Where to Open What Type of Restaurant

IBM Data Science Professional Certification: Final Assessment

Nik Caine

16 JUN 2020

Introduction

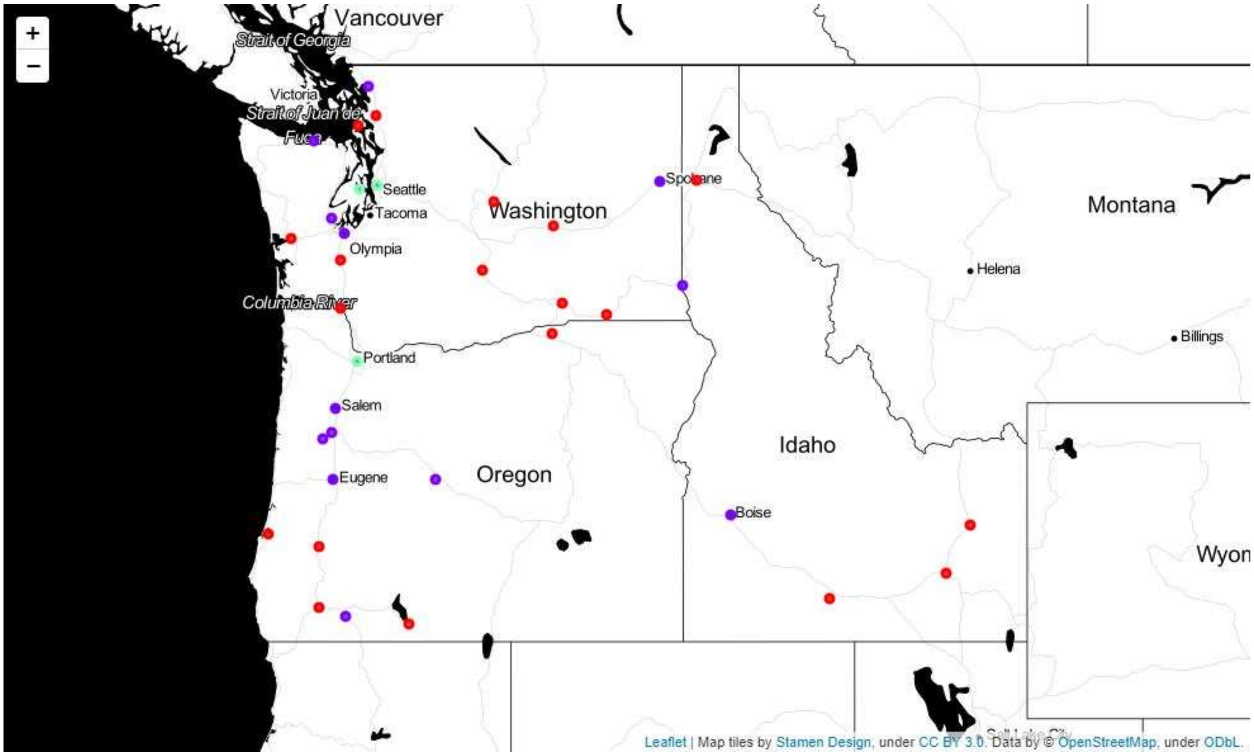
Restaurants are a competitive and risky business; 60% of restaurants close in their first year and 80% within five years of opening. A tool that could help assess the viability of a restaurant, or even recommend where and what type of restaurant to open, would be invaluable to the right client.

Use-case: The client is a restauranteur who wants to open a restaurant in the Pacific Northwest region (PNW: Washington, Oregon, Idaho) of the United States. The client needs to know what market gaps exist in the region’s various metropolitan areas.

The final deliverable to the client will be a recommendation of three underserved metro areas, the restaurant types with the largest market gap in the respective metros, and appropriate locations for that restaurant type within the metros.

Data

Three main sources of data were used. Location data and restaurant types were provided via Foursquare’s Places API. Geolocation was facilitated by ArcGIS. Demographic data as well as the names and states of the different metro/micropolitan areas were obtained from the US Census Bureau.



Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under ODbL.

Figure 1: A map depicting the metro areas of the PNW, clustered according to their ratio of different restaurant types. Each of the three clusters are represented by a different color.

Methodology

The US Census Bureau is a trove of accessible data on various topics, some of which were used in deciding which metro areas to recommend to the client. A CSV file containing the cities, states, and demographics of each metro area in the US was downloaded from the bureau's website and cleaned in MS Excel and using the Python Pandas library. Only metros with the state codes WA, OR, and ID were used. The latitude and longitude of the cities in the different metros was obtained from ArcGIS, accessed through geopy, by feeding in the names and states of the cities from the above CSV file. Foursquare's Places API used the geolocations of the cities, to sweep a 25km radius around the geolocations to search for different restaurants. Since the API can only return 50 venues per call (see also, Discussion: Shortcoming #1), one API call was made per restaurant type, per city, per metro. The name, restaurant type, and geolocation of each restaurant in the radius was extracted from the JSON that the Foursquare API returned. Fast food restaurants and coffee shops were deliberately left out of the study, as while they technically can be considered restaurants, they serve a different market space.

The frequency of each restaurant type in each metro was calculated and scikit-learn's MinMaxScaler was used to normalize the restaurant frequencies of each metro. Using the above restaurant-type counts, scikit-learn's k-means clustering algorithm clustered the metros, returning three groups of metros with similar tastes in restaurants (see Figure 1).

The dissimilarity of the frequency of restaurant types within the clusters was then used to determine the size of any market gap within the metro. First, the mean and standard deviation of each restaurant type in each cluster was calculated. The amount of standard deviations between the frequency of each restaurant type in each metro and the respective cluster's restaurant type mean

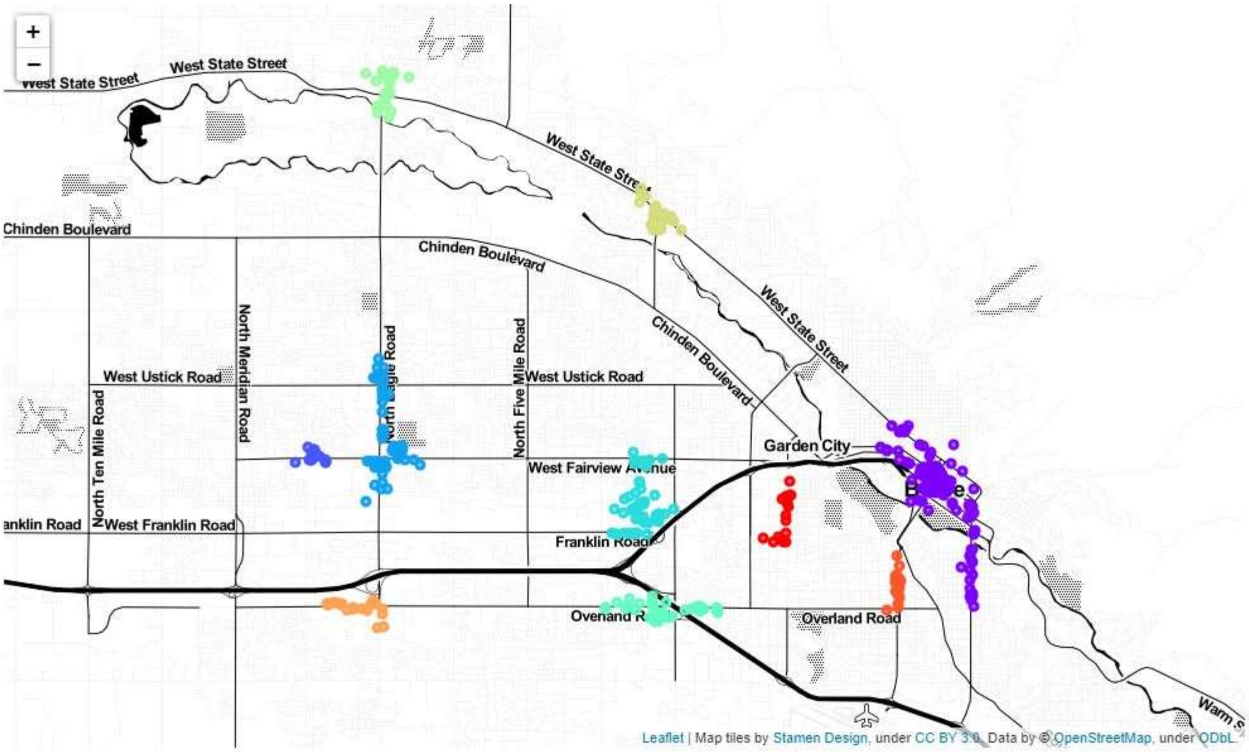


Figure 2. A map showing the geographic clusters of restaurants in Boise, ID.

was then calculated. The more negative the number, the greater the underrepresentation of that restaurant type in that metro, compared to metros with similar restaurants.

The adult population (at least 16 years old) to total restaurant count ratio was calculated. This ratio indicates what metros are most underserved by the restaurant industry in general. The most underrepresented restaurant type from each of the top three underserved metros were found. The restaurants in those metros were then clustered with scikit-learn's DBSCAN function, using only geolocation data as inputs. The DBSCAN method was used because it handles non-random distribution of clusters (like restaurants distributed along a street) very well. The centroid of those clusters was then determined. This provided a general area where would-be patrons expect to find a restaurant, and were restaurants are already shown to stay operational.

The clusters that didn't contain the underserved restaurant type were noted. The distance between the centroid of these clusters and the nearest restaurant of the underserved type was calculated. The location of the metro's clusters without the underserved restaurant type would be given to the client as a recommended location to open a restaurant of the underserved type.

Results

Table 1: A table showing the 3 most underserved metro areas in the PNW, according to their respective population/restaurant-count ratios. The column on the right shows the most underrepresented restaurant type in that metro.

Underserved Metro Area	Adult Pop. to Total Restaurant Count	Most Underrepresented Restaurant Type
Seattle-Tacoma-Bellevue, WA	790.6	Pizza Place
Moses Lake, WA	764.0	American Restaurant
Boise City, ID	650.7	Sandwich Place

Table 2. Showing the geolocation of the general area recommended to the use-case's would-be client for the opening of the metro's most underrepresented restaurant type.

Metros, restaurant type	Recommended General Location of Restaurant	Distance To Nearest Restaurant Type (meters)
Seattle-Tacoma-Bellevue, Pizza Place	47.667749, -122.38402	6114
	47.661572, -122.31382	4346
	47.629391, -122.14545	4566
	47.673930, -122.12157	8728
	47.454746, -122.25918	5668
Moses Lake, American Restaurant	47.124907, -119.29006	4337
	47.130477, -119.27510	3961
	47.140721, -119.27757	5014
Boise, Sandwich Place	43.692205, -116.35285	5904
	43.619767, -116.37227	2569

Table 1 shows the Seattle-Tacoma-Bellevue metro area has the most adults per restaurant, making it the most underserved metro in the PNW. Second and third are Moses Lake and Boise respectively.

Using the above methodology, Figures 3 to 5 in the appendix show the location recommendations for each metro’s most underrepresented restaurant type. The red dots show the location of existing restaurants of the underrepresented type. The blue circles show a recommended general area to open such restaurant. It should be noted that each of the three analyzed metros had clusters that did not contain a restaurant of the underrepresented type. If no such cluster existed, the centroids of the clusters with the fewest restaurants of that type would be used to make a location recommendation.

Discussion

After cross checking the results with a Google Maps search, the results are as accurate as can possibly be known without actually visiting the locations in person. While the tool was successful in recommending locations and restaurant types to the would-be client, some shortcomings are noted below.

An interesting point for possible future analysis would be that while the k-means clustering of the metros used only the proportions of restaurant types in each metro, there is a correlation between the clusters and population size of the metros. This may indicate the diversity generally associated with larger cities is reflected in their restaurant types. It may also be affected by shortcoming #1 below.

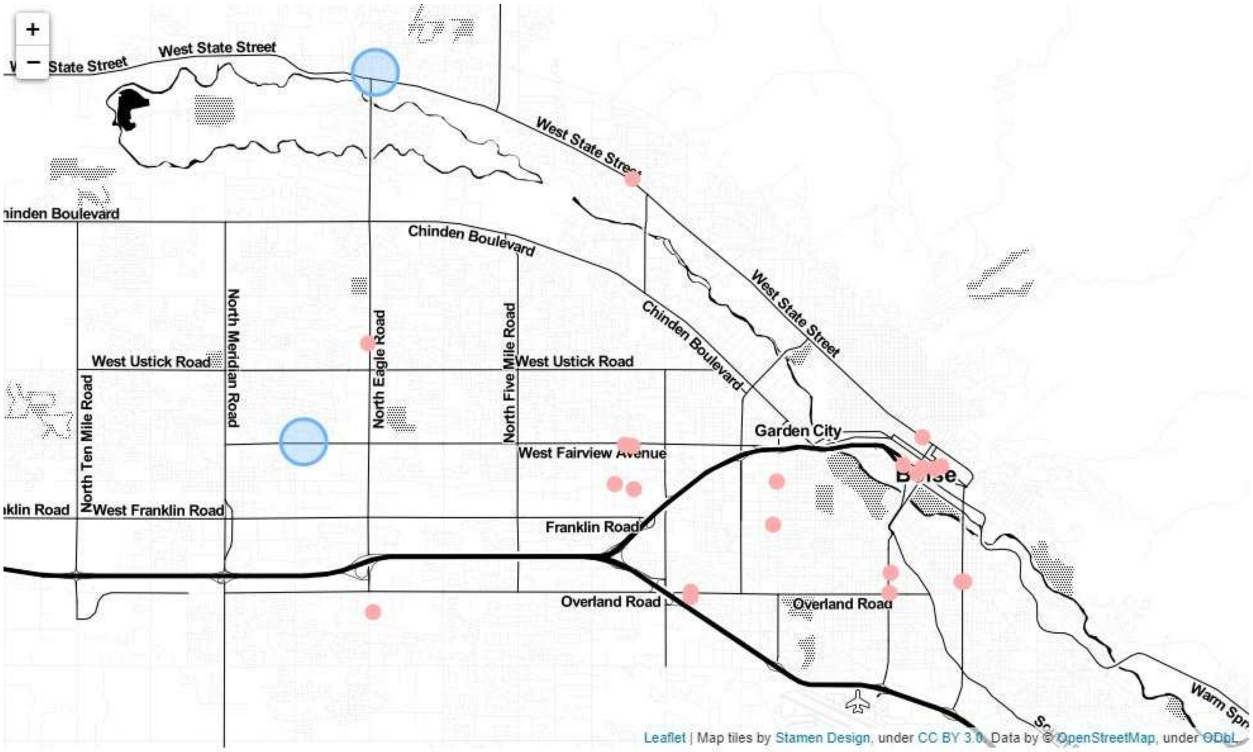
Shortcomings and possible improvements

1. *(EDIT: 17 JUN 2020. Fixed)* Foursquare's Places API can only return a maximum of 50 venues per API call. This means popular restaurant types in large metros could be drastically undercounted. This has at least two major ill-effects on the study. Firstly, the tool's output of Pizza Places being underrepresented in the Seattle-Tacoma-Bellevue metro area (STB) has a strong chance of being inaccurate. Secondly, the population to restaurant count ratio, used to identify the STB metro as underserved by the restaurant industry, would tend to be higher in large metros where so many popular restaurant types are undercounted. It's no coincidence that STB is the most populous metro in the region. This renders the recommendation of Pizza Place in STB to be unreliable. But, since Boise and Moses Lake are smaller metros and are clustered with other smaller metros, this shortcoming doesn't have a strong effect (if any) on their.
   One solution would be to simply use multiple locations per city instead of just the one, when the API returns 50 venues of a particular restaurant type. This solution would be easily implemented using the geopy library and will most likely be in the next version of the tool.
2. One cluster of metros only contained three different metros (STB, Portland-Vancouver-Hillsboro, and Bremerton-Silverdale). A dataset of three is not large enough for the purposes of this study.
   A possible solution would be to include metros from neighboring states to create the clusters of metros with similar proportions of restaurant type (see Figure 1).
3. The tool relies on Foursquare's data, and thus Foursquare's restaurant types. While their list of different restaurant types is extensive, not all niche venues will have a specific category. For example, Meat Pie shops would typically be categories as "English Restaurant" or "Bakery". Neither of which is as a precise definition as it could be.
4. Sticking with the relying-on-Foursquare's-data theme, some restaurants were clearly miskeyed. McKenzie River Pizza, Grill & Pub (a chain in the PNW) was often labeled as an "African Restaurant". The author of this report and creator of the tool, being both African and familiar with the restaurant was in a unique position to spot the obvious flaw. Similarly, many restaurants that were clearly nothing but pizza joints, many being well known chains, were categorized as "Italian Restaurants". This called for a manual edit to the collect data. Other categories may also suffer from this issue. Checking each restaurant individually would be impractical, but crosschecking the data with a different service, like Google Maps, would help reduce the occurrence of this error.

Conclusion

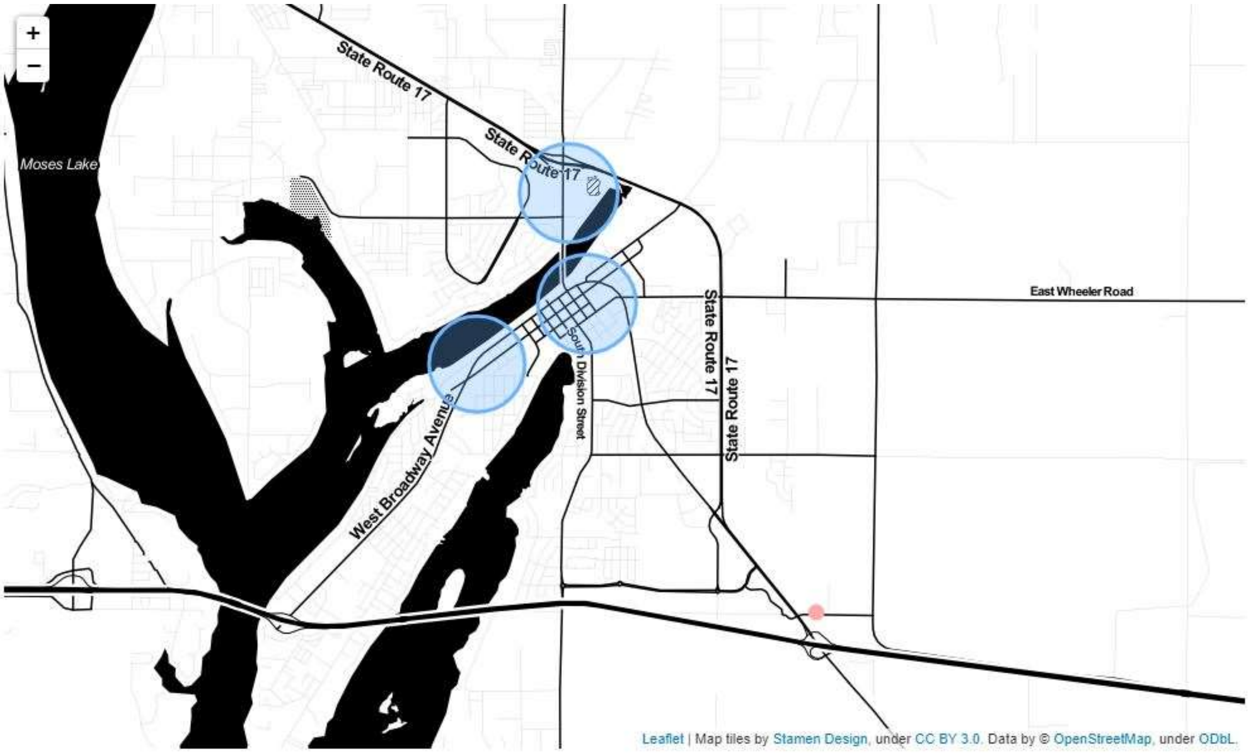
With the above improvements, this would be a useful forward analysis tool for entrepreneurs and big business alike. Especially in a large, diverse country like the US. While this project has focused on restaurants, the tool could be used to help decide the location of any brick-and-mortar, customer facing establishment.

Appendix



Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under ODbL.

Figure 3. A map showing all the sandwich places in Boise as pink dots. The blue circles represent the recommended areas to open a sandwich place.



Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under ODbL.

Figure 4. A map showing all the American restaurants in Moses Lake as pink dots. The blue circles represent the recommended areas to open an American Restaurant.

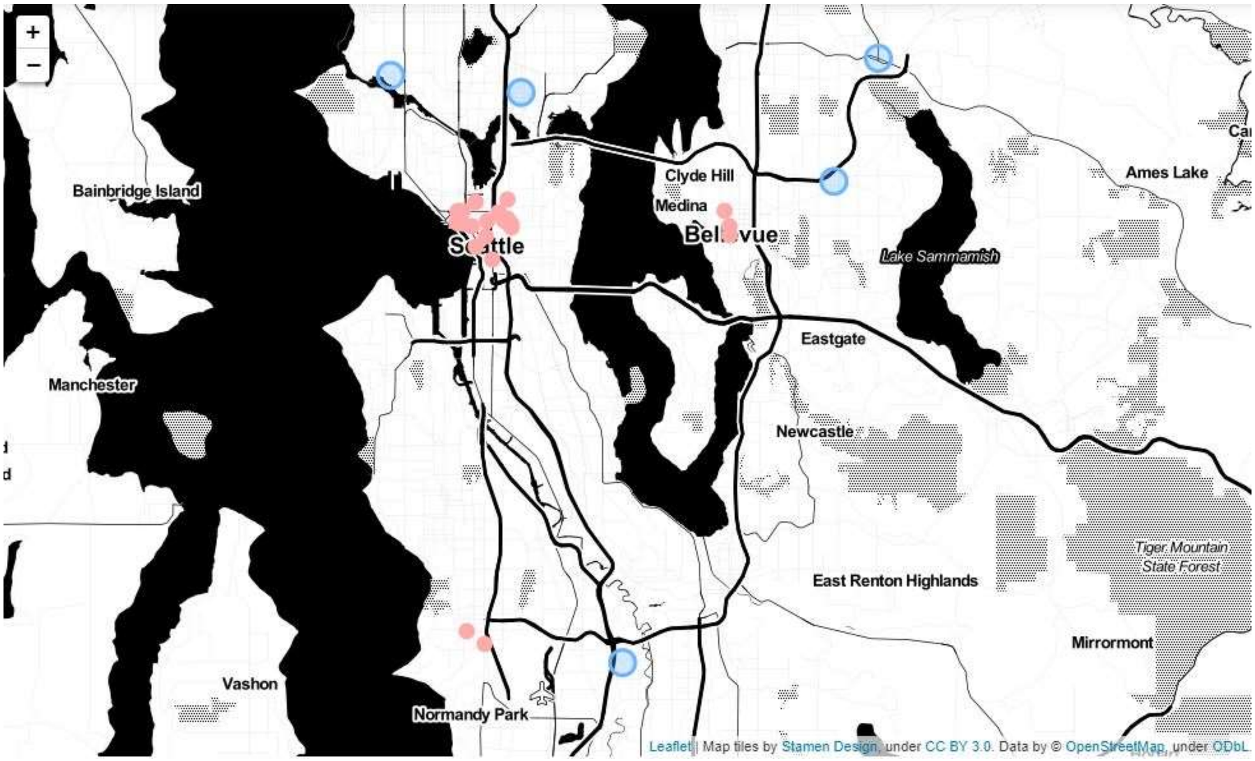


Figure 5. A map showing pizza places in the Seattle-Tacoma-Bellevue as pink dots. The blue circles represent the recommended areas to open a pizza place. See Discussion: Shortcoming #1.

The project's notebook can be found at *https://github.com/NikCaine/IBM_DS_capstone_final/blob/master/where_to_open_what_restaurant.ipynb*