**Optimized Rating Model Part I**

**An Exploratory Analysis**

**Bradley Droegkamp**

**Introduction:**

Before grabbing a bite to eat, restaurant patrons often reference user ratings to make decisions on where they go. The problem is these raw ratings can at times be incomplete or biased, which is a disservice to the restaurant owner and the patron. Foursquare and other clients that publish user ratings for restaurants have interest in building an optimized rating model to remedy this problem. As a first step, we would like to have available an exploratory analysis on which variables would be most useful to include in this model.

The end-result of an optimized rating model is likely a classification model that scales from 0 to 100% (5/5 stars) that closes in on the actual user rating as the number of reviews approaches a certain threshold. The scope of this project is to define which independent variables to consider in this model. A/B testing for all potential variables with an incomplete dependent variable could be tedious and incomplete. Alternatively, we can use K-means clustering to give us more insight into which variables to consider in the final model.

**Data:**

We will focus our research on restaurants listed on **Foursquare** ([developer.foursquare.com/](https://developer.foursquare.com/)) in the Chicago major metropolitan area. Along with user ratings, number of ratings, cuisine type, location, and other variables we can obtain from Foursquare, we will reference demographic information about the tracts in which the restaurants are located from the **Federal Financial Institutions Examination Council** (FFIEC - [www.ffiec.gov/](http://www.ffiec.gov/)). A census tract *is a geographic region*defined*for the purpose of taking a*census (<https://en.wikipedia.org/wiki/Census_tract>). FFIEC search results are based on 2015 Census data.

For example, one data point is **Chapati Mediterranean Restaurant** with a user rating of **6.5**/10 given **8** total ratings and **Moderate** affordability according to Foursquare. The tract this restaurant is located in has a population of **4,888**, an **Upper** income level (definitions for income levels here: <https://www.ffiec.gov/census/htm/2015CensusInfoSheet.htm>), and a median household age of **30** according to FFIEC.

**Methodology:**

1. **Gather Restaurant Data from Foursquare**

Foursquare limits total number of premium API calls to 50. We are thus limited to a sample of 50 in order to get User Ratings and Pricing detail of each Restaurant.

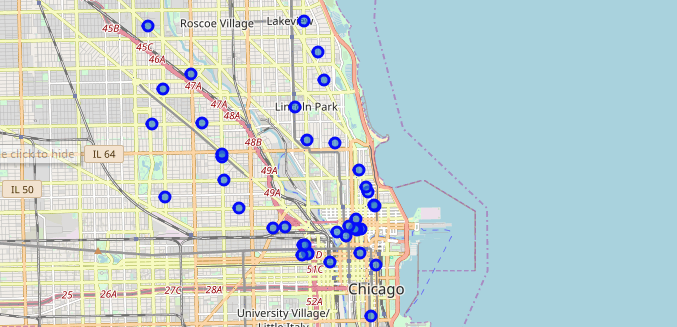


Fig 1: Map of all Restaurants in the Dataset

1. **Gather Demographic Data from (FFIEC -** [**www.ffiec.gov/**](http://www.ffiec.gov/)**) and Merge with Foursquare Data**

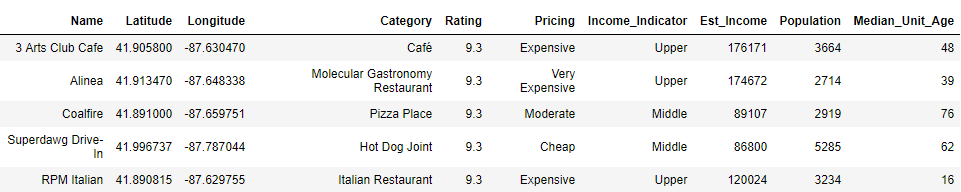
****

Fig 2: Example of 5 samples from Dataset

1. **Explore Data**

* The majority of our sample of restaurants reside in **Upper** Income Bracket areas.
* We can remove **Income Indicator** since Estimated Income provides a more granular detail.
* **Median Unit Age** gives us an idea of how mature the area is. It appears most areas are either newly rebuilt or historic, with less in-between.
* **Ratings** from Foursquare are sorted by descending order, so it is a challenge to obtain a wider variety of ratings.

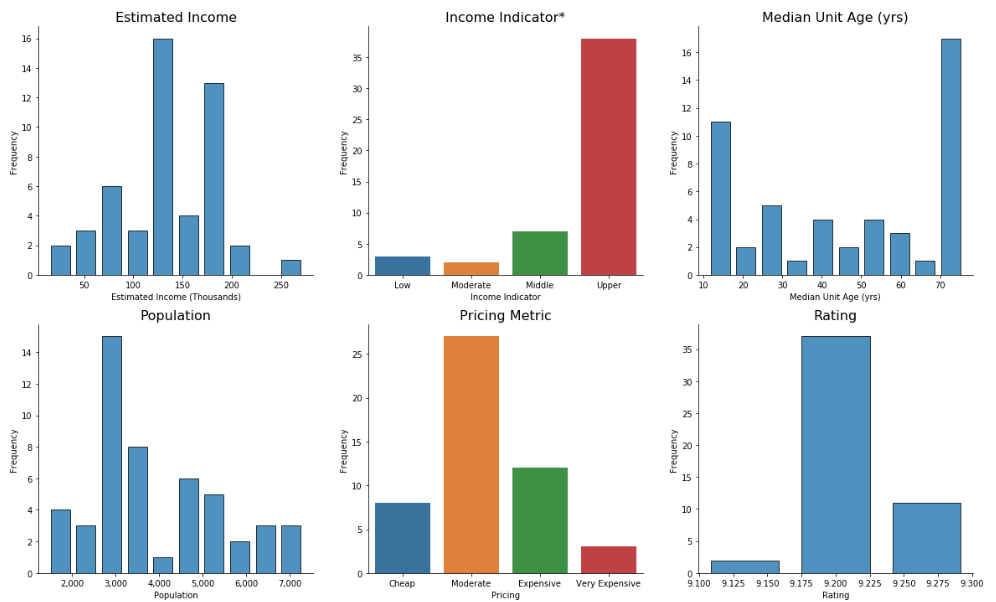


Fig 3: The Dataset broken down by several fields

\*Income Indicator is defined as:

**Tract Income Level** - This corresponds to tract classifications as defined by the HMDA and CRA regulations. This field is based on the Tract Median Family Income %: If the Median Family Income % is < 50% then the Income Level is Low. If the Median Family Income % is >= 50% and < 80% then the Income Level is Moderate. If the Median Family Income % is >= 80% and < 120% then the Income Level is Middle. If the Median Family Income % is > =120% then the Income Level is Upper.   
<https://www.ffiec.gov/census/htm/2015CensusInfoSheet.htm>

****

Fig 4: Restaurant Category Breakdown

1. **K-means Clustering**

First encoded categorical data. Then after running with several different cluster counts, 3 was most logical given the small size of the dataset.

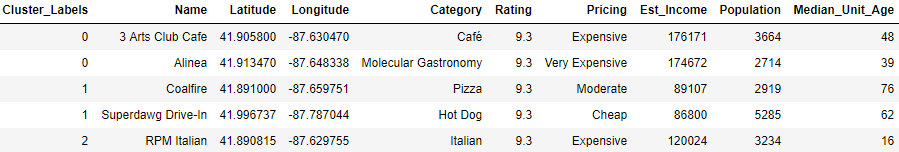


Fig 5: 5 samples from the data with clusters identified

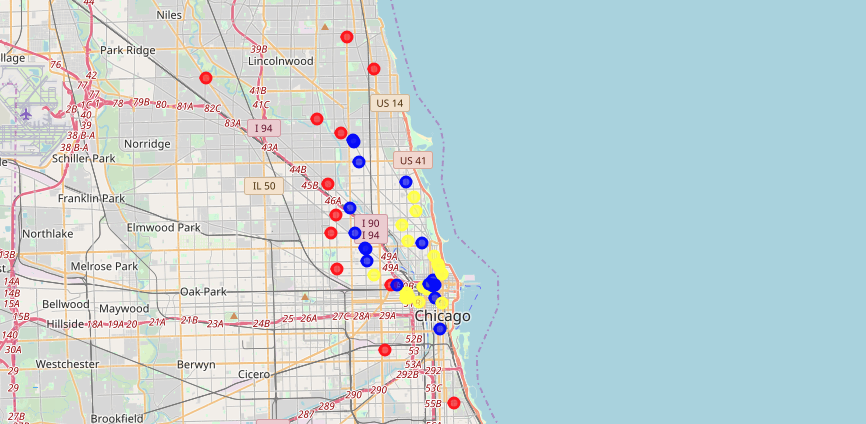


Fig 6: Restaurants by Cluster. Cluster 0 = Yellow, Cluster 1= Red, Cluster 2 = Blue

**Results:**

#### Estimated Incomes and Median Building Age vary significantly, while Population and Rating do not. Also note on the map above that cluster 1 is further from the city center than the other cluster.

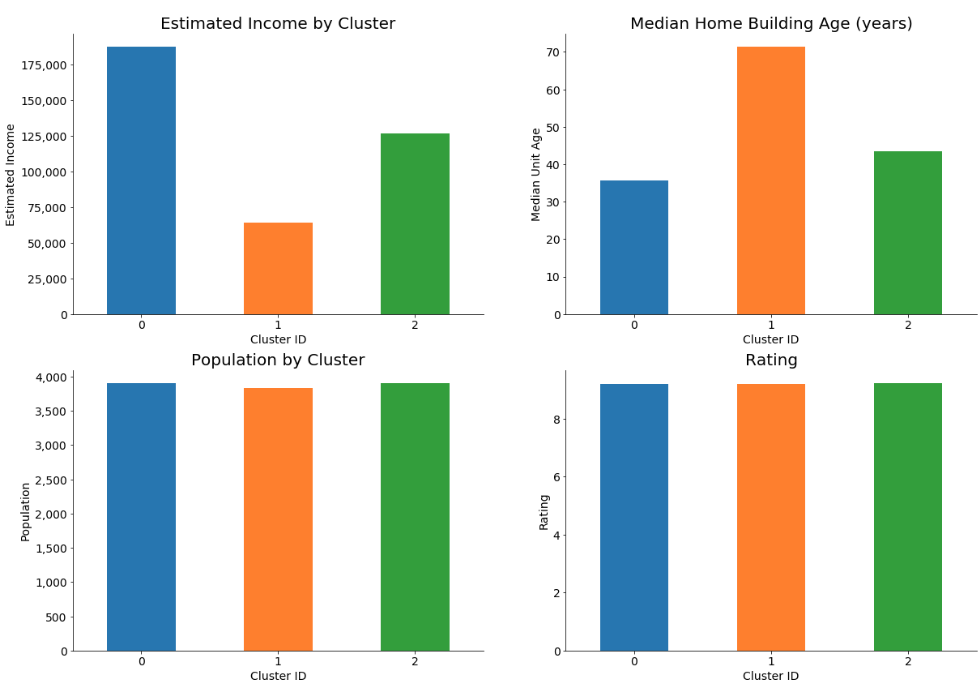


Fig 7: Cluster Breakdown

### **Top Restaurant Categories and Price Ratings by Cluster:**

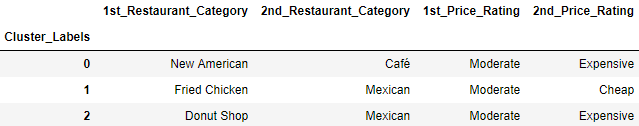


Fig 8: Top Restaurant Categories and Price Ratings by Cluster

**Discussion:**

Older neighborhoods (based on median house unit age) with lower average estimated incomes and located further from the city center (cluster 1) were distinctly different than the other clusters. Beyond this, limitations of Foursquare's basic API access hasn't permitted a robust dataset with varying User Ratings. As a result, we are unable to utilize a larger dataset to assess our objective of observing how demographic information and other Foursquare information affects User Ratings.

Variables such as Foursquare price rating, neighborhood estimated income, and median house unit age appear to be correlated and potentially redundant. In addition to a more robust Foursquare dataset, obtaining more relevant demographic information such as neighborhood age should enrich the analysis. Furthermore, demographic information of the Foursquare users should be added to any future analysis as well.

**Conclusion:**

A revision of this User Rating exploratory analysis will be required once we have premium Foursquare access. More clusters in a larger dataset (at least 500 restaurants) with a random distribution (rather than sorted by User Rating) may provide more segmented, meaningful results.