Project 2 Proposal
Lina Sheremet, Ernesto Oropeza, Jaclyn Andrews
Repository Name: Project2_Sheremet_Oropeza_Andrews

The Data

The primary dataset we are analyzing was obtained from Yelp and contains data on 192,609 businesses with 6,685,900 reviews from customers spanning from 2005 to 2017. In order to keep our analyses more focused, we have decided to only analyze restaurants in Madison, Wisconsin. After this filtering, our dataset now contains information on 1024 businesses with 59,930 total reviews. The dataset is split into 4 groups of differing data: businesses, business attributes, check-ins, and reviews. Business data contains columns like the business id (which we will be using if we need to join), the location of the business (lat, lon), the average number of stars that business got, and some of its characteristics/categories. The business attribute data contains the business id, as well as many columns of attributes about that specific business, such as whether they accommodate vegetarians, have a TV, etc. The check-in data contains business id and how many check-ins (on Yelp) that business had on any given day and on any given hour. This could be useful for gauging the popularity of a restaurant. Lastly, review data is row-by-row reviews of each business on any given data by any given user. There is a rating, as well as a free text field, which we may try to conquer if we are feeling brave.

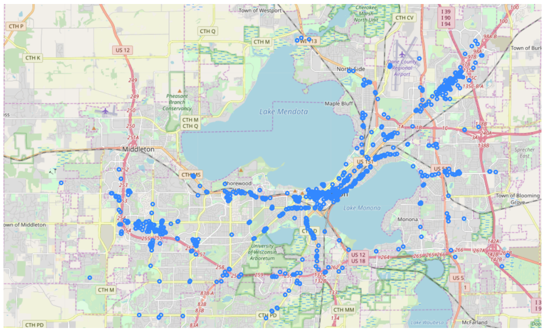


Figure 1: A map of all the businesses in Madison, Wisconsin in our dataset

In order to compliment our analyses of the businesses, we have decided to supplement with data on the weather conditions of Madison during the time of these reviews. We obtained the data on weather conditions from the National Oceanic and Atmospheric Administration, and it has the weather conditions, as well as the minimum and maximum temperature for any given date. We are interested in whether certain weather conditions are correlated with what kind of food people eat and how they rate the restaurant, so we will join the weather data to our yelp data on date. This way, we will have the exact weather conditions on any given data and can use that as one of our predictors.

Digging In

The project objective is to analyze the differences of the perceptions of the clients of the restaurants with time, weather conditions and location. The main variable that is intended to be explored is the rate given by clients after visiting a restaurant. In the data set this rate is represented by the “STARS” where 5 is the highest score and 1 is the minimum. The assumption in this case is that the reviews are written the same date of the restaurant visit. The figure below shows the distribution of star restaurant ratings over the span of 13 years.

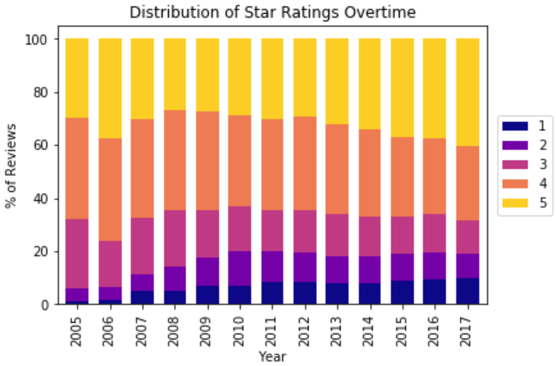


Figure 2: Distribution of restaurant ratings over time

This data set also has the reaction of users to each review by finding them useful, funny or cool. Also, other available information in this data set is the type of restaurant, the number of check-ins and geographical location.

Among the secondary available data, whether information in the area is available for the same time period of all reviews. These data were downloaded from the National Oceanic and Atmospheric Administration (NOAA, https://www.ncdc.noaa.gov/).

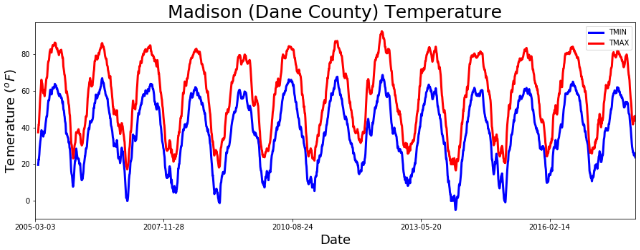
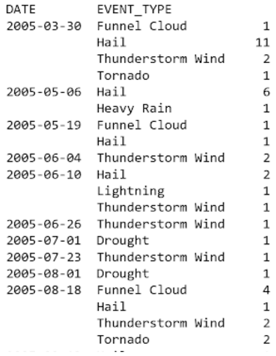


Figure 3: Temperature in Madison over time

The figure above shows the minimum and maximum temperature recorded in Madison for the period. Another information that will be analyzed is the severe weather information. A filtered version of these data is as follows:



```
DATE        EVENT_TYPE
2005-03-30  Funnel Cloud          1
            Hail                 11
            Thunderstorm Wind     2
            Tornado               1
2005-05-06  Hail                  6
            Heavy Rain            1
2005-05-19  Funnel Cloud          1
            Hail                  1
2005-06-04  Thunderstorm Wind     2
2005-06-10  Hail                  2
            Lightning             1
            Thunderstorm Wind     1
2005-06-26  Thunderstorm Wind     1
2005-07-01  Drought               1
2005-07-23  Thunderstorm Wind     1
2005-08-01  Drought               1
2005-08-18  Funnel Cloud          4
            Hail                  1
            Thunderstorm Wind     2
            Tornado               2
```

The complete list of weather events is: Hail, Tornado, Funnel Cloud, Thunderstorm Wind, Heavy Rain, Lightning, Drought, Dense Fog, Flash Flood, Winter Weather, Cold/Wind Chill, Strong Wind, Winter Storm, Blizzard, Heavy Snow, Extreme Cold/Wind Chill, Heat, High Wind, Dust Devil and Excessive Heat.

Our Questions & Structure

Our goal is to explore trends in the joined business review and weather data from the Madison, Wisconsin area over a 13-year span. Some questions we plan to explore are:

- Has the perception of restaurants or any subgroups of restaurants by location or cuisine in Madison changed over time?
- Are certain weather conditions correlated with the types of food people eat?
- Is there a significant difference in the distribution of ratings for restaurants during inclement weather?
- What businesses were visited more or less during inclement weather?

After our questions, our report will include a description of our source data with an exploratory analysis to learn more about the contents of our data sets. Next we will share our data cleaning and joining process as well as the assumptions we made about our data that lead to our analysis. For each question, we will describe our analysis and show visuals that convey our findings. And finally, we will share a summary of our largest takeaways and most interesting information we learned.