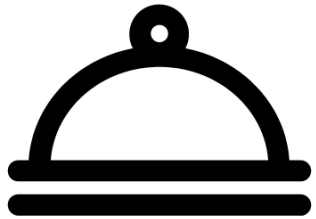
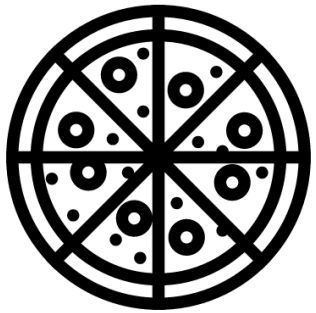
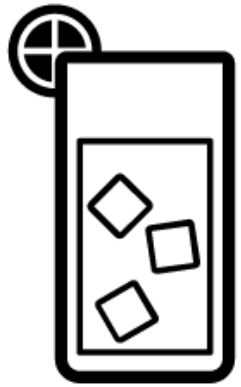






Group 7

JASON CHANG

CHEN-YUAN HO

MARGARET CHUNG

PROJECT GOALS

- Leverage yelp data set to understand the Toronto Food Scene
  - **Cuisines by Neighborhood**
    - What are the top cuisines, and what neighborhoods are they found?
    - Is there correlation between the concentration of these cuisines & the average rating?
  - **Where are the *"hidden gems"*?**
    - How to we create a tool to identify great restaurants that are not popular/well-known?

Note: This can be applied to any city, but we focused on Toronto for our project

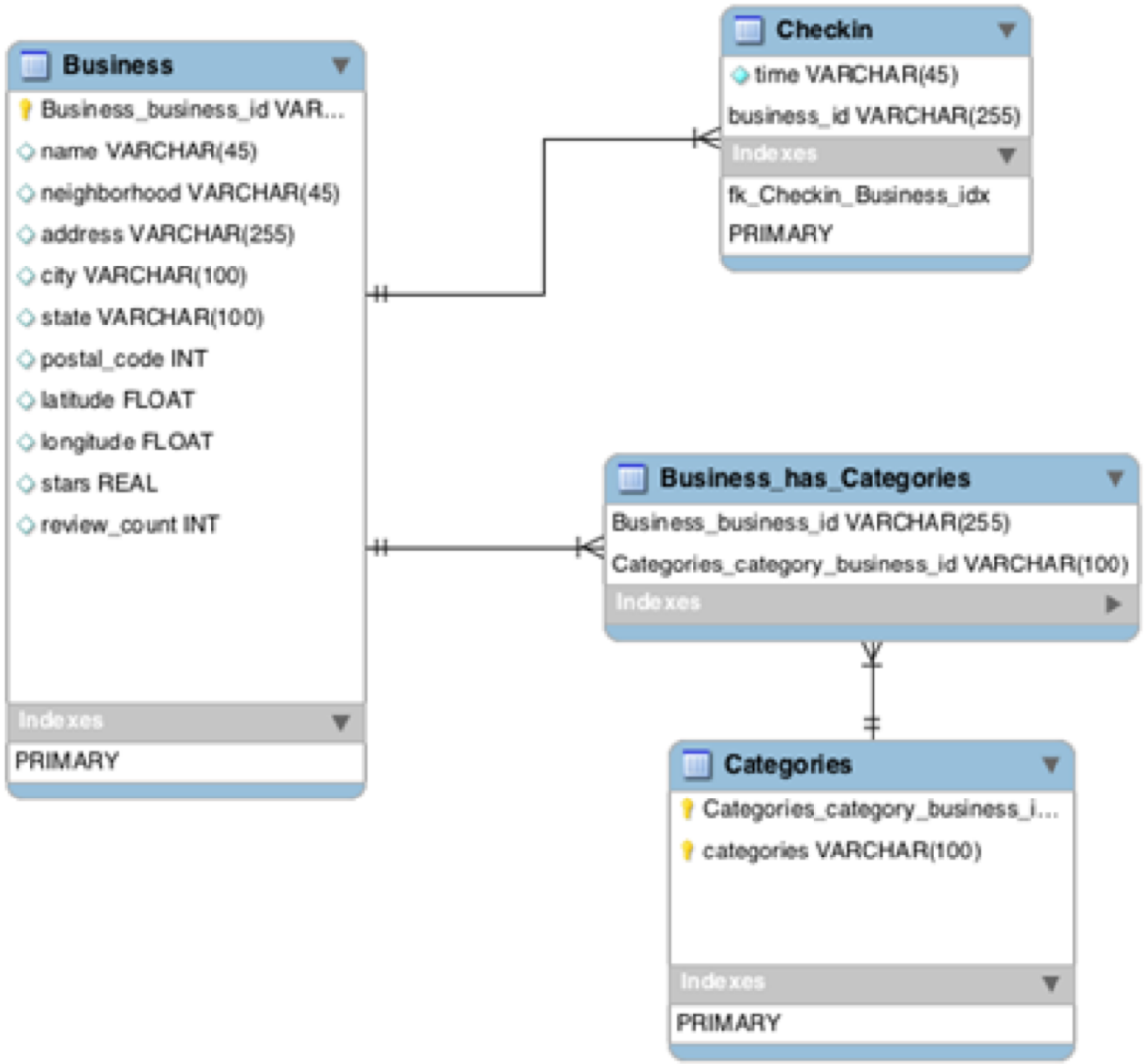
DATA SETS

Table	Contents
yelp_academic_dataset_business.json	Business/Company information • Address (City, State, Postal Code, Longitude/Latitude) • Reviews (# of Reviews, Star Rating) • Attributes • Categories • Business Hours
yelp_academic_dataset_checkin.json	• Check-ins with date and time

SCOPE:

- Location: Toronto, Canada
- Restaurants Only
- Top 5 Cuisines

SQL SCHEMA


Business
Business_business_id VAR...
name VARCHAR(45)
neighborhood VARCHAR(45)
address VARCHAR(255)
city VARCHAR(100)
state VARCHAR(100)
postal_code INT
latitude FLOAT
longitude FLOAT
stars REAL
review_count INT
Indexes
PRIMARY
Checkin
time VARCHAR(45)
business_id VARCHAR(255)
Indexes
fk_Checkin_Business_idx
PRIMARY
Business_has_Categories
Business_business_id VARCHAR(255)
Categories_category_business_id VARCHAR(100)
Indexes
Categories
Categories_category_business_i...
categories VARCHAR(100)
Indexes
PRIMARY

ORGANIZING THE DATA

1. Cleaning up the data

business_id	name	neighborhood	city	latitude	longitude	stars	review_count	attributes	categories
0DI8Dt2PJp07XkVvIElIcQ	Innovative Vapors		Tempe	33.37821	-111.936	4.5	17	['BikeParking: True	['Tobacco Shops', 'Nightlife', 'Vape Shops', 'Shopping']
LTlCaCGZE14GuaUXUGbamg	Cut and Taste		Las Vegas	36.19228	-115.159	5	9	['BusinessAcceptsB	['Caterers', 'Grocery', 'Food', 'Event Planning & Services', 'Party & Event Planning', 'Specialty Food']
EDqCEAGXVGCH4FJXgqtjqg	Pizza Pizza	Dufferin Grove	Toronto	43.66105	-79.4291	2.5	7	['Alcohol: none', "A	['Restaurants', 'Pizza', 'Chicken Wings', 'Italian']
cnGlivYRLxpF7tBVR_JwWA	Plush Salon and Spa		Oakdale	40.44454	-80.1745	4	4	['AcceptsInsurance	['Hair Removal', 'Beauty & Spas', 'Blow Dry/Out Services', 'Hair Stylists', 'Hair Extensions', 'Massage', 'Permane

2. Input json file to CSV
   - Everything in CSV is a string
   - Clean out “NAN” and separate off non-restaurant businesses
3. Leveraged yelp category filter list to select all the types of cuisines possible
   - Re-categorized the cuisines
4. Formatted CSV into Pandas

Assumptions:

- The first food category listed in the category description is the most relevant to that institution.
  - For example: “Restaurant, Pizza, Chicken Wings” – we classified this as a Pizza Restaurant

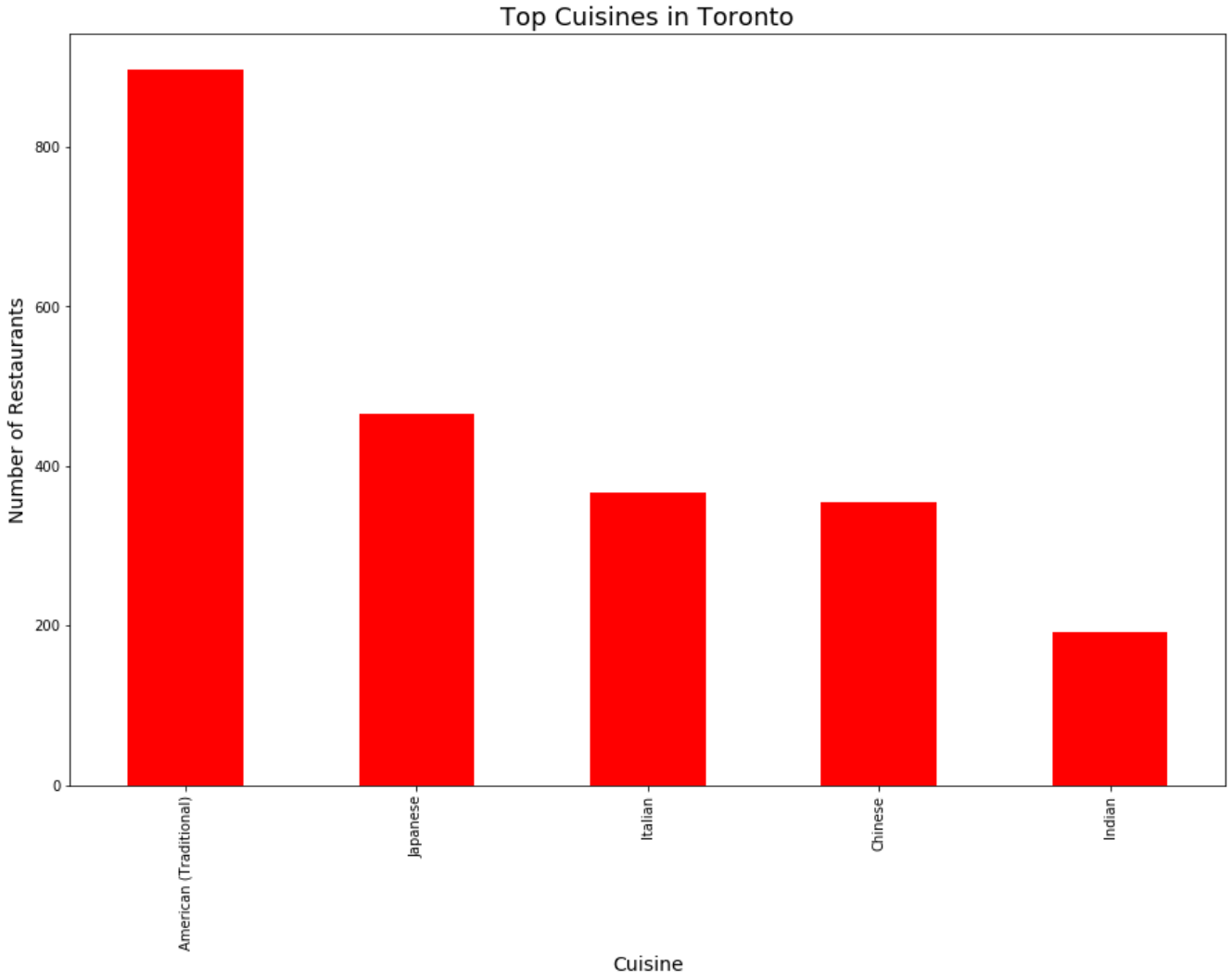
CUISINES BY NEIGHBORHOOD

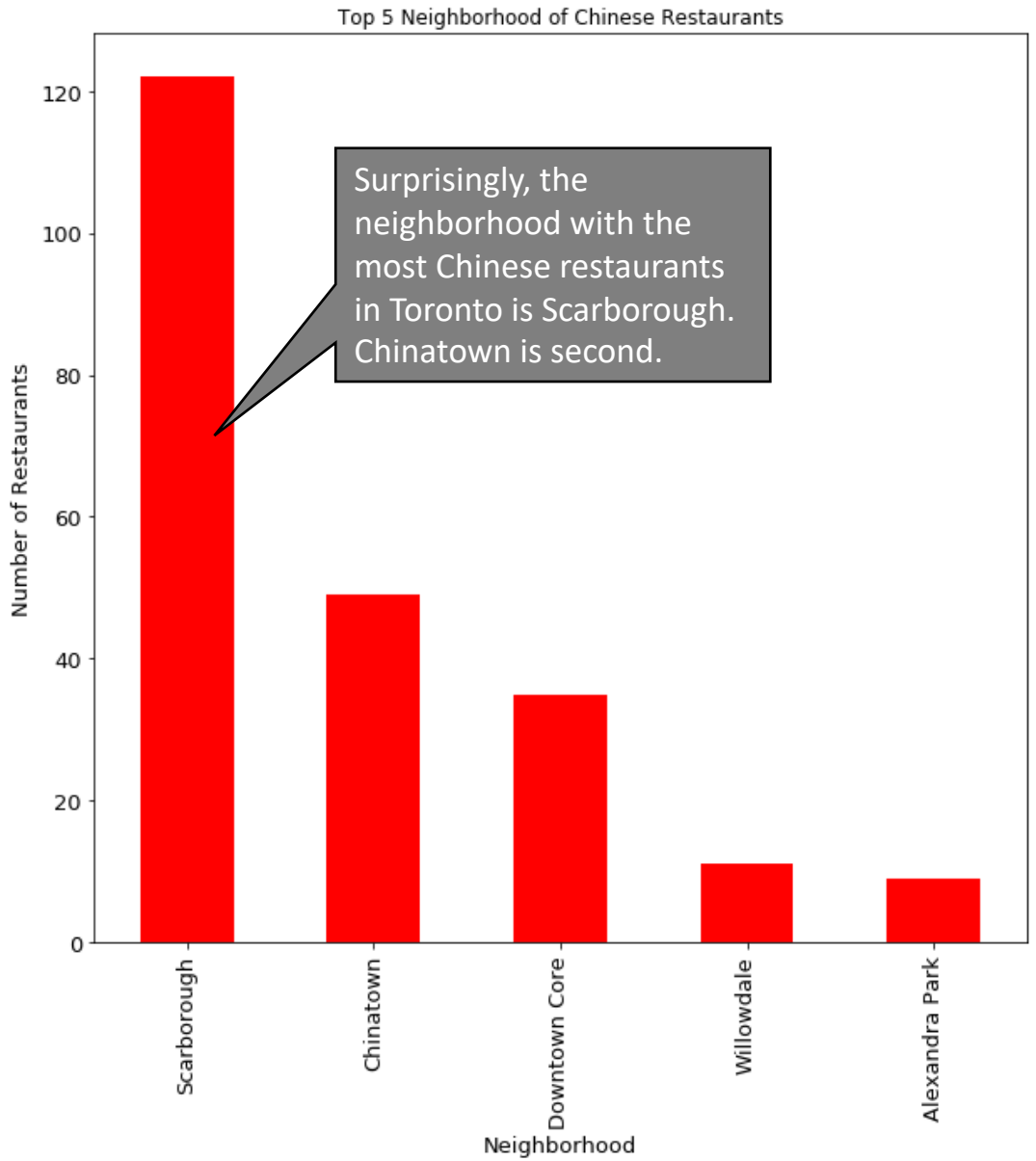
GOALS:

- Find the top 5 cuisines offered and neighborhoods they are most prominent
- Is there a correlation between the concentration of the cuisine and the average rating in that neighborhood?

  Note: Only plotted the top 5 neighborhoods of each cuisine

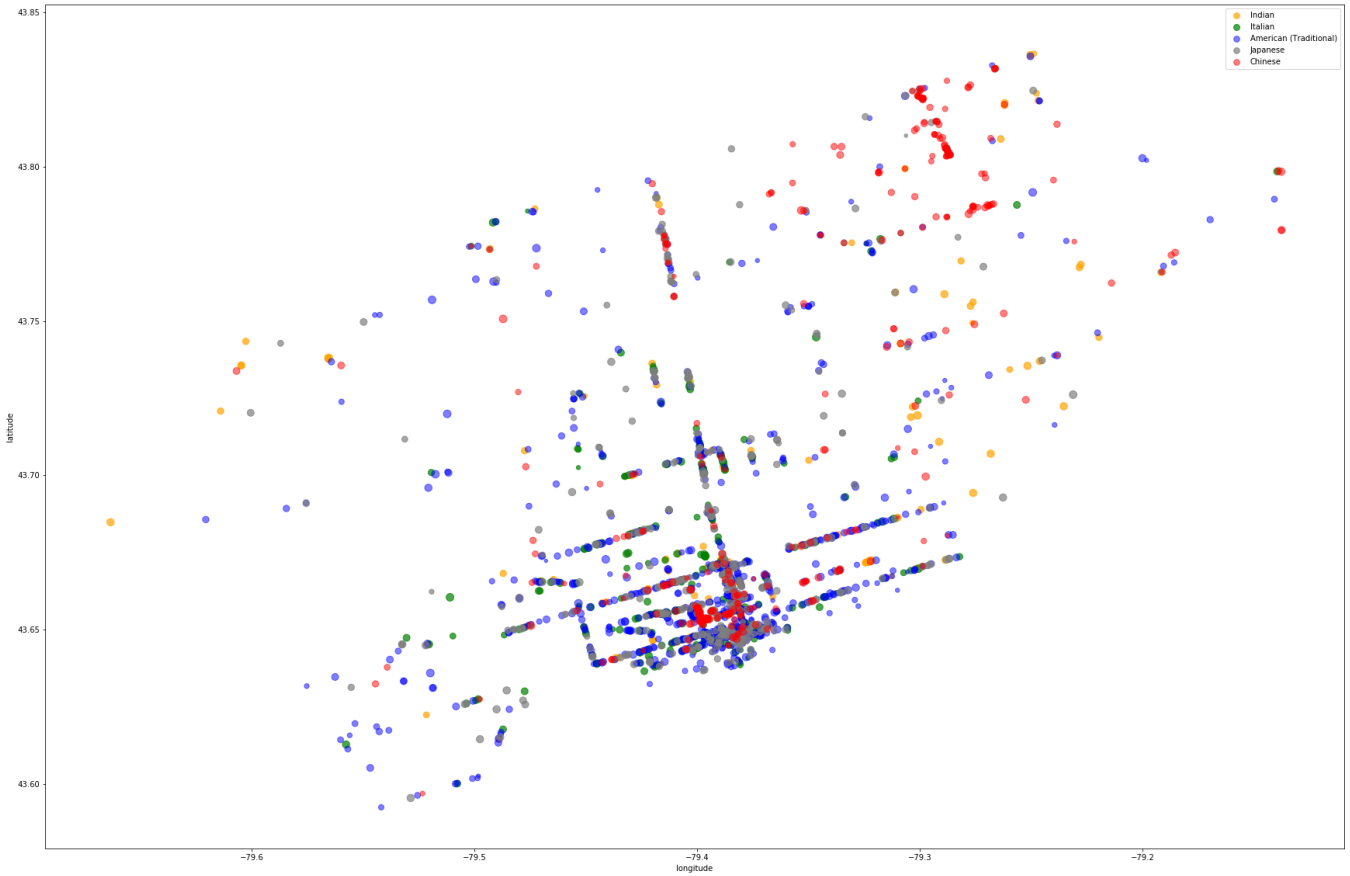
TORONTO FOOD SCENE

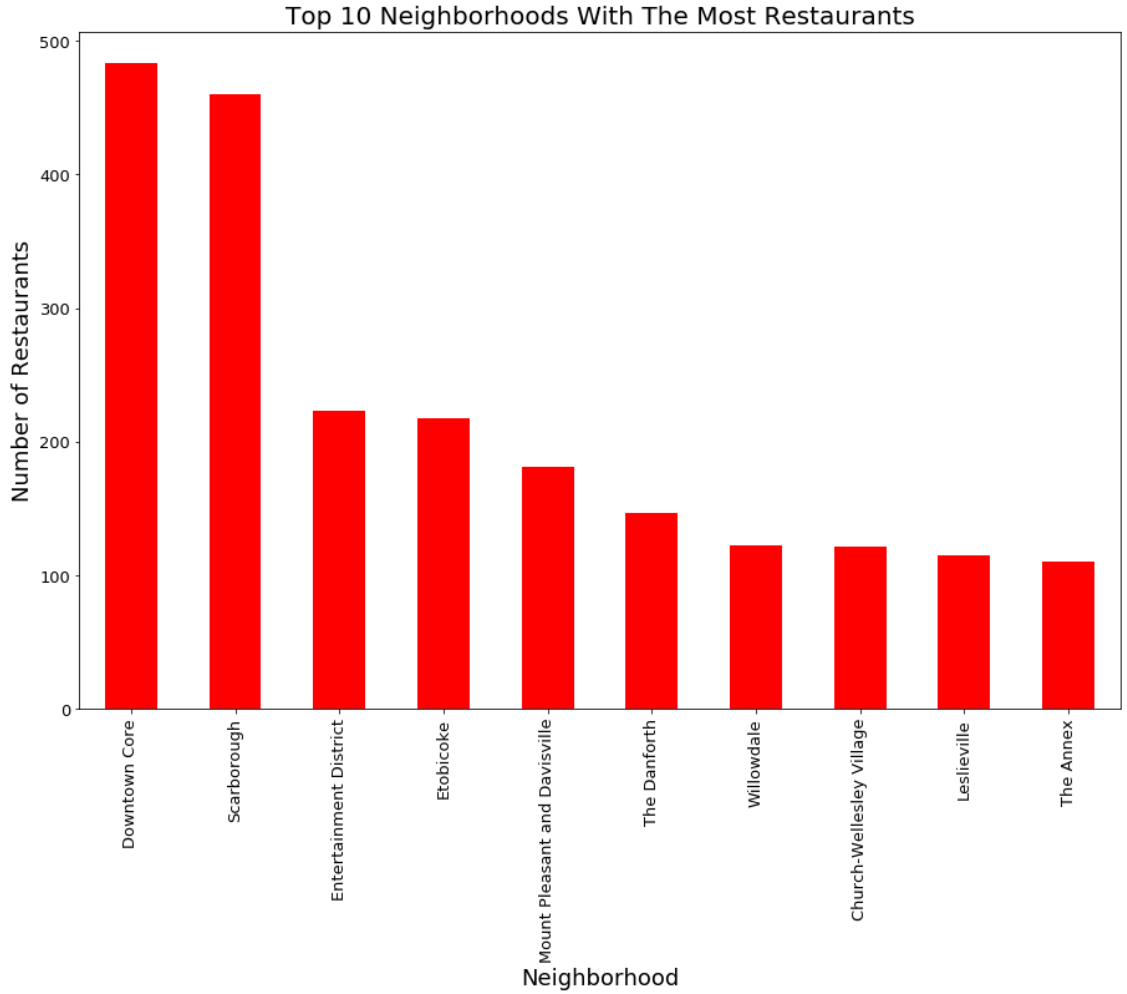




TORONTO FOOD SCENE

Locations Of Top 5 Cuisines





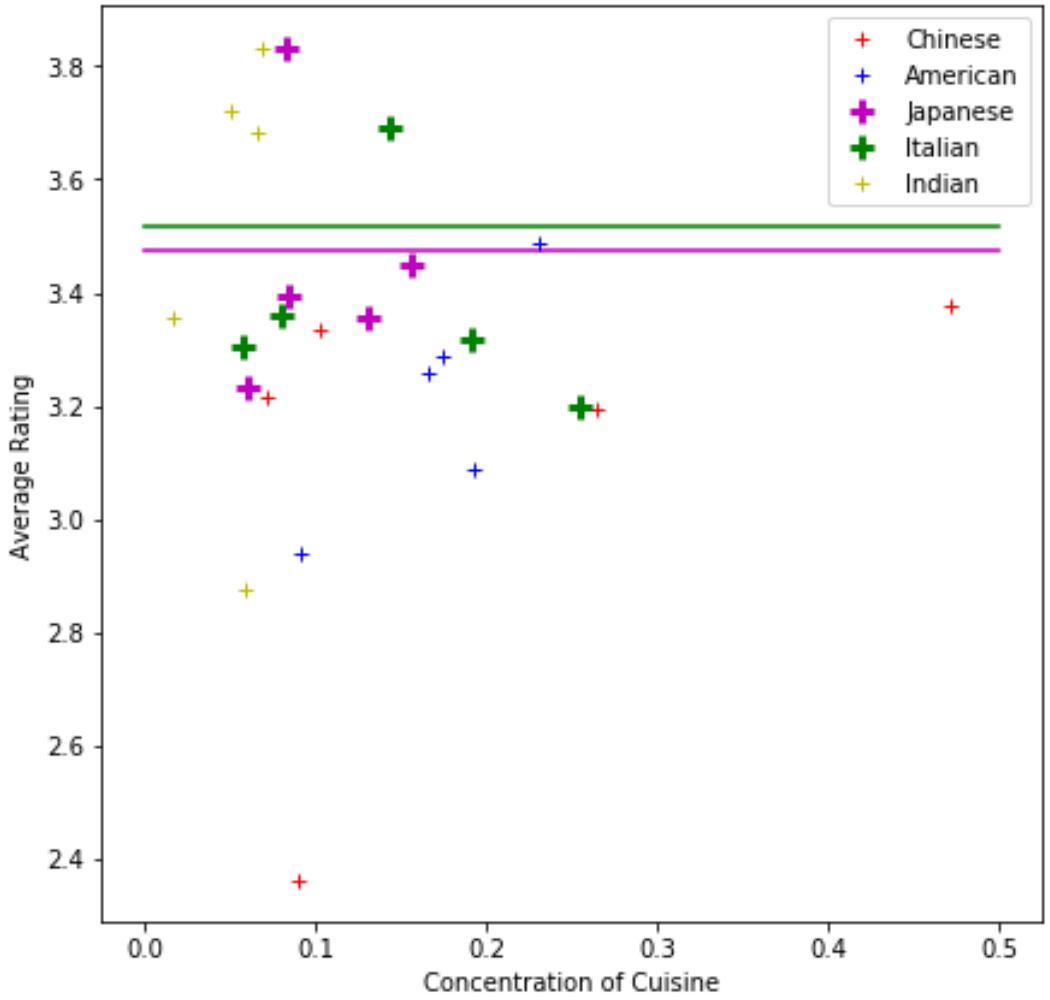
CORRELATION TEST:

CONCENTRATION OF CUISINE VS. AVERAGE RATING

- Concentration of cuisine does not mean better reviews.
- Lower concentrations of a cuisine could be favorable for a restaurant's review!

Our Theory: Lack of comparison in cuisine to nearby restaurants may be helping a restaurant's review (relative standpoint)

$$\text{CONCENTRATION} = \frac{\text{\# OF SPECIFIC CUISINE IN NEIGHBORHOOD}}{\text{TOTAL RESTAURANTS IN THAT NEIGHBORHOOD}}$$

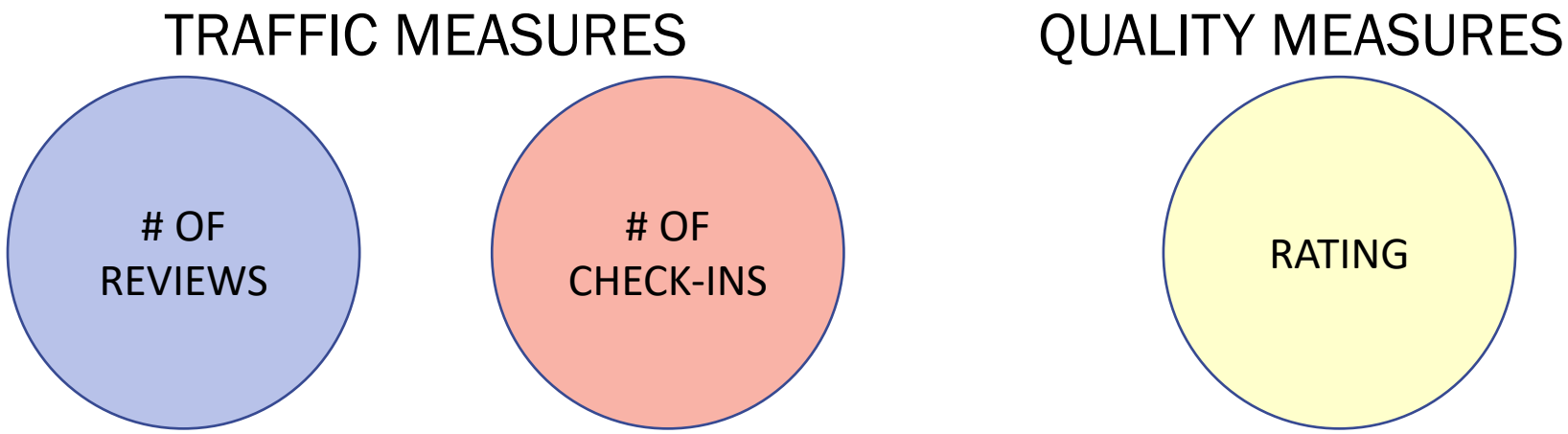


Note: Only top 5 neighborhoods for each cuisine are shown

HIDDEN GEMS

GOAL: To have a great dining experience without the wait

1. Identify the rule sets to classify a hidden gem



2. Apply rule sets to our data to create the list of restaurants
3. How do we use this in practice?

SETTING THE CRITERIA

	latitude	longitude	stars	review_count	is_open	Checkin_Frequency
count	5222.00	5222.00	5222.00	5222.00	5222.00	5222.00
mean	43.681415	-79.391006	3.419667	40.230372	0.732861	65.867101
std	0.047025	0.059862	0.716932	65.978974	0.442508	120.911681
min	43.592484	-79.663413	1.000000	3.000000	0.000000	1.000000
25%	43.651033	-79.416209	3.000000	8.000000	0.000000	8.000000
50%	43.663863	-79.393125	3.500000	18.000000	1.000000	25.000000
75%	43.690747	-79.373408	4.000000	45.750000	1.000000	70.000000
max	43.876501	-79.137540	5.000000	1145.000000	1.000000	1656.000000

- Described the CHECK-INs dataset.
- Use this information to decide criteria.

We considered 4.5+ Stars to be the threshold

The lower 25% had 8 reviews. We used 5-10 range for consideration set

The lower 25% had 8 check-ins. We looked for locations with less than 8.

HIDDEN GEMS

- DEMO

USE CASES

- From restaurant-owner standpoint, this could be a helpful tool in location scouting – seeing who else is there and how that may affect their reviews.
- From a diner standpoint, they can use this tool when traveling to a new area if they want options for good restaurants without a long wait. They can also get credit for being “in-the-know”.

THANK YOU!