ANALYZING RESTAURANT TRENDS IN SAN FRANCISCO

PRESENTED BY MAYA REISS AND BONNIE SHEN

HEY THERE!

Nice to meet you.

MAYA

- Keeps a pulse on the newest, trendiest eateries in SF
- Canadian, currently lives in SF
- Highlight of her day is lunch

BONNIE

- Yelp Elite Community member
- Loves to take photos of her food but never posts them
- Bay Area native

THE STEPS TO OUR PROCESS

A story about loss and triumph.

Concept

Gather + Clean Up Data

Analyze

Cry

Reflect

WHAT WE WANT TO KNOW

Being food enthusiasts, we set out to find the following information about restaurant trends and lifecycles in the different neighborhoods in San Francisco.

1. How many restaurants open in San Francisco per year versus how many close?
2. What neighborhoods see the most turnover?
3. What is the breakdown on cuisines of restaurants in San Francisco?
4. Popularity?

THE STEPS TO OUR PROCESS

A story about loss and triumph.

Concept

Gather + Clean Up Data

Analyze

Cry

Reflect

Our Data Sources

1. We first combed through San Francisco city public records which was very DIRTY (more on that later)
2. After we gathered a list of businesses that applied for permits to operate in San Francisco, we crossed that with an API call to Yelp to clean the data further and patch missing fields such as popularity, categories, and average price

City Data

SF CITY DATA

We downloaded the Registered Business Locations from SFgov.org. This dataset includes the locations of businesses that pay taxes to the City and County of San Francisco. Each registered business may have multiple locations and each location is a single row. The Treasurer & Tax Collector's Office collects this data through business registration applications, account update/closure forms, and taxpayer filings.

DIRTY AS F*CK

We started with a .CSV containing ~250,000 businesses that we narrowed down to 12,000 food services, but it was very dirty. How dirty? Fields were incomplete, inconsistent, incorrectly categorized and misspelled. Ex: these city records for San Francisco had "San Francisco" spelled 30+ different ways...

CLEANING UP...

We thought this was good source to use to pull business name, business type, start date, close date, neighborhood, and zipcode from the CSV, but WE WERE WRONG!!!!

We concluded that we could not determine the following from this dataset:

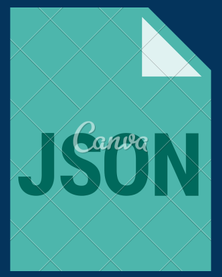
- Business closure date - not reliable (duplicates, wrong dates)
- Neighborhood - inconsistent

Yelp API Calls

RESTAURANT DATA

We wanted to run the API call by "Business Name" and "Street Address" to determine more information about the restaurants in our CSV.

We were limited to 5000 API calls per person per day, so getting our API information for 12,000 queries took about 2 days between the two of us



HOPEFULLY...

We intended to collect the following information from the Yelp API:

- Confirmation that the restaurant was still located at the address queried
- Longitude/Latitude
- Price Range
- Rating
- Whether the restaurant was still open
- What top category was the restaurant, or what type of food was served

99 PROBLEMS

Another major issue was that Yelp would fill the first match for Business Name + Street Address. This led to a lot of mismatches which we had to clean, leaving us with ~3000 lines of data to analyze.

After we completed our API calls, we found out we still could not match up the restaurant to the SF neighborhood due to insufficient map information.

PROBLEMS WE ENCOUNTERED

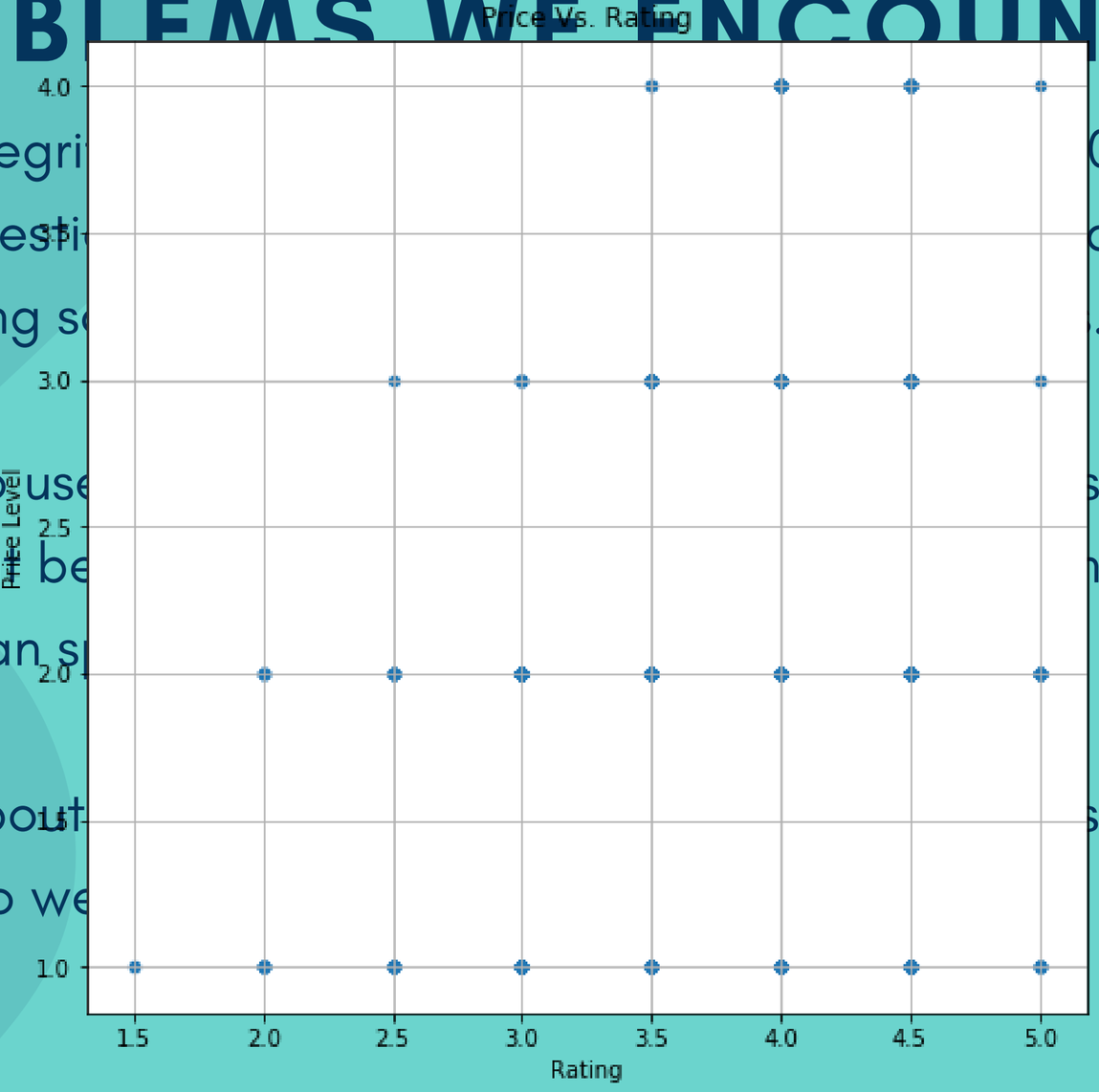
- Because of the integrity of our data - we were left with about 3000 lines of data
  - Some of the questions we had originally laid out couldn't be answered because our data was missing so much, so we had to adjust our questions.

- We were hoping to use a restaurant's zip code to pinpoint its respective neighborhood, but that didn't work out because a zip code can have multiple neighborhoods - and vise versa: a neighborhood can spread across multiple zip codes.

- We had to think about how to analyze and provide visualizations of the data in a meaningful way, so we wouldn't end up with graphs like this:

PROBLEMS WE ENCOUNTERED

- Because of the integri 000 lines of data
  - Some of the questi answered because our data was missing s .
- We were hoping to use spective neighborhood, but that didn't work out be hborhoods. And vise versa: a neighborhood can s
- We had to think about s of the data in a meaningful way, so we



THE STEPS TO OUR PROCESS

A story about loss and triumph.

Concept

Gather + Clean Up Data

Cry

Analyze

Reflect

THE STEPS TO OUR PROCESS

A story about loss and triumph.

Concept

Gather + Clean Up Data

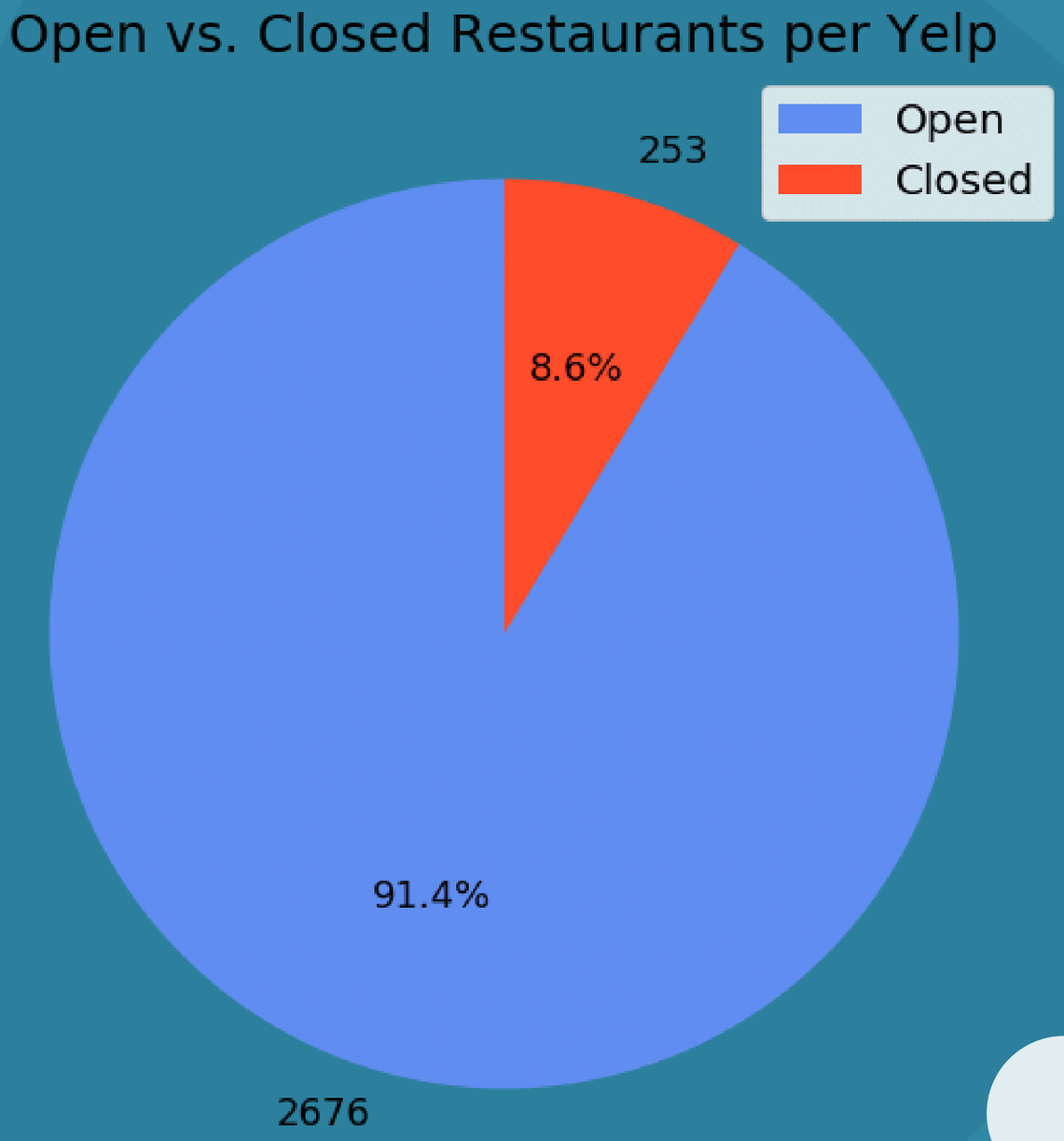
Cry

Analyze

Reflect

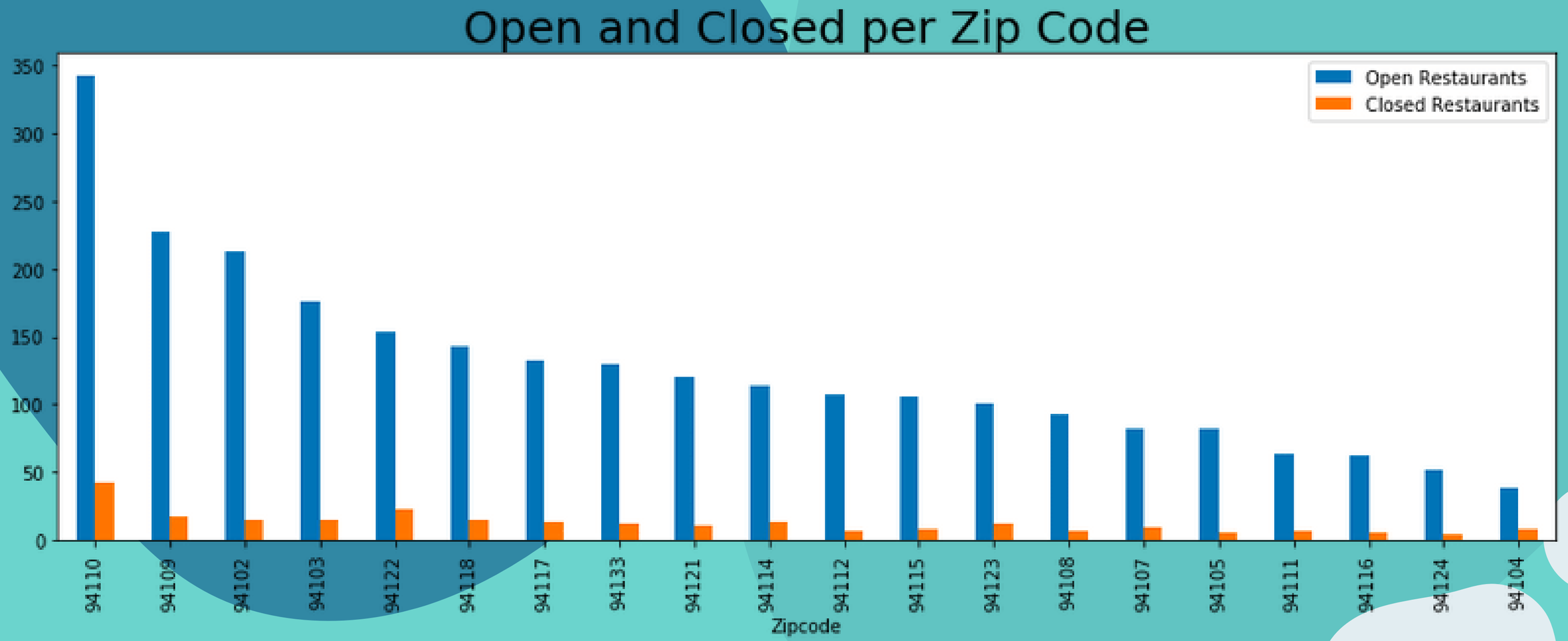
OF THE RESTAURANTS OPEN IN SAN FRANCISCO, WHAT PERCENT HAS RECENTLY CLOSED?

Yelp helped us cross reference which names had closed.



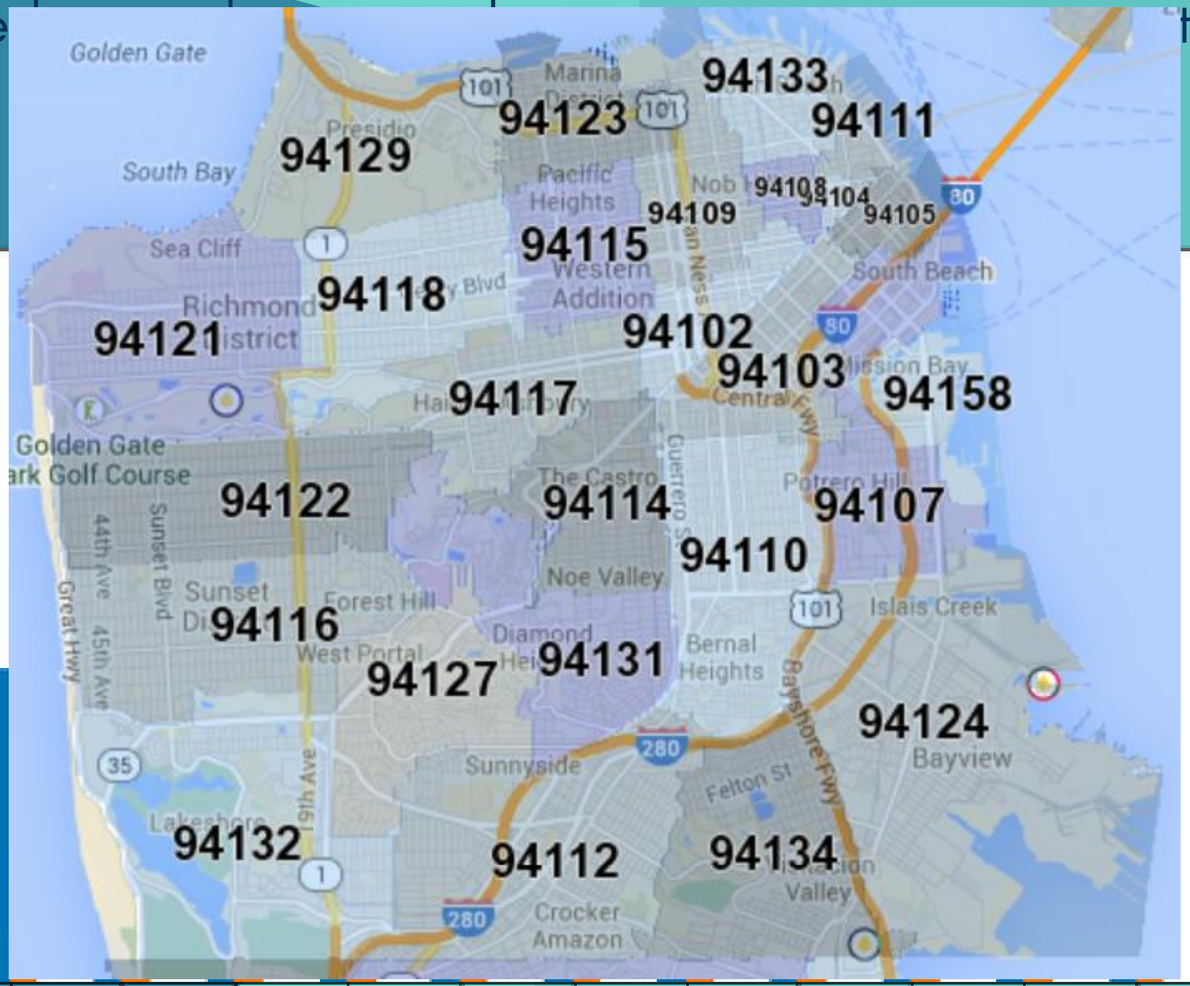
WHICH WERE THE TOP 25 ZIP CODES WITH THE MOST RESTAURANTS?

We also compared to see if having more open restaurants correlated with having more closed restaurants.



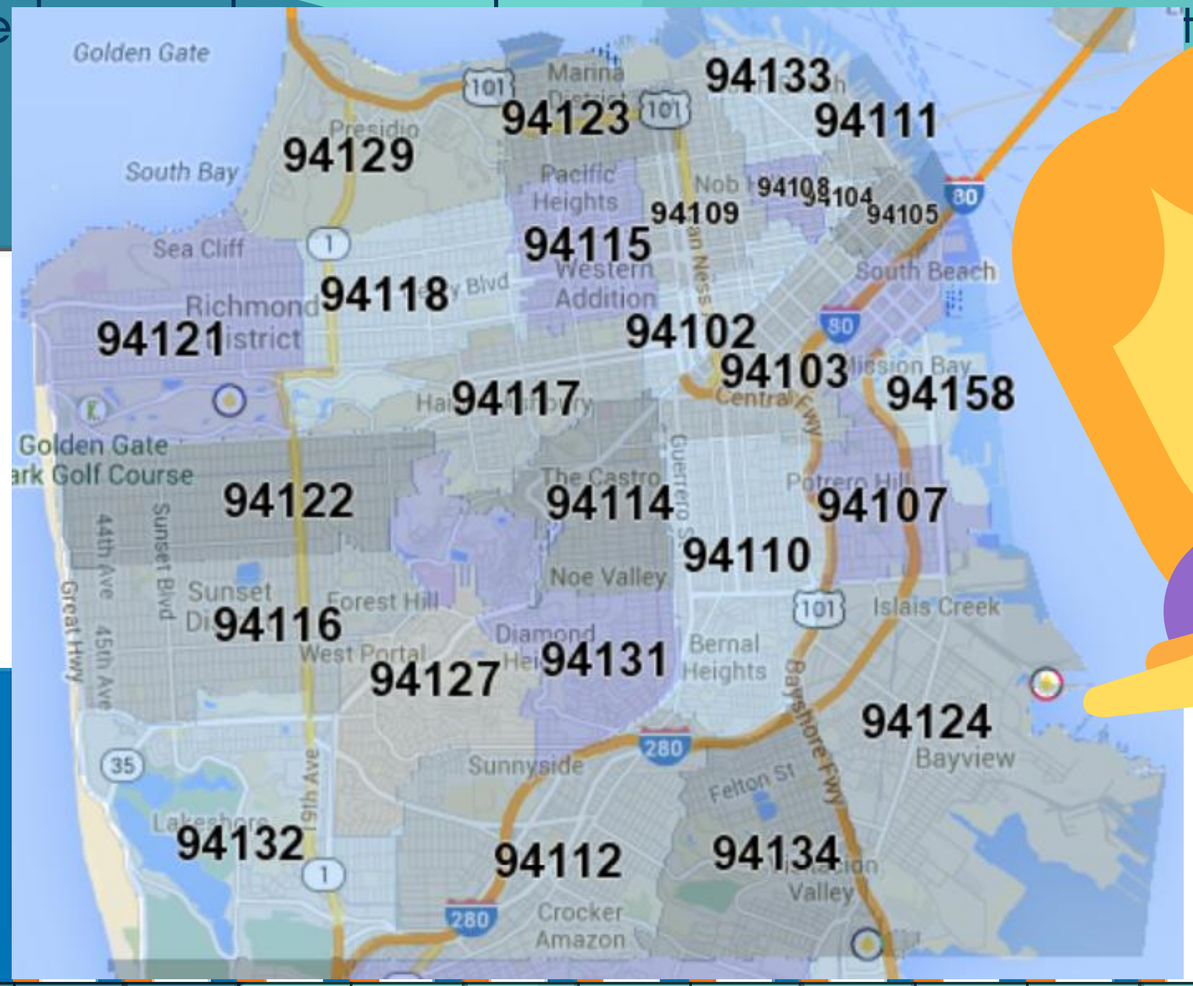
WHICH WHERE THE TOP 25 ZIPCODES WITH THE MOST RESTAURANTS?

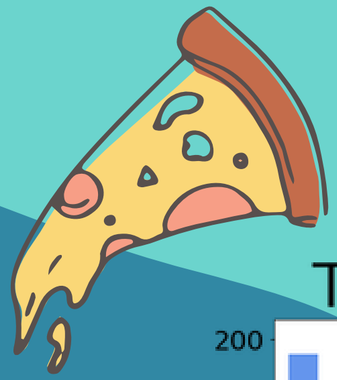
We also compare [illegible] ts correlated with



WHICH WHERE THE TOP 25 ZIPCODES WITH THE MOST RESTAURANTS?

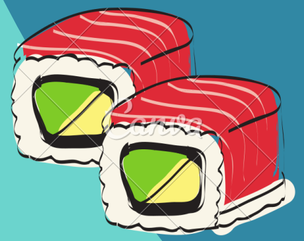
We also compare [illegible] its correlated with

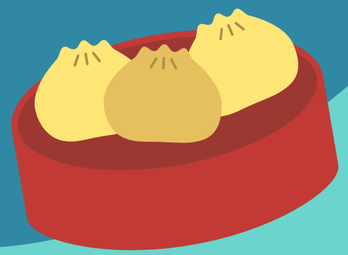


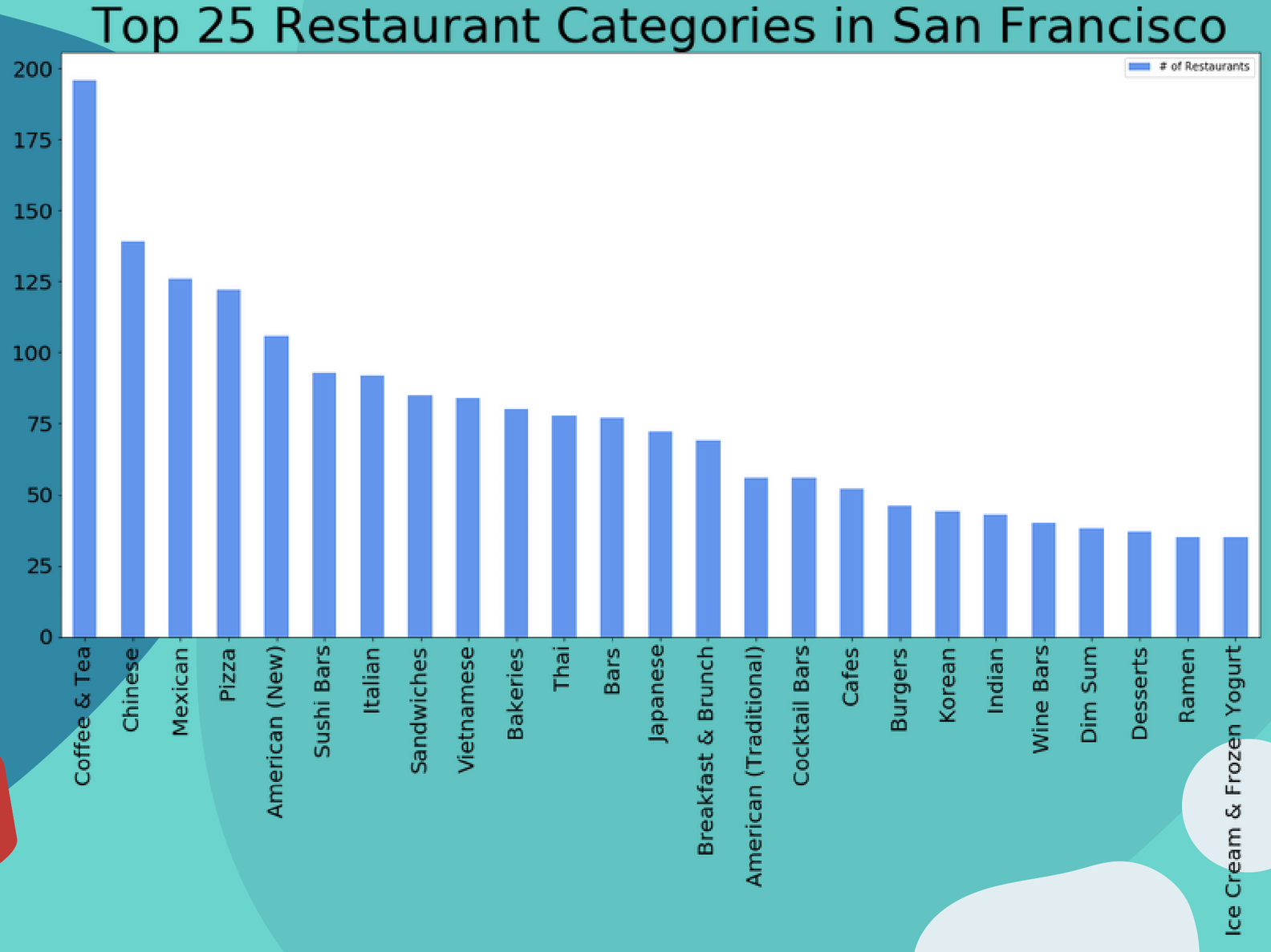


WHAT IS THE BREAKDOWN ON THE CUISINES?

We pulled the Top 25 Categories from Yelp.

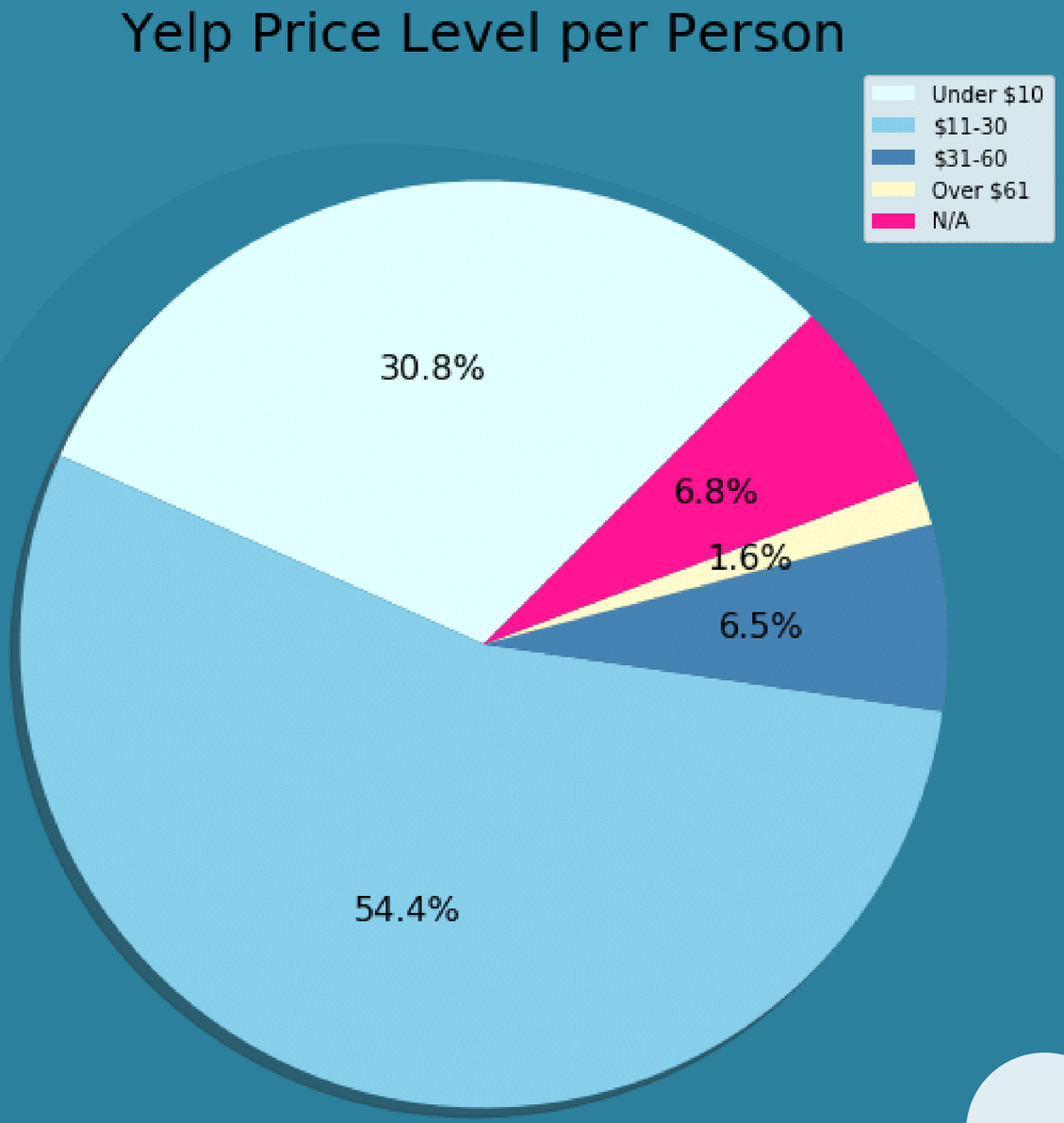






WHAT IS THE PRICE RANGE FOR THE RESTAURANTS?

How much was the average tab per person, and did price range affect the number of restaurants open?



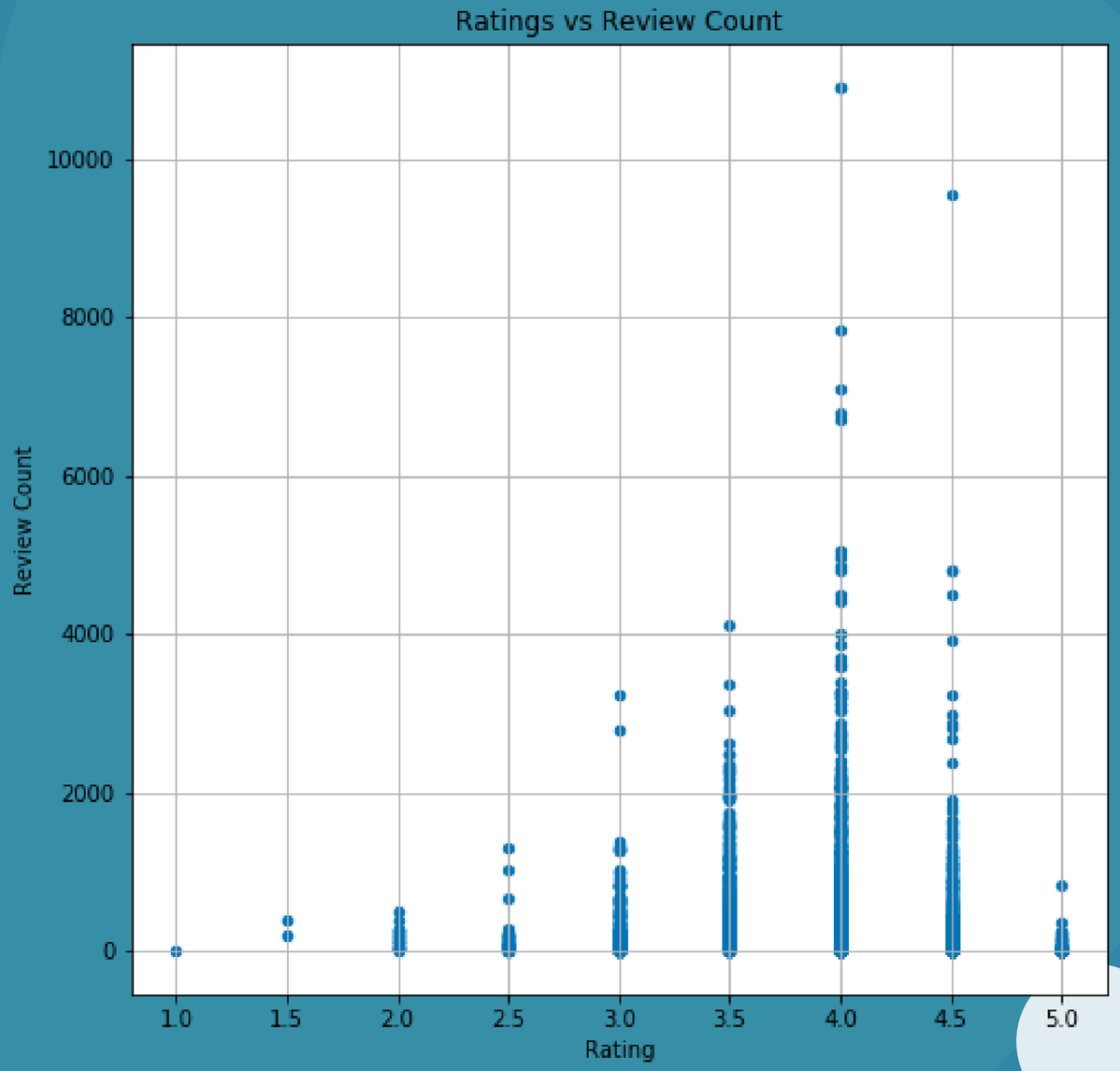
POPULARITY OF RESTAURANTS

WE BASED POPULARITY ON HOW MANY REVIEWS IT RECEIVED.

	Restaurants	Percentage of Total Reviews
0 - 50 Reviews	420	14.34%
51 - 100 Reviews	346	11.81%
101 - 200 Reviews	504	17.21%
201 - 500 Reviews	815	27.83%
501 - 1000 Reviews	476	16.25%
1001 - 2000 Reviews	264	9.01%
2001 - 3000 Reviews	65	2.22%
3000+ Reviews	39	1.33%

WHAT IS THE RATING DISTRIBUTION COMPARED TO NUMBER OF REVIEWS?

Do people like to write good reviews or bad reviews? The general consensus is that people like to give positive reviews



Top Yelp Category	Restaurant Count	Average Rating	Total Reviews	Average Reviews
Coffee & Tea	196	4.03	61663	314
Chinese	139	3.52	53032	381
Mexican	126	3.80	61625	489
Pizza	122	3.60	50145	411
American (New)	106	3.91	69666	657
Sushi Bars	93	3.89	45618	490
Italian	92	3.85	68609	745
Sandwiches	85	3.94	29453	346
Vietnamese	84	3.76	37905	451
Bakeries	80	3.96	47328	591
Thai	78	3.82	38274	490
Bars	77	3.90	43698	567
Japanese	72	3.89	28595	397
Breakfast & Brunch	69	3.99	73450	1064
American (Traditional)	56	3.81	49178	878
Cocktail Bars	56	4.11	22491	401
Cafes	52	4.07	8321	160
Burgers	46	3.67	23100	502
Korean	44	3.89	22220	505
Indian	43	3.90	23911	556
Wine Bars	40	4.22	10941	273
Dim Sum	38	3.63	20838	548
Desserts	37	4.11	14389	388
Ramen	35	3.89	20573	587
Ice Cream & Frozen Yogurt	35	4.03	28842	824

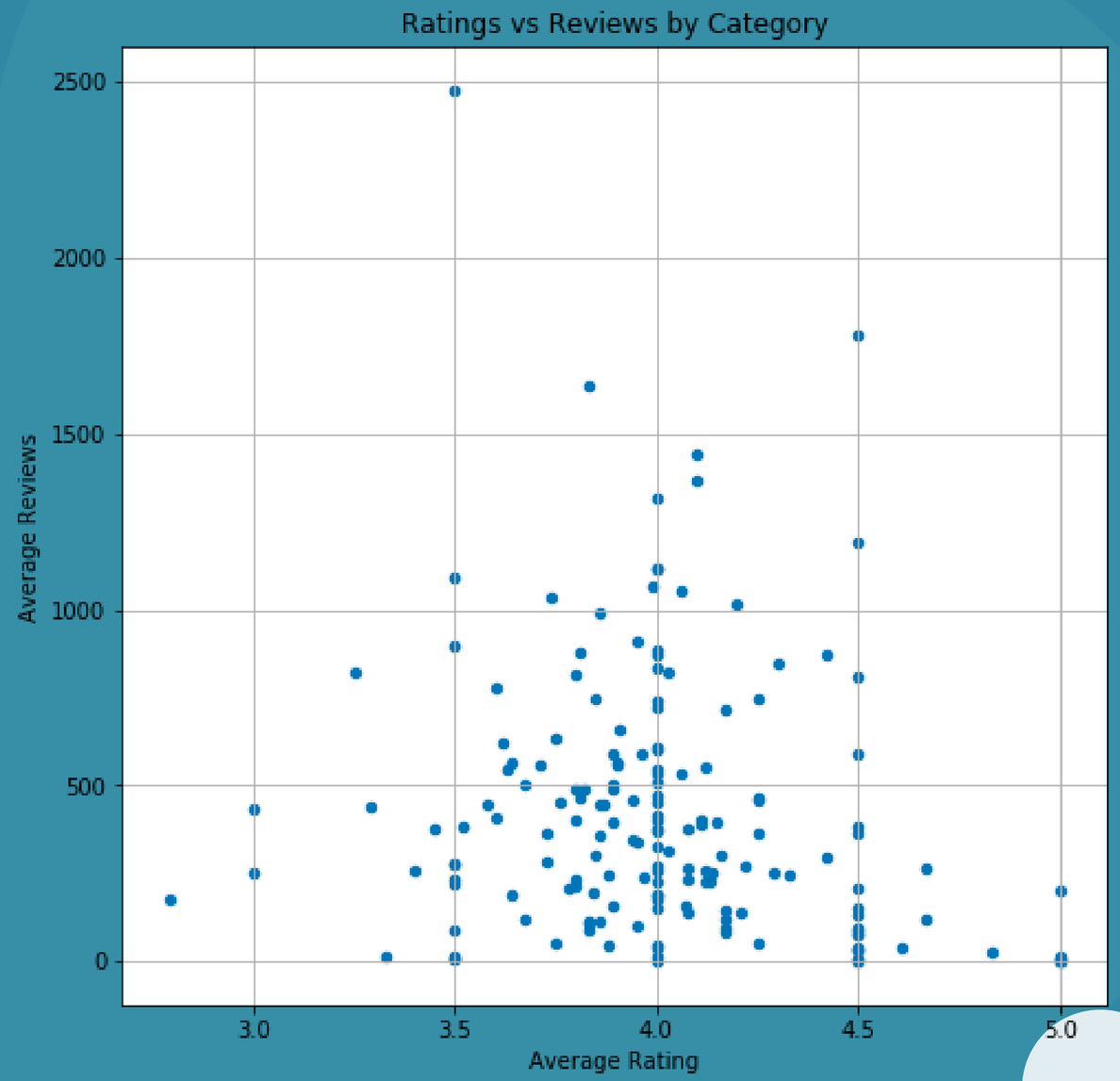
RATING DISTRIBUTION FOR MOST POPULAR CATEGORIES

WE WERE ABLE TO TAKE THE AVERAGE RATING FOR THE MOST POPULAR CATEGORIES AND PLOT ON A SCATTERPLOT

OVERALL, SAN FRANCISCANS REALLY LOVE THEIR BREAKFAST, BRUNCH AND COFFEE!

WHAT IS THE AVERAGE RATING DISTRIBUTION COMPARED TO THE AVERAGE NUMBER OF REVIEWS?

What was the average rating per category?



THE STEPS TO OUR PROCESS

A story about loss and triumph.

Concept

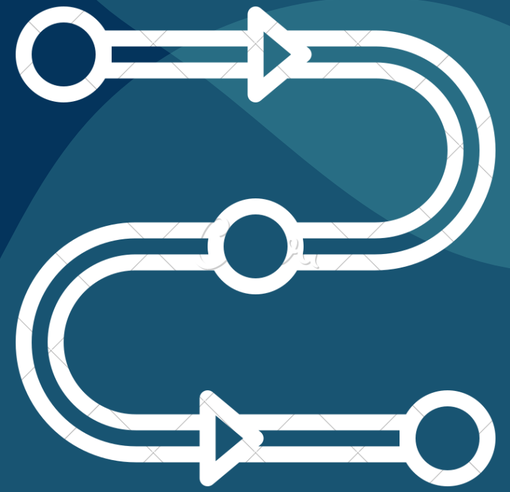
Gather + Clean Up Data

Cry

Analyze

Reflect

HOW WE COULD HAVE IMPROVED



GROUP CATEGORIES

We could have combined different types of cuisines or pulled secondary categories from Yelp

Ex: Sushi Bars with Japanese or Wine Bars/ Speakeasies/ Cocktail Bars with Bars

DIG DEEPER ON NEIGHBORHOODS

The neighborhood lines can be blurry - even to Bay Area natives. We didn't have time to do another API pull with Google Maps.

MORE 4-1-1

We could have looked at other sources like Google Reviews, Foursquare APIs to cross-reference.

THE STEPS TO OUR PROCESS

A story about loss and triumph.

Concept

Gather + Clean Up Data

Cry

Analyze

Reflect

Don't cry because it's over,
smile because it's over.

BONNIE SHEN

QUESTIONS?