BATTLE OF THE NEIGHBORHOODS

IBM Data Science: Capstone Project
Presented by Claire Li

INTRODUCTION/BUSINESS PROBLEM

The client is the owner of a fast-casual Chinese restaurant in Irvine, California, which is located in the suburb of Orange County. As the restaurant has proven to be a success in its years of business, the client is seeking to take the next step and expand to the Los Angeles metropolitan area.

As a bustling urban area with some of the nation's highest rent costs, it is a considerable investment to open a storefront in Los Angeles, but also promises the opportunity for high returns and rapid growth. The client is seeking a location for their new venture that shares similar characteristics with its existing location in Orange County. Specifically, they are looking for a similar demographic profile as well as a similar competitor landscape. Therefore, the aim of this study is to perform a preliminary analysis to research and suggest neighborhoods for the client's proposed new venture. Specifically, the business problem is: **which neighborhoods in Los Angeles have a similar demographic profile and competitor landscape to Irvine, California, and are potential locations for the client's proposed new restaurant?**

DATA

The features that we seek in a potential location include:

- A similar demographic profile to the original Irvine location. Factors to be included in the analysis are: total population, age distribution, median income, and race/ethnicity.
- Similar consumer preferences. This will be measured by the types of businesses and venues in each area.
- A balance between a proven consumer demand for Chinese food and avoiding oversaturated areas

The data comes from the following sources:

- **List of Los Angeles neighborhoods:** We used the list of L.A. County neighborhoods as defined by the Los Angeles Times, available here: http://boundaries.latimes.com/set/la-county-neighborhoods-current/
- **Los Angeles demographic data:** We were interested in identifying a neighborhood of Los Angeles with comparable characteristics to the original restaurant location in Orange County. To do this, we used data about age distribution (https://usc.data.socrata.com/Los-Angeles/Age-Distribution-LA-/rqg9-k6ju/data), median income (https://maps.latimes.com/neighborhoods/income/median/neighborhood/list/), and race/ethnicity (https://usc.data.socrata.com/Los-Angeles/Race-Ethnicity-LA-/jxw5-xxv5/data).

- **Orange County (Irvine) demographic data:** We were specifically interested in the demographic data of Irvine, California, as provided by the US Census Bureau: https://www.census.gov/quickfacts/fact/table/irvinecitycalifornia/PST045219 and https://data.census.gov/cedsci/table?q=%20irvine&g=1600000US0636770&tid=ACSDP1Y2018.DP05&hidePreview=true
- **Foursquare API:** Foursquare is a search-and-discovery platform that aims to help users discover and share information about local businesses and attractions. We utilized the Foursquare API for location-based insights on local venues in order to (1) determine the top types of venues for each neighborhood, to get a sense for consumer preferences, (2) find a neighborhood with a similar business landscape as Irvine, and (3) research the relative frequency .

Methodology

This study was conducted in two parts. First, we examined the demographic composition of various LA neighborhoods in order to find a similar demographic area to the original Irvine location. Second, we studied competitor landscape and consumer preferences via the Foursquare API.

Part 1: Demographic Data

Data Acquisition and Cleaning

To delineate our neighborhoods of interest, we used the list of Los Angeles County neighborhoods as designated by the *Los Angeles Times*. Altogether, there are 272 individual neighborhoods. To obtain geographical coordinates for each neighborhood, we took the centroid of the geographical region given by the data (see **Figure 1**) .

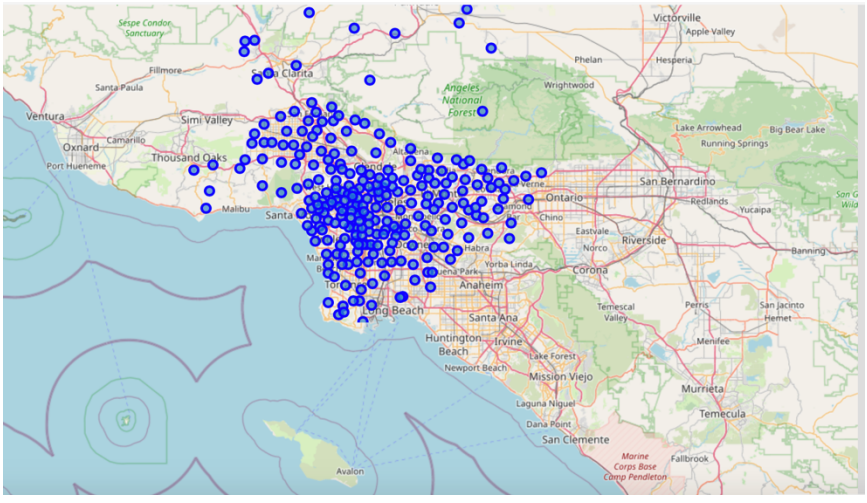

FIGURE 1: MAP OF LOS ANGELES NEIGHBORHOODS

We then examined data about various demographic indicators, including total population, age distribution, median income, and race and ethnicity. Missing values were dropped from the

dataset in order to avoid skewing the analysis. Combining this information with our list of Los Angeles neighborhoods, we created a table displaying the statistics for each neighborhood. This table provides us with a snapshot of the demographic composition of each neighborhood (see **Figure 2** for an example).

	Neighborhood	Longitude	Latitude	Median Income	Under Age 18	Ages 18-24	Ages 25-34	Ages 35-44	Ages 45-54	Ages 55-64	Ages 65 & Older	Total Population
0	Acton	-118.185799	34.495516	83983	17.535794	8.340996	9.667674	10.271903	17.588336	20.596348	15.998949	7613.0
1	Adams-Normandie	-118.300288	34.031411	29606	20.946428	18.297966	16.027855	12.304655	11.443768	11.860503	9.118824	18237.0
2	Agoura Hills	-118.760944	34.150734	117608	22.520192	6.799345	9.771420	13.350578	17.763818	15.382991	14.411656	18943.0
3	Agua Dulce	-118.313371	34.508909	106078	17.722140	8.492882	12.714777	7.756505	15.635739	20.250368	17.427590	4074.0
4	Alhambra	-118.135494	34.083967	53224	17.292348	8.549674	16.339115	13.664180	14.331443	12.968673	16.854567	84974.0

FIGURE 2: SAMPLE OF FINAL DEMOGRAPHIC DATA TABLE
(NOTE THAT RACE/ETHNICITY DATA, WHICH WAS INCLUDED IN THE TABLE, IS NOT SHOWN IN THE FIGURE)

Next, we retrieved data for the same metrics for the city of Irvine, where the client's original storefront is located. We combined this with the data on LA neighborhoods to create a table containing all the demographic information, which we used for our clustering procedure.

DATA ANALYSIS

Since the client aims to find similar neighborhoods to Irvine, we performed a clustering analysis, which aggregates data points (in our case, neighborhoods) together based on certain similarities. Specifically, we chose to perform K-means clustering on the data.

In our clustering procedure, the data was normalized using StandardScaler from the scikitlearn package, which removes the mean and scales to unit variance. We performed the analysis using k = 15 (i.e., 15 clusters). After running K-means, the neighborhoods grouped in the same cluster as Irvine were found to be Glendale, Lancaster, Long Beach, Palmdale, Pasadena, Pomona, Santa Clarita, and Torrance (see **Figure 3**). This result was robust for k = 12 – 17. These neighborhoods represent potential locations, based solely on the demographic data.

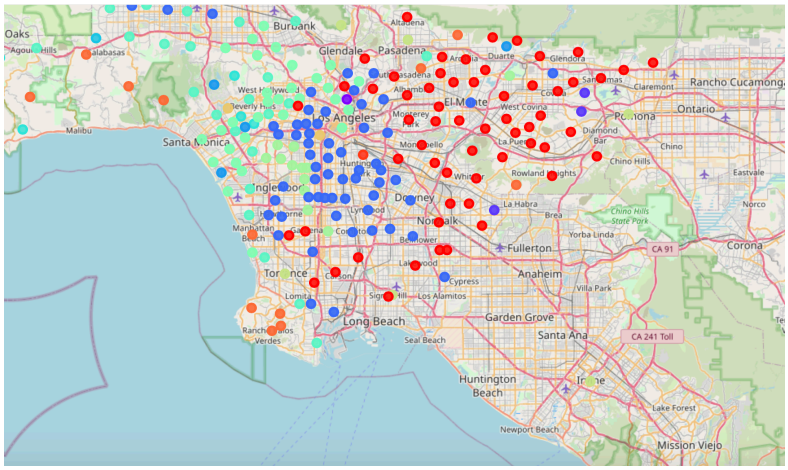

FIGURE 3: CLUSTERING WITH DEMOGRAPHIC DATA

PART 2: BUSINESS LANDSCAPE

DATA ACQUISITION AND CLEANING

Foursquare is a search-and-discovery platform where users can discover about local businesses and attractions, and share information with others. Based on their location, users can discover local venues (e.g., restaurants, attractions, recreation centers) around them, as well as view ratings and reviews from other users. For the purposes of this study, we focused only on capturing the most common types of venues for each neighborhood. We used the Foursquare API to retrieve this information by getting a list of venues and their categories (e.g., "Mexican Restaurant", "Park", "Café") that are located within 3km of each neighborhood's calculated centroid (which was obtained in Part 1). The number of venues returned for each neighborhood was limited to 100[1]. Overall, we found 22381 venues, with 431 unique categories. The top 10 most common venue categories overall are Mexican Restaurant (1213), Coffee Shop (983), Fast Food Restaurant (880), Pizza Place (711), Sandwich Place (656) , Grocery Store (648), Burger Joint (591), Park (551), American Restaurant (513), and Convenience Store (465).

DATA ANALYSIS

After obtaining the data from the Foursquare API, we moved to analyzing and preparing the data for K-means clustering. Since our aim was to determine which neighborhood had a similar business landscape to Irvine, we calculated the 10 most common categories for each neighborhood. This was accomplished by first performing one-hot encoding on the data and calculating the mean of occurrence of each venue category by neighborhood. The resulting grouped table listed the neighborhoods with the relative scaled frequency of occurrence of each venue category, across all categories; we were then able to use this table to rank the venue categories in order of frequency of occurrence for each neighborhood. For Irvine, we found that the most common venue categories are Sandwich Place, Café, Japanese Restaurant, Fast Food Restaurant, Grocery Store, Coffee Shop, Dessert Shop, Shopping Mall, Ice Cream Shop, and Mexican Restaurant.

Using the grouped table with the relative scaled frequency of occurrence of each venue category, we performed K-means clustering. For our analysis, we once again took k=15. After clustering, we found 84 other neighborhoods grouped in the same cluster as Irvine. This large number can be attributed to the sparsity of features used for the analysis. To narrow down the results, we created a table combining the demographic data and the Foursquare data, and performed K-means clustering again (k = 15). After clustering, we found that Glendale, Lancaster, Long Beach, Palmdale, Pasadena, Pomona, Santa Clarita, and Torrance were in the same cluster with Irvine—the same neighborhoods as before (see **Figure 4**).

Finally, we examined the number of Chinese restaurants in each of the potential neighborhoods. We judged the relative frequency of Chinese restaurants by calculating the

[1] We acknowledge that retrieving 100 venues within the given radius of the geographical centroid of a neighborhood may not necessarily be representative of the business landscape of the overall neighborhood. Additionally, it is possible that some of the search areas for neighborhoods may have overlapped. However, these factors should not prevent us from getting an general picture of consumer preferences for each neighborhood. See Discussion section for more information.

proportion of Chinese restaurants over total venues returned[2] for each neighborhood. The neighborhood with the lowest relative frequency was Long Beach, while the neighborhood with the highest was Palmdale.

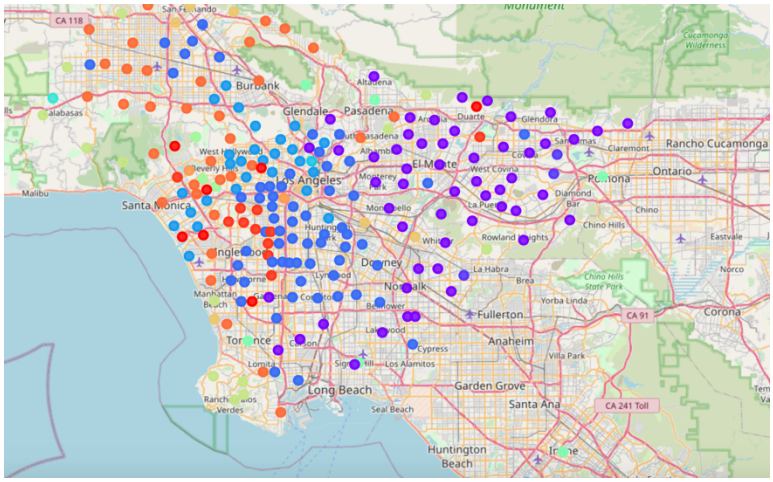

FIGURE 4: CLUSTERING WITH FOURSQUARE DATA AND DEMOGRAPHIC DATA

RESULTS

Based on the demographic data, which included total population, age distribution, median income, and race/ethnicity, it was determined that the LA neighborhoods most similar to Irvine are Glendale, Lancaster, Long Beach, Palmdale, Pasadena, Pomona, Santa Clarita, and Torrance.

The data from Foursquare gave us insight on the most common types of venues in each neighborhood. Among all neighborhoods, the top five most common types of venues were Mexican restaurants, coffee shops, fast food restaurants, pizza places, and sandwich places. In Irvine, the most common types of venues were Sandwich Place, Café, Japanese Restaurant, Fast Food Restaurant, Grocery Store, Coffee Shop, Dessert Shop, Shopping Mall, Ice Cream Shop, and Mexican Restaurant. Using K-means clustering produced 84 other neighborhoods judged to be similar to Irvine based on common venues. To narrow this down, we combined the Foursquare data with the demographic data, and performed K-means clustering again. This time, we again found that the LA neighborhoods in the same cluster as Irvine were the same as the ones from clustering just the demographic data: Glendale, Lancaster, Long Beach, Palmdale, Pasadena, Pomona, Santa Clarita, and Torrance. Based on these results, we selected these eight neighborhoods as the locations that would be most similar to Irvine. We next examined the frequency of Chinese restaurants in these neighborhoods and found Pomona to have the most and Long Beach the least among the seven potential neighborhoods.

DISCUSSION

In this section, we address a few issues regarding the data analysis. First, regarding the Foursquare data, we acknowledge that retrieving 100 venues within the given radius of the geographical centroid of a neighborhood may not necessarily be representative of the business

[2] Though the limit was capped at 100 venues per neighborhood, each neighborhood did not necessarily return 100 results.

landscape of the overall neighborhood. However, since the 3km radius is relatively large, this is sufficient to capture a general picture and provide the overall sense of consumer preferences that we seek. Additionally, it is possible that some of the search areas for neighborhoods may have overlapped. However, searching for venues within our 3km radius of the centroid should still give a general sense of consumer preferences for the area. Thus, our use of the Foursquare data for analyzing the business landscape maintains its validity.

The second issue we wish to address relates to the analysis of relative frequency of each venue category. In general, this information provides a sense of consumer preferences for each neighborhood: if a certain category of venue is more frequent, it can be concluded that there is greater consumer demand. However, using this information represents a trade-off with the potential gains from opening a type of venue in an area where there are fewer competitors of the same category (e.g., opening a Chinese restaurant in an area with many Chinese restaurants, demonstrating a proven consumer demand vs. opening a Chinese restaurant in an area without any options for Chinese food). Our recommendation takes into consideration both of these opposing forces to choose a neighborhood that is not oversaturated yet has proven consumer demand: we eliminate the neighborhoods with the highest and the lowest frequencies.

Therefore, with all these factors taken into consideration, the neighborhoods we recommend to the client based upon this preliminary analysis are: **Pasadena**, **Torrance, Glendale, Lancaster, Santa Clarita,** and **Pomona.**

CONCLUSION

In this study, we analyzed the 272 Los Angeles neighborhoods in order to find a potential candidate for the client's second restaurant location. Based on our clustering analysis, we found that the neighborhoods judged to be most similar to the client's original storefront in Irvine are: Glendale, Lancaster, Long Beach, Palmdale, Pasadena, Pomona, Santa Clarita, and Torrance. After analyzing the relative frequency of existing Chinese restaurants in these eight neighborhoods, we selected Pasadena, Santa Clarita, and Torrance as our recommendation.

For future more in-depth analyses, more demographic and geographic metrics could be used (e.g., sex, median household size, proximity to population hubs). Additionally, a ranking system could be created based on the features that the client most prioritizes (e.g., if the client wishes to shift their focus to college-age students, or if the client decides to cater towards office lunch crowds). As this was a preliminary analysis, this study may be further developed to incorporate future directions.