San Francisco Restaurants

Data Visualization Final Project

tinyurl.com/cs560-restaurants

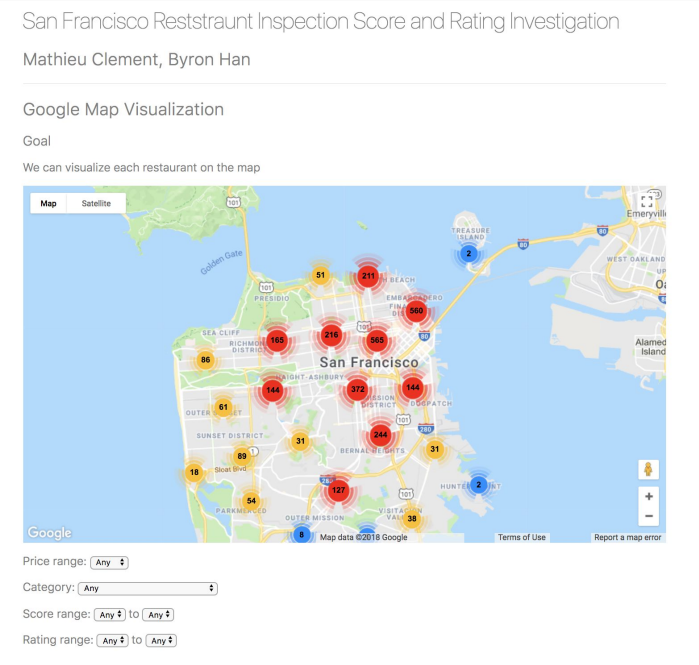
Mathieu Clément, Byron Han

CS360/CS560 USF

Objectives

- Visualize restaurants inspection scores and violations of all San Francisco
- Visualize changes of inspection scores over time
- Visualize average inspection score of SF neighborhoods
- Tell which type of restaurants (Chinese, French, etc.) has the cleanest, highest rated establishments
- Tell which type of restaurants has the best ratings

Visualize restaurants on Google Maps API



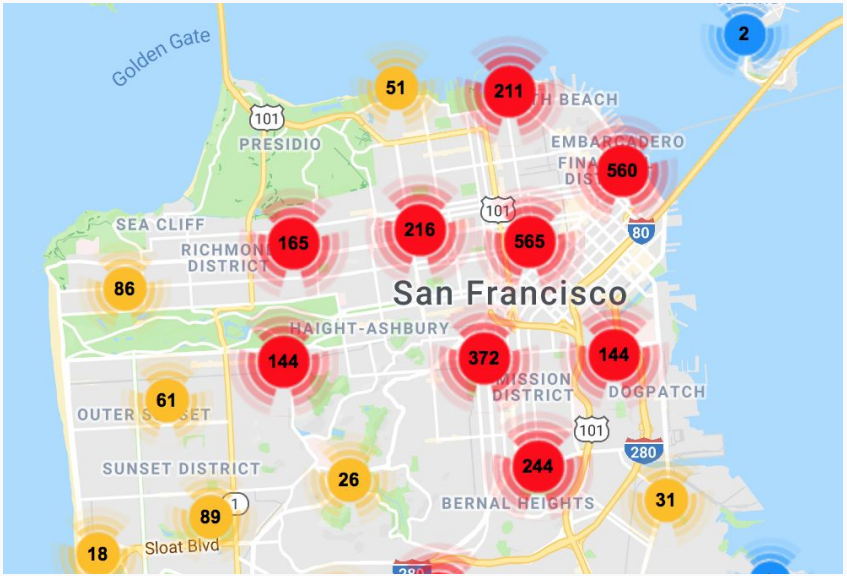
-Clustering used due to high number of establishments in the city

-Color of discs shows density (number of restaurants in the area)

-After zooming in, you will see a marker for each restaurant. The color represents the inspection score.

-Filters such as category, price range, score and rating can be used to search restaurants

Legend (clusters)

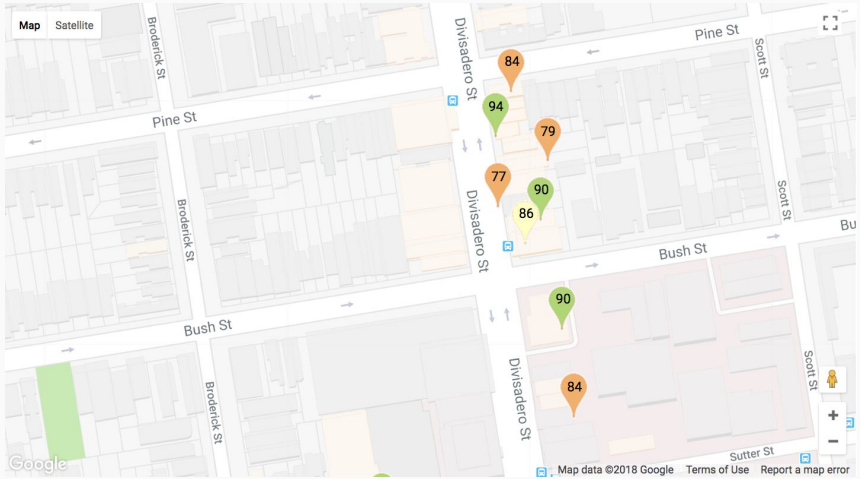


-Each disc represents 2 or more restaurants

-The color represents the density

Blue:	2 - 9
Yellow:	10 - 99
Red:	100 - 999
Purple:	1000+

Legend (inspection scores)



-Each marker represents a restaurant. The text is the most recent inspection score.

-Each one has a color representing the cleanliness of the restaurant:

Dark green:	98 - 100
Light green:	90 - 97
Cream:	86 - 89
Orange:	71 - 85
Red:	50 - 71

Take Home message: *dirty and clean restaurants right next to each other*

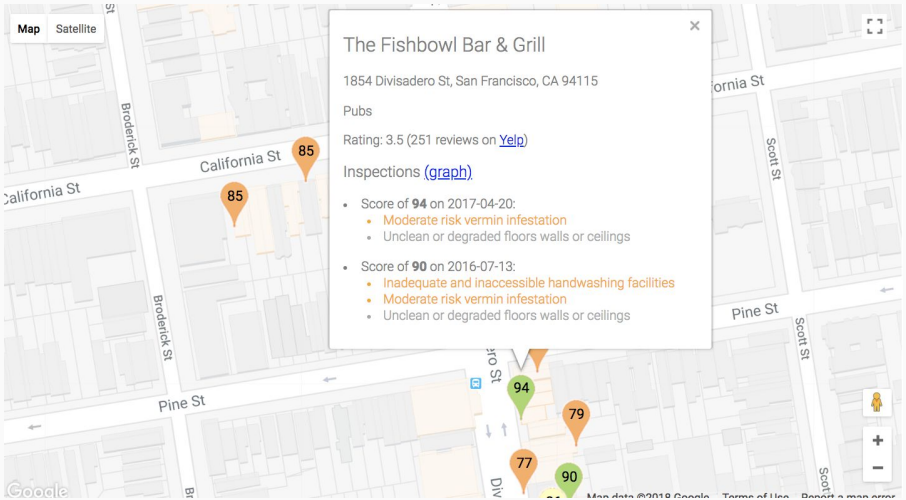
SF Health Department score and risk definitions

- **High Risk:** Violations that directly relate to the transmission of food borne illnesses, the adulteration of food products and the contamination of food-contact surfaces.
- **Moderate Risk:** Violations that are of a moderate risk to the public health and safety.
- **Low Risk:** Violations that are low risk or have no immediate risk to the public health and safety.

Food Safety Score Categories and Interpretation

Score	Operating Condition Category	Inspection Findings
>90	Good	• Typically, only lower-risk health and safety violations observed • May have high-risk violations
86-90	Adequate	• Several violations observed • May have high-risk violations
71-85	Needs Improvement	• Multiple violations observed • Typically, several high-risk violations
Less than or equal to 70	Poor	• Multiple violations observed • Typically, several high-risk violations

Visualize restaurants on Google Maps API



-When you zoom in you will see the detailed inspection record for this restaurant, with:

- a link to the page of the restaurant on Yelp

- a link to the line chart showing the inspection score over time

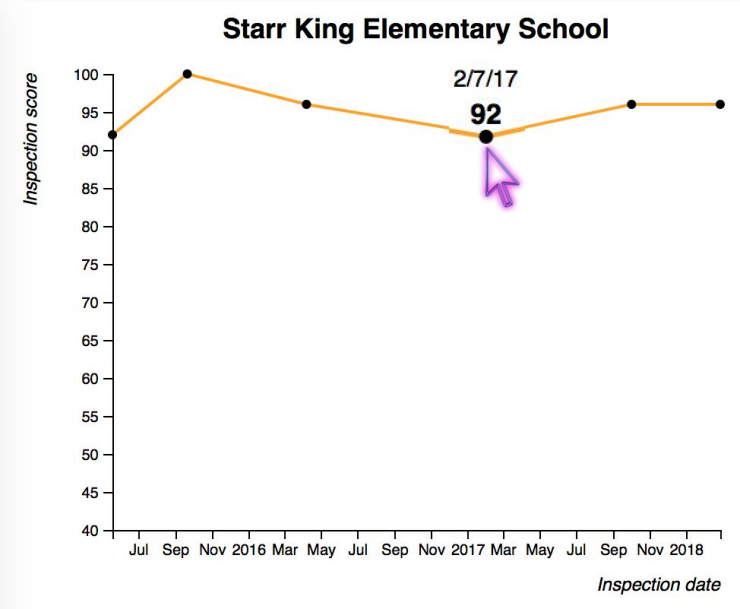
- the list of violations

red: high risk,

orange: moderate risk

gray: low risk

Inspection score over time



-The y-axis is the the inspection score

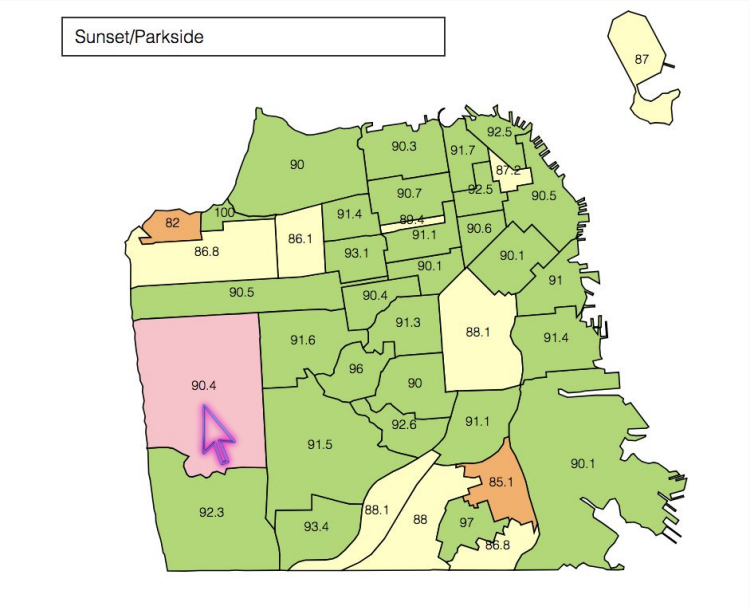
-The x-axis is the time

-If the restaurant only has one inspection then we will not display the line chart

- Interaction: score and date shown on mouse over event

Take Home message: *restaurants seem to stay within same risk category = same color on the map.*

Average inspection score by neighborhood

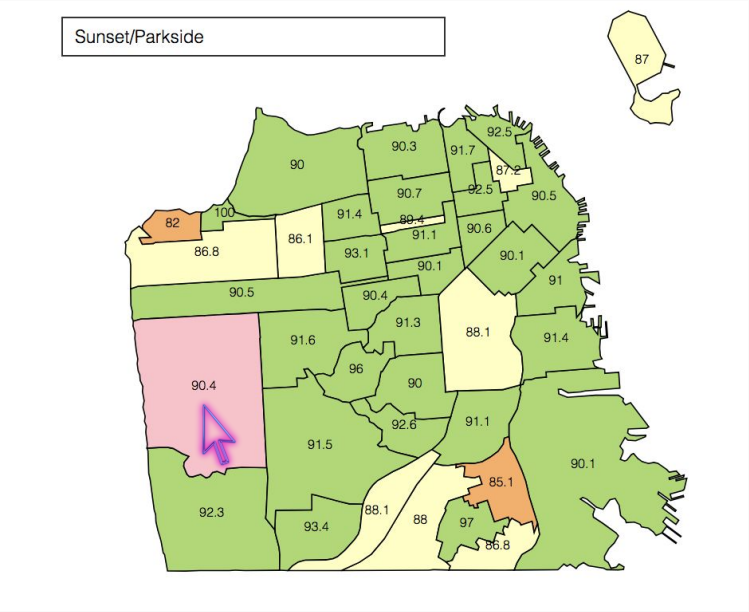


- The average inspection score is displayed for each district considered by the SF Health Dept.

Legend:

Green:	90 - 100
Cream:	86 - 89
Orange:	71 - 85
Red:	50 - 70

Average inspection score by neighborhood



- Moving the mouse over the map highlights the districts. The name smoothly appears and disappears in the top left corner.

Take Home message: *density has substantial effect on an average. Only Portola and Lincoln need improvement. On average SF is safe.*

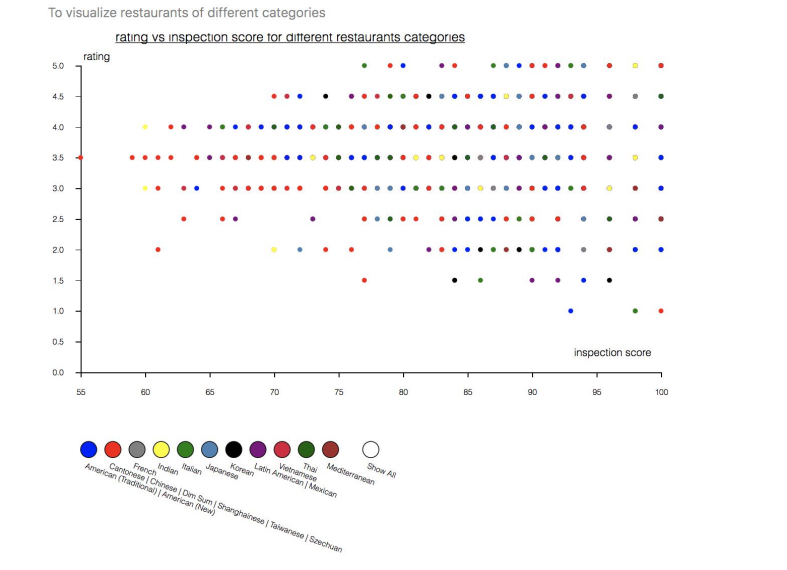
SF Health Department score and risk definitions

- **High Risk:** Violations that directly relate to the transmission of food borne illnesses, the adulteration of food products and the contamination of food-contact surfaces.
- **Moderate Risk:** Violations that are of a moderate risk to the public health and safety.
- **Low Risk:** Violations that are low risk or have no immediate risk to the public health and safety.

Food Safety Score Categories and Interpretation

Score	Operating Condition Category	Inspection Findings
>90	Good	• Typically, only lower-risk health and safety violations observed • May have high-risk violations
86-90	Adequate	• Several violations observed • May have high-risk violations
71-85	Needs Improvement	• Multiple violations observed • Typically, several high-risk violations
Less than or equal to 70	Poor	• Multiple violations observed • Typically, several high-risk violations

Cleanest and highest rated cuisines



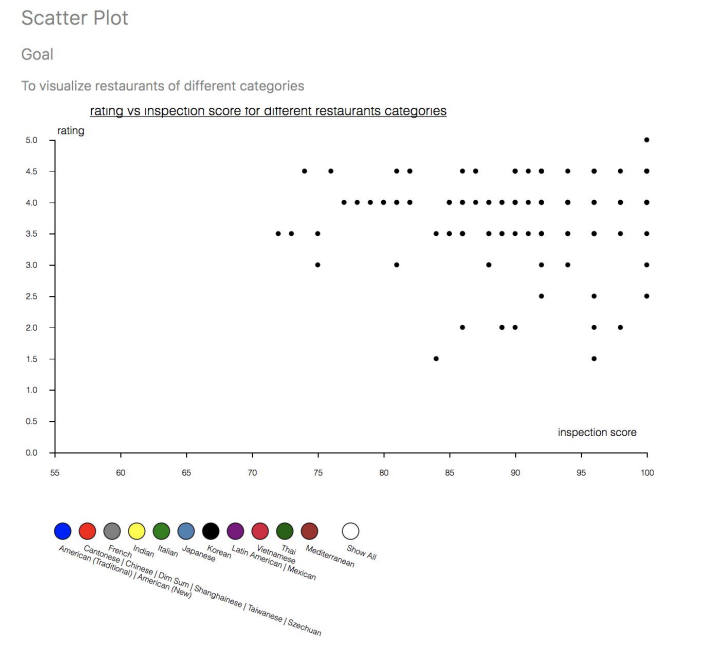
- Intuitive visualization to see if there is a correlation between inspection score and Yelp ratings

- Cleaner = better rated?

=> Clearly this is not true (no visible tendency)

-No strong positive correlation between score and rating

Cleanest and highest rated cuisines



-We can also select different category

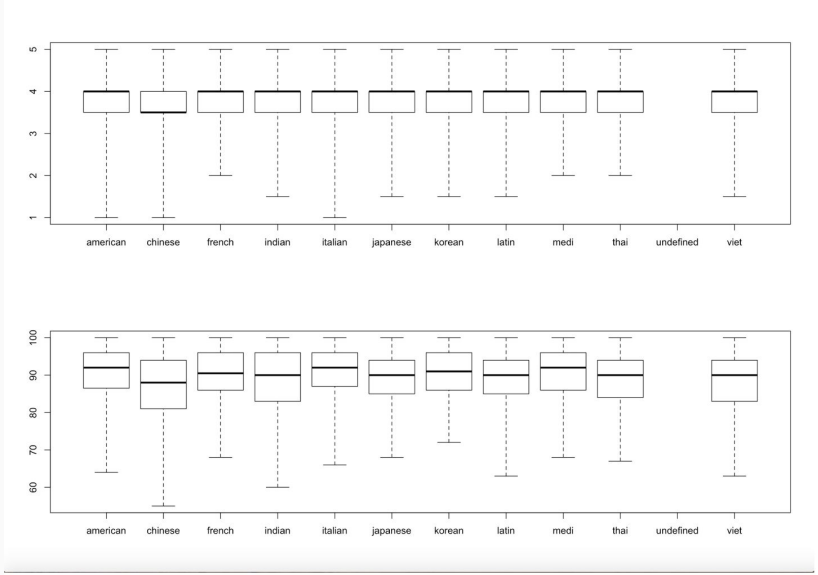
-To see if within each category there is any correlation

-Clearly it doesn't either

-There might exist positive correlation, but not quite strong

-Mouse is not there because taking screenshots

Cleanest and highest rated cuisines



-Boxplot is often a powerful way to visualize these difference

-It shows the mean and quartile of the numeric data

-We specify no outliers as whisker extend to infinity

-All means and quartiles are very close

-(Chinese has relatively lower)

Distribution of inspection scores

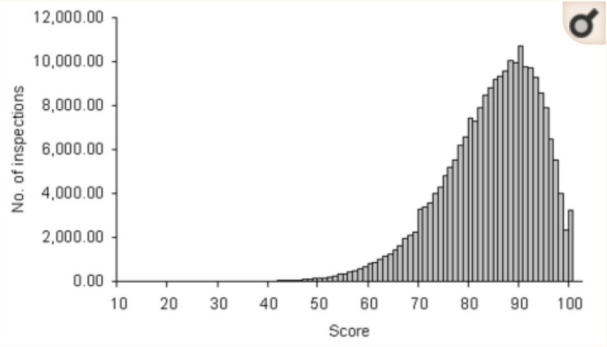


Figure 1

Distribution of scores of restaurants inspected statewide from July 1993 to June 2000, based on a standardized inspection with 44 scored items and a maximum score of 100.

-So it is better to use probability tool

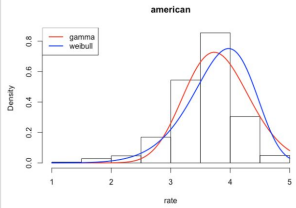
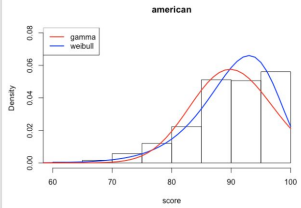
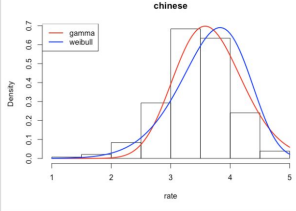
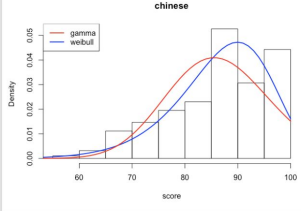
-In this paper researcher investigate all restaurants in Tennessee from 1993 to 2000

-They have put all category in one basket and came up with this distribution for the inspection score

-They didn't specify the distribution type, only says it is skewed and calculated the mean

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3323064/

Cleanest and highest rated cuisines

	Rating	Score
American		
	Average 3.823903	Average 90.89365
Chinese		
	Average 3.823899	Average 90.91713

-Used two most common skewed distribution:

Gamma and Weibull

-Then find MME of the distribution and use it as a rough estimation

-Then use built-in maximize log-likelihood function to find MLE estimation of the distribution

-Result shows that indeed each category has very close average of rating and cleanliness.

Preprocessing

Main map

- Use CSV from SF Health Department with name, address, inspections
- Exclude incomplete records
- Retrieve Yelp Data JSON with Yelp rating, cuisine
- Merge all into a single JSON file with Python

Neighborhood map

- Convert shapefiles into GeoJSON with MapShaper
- Match geo coordinate to Polygon
- Calculate average per polygon a.k.a. district/neighborhood

Preprocessing

Scatter-plot

- Use Javascript to parse data into csv file
- Remove unwanted columns and only save id, category, rating and score
- Remove NAs and empty data

Box-plot and Bar-chart with Distribution Curve

- Use raw csv for box-plot
- Use method of moments estimation as initialization
- Iteratively find the optimize average by using built in NR method

Conclusion (Take Home Messages compilation)

-On average all neighborhoods are safe, but more variation exists within some of them (e.g. Chinese).

-Clean and dirty restaurants are often next to each other.

-Restaurant cleanliness tends to stay the same over time.

-Scores only for SF and NYC on Yelp, but cities often have information available online.

-No strong positive correlation between rating and cleanliness overall or

- at least within each category
- at least within San Francisco

Thank You !