Clustering the biggest tech hub cities

Evgueni Stoilkov

May 11, 2021

1. Introduction/Business Problem

This project will try to compare a group of some of the best known world cities that are considered to be the most promising tech hubs of the future. The choice of category of cities is not an essential part of the methodology, but this may help in identifying stronger relations when working with a relatively homogenous group. Experiments with other categories will indeed help find the best approach.

The comparison will run across several completely different datasets of features for all of the cities chosen and will try to investigate whether the different feature sets may lead to similar cities clusters. K-means will be used to analyze data.

This approach may be of interest for a number of reasons. On the one side we may discover that somehow a group of features correlates with a completely different other group of features if final city clusters overlap. On the other side we may use this approach for feature selection to use later with other machine learning algorithms.

If we discover a strong relation between the two groups of features we may analyze other possible practical applications when comparing cities within a cluster. For example, if we identify typical features within some clusters, business opportunities may become obvious in cities with strong outliers as compared to other cities of the same cluster.

2. Data for the analysis

We will study the following cities: New York, London, Beijing, Boston, Tel Aviv, Los Angeles, Shanghai, Paris, Berlin, Stockholm, Seattle, Toronto-Waterloo, Singapore, Amsterdam, Austin, Chicago, Bangalore, Washington, San Diego, Lausanne, Bern, Geneva, Sydney, Vancouver, Hong Kong, Atlanta, Barcelona, Dublin, Miami, Munich.

The first group of features consists of generally available data on population, economy, development statistics, health, environment. Data was obtained online from sources like NYC Global Partners, Startup Genome, OECD, UBS.

Data used:

- City density
- GDP per capita
- Employment
- Environment
- Transportation

- Education
- Health

The second set of features was taken from Startup Genome. They provide indicators around innovation and the startup development of the cities, ecosystems performance, funding, market reach, connectedness.

The third set of data is taken from Foursquare API consisting of registered venues around central parts of the cities. Coordinates of locations for data collection were manually selected using Google maps.

The choice of data may be discussed at length in a separate paper. The best approach would be to research available datasets and select among the most consistent datasets, probably coming from different sources and covering different aspects of the city environment. It took me a relatively high amount of time to scrape basic data which may not be the most appropriate for the purpose of this analysis, but due to time constrains we prefer to focus here on methodology. Anyway, such data should be much more easily available.

As k-means was the adopted algorithm for most of the analysis numerical data was standardized and categorical features, one-hot encoded.

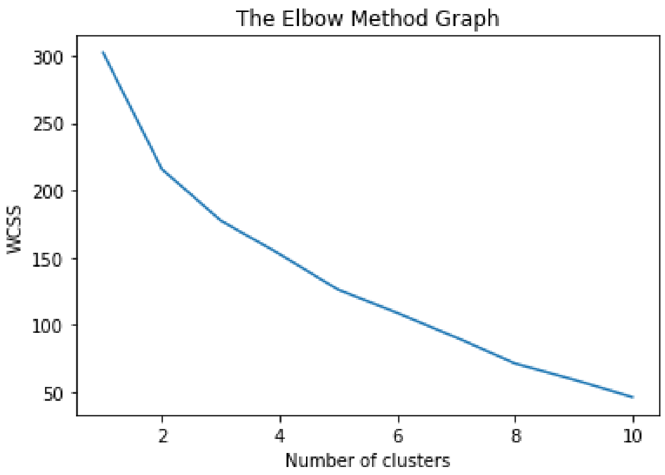
3. Methodology

The approach taken here consisted of identifying clusters of cities based on different datasets or combinations of datasets.

For each dataset chosen, the optimal number of clusters was chosen based on the elbow approach. We use a chart where y is a representation of the intra-centroid distance, which we would like to be minimal. It would naturally decline if we increase the number of clusters in the model. However, we prefer a smaller number of clusters, because at the limit the smallest number is always obtained when clusters are equal to the number of observations.

It is considered that a good selection for n-clusters is obtained at the point where there is an inflection and the curve becomes flatter.

a. **1st set of data chosen:** 'foborn', 'anpopgrowth', 'share500', 'povrate', 'masstransit', 'higher_educ', 'fo_tourists', 'dom_tourists', 'inf_mortality', 'life_exp_m', 'life_exp_f', 'physicians', 'anti_smoking', 'air_quality', 'retrofitted_city_vehic', 'bike_share', 'Wholesale', 'Real Estate & Rental & Leasing', 'Wholesale/Retail Trade', 'Business Services', 'Professional Services', 'Manufacturing', 'Services', 'Financial Operations', 'Finance, Rental & Leasing', 'Research and Advising', 'Trading and Logistics/Finances',   'Finance and Insurance', 'Wholesale & Retail', 'Professional Scientific and Technical Services'



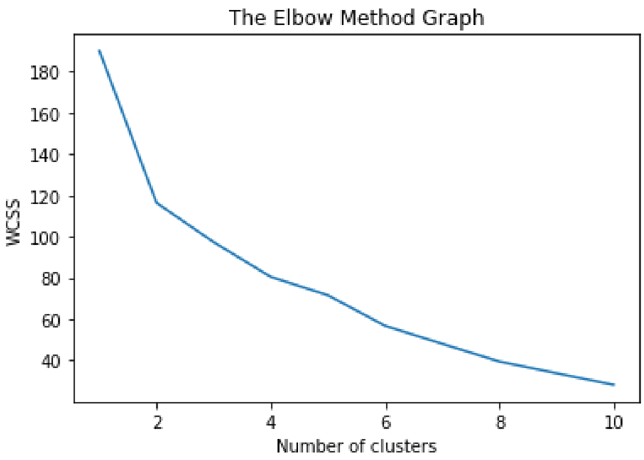
The elbow curve flattens more significantly when the number of clusters is 5, so we take this number as optimal.

The labels obtained are as below:

LABELS	CITY	
0	Dublin	1
1	Barcelona	1
	Berlin	1
	Geneva	1
2	Hong Kong	1
	London	1
	Los Angeles	1
	New York	1
	Paris	1
	Singapore	1
	Stockholm	1
	Sydney	1
	Toronto	1
	Vancouver	1
3	Beijing	1
	Shanghai	1
4	Amsterdam	1
	Boston	1
	Chicago	1

The map later will let us identify patterns more clearly.

2nd set of data chosen: 'ranking', 'change', 'performance', 'funding', 'marketReach', 'connectedness', 'talent', 'experience', 'k0wledge', 'growthIndex'

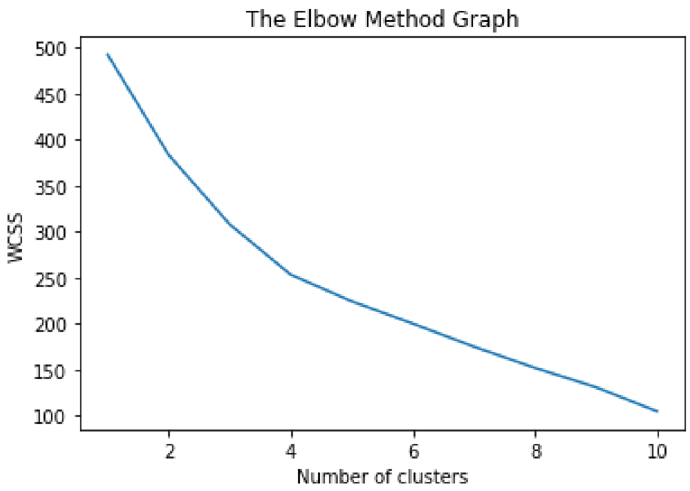


We see that the results are quite different. The curve becomes less steep at a later point with clusters increase and the WCSS is generally lower which is expected given the lower number of features.

We would chose a number of clusters of 4.The labels obtained are as below:

```
labels_2  city
0         Barcelona      1
          Chicago        1
          Dublin         1
1         Beijing        1
          Boston         1
          London         1
          Los Angeles    1
          New York       1
          Shanghai       1
2         Amsterdam      1
          Berlin         1
          Geneva         1
          Paris          1
          Singapore      1
          Stockholm      1
          Toronto        1
3         Hong Kong      1
          Sydney         1
          Vancouver      1
```

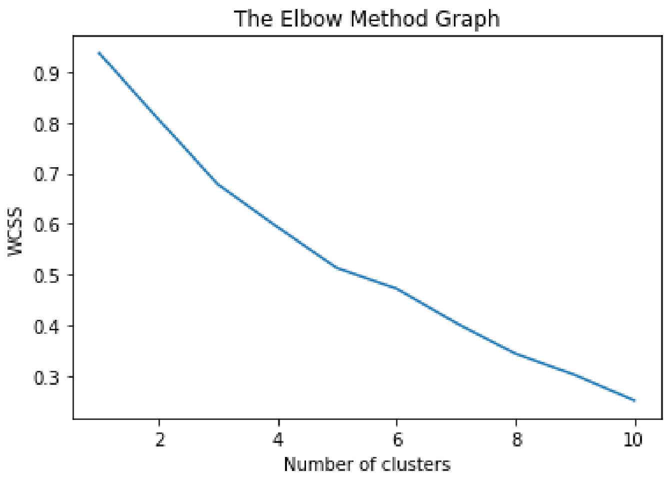
3rd set of data chosen: combine the first two datasets and run k-means



There is some inflection at 4 clusters, so we take that number. The labels obtained are as below:

```
labels_3  city
0         Boston          1
          London          1
          Los Angeles     1
          New York        1
          Paris           1
          Stockholm       1
          Toronto         1
1         Berlin          1
          Hong Kong       1
          Singapore       1
          Vancouver       1
2         Beijing         1
          Shanghai        1
3         Amsterdam       1
          Barcelona       1
          Chicago         1
          Dublin          1
          Geneva          1
          Sydney          1
```

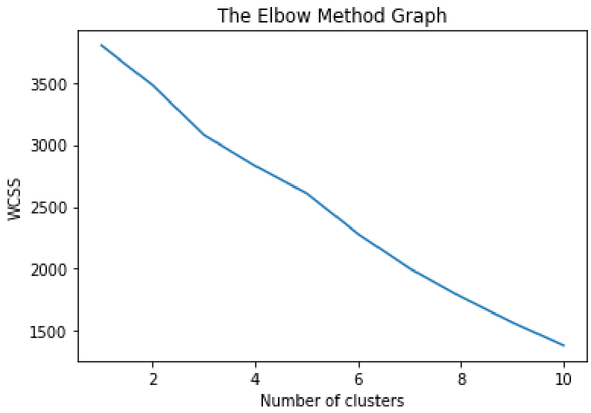
4th set of data chosen: we try k-means with summarized data from Foursquare



There is some inflection at 5 clusters, so we take that number. The labels obtained are as below:

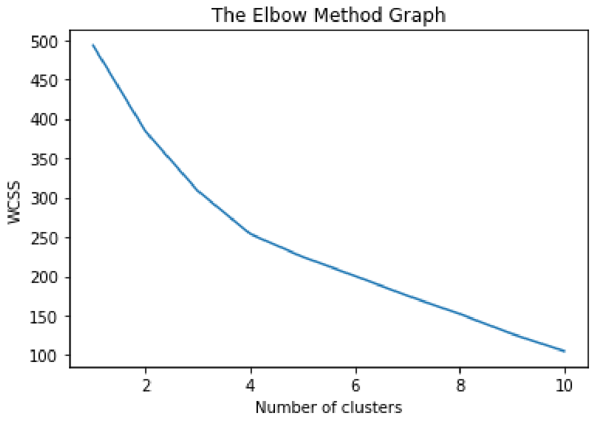
```
labels_4  city
0         Geneva          1
          Sydney          1
1         Barcelona       1
          Berlin          1
          Boston          1
          Chicago         1
          Dublin          1
          London          1
          Los Angeles     1
          New York        1
          Paris           1
          Singapore       1
          Stockholm       1
          Toronto         1
          Vancouver       1
2         Shanghai        1
3         Amsterdam       1
          Beijing         1
4         Hong Kong       1
```

5th set of data chosen: we try using standardized data from Foursquare venues and number of venues



We get a significantly higher number for WCSS so this route is apparently not a good one.

6th set of data chosen: we combine 3rd and 4th datasets



It does not look bad especially with the high number of features, without standardizing, so we will keep it. There is some inflection at 4 clusters, so we take that number.

```
labels_6  city
0         Boston        1
          London        1
          Los Angeles   1
          New York      1
          Paris         1
          Stockholm     1
          Toronto       1
1         Berlin        1
          Hong Kong     1
          Singapore     1
          Vancouver     1
2         Beijing       1
          Shanghai      1
3         Amsterdam     1
          Barcelona     1
          Chicago       1
```

```
DUBLIN         1
GENEVA         1
SYDNEY         1
```

It is not easy to draw significant conclusions from these results. Beijing and Shanghai seem to get often in the same group.

In the end, the fact that we get different groups of clusters may only indicate that these clusters may be good for different purposes.

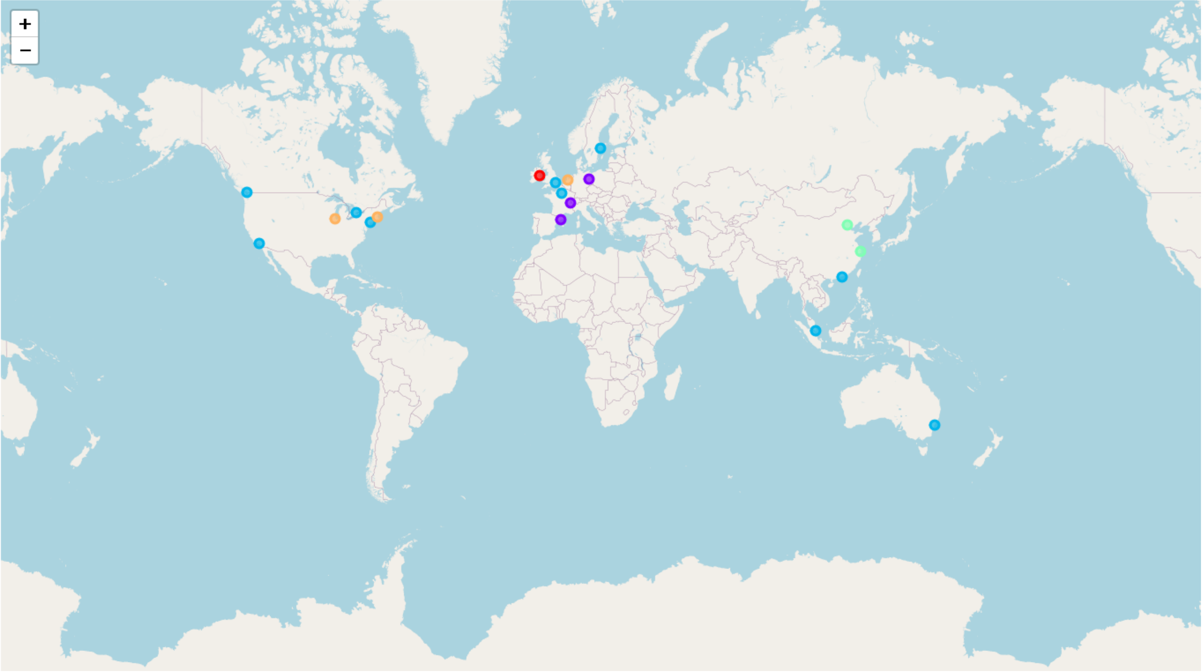
Let`s try a combined table of labels:

	city	try_1	try_2	try_3	try_4	try_6
0	New York	2	1	0	1	0
1	London	2	1	0	1	0
2	Beijing	3	1	2	3	2
3	Boston	4	1	0	1	0
4	Los Angeles	2	1	0	1	0
5	Shanghai	3	1	2	2	2
6	Paris	2	2	0	1	0
7	Berlin	1	2	1	1	1
8	Stockholm	2	2	0	1	0
9	Toronto	2	2	0	1	0
10	Singapore	2	2	1	1	1
11	Amsterdam	4	2	3	3	3
12	Chicago	4	0	3	1	3
13	Geneva	1	2	3	0	3
14	Sydney	2	3	3	0	3
15	Vancouver	2	3	1	1	1
16	Hong Kong	2	3	1	4	1
17	Barcelona	1	0	3	1	3
18	Dublin	0	0	3	1	3

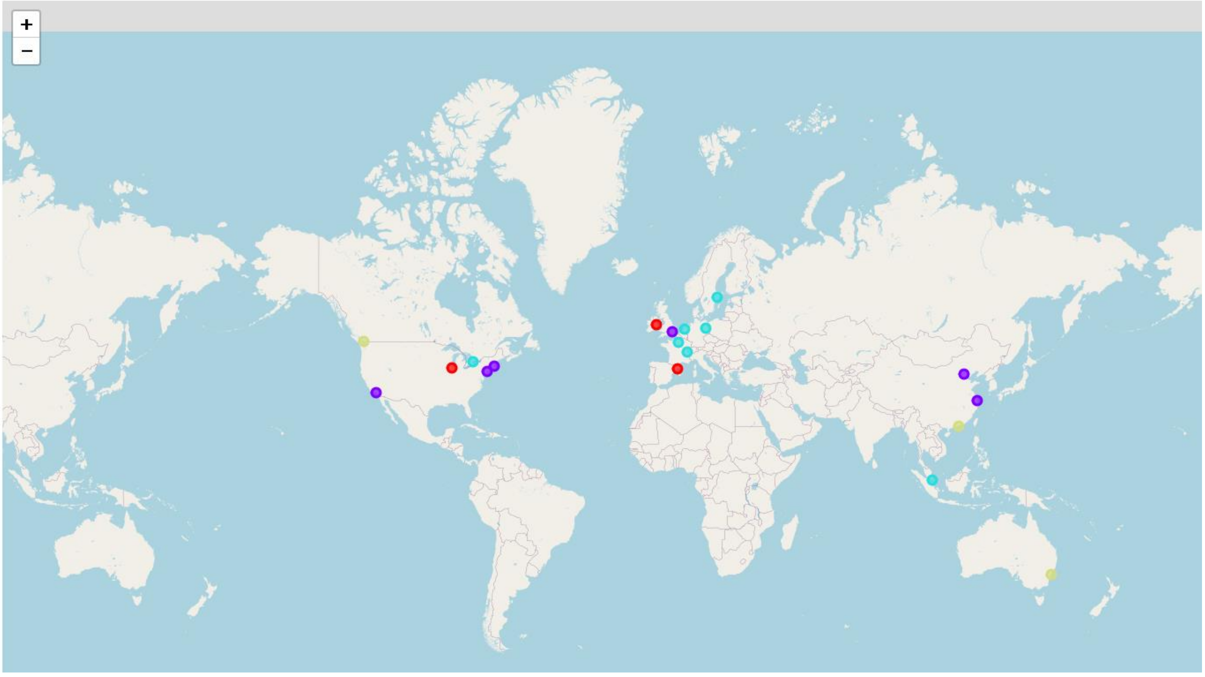
It would be hard to draw conclusions from this table except that 3rd and 6th datasets are identical.

B. Let`s see the clusters on a map:

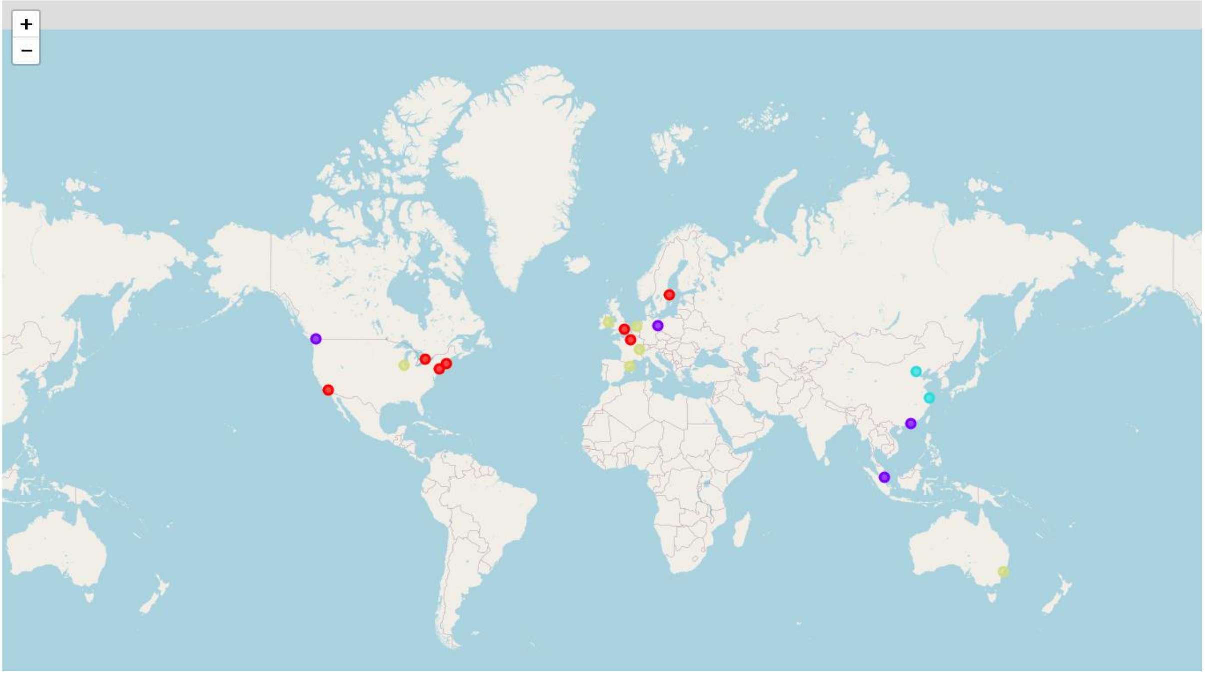
1st set of data chosen:



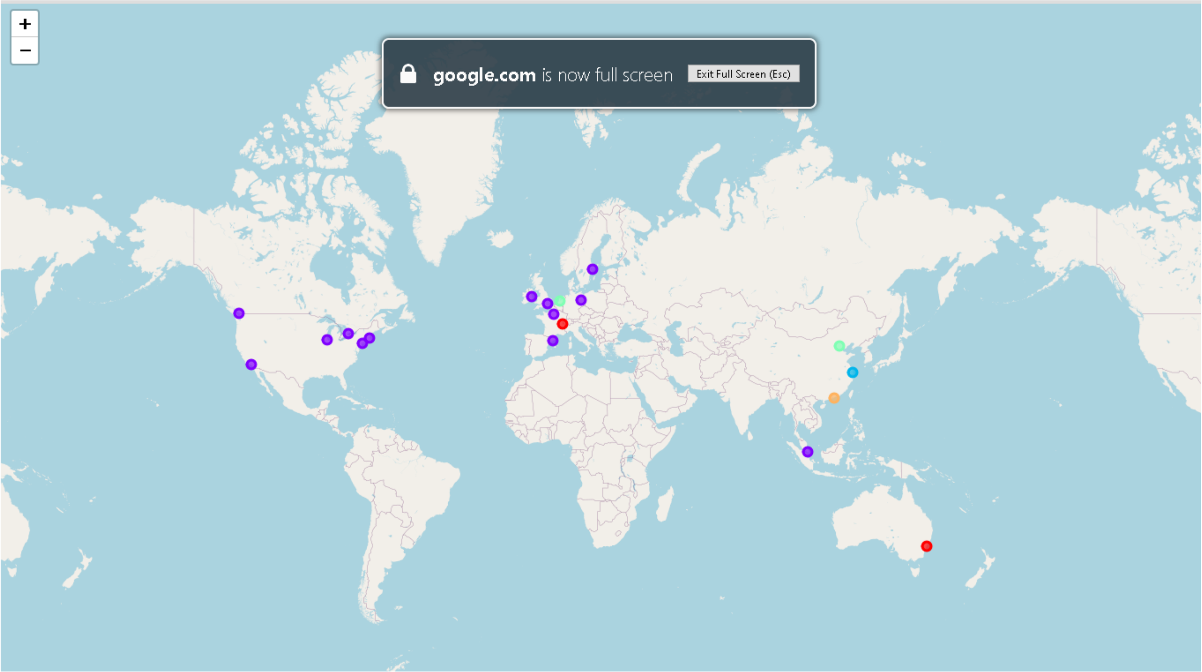
2nd set of data chosen:



3rd set of data chosen:

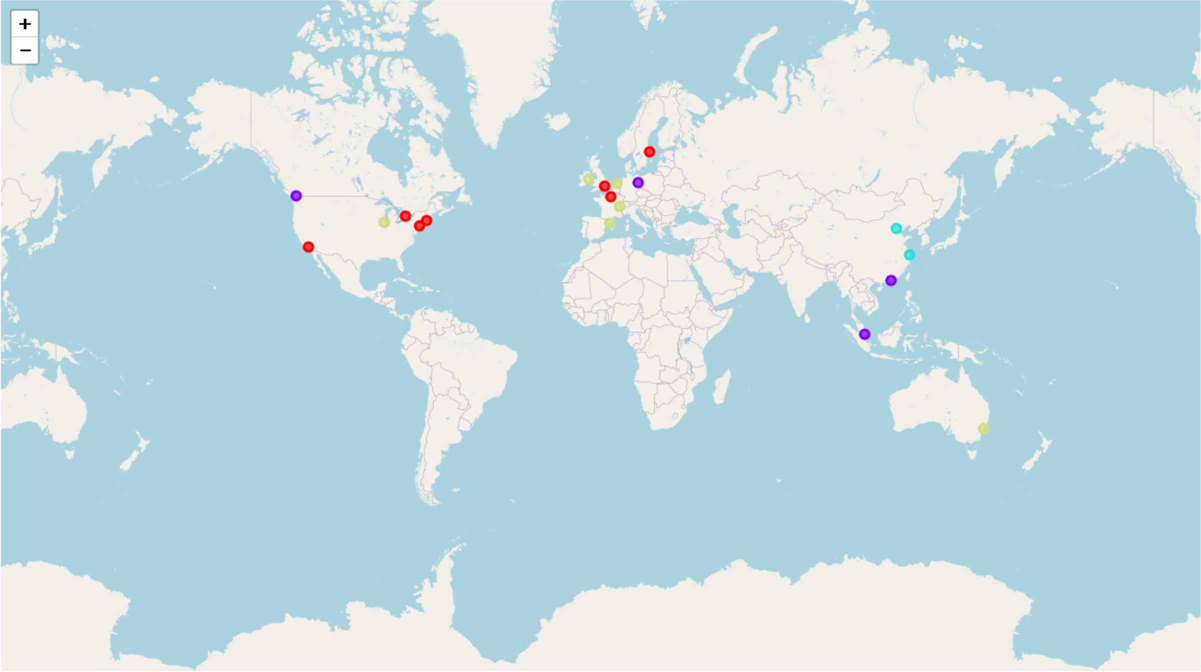


4th set of data chosen:

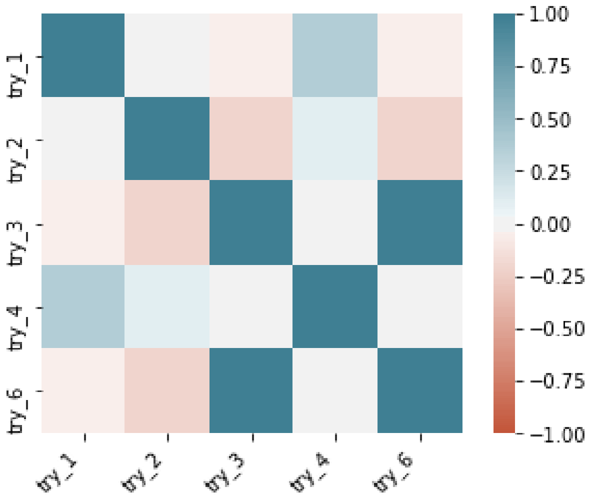


google.com is now full screen

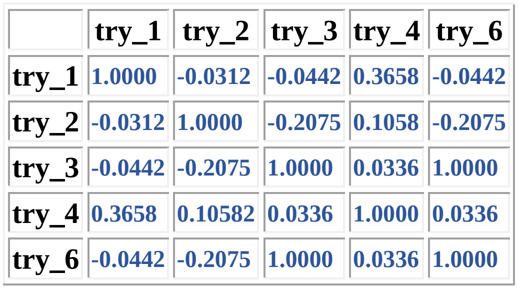
6th set of data chosen:



We may try to check how similar are the different clustering produced. We may use different measures, here let`s try correlation:



And the correlation matrix is:



	try_1	try_2	try_3	try_4	try_6
try_1	1.0000	-0.0312	-0.0442	0.3658	-0.0442
try_2	-0.0312	1.0000	-0.2075	0.1058	-0.2075
try_3	-0.0442	-0.2075	1.0000	0.0336	1.0000
try_4	0.3658	0.10582	0.0336	1.0000	0.0336
try_6	-0.0442	-0.2075	1.0000	0.0336	1.0000

4. Results

We would not pretend to have discovered a big pattern within the data observed and have a clue why some clusters were obtained with some features while not with some others. We can make several observations:

- The most valuable finding is the proposed algorithm of comparing different datasets
- We need to investigate the quality of clusters obtained with k-means to improve the precision of results and be more confident in making conclusions based on the results
- If we observe the data, apparently, there is a perfect correlation between try_3 (complete 1st set of features and try_6 (combination between try_3 and Foursquare data) which suggest that the features somehow overwhelm the features with Foursquare. This is supported by the fact that Foursquare features (try_4) have low correlation with any other group. There is need to further refine the methodology, but this would require further and more detailed investigation. There may also be a hidden error with calculations.
- Clear conclusions are hard to draw from these results apart from the fact that East-Asian/Australian clusters rarely overlap with other clusters.
- Other relations and patterns should be investigated explain what causes different clustering configurations

5. Discussion section

We investigated 6 sets of features in order to identify way to cluster 19 high technology hubs. We would not be able to draw strong conclusions on any of the initial assumptions for various reasons including the quality and reach of data used, the necessity to perform a long list of testing task to evaluate different approaches, datasets and assumptions, however we sketched a possible way to investigate relations between data pools by calculating similarity between clusters within different datasets.

It is good to note that it seems not trivial to obtain data on not so obscure sets such as the world biggest cities, despite being told that we are every day submerged in a deluge of data like never before.

Additionally, this may be used to identify sets of features with low mutual correlation for other purposes.

6. Conclusion

This project led us to delve in a variety of data science subjects such as data collection, cleaning, methodology investigation and, uncovering practical applications of a research. We tried to deal with a good number of research and analysis challenges and while mentioning a number of results, the most important contribution is in proposing an approach to evaluate datasets which may not have been discussed by analysts. The big conclusion is that we would need a much greater variety of data which consistency and integrity we may trust to make further evaluations.