Capstone Milestone Report

Student: Christopher Jose
Mentor: Raj Bandyopadhyay

I would like to determine the main factors that affect Starbucks yelp scores in the United States. By knowing which factors customers care about the most, Starbucks can focus on improving these factors to improve the customer experience at existing shops, as well as keep these factors in mind when building new shops so as to best position them for success.

While Starbucks has its own internal data that indicates how well each store is doing, yelp data may provide additional insights. Yelp users can provide text reviews, and the content of these reviews may indicate factors that are affecting the customer experience that Starbucks was not as aware of through its own collection of customer feedback. And even if Starbucks is aware of the presence of every factor indicated by the yelp data (which is likely), the data may indicate an order of importance of the factors, meaning which factors yelp customers care most about, that Starbucks was previously unaware of.

Also, knowing about the preferences of yelp customers compared with all customers (information that Starbucks contains about all of its customers in its databases), may provide new and useful information. Perhaps factors that yelp values most highly differ from those in the overall customer base. Therefore, if Starbucks wants to improve its yelp scores over the long run, then it'd focus on factors most important to yelp.

I am going to make a determination about the main factors affecting yelp scores by building mathematical models to predict a location's overall yelp score from factors I've looked into through exploratory data analysis. I will regard factors the models deem as having a statistically significant relationship with the overall yelp star rating as the main factors affecting Starbucks yelp scores.

What is Yelp?
Yelp is a website where users can assign star ratings on a scale from one to five to businesses (coffee shops, restaurants, hair salons, etc). A star rating of one is the worst possible rating, while a star rating of five is the best possible rating. Users can also write text reviews. Yelp assigns an overall star rating to each business using automated software that filters out reviews it essentially deems as junk reviews. Yelp claims that about one quarter of reviews are not factored into the overall star rating.

What does the yelp data look like?
The yelp data set consists of multiple files. I use the yelp_academic_dataset_business.json and yelp_academic_dataset_review.json files. Fields that I make use of in the business.json file include: "business_id" which uniquely identifies each store, store name, several location fields, the overall yelp star rating assigned to the store, and number of reviews. Fields that I make use of in the review.json file include: "business_id", individual reviews (their text content) for each business id, and the number of stars that each user assigned to their respective review. Thus, this

file contains multiple records for each store, where each record is a review and the associated number of stars that the user gave that store location.

Files that I do not use include: a file containing information about each user, a file containing the total number of times yelp users indicated that they checked into the store during each hour of each day of the week, and a file containing photos.

Where did I get the yelp data from?
https://www.yelp.com/dataset_challenge

Limitations and questions that I cannot answer based on the available data
Is the overall yelp star rating for each Starbucks location reflective of all customer's perceived opinion of that location (or is the rating biased based on characteristics of the customers that use yelp that cause them to have a different opinion from all customers)? Starbucks would know the answer to this based on their collection of feedback data.

In general, non-public information that Starbucks maintains, such as metrics that indicate the success level of each location, would be useful. This information for each location could be compared with the yelp star rating of each location to determine if the opinion of yelp users is in alignment with how well the store is doing, as measured by Starbucks' data.

Both files that I am using lack data for many of the states. Data is only provided for the following locations: AZ, NV, NC, PA, WI, IL, SC and Canada.

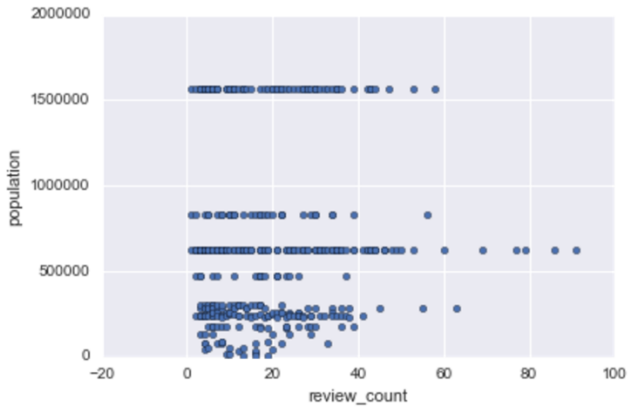
Data wrangling
The data was available in the json format. I imported the data into pandas and converted it to pandas DataFrames. I imported city population data and read it as a pandas Dataframe. The data is from: http://www.census.gov/popest/data/cities/totals/2015/SUB-EST2015.html. I converted a date field in the review content Dataframe to datetime format.

Exploratory Data Analysis

As a result of data exploration, I've decided to consider the following predictor variables (as detailed below): review year, region, state, and three indicator variables to account for store cleanliness, barista friendliness, and the presence of homeless people

Investigating the data - looking at review count by location and population size
In understanding the makeup of the data, and to check that the data makes sense, I investigate the number of reviews for each location. The median number of reviews per location is 13. The mean number of reviews is 17 and is greater than the median because of some locations having a large number of reviews (max is 91). The typical amount of deviation from the mean (the standard deviation) is 14.5 reviews.

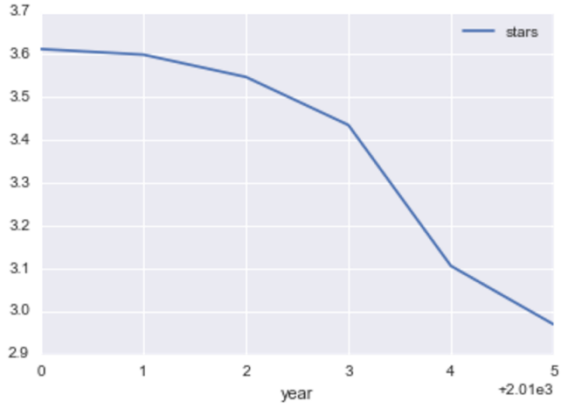
Why might review count differ substantially by location? I look at whether locations with a lot of reviews are located in cities with larger populations, meaning, whether there is a positive correlation between review count and population. While correlation does not imply causation, a greater population would be a sensible reason as to why certain stores have a greater number of reviews. The below scatterplot shows that Starbucks locations within the same cities having differing review counts, as shown by the horizontal bands of blue dots. The spearman correlation coefficient is only .247, indicating a weak positive correlation. There does appear to be a stronger positive correlation for cities with lower populations. Perhaps if more data was included in the yelp data set, there would be a more noticeable correlation between review count and population for smaller cities/towns/regions. Both the distribution of review count by population and review count by location seem plausible, and so I do not see any obvious absurdities in the data with respect to review count.



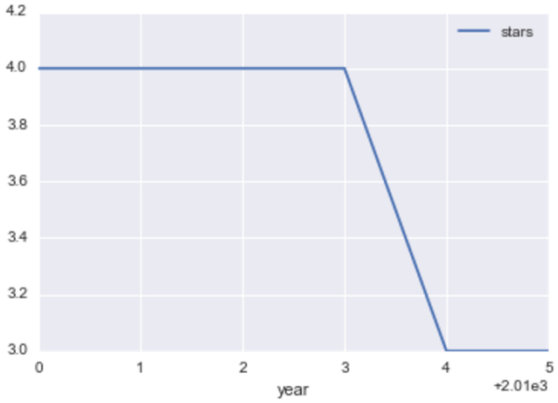
Investigating stars by review year

I investigate the star ratings by review year. I want to know whether review year deserves consideration as a predictor variable of overall star rating. For any given year in which users give Starbucks star ratings, I look at the average and median star rating for that year. The average and median star rating decreases as review year increases (mean ratings are graphed on the left, and median ratings are graphed on the right). I will thus give review year consideration as a predictor variable.

Avg star rating by year

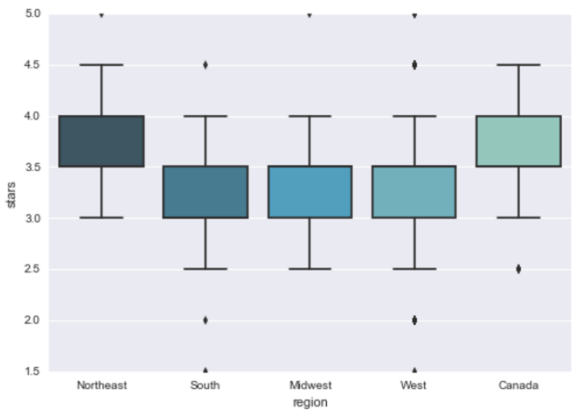


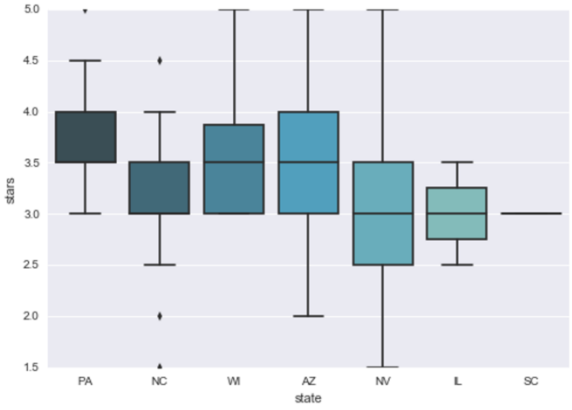
Median star rating by year



Investigating whether stars vary by region and state

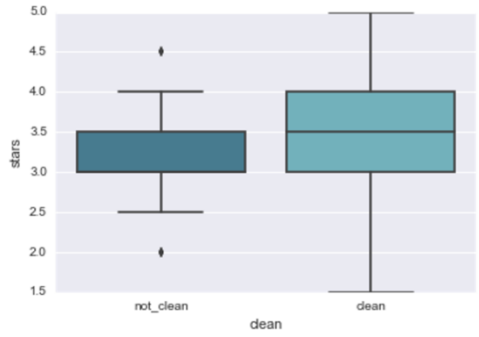
I investigate whether either region (Northeast, West, Midwest, South, Canada) or state deserve consideration as predictor variables. Does the average overall star rating differ across regions and states? Judging by the boxplots below, region does not appear to have differing average overall star ratings, while state does appear to have differing average overall star ratings. However, I will try using both as predictor variables.



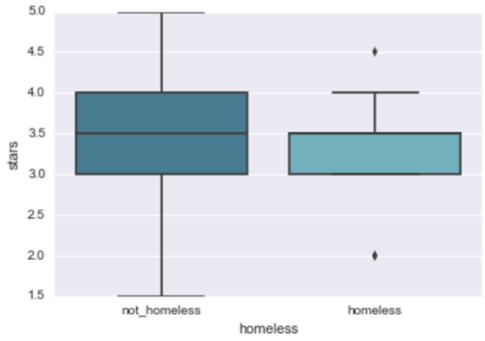


Investigating text review content

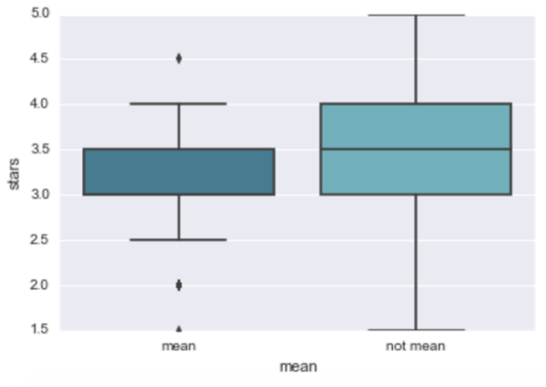
I investigate whether the presence of certain words in the text reviews changes the overall star score. I look into overall star score distributions for stores whose reviews contain the words "dirty", "filthy", and "messy" versus stores that do not have these words in their reviews. The presence of these words indicates that the store is not regarded as clean and so I'm considering store cleanliness as a predictor variable. The below box plots indicate that store cleanliness should be considered as a predictor variable, since the distribution of star scores for perceived non-clean stores is lower than that of clean stores.



Similarly, I investigate whether the presence of the word "homeless" in text reviews affect the overall yelp score of a location. The below boxplot suggests that it does, so I will consider the presence of this word as a predictor variable:



I investigate whether the presence of the words "mean" or "rude" in text reviews affect the overall yelp score of a location, and they appear to, so I will consider the presence of these words as a predictor variable:



Investigating review count

Regarding the dummy variables, a potential issue is that the greater the number of reviews, the more opportunities for reviewers to use any of the words indicating homeless problems, uncleanliness, or unfriendliness. Thus, it seems plausible that a store with more reviews will also have a greater chance of having dummy variable values indicating the presence of the above problems. This would mean that the dummy variables are correlated with review count, which, if excluded from the model, would be a lurking variable that is contributing to the change in star scores associated with a change in the dummy variable values.

I find that review count is strongly positively correlated with the unfriendliness variable, mildly positively correlated with the uncleanliness variable, and that there are no conclusions to be drawn on the homeless variable due to too few stores labeled as having a homeless problem. The spearman correlation coefficients are .784, .535, and .46, respectively. The scatterplots between review count and a count of the number of reviews containing the words indicating problems is shown below. Least squares regression lines are shown as well, along with 95% confidence intervals of the predicted y values surrounding the line (the shaded area).

I am going to included review count as a predictor variable due to its correlation with some of the dummy variables.