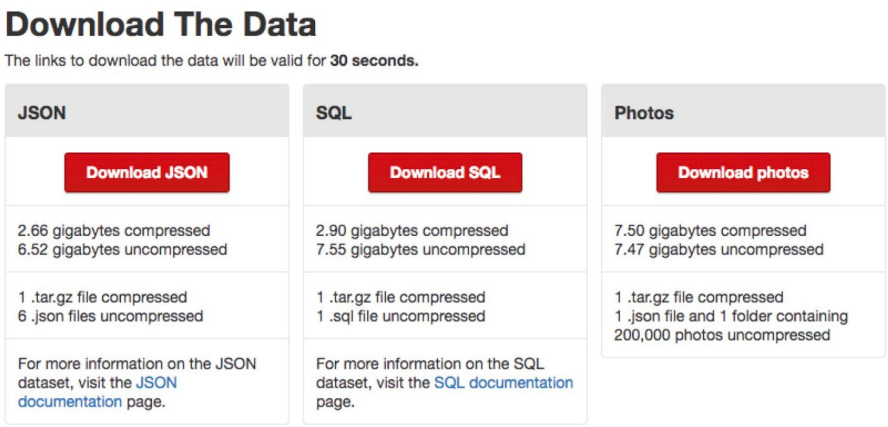
CMSC12300 Group Project Proposal

Group name: WACC
Group member: Shuting Chen, Ruxin Chen, Mengchen Shi, Lerong Wang
Dataset: Yelp dataset



Download The Data

The links to download the data will be valid for **30 seconds.**

JSON	SQL	Photos
Download JSON	Download SQL	Download photos
2.66 gigabytes compressed 6.52 gigabytes uncompressed	2.90 gigabytes compressed 7.55 gigabytes uncompressed	7.50 gigabytes compressed 7.47 gigabytes uncompressed
1 .tar.gz file compressed 6 .json files uncompressed	1 .tar.gz file compressed 1 .sql file uncompressed	1 .tar.gz file compressed 1 .json file and 1 folder containing 200,000 photos uncompressed
For more information on the JSON dataset, visit the JSON documentation page.	For more information on the SQL dataset, visit the SQL documentation page.	

Yelp dataset contains information for 174,000 businesses(Over 1.2 million business attributes like hours, parking, availability, and ambience), 5,200,000 reviews contributed from 1,300,000 users.

Hypothesis One: Economics of agglomeration is a theory says that similar businesses tend to locate nearby so that they can enjoy the economies of scale, share resources and create synergies. A real-world example of this theory is that, Mcdonald's and Kentucky Fried Chicken are often open near to each other. One of the questions our project wants to investigate is to explore the spatial distribution of different businesses. Yelp classifies local businesses into several categories: shopping, restaurants, home and local services, beauty and fitness, arts, entertainments and events, health, Auto, Nightlife and Travel & Hotel. Specifically, we want to explore the underlying patterns of the locations of these businesses and examine whether there is a demonstration of economics of agglomeration in each category. This investigation can be conducted at both national level as well as city level. For example, we are expected to find out Travel & Hotel businesses are more concentrated in popular travelling destinations, say California, than some remote, less scenery-attractive states; also, within city, it is often seen that Chinese restaurants open together and form as a little community to attract customers.

Hypothesis Two:
On Yelp, a user' reviews can be read by other users and can be read by other users. Other users can vote one review as "useful", "cool", or "funny", and can rate it from one star to five stars. My hypothesis is that the number of "useful" votes a reviews gain is related to several features: length, with or without a picture, the user's review counts, average stars and the

number of the user' fans. Longer reviews and reviews with pictures tend to contain more information so that are more likely to be useful. The more fans a user has, the higher her average star is, the more reviews she has written, indicating the user's rich experience so that her reviews are more likely win more "useful" votes.

Hypothesis Three:

Another question that we would like to explore is what is the relationship between trending businesses and successful businesses. By using the review dataset, we can determine trending businesses as those receiving most reviews in recent years (probably one or two years). However, they are not necessarily the most successful ones with most number of positive reviews (more than 4 stars). In this project, we can investigate this relationship in two dimensions. The first dimension focus on figuring out the most trending businesses and successful businesses across different categories. Yelp has data not only for restaurants, but also for nightlife, event planning and services, hotels, arts and entertainment, etc. Moreover, we intend to identify the most trending and successful businesses within each category. This provides the possibility for us to explore the heterogeneity of the relationship between trending and successful businesses from category to category.

Hypothesis Four:

Exploring successful business is important, but analyzing factors lead to failure is also important for future business and investors. We can also use yelp data to analyze factors that lead to restaurant closure from the perspectives of restaurant characteristics (rating, price, number of reviews) and spatial characteristics such as how many other restaurants are nearby and how does it compare to nearby restaurants based on price, ratings, etc. One possible hypothesis may be that some restaurant characteristics from yelp data such as ratings, locations are significant for explaining restaurant closure.