# IBM Data Science Capstone Project

## Detecting Food Deserts in Los Angeles County

## Introduction

Food deserts are areas where access to affordable, nutritious food is limited. Access can be defined as inadequate transportation, inadequate income, or even as simple as not having a farmer's market or grocery store within a certain radius.

Food deserts can also be associated with multiple health problems such as diabetes, high blood pressure, and other components of metabolic syndrome. Local governments are now trying to implement programs that mitigate food deserts.

## Business Problem

Los Angeles County has a high unemployment rate and a high disparity in incomes among zip codes. Given these characteristics, it may become prone to food deserts. As most city and county budgets are already strained, proliferation of food deserts could also lead to increased public health problems, thus creating a cascade effect of strained public health and social resources.

Identification of food deserts could assist city and county governments during permitting and development processes, helping to direct new investment from grocery chains and farmers markets to areas where these resources are needed most. Local governments could also offer incentives to appropriate vendors in the forms of tax benefits, tax credits for employment, or other expenditure offsets.

## Data

Data for this project will be aggregated from several sources:

1. Foursquare API – Location data for grocery stores and farmers markets to identify the current status for access to nutritious foods
2. LA Almanac – Median income data by zip code in Los Angeles County and all zip codes in LA County; identify areas at or below the poverty line or below the overall median income in LA County and/or California. Also used to obtain population data by zip code for entirety of LA County.
3. LA Times – Zip code boundary data for use in choropleth mapping of median income in LA County

## Methodology

Data for the study was ingested using various methods, including directly from files and scraped from websites. Data cleaning and exploratory analysis was performed using Python's pandas module, evaluating shape and composition of the data, including evaluating for any missing data in columns.

Zip code data from the LA Times was in JSON format and read into a pandas DataFrame. Three columns were kept to populate the initial DataFrame: zip code, latitude, and longitude of a central point in the zip code. Latitude data was then cleaned into a useable format.

Income data for Los Angeles County was retrieved from LA Almanac's website. This was read directly into a DataFrame and then currency formatting removed and converted to an integer for use later during normalization and machine learning applications. The poverty ratio was then calculated by dividing the median income of each zip code by the federal poverty level for a 4-person household and added to the DataFrame. Scikit-learn's MinMaxScaler function was then used to normalize the median income for all counties and then also added to the income DataFrame.

Population data for each zip code in Los Angeles county was also retrieved from LA Almanac's website. After being read into a pandas DataFrame, columns were cleaned and only the data for 2010 was kept, as 2019 was only estimated data. Population data was then normalized using scikit-learn's MinMaxScaler function. The resulting series was then added back into the population DataFrame.

A function was created to get any venue in the "grocery" category from the Foursquare API. The function used a 3000 meter search radius from the central point that was supplied in the geojson file from the LA Times.

The resulting DataFrame returned by the function was then evaluated for uniformity and acceptability. Some of the venues were deemed unsuitable, as they categories that ended in "Restaurant". Venues with unsuitable categories were dropped from the DataFrame. Food stores that were assumed to be able to supply nutritious foods were then flagged as "Qualifying Food Store". Those categories were:

1. Grocery Store
2. Supermarket
3. Food
4. Market
5. Big box Store
6. Salad Place
7. Gourmet Shop
8. Health Food Store

One hot encoding was then performed using the store category and the resultant DataFrame was grouped by zip code with mean of each category being calculated for each zip code. The sum of the qualifying food score flag was then added back into the grouped DataFrame. The grouped DataFrame also had median income, poverty ratio, normalized median income, 2010 population, and normalized 2010 population data added back in.

K-means clustering was applied to the DataFrame. An initial evaluation of various numbers of clusters was performed, applying the elbow method. The optimal cluster number was chosen for the final model, and the final model was then run and results used to add a "Cluster Label" column back into the DataFrame.

The final DataFrame and geojson data were used to create a choropleth map with cluster markers plotted at each central coordinate within the zip code.

Each cluster was evaluated for number of zip codes, and mean values per zip code of qualifying food stores, population, income, and poverty ratio. Data was then sorted by number of qualifying food stores and estimated median income in ascending order and 2010 population in descending order, in order to prioritize large groups of people with poorer income levels and fewer quality food options.

## Results

The total number of zip codes in Los Angeles County is 282. After all data was pooled and aggregated, 213 zip codes remained. In those zip codes, 810 venues were returned from the Foursquare API using the search parameters of "grocery" venue type within a 3000-meter radius of the center of each zip code. After removing erroneous categories, 600 venues remained over the resultant 213 zip codes.

During evaluation of k-means clustering, it was determined via the elbow method that the optimum number of clusters for this model was four. After running k-means clustering and evaluation of the clusters, the 5 optimal zip codes and supporting data for each cluster are:

| **Cluster 1** | | | |
|---|---|---|---|
| **Zip Code** | **Qualifying Food Stores** | **Median Income** | **2010 Population** |
| 90037 | 1 | 40,598 | 66,266 |
| 90011 | 1 | 49,675 | 62,180 |
| 90255 | 1 | 37,072 | 59,185 |
| 90022 | 1 | 40,940 | 103,892 |
| 90221 | 1 | 56,389 | 64,458 |

| **Cluster 2** | | | |
|---|---|---|---|
| **Zip Code** | **Qualifying Grocery Stores** | **Median Income** | **2010 Population** |
| 90502 | 1 | 73,826 | 18,010 |
| 90604 | 1 | 74,944 | 39,407 |
| 91345 | 1 | 77,273 | 18,496 |
| 90605 | 1 | 78,297 | 40,331 |
| 90807 | 1 | 78,948 | 31,481 |

| Cluster 3 | | | |
|---|---|---|---|
| Zip Code | Qualifying Grocery Stores | Median Income | 2010 Population |
| 91046 | 1 | 0 | 156 |
| 90058 | 1 | 21,964 | 3,223 |
| 91731 | 1 | 42,293 | 29,591 |
| 90008 | 1 | 43,364 | 32,327 |
| 90270 | 1 | 44,124 | 27,372 |

| Cluster 4 | | | |
|---|---|---|---|
| Zip Code | Qualifying Grocery Stores | Median Income | 2010 Population |
| 91354 | 1 | 128,308 | 28,722 |
| 91105 | 1 | 131,158 | 11,254 |
| 90275 | 1 | 138,293 | 41,804 |
| 90094 | 1 | 155,417 | 5,464 |
| 91108 | 1 | 165,765 | 13,361 |

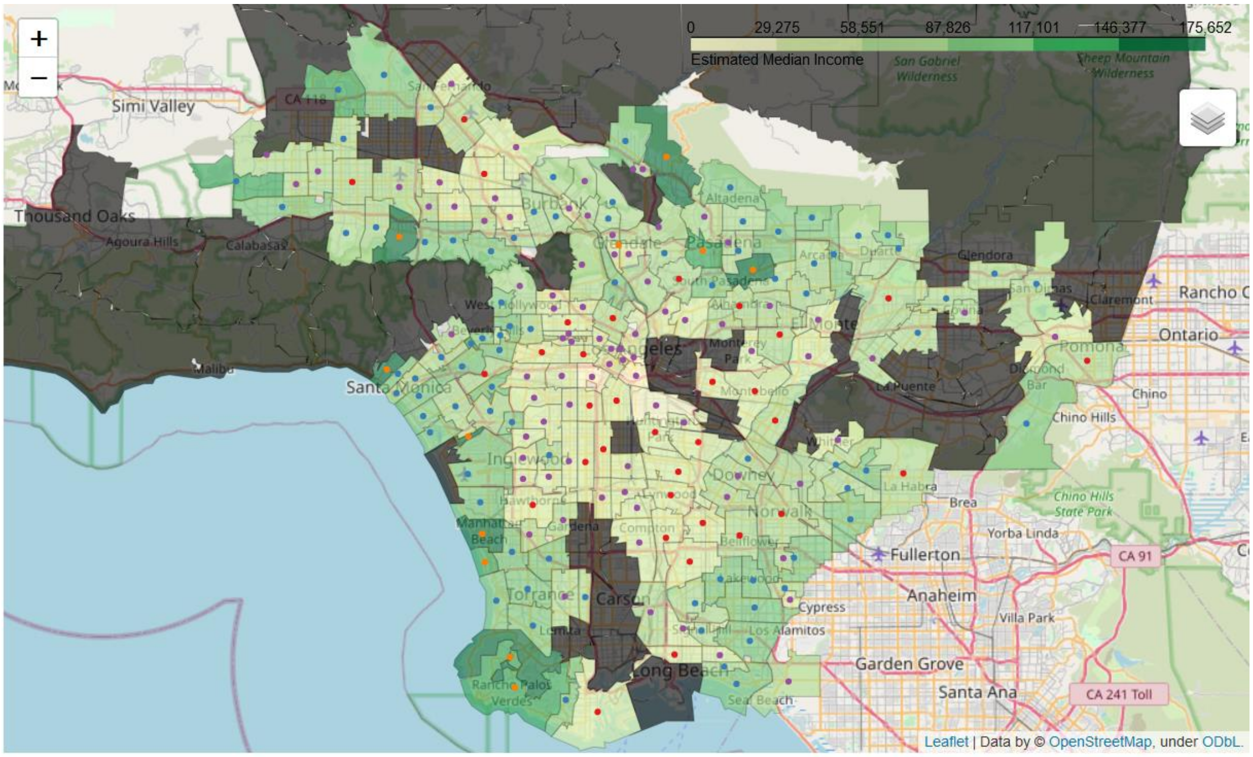

*Figure 1. Choropleth map of Los Angeles County. Zip code colors are median income. Colored points are assigned clusters of each zip code.*

## Discussion

As seen in the cluster data above, Cluster 1 or Cluster 3 would be leading candidates for evaluation of grocery store placement for disadvantaged residents. These assumptions are made based on the lower median income and higher population. Zip code 91046 likely should have been removed from the data set given the zero population. The effect this may have had on clustering was not evaluated.

The results observed during this study may need further evaluation. Feature selection may be imperfect and/or incomplete, and multiple different machine learning models were not evaluated. There are newer packages in Python, such as PyCaret, that could evaluate multiple different algorithms and models with far less code. However, that was not in the scope of this study.

## Conclusion

Upon evaluation of the results, I would recommend the evaluation of Cluster 1 or Cluster 3, specifically focusing on 90255 from Cluster 1 and 90008 from Cluster 3.