Clustering cities for similar liveability

HUNG DINH

FEBRUARY, 2020

1. Business problem

- Human’s need to relocate:
  - Individual move for career, family, leisure
  - Business plan to open new branch
  - Administrative management for city expansion/merge
- What indexes to know about the change:
  - Crime rate
  - Nature
  - Culture
  - Health
  - Education
  - Infrastructure
- Project ojective: group cities based on liveable indexes

2. Data

- City list:
  - Based on economy impact
  - https://en.wikipedia.org/wiki/Globalization_and_World_Cities_Research_Network
- Coordinates
  - https://geodatos.net
- Crime index
  - https://numbeo.com
- Other indexes:
  - Based on number of venues in each criteria
  - https://foursquare.com

3. Methodology

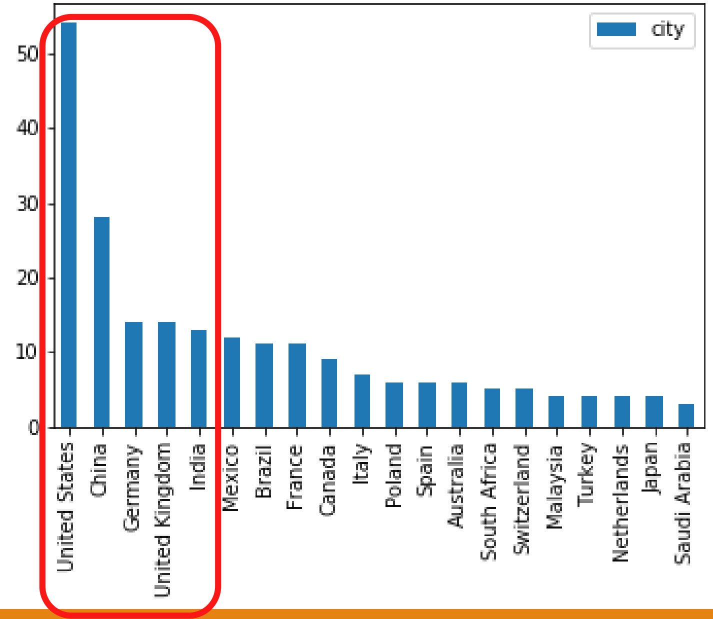
- Acquire data:
  - City list
  - Acquire their coordinates (for later Foursquare queries)
  - Acquire crime rate
  - Acquire Foursquare categories and put them into 5 index criteria groups
  - Acquire popular venues for each city
- Count venues in each of the 5 groups
- Run k-means clustering on final data
- Analyze cluster results
- Give recommendation

3.1. City list

- Packages used: request, beautifulsoup

- Total entries after nan removal: 365

- Most city countries:
  - US
  - China
  - G8 countries



3.2 – 3.3 Crime data and coordinates

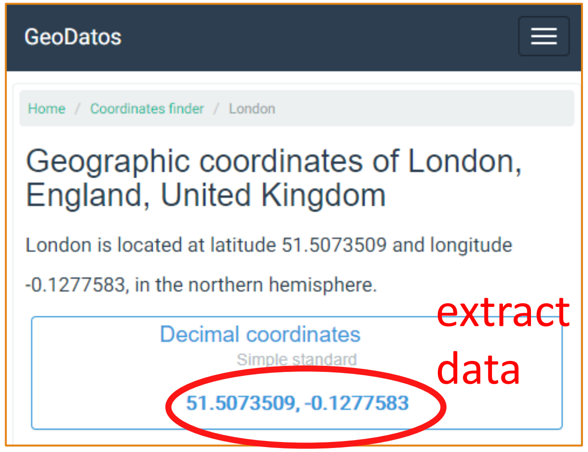
- Packages used: read_html (pandas)
- Total entries after nan removal: 358
- Most dangerous country: China
- Safest countries: South America, South Africa

Highest crime rate

	city	safe	nation
129	Hefei	88.97	China
91	Doha	88.52	Qatar
2	Abu Dhabi	88.51	United Arab Emirates
219	Nagoya	87.82	Japan
222	Nanjing	86.68	China
316	Taipei	85.89	Taiwan
236	Ningbo	85.25	China
267	Quebec City	84.95	Canada
143	Jinan	84.07	China
312	Suzhou	83.38	China

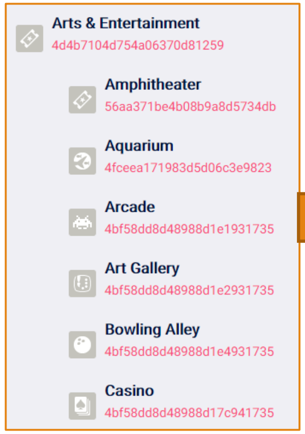
Lowest crime rate

	city	safe	nation
275	Rio De Janeiro	22.68	Brazil
93	Douala	22.52	Cameroon
229	Natal	21.13	Brazil
290	San Pedro Sula	19.30	Honduras
144	Johannesburg	19.25	South Africa
97	Durban	19.08	South Africa
262	Pretoria	18.31	South Africa
336	Valencia-Venezuela	15.61	Venezuela
59	Caracas	15.03	Venezuela
213	Mosul	1.96	Iraq



3.4. Foursquare category grouping

- Packages used: request from foursquare api
- Total entries: 941
- Health:
  - Medical Center (6.22)
- Culture:
  - Art and Entertainment (0)
  - Nightlife Spot (4)
- Environment:
  - Outdoors & Recreation (5) except Athletics & Sports (5.0) and States & Municipalities (5.52)
- Education: school and the like in Foursquare categories
  - College & University (1)
  - School (6.34)
  - Daycare (8.25)
- Infrastructure: focus on public transportation (bus stop, train station), commercial offices and the like in Foursquare categories
  - Travel & Transport (9) except Hotel (9.13)



	categories
0	cat.0_Arts & Entertainment
1	cat.0.0_Amphitheater
2	cat.0.1_Aquarium
3	cat.0.2_Arcade
4	cat.0.3_Art Gallery
5	cat.0.4_Bowling Alley
6	cat.0.5_Casino

3.5-3.6 Foursquare venues and group

- Packages used: request from foursquare api explore call
- Total entries for each city: 100
- Query radius: 5 km
- Final data entries: 358 cities

	lat	lon	safe	catCulture	catEducation	catHealth	catEnvironment	catInfrastructure
count	358.000000	358.000000	358.000000	358.000000	358.000000	358.000000	358.000000	358.000000
mean	28.120047	3.577499	55.840503	24.357542	2.958101	0.030726	8.905028	5.047486
std	24.224879	75.086607	16.836823	10.390180	2.374215	0.188329	5.737188	3.259201
min	-43.532054	-157.858333	1.960000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	19.128721	-73.343785	44.395000	17.000000	1.000000	0.000000	4.000000	3.000000
50%	35.206326	8.611910	56.940000	26.000000	3.000000	0.000000	9.000000	5.000000
75%	44.942762	47.299753	69.765000	31.000000	4.000000	0.000000	12.000000	7.000000
max	64.146582	174.776236	88.970000	49.000000	13.000000	2.000000	28.000000	15.000000

low value, potentially not contribute to classification

3.7. k-means clustering

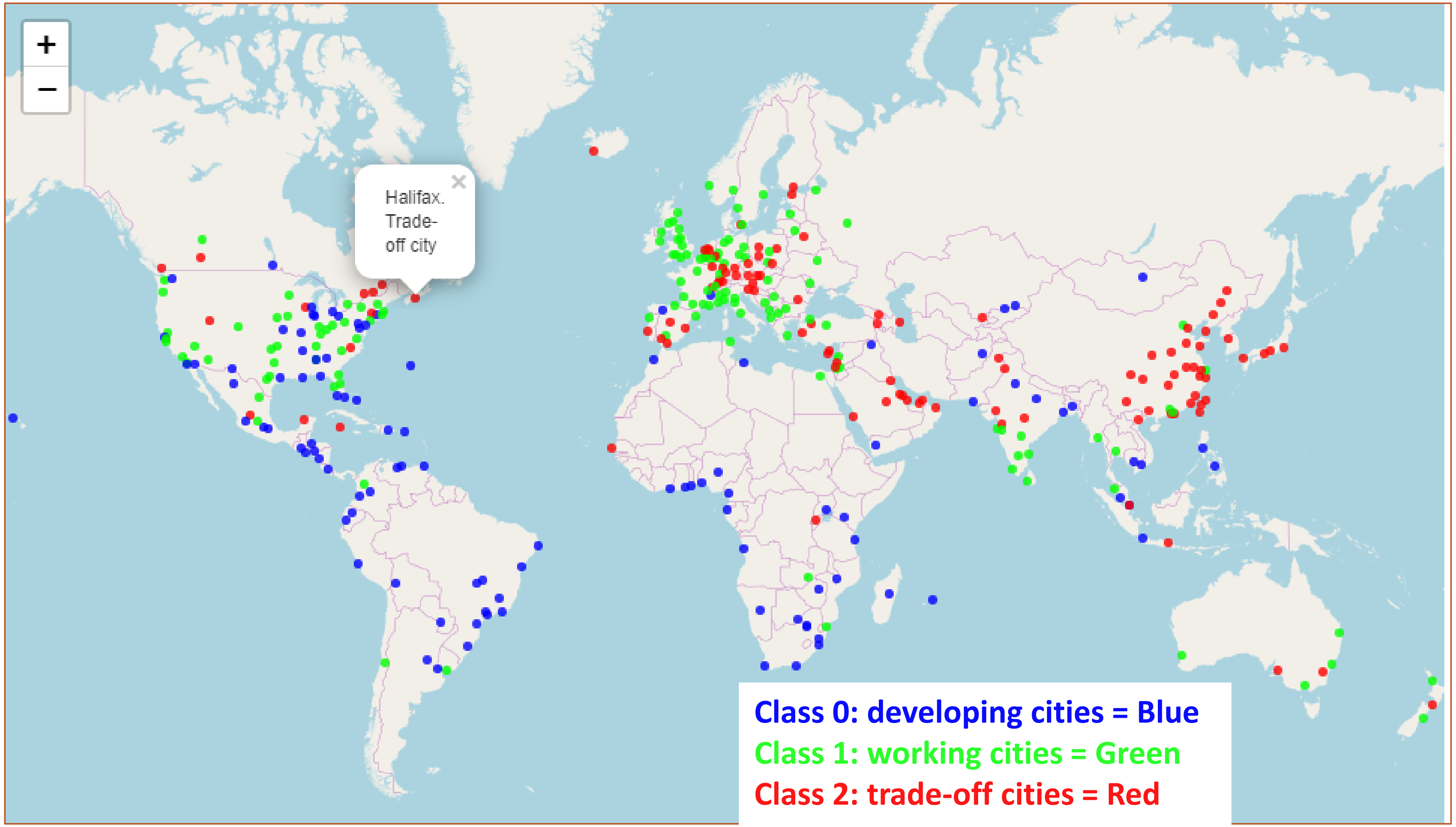
- unsupervised learning to discover in-depth data distribution
- k-means: runs quickly, results easy to analyze
- k=3 for low-mid-high numerical range
- data normalization is not required because crime data is in 100 scale, other 5 indexes are counted from 100 venues

4. Results

- Class 0 - developing cities: lowest crime rate and other living indexes except education (for development).
- Class 1 - working cities: best for democratic indexes but medium crime rate - best for career and financial dynamic life (high crime rate is quite common for such cities).
- Class 2 - trade-off cities: average in living indicators but high crime index.

Cluster Labels	lat	lon	safe	catCulture	catEducation	catHealth	catEnvironment	catInfrastructure
0	11.163473	-26.307907	36.107850	18.177570	2.943925	0.065421	5.822430	3.644860
1	35.763806	-6.082107	56.007226	31.080292	3.510949	0.029197	11.540146	6.072993
2	34.849506	43.236310	74.161140	22.078947	2.307018	0.000000	8.631579	5.131579

4. Results



Halifax. Trade-off city
Class 0: developing cities = Blue
Class 1: working cities = Green
Class 2: trade-off cities = Red

5. Discussion

- US mainly has developing or full work cities, according to its strong economic system
- China with the reputation of pollution, low living conditions and closed politics has highest crime rate and ranks 1st in the trade-off cities
- Most UK and EU cities are good for work, except Central European cities
- South America and South Africa are best for development

Class 0: developing cities

	city
nation	
United States	19
Brazil	11
Mexico	8
South Africa	5
India	3

Class 1: working cities

	city
nation	
United States	31
United Kingdom	14
France	10
Germany	9
Italy	6

Class 2: trade-off cities

	city
nation	
China	22
Canada	6
Germany	5
Poland	4
United States	4

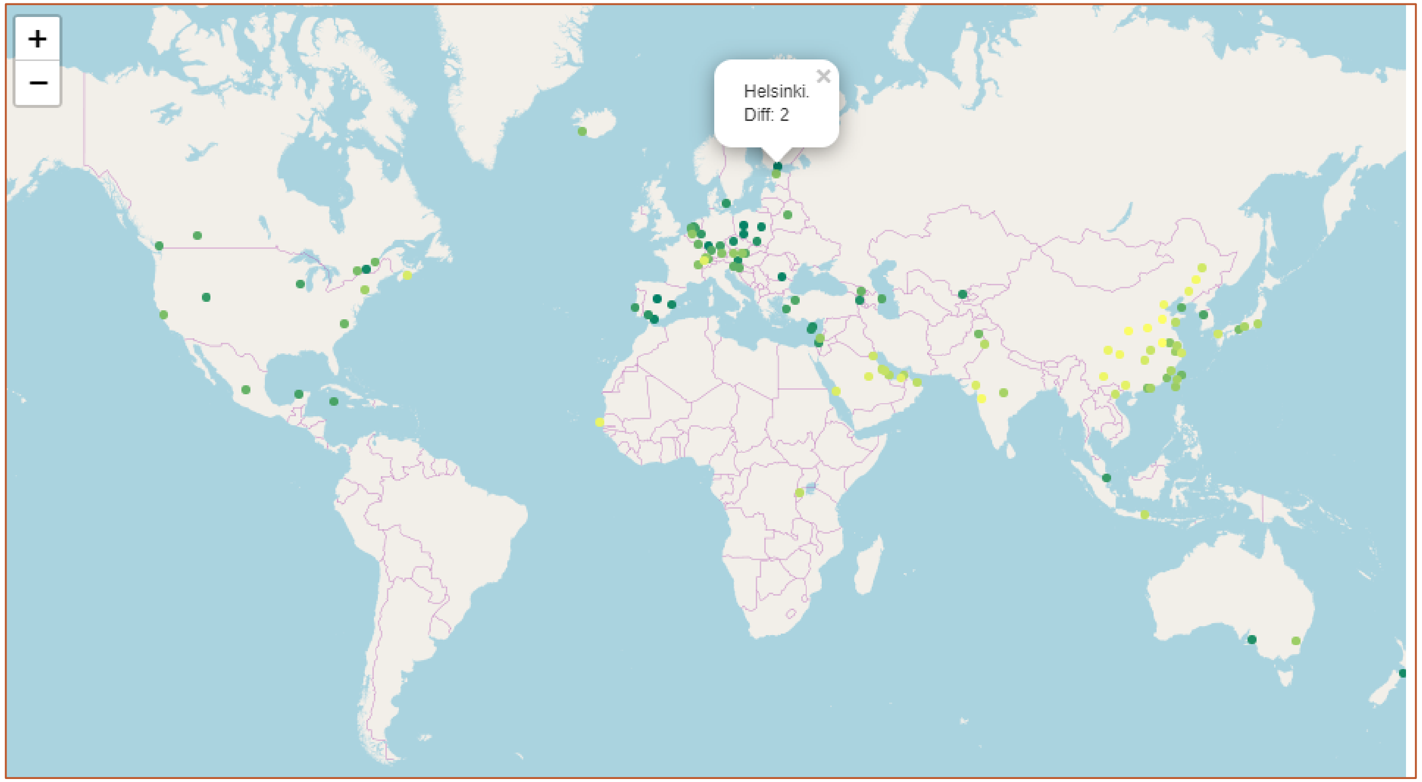
6. Application

- Suggest similar cities to one from user input

```
ctest = input("Your reference city: ")
if ctest.lower() in set(g.city.str.lower()):
    idex = g.index[g.city.str.lower() == ctest.lower()].tolist()
    clustervalue = g.iloc[idex]["Cluster Labels"].tolist()
else:
    print('Your city is not in the list. Try another.')

Your reference city: madrid
```

city	lat	lon	diff	stt
Madrid	40.416775	-3.703790	0.000000	1
Helsinki	60.169856	24.938379	0.407502	2
Poznan	52.406374	16.925168	0.444031	3
Mannheim	49.487459	8.466040	0.495382	4
Bucharest	44.426767	26.102538	0.498187	5
Wroclaw	51.107885	17.038538	0.570776	6



7. Conclusions

- The project has grouped most 358 popular cities in the world.
- The similarity between these groups are measured based on the 6 main living indexes.
- Three main city groups are determined associating to to work, to develop and to beware of a possible trade-off with crime.
- Closer examination into each group and their international relationship also reveals interesting findings between countries. Possible explanations are provided to support them.
- An applied example is shown to demonstrate how one can use the result to get meaningful information.

8. Acknowledgements

- I acknowledge Anton Biryukov for inspiring me to pursue the Machine learning and Data science path since October 2017.
- Special thanks to Si Le who have been always supported me and, in this project, graciously assisted me playing around with Foursquare category data.
- Final thanks to IBM Data Science Professional Certificate lecturers and classmates who have helped me follow this intensive specialization