Critics: Study of TV Pilots

A. L. S.

May 31, 2015

1 Introduction

Two people developed a quantitative method to determine which TV shows to watch. The goal was to watch a battery of shows, score each on a variety of criteria or features, and then watch the entire season of the top six shows. Fourteen TV pilots, both current and older shows, were watched and scored them based on the following five features:

- Acting
- Characters
- Curiosity
- Originality
- Script

For this round of pilots, the scoring metric was a range of 0 to 10 for each feature. A previous round only used scores between 0 and 5, which was found to be an insufficient range. Scores were immediately tabulated after viewing each pilot and were kept secret until all pilots were scored. This report first discusses the total score, summed over all features, followed by analysis of the features separately. Then the methods to tally the results to determine the top six shows are investigated, and correlations among features are explored.

The combined scores for each show in this round as a function of critic are shown in Figure 1. One critic, Avery, scored almost every show higher than the other critic, Anne (exception: House of Lies) and also had a much more narrow spread of combined scores: 11 %, as compared to 23 % (from the standard deviations). A breakdown of the spread in scores for each show is shown in Figure 2. This illustrates variability among features for a given show and critic. The whiskers represented by points indicate major outlying features. Some boxes are represented only by a short vertical line, as there was no spread in their scores. This is observed more often in Avery's scores, as his overall spread is more narrow than those of Anne.

1.1 Hypothesis Testing to Investigate Differences in Critics

In order to determine if the average scores per critic were significantly different from a statistical point of view, several hypothesis tests were used. This is an indication that the critics were using the scoring scale in different ways, without equally distributing scores across the range.

The first two hypothesis tests assumed the sets of data from each critic were not paired by show, and simply investigated the differences in the distributions of total scores. Details on the calculations are discussed in Appendix A. The null hypothesis assumed to be true, H_0, is that the differences in total scores per critic are *not* statistically significant. That is, both critics utilized the same range of allowed scores. The alternative hypothesis, H_α, is that they did not:

$$\begin{aligned} H_0 &: x_{\text{Avery}} = x_{\text{Anne}} \\ H_\alpha &: x_{\text{Avery}} \neq x_{\text{Anne}} \end{aligned} \tag{1}$$

Keywords: TV shows, statistics, critics

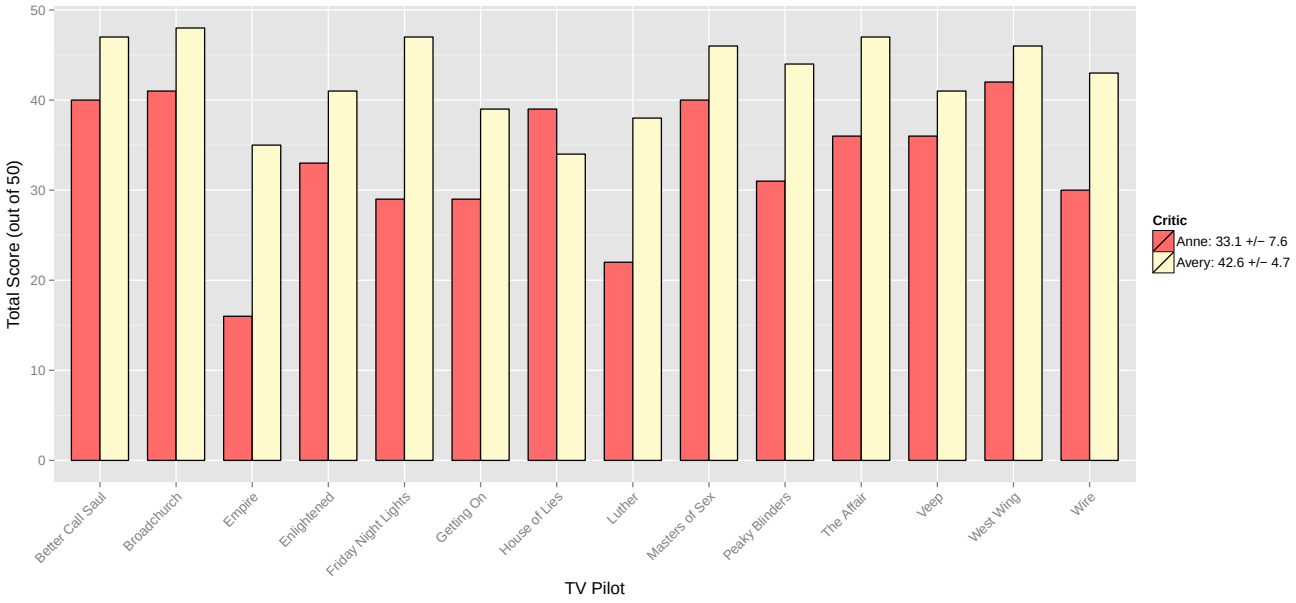

Figure 1: Total scores as a function of TV pilot. The average score and its associated standard deviation are shown in the legend.

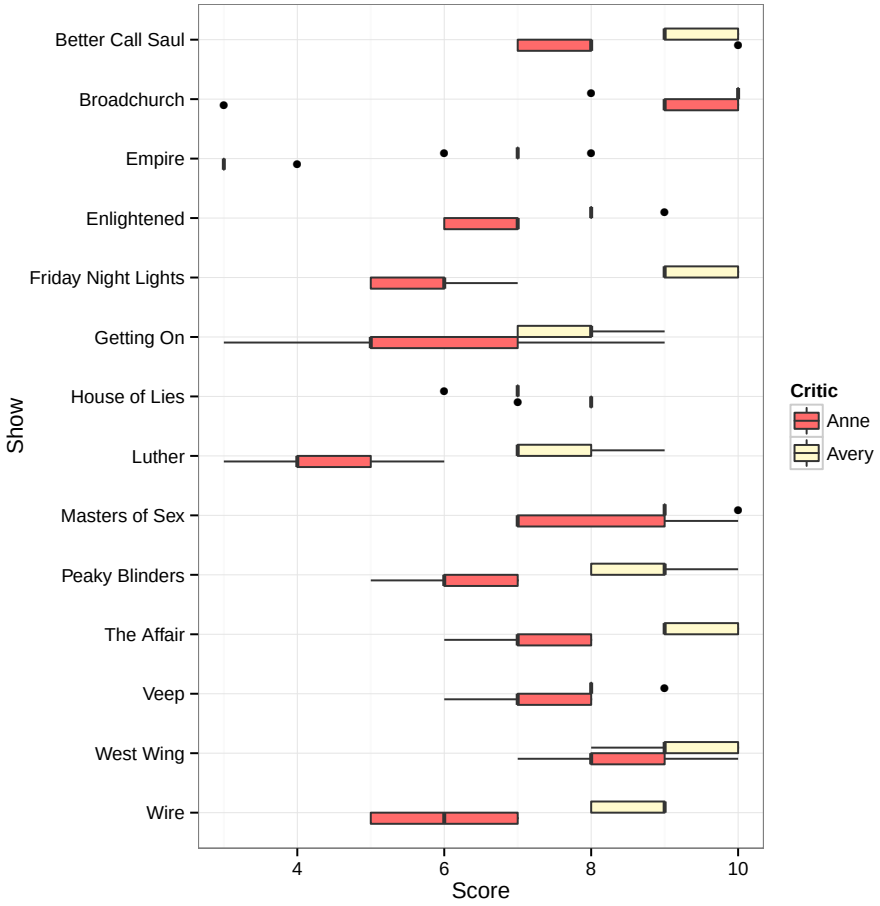

Figure 2: Distribution of scores across features as a function of TV pilot.

where x_i is the average total score per critic i. In order to reject the null hypothesis, the p-value from each test should be less than the significance level of, say, 0.05. That would indicate that, assuming the null hypothesis is actually true, there would only be a 5 % chance that it would be falsely rejected. This is also known as the type I error rate. An equally valid way to test if the differences are significant is to calculate the 95 % confidence interval and observe if it includes zero, indicating that there is zero difference in the average scores with a confidence of 95 %. In this case, within statistical uncertainty, the average scores from each critic would be the same.

Assuming the variances for the two sets of data were unequal, a p-value of 7×10^{-4} is calculated, and the confidence interval for the difference in scores is 4.5 to 14.4, with a 95 % confidence. Because this p-value is much smaller than 0.05 and because the confidence interval does not include zero, the null hypothesis can be rejected, and the average combined scores per critic are statistically quite different from one another. Using an average variance between the sets can be useful if the number of data points in each set is vastly different. In this case, the number of shows per critic is identical, and executing a hypothesis test with this average variance yields very similar results (see Appendix A for details).

Because the scores from each critic are paired to a specific show, a paired hypothesis test can also be performed. Each of Anne's scores is subtracted from the respective values from Avery's scores for the same show, resulting in a single set of data. For this test, the p-value is even smaller at 9×10^{-5}. The confidence interval of 5.8 to 13.1 still does not encompass zero. Again, the null hypothesis is rejected. Although it is possible to pair the results per show in this way, it is more informative to use the non-paired, aggregated data to investigate whether or not the use of the scoring metric was different, as personal preference would skew the results of a single show and not necessarily be an indicator that the use of the scoring system by each critic was significantly different.

1.2 Distribution of Scores for Features

Different ways of displaying the distributions of scores across the five features are shown in Figures 3 and 4. Average scores for each feature are shown in Table 1. Similar to the total scores, the distribution of scores were different depending on the critic. However, the distribution across features of one critic appear to be quite homogenous. For Anne's scores across features, most distributions are mostly Gaussian-like, with the exception of Curiosity, which is skewed toward the higher scores. For Avery's scores, they could appear half-Gaussian, but the ceiling of the score (a value of 10) cuts off any evidence of possibly symmetric behavior.

Table 1: Average scores in each feature for each critic. Uncertainties are standard deviations.

Feature	Anne	Avery
Acting	7.0 ± 1.6	8.9 ± 0.9
Curiosity	6.8 ± 2.3	8.6 ± 1.2
Script	6.5 ± 2.0	8.3 ± 1.3
Characters	6.7 ± 1.6	8.6 ± 1.0
Originality	6.1 ± 2.1	8.1 ± 1.0
Overall	33.1 ± 7.6	42.6 ± 4.7

To check whether the distributions across features were significantly different within a critic, a matrix of hypothesis tests were run, pairing each feature with the others, for each critic. The null hypothesis, similar to the total scores, was to assume there is no difference in the distributions between two sets of features. The alternative hypothesis was that there is a significant difference. Results are shown in Figure 5 for each feature pairing and critic. Values along the diagonal are all equal to 1.0, when the same features are compared (i.e., there is zero probability that the data sets are different). The matrix is also symmetric since p-values are insensitive to reversing the order of the feature pair.

For Anne, none of the feature pairings produced a p-value smaller than the significance level of 0.05, with a minimum p-value of 0.23 for Acting verses Originality. In this critic's case, the null hypothesis cannot be rejected for any feature pairing, and thus, no distributions are significantly different. On the other hand, the null hypothesis *is* rejected for the pairing of Originality and Acting for Avery with a p-value of 0.035, though the p-values for the rest of the pairings were all above 0.05. It might be expected that the features of Originality and Acting would reject the null hypothesis because those distributions appear the most different in Figure 4.

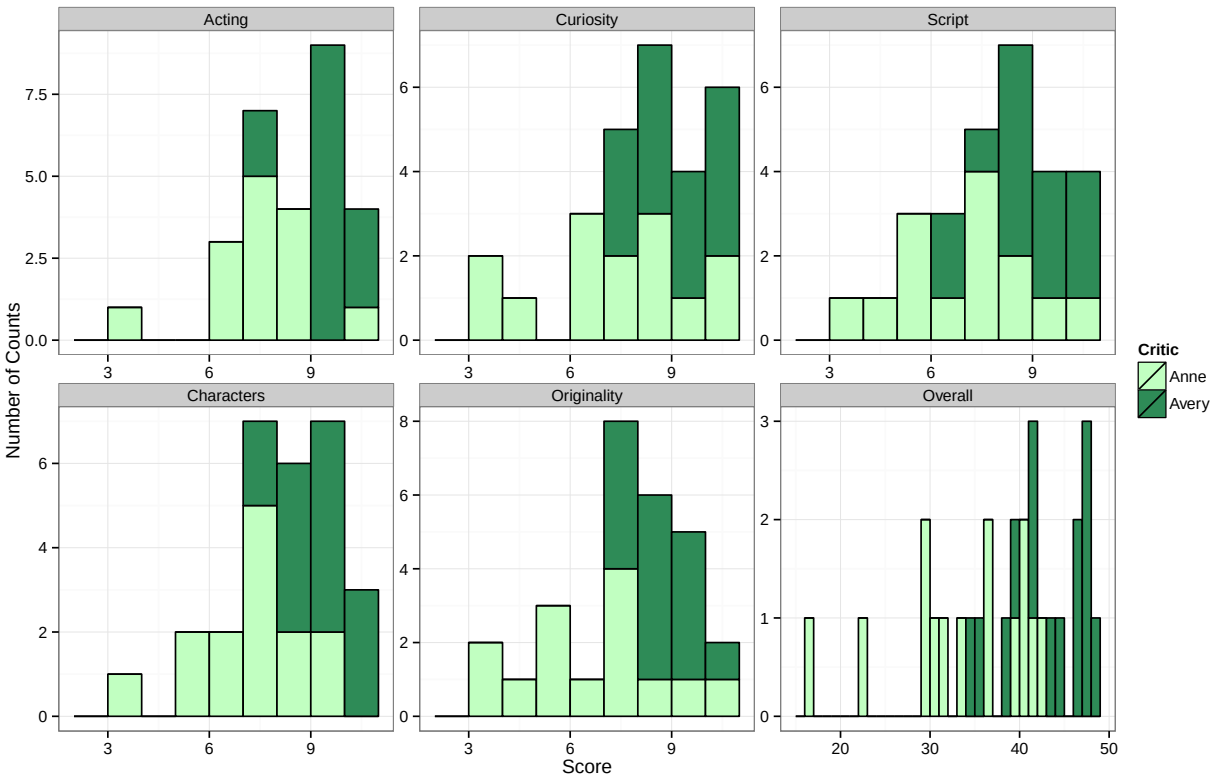

Figure 3: Distributions of scores as a function of feature. Color denotes critic.

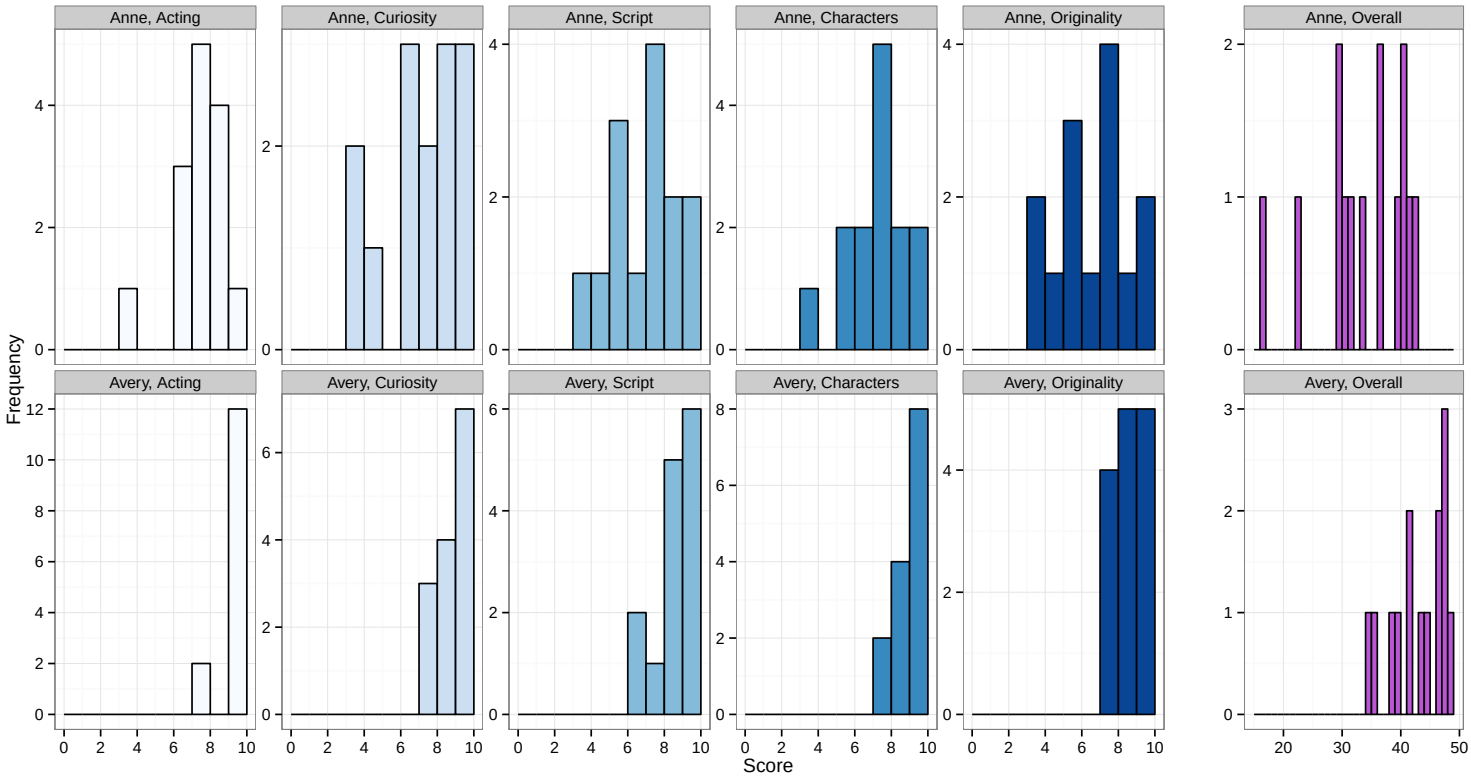

Figure 4: Distributions of scores as a function of feature, with rows split by critic.

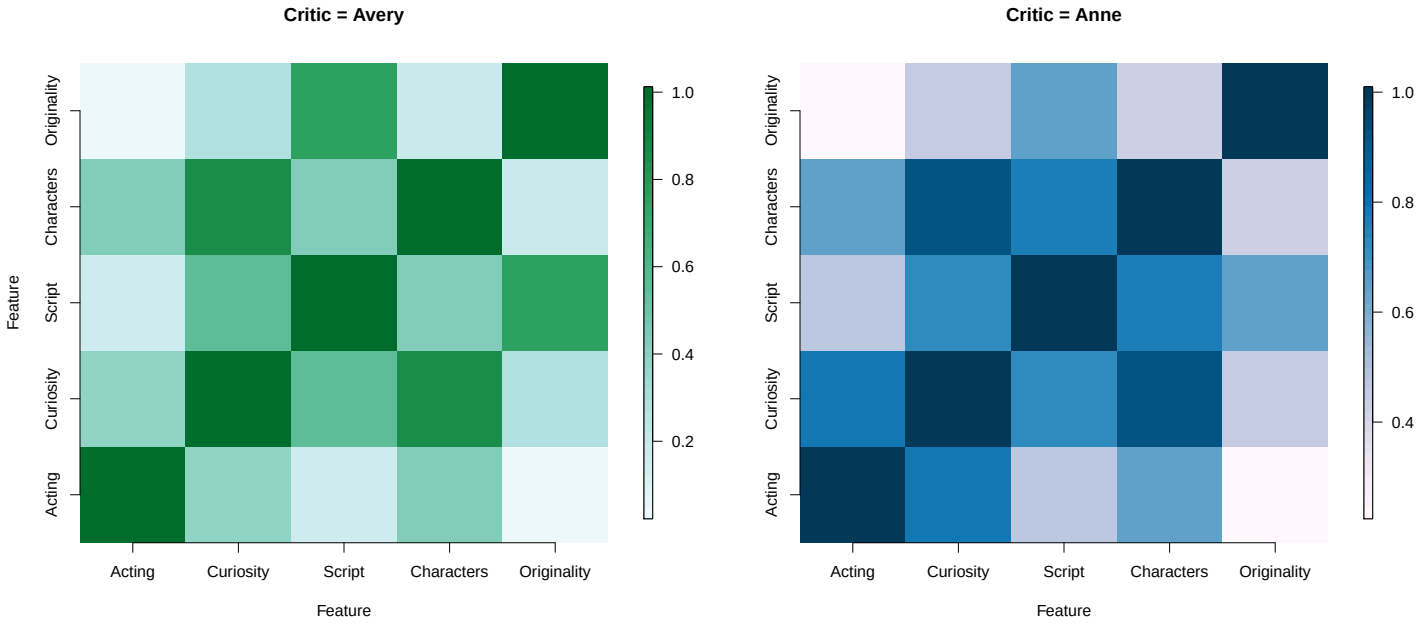

Figure 5: P-values as a function of feature pairing for each critic. Only the case of Originality compared with Acting for Avery is below the significance level of 0.05.

2 Tallying the Scores

Because the critics' use of the scoring metric and the subsequent spread in data was so vastly different, in order to determine which shows constituted the "top six", two methods were employed and are compared in Table 2. Features were each given equal weight. Analyzing this data using different weights is discussed in Section 3.

Table 2: Total scores and ranks for each show, by critic. The total is the sum of the individual feature scores, and the mean rank is the average of the ranks given by each critic. Data is sorted in order of ascending mean rank. The shows above the dashed line were selected.

	Anne		Avery			Mean
Show	Score	Rank	Score	Rank	Total	Rank
Broadchurch	41	2	48	1	89	1.5
Better Call Saul	40	4	47	2	87	3.0
West Wing	42	1	46	6	88	3.5
Masters of Sex	40	3	46	5	86	4.0
The Affair	36	6	47	3	83	4.5
Friday Night Lights	29	12	47	4	76	8.0
Peaky Blinders	31	9	44	7	75	8.0
Veep	36	7	41	10	77	8.5
Enlightened	33	8	41	9	74	8.5
Wire	30	10	43	8	73	9.0
House of Lies	39	5	34	14	73	9.5
Getting On	29	11	39	11	68	11.0
Luther	22	13	38	12	60	12.5
Empire	16	14	35	13	51	13.5

The first method was simply to add up the scores from each feature and select the top highest scoring shows to continue watching. However, this could potentially bias the results, as one critic continually scored shows quite high across the spectrum. The second method assigned a rank per critic based on descending score. Then the two ranks were averaged, as a way to normalize the scores and eliminate any dependence on different utilization of the scoring system. In many of Avery's cases, combined scores for multiple shows were sometimes identical, as he used a much narrower range of scoring. For shows with the same score, the critic split the degeneracies in order of personal preference. For this data set, both methods yielded similar results. In the case of the sixth and final show, there was a tie in the mean rankings between Friday Night Lights and Peaky Blinders. The critics then deferred to the total score to break the tie.

3 Outliers and Refinements to the Algorithm

The British show Broadchurch (Season 1), now available on Netflix, is an interesting outlier, as evidenced by its wide distribution of scores across features seen in Figure 4. Although the critics' rankings were similar (Anne: 2, Avery: 1), the spread in the data is quite different across the features. This ranking would also suggest that this was not Anne's favorite show of the group, which is incorrect. Table 3 illustrates the quantiles for this show. The median score for each feature is 9 or 10, depending on critic. This close matching of scores extends even to the first quartile. However, the minimum score for Anne, a 3 for Originality, diverges wildly from the rest of the feature scores. This is a show that depicts how a family and town cope with the violent loss of a child, which is not a new theme. It was even displayed in a show from the previous round of pilot scoring: AMC-turned-Netflix

Table 3: Quantiles for the distribution of scores across features for Broadchurch (Season 1).

	0.0	0.25	0.50	0.75	1.0
Avery	8	10	10	10	10
Anne	3	9	9	10	10

show The Killing. Thus, the scoring of a stellar show can be sullied by a low score in a single feature, one that may or may not have any bearing on the enjoyment of the show.

The issue of potentially weighting features then arises. The analysis presented in this report includes a uniform weight to each feature, but this may be an oversimplification. Originality, for example, may be irrelevant and could be assigned a smaller weight in comparison to other more important features. However, it is difficult to implement this in a broad stroke because a selling point of some shows may indeed be their originality. This is the case of a show included in the previous round of scoring, The Leftovers.

One may also argue that, regardless of whether some features are slightly more important than others, there is one that far outweighs the rest: Curiosity. A show may have the best acting available and award-winning writers, but if the show does not imbue the viewer with the innate desire to continue watching and discover what is beyond the next episode, then it is not worth watching.

3.1 Results with Weights

To test if weighting features affects the outcome (i.e., the top six selected shows), a relative weight of 2.0 was assigned to Curiosity and 0.5 to Originality, with respect to the other features (properly normalized). Although the values of the mean ranks changed (with Broadchurch's mean rank increasing from 1.5 to 1.0), the ordering of the top six shows did not change. The degeneracy of the mean rank of Friday Night Lights and Peaky Blinders (both with a previous mean rank of 8.0) was broken with the new weights: Friday Night Lights shifted to a mean rank of 7.0, and Peaky Blinders dropped to 8.5.

4 Correlations between Features

The interaction and correlation between features was also explored. Figure 6 shows the strong correlation between scoring of Acting and Script. This correlation is blind to critic, even though the spread of Avery's scores is narrower. This behavior could be expected, as acting ability is easier to exhibit with better source material.

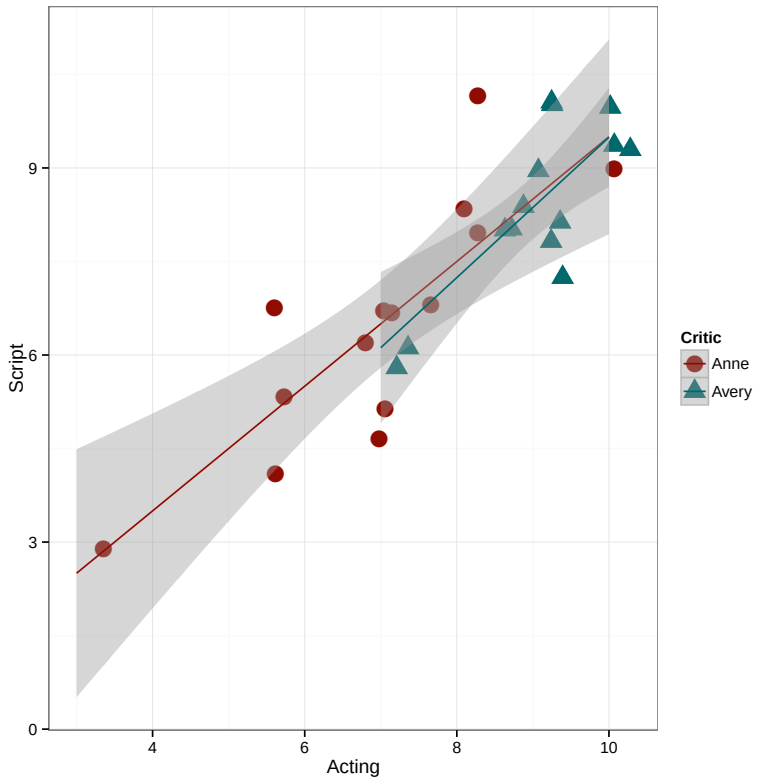

Figure 6: Script score as a function of Acting score for each critic. Points are shifted slightly to avoid overlap. Lines are a least squares fit of each data set, and the grey bands are 95 % confidence intervals.

Figure 7 show the paired correlations among all features. One can clearly observe the highly correlated nature of each pair, as data clusters around a 45° line. This is also expected: with a good script, the characters are better crafted. With better characters, curiosity increases. The flattest fit would be curiosity as a function of acting, though this response is driven by only one of the two critics. This would indicate that, even if the acting is top notch, a viewer may or may not want to see more, dependent upon other factors.

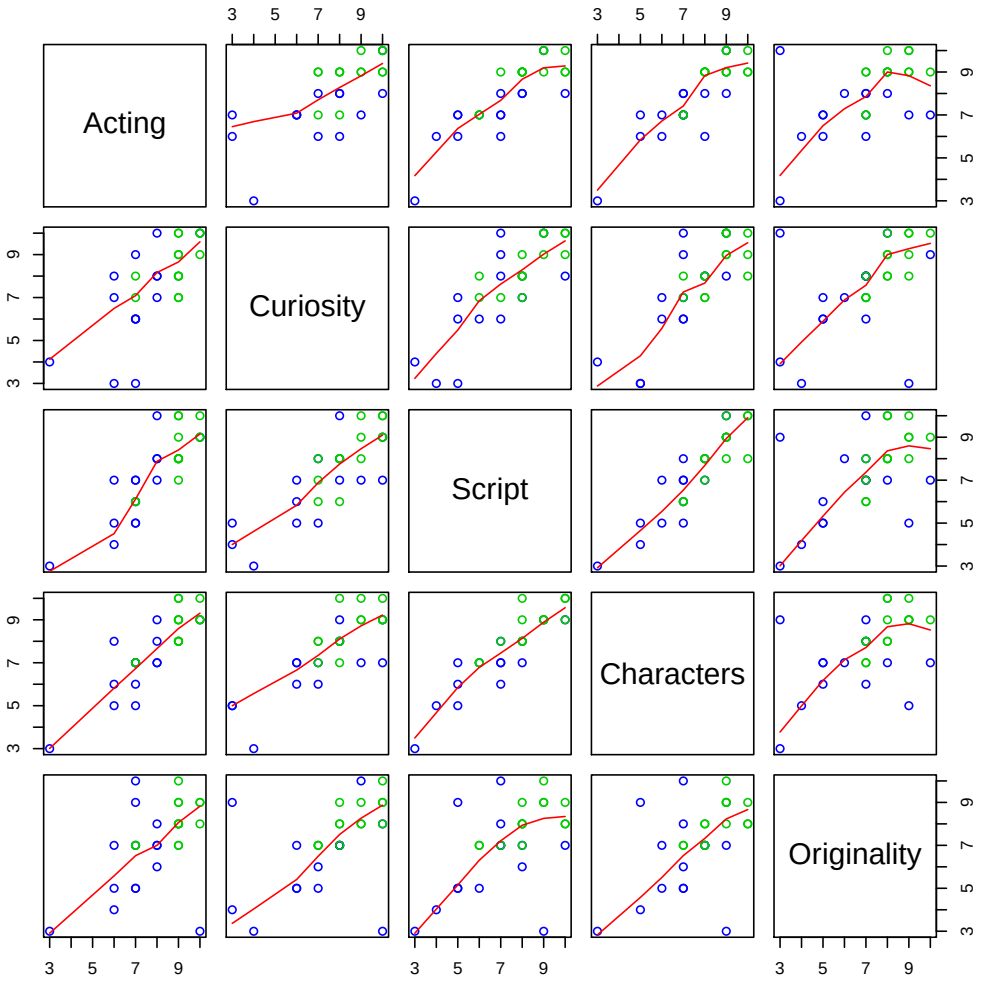

Figure 7: Correlations between each set of features. Color indicates critic: Avery–green, Anne–red. A lowess fit to data from both critics is shown in red. For each column, the x-axis is constant and given by the feature in the box, whereas the y-axis varies down the column by the feature indicated in each row. This structure is simply flipped if reading across rows.

5 Summary, Observations, and Future Prospects

A battery of TV pilots were watched and scored by two (novice) critics for five features. Although the allowed scoring range was identical, the distribution of scores varied between the two critics and were statistically significant, as shown by multiple hypothesis tests. This is an indication that the two critics were not utilizing the same range of scores, which could potentially change the top six shows. Distributions of scores across features was fairly constant within each critic, but quite different between critics. Because of this observed difference in critics, in order to determine the top six shows to continue watching, both methods of simply taking the cumulatively highest scores along with the highest average ranked shows were tested. For this small data set, there was no difference in the result for each method. Weighting of the importance of each feature was also examined, and the

results still reproduced the same selection of TV pilots. The features also appear to be highly correlated, as one would expect.

Potential challenges include scoring drift, whereby shows watched early in the data set may be scored more easily or more harshly as the critics work out their scoring systems. This systematic uncertainty would be very difficult to account for, though one might expect that as additional rounds of TV pilots are watched, potential fluctuations would decrease and begin to normalize. Of course, watching only the pilot for each show may also skew results, as more data is usually preferable before making a determination of score. However, the time to watch more than the pilot is prohibitive. Additional changes to future rounds could include increasing the number of pilots or expanding features to probe a deeper data space.

A Appendix

This appendix details the calculations used in Section 1.1.

A.1 Non-Paired Hypothesis Tests

For average combined scores, $\bar{x}_i$ with a standard error, SE and number of degrees of freedom, df, the test statistic, t_{stat}, p-value, and confidence interval, CI, are calculated as:

$$t_{stat} = \frac{\bar{x}_{\text{Avery}} - \bar{x}_{\text{Anne}}}{SE} \tag{2}$$

$$\text{p} - \text{value} = \texttt{pt}(t_{stat}, df, \text{lower.tail} = \text{F}) * 2 \tag{3}$$

$$\text{CI} = (\bar{x}_{\text{Avery}} - \bar{x}_{\text{Anne}}) \pm \texttt{qt}(0.975, df) * SE \tag{4}$$

For unequal variances, df and SE are determined via:

$$df = \frac{\frac{sd^2_{\text{Avery}}}{n_{\text{Avery}}} + \frac{sd^2_{\text{Anne}}}{n_{\text{Anne}}}}{\frac{(sd^2_{\text{Avery}}/n_{\text{Avery}})^2}{n_{\text{Avery}}-1} + \frac{(sd^2_{\text{Anne}}/n_{\text{Anne}})^2}{n_{\text{Anne}}-1}} \tag{5}$$

$$SE = \sqrt{\frac{sd^2_{\text{Avery}}}{n_{\text{Avery}}} + \frac{sd^2_{\text{Anne}}}{n_{\text{Anne}}}} \tag{6}$$

where n_i is the number of shows for each critic (14 for this data set), and sd_i is the standard deviation of each critic's total scores. These results are also calculated using the R `t.test` function: $\texttt{t.test}(x_{\text{Avery}}, x_{\text{Anne}})$ with x_i is a vector of scores for each critic.

For equal variances (i.e., calculating an average variance between the data sets), df and SE are computed as:

$$df = n_{\text{Avery}} + n_{\text{Anne}} - 2 \tag{7}$$

$$SE = \Big(\frac{(n_{\text{Avery}} - 1) * sd^2_{\text{Avery}} + (n_{\text{Anne}} - 1) * sd^2_{\text{Anne}}}{n_{\text{Avery}} + n_{\text{Anne}} - 2}\Big) * \sqrt{\frac{1}{n_{\text{Avery}}} + \frac{1}{n_{\text{Anne}}}} \tag{8}$$

Similarly, this can be calculated using $\texttt{t.test}(x_{\text{Avery}}, x_{\text{Anne}}, \text{var.equal} = \text{T})$.

A.2 Paired Hypothesis Tests

Instead of using separate vectors, x_i, for each critic the difference in the vectors, $x_{\text{diff}} = x_{\text{Avery}} - x_{\text{Anne}}$ is used, where the score for each show is subtracted pairwise. Then the following statistics are calculated:

$$t_{stat} = \frac{\bar{x}_{\text{diff}}}{SE} \tag{9}$$

$$\text{p} - \text{value} = \texttt{pt}(t_{stat}, df, \text{lower.tail} = \text{F}) * 2 \tag{10}$$

$$\text{CI} = \bar{x}_{\text{dif}} \pm \texttt{qt}(0.975, df) * SE \tag{11}$$

The degrees of freedom and standard error are:

$$df = n - 1 \tag{12}$$

$$SE = \frac{sd_{\text{diff}}}{n} \tag{13}$$

where sd_{diff} is the standard deviation for x_{diff}, and n is the number of paired shows (14 for this set). The calculation can also be performed using $\texttt{t.test}(x_{\text{Avery}}, x_{\text{Anne}}, \text{paired} = \text{T})$ or $\texttt{t.test}(x_{\text{Avery}} - x_{\text{Anne}})$.