# Project Yelp Datasets

Team Members:

* LaKeisha Ringo
* John Giglio
* Pride Hawkins
* Stephen Wray

Scope:

For this project, we will be reviewing Yelp Datasets to answers the questions and hypothesis below:

Sources of data:

From Kaggle there was a competition that pulled data from Yelp and the dataset was converted from JSON to CSV format. This dataset is a subset of Yelp's businesses, reviews, and user data. It was originally put together for the Yelp Dataset Challenge which is a chance for analysts and students to conduct research or analysis on Yelp's data and share their discoveries. In the dataset you'll find information about businesses across 11 metropolitan areas in four countries.

**In total, there are**:

5,200,000 user reviews

Information on 174,000 businesses

The data spans 11 metropolitan areas

https://www.kaggle.com/yelp-dataset/yelp-dataset/home

Data Items will include:

* Yelp Business & attributes
* Yelp Users
* Yelp reviews
* Yelp tips

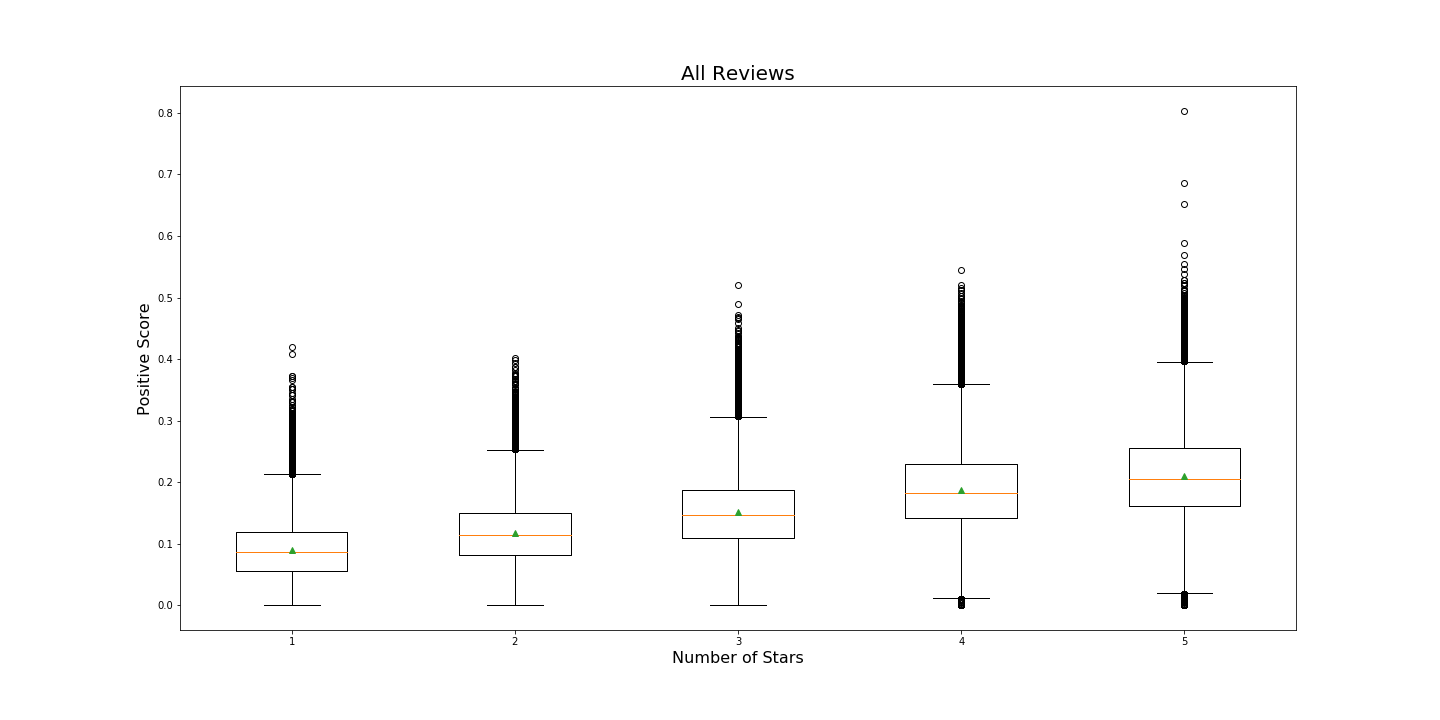
## Hypothesis

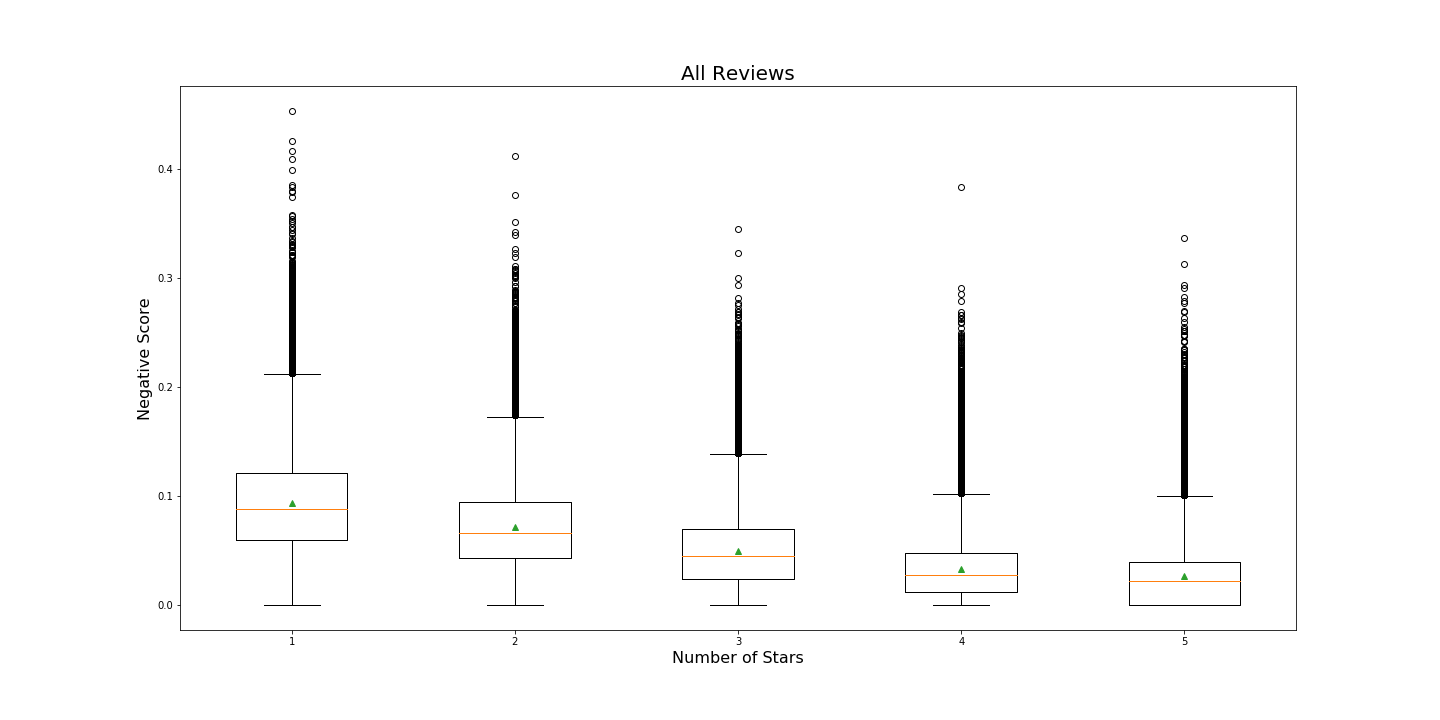
Elite users write reviews that more closely reflect the star ratings than the total population of reviewers.

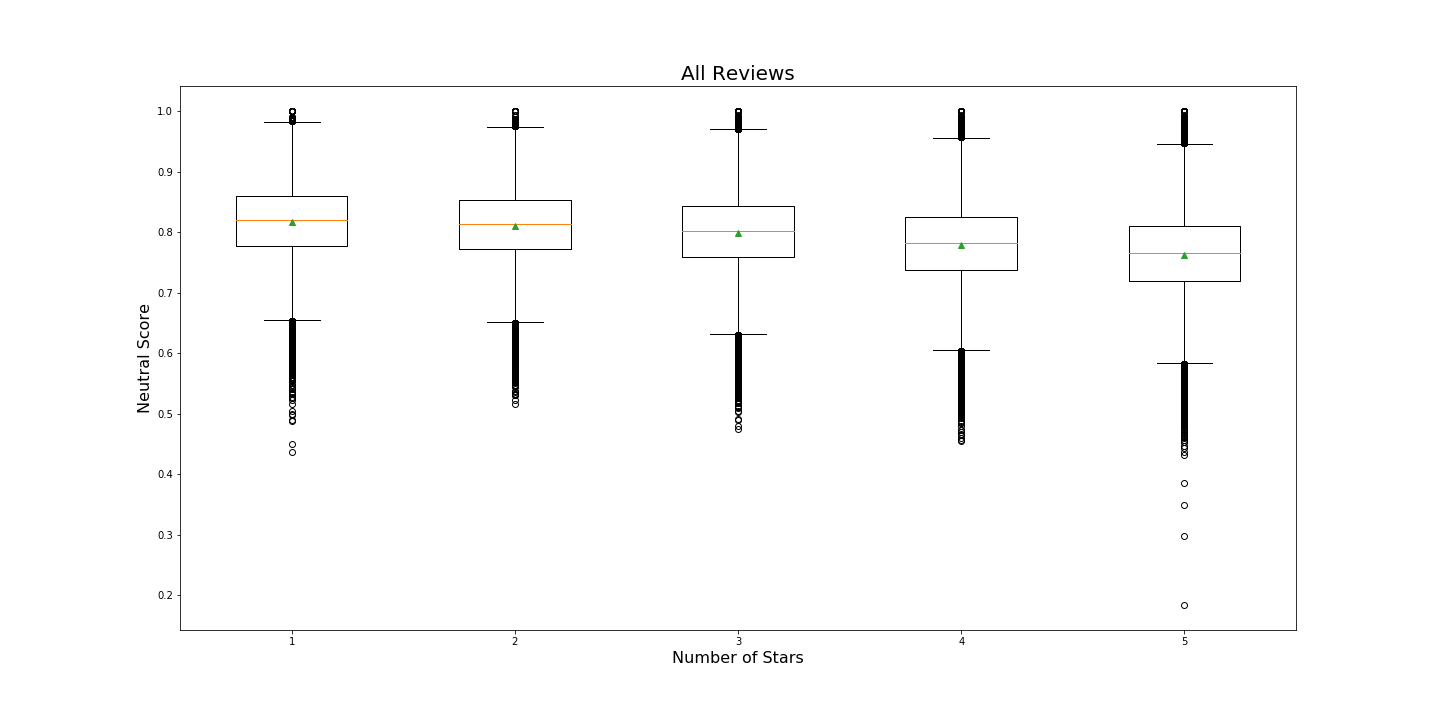
Key Questions:

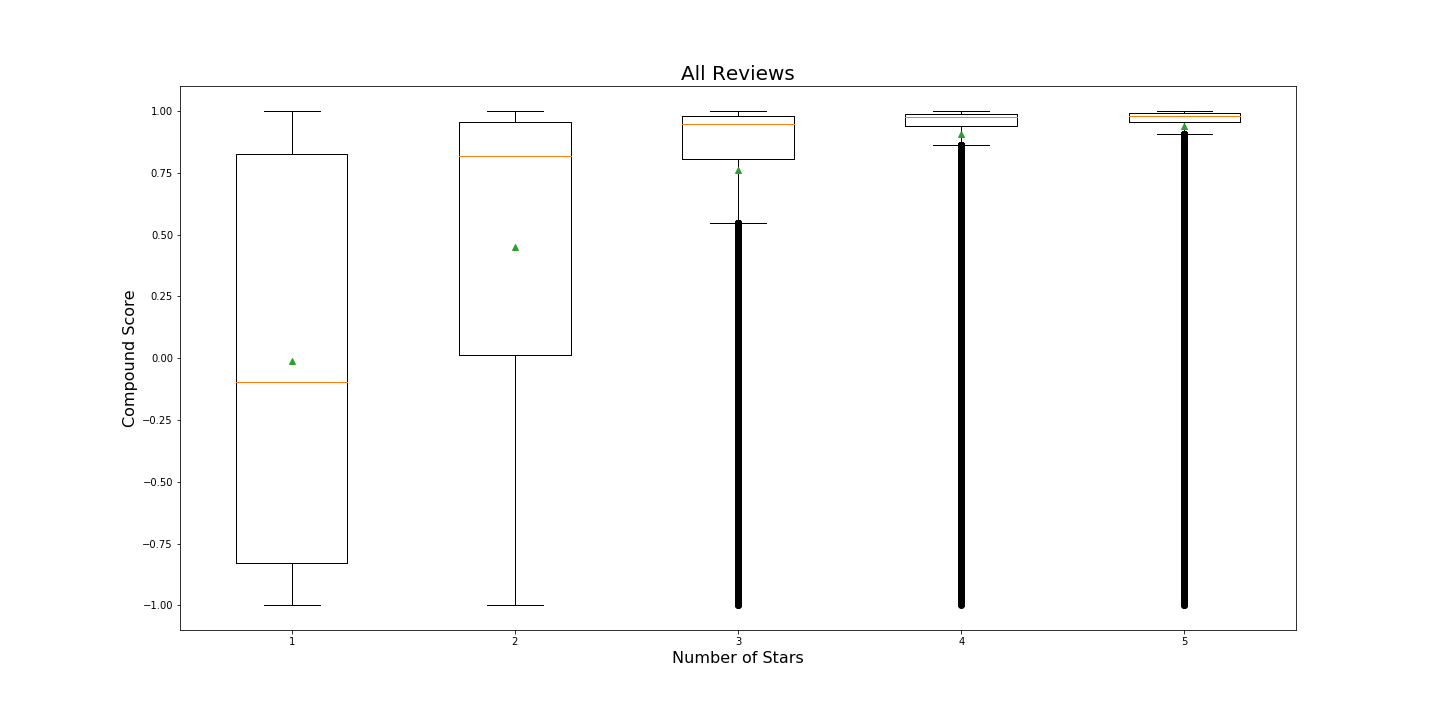
1. Are the sentiment analysis of the individual reviews different by star ratings?

The star rating does not appear to have a significant relationship with any of the four scores that were returned by the sentiment analysis. There were slight trends in the expected directions (upward for positive and compound scores, downward for negative score, flat for neutral score) as the star rating increased, but there are not significant differences that were detected through visual inspection. The box and whisker plots below showed that there does not appear to be any distinction between these star-rating subsets.









There were 95,286 1-star reviews, 100,673 2-star reviews, 159,025 3-star reviews, 295,404 4-star reviews, and 307,680 5-star reviews. Some characteristics of the above box-and-whicker plot are aligned with expectations. 4 and 5-star reviews make up the majority of the data set, and are highly concentrated in a relatively small area at the top of the compound score range. Additionally, 1 and 2 star ratings represent the majority of the negative compound scores in the dataset. However, the upper quartiles of the 1 and 2-star reviews are both over 0.82, meaning that 25% of each of these low-rating subsets had written reviews that scored very positively.

1. Is the individual elite review sentiment better correlated with the corresponding elite star rating than the reviews and star ratings of the total population?

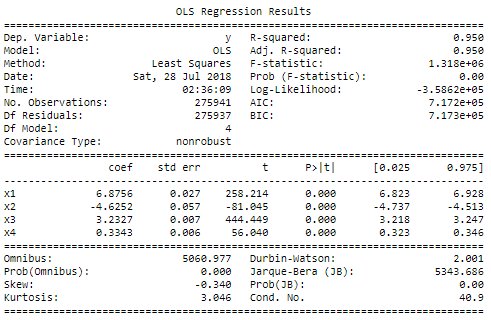
No, we created linear regression models for the elite and total population. We could not show that either was particularly good for showing a better correlation.

For both the elite user subset and the total population of users, we created linear regression models using all four sentiment scores as predictors and star ratings as the response. Summaries of these models are below.

Elite Users:

MSE: 0.789

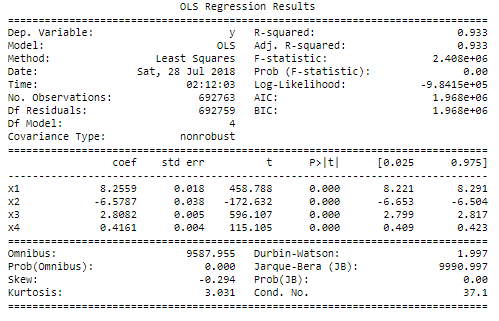
Variance Score: 0.253



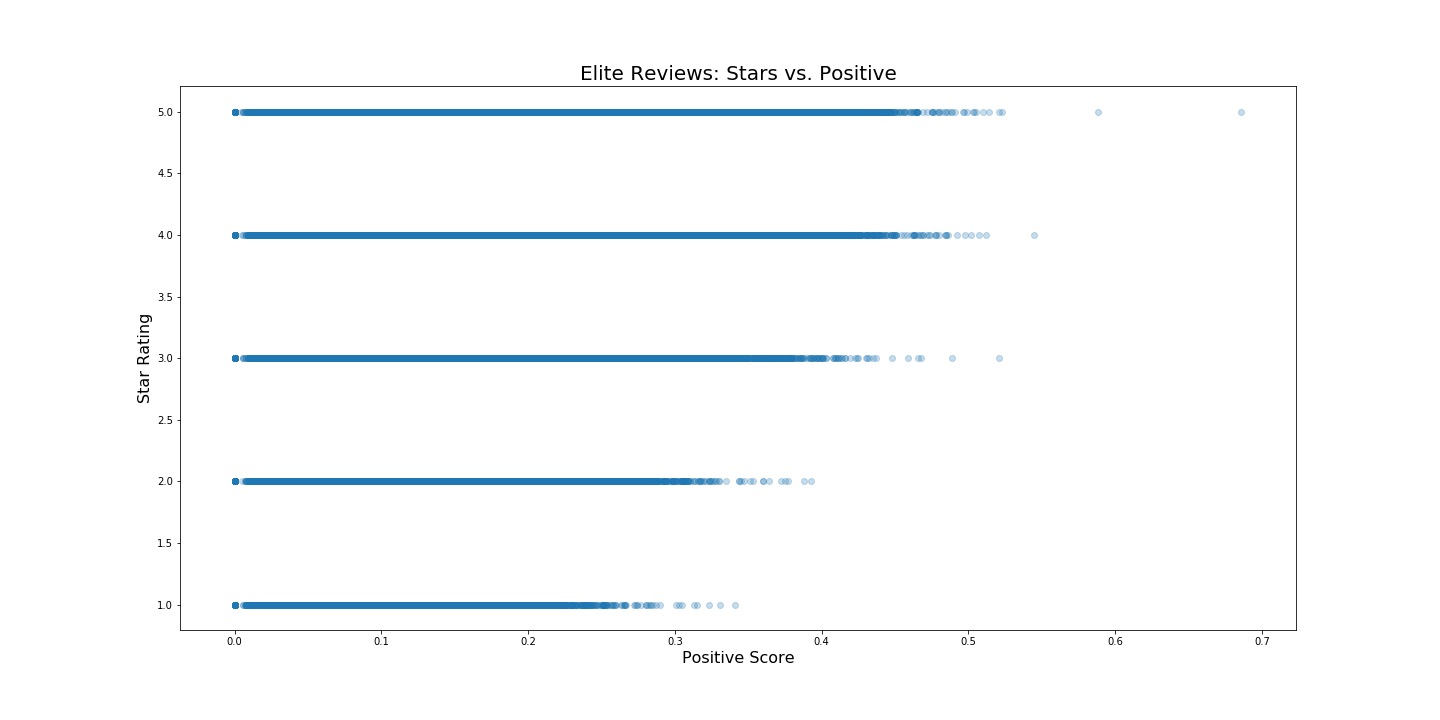
All Users:

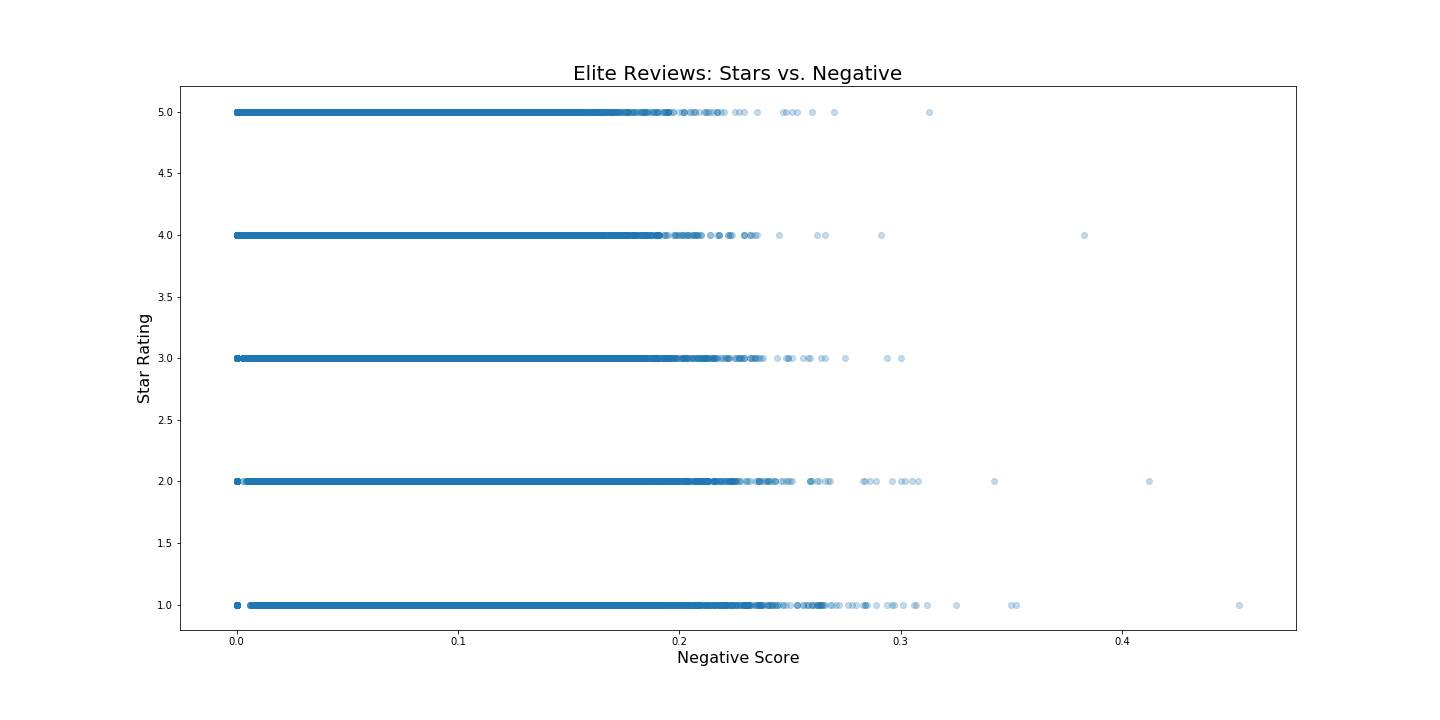
MSE: 1.005

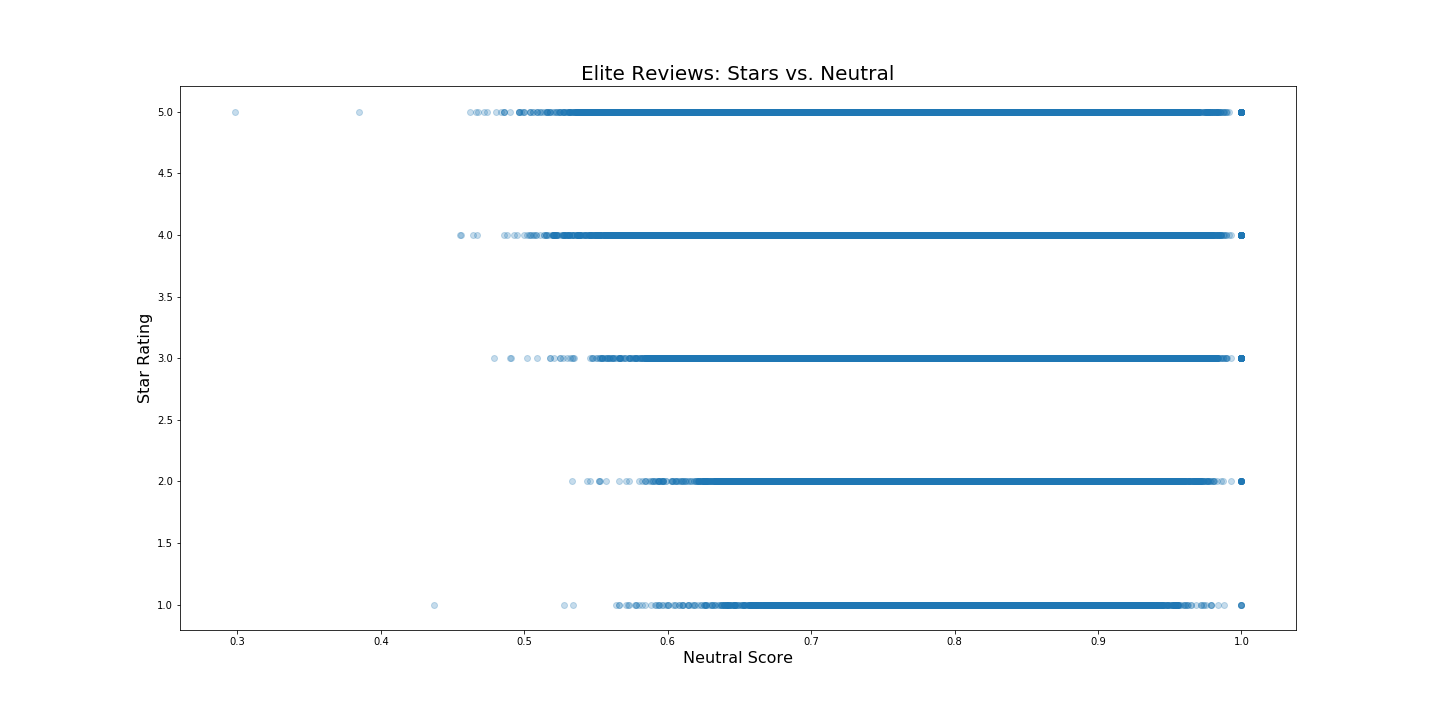
Variance Score: 0.406

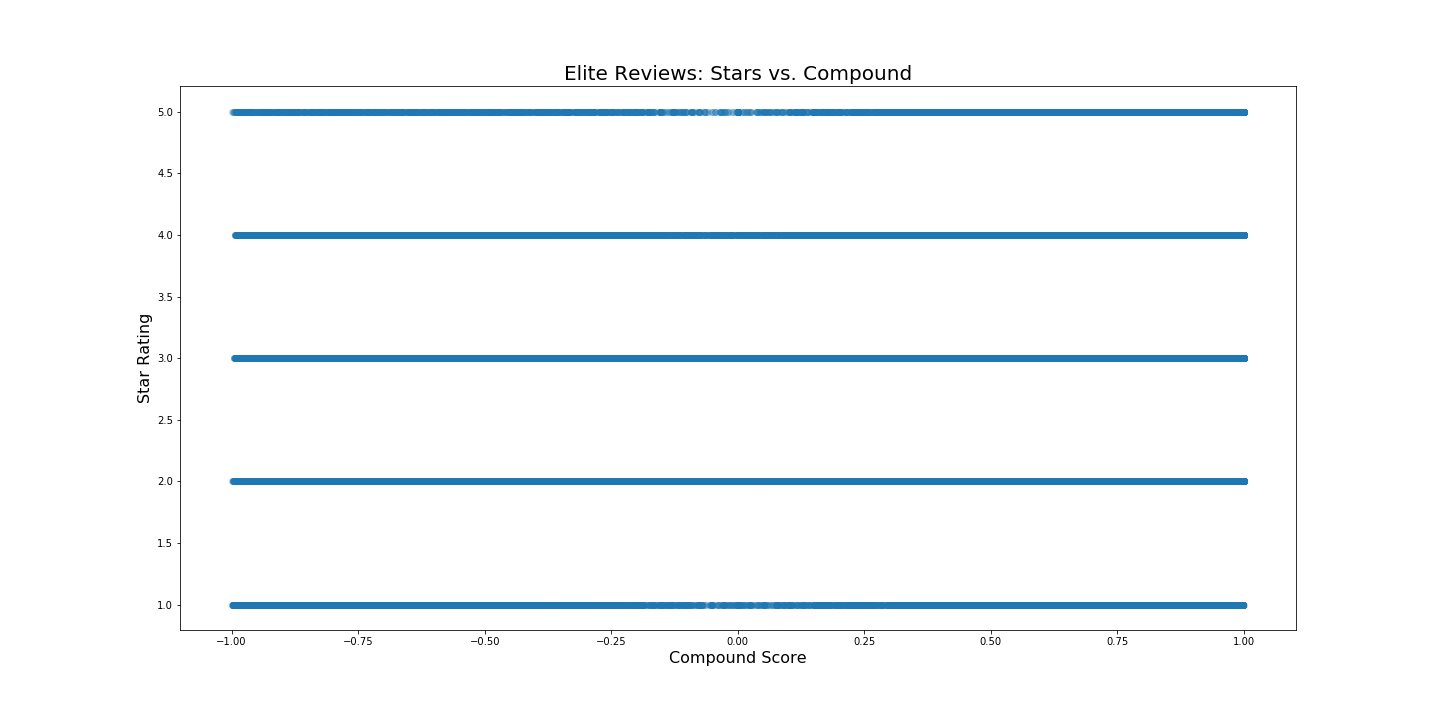


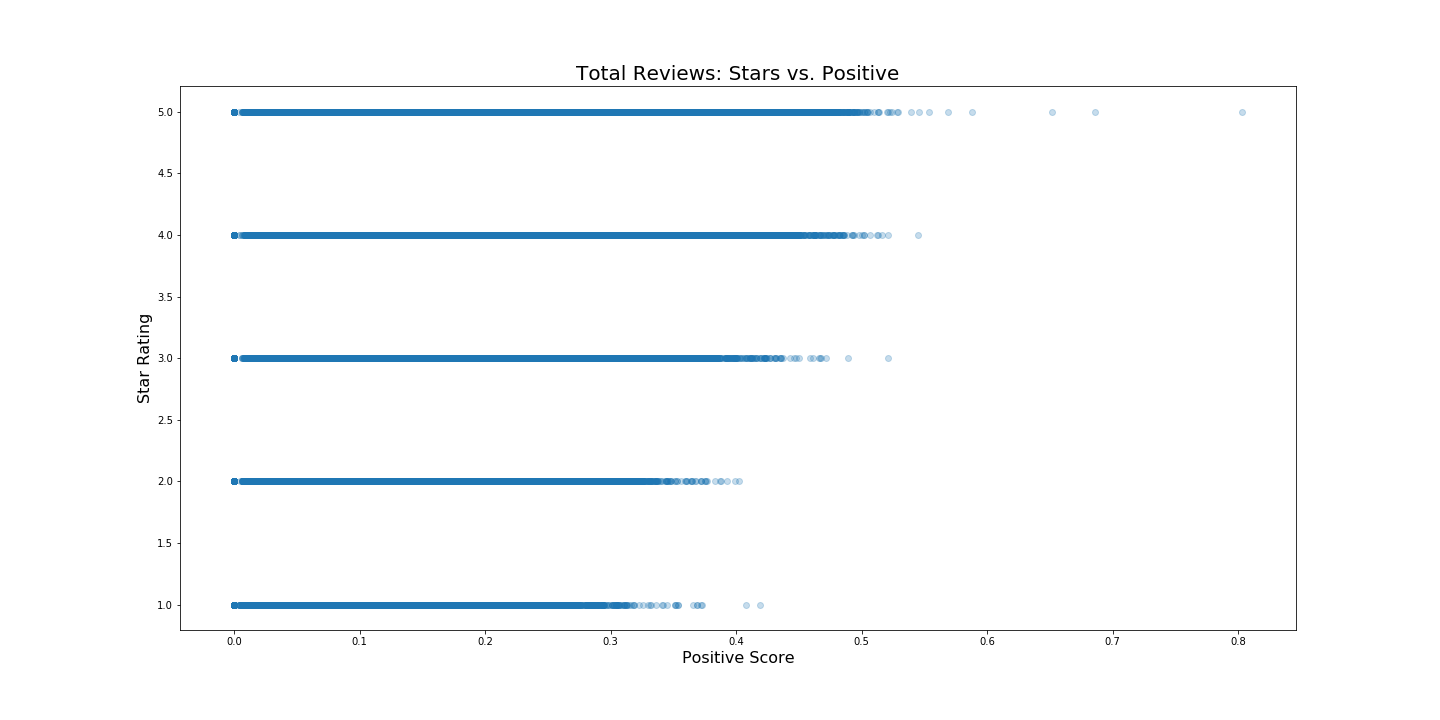
In both of the above models, the summaries show that each coefficient has a standard error and p-value that are sufficiently low to justify the assumption that each coefficient is significant to the model. However, for the sake of exploration, we let our intuition guide our feature selection to create additional models. Since the composite score is an aggregate of the other 3 scores, we chose to evaluate models that excluded this score. Conversely, we also chose to create models that used the composite score as the lone predictor. In each case, the mean squared error increased and the variance score decreased from the original model.

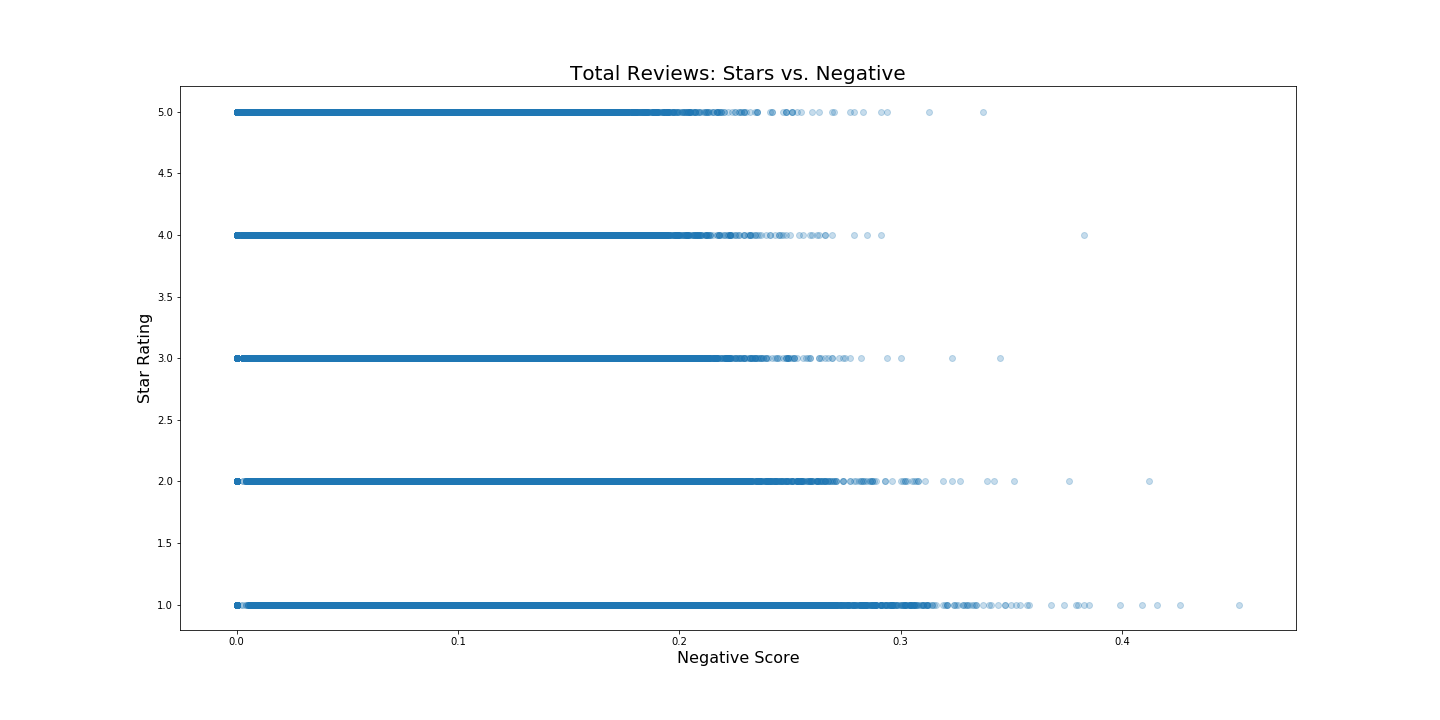


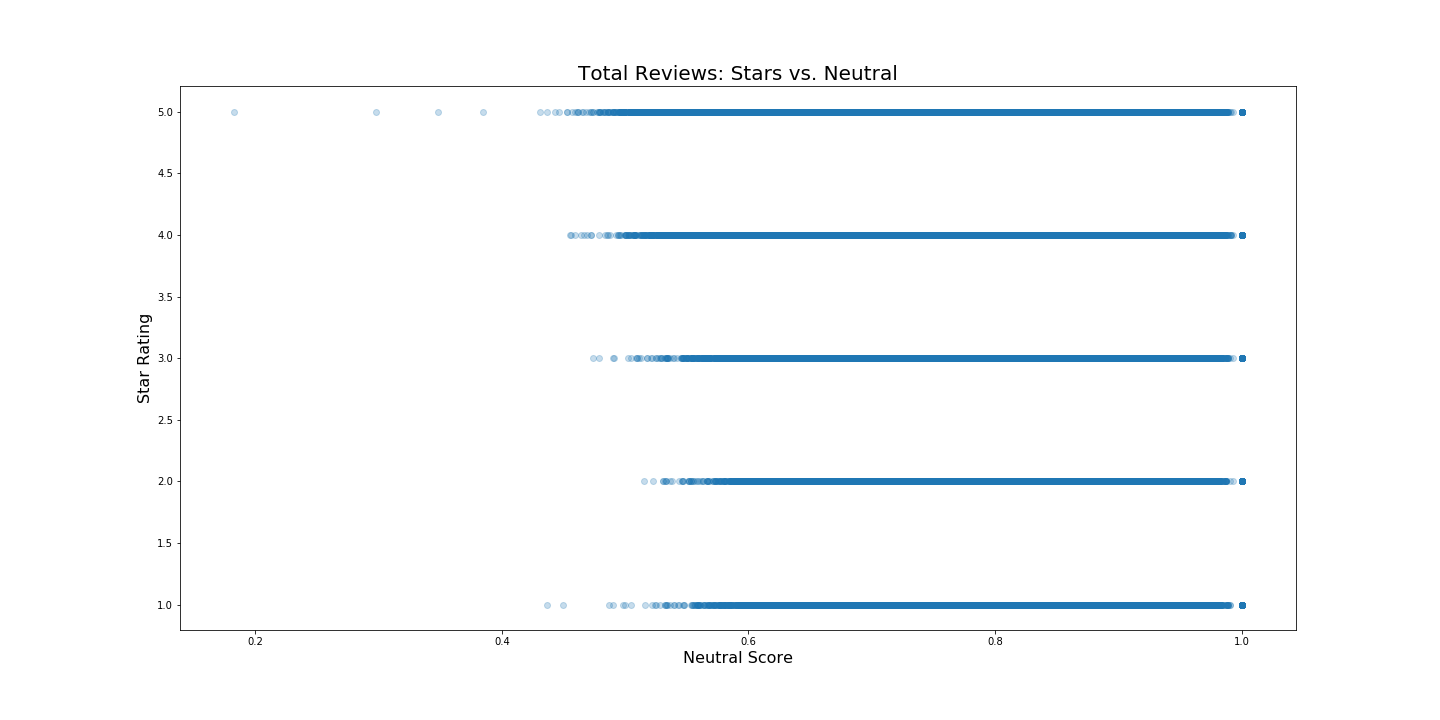


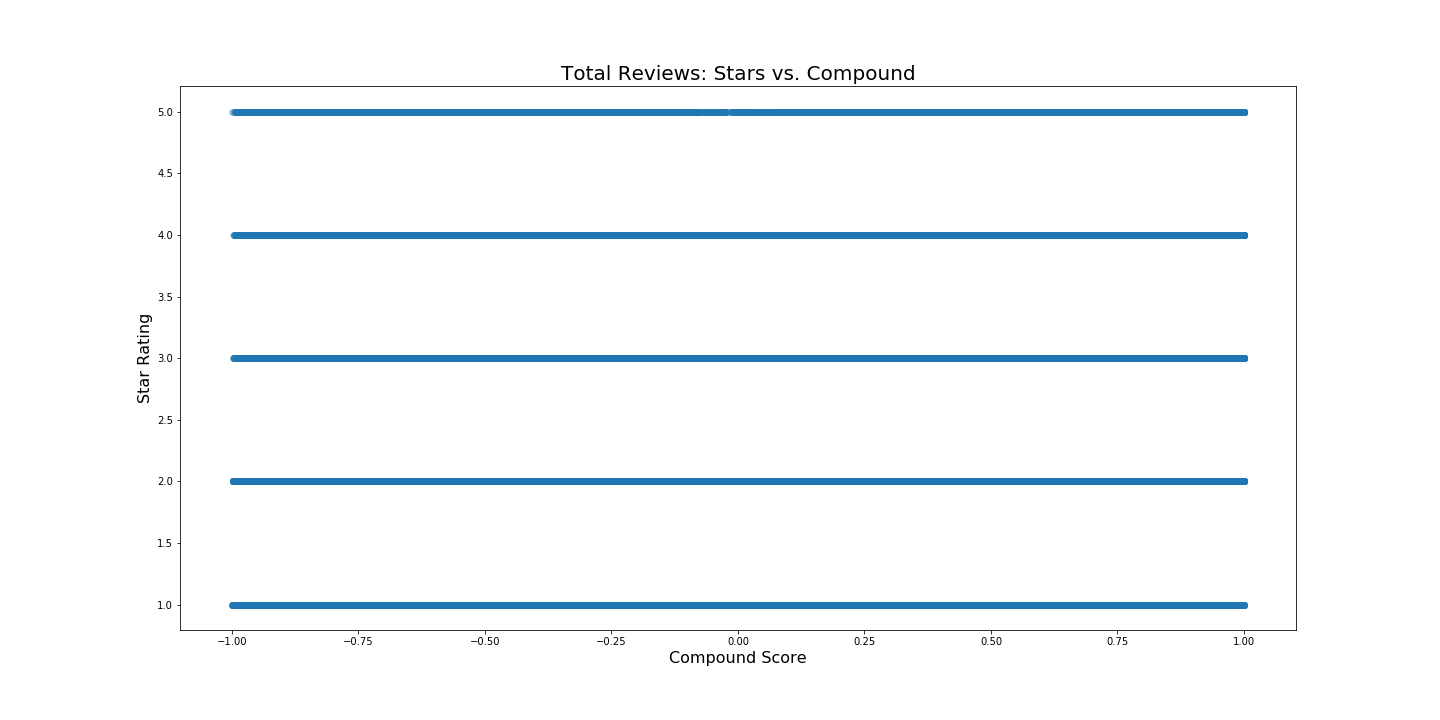








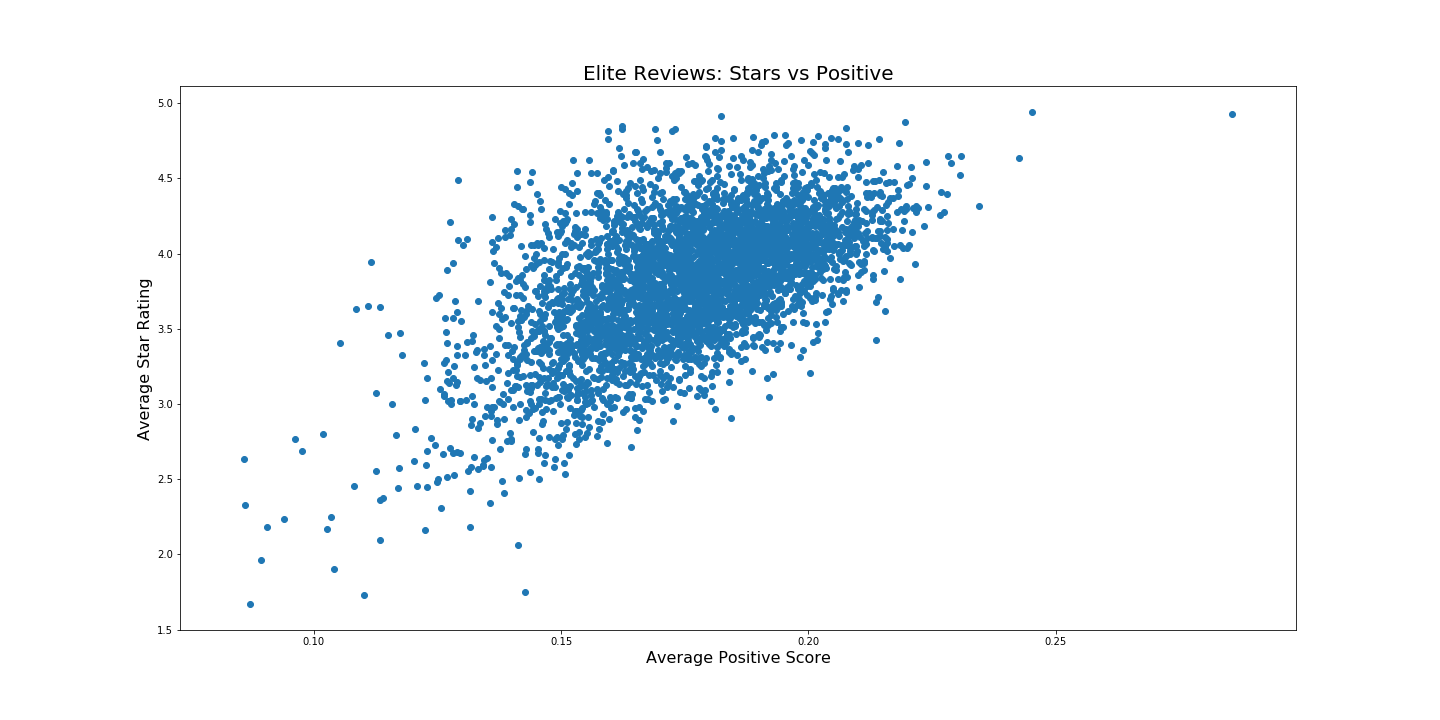


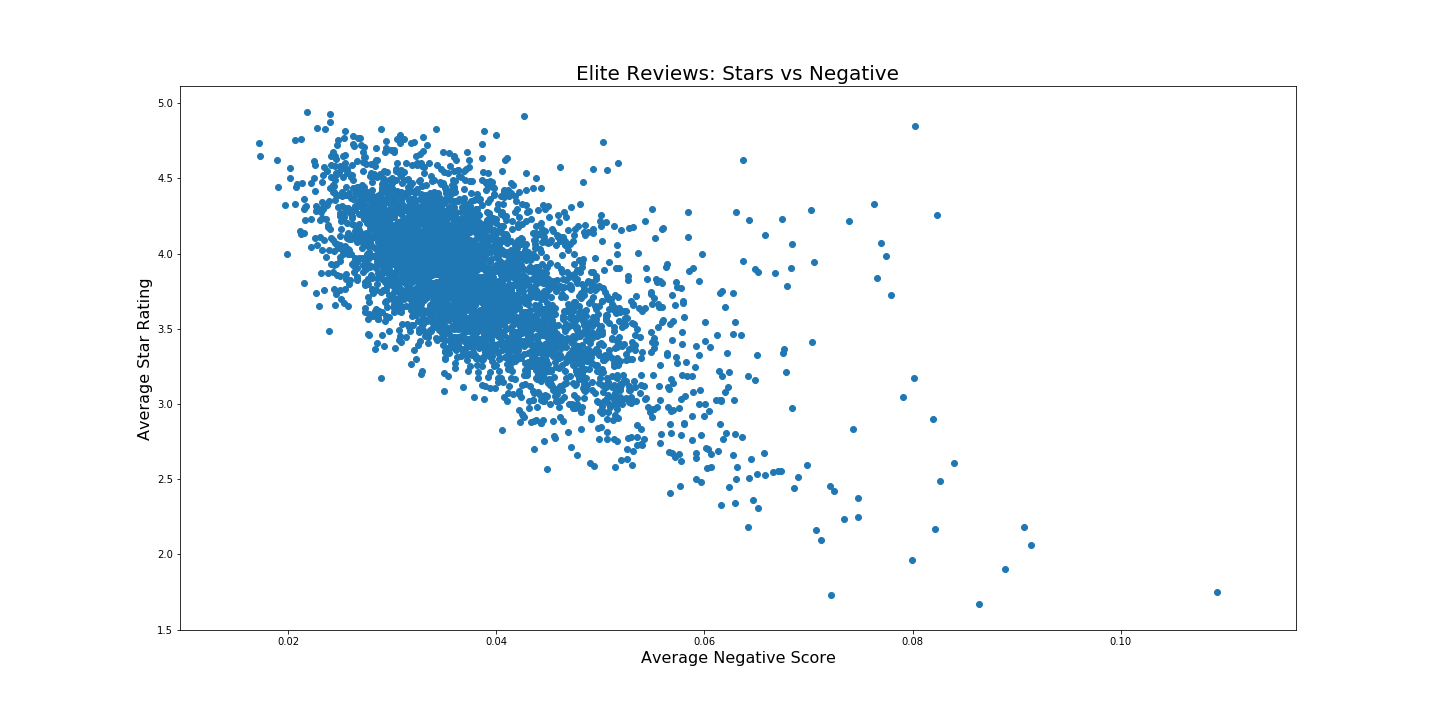


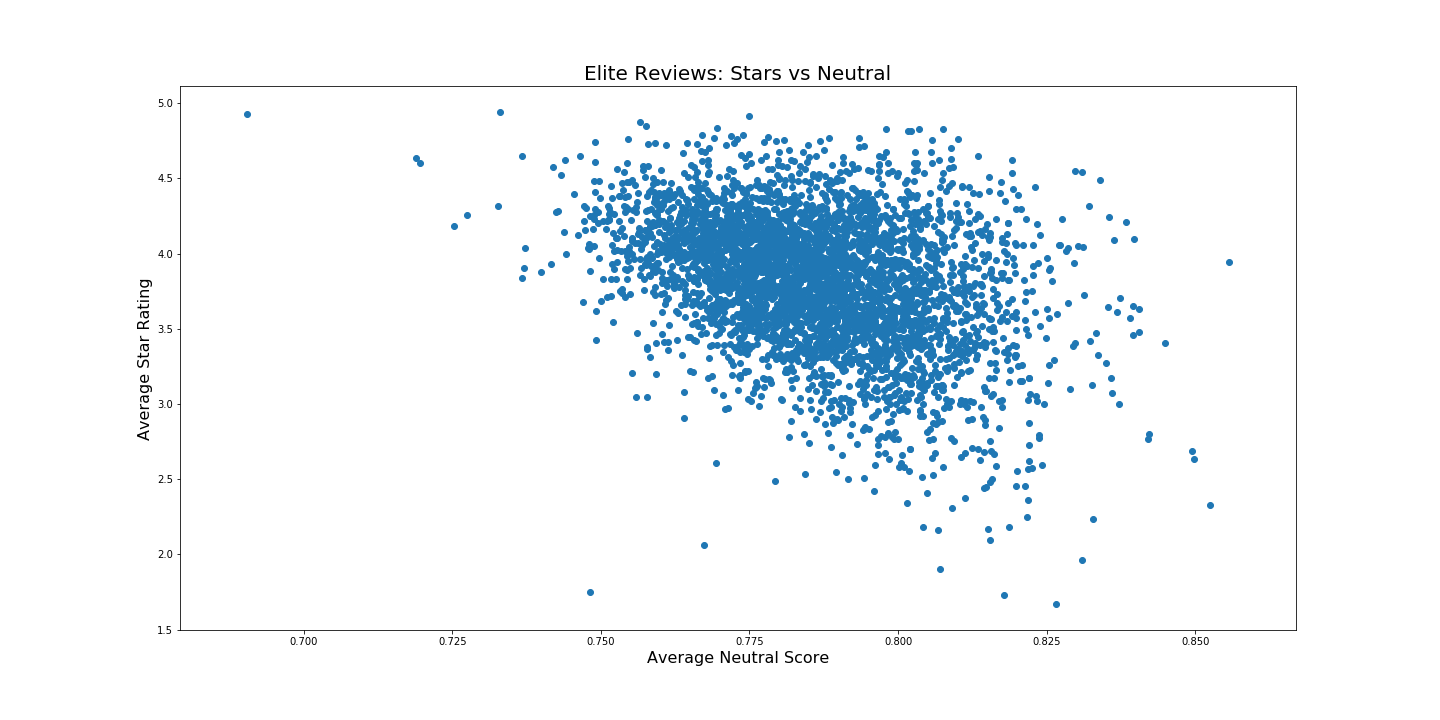
The above plots further illustrate the previous finding that there is no strong relationship between any of the sentiment scores and the star rating. Breaking out the elite users did not help to differentiate the similarly rated reviews. The plots also highlight the lack of granularity in the response variable, as each individual reviewer is only able to select from 5 possible choices. Although the MSE of the Elite User model is lower than the Total Population model ( 0.789 vs 1.005) it is not low enough relative to the magnitude of the response variables to absolutely conclude that it would be a good predictive model.

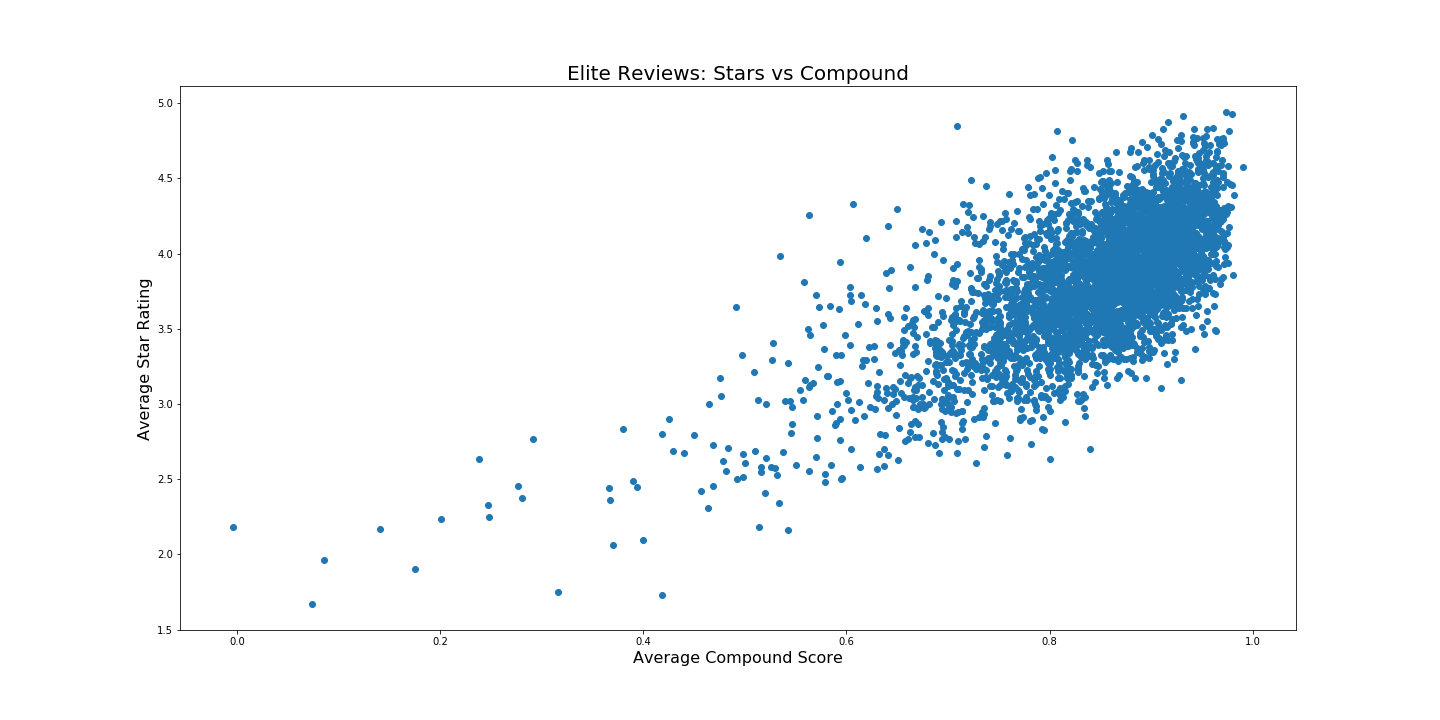
1. Is the average elite review sentiment better correlated with the average elite star rating then the reviews and star ratings of the total population.

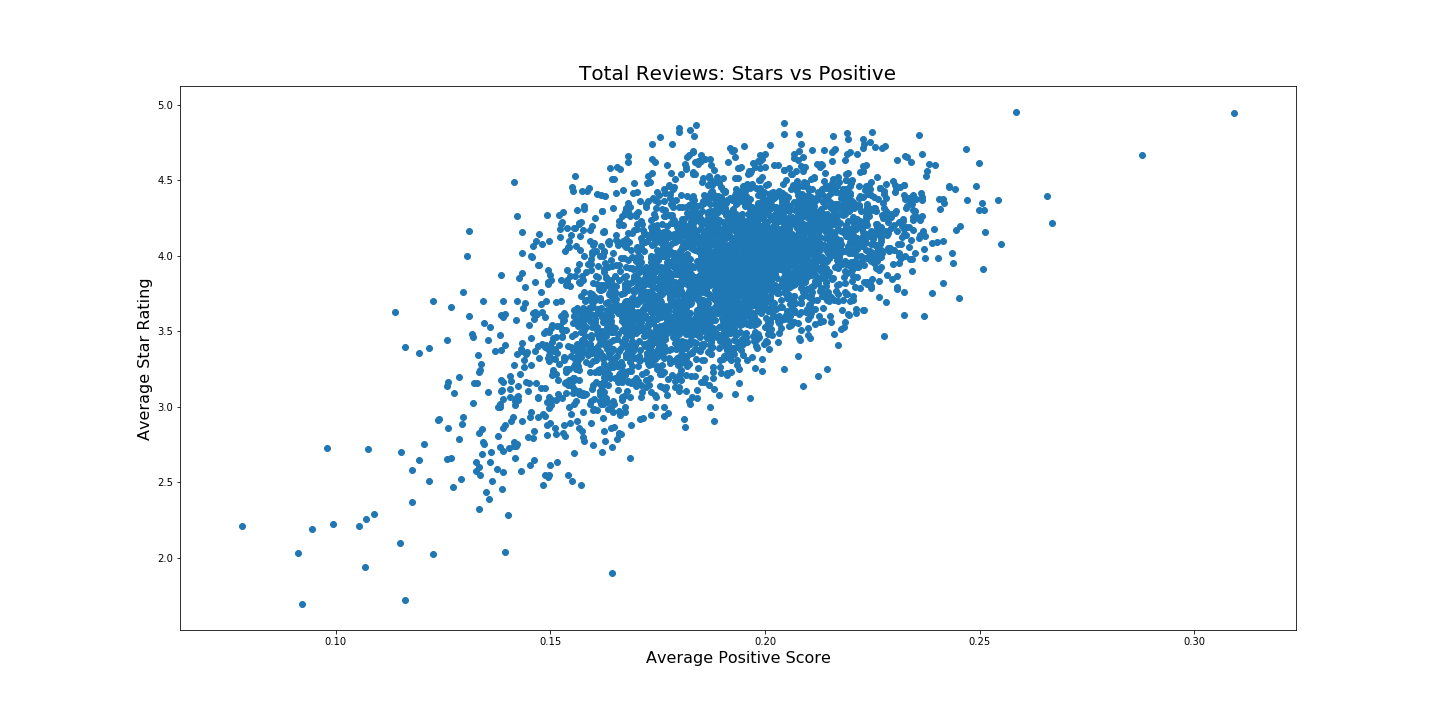
Again, no we created models for the elite and total population. We could not show that either was particularly good for showing a better correlation. To increase granularity of the response variable, we decided to group the reviews by business and calculate the average of each predictor variable and the response variable. This forced the response to be more of a continuous variable in hopes of creating more useful linear models. Whereas the previous analysis efforts modeled each review as a data point, this transformation results in each business becoming a data point.

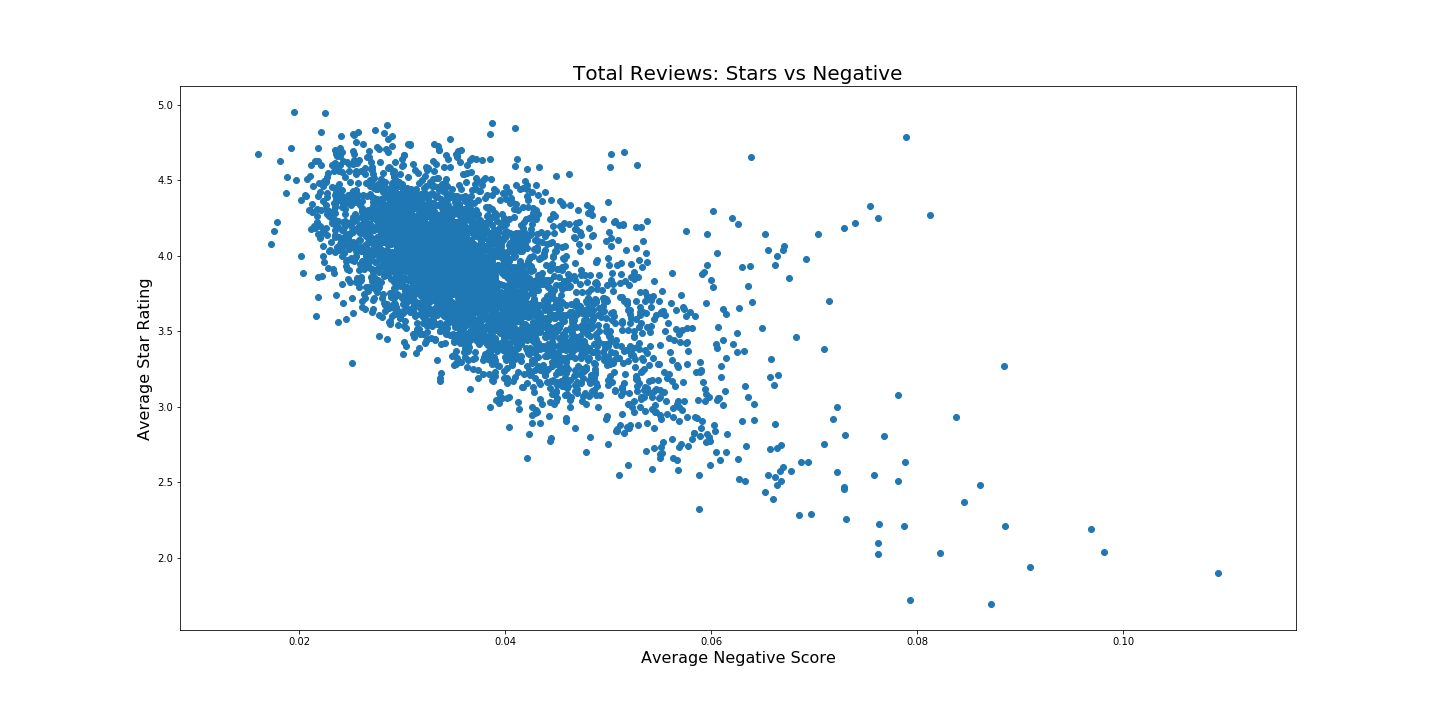


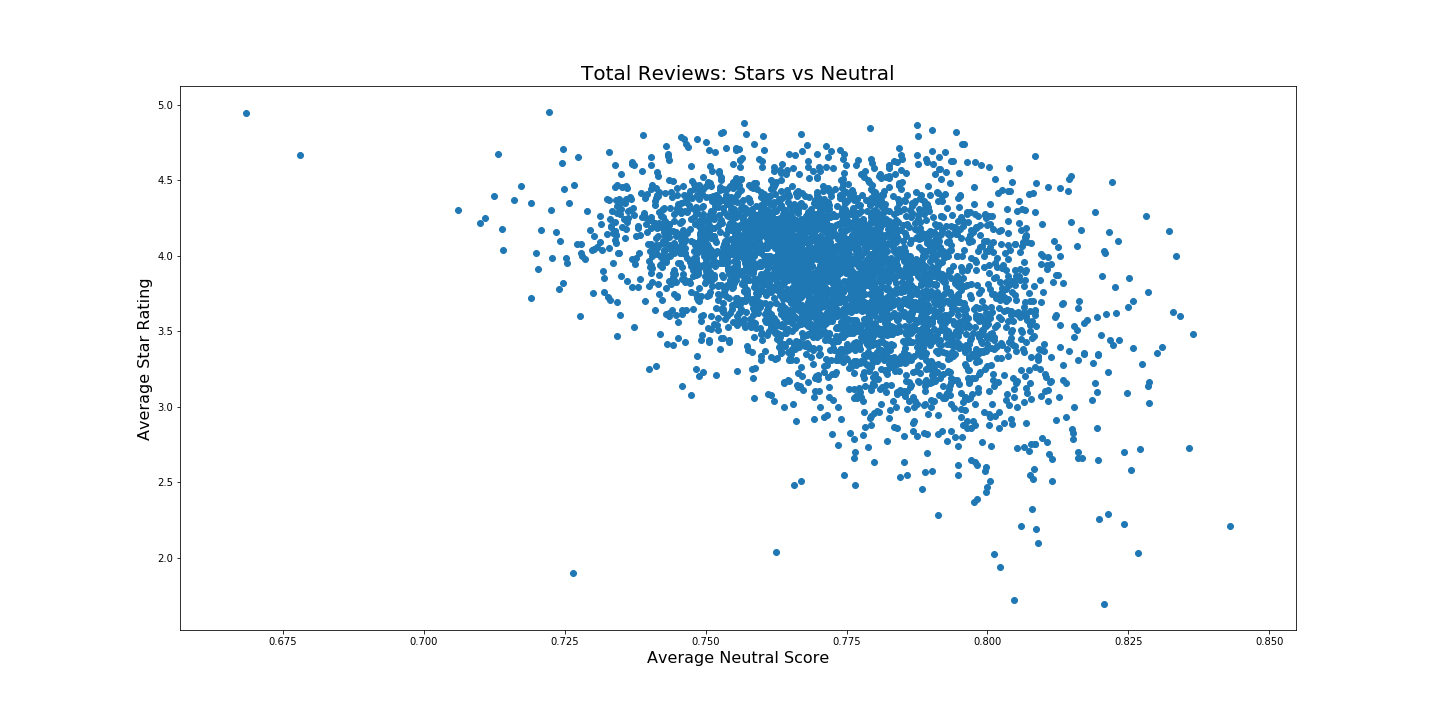


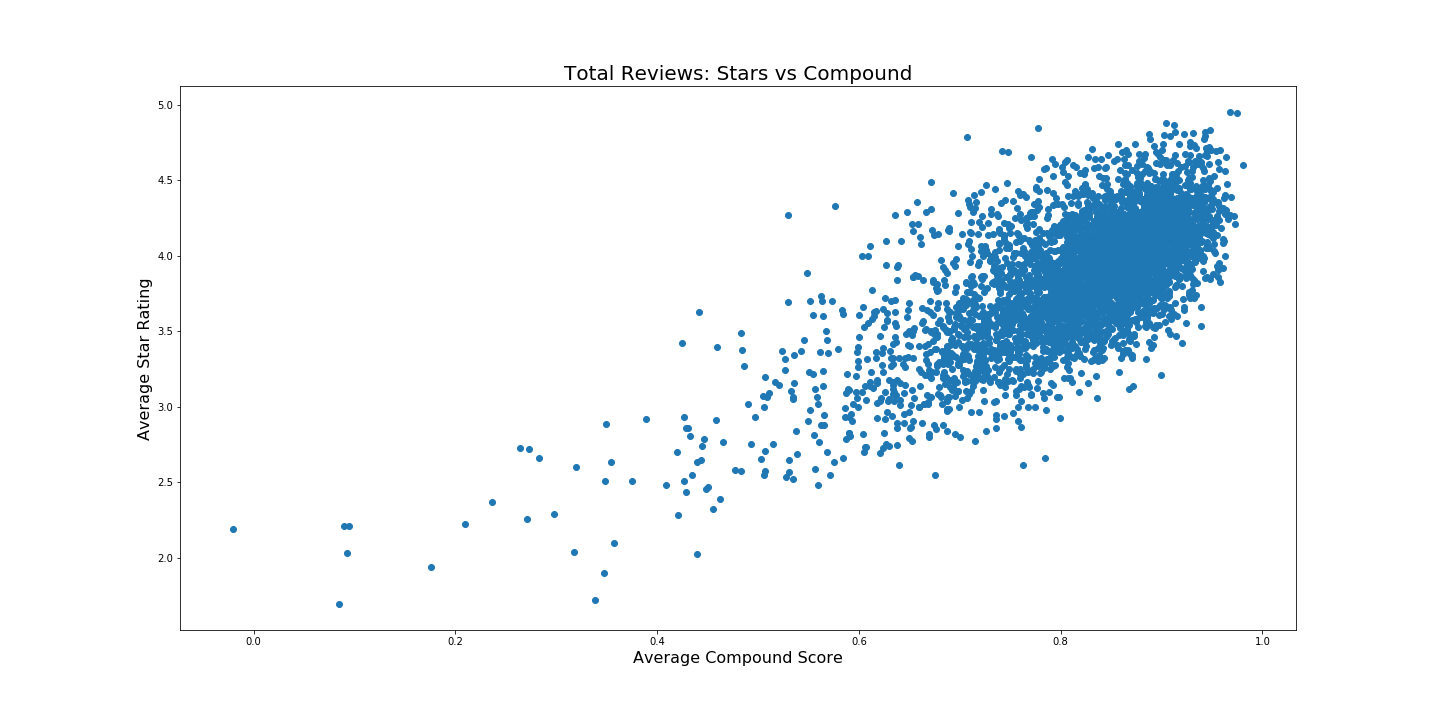












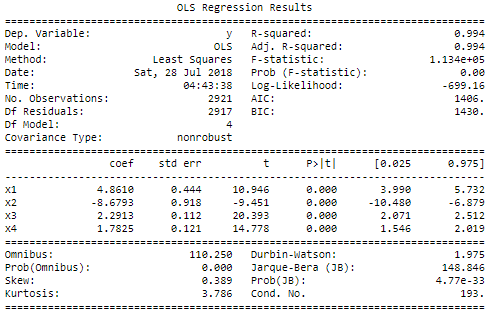
The above plots show that the transformation resulted in relationships that appear much more linear than what was observed in the unaggregated data. Though the data points are not strictly distributed linearly, we are seeing general trends in the expected direction for the positive, negative, and compound scores. There appears to be less of a relationship in the neutral score plots. Visual inspection of the Elite Reviews plots compared to the Total Reviews plots do not seem to show a significantly better relationship.

As we did with the unaggregated data, we created several linear models to describe the elite subset as well as the total set of average reviews. Again we found that the models for both data sets that included all four sentiment scores provided the best results. Summaries of these models are below.

Elite Reviews:

MSE: 0.087

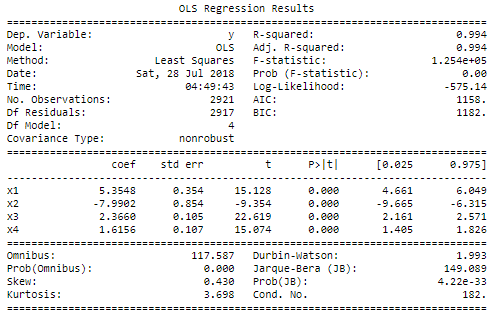
Variance Score: 0.518



All Reviews:

MSE: 0.078

Variance Score: 0.546



The mean squared errors produced by the above average review models are significantly lower than their respective unaggregated counterparts. Unlike before, the model that included all reviews has a lower mean squared error than the model of only elite reviews.

In conclusion, we attempted to answer several questions that we documented here

the analysis we did we think we can provide no statistical relation between the sentiment analysis of the text reviews and the stars averages.

If we were to continue this project we believe that we would want to investigate other options for transforming the data to be able to more analysis other than linear regression. Some other techniques that might provide better results could include classification models, decisions trees, principal component analysis.

https://docs.google.com/presentation/d/1nePg6pjavWmxLhki0A\_Ct\_VN9NwXtXJdEtzfJUe6Ut0/edit?usp=sharing