Battle of the LA Cities

IBM Coursera Capstone Project - 2020

Eric Jung

I. Introduction

1. Background

Ice cream is America’ s favorite dessert. According to IDFA, the average American consumes more than 23 pounds of ice cream per year [1]. With consumption levels this high, an ice cream business is bound to bring in massive profit.

2. Problem Statement

Due to its high population and scorching summer heat, Los Angeles (LA) County is an ideal place for entrepreneurs to start an ice cream shop. However, LA County stretches far, and a subpar business location will result in low revenue at best. This project will demonstrate the best cities within LA county for starting an ice cream business.

3. Interest

The project is dedicated to stakeholders who are interested in starting an ice cream business, but are unable to decide on the location. After reviewing the results, stakeholders can confidently take one step forward into starting their business.

II. Data

1. Data Acquisition & Description

The dataset was obtained through web scraping and various APIs.

- **City** - A table with the city name, incorporated date, and population as of 2010 can be found here [2]. Only the city name was scraped for the dataset. The incorporated date would not have been a factor in choosing the best city. Although the population would have an impact on the decision, the numbers from 2010 are outdated and can provide inaccurate results.
- **Population** - A densely populated area means larger customer base. Previously mentioned, the data from 2010 is not an accurate representation of a city's population. Therefore, the US Census API was used in place of the 2010 data. The population is derived from a 5-year average, resulting in more reliable estimates. For this project, years ranging from 2014 to 2018 were used to approximate the cities' population. More information on the US Census API can be found here [3].
- **Income** - High-income residents have more to spend on luxurious items. Similar to the population data, the US Census API was utilized to obtain the median household income for each city.
- **Latitude & Longitude** - Google's Geocoding API was used to retrieve the cities' latitude and longitude. The coordinates will be utilized to visualize city location

as well as find nearby competitors. More information on Google's Geocoding API can be found here [4].

- **Area** - A table with the city name, land area, and other features can be found here [5]. Only city name and land area data were scraped. The table has land area in both square mile and square kilometer. In order to avoid unnecessary future conversions, only the land area in square kilometer was retrieved. The remaining features either had no impact on the business problem or contained outdated information. By itself, city area has little meaning to choosing the optimal location. However, the radius can be calculated from the land area.
- **Radius** - The radius was used to approximate a circle that envelops the city. Due to a city's irregular shape, this method is the best approach to find competitors within city limits.
- **Mean** - The Foursquare API was utilized to search for nearby dessert shops. For each city, the number of ice cream shops are divided by the total number of venues returned. This is to take into account the popularity of dessert shops within the vicinity. For example, a city with only two ice cream shops may seem ideal at first. However, absence of other dessert shops indicates residents' distaste towards desserts. Getting the percentage of the number of ice cream shops in relation to other dessert venues addresses these deceitful cases. More information on Foursquare API can be found here [6].

	city	population	income	latitude	longitude	area	radius	mean
0	Agoura Hills	20636	121896	34.153340	-118.761676	20.2	2535.72	0.333333
1	Alhambra	84974	57117	34.095287	-118.127015	19.8	2510.49	0.295455
2	Arcadia	58207	92102	34.139729	-118.035345	28.3	3001.36	0.428571
3	Artesia	16817	63708	33.865848	-118.083121	4.2	1156.24	0.382353
4	Avalon	3763	69440	33.342819	-118.328228	7.6	1555.36	0.750000

Table 1: Dataset First Five Rows

III. Methodology

1. Feature Analysis

To gain insight, the population, income, and mean data were visualized.

A. Population

Los Angeles is the most populated city in LA county. Within Los Angeles City, there are approximately four million residents. The city with the second highest population is Long Beach with just under half a million residents. This means all other cities in LA county have less than half a million residents. Certainly, Los Angeles city has the highest population by a wide margin.

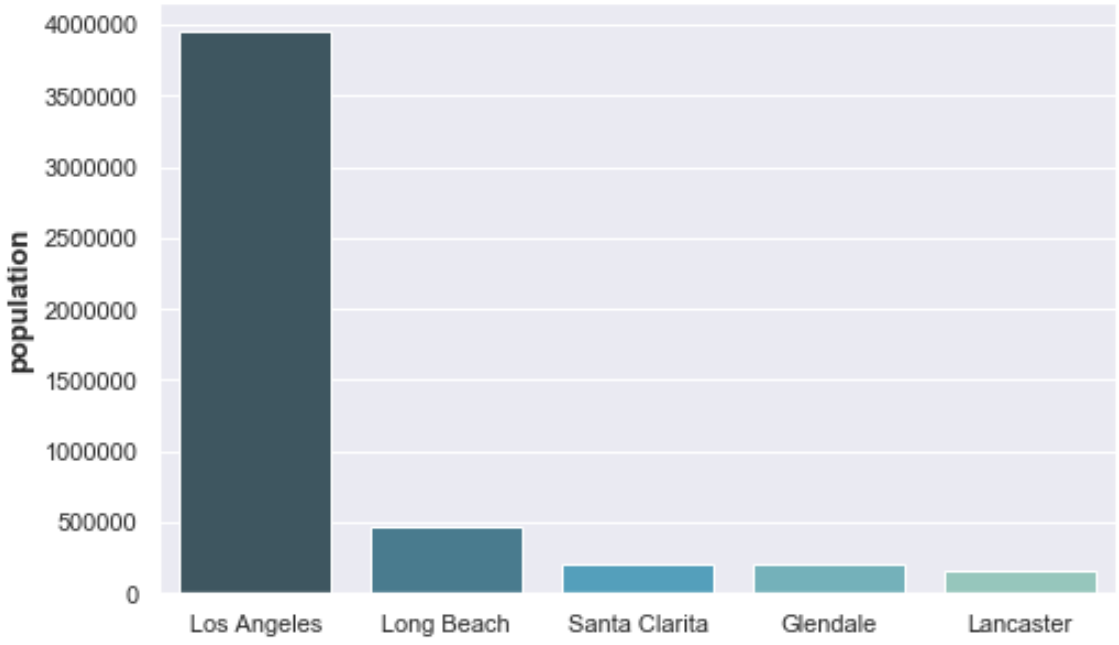

Figure 1: Top-five most populated cities

Vernon City has the smallest population in LA county. There are only around 100 people living in the city of Vernon. It is interesting that a city can contain such a small number of residents.

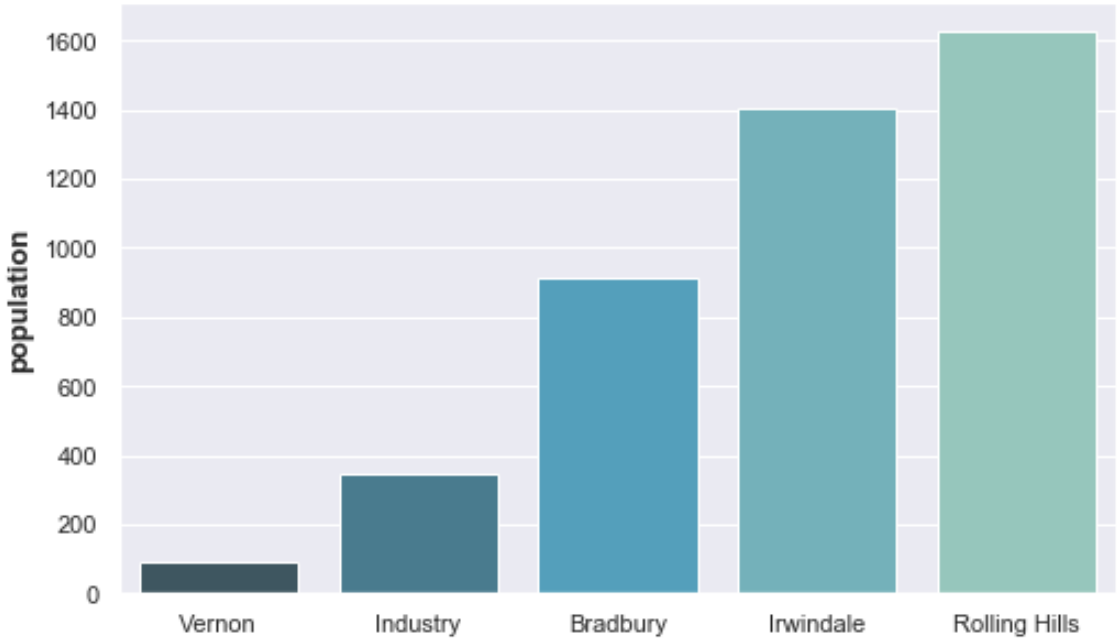

Figure 2: Top-five least populated cities

B. Income

A similar approach was taken to analyze the income data. Rolling Hills has the highest median household income, followed the Hidden Hills, Palos Verdes Estate, La Cañada Flintridge, and San Marino. These cities all have a median household income of over $150,000. A shop selling ice cream products topped with gold flakes and powder may thrive in these areas.

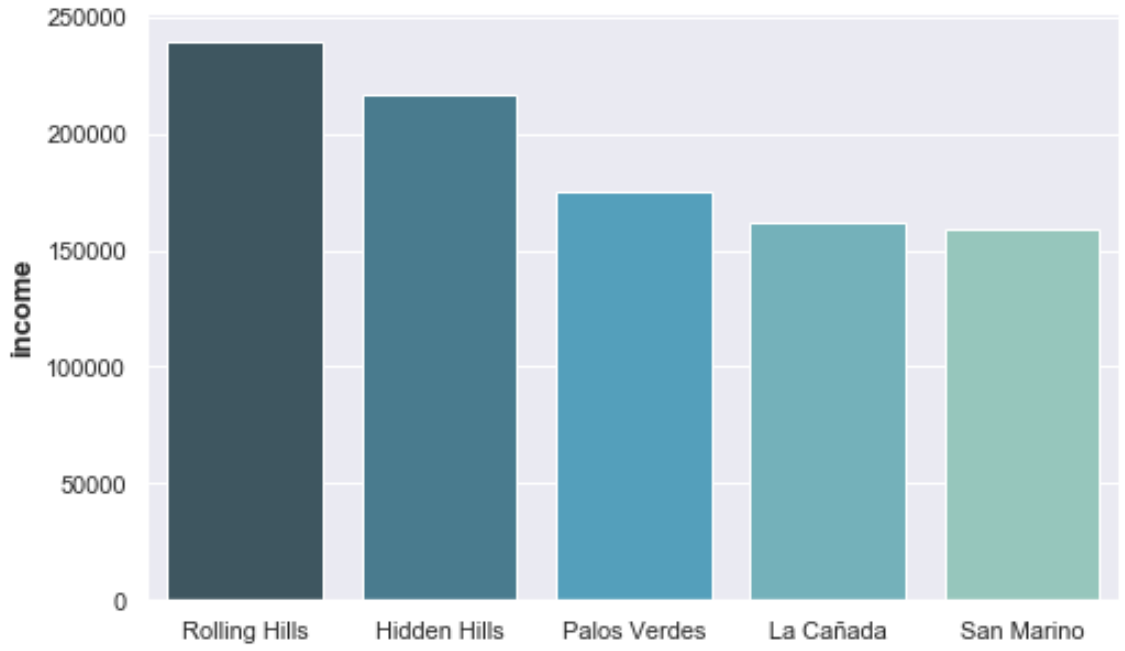

Figure 3: Top-five cities with the highest median household income

Compared to residents in other cities, residents in Maywood, Huntington Park, Bell Gardens, Bell, and Cudahy have the lowest median household income compared to other cities. Their median household income comes out to be around $40,000, well below the median household income in LA county. Starting an ice cream business in these cities may not be ideal.

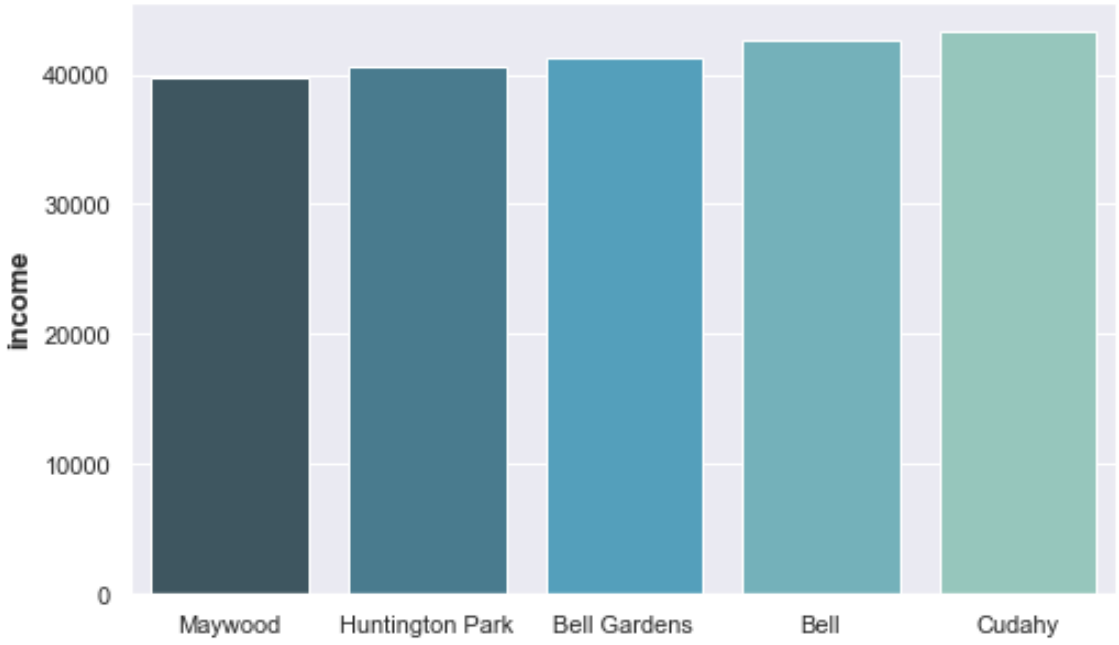

Figure 4: Top-five cities with the lowest median household income

C. Multivariate Analysis

For further research, the relationship between population and income was plotted. Two mean ranges were established to separate the data and create scatterplots.

Cities that have a low number of ice cream shops relative to the number of dessert venues may seem like a great location; however, low city population or resident income results in small customer base. For this reason, all three features are analyzed in conjunction. The marker the arrow points to in *Figure 5* has great characteristics for starting an ice cream business.

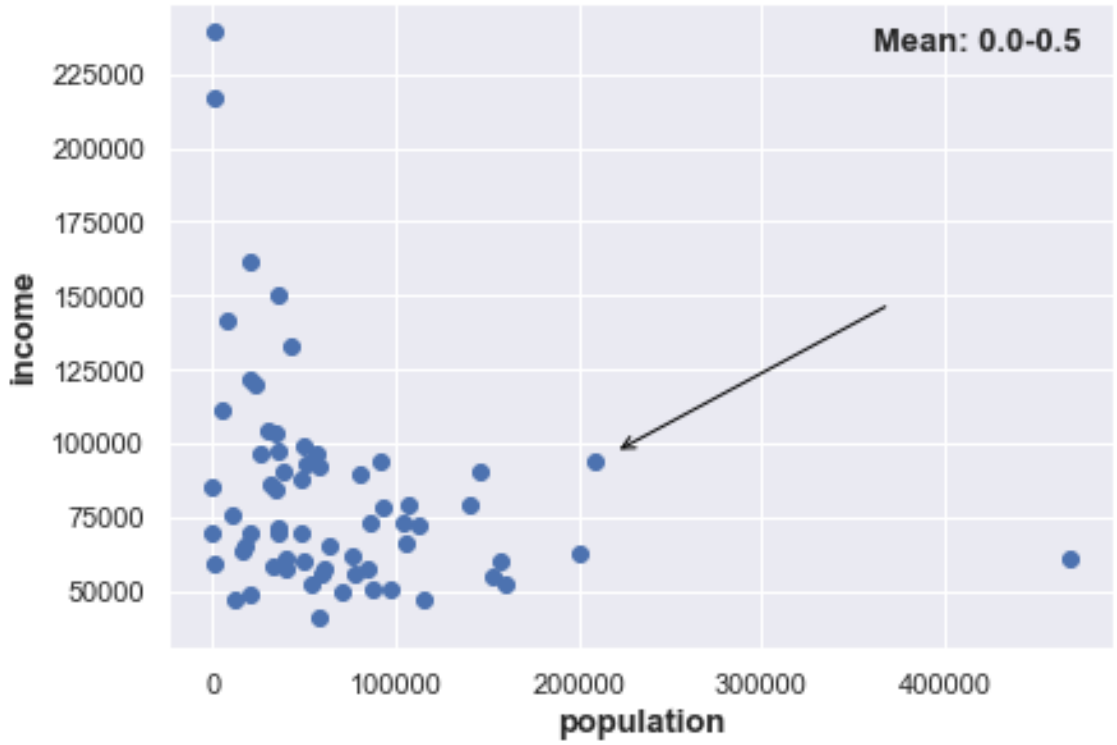

Figure 5: Relationship between population and income for cities with a small percentage of ice cream shops

For cities that have a high percentage of ice cream shops, the city population is generally low. There are some cities with a high percentage of ice cream parlors, low population, and high-income residents. If the ice cream quality can outcompete other venues', then locations with these characteristics may be applicable.

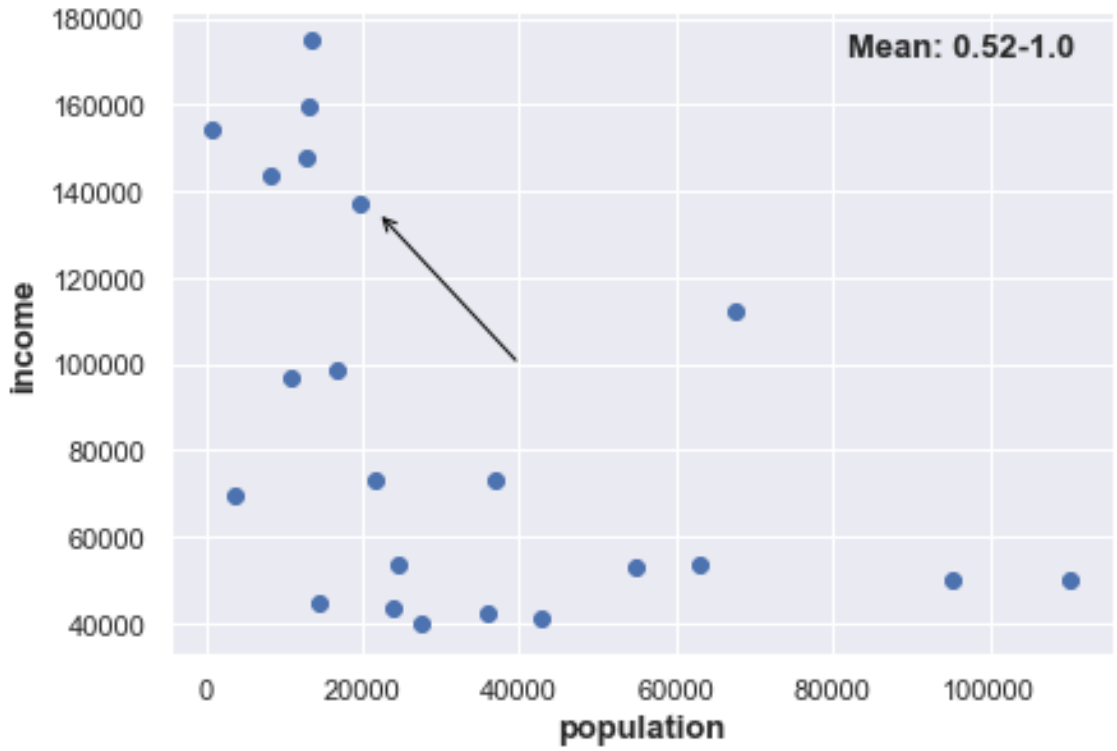

Figure 6: Relationship between population and income for cities with a high percentage of ice cream shops

Finally, a separate scatterplot was created for Los Angeles City. Los Angeles City's high population affected the visualization shown in *Figure 5* and prevented the ability to gather insight for other cities' population and income relationship. Los Angeles City has a medium percentage of ice cream shops, an approximate median household income of $58,000 and a population of nearly 4 million people.

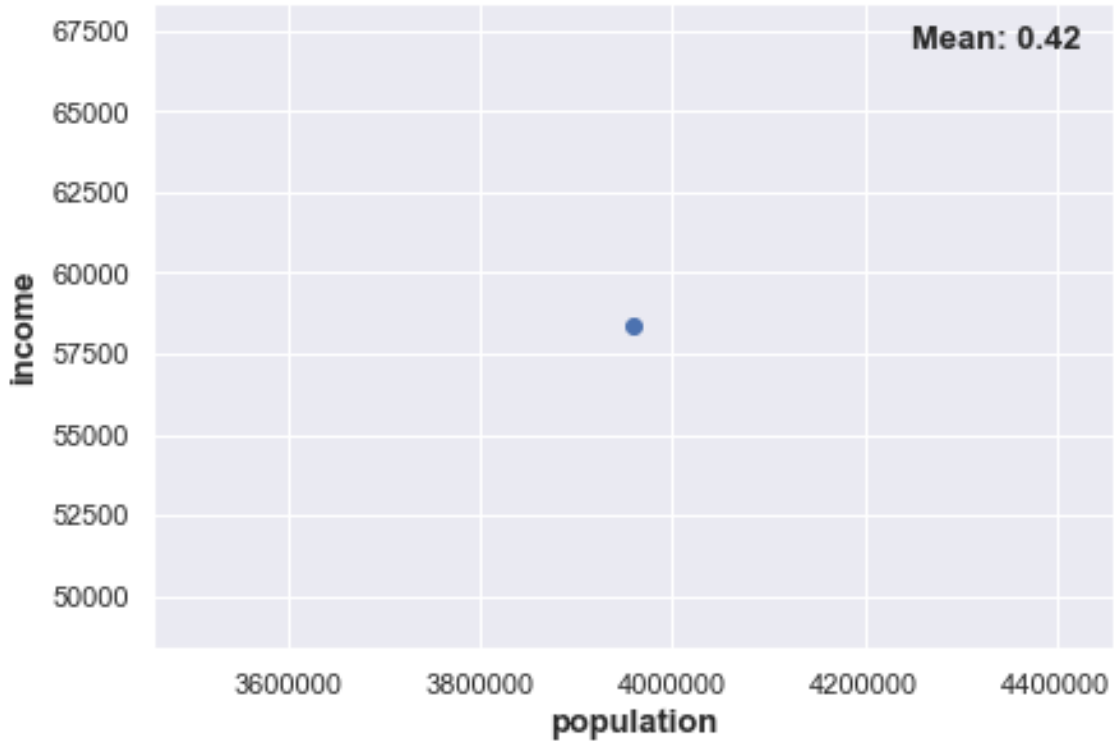

Figure 7: Relationship between population and income for Los Angeles City

2. Modeling

A. Clustering Model

Since there is no historical data for the target variable, an unsupervised machine learning approach must be taken. A K-Means Clustering model was implemented to segment similar cities into groups. The K-Means algorithm creates initial centroid points equal to the number of clusters defined by the user. Next, the distance from every data point to each centroid is calculated using the Euclidean Distance Formula. The Euclidean Distance formula is

$$D(x, c) = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2}$$

where n is the number of dimensions, x_i is the value on the i^{th} dimension, and c_i is the centroid value on the i^{th} dimension. Each data point is then assigned to a cluster containing the closest centroid, and the Sum of Squared Error (SSE) is retained. The SSE function is

$$SSE = \sum_{j=1}^{m} (\boldsymbol{x_m} - \boldsymbol{y_m})^2$$

where m is the number of observations, $\boldsymbol{x_m}$ is vector containing values on observation m, and $\boldsymbol{y_m}$ is a vector containing the closest centroid values to x_m. The average of all observation coordinates within each cluster is used to update the centroid positions. This process is repeated until the minimum SSE is found.

B. Feature Transformation & Scaling

Further analysis revealed abnormal distributions and outliers for the population and income data.

To address these issues, the data was transformed using the Box-Cox method. A Box-Cox transformation changes non-normal features into an approximate normal distribution. Applying the Box-Cox transformation to the population data resulted in a

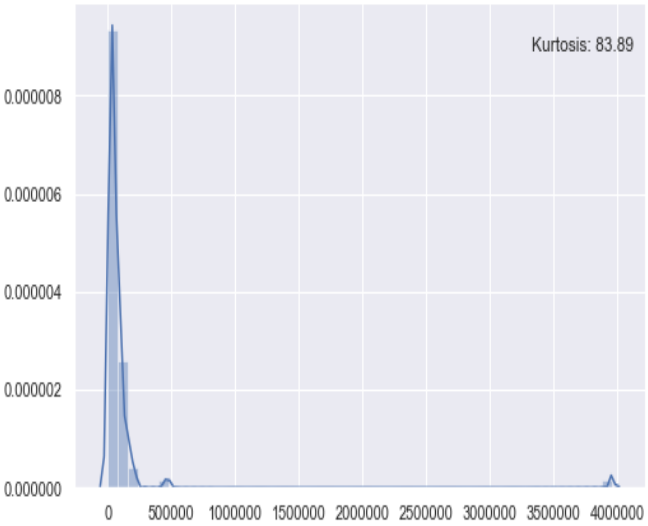

Figure 8: Population distribution before transformation

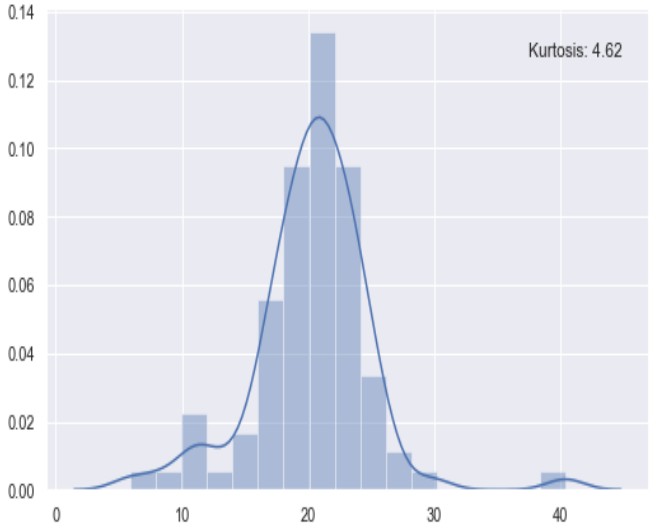

Figure 9: Population distribution after transformation

significant reduction in extreme values.

For the income data, outliers were removed due to the transformation.

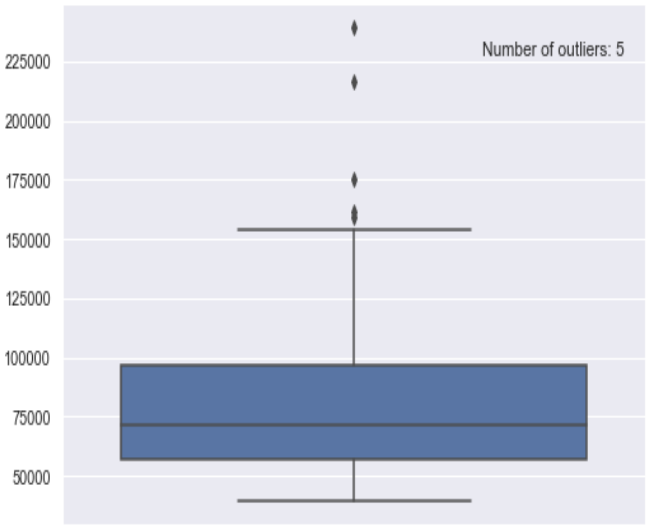

Figure 10: Income outliers before transformation

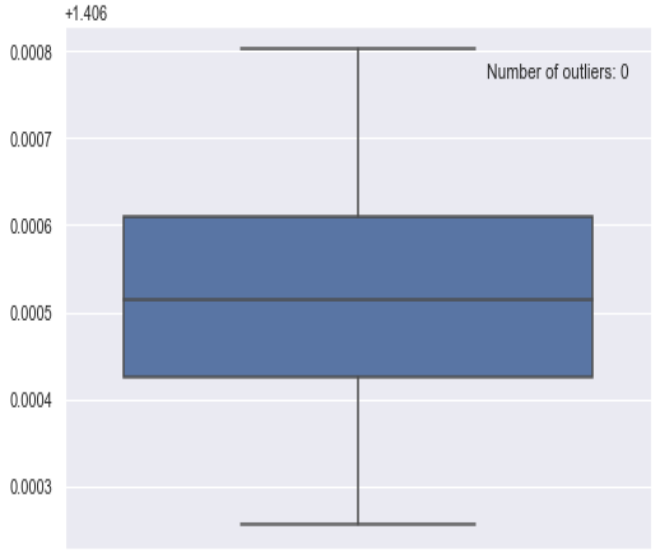

Figure 11: Income outliers after transformation

Since population and income have different unit of measures, the two features must be normalized. This will cause the values to have equal weight and improve the model's clustering ability. The transformed features were rescaled using min-max normalization method. The min-max formula is

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x' is the normalized value, x is the original value, x_{min} is the minimum value of the column data, and x_{max} is the maximum value of the column data.

Because the Box-Cox transformation and normalization method had minimal effect on the mean data, the feature was left the same.

C. Elbow Method

The number of clusters was determined by the elbow method. The abrupt change in slope indicates on the optimal number of clusters. For this project, the number of clusters will be five.

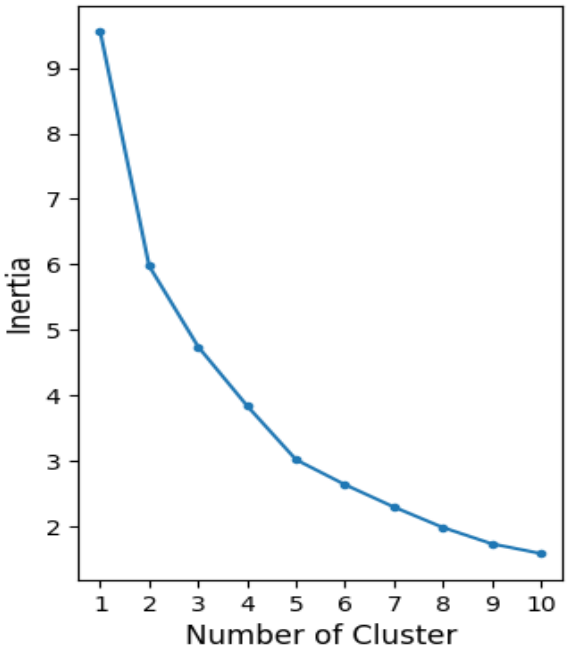

Figure 12: Elbow method

IV. Results

The cities were assigned to cluster groups by running the K-Means algorithm with five clusters as its parameter. Cities were plotted based on their coordinates using the Folium library. A marker for each city was color coded dependent on the cluster the city belonged to.

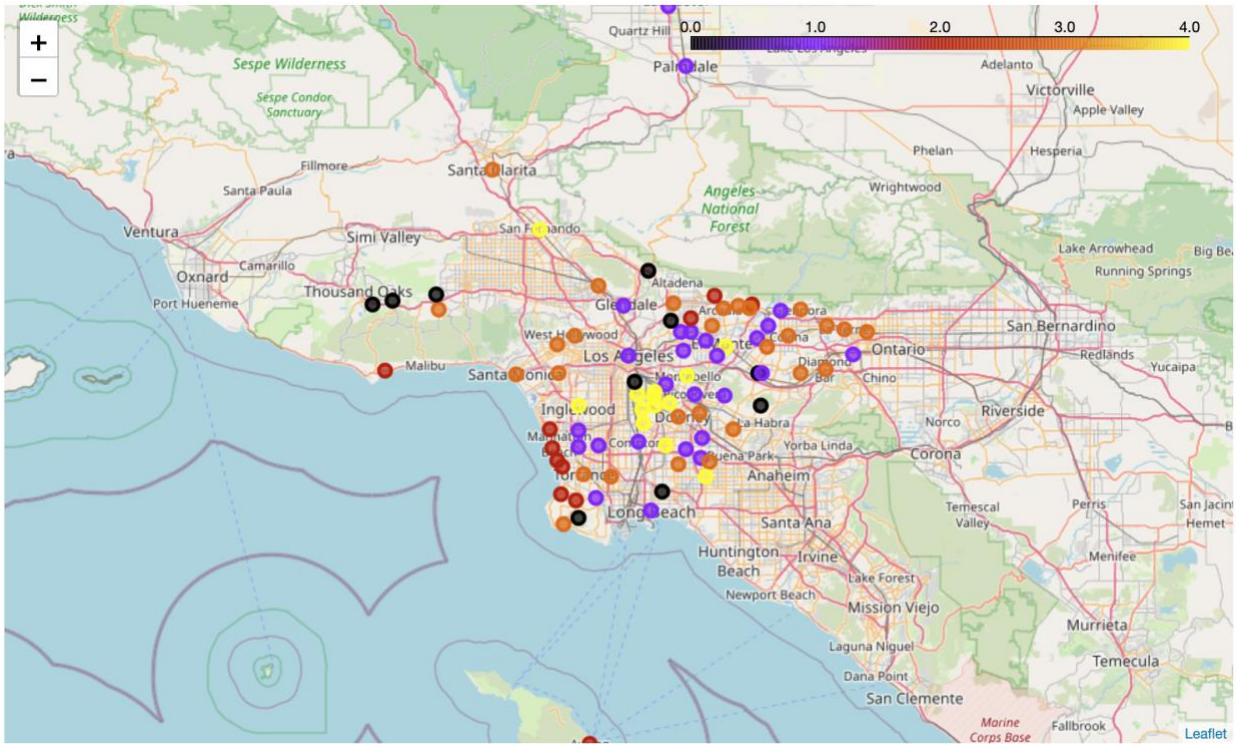

Figure 13: Map of color-coded cities

The population, income, and mean minimum and maximum values were obtained for each cluster.

	population		income		mean	
	min	max	min	max	min	max
cluster						
0	916	67700	69440	175000	0.500000	0.800000
1	1405	3959657	47083	73517	0.000000	0.500000
2	17791	209478	65518	133286	0.266667	0.545455
3	14411	115669	39738	53677	0.380952	1.000000
4	90	25824	70000	239375	0.000000	0.416667

Table 2: Clusters' minimum and maximum values

V. Discussion

Cluster 0 contains cities with low to medium population, medium to high resident income, and medium to high percentage of ice cream shops. If the ice cream quality can outcompete other venues' ice cream quality, then starting a business in these locations may be a decent decision. Cluster 1 contains cities with medium to high population, but low resident income. The percentage of ice cream parlors are average at best. If the ice cream will be affordable to customers, then setting up an ice cream shop in these areas is a good choice. Cluster 2 contains cities with medium population, medium to medium-high resident income, and low to medium percentage of ice cream shops. Cities in cluster 2 are the ideal location to start an ice cream business. Cluster 3 contains cities with decent population, low resident income, and many competitors. Setting an ice cream shop in these cities should be avoided. Cluster 4 contains cities with low population, medium to high resident income, and low percentage of ice cream shops. If the goal is to sell premium ice cream like Serendipity 3 ice creams, these cities will be great locations.

VI. Conclusion

Ice cream businesses have the ability to bring in high cash flow. LA County is a great place to start an ice cream shop due to the county's high population and extreme summer heat. However, an ice cream parlor in a poor location can result in bankruptcy. Therefore, population, resident income, and the percentage of ice cream businesses for each city in LA county were obtained. The data was then used to cluster similar cities. After analyzing each cluster's characteristics, cities within cluster 2 was determined to be the ideal locations to start an ice cream business.

VII. Room for Improvement

After choosing the city, the next step would be to analyze each census tract within city limits using the similar approach taken for this project. Doing so will pinpoint exactly where in the city will be the best area to start an ice cream business.

References

[1] International Dairy Foods Association. (2020). Ice Cream Sales & Trends. Retrieved from https://www.idfa.org/ice-cream-sales-trends

[2] Wikipedia. (2020). List of cities in Los Angeles County, California. Retrieved from https://en.wikipedia.org/wiki/List_of_cities_in_Los_Angeles_County,_California

[3] United States Census Bureau. (2019) American Community Survey 5-Year Data (2009-2018). Retrieved from https://www.census.gov/data/developers/data-sets/acs-5year.html

[4] Google. (2020). Get Started. Retrieved from https://developers.google.com/maps/documentation/geocoding/start

[5] Wikipedia. (2020). List of cities and towns in California. Retrieved from https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_California#cite_note-Census_2010-1

[6] Foursquare. (2020). Venue Search. Retrieved from https://developer.foursquare.com/docs/api-reference/venues/search/