Clustering Los Angeles neighborhoods

IBM Data Science Professional Certificate
Coursera Course #9 - Capstone Project

Week 5 assignment

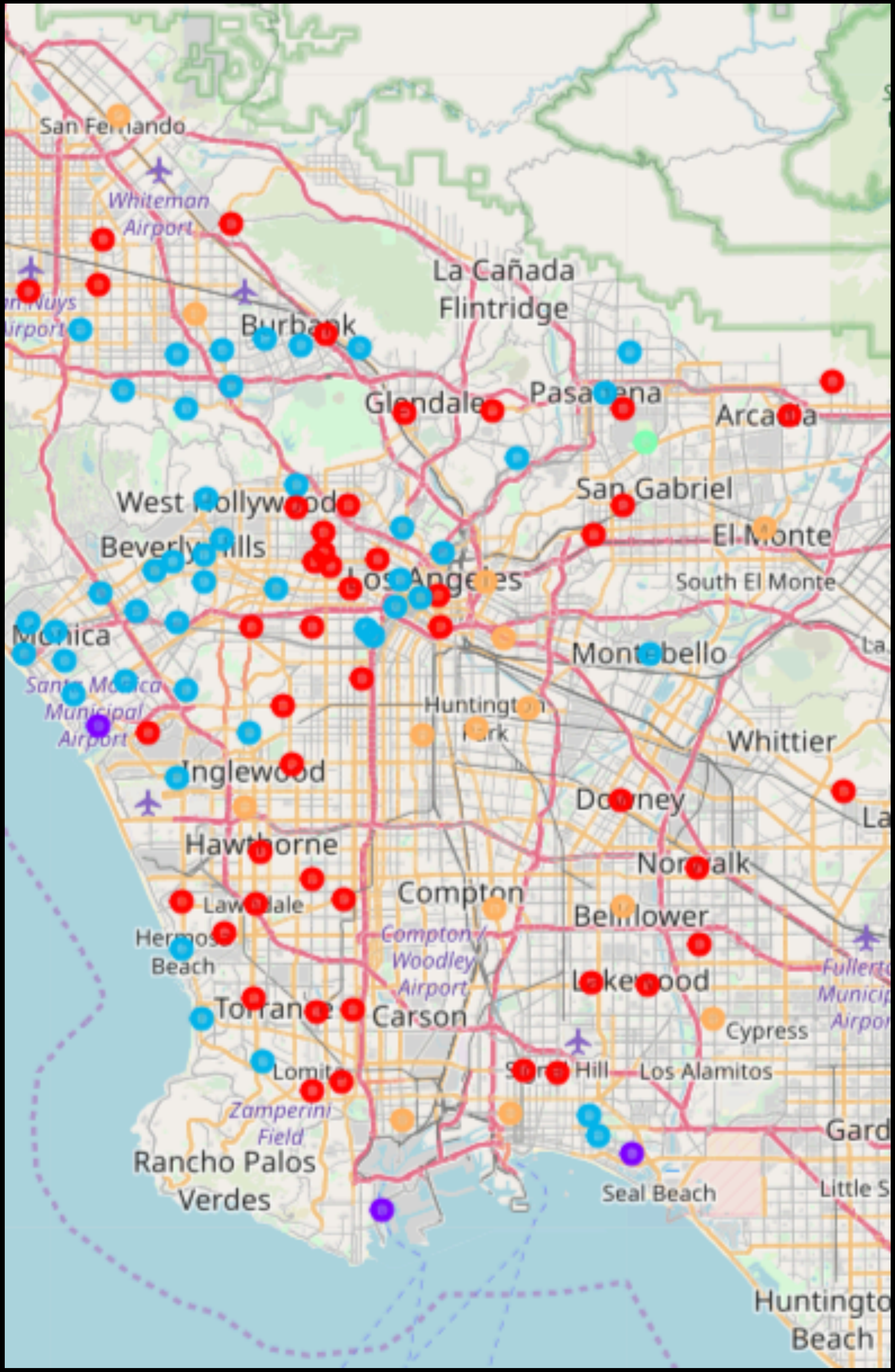
Gerald NACACHE
geraldnacache@gmail.com
Feb 2019

Report Presentation

This report is a part of the project capstone deliveries.

It completes a project notebook and a full report.

Full report is accessible here.
Note book is accessible here.



Report Presentation

- Introduction
- Data
- Methodology
- Results
- Conclusion

Introduction

- A European company plans to deploy her hair dresser brand in the US and starts with 20 new stores in Los Angeles, CA.
- In order to provide guidance to her preferred real estate consultant and target areas for prospecting new store locations, the company asked us to study LA neighborhoods and recommend a priority list of areas.
- The client is interested our expertise in solving problems via classifying techniques. We decide to launch a study of LA neighborhoods via exploring existing venues for each area we will consider and review clusters of neighborhoods generated from unsupervised learning on large data sets.
- Description of the data used to solve the problem as well as methodology are described in this presentation and detailed in our report.
- Results and conclusions will then end this presentation.

Data

- Source #1 : database with all zip codes in California (from https://simplemaps.com/data/us-zips), latitude, longitude and more.

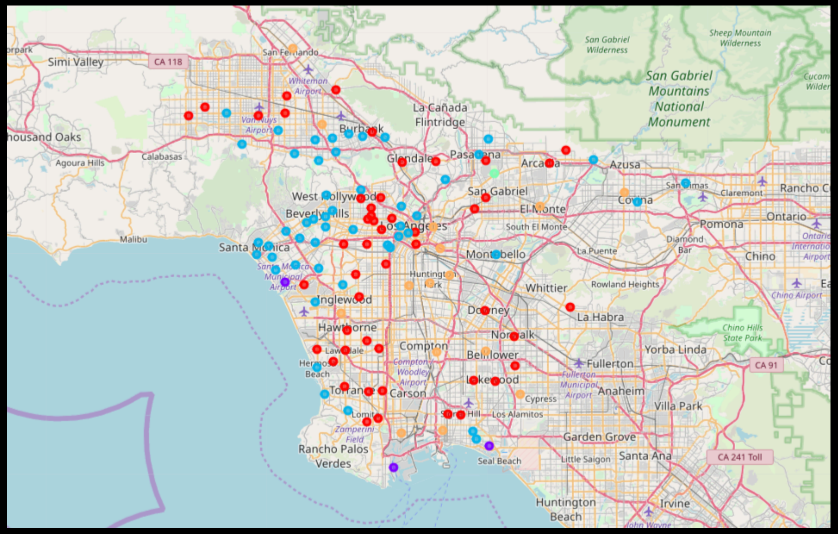
	zip	lat	lng	city	state_id	state_name	population	density	county_name
23	92280	34.1256	-114.7779	Vidal	CA	California	14	0.0	San Bernardino
25	92304	34.5415	-115.6445	Amboy	CA	California	17	0.2	San Bernardino
27	92332	34.9127	-115.3417	Essex	CA	California	65	0.0	
30	93519	35.2975	-117.9294	Cantil	CA	California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


```
{'meta': {'code': 200, 'requestId': '5c61054f1ed2192b556fb895'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': '$-$$$$', 'key': 'price'},
    {'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Mid-City West',
  'headerFullLocation': 'Mid-City West, Los Angeles',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 95,
  'suggestedBounds': {'ne': {'lat': 34.0775000045, 'lng': -118.36717748404739},
   'sw': {'lat': 34.0684999955, 'lng': -118.37802251595262}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '49e3aa19f964a520aa621fe3',
       'name': "Joan's on Third",
       'location': {'address': '8350 W 3rd St',
        'crossStreet': 'at Kings Rd',
        'lat': 34.072820607435126,
        'lng': -118.37184744336331,
        'labeledLatLngs': [{'label': 'display',
          'lat': 34.072820607435126,
          'lng': -118.37184744336331}],
```


- Source #2 : we collect data thanks to Foursquare API for location data. Inputs are zipcode(lon,lat) and outputs are json files with all venues and categories in a radius.

- Areas and clusters are visualized on maps.

- Numpy
- Pandas
- Matplotlib
- Scikit-Learn
- Scipy




```
[9]: LA_zipc = LA_zipc.drop(LA_zipc.index[LA_zipc['state_id'] != 'CA'])
     LA_zipc.head()
```

[9]:

	zip	lat	lng	city	state_id	state_name	zcta	parent_zcta	population	density	county_fips	county_name
23	92280	34.1256	-114.7779	Vidal	CA	California	True	NaN	14	0.0	6071	San Bernardino
25	92304	34.5415	-115.6445	Amboy	CA	California	True	NaN	17	0.2	6071	San Bernardino

- All data preparation, normalization tasks as well as machine learning for clustering are coded in python and specific libraries. All code is visible in Jupiter notebooks.

Methodology

- A zipcode is a vector and categories of venues in the area generate our features.
- Frequency of categories in the area are measured, normalized and become inputs for kmeans algorithm.
- Review clustering results with number of clusters from 3 to 18.
- Challenge features filtering and numbers, as well as data collection methods.
- Cluster shapes and differences; maps visualization

Methodology

- Because data collection with foursquare generates too many categories of venues with some consistency issues, we had to work on grouping categories to reduce the number of features, so that normalization and clustering can work correctly.

    - *Number of categories of venues reduced from >300 to 38.*

- Then clustering with k-means and k=5,6,7,8 became relevant.

    - *First results shew a similar trend with many values of k, that was creating one cluster gathering more than 80% of the points, and much smaller clusters with only few points.*

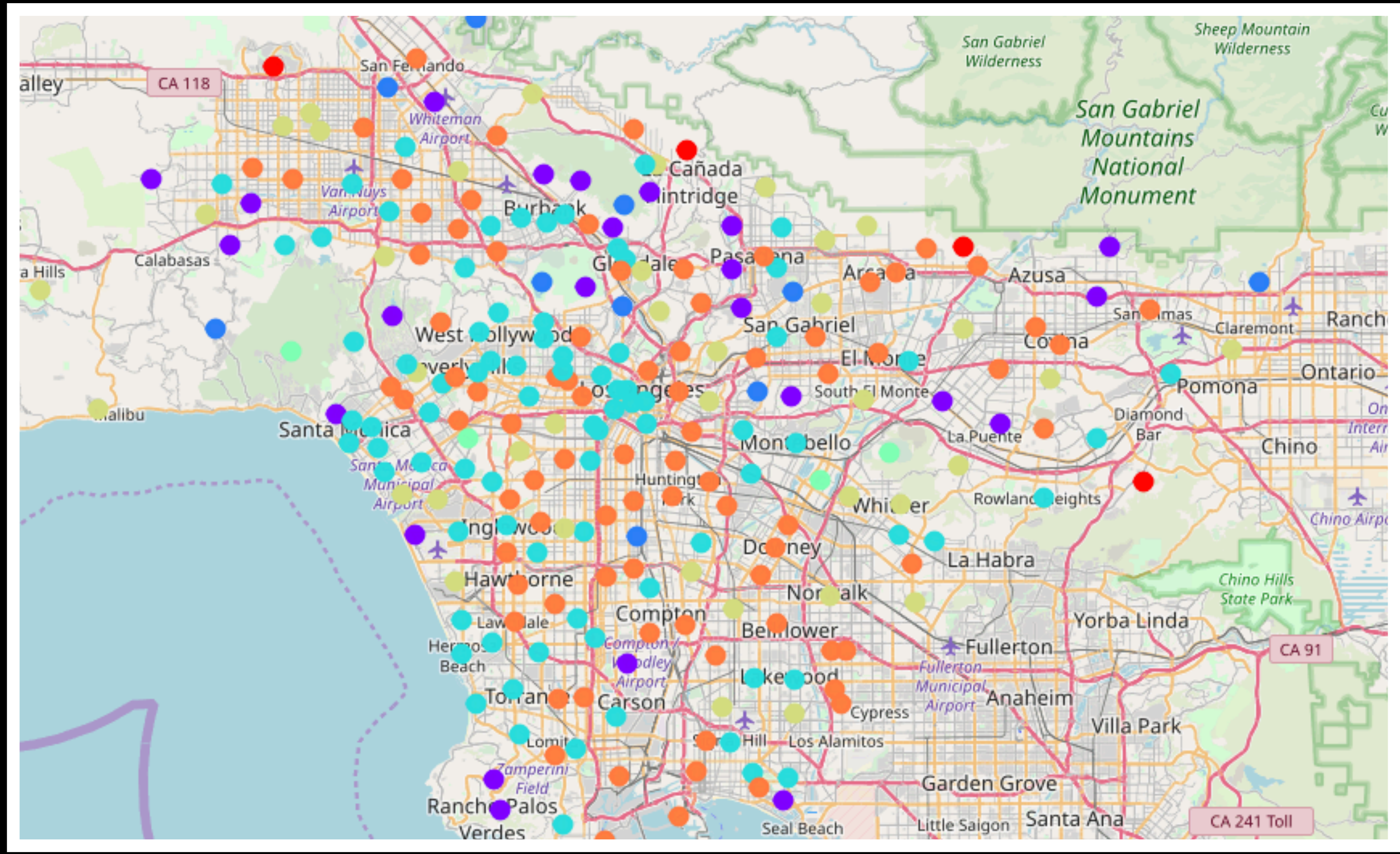
- Results with reduced features and k=7 are presented in th next slide

Results

```
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}
```

- *With k=7, we choose a cluster of 85 zipcodes with well-proportioned list of restaurants, food stores, convenient stores and cosmetics stores that will fit well with the target customers of our client.*

	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



Clustering Los Angeles neighborhoods

Selection for our client of a cluster of 85 zip codes among 257 initially.
Value generated is
a) huge time saving during prospecting tasks for new store locations
b) qualitative selection of areas matching with customers and neighborhoods targets

Conclusion

- Determining the right number of clusters was crucial for kmeans algorithm efficiency

- Exploring deeper and working on the features extracted from foursquare api and creating more relevant features was a key step to solve the problem

- Our client has accepted the results and leverages them to provide guidance and clear road maps for his real estate consultant to find relevant locations for their next stores in Los Angeles, CA.