Capstone Project - The Battle of Neighborhoods - Public
August 17, 2020
Capstone Project - The Battle of Neighborhoods
Project for Applied Data Science Capstone by IBM on Coursera
By: Eric J. Puttock (August 16, 2020)

This is an informal report outlining my detailed project. This was made to compliment the notebook to quickly outline/sketch the project. More details are presented within my Jupyter Notebook.

Introduction.

In the changing world today, we are always interested in venues around us. People are often curious to the kinds of restaurants, shops, or entertainment venues exist in their area to spend time with friends and family. This project aims to identify the most common venues of each city within Los Angeles County and to determine whether certain cities can provide similar experiences. The project will also attempt to determine whether more population of a city show differences or similarities among different cities. The audience for this project is primarily for people who are planning to live within Los Angeles County, or for beginner data scientists interested in getting started.

Data.

This project utilizes the 2010 Census data of each city of Los Angeles County obtained from Wikipedia (https://en.wikipedia.org/wiki/List_of_cities_in_Los_Angeles_County,_California). Geospatial coordinates of each city was used together the help of Foursquare API (https://foursquare.com/) to obtain top venues. Finally, a publicly available GeoJSON file from Geohub LACity (https://geohub.lacity.org/datasets/lacounty::city-boundaries-3) was used to outline the boundaries of each city in the county on OpenStreetMap. Python and several popular libraries (pandas, sklearn, matplotlib, etc.) were used to complete this project. Additionally, folium was used to help visualize the data on OpenStreetMap.

Methodology.

The pandas library was used to extract 2010 Census population data of each city of Los Angeles County from Wikipedia (see link at the end of the document). I broke up the cities by population size into four distinct sizes (Low, Medium, High, Extreme). Two outliers (Los Angeles & Long Beach) were placed in their own Extreme group. Low, Medium, and High were split into three sets of equal number of cities. Additional information regarding when the city was established was also available from the Wiki page but was not used in this study. Nominatim, a search engine for OpenStreetMap data was used to identify the geospatial coordinates of each city. Using the geospatial coordinates and the Foursquare API, we explored venues within 2-miles of each city (and restricted to first 100 venues obtained). To do this, we used a regular endpoint get request to venues/explore (https://api.foursquare.com/v2/venues/explore). After extracting and cleaning up the data, top 10 most common venues of each city was obtained. Using a common unsupervised learning algorithm called K-Means, we grouped cities together based on their similarities to most common venues. This helps identify venues nearest to your current location and provide ideas

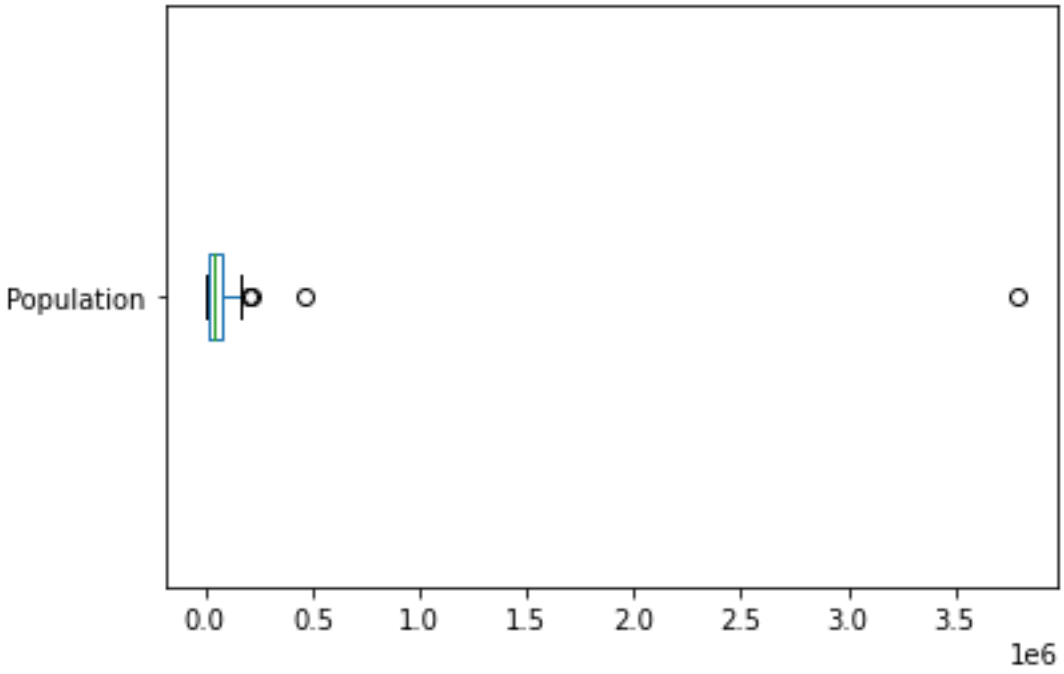
for future planning. Additionally, simple bar graphs demonstrating which venues were common grouped by population or the K-Means clustering were used to visualize. Finally, using Folium we produce a world map and choropleth map showing the geographical location of each city, how they clustered together (shared colors), and their population groups. The codes used for this project is publicly available for download from my github account (https://github.com/EJPanda/Coursera_Capstone).

Results.

After reading in the data, we formed a data table.

	City	Latitude	Longitude
83	Walnut	34.020289	-117.865339
84	West Covina	34.068621	-117.938953
85	West Hollywood	34.092301	-118.369289
86	Westlake Village	34.146023	-118.806179
87	Whittier	33.970878	-118.030840

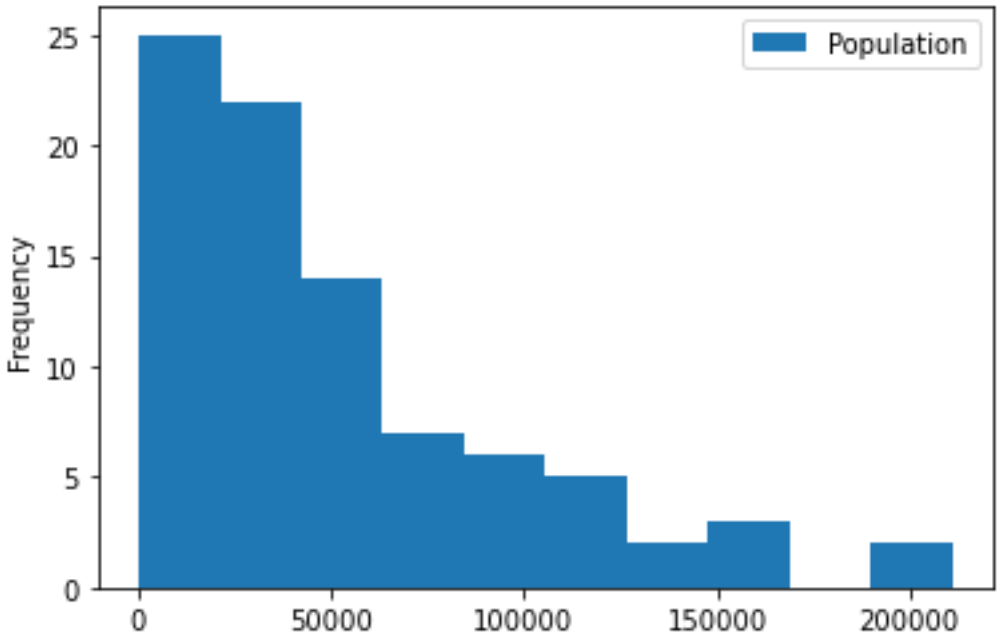
Remark: The population data was highly skewed due to Los Angeles and Long Beach as you will see the following box plot.



Let's see what some of those outliers are by sorting the population.

	City	Population
48	Los Angeles	3792621
47	Long Beach	462257
72	Santa Clarita	210888
28	Glendale	203054
44	Lancaster	160316
57	Palmdale	152750
62	Pomona	149058
81	Torrance	145438
60	Pasadena	137122
25	El Monte	113475

As one could see, there's a significant population gap between Los Angeles and Long Beach.
Let's see how the distribution looks like without these outliers.



Let's bin these into three population groups of roughly the same size.
Long Beach and Los Angeles will be grouped together to form its own group.
We assign to each city an interval group relative to their population using pandas qcut method.
This divides up the data into q many intervals of equal size (as much as possible).

```
Group 3 (Extreme) corresponds to interval (462256.999, 3792621.0].
Group 2 (High) corresponds to interval (58590.667, 210888.0].
Group 1 (Medium) corresponds to interval (24409.667, 58590.667].
Group 0 (Low) corresponds to interval (111.999, 24409.667].
```

	City	Population	Group #	Group
0	Agoura Hills	20330	0	Low
1	Alhambra	83653	2	High
2	Arcadia	56364	1	Medium
3	Artesia	16522	0	Low
4	Avalon	3728	0	Low

	Group	Counts
0	Extreme	2
1	High	29
2	Low	29
3	Medium	28

Utilizing Geocoder - Nominatim OpenStreetMap and Foursquare API:

Using Geocoder - Nominatim OpenStreetMap we find the geospatial coordinates of each city name in Los Angeles County. Using the Foursquare API, we determine the top 10 most common venues of each city of Los Angeles County. We do this by finding 100 venues within a 2-mile radius (roughly 2319 meters) from the geographical coordinates of each city. For each city, we extract the top 10 venues category. Note that this doesn't necessarily mean the venue is necessarily popular but exist in greater quantities in comparison to other venues within the city range.

Identify the coordinates of Los Angeles, California. The coordinates here will be used for our Folium world map plots. Originally, Los Angeles County coordinates were used, but the map becomes off center since the county is wide. We center the map around Los Angeles. The geographical coordinate of Los Angeles, Los Angeles County, California is 34.0536909, -118.2427666.

After a lot of data wrangling and cleaning of the data, we formed two data frames.
One containing top 10 most common venues of each city, and a data frame containing all venues from each city up to a 2-mile radius with a 100-venue cap set for the Foursquare API.

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
83	Walnut	Fast Food Restaurant	Pizza Place	Park	Coffee Shop	Japanese Restaurant	Mexican Restaurant	Asian Restaurant	Furniture / Home Store	Sandwich Place	Liquor Store
84	West Covina	Mexican Restaurant	Pharmacy	Coffee Shop	Fast Food Restaurant	Ice Cream Shop	Grocery Store	Burger Joint	Thai Restaurant	Asian Restaurant	Bubble Tea Shop
85	West Hollywood	Hotel	Café	New American Restaurant	Coffee Shop	Gym	Gym / Fitness Center	Cocktail Bar	Clothing Store	Boutique	French Restaurant
86	Westlake Village	Italian Restaurant	American Restaurant	Hotel	Mexican Restaurant	Sushi Restaurant	Brewery	Grocery Store	Park	Burger Joint	Seafood Restaurant
87	Whittier	Mexican Restaurant	Fast Food Restaurant	Coffee Shop	Trail	Donut Shop	American Restaurant	Bakery	Pharmacy	Convenience Store	Rental Car Location

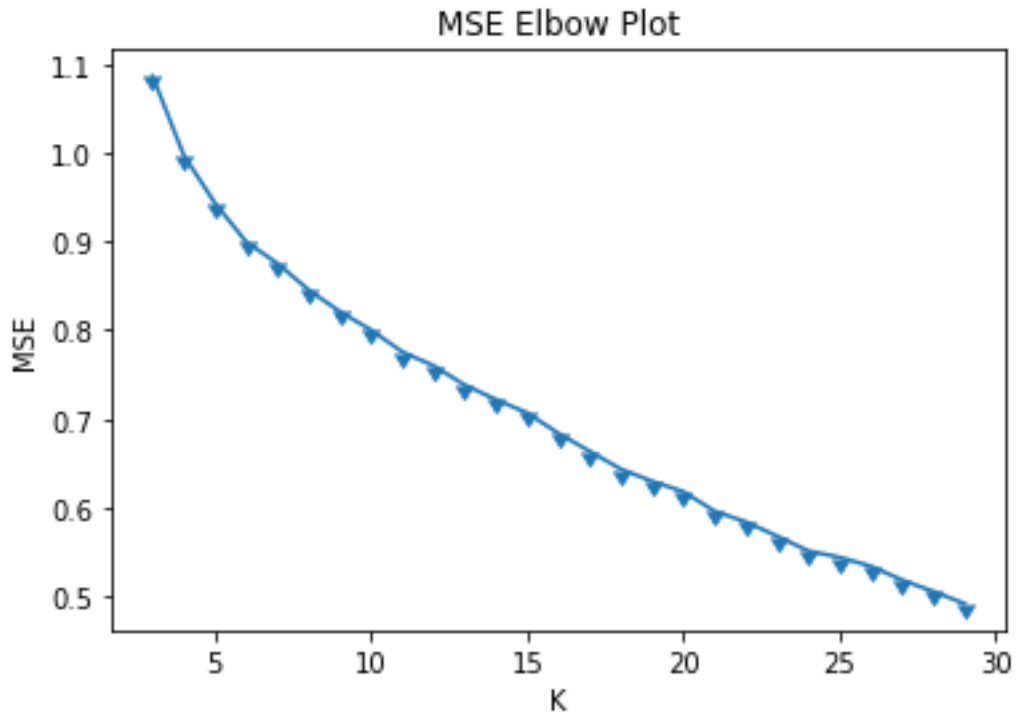
	City	Venue Category	Venue	Venue Latitude	Venue Longitude
0	Agoura Hills	Fast Food Restaurant	El Pollo Loco	34.144732	-118.761088
1	Agoura Hills	Deli / Bodega	Italia Deli & Bakery	34.153099	-118.759037
2	Agoura Hills	Brewery	Ladyface Alehouse & Brasserie	34.143834	-118.762823
3	Agoura Hills	Grocery Store	Trader Joe's	34.146241	-118.756609
4	Agoura Hills	Dessert Shop	Tifa Chocolate & Gelato	34.144586	-118.754595

K-Means unsupervised clustering to group cities by similarity.

When determining the top 10 most common venues, data frame containing the proportion of each city's venues were made. We train using this data.

	City	ATM	Accessories Store	Advertising Agency	African Restaurant	Airport	Airport Lounge	Airport Terminal	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Arcade
83	Walnut	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
84	West Covina	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.020000	0.0	0.000000	0.0	0.0
85	West Hollywood	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.020000	0.0	0.000000	0.0	0.0
86	Westlake Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.050000	0.0	0.000000	0.0	0.0
87	Whittier	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.035714	0.0	0.011905	0.0	0.0

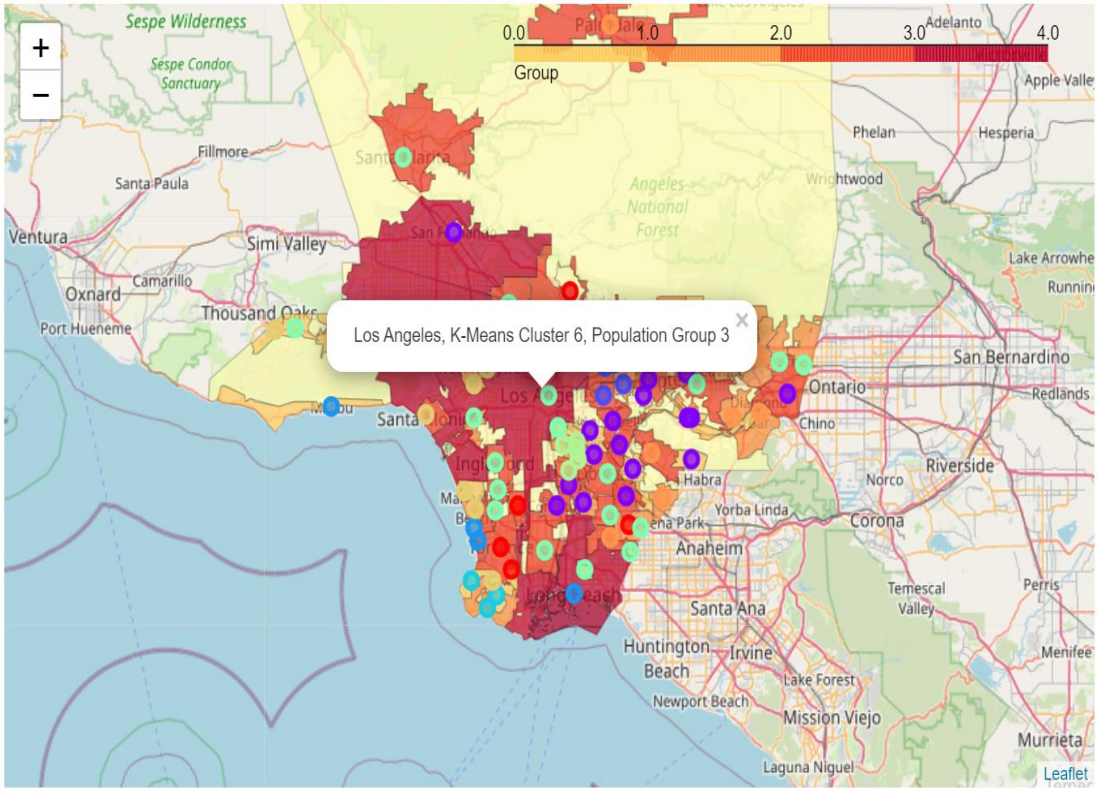
We use the elbow method to visually estimate what an appropriate cluster count should be. We measure the Mean-Square-Error (MSE) and determine a cluster size where decrease of MSE slows down. We run 12 different initializations for each cluster size, then compare the minimum MSE for each cluster size.



We estimate around cluster size K = 11 there's a slight shift in MSE reduction.

Visualization using Folum & Choropleth map

We visualize our clusters using the Folium package. We also provide a choropleth map of the population groups to see if we can visually see a pattern.



Discussion

We studied the cities within Los Angeles County using the 2010 census data and their most common venues using the foursquare API. We attempted to understand whether any of the cities share common venue characteristics by population size, or geographical location. We applied K-means to cluster cities in similar in terms of top 10 venues, and we applied a simple 3-fold population size change (excluding the Los Angeles and Long Beach). There were ~60 unique venue categories in each of the Low, Middle, and High population groups. Among these venue categories the data shows that regardless of population size, people are often surrounded by coffee shops and Mexican restaurants.

We identified some of the common trends among the clustered cities, and list out a few of them below.

Cluster 0: Bakery & Japanese Restaurant & Coffee Shop
Gardena, Glendale, Lomita, Torrance
Cluster 1: Fast Food & Mexican Restaurant & Convenience Stores & Sandwich Places
Baldwin Park, Irwindale, Montebello, Pomona
Cluster 2: Chinese & Vietnamese Restaurants
Alhambra, San Gabriel, Temple City
Cluster 3: Seafood Restaurants & American Restaurants throughout the city.
Long Beach, Redondo Beach
Cluster 4: Trail & Coffee Shop & Mexican Restaurant
Palos Verdes Estates, Rolling Hills
Cluster 5: Inconclusive. Only two cities. Share Cosmetics Shop and a Bank.
Calabasas, Hidden Hills
Cluster 6: Coffee Shop & Mexican Restaurant & Fast Food & Pizza, Sandwich Places & Parks
Arcadia, Burbank, Inglewood, Los Angeles,
Cluster 7: Coffee Shop & Fast Food Restaurant & Pizza Place & Mexican Restaurant
Bell, Huntington Park
Cluster 8: Coffee Shop & American Restaurant & Burger Joints & Hotels
Beverly Hills, Pasadena, West Hollywood
Cluster 9: Coffee Shop & Fast Food Restaurant & Pizza Place & Mexican Restaurant (Similar to Cluster 7)
Diamond Bar, La Mirada, Whittier
Cluster 10: Inconclusive. Only one city. Have a bar, beach, and history museum.
Avalon

The prevalence of coffee shops and Mexican restaurants was also seen in most K-Means clusters (groups of cities). Interestingly, one cluster had Japanese foods being one of the most common venues. Few other clusters preferred Chinese/Vietnamese foods and while another preferred Indian food. Some clusters did demonstrate (not all) similarities in geographical locations. For example, cities in cluster 1 seemed to be concentrated in mid-east region of Los Angeles County in high population regions.

By expanding the number of venues for each city, rather than 100, a better understanding of the top venues is obtainable. Due to the limitations of the Foursquare API for free accounts, additional studies such as venue ratings, menu items, and prices of each restaurant were not obtainable. It would be interesting to see whether those data sets can help us paint the bigger

picture of the city. Another thing to consider is that this is a snapshot of the city when the data was collected. These popular venues may not shift too much, but as different venues become available and listed on the Foursquare API, the top venues change. This could lead to changes in cluster groupings as the world changes. It would be interesting to see how these changes in common venues occur over time.

Conclusion

The project provides an informative broad picture of Los Angeles County. Some unexpected outcomes when doing this project was finding Japanese restaurant chains I didn't know existed in the United States! Some cities that came up popular for Japanese foods was Glendale and Gardena regions. The study shows Alhambra regions have many Chinese restaurants to explore.
Finally, if you love coffee or Mexican foods, they're everywhere within Los Angeles County (regardless of population, or city clustering).
Thanks to this project, I now have an interest to explore more venues around Los Angeles County.
Let's grab some tacos/burritos while exploring Los Angeles County!

References & Acknowledgements:

Wiki Page Link:
https://en.wikipedia.org/wiki/List_of_cities_in_Los_Angeles_County,_California

Thanks for the IBM Coursera courses and instructors for this fun opportunity to explore Los Angeles County!

Additionally, I thank the readers/learners for taking the time to read my project.

The codes used for this project is publicly available for download from my github account (https://github.com/EJPanda/Coursera_Capstone).