Classification of Seattle Neighborhoods by Venue and Housing Price Data

By Alexey Rybak

May 19th, 2019

Contents

Introduction

Background

Seattle is the fastest-growing big city in the United States of America[1]. Having gained over 100,000 new residents over the past decade, it is rapidly becoming one of the key West Coast business and cultural centers.

This growth has transformed the city. Rapid rise in housing prices is accompanied by overall increase in business activity. At the same time, not all parts of the city are affected by these trends to the same degree: some neighborhoods, especially those experiencing rapid gentrification, may lack certain amenities compared to their more established counterparts.

There are several existing services that evaluate neighborhoods based on various criteria (e.g. Areavibes, Walkscore, etc.), but they mostly focus on providing data per each neighborhood, without attempting to uncover city-wide patterns.

Problem

There are two target customer groups for this research, each with its own need:

- *Prospective homebuyers* are looking for neighborhoods with certain amenities (depending on their demographics and lifestyle) and lowest possible property prices;
- *Business owners* about to open or expand a business are looking for neighborhoods with sufficiently affluent population (as reflected by median house prices) and a relative lack of competition.

Solution

This project will cluster Seattle neighborhoods by availability of various venues using machine learning techniques, and then rank the neighborhoods within each cluster by median housing prices. Essentially, we want to understand if there are significant differences between 'similar' neighborhoods in terms of real estate pricing.

Prospective homebuyers could use this information to identify the most affordable neighborhood for a given set of venue features; business owners could identify the most lucrative neighborhood lacking sufficient venues of the type they would be interested in opening.

[1] https://www.seattletimes.com/seattle-news/data/114000-more-people-seattle-now-this-decades-fastest-growing-big-city-in-all-of-united-states/

Data

Sources

The following data sources will be used for the research project:

- Geographical data on Seattle neighborhoods from the Seattle Open Data program (https://data.seattle.gov).
- Data on median housing prices for each neighborhood from Zillow using Zillow API (https://www.zillow.com/howto/api)
- Data on various Seattle venues from Foursquare using Foursquare API (https://developer.foursquare.com/).

Data strategy

First, geo data from Seattle Open Data will be used to create a map of Seattle neighborhoods. This data will then be combined with median housing prices for each neighborhood. Finally, each neighborhood will be populated with venue information from Foursquare, which will then be used for neighborhood clustering.

Data cleaning and availability

All data providers mentioned above provide data in a format ready for analysis:

- Seattle Open Data project provides a GeoJSON containing information on all neighborhoods, including boundary data;
- Zillow provides a single XML file with median housing prices for each neighborhood via its API;
- Foursquare API provides several endpoints for venue information, returning a JSON for each query. For this research, the 'search' endpoint will be used.

There are several data limitations to keep in mind:

- Foursquare limits the number of requests to 99,500 regular API calls and 500 premium API calls per day. 'search' is a regular endpoint, and the algorithm will run 100 queries per neighborhood, so there is no risk of running over the limit.
- Zillow limits the number of requests to 1,000 calls per day. A single call using 'GetRegionChildren' API will be made to obtain median house prices for all neighborhoods, so there is no risk of running over the limit, either.
- Finally, a single call to the Seattle Open Data portal will be made to retrieve the GeoJSON file.