FinalProject

I. Introduction

II. About the data source

The data of this project is scraped and converted to csv file in Python from Yelp API's business search endpoint. With the API Key and URL 'https://api.yelp.com/v3/businesses/search', it could be easy to make a request to Yelp API for scrapping and do not need to solve those annoying problems like cookies.

Because of the maximum limit of Yelp API, the data was scraped twice accroding to the "`category`". Also the paratemer of location is '`New York`'. All the data are originally scrapped except we add one item '`cuisine`' which is also the searching category in scrapping result. What's more, it's important to avoid duplication in the results and it could be easily operated by the `df.drop_duplicates` using `pandas` in Python.

For the purpose that the scrapping results would be more uniform distributed, this project choose five western cuisines and five eastern cuisines that people may be more interested in. In the first file "business_data_western.csv", there are five western cuisines: '`italian`', '`spanish`', '`greek`','`french`', '`mexican`'. In the second file "business_data_eastern.csv", there are five eastern cuisines: '`chinese`', '`korean`', '`japanese`', '`indian`', '`thai`'. There are total 8814 records.

Here is the link[https://github.com/zhangyue9966/Final_Project_STAT5293/blob/master/scrape_yelp/scrape_yelp_western.py] of the Python code for getting western restaurants information.

III. Data cleaning

The original data's dimension is 8814*18. Our steps for data cleaning is shown as below:
1. Choosed features that would be useful for further analysis: `coordinates`, `cuisine`, `cuisine_type`, `id`, `name`, `price`, `rating`, `review_count`.
2. For missing value, it is only necessary to delete the records containing missing value in the features that we interested. Thus, we deleted records with missing value in `price`, `coordinates`, `rating` and `review_count`. There are total such 351 records.
3. For error values, as there are only two records containing error(false dollar sign), we directly deleted them.
4. Last step is to format data. Firstly, '`price`' is factorred and its levels are resetted. Secondly, as we might need '`latitude`' and '`longitude`' in further analysis, we extracted them from '`coordinates`' and created new columns for them, then delete the original '`coordinates`' column.

Cuisine	Number
French	550
Greek	824
Italian	1000
Mexican	792
Spanish	916
Chinese	884
Indian	726
Japanese	535
Korean	705
Thai	523

```
##  [1] "alias"         "categories"    "coordinates"   "cuisine"
##  [5] "display_phone" "distance"      "id"            "image_url"
##  [9] "is_closed"     "location"      "name"          "phone"
```

```
## [13] "price"         "rating"        "review_count"  "transactions"
## [17] "url"           "cuisine_type"
```

```
## # A tibble: 10 x 2
##    cuisine  norows
##    <fct>     <int>
##  1 french      556
##  2 greek       836
##  3 italian    1003
##  4 mexican     819
##  5 spanish     930
##  6 chinese     909
##  7 indian      763
##  8 japanese    555
##  9 korean      747
## 10 thai        530
```

Besides the data we get directly from Yelp, we also want to explore the relationship of differnt crusine and the boroughs they belong to. Because Yelp does not prodive the details, we used the latitude and longitude information to join our dataset with the borough infromation. As the result we have the all_data with the following vairables.

```
## [1] "latitude"     "longitude"    "name"         "cuisine"
## [5] "price"        "rating"       "review_count" "neighborhood"
## [9] "borough"
```

IV Analysis of missing values

All the records are independent and most of the data are categorical. Moreover, the missing value is not extensive compared with the whole dataset. Thus, it is not meaningful to use methods like listwise, pairwise or regression to fill up missing valu. In this project, records contaning missing value are deleted directly.

While matching df with the neighbourhood, many unmatched records occured. For dataset all_data, there're 8478 rows and amony them 1010 rows contain missing values. Compared to the total amout of data, we decide to remove these missing data later while ploting.

```
## [1] 8478
```

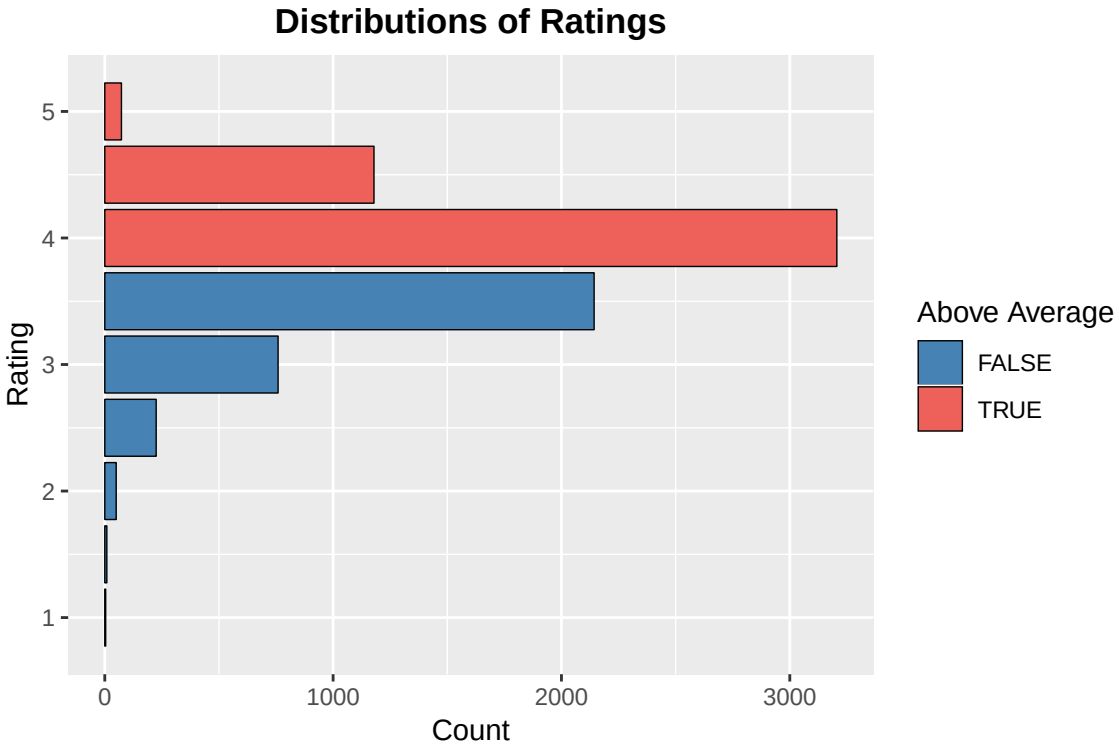
```
## [1] 1010
```

```
## [1] 1010
```

V Results

What can affect the rating?

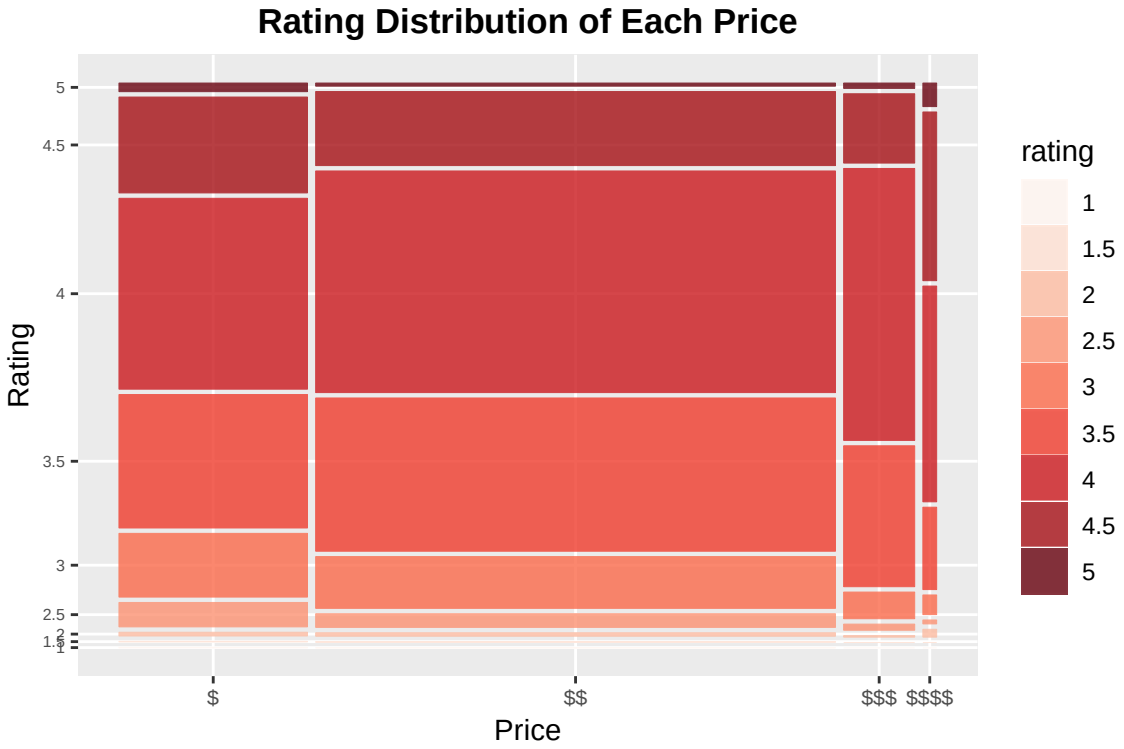
Firstly, let's get some insights by looking at the distribution of rating.



The mean of rating is 3.79. Thus, it would be reasonable to consider the restaurants with rating queals to or higher than 4 to be "good" restaurants. They are represented by the red bars in the bar chart. The blue bars represent "bad" restaurants. The distribution of ratings is right-skewed. People tends to give a neutral score, and a few people would give ratings lower than 3.

1. Rating & Price

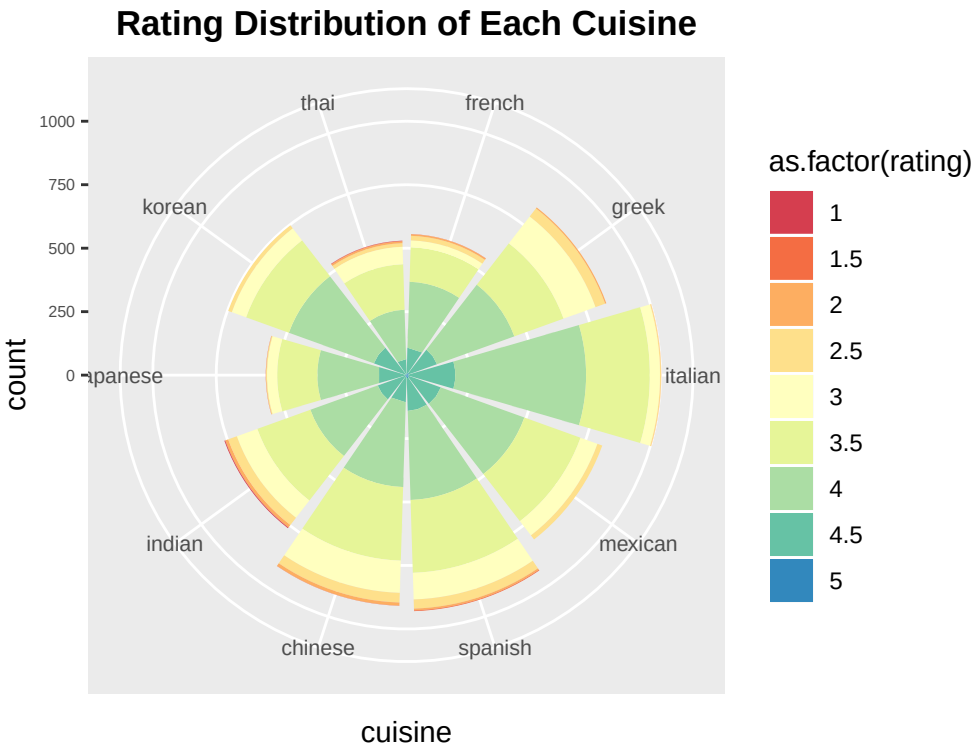
The second-level price takes the largest proportion. It also owns the most centralized rating. Overall, The rating present the postive relationship with price. Especially in the higheset-price restaurants, the rating distribution is obviously different with the other levels. There are almost 80% restaurants whose rating are 4 or higher than 4.
Howerver, outliers exist. In the cheapest restaurants, rating of 4.5 and 5 owns a high proportion. This is reasonable that we must have experienced surprise that some cheap food have the taste which are out of expection. People would give their highese commendation in such situation.



For further explory, it could be helpfule to add cuisine as a feature.

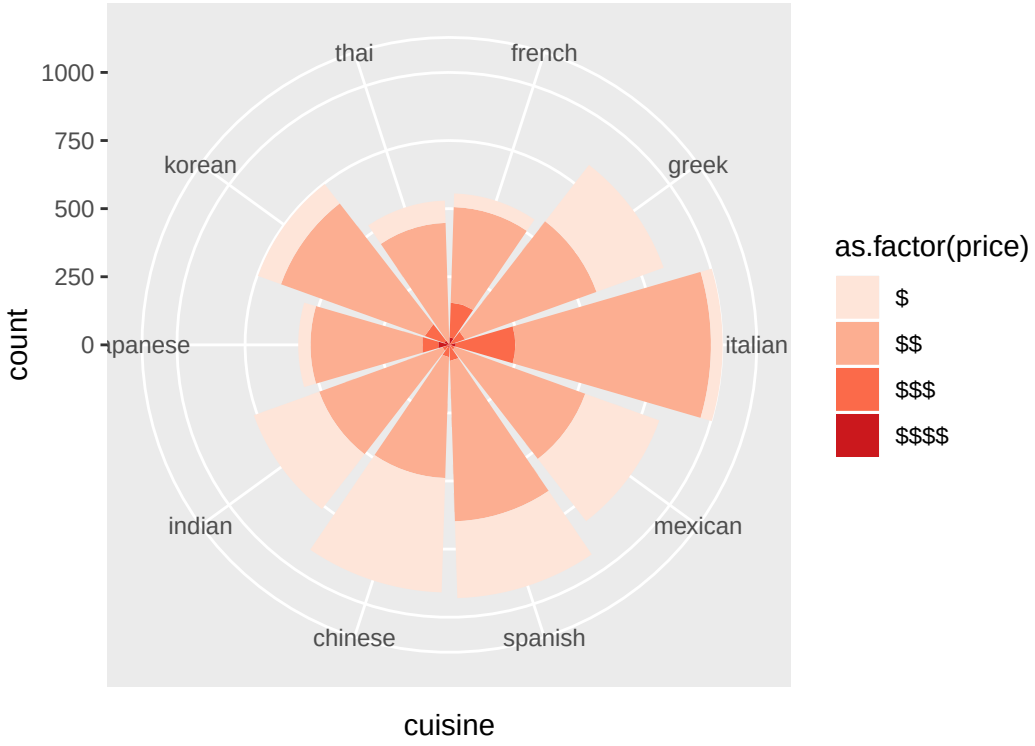
2. Cuisine & Rating

Beauser there are nine levels, it would be confusing to using sequential color, instead we used diverging color to show the pattern. What's more, it's better to compare when using `coord_polar` to convert parallel bars into pie chart. The color closer to blue, the higher the rating. The area represents the count.
It is very clear that italian and korean cusine has the relatively larger proportion of high rating. At the meantime, italian restaurants has the largest amount. Also, there's no distinct difference between western and eastern cuisines.



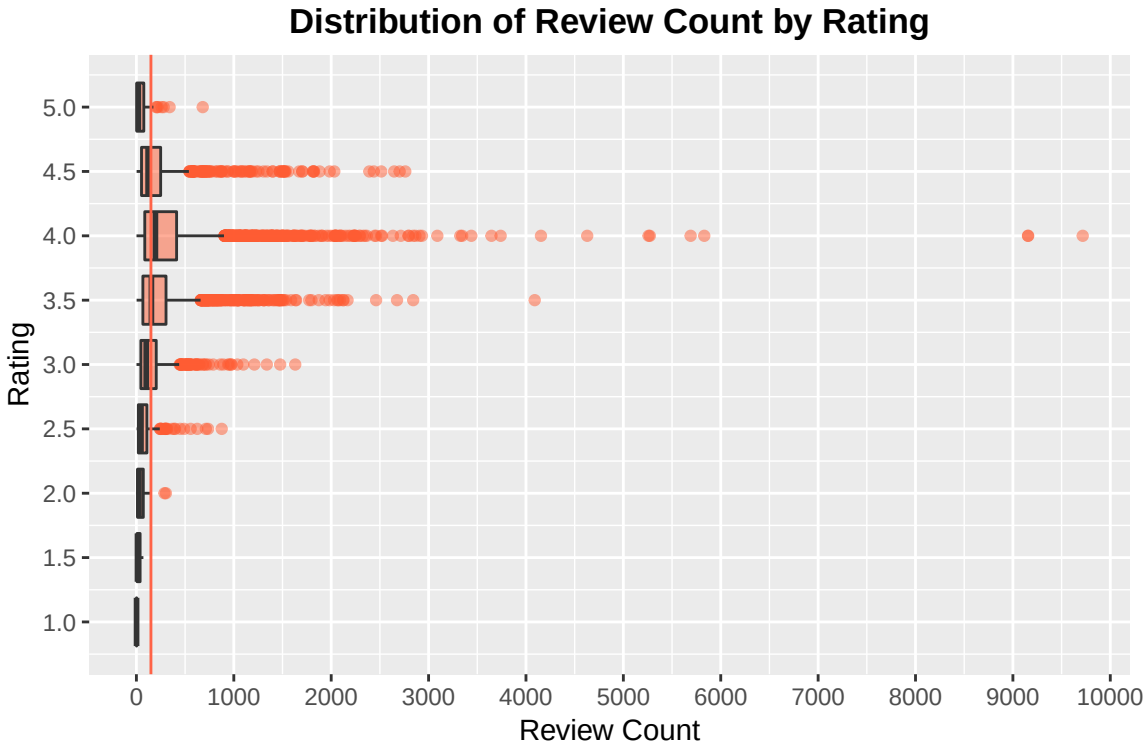
3. Cuisine & Price

Because of less levels, it would be better to use sequential palettes to show the distribution. Darker color represent the higher price. Appreantly, Chinese restaurants has the relatively lowest price while italian restaurants has the highest. While Chinese restaurants do not own the lowest rating, thus, there's no direct relationship between price and rating. This result is consistent with the conclution we drawed before when compared Rating & Price.



4. Rating & Review Count

The box plot clearly show that how much outliers could be of review count. The mean of review count is 270, the median of review count is 153, while the largest review count could be close to 6000. This exposured the high bias in this market. The most focus are on a few popular restaurants. Interestingly, people tends to give review for those low-rate restaurants than high-rate restaurants. People may cannot help expressing anger than happiness.



VI Interactive Component

While analyzing data, we found it hard to compare more than 3 elements. So we built R shiny app which contains 3 interavtive parts.

Is the distribution of restaurant in different cuisine related to borough?

The first interactive part is an interactive map which show the distribution of the location of restaurants given cuisine and rating value. We can explore the distribution change of these restaurants when the cuisine or rating changes. For example, for the most of restaurants in every cuisine, the higher rating is, the closer restaurants are to midtown. This shows that midtown in Manhattan is a great place to dine in if you prefer high rating.

Unsurprisingly, some cuisine like Chinese restaurants with ratings 4 are located near the China Town area in Mahanttan. However, Itanlian food does not have this trend. From the screenshot below we can see that it's more evenly distributed in Manhattan and do not have specific prefenerce in locations.

Will the relationship of price and borough change a lot given differnt cuisine and rating?

Interactive plot give us an opportunity to explore more than 3 elements at the same time. From common sense, we will consider price highly related to borough as the rent for restaurants in Manhanttan will definitely be higher than that in others boroughs. But is this really the case? And what about other boroughs? From the app we can see that although this result depends on crusine and vary from case to case, for most cuisines the price in Manhattan has more different price level. Especially for Chinese food.

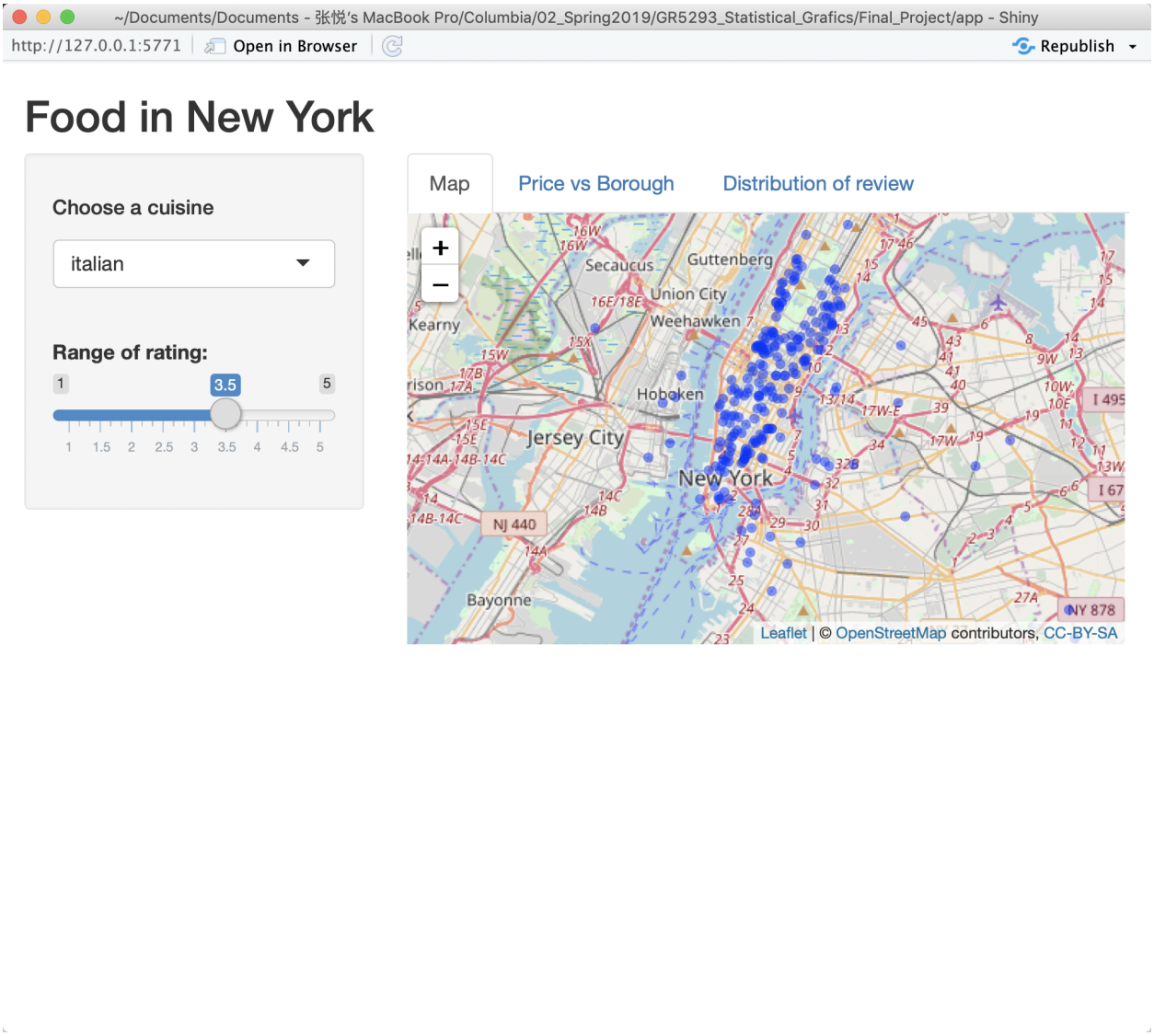

~/Documents/Documents - 张悦's MacBook Pro/Columbia/02_Spring2019/GR5293_Statistical_Grafics/Final_Project/app - Shiny
http://127.0.0.1:5771
Open in Browser
Republish
Food in New York
Choose a cuisine
italian
Range of rating:
1
3.5
5
1 1.5 2 2.5 3 3.5 4 4.5 5
Map
Price vs Borough
Distribution of review
Secaucus
Guttenberg
Union City
Weehawken
Kearny
Hoboken
Jersey City
New York
Bayonne
NJ 440
NY 878
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 1:

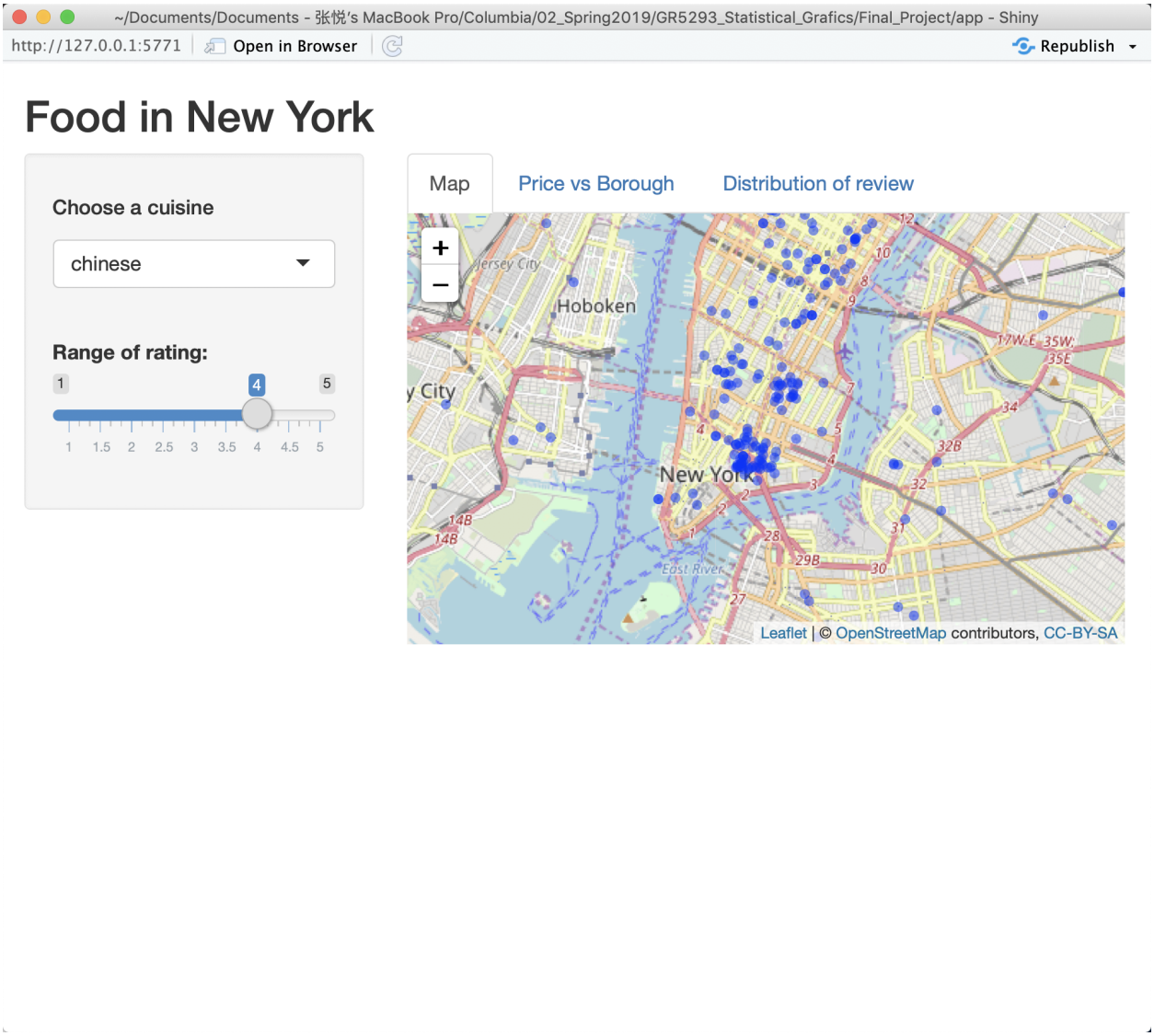

~/Documents/Documents - 张悦's MacBook Pro/Columbia/02_Spring2019/GR5293_Statistical_Grafics/Final_Project/app - Shiny
http://127.0.0.1:5771
Open in Browser
Republish
Food in New York
Choose a cuisine
chinese
Range of rating:
1
4
5
1 1.5 2 2.5 3 3.5 4 4.5 5
Map
Price vs Borough
Distribution of review
Hoboken
New York
East River
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 2:

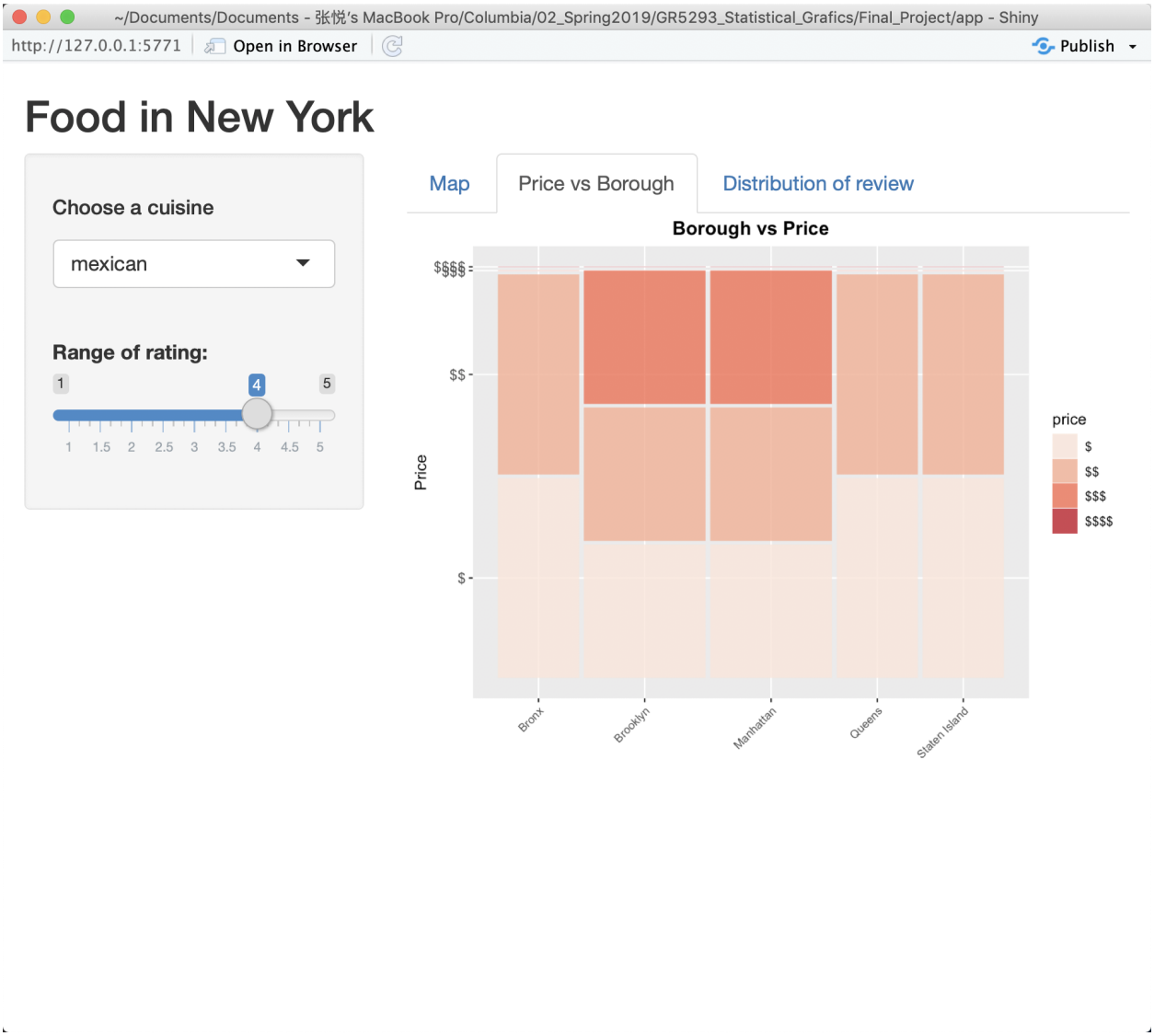

~/Documents/Documents - 张悦's MacBook Pro/Columbia/02_Spring2019/GR5293_Statistical_Grafics/Final_Project/app - Shiny
http://127.0.0.1:5771
Open in Browser
Publish
Food in New York
Choose a cuisine
mexican
Range of rating:
1
4
5
1 1.5 2 2.5 3 3.5 4 4.5 5
Map
Price vs Borough
Distribution of review
Borough vs Price
Price
$$$$
$$
$
Bronx
Brooklyn
Manhattan
Queens
Staten Island
price
$
$$
$$$
$$$$

Figure 3:

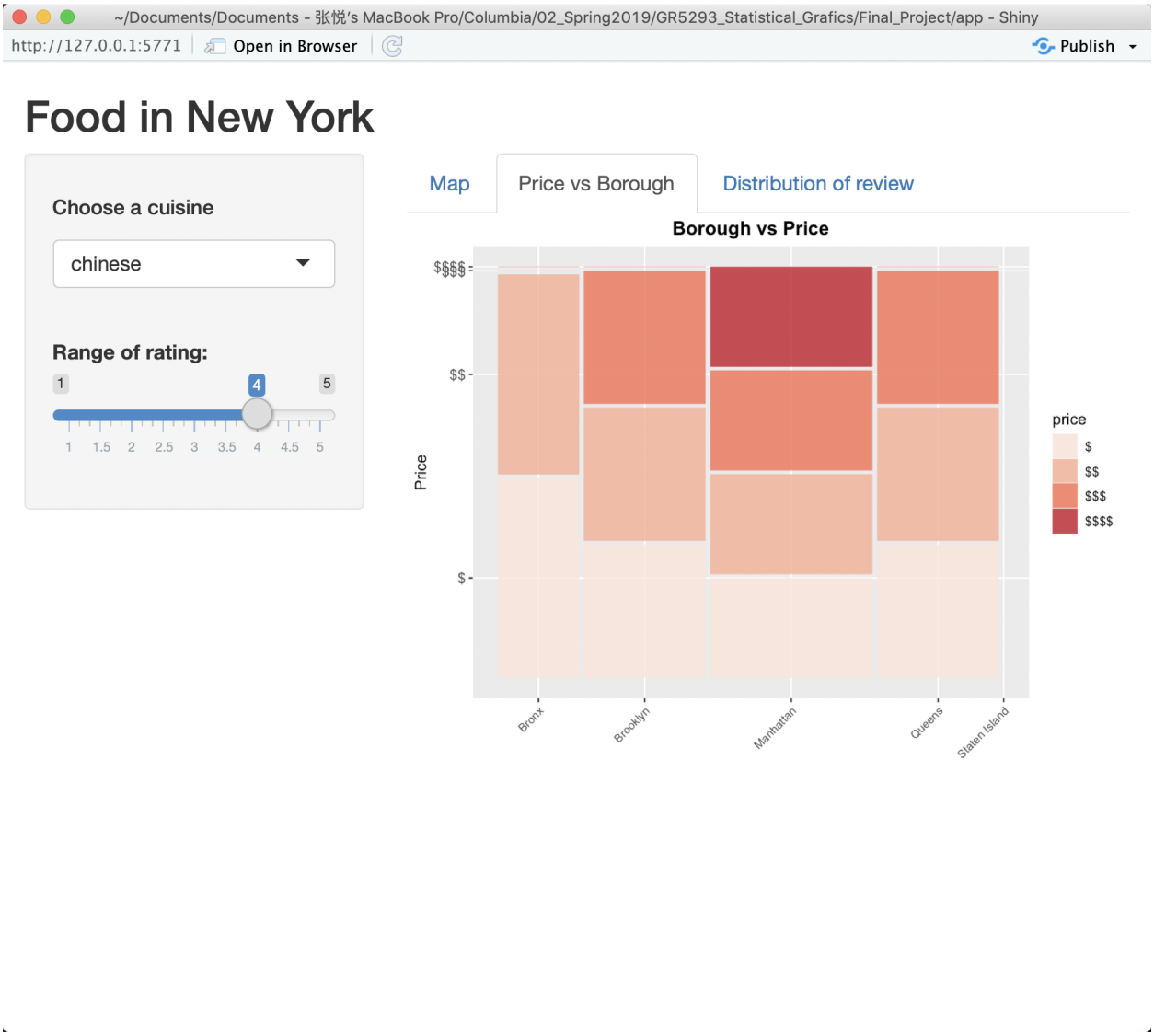

~/Documents/Documents - 张悦's MacBook Pro/Columbia/02_Spring2019/GR5293_Statistical_Grafics/Final_Project/app - Shiny
http://127.0.0.1:5771
Open in Browser
Publish
Food in New York
Choose a cuisine
chinese
Range of rating:
1
4
5
1 1.5 2 2.5 3 3.5 4 4.5 5
Map
Price vs Borough
Distribution of review
Borough vs Price
Price
$$$$
$$
$
Bronx
Brooklyn
Manhattan
Queens
Staten Island
price
$
$$
$$$
$$$$

Figure 4:

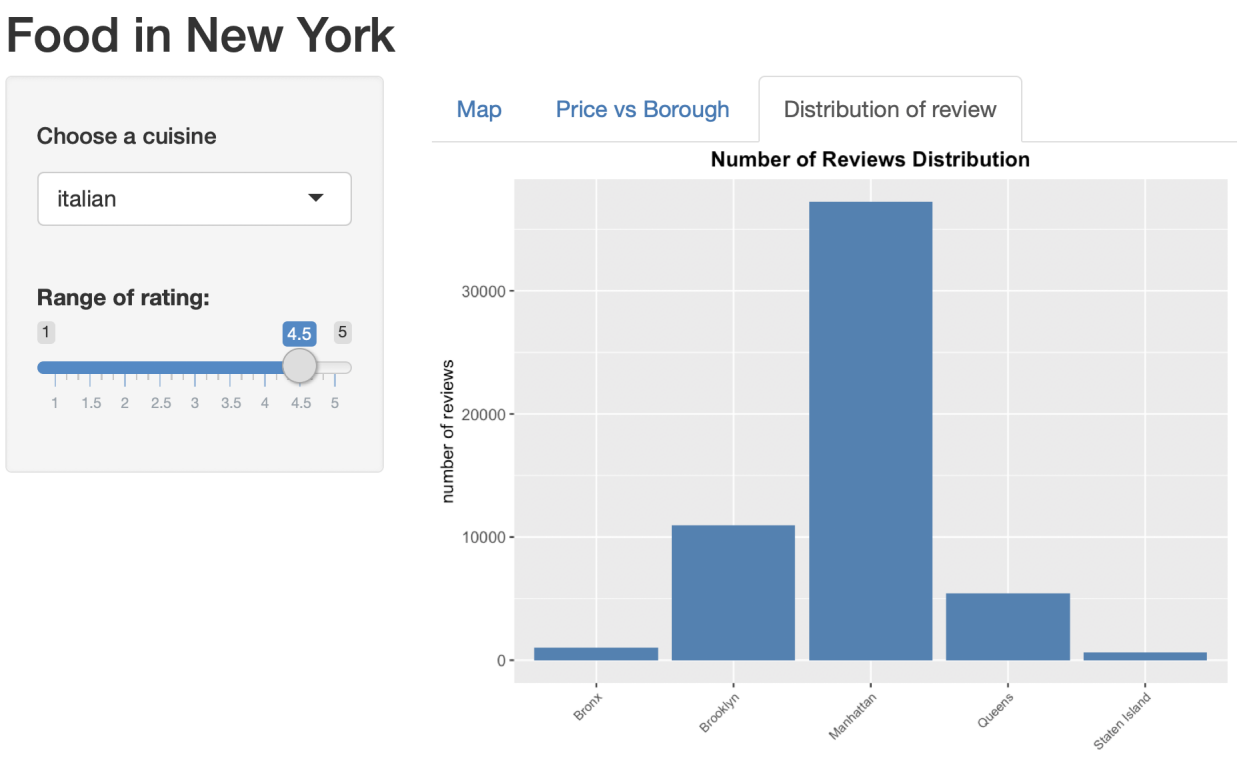


Figure 5:

Is the number of reviews matters?

Interested in 5 star resaurants only? No problem at all. The plot below is the distribution of number of reviews we would expect for all cuisine.

But for some cuisines, this is not case for some cuisine like Greek and Spanish. Most of the Greek restaurants with 5 star rating are located in Queens instead of Manhanttan. Janpaness food, however, are located in Brooklyn. From the Price vs Borough we can also find that the price of Japaness restaurant in Brooklyn is not as high as the price in Manhanttan. So we can assume that the price is also an important part which affect the rating.

All data and code documents can be found here. Our R shiny app works fine locally, but somehow when we tried to publish it caused problems. We just can't figure out how to solve it. We talked to the professor about this problem and she said it's fine as long as our app can work. Our R shiny app is located at the app.R folder on the GiHub above.

VII Conclusion

Due to the limit of our access to the data from Yelp, most of our data are discrete variables, which also limit our ways to explore the data.

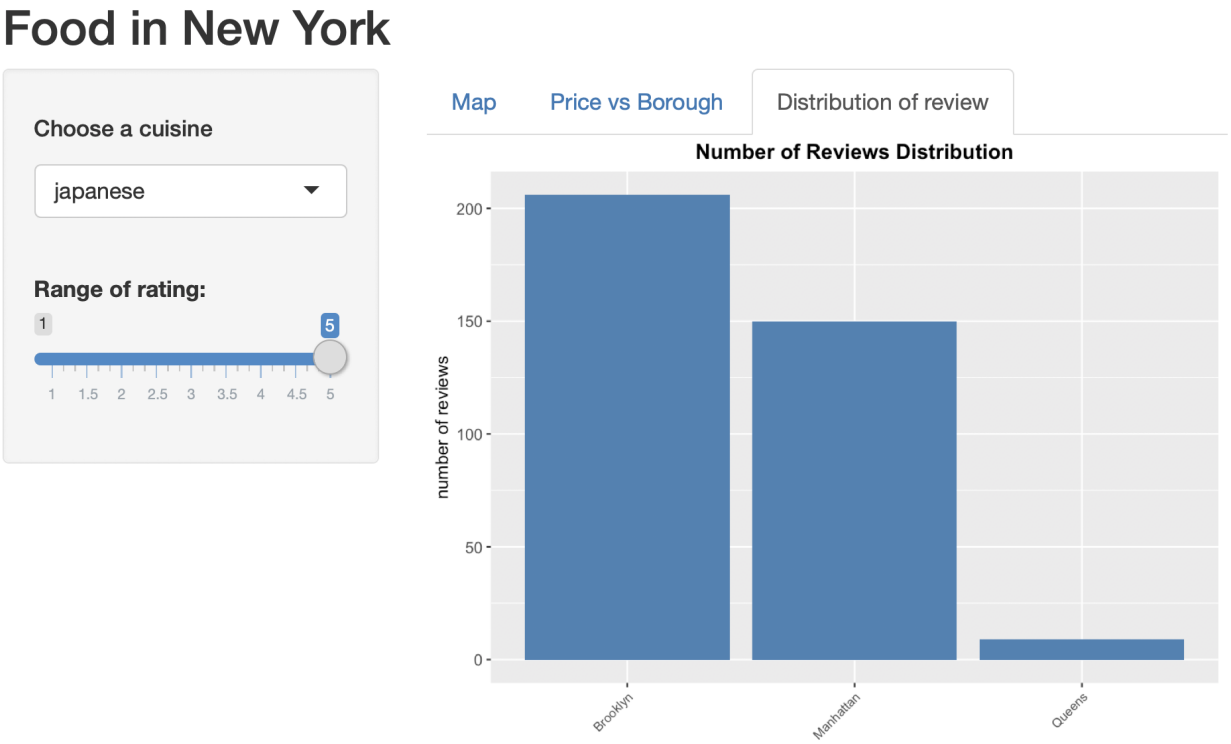

Food in New York
Choose a cuisine
japanese
Range of rating:
1
5
1 1.5 2 2.5 3 3.5 4 4.5 5
Map
Price vs Borough
Distribution of review
Number of Reviews Distribution
number of reviews
200
150
100
50
0
Brooklyn
Manhattan
Queens

Figure 6: