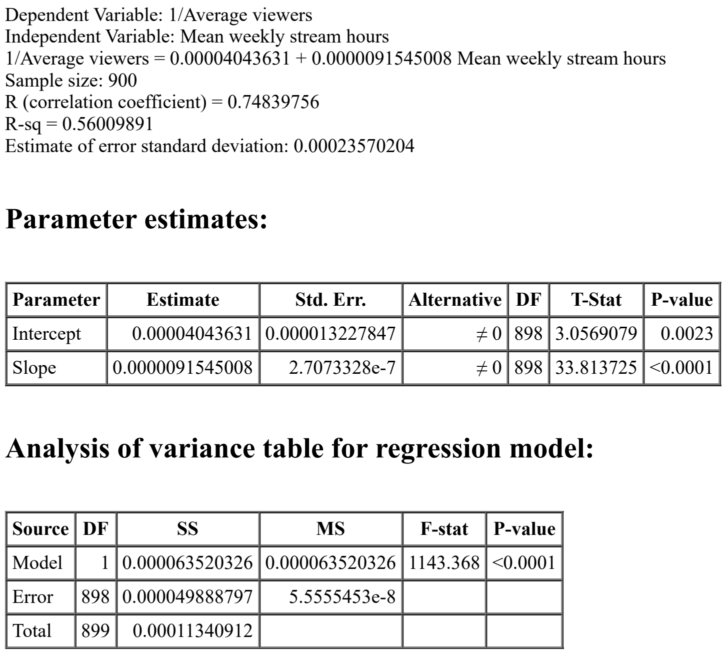
where i is a Twitch account. *‘Average viewers’* is the average number of viewers that watched the respective Twitch account; *‘Mean weekly stream hours’* is the number of hours that the respective Twitch account streamed over the year divided by 52.

4.1.2 Simple Linear Regression Results

R-squared is moderate (0.56), 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 transformed; Equation 5) will diminish the predicted number of average viewers.

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 symmetric, 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 improves this, but it is unlikely to be helpful since inference is not being conducted, and fitting data to your model is not the best practice.[2] Heteroskedasticity also was not corrected, but it is not an issue

[2] It is better to adjust your model to fit the data than the other way around.



Dependent Variable: 1/Average viewers
Independent Variable: Mean weekly stream hours
1/Average viewers = 0.00004043631 + 0.0000091545008 Mean weekly stream hours
Sample size: 900
R (correlation coefficient) = 0.74839756
R-sq = 0.56009891
Estimate of error standard deviation: 0.00023570204

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	0.00004043631	0.000013227847	≠ 0	898	3.0569079	0.0023
Slope	0.0000091545008	2.7073328e-7	≠ 0	898	33.813725	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	0.000063520326	0.000063520326	1143.368	<0.0001
Error	898	0.000049888797	5.5555453e-8		
Total	899	0.00011340912			

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

as no hypothesis testing is being conducted, nor are confidence intervals being calculated.[3] The model with coefficients is as follows:

$$\frac{1}{Average\ viewers_i} = 4.04e^{-5} + 9.15e^{-6} * Mean\ weekly\ stream\ hours_i \tag{4}$$

[3] See Section 1 for an explanation of why I chose not to conduct inference.

4.1.3 Scatter Plots

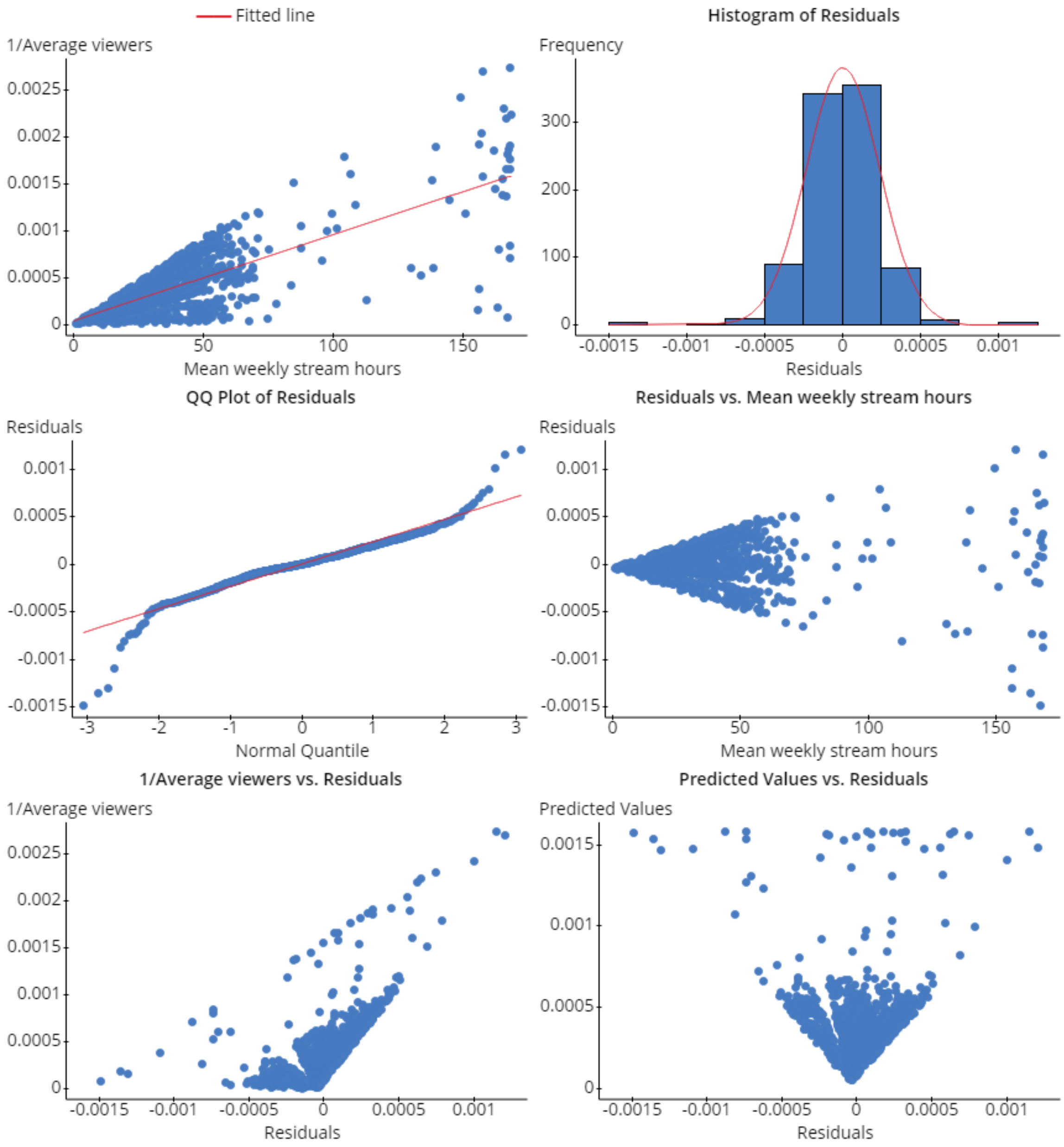


Figure 3: Model plots. The residuals are symmetric, making the model a decent fit, notwithstanding the non-constant variance.

Since one usually does not have a particular interest in knowing what the reciprocal of the average number of viewers is, `‘Average viewers’` was reverse transformed, and the transformed equation was plotted over the data on their original scale (Fig. 4).

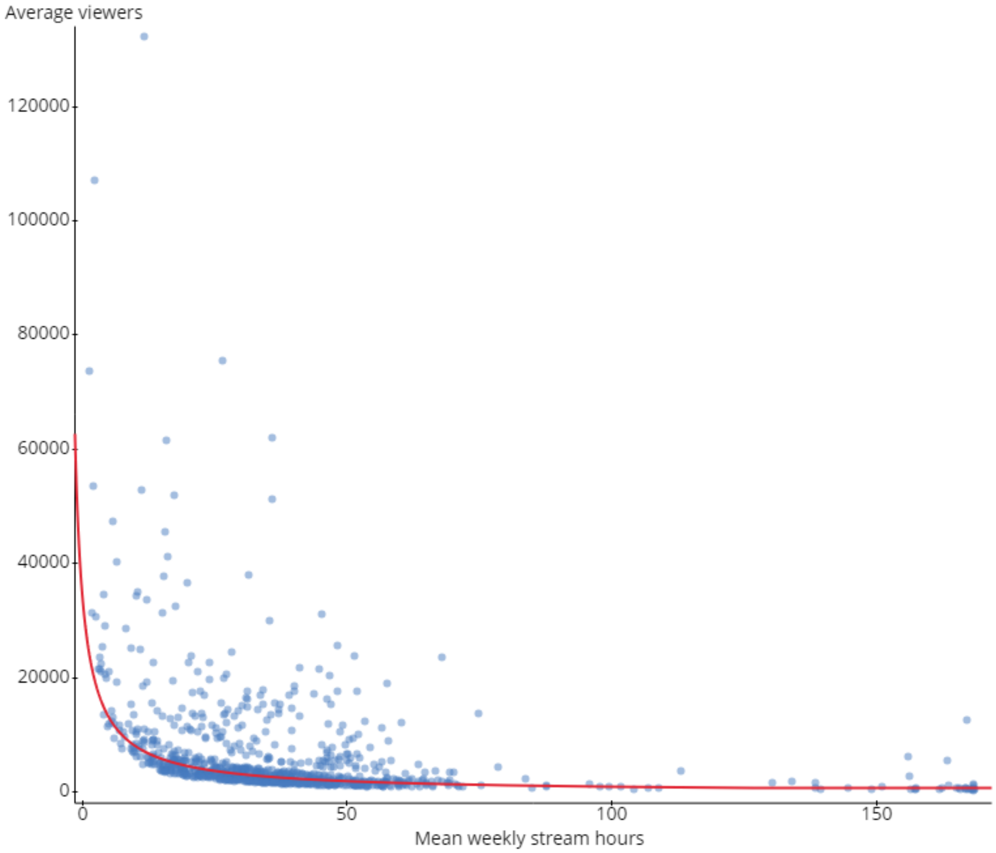


Figure 4: Back Transformed Average Viewers Predicted by Stream Hours. The fitted line is Equation 5

.

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

4.1.4 Discussion

Contrary to what one might think, the amount of time that a Twitch account streams has a negative association with the average number of viewers. So, if one wants to increase their viewership, streaming more does not seem to be a good strategy. A quick illustration is helpful. The median weekly stream hours for all accounts is 34.23 hours per week. Using Equation 5, we can estimate that the mean average number of viewers[4] for accounts that stream the median number of hours per week is ≈ 2826. Accounts that stream ten hours less than the median per week have a mean average viewership of 3813, which is 987 more

[4] The formulation appears awkward, but the variable is `'Average viewers'`, and the predicted value represents the mean `'Average viewers'` for that particular `x` value.

viewers than accounts that stream the median number of hours per week (See Table 7 for predicted values).

4.2 Stream Hours to Predict Followers

Do accounts that stream more have more followers? Intuitively, it would make sense. A simple linear regression was run to assess the relationship between streaming time, `‘Mean weekly stream hours’`, and the number of `‘Followers’` Twitch accounts have. The initial model revealed a highly non-linear relationship with highly non-normal residuals. A log transformation of `‘Followers’` was conducted to correct for this and the models subsequently fit the data much better.

4.2.1 Model Specification

The regression model is as follows:

$$ln(Followers_i) = \beta_0 + \beta_1 * Mean\ weekly\ stream\ hours_i \tag{6}$$

where i is a Twitch account; `‘Followers’` is the number of followers that a Twitch account has; `‘Mean weekly stream hours’` is the amount of time a Twitch account streamed over the year divided by 60 and then divided by 52.

4.2.2 Simple Linear Regression Results

The regression reveals a rather weak negative correlation between the mean weekly stream hours and the number of followers that an account has. R-squared is only 0.056. The coefficient for `‘Mean weekly stream hours’` (Table 4) was exponentiated ($e^{-0.0088658957} = 0.99117$), indicating that for every additional weekly streaming hour, an account should expect 0.88 percent fewer followers. (See Table 4 for all model parameters.)

4.2.3 Scatter Plots

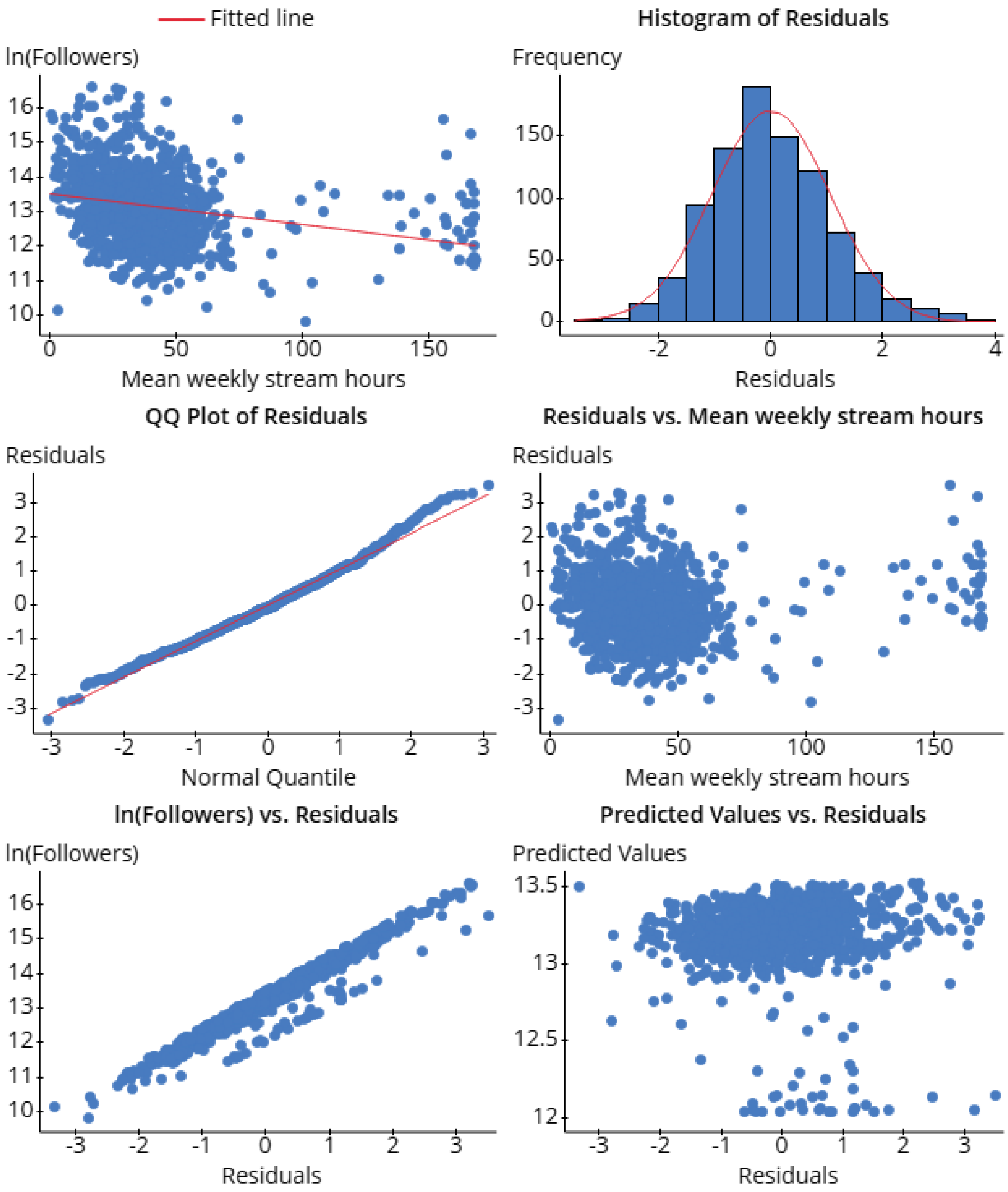


Figure 5: ‘Followers’ regressed against ‘Mean weekly stream hours’.

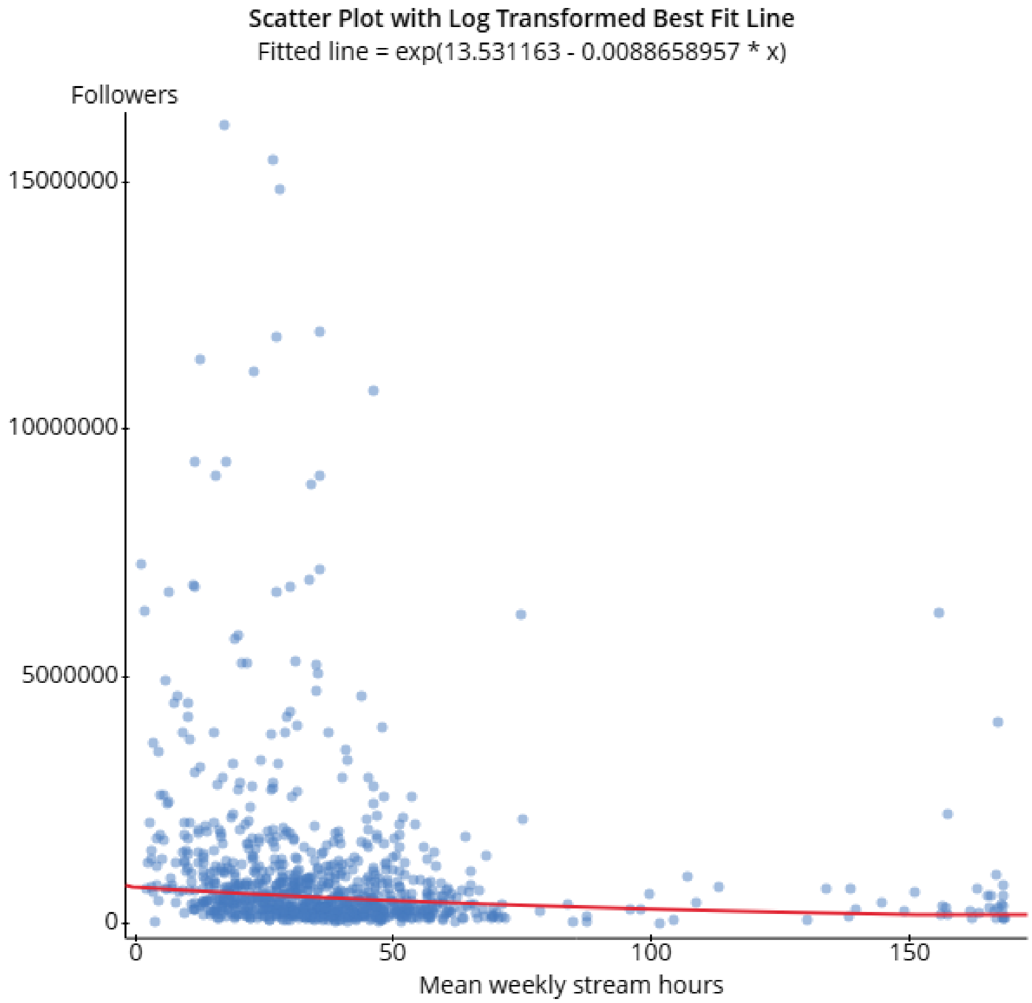
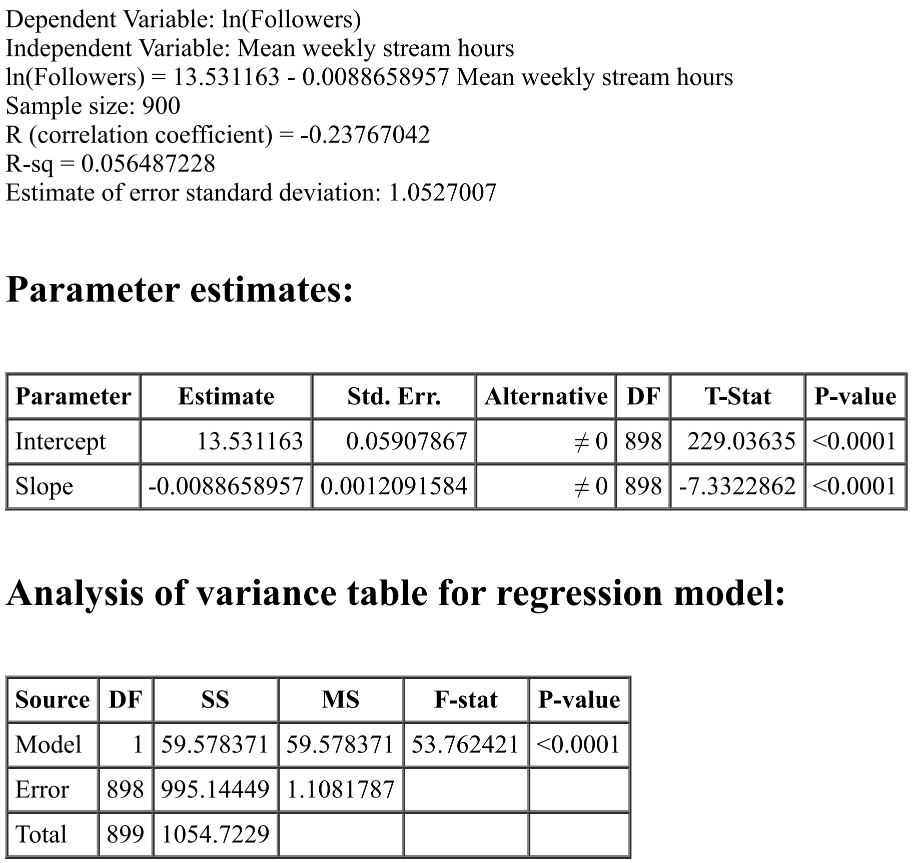


Figure 6: `'Followers'` regressed against `'Mean weekly stream hours'`.

4.2.4 Discussion

The results, again, are counterintuitive. Rather than observing more followers with accounts that stream more frequently, we see the opposite trend, albeit weakly. Again, this suggests that simply streaming more frequently or for longer durations is not sufficient to build a more successful Twitch account.



Dependent Variable: ln(Followers)
Independent Variable: Mean weekly stream hours
ln(Followers) = 13.531163 - 0.0088658957 Mean weekly stream hours
Sample size: 900
R (correlation coefficient) = -0.23767042
R-sq = 0.056487228
Estimate of error standard deviation: 1.0527007

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	13.531163	0.05907867	$\neq 0$	898	229.03635	<0.0001
Slope	-0.0088658957	0.0012091584	$\neq 0$	898	-7.3322862	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	59.578371	59.578371	53.762421	<0.0001
Error	898	995.14449	1.1081787		
Total	899	1054.7229			

Table 4: `‘Followers’` regressed against `‘Mean weekly stream hours’`

5 Multiple Linear Regression

5.1 Mature Accounts and Followers

How much does the difference in watch time explain the difference in the number of followers for accounts categorized as mature or not mature? Twitch accounts classified as `‘Mature’` have fewer followers than those not classified as `‘Mature’`. At the same time, mature accounts are watched fewer hours per week (Figure 9).[5] While it is clear that mature accounts are watched less, it is interesting to ask how much that disparity in watch time can “explain”[6] the disparity in the number of followers.

[5] Mature accounts also tend to have fewer peak and average viewers, and they also have not gained as many followers over the period. See Figure 9.

[6] This is observational data, so causal claims cannot be made here.

5.1.1 Difference in Means

Since `‘Followers’` is highly skewed, the log of `‘Followers’` was taken. A difference in means of the log followers was calculated. (This also allows for a better “apples-to-apples” comparison in the subsequent multiple linear regression.) The difference in log means (-0.13351611) is exponentiated to get the ratio (0.875) between the two means (Equation 7). Thus, mature Twitch accounts have 12.5 percent fewer followers than accounts that are not classified as mature.

$$\begin{aligned}\mu_1 &= \text{Mean of } \ln(\text{Followers}) \text{ where Mature}\\ \mu_2 &= \text{Mean of } \ln(\text{Followers}) \text{ where not Mature}\\ \mu_1 - \mu_2 &= -0.13351611\\ e^{-0.13351611} &= 0.875\end{aligned} \tag{7}$$

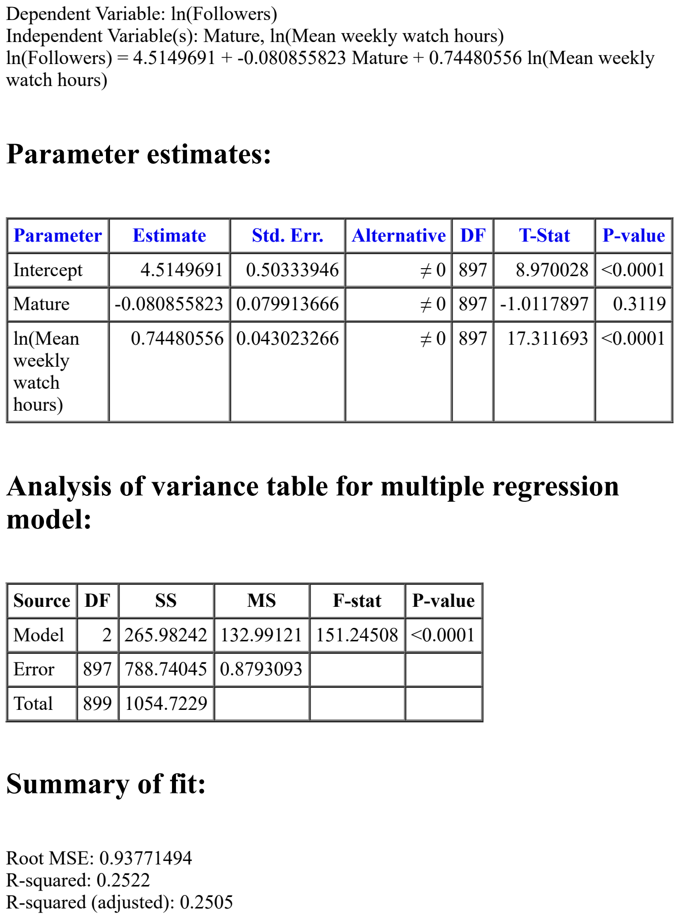
5.1.2 Model Specificiation

Next a multiple linear regression model was run to assess how much differences in the number of watch hours can explain the differences in the number of followers between mature and non-mature accounts. Log transformations on both the independent (`‘Mean weekly watch hours’`) and dependent variable (`‘Followers’`) were conducted to correct for the non-linear nature of the relationship. The model is as follows:

$$ln(Followers_i) = \beta_0 + \beta_1 * Mature_i + \beta_2 * ln(Mean\ weekly\ watch\ hours_i) \tag{8}$$

where i is each Twitch account in the dataset; `‘Followers’` is the number of followers; `‘Mature’` is a binary variable where `1` is when the account has been classified as mature, and `0` is when it has not; `‘Mean weekly watch hours’` is the total watch time in minutes for a Twitch account divided by 60 and divided by 52.

5.1.3 Multiple Linear Regression Results



Dependent Variable: ln(Followers)
Independent Variable(s): Mature, ln(Mean weekly watch hours)
ln(Followers) = 4.5149691 + -0.080855823 Mature + 0.74480556 ln(Mean weekly watch hours)

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	4.5149691	0.50333946	≠ 0	897	8.970028	<0.0001
Mature	-0.080855823	0.079913666	≠ 0	897	-1.0117897	0.3119
ln(Mean weekly watch hours)	0.74480556	0.043023266	≠ 0	897	17.311693	<0.0001

Analysis of variance table for multiple regression model:

Source	DF	SS	MS	F-stat	P-value
Model	2	265.98242	132.99121	151.24508	<0.0001
Error	897	788.74045	0.8793093		
Total	899	1054.7229			

Summary of fit:

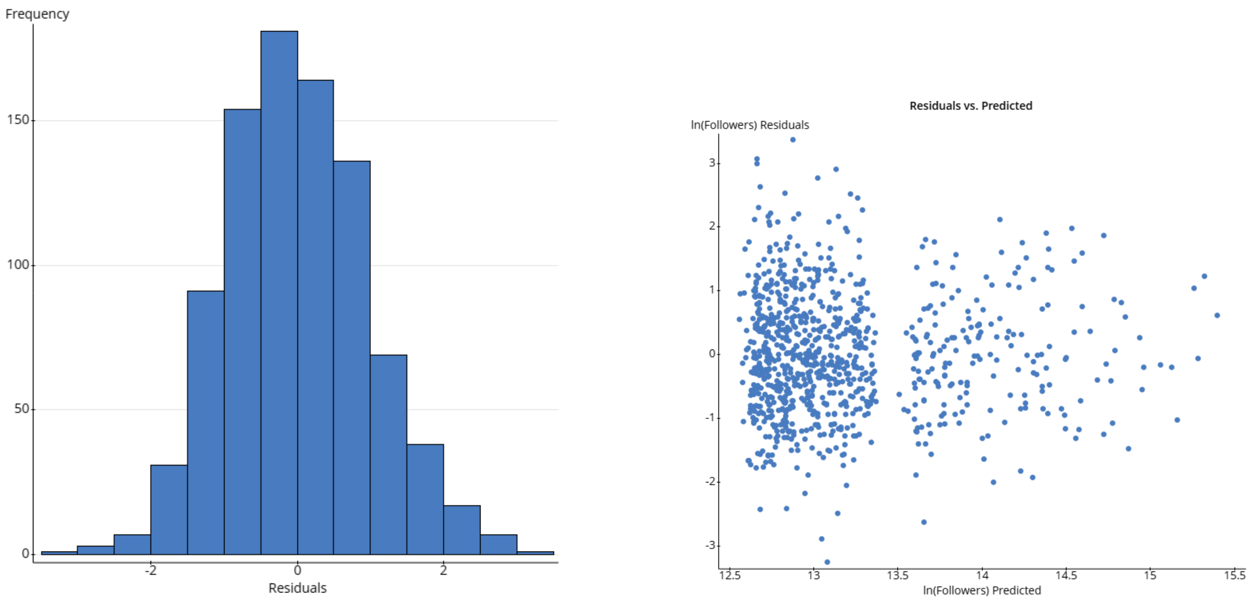
Root MSE: 0.93771494
R-squared: 0.2522
R-squared (adjusted): 0.2505

Table 5: Followers regressed against Mature and Mean weekly watch hours

R-squared is rather weak (0.25), but the difference in watch time (mean weekly watch hours) does explain some of the differences observed between mature and non-mature accounts. The coefficient for `'Mature'` is -0.080855823. Exponentiating the coefficient ($e^{-0.080855823}$) results in 0.9223. This is the ratio of followers for mature accounts to non-mature accounts when the natural log of mean weekly watch hours is held constant. In other words, mature accounts have 7.7 (1- 0.9223) percent fewer followers than non-mature accounts when controlling for differences in watch time.

5.1.4 Discussion

The results suggest that the difference between the number of followers that mature and non-mature accounts have is only slightly "explained" by differences in the number of



(a) Histogram: residuals are nearly normal

(b) No non-linear trends detected in the data

Figure 7: Plots of residuals for the multiple linear regression model in Equation 8 and Table 5.

mean weekly watch hours between the two types of accounts. In a way, this is not totally surprising. Figure 1 shows that there is only a moderate correlation between watch time and the number of followers an account has. It could be that there is some other unobserved variable out there that explains the difference in followers, or, perhaps just as likely, mature accounts are less popular in general and struggle to gain as many followers. Of course, identifying a mechanism like this is beyond the scope of this paper, as that would require more data.

6 Logistic Regression

6.1 Predicting Partnered Status

Which variable in the dataset is the best predictor of whether an account is partnered? To answer this, I ran a logistic regression on each of the predictor values in the dataset. Using the log-likelihood from the output, I calculated the Akaike information criterion (AIC), which estimates the quality of a statistical model relative to other models. The formula for AIC is as follows:

$$AIC = 2k - 2 * ln(Likelihood) \tag{9}$$

where k is the number of parameters. When comparing models with the same number of predictor variables, k remains constant; thus, the model with the greatest ln(likelihood) is the model with the lowest AIC. The model with the smallest AIC is usually preferred over more complicated models unless there is a good theoretical reason to include those additional variables.[7]

6.1.1 AIC Values

After running each model, `‘Followers’` was the best model to predict `‘Partnered’` status (AIC = 4 - 2 × -116.78416 = 237.5683).

AIC values for other models with a single predictor variable:

Variable(s)	AIC
`‘Peak viewers’`	245.79
`‘Average viewers’`	246.42
`‘Mature’`	246.22
`‘Mean weekly watch hours’`	245.77
`‘Mean weekly stream hours’`	246.36

Adding additional variables beyond `‘Followers’` did not improve the AIC:

Variable(s)	AIC
`‘Followers + Peak viewers’`	238.80
`‘Followers + Average viewers’`	237.67
`‘Followers + Mature’`	239.07
`‘Followers + Mean weekly watch hours’`	239.07
`‘Followers + Mean weekly stream hours’`	238.32

Thus, I opted to only use `‘Followers’` to predict partnered status in my logit model.

[7] To be clear, one should be guided by theory first; AIC is only a tool.

6.1.2 Model Specification

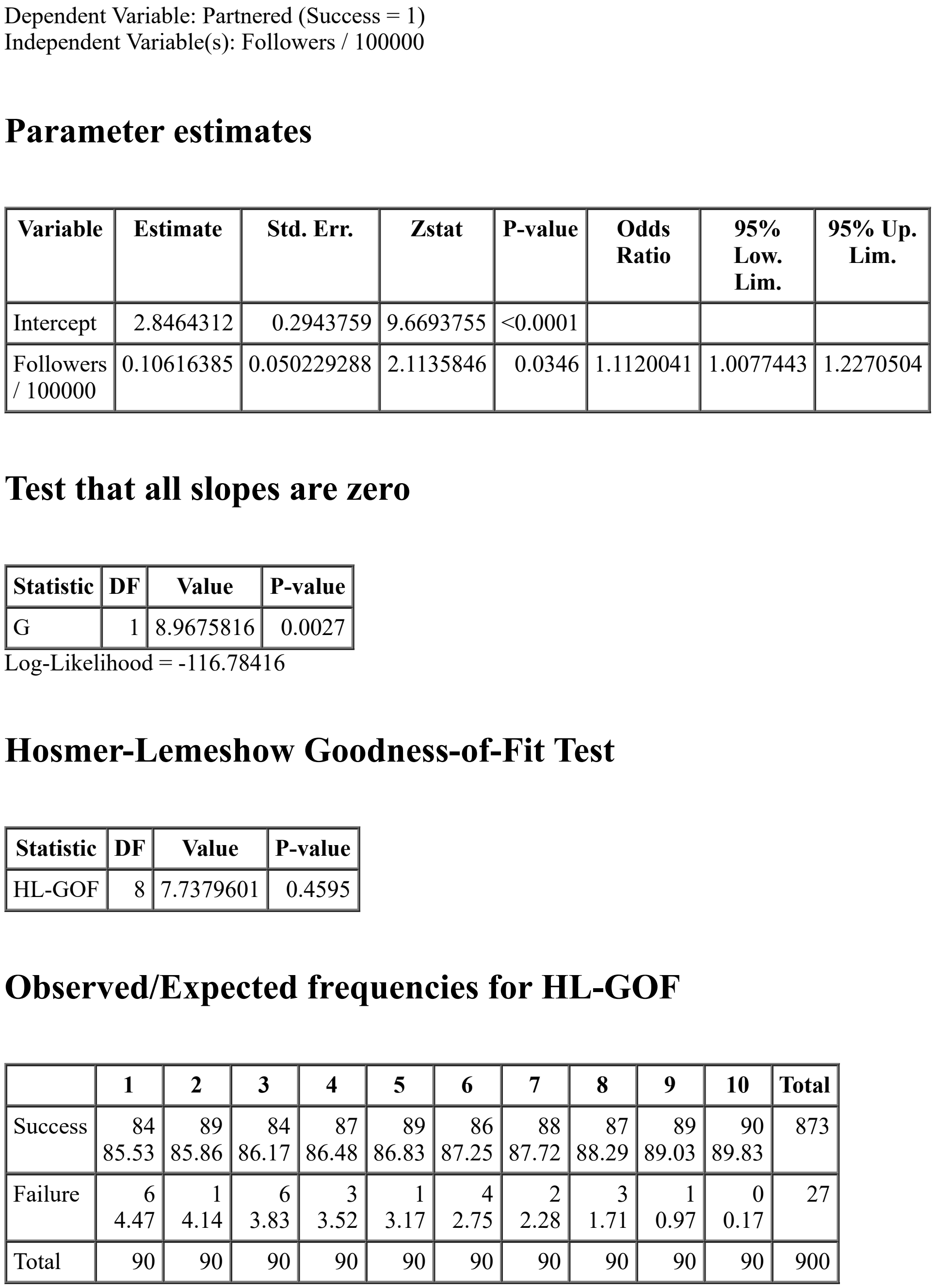
I decided to rescale `‘Followers’` by dividing it by 100,000 so that the new variable is 100,000 followers. This will make the coefficient easier to interpret.

$$\begin{aligned} ln(\frac{p_i}{1-p_i}) &= logit(Partnered_i) = \beta_0 + \beta_1 * Followers/100,000_i \\ \frac{p_i}{1-p_i} &= odds \end{aligned} \tag{10}$$

where i is a Twitch account; `‘Partnered’` is a binary variable indicating whether the respective account has attained partnered status or not; `‘Followers / 100,000’` is the number of followers that an account has rescaled to 100,000 followers; p is the probability of an account being achieving partnered status.

6.1.3 Logistic Regression Results

The logistic regression shows a positive relationship between the number of followers a Twitch account has and the log odds of being a partnered account (Table 6). Exponentiating the coefficient for `‘Followers / 100,000’` ($e^{0.10616385}$) results in 1.112, which means that for every 100,000 followers that an account has, the odds of being a partnered account increase 11.2 percent.



Dependent Variable: Partnered (Success = 1)
Independent Variable(s): Followers / 100000

Parameter estimates

Variable	Estimate	Std. Err.	Zstat	P-value	Odds Ratio	95% Low. Lim.	95% Up. Lim.
Intercept	2.8464312	0.2943759	9.6693755	<0.0001			
Followers / 100000	0.10616385	0.050229288	2.1135846	0.0346	1.1120041	1.0077443	1.2270504

Test that all slopes are zero

Statistic	DF	Value	P-value
G	1	8.9675816	0.0027

Log-Likelihood = -116.78416

Hosmer-Lemeshow Goodness-of-Fit Test

Statistic	DF	Value	P-value
HL-GOF	8	7.7379601	0.4595

Observed/Expected frequencies for HL-GOF

	1	2	3	4	5	6	7	8	9	10	Total
Success	84 85.53	89 85.86	84 86.17	87 86.48	89 86.83	86 87.25	88 87.72	87 88.29	89 89.03	90 89.83	873
Failure	6 4.47	1 4.14	6 3.83	3 3.52	1 3.17	4 2.75	2 2.28	3 1.71	1 0.97	0 0.17	27
Total	90	90	90	90	90	90	90	90	90	90	900

Table 6: Logistic Regression Table. The results indicate a positive relationship between the number of followers an account has and whether it has achieved partnership status.

6.1.4 Figures

A scatter plot was generated to visualize the relationship. Since all values for `'Partnered'` take either a zero or one, random noise using the normal distribution simulation function centered on zero with a standard deviation of 0.01 — $\mathcal{N}(0, 0.01)$ — was generated in StatCrunch to make visualization clearer. Then, the absolute value of that simulated normal distribution was either added or subtracted from the `‘Partnered’` value depending on if the value was a one or zero, respectively.

The equation representing the generation of the `‘PartneredNoise’` variable is:

$$\text{PartneredNoise} = \begin{cases} \text{Partnered} + |\text{Normal}| & \text{if Partnered} = 1 \\ \text{Partnered} - |\text{Normal}| & \text{if Partnered} = 0 \end{cases} \tag{11}$$

where `‘Partnered’` is the binary variable for whether an account is partnered; `‘Normal’` is the simulated normal distribution based on $\mathcal{N}(0, 0.01)$.

The model was exponentiated and plotted over the scatter plot. It represents the probability that an account will be partnered, as predicted by the number of followers that that account has.

$$\hat{p_i} = \frac{e^{2.8464312+0.10616385*(Followers/100{,}000)}}{1 + e^{2.8464312+0.10616385*(Followers/100{,}000)}} \tag{12}$$

This resulted in the following plot:

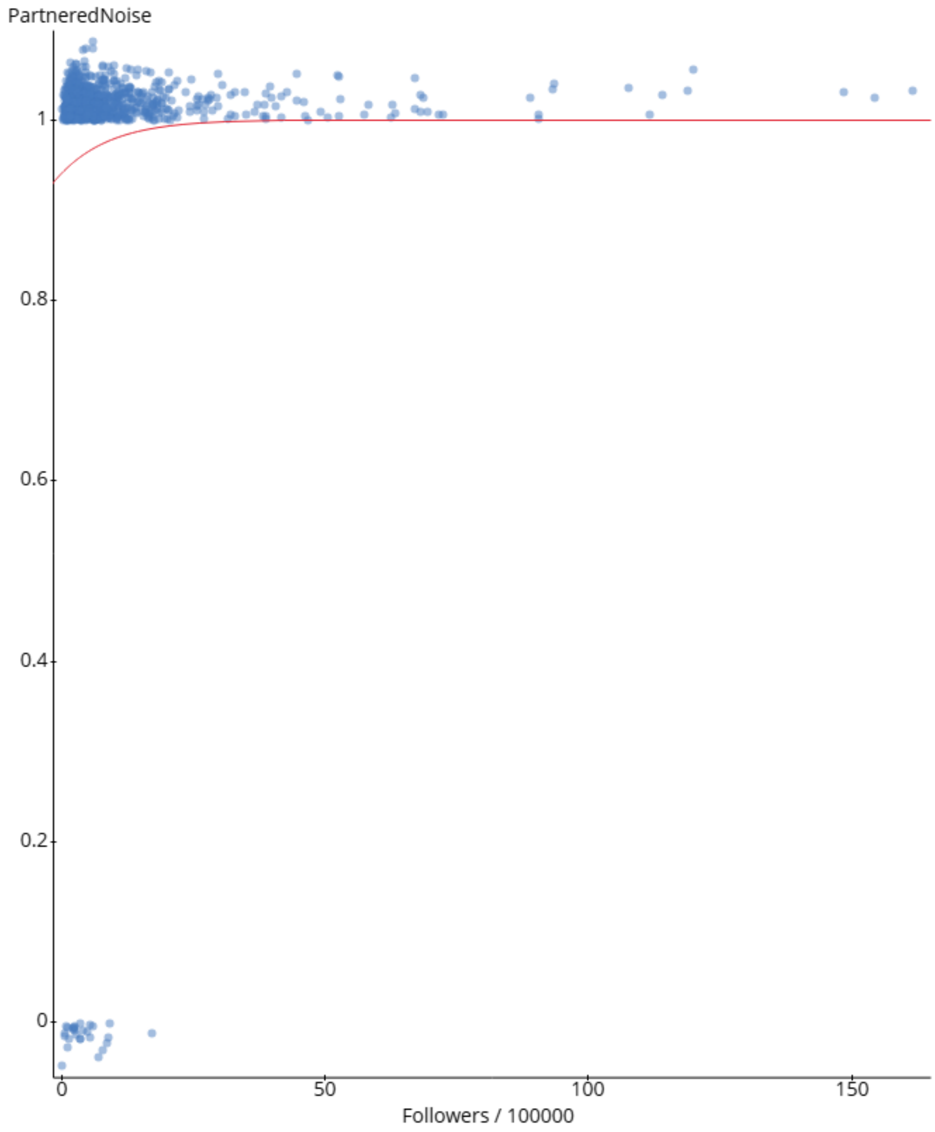


Figure 8: Scatter Plot: Partnered Status Predicted by Number of Followers.

6.1.5 Goodness of Fit

The Hosmer-Lemeshow Goodness-of-Fit observed vs. expected successes for partnered status suggest that, in general, the model is reasonably calibrated (Table 6). The Hosmer-Lemeshow Goodness-of-Fit Test assesses how accurately the model predicts the success rate in each decile of the independent variable (in this case, `'Followers'`). The model predicts successful partnered status reasonably well, but it is more limited in its ability to predict non-partnered status (failures). This could be because of the extreme class imbalance in the data (success-to-failure ratio of 873:27).

6.1.6 Discussion

The number of followers is the best predictor of the probability that an account will be partnered. However, this analysis is hampered because of the severe class imbalance in the data. That notwithstanding, the disparity in followers between partnered and non-partnered accounts is striking. 123 of the 873 partnered accounts have more followers than the maximum number (1714324) of followers that non-partnered accounts have.[8] Future analyses could randomly sample Twitch data using a blocking technique to ensure that success and failure conditions are more equally represented in the data.

7 Conclusion

The Twitch dataset analysis stands out primarily due to the unexpected findings it yielded. There is practically no relationship between the amount of time that an account streams and the amount of time that the account is watched. And contrary to what one might expect, streaming more frequently has a strong *negative* relationship with the average number of viewers an account has. In short, if you stream more, fewer people are watching you (See Figure 1 and 3.1: Stream Hours to Predict Average Viewers). Similarly, the number of stream hours has a negative relationship with the number of followers that an account has (see 3.2: Stream Hours to Predict Followers), though the relationship is relatively weak.

In general, Twitch accounts classified as mature had fewer followers than accounts not classified as mature; mature accounts are also watched less than non-mature accounts. While it was hypothesized that the disparity in watch time could explain the difference in the number of followers, the multiple linear regression model (4.1: Mature Accounts and Followers) revealed only a small R-squared, suggesting that there are probably other unobserved variables influencing the differences in the number of followers.

Lastly, whether an account is partnered on Twitch is best predicted by the number of followers it has. Additional variables did not add enough predictive power to warrant their inclusion in the model. However, this analysis is hampered by the extreme class imbalance in partnered vs. non-partnered status. Future research could randomly sample while maintaining a better class balance to assess which variables best predict partnered status.

[8] Author's calculations in StatCrunch.

8 Appendix

8.1 Additional Figures

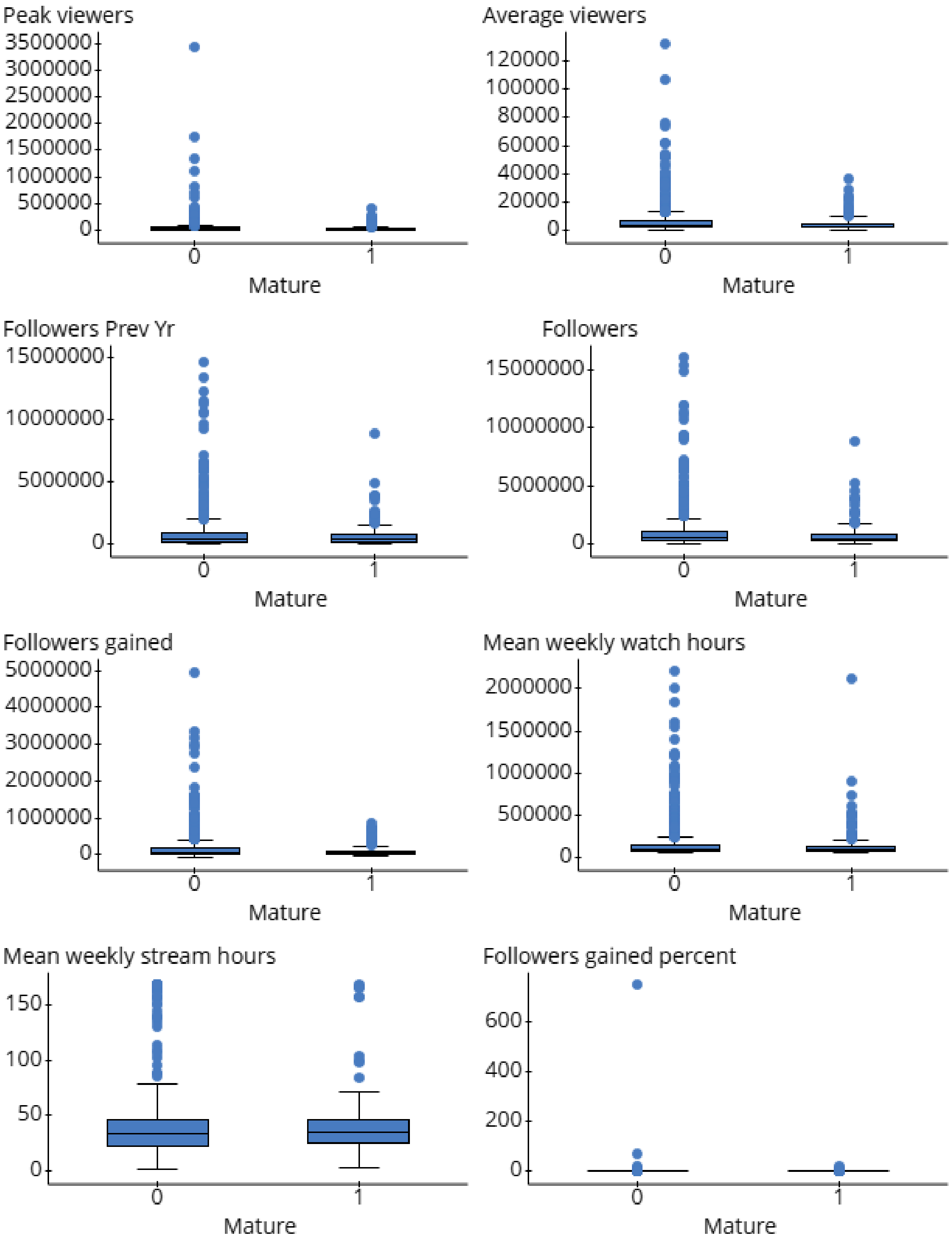


Figure 9: Box Plots of Twitch Accounts by Mature Calssification

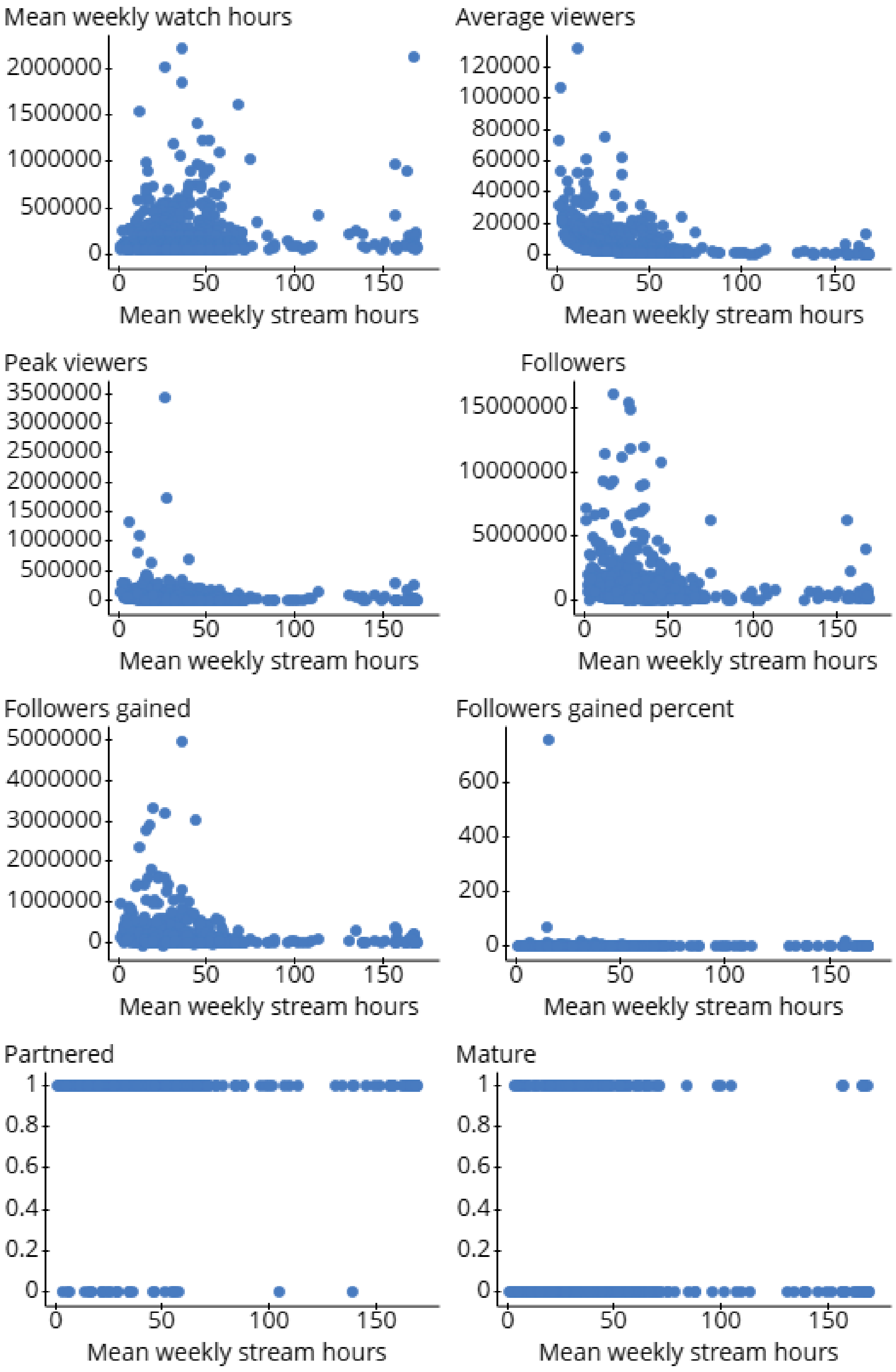


Figure 10: Relationships between '`Mean weekly stream hours`' and other variables.

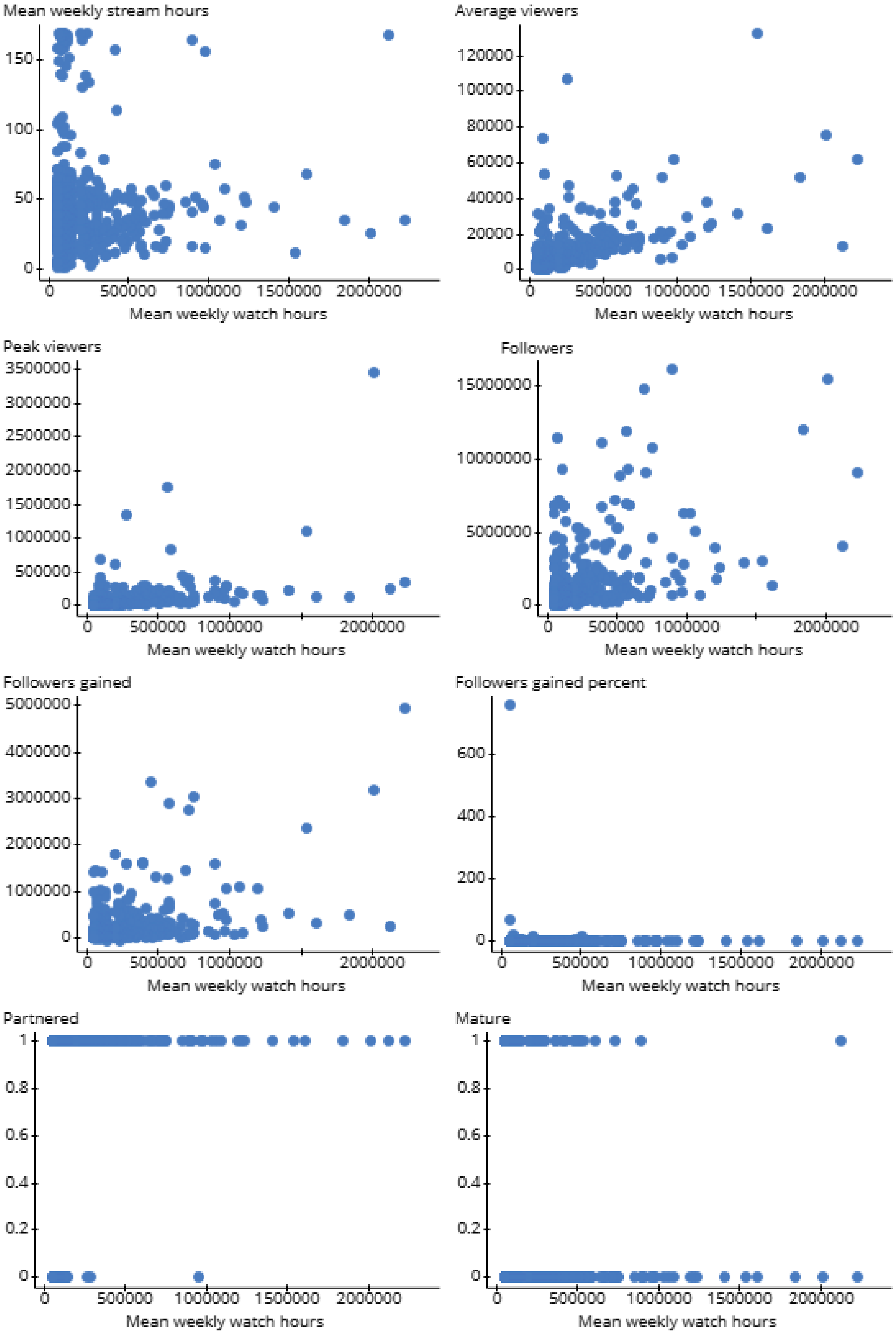


Figure 11: Relationships between '`Mean weekly watch hours`' and other variables.

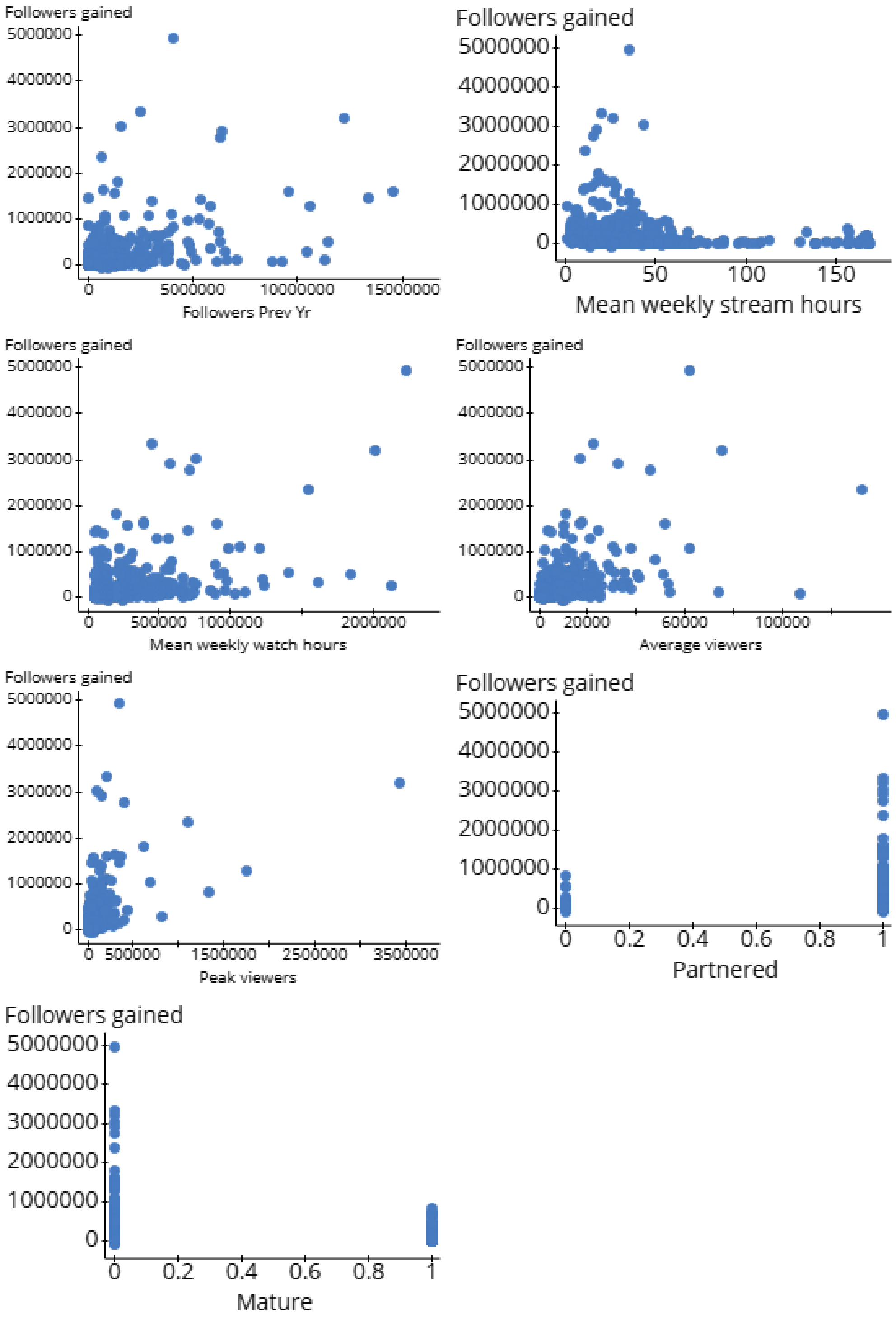


Figure 12: Relationships between ‘`Followers gained`’ and other variables.

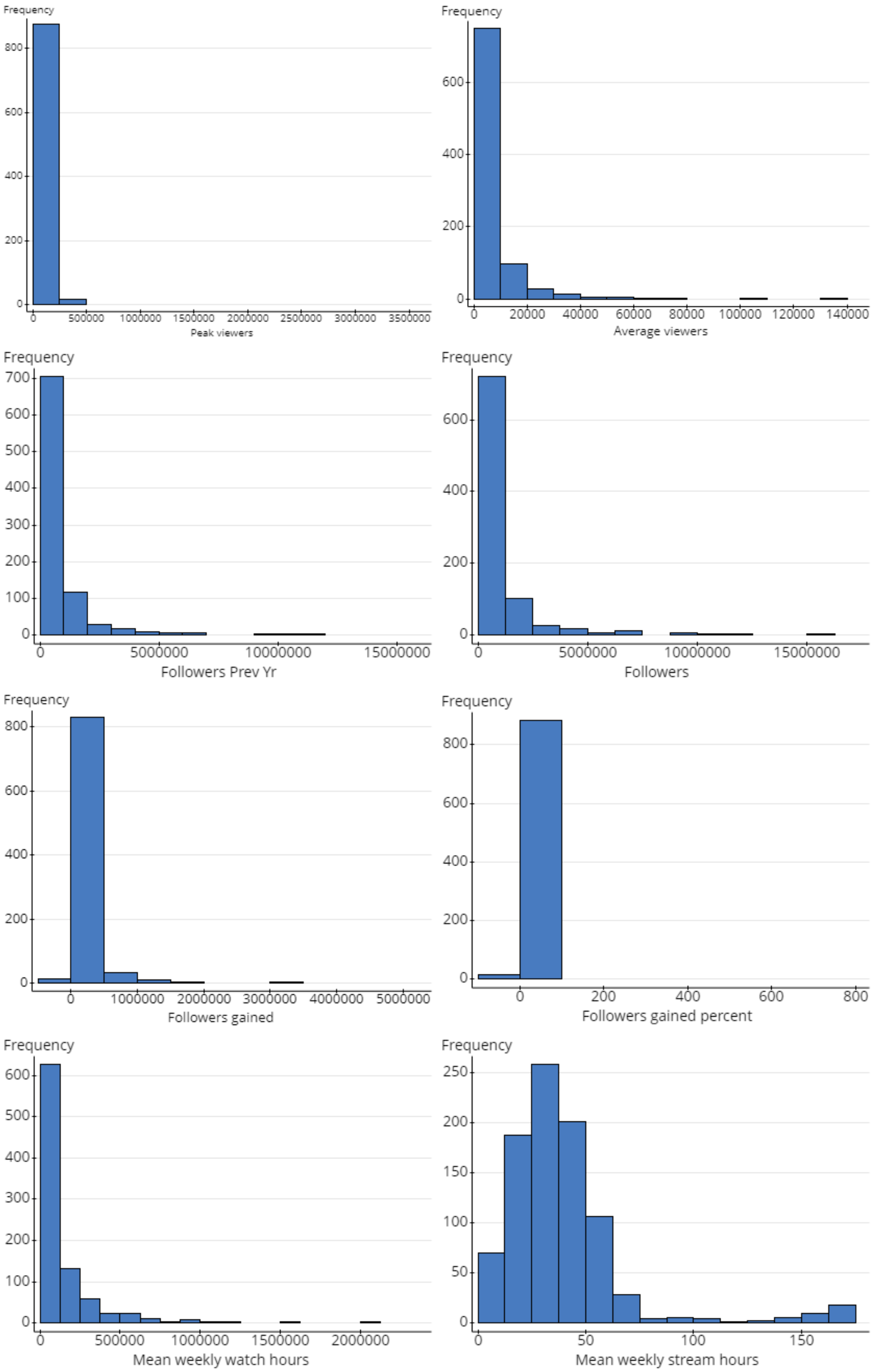
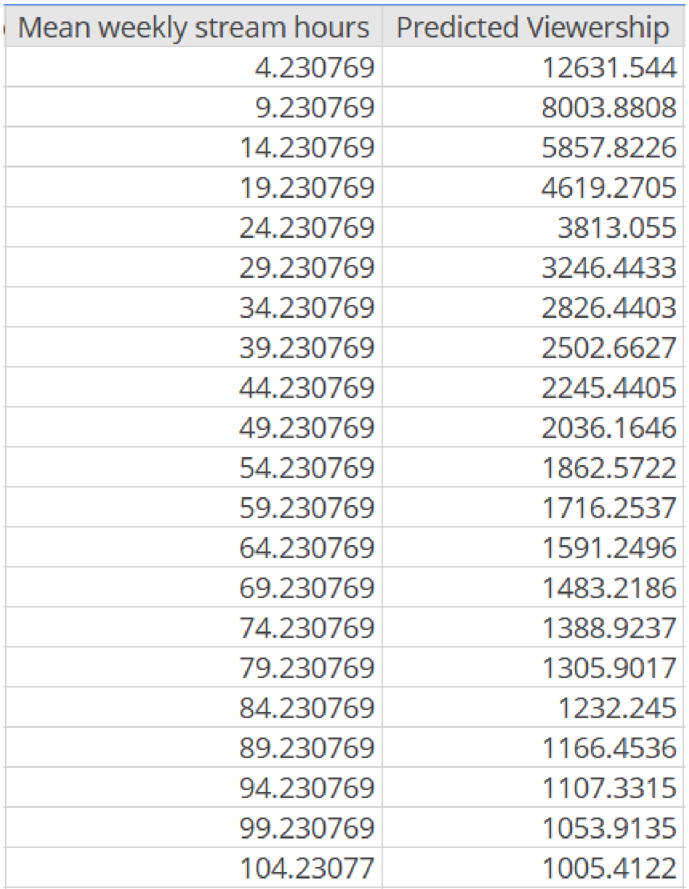


Figure 13: All distributions are heavily skewed and non-normal.



Mean weekly stream hours	Predicted Viewership
4.230769	12631.544
9.230769	8003.8808
14.230769	5857.8226
19.230769	4619.2705
24.230769	3813.055
29.230769	3246.4433
34.230769	2826.4403
39.230769	2502.6627
44.230769	2245.4405
49.230769	2036.1646
54.230769	1862.5722
59.230769	1716.2537
64.230769	1591.2496
69.230769	1483.2186
74.230769	1388.9237
79.230769	1305.9017
84.230769	1232.245
89.230769	1166.4536
94.230769	1107.3315
99.230769	1053.9135
104.23077	1005.4122

Table 7: Predicted Values.

8.2 StatCrunch Dataset

https://www.statcrunch.com/app/index.html?dataid=4597814&token=OTI3Z8%2FON6hSC1KVw9hTXmyjHLnuZvqCMyuxkgn1QRYPhIQfDVLUFClF3Y41ShOi4C%2BMKL5%
2FHgpBTXKukjWOPGD4pN%2FCkiobeyKouIjPB7LoPvHOTDN7wUNtPZQd2%2BNjOwAtSMQl9aKQrbthjCSuuSihsliiToOMvakPDYNOlwiE8N11ITBSTS9QJH9QgHEmO4ahoV6IkASuVdKosV%2FJSQ%
3D%3D