Elliott Rose

1-24-20

Oakland Zip code analysis for Site Location and Marketing of Wellness Spa

Introduction

For this report, I will be utilizing data to research the business environment in Oakland, California for a friend who is starting a wellness spa. To assist her in starting this business, I will use Foursquare venue and location data to examine the Zip codes in Oakland to find business environments that would be fitting for her first location. Further, I will use demographic data scraped from the web detailing the makeup of the zip codes and neighborhoods in Oakland. With this data, we'll be able to examine the demographic makeup for marketing to her target audience. This paper will detail the process of wrangling, cleaning, visualizing, and analyzing this data.

Background

The wellness spa will provide services like yoga and meditation classes, as well as homeopathic products. The target audience is primarily people between working age adults, majority female, with household incomes greater than $60,000. I will assume that areas with makeups similar to this demographic would be optimal for location placement and marketing. Further, I believe it would be beneficial for location placement to find areas that contain venues that have a similar recreational and health-conscious tilt to provide a synergistic environment for our location. These commonalities could provide for walk or drive-by customer opportunities that may not be as present in areas with disassociated business profiles.

Data

I will be sourcing data from two sources: the Foursquare API for location and venue data, and zip-codes.com for demographic data. Zip-codes.com does not have an API, but stores the data in html tables, which I will use to scape the data using Beautifulsoup with Python. Each zip code is stored on a different page so I will have to use a loop to scrape zip code page to collect the data. From zip-codes.com, I will collect population, gender profile, average income, housing value, persons-per-household, and households per zip code.

The Foursquare API returns venue and location data that will be relevant to our search. We want to understand the types of venues that are in each zip code to get a picture of the business environment. For example, we would want to avoid and area that is mostly industrial warehouses, or fast-food type restaurants. The location data that is provided by the Foursquare app is particularly useful as it returns the mapping coordinates of each location, which will allow us to plot the venues on the map to visualize each zip code.

Methodology

As previously specified, I used Beautifulsoup to scrape the demographic data from the web. To make this more easy to accomplish, I created a for-loop to iterate over a list of Oakland zip codes, adding the current zip code into the web address, and then scraping the data fo the new and related page. This allowed me to scrape the data from each appropriate zip code for the Oakland area, and then store that data in Python lists. I then collected these lists into another list, creating a 2D array that made it easy to convert the information into a Pandas data frame for analysis and processing.

I initially found 4 zipcodes that had missing data, which some further googling allowed me to find that these were cancelled and out-of-use zip codes. With some further manipulation to acquire only the fields that were required for analysis, I was then left with a data frame of 21 columns and 21 rows.

Next, I set up the Foursquare API to pull the venue information about the zip codes in question. This data is retrieved in JSON, allowing me to filter the categories that were most relevant to the analysis. After filtering, I ended up with 7 columns (zip, zip latitude, zip longitude, venue name, venue category, venue longitude, and venue latitude) and 450 rows.

Next, I wanted to set up a data frame that would give me the 10 top venue types in each zip code, that would allow me to look for the most appropriate neighborhoods for our proposed venue. This was done by one-hot encoding the venue types and then locating the ten highest occurring venue types in each zip code.

The next step was to cluster the zip codes so as to find our most similar neighborhoods and give us insight into possible locations and customer types. I found that clustering the zip codes into five clusters yielded the most useful results, accomplishing this task through trial and error. Once clustered, I merged the cluster labels to our 10 top venues data frame to map the clusters onto a map of Oakland using Folium. This initial analysis pointed out helpfully that cluster 1 was not only curiously full of the identical top 10 venues, but also missing from our map. Further investigation revealed that these zip codes were actually P.O. Boxes, which would be unhelpful for our analysis. Also, I found that the most suitable areas were found in clusters 2,3, and 4, with the majority venues consisting of food, recreation, and transit.

I then performed clustering on the demographic data to compare with the venue clusters. This quickly helped me isolate areas that would not be appropriate for our studio. As this studio is

primarily targeted at higher earners, with household incomes greater that $60,000 dollars a year, cluster number 1 quickly captured the highest earning areas, leaving us with 4 zip codes with average household incomes greater than $60,000 with high household values (greater than $650,000), and significant populations (greater than 9,000 people). Finally, I created a new row for our data frame that divided our average household salaries by the average number of occupants in the household, to identify areas that may have lower incomes but be hidden because their average household is high, while having a high number of household members. I then mapped this data, finding that one of our cluster 1 zip codes would not be suitable as it was located too far out of range in Pleasanton, leaving us with 3 appropriate zip codes found in the south eastern region of Oakland.

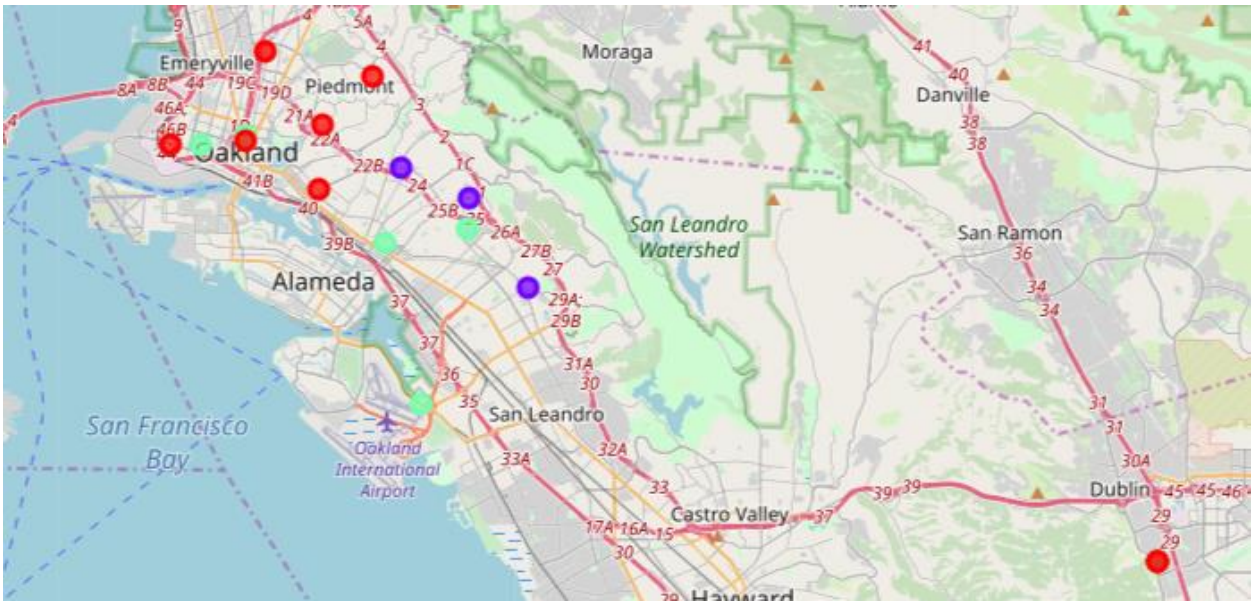


Figure 1 Note the three purple points, these are our high income, well-populated areas.

Finally, with these three zip codes identified, I was able to map the local and competing venues with Folium. Folium allows the user to move the map and zoom in on areas in question. I

also added venue data to the venue markers, so the user can click on the local venue to find its zip code, neighbor hood, venue name, and venue type. Then I created bar graphs to visualize the make up of the neighborhoods found to be most suitable for our new venue's location and marketing.

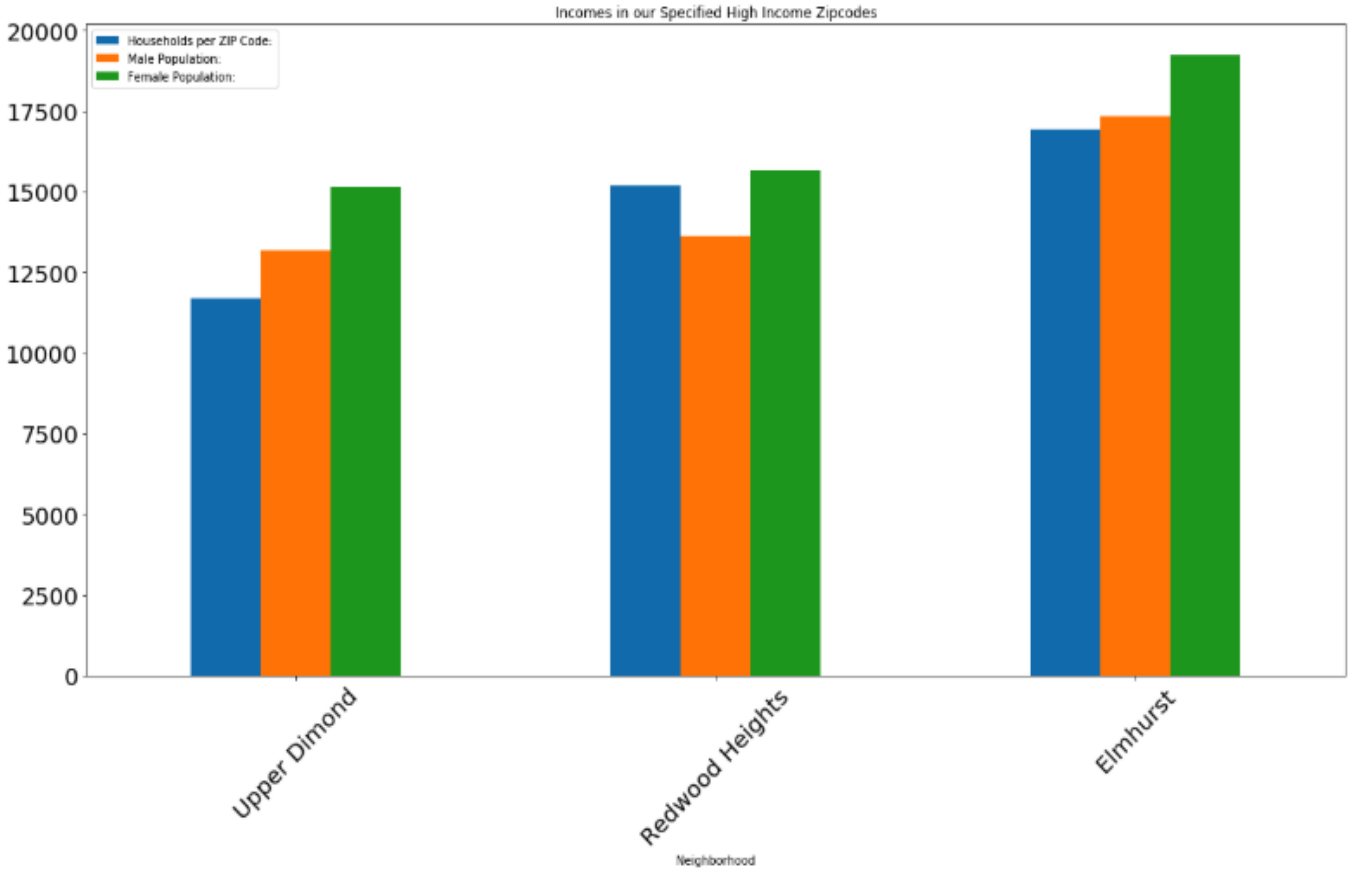
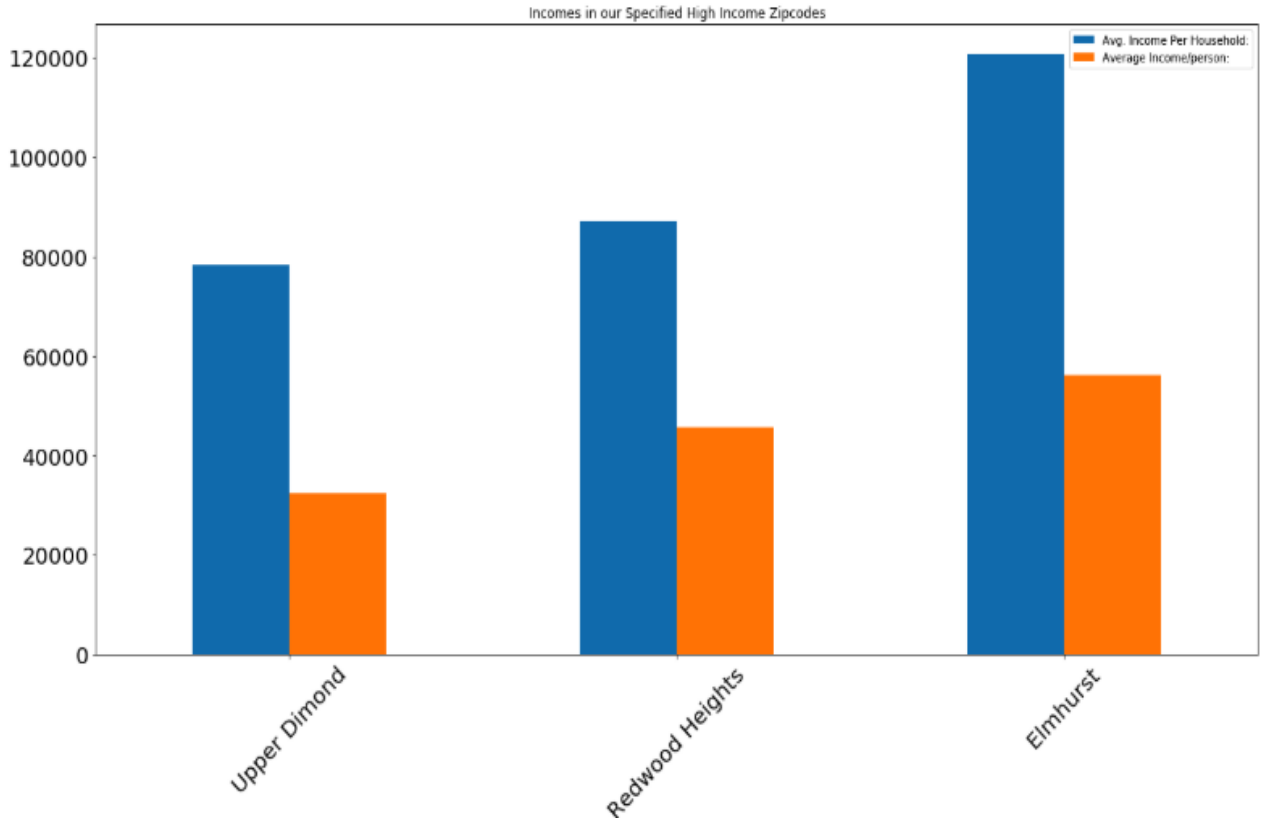


Figure 2 Well populated, wealthier neighborhoods.

Results

In the end, I found the most suitable neighborhoods for the new wellness center to be Upper Dimond (correct spelling), Redwood Heights, and Elmhurst. These areas all had significant populations, local venues that could be conducive the success of our wellness studio, and populations that had incomes sufficient to fit our potential customer profile. It is worth noting, that there is one other possible option that did not appear in our clustering due to its lack of population

density. The Temescal area is large business environment bordering on the north western side of the city, bordering Emeryville that could also prove promising.



Discussion

This analysis was done using the Foursquare API and web scraping technologies. Because Oakland is a smaller city, the amount of data for the analysis was not restrictive and could be performed easily on a local computer. For the future, larger environments could also be tackled using similar techniques, extending the boundaries of our area in question to larger geographical areas like the Bay Area, or even state or country wide. Data quality is one of the largest limiting factors in this type of analysis, with lack of data, or insufficient data quality being some of the biggest limiting factors. In this analysis, I was unable to find coordinated data for the individual neighborhoods in Oakland, which limited me to solely utilizing the zip code data for these areas.

It should be noted that there are many factors that guide the success of a new business, and that the analysis performed here only provides information that may be useful to the individuals or groups interested in starting the new business. Many factors contribute to the success and some are difficult if not impossible to identify. Further analysis of traffic patterns, area safety, pricing, and community development would also be useful in creating a more filled out picture of any potential environment proposed for the start of the new business. All that being said, the analysis here should provide a leg up for the potential business creator.

Conclusion

In conclusion, the analysis performed here has given us insight that out of the 21 zip codes analyzed, there are 3 promising possibilities available. These areas have local businesses that can provide the right environment for a synergistic effect of customer visibility. The local businesses and environment line up with the type of business proposed, and thus may provide opportunities for local advertising to be effective. In addition, these areas are well populated, with 9,000 or more people residing in each zip code. These zip codes are also neighboring, adding a cumulative effect to the 3 zip codes, creating a local potential customer base that may number in the tens of thousands. These zip codes also house some of the highest income residents in Oakland, which fits the business owners potential customer profile.