Battle of the Neighborhoods:

Seattle Neighborhood Classification 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.

Methodology

Feature Selection

We will pull venue data from Foursquare for each neighborhood and create a one-hot model of all venues across all neighborhoods. We will then aggregate venue information for each neighborhood and calculate mean values for each feature.

	Neighborhood	ATM	Adult Boutique	Advertising Agency	African Restaurant	Airport	Airport Terminal	Alternative Healer	American Restaurant	Animal Shelter	Antique Shop
0	Adams	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00
1	Alki	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00
2	Arbor Heights	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00
3	Atlantic	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00
4	Belltown	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.010000	0.0	0.00
5	Bitter Lake	0.040000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.04
6	Briarcliff	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00
7	Brighton	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00
8	Broadview	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00
9	Broadway	0.019231	0.0	0.0	0.000000	0.0	0.0	0.0	0.057692	0.0	0.00
10	Bryant	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00
11	Cedar Park	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.00
12	Central Business District	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.050000	0.0	0.00

We will also create a dataframe with top 10 most common venues in each neighborhoods, which we will later use for result interpretation.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adams	Mexican Restaurant	Cocktail Bar	Ice Cream Shop	Coffee Shop	Burger Joint	Bar	Vietnamese Restaurant	Sushi Restaurant	Bakery	BBQ Joint
1	Alki	Trail	Boarding House	Park	Scenic Lookout	Playground	Brewery	Falafel Restaurant	Dumpling Restaurant	Electronics Store	Ethiopian Restaurant
2	Arbor Heights	BBQ Joint	Music Venue	Zoo Exhibit	Field	Ethiopian Restaurant	Event Space	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Filipino Restaurant
3	Atlantic	Bus Station	Scenic Lookout	Rental Service	Trail	Skate Park	Tunnel	Dry Cleaner	Park	Café	South American Restaurant
4	Belltown	Bar	Coffee Shop	Bakery	Sushi Restaurant	Pizza Place	Cocktail Bar	Breakfast Spot	Hotel	New American Restaurant	Middle Eastern Restaurant

Modeling

We will apply several unsupervised machine learning classification algorithms to identify clusters of neighborhoods based on these features, in particular:

- K-means clustering
- Agglomerative (hierarchical) clustering
- DBSCAN

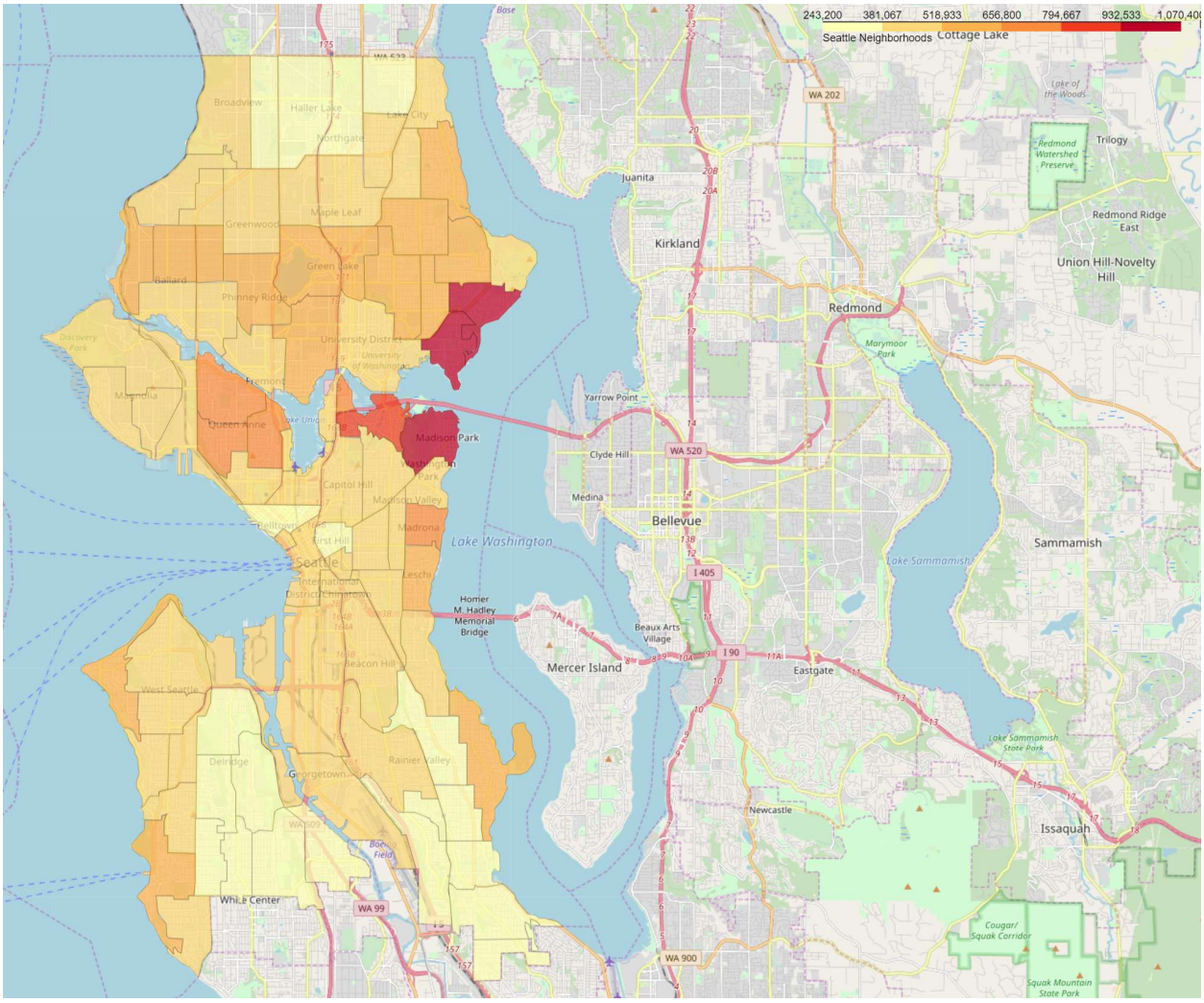
Data Visualization

We will show Seattle neighborhoods on the map using boundary data from the Seattle Open Data project, and add a choropleth layer to show housing prices for each neighborhood.

We will then display clusters on the map for each machine learning algorithm.

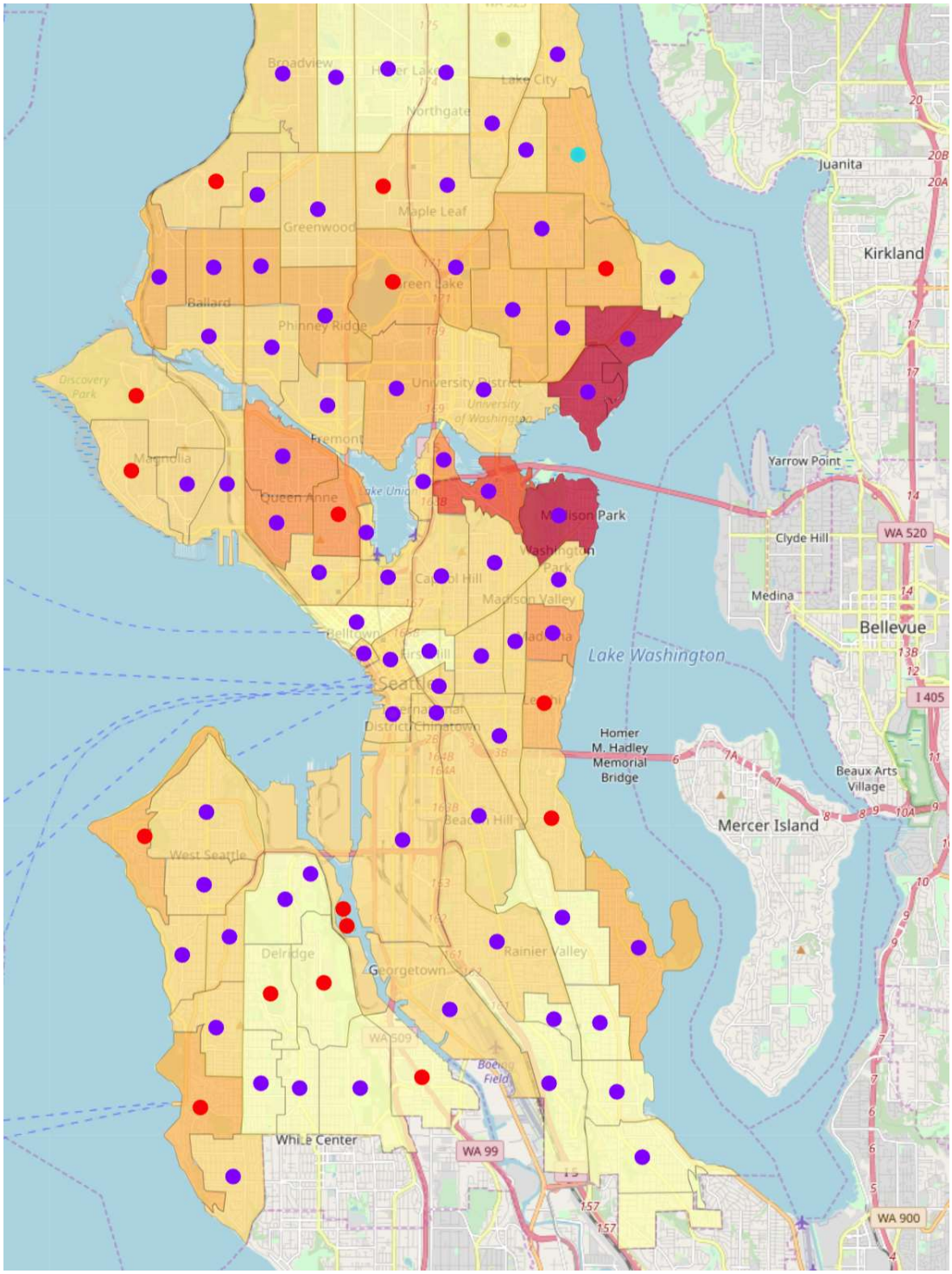
Results

Based on the data from Seattle Open Data and Foursquare, there are 89 neighborhoods and 276 features for the model. Unfortunately, Zillow API did not return median housing price values for each neighborhood, so mean city-wide house price was calculated to replace missing data.

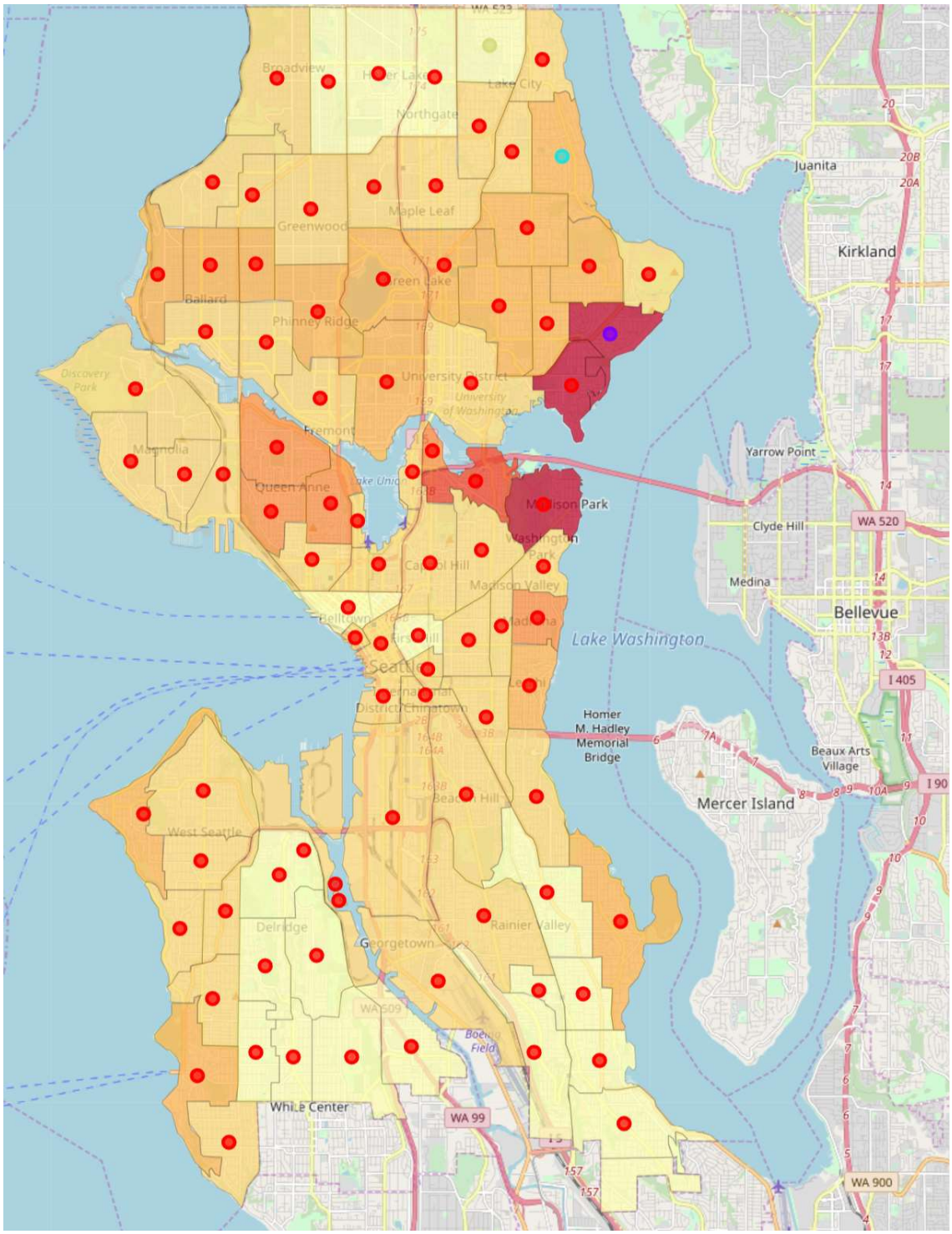


The machine learning models generated the following output:

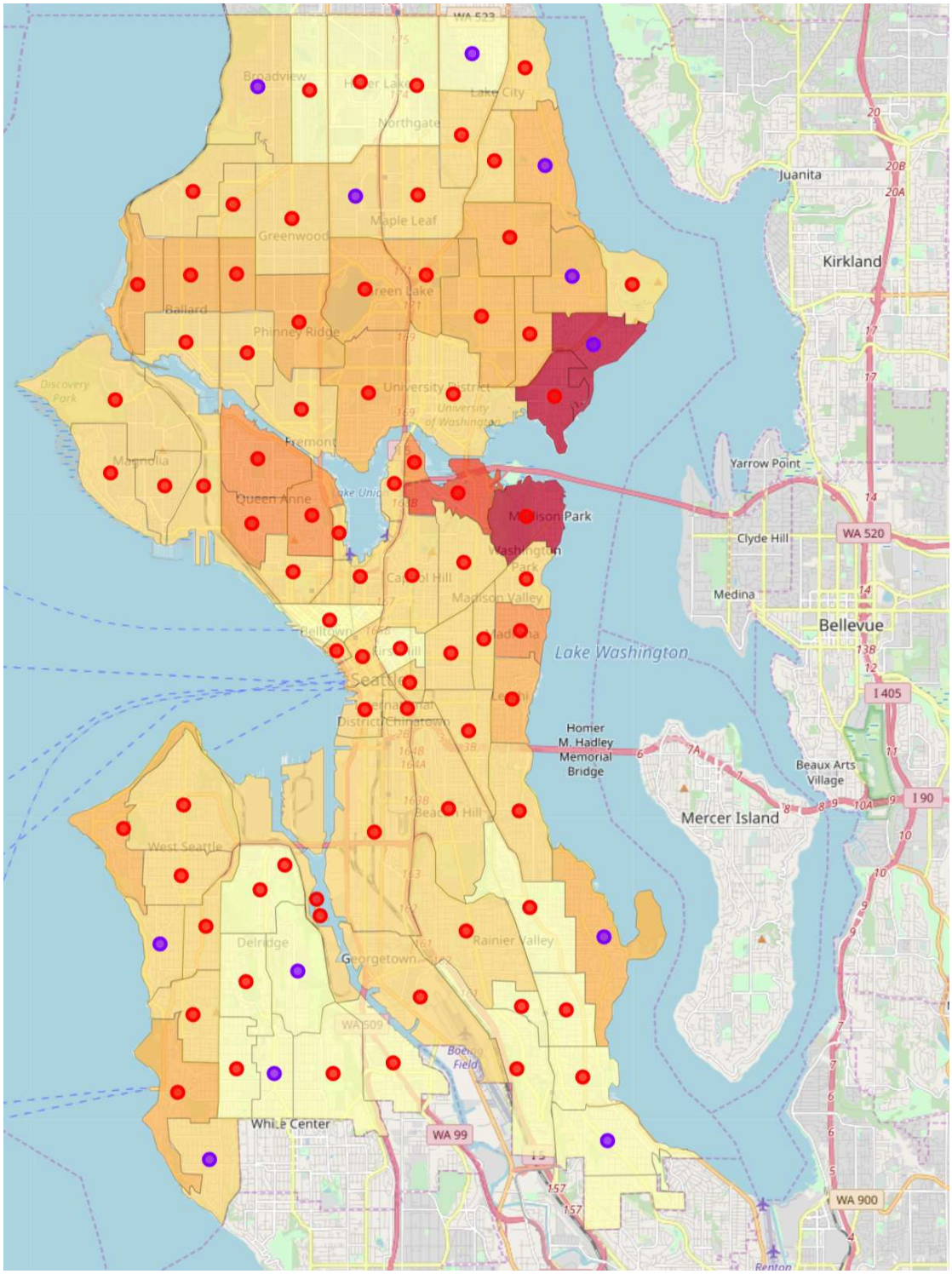
- **K-Means clustering**. One cluster of 16 neighborhoods, one cluster of 73 neighborhoods, and two clusters of one neighborhood each.



- **Agglomerative clustering**. One cluster of 88 neighborhoods, and three clusters of one neighborhood each.



- **DBSCAN**. One cluster of 79 neighborhoods, and 12 outliers.



Discussion

Out of the three algorithms applied to the model, k-means clustering seems to provide the most meaningful result, so we will use it to interpret the data.

The first cluster (73 neighborhoods) comprises the majority of Seattle, and is characterized by a mix of urban venues: coffee shops, bars, various restaurants and grocery chains. These neighborhoods also feature public transit infrastructure (bus stops and light rail stations).

Variance in housing prices within this cluster is substantial: from over 1 million in Laurelhurst to sub-300 thousand in neighborhoods like Rainier Beach, Highland Park and Dunlap. Most neighborhoods on the lower end of the spectrum are rapidly gentrifying, with newly constructed townhomes dominating the residential market. Families looking to move into Seattle are likely to consider these neighborhoods as their first choice.

The second cluster (16 neighborhoods) is characterized by close proximity to parks, trails and playgrounds. Unsurprisingly, being quiet residential areas that they are, these neighborhoods tend to be located off city center, with 6 neighborhoods in North Seattle, 6 in West Seattle, and two in South Seattle.

The most expensive neighborhoods in this cluster are East Queen Anne, View Ridge and Leschi: all old, established neighborhoods with plenty of single-family homes. On the other hand, neighborhoods like Riverview, South Park and High Point present a similar mix of amenities, yet have much lower housing prices on average, and as such may be appealing to families looking at inexpensive homes with close proximity to parks.

It is interesting to look at the two outliers generated by the model.

Matthews Beach and Olympic Hills are two neighborhoods in North Seattle. WIth median house prices of 556,700 and 365,200, respectively, and relative proximity to the University of Washington, they seem to be attractive destinations for prospective homeowners. Both have access to parks, gyms and pools, and yet seem to be lacking in venues that distinguish neighborhoods in immediate proximity: e.g. coffee shops, bars and cafes. This present a potential opportunity for business owners looking to expand into new neighborhoods.

Conclusion

This model leaves ample room for improvement and refinement.

First of all, model accuracy can be significantly improved by better feature selection. Currently, all venue data from Foursquare is treated equally, without any attempt at categorization. It might make sense to separate venues into public amenities and transit, and commercial venues.

Second, a better algorithm to select venues by neighborhood can be implemented. Right now, the algorithm pulls all venues within a 500-meter radius of each neighborhood's centroid; it would be better to consider a more accurate approach:

- Create a grid of evenly spaced points (e.g. 100 meters apart) spanning the entire area of Seattle;
- Allocate each point to one neighborhood based on boundary data;
- Run a Foursquare query for each point, so that the entire area is covered by overlapping circles;
- Remove duplicate results;
- Assign venues to each neighborhood.

Third, incorporating other data sources (e.g. crime statistics and school ratings) would present a much better picture of each neighborhood's desirability.

Finally, expanding the model to the greater Seattle area (including Renton, Bellevue, Redmond, etc.) would greatly increase its practical value. However, these municipalities may not have neighborhood data as readily available as Seattle, so further research into data collection is needed.

Nevertheless, even in its most basic form, this model provides some interesting insights into the Seattle neighborhoods.