What Makes a Highly-Rated Restaurant on Yelp?

Final Paper for ECO225 - Big Data Tools for Economists

Yexuan Shen

Supervised by Professor Nazanin Khazra

Department of Economics

University of Toronto

April 16, 2022

1 Introduction

1.1 Research Question

Online ratings and reviews have increasingly become important either to consumers and businesses. For consumers, it is convenient to check online ratings anywhere and then make a best decision before shopping or eating out. For businesses, ratings and reviews are critical to their reputation, sales and future prospects. Ratings can both help consumers know more about what they are going to spend money on and give businesses feedback for possible future improvements.

Yelp has became a popular app for many people to research business ratings and reviews as part of their decision-making process before shopping or eating out. Thus, it's critical for businesses on Yelp to get high ratings from its previous customers so that they will have more new customers in the future.

This research project aims to study factors which can possibly affect the ratings of restaurants on Yelp. Studying what could make a highly-rated restaurant on Yelp can either help people explore what is important for customers to rate on restaurants or help restaurants improve their services efficiently by sharing insights and data-driven recommendations. This is the main reason why I believe my project is meaningful and important to do.

The outcome variable (Y) of my interest in this research is *stars*, which represents the rating level of a restaurant on Yelp, ranging from 1 to 5 stars indicating 'not good' to 'great'. Three multiple predictive covariates (X's) are *regional_food*, *open_saturday*, *open_sunday*, which are all new variables created out from original dataset variables. *regional_food* represents the regional cuisine the restaurant has, including American, Mexican, Thai, Chinese, Japanese, Korean food and Other food. I mainly focus on these popular and common regional food in order to simplify the comparison of different typical categories of restaurants. *open_saturday* indicates whether the restaurant open on Saturday while *open_sunday* indicates whether the restaurant open on Sunday.

1.2 Literature Review

One previous research on Yelp has pointed out that online consumer reviews and ratings could affect restaurant demand [1]. One of its key findings states that a one-star increase in Yelp rating results in a 5% to 9% increase in revenue [1]. Another research on the effects of online ratings found that higher ratings have a substitute effect on restaurants' advertising, which is likely to reduce the cost of advertising [2]. These all indicate the advantages of high online ratings and good online reviews for businesses.

Previous findings on the effects of ratings also inspire me to explore the reasons for high ratings so that consumer behaviours could be analyzed and businesses could possibly learn from my research and get improved on themselves.

1.3 Summary

In the project, I am going to clean and merge useful datasets, summarize key variables, visualize meaningful information about Yelp restaurants' ratings and its correlation with important factors including regional food types and whether open on weekends. I will also use web scraping techniques to add new data to the original dataset and study the correlation between average stars of ratings and U.S. state per capita income. Finally, I will run OLS regressions as well as a regression tree to deeply explore the linear relationship under this research topic.

One key finding of my research is that opening on Saturday has a positive effect on the ratings for restaurants on Yelp. Another finding is that Thai food restaurants have highest average ratings among all common regional food restaurants, while Chinese Food restaurants have relatively lowest average ratings. In addition, the regression results show significance on the effects of opening on weekends. Unexpectedly, Thai food is not a significant factor affecting ratings while some other regional foods have significant effects on ratings including Chinese food, Mexican food, and American food but all with negative effects.

In the following sections, I will introduce the data source, show summary statistics

and initial data visualization, and analyze regression results.

2 Data

The main data I am going to use in this project is Yelp's business data [3]. These Yelp datasets can be found on Kaggle. In these datasets, we can find thousands of data points on useful information about restaurants on Yelp with their ratings and some factors which might be correlated with their ratings. However, the limitation is that there are many useless variables to be deleted from the dataset and not enough meaningful variables to do a deeper analysis. Thus, I will use web scraping techniques to scrape U.S. state per capita income data on Wikipedia [4]. United States Census data on states location will also be used to plot maps [5].

3 Summary Statistics & Visualization

3.1 Yelp Data

The data has been cleaned, merged, and renamed in order for future analysis and visualization. In this section, key information of data will be summarized in table and visualized through figures.

Table 1: Summary Statistics

	stars	**review_count**
count	6928.000000	6928.000000
mean	3.763712	278.393043
std	0.522650	295.422508
min	1.500000	100.000000
25%	3.500000	131.000000
50%	4.000000	187.000000
75%	4.000000	307.000000
max	5.000000	4869.000000

From the above *Table 1: Summary Statistics*, we can see that for our main dataset of restaurants on Yelp after data cleaning process, there are 6928 observations in total.

The overall mean stars of ratings is 3.76 and the mean counts of reviews is 278. The minimum stars of ratings is 1.5 and the maximum stars of ratings is 5. Since we droped the extreme values of number of review counts before, now the minimum counts of reviews is 100 and the maximum counts of reviews is 4869.

Table 2: Summary by State

state	latitude	longitude	stars	review_count
ON	43.692915	-79.404918	3.644231	193.680556
BW	48.776986	9.176954	3.727273	111.909091
SC	35.035569	-80.954050	3.730769	159.692308
NV	36.113527	-115.182587	3.749349	403.873372
PA	40.443329	-79.979507	3.775766	213.289694
AZ	33.485421	-111.984374	3.783800	259.983100
NC	35.206485	-80.843005	3.792398	215.005848
OH	41.458666	-81.678710	3.809038	193.206997
WI	43.075003	-89.403422	3.819767	207.796512
IL	40.114661	-88.237532	3.840426	195.127660
QC	45.507422	-73.576313	3.982456	219.643275
EDH	55.950464	-3.190302	4.111111	152.944444

In *Table 2: Summary by State*, grouping by state, it is interesting to see average ratings and counts of reviews of each state. We can tell that the state ON has the lowest average stars of ratings being 3.644 with 194 average counts of reviews, while the state EDH has highest average stars of ratings being 4.111 with 153 average counts of reviews.

From the above *Figure 1: Histogram of Stars of Ratings*, we can see the distribution of stars of ratings of restaurants in the target dataset. It is roughly normally distributed with a little bit left-skewed. The most frequencies of stars of ratings occur at 4, which indicates a good rating. There are just a few extremely low ratings and high ratings, while most restaurants' ratings range from 3 to 4.5, which means most restaurants are considered okay or good to customers on Yelp.

According to *Figure 2: Boxplot of Ratings by Whether Open on Saturday*, restaurants open on Saturday have a higher average stars of ratings than those not open on Saturday, with a difference of over 0.3 stars on average. It is reasonable to guess that whether open on Saturday may affect people's rating of restaurants because people are very likely to eat out on weekends. Some people might not be very happy with those restaurants not

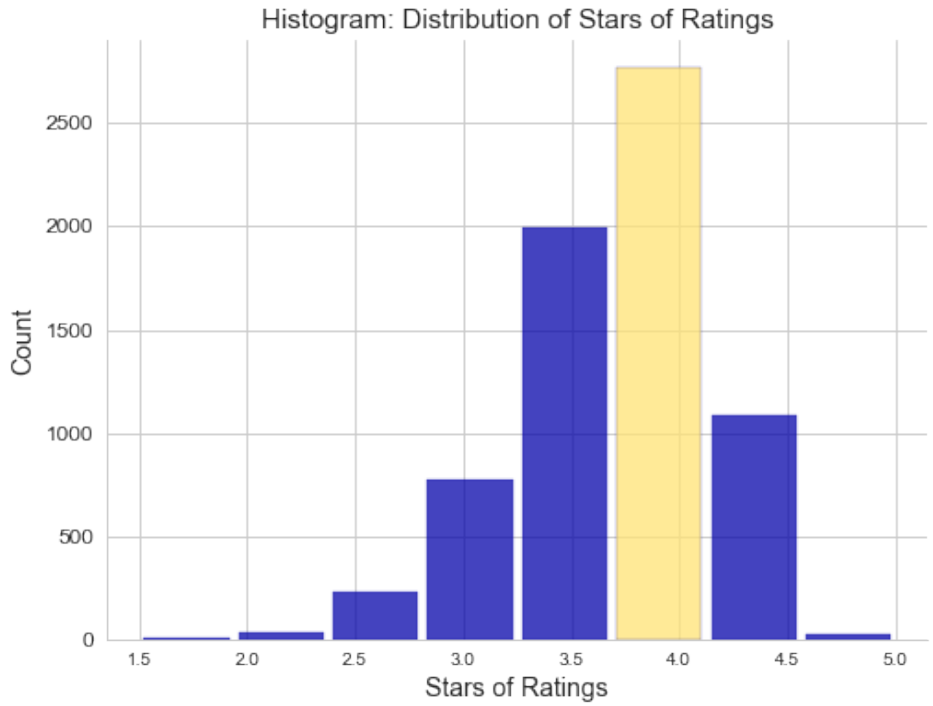

Figure 1: Histogram of Stars of Ratings

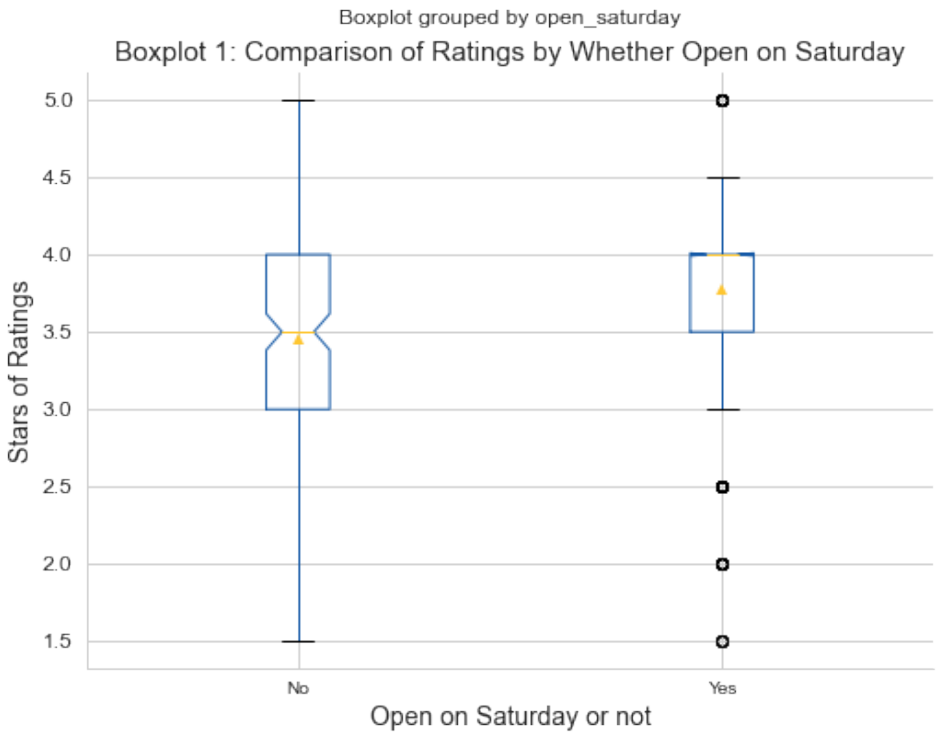

Figure 2: Boxplot of Ratings by Whether Open on Saturday

open on Saturday.

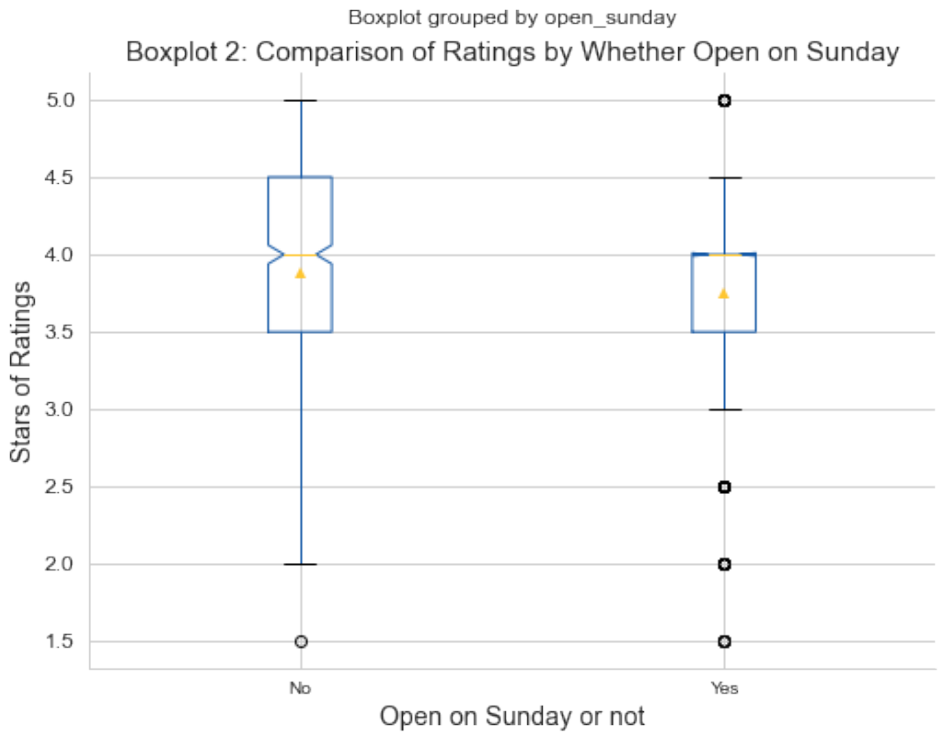

Figure 3: Boxplot of Ratings by Whether Open on Sunday

According to *Figure 3: Boxplot of Ratings by Whether Open on Sunday*, restaurants open on Sunday have a little bit lower average stars of ratings than those not open on Sunday, with a difference of about 0.1 stars on average, which is surprising. It is different than what we saw for Saturday previously. It is reasonable to guess that whether open on Sunday does not matter a lot because it is very common for restaurants to close on Sunday so people would not be unhappy that much if a single restaurant closes on Sunday. The reason why I choose to study whether the restaurant open on Saturday and Sunday as key covariates to ratings because a lot of restaurants close on Weekends so whether open on weekends could probably make a difference among restaurants to customers.

From the above *Figure 4: Barplot of Retional Food Ratings*, we can tell that Thai Food restaurants have highest average stars of ratings close to 3.9, while Chinese Food restaurants have relatively lowest average stars of ratings close to 3.6. The reason why I choose the regional category as a covariate is that it is meaningful to see which regional food has higher ratings and which has lower ratings and needs improvement. The market

Figure 4: Barplot of Retional Food Ratings

competition of different categories of restaurants can also be shown through this covariate. Although the average ratings are not bad for each category, it is good to compare them and learn which categories of restaurants should improve themselves to get higher ratings.

In conclusion to the visualizations above, a main message can be conveyed that business hours and type of food could both possibly affect a restaurant's rating on Yelp. Restaurants opening on Saturdays are more likely to get higher ratings on Yelp than those not opening on Saturdays, and Thai food restaurants also have higher average ratings on Yelp than other regional foods.

3.2 Geo-spatial Analysis & Maps

I am going to show three U.S. maps as the illustration of my main message mentioned above. With comprehensive comparison and analysis of these informative maps, the main message will be clearly explained.

The first map I plotted above (*Figure 5*) indicates the average stars of each state

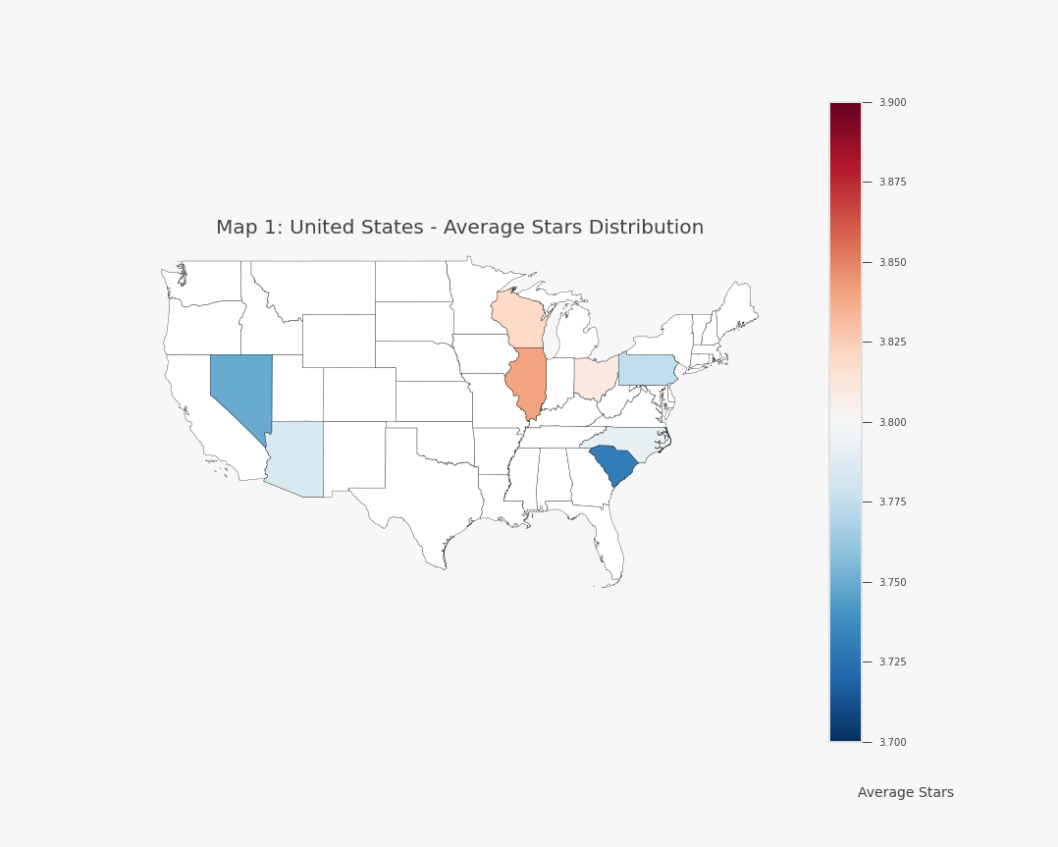

Figure 5: U.S. Average Stars Distribution

among the 8 states in U.S. available in Yelp data, with the average stars ranging from 3.7 to 3.9. Darker red color means relatively higher average stars, while darker blue color means relatively lower average stars. From this map, we can see the geo-spatial distribution of the outcome variable, i.e., stars of ratings.

The second map I plotted above (*Figure 6*) illustrates the percentage of restaurants opening on Saturdays in each state among the 8 states in U.S. available in Yelp data, with the percent ranging from 0.95 to 1. Darker red color means relatively higher percentage of restaurants opening on Saturday, while darker blue color means relatively lower percentage. From this map, we can see the geo-spatial distribution of an important independent variable, i.e., whether open on Saturday.

Comparing the first map (*Figure 5*) and second map (*Figure 6*) above, we can see the same relation between the outcome variable *stars* and one of the independent variables *open_saturday*. The comparison of maps demonstrates that states with higher percentage of restaurants opening on Saturday are likely to have higher average stars of ratings,

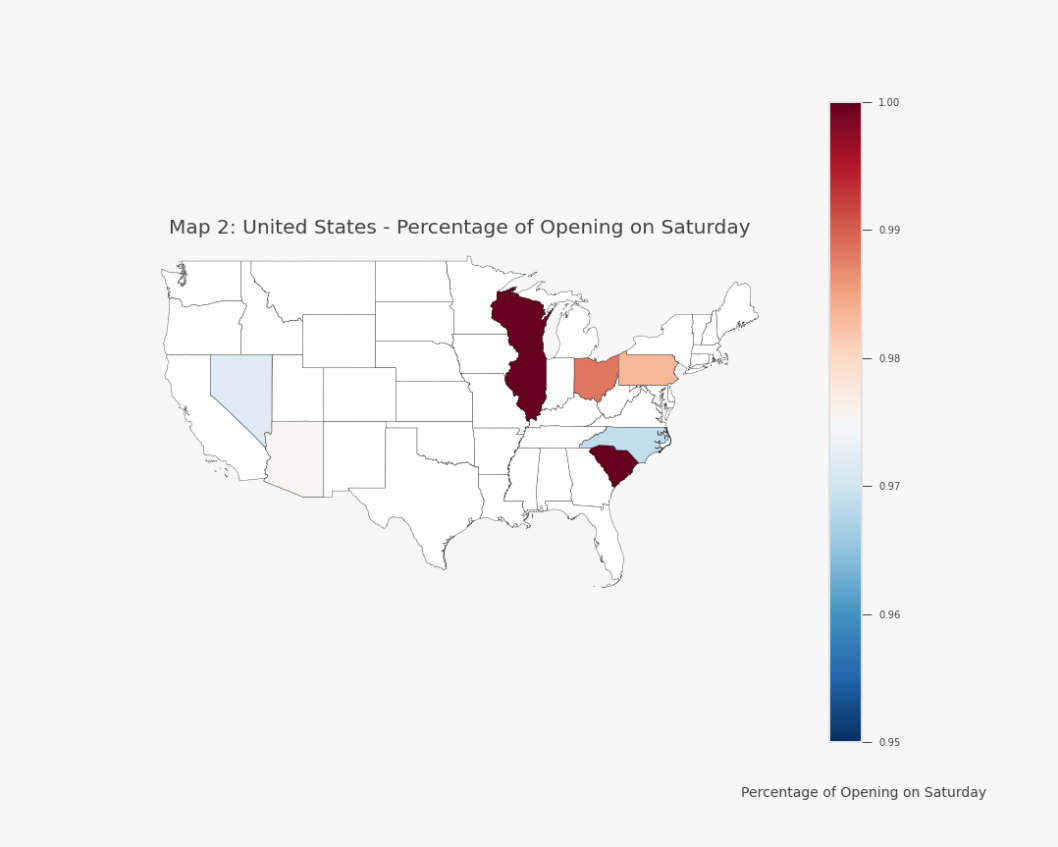

Figure 6: U.S. Percentage Opening on Saturday

which provides support for one previous key visualized finding that restaurants open on Saturday tend to have a higher average stars of ratings than those not open on Saturday.

The third map (*Figure 7*) above shows the percentage of Thai restaurants in each state among the 8 states in U.S. available in Yelp data, with the percent ranging from 0 to 0.09. Darker red color means relatively higher percentage of Thai restaurants, while darker blue color means relatively lower percentage. From this map, we can see the geo-spatial distribution of another important independent variable about the regional food.

Comparing first map (*Figure 5*) and third map (*Figure 7*), we can see a similar relation between outcome variable *stars* and another independent variable *thai_food*. The comparison of these two maps demonstrates that states with higher percentage of Thai restaurants are likely to have higher average stars of ratings, which provides support for another previous key visualized finding that Thai restaurants have a higher average stars of ratings than other regional foods.

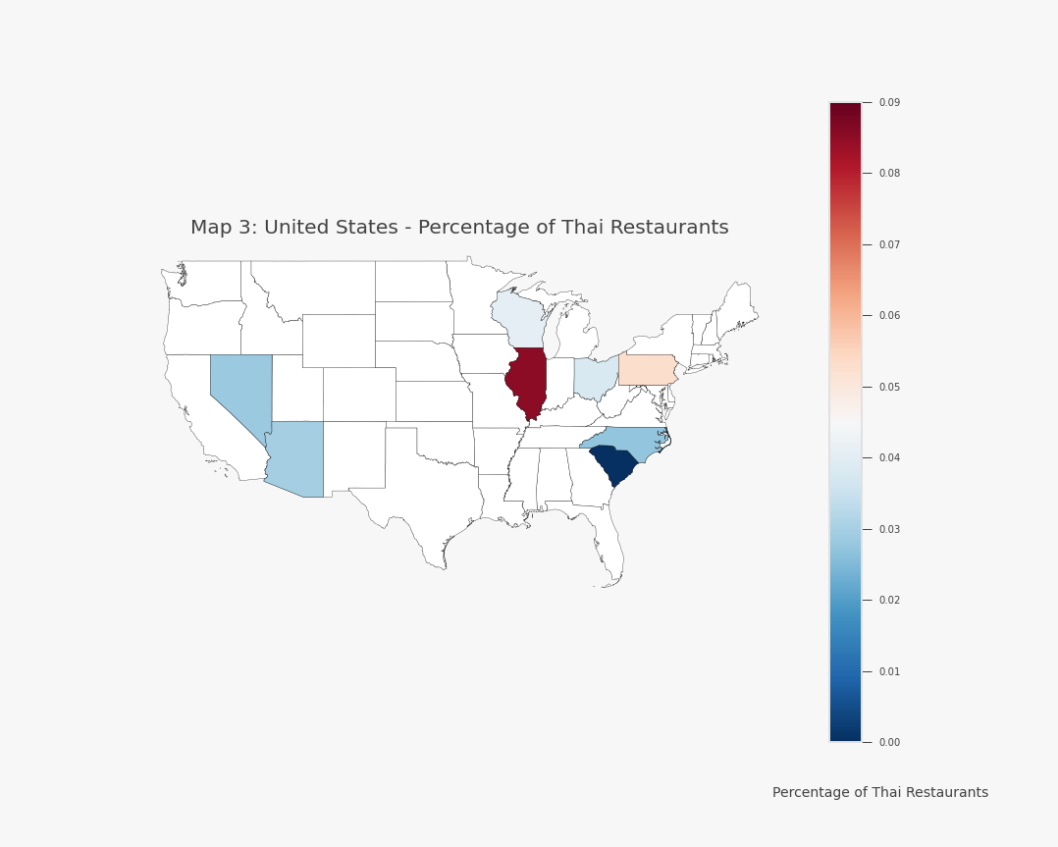

Figure 7: U.S. Percentage of Thai Restaurants

3.3 Web Scraping Challenges & Practice

To better investigate factors related to ratings for restaurants, only using Yelp dataset is too limited. Through web scraping, I might be able to acquire any data needed and add more diversity to the topic.

Google Maps could be a very useful website to scrape data and improve my research results for the following reasons. First of all, almost every existing business can be searched on google map, with detailed information about business hours, location, ratings and reviews, etc. It has similar accessible information as Yelp, but more registered businesses than on Yelp. Second, traffic factor can be taken into consideration if we can scrape some traffic data from google map. It can be important because customers are likely to rate a restaurant based on the traffic status around it. Third, based on my current location, I can search for all restaurants nearby in a certain distance on google map and get all their information to compare. In addition, we can track the ratings on

Google Map over time to analyze the change of ratings. Specifically, we can scrape data from *https://www.google.com/maps* using available APIs.

There are two ways to use the new data from Google Maps to enhance my current research. One way is to add new information to existing restaurants in my Yelp dataset. I will merge the new data, which has more information about existing restaurants, new ratings and reviews, with my original data. Thus, I could investigate new factors such as traffic for each restaurant on Yelp, and I could also be able to compare ratings on Yelp and ratings on Google Map to see if there exists an unexpected difference and explore the possible reasons. The Other way is to add new restaurants to my Yelp dataset with similar aspects of information, so that I will be able to analyze on more data and get improved results.

However, there are possible challenges which make it hard for me to scrape data from Google Maps in this project. Firstly, it might take a long time to get APIs from Google, hence I am not allowed to scrape data from Google Maps for now. Secondly, if I want to use the data to track the rating change over time, the program would take months or years to generate the data. Moreover, my current programming knowledge about web scraping does not allow me to get the traffic information on maps, which means I need to learn more or acquire help from programming professionals. In the future, I need to learn more long-term API-based scraping to work on these challenges. It will also take a long time to complete the future project with these ideas.

In practice, I did some HTML-based web scraping to get income per capita data for every state and territory in the U.S. and merge it with U.S. state postal codes for later use. After cleaning and merging target datasets, I am going to do some U.S. income visualization based on the scraped new data.

We can see from the histogram (*Figure 8*) that U.S. state per capita income is roughly normally distributed. The highest frequencies are between 30000 to 35000 dollars, which are highlighted in yellow color.

Then, I made two barplots to show per capita income of each state in 2019 and average rating of stars of each state separately. Thus, we can compare these two barplots to see if

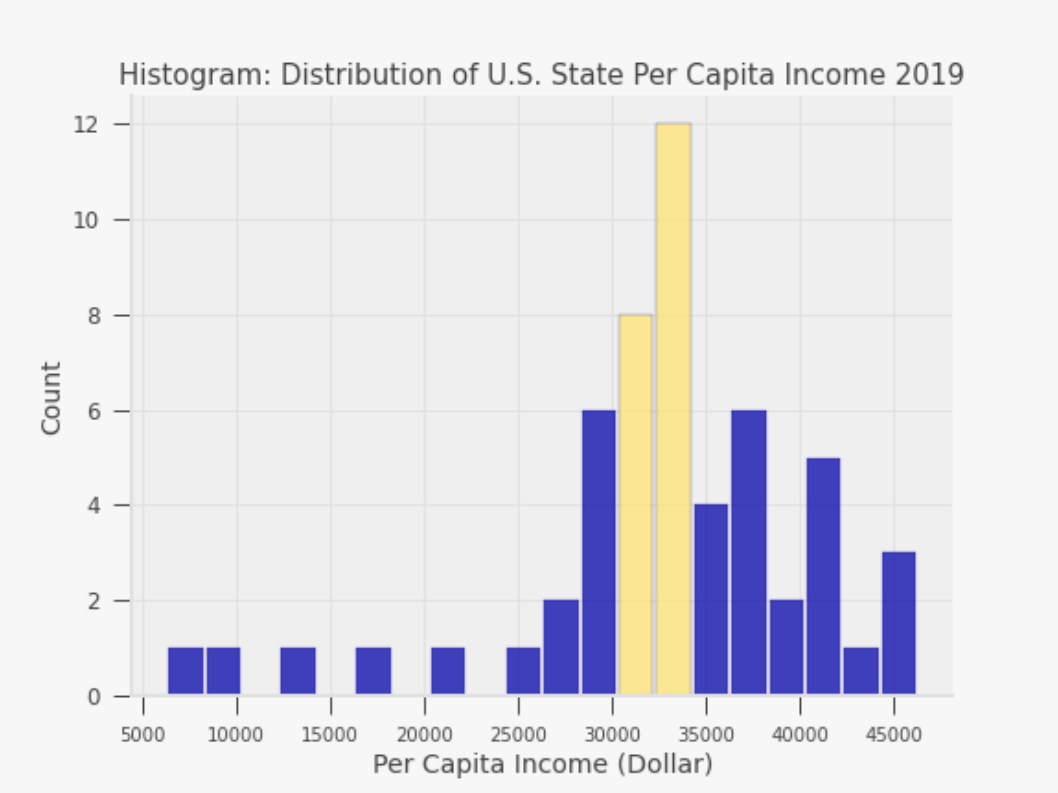

Figure 8: U.S. State Per Capita Income 2019

states with higher per capita income also have higher average stars. The bars are ordered by state with highest per capita income to state with lowest per capita income. In the current dataset, there are only 8 U.S. states to be plotted, including IL, PA, WI, NV, OH, AZ, NC, and SC.

According to the two barplots in *Figure 9*, it is obvious that the state IL has highest per capita income in 2019 (near 38000 dollars) as well as highest average stars of rating (around 3.84) for restaurants on Yelp. Also, the state SC with lowest per capita income in 2019 (around 31000 dollar) has lowest average stars (around 3.73). Nevertheless, from comparison of two barplots, it is not always the case that higher per capita income is corresponding to higher average stars of ratings. For instance, the state PA with second highest per capita income only gets around 3.78 average stars of ratings, which can only be ranked sixth in total 8 states.

To better visualize the relationship between state per capita income and average stars, I also made a line plot, using average stars of ratings as a function of per capita income in 2019.

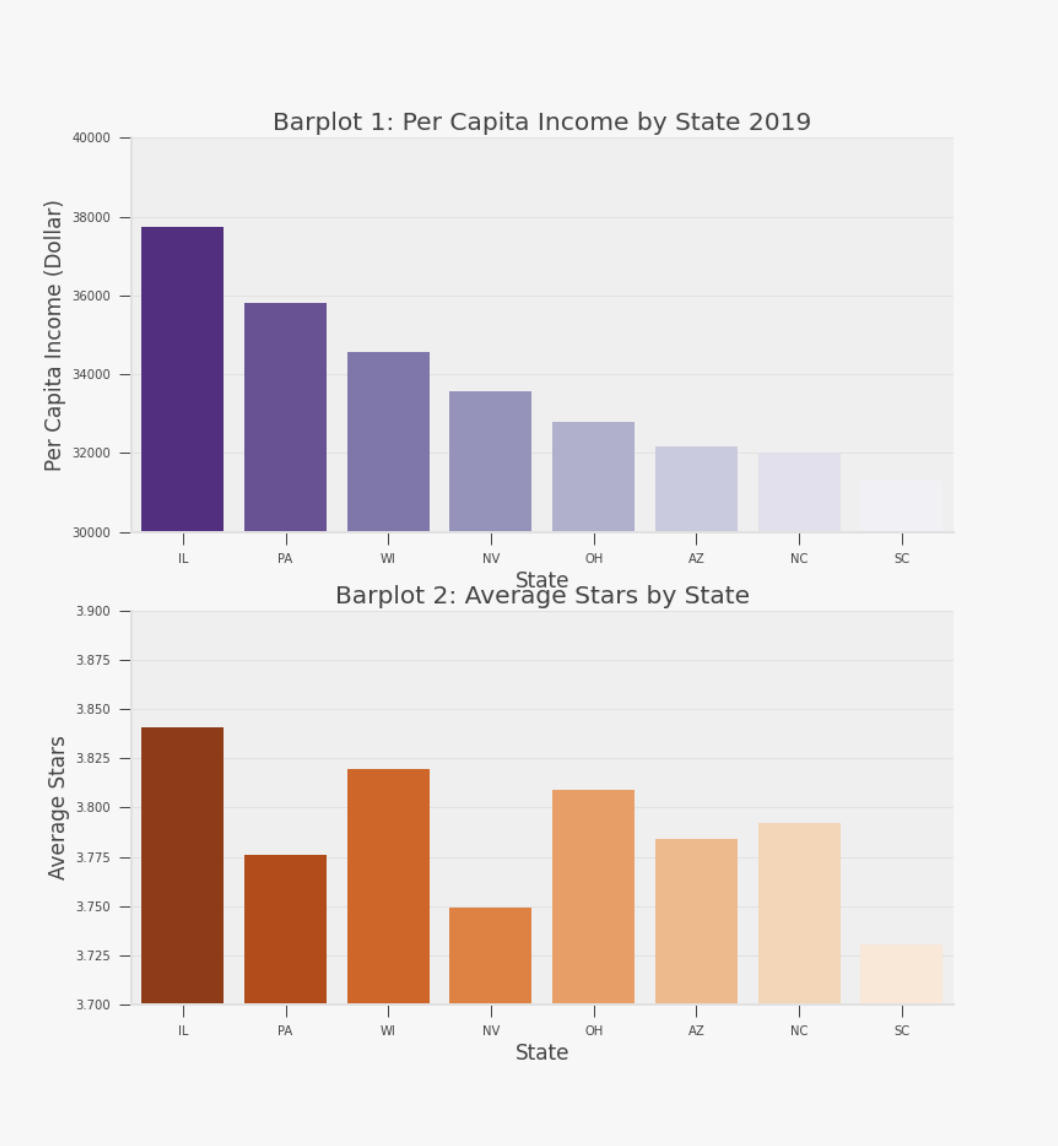

Figure 9: Comparison of Two Barplots

From the *Figure 10* line plot, it seems to be a very roughly positive relationship between state average stars of ratings and state per capita income in 2019. Therefore, in general cases, higher income could indicate higher average stars. However, the line looks fluctuating because some states are exceptions which have high per capita income but low average stars of ratings.

From economics perspective, the general roughly positive relationship could be attributed to people's behaviours. Specifically, high income individuals may be more educated and more inclusive, hence more likely to give high stars of ratings to restaurants. Yet, a possible reason for exceptional cases is the lack of data and information, i.e, the

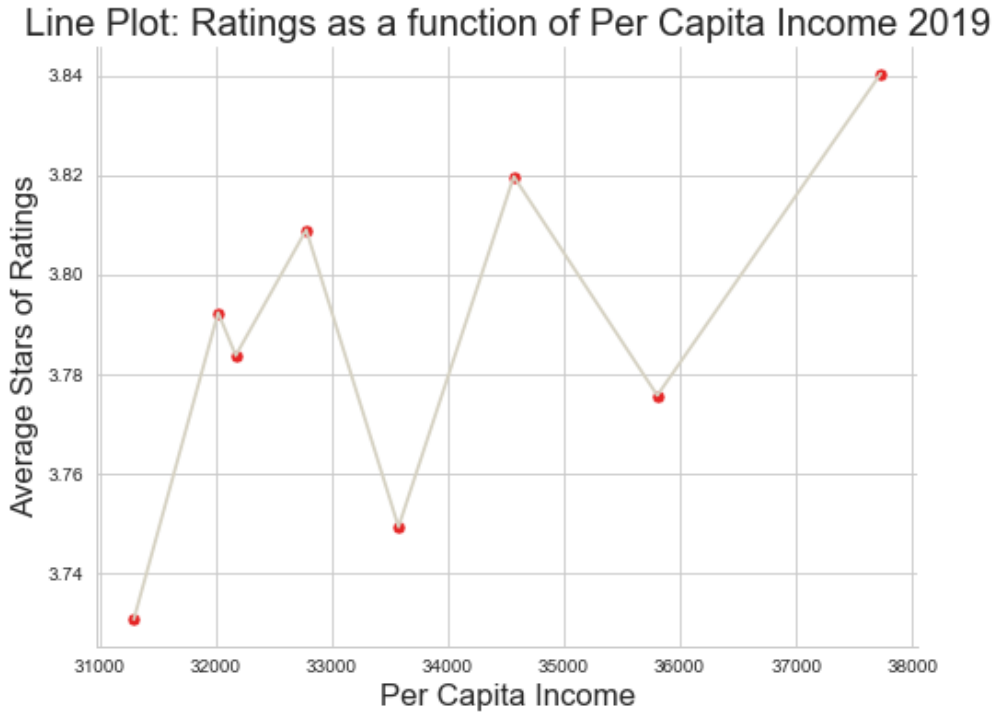

Figure 10: Lineplot of Ratings v.s.Per Capita Income

number of states in my dataset is too small which could limit me from finding the real relationship between stars and income. Also, there are many other factors which could affect the stars of ratings for restaurants on Yelp. I will need more information to generate a more reasonable conclusion about the relationship between per capita income and average stars of ratings, which could be one of the limitations to be resolved in the future.

4 Regression Results

In the previous parts, I determined my dependent variable to be stars of ratings on Yelp, and three independent variables to be regional cuisine type, whether open on Saturday, whether open on Sunday. Based on previous data visualization, there might exist both linear and non-linear economic relationship between the dependent variable and independent variables except for whether open on Sunday on which I have not found any obvious relationship with stars of ratings. Regional cuisine type may have a non-linear relationship with ratings because multiple categories exist in this independent variable

but it is hard to rank them so that a linear trend is hard to find. Whether open on Saturday is likely to have a linear relationship with ratings. Due to people's behaviours and businesses' competition, restaurants open on Saturday are more likely to get higher ratings.

I believe the chosen independent factors can explain my outcome variable, i.e., stars of ratings, because the visualization evidence in previous parts indicates some relation between these factors and stars of ratings.

In another aspect, the dependent variable could be the average stars of ratings by state and the independent variable could be per capita income by state. They might also have a linear relationship because higher income might indicate higher education level and higher possibilities of giving good ratings.

4.1 Four OLS Regressions

For the first regression, I intend to include all the possible terms to build a full model. For the independent variable indicating regional food type, I will use multiple dummy variables to represent different common types of regional food including thai, chinese, american, japanese, korean and mexican, and the reference group will be other types. Since the effect of each possible factor may differ depending on the number of reviews, I will include the interaction terms in this regression model.

The first regression model can be written as

$$stars = \beta_0 + \beta_1 open\_saturday + \beta_2 open\_sunday + \beta_3 thai\_food + \beta_4 chinese\_food + \beta_5 american\_food + \beta_6 mexican\_food + \beta_7 japan\_food + \beta_8 korean\_food + \beta_9 open\_saturday \times review\_count + \beta_{10} open\_sunday \times review\_count + \beta_{11} thai\_food \times review\_count + \beta_{12} chinese\_food \times review\_count + \beta_{13} american\_food \times review\_count + \beta_{14} mexican\_food \times review\_count + \beta_{15} japan\_food \times review\_count + \beta_{16} korean\_food \times review\_count + u$$

where:

- β_0 is the intercept of the linear trend line on the y-axis

- β_1 represents the *difference of stars* between open on Saturday and not open on Saturday

- β_2 represents the *difference of stars* between open on Sunday and not open on Sunday

- β_3 to β_8 represents the *difference of stars* between each corresponding regional food type and the reference group which represents other regional foods

- β_9 to β_{16} represents each corresponding *interaction effect* with number of review counts

- u is the error term (deviations of observations from the linear trend due to factors not included in the model)

The regression results are shown in the following *Table 3*.

Table 3: OLS Regression 1 Results

report booktabs

Dep. Variable:	stars	**R-squared:**	0.074
Model:	OLS	**Adj. R-squared:**	0.072
Method:	Least Squares	**F-statistic:**	36.69
Date:	Sat, 16 Apr 2022	**Prob (F-statistic):**	2.47e-103
Time:	09:05:05	**Log-Likelihood:**	-5069.3
No. Observations:	6928	**AIC:**	1.017e+04
Df Residuals:	6912	**BIC:**	1.028e+04
Df Model:	15		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	3.4700	0.078	44.419	0.000	3.317	3.623
C(open_saturday)[T.1]	0.5174	0.085	6.120	0.000	0.352	0.683
C(open_sunday)[T.1]	-0.2188	0.035	-6.198	0.000	-0.288	-0.150
C(thai_food)[T.1]	0.0671	0.046	1.454	0.146	-0.023	0.158
C(chinese_food)[T.1]	-0.2005	0.035	-5.656	0.000	-0.270	-0.131
C(american_food)[T.1]	-0.1839	0.020	-9.380	0.000	-0.222	-0.145
C(mexican_food)[T.1]	-0.2088	0.034	-6.182	0.000	-0.275	-0.143
C(japan_food)[T.1]	-0.0440	0.039	-1.131	0.258	-0.120	0.032
review_count	0.0005	0.000	1.101	0.271	-0.000	0.001
C(open_saturday)[T.1]:review_count	-3.088e-05	0.000	-0.073	0.942	-0.001	0.001
C(open_sunday)[T.1]:review_count	-0.0002	0.000	-2.169	0.030	-0.000	-2.32e-05
C(thai_food)[T.1]:review_count	-4.736e-05	0.000	-0.429	0.668	-0.000	0.000
C(chinese_food)[T.1]:review_count	-9.951e-05	0.000	-0.900	0.368	-0.000	0.000
C(american_food)[T.1]:review_count	0.0001	4.55e-05	3.047	0.002	4.94e-05	0.000
C(mexican_food)[T.1]:review_count	0.0002	9.37e-05	1.964	0.050	3.54e-07	0.000
C(japan_food)[T.1]:review_count	0.0001	0.000	1.042	0.298	-9.37e-05	0.000

Omnibus:	320.864	**Durbin-Watson:**	1.986
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	372.996
Skew:	-0.525	**Prob(JB):**	1.01e-81
Kurtosis:	3.435	**Cond. No.**	1.33e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.33e+04. This might indicate that there are strong multicollinearity or other numerical problems.

For the second regression, I plan to delete the term *review_count* because the first full model's multi-collinearity problem related to this term.

The second regression model can be written as

$stars = \beta_0 + \beta_1 open\_saturday + \beta_2 open\_sunday + \beta_3 thai\_food + \beta_4 chinese\_food + \beta_5 american\_food + \beta_6 mexican\_food + \beta_7 japan\_food + \beta_8 korean\_food + u$

where:

- β_0 is the intercept of the linear trend line on the y-axis

- β_1 represents the *difference of stars* between open on Saturday and not open on Saturday

- β_2 to β_8 represents the *difference of stars* between each corresponding regional food type and the reference group which represents other regional foods

- u is the error term

The regression results are shown in *Table 4*.

Table 4: OLS Regression 2 Results

report booktabs

Dep. Variable:	stars	**R-squared:**	0.051
Model:	OLS	**Adj. R-squared:**	0.050
Method:	Least Squares	**F-statistic:**	46.88
Date:	Sat, 16 Apr 2022	**Prob (F-statistic):**	4.65e-74
Time:	09:25:28	**Log-Likelihood:**	-5151.9
No. Observations:	6928	**AIC:**	1.032e+04
Df Residuals:	6919	**BIC:**	1.038e+04
Df Model:	8		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	3.5386	0.039	91.054	0.000	3.462	3.615
C(open_saturday)[T.1]	0.5452	0.044	12.376	0.000	0.459	0.632
C(open_sunday)[T.1]	-0.2650	0.024	-11.180	0.000	-0.311	-0.219
C(thai_food)[T.1]	0.0539	0.036	1.508	0.132	-0.016	0.124
C(chinese_food)[T.1]	-0.2324	0.025	-9.166	0.000	-0.282	-0.183
C(american_food)[T.1]	-0.1316	0.015	-8.987	0.000	-0.160	-0.103
C(mexican_food)[T.1]	-0.1596	0.023	-6.923	0.000	-0.205	-0.114
C(japan_food)[T.1]	-0.0103	0.026	-0.395	0.692	-0.061	0.041
C(korean_food)[T.1]	-0.0594	0.047	-1.263	0.206	-0.152	0.033

Omnibus:	341.884	**Durbin-Watson:**	1.986
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	397.458
Skew:	-0.552	**Prob(JB):**	4.93e-87
Kurtosis:	3.397	**Cond. No.**	15.9

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In the third regression, based on the first full model, I am going to delete other types of regional food categories except for Thai food and Chinese food which have highest and lowest average stars of ratings respectively. I believe these two groups are two most

interesting ones to study, among all regional food groups.

The third regression model can be written as

$stars = \beta_0 + \beta_1 open\_saturday + \beta_2 open\_sunday + \beta_3 thai\_food + \beta_4 chinese\_food + u$

where:

- β_0 is the intercept of the linear trend line on the y-axis

- β_1 represents the *difference of stars* between open on Saturday and not open on Saturday

- β_2 represents the *difference of stars* between open on Sunday and not open on Sunday

- β_3 represents the *difference of stars* between Thai food group and the reference group which represents other regional foods

- β_4 represents the *difference of stars* between Chinese food group and the reference group which represents other regional foods

- u is the error term (deviations of observations from the linear trend due to factors not included in the model)

The regression results can be viewed in *Table 5*.

Table 5: OLS Regression 3 Results

report booktabs

Dep. Variable:	stars	**R-squared:**	0.037
Model:	OLS	**Adj. R-squared:**	0.036
Method:	Least Squares	**F-statistic:**	65.60
Date:	Sat, 16 Apr 2022	**Prob (F-statistic):**	1.50e-54
Time:	09:26:43	**Log-Likelihood:**	-5205.8
No. Observations:	6928	**AIC:**	1.042e+04
Df Residuals:	6923	**BIC:**	1.046e+04
Df Model:	4		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	3.4733	0.039	90.091	0.000	3.398	3.549
C(open_saturday)[T.1]	0.5562	0.044	12.542	0.000	0.469	0.643
C(open_sunday)[T.1]	-0.2692	0.024	-11.285	0.000	-0.316	-0.222
C(thai_food)[T.1]	0.1122	0.035	3.164	0.002	0.043	0.182
C(chinese_food)[T.1]	-0.1739	0.025	-7.039	0.000	-0.222	-0.125

Omnibus:	365.114	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	428.653
Skew:	-0.573	**Prob(JB):**	8.30e-94
Kurtosis:	3.417	**Cond. No.**	15.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In the fourth regression, I will study another aspect of this topic discussed in Project

3, which investigates the relationship between average stars of ratings and per capita income by state. Here, the dependent variable becomes average stars and the independent variable becomes per capita income. This aspect is also interesting to explore because income might be related to ratings of restaurants indirectly. However, the limitation is that the data points are not enough to ensure the accuracy. We can only have a rough look an their relationship because the dataset size is small.

The fourth regression model can be written as

$average\_stars = \beta_0 + \beta_1 per\_capita\_income\_2019 + u$

where:

- β_0 is the intercept of the linear trend line on the y-axis

- β_1 represents the *marginal effect* of per_capita_income 2019 on average stars by state

- u is the error term

The regression results are in *Table 6* as follows.

Table 6: OLS Regression 4 Results

report booktabs

Dep. Variable:	average_stars	**R-squared:**	0.370
Model:	OLS	**Adj. R-squared:**	0.265
Method:	Least Squares	**F-statistic:**	3.523
Date:	Sat, 16 Apr 2022	**Prob (F-statistic):**	0.110
Time:	09:26:54	**Log-Likelihood:**	17.588
No. Observations:	8	**AIC:**	-31.18
Df Residuals:	6	**BIC:**	-31.02
Df Model:	1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	3.4468	0.182	18.944	0.000	3.002	3.892
per_capita_income_2019	1.01e-05	5.38e-06	1.877	0.110	-3.07e-06	2.33e-05

Omnibus:	3.826	**Durbin-Watson:**	2.658
Prob(Omnibus):	0.148	**Jarque-Bera (JB):**	1.119
Skew:	-0.383	**Prob(JB):**	0.571
Kurtosis:	1.335	**Cond. No.**	5.61e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.61e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

4.2 Choice of Regression

I evaluate my regressions based on R-squared, AIC, and BIC values, as well as whether the variables are most economically meaningful. Specifically, R-squared indicates the proportion of the variance for dependent variable explained by independent variables in the regression model. We prefer larger R-squared. AIC and BIC stand for Akaike's Information Criteria and Bayesian Information Criteria, which are common model selection criteria. AIC and BIC could measure model performance, namely, how well the model fits the data. We want AIC and BIC to be small. Therefore, I will choose the best model with smaller AIC, smaller BIC, larger R-squared. Based on this criteria, the first regression and fourth regression have relatively larger R-squared and smaller AIC and BIC, but they have multi-collinearity problem. Thus, the best model I chose was the second regression with R-squared equal to 0.051 and smaller AIC (1.032e+04) and BIC (1.038e+04) than the third regression (*Table 7*).

Table 7: Comparison of 4 Regressions

Regression	R-squared	AIC	BIC	Multicollinearity
Regression 1	0.074	1.017e+04	1.028e+04	Yes
Regression 2	0.051	1.032e+04	1.038e+04	No
Regression 3	0.037	1.042e+04	1.046e+04	No
Regression 4	0.370	-31.18	-31.02	Yes

The best model is the second regression model, written as

$stars = \beta_0 + \beta_1 open\_saturday + \beta_2 open\_sunday + \beta_3 thai\_food + \beta_4 chinese\_food + \beta_5 american\_food + \beta_6 mexican\_food + \beta_7 japan\_food + \beta_8 korean\_food + u$

From the second regression results, we can see that

- The intercept $\hat{\beta}_0 = 3.5386$.

- The slopes $\hat{\beta}_1 = 0.5452$, $\hat{\beta}_2 = -0.2650$, $\hat{\beta}_3 = 0.0539$, $\hat{\beta}_4 = -0.2324$, $\hat{\beta}_5 = -0.1316$, $\hat{\beta}_6 = -0.1596$, $\hat{\beta}_7 = -0.0103$, $\hat{\beta}_8 = -0.0594$.

- The positive $\hat{\beta}_1$ and $\hat{\beta}_3$ parameter estimates imply that whether open on Saturday and whether is Thai food restaurant have positive effects on stars of ratings. In detail, restaurants open on Saturday are likely to get 0.5452 extra stars on average than those

not open on Saturday. Thai restaurants are likely to get 0.0539 extra stars than other regional food restaurants.

- The negative $\hat{\beta}_2$, $\hat{\beta}_4$ to $\hat{\beta}_8$ parameter estimates imply that whether open on Sunday and whether is Chinese/American/Mexican/Japanese/Korean food restaurant have negative effects on stars of ratings.

- The p-values are smaller than 0.05 for $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_4$, $\hat{\beta}_5$, $\hat{\beta}_6$, which implies that these effects are statistically significant (using p ¡ 0.05 as a rejection rule); the p-values for $\hat{\beta}_3$, $\hat{\beta}_7$, $\hat{\beta}_8$ are larger than 0.1, which indicates that these effects are not significant.

- The R-squared value of 0.051 indicates that around 5.1% of variation in stars of ratings is explained in this model.

Using our parameter estimates from above regression results, we can write the estimated relationship as

$\widehat{stars} = 3.5386 + 0.5452open\_saturday - 0.2650open\_sunday + 0.0539thai\_food - 0.2324chinese\_food - 0.1316american\_food - 0.1596mexican\_food - 0.0103japan\_food - 0.0594korean\_food$

4.3 Machine Learning

In addition to common OLS regressions, I am also going to use some machine learning techniques to analyze the relationship between outcome variable and independent variables.

First of all, we need to be clear that the objective function here is to minimize the error of prediction. Parameters in linear regressions are chosen to minimize the mean square error (MSE) function.

Thus, the objective function can be written as

$\frac{1}{N}\sum_{i=1}^{N}(\text{stars}_i - (\beta_0 + \beta_1 open\_saturday_i + \beta_2 open\_sunday_i + \beta_3 thai\_food_i + \beta_4 chinese\_food_i + \beta_5$

In other words, we want to minimize the difference between real value and predicted value of outcome variable, i.e., stars of ratings in this case.

Then I kept only significant variables in the previous second OLS regression to run

a regression tree. The MSE here is approximately 0.2594. The regression tree can be displayed as following. The regression model is displayed graphically in the regression tree (*Figure 11*), showing each prediction level of key variables. We can tell from the regression tree that restaurants open on Saturday are predicted to have relatively higher stars of ratings.

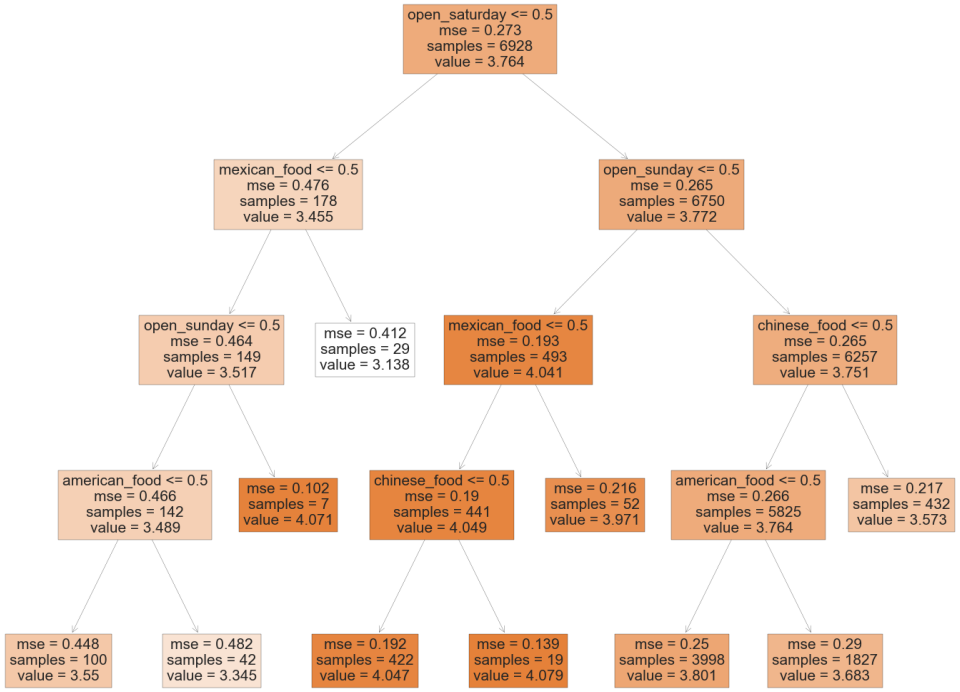

Figure 11: Regression Tree

4.4 Regression v.s. Regression Tree

Compared to regression results, regression tree gives a much clearer graphical view of the regression model. It is easier to view the prediction levels and decision processes in a regression tree. In regression results, there are estimates of effect or dummy group difference, while in regression tree we can directly see the prediction value of each possi-

bility. There are other extra information we can get from the regression tree, including MSE on the each level and sample size on the each level, which is not possible to see in OLS regression.

5 Conclusion & Future Steps

To conclude, I mainly studied some possible factors related to stars of ratings on Yelp restaurants and got a few interesting data-based findings through summary statistics, visualization of various graphs and maps, and regression results.

First of all, most restaurants' ratings on Yelp range from 3 to 4.5 (out of 5), which means most customers feel okay or good to restaurants on Yelp. The Illinois state has the highest average rating of around 3.84 stars, among all available states in the U.S., which can also be seen from the U.S. map.

Secondly, whether open on Saturday seems to be related to ratings, as the restaurants open on Saturday have a 0.3 higher average stars of ratings than those not open on Saturday. Comparison between maps also supports this finding since we can see that U.S. states with higher percentage of restaurants opening on Saturday tend to have higher average ratings. However, whether open on Sunday does not seem to make a big difference to ratings. The reason of the difference might be that people care more about whether restaurants open on Saturday when rating them but more likely to accept that restaurants close on Sunday.

Thirdly, I found Thai food restaurants have relatively highest average ratings close to 3.9 while Chinese food restaurants have lowest average ratings close to 3.6 which needs improvement compared to other regional categories of restaurants. By comparing maps, there is more evidence that Thai food has higher ratings than others because we can see from the maps that states with more percentage of Thai restaurants usually also have higher average ratings.

Lastly, there might exists a rough positive relationship between U.S. state per capita income and state average stars of ratings for Yelp restaurants. It means that a U.S.

state with higher per capita income is likely to have higher average stars, which might be attributed to different behaviours between high income people and low income people. However, the dataset is too small and there are only 8 available U.S. states in the dataset to study, so that I have not found the real accurate relationship yet. I may need more data and more factors to do further study on their correlation.

To further investigate what makes a highly rated restaurant on Yelp, I will explore more possible data existing and methods to analyze the correlation between those factors and stars of ratings. More machine learning techniques such cross validation might be utilized to improve this research. This study will be largely improved if more characteristics data and more restaurant data are available on Yelp, such as whether close to downtown or subway station, whether pet-friendly, etc., which I may also investigate for in the future steps.

References

[1] Luca, Michael, Reviews, Reputation, and Revenue: The Case of Yelp.Com (March 15, 2016). Harvard Business School NOM Unit Working Paper No. 12-016, Available at SSRN: https://ssrn.com/abstract=1928601 or http://dx.doi.org/10.2139/ssrn.1928601

[2] Lei, Ying, Local Restaurants' Advertising Response to Better Online Ratings (November 1, 2017). Available at SSRN: https://ssrn.com/abstract=3059763 or http://dx.doi.org/10.2139/ssrn.3059763

[3] Yelp, Yelp Dataset. Available at Kaggle:

[4] Wikipedia, List of U.S. states and territories by income. Available at Wikipedia: https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_income

[5] United States Census Bureau, cb_2019_us_state_5m.zip. Available at: https://www2.census.gov/geo/tiger/GENZ2019/shp/cb_2019_us_state_5m.zip