Battle of the LA Cities

Objective

Ice cream is American's favorite dessert. Due to its high population and scorching summer heat, Los Angeles County is a great area to start an ice cream business. However, entrepreneurs are having trouble choosing which city to start their business. In order to find the best locations, cities will be grouped based on similar characteristics.

Figure 1: TasteofHome. (2020).Chocolate Crunch Ice Cream. Retrieved from https://www.tasteofhome.com/recipes/chocolate-crunch-ice-cream/

Data Description

- City: City name
- Population: Number of residents
- Income: Median household income
- Latitude: City center latitude
- Longitude: City center longitude
- Area: Land Area of the city in km^2
- Radius: Value to approximate a circle
- Mean: Number of ice cream shops divided by the total number of dessert venues.

	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

Data Acquisition

Web scraping

- Connected to web page
- Utilized BeautifulSoup to scrape desired data

```
def scrape(url):

    r = requests.get(url)
    html = r.content
    soup = BeautifulSoup(html, 'html.parser')

    return soup
```

```
list_of_cities = []

table = page.find('tbody')
for data in table.find_all('td'):
    for tag in data.find_all('a'):
        list_of_cities.append(tag.get_text())
```

Data obtained

- City Name
- Land Area

```
list_of_area = []

table = page.find('table', attrs = {'class': 'wikitable plainrowheaders sortable'})

for tag in table.find_all('th', attrs = {'scope': 'row'}):
    list_of_cities.append(tag.get_text().rstrip())
for area in table.find_all('td')[4::6]:
    list_of_area.append(area.get_text().rstrip())
```

Data Acquisition

Application Programming Interface (API)

- US Census API
- Google API
- Foursquare API

API usage

- Received Developer Keys
- Connected to API using request library
- Parsed html content to retrieve desired data

Data obtained

- Population
- Median household income
- Latitude
- Longitude
- Mean

```
def find_population(city_name):

    city_code_dict = censusdata.geographies(censusdata.censusgeo([('state', '06'),('place', '*')]), 'acs5', 2018)
    place_code = city_code_dict[city_name + ' city, California'].params()[1][1]

    url = 'https://api.census.gov/data/2018/acs/acs5/profile?get=DP05_0001E&for=place:{}&in=state:06&key={}'.format(
        place_code, Keys.CENSUS_KEY)

    r = requests.get(url)

    return r.json()[1][0]
```

```
def find_coord(name):
    g = geocoder.google('{}, California'.format(name), key = Keys.API_KEY)
    latitude, longitude = g.latlng

    return latitude, longitude
```

Feature Engineering

Calculated radius by using the Area Formula

$$A = \pi r^2$$

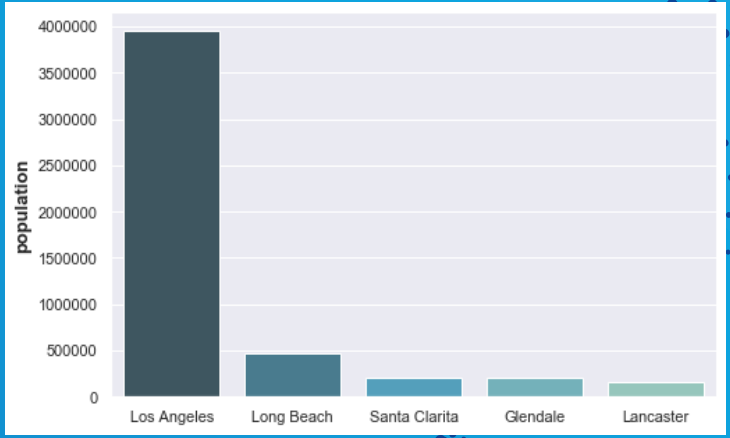
where A is the land area of the city.

Note: The radius is used to create a circle that approximately envelops the city.

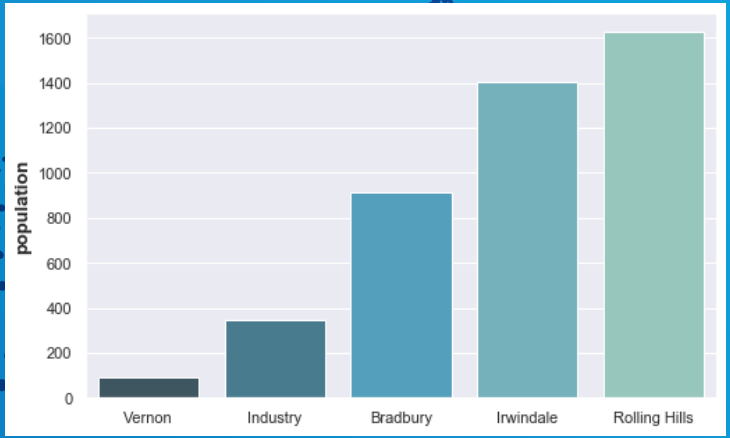
```
def find_radius(area):

    r = np.sqrt(area / np.pi)
    r = r * 1000 # convert to meters

    return r.round(decimals = 2)
```

Feature Analysis

Top-five most populated cities
- Los Angeles
- Long Beach
- Santa Clarita
- Glendale
- Lancaster



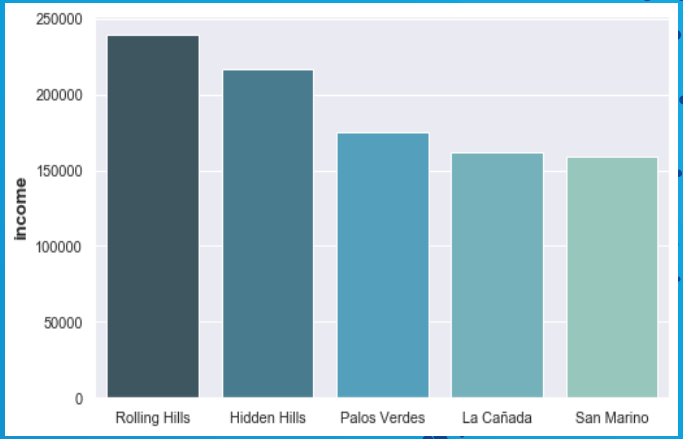
Top-five least populated cities
- Vernon
- Industry
- Bradbury
- Irwindale
- Rolling Heights



Feature Analysis

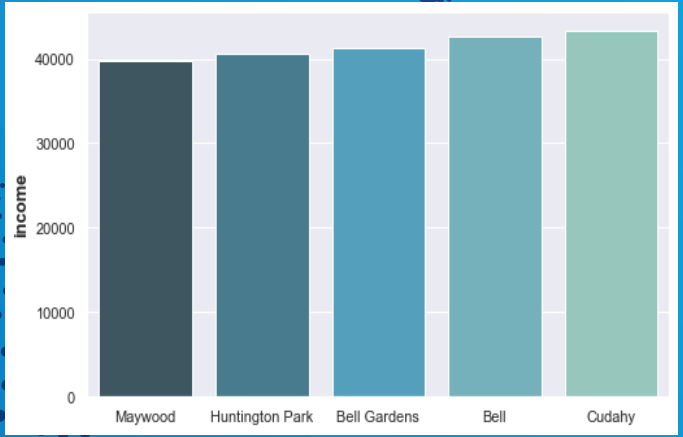
Top-five cities with the highest median household income

- Rolling Hills
- Hidden Hills
- Palos Verdes Estate
- La Cañada Flintridge
- San Marino



Top-five cities with the lowest median household income

- Maywood
- Huntington Park
- Bell Gardens
- Bell
- Cudahy

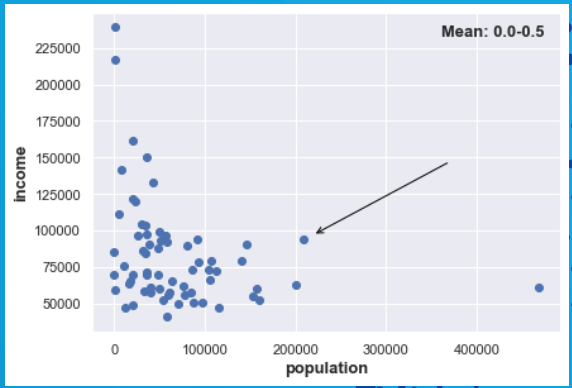


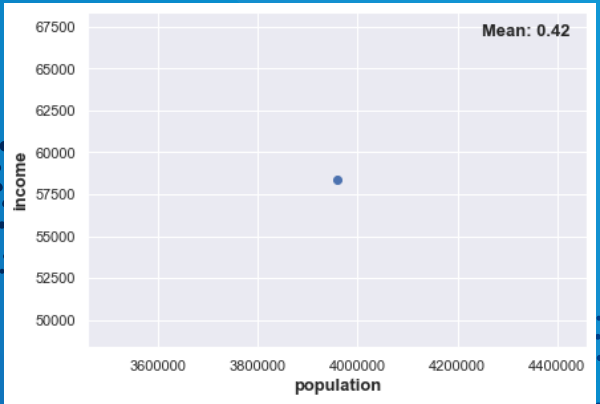
Feature Analysis

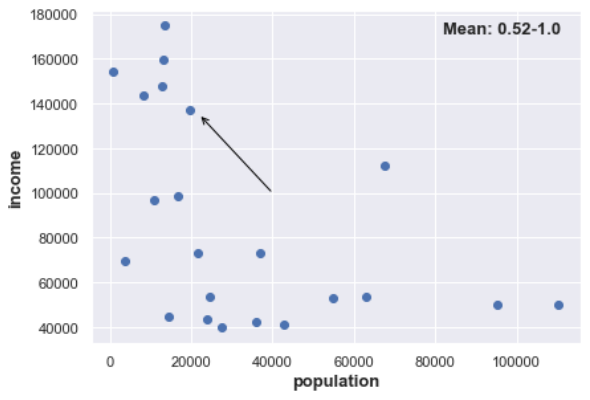
Separated into three data frames

- Cities with mean less than or equal to 0.50
- Cities with mean greater than 0.50
- Los Angeles City

Reviewed relationship between population and income







Modeling

K-Means

- No historical data for independent variable
- Clusters similar cities together.
- Optimal location can be decided based on cluster with appealing characteristics.

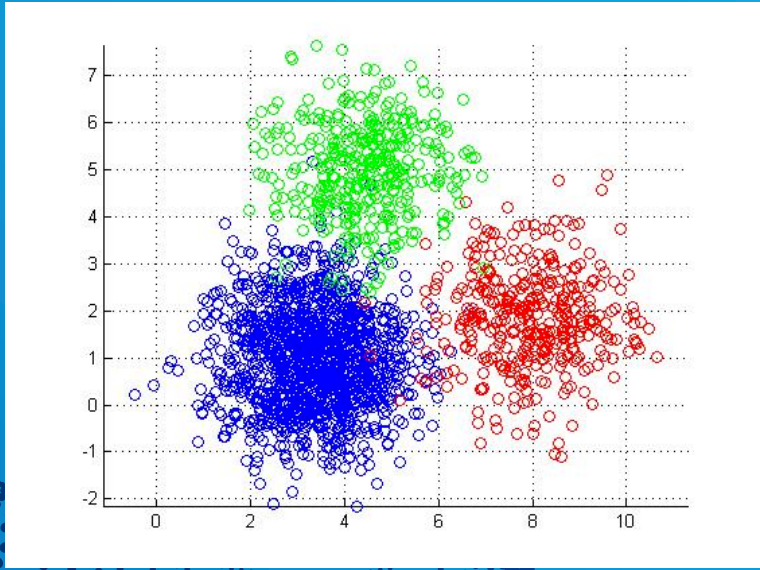


Figure 2: Mo Chen (2020). Kmeans Clustering (https://www.mathworks.com/matlabcentral/fileexchange/24616-kmeans-clustering), MATLAB Central File Exchange. Retrieved September 7, 2020.

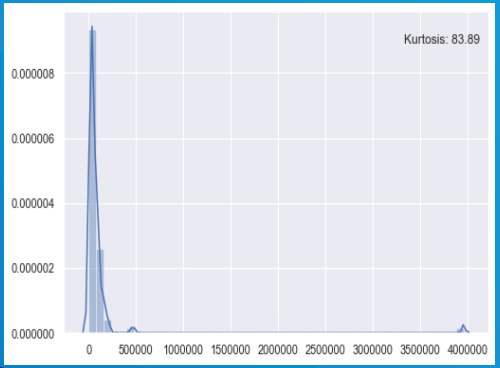
Feature Transformation & Scaling

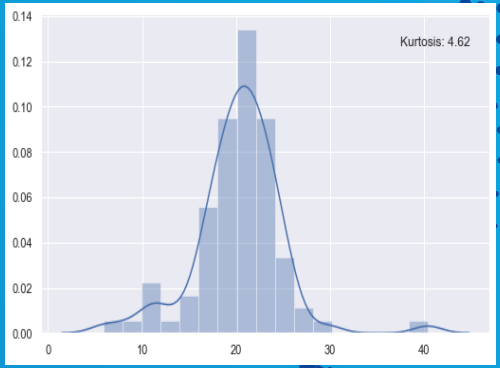
Feature transformation

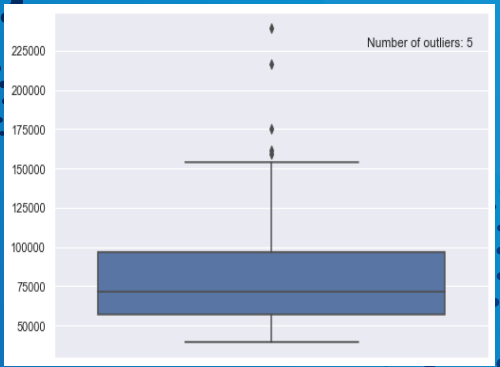
- Box-Cox Transformation on population data significantly reduced the kurtosis
- Box-Cox Transformation on income data removed outliers

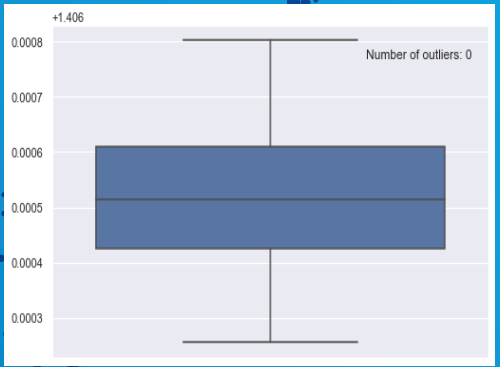
Feature normalization

- Min-max method applied to population and income data
- Removed unequal variances







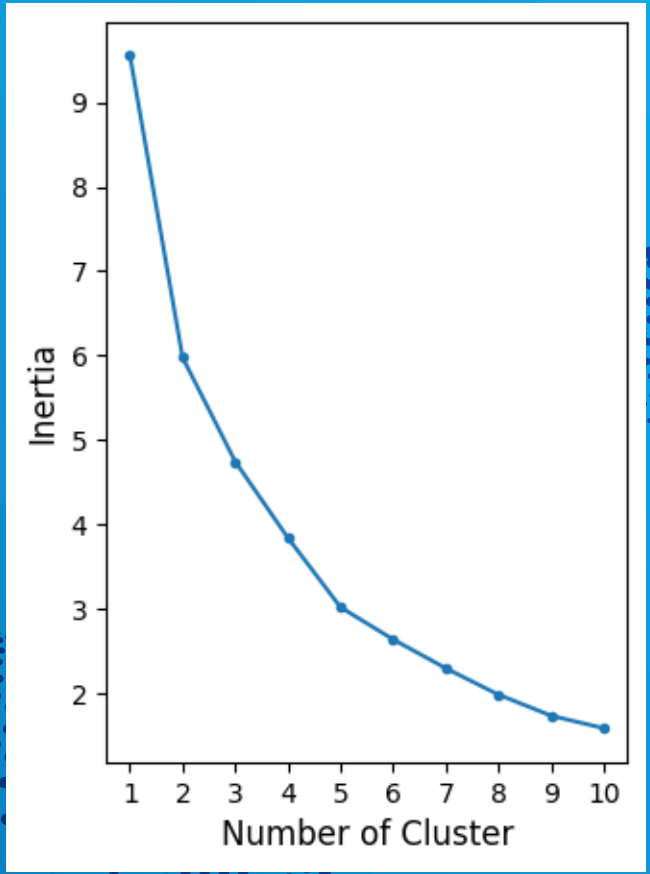


Elbow Method

Optimal number of clusters was chosen based on the elbow method.

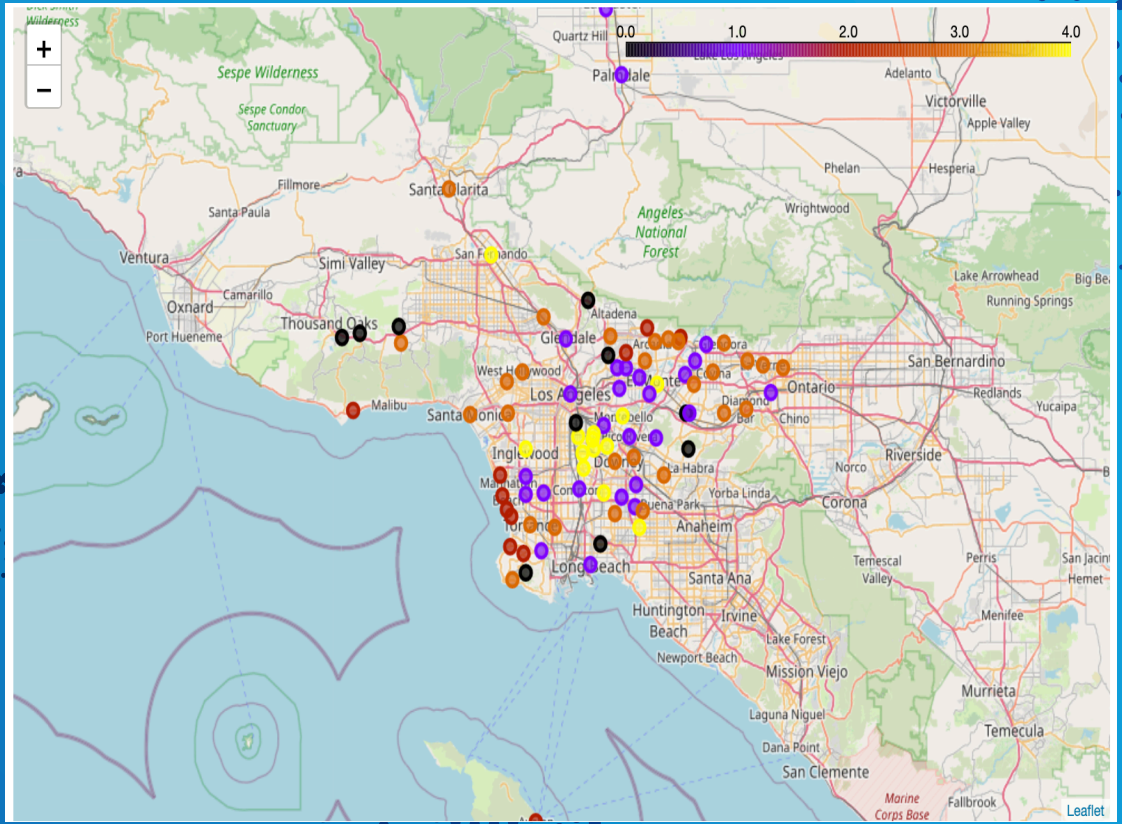
Sharp change in slope dictates the number of clusters.

- 5 clusters is optimal



Results

Cities are assigned to clusters



Business Decision

- *Cluster 0*: Can be good if ice cream quality outcompetes.
- *Cluster 1*: Good for selling affordable cheap ice cream.
- *Cluster 2*: Ideal location to start an ice cream business.
- *Cluster 3*: Avoid starting a business in these cities.
- *Cluster 4*: Good for selling expensive ice cream.

	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

Conclusion & Future Improvement

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.

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.