# Group 17

Nick Ngare, Yuliya Kozina, Jimmy Zhang, Wendy Lu, Angie Shen, Bao Doan

# The Motivation

- We all loved food and wanted to learn how others thought about it
- What can we learn by analyzing people’s reviews?
    - Could predict the ratings of a review based on the words and the sentiments they convey?
- How do these findings vary geographically?

# Getting the Data

- We downloaded the JSON files from yelp and then converted them into CSV files or SQL tables using Python scripts
- limit our analysis to cities in the United States, which is a sample of 7 cities (Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, Cleveland).

# Ratings Prediction

- Goal: predict the ratings of a review based on the text of the review.
- Scope: restaurants in Charlotte, 122,322 reviews
- Features: 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”.
- Response: rating of the review (1-5)

# Data Set-up

- Extract salient words from all reviews by calculating the TF/IDF score for all words
- Match 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
- 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.
- Create a matrix of counts, i.e. count the number of times each word appears in each review. We end up with a 122,322 by 1,000 sparse matrix.

# Word Clouds

200 Negative Words with Highest TF/IDF


200 Positive Words with Highest TF/IDF


Highest TF/IDF

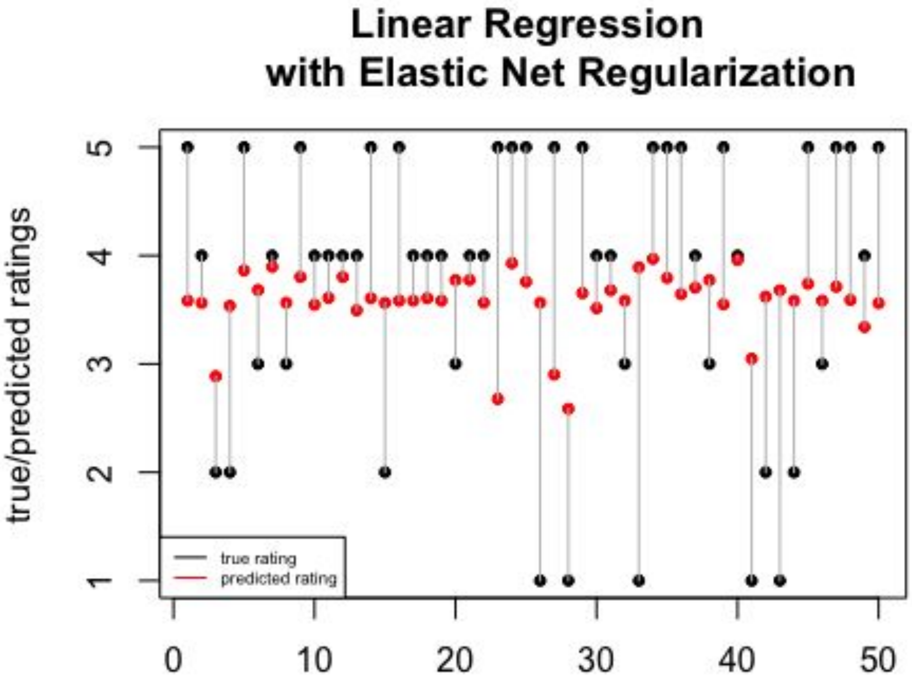
# Algorithms

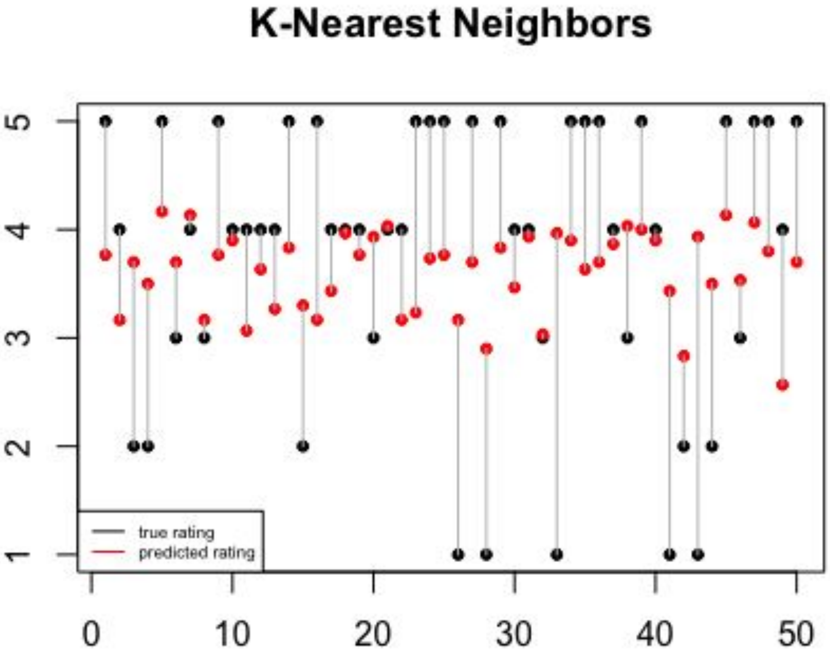
- 34K training set,  3K test set
- Linear Regression
- Linear Regression with Regularization (elastic net)
- Decision Trees (CART)
- K Nearest Neighbors
- Naive Bayes
- Support Vector Machines
- Random Forests

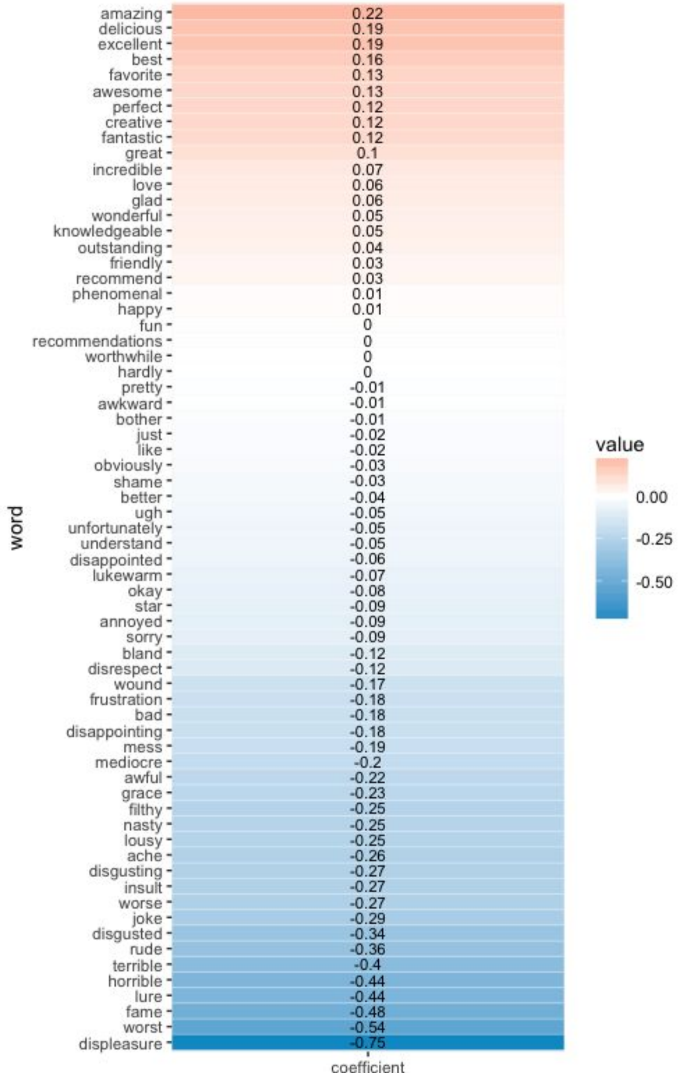
## Predictive Performance

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

# Visualize Prediction Performance






## Model Parameters of Linear Regression with Elastic Net Regularization

- Induce sparsity (selected 67 features out of 1000)
- Top positive words: Amazing, Delicious, Excellent, Best, Favorite, Awesome, Perfect, Creative, Fantastic, Great, Incredible, Love
- Top negative words: Displeasure, Worst, Lure, Horrible, Terrible, Rude, Disgusted, Joke, Worse, Insult, Ache, Lousy, Nasty, Filthy, Awful
- Note that the regularization resolves multi-colinearity: words that are highly correlated are removed

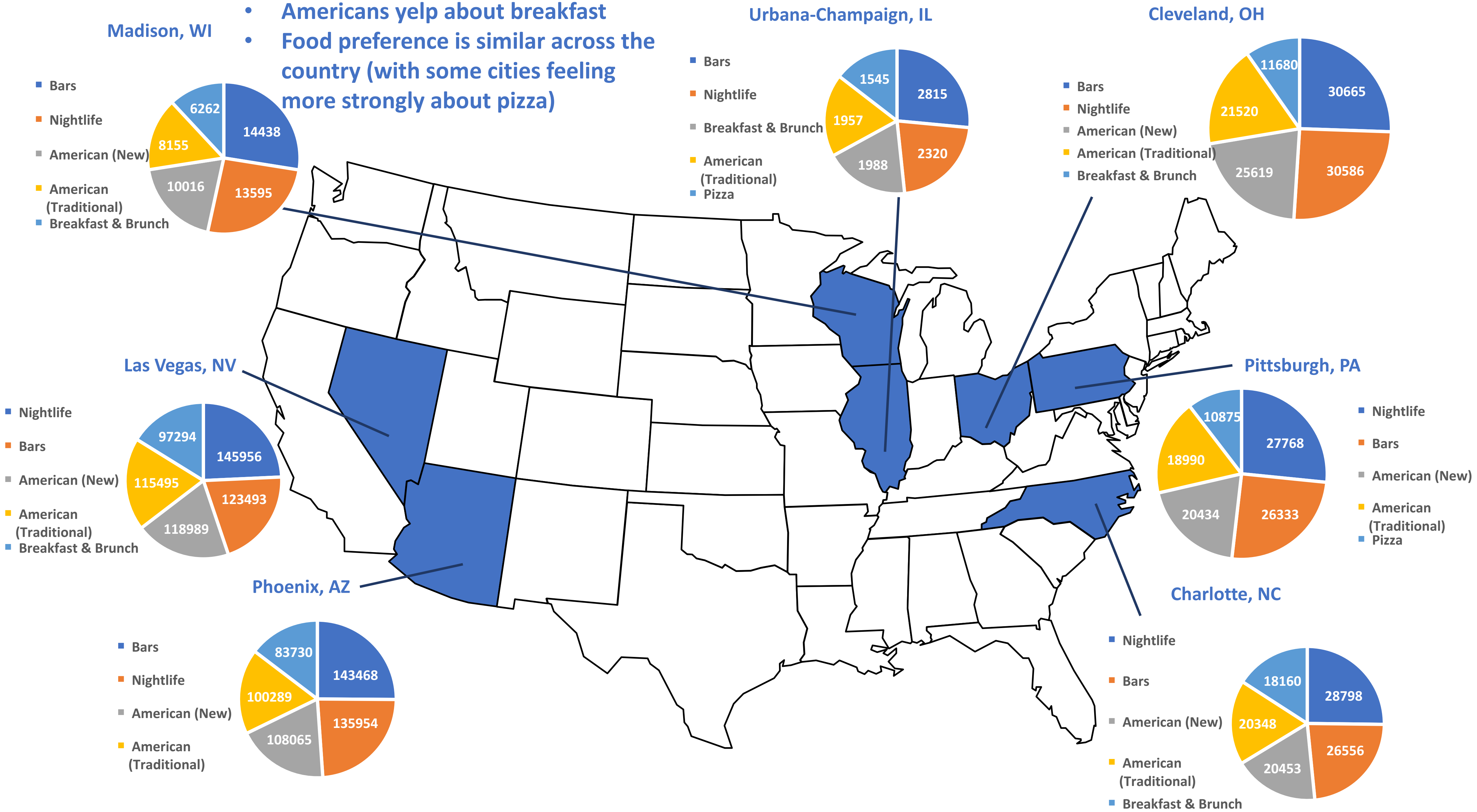
# More Insights on food preference

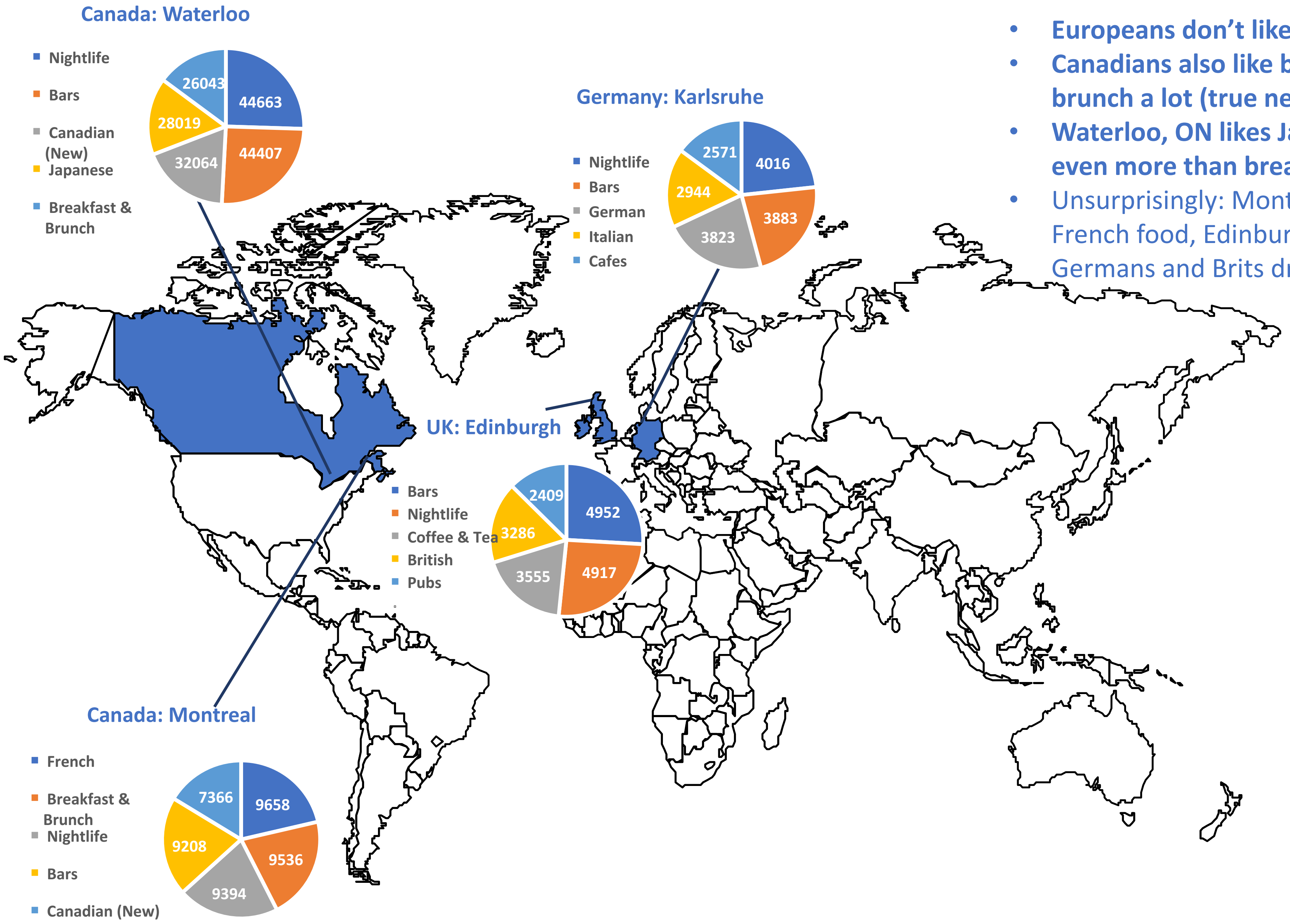
- We investigated the most popular categories of each state (international cities included)
- We defined most popular as having the most reviews and also the highest average stars ratings
- Interesting findings listed on following slides

| | state | newc | total_reviews |
|---|---|---|---|
| 1 | AZ | Bars | 143468 |
| 2 | BW | Nightlife | 4016 |
| 3 | EDH | Bars | 4952 |
| 4 | ELN | Cafes | 40 |
| 5 | ELN | Coffee & Tea | 40 |
| 6 | ESX | Pakistani | 5 |
| 7 | ESX | Indian | 5 |
| 8 | FIF | Bars | 21 |
| 9 | HLD | British | 102 |
| 10 | IL | Bars | 2815 |
| 11 | KHL | Coffee & Tea | 7 |
| 12 | KHL | Sandwiches | 7 |
| 13 | KHL | Soup | 7 |
| 14 | MLN | Nightlife | 205 |
| 15 | NC | Nightlife | 28798 |
| 16 | NI | German | 24 |
| 17 | NV | Nightlife | 145956 |
| 18 | NY | Pizza | 21 |
| 19 | OH | Bars | 30665 |
| 20 | ON | Nightlife | 44663 |
| 21 | PA | Nightlife | 27768 |
| 22 | PKN | Italian | 24 |
| 23 | QC | French | 9658 |
| 24 | SC | Nightlife | 813 |
| 25 | WI | Bars | 14438 |
| 26 | WLN | Fast Food | 15 |

Top one food categories in each state based on number of reviews

- Americans yelp about breakfast
- Food preference is similar across the country (with some cities feeling more strongly about pizza)






- **Europeans don't like breakfast at all**
- **Canadians also like breakfast and brunch a lot (true neighbor!)**
- **Waterloo, ON likes Japanese food even more than breakfast**
- Unsurprisingly: Montreal favors French food, Edinburgh likes tea, Germans and Brits drink a lot…

# Conclusion

- Linear regression with regularization performs the pest in terms of Root Mean Square Error (RMSE)
- Top positive words: Amazing, Delicious, Excellent, Best,
- Top negative words: Displeasure, Worst, Lure, Horrible,
- Analysis can be improved by including more reviews--we were only able to include 40K reviews due to time constraint
-
-