CS573 Data Visualization Final Project Proposal

Project title: Visualizing Yelp Dataset
Project repo: https://github.com/stels07/DataVisFinal
Team:

Yihao Zhou, yzhou2@wpi.edu, iihaw
Hongzhang Cheng hcheng3@wpi.edu, hcheng3
Shi Wang, swang11@wpi.edu, stels07

Background and motivation

Our team members have been yelp users for a long time and found the platform very useful. The app offers list view and map view of the businesses around, and users can apply different filters according to their preference. The mobile app seems to be more useful, but there is limited screen for displaying data. We thought it'd be interesting to explore how data could be better visualized on a bigger screen. We'd like to help user find what they are looking for more quickly, and at the same time give more information about rating and an overall picture of the businesses around. Based on our own experiences, we concluded that the key factors for decision making are: average rating of the business, category, location, price, number of reviews. Number of reviews is a strong indicator to how reliable the average rating is. We design to encode all key factors, and hope our design can provide a better user experience.

Project objectives

- Include average rating, category, location, price, busy hour for a business in our design
- Show details on rating
- Provide more general picture
- Easy filtering

Data

We use the data from yelp's data challenge: https://www.yelp.com/dataset_challenge

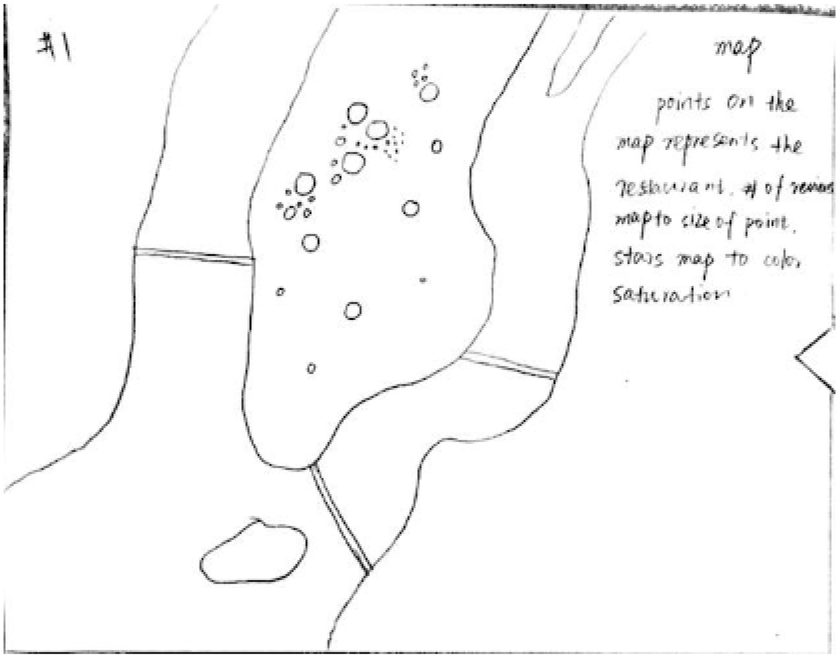
Data processing

The data we get from Yelp includes multiple large json files. It contains data for selected cities. The dataset is pretty clean. We just need to filter out data for a certain city (Montreal for instance). After using R for some preliminary assessment, we found out there were 4371 businesses in Montreal. One of the concerns is to show all the businesses on map. We plan to only use restaurant data if 4371 is too much, or possibly businesses are very close to each other in dense commercial areas.

Business, review and user data are in different json files. So the next step is to join the data according to business ID and user ID. We'll get map data from
http://geojson.io/#map=12/45.5043/-73.6264
At current stage, data will be processed using R. We'll also consider using Python.

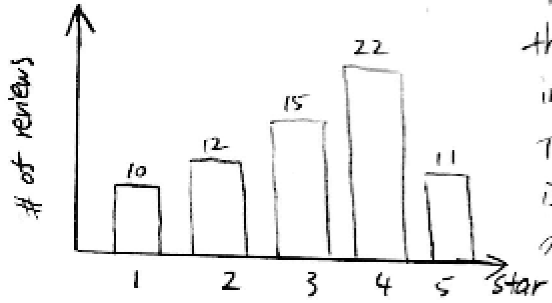
Visualization design

1. At first we represent all the restaurants on the map (see the graph #1), mapping the number of the reviews for each restaurant to the size of the point and map the average star of the restaurant to the saturation of the color. From this encoding readers could rapidly perceive the location information about different restaurants, and restaurants with more reliable rating in the map,(number of reviews is a strong indicator to how reliable the average rating is)which is the point with the bigger size and high saturation. This design gives readers a general information about the topic of our data visualization project.


#1
map
points on the map represents the restaurant, # of reviews map to size of point, stars map to color saturation

#1 Map Area

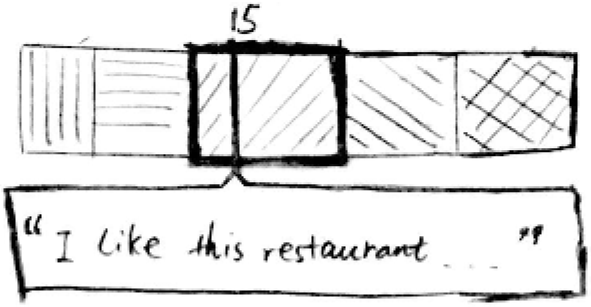
2. When the user mouse over the restaurant on the map, the charts about this restaurant will show up on the right side of the map. One chart is showing the number of reviews that made for each star. The first version of design is simply shows a bar chart as seen in graph #2.1. The x axis is the star and the y axis is the number of reviews.



This design simply shows the number of reviews in each star.
The average star is the sum of star multiply by the number of reviews and divided by the totoal number of reviews.

#2.1 Bar chart for number of reviews of each star

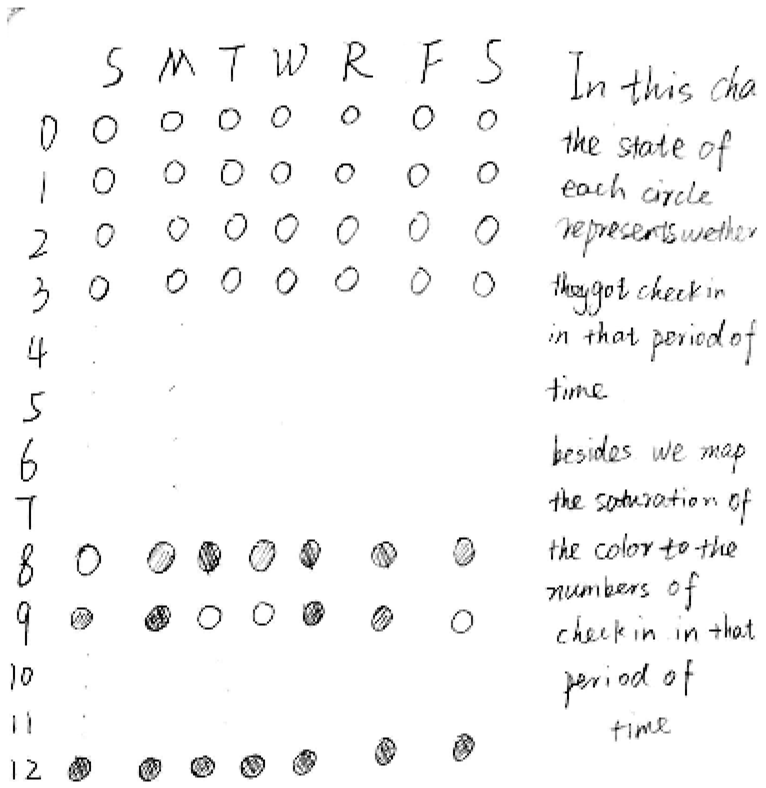
The second version of the design to show a single horizontal bar representing the total number of reviews for the restaurant. We then map the star to five colors on the bar. When mouse over the bar, the area where mouse is located will be highlighted and a number of reviews for this star will display on the top of the bar. When mouse over the bar, a vertical line will display on the bar and follow the mouse movement. A text box will show up below the vertical line with the actual review sentences.



colors map to star.
when mouse over the bar, shows a red line and highlight the color region and show comments.

#2.2 Single bar chart showing number of reviews

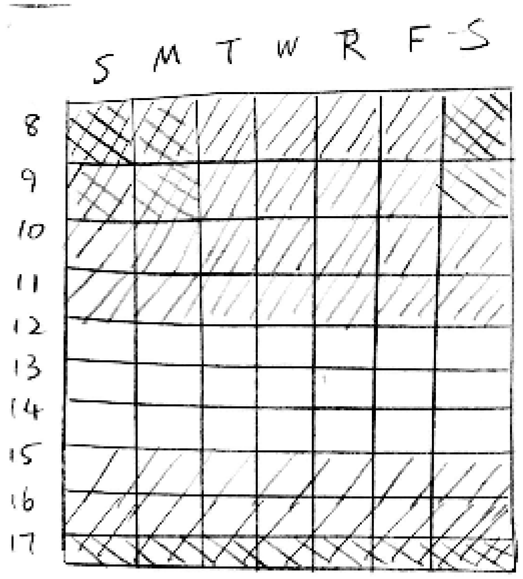
3. We also designed the heatmap that shows the number of check-in for each hour in a day and seven days a week. One version of the heatmap shows in graph #3.1. The saturation of the color in each circle represents how many check-ins in that period of time.



In this cha
the state of
each circle
represents wether
they got check in
in that period of
time
besides we map
the saturation of
the color to the
numbers of
check in in that
period of
time

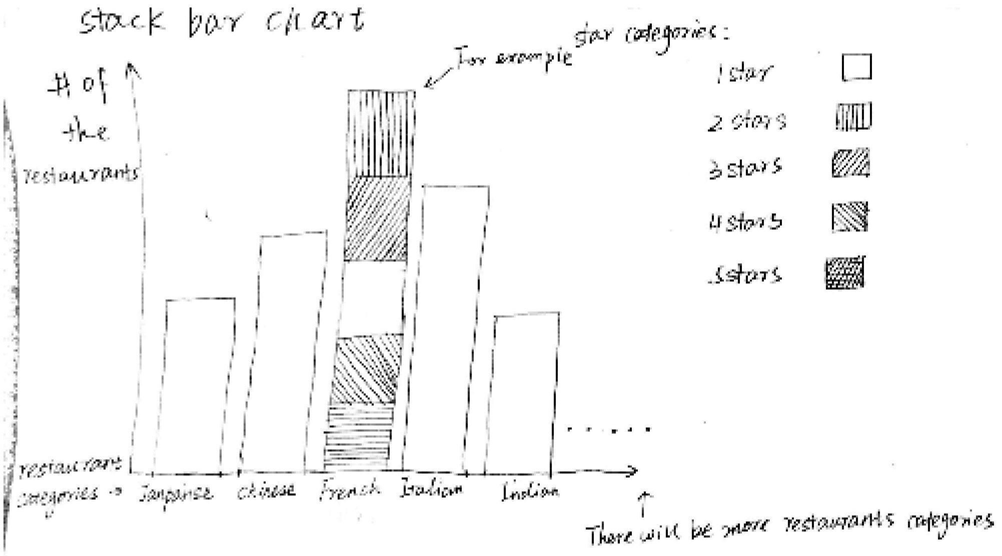
#3.1 Heatmap showing numbers of check-in

Another version (see in graph #3.2 is to use rectangle to represent the number of check-in in one hour period of time and the overall graph is a square with blocks of different color saturation. Using heatmap allows to visually show at what time in a day and in what day in a week this restaurant is busy.



#3.2 Rectangle Heatmap showing number of check-in

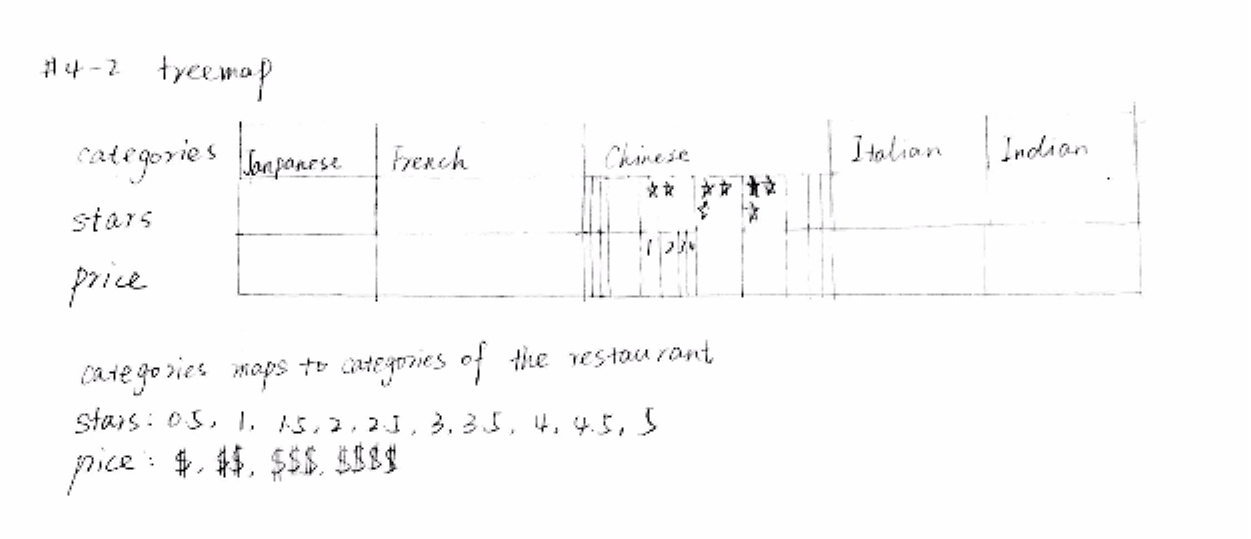
4. This is a stacked barchart (see graph #4.1) to show information about in different restaurant categories the distribution of the average stars. In this chart, we grouped the restaurants in several categories,such as japanese, chinese, italian, french, indians and so on, then we map the average star to the color to show the star distribution of different restaurant categories.This chart is also an interactive chart with the map(see in graph#1), when we click somewhere inside a bar,the related restaurants in that categories and with that average star will highlight in the map.Besides, it also interactive with the chart # 5.1, This design helps reader more easily find their desired restaurant. But eventually, we give up this design because we want to integrate other information in our design: the suggest price of the restaurant, so we chose the tree map (see graph #4.2 or #4.3)instead.



#4.1 Stack bar chart to show number of restaurants in different categories

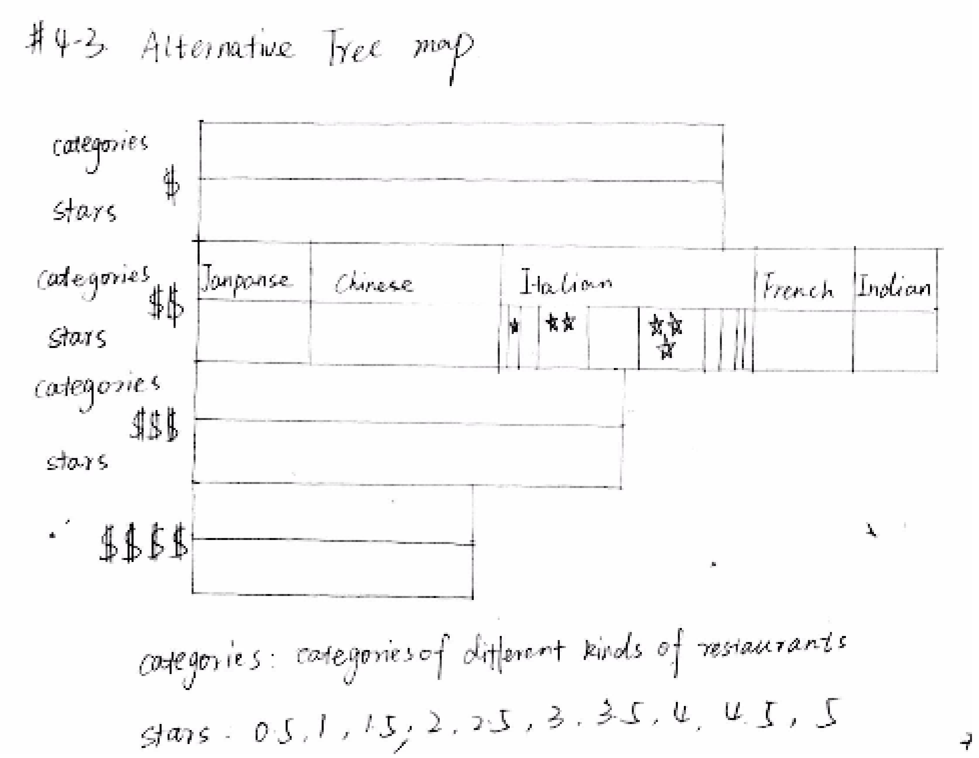
5.Below (graph #4.2) is our treemap design, In this design, we encoded the categories of the restaurants in the first line, average stars in the second line and its suggested price in the third line.This design helps the reader get the information about the distribution of the price in a certain star category which in a certain restaurant category.This tree map also interactive with the map (graph#1), when we mouse over one point in the map, simultaneously the related category, star, price will highlight in the this treemap, and when we click somewhere in the treemap, the related restaurants will highlight in the map.

This map also interactive with the graph #5.1 and graph #5.2, when we click somewhere in the treemap, graph #5.2 will show us 10 restaurants with largest numbers of reviews in that category, and the way it interactive with graph #5.1 is when we click one point in the map,not only the treemap will show its related category information, graph #5.1 will also show its number of reviews distribution. This design makes the way reader choose restaurant easier, all kind of information is categorized, and from the distribution, reader could know more accurate information about the quality of this restaurant. For example, when user filter the chinese food, 4 star, two price sign, all the restaurants qualified for those filter will appear in the map, then user could click the one with largest number of reviews in graph #5.2, that restaurant will highlight in the map. When the reader mouse over one point in the map, the related information for that restaurant will highlight in the treemap and also its distribution in the graph #5.1 will appear.


#4-2 treemap
categories
stars
price
Japanese
French
Chinese
Italian
Indian
categories maps to categories of the restaurant
stars: 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5
price: $, $$, $$$, $$$$

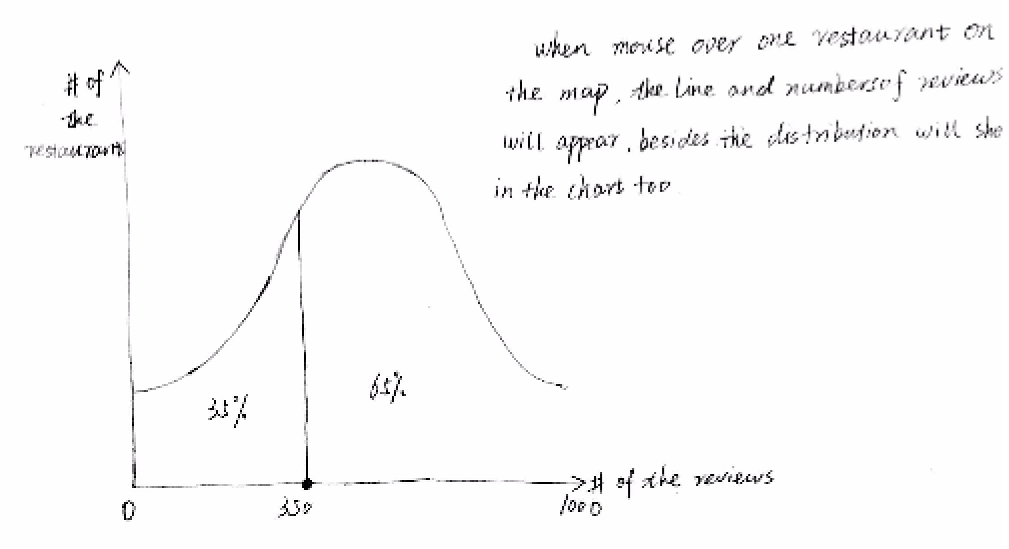
#4.2 Treemap to filter restaurants

6.This (graph #4.3) is an alternative treemap we designed, the only difference between this one and last one is we take the suggested price out, so the there will only be 2 layer in the treemap,but there will be four tree maps. the reason we did this is because in before one （ graph#4.2）, there are three layers, and in total, the bottom layer will be the numbers of the restaurant categories times categories of stars times price categories, that will be a huge number and we worried about the resolution will be not enough and the each single element will be too tiny in the last layer for reader to choose. So in this case, the number of the bottom layer will significantly decrease, but this graph will become bigger and hard to integrated in final design, now we are still thinking about which one to use.


#4-3. Alternative Tree map
categories
$
stars
categories
$$
stars
Janponse
Chinese
Italian
French
Indian
★
★★
☆☆☆
categories
$$$
stars
$$$$
categories: categories of different kinds of resiaurants
stars: 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5

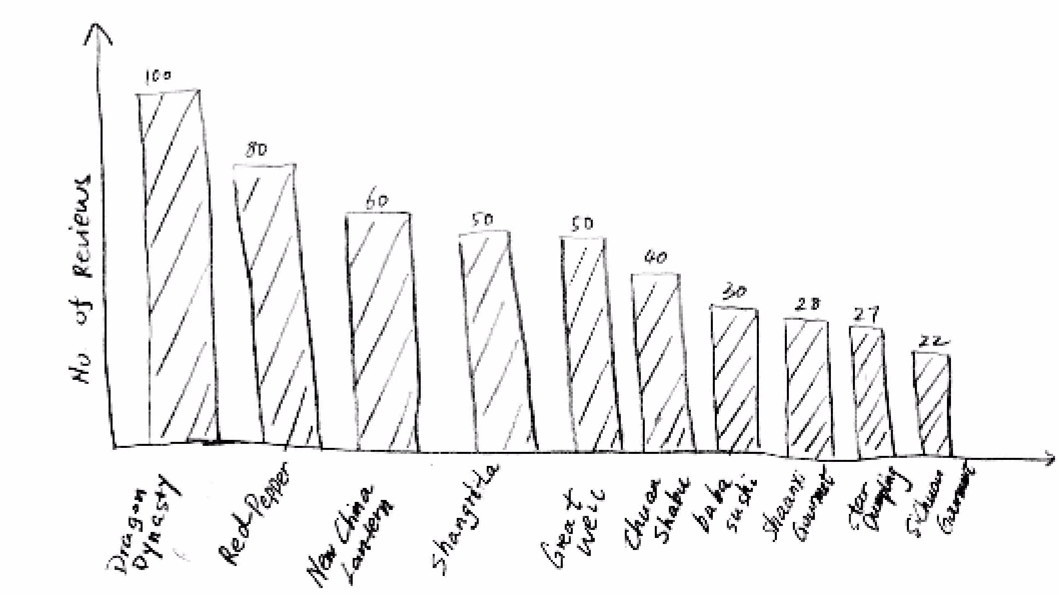
#4.3 Alternative Treemap

7. This graph shows the distribution of the number of restaurants over the number of reviews. We believe the graph will show as a normal distribution. Therefore we will curve-fitting

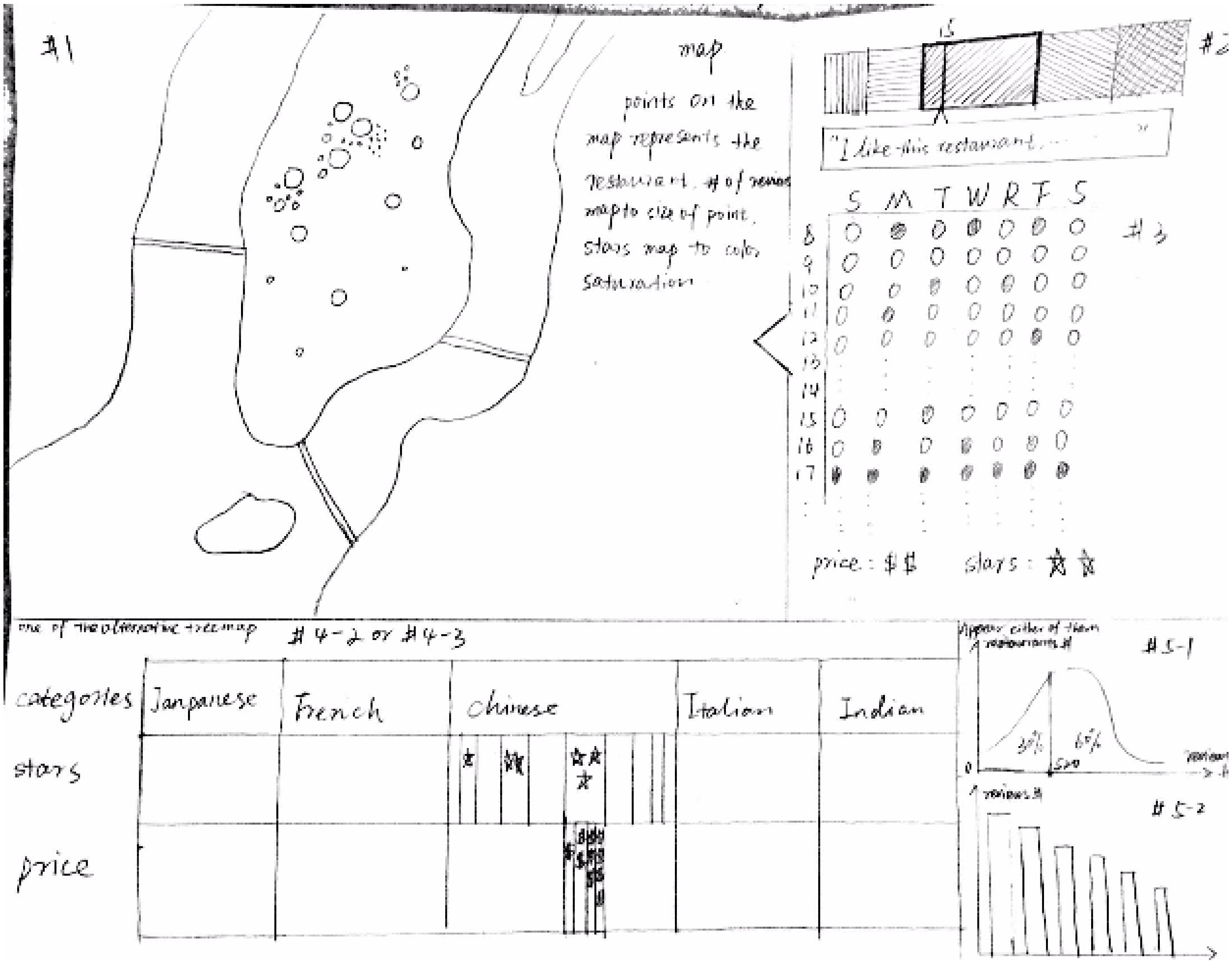


#5.1 Number of restaurants distribution over number of reviews

8.This graph shows the list of 10 restaurant with the largest number of reviews after filtering in tree map. It interacts with the treemap and map as described in the treemap chart.

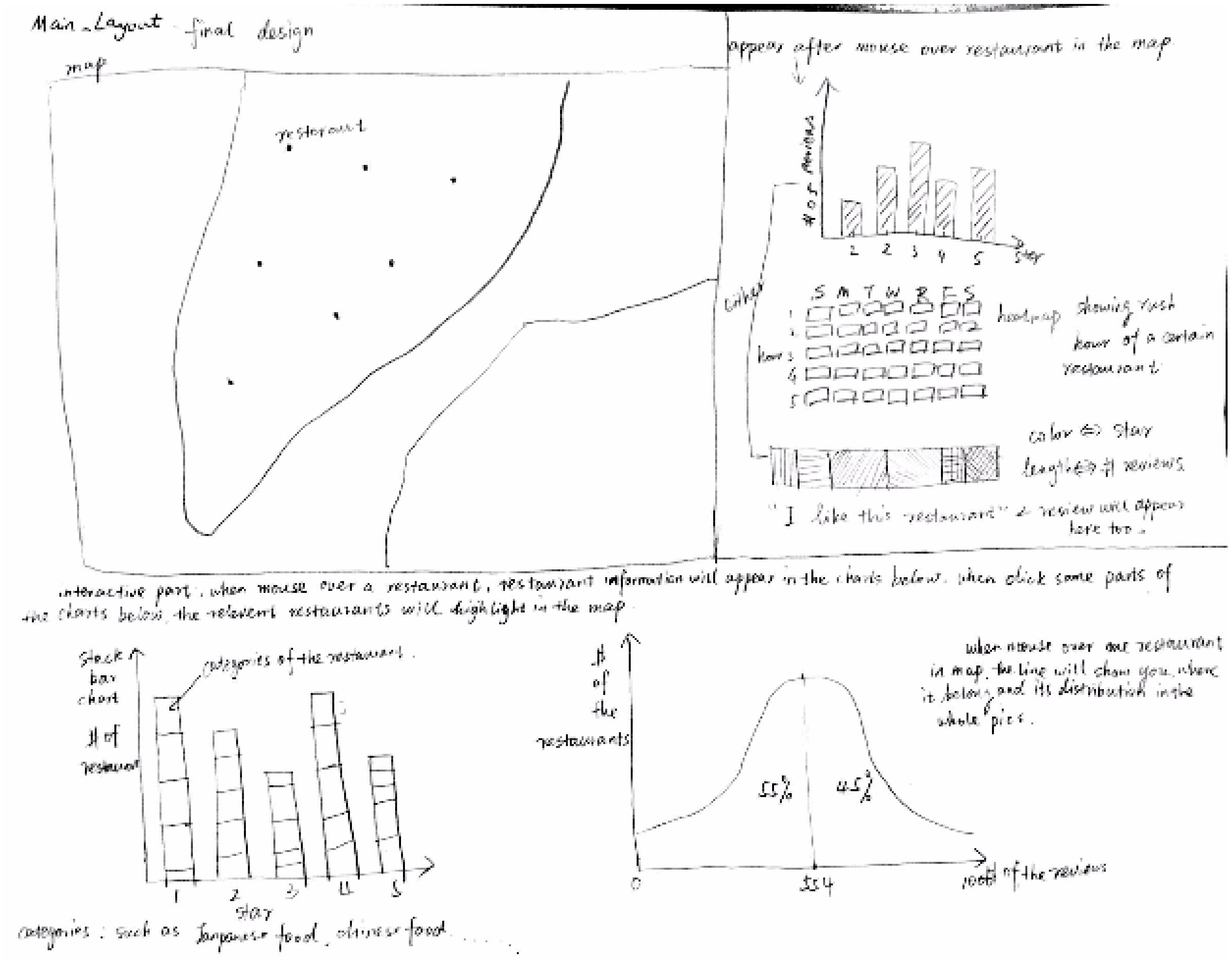


#5.2 List of first 10 business after filtering in Treemap



#6.1 Final Design Graph

The final design put together #1-5, and adds interactions among the graphs. Detailed interactions are listed in the must-have and optional features section. Above sketch shows the layout of the whole visualization. #1 is the main graph that have business location, average rating and popularity (reflected by number of reviews it get) encoded. So it occupies most area. #2 and #3 shows more specific information of one business if the user is interested. #4 and #5 provide more information of the whole picture. #4 shows the category and price information of that chosen business. It can also function as a filter to #1. #5-1 and #5-2 occupy the same space. #5-2 will show when #4 is used as a filter.


Main Layout final design
map
restaurant
appears after mouse over restaurant in the map
# of reviews
2
2
3
4
5
star
other
S M T W R F S
1
2
hours 3
4
5
heatmap showing rush hour of a certain restaurant
color ⇔ star
length ⇔ # reviews
"I like this restaurant" a review will appears here too.
interactive part, when mouse over a restaurant, restaurant information will appear in the charts below. when click some parts of the charts below, the relevent restaurants will highlight in the map
stack bar chart
# of restaurant
categories of the restaurant.
1
2
3
4
5
star
categories: such as Japanese food, chinese food......
# of the restaurants
55%
45%
0
554
1000 # of the reviews
when mouse over one restaurant in map, the line will show you where it belong and its distribution in the whole pics.

#6.2 Initial Final Design

This is the original final design we discarded, the reason we discarded this design is it didn't encode the price information in the design, we thought we could put a filter button on the map to let readers choose which price category they want to choose at first, but then we felt after the readers choose one price category, all the information will be filtered in that price category, then when the readers want to change a price category, they need to redo all the work. Besides, the information in stack bar chart is not as explicit as the tree map.

Must-have features

Our primary goal is to present the business data to Yelp users to help them find the business they want. So we must have a map and business locations (with average star mapped to color and bin of number of reviews mapped to circle size) on it. The must-have features regarding each of the graph are:

#1:

Mouse over an restaurant on the map will display graph #2 and #3 on the right side

Mouse over an restaurant on the map will highlight the corresponding rect in #4 according to its average star, category and price

Mouse over an restaurant on the map will display a vertical line in #5-1 to show how many reviews the business has and where it is at in the distribution

#4

Mouse over a rect on the treemap will highlight in #1 the businesses of that category

Mouse over a rect on the treemap will display #5-2, first 10 businesses with the most review

#5

Mouse over graph will display a vertical line, and highlight all the businesses on the right hand side of the line on #1

Optional features

To make our design more interactive, and offer better filtering to users, here are some nice to have features:

#2:

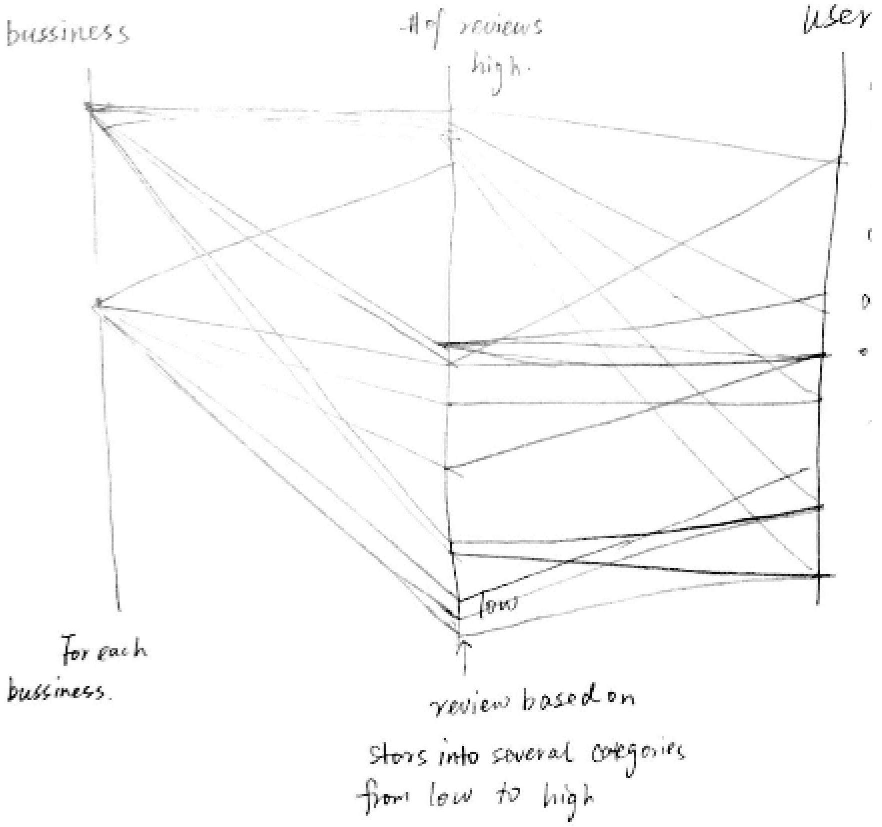
Mouse over the horizontal stacked bar will display below the chart 3 pieces of randomly selected review texts of that star category.

#4

Click on the rect on the treemap will do exactly the same as mouse over the rect, but the view stays. Multiple clicks on multiple rects will highlight all the selected businesses on #1, and display the first 10 businesses on #5-2. This enters the multiple clicks state. In this state, mouse over certain business on #1, the same as mouse over businesses on #5-2, will also display #2 and #3. Click anywhere else exit this state.

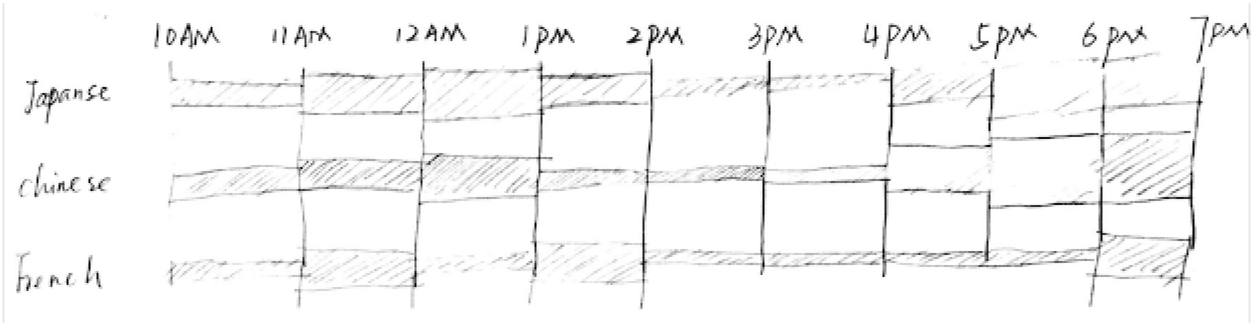
Other ideas

After deciding using the Yelp dataset, we had several discussions. There are two directions we can focus on: the business and the user. The above design mainly focusing on visualizing business information. If we were to focus on the users, we can show how active the users are, or the network among users. One idea is to see if user has a tendency to rate more often 1 star or 5 star, that is, if people tend to speak up more when they have extreme experience, either good or bad. And how this tendency, if there is any, relate to the activeness of the user.



#7.1 slope chart

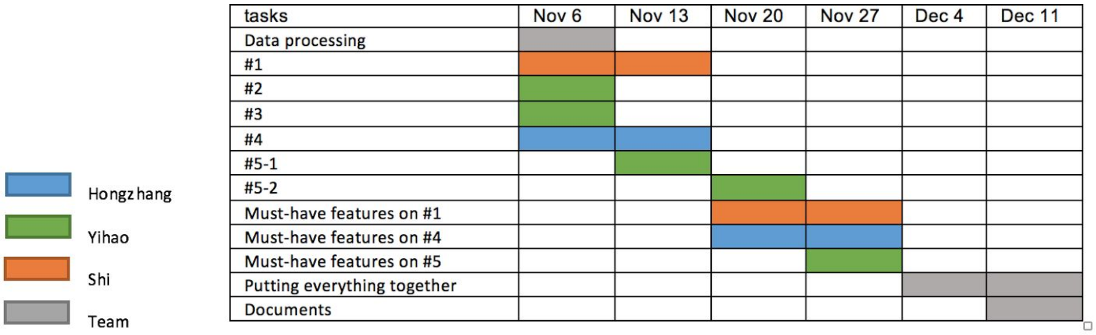
This is another original design which we discard. It connects the business to number of reviews and then connects the number of reviews to users. In this chart we could know numbers of stars on user give different restaurants in one category and the distribution of the reviews for each restaurant. The reason we discard the chart is because we believe there will be too much data to show in the chart, which makes the chart look too complicated.



#7.2 Timeline for restaurant categories

This visualization design tries to find the most busy time of each restaurant category. In this design, we want to find out whether there is a trend that when which category of restaurant will be most popular.For example, french restaurant might be busy in after 8pm due to most of the people will choose french restaurant for formal dating, chinese restaurant will be earlier, because people tend to think chinese restaurants are more casual here.

Project schedule (Timeline)



🟦 Hongzhang
🟩 Yihao
🟧 Shi
▨ Team

tasks	Nov 6	Nov 13	Nov 20	Nov 27	Dec 4	Dec 11
Data processing	▨					
#1	🟧	🟧				
#2	🟩					
#3	🟩					
#4	🟦	🟦				
#5-1		🟩				
#5-2			🟩			
Must-have features on #1			🟧	🟧		
Must-have features on #4			🟦	🟦		
Must-have features on #5				🟩		
Putting everything together					▨	▨
Documents						▨