Using Foursquare data to predict restaurant ratings in Boston, MA

By McKenzie Mandich

Introduction

The goal of this project is to analyze restaurants in Boston, MA for stakeholders interested in opening a restaurant. The restaurant industry in North America is extremely competitive, and it is difficult to become successful as a small business owner in this industry.

Understanding what type of restaurant to open and where is critical to success. Furthermore, while finding the optimal location for a restaurant is just the first step, it is complex. For example, a neighbourhood with no Indian restaurants could be an excellent location for a new Indian restaurant because there would be little competition. However, it could also indicate that Indian food does not appeal to restaurant-goers in this neighbourhood.

Business Problem

In order to help stakeholders find an appropriate location for their business venture, we have created a machine learning model that considers the location (as lat and lon), price tier, tip count, likes, and genre of cuisine as predictors for its rating (with rating as a proxy to success).

This project deliberately does not limit itself to stakeholders interested in opening a particular type of restaurant. Because choosing the right location takes into account a complex and subjective balance of factors, there is no one 'perfect' location for all stakeholders. The goal of this project is therefore to detect and analyze geographic trends in Manhattan restaurants so that stakeholders can decide exactly how they want to balance between 'fitting in' and 'standing out'.

Data Selection

Data from foursquare:

- price tier
- genre of cuisine (venue category)
- name of restaurant
- restaurant rating
- tip count
- likes
- location (lat and long)

The number of restaurants per category was limited to 10. This reduces the bias towards a certain category during ML algorithm development. For example, if 400 out of the 500 restaurants were fast food, this would bias the algorithm towards that restaurant category.

'Restaurant' is be defined as any venue that primarily serves prepared meals. Bakeries, cafes, bars, and food shops are not included in the definition.

Data Methodology

- All data from foursquare
- Pulled multiple datasets from foursquare including venue category names/ids, venue names and categories, and venue details
- Some calls were premium, which has limits on quantity, so data collection took several days

Sample of Final Table

	venue_name	tipCount	pricetier	likes	rating	lat	lon	category
0	Ariana Restaurant	20	2.0	38	8.3	42.362593	-71.138953	Afghan Restaurant
1	Helmand Restaurant	71	2.0	137	8.4	42.366529	-71.078079	Afghan Restaurant
4	The Blue Nile	22	2.0	38	8.7	42.322086	-71.109521	African Restaurant
5	Ideal Sub Shop	2	2.0	10	8.0	42.322339	-71.072645	African Restaurant
6	Teranga	22	2.0	22	7.7	42.336642	-71.076827	African Restaurant

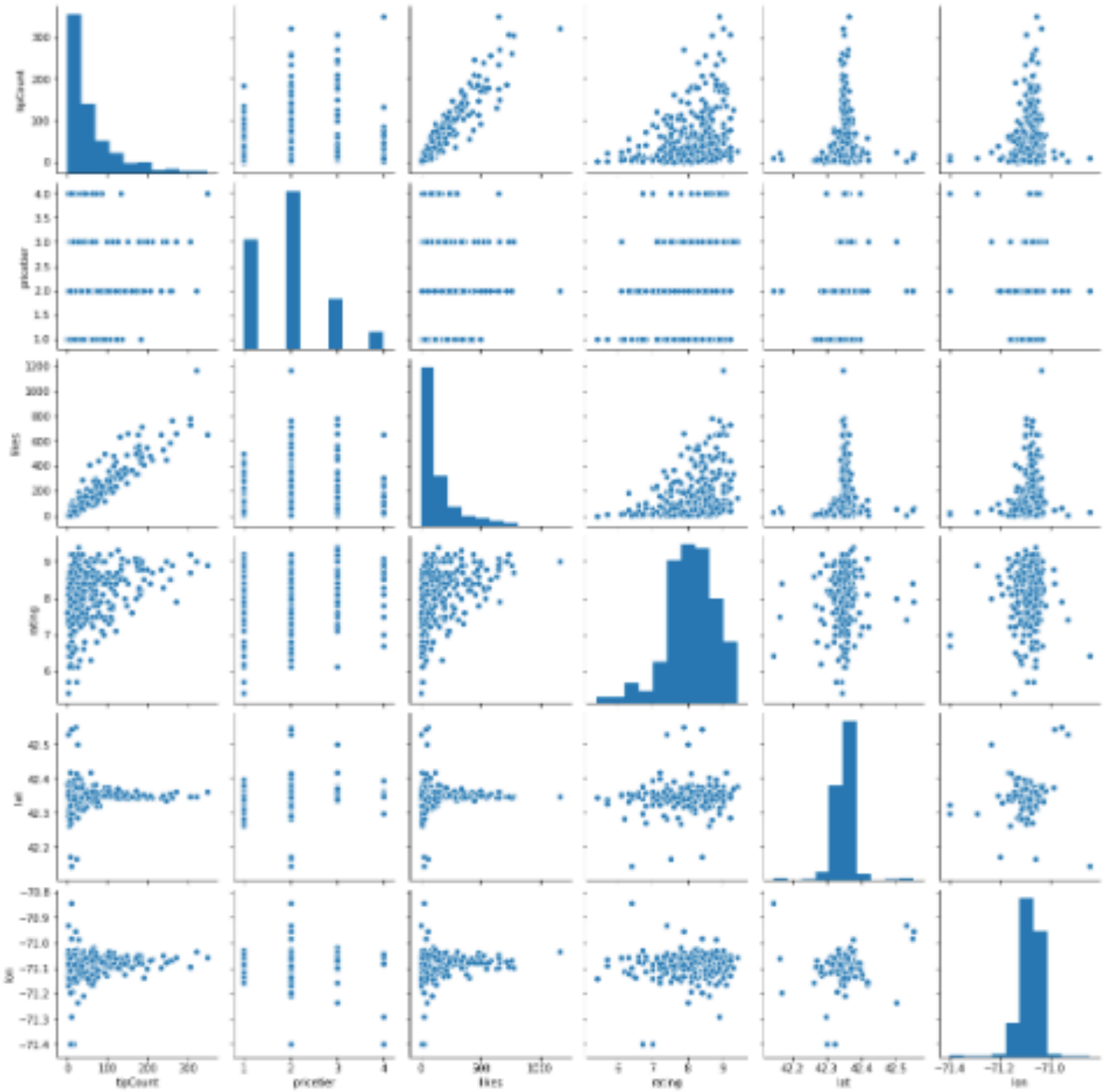
Exploratory Data Analysis

Summary Statistics for Numerical Variables

	tipCount	pricetier	likes	rating	lat	lon
count	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000
mean	54.100840	1.932773	140.445378	8.090196	42.349145	-71.080535
std	61.165377	0.828540	167.472169	0.702115	0.032583	0.045064
min	0.000000	1.000000	0.000000	5.400000	42.145341	-71.399306
25%	11.000000	1.000000	29.000000	7.700000	42.342649	-71.097536
50%	32.000000	2.000000	80.000000	8.200000	42.350013	-71.072997
75%	71.000000	2.000000	192.000000	8.600000	42.357752	-71.057719
max	350.000000	4.000000	1165.000000	9.400000	42.549417	-70.842051

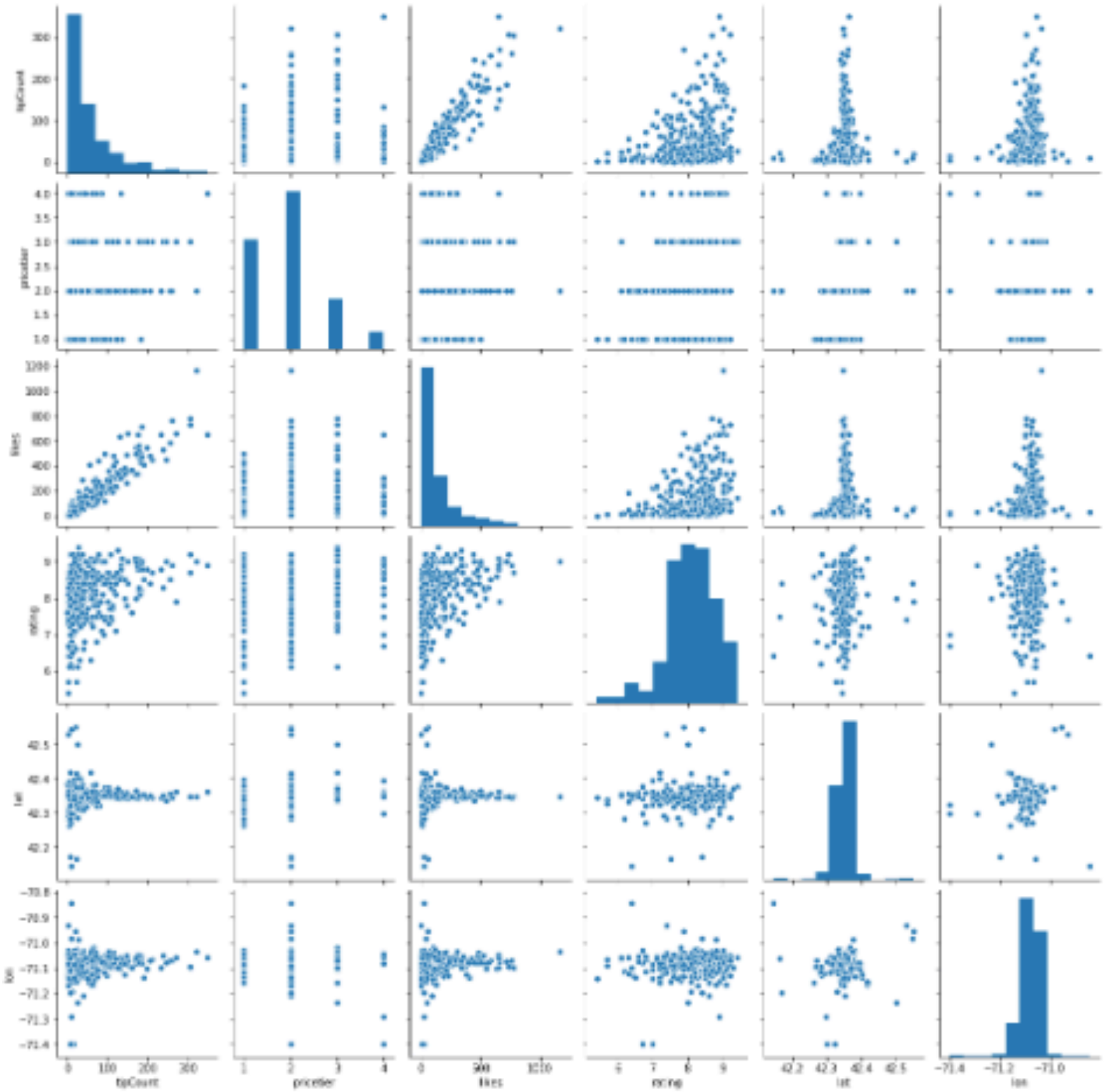
Comparison Using Seaborn's pairplot

(A bit overwhelming!)

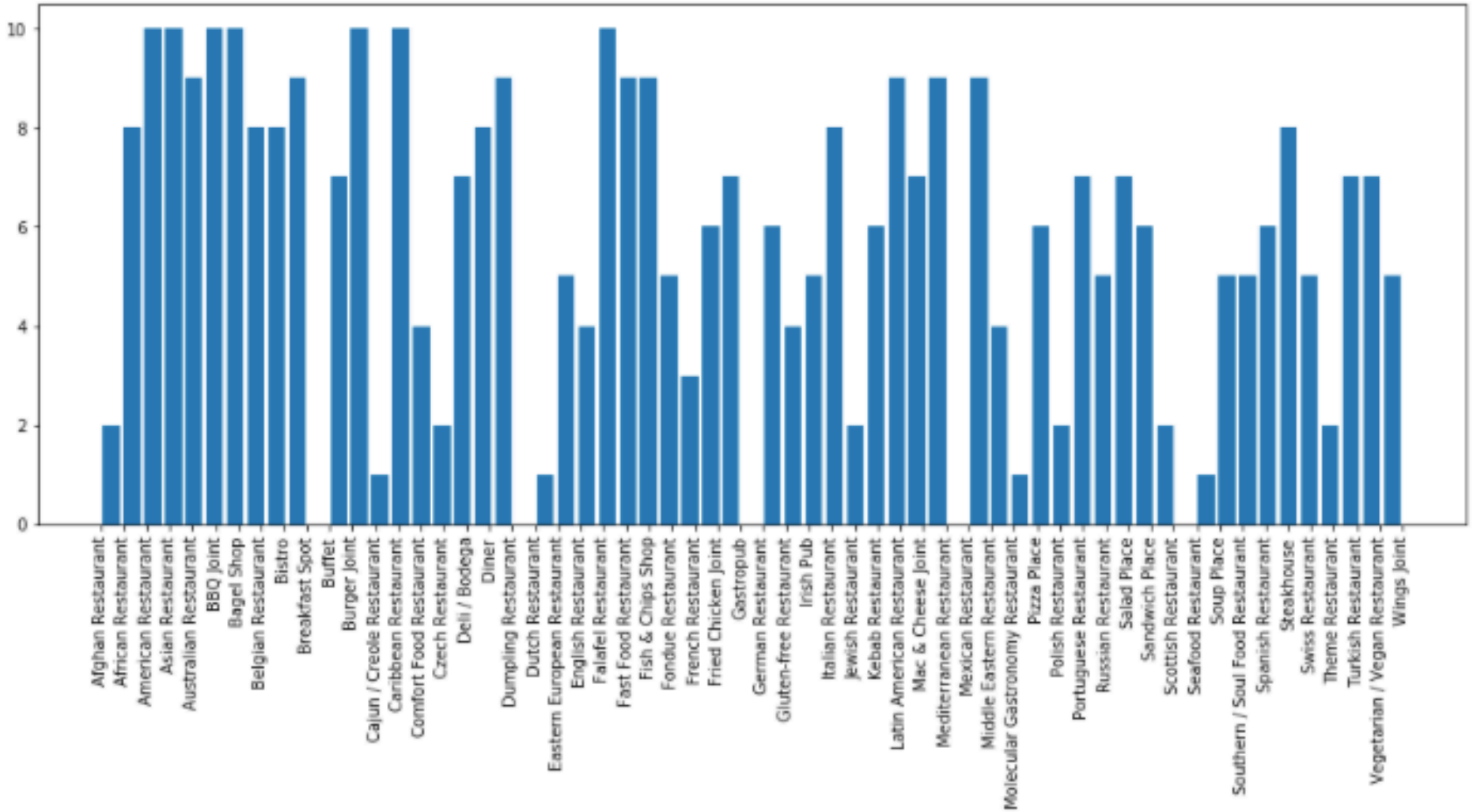


Some details to Note:

- Tip count and likes are the most strongly correlated variables
- Restaurants with lower price tiers are more common
- Latitude and longitude show very little correlation with other variables
- biased correlation between tip count/likes and rating: no restaurants with high tip counts/likes but low ratings, but many with low tip counts/likes but high ratings



Distribution of Restaurants by Category



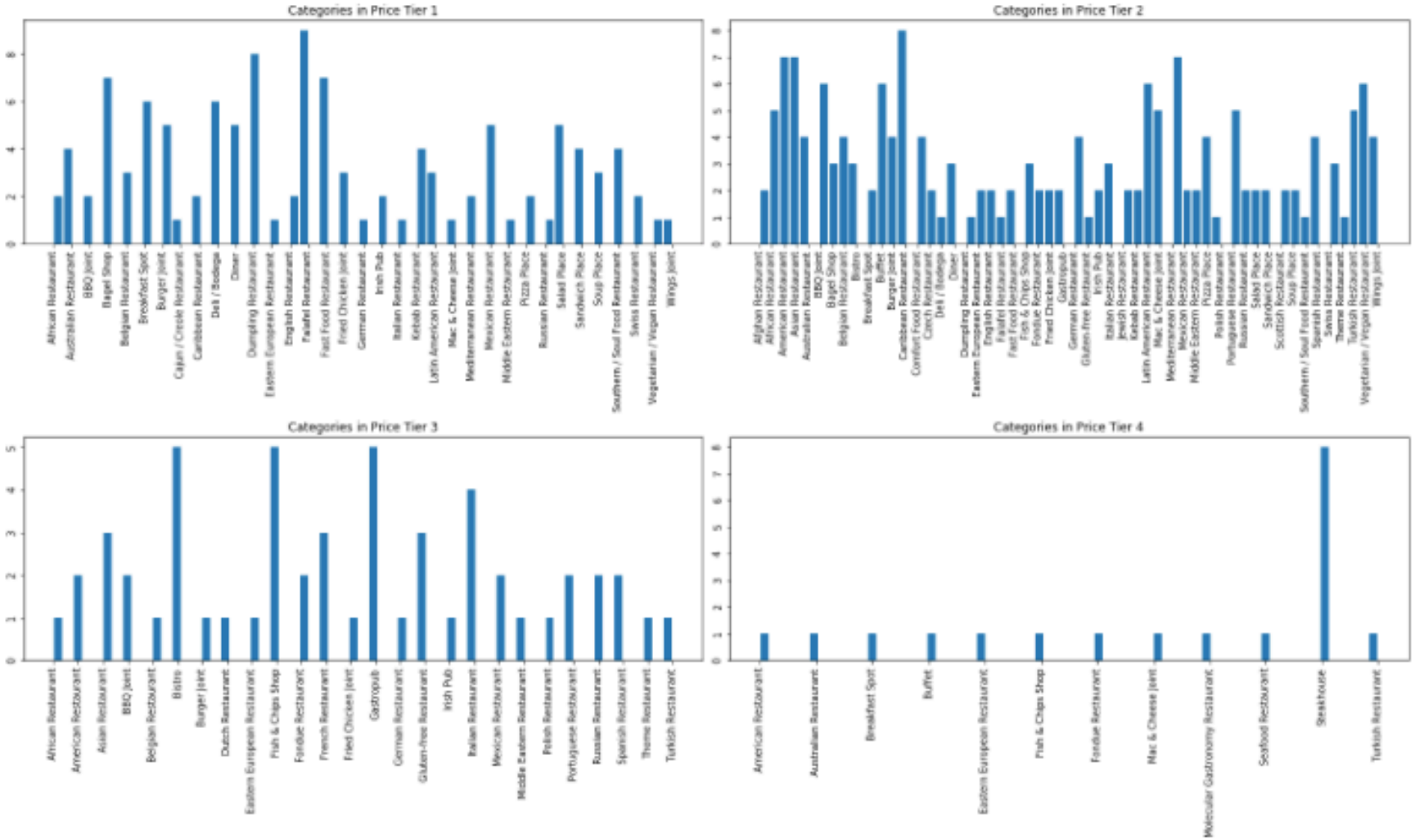
Distribution by Category and Price Tier

Most common price tier:2

Most common category in highest price tier: Steakhouse

Most common category in lowest price tier:: bagel shops, falafel restaurants, and dumpling restaurants.

Most unusual category in highest price tier (4): Mac & Cheese place!



Inferential Statistical Testing

- One way ANOVA test for restaurant rating by category (to determine that sub-populations are different)
  - statistic: 4.384602586305413
  - p-value: 2.4104321563429 97e-17

- Shapiro test to determine normality of ratings
  - statistic: 0.9505376815795898
  - p-value: 0.9505376815795898

Building A Linear Regression Algorithm

1. Use one hot encoding to transform venue categories into numerical values.
2. Split data into predictor variables and what we want to predict (the rating) .
3. Split data into train (70%) and test (30%) subsets.
4. Define and fit a model. A linear regression model was chosen because it is simple an robust.

Comparison of Algorithm Results to Data

Actual mean of population:8.09
Actual std of population:0.702

Encoding(mean) :8.12
Encoding(std) :0.564

Not bad!

Conclusions

- Overall, the model is a good first step in constructing a predictive algorithm for restaurant success based on data available from foursquare
- a full scale project of this nature would involve many more interactions and more in depth data collection.
- However, even with our limited data and simple analysis, we were able to construct an algorithm that correlated well with testing data.

What Next?

- Reassess whether rating is the optimal indicator to predict restaurant success, as we saw during the data analysis step that few restaurants had low ratings, and many restaurants with high ratings had a low number of likes/tips
  - Likes or tip counts could be considered as an alternative indicator

- categorize locations into neighbourhoods based on their latitude and longitude
  - By transforming location into a categorical variable then using one hot encoding to include this in the analysis, the importance of location choice for restaurant success may become more apparent.