1. Research question (2 points) – provide the following:
   * at least one concrete research question or objective
   * a concrete hypothesis
   * a statement of interest and impact: why is this project interesting and useful, and to whom? Who will be impacted by the results or findings of this work?
2. Data (2 points) – provide the following:
   * at least one appropriate data source.
   * a concrete plan for how to obtain and preprocess this data. If there are access/API restrictions, please make a note of this and plan accordingly.
   * an example of a data point from your data: what attributes are provided in the dataset and how are these attributes represented?
   * a brief discussion of why this data source is appropriate
3. Methods (2 points) – address the following points:
   * list the aspects of the setting you will characterize/measure, and propose methods for measuring these aspects.
   * how will you analyze/measure the relationship between these aspects?
   * What will your measurements of these aspects and the relationships between them tell us about the research question?
4. Related work (2 points)
   * at least four references. These can be academic papers, news articles, blog posts … but they should give a sense of what existing work has been done in this space that you will borrow from and build on.
   * For each reference, briefly (1-2 sentences is ok) explain how the work relates to your project, and how your project differs from/builds on/extends the work. Does it present a method that is useful? Does it provide existing insights on a domain? Why aren’t you simply redoing what’s already been done in this work?
5. Feasibility test (2 points)
   * What is the first experiment/set of experiments you will do to confirm that your project is doable?
6. The names and IDs of each team member (these should already be there, but be sure to update as needed).

Your post should incorporate any feedback that was given in class and on Canvas.

Note that at this stage, your choice of research question and dataset should have somewhat converged, contingent on the results of your feasibility test (see next assignment). You should have a good-enough sense of what methods to use such that you can address this point in writing. However, you should expect that as you explore the data and try things that may or may not work, your thoughts on methodology will evolve. Finally, as you work on the project, you should continue to accumulate related work–this will aid in your efforts to complete the project, and you will be required to have a more extensive collection of related work for your final report.

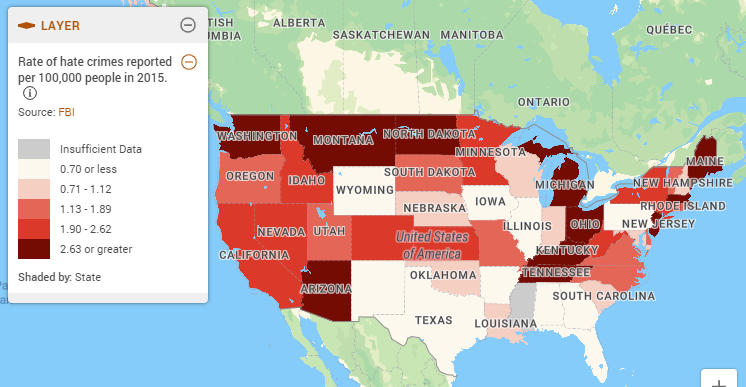
[Proposal] Yelp restaurant reviews analysis

## Research questions

* What makes a restaurant successful? (e.g., service, environment, or price)
* Is there any regional bias / ethnicity bias when users post reviews?

### Hypothesis

* As a successful restaurant, there must be some key factors that make it popular. A reasonable assumption is that we can find out those factors from the positive reviews of this restaurant. For example, most of the five stars reviews of “Evergreen Restaurant” are about “Chinese”, “environment”, and “service”. We can infer that these are why people in Ann Arbor love it.
* Based on the previous research and statistics on hate crimes and racial discrimination, we assume that restaurants under minority groups would be more likely to have biased reviews in certain regions.



### Statement of interest and impact

Understanding restaurant success factors can give restaurant owners clear goals to achieve. Moreover, the analysis result can be used to evaluate whether a specific type of restaurant is easy/hard to succeed in a specific area.

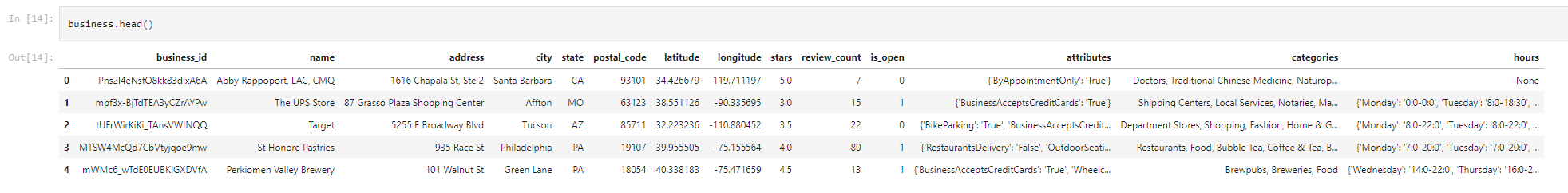
(source: https://www.policymap.com/blog/mapping-hate-crime)

## Data

The objective of the analysis is to provide insights into the success of the restaurants based on the total number of reviews and ratings. The ideal dataset should include information about restaurants, users, and reviews/comments. We plan to use [Yelp dataset](https://www.yelp.com/dataset) for this problem. This dataset has the following json format files that we need for our text analysis.

(1) business.json

It contains data about the business, including location data, attributes, and categories. Since we’re focusing on the restaurant reviews, we filter the dataset based on the categories and all of the businesses include “Restaurant” as one of their categories. We have 52k businesses after filtering.



To study further the regional bias, we plan to categorize the locations into several groups (West, Mid-West, South, East, etc.). And we will also categorize the restaurants based on the categories that represent nationalities. (e.g. Asian fusion, Japanese, American (Traditional), Italian, etc.)

The “attributes” column contains more information about the business. But not all restaurants have all the attributes. We currently plan to include “RestaurantsPriceRange2” as an indicator of the price. And we’ll check the completeness of each attribute to decide which to use.

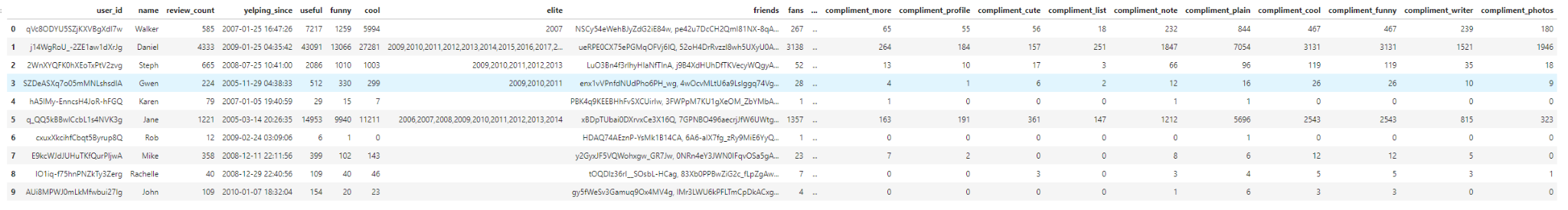
(2) review.json

It contains full review text data, including the user\_id that wrote the review and the business\_id the review is written for. After filtering based on the business\_id, we have 4.7M reviews.



(3) user.json

It contains the user's friend mapping, and history reviews stats. There are 2M users in the dataset.



## Methods

### Sentiment Analysis

There are several out-of-box Python libraries available. We’ll use NTLK for text preprocessing and sentiment analysis. Specifically, we will attempt to use this two analyzers:

SentiWordNet: enhanced lexical resource for sentiment analysis and opinion mining (https://www.nltk.org/howto/sentiwordnet.html)

SentimentIntensityAnalyzer: This method has a predefined list of words with sentiment scores and it matches words from the lexicon with words from the text. The VADER analyzer in the NLTK package provides four scores: negative, neutral, positive and compound for each piece of text.

### Regional Metrics

In light of [1], we will include the following metrics in our analysis:

Density: calculated as the number of shops in the restaurant’s neighborhood

Neighbor Entropy: measures the diversity in terms of facilities in the surrounding area. Lower entropy indicates that the area has more similar businesses.

Completeness: measures the proportion of neighboring restaurants of the same category.

EDA

We can visualize the data with respect to time to keep the reviews in a certain time frame and remove the potential noise. For data cleaning purposes we will filter out all of the stop words and unreadable texts, and use detect\_langs to select reviews written in English only.

Text Mining

We would like to use Bag-of-Words based on the tokenized stemmed and lemmatized words to initially calculate the word frequency and use n-grams to calculate the frequency of the pairs or triples of words that appear next to each other to get the context.

## Related Work

[1] Lian, Jianxun, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. "Restaurant survival analysis with heterogeneous information." In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 993-1002. 2017.

This paper focuses on building a predictive model to find out whether a restaurant is still open 4 years later. Though the topic is different, the paper used a similar dataset from Dianping (a Chinese app similar to Yelp) and created regional metrics for their analysis. Those metrics take into account the similar surrounding restaurants and also the density of restaurants nearby. The Yelp dataset also includes the longitude and latitude information. These metrics can also be used as regional factors in our analysis.

[2] Narrative framing of consumer sentiment in online restaurant reviews by Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith. *First Monday*, Volume 19, Number 4 - 7 April 2014 <https://firstmonday.org/ojs/index.php/fm/article/download/4944/3863> doi: <http://dx.doi.org/10.5210/fm.v19i4.4944>.

This paper focuses on the narratives that consumers use to frame positive and negative sentiment online. In addition to analyzing the successfulness of the restaurants by the number of reviews and the ratings, we can also take word and review lengths into consideration since the more pricey the restaurant is, the fewer reviews it might have, and the paper indicates that the more pricey the restaurant is, the longer the reviews it tend to receive. Thus, for pricey restaurants, we can calculate the word frequency to measure the successfulness instead of using the total number of reviews.

[3] Muchnik, Lev & Aral, Sinan & Taylor, Sean. (2013). Social Influence Bias: A Randomized Experiment. Science (New York, N.Y.). 341. 647-51. 10.1126/science.1240466.

This paper indicates that people tend to give ratings differently based on the treatments they received (eg. no treatment, up-treated by positive social influence, etc.). A similar pattern is found by Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science, 311*(5762), 854–856. [https://doi.org/10.1126/science.1121066](https://psycnet.apa.org/doi/10.1126/science.1121066). Although we may not be able to perform similar experiments on the users due to reality constraints, they do provide support for our hypothesis.

[4] [Text Mining and Sentiment Analysis for Yelp Reviews of A Burger Chain](https://towardsdatascience.com/text-mining-and-sentiment-analysis-for-yelp-reviews-of-a-burger-chain-6d3bcfcab17b)

In this article, the author attempted to figure out the success factors of a restaurant by analyzing the reviews of a burger chain. In our project, we also want to explore the success factors of restaurants. The difference is that we plan to analyze restaurants across several regions and if these factors vary by region.

[5] Sentiment Analysis of Yelp Reviews: A Comparison of Techniques and Models, Siqi Liu

University of Waterloo. arXiv:2004.13851

The author introduced the basic methods of sentiment analysis on the Yelp dataset and built machine learning models to further classify the reviews and test their performances. However, the main focus of the paper is about the restaurants in Toronto, and we would like to further expand to several regions in the US and have a larger dataset for big data analysis. In addition, since our focus is mainly on the analysis, we would like to skip the model building part.

## Feasibility Tests

Our project should be feasible since we use similar methods in the related works above. The difference is the dataset, and we incorporate regional analysis into it.

* We plan to first use a subset of the dataset containing a certain region of restaurants and a certain time frame of reviews to test the feasibility of our proposed methods.
* We’ll check the attribute completeness in the business.json file to decide which attributes we be used for our analysis
* We’ll check the categories in business.json and try to categorize them according to the nationalities. There is a risk that we can’t categorize them.
* We will further discuss any questions we may encounter during this process and seek new methods if needed.