

Battle of cities

EUROPEAN SOFTWARE COMPANY WANTS TO EXPAND INTO THE U.S.

Introduction/Business Problem

The management of a successful European software firm decides that in order to further advance the company's market share, a new R&D office is needed in the United States.

Being European, they have little knowledge about the U.S: they can't decide which city to target. What they do know is that quality coding requires quality coders, and quality coders are only happy if they can work and live in an ideal environment.

What makes an environment ideal? Cities with high quality of life, low crime, good weather, good healthcare, low pollution all comes to mind.

The goal of this research is to use data to find cities within the United States best suited for being the location of the new R&D office, keeping in mind the factors mentioned above. A recommendation of neighborhood within the highest ranked cities should also be part of the final report: the neighborhood chosen should be rich in amenities for the developers to spend their free time and hard-earned pay: restaurants, cinemas, parks all come to mind.

Data

The following data sources should be considered:

Foursquare API¹

We'll use Foursquare for getting info about the venues in cities such as trending venues and venue exploration, where we'll input city neighborhood coordinates and receive a list of venues nearby.

Wikipedia

We'll use Wikipedia as our source for weather-related data, such as average temperature, average rainfall and so on. An example of the data we'll be working with:

¹ Endpoint documentation at: <https://developer.foursquare.com/docs/api/endpoints>

Saint Pierre and Miquelon	Saint-Pierre	-2.6 (27.3)	-3.2 (26.2)	-1.4 (29.5)	2.0 (35.6)	5.6 (42.1)	9.6 (49.3)	14.1 (57.4)	16.2 (61.2)	13.5 (56.3)	8.9 (48.0)	4.5 (40.1)	0.4 (32.7)	5.7 (42.3)
United States	Albuquerque	2.4 (36.4)	5.2 (41.4)	8.9 (48.1)	13.3 (56.0)	18.7 (65.6)	23.8 (74.9)	25.7 (78.3)	24.6 (76.2)	20.7 (69.3)	14.2 (57.5)	7.2 (44.9)	2.4 (36.3)	14.0 (57.2)
United States	Anchorage	-8.3 (17.1)	-6.6 (20.2)	-3.0 (26.6)	2.7 (36.8)	8.8 (47.8)	12.9 (55.2)	14.9 (58.8)	13.7 (56.7)	9.2 (48.6)	1.6 (34.8)	-5.4 (22.2)	-7.2 (19.0)	2.8 (37.1)
United States	Atlanta	6.3 (43.3)	8.4 (47.1)	12.3 (54.2)	16.4 (61.5)	21.1 (69.9)	25.2 (77.4)	26.9 (80.5)	26.3 (79.4)	22.9 (73.2)	17.1 (62.8)	12.0 (53.6)	7.3 (45.1)	16.8 (62.3)
United States	Austin	10.8 (51.5)	12.8 (55.0)	16.5 (61.7)	20.7 (69.2)	24.8 (76.6)	27.9 (82.2)	29.4 (85.0)	29.9 (85.8)	26.7 (80.0)	21.8 (71.2)	16.1 (61.0)	11.4 (52.5)	20.7 (69.3)
United States	Baltimore	0.8 (33.5)	2.4 (36.4)	6.8 (44.2)	12.4 (54.3)	17.6 (63.6)	22.8 (73.0)	25.3 (77.6)	24.3 (75.7)	20.2 (68.4)	13.7 (56.7)	8.3 (47.0)	2.9 (37.3)	13.1 (55.6)
United States	Boise	-0.4 (31.3)	2.5 (36.5)	6.9 (44.5)	10.4 (50.8)	15.1 (59.1)	19.7 (67.5)	24.3 (75.8)	23.7 (74.7)	18.3 (64.9)	11.6 (52.8)	4.4 (40.0)	-0.7 (30.7)	11.4 (52.5)
United States	Boston	-1.5 (29.3)	0.0 (32.0)	3.7 (38.6)	9.1 (48.4)	14.6 (58.2)	20.0 (68.0)	23.2 (73.7)	22.4 (72.4)	18.4 (65.2)	12.4 (54.3)	7.2 (45.0)	1.7 (35.0)	10.9 (51.7)
United States	Charlotte	5.1 (41.2)	7.2 (45.0)	11.3 (52.3)	15.8 (60.5)	20.3 (68.5)	24.7 (76.4)	26.4 (79.6)	25.8 (78.4)	22.2 (71.9)	16.3 (61.3)	11.1 (51.9)	6.3 (43.4)	16.1 (60.9)
United States	Chicago	-5.6 (21.9)	-2.9 (26.7)	5.4 (41.7)	12.6 (54.6)	16.5 (61.7)	21.8 (71.2)	25.3 (77.6)	24.8 (76.7)	18.4 (65.1)	13.3 (56.0)	5.3 (41.5)	-5.3 (22.5)	10.8 (51.4)

Figure 1 Average temperature data ²

Kaggle

The Movehub City Rankings³ will be one input. This dataset has valuable features such as Health Care index, Pollution index and a more general and ambitious-sounding “Quality of Life” index for major cities around the world. We’ll just concentrate on U.S. cities.

The “Hurricanes and Typhoons, 1851-2004” dataset⁴ will be important as well, Europeans are afraid of such extreme weather phenomena, and high frequency of these can be a reason for vetoing a candidate city.

Additionally, we’ll grab a GeoJSON file from the internet

All datasets shall be first explored via descriptive statistics and data visualization techniques and then cleaned and transformed into a format where machine learning algorithms can take them as input for deriving models that can help quantify differences, show similarities between cities.

² Source: https://en.wikipedia.org/wiki/List_of_cities_by_average_temperature

³ <https://www.kaggle.com/blitzr/movehub-city-rankings>

⁴ <https://www.kaggle.com/noaa/hurricane-database>

Methodology

We will be using Python in the Jupyter Notebook environment. Packages pandas and numpy are standard in this scenario, The sklearn package will help with data preprocessing/cleaning, model selection and modeling. We will use the folium package to generate maps to visualize some of our findings. No special package will be needed for communicating with Foursquare, we'll use simple HTTP via requests to exercise its REST API.

At the first part of our research we will look at climate: using the average temperature data mentioned above we will collect the necessary data points, and format the resulting pandas dataframe so that we can turn to machine learning models.

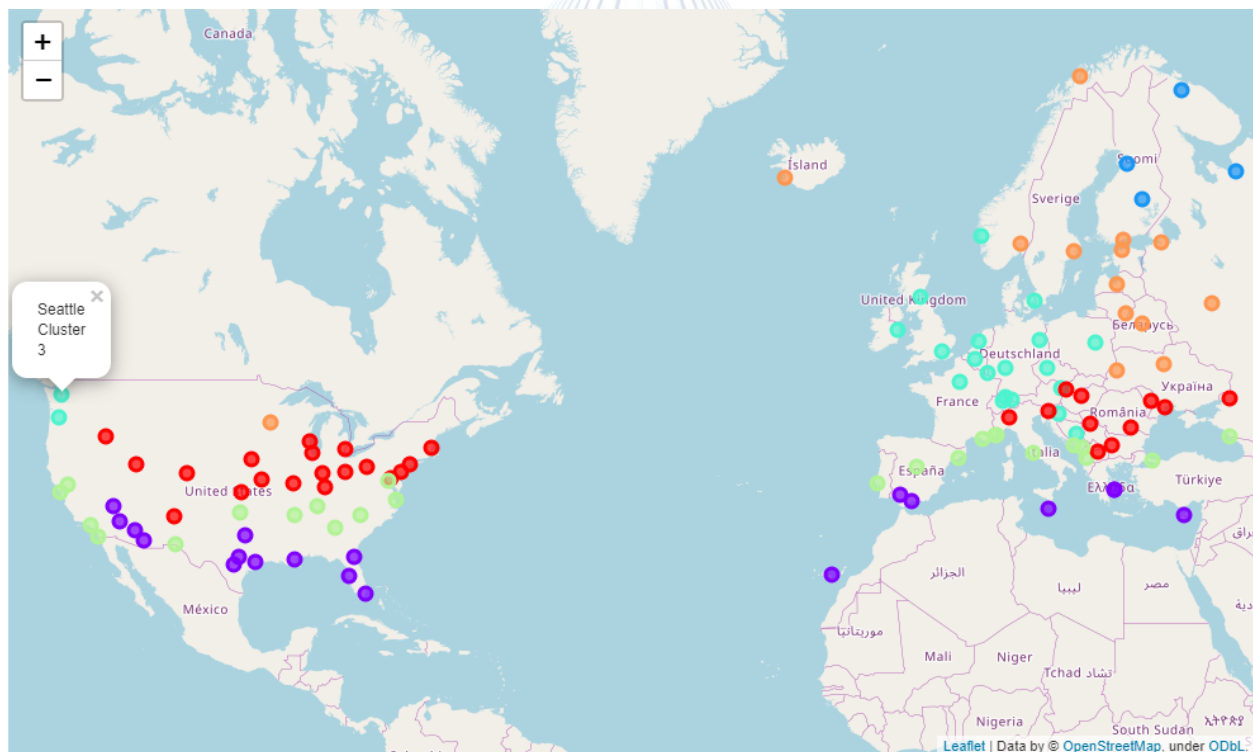
The model first used is K-Means Clustering, that puts similar data points together into a cluster. In our case, the data points are U.S. and European cities, and forming the clusters will provide key insight on the climate of U.S. cities for an European audience.

This will allow us to narrow down the set of candidate cities. This narrow list will be further probed using the Movehub City Rankings, and one top candidate will be selected.

Having selected the city, we will dig deeper: analysis of the neighborhoods will follow. Foursquare will be used to fetch a set of venues and we will visualize their frequency across neighborhoods using choropleth maps.

Results

K-Means Clustering identified the following clusters, each cluster shown with a distinct color:



Cluster 0 (red) is Central-Eastern Europe and many U.S. cities from Salt Lake City through New York until Boston. **Cluster 1** is the South-Mediterranean from Seville, Malaga on the West to Nicosia on the East, plus the Canaries. In the U.S. most southern cities (except California) also fall into this cluster which might be surprising. **Cluster 2** is snow and ice: Oulu, Murmansk, Arhangelsk in Europe, Alaska in the U.S. **Cluster 3**: A large part of Western Europe belongs here, the UK, Northern France, Germany. Interesting that very few cities in the U.S. have this climate: only Seattle and Portland in the North-West. **Cluster 4** is nice weather: the North Mediterranean plus Istanbul, Lisbon and Madrid in Europe, California, and the cities north to Cluster 1 and south to Cluster 0 in the US from El Paso through Memphis to Virginia Beach. **The last cluster (#5)** is Scandinavian cities Reykjavik, Oslo, Stockholm plus the cold Northeastern parts of Europe: the Baltic states, Ukraine, Belarus and Moscow. Not surprising that the U.S. doesn't have many entries here: only Minneapolis.

This allowed to narrow down candidates to **Cluster 3** and **Cluster 4**, namely the following cities:

- Portland, Oregon
- Seattle
- Sacramento
- San Francisco
- Los Angeles
- San Diego
- El Paso
- Oklahoma City
- Memphis
- Nashville
- Atlanta
- Charlotte
- Washington D.C.
- Virginia Beach

Next, we selected these cities in the Movehub city rankings, and ordered them by “Quality of life”:

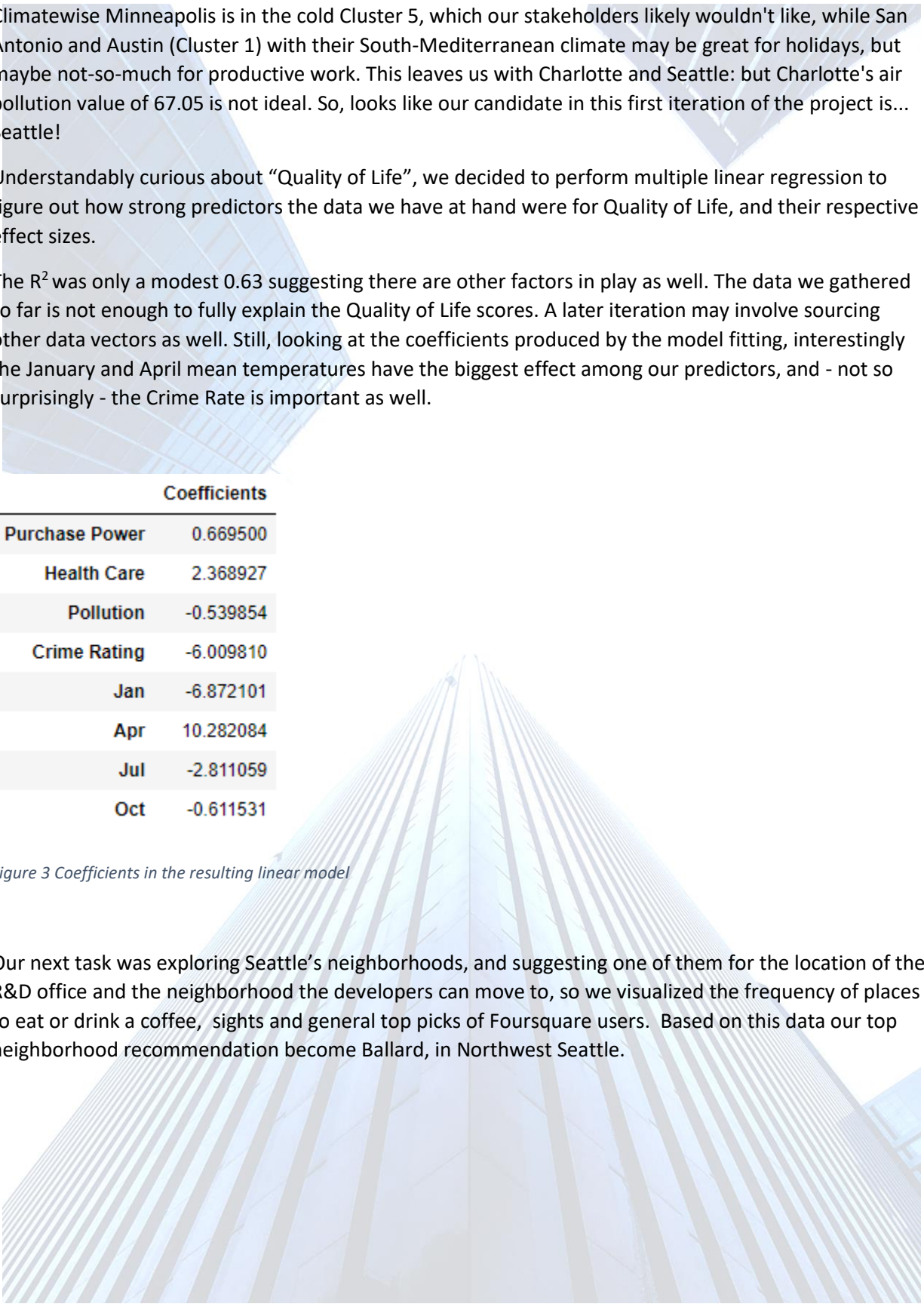
City	Movehub Rating	Purchase Power	Health Care	Pollution	Quality of Life	Crime Rating
Austin	84.86	69.22	73.61	28.84	86.51	42.50
San Antonio	83.76	74.78	60.97	59.19	84.88	51.41
Charlotte	84.46	77.18	72.08	67.05	84.39	30.21
Seattle	85.38	78.46	75.46	32.90	84.10	42.03
Minneapolis	83.47	69.91	62.35	77.94	83.79	40.36

Figure 2 Movehub city rankings data for candidate cities

Climatewise Minneapolis is in the cold Cluster 5, which our stakeholders likely wouldn't like, while San Antonio and Austin (Cluster 1) with their South-Mediterranean climate may be great for holidays, but maybe not-so-much for productive work. This leaves us with Charlotte and Seattle: but Charlotte's air pollution value of 67.05 is not ideal. So, looks like our candidate in this first iteration of the project is... Seattle!

Understandably curious about “Quality of Life”, we decided to perform multiple linear regression to figure out how strong predictors the data we have at hand were for Quality of Life, and their respective effect sizes.

The R^2 was only a modest 0.63 suggesting there are other factors in play as well. The data we gathered so far is not enough to fully explain the Quality of Life scores. A later iteration may involve sourcing other data vectors as well. Still, looking at the coefficients produced by the model fitting, interestingly the January and April mean temperatures have the biggest effect among our predictors, and - not so surprisingly - the Crime Rate is important as well.



Coefficients	
Purchase Power	0.669500
Health Care	2.368927
Pollution	-0.539854
Crime Rating	-6.009810
Jan	-6.872101
Apr	10.282084
Jul	-2.811059
Oct	-0.611531

Figure 3 Coefficients in the resulting linear model

Our next task was exploring Seattle’s neighborhoods, and suggesting one of them for the location of the R&D office and the neighborhood the developers can move to, so we visualized the frequency of places to eat or drink a coffee, sights and general top picks of Foursquare users. Based on this data our top neighborhood recommendation became Ballard, in Northwest Seattle.

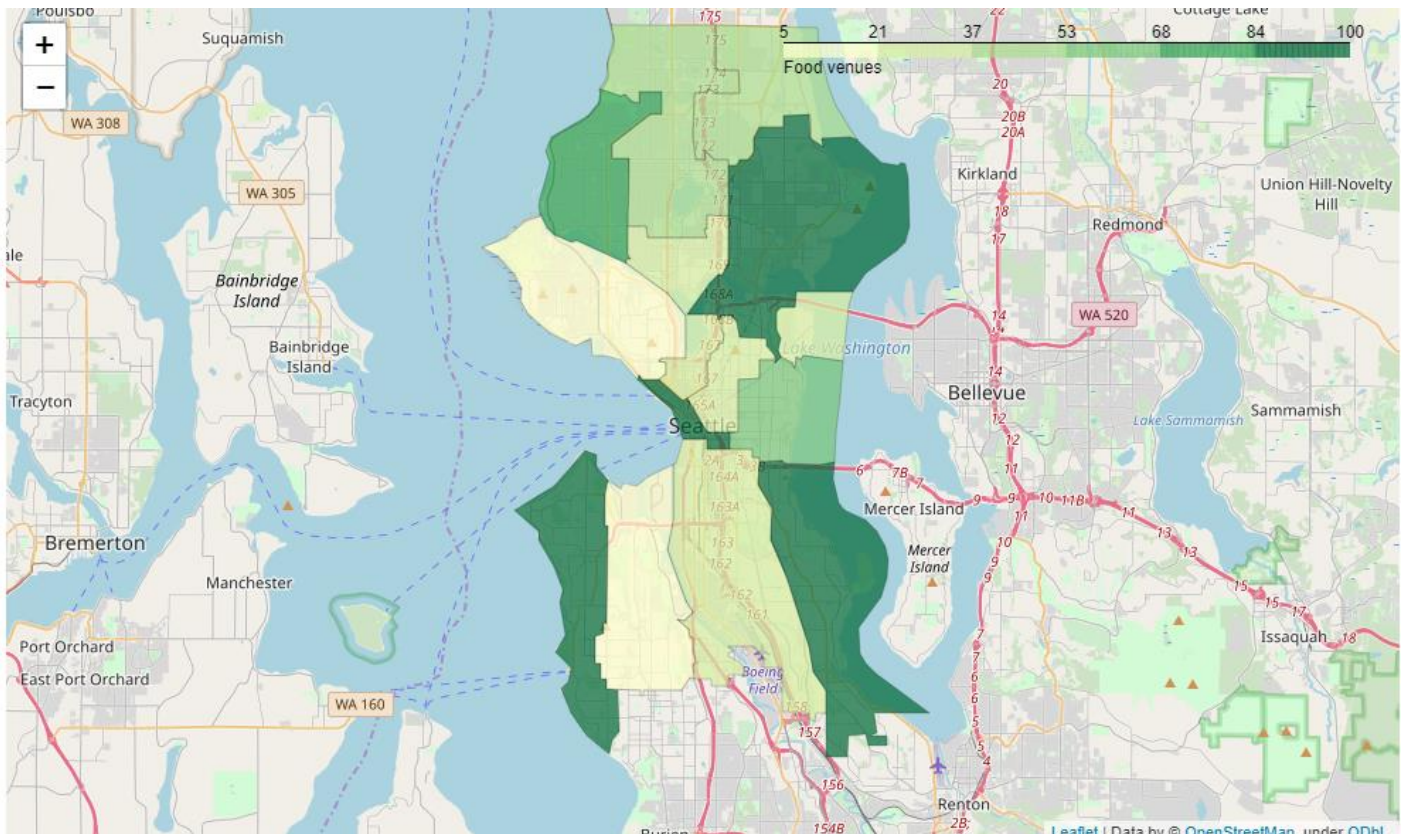


Figure 4 Seattle: Places to eat by neighborhood

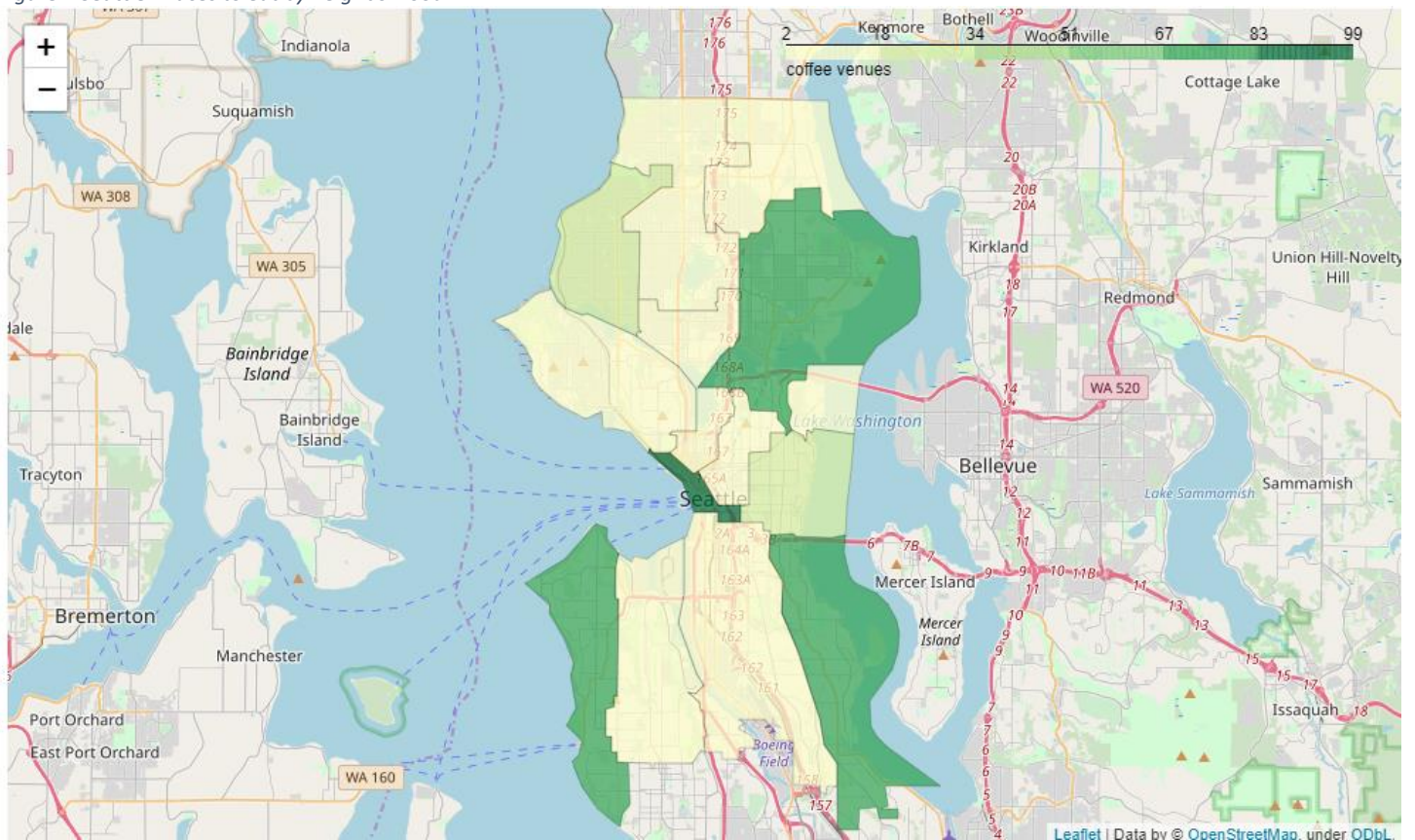


Figure 5 Seattle: Coffee shops by neighborhood

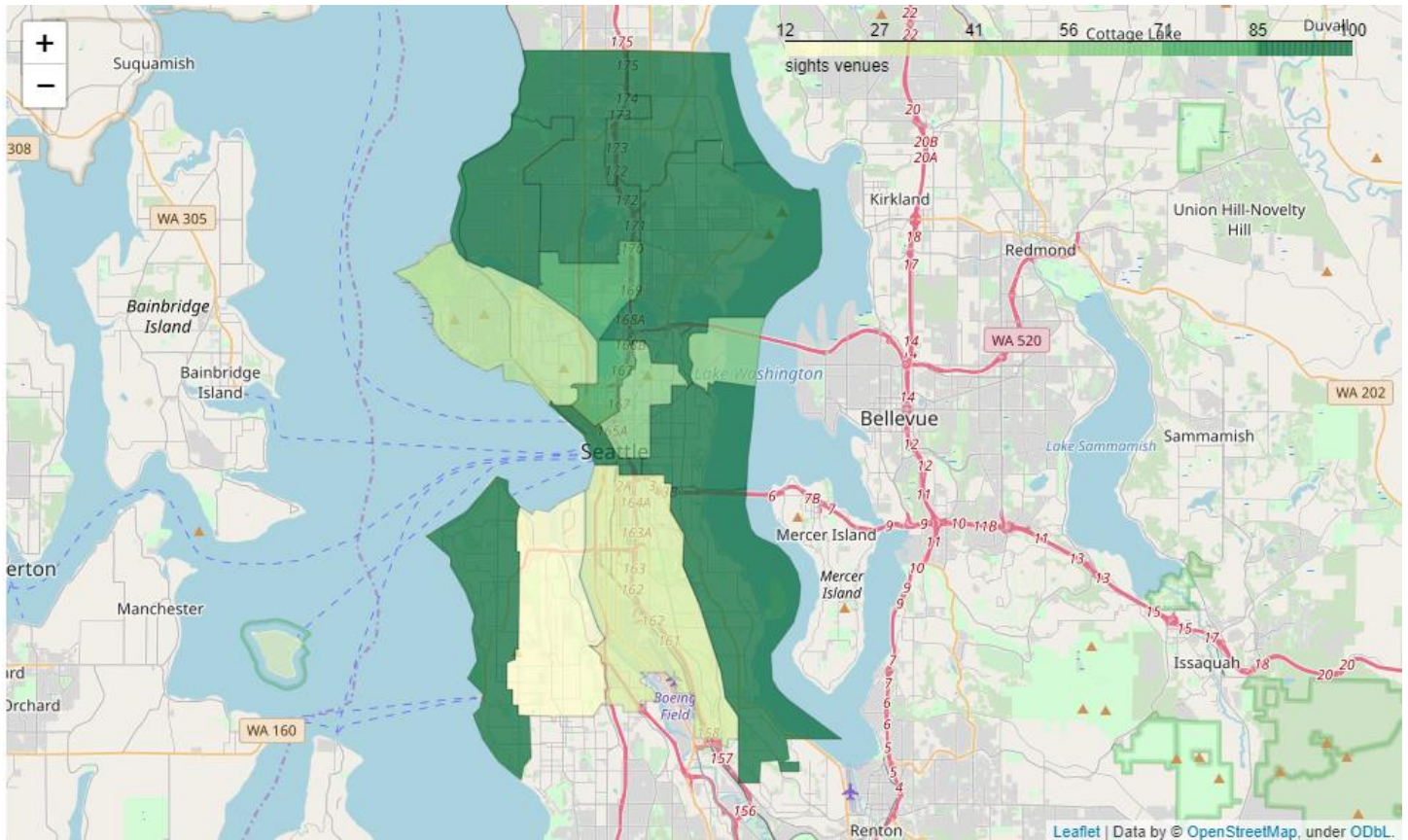


Figure 6 Seattle: notable sights by neighborhood

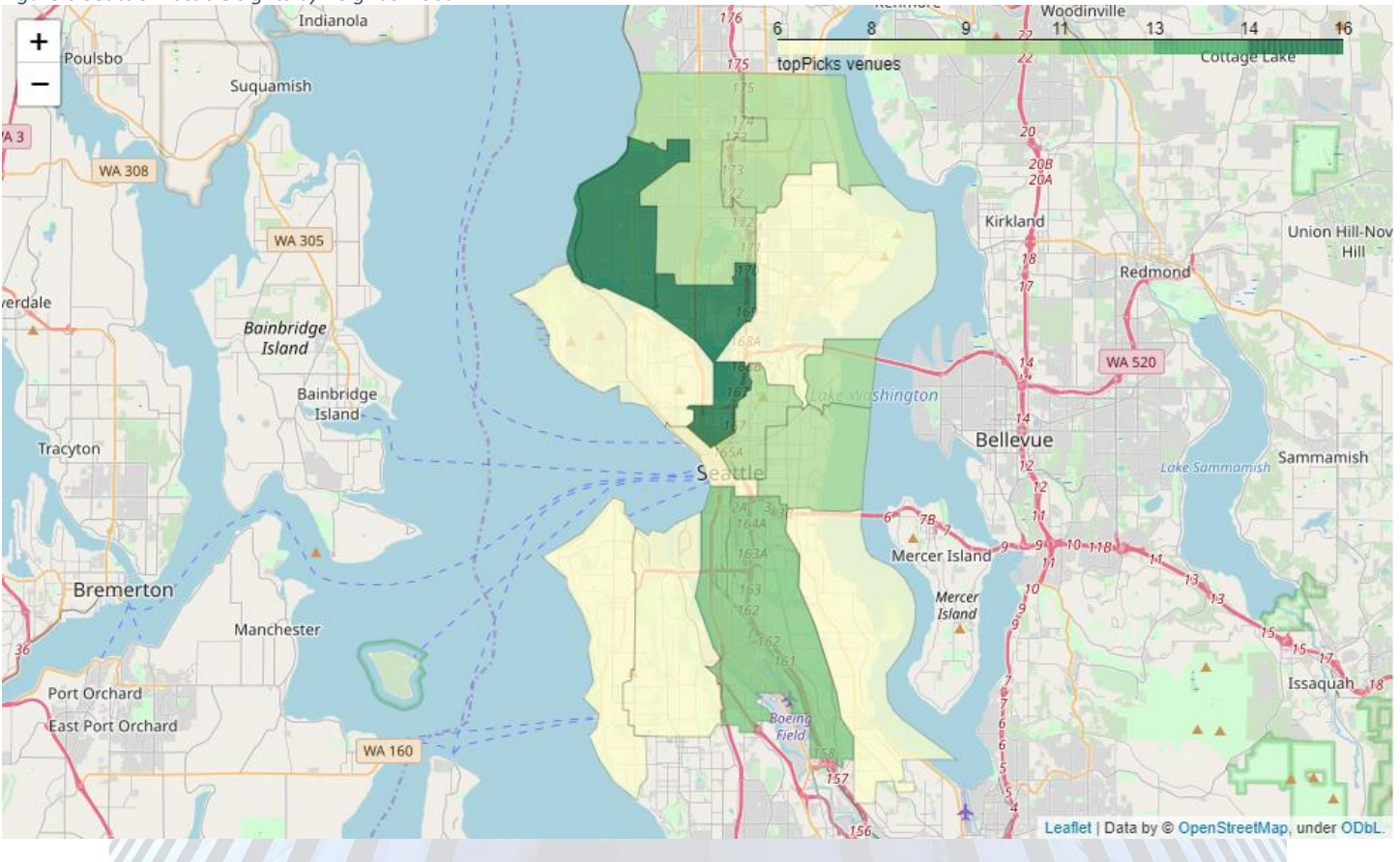


Figure 7 Seattle: Number of 'Top Picks' of Foursquare users in each neighborhood

Discussion

Providing rigorous recommendations on such business problems is always a challenge: what data sources to consider, how to weigh results on different aspects of a city. An obvious limitation preventing surefire application of Machine Learning techniques is the relative scarcity of data, for example we only had a few dozen data points to work with when trying to model “Quality of Life”. Still the insights gained during this research, though far from domain-complete, are hopefully helpful in steering the stakeholders into a right direction when they decide on their choice of location for their new office.

Conclusion

This research iteration ends with recommendation of Seattle. Feedback is welcomed and if there’s a need for additional considerations, modeling can be refined to provide a more complete picture.

