MACS 30200
Initial Result Section
Fangfang Wan

In my research project, I aim to determine what factors can influence the Yelp ratings of business, and then test my models. Previously I planned to investigate this research question solely on restaurants, but it is hard to isolate restaurants from other businesses from this dataset (I may add this isolation process later if there is enough time), so I run linear regression on all businesses (restaurants, museums, etc.) included in the dataset. There are two sets of data that I have used. One is “Yelp Dataset” on Kaggle.com, which was uploaded by Yelp, Inc. I used this dataset to build 2 versions of OLS linear regression models. Another is a dataset that I scraped from top businesses in Chicago, and I used this dataset to test the linear regression models that I have built using the previous dataset.

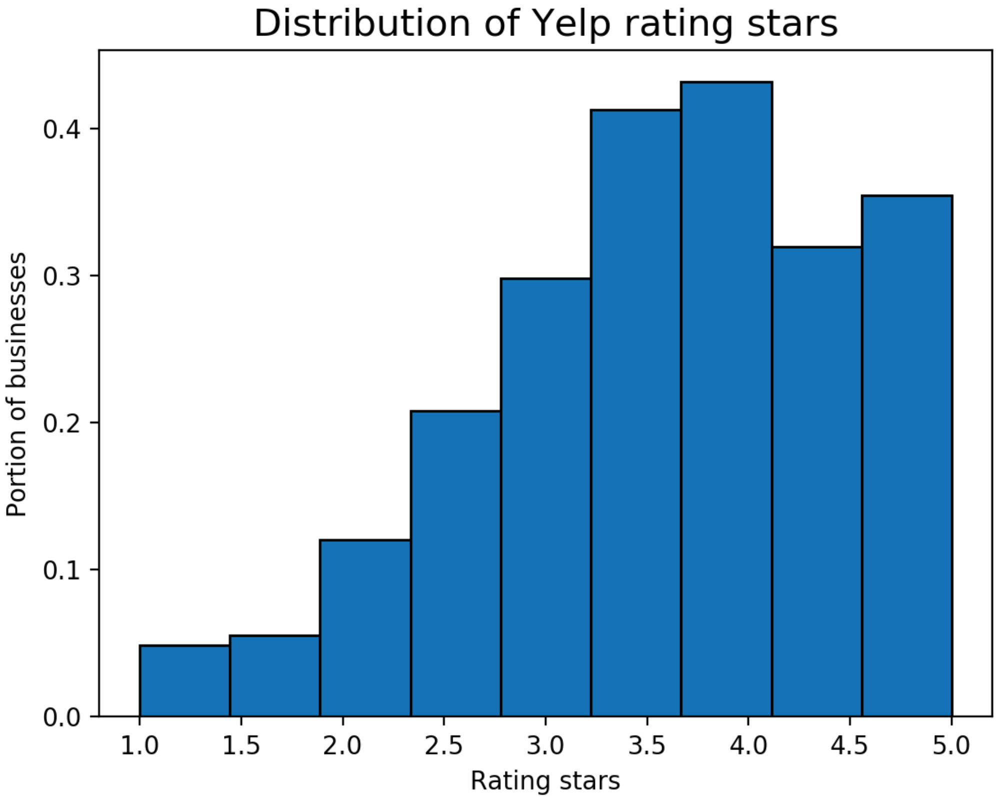
Section 1. Yelp, Inc. dataset from Kaggle.com

This dataset contains attributes and information of over 170 thousands businesses on Yelp.com, and it also includes review and user information. Below is summary statistics information about Yelp rating scores in this dataset. From Table 1 and Figure 1, we can see that most businesses have 3.5 or 4 stars on Yelp, which is consistent with our daily life experience when using Yelp.

Table 1: Number of businesses in each rating level

Yelp rating	Number of businesses
5.0	27540
4.5	24,796
4.0	33,492
3.5	32,038
3.0	23,142
2.5	16,148
2.0	9,320
1.5	4,303
1	3,788

Figure 1. Distribution of Yelp ratings



I originally planned to include information such as parking, noise level, ambience, etc. as independent variables, but data on those factors is highly incomplete, so I only included some of them in my linear regression model. I have run two versions of linear regression. In both versions, the dependent variable is aggregate Yelp rating stars of those businesses. In the first version, I have included opening hours throughout every of seven days of a week, review count, and the top 10 cities included in this dataset, including Las Vegas, Phoenix, Toronto, Charlotte, Scottsdale, Pittsburgh, Mesa, Montreal, Henderson, and Tempe. Below is a form of top 10 cities and the number of businesses included.

Table 2. Number of businesses in the most frequently occurring cities in Yelp, Inc. Dataset from Kaggle.com

City	Number of businesses included

Las Vegas	26775
Phoenix	17213
Toronto	17206
Charlotte	8553
Scottsdale	8228
Pittsburgh	6355
Mesa	5760
Montreal	5709
Henderson	4465
Tempe	4263

Chicago businesses are not included in this dataset. In the second version of my linear regression model, I have excluded city dummies in the previous version, and included all others. I have included the regression result tables of 2 versions of model below.

Model Version 1:

Table 3. OLS regression result of model version 1

OLS Regression Results

	coef	std err	p>|t|
constant	3.4410	0.004	0.000
mondayhrs	-0.0245	0.001	0.000
tuesdayhrs	0.0130	0.002	0.000
wednesdayhrs	0.0221	0.002	0.000
thursdayhrs	0.0123	0.002	0.000
fridayhrs	0.0150	0.001	0.000
saturdayhrs	-0.0085	0.001	0.000
sundayhrs	-0.0175	0.001	0.000
Las Vegas	0.1224	0.007	0.000

Phoenix	0.0660	0.008	0.000
Toronto	-0.0722	0.008	0.000
Charlotte	-0.0240	0.011	0.033
Scittsdale	0.3302	0.011	0.000
Pittsburgh	0.0563	0.013	0.000
Mesa	0.0158	0.014	0.243
Montreal	0.1654	0.014	0.000
Henderson	0.1772	0.015	0.000
Tempe	0.1262	0.016	0.000
review_count	0.0003	2.44e-05	0.000

Number of observations: 174567

R-squared: 0.040

Adjusted R-squared: 0.040

Model Version 2:

Table 4. OLS regression result of model version 2

OLS Regression Results

	coef	std err	p>|t|
constant	3.4831	0.003	0.000
mondayhrs	-0.0235	0.001	0.000
tuesdayhrs	0.0127	0.002	0.000
wednesdayhrs	0.0221	0.002	0.000
thursdayhrs	0.0122	0.002	0.000
fridayhrs	0.0164	0.001	0.000
saturdayhrs	-0.0099	0.001	0.000
sundayhrs	-0.0180	0.001	0.000
review_count	0.0004	2.42e-05	0.000

The first thing to point out is that the R-squared of both models are very low, so both models does not well explained heterogeneity of Yelp ratings of different businesses. This is because that business information from the Yelp, Inc. dataset from Kaggle.com is highly incomplete, and if I want to run regressions on more variables, it can result in a small sample, and therefore causing the problem of over-fitting and even problem in degrees of freedom. However, we can still detect certain patterns from such results.

From both models, we can see from the common independent variables that businesses with longer operating time on Saturday, Sunday and Monday tend to have lower Yelp ratings, while longer operating time on other days of a week is associated with higher Yelp rating. The reason behind such phenomenon is unclear, but my guess is that on weekends, there are generally more people visiting various kinds of businesses, while on weekdays when people are busy, those who visit business tend to have greater needs from these businesses, and they may also have higher expectation on them. Such phenomenon tend to extend to Mondays, while on Fridays people feel relaxed from a week's work and tend to visit places that they hold higher expectations. As for common independent variables of two models, businesses with more number of reviews tend to have slightly higher Yelp ratings.

In the first version of my linear regression model, I also included top 10 cities included in the dataset. Results indicate that businesses in Toronto and Charlotte tend to have lower Yelp ratings. The reason behind such phenomenon is not clear, and it is probably not because businesses in Toronto are of lower quality. Rather, there are many other possibilities. One is that people in Toronto and Charlotte tend to rate businesses more strictly than the other 9 cities. Another possible reason is that there might be that since the dataset cannot be considered as "big" (174,000 ~ businesses in total, 17,000 ~ for Toronto, and 8500 ~ for Charlotte), so it is highly possible that some randomness is included.

Section 2. Dataset scraped from Yelp businesses in Chicago

In this dataset, due to time limitation I have included top 200 businesses in Chicago (I may extend to top 1000 later if there is enough time). Below is a snapshot (information of the first five businesses, including "Girl & the Goat", "The purple pig", “Wildberry pancakes and café”, “Au Cheval” and Art Institute of Chicago) of my resulting data frame.

Table 5. Data frame scraped from Yelp pages of Chicago businesses

	stars	review_count	mondayhrs	tuesdayhrs	wednesdayhrs	thursdayhrs	fridayhrs	saturdayhrs	sundayhrs	const
0	4.5	7115	7.5	7.5	7.5	7.5	8.5	8.5	7.5	1
1	4.0	5848	12.5	12.5	12.5	12.5	12.5	12.5	12.5	1
2	4.5	4627	9	9	9	9	9	9	9	1
3	4.5	5017	15	15	15	15	15	15	15	1
4	4.5	1505	7.5	7.5	7.5	10.5	7.5	7.5	7.5	1

I fit this data to the version 2 model derived from the Yelp, Inc. dataset from Kaggle.com. I have calculated the Mean Squared Error, and concluded that the previous models are not accurate for predicting Yelp rating stars, but the prediction is not bad. The mean-squared error is 0.26, which means that on average, the difference between predicted and real stars is 0.5, which is a significant error, but not exceptionally high.

To improve my data analysis part, I will do part or all of the following if there is enough time, and they are listed according to my priority.

1. Scrape data from top businesses and restaurants from Yelp.com in Chicago, include more independent variables such as parking availability, noise level, price level and ambience, split them into training and testing sets, and see if including more independent variables and deriving an OLS linear regression model based on training set will fit data better.
2. Isolate restaurant from the Yelp, Inc. dataset from Kaggle.com and run 2 regression models again to see if results differ for restaurant.

3. Scrape data from top 1000 businesses and restaurants from Yelp.com in Chicago, and fit models on them respectively.