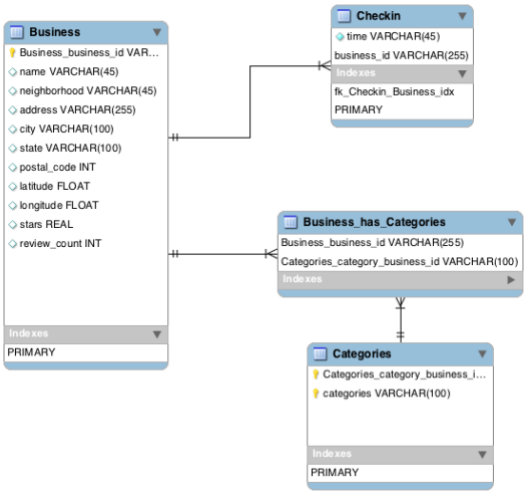
YELP

Jason Chang, Chen-Yuan Ho, Margaret Chung

Data: We used two databases provided in the yelp dataset challenge: yelp_academic_dataset_business.json + yelp_academic_dataset_checkin.json to retrieve company and check-in data, respectively.

To store it in MySQL, this is the ER Model we created:

We connected the data on businessID.



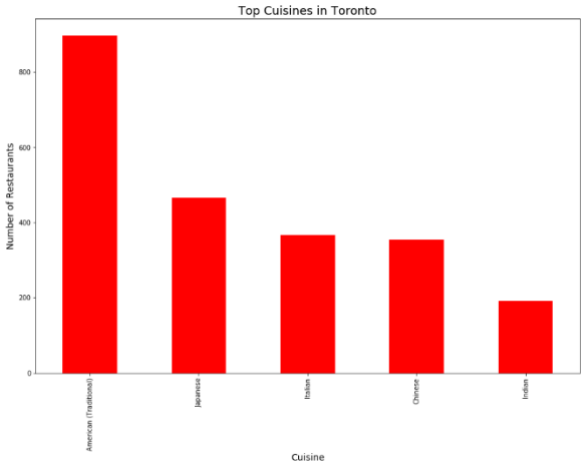
Problem: There were two main goals of the project:

1. Does the average rating of a cuisine change with the concentration of that cuisine within a neighborhood? For example, are Chinese restaurants outside of Chinatown rated higher than those inside Chinatown?
   - If you are traveling to a new city, you don't want to be fooled into eating a cuisine in a neighborhood, just because it's prominent! Our results are interesting.
   - If you are opening up a restaurant, you may want to factor in our research in your location scouting.
2. How do we define and find underrated restaurants?
   - Nobody enjoys lines. How do you still get good food, without the wait or reservations that most of the best restaurants have?

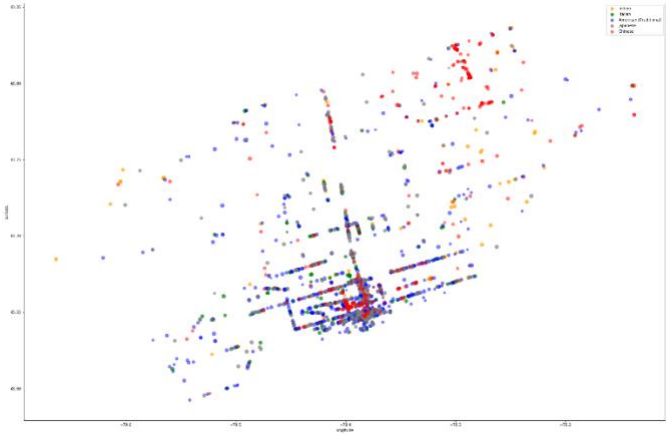
Approach:

1. Clean the data: A challenging component of the Yelp dataset was deciphering the "categories" list in the business dataset (ex. ["Restaurants", "Pizza", "Chicken Wings", "Italian"] would be components of one entry). We converted the json file into a CSV.
2. Next, we removed the "Nan" entries and all the non-restaurants – we ignored this data in our project. We also re-categorized all the restaurants into a cuisine type, using the yelp category filter list (found on this website: https://goo.gl/EL30xS). (Ex. "Sushi bars" → "Japanese")
3. After looking at the data, we decided to focus on Toronto because that city had the most restaurants. We formatted the CSV into Pandas.

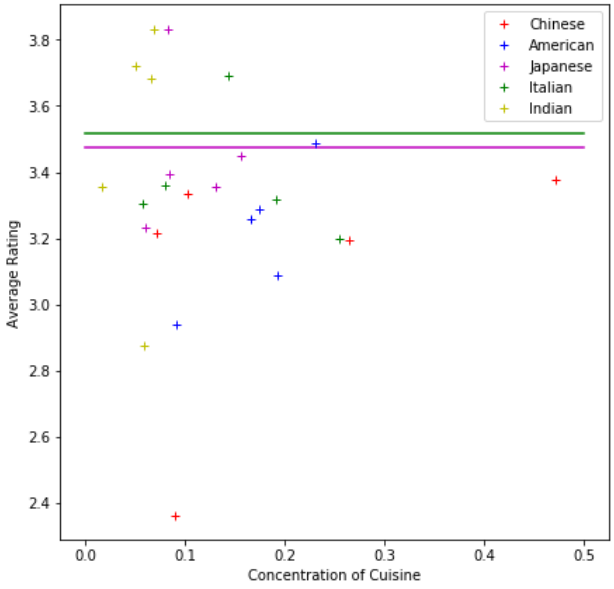
Part I: We counted the number of each cuisine and figured out the top 5, as well as each cuisine's top neighborhoods. (shown below)



We plotted these cuisines onto a scatterplot to visualize the concentrations.



Then we calculated the concentration of the cuisine in each of these neighborhoods, and the average rating of the cuisine in that neighborhood. We used matplotlib to plot these results:



It seems that the higher the concentration of a cuisine, the lower the average rating. Perhaps this is because there are comparison opportunities nearby.

Part II: To identify the "hidden gems", we first had to identify the parameters of an underrated, but good restaurant. We used the describe function to see the average rating, review count, and check-in frequency and determined the criteria from there.

	stars	review_count	Checkin_Frequency
count	5222.00	5222.00	5222.00
mean	3.419667	40.230372	65.867101
std	0.716932	65.978974	120.911681
min	1.000000	3.000000	1.000000
25%	3.000000	8.000000	8.000000
50%	3.500000	18.000000	25.000000
75%	4.000000	45.750000	70.000000
max	5.000000	1145.000000	1656.000000

Hidden Gem Criteria:
Stars = Min 4.5 **AND**

Reviews = 5-10

OR Check-ins = Max 8

To make this more user-friendly, we created a textbox in which users can type in a neighborhood to search for hidden gems!

Course Topics Applied: Pandas, Visualization, SQL, CSV, Json, Load/Store Files, Loop, If Statements, Functions

Assumptions: The first category listed in the categories string, outside of "Restaurants", was the most relevant to that institution.