FINDING A LOCATION AND TYPE FOR A RESTAURANT IN LOS ANGELES COUNTY, CA

By Felix Fichtner

Table of Contents

1 Introduction

In this capstone project I will try to give a recommendation on where to open a restaurant in Los Angeles County, CA. The location will be a community or area in the county, that is defined by its postal code. In addition, there shall be given a recommendation of which type of food venue could be opened, based on existing restaurants in the area and generally popular restaurants in the whole county.

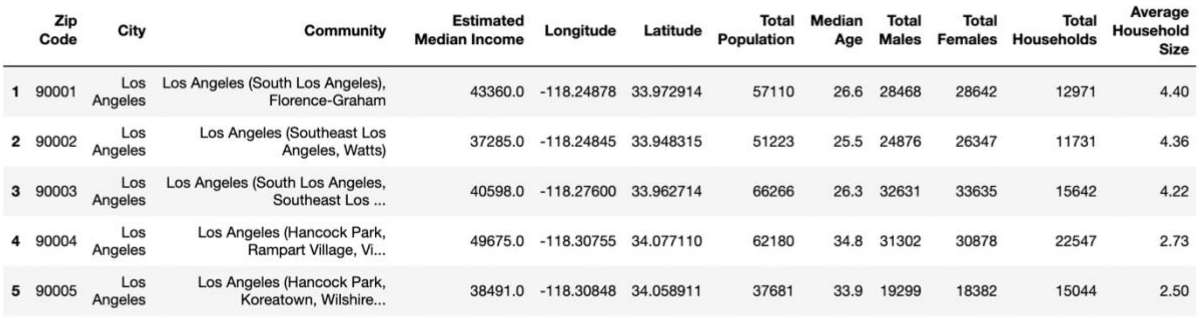
The decision on where to open a restaurant can be based on many factors, depending on the target group. For example, one could look for very dense populated areas, or areas with lots of wealthy citizens. Even the median age of the population can play a role. Because of this, I will include census and economic data into the analysis.

I will make a final recommendation based on the following conditions, that:

- Find the area with a good balance between number of possible customers and a high median income (population is slightly more important than income).
- The type of restaurant will be determined by the most recommended categories of food venue in Los Angeles County, CA and the number of already existing venues in the area, grouped by their categories.

2 Data

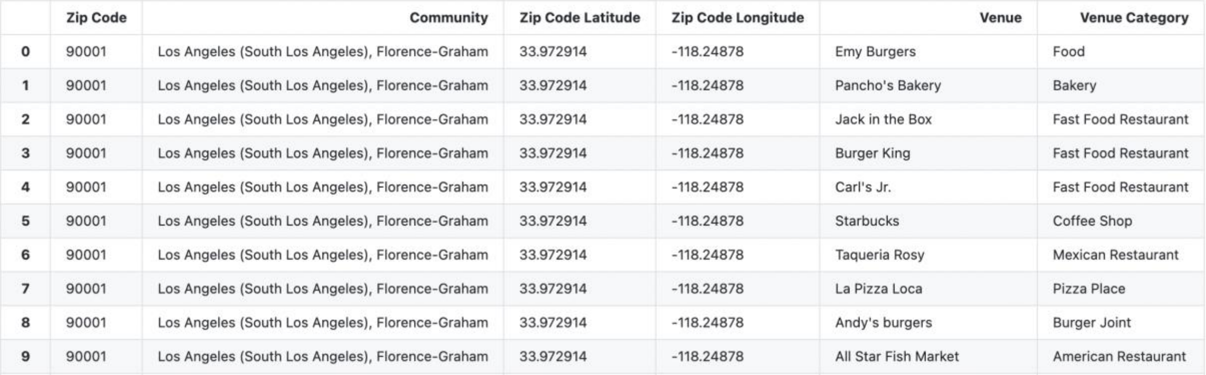
The basis for the analysis will be the 2010 census data of all zip codes in Los Angeles County, CA. This census data provides information on the composition of the population, age distribution, and the total numbers. This dataset will be enhanced with the 2019 median household income per zip code. The income can also serve as factor in the recommendation on where to open a restaurant, how it should be set up, and what the target audience could be. As a third part geospatial data is added for every zip code, to be able to use Foursquare to explore existing food offerings in the communities of Los Angeles County, CA.



	Zip Code	City	Community	Estimated Median Income	Longitude	Latitude	Total Population	Median Age	Total Males	Total Females	Total Households	Average Household Size
1	90001	Los Angeles	Los Angeles (South Los Angeles), Florence-Graham	43360.0	-118.24878	33.972914	57110	26.6	28468	28642	12971	4.40
2	90002	Los Angeles	Los Angeles (Southeast Los Angeles, Watts)	37285.0	-118.24845	33.948315	51223	25.5	24876	26347	11731	4.36
3	90003	Los Angeles	Los Angeles (South Los Angeles, Southeast Los ...	40598.0	-118.27600	33.962714	66266	26.3	32631	33635	15642	4.22
4	90004	Los Angeles	Los Angeles (Hancock Park, Rampart Village, Vi...	49675.0	-118.30755	34.077110	62180	34.8	31302	30878	22547	2.73
5	90005	Los Angeles	Los Angeles (Hancock Park, Koreatown, Wilshire...	38491.0	-118.30848	34.058911	37681	33.9	19299	18382	15044	2.50

Figure 1 - subset of the prepared dataset

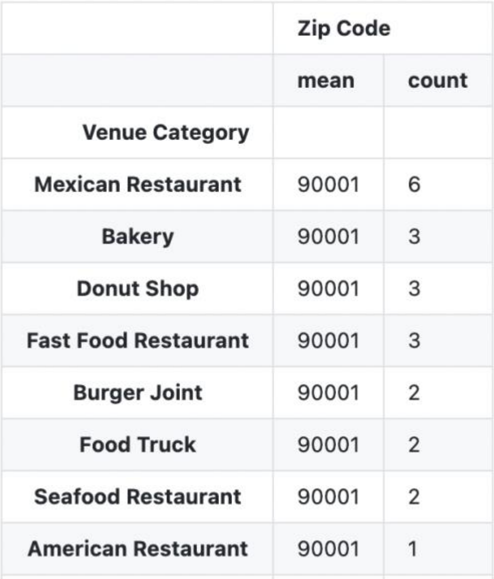
Using Foursquare, already existing food venues can be explored, and grouped. This way the competitors can be explored, for example by the food type or rating. It is also possible to identify market niches and give a recommendation for a type of restaurant that could be opened.



	Zip Code	Community	Zip Code Latitude	Zip Code Longitude	Venue	Venue Category
0	90001	Los Angeles (South Los Angeles), Florence-Graham	33.972914	-118.24878	Emy Burgers	Food
1	90001	Los Angeles (South Los Angeles), Florence-Graham	33.972914	-118.24878	Pancho's Bakery	Bakery
2	90001	Los Angeles (South Los Angeles), Florence-Graham	33.972914	-118.24878	Jack in the Box	Fast Food Restaurant
3	90001	Los Angeles (South Los Angeles), Florence-Graham	33.972914	-118.24878	Burger King	Fast Food Restaurant
4	90001	Los Angeles (South Los Angeles), Florence-Graham	33.972914	-118.24878	Carl's Jr.	Fast Food Restaurant
5	90001	Los Angeles (South Los Angeles), Florence-Graham	33.972914	-118.24878	Starbucks	Coffee Shop
6	90001	Los Angeles (South Los Angeles), Florence-Graham	33.972914	-118.24878	Taqueria Rosy	Mexican Restaurant
7	90001	Los Angeles (South Los Angeles), Florence-Graham	33.972914	-118.24878	La Pizza Loca	Pizza Place
8	90001	Los Angeles (South Los Angeles), Florence-Graham	33.972914	-118.24878	Andy's burgers	Burger Joint
9	90001	Los Angeles (South Los Angeles), Florence-Graham	33.972914	-118.24878	All Star Fish Market	American Restaurant

Figure 2 - Foursquare Data for Zip Code 90001

The following example illustrates how the data from Foursquare helps to analyze the food venues in a community. Using this API, one can determine that Mexican restaurants are the dominant food venue category in Florence-Graham, South Los Angeles.



	Zip Code	
	mean	count
Venue Category		
Mexican Restaurant	90001	6
Bakery	90001	3
Donut Shop	90001	3
Fast Food Restaurant	90001	3
Burger Joint	90001	2
Food Truck	90001	2
Seafood Restaurant	90001	2
American Restaurant	90001	1

Figure 3 - Food Venues grouped by Category in Florence-Graham

In detail, we will try to answer the following questions:

- What are the most recommended types of restaurants in the county?
- What is the distribution of the existing food venues categories in the area where we want to open a restaurant in?

The datasets are publicly available at:

- 2010 Los Angeles Census Data
  - https://www.kaggle.com/cityofLA/los-angeles-census-data
- Median Household Income by Zip Code in 2019
  - http://www.laalmanac.com/employment/em12c.php
- US Zip Code Latitude and Longitude
  - https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/information/

3 Methodology

In this project we will focus on finding a suitable area for a new restaurant in Los Angeles County, CA. The areas are defined by their US postal code. In addition, we will look at the most recommended food venue categories throughout the country, to suggest which type of restaurant could be opened. There won’t be a specific location in the chosen area recommended.

In the first step we have merged three different datasets, that provide data on the different areas in Los Angeles County. With this data it is possible to group or cluster the areas using information like median income, number of households and number of inhabitants.

The second step in the analysis is to cluster (using k-means clustering) the areas in the county and to describe the individual clusters. Using this method we support the process of finding a single area that looks promising for a new restaurant.

The third step is to pick a cluster that fits the chosen criteria most. The area shall be chosen under the premise of finding a good balance between estimated median income and number of potential customers, where population is rated slightly more important than income. So, the target is to find an area that has as many citizens as possible with the highest income possible. After an area was chosen, the distribution of local restaurant types in this area will be analysed.
Combining this information with the categories of food venues that are popular throughout the whole county, a recommendation of the restaurant to open can be given.

3.1 Initial Analysis

First, let us see what we can find out by analyzing the dataset. As population per Area and income are most important. I have sorted the dataset by these features, to make out some potential candidates for opening a food venue, according to our criteria.

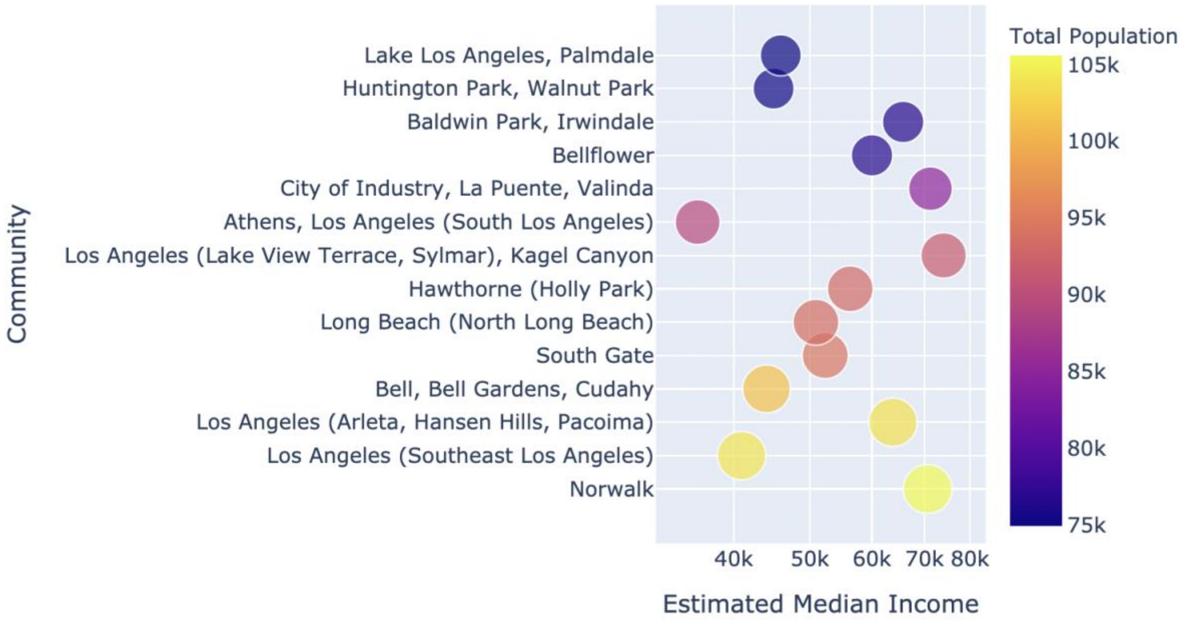


Figure 4- Income and Population per Area

In Figure 4 we find a combined representation of the top 15 areas by their population (bubble size and color), combined with their estimated median income (x axis). In this graph there are four areas that stand out:

City	Population	Median Income
Norwalk	105,600	70,700
Lake View Terrace, Sylmar	91,700	74,000
La Puente, Valinda	85,000	71,200
Hansen Hills, Pacoima	104,000	64,000

Table 1- Areas with a good combination of income and population

These four cities / neighbourhoods seem to be suitable areas, based on their combination of population and median income. We are going to see if this assumption is confirmed going forward.

In preparation for further analysis, the most recommended food venue categories in Los Angeles County were identified, using the Foursquare API. This was done by going through all available postal codes and querying up to 50 of the most recommended food venues in every area (sorted by their popularity) in a 1 km radius. The food venues were then grouped by their category, to see which are most often recommended.

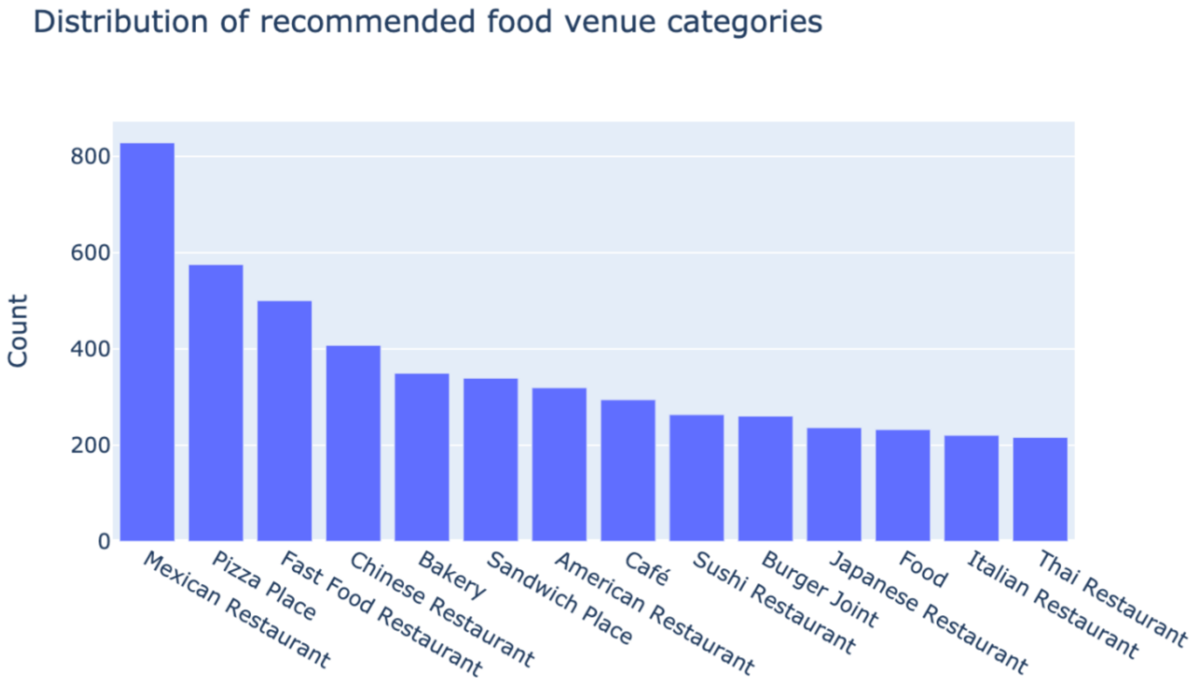
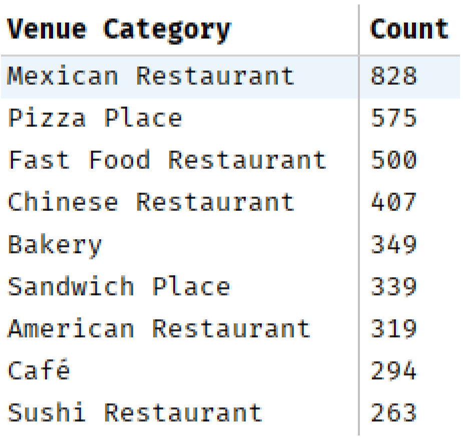


Figure 5- Distribution of recommended Food Venue Categories

Mexican restaurants are by far the most often recommended type of food venue in Los Angeles County, CA. They are followed in some distance by pizza places, fast food and Chinese restaurants and bakeries.

The exact numbers of the top 10 categories are:



Venue Category	Count
Mexican Restaurant	828
Pizza Place	575
Fast Food Restaurant	500
Chinese Restaurant	407
Bakery	349
Sandwich Place	339
American Restaurant	319
Café	294
Sushi Restaurant	263

Figure 6- Top 10 Food Venue Categories

3.2 Clustering the Dataset

In the second step I chose to cluster the areas, based on the available census and income data. This was done to group similar areas and to identify the cluster that fits the criteria for this project most. The areas from this cluster are then all possible fits for opening a new restaurant and can be compared against each other. This is a useful method to make it easier to select areas, because there are 227 areas in our dataset. I chose k-Means as clustering method.

For identifying the optimal k value, I used the elbow method. So, the clustering algorithm was run multiple times with different k values (1 to 14) and extracted the sum of squared distances of samples to their closes cluster center. Based on the results I chose a k value of 6 going forward.

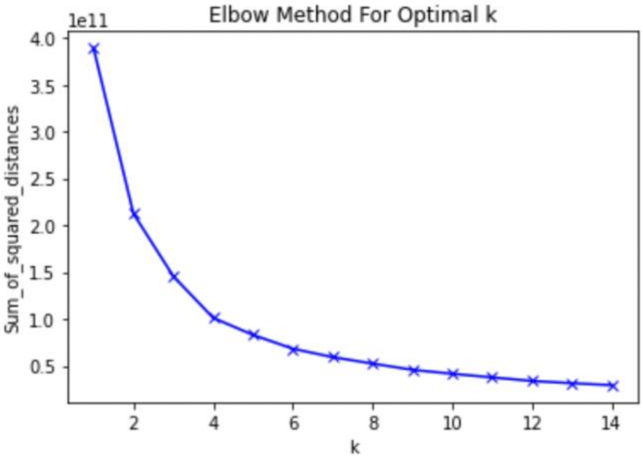


Figure 7- Results of the Elbow Method to choose a k value

I provided five numeric features of the dataset for the clustering algorithm,

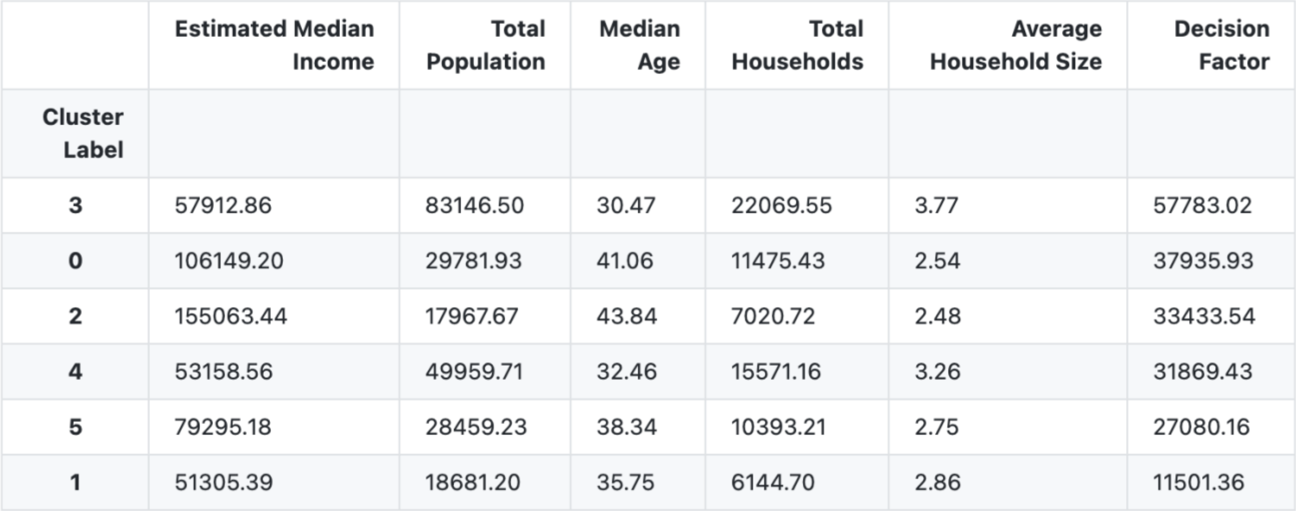
- Estimated Median Income,
- Total Population,
- Median Age,
- Total Households and
- Average Household Size.

After the clustering was done, I calculated a value called “decision factor” for every cluster. This was done to make it easier to compare the clusters and to reproduce the criteria of our stakeholders. This factor is calculated the following way:

$$Decision\ Factor\ = \frac{(Total\ Population * 1.2) * Estimated\ Median\ Income}{100000}$$

By adding the result of this calculation to every cluster, it is easy to compare them and find the one that fits best to our criteria.

Now let us have a look at the six clusters that were formed by the algorithm:

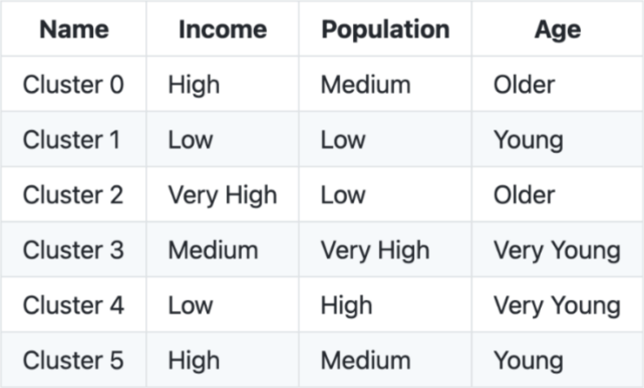


	Estimated Median Income	Total Population	Median Age	Total Households	Average Household Size	Decision Factor
Cluster Label						
3	57912.86	83146.50	30.47	22069.55	3.77	57783.02
0	106149.20	29781.93	41.06	11475.43	2.54	37935.93
2	155063.44	17967.67	43.84	7020.72	2.48	33433.54
4	53158.56	49959.71	32.46	15571.16	3.26	31869.43
5	79295.18	28459.23	38.34	10393.21	2.75	27080.16
1	51305.39	18681.20	35.75	6144.70	2.86	11501.36

Figure 8 - Clusters that were formed from the Dataset

3.3 Selecting a Cluster and three of the best fitting areas

By comparing the mean values of the numeric data that was used for clustering, I found that income, population and age were the main influences for the cluster that is assigned. I have applied the following labels to every cluster:



Name	Income	Population	Age
Cluster 0	High	Medium	Older
Cluster 1	Low	Low	Young
Cluster 2	Very High	Low	Older
Cluster 3	Medium	Very High	Very Young
Cluster 4	Low	High	Very Young
Cluster 5	High	Medium	Young

Figure 9- Labels for the Clusters

So based on this information I am going to choose Cluster 3 for further evaluation and as the cluster where I will pick an area from. This cluster has a medium income combined with a very high population. It also contains the youngest median age, which also can be considered for choosing the food venue category. So, it should contain the areas that fit best for further analysis.

Cluster 3 contains 22 areas in Los Angeles County, CA. First, let us see where exactly they are in the County.

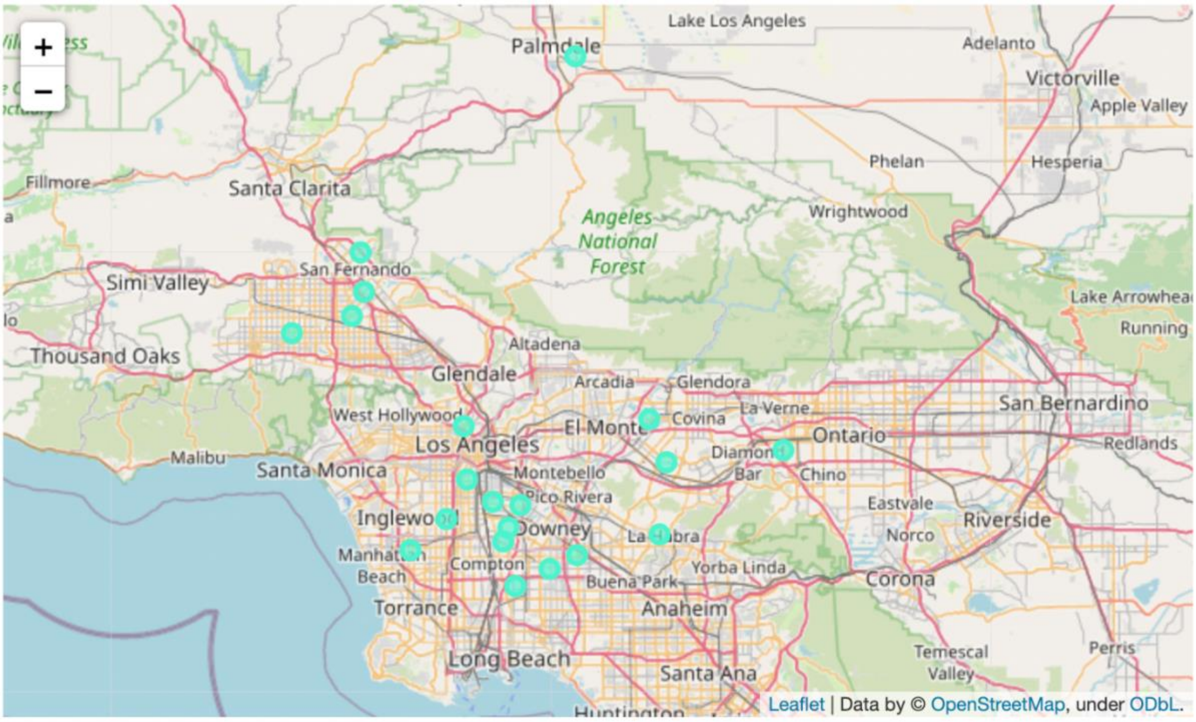


Figure 10- Geographical Locations of the Areas of Cluster 3

Most of the locations are in very densely populated areas, most of them in the vicinity of the city of Los Angeles. I applied the decision factor again to every of the 22 areas, to identify the best fitting areas and to further investigate them. The following graph shows the areas that have the best distribution of income and population, according to the decision factor used as metric. What the data also showed, that as the average household size increases, the income also decreases.

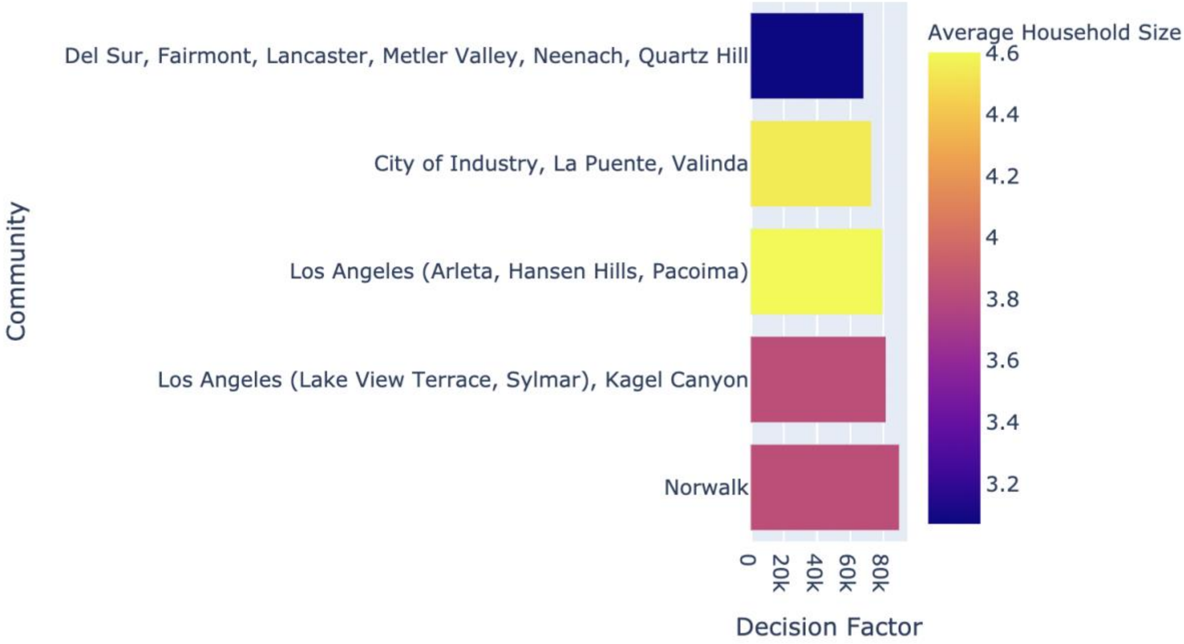


Figure 11- Five best fitting Areas

According to these metrics there are three areas that stand out:

- Norwalk, with a population of 105,600 and an income of $70,600; it has the oldest median age of the three communities and the largest population.
- Lake View Terrace in Sylmar, with a population of 91,700 and an income of $74000; it has the highest income of the group and the lowest population.
- Hansen Hills in Pacoima, with a population 104,700 and an income of $63,800. It has the youngest median age of the three and is very close to Norwalk in terms of population, but it has the lowest income.

Looking back at the initial analysis we did for all areas in Los Angeles County, we find that these three areas were also part auf the group we identified, based on their features. Through the clustering we could confirm our initial findings.

3.4 Exploring the local Food Venues and identifying Market Gaps

Using the Foursquare I explored the local food venues in every of the three food venues. I analyzed all venues of a location in a 4 km radius that belong to the postal code of the area. Overall, I found 167 food venues across all 3 areas. The next step was to group the food venues by their categories and cities. Let us look at the visualization of the results:

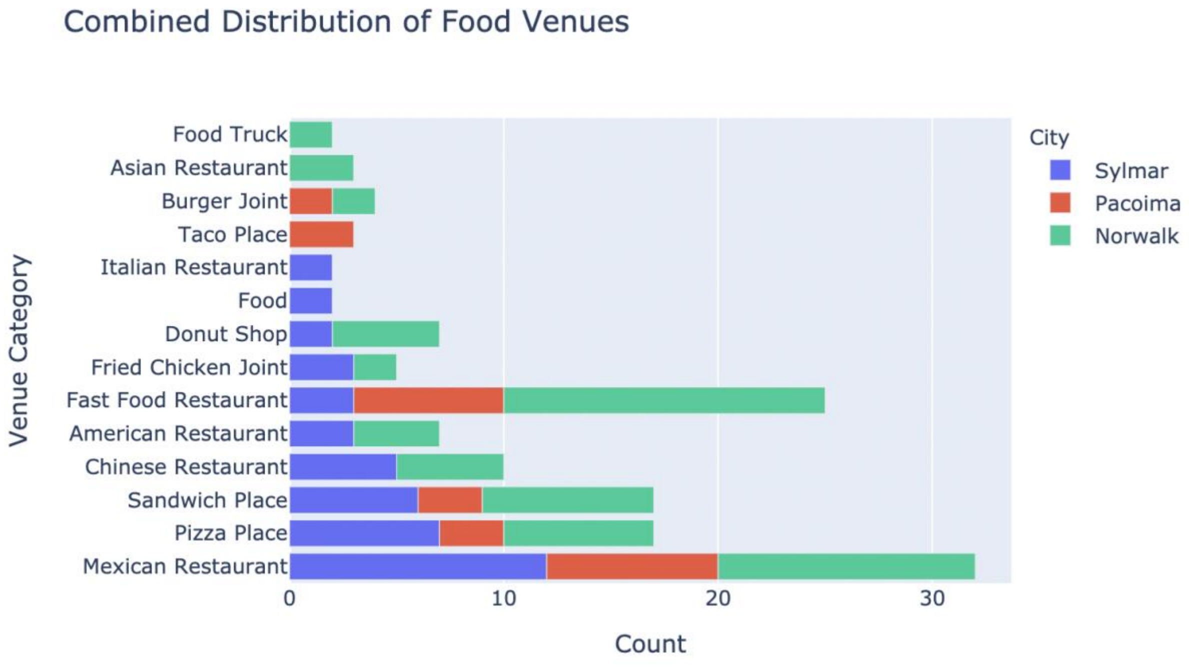


Figure 12 - Distribution of Food Venues in the 3 chosen Areas

Having a look at the distribution of food venues across the three areas, we can make the following observations:

- Mexican restaurants are the most common food venue overall.
- Fast Food restaurants are the second most common venue, but Norwalk has by far the most.
- Pacoima has the least restaurants overall and Norwalk the most, despite both are very close in population.
- There seems to be space for a Chinese or American Restaurant or a Donut Shop in Pacoima.
- Sandwich and pizza places are not common in Pacoima.
- There is a maximum of one bakery per area, although it is an often-recommended venue type in the county.
- Opening another fast-food restaurant or Mexican restaurant does not seem like a good idea.
- There are also multiple options for sushi restaurants, burger joints, Japanese or Thai restaurants in all three areas.

4 Results and Discussion

Our analysis shows that there is a high variability in population, income and median age in the different areas of Los Angeles County, CA. So it was possible to identify multiple areas that fit the criteria of a relatively high median income and a high population. Using census and income data of all areas in Los Angeles County, we did a clustering to identify similar areas. By analysing the formed clusters there have been three areas identified that fit the criteria best. Norwalk, Lake View Terrace in Sylmar and Hansen Hills in Pacoima.

They are slightly different in income, population and age and also very different in their local distribution of available food venues. The final area could be picked on which of these factors matters to the stakeholders most. I will pick Norwalk as the final area for a food venue, because it offers the best combination of a high population and a good income out of these three areas.

By analysing the most recommended food venue categories across the whole county, we found that mexican restaurants are by far the most often recommended venue. After them pizza places, fast food restaurants, chinese restaurants and bakeries follow in that order. To give a recommendation for a food venue to open in Norwalk, we can compare the local distribution of food venues what was recommended the most in the county. By looking at Norwalk we found that there are already many mexican restaurants (12) and fast food restaurants (15). Pizza places (7) and chinese restaurants (5) area also recommended in a higher number, so opening a restaurant in one of those categories would be better, but there is still some competition. What stands out is that there is currently only one bakery recommended by Foursquare in Norwalk. Looking at the distribution of recommendations in the county, bakeries are the fifth most recommended venue category. Because of this, I would recommend opening a bakery in Norwalk, CA to the stakeholders.

Purpose of this analysis was to identify a possible area and food venue type for a new restaurant in Los Angeles County, CA based on a very limited amount of factors. Analysing census data and existing food venues is only one part on the way to find a location for opening a new restaurant. Other factors that also play a role are for example available spaces, rent costs, other venues in the area. This analysis serves as a starting point for finding possible locations, but further analysis needs to be done by the stakeholders.

5 Conclusion

The purpose of this project was to find a possible location for a new restaurant in Los Angeles County, CA. The desire from the stakeholders was to identify locations that offer a good balance between median income and number of inhabitants, although income shall be rated slightly more important than population. In addition, the idea was to identify possible food venue categories by comparing recommended venues across the country with the local venues in the different areas. So for this there were census and income data combined to identify areas that fit the criteria. A clustering was performed, to group the communities in Los Angeles County using their income and population. Then the cluster was chosen that fit the former mentioned criteria the most. From this cluster the top 3 areas were chosen, that had the best balance between income and population. This way the best three candidates for a new restaurant location were identified.

The next step was to analyze the local food venue categories. For this the local venues in a 4 km radius were identified and grouped. This grouping was then compared with the distribution of the most recommended food venue categories across the whole county. In doing so, opportunities for new restaurants in any of the three chosen communities have been identified.

The final decision can be made by the stakeholders, based on the recommendations given in this project. This decision for a locality can be based on income, population or median age of the areas. The decision for a venue category can be based on popular venues across the county, and the gaps in the local food offerings that have been identified.