EDA and Revised Project Statement

Final Project Milestone #3
APCOMP 209A
Anthony Sciola, Maia Woluchem, and Jay Dev

Description of the data

The Yelp Dataset Challenge data consisted of six separate data files: check-ins, photos, tips, reviews, users, and businesses. Each of these data sets contain fields relevant to the particular population it describes, either users, establishments, or reviews. We are most interested in the data sets on reviews (which is at the review-level), users (which is at the user-level), and businesses (which is at the establishment-level).

The reviews data set contains 2,927,859 records. Among the fields provided by Yelp, we were most interested in two: the star rating given in each review (*rating*) and the date that the review was logged (*review_date*). We immediately noticed that the mean of *rating* is 3.70 and the median is 4.00, indicating that the distribution of reviews is skewed towards higher star ratings. The users data contains 1,183,362 user records, with variables describing average rating (*average_stars*), total number of reviews (*review_count*), years in which the user had 'elite' status (which we aggregate to the number of years with elite status: *elite_count*), and the date that the user joined Yelp (*join_date*). In reviewing this user data, we notice that *average_stars* match the ratings seen in the reviews data set (with mean 3.71 and median 3.89). We can also see that *review_count* follows an exponential decay function with a long tail—many users post just one or a handful of reviews with a small segment of very active users. This is reflected in the relatively small number users that have ever achieved elite status: less than 5 percent have ever held elite status.

As the Yelp business data set included records beyond the intended scope of our project, we filtered it prior to EDA. According to the project guidelines, we focused our analysis of businesses to those that were categorized as either 'restaurants,' 'cafes,' or 'coffee/tea' establishments. We were left with 38,668 business records. While the data set included a large array of features on business characteristics, we found that many of those features had a large share of missing values. As we could not confidently impute these values as False, we determined to minimize the level of missingness among features that we retain for regression analysis. After merging with the reviews data, we removed all variables with more than 50 percent missingness (leaving us with 49 characteristic variables of a possible 93). We were interested in conducting EDA that revealed patterns in ratings based on the cuisine of the establishment, but found 628unique values for types of cuisine in our dataset. Even when restricting for the most common cuisines, the visualizations were quite cumbersome, so we have forgone including them in this brief. After visualizing the remaining features, we found several which seemed important to include in our regression—particularly location of the restaurant; whether the restaurant takes reservations; is good for breakfast, brunch, lunch, dinner, dessert, and late-night; has delivery; has parking; type of cuisine; and whether the location has Wi-fi.

Within our initial analysis, we looked at the number of check-ins, photos, and tips at each establishment, as well as the number of tips provided by each user, but found that these data were missing for a majority of businesses and users, so we have opted not to ultimately use them.

Visualizations of Noteworthy Findings

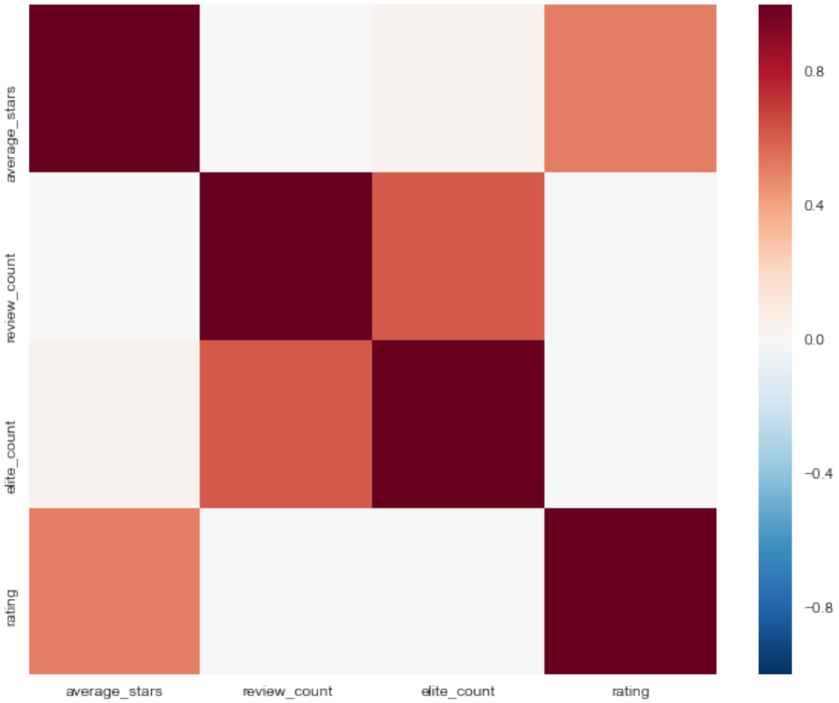

Figure 1. We see here that among users, *average_stars* is highly correlated with *rating*

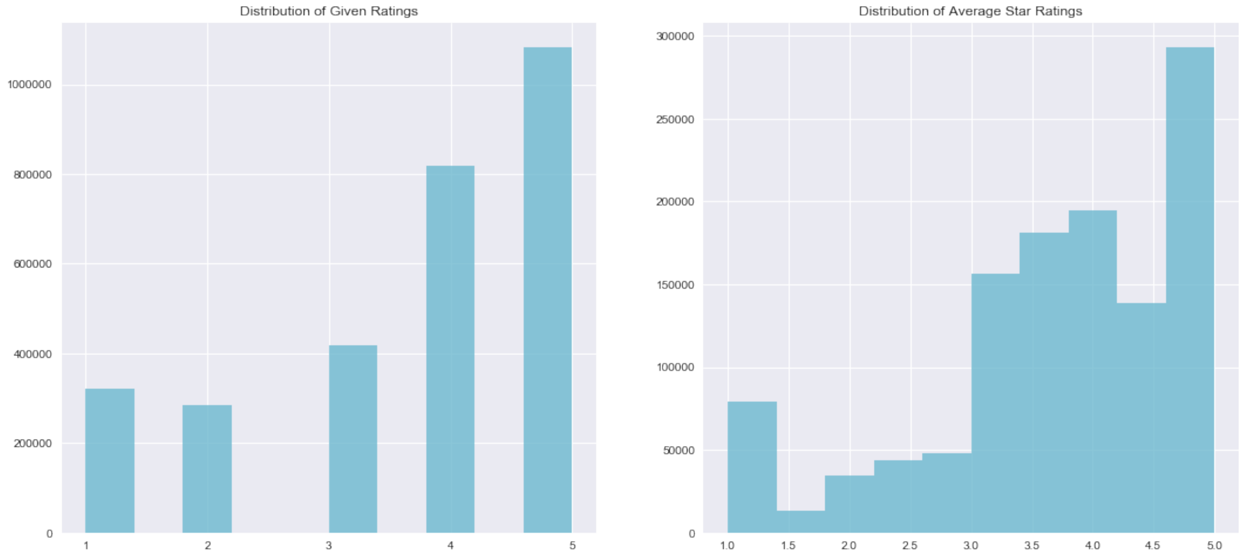

Figure 2. Corroborating the result above, the distribution of *ratings* and *average_stars* match one another quite closely.

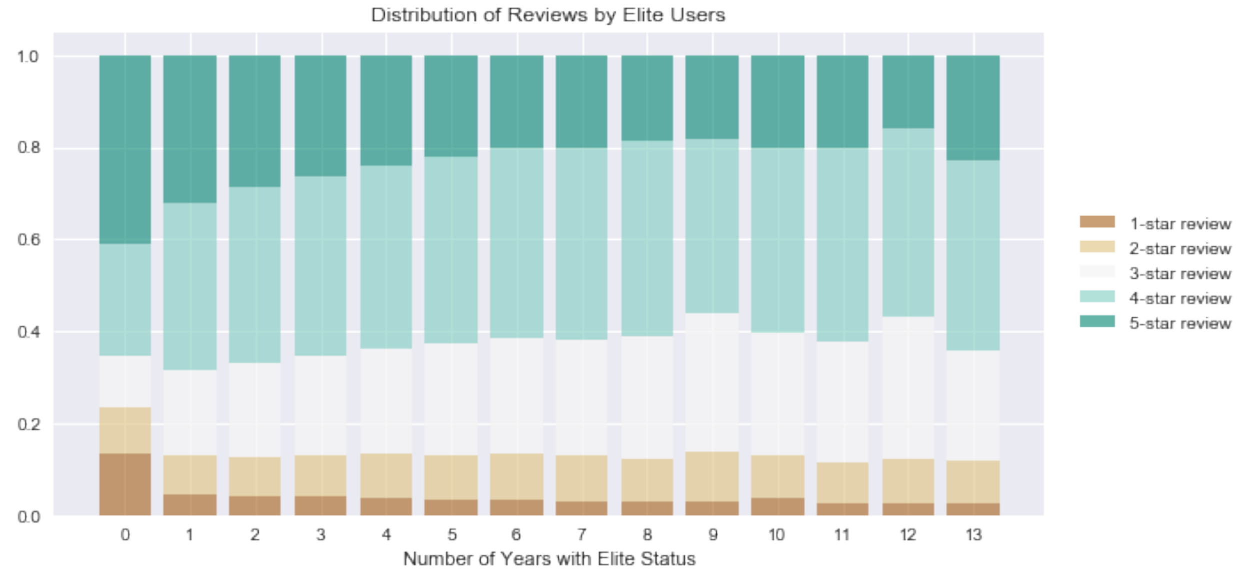

Figure 3. We find that those without Elite status are more likely to give 1- or 5-star reviews (which follows one-and-done angry/happy reviewers). We also see that more active users with a higher number of reviews are more likely to give 3 and 4-star reviews (visualization not shown).

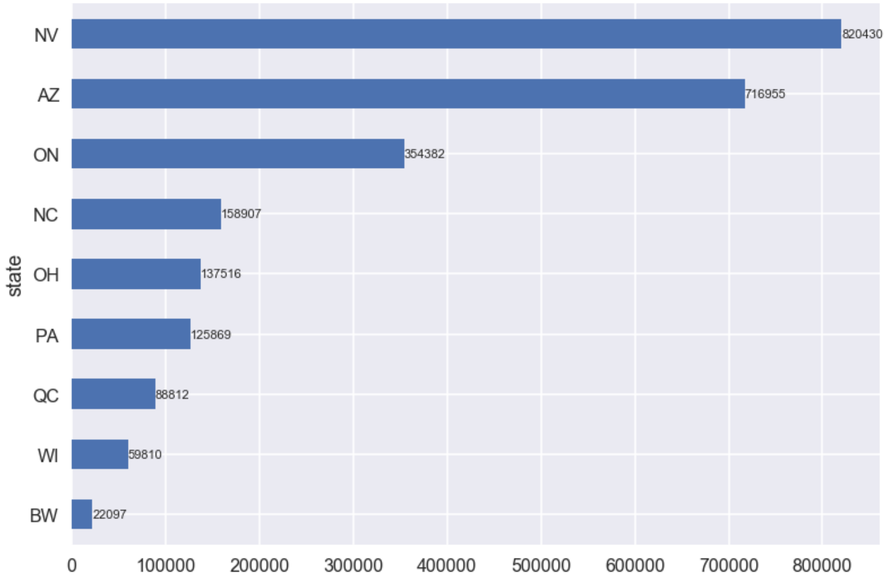

Figure 4. In this abridged visualization of reviews per state, we see a clear divide between the top eight states in North America and businesses in other states. We will therefore select only reviews from these states.

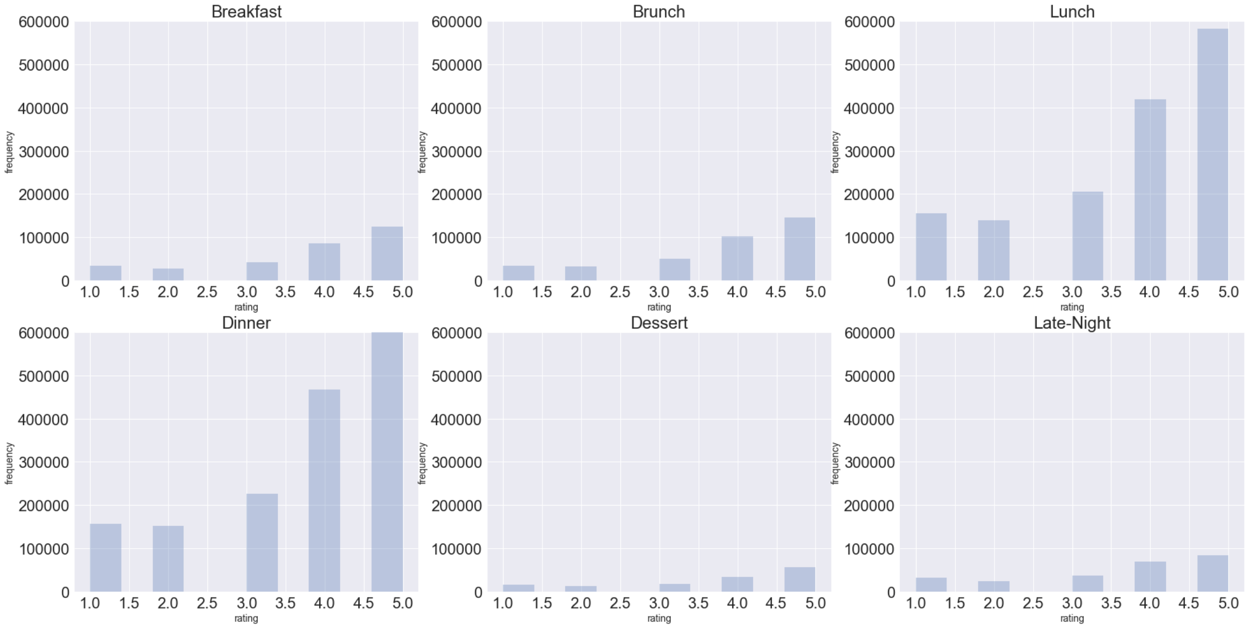

Figure 5. In comparing each meal that a restaurant is known for, we see that the distribution of star ratings is roughly similar across meals. However, many more restaurants are known for lunch and dinner than any other meal, potentially making them the most reliable predictors to include in our analysis.

Revised Project Question

Based on our exploratory data analysis findings, we are interested in understanding the impact of local market on prediction power. We plan to run the model separately for each of the eight cities that we have identified above. We will also run it with records from all of the cities together and compare results across the models. We are also interested to see if cuisine plays a major role in predicting ratings.