Zhiling Hu (zh347)
Irene Wei (zw575)
CS5304 Data Science in the Wild

Did diners make a better choice?

Introduction

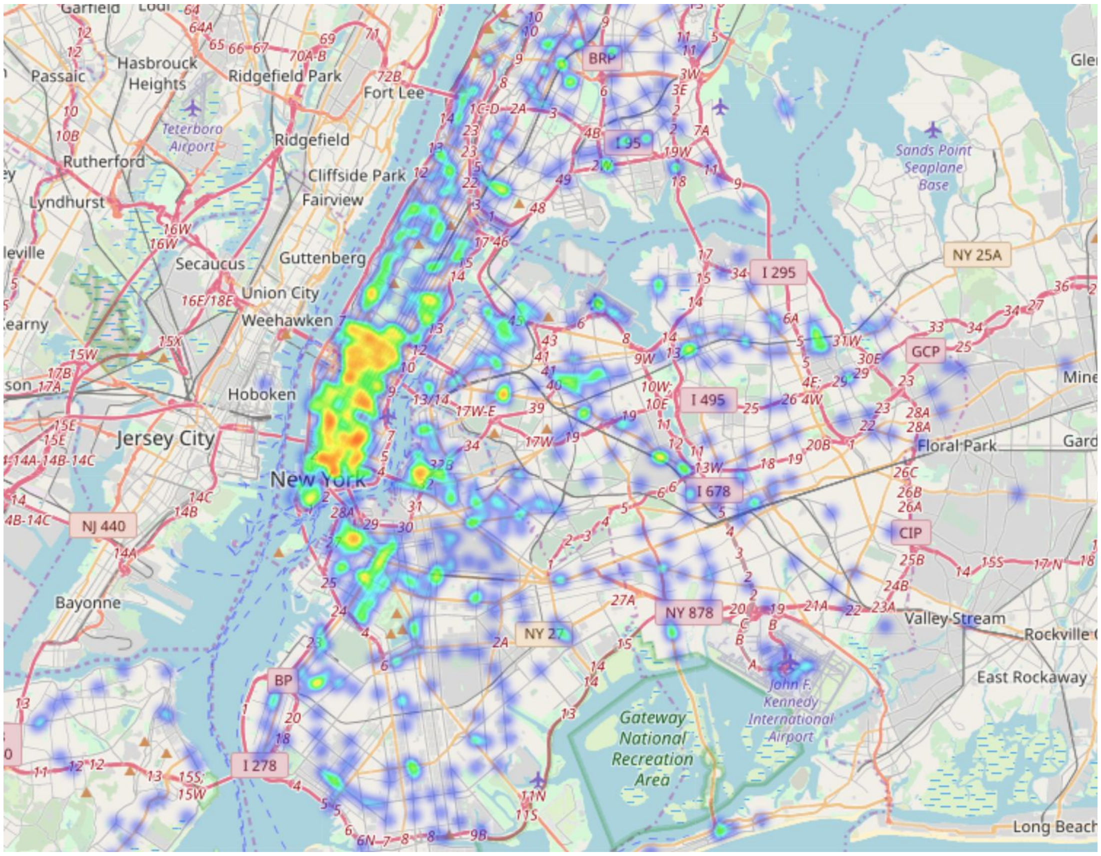
Since the summer of 2018, Yelp, the crowd source review forum has added health inspection score issued by government on their restaurants' webpages, trying to better inform the users. However, this feature is controversial, since many restaurants do not want the inspection grade affect their customers' decision making. Therefore, it is interesting to see given the health inspection, do diners make a better choice in selecting restaurants?

Data

- NYC Restaurant Inspection Results from NYC OpenData
- Yelp Fusion API

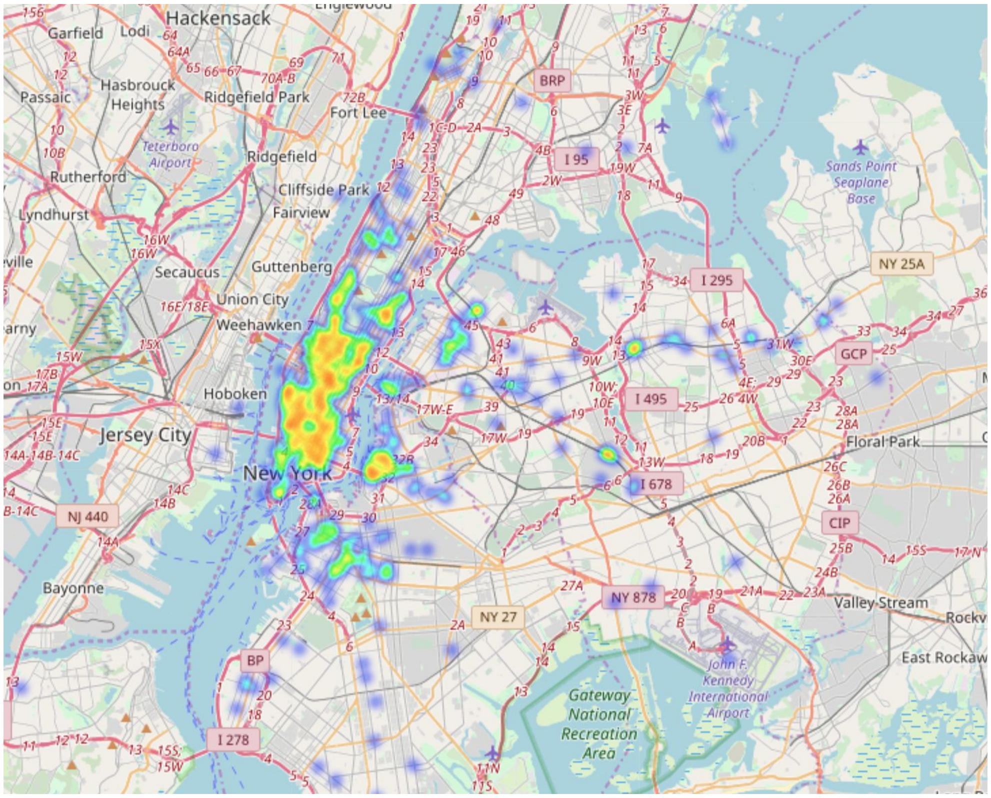
With the restaurants listed in the inspection dataset, we wrote a python script to query from Yelp Fusion API on each restaurant's location and name. We extracted information including their ratings, price levels, categories, and text reviews. The inspection dataset contains grade and score. We keep the most recent inspection score for each restaurant, which was all conducted in 2018 and 2019.

An overview of each dataset:
The inspection dataset contains the location of each restaurant in New York (Borough and Street), cuisine description, Grade, Violation Code, etc. To explore this dataset, we draw a heatmap of the clean restaurants in New York City (see Figure1). They are the restaurants with inspection grade A, which also do not have more than one violation and critical violation.

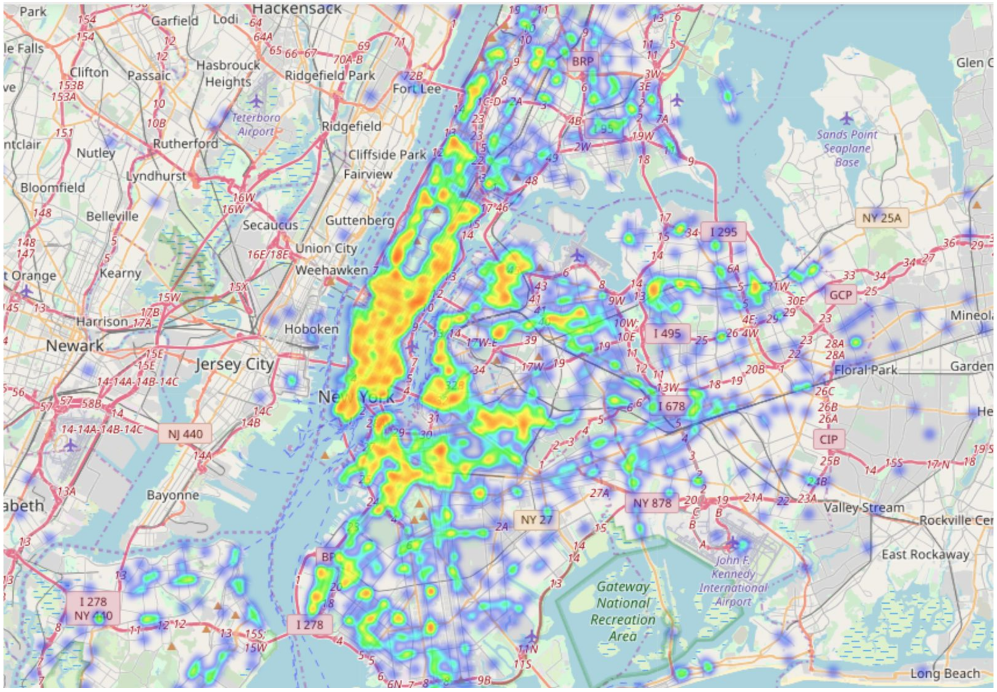


(Figure 1: Restaurants for Clean Eating)

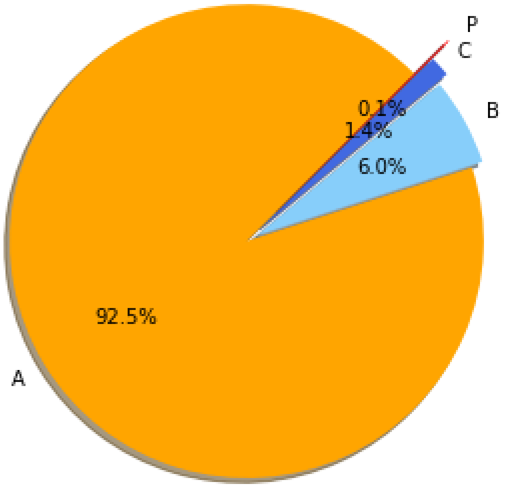
The Yelp restaurants data has their ratings, reviews, prices, check-ins etc. To see the distribution of popular restaurants and the top rated restaurants, we also draw heat maps Figure and Figure 3. We define popular restaurants as having check ins more than 500, and the top rated restaurants as having Yelp rating over 4.0.



(Figure 2: Popular Restaurants)



(Figure 3: The top rated restaurants)



(Figure 4: The distribution of restaurants in terms of major grades)

From the visualization, we could get a better sense of what these dataset look like, but it is hard to conclude any insights from the inspection scores and the popularity of NYC restaurants, so we need to analyze using models and statistical methods.

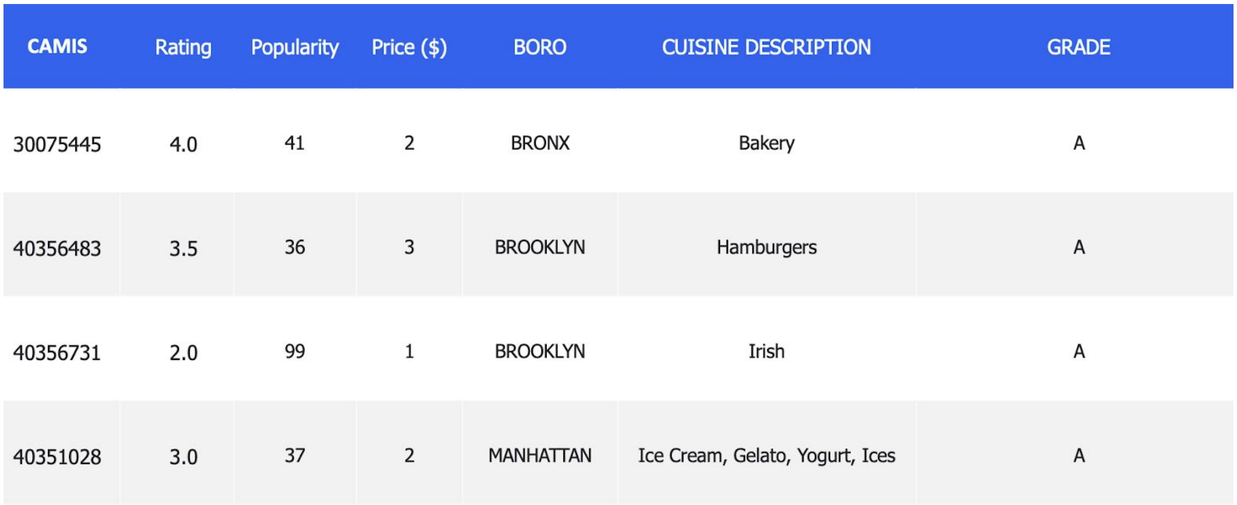
Tools

Our Python script for scraping from Yelp Fusion API used the requests library of Python, Requests and JSON library to parse the JSON response of restaurants data.

Our list with the libraries for scientific applications includes numpy, Statsmodels, Seaborn, pandas, sklearn, and matplotlib.

Data Processing and Modeling

After some simple manipulations and loading of the csv data into pandas DataFrame, we have the following example dataset, where rating, popularity, price, boro, cuisine description and grade on each restaurant are shown.



CAMIS	Rating	Popularity	Price ($)	BORO	CUISINE DESCRIPTION	GRADE
30075445	4.0	41	2	BRONX	Bakery	A
40356483	3.5	36	3	BROOKLYN	Hamburgers	A
40356731	2.0	99	1	BROOKLYN	Irish	A
40351028	3.0	37	2	MANHATTAN	Ice Cream, Gelato, Yogurt, Ices	A

(Figure 5: Example columns of entries in merged dataframe)

Then we started investigations on missing values. There are around 1,000 rows containing NA entries in the dataset of over 20,000 entries. The NA entries fall under grade or rating, which we assume to be hard to predict based on other rows. We decided to drop them.

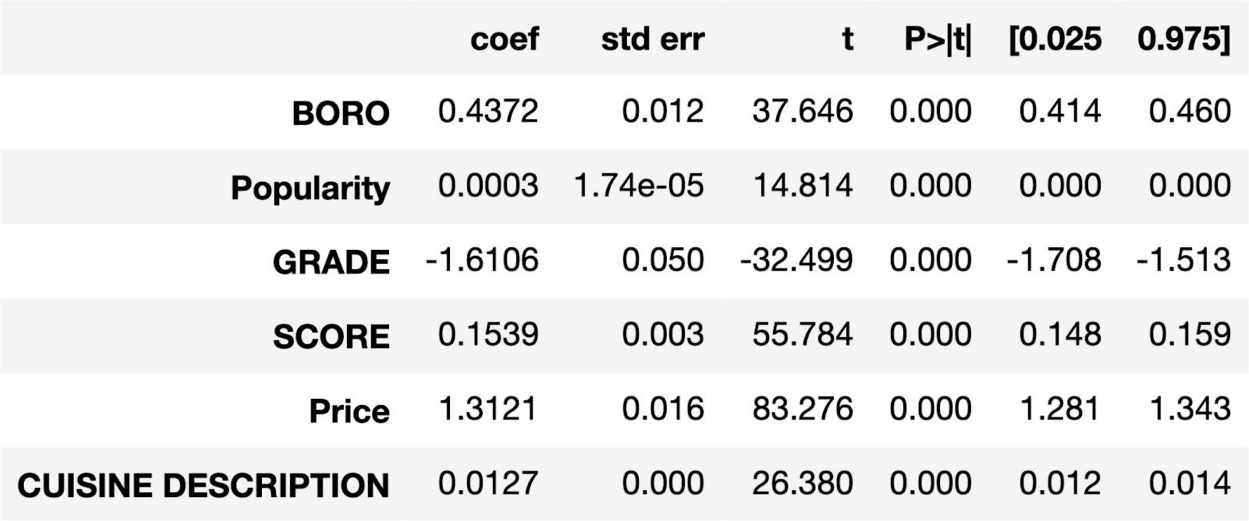
As most of machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers. Thus we used the encoders from the scikit-learn library. Specifically, the LabelEncoder of creating an integer encoding of labels and the OneHotEncoder for creating a one hot encoding of integer encoded values. The list includes “CUISINE DESCRIPTION”, “BORO”, “GRADE.”

Next we’ll want to fit a linear regression model.

As we pulled out the most updated Yelp ratings, there is no possibility to create a comparison between the data dated before adding inspection score and the data dated after. We decided to move forward with creating two models with different sets of variables, one containing grade while the other one’s not.

We chose variables that we think we’ll be good predictors for the predicted variable, “rating.”

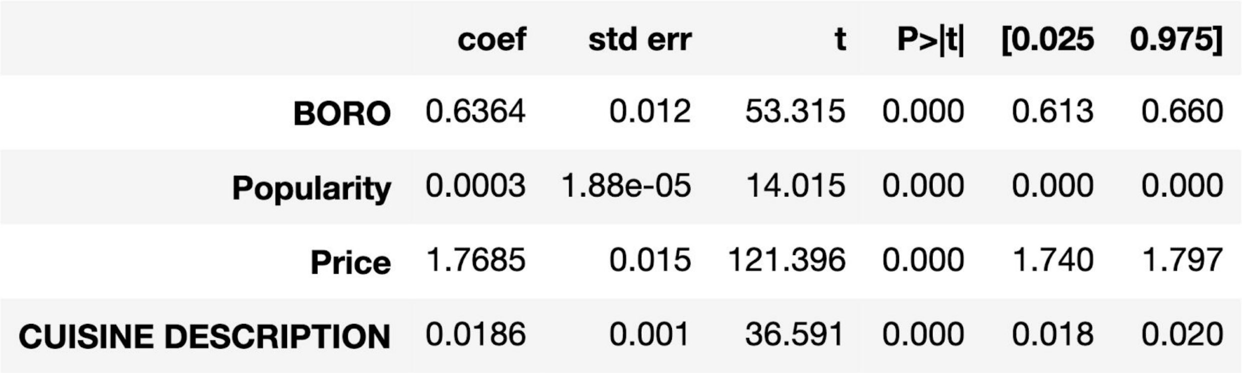
We got our first table for the model with inspection grade and score. This model has a much higher R-squared value — 0.901, meaning that this model explains 90.1% of the variance in our dependent variables. We can see that all the variables are statistically significant in predicting the rating.



	coef	std err	t	P>|t|	[0.025	0.975]
BORO	0.4372	0.012	37.646	0.000	0.414	0.460
Popularity	0.0003	1.74e-05	14.814	0.000	0.000	0.000
GRADE	-1.6106	0.050	-32.499	0.000	-1.708	-1.513
SCORE	0.1539	0.003	55.784	0.000	0.148	0.159
Price	1.3121	0.016	83.276	0.000	1.281	1.343
CUISINE DESCRIPTION	0.0127	0.000	26.380	0.000	0.012	0.014

(Figure 6: Results for Linear Regression model w/ Inspection results)

The second table is for the model without inspection grade and score.



	coef	std err	t	P>|t|	[0.025	0.975]
BORO	0.6364	0.012	53.315	0.000	0.613	0.660
Popularity	0.0003	1.88e-05	14.015	0.000	0.000	0.000
Price	1.7685	0.015	121.396	0.000	1.740	1.797
CUISINE DESCRIPTION	0.0186	0.001	36.591	0.000	0.018	0.020

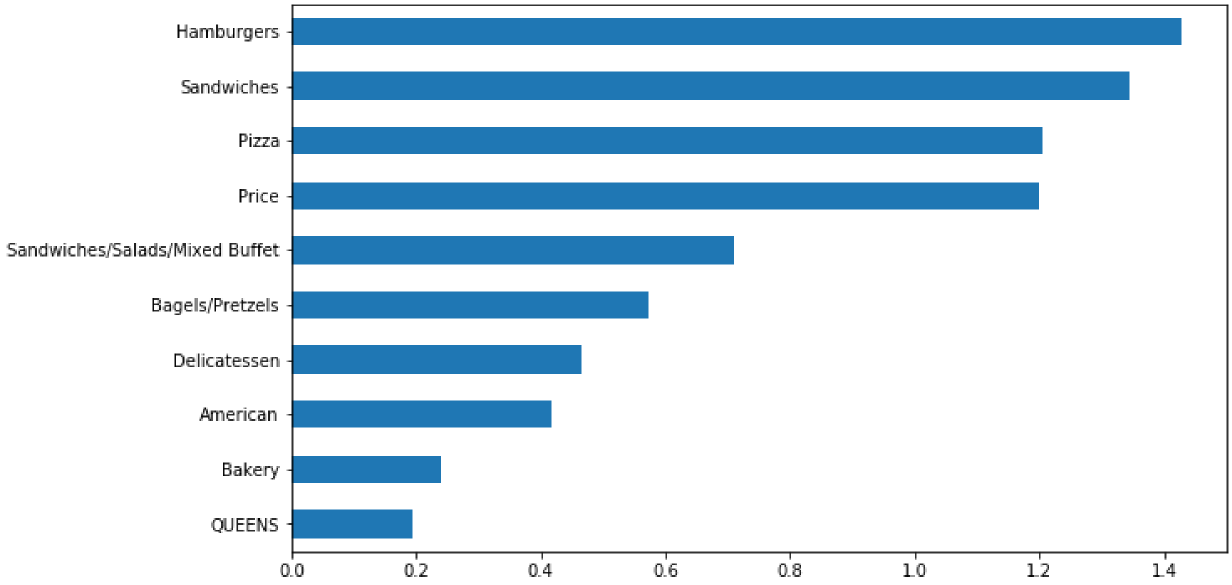
(Figure 7: Results for Linear Regression model w/o Inspection results)

This model also features a high R-squared value — 0.884, slightly lower than 0.901. This doesn't show a strong evidence that inspection grade and score will influence the results with other variables provided together. So we move forward to build up predictive models, using Logistic Regression.

After test train splitting, we trained models on the two datasets we mentioned above. The one with inspections scores achieved 63%. The other one achieved 65%. The slight different couldn't help us gain new insights into our problem.

Reflections

We re-evaluated the whole project. First of all, the "inspection grade" listed on the Yelp interface is not easily recognizable. So it might be possible that few people realize there is a grade variable to be considered. Besides, we looked at the top important features in predicting the rating. They are mainly under food categories. Some are location-wise. None of the grade show up on the list.



(Figure 8: Top 10 features for Logistic Regression model w/ Inspection results)

During the preprocessing, the grade contains some variables that are either not explainable or hard to be categorized. We kept the most we can, but also dropped some of them since there were not many entries associated with. This also could influence the results.

<u>List of code:</u>
Yelp_API_Data_Matching.ipynb - scraper for Yelp Fusion API
Preprocessing.ipynb
Merging.ipynb
Model.ipynb