Predicting prices on FourSquare venues

Surya Prerapa

8th March 2020

1. Introduction

On inspecting New York and Toronto's FourSquare venue data, it was evident that customers can derive real value out of getting an approximate price estimate at a glance on a given venue. With this price information, they can quickly assess if a place fits into their budget and occasion before they spend time reading the menu and reviews. Adding pricing information to a venue can be done either directly or indirectly using the FourSquare Premium API calls or some other platforms such as TripAdvisor. But what about locations such as in developing countries or redeveloped/rejuvenated neighbourhoods where quality data is not available to make such qualitative assessments? Is there a simple intuitive approach to qualitatively assign price ranges to venues ? This project is an attempt at such an undertaking.

Potential beneficiaries of such a tool would be travel, tourism and hotel booking portals (& aggregators), home delivery platforms, etc. This is because, customers visiting such websites would have a vast collection of venues to choose from especially in big cities, even in narrow categories such as Italian Restaurant or Pizza Place. Being able to filter out venues based on price levels would be a great value addition for at least some of them.

With this in mind, I proposed a simple intuitive approach below to assess price ranges of venues in Toronto city and compared against real-world price values to assess the accuracy. The central idea underpinning this project is to obtain freely available demographic data with indicators for disposable income in each neighbourhood of Toronto. Then cluster these neighbourhoods into low, medium and high income neighbourhoods. Based on these cluster labels, query FourSquare venues in each neighbourhood and apply these cluster labels (based on neighbourhood). Finally, pick a few random samples from this enhanced dataset and compare the computed price range against their real price ranges.

2. Data Sources, cleaning and transformation

Three datasets have been used for this project:

a. **Neighbourhoods geographical data:** Toronto's geographical data has been sourced from this free data platform https://open.toronto.ca/dataset/neighbourhoods/. CSV file formats containing WGS84 coordinates (latitude and longitude) of neighbourhoods has been downloaded. GEOJSON file format containing the geographical boundaries of all neighbourhoods has also been obtained to view the neighbourhood clusters in Choropleth map.

b. **Neighbourhoods demographical data:** Toronto's demographical data has been sourced from this free data platform https://open.toronto.ca/dataset/neighbourhood-profiles/. CSV file formats containing ethnicities, spoken languages, age-group, income, population density, benefits entitlement, etc has been downloaded.

c. **FourSquare regular API calls:** 100 venues within 500Metre radius of the neighbourhoods from the above are used.

Area code had to be cleaned out of *Area name* field using regular-expression-replace operation for joining the geographical and demographical data. In the demographical data only the rows corresponding to Population density and Average after-tax income fields per neighbourhood are retained. The demographical data is transposed to align with geographical data. Finally the values have been converted to integer format after removing commas from the entries.

Descriptive statistics have been obtained on demographical data to see if three classes can be broadly defined within the data. This has been confirmed to be the case.

Since this sort of classification of venues into bargain, medium and premium doesn't apply to all sorts of venues given by FourSquare we have restricted this project only to restaurants, pubs, cafes, coffee and dessert shops. Remainder of the categories have been filtered out of the dataset.

3. Feature engineering

As stated above, we are aiming to achieve an indicator for disposable income of the people in a given neighbourhood which in turn dictates how much people are willing to spend for dining and drinking out. This will in turn determine the price establishments in this neighbourhood will set. We picked Population density and Average after-tax income per neighbourhood as this indicator. This is based on the following assumptions:

1. Affluent areas have low population density (spacious accommodation) and high average tax-free incomes. Median (instead of the average) would have been a better indicator but the datasets do not have this data readily available.
2. Establishments cater to local area closely and their prices reflect affordability of locals
3. Effects of transient population such as tourists and commuters ignored
4. Fixed priced models of local and international chains ignored.

Since these columns are at different scales and we don't want the income to dominate the population density variable, we scale the data using min-max scaler. The choice of scaling itself has minimal effect on the results.

4. Finding patterns in the data

We have tried to find the optimal number of clusters using the K-Means algorithm in the dataset employing the elbow method. The results are shown in Figure 1 below. It demonstrates that either 2 or 3 clusters can be optimal for the dataset under consideration. We pick K=3 so that we can divide venues into bargain, medium and premium categories. When K is set to 3, we get the three clusters, with centers shown in Table 1 below:

Population density	Average tax-free income	Neighbourhood income	Venue category
21513	37559	Low income	Bargain
5537	36850	Medium income	Medium
4791	101460	High income	Premium

Table 1: Cluster centers and their interpretation.

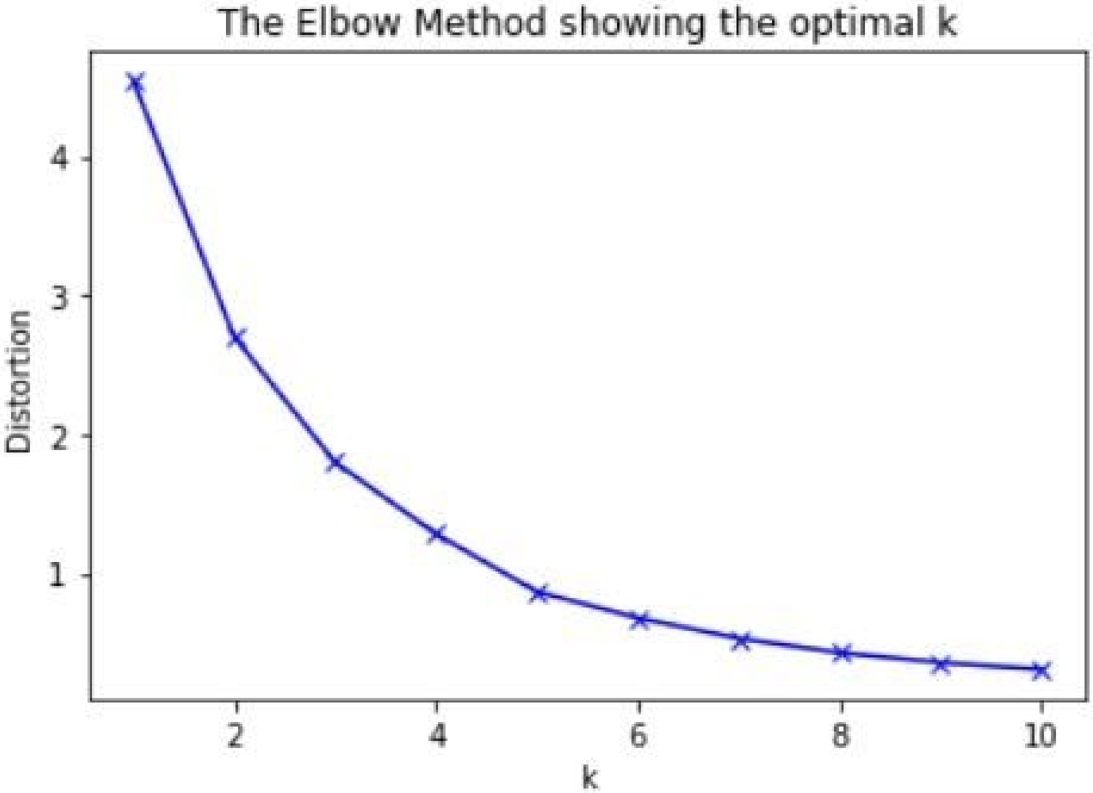

Figure 1: Elbow method to determine the optimal number of clusters

In Table 1, only the first two rows merit explanation. Note that although the income in row #1 is larger than #2, since it is divided over a much larger population density (proportionally assigned to each household or income), it is categorised as low income and vice-versa. For the high income, both the population density and income stand hand and shoulders above the rest making the choice easy for us.

Visualising the cluster labels over each neighbourhood as a Choropleth map, we see Figure 2 below where the majority of neighbourhoods are medium income and only a small proportion belonging to low and rich incomes as expected in a rich city like Toronto. These cluster labels have been assigned to the FourSquare venue data filtered over restaurants, pubs, cafes, coffee and dessert shops and shown as 6 dimensional data plot in Figure 3. Dimensions contain 5 venue categories listed above plus the cluster label shown in red, green and blue corresponding to bargain, medium and premium respectively.

5. Results and discussion

For rudimentary validation with real-world data, I drew 20 random samples from the above data, assigned the price rating sourced from TripAdvisor to each establishment. Note that TripAdvisor rates places on the scale of £, ££, £££, ££££ with progression from cheap eats to expensive options. For the purpose of our study £ corresponds to bargain, ££££ corresponds to premium and the middle two both are

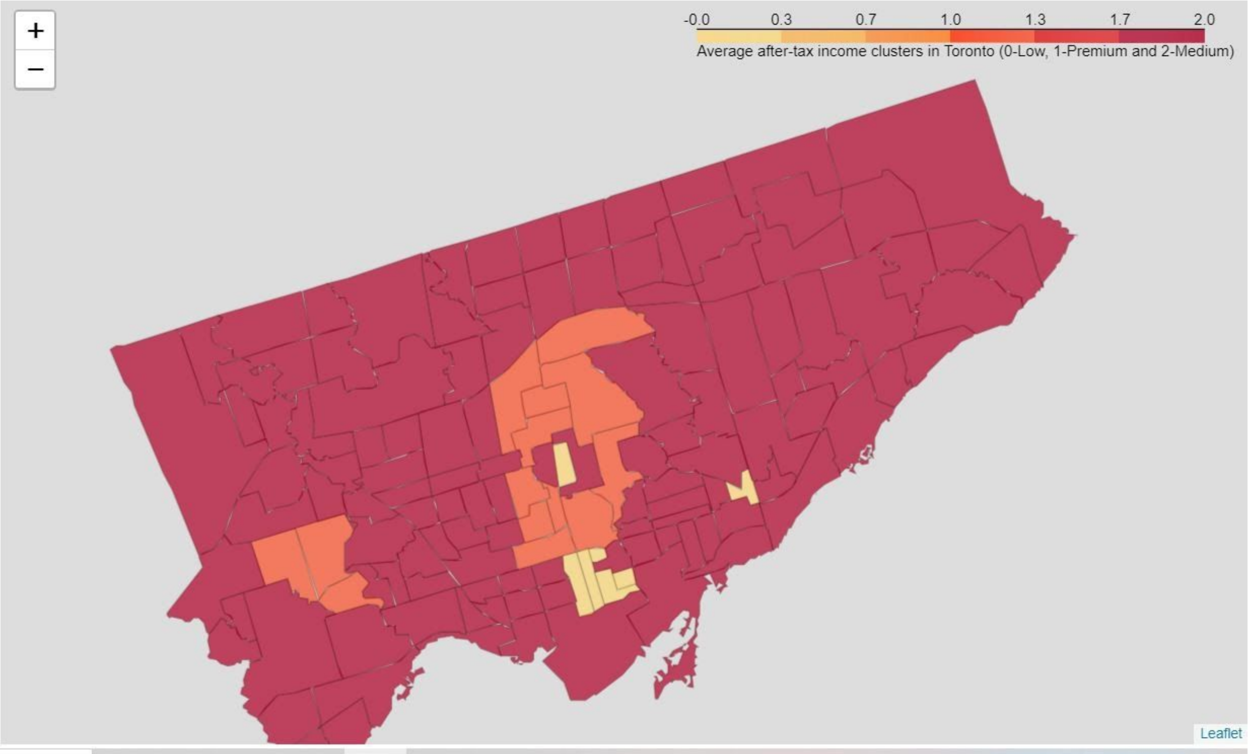
+
−
-0.0
0.3
0.7
1.0
1.3
1.7
2.0
Average after-tax income clusters in Toronto (0-Low, 1-Premium and 2-Medium)
Leaflet

Figure 2: Toronto neighbourhood per-person income levels

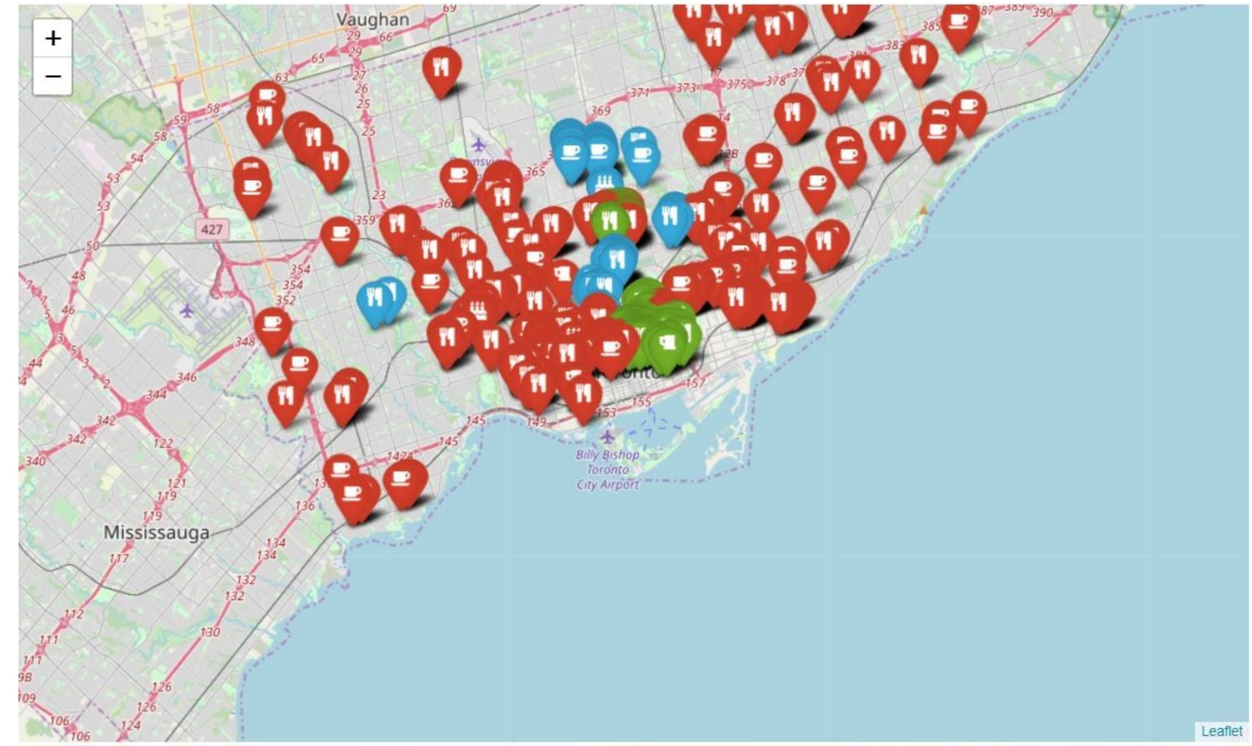

Figure 3: Toronto restaurants (cutlery), pubs (beer-mug), cafes & coffee shops (coffee cup) and dessert shops (birthday-cake) shown in red, green and blue denoting bargain, medium and premium price ranges.

mapped to medium. With this mapping, 11 out of 20 venues mismatched with the computed price category. Out of these 11, at 2 venues there was a drastic mismatch where we have predicted premium but they turned out to be bargain-eats in reality. The other 9 were closer mismatches such as bargain-medium or medium-premium and vice-versa. So for such a time and resource bound academic case study, this model provided reasonable results. Although it is far from ideal for real-world applications. However, this very limited validation exercise gave valuable pointers for future improvements namely:

1. A couple of mismatches happen at chain restaurants. These usually follow fixed priced models (with very few exceptions such as airports or train terminals) across a country or large region. Hence they should be treated separately.
2. Tourist hotspots, downtown or city centre locations with large transient or commuter populations have skewed income statistics. So they should be dealt with separately.
3. Certain very rich or very poor individuals can skew means. Therefore source and use median after-tax income as a better indicator of disposable income.
4. Also remove cost of living such as average rental or mortgage payments from median incomes to calculate disposable income.
5. Consider other demographics such as age-group statistics, ethnicities, etc for accurate price estimates.

6. Concluding remarks

I have attempted to address an important business challenge using freely available data with reasonable degree of success. Real-world businesses with such a need can easily address this challenge by procuring solutions such as FourSquare premium API calls or TripAdvisor data. But where such data is not available for example in developing countries or redeveloped/re-settled/rejuvenated communities, techniques such as shown above could be good starting points (or solutions if well designed and executed).