Comparing a set of neighborhoods world-wide by Livability score

Mohammed Elwardi Fadeli

10/30/2019

Contents

1 Introduction

1.1 Background

Many of us have an intuitive notion of how **easy** it is to live in a certain neighborhood or city we're well informed about. Typically, one can classify few cities (by pure intuition) as "harder", "easier", or "the same" if he compares these cities to his current living neighborhood. This notion of (intuitively) "clustering cities" can be even extent to predicting which city will leave its original cluster in the next few years.

If one takes New York City as the reference city, he can compare the current state of its neighborhoods with each other, or compare the city to other cities. After that, and by acquiring the history of the features making up his model, he can even predict which cities will the New York cluster in the next few years (cities become much harder or easier to live in)

Such clustering and prediction operations should be based on past history of different factors which contribute to the "easiness of living in a city". This information constitutes an important factor in decisions taken by a wide range of companies world-wide and it's expressed as a Livability score.

The Livability score measurement is not standardized world-wide, but this project focuses on five (5) factors:

- Amenities availability
- Cost of living
- Crime rate
- Education level
- Employment status

1.2 Problem

Data that might determine the livability score of a city or a neighborhood may include the number of Amenities (Groceries, schools, shopping, fitness facilities, libraries … etc) that are available, transportation, health care costs, and poverty rate, different kinds of crimes rates, unemployment rate and even the percentage of the population with less-than-high-school level of education.

This project aims to cluster some neighborhoods and cities world-wide, focusing on New York City neighborhoods, so that similar cities and neighborhood would belong to the same cluster.

1.3 Interest

If a city's cluster can be identified; predicting when the city might leave its cluster is a matter of repeating the process described in this report while building a predictive model (eg. linear regression). Hence this project is an important step in an even more-crucial workflow.

2 Data acquisition and cleaning

2.1 Data sources

For this project, we rely on two main sources of data:

1. A huge data set of NYC neighborhood stats
2. Foursquare API

Of course, the NYC neighborhood stats data set was last updated in December, 2017, but the foursquare data are fetched at a later date (November, 2018). We assume an offset of one year wouldn't affect the results that much.

Foursquare also sets some limitations on how much we can get per day. Thus, the code assumes sandbox accounts are used by default but also works in a much more accurate way if premium accounts are used.

NYC is not the only city studied in this project, non-foursquare data for other cities was acquired manually and added to the data frame.

This webpage was used as a base t construct our own Livability score.

2.2 Data Cleaning

The following features from the NYC neighborhood stats data set were chosen to calculate the livability score:

- Poverty
- Violent Crime
- Property Crime
- EduLessThanHS
- Crowded Housing
- Health Ins
- Unemployment Rate

All these metrics were then normalized so they becomes indices in the range [0, 1]

Also, data rows are cleaned while looking for Foursquare venues (if a Neighborhood fails to be found, it's dropped) for the following search queries:

- Groceries
- Food & Drink
- Shopping
- Schools
- Entertainment
- Fitness Facilities
- Transportation
- Libraries
- Goods & Services

If the user account is a sandbox one, only counts of these venues are used to cluster neighborhoods (due to Foursquare limitations on premium calls), but if a premium account is used, the code fetches "likes" for each venue and use that instead. Of course, this information is normalized over the Limit set for foursquare queries (20).

Also, all queries results are saved into JSON files so we can retrieve latitude/longitude data from Foursquare searches.

2.3 Feature selection

After cleaning, the data set has 20 features in addition to neighborhood names. But some of these features were not available for the majority of neighborhoods, so they were dropped (Only 16 remained).

3 Exploratory Data Analysis

3.1 Calculation of target variables

To be able to estimate the livability of a neighborhood, five indices must be calculated using existing data:

- Amenities availability
- Cost of living
- Crime rate
- Education level
- Employment status

Each index is calculated by multiplying the value of the feature by a certain coefficient (all coefficients for each index add up to 1, so they are a measure of the impact of a feature on the index's value; These can be estimated easily using a survey for example).

The contribution of data frame columns are shown in the following table [1] :

	Amenities	Cost of Living	Crime	Education	Employment
Crowded Housing	0.00	0.25	0.00	0.00	0.0
EduLessThanHS	0.00	0.00	0.00	0.85	0.0
Entertainment	0.16	0.00	0.00	0.00	0.0
Fitness Facilities	0.06	0.00	0.00	0.00	0.0
Food & Drink	0.17	0.00	0.00	0.00	0.0
Goods and Services	0.00	0.30	0.00	0.00	0.0
Groceries	0.17	0.10	0.00	0.00	0.0
Health Ins	0.00	0.15	0.00	0.00	0.0
Libraries	0.03	0.00	0.00	0.00	0.0
Poverty	0.00	0.10	0.00	0.00	0.0
Property Crime	0.00	0.00	0.35	0.00	0.0
Schools	0.12	0.00	0.00	0.25	0.0
Shopping	0.26	0.00	0.00	0.00	0.0
Transportation	0.03	0.10	0.00	0.00	0.0
Unemployment Rate	0.00	0.00	0.00	0.00	0.1
Violent Crime	0.00	0.00	0.65	0.00	0.0

[1] These coefficients are simplified. The table can have much more columns and rows!