Yelp Data Analysis

Scott Small

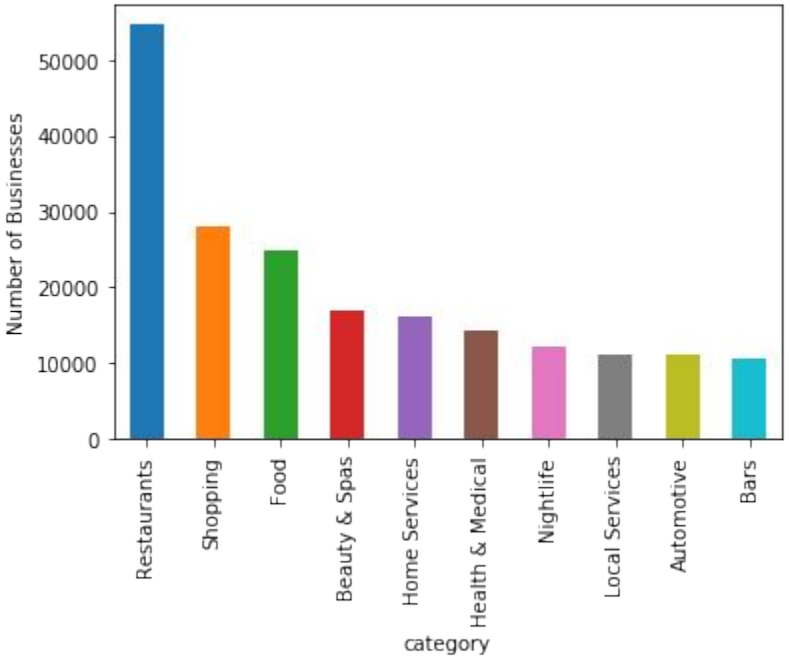
Data Overview

The Yelp dataset contains information about the reviews of users and businesses. The idea is that users of Yelp:

- visit a business
- rate their experience with the business from 1-5 stars
- leave comments about their experience

Users can search for a business and learn from the experiences of other users.

Data Overview - Categories

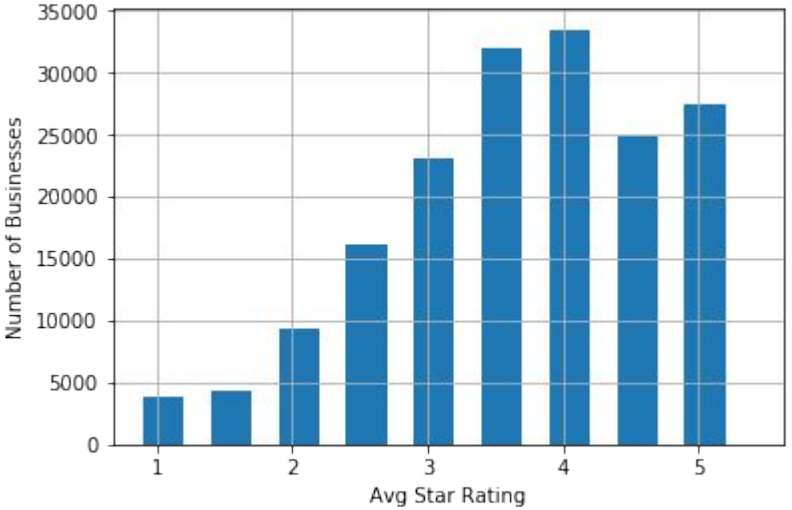


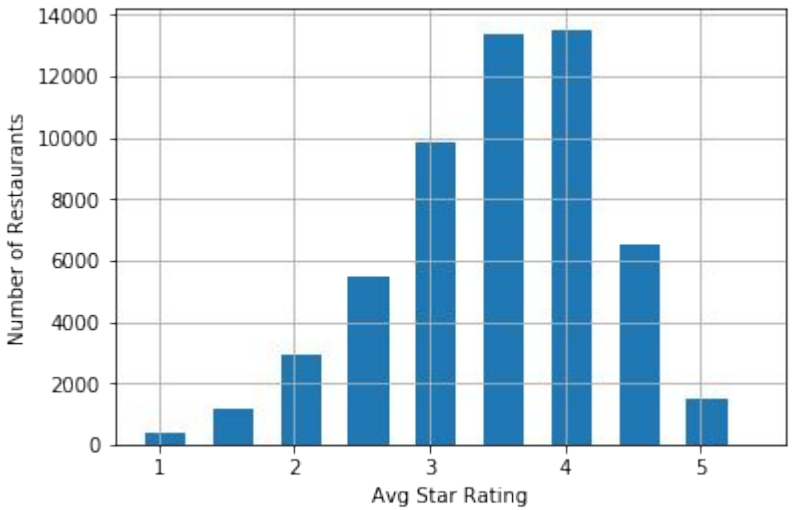
174,567 Total Businesses
1,293 Total Categories

Restaurants makes up the largest category with 54,618 businesses

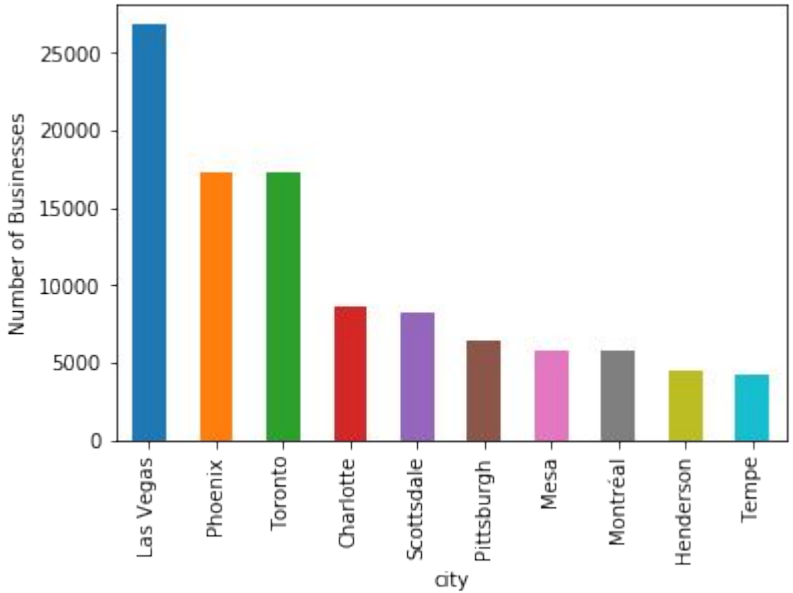
Note that each business can fall under multiple categories

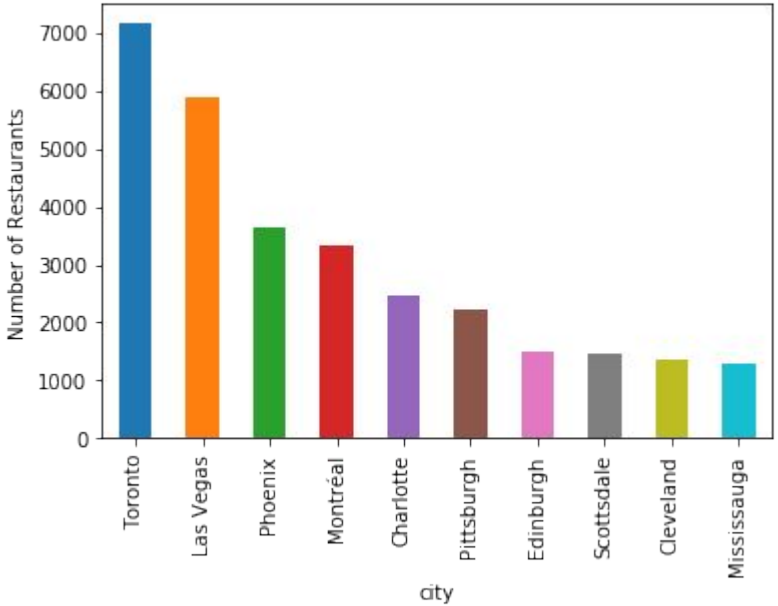
Data Overview - Average User Ratings



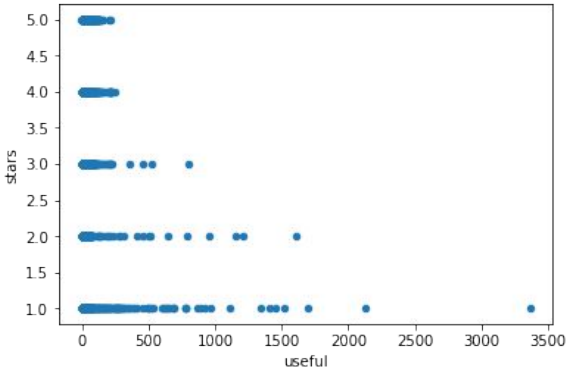


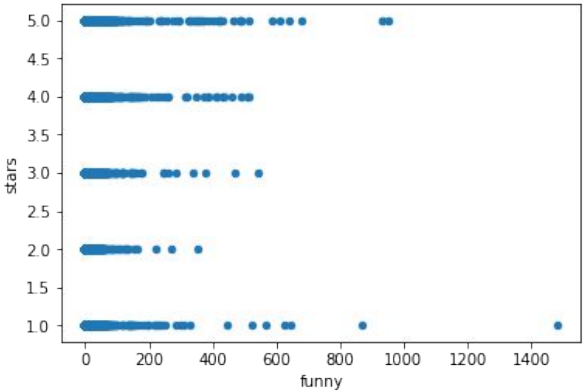
Data Overview - Geographical Distribution

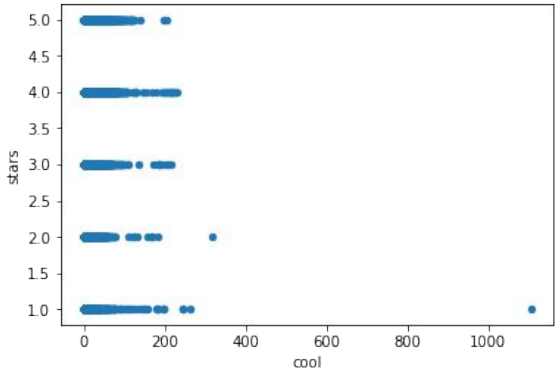




Data Overview - User Comments - For Restaurants







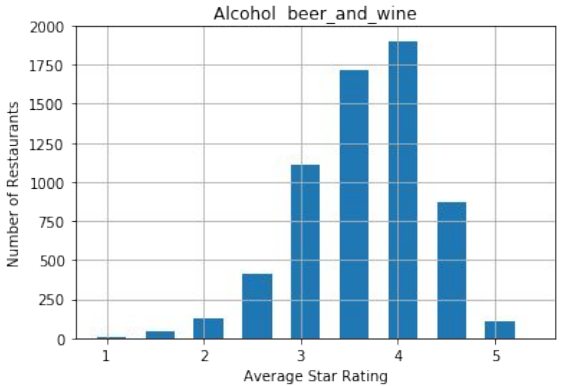
What gives a restaurant its average rating?

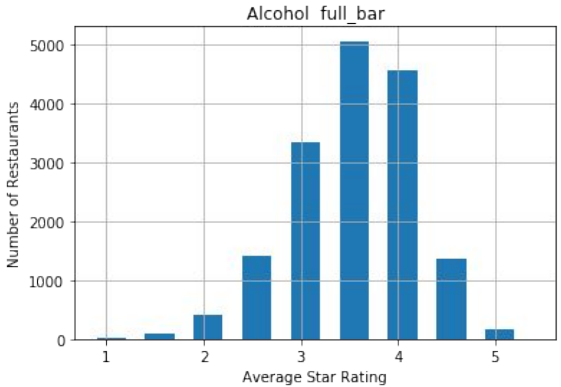
Businesses come with attached “Attributes”

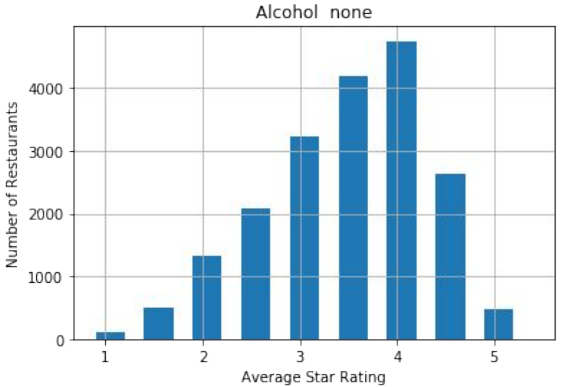
- 39 attributes in total
- A business can take many attributes
- Each attribute for a business may take a value
    - For example, a company with the “WiFi” attribute can take a value of either “free”, “paid”, or “no”

#	name	total_businesses
1	BusinessAcceptsCreditCards	131714
2	RestaurantsPriceRange2	102754
3	BusinessParking	96421
4	BikeParking	79014
5	GoodForKids	61625
6	RestaurantsTakeOut	57786
7	OutdoorSeating	52468
8	RestaurantsGoodForGroups	51888
9	RestaurantsDelivery	49168
10	RestaurantsReservations	48774
11	WheelchairAccessible	47910
12	RestaurantsAttire	46707
13	WiFi	46210
14	Alcohol	46167
15	HasTV	45523

Attribute Data - Restaurants with Alcohol







The “Alcohol” attribute can take one of three values: “beer_and_wine”, “full_bar”, and “none”.

Categories Beyond “Restaurants”

A restaurant can have different categories

Some categories appear that are not “expected” for a restaurant:

- Office Cleaning
- Estate Planning Law
- Preschools
- Currency Exchange

- 'American (New)'
- 'American (Traditional)'
- 'Arts & Entertainment'
- 'Asian Fusion'
- 'Bakeries'
- 'Barbeque'
- 'Bars'
- 'Beer'
- 'Breakfast & Brunch'
- 'Buffets'
- 'Burgers'
- 'Cafes'
- 'Canadian (New)'
- 'Caribbean'
- ...

KPI - How can accuracy be measured?

Some average ratings are closer to the “true” rating than others.

Both of these restaurants in Las Vegas have an average star rating of 3.
Which of these two restaurants has a more reliable average rating?

Wildburger
3 ratings

Bachi Burger
3,065 ratings

KPI - How can accuracy be measured?

$$\frac{\left(\sum_{i \in restaurants} (predicted_i - actual_i)^2 \times reviews_i\right)^{1/2}}{\sum_{i \in restaurants} reviews_i}$$

The idea is that restaurants with more reviews recieve more consideration

Modeling - Features

Categories: Only consider categories attached to more than 500 restaurants, and do not consider restaurants without one of these features

- 64 categories remain
- There are still 53,373 restaurants

Attributes: Only consider categories that exist for at least 50% of all restaurants.

- 18 attributes remain

Modeling - Features

Categories: Use 0/1 encoding

- For example, every restaurant either belongs to the category “Bakeries”, or it does not.

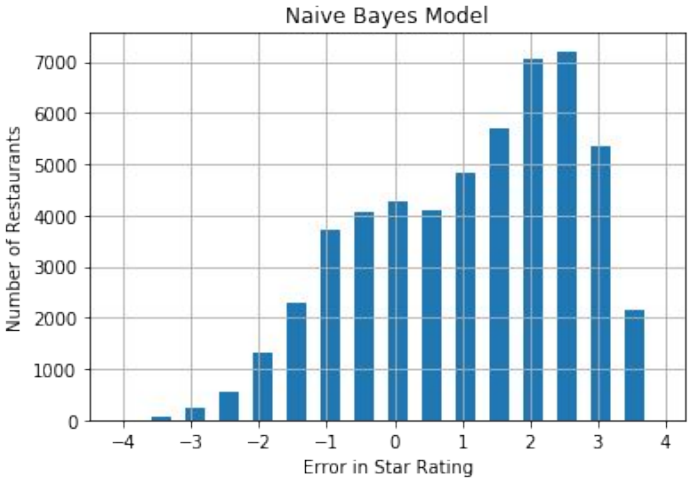
Attributes: Encode each attribute (as an integer). If a restaurant is missing an attribute, use a special “-1” code.

- An alternative approach is to replace missing values with the mode of the attribute, but this creates worse estimations here!

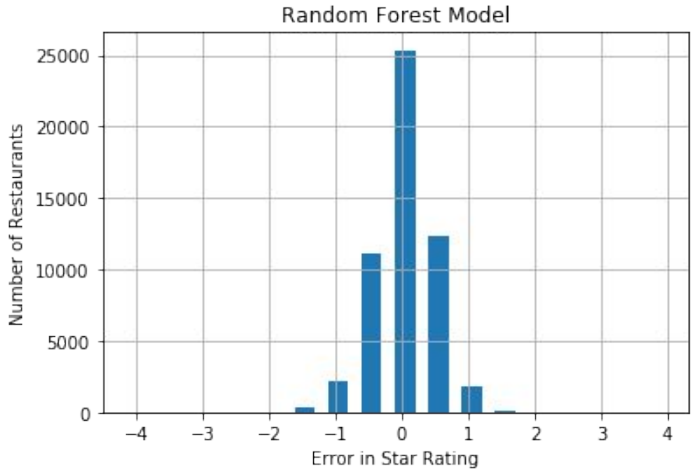
Modeling - Naive Bayes and Random Forest

- Allow for weighted training data
- Use the nearest “half-star” rating (so if the model predicts a rating of 3.68 stars for a restaurant, then we take it to be 3.5 stars)
- Select optimum hyperparameters (for the random forest) with a randomized grid search with 5-Fold cross validation

Modeling - Naive Bayes and Random Forest



KPI on full data 9.578 x 10e-4



KPI on full data 1.736 x 10e-4

Important Features

Based upon the impurity loss in the random forest model, we can see which features were important for that model.

Attributes

- BikeParking
- AgesAllowed
- BYOB
- BusinessAcceptsCreditCards
- Alcohol

Categories

- Fast Food
- Food
- American (Traditional)
- Burgers

Further Questions

- Location was not considered. Is there a geographical influence on ratings?
- Some attributes and categories can be used to get a feel for the average rating of a restaurant. But what about the distribution of ratings?
- What features can be used to predict the ratings of businesses in areas besides restaurants?