San Francisco Restaurants

Project Proposal (revised)

Mathieu Clément <mclement2@usfca.edu>
Byron Han <zhan12@usfca.edu>

Project repository: https://github.com/tiktaktok/cs560-restaurants
Project website: https://tiktaktok.github.io/cs560-restaurants

CS 360/560 Data Visualization, University of San Francisco
April 11, 2018

Background and motivation

If you have ever eaten out in San Francisco, chances are that you have seen a document from the San Francisco Health Department displayed prominently in the restaurant, reporting the results of their announced and unannounced inspections. City officials will assess whether there are violations regarding public health and safety, in order to reduce food borne illnesses. Inspections receive a score from 0-100 where 100 means no violations have been observed.

While the health score and violations are interesting in themselves, the data is relatively dry. Not only do we want to visualize that data (we can easily plot it on a map), but it might be interesting to fuse this data with information fetched from Yelp, such as reviews, category of restaurant, prices, etc.

We have found a relevant dataset containing inspections including health scores from the SF Health Department on Kaggle (detailed below) and successfully used the Yelp API through Postman, which led us to believe that this project was feasible and that the information is available. We have also chosen this project because we were not able to find any insights regarding the subject matter, and we are interested in the subject. We hope to find correlations in the data, and allow the public to make smart decisions when picking up a place to eat.

Currently Yelp does not offer to filter results based on the health score. And we believe this is a shame.

Related Work

Related works which we can learn from:
*:interaction included
*bar chart: previous assignment
*bi/tri/multivariate scatter plot: previous assignment
*line chart: previous assignment
tree: for tree we can do regression tree or classification tree so that we can predict values
tree graph: https://cran.r-project.org/web/packages/data.tree/vignettes/data.tree.html
boxplot: https://www.statmethods.net/graphs/boxplot.html
*Geo map: http://bl.ocks.org/micahstubbs/8e15870eb432a21f0bc4d3d527b2d14f
Google map api: For pin point the restaurants
*Hex tile map: https://bl.ocks.org/eesur/8678df74ee7efab6d645de07a79ebcc5

Project objectives

We would like to be able to answer the following questions:

- What Chinese restaurant has the highest rating on Yelp and is the cleanest?
- Are Indian restaurants cleaner than their Chinese counterpart?
- Is there a correlation between: health score, price, Yelp, rating, cuisine, ZIP code, etc. ?
- What was the health score and which violations were observed at restaurant X?
- What is the distribution of health scores?
- What is the average for each ZIP code (941xx) / neighborhood in San Francisco?
- Has the health score improved since the last inspection of restaurant X?
- What are the restaurants with High Risk, Moderate Risk, Low Risk?

We would also like users to be able to explore the restaurant in the city using a map, markers and tooltips/detail page.

Trivariate/Multivariate scatter plot:

Design Objectives:

1. Depends on how many data we have, 4 variate if data from Yelp is large enough
2. Rating vs. Inspection Score vs. Restaurant Category vs. Price
3. Primary choice: Rating vs Inspection score, category as color, and price as menu to choose from.

User Objectives:

1. User can select which one to have on the graph and which one to be legend
2. Find out if there are any variable has clusters
3. See if any type of restaurant has cluster in cleanness, price etc.

Regular Map

Design Objectives:

1. Either D3 package or Google API or other tools to possibly enable user interaction
2. Able to filter based on risk level, we define certain inspection score for different level
3. Maybe color filtering added, mouse over interaction
4. Tooltips to display certain restaurant info by the side
5. Possible link to timeline of this restaurant's inspection score

User Objectives:

1. Able to see restaurants risk category
2. Show user more info in one page, like what violation

Timeline Graph

Design Objectives:

1. Inspection score vs. Time
2. Mouse over interaction

User Objective:

1. User can see how this restaurant is doing over time

Inspection score distribution bar chart

Design Objectives:

1. Based on the area and zip code we can see the distribution of inspection score

User Objective:

1. See if outlier is contributing to the overall score or not
2. Explore to see the distribution type

Tree plot

Design Objective:

1. Use tree package in R to analyze and visualize the tree structure of regression and classification
2. Rating vs price, zip code, etc.
3. Healther vs category, etc.

User Objectives:

1. Find out if it is easier to encounter a clean restaurant or ratings given specific conditions

2. Predict future restaurant ratings or cleanness

Box Plot

Design Objectives:
1. Inspection score vs price
2. Use R code to analyze and find out percentile and mean of different category,

User Objectives:
1. Have user visual the correlation between price and cleanness
2. Better statistical understanding about the mean and variance and percentile

Hex tile Graph (and Geo Map maybe)

Design Objectives:
1. Geomap usually will affect visualization based on the area of the zip
2. We normalize each zip code and assign average score as a color to the zip code

User Objectives:
1. Have user visualize the correlation of cleanness vs area

Data

Health scores

The relevant dataset for health scores from the SF Health Department is available on Kaggle: https://www.kaggle.com/datasf/sf-restaurant-inspection-scores

It takes the form of a table (CSV file) and uses the following data dictionary: https://docs.google.com/document/d/1eeO5T_lt8QHGHMjpj6M9071y5OFOXcADFI4F7AIZ2iE

The file contains information such as:
- Business name
- Business address
- Business geo coordinates (WGS 84)
- Inspection date
- Inspection type (announced/unannounced)
- Inspection score
- Violations

Additionally this dataset has shapefiles describing the supervisor and districts, as well as the neighborhoods (Bernal Heights, Castro, Inner Richmond, Japantown, etc.)

Yelp

Data from Yelp, to our knowledge, can only be extracted through their public REST API, documented here:
https://www.yelp.com/developers/documentation/v3/business_search

Parameters such as the following can be used for queries:

- Geo coordinates (WGS 84)
- Search term (e.g. business name)
- Phone number
- Street address
- Radius (to limit the search results to a certain area)

The response is a list of restaurants that match completely or partially the search criteria, with the most relevant business listed first.

Here's one example of such a response:

```
{
    "businesses": [
        {
            "id": "1428-haight-patio-cafe-and-crepery-san-francisco",
            "name": "1428 HAIGHT Patio Cafe & Crepery",
            "image_url": "https://s3-media2.fl.yelpcdn.com/bphoto/VHf7PihAjQtfin0BEl-CZw/o.jpg",
            "is_closed": false,
            "url": "https://www.yelp.com/biz/1428-haight-patio-cafe-and-crepery-san-francisco?adjust_creative
                =bs0Nt1SUQSRjLm6j3N3CNA&utm_campaign=yelp_api_v3&utm_medium=api_v3_business_search&utm_source
                =bs0Nt1SUQSRjLm6j3N3CNA",
            "review_count": 280,
            "categories": [
                {
                    "alias": "tradamerican",
                    "title": "American (Traditional)"
                },
                {
                    "alias": "creperies",
                    "title": "Creperies"
                },
                {
                    "alias": "breakfast_brunch",
                    "title": "Breakfast & Brunch"
                }
            ],
            "rating": 4,
            "coordinates": {
                "latitude": 37.7701485211239,
                "longitude": -122.445868540619
            },
            "transactions": [
                "pickup",
                "delivery"
            ],
            "price": "$$",
            "location": {
                "address1": "1428 Haight St",
                "address2": "",
                "address3": "",
                "city": "San Francisco",
                "zip_code": "94117",
                "country": "US",
                "state": "CA",
                "display_address": [
                    "1428 Haight St",
                    "San Francisco, CA 94117"
                ]
            },
            "phone": "+14158648484",
            "display_phone": "(415) 864-8484",
            "distance": 17.47225917657
        },
```


Most notably we find:

- the business name
- the business address and geo coordinates
- the business phone number
- the average rating
- the number of reviews
- the types of restaurant / cuisine (“Breakfast & Brunch”, “American Traditional”)
- the price range

Data Processing

The Health Score dataset has many rows missing the health score. These rows would be removed.
It also contains multiple inspections for the same restaurant; we would most likely keep only the results of the most recent unannounced visit for many of our visualizations.
Before using it in D3, we would also need to remove useless columns to speed things up.

Data from Yelp can only be queried from their API. We would need to write a script that will read the Health Score dataset, try to find the matching business on Yelp, and save the JSON response to a file.
Then we would read those files and augment the original dataset so that it contains the information fetched from Yelp.

Note: we could in theory avoid writing the JSON response to the disk, but we prefer to do so as Yelp restricts the number of queries to 5000 per day per application (API KEY).
This way we can perfect our matching algorithm (from one dataset to the other) and fix possible bugs without wasting Yelp API "credits."

At the end of the day, we would have a JSON file containing one entry for each restaurant, and each of these entries would have a list of inspections (including date and health score), which would have a list of violations. The main restaurant entry would have a name, address, geo coordinates, Yelp rating, number of Yelp reviews, types of restaurant/cuisine, price category.

We will use Python to prepare the data.

Visualization Design

Our top picks

As “must-have” visualizations, we will implement the map with individual markers for each restaurant (fig. 1), the line chart showing the change of the inspection score for a particular restaurant (fig. 2), a scatterplot showing the relationship between ratings and health score (fig. 4) highlighting a few types of restaurant.

Bonus graphs

If they give any interesting insight we can also look at the average inspection for each neighborhood of San Francisco (fig. 3) and the distribution of scores in general (fig. 6).
Fig. 5 compares health score and price in the form of a Box Plot, while fig. 7 is the representation of a regression tree which can be used to predict the score based on variables such as price, type of cuisine, location, etc.

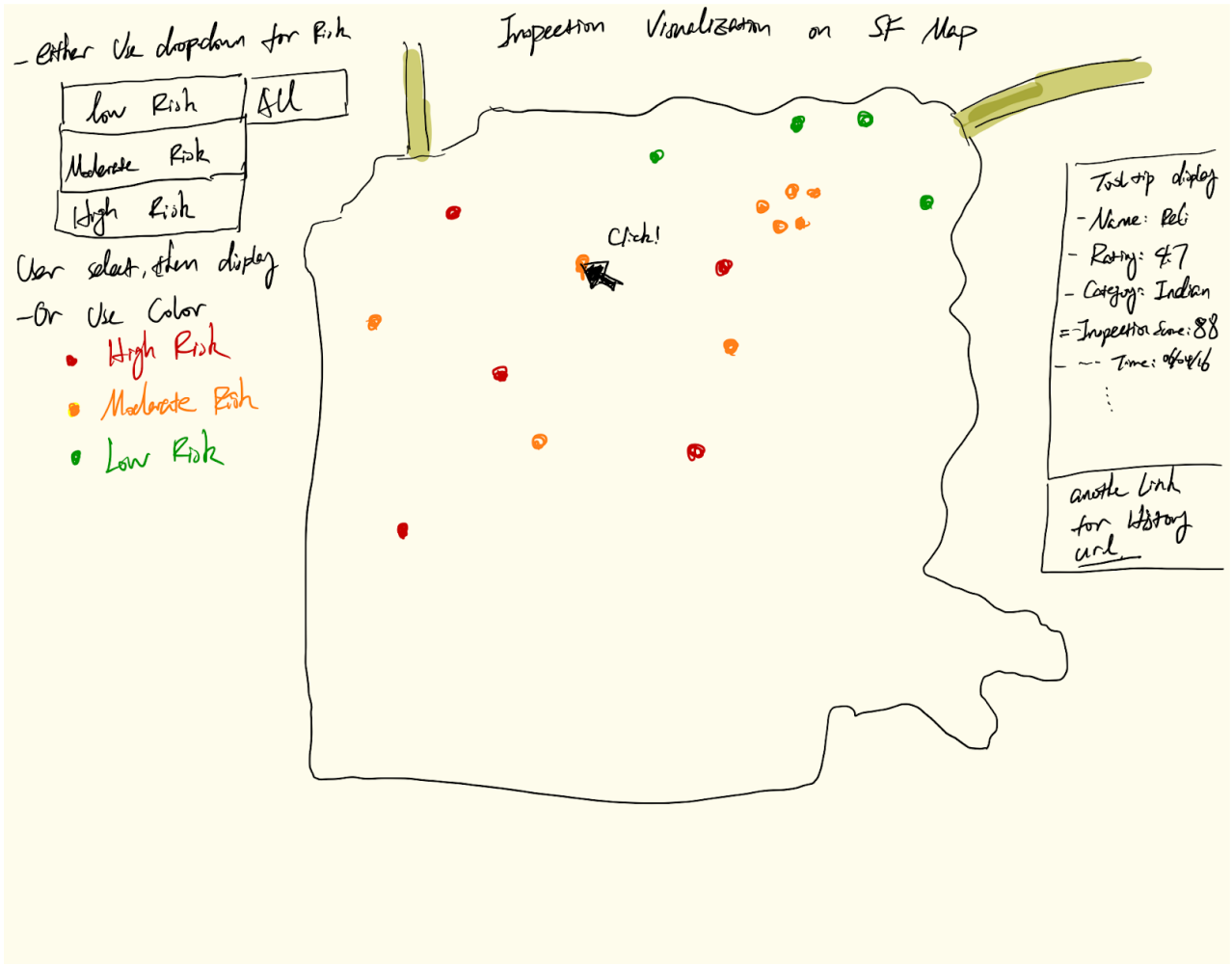

Fig. 1 - Map with individual markers for each restaurant, and filters

Answers the following questions:

- What Chinese restaurants have the highest rating on Yelp and are the cleanest?
- What are the restaurants with High Risk, Moderate Risk, Low Risk?

Interaction:

- If using Google Maps or similar: pan/zoom
- Specifying filters makes markers appear/disappear
- Clicking on a marker brings up details about a restaurant

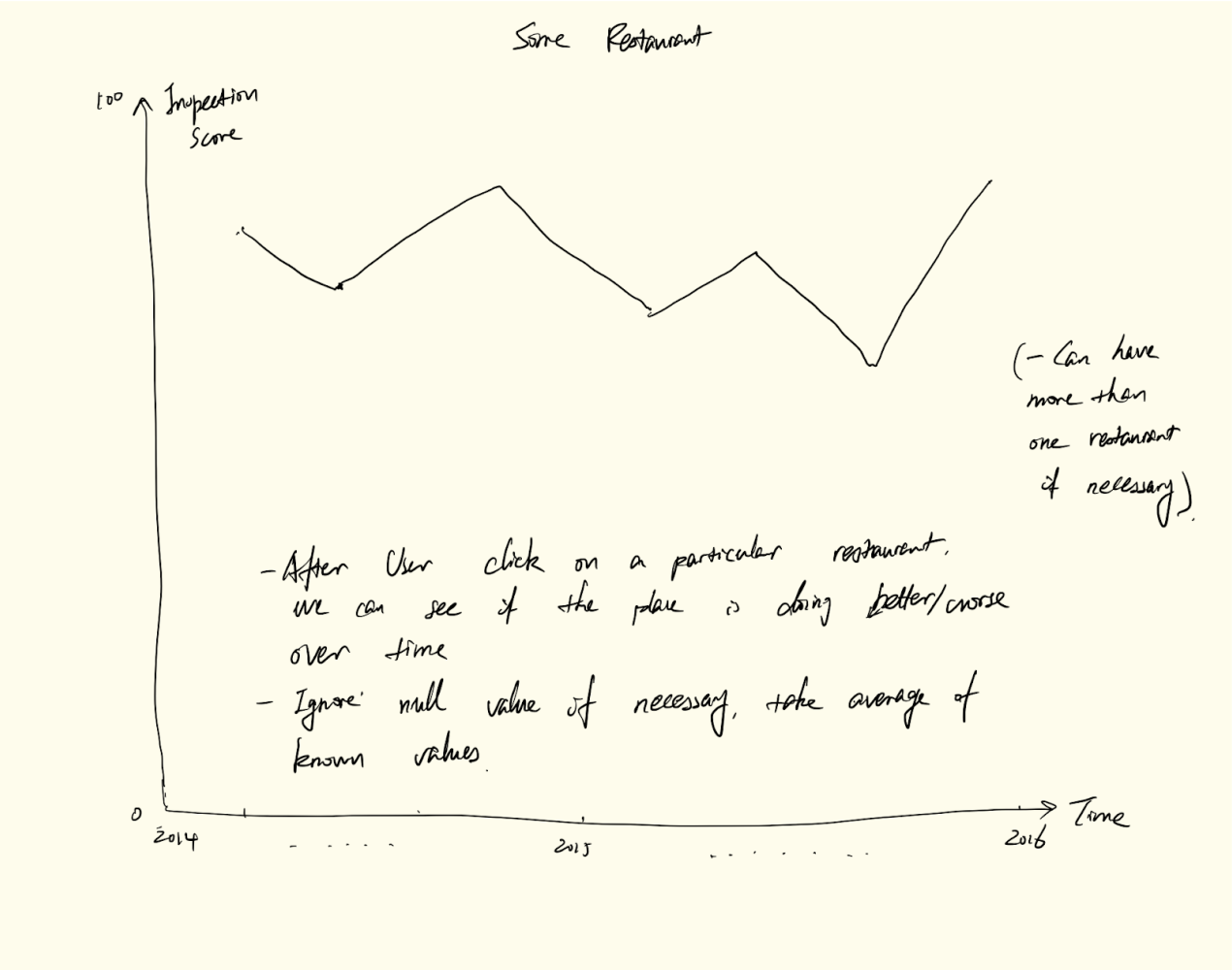

Fig. 2 - Change of inspection score over time for a particular restaurant

Answers the following question:

- What is the change of the health score for a particular restaurant over time?

Interaction:

- Hovering over data points shows values for those points

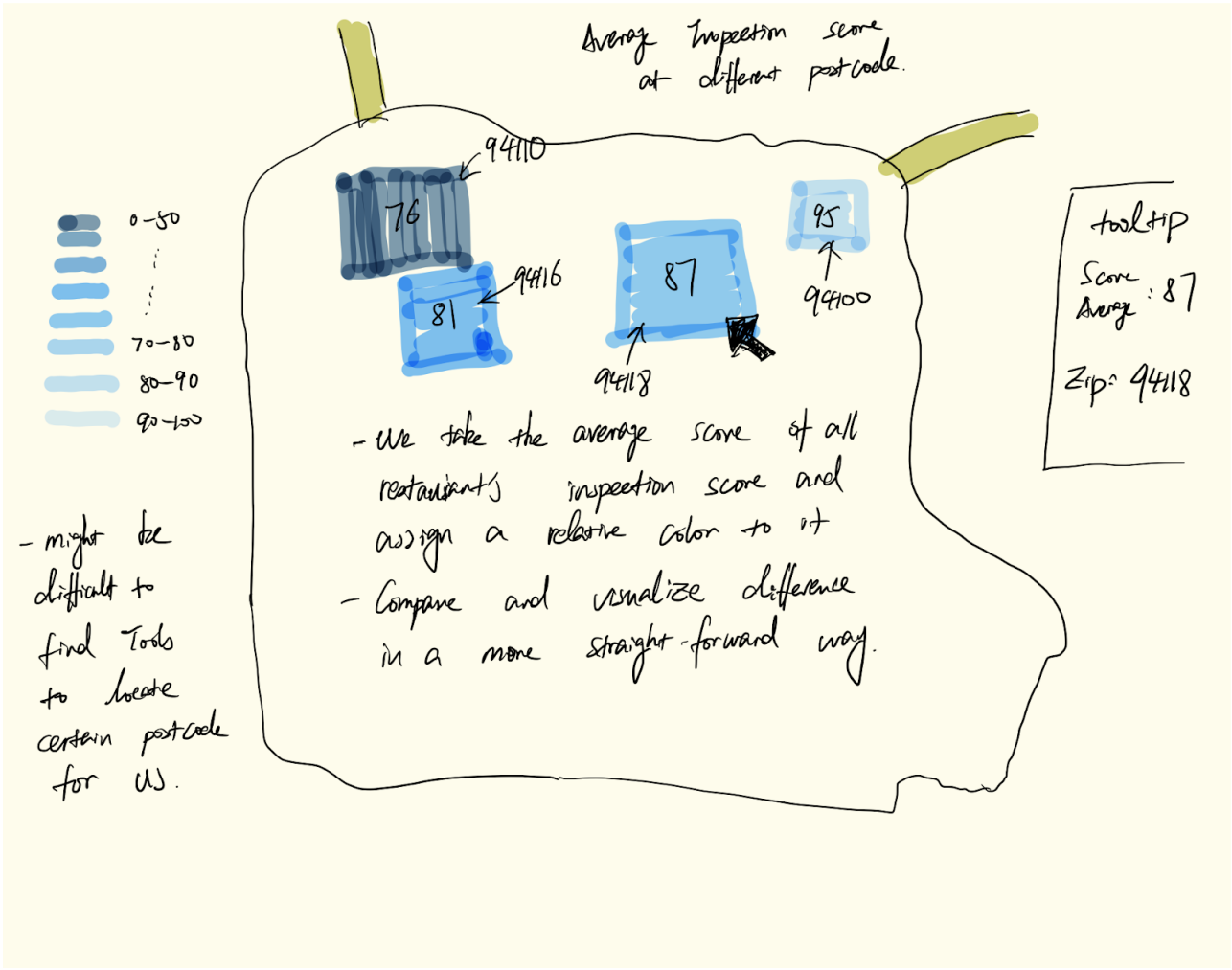

Fig. 3 - Map showing the average rating for each neighborhood

Answers the following question:

- What is the average for each ZIP code (941xx) / neighborhood in San Francisco?

Interaction:

- probably none, but could have some

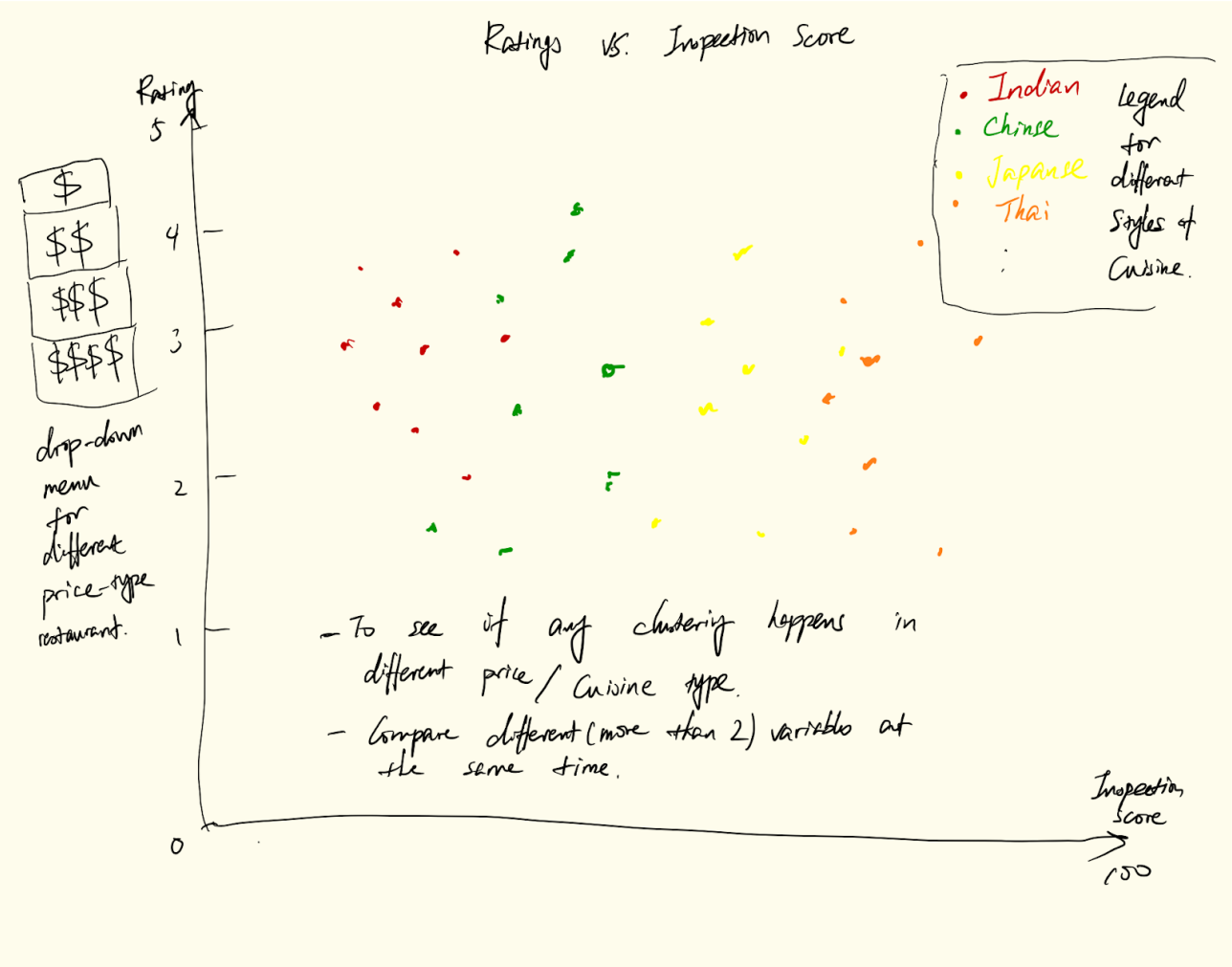

Fig. 4 - Scatterplot comparing ratings and inspection score

Answers the following questions:

- What Chinese restaurant has the highest rating on Yelp and is the cleanest?
- Are Indian restaurants cleaner than their Chinese counterpart?
- Is there a correlation between: health score, cuisine, and rating?

Interaction:

- Hovering over data points shows values for those points
- Optionally, restaurant categories could be hidden using checkboxes or another user interface component achieving the same effect.

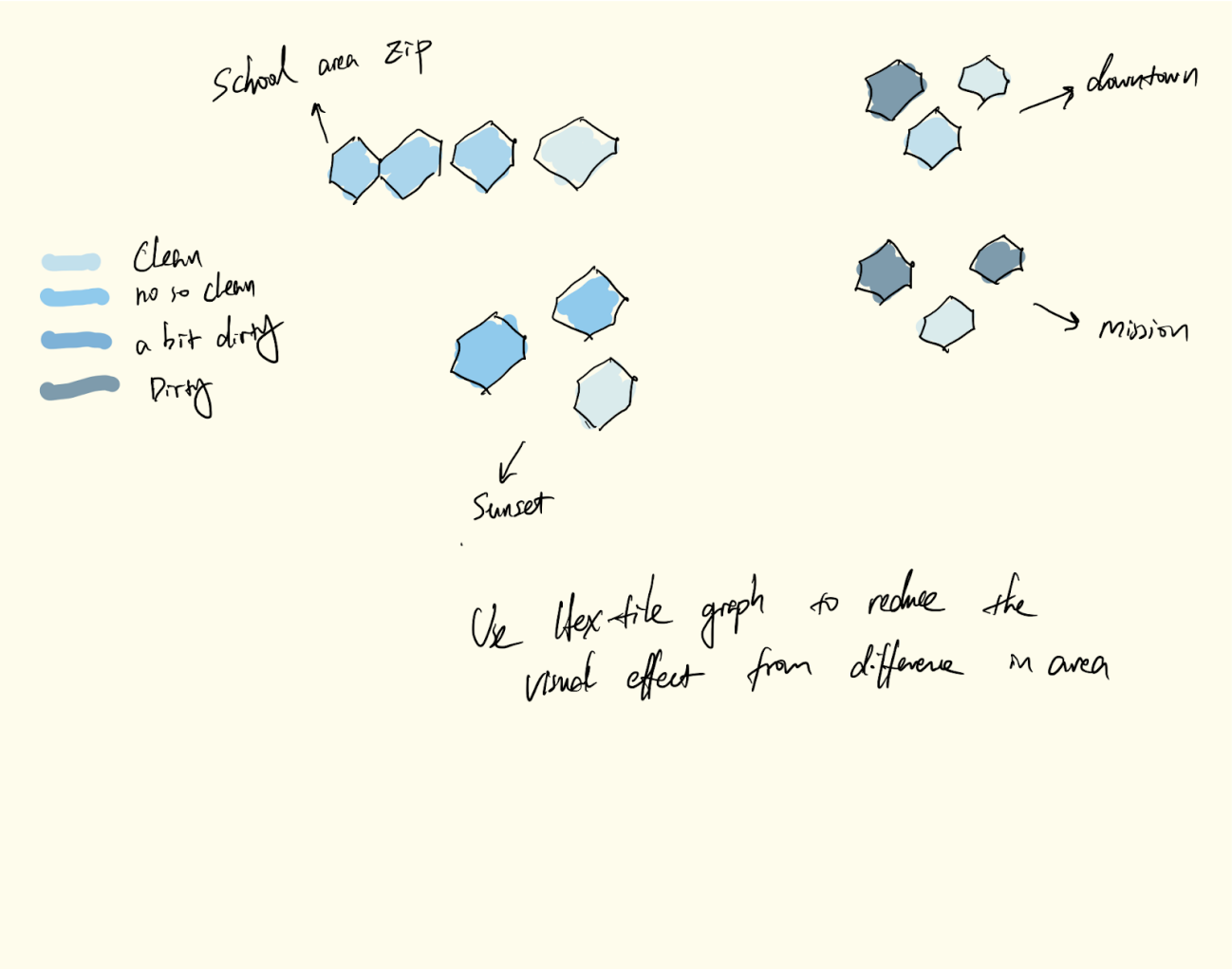

Fig. 5 - Hex-tile plot showing the correlation between zip code and inspection score

Answers similar questions as Fig. 1.

Interaction:

- Due to limited screen estate, hovering over tiles will most likely bring up the complete name of the neighborhood, number of restaurants in the area, etc.

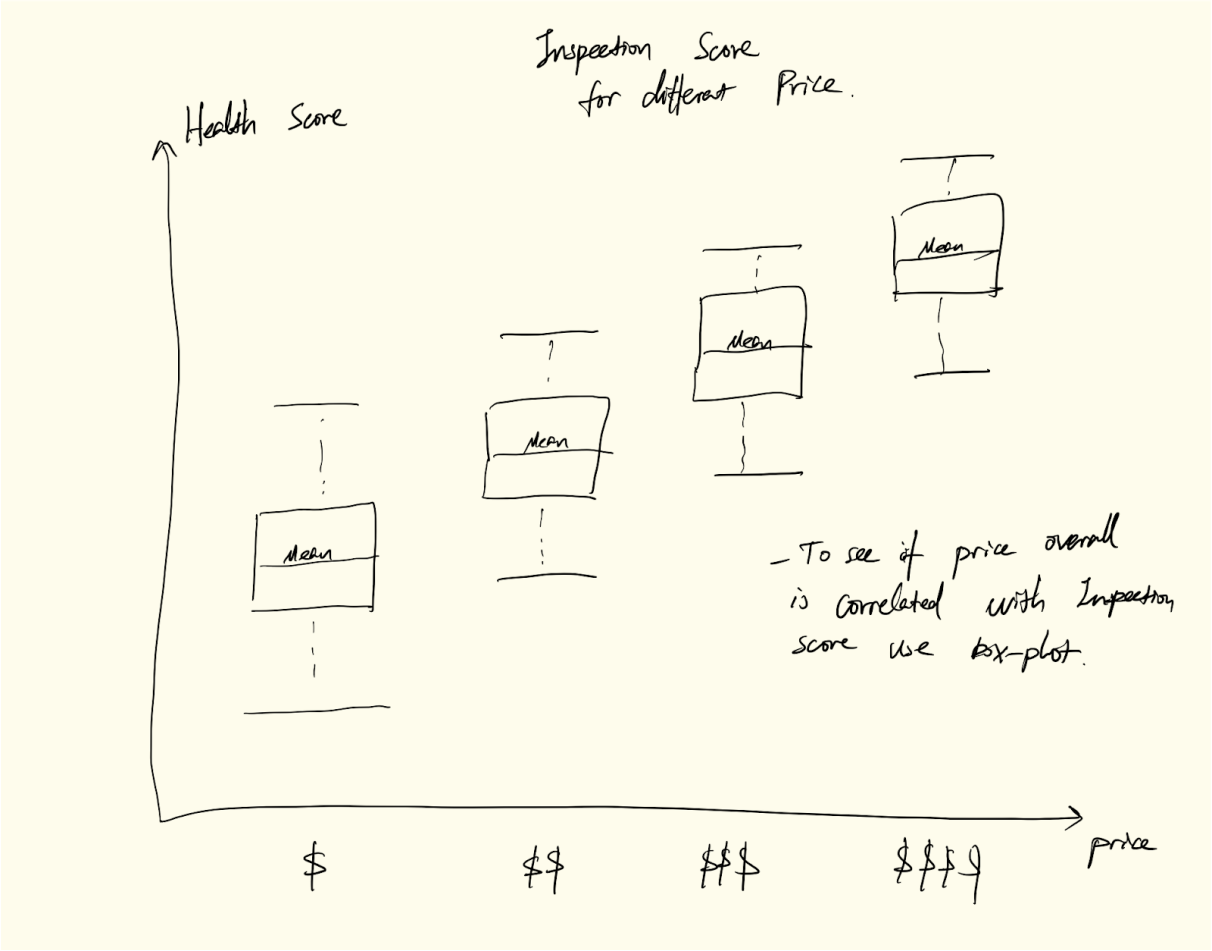

Fig. 6 - Box Plot showing the correlation between price and ratings

Answers the following question:

- Is there a correlation between: price and Yelp rating?

Interaction:

- Most likely, no interaction needed. All information should be on the screen.

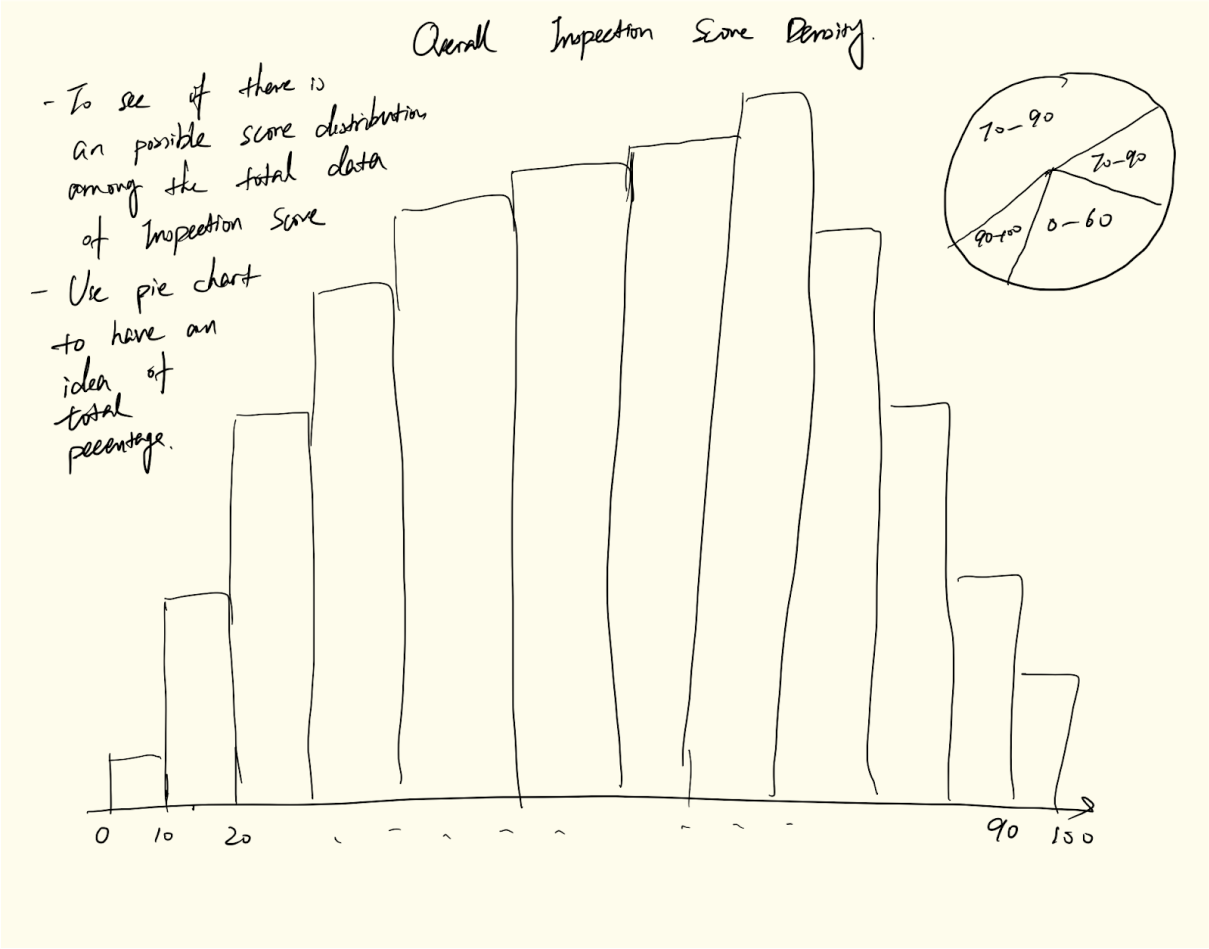

Fig. 7 - Distribution of inspection scores

Answers the following question:

- What is the distribution of health scores?

Interaction:

- Most likely none, unless we decide not to show the values for each bar immediately, in which case hovering over bars will bring up the values.

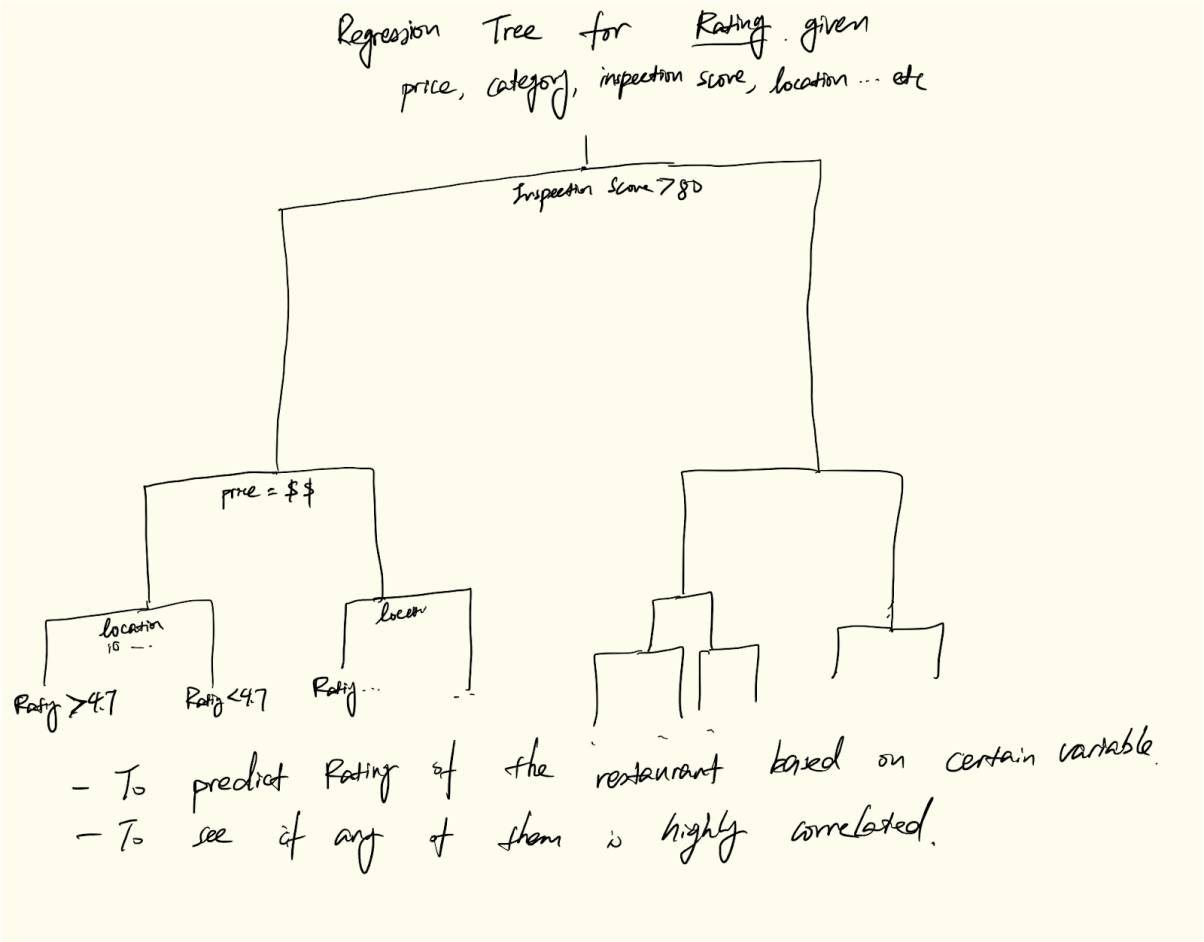

Fig. 8 - Regression Tree, to try predicting the health score based on select variables

Answers the following question:

- Is there a correlation between: health score, price, Yelp, rating, cuisine, ZIP code, etc. ?

Interaction:

- None

Must-Have Features

- Map with filters (might not use D3). i.e. when the user selects a value in one of the drop down menus or checks a checkbox, some of the markers appear or disappear as a result of filtering. Or we could design color into the visualization such that different category will has different color associated with it.
- Tooltip at the side. i.e. Some points may have multiple pieces of information associated with it. Instead of having a tooltip by the side of mouse, we might leave some space by the side of the visualization to display information of this map point.
- Line chart with health score vs time for one restaurant. After user click on one data point, we may be able to show the inspection score with respect to time if there are enough data points associated

- Trivariate scatter plot with Yelp rating, health score, and cuisine. i.e. As explained in the graph, we will have different color for different cuisine type.
- Histogram showing the distribution of health scores. This is to explore the overall distribution type associated with the inspection data points.

Optional Features

- Regression tree. To explore if any variable is most associated with predicting the rating of a particular restaurant. i.e. given the inspection score, location, cuisine type, price, we see if it is possible to predict the rating associated with it.
- Health score by neighborhood/ZIP code: there would be a map showing neighborhoods or ZIP codes of San Francisco with the average health score for that area. The neighborhoods/ZIP code regions would be shaded accordingly, e.g. in different shades of red, yellow and green depending on the health score (green = 100, red = 0)
- Box Plot with the price on the X axis and the health score on the Y axis, whose goal is to determine if there is a relationship between health score and price.
- Users can change the axes of the scatterplot to his likings, using different columns than originally intended by the authors of the visualization.
- Full text search filter on the map (Fig. 1)

Project Schedule

Date	Objectives
4/6	**Milestone: Proposal**
4/11	**Milestone: Revised Proposal, Project Website** Revise proposal, create project website, create process book Process the health score dataset to remove irrelevant rows and columns. Query the Yelp API to obtain review, cuisine and price information for each restaurant found in the health score dataset. Transform the data to a CSV or JSON file readable by D3, and write the appropriate parsing functions (a.k.a. rowConverter in D3.js v4)

	Explore the original and fused data using Tableau or other tools
4/16	**Milestone: Alpha Release** Have the histogram working but without interaction Be able to display markers representing restaurants on a map, without pop-ups Some of the filters implemented Line chart of health score over time, without interactivity and assuming all restaurants have at least one data point No colors Scales might be a little rough, too many fraction digits, etc.
4/25	**Milestone: Beta Release** Interactivity should be working by then, such as popovers or detailed page for each restaurant on the map Most filters implemented Color and finer presentation details taken care of, all relevant labels present, meaningful scales Time permitting: optional features
5/7 or 5/9	**Milestone: Project Presentation** Time permitting: optional features
5/16	**Milestone: Presentation slides, Code, Data, Process Book, README**