IMT 575, Autumn 2018

Preliminary Analysis and Project Plan

Team - Sahil Aggarwal, Amir Ali, Richard McGovern, Shreya Sabharwal

Question

Yelp gathers a wide set of variables about businesses in metropolitan areas including restaurants and other services. They also gather 4-star ratings from reviewers who check-in to these establishments and rate the quality of their experience. These ratings can have a significant impact on future business performance when prospective patrons first use Yelp's service and see the highly rated ones first. So it is worth investigating which business characteristics correlate with a high Yelp score, but what about data unaccounted for by Yelp such as proximity and geographic characteristics?

We decided to investigate the ratings of restaurants in the Las Vegas metropolitan area. We set out to first explore relationships between various restaurant characteristics such as whether or not it provides take-out, delivery, parking, credit card transactions, if they're good for kids, and if there is outdoor seating. We also looked at the number of Yelp check-ins, number of reviews, and price range. Beyond these variables intrinsic to a business, we examined features related to the neighborhood to which a restaurant belonged, also collected by Yelp. We examined correlations between these separately, and with our target variable: the Yelp star ratings.

The question we are aiming to answer is: **to what extent does the knowledge of a business's proximity and other geographic variables improve the accuracy of a model predicting its Yelp score?** For now we present an exploratory analysis comparing different characteristics of a restaurant, properties of the neighborhoods, and their Yelp ratings, number of checkins, and reviews.

Feature Extraction

The Yelp dataset was originally in a JSON format, which we flattened out into a Postgres instance extracting multiple features pertinent to our research. We included only businesses in Las Vegas, and focused our prediction of Yelp score to restaurants.

In summary, we were able to extract features like the business_id, city, state, neighborhood, category, accepts_credit_cards, good for kids, outdoor_seating, delivery option, price range (1-4),

reservations facility, takeout facility, parking facility, count of businesses in the neighborhood, diversity of business in the neighborhood', number of check ins, number of reviews posted and the Yelp Rating (stars) of a restaurant.

The number of reviews and check ins are continuous variables, neighborhood and price range are multiclass and all our other variables are categorical binary variables.

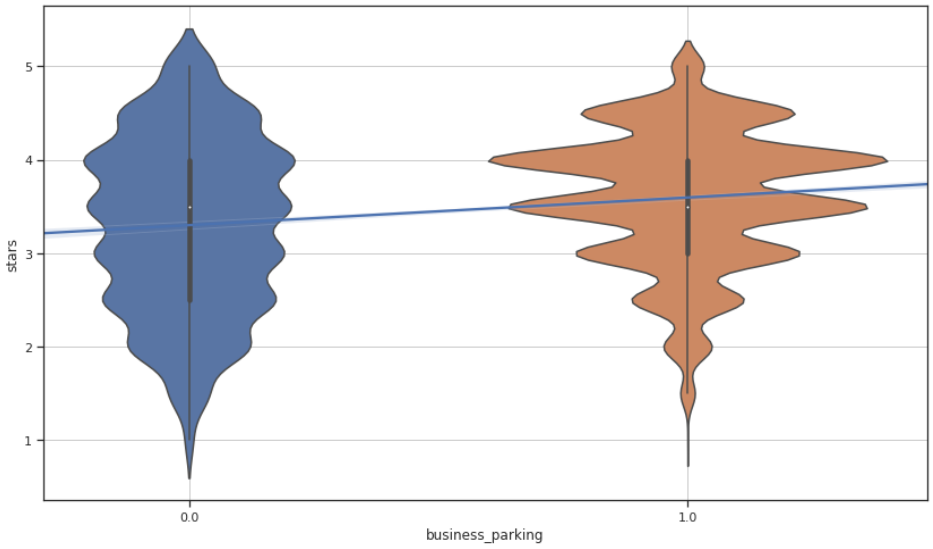
To handle the null values, we converted the categorical features into binary, and replaced the missing values with the mean of the column values. In the case of all 14 features we examined from Yelp, 700 out of the 6446 restaurants in Las Vegas were missing values.

All of these features were provided in the Yelp dataset, but we also computed the "business diversity" of each neighborhood. This was an aggregate of business category we defined using the gini index coefficient. The coefficient ranges from 0 to 1 and we normalized it so that 1 represents a perfectly diverse neighborhood (technically 1 - gini) where there are an equal number of businesses in each of the 22 different business types.

Preliminary Analysis:

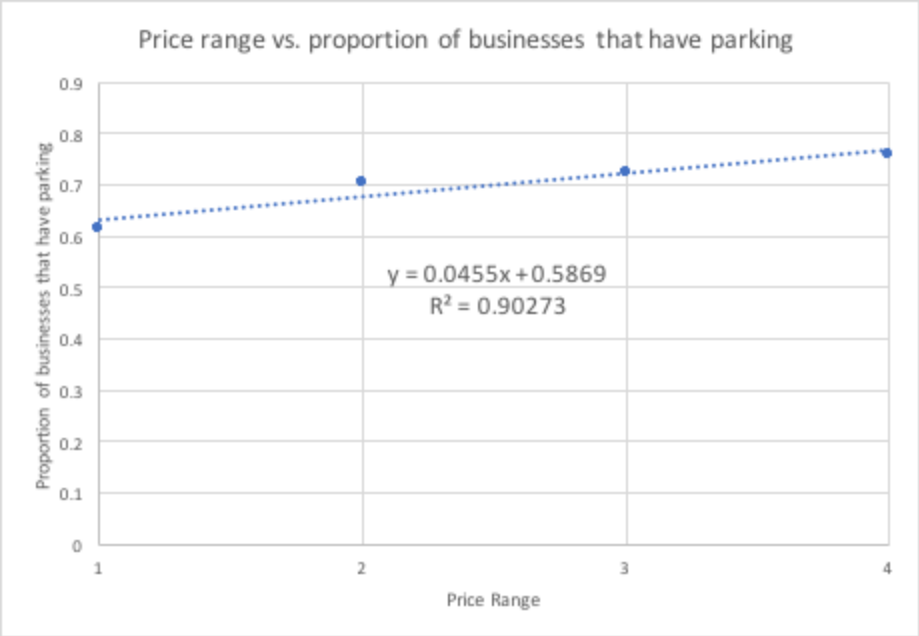
a) Relationship of Yelp Rating with business characteristics:

1. Yelp Rating vs Business Parking



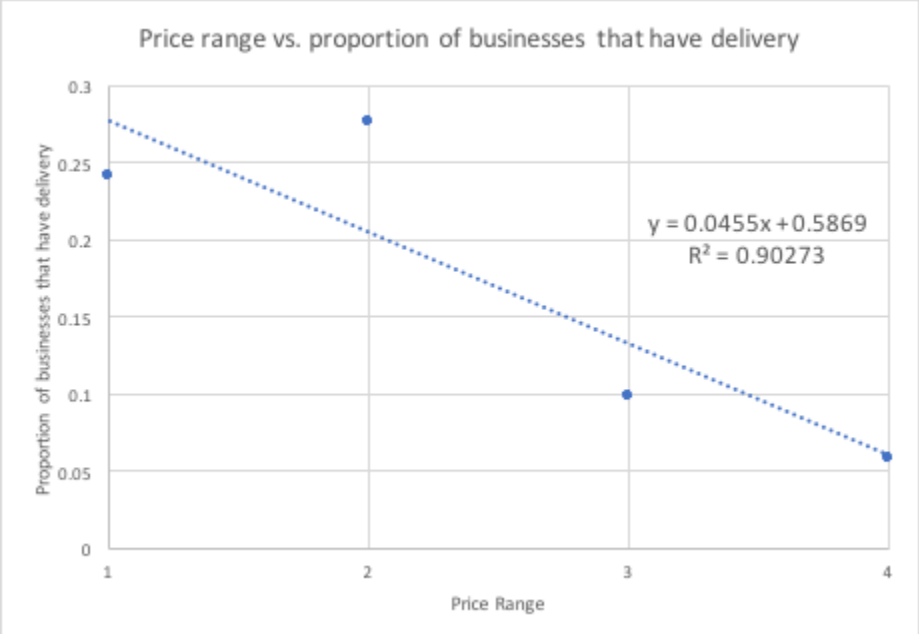
Businesses that have parking facilities tend to have a slightly higher yelp rating.

2. Proportion of businesses that have parking vs Price Range



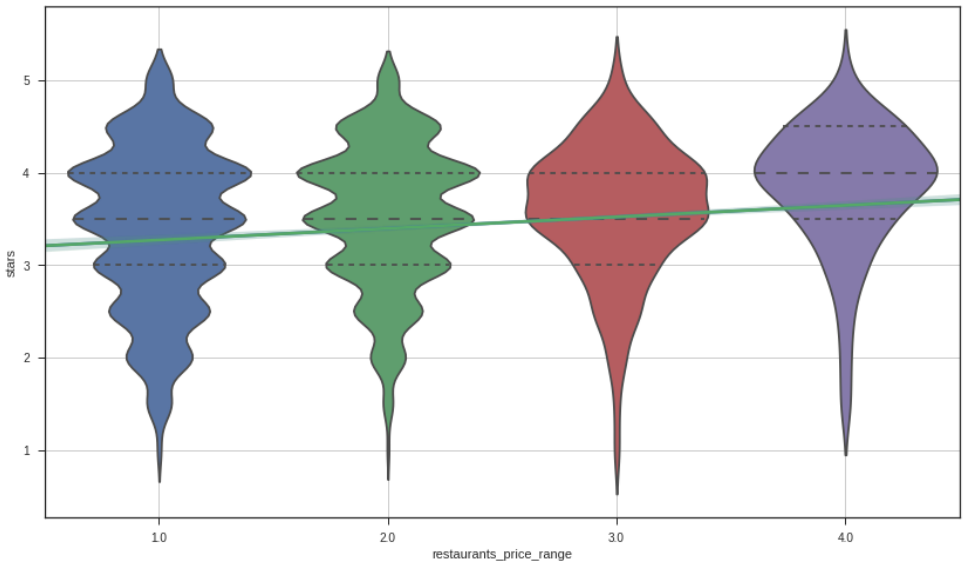
Businesses that have higher price range on average tend to have available parking.

3. Proportion of businesses with delivery vs Price Range



On average businesses that have lower price range tend to have delivery option available.

4. Variation of Business price range with Yelp Rating

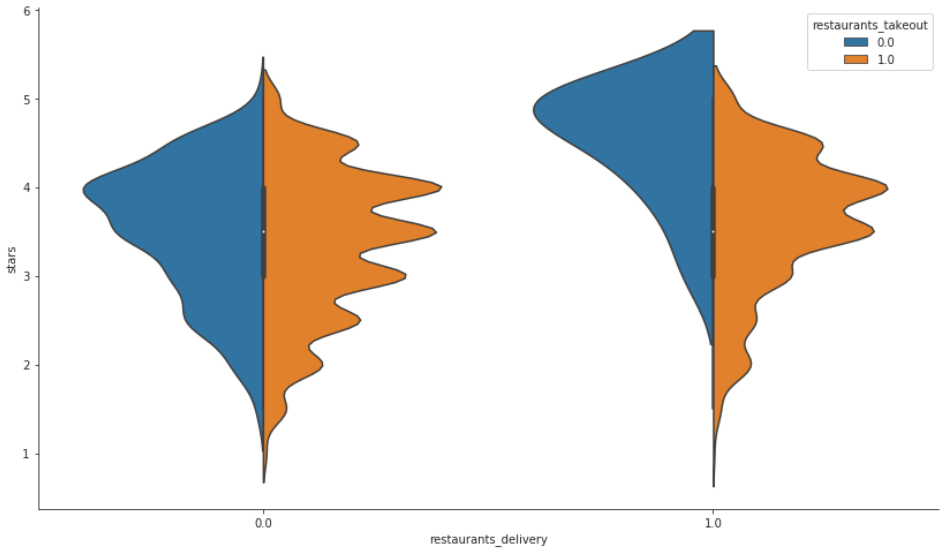


Linear Regression Coefficients:

slope: **0.12**
intercept: **3.27**
R-squared: 0.0096

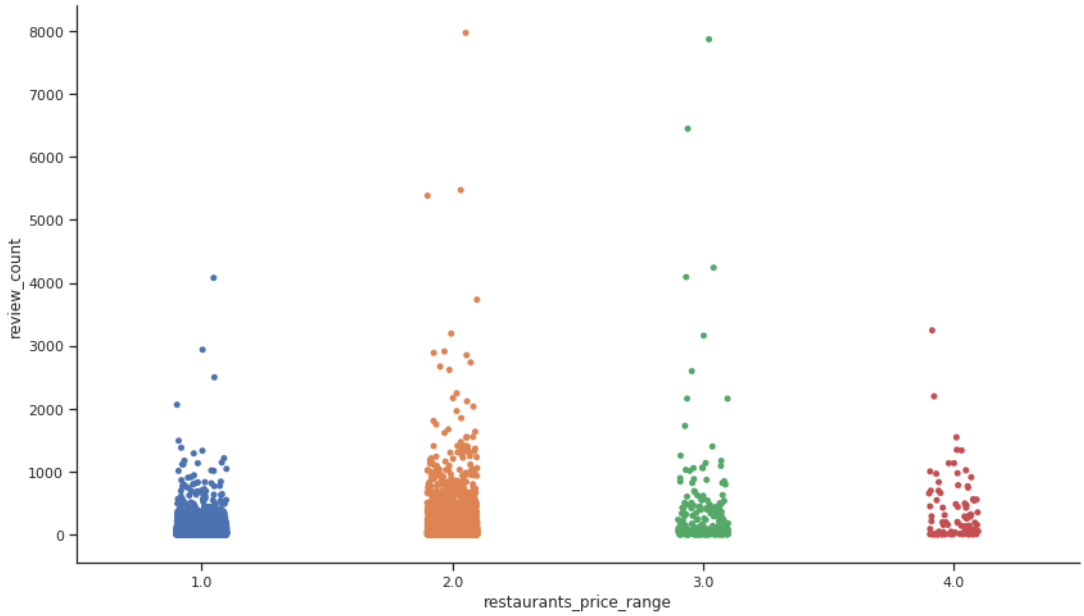
Restaurants that have a higher price range tend to have a higher average star rating, while lower price ranges tend to yield lower average star ratings.

5. Yelp Rating vs Delivery and Takeout option



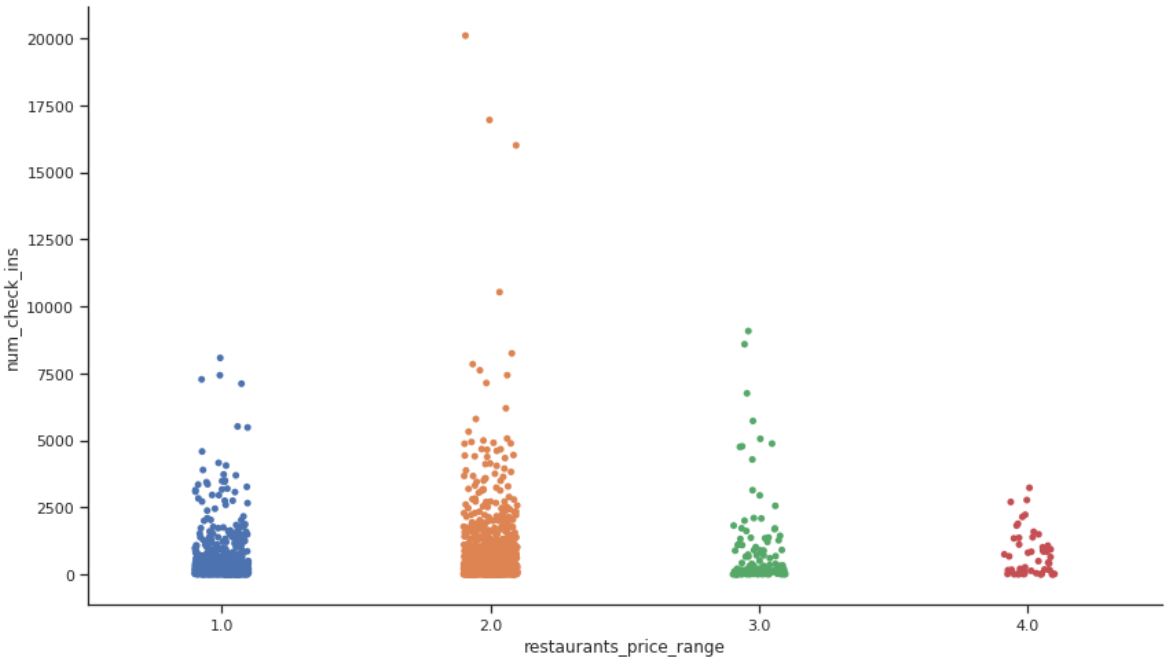
Businesses that have a delivery option and also a takeout option tend to have lower yelp rating, probably because it messes things up. Restaurants with no takeout but a delivery option have a higher yelp rating.

6. # reviews vs price range



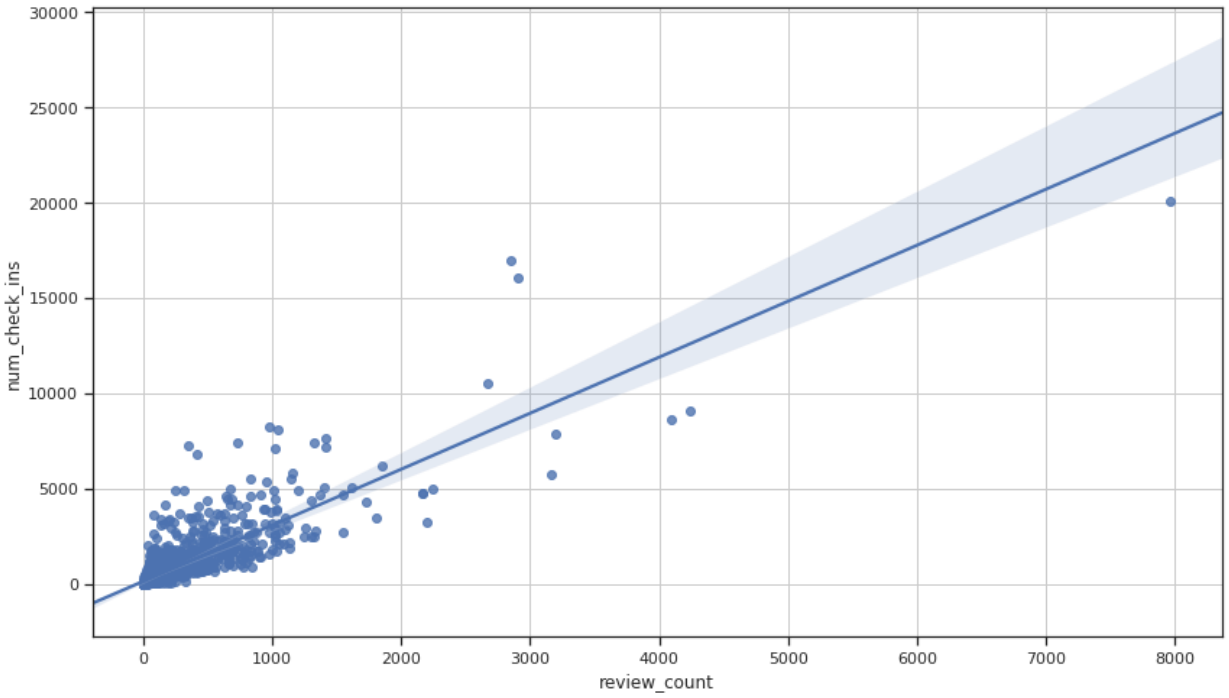
The count of number of reviews posted for each restaurant seems like a normal distribution over the price range. We see that most people post reviews for mid-segment restaurants, whereas for very cheap or very expensive restaurants, the number of reviews are comparatively less.

7. # check ins vs yelp rating



The number of checkins has a similar distribution over the price range as the number of reviews posted. It would be interesting to investigate the relationship between the number of reviews and number of checkins for a restaurant.

8. # check ins vs # reviews

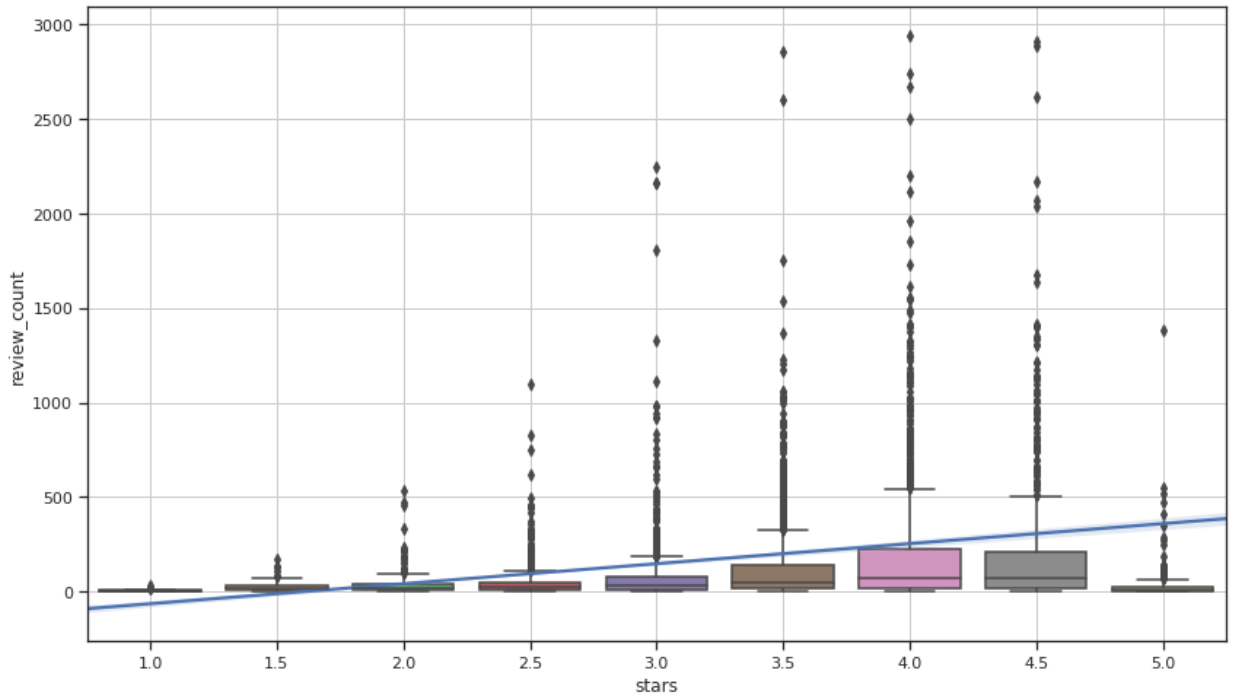


Linear regression Coefficients:

slope: **2.94**
intercept: **127.89**
R-squared: 0.7198

As expected, we observe a linear relationship between the number of checkins and the number of reviews. With each review, the number of checkins increase by ~3.

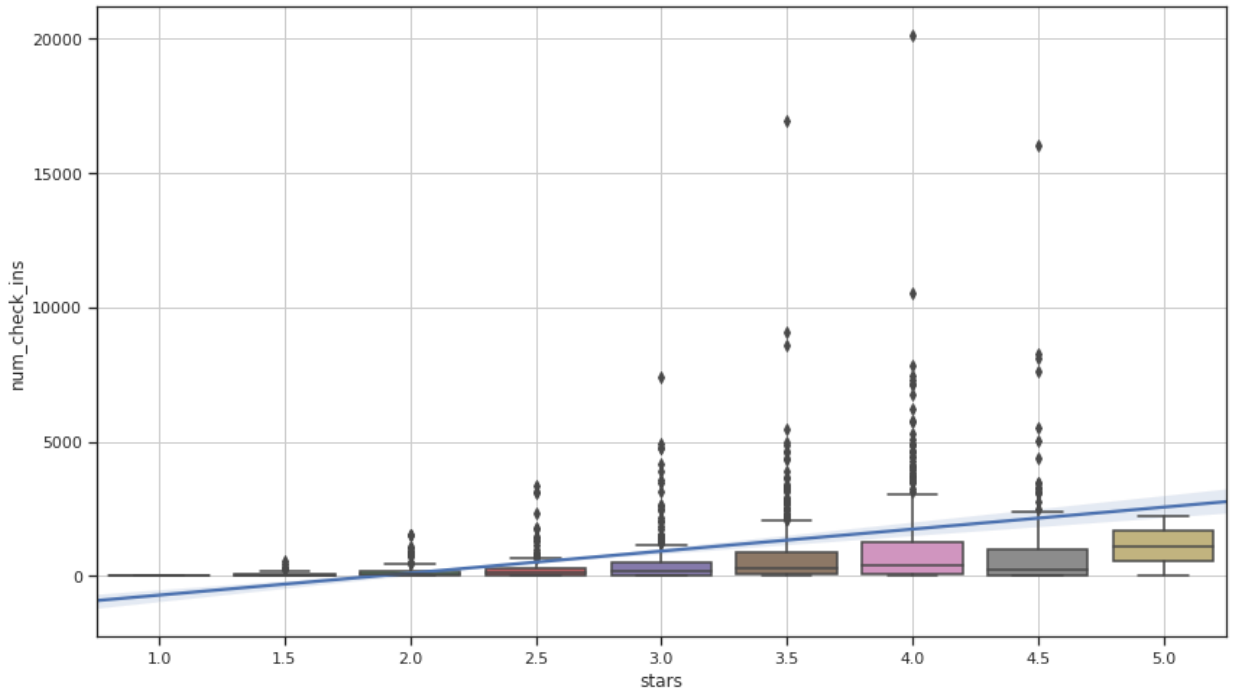
9. # reviews vs yelp rating



slope: **52.87**
intercept: **-62.75**
R-squared: 0.0355

To avoid outliers, we limited ourselves to restaurants that have less than or equal to 3000 reviews. However, most restaurants have reviews less than ~250. Interestingly, we observe that better restaurants (more yelp rating) have higher number of reviews.

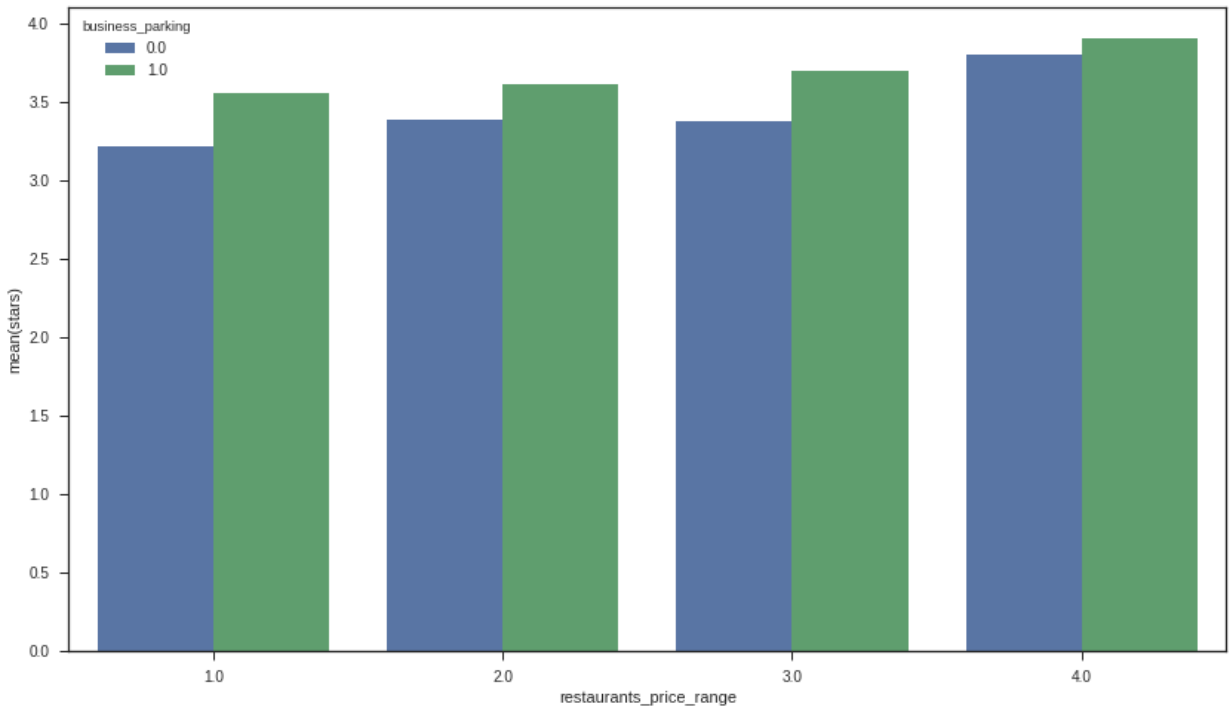
10. # checkins vs price range



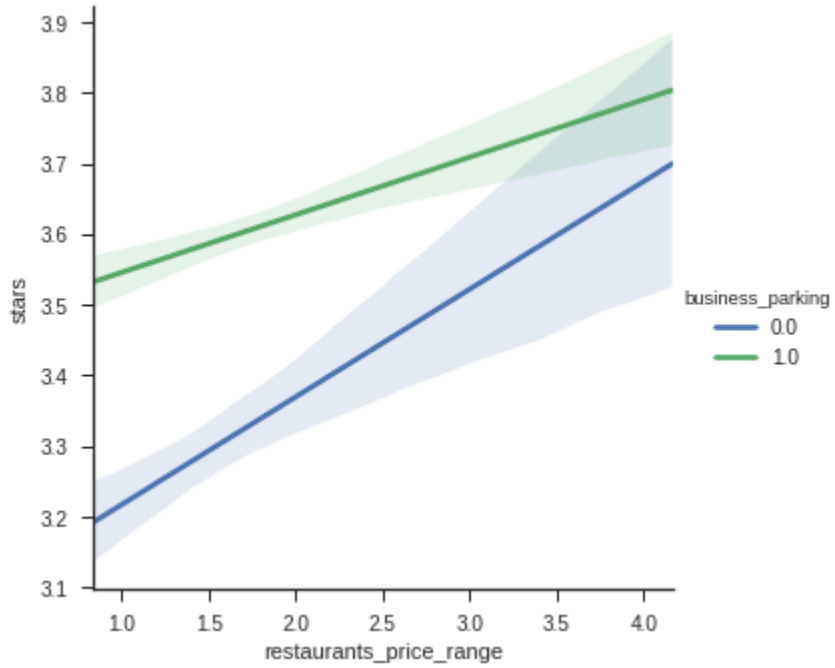
slope: **408.66**
intercept: **-706.83**
R-squared: 0.0535

If the average yelp rating goes higher by 1 point, we see about the number of checkins for a restaurant increases by ~400.

11. Variation of Price Range and Business Parking with Stars



Restaurants with higher price range tend to have higher yelp rating. Also, restaurants having business parking have slightly higher rating than the ones without parking for all price ranges.



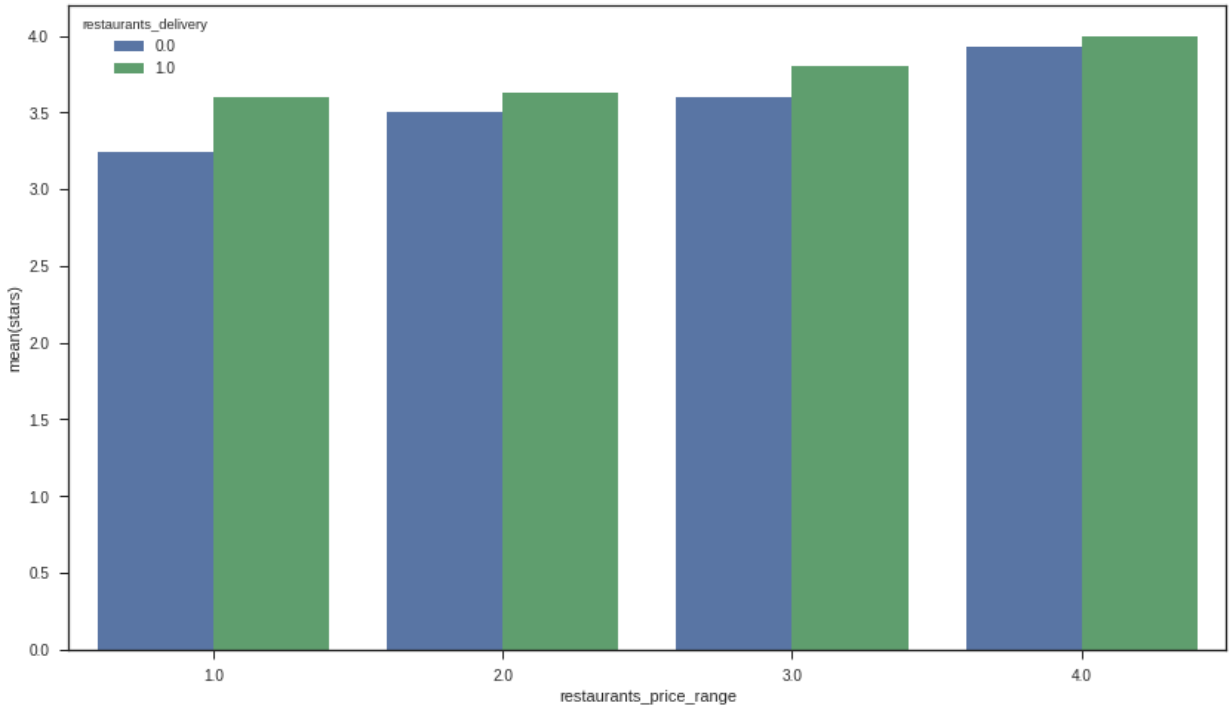
business_parking: **YES**
slope: **0.12**
intercept: **3.41**

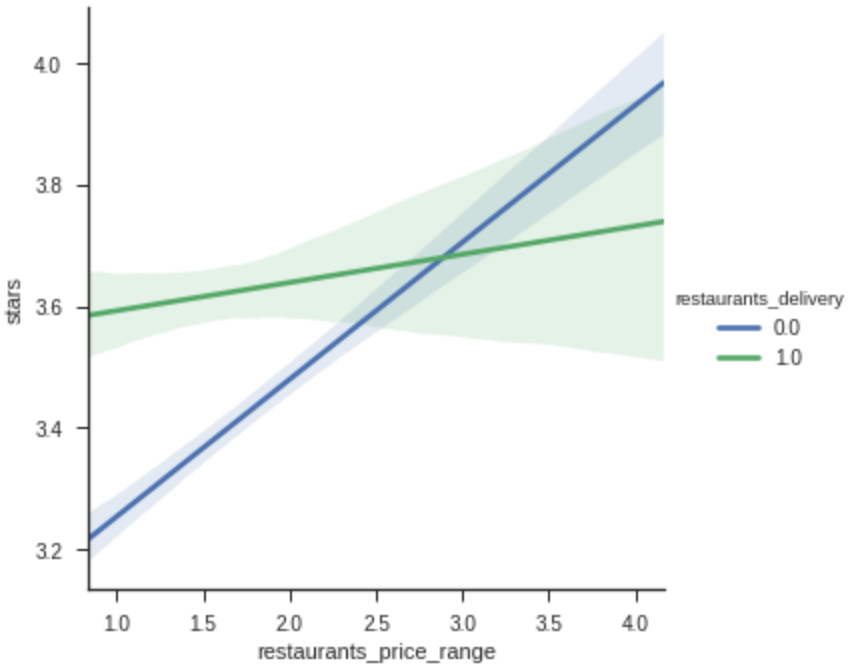
R-squared: **0.9122**

business_parking: **NO**
slope: **0.17**
intercept: **3.01**
R-squared: 0.8135

We can see from this regression that restaurants which have parking facility are to likely have a significantly higher star rating at the lower end of price ranges. But as you limit your attention to only more expensive restaurants, this gap narrows. High price range generally indicates a higher Yelp rating.

12. Variation of Price Range and Restaurants Delivery with Yelp Rating





Linear Regression

Delivery

slope:	0.14
intercept:	3.41

R-squared: 0.9273

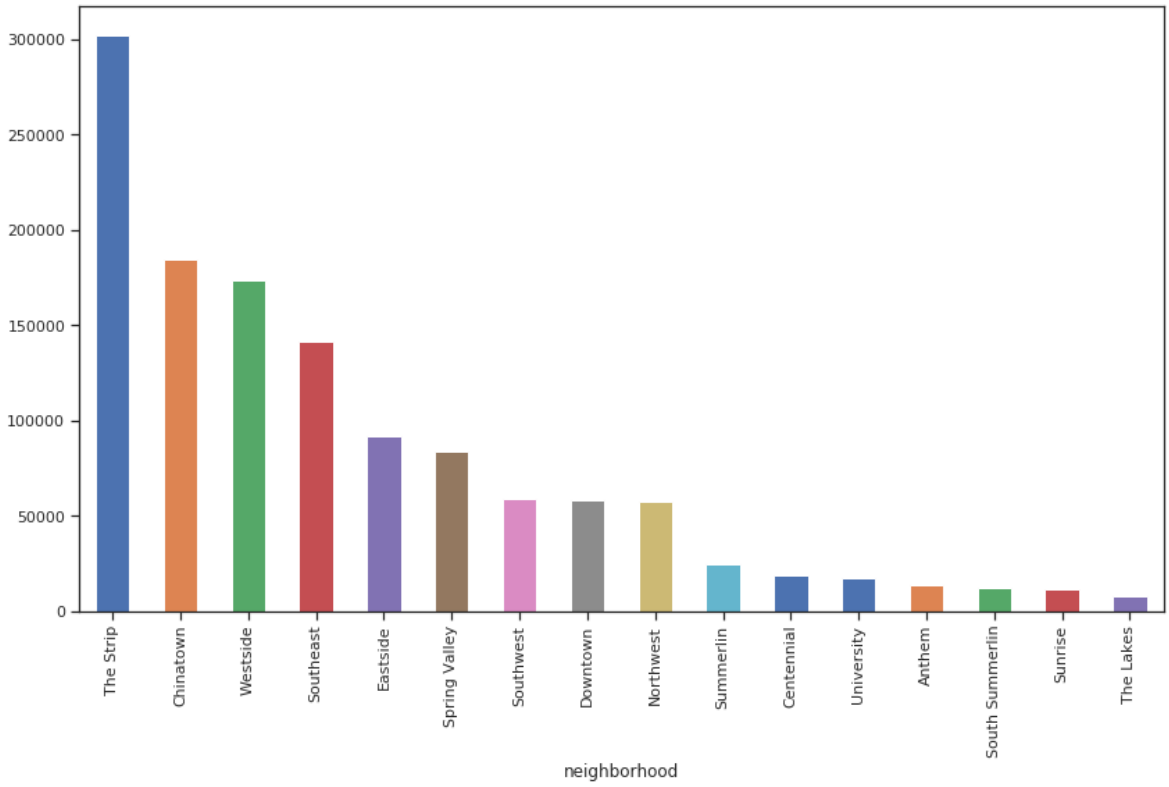
No	**Delivery**
slope:	0.21
intercept:	3.03

R-squared: 0.9663

We can see from this regression that restaurants which don't offer delivery service and have a high price range are likely have a significantly higher star rating. But there is little star rating variation across the price ranges for restaurants that have delivery.

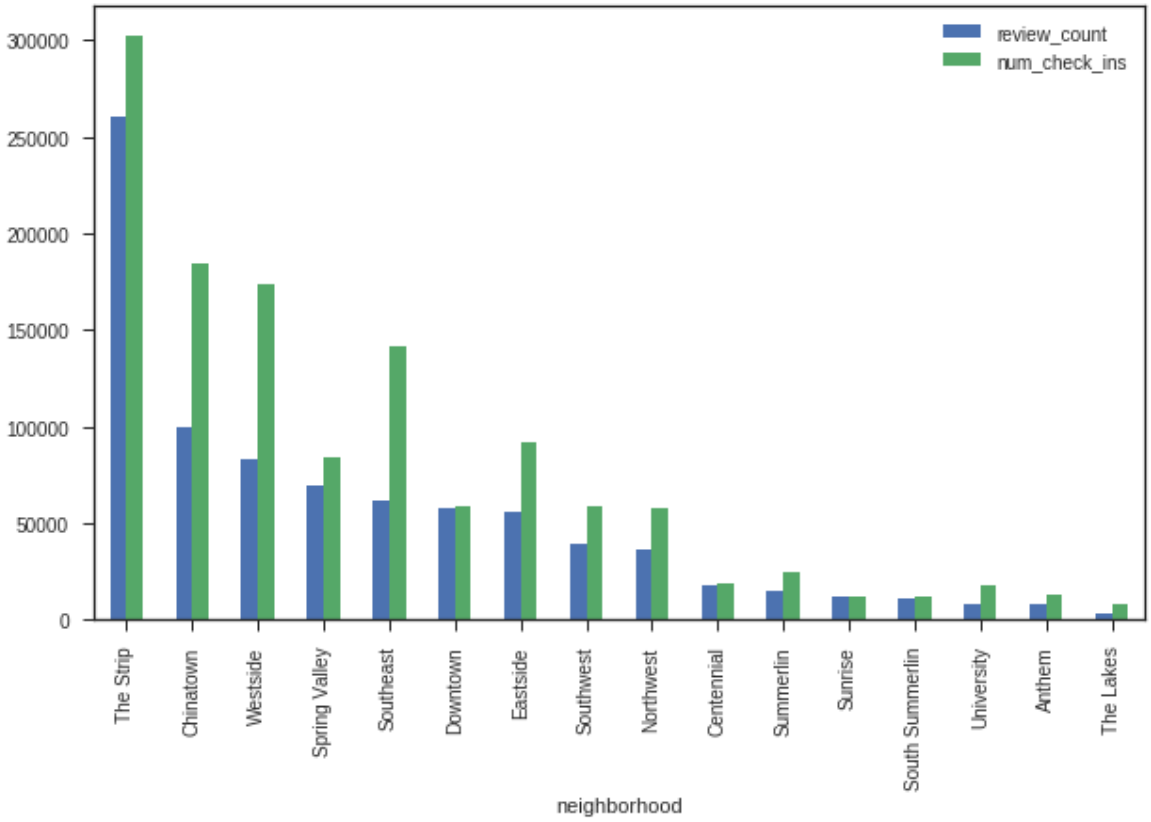
b) Relationship of Yelp Rating with neighborhood characteristics:

1. Neighborhood vs # businesses



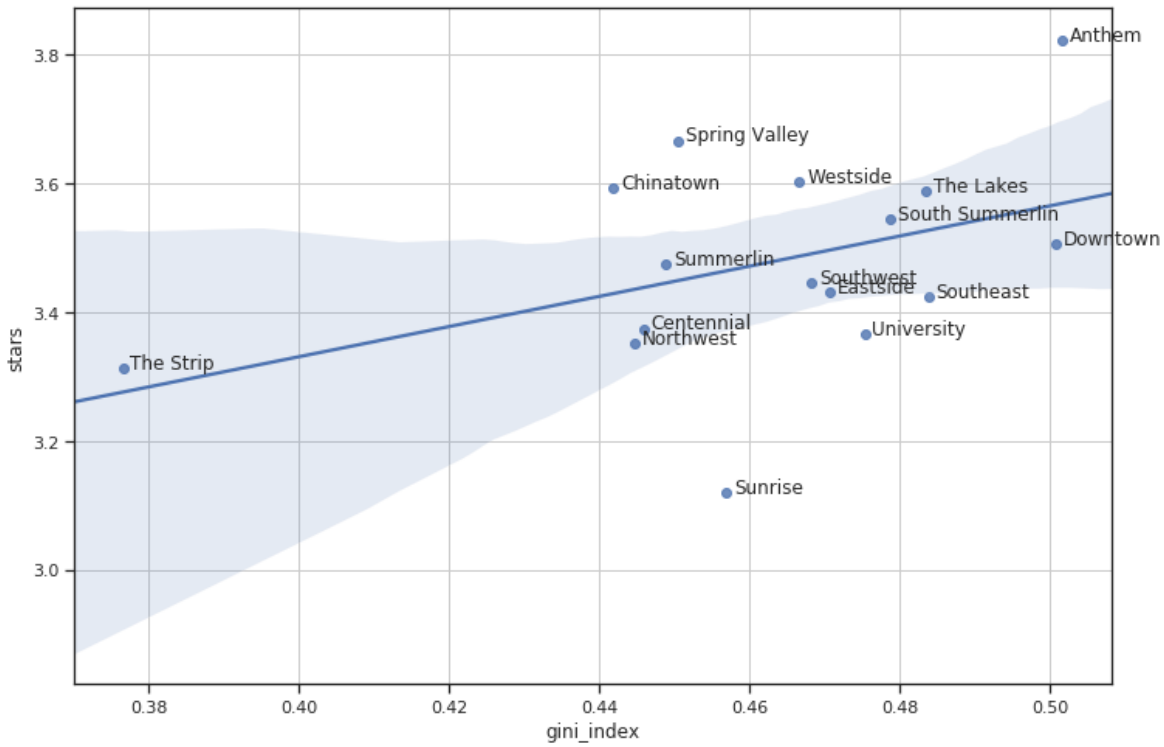
'The Las Vegas Strip' has the maximum number of restaurants (by almost a 100,000). It would be interesting to investigate the number of reviews and the check ins for these neighborhoods.

2. Neighborhood vs # check ins and reviews



Interestingly, the number of reviews and check ins also has a similar trend.

3. Neighborhood vs Diversity (Gini Index)



slope: **2.34**
intercept: **2.39**
R-squared: 0.1807

The gini_index is a measure of the diversity of a business in a certain neighborhood which is measured by computing the gini index. We observe that neighborhoods that have higher gini index (more diverse business types) tend to have higher star ratings. Previously, we observed that 'The Strip' has the highest number of restaurants, reviews and check ins, but it is not very diverse in terms of other businesses in the neighborhood. As a result, its collective yelp rating is almost the second lowest in Las Vegas.

Project Plan:

Some additional data modeling, analysis, and model comparison is necessary for finishing this project. Presently, Yelp's provided assignments of each business to a "neighborhood" is our only measurement of proximity. We plan on expanding upon this by projecting the latitude and longitude coordinates of each business into projected X and Y coordinates in the East Nevada Stateplane coordinate reference system. Then we can run a *k-nearest neighbors* or similar clustering algorithm, and investigate the variance in Yelp score of such clusters and relationships with other covariates.

When developing k-nearest neighbors we will choose an optimal k based on the "elbow method." When all businesses have been assigned to an unsupervised knn cluster, we will compute the aggregate business level features, counts of each business category, check-ins, reviews, and the gini index of business category diversity. **Richard** will handle transformations, geospatial mapping by November 16, and the k-NN clustering by November 23rd.

Our question, restated:

To what extent does the knowledge of a business's proximity and other geographic variables improve the accuracy of a model predicting its Yelp score?

In order to answer this we will build three models:

1. Linear regressions of Yelp score on all business characteristics with interactions
2. … with aggregate proximity features based on Yelp's neighborhood labels
3. … with aggregate proximity features based on the k-NN unsupervised labels (using an optimal k)

Shreya will compute and maintain the implementations of each of these three models and compare their results. This will be ongoing after completion of k-NN, but models 1 and 2 will be complete by November 23rd, and model 3 will be complete by November 28th.

We'll also try building decision tree classifiers and create decision tree diagrams showing the feature splits and their proportions at each step. Many of our features are categorical, so decision tree classification makes sense. **Amir** will build these decision trees by November 28th.

Using Tableau, we will make colored scatter plots overlaying the business point locations on top of a map of Las Vegas to compare the Yelp labeled neighborhoods to our knn resulting clusters. Some of these neighborhoods will likely be overlapped by multiple clusters. Further analysis may reveal *why* these clusters have their distributions in the response variable and aggregate business level variables. An example might be: "cluster A is so heterogeneous with a low diversity score because it lies entirely on the strip, which is filled with mostly restaurants all along a thin width." **Sahil** will conduct this analysis by November 28th.

-- Visualize diversity of different nbds. Hover over each nbd, see diversity index
Keep diversity for each nbd? Or for each cluster?

Discrete stars