A preliminary analysis of Foursquare ratings of sushi restaurants at New York City

Hugo Tadashi

April 2020

Introduction

- Sushi restaurants industry has a market size (measured by revenue) of $22.1 billion.

Photo by `https://www.flickr.com/photos/chidorian/`

Introduction

- **Problem:** Some features obtained by the Foursquare site can be used to predict the Foursquare rating of sushi restaurants in New York City? Which are more important to increase the rating?
- **Motivation:** Useful information for decision making to owners and prospective owners of sushi restaurants.

Data acquisition

- All the data were extracted from the Foursquare site and retrieved by using Foursquare Places API.[1]

[1] `https://developer.foursquare.com/places`

Data description

Data	Description
Latitude	Latitude of restaurant.
Longitude	Longitude of restaurant.
Price tier	An integer from 1 (least pricey) to 4 (most pricey).
Reservations	1 if restaurant has reservations, 0 otherwise
Credit cards	1 if restaurant accepts credit cards, 0 otherwise
Outdoor seats	1 if restaurant has outdoor seats, 0 otherwise
Delivery	1 if restaurant has delivery service, 0 otherwise
Created at	When restaurant page was created
Likes	Number of users who have liked this restaurant.
Number of photos	Number of photos of this restaurant.
Rating	Rating of the restaurant (0 through 10).

Methodology

- The dataset was obtained by using the Python request package to communicate with Foursquare API.
- 282 different sushi restaurants were retrieved.

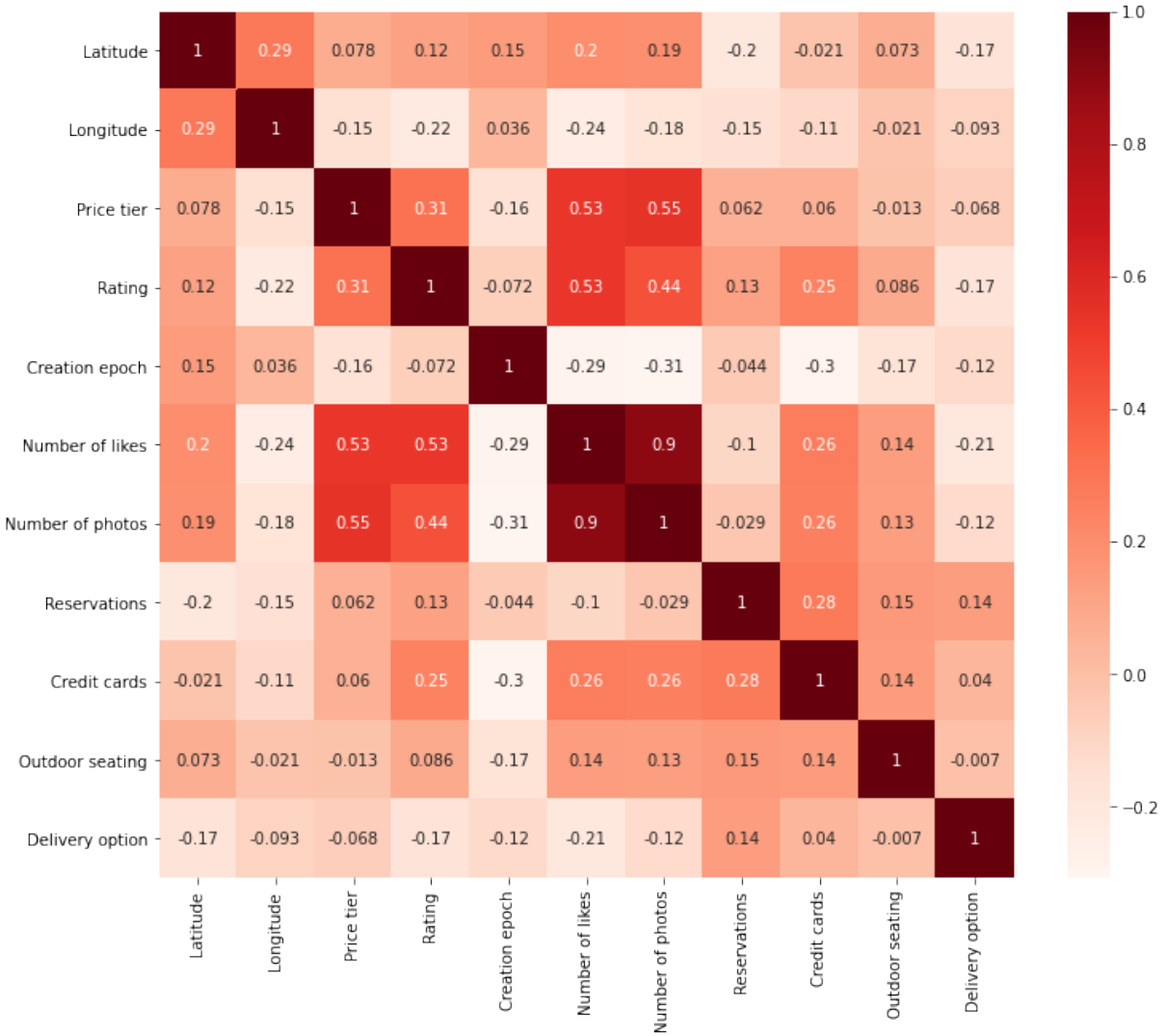
Methodology

Unfortunately, the data was not complete:

- 100 restaurants did not have rating and/or price tier information
- 125 restaurants did not have information about credit card acceptance
- 143 restaurants did not have information if accept reservations
- 131 restaurants did not have information if outdoor seats were available
- 128 restaurants did not have information if a delivery service was available

Exploratory data analysis

Correlation



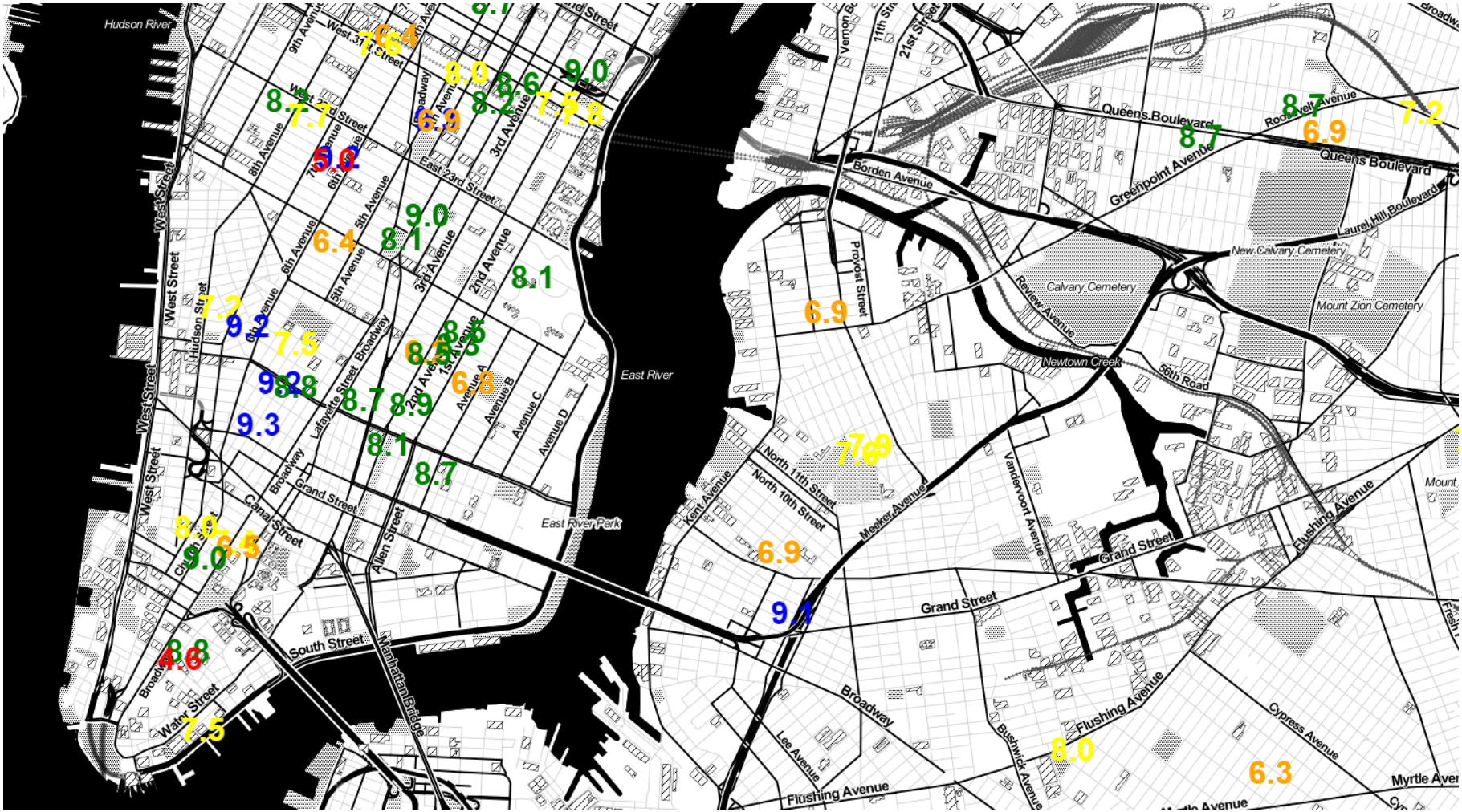
Exploratory data analysis

Correlation

- Most promising features for predicting the rating are the **number of likes** and the **number of photos** in Foursquare.
- **Creation epoch**, **the delivery option** and the **longitude** are unpropitious as features.

Exploratory data analysis

Map plot



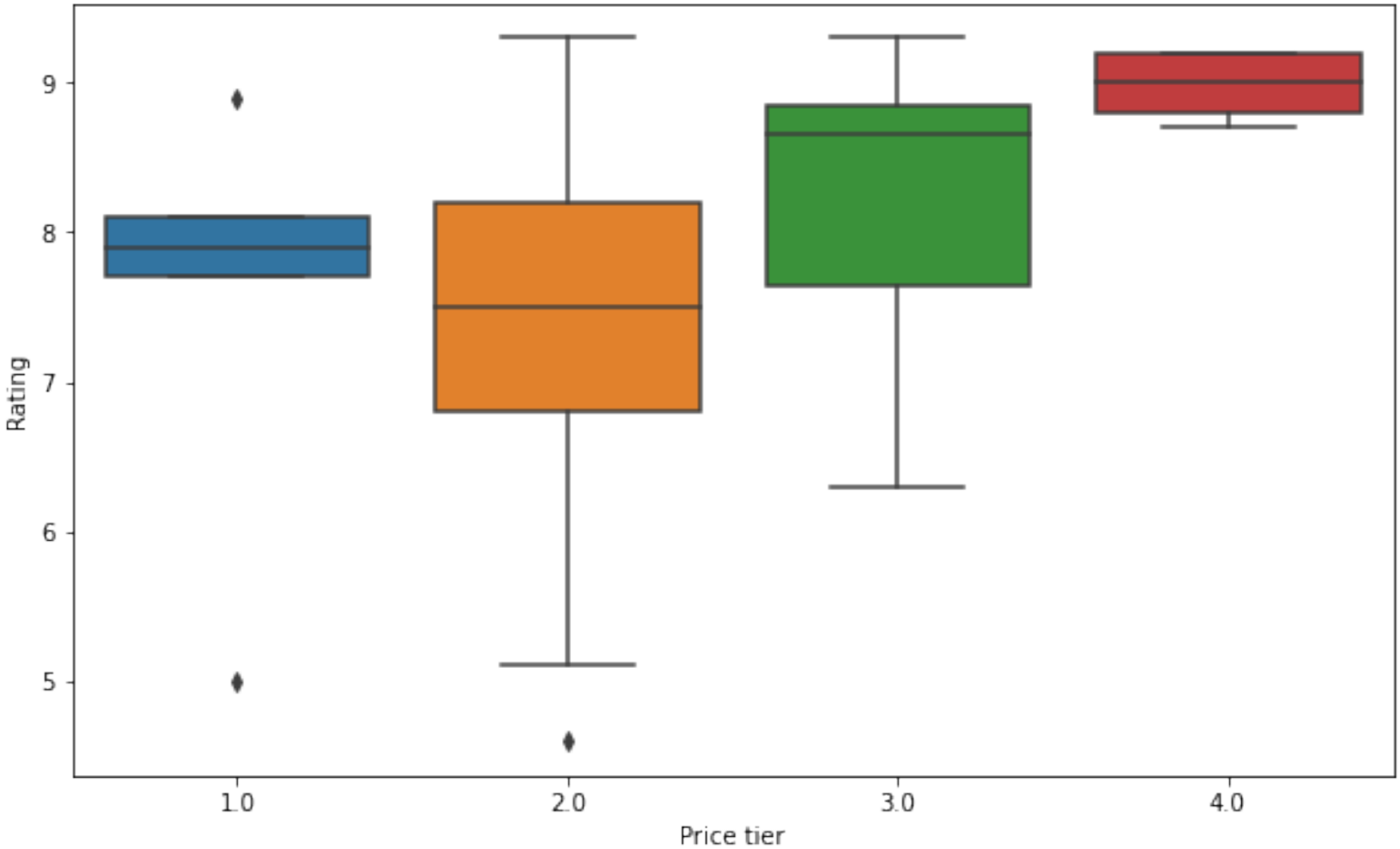
Exploratory data analysis

Map plot

- No obvious relation between the rating of the restaurant and its location.
- Surprising because it was expected that the location of the restaurant would influence the price tier and consequently, influence the rating of the restaurant.

Exploratory data analysis

Price tier



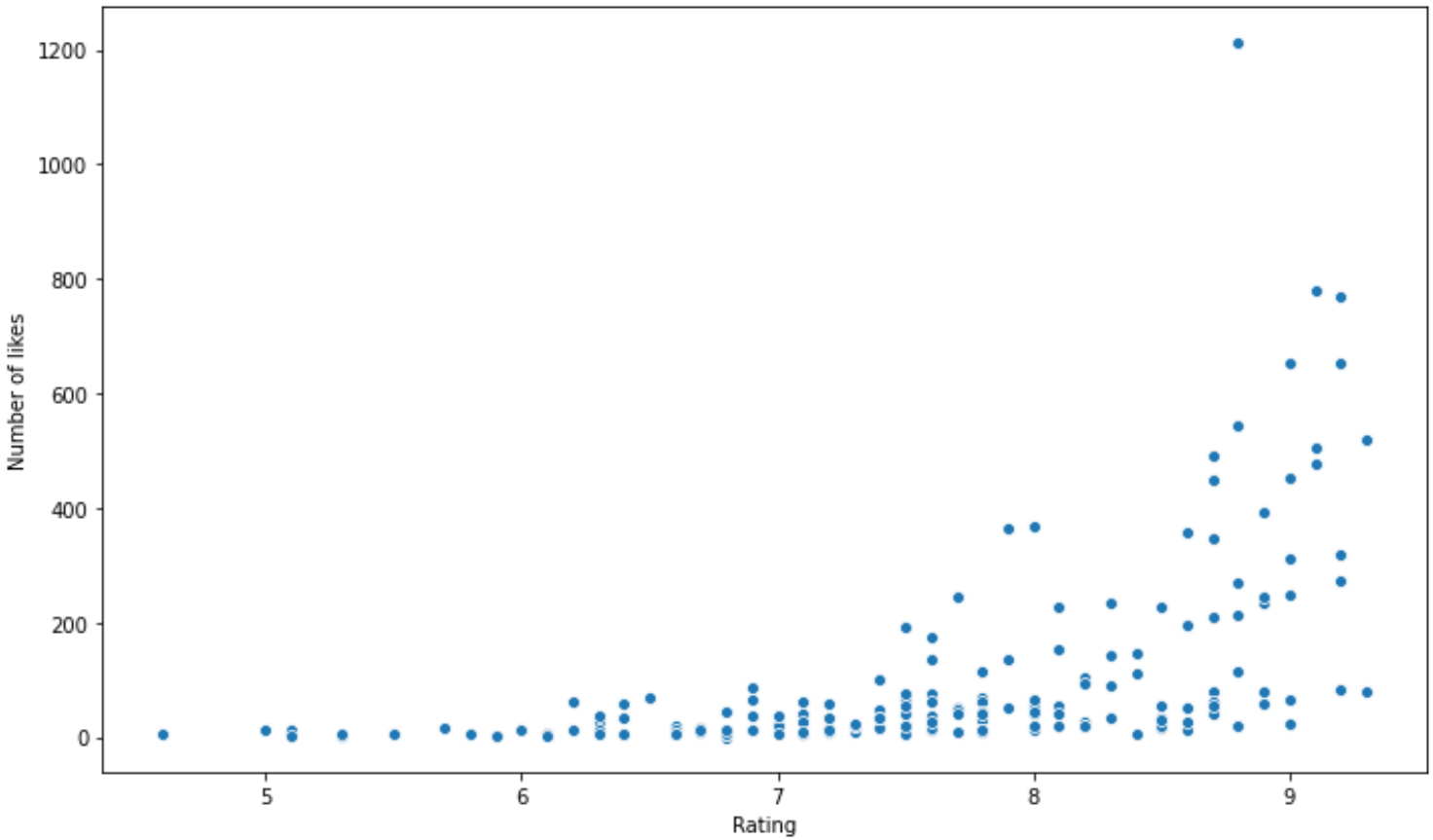
Exploratory data analysis

Price tier

- As expected, the restaurants with the highest price tier (4.0) have a consistent high rating, between 8.7 and 9.2.
- Great variation of ratings for restaurants with price tier 2.0.
- Doing a checkup on the dataset it was found that **84.07**% of the restaurants have price tier 2.0. The data obtained was too much unbalanced to be useful.

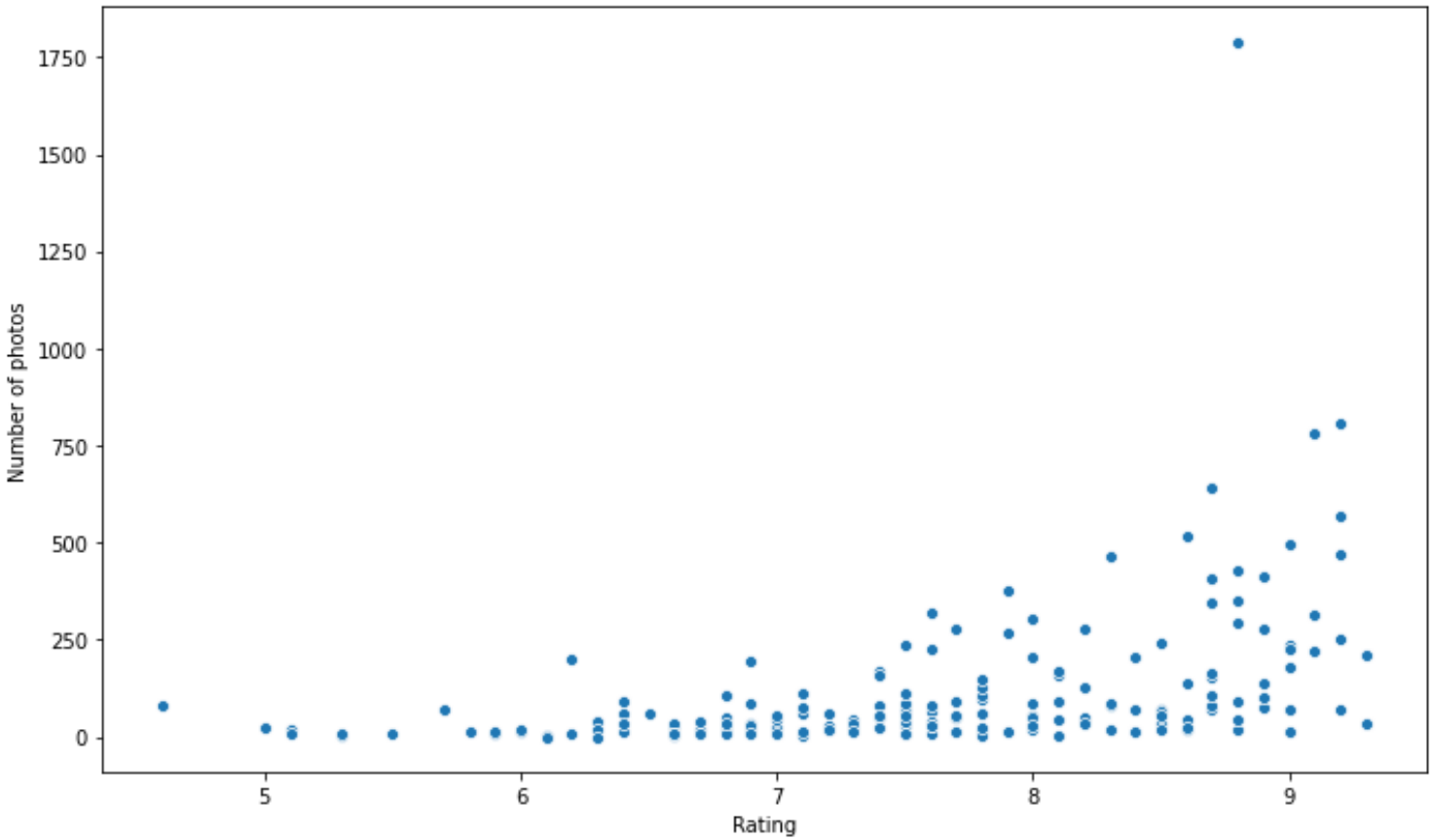
Exploratory data analysis

Number of likes



Exploratory data analysis

Number of photos



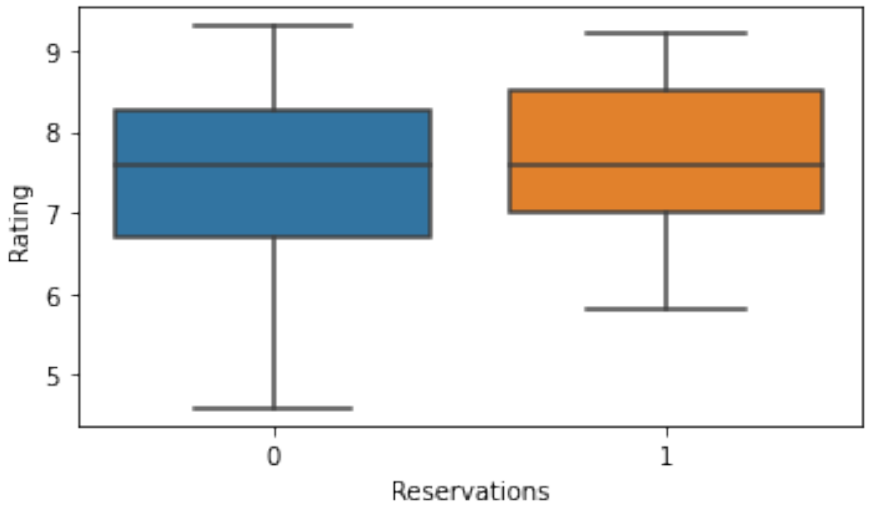
Exploratory data analysis

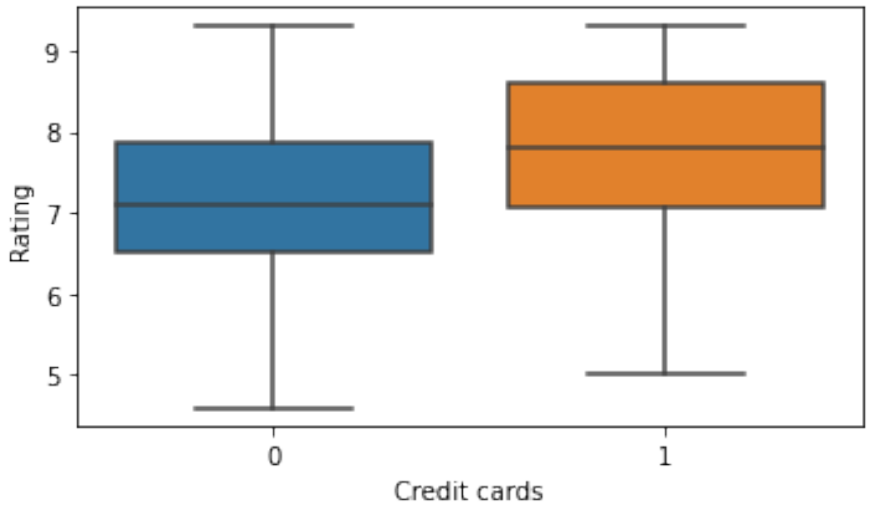
Number of likes and photos

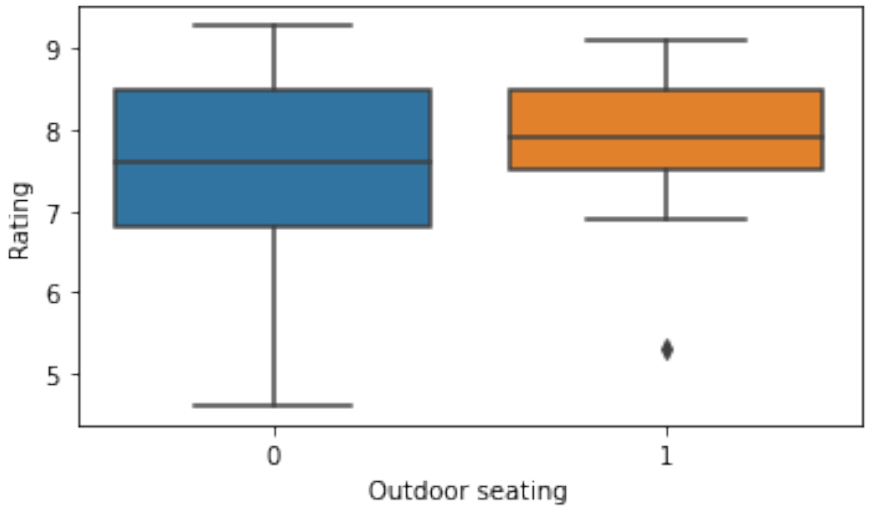
- Good features to develop a regression model
- Restaurants with a few number of likes and/or photos have a broad spectrum of ratings
- Less than 100 likes or 100 photos were considered as noise
- **Only 39 restaurants** after filtering noise

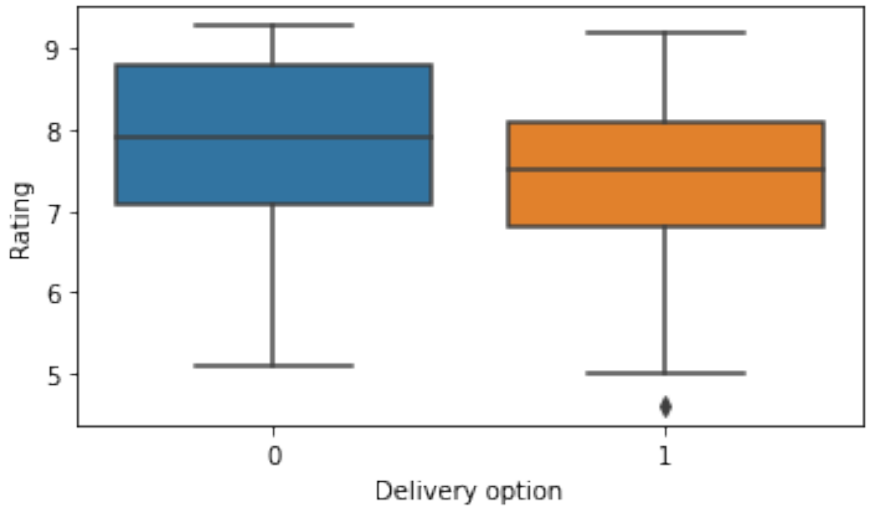
Exploratory data analysis

Other categorical variables









Exploratory data analysis

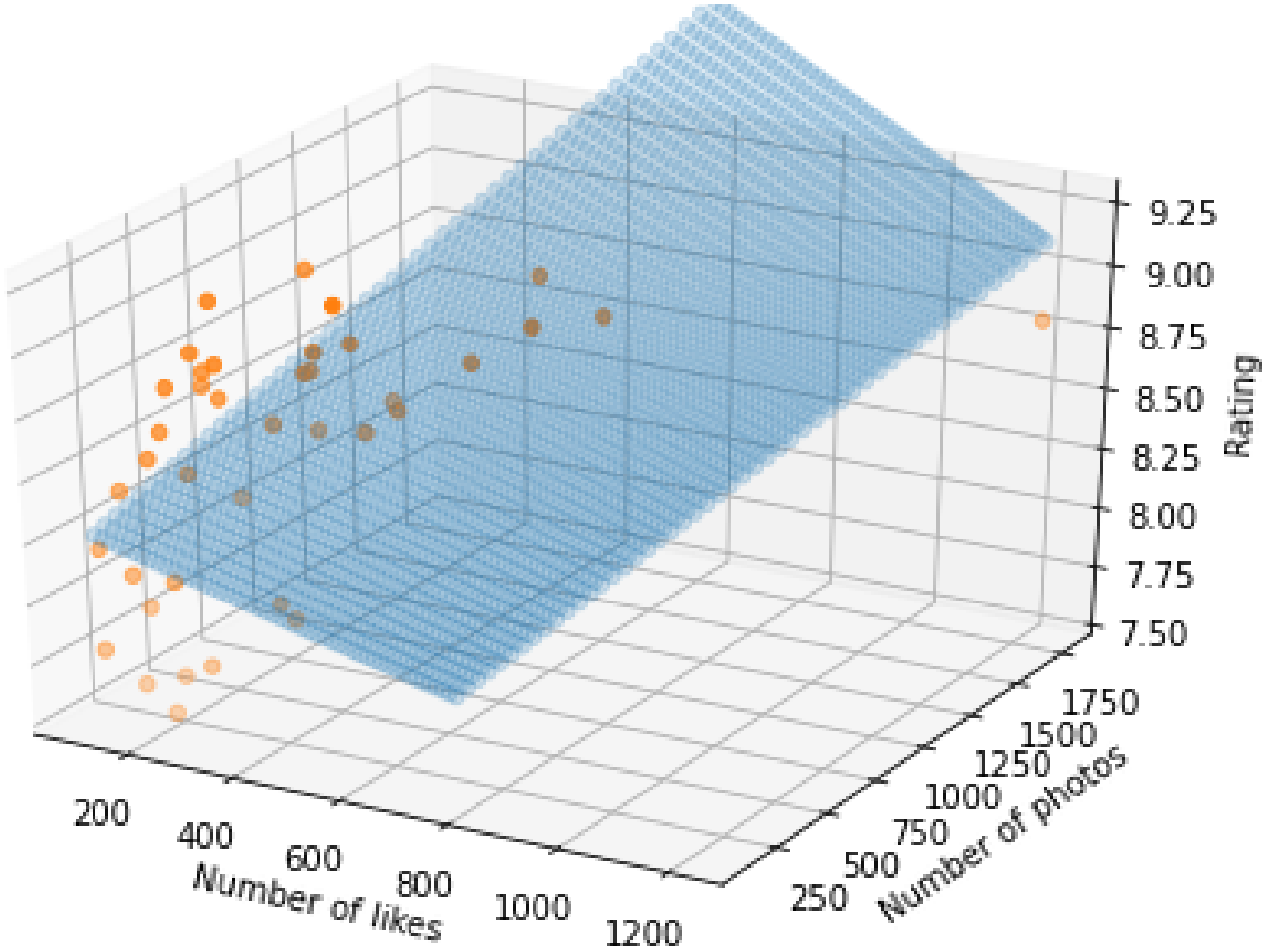
Other categorical variables

- Restaurants with lower ratings does not accept reservations nor have outdoor seating. A possible explanation for this is that these services can satisfy some costumers and increase their perception of the rating of the restaurant.
- No obvious relation between the availability of any of these services and the restaurant rating in the extracted data, thus these features were not used in the model.

Regression model

- The proposed model for prediction of the rating based on the features of the number of likes and the number of photos is a linear model fitted by Ridge regression, which is a variant of ordinary least squares where the cost function has an additional regularization term with weight α.
- Since the final number of samples was small (39 restaurants), the Scikit learn default split of 75% for the training set and 25% for the testing set was not used. Instead, it was decided to split the data such that 90% of these restaurants were used to train a model and the remaining 10% to test and evaluate this model.

Regression model



- Obtained R^2 score of 0.78

Conclusion

- This project proposed the analysis of the Foursquare ratings of sushi restaurants located at New York City based on the other features available on Foursquare database.
- Another interesting future direction is to add more details to the credit cards accepted (e.g.: only Visa, only Visa and Mastercard, etc.) to see if the brand of credit card influences on the rating.
- Sentiment analysis algorithms could be used to automatically classify the reviews in Foursquare as positive and negative reviews, giving an interesting feature to add to the model.
- Recent works on deep neural networks could help to classify the goodness of Foursquare photos, and the goodness of these photos could also be another interesting feature to analyze.