Course 9 - W5 - Report Only

February 11, 2019

1 CLUSTERING LOS ANGELES NEIGHBORHOODS

Gerald Nacache geraldnacache@gmail.com

Project link :

This notebook is submitted as a part of final assignment of Coursera Course 9 (week 5) of the IBM Professional Data Science Certificate.

As required, please let me submit the following:

A full report consisting of all of the following components: . Introduction where we discusse the business problem and who would be interested in this project.

. Data where we describe the data that will be used to solve the problem and the source of the data.

. Methodology section which represents the main component of the report where we discuss and describe any exploratory data analysis that we did, any inferential statistical testing that we performed, and what machine learnings were used and why.

. Results section where we discuss the results.

. Discussion section where we discuss any observations we noted and any recommendations we can make based on the results.

. Conclusion section where we conclude the report.

This notebook only contains markdown cells and another notebook is provided with all the coding plus detailed explanations of it.

2 Introduction

2.1 Description of the Problem and Context:

An investment company invested in and took control of "Best Hair Style Ever" (BHSE) a hair dressing brand currently operating several hundreds hairdresser stores in Europe. One of the decision made during last business strategy company meeting was to confirm a plan to establish and deploy the brand in the USA via launching stores in targeted cities. As a consequence, a first pilot project is launched, and Dan, project leader was nominated with first objective to launch 20 new stores in Los Angeles, CA. The success of this initiative is crucial for the next phases of the project and for the development of the whole company.

Among Dan's key tasks, he needs to work with a real estate consultant to identify available venues in Los Angeles and close relevant deals as soon as possible to deploy the brand. After a first discussion with Dan's preferred real estate broker in LA, he realized that the criteria he defined for finding the stores are not accurate enough and his broker told him a) he may spend too much time

finding the right places in such a large city like LA, so he needs to identify some preferred zones within Los Angeles to focus his search, and b) he may not be the best person to define priorities and most relevant areas for BSHE criteria, so he needs more guidance with targeted areas.

As a consequence, Dan contacted me and asked for some help on where I would recommend they should open the first BSHE stores in LA. Dan and I discussed and we came to a conclusion that the problem could be solved with defining a list of preferred areas in Los Angeles issued from classifying neighborhoods based on exploring existing venues and most frequent categories of venues in each candidate zone. This way, we can identify similar neighborhoods, gather them within several clusters and choose the right cluster of areas within Los Angeles to focus on. Such output will serve as a guidance and a list of target zones for Dan's real estate broker.

2.2 Description of the data that will be used to solve the problem, and how:

As Los Angeles is a large city, we have many neighborhoods to explore within the city itself (almost 100 zip codes). During the study, we will exchange on a regular basis together with Dan on first data extracted and first explorations in order to decide if we need to extend to Los Angeles county or keep the research within Los Angeles city boundaries. Los Angeles county will be a larger data of more than 400 zip codes.

Online search for location data let us find a database gathering all zip codes in Los Angeles, CA and each corresponding latitude, longitude coordinates. This database use is free to download and use; the only request is to clearly mention the link it is coming from https://simplemaps.com/data/us-zips so we will mention on each notebook or page or presentation related to this assignment. At this stage, the main info we plan to extract from this database are zip code, latitude, longitude, city, county, state. We also have access to the population and density for each zip code. We will extract such information as potential additional data to be used during the study, if needed. Such parameter like population and density are valuable info as they may come up in some discussions with the client during the project.

Knowing each zip code latitude and longitude, we can then explore categories of venues thanks to Foursquare API. We use the Foursquare location data to explore neighborhoods of LA, specifically categories of venues, in a similar way we did with Toronto area in the previous lab and assignment. We plan to use unsupervised machine learning method for classification, like k-means algorithm. This topic will be developed further in the next Methodology section of this report.

Some additional kind of data may be relevant to add for a deeper exploration: the foot traffic data. Such data is available if we build a process to get it for each location and we can even explore the foot traffic depending on the time in the day and the day in the week. Adding such data would mean defining the right method of usage of Foursquare API and would mean more time and effort, whereas the method we are using has been proven with previous examples like the one in Toronto during the lab. Knowing this, Dan asked me to focus for now on the current method without foot traffic data and we can extend our study further if needed. Another reason why we did not add such data is that the consumer habits in Europe and in California are fairly different. As an example, the stores BHSE operates in Paris are in zones with very high foot traffic. But this does not mean Dan wants to follow a location strategy in LA necessarily similar to the location strategy BHSE used in Paris. So we focus for now on unsupervised clustering method based on venues around candidate locations, and within each zip code in Los Angeles, CA area.

2.2.1 Additional tools for solving the problem:

In addition to Foursquare API we already mentioned, we will use jupyter notebook for all the coding and explanations of our method, process and computations.

Coding will be done in python 3, and leveraging usual libraries: NumPy and SciPy for scientific computing, Pandas for data extracting, cleaning and analysis, Matplotlib and Folium for figures, plots, maps and visualization, Scikit-learn for machine learning, in particular we plan to use clustering k-means algorithm

In this report, we mainly present explanations, notes and comments as described above, but we don't insert actual code. Another note book is provided as part of the assignment and gathers all the python code, dataframes, plots and maps involved in the capstone project.

As a summary of Intro and Data section, The objective is to build clusters to partitions Los Angeles, CA in similar areas and identify the most suitable areas for launching new stores for BHSE. We will leverage unsupervised method like k-means clustering algorithm to classify all Los Angeles, CA zip codes. For that, we will clean and leverage Foursquare location data - specifically to explore categories of venues in the neighborhood - as well as Los Angeles zip code location database.
Once clusters are created, we will review the clusters and identify similarities within a given cluster and unsimilarities between two different clusters. Then the target cluster of areas (zip codes) will be defined and validated with the client.

2.3 Methodology

Review our sources of data Let's review our data source and see what data preparation process is needed. As said above, we have two main sources of data: a) one database gathering all zip codes in Los Angeles, CA and each corresponding latitude, longitude coordinates. This database is coming from https://simplemaps.com/data/us-zips b) another database we'll build thanks to Fouresquare API to get neighborhood information for each (latitude, longitude) point that will be considered, and coming from the first source. Foursquare requests will help gather venues information and categories of venues in each considered area. In particular, computing the n most frequent categories of venues in a predefined radius will help build main features used by our clustering algorithm, and for achieving our classification objective of all considered areas (zip codes).

The first source of data (from simplemaps) is easily downloadable as a .csv file and transformed in a pandas dataframe.

```
In [14]: from IPython.display import Image
         Image("LAareas-zipcodes-image.png")
```

Out[14]:

	zip	lat	lng	city	state_id	state_name	population	density	county_name
23	92280	34.1256	-114.7779	Vidal	CA	California	14	0.0	San Bernardino
25	92304	34.5415	-115.6445	Amboy	CA	California	17	0.2	San Bernardino
27	92332	34.9127	-115.3417	Essex	CA	California	65	0.0	San Bernardino
30	93519	35.2975	-117.9294	Cantil	CA	California	101	0.5	Kern
48	95226	38.2286	-120.8581	Campo Seco	CA	California	84	5.7	Calaveras

Our database of zipcodes needs some reduction as it includes columns we won't use, plus we just want to keep Los Angeles area zip codes. We keep the columns refering to zip code, latitude, longitude, city, state name, population, density and county name. We can remove all other columns.

We end up with 284 zip codes in Los Angeles county - after we clean few of them. As an example, we found 6 zip codes lines where population is 0 in the database. This is just very special areas or administrative zip codes that we'll drop from the database. If we want to explore deeper, we can actually look at the zip codes on a map.

This how the zip codes data base looks like.

2.3.1 Let's visualize the areas we consider on a map of Los Angeles, CA

After we've cleaned our data base of zip codes, we want to visualize the areas on a map with zipcodes superimposed on it. Let's center on an area in West LA that corresponds to a good central point for our research. We choose Carthay Square neighborhood at zip code = 90048. This is close to Miracle Mile and the LACMA (LA County Museum of Art).

```
In [ ]:
```

First presentation for map visualization with the client From a discussion with the client, some questions were asked about the size of the area of research. In our database, we have 284 zip codes in Los Angeles county, spread among 126 cities. (as the two cells above show) The city of Los Angeles itself has 64 zip codes, among the 284 zip codes in the database. This could be a fairly large enough area for our study, especially knowing that the total population in these 64 zip codes is > 2.3M people, as the cell above is showing. But on the other hand, looking at the map of Los Angeles county, filtering on the city is not very relevant because many cities surrounded by parts of Los Angeles city would be removed from the study just for administrative reasons and it does not make sense if we look at how the neighborhoods are organized. As an example, cities like Culver City, Beverly Hills, Santa Monica or West Hollywood would be removed, whereas they can be considered on the field like boroughs of Los Angeles, even if they have their own mayors and their own school systems. So, we decided to work with the whole LA county data. The zip codes and areas we have then in the data base will be more diverse and this is actually the goal here to determine clusters of areas based on their similarities and differences.

At this stage, we are considering 284 zip codes in Los Angeles county, CA and the client is comfortable with the global area and with the numbers before clustering work. The global population within the global area considered is more than 9.8M.

2.4 Work on our second database source

We build our second source of data thanks to Foursquare API for location data. Inputs will be our list of coordinates for our zip codes and outputs will be data structures with information on venues - including categories - close to each point. We first choose one location in our list and send our request to Foursquare, to explore how information is organized in the data structures we receive. Once we clearly see how to leverage such info and how we want to organize it, we can run our requests on all our points and create our second data frame. This will be the data base we

will use to choose and prepare our features - this will include some normalization tasks - and then process our clustering work with.

IN order to explore one first area, we choose our central point in our map (that is zip code = 90048). And we request the top 100 venues around our central point in a radius of 500 meters. To do that, we first create the url with our credentials and other parameters, and then we send a request. We get a json file as a result to our request.

It is no surprise we get a large number of venues in this particular area. We can see at the very end of the json file the total number is 94. Now, let's work on how to parse our json files to extract the categories of venues, as our classification process will use such information to find similarities between areas.

Now let's explore for all the areas. The function below will append all the useful results for all the areas we want to explore. We add a parameter min_venues to check if any requests may return a small (maybe suspicious?) number of venues. If that is the case for one or few areas, we can double check the area on the map to see if this number makes sense or if there is something more to explore about that. The idea is to make sure the data we collected from the database online, and from foursquare API is reliable.

Earlier in the study, and in a previous version of our notebooks, we also used the same parameter to filter the data we collect from Foursquare. In particular, we would not keep and append venues from a zip code that would gather - according to Foursquare - less than min_venues in total, in the radius chosen. In this version of the study, we don't filter because we want to let the algorithm measure how different such areas are, and as a consequence, do the actual work of clustering.

In our newly created dataframe LA_venues, we have 4276 venues identified, spread among 257 zip codes in Los Angeles county. We can note that 27 zip codes were removed by the process of data extraction from the json file received from Foursquare. The way the function extracts and appends venues and categories generates no line for a zip code that would not get venues from the Foursquare request.

It is important to see how many unique categories we have. We have 364 unique categories.

2.4.1 Each area in our data base #1 is being analysed

The process is the following:

. we apply one hot encoding to the data frame we created and gathers all the venues from each area (and their category)

. we get rid of the actual lat,lon coordinates of the venues as we don't exploit it in the classification algorithm. In term of lat,lon coordinates, we stay at the level of the zipcode. So, we make sure we add the zipcode lat,lon info to the one hot dataframe (we place the columns on the left).

. Next, we group rows by postal code and by taking the mean of the frequency of occurrence of each category

. Once this is done, we can see (and print) the 5 most common categories of venues in each area

. Then we can put that information for all areas into a dataframe. Using a function to sort the venues categories in descending order, then we create the new dataframe and display the top 10 venues (categories) for each zip code we consider.

2.4.2 Clustering work

A this stage, we have our database ready to apply clustering methods. We run k-means to cluster the areas in 5 clusters

Based on the total number of areas we have (257), we consider 5 clusters is a good number for classification purposes. As a reminder, our objective is to identify the best cluster to select and use as a guidance and as a list of zip codes to target for our real estate consultant to start his prospecting work. So, via reviewing the clustering results, we will estimate the similarity inside each cluster and unsimilarities between two clusters, and we will give an opinion based on our global knowledge of the region. It will be a key phase when we work with the client and have a shared understanding of the results. In case of more clarity needed, or if we need to get our results more solid, we can run k-means for other values of k and explore further.

For the 257 zipcodes we consider, our clustering work for k=5 did not work well. One cluster gathers most of the zip codes (228) and the other clusters respectively contains only 11, 8, 8 and 2 zip codes.

2.4.3 Visualization of clusters we generated on our map

This is the usual next step once we get our clustering work done. But here with the result we got and the split of point between clusters, we know this is not satisfying as we have almost 90% of the points in one particular cluster, so we don't even need to visualize at this stage.

In other words, in such conditions, we cannot determine the discriminating venue categories that distinguish each cluster.

With k=10, we have a similar issue. One cluster is too large and gathers most of the points (again almost 90% of all the points). This leaves no visible way to discriminate among more frequent categories of venues. We run our kmeans algorithm with many values of k (from k=2 to 20 !) and we see the same behavior on how clustering is done. ### We had to improve our method...

Si it is clear now that changing the number of k clusters does not solve the problem. We tried many values of k with no change in the clustering results we cannot exploit.

A new axis of research for our solution is needed. Let's have a look at the features (the categories). The idea is that we got all these categories directly from Foursquare and we did not explore deep enough what we received. It looks like we noeed to group some of the categories together. This is how we discovered that:

All the categories of venues that we received from Foursquare request have been listed in a database. There are more than 300 of them and we can easily see two things: a) there is no consistency in the naming as many categories fields have been filled by the Foursquare end users with much freedom on the strings actually chosen. b) some categories should be grouped in meta categories to stay at the right level of details on how to group the venues. We have information on 4276 venues total, and 365 categories were generated, knowing one venue has only one category. As an example, "Frozen Yogurt Shop" category counts 20 venues, and "Ice Cream shop" counts 49 other venues. It seems relevant to have these two categories of venues grouped into one new unique category. Also, there are many restaurants in our venues, as we can imagine. Many styles of restaurants generated many different categories. Among our 365 categories, we find 73 categories of restaurants! As an example, one Australian Restaurant venue generated one category of venue, just for one unique venue.

So, we decided to work on this list of categories also because they are actually the features used to run kmeans once they get normalized. The objective is to build a table of correspondence

between current categories and new categories we will use - or we could call meta categories. The objective is to reach a number of categories under 40, as we have around 4000 venues. It is also possible to remove some categories with very small number of venues each time it makes sense. As an example, a category "Intersection" counts 8 "venues". We will remove them, and as a consequence, the "Intersection" category as well. Another example, one category "Road" counts only one venue. Before we run the process that will clean and create a new list of (venue, new category) we estimate around 20 categories should be removed.

We could generate manually and with low effort a table just by extracting and reviewing the 360 categories from our dataframe. The number of new categories (or meta categories) is 38.

Our data preparation for clustering work is done with our updated database.

2.4.4 New Clustering work

A this stage, we have our new database ready to apply clustering methods. We have meta categories for each venue instead of gross categories as received from Foursquare. We run k-means to cluster the areas in 5 clusters, and we will try with more values of k too.

As a reminder, our objective is to identify the best cluster to select and use as a guidance and as a list of zip codes to target for our real estate consultant to start his prospecting work.

After we run our kmeans algo for many values of k (from k=2 to k=20), we see that the more discriminating result is with k=7. Number of points per cluster for k=7 is {0: 7, 1: 26, 2: 10, 3: 83, 4: 4, 5: 42, 6: 85}

We'll explore the results in the Results section of this report.

Below is the results we get as a list of points per cluster for many values of k.

```
In [17]: Image("NbOfPointsPerClusters-image.png")
```

Out[17]:

```
Nb of points per cluster for 2 clusters:
{0: 171, 1: 86}
Nb of points per cluster for 3 clusters:
{0: 73, 1: 158, 2: 26}
Nb of points per cluster for 4 clusters:
{0: 68, 1: 26, 2: 11, 3: 152}
Nb of points per cluster for 5 clusters:
{0: 11, 1: 139, 2: 26, 3: 5, 4: 76}
Nb of points per cluster for 6 clusters:
{0: 75, 1: 138, 2: 11, 3: 22, 4: 7, 5: 4}
Nb of points per cluster for 7 clusters:
{0: 7, 1: 26, 2: 10, 3: 83, 4: 4, 5: 42, 6: 85}
Nb of points per cluster for 8 clusters:
{0: 12, 1: 12, 2: 88, 3: 11, 4: 5, 5: 32, 6: 7, 7: 90}
Nb of points per cluster for 9 clusters:
{0: 4, 1: 14, 2: 75, 3: 11, 4: 10, 5: 82, 6: 34, 7: 5, 8: 22}
Nb of points per cluster for 10 clusters:
{0: 3, 1: 10, 2: 77, 3: 10, 4: 4, 5: 18, 6: 7, 7: 89, 8: 37, 9: 2}
Nb of points per cluster for 11 clusters:
{0: 14, 1: 5, 2: 73, 3: 7, 4: 11, 5: 12, 6: 83, 7: 3, 8: 23, 9: 2, 10: 24}
Nb of points per cluster for 12 clusters:
{0: 33, 1: 86, 2: 14, 3: 77, 4: 5, 5: 12, 6: 10, 7: 4, 8: 2, 9: 7, 10: 5, 11: 2}
Nb of points per cluster for 13 clusters:
{0: 1, 1: 75, 2: 20, 3: 4, 4: 10, 5: 5, 6: 14, 7: 81, 8: 10, 9: 2, 10: 2, 11: 32, 12: 1}
Nb of points per cluster for 14 clusters:
{0: 73, 1: 23, 2: 6, 3: 23, 4: 12, 5: 3, 6: 4, 7: 83, 8: 1, 9: 2, 10: 14, 11: 5, 12: 7, 13: 1}
Nb of points per cluster for 15 clusters:
{0: 81, 1: 7, 2: 10, 3: 23, 4: 5, 5: 73, 6: 12, 7: 3, 8: 13, 9: 13, 10: 11, 11: 1, 12: 1, 13: 2, 14: 2}
Nb of points per cluster for 16 clusters:
```

At this stage we realized how the number of clusters was crucial in the way our kmeans algorithm could efficiently do the classifying tasks needed.

We've also seen how we had to work on the features we extracted got with the Foursquare API to make more relevant features to solve our problem.

It is both thanks to working on many values of k for kmeans and thanks to features modifications that we could reach the results we present in the next section of our report.

2.5 Results

Most frequent venues in 'cluster 0' areas are businesses, services, business hotels and transportation.

Most frequent venues in 'cluster 1' areas are sport stores oriented and venues, as well as transportation.

Most frequent venues in 'cluster 2' areas are parks and transportation.

Most frequent venues in 'cluster 3' areas are restaurants and bars. This is a good candidate for our priority cluster of areas, with 83 zip codes. We want our result to find specific but we also want our list to be broad enough for our client's broker prospecting work.

Most frequent venues in 'cluster 4' areas are cultural venues, like museum and music venues.

Most frequent venues in 'cluster 5' areas are parks and some restaurants and markets.

In most frequent venues in 'cluster 6' areas, we find a well-proportioned list of restaurants, food stores, convenient stores and cosmetics stores that fits well with the target potential customers of BHSE. This is a perfect candidate for our priority cluster of areas, with 85 zip codes.

This is an extract of the table used to examine our clustering results.

```
In [20]: Image("ChosenClusterTableImage.png")
```

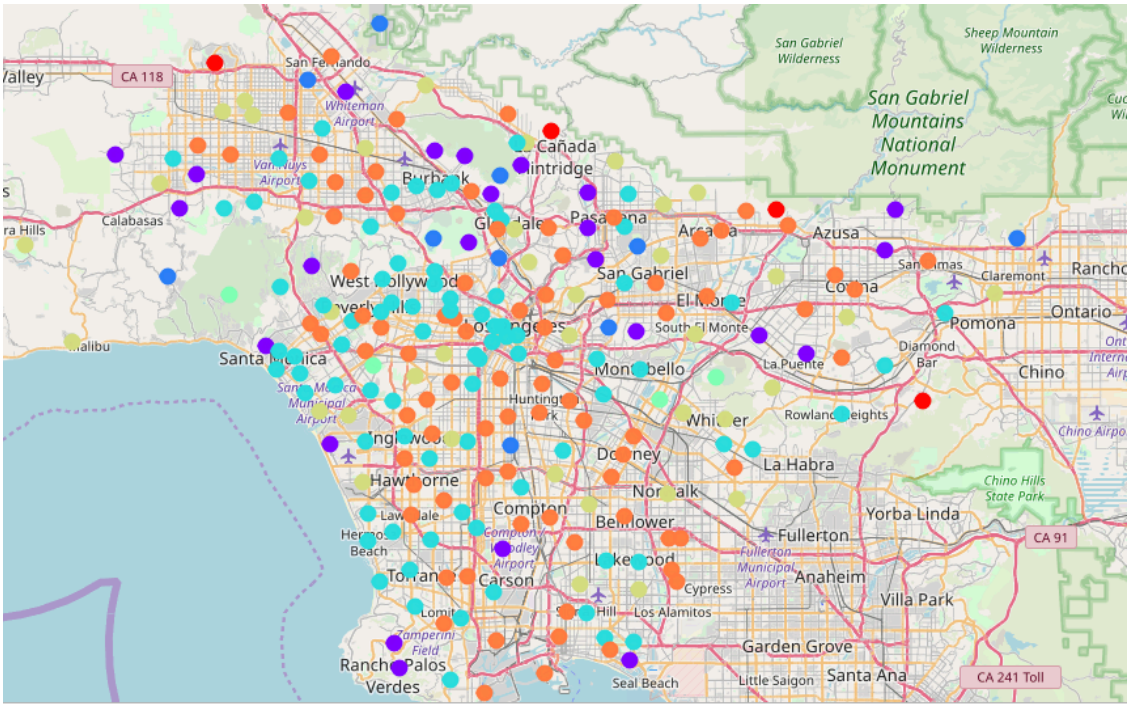
Out[20]:

	zip	population	density	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
135	90073	539	613.7	6	Bank	Restaurant	Transportation	Cosmetics Store	Food Store	Flowers Store
1187	90012	31103	3711.6	6	Restaurant	CafeBar	Food Store	Cultural	Stores	Club
5899	90802	39347	2343.1	6	Sport	Restaurant	Transportation	Convenience Store	Food Store	Flowers Store
8266	90260	34924	5031.3	6	Restaurant	Transportation	Health	Cosmetics Store	Food Store	Sport
9902	91306	45061	4204.2	6	Restaurant	Convenience Store	Food Store	Ice Cream	CafeBar	Cosmetics Store
10196	90029	38617	10949.3	6	Restaurant	Food Store	Convenience Store	CafeBar	Health	Cultural
10809	90001	57110	6295.9	6	Restaurant	Food Store	Liquor Store	Bakery	Health	Shoe Store
10811	90034	57964	7194.2	6	Restaurant	Food Store	Sport	Health	CafeBar	Cosmetics Store

2.5.1 This is an image of a map of our clustering work.

```
In [15]: Image("map-chosen-cluster.png")
```

Out[15]:



3 Conclusion

Determining the right number of clusters was crucial for our kmeans algorithm to efficiently do the classifying tasks.

Working on the features we extracted from the Foursquare API and making more relevant features was key to solve our problem.

It is both thanks to working on many values of k for kmeans and thanks to features modifications that we could reach the results we presented.

Our client has accepted the result and is using it to provide guidance and a list of target areas to his real estate consultant to find relevant locations for his next stores in Los Angeles.