Group 17 Project Report: Yelp Data Challenge

Our analysis consists of two parts: ratings prediction and visualizing food trends around the world.

Data Cleaning

We have been primarily working with two parts of the Yelp dataset: the reviews and the businesses. These two files were originally in JSON format, so we first had to convert them into csv file format to be able to work with R for the analysis portion of the project. To do this conversion, we used Python scripts. In addition to converting the JSON files into csv format, we also converted the business JSON file into a SQL table with another Python script in order to run queries on it. For the business SQL table, we focused on businesses that were restaurants. Even when controlling for only restaurants, we still had over 40,000 businesses in the table and were able filter out restaurant by food types, cities, and price ranges for further analysis.

Because of the large size of the reviews JSON files, we decided that we would limit most of our analysis to cities in the United States, which is a sample of 7 cities (Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, Cleveland). However, for other parts of the project such as most the popular foods reviewed in each territory and the price ranges of foods in the most popular reviews, we were able to incorporate other countries as well since these kinds of analysis could be done from just the business table.

Analysis

1. Ratings Prediction

We aim to predict the ratings of a review based on the text of the review. As the data is very big, we restrict our analysis to restaurants in Charlotte, which has a total of 122,322 reviews.

We choose our features to be words that are the most predictive of the rating, i.e. words that convey a strong negative or positive sentiment such as “great” and “awful”.

First, we extract salient words from all reviews by calculating the TF/IDF score for all words.

Second, we matched our list of words with a list of words which convey strong positive and negative sentiments provided by the Multi-Perspective Question Answering (MPQA) Subjectivity Lexicon at University of Pittsburgh (http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/). There are a total of 2,874 words in our reviews that convey strong positive and negative sentiments. To reduce the dimension of our feature space,we choose 1,000 words with the

highest TF/IDF score as some of the words with low TF/IDF score, such as “ignominiously” and “sanguine” do not appear in many reviews would not be good features for prediction. **Figure 1** shows word clouds of 200 positive words and 200 negative words with the highest TF/IDF. Some of the highest score negative words were: *little, bad, disappointed, long, extremely, bland, rude, terrible, cheap.* Some of the highest score positive words were: *great, amazing, awesome, love, best, friendly, nice, delicious.*

Figure 1: Word clouds of 200 positive words and 200 negative words with the highest TF/IDF.

Third, we use this list of 1,000 words to create a matrix of counts, i.e. we count the number of times each word appears in each review. We end up with a 122,322 by 1,000 sparse matrix. We filter out the all-zero rows. **Figure 2** shows a histogram of the numbers of feature words each review contains. We can see that most review contains fewer than 10 of the feature words. The feature matrix is sparse.

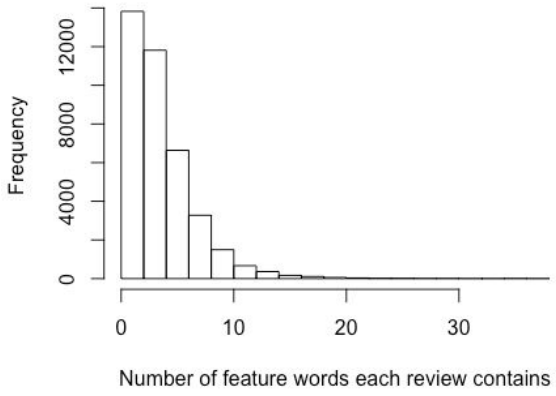


Figure 2: Histogram of the number of feature words each review contains

We divide the data into 90% training set and 10% test set. We end up with 34,638 observations in the training set and 3,848 observations in the test set. We used the following regression algorithms:

(A) Linear regression

(B) Linear regression with elastic net regularization. We use cross validation to choose the optimal penalty coefficient lambda 0.015 (See *Appendix A*). We use a mix of L1 and L2 (alpha=0.7). The algorithm selects 67 features out of the total of 1000 features. Some of the words with the highest coefficients are Amazing, Delicious, Excellent, Best, Favorite, Awesome, Perfect, Creative, Fantastic, Great, Incredible, Love. Some of the words with the lowest coefficients are Displeasure, Worst, Lure, Horrible, Terrible, Rude, Disgusted, Joke, Worse, Insult, Ache, Lousy, Nasty, Filthy, Awful. **Figure 3** shows all 67 features.

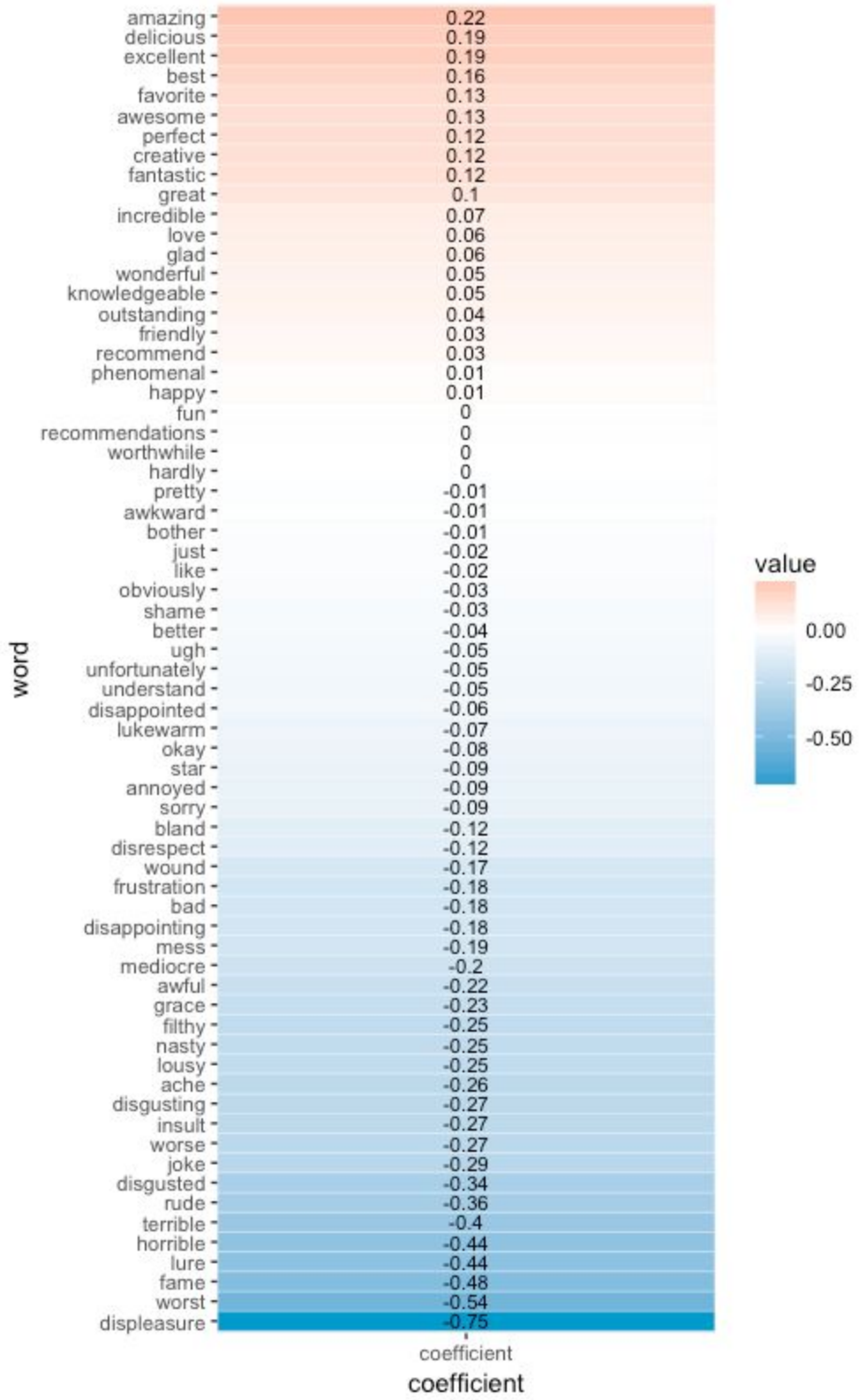


Figure 3: coefficients from linear regression with regularization

(C) Decision Tree (CART). We use the CART algorithm. The algorithm produced the following splits:

Primary splits:

worst < 0.5 to the right, improve=0.010449640, (0 missing)

great < 0.5 to the left, improve=0.007582825, (0 missing)

amazing < 0.5 to the left, improve=0.007376710, (0 missing)

delicious < 0.5 to the left, improve=0.007313671, (0 missing)

terrible < 0.5 to the right, improve=0.006925437, (0 missing)

Surrogate splits:

kindly < 1.5 to the right, agree=0.983, adj=0.003, (0 split)

(D) **K-Nearest Neighbors.** We use cross validation to choose the optimal number of neighbors considered. We choose k=30 (See *Appendix B*).

(E) **Naive Bayes**. We treat the ratings as categories 1-5 and use Naive Bayes to classify the ratings.

(F) **Support Vector Machines**: We use the radial basis kernel radial basis *exp(-gamma*|u-v|^2)* where gamma=1/data dimension=0.001, and cost coefficient C=1.

(G) **Random Forest**: We treat the outcome as a 5-level categorical variable with 500 trees. RF outputs variable importance measure, defined as the total decrease in node impurities from splitting on the variable, averaged over all trees. For classification, the node impurity is measured by the Gini index. **Figure 4** shows the 50 variables that produced the biggest mean decrease in Gini index. The most important variables are “great”, “like”, “just”, “best”, “amazing”.

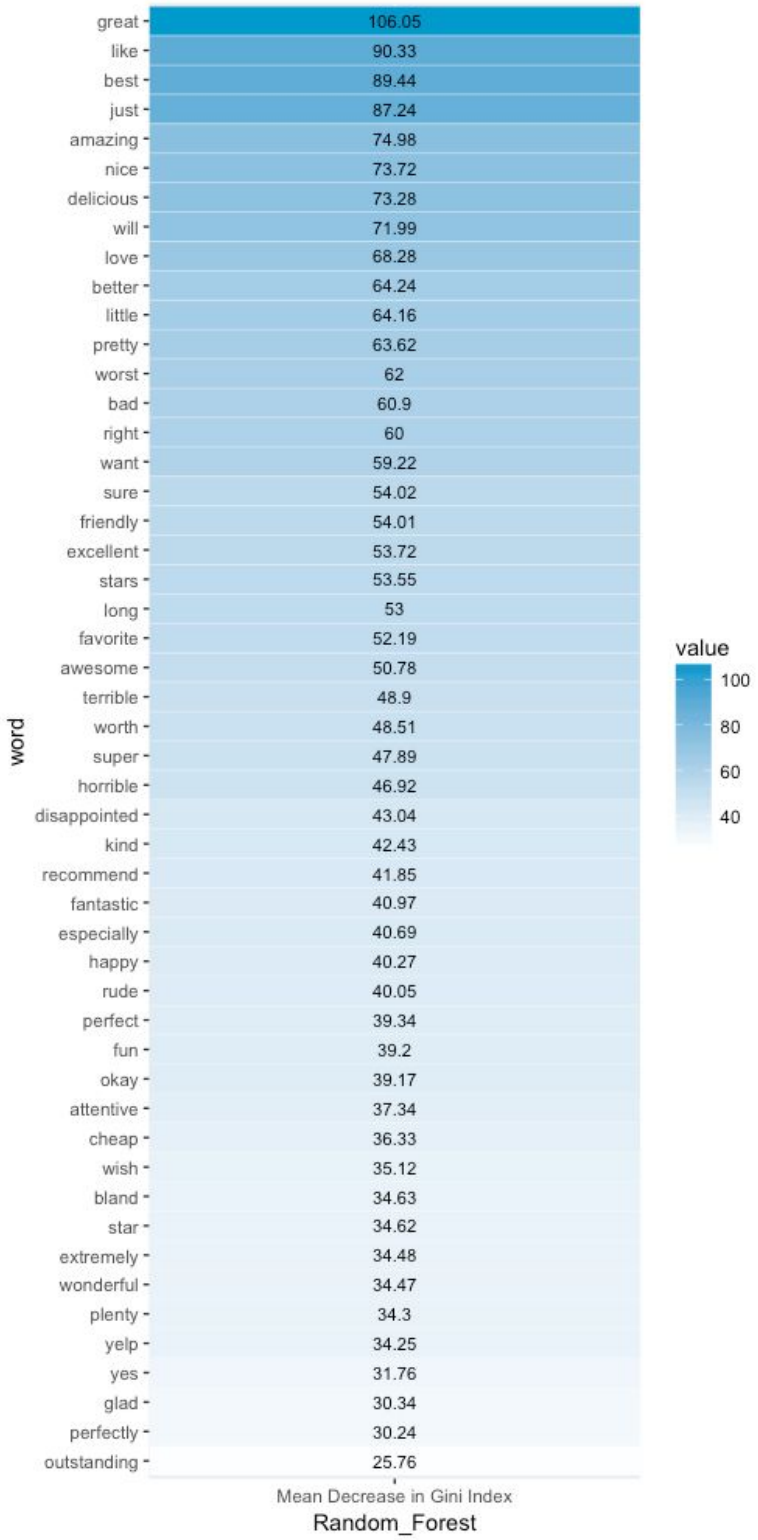
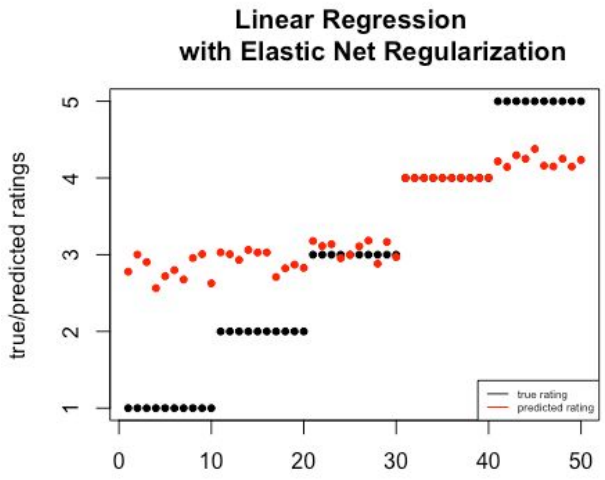


Figure 4: 50 variables that produced the highest mean decrease in Gini Index (Random Forest)

We assess the performance of the algorithms by calculating the Root Mean Square Error (RMSE) between predicted outcome and observed outcome. The result is summarized in the table below. The table shows that regression with regularization performs the best, i.e. has the lowest RMSE.

For Linear Regression with Regularization and Random Forest, we plot the predicted and observed values for randomly sampled 50 test data points, 10 from each rating category 1-5 in **Figure 5.** The figure shows that LR can predict for categories 3 and 4 very well, but not for 1, 2 and 5. RF generally overestimates the rating category. The reason might be that, in our dataset, we have more ratings of 3,4 and 5. See **Figure 6** for counts of reviews for each rating category.

	LR	LR (rglr)	CART	KNN	NB	SVM	RF
RMSE	1.29	1.28	1.31	1.29	2.19	1.31	1.62



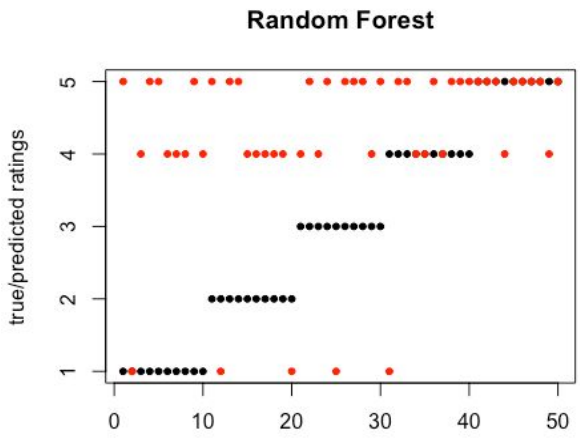


Figure 5: Predicted and observed values for 50 test data points

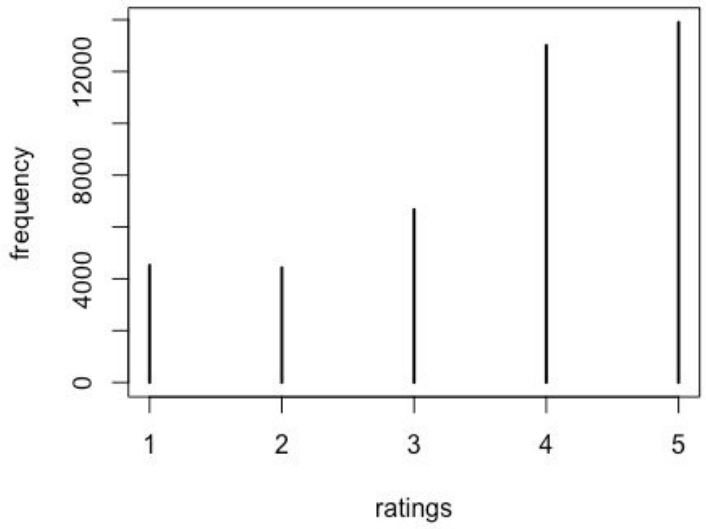


Figure 6: Number of reviews for each rating category

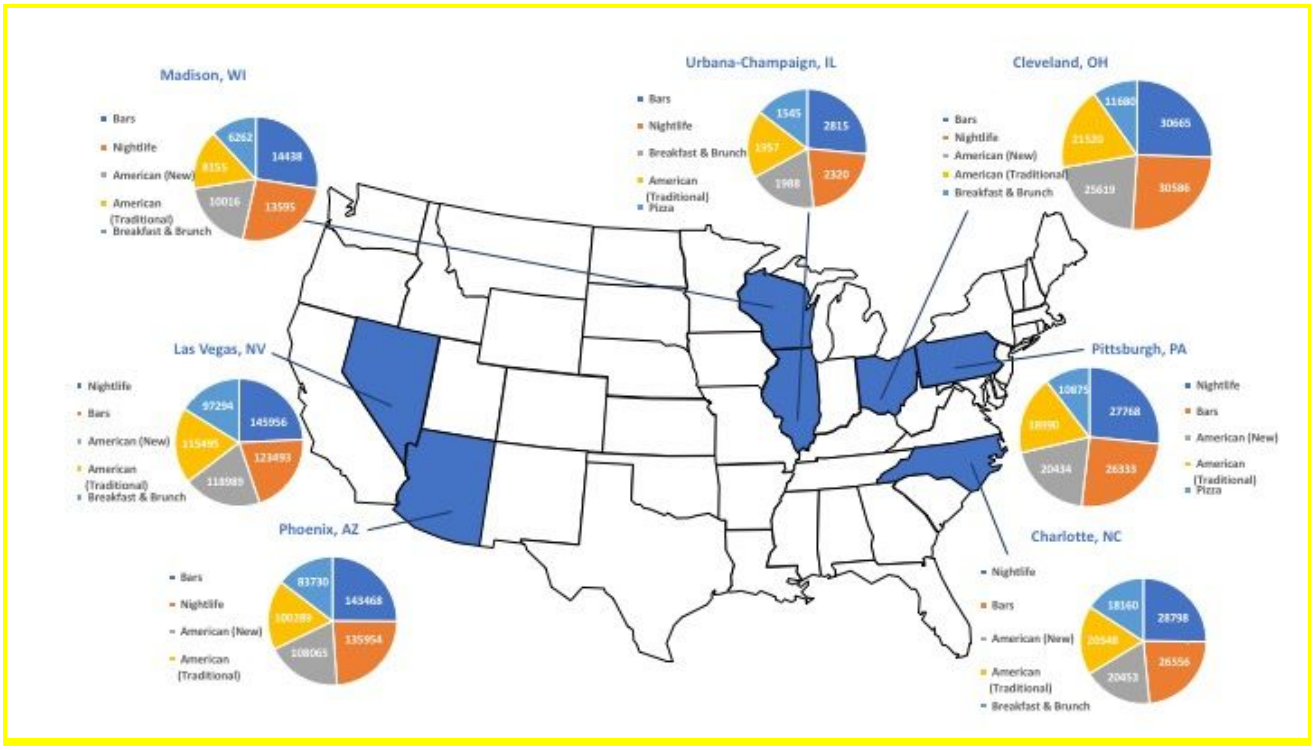
To conclude findings of our ratings prediction, Linear regression with regularization performs the best in terms of Root Mean Square Error (RMSE). It can predict very well for categories 3 and 4, but not for 1, 2 and 5. Our analysis can be improved by using classification for multiple labels, treating the outcome as categorical rather than numeric variable. Our analysis can also be improved by including more reviews--we were only able to include 40K reviews due to time constraint.

2. Visualize food trends around the world

We visualize what food categories were most popular for each state, as well we for different countries in the Yelp dataset. We visualize the food preference in US cities (by state) and international cities (by country) through pie charts and geographical maps. We define most popular as having the most reviews and also the highest average stars ratings. We also visualize how many people review cheap and expensive restaurants by state and country.

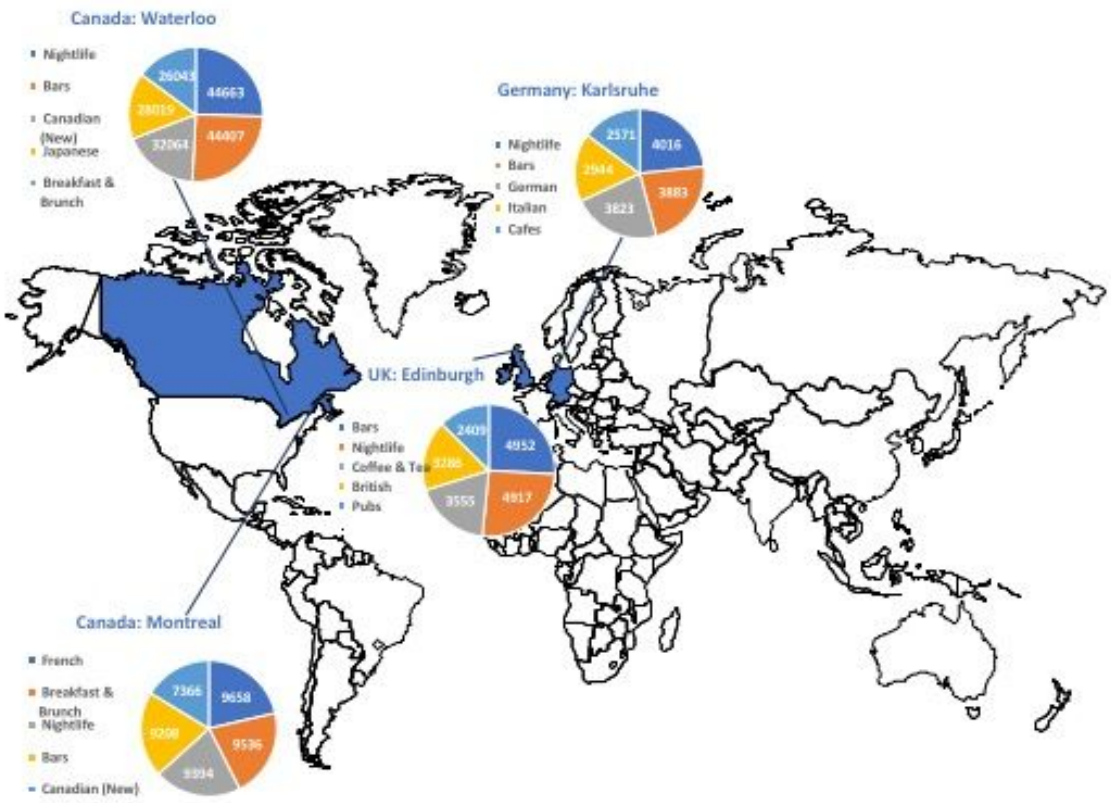
Here is a summary of interesting findings:
US:

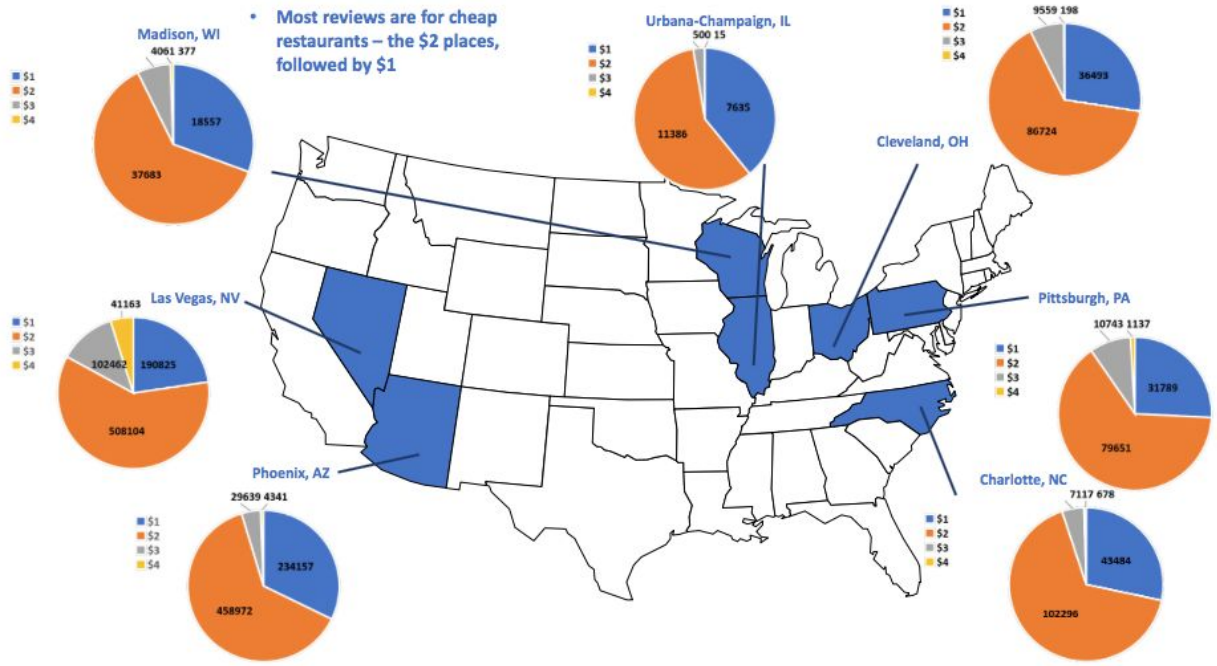
- Americans Yelp a lot about “breakfast & brunch”
- Food preference is relatively similar across the US (compared to food preference across different countries) - favorite categories are “bar”, “nightlife”, “American (new and traditional)”, “breakfast & brunch”, “pizza”
- Urbana-Champaign (Chicago area) and Pittsburgh like Pizza more than other cities
- “Nightlife” is more well-liked than “Bar” in Las Vegas, Charlotte, and Pittsburgh

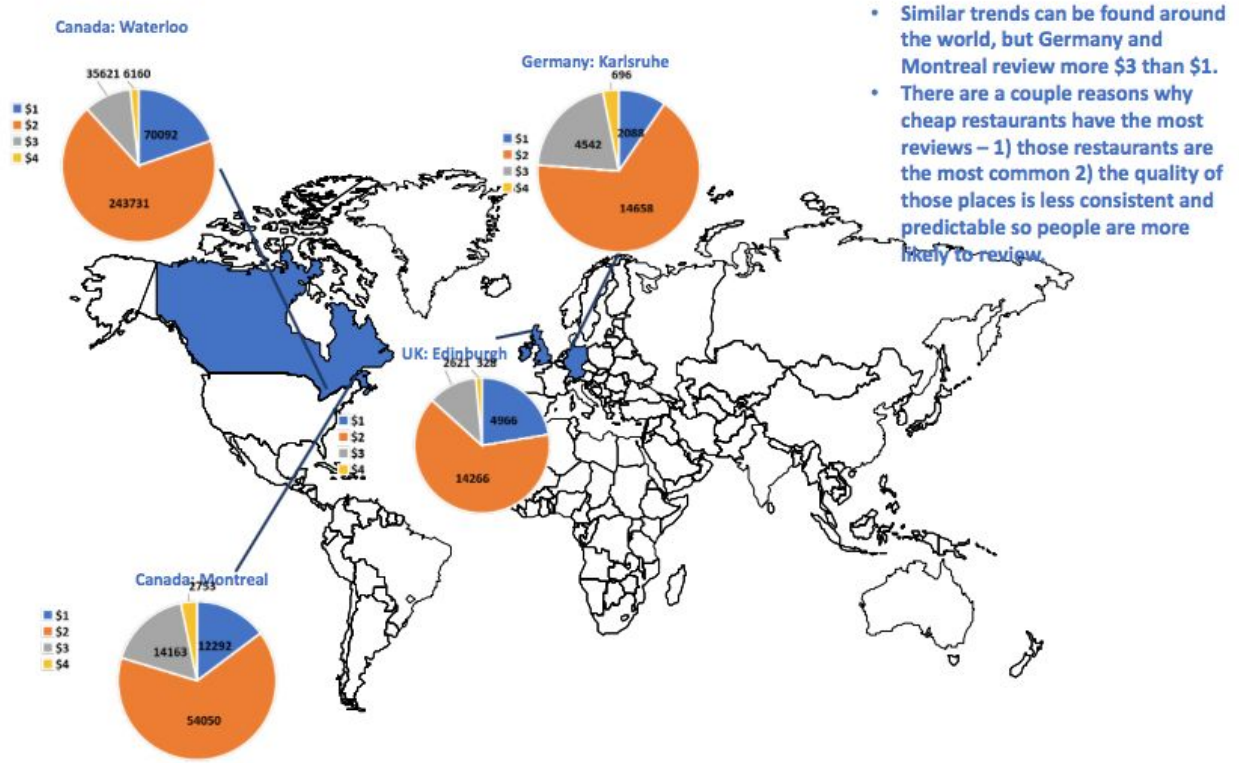


International:

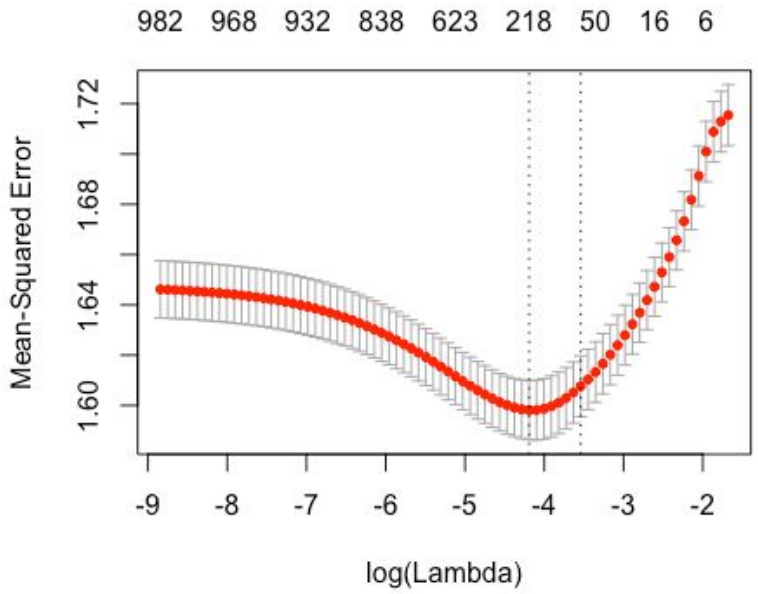
- Nightlife is popular in all the cities except Montreal
- Europeans don't like breakfast at all
- Canadians also like breakfast and brunch a lot (similar to the US)
- Waterloo, ON likes Japanese food even more than breakfast
- Germans like Italian food (but not as much as German food)
- Unsurprisingly: Edinburgh has a separate "pub" category besides "bars", Montreal favors French food, Edinburgh likes tea, Germans and Brits drink a lot…




Most reviews are for cheap restaurants – the $2 places, followed by $1
Madison, WI
4061 377
18557
37683
Urbana-Champaign, IL
500 15
7635
11386
9559 198
36493
86724
Cleveland, OH
Las Vegas, NV
41163
190825
102462
508104
Pittsburgh, PA
10743 1137
31789
79651
Phoenix, AZ
29639 4341
234157
458972
Charlotte, NC
7117 678
43484
102296
$1
$2
$3
$4


Similar trends can be found around the world, but Germany and Montreal review more $3 than $1.
There are a couple reasons why cheap restaurants have the most reviews – 1) those restaurants are the most common 2) the quality of those places is less consistent and predictable so people are more likely to review.
Canada: Waterloo
35621 6160
70092
243731
Germany: Karlsruhe
696
4542
2088
14658
UK: Edinburgh
4966
14266
Canada: Montreal
2733
14163
12292
54050
$1
$2
$3
$4

Appendix A: Choose the optimal lambda for linear regression with regularization



Appendix B: Choose the optimal K for KNN algorithm