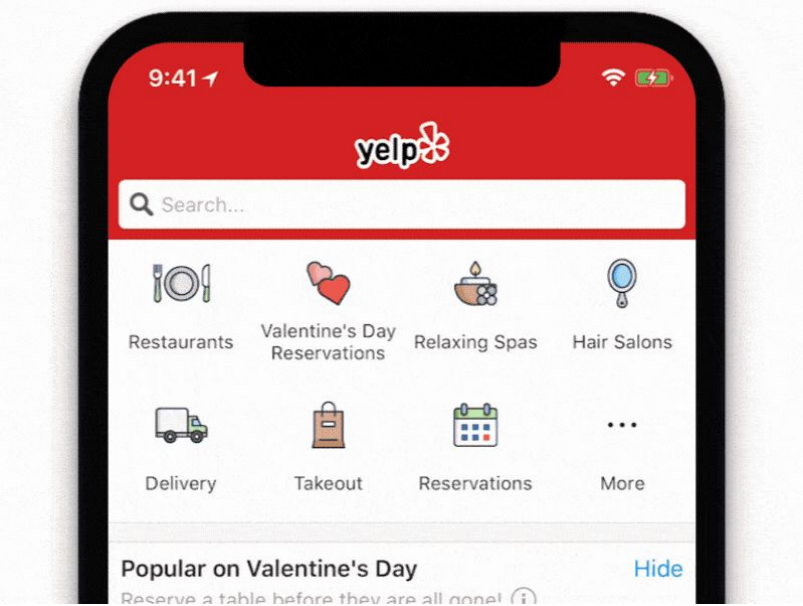
What's in your lunch?



Nicole Imbriaco, Sebastian Hooker, Ethan Kessinger, Ross Memole

Agenda

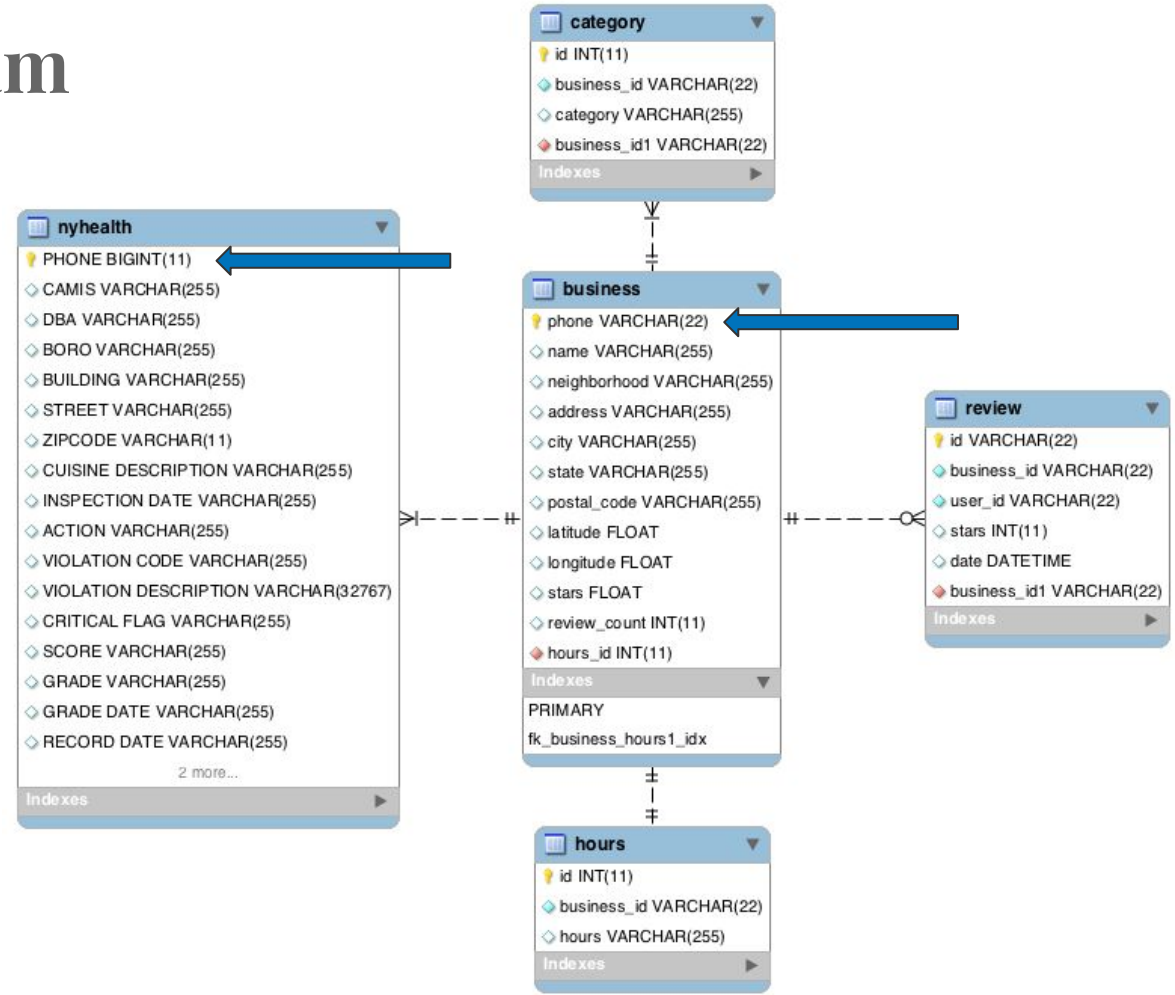
- Datasets
- Dataset Questions
- Findings
- Questions?

Datasets

Selecting a Dataset

1. Started with the **New York Health Department** and **Yelp** databases
2. Decided to focus only on **Manhattan**
3. Used distinct **Phone Numbers** to join databases
4. Deleted duplicates and pared down to **9,235 Restaurants**

ER Diagram


category
id INT(11)
business_id VARCHAR(22)
category VARCHAR(255)
business_id1 VARCHAR(22)
Indexes
nyhealth
PHONE BIGINT(11)
CAMIS VARCHAR(255)
DBA VARCHAR(255)
BORO VARCHAR(255)
BUILDING VARCHAR(255)
STREET VARCHAR(255)
ZIPCODE VARCHAR(11)
CUISINE DESCRIPTION VARCHAR(255)
INSPECTION DATE VARCHAR(255)
ACTION VARCHAR(255)
VIOLATION CODE VARCHAR(255)
VIOLATION DESCRIPTION VARCHAR(32767)
CRITICAL FLAG VARCHAR(255)
SCORE VARCHAR(255)
GRADE VARCHAR(255)
GRADE DATE VARCHAR(255)
RECORD DATE VARCHAR(255)
2 more...
Indexes
business
phone VARCHAR(22)
name VARCHAR(255)
neighborhood VARCHAR(255)
address VARCHAR(255)
city VARCHAR(255)
state VARCHAR(255)
postal_code VARCHAR(255)
latitude FLOAT
longitude FLOAT
stars FLOAT
review_count INT(11)
hours_id INT(11)
Indexes
PRIMARY
fk_business_hours1_idx
review
id VARCHAR(22)
business_id VARCHAR(22)
user_id VARCHAR(22)
stars INT(11)
date DATETIME
business_id1 VARCHAR(22)
Indexes
hours
id INT(11)
business_id VARCHAR(22)
hours VARCHAR(255)
Indexes

Downloading the Data

1. Queried the Yelp API by Phone Number and stored the responses as JSON files
   (Required approximately 9,500 API requests and 2 API keys to sustain request volume.)
2. Read JSON Files Using Pandas & Appended Each Entry into a DataFrame, then deleted the file

```
appended = pd.DataFrame(data['businesses']).append(appended, sort=False)
```

API Request:
URI:
https://api.yelp.com/v3/businesses/search/phone
Method: GET
Parameter: '+12128658777

API Response:
{'businesses': [{'alias': 'jimbos-hamburger-palace',
'categories': [{'alias': 'burgers', 'title': 'Burgers'}],

Sanitizing the Data

1. Cleaned DataFrame using Pandas by applying pd.series, dropping columns, etc.

```
coordinates_df = appended.coordinates.apply(pd.Series)appended.drop('image_url',axis='columns', inplace=True)
```

2. Downloaded the final DataFrame as a CSV

```
final_appended.to_csv('yelp.csv',index=False)
```

Importing the Data

1. Created an RDS instance at AWS to store all of our data in one database
2. Imported CSVs from NY Health Department and our final Yelp CSV using MySQL Workbench

Focusing only on the data we need

```
DELETE FROM NYHEALTH WHERE BORO NOT IN ('MANHATTAN');

DELETE FROM NYHEALTH WHERE PHONE NOT IN (SELECT Y.phone FROM YELP Y);
```

A Note About Code Reusability

- The data we analyzed was not live from the API and required sanitization, appending, and downloading
- As soon as we download the CSVs, the information is out of date

Dataset Questions

Dataset Questions

1. What is the average health score of restaurants in NYC area with a rating of 4.5 stars or higher?

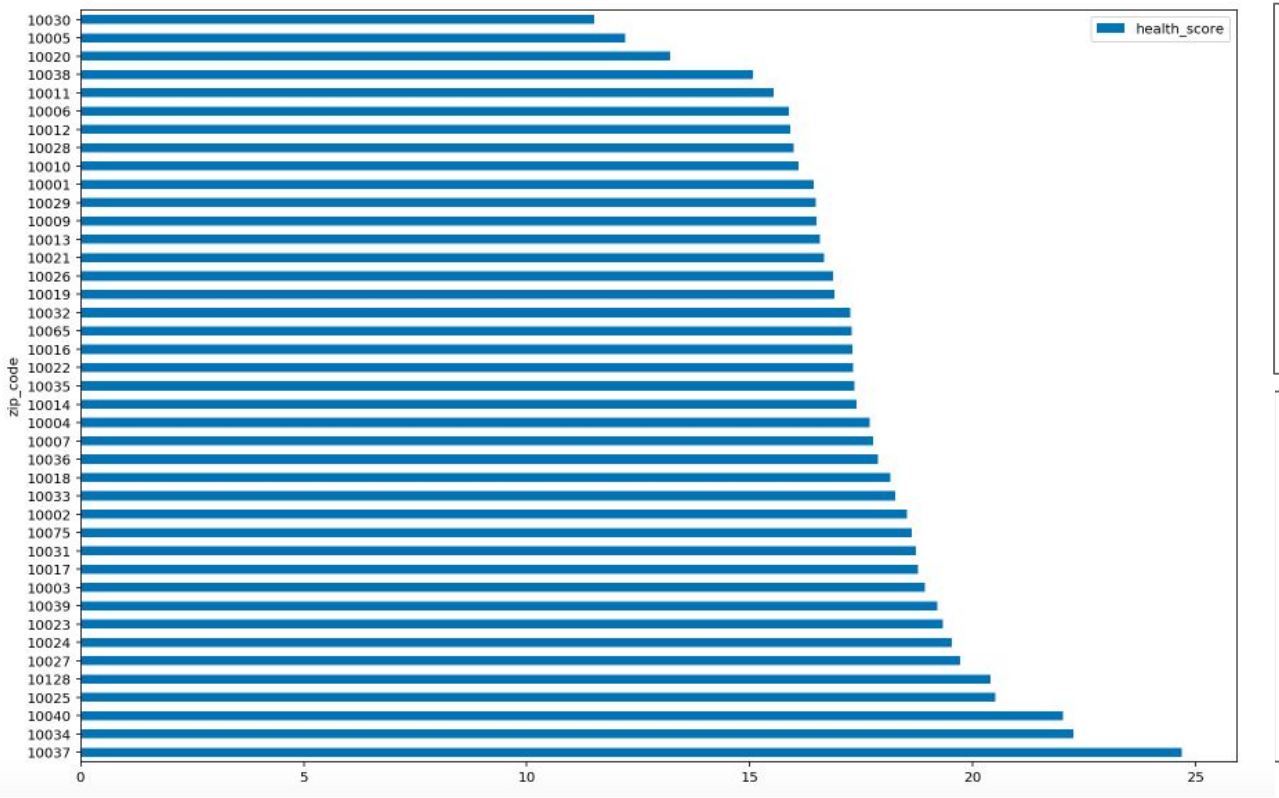
2. Is there a relationship between the number of reviews and number of inspections?

3. What are the top 5 violations amongst each price category? Does a higher $ range mean less rat violations?

4. Correlation matrix for highly rated yelp versus highly rated health score

Findings

4.5+ star rated restaurants average a B health grade



Findings:

1. Average = 17.62
2. Median= 17.35
3. Best zip code = 10030
4. Worst zip code = 10037

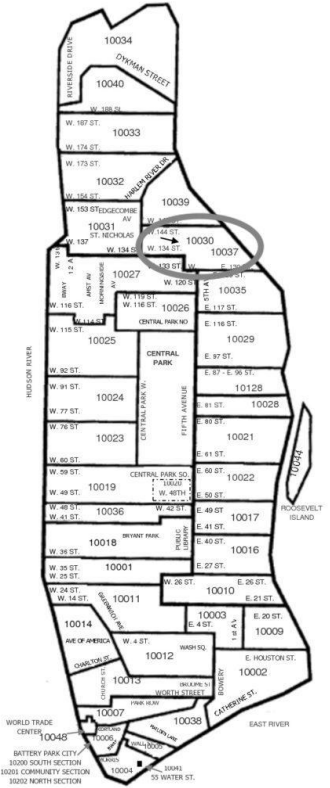
Scale:

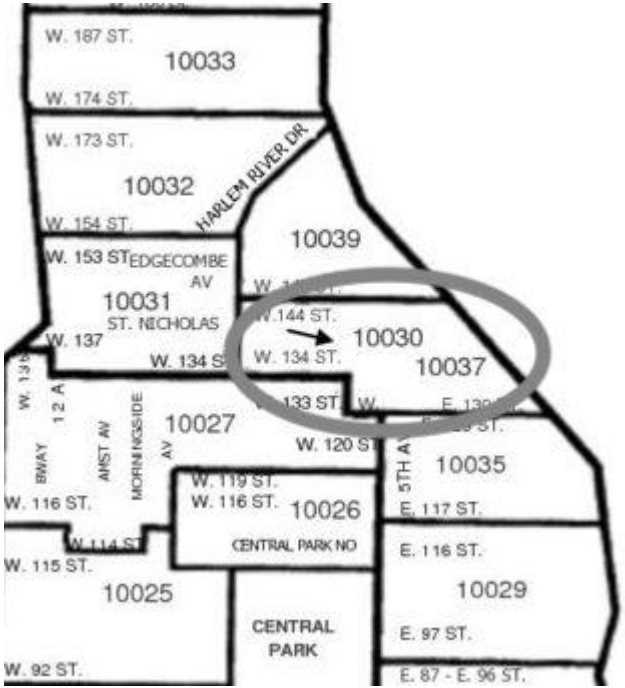
0 - 13 points = A

14 - 27 = B

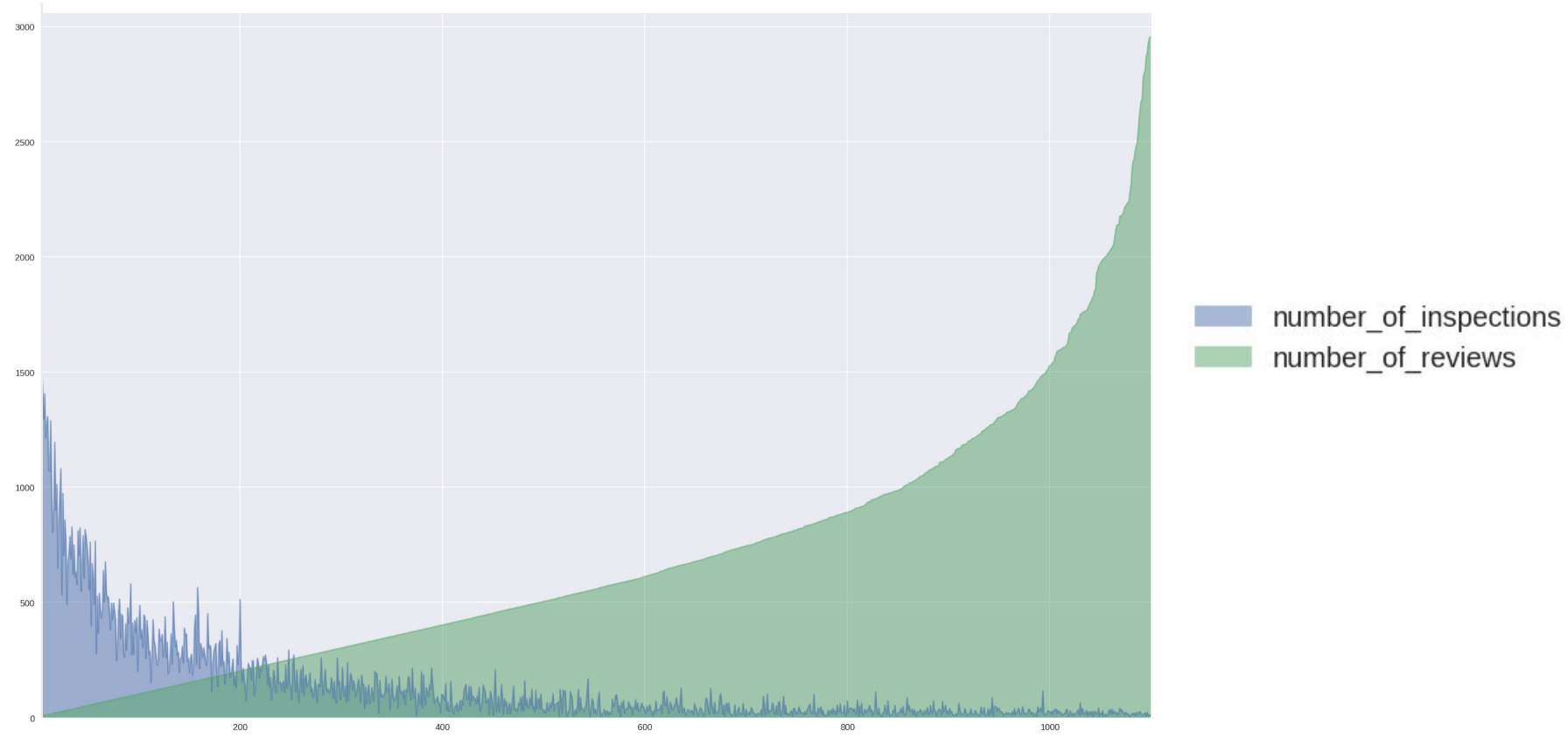
28+ = C

Best and worst scoring zip codes are geographically next to one another:





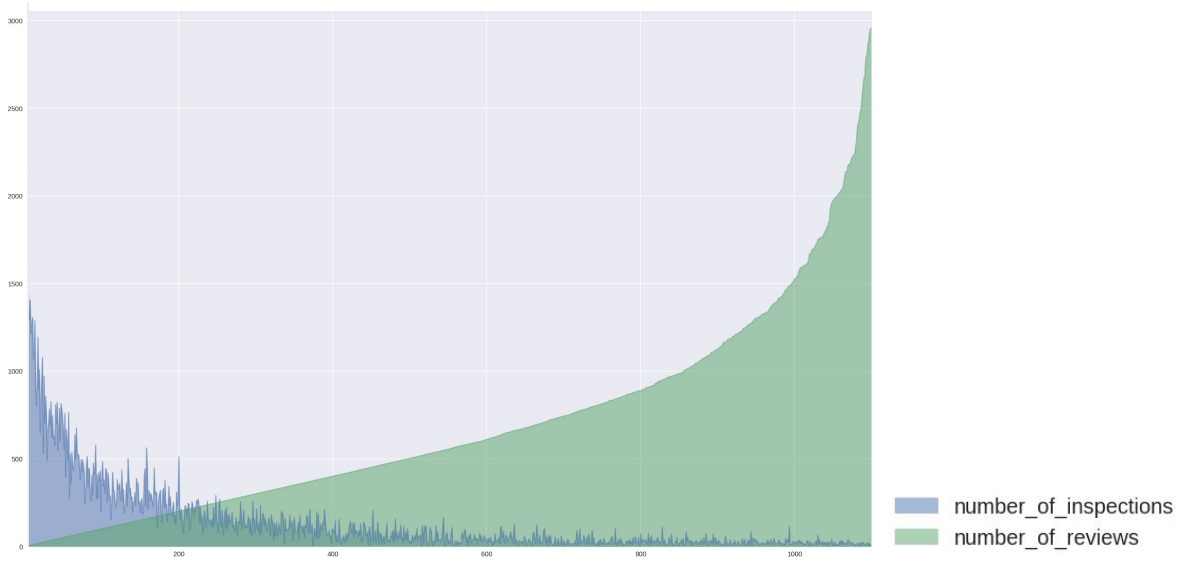
The Relationship Between Reviews & Inspections



There is an inverse relationship... why?

Our hypotheses:

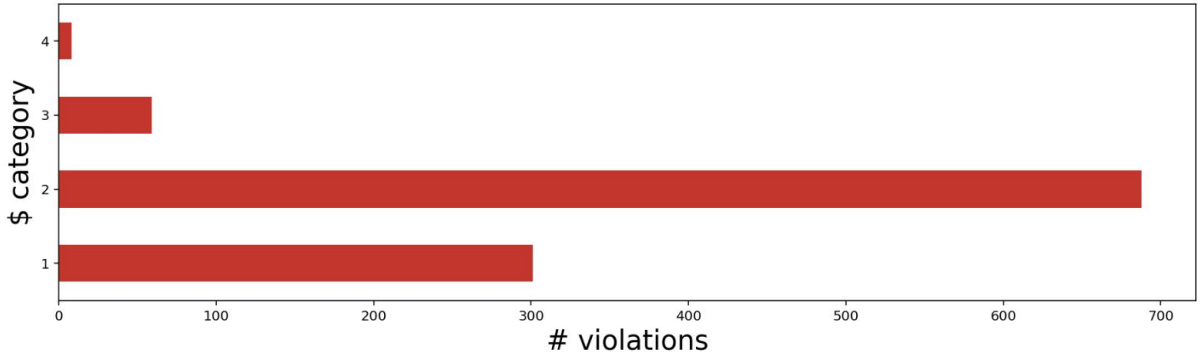
- New restaurants are under more scrutiny
- Popular restaurants maintain higher quality standards, but how can we know if we don't inspect them?

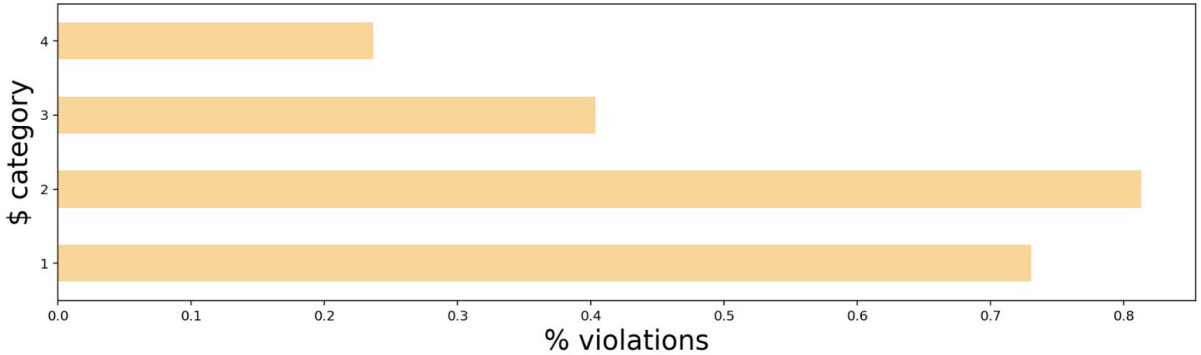


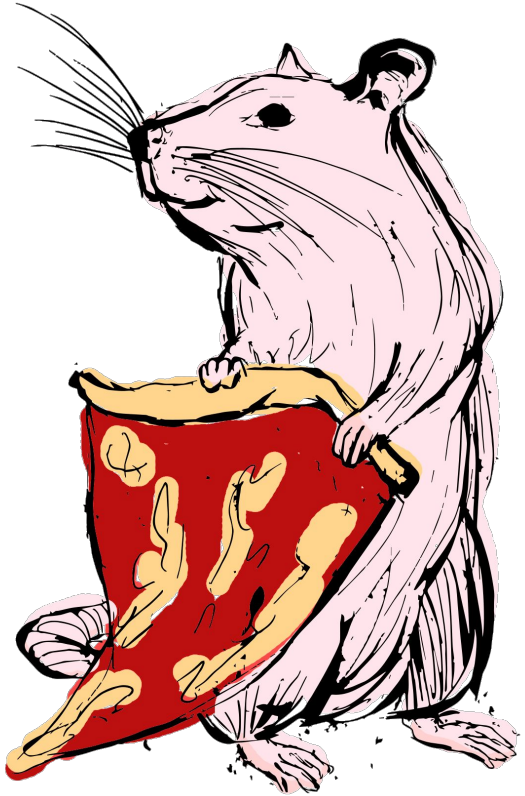
Top 5 health violations across Yelp price categories

Violation Description	% total inspections within ($) range			
	$	$$	$$$	$$$$
Non-food contact surface or equipment improperly maintained and/or not properly sealed, raised, spaced or movable to allow accessibility for cleaning on all sides, above and underneath the unit.	14	13	13	15
Facility not vermin proof / Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.	10	10	9	8
Food contact surface not properly washed, rinsed and sanitized after each use and following any activity when contamination may have occurred.	NA	7	9	8
Food not protected from potential source of contamination during storage, preparation, transportation, display or service.	7	7	7	NA
Evidence of mice or live mice present in facility's food and/or non-food areas.	7	7	NA	NA
Cold food item held above 41° F, except during necessary preparation.	7	NA	6	7
Plumbing not properly installed or maintained; equipment or floor not properly drained; sewage disposal system in disrepair or not functioning properly.	NA	NA	NA	7

Does more $$$ = less rats?





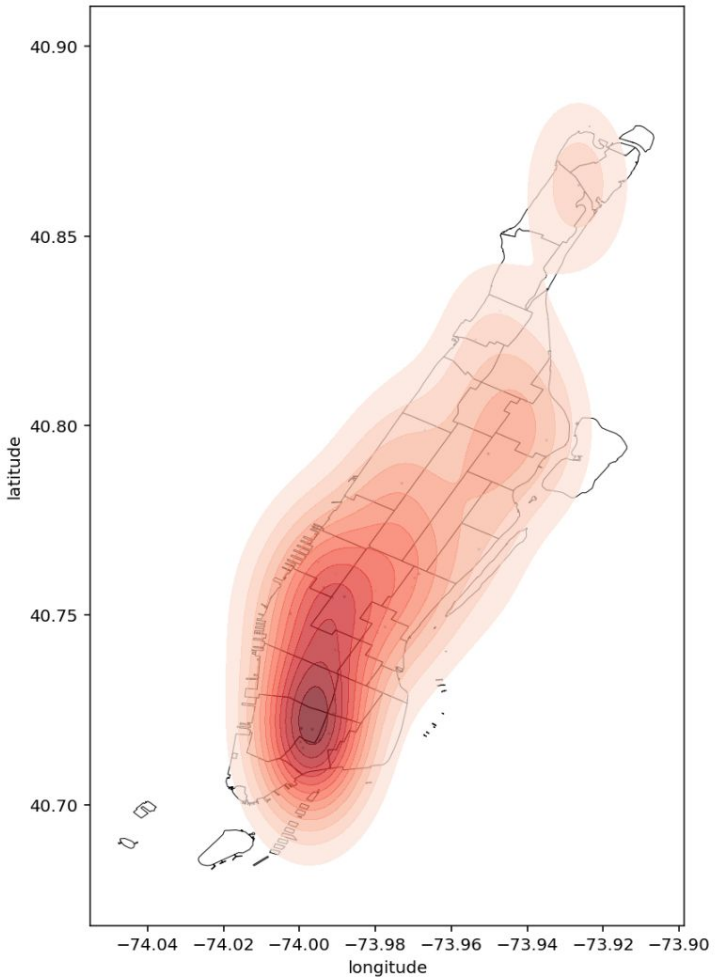


Where to Avoid

High Yelp Ratings (4.5+ stars)

BUT

Low Health Ratings (At least 1 C rating)

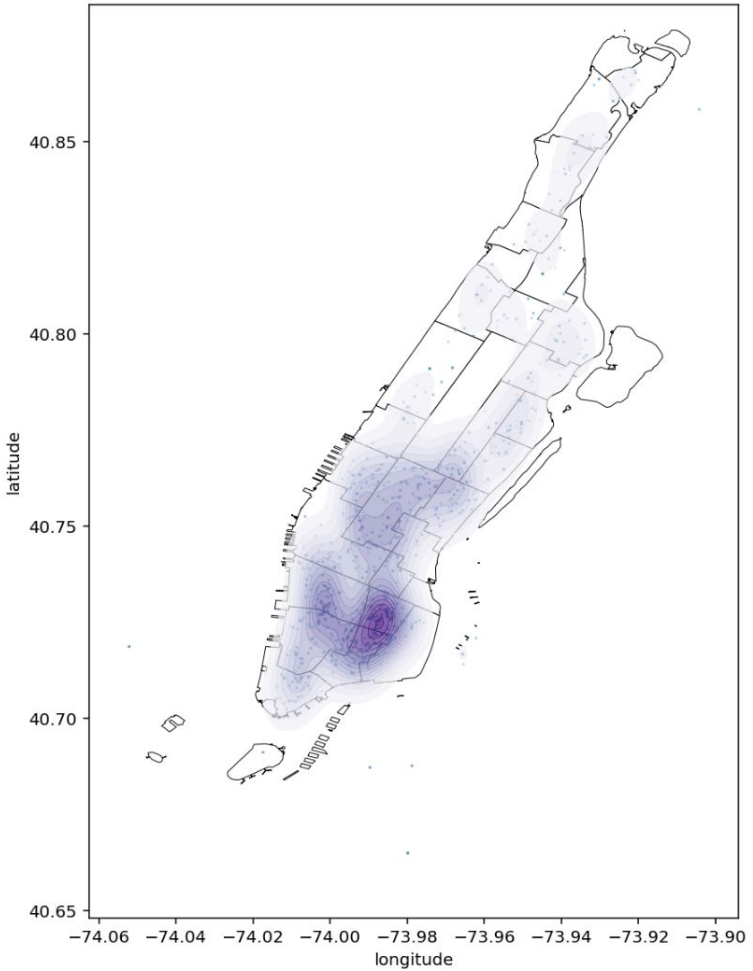


Where to Go

High Yelp Ratings (4.5+ stars)

AND

A Ratings



Questions?