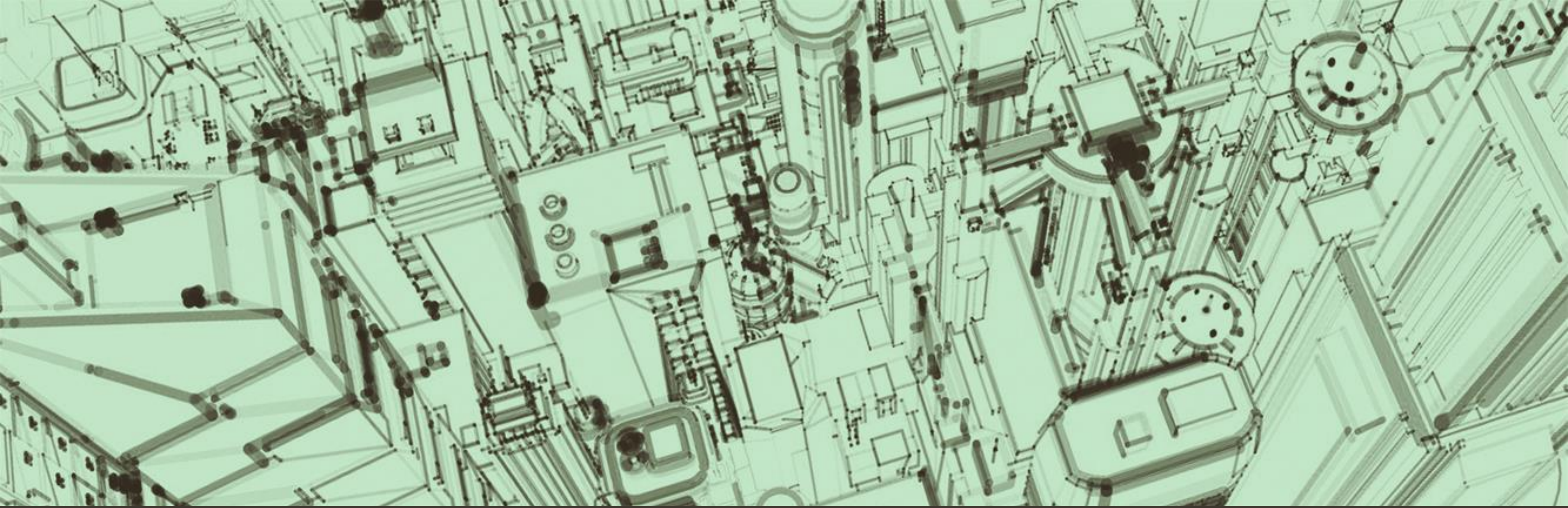


Finding Similar Cities in São Paulo State of Brazil

Luiz Alberto Hiroshi Horita

Finding where are the opportunities is valuable for entrepreneurship

- Since Industrial Revolution there is the illusion the best opportunites are at the biggest towns.
  - However, there are several products and services that still do not exist, or are scarce, as far as the cities are from these big towns.
- Understanding it, and tracing a region's cities profiles in terms of their population and most common venue categories to find where the best opportunities are is valuable for entrepreneurship.
- This kind of analysis is also useful for people needs to move to another city, but keeping accessibility to same facilities they are used to in current living town.

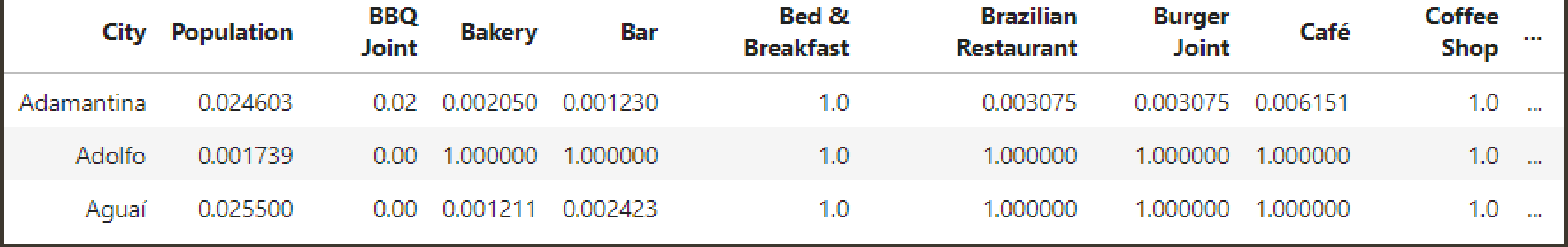
P.S.: in this Project we have analysed the cities of São Paulo (SP) state of Brazil.

Data acquisition and cleaning

- List of cities and their respective population data: scraped from Wikipedia.
- Geographical position (latitude, longitude) of each city: retrieved from GeoPy module from Python.
- List of venues in each city: retrieved from Foursquare API.
- In total, 643 cities, and 14,757 venues listed (rows) with 7 features (columns) in the raw dataframes.
- Eliminated outlier city, droped irrelevante features from both dataframes, one-hot-encoded categorical feature, grouped venues by category, joined similar categories, droped categories with too few and too much occurences.
- Cleaned dataframes contain 643 cities (rows) with 4 features (columns), and 14,767 venues (rows) with 23 features (columns).

Exploratory Analysis

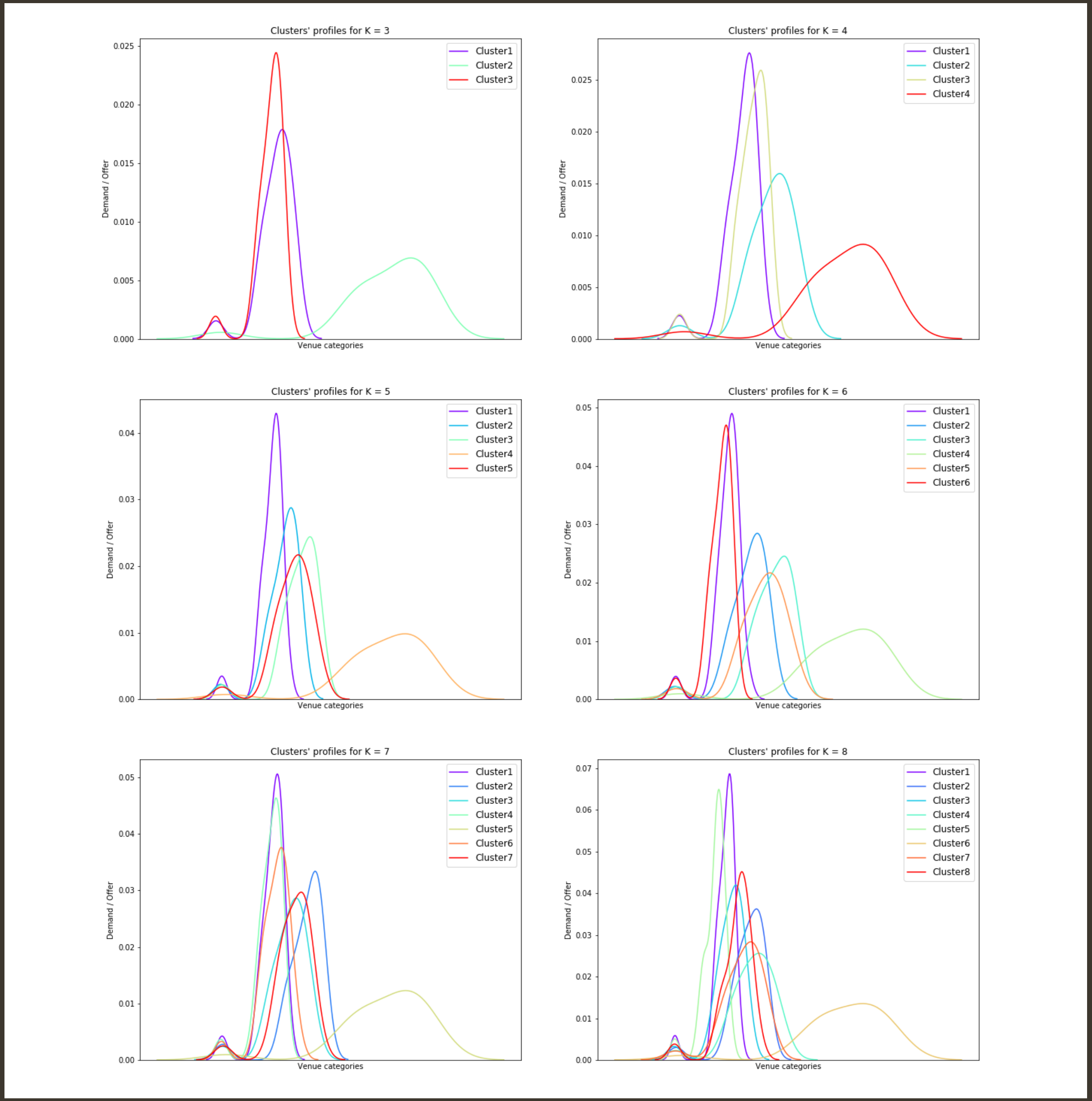
- Grouped the venues dataframe by city name, and computed mean occurrences of each venue category.
- Computed the reason of population with each veneu category, to trace each city's profile in terms of demand vs. offer estimation.
- Treated resulted NaN values.
- Normalized, by column, all features.



City	Population	BBQ Joint	Bakery	Bar	Bed & Breakfast	Brazilian Restaurant	Burger Joint	Café	Coffee Shop	...
Adamantina	0.024603	0.02	0.002050	0.001230	1.0	0.003075	0.003075	0.006151	1.0	...
Adolfo	0.001739	0.00	1.000000	1.000000	1.0	1.000000	1.000000	1.000000	1.0	...
Aguaí	0.025500	0.00	0.001211	0.002423	1.0	1.000000	1.000000	1.000000	1.0	...

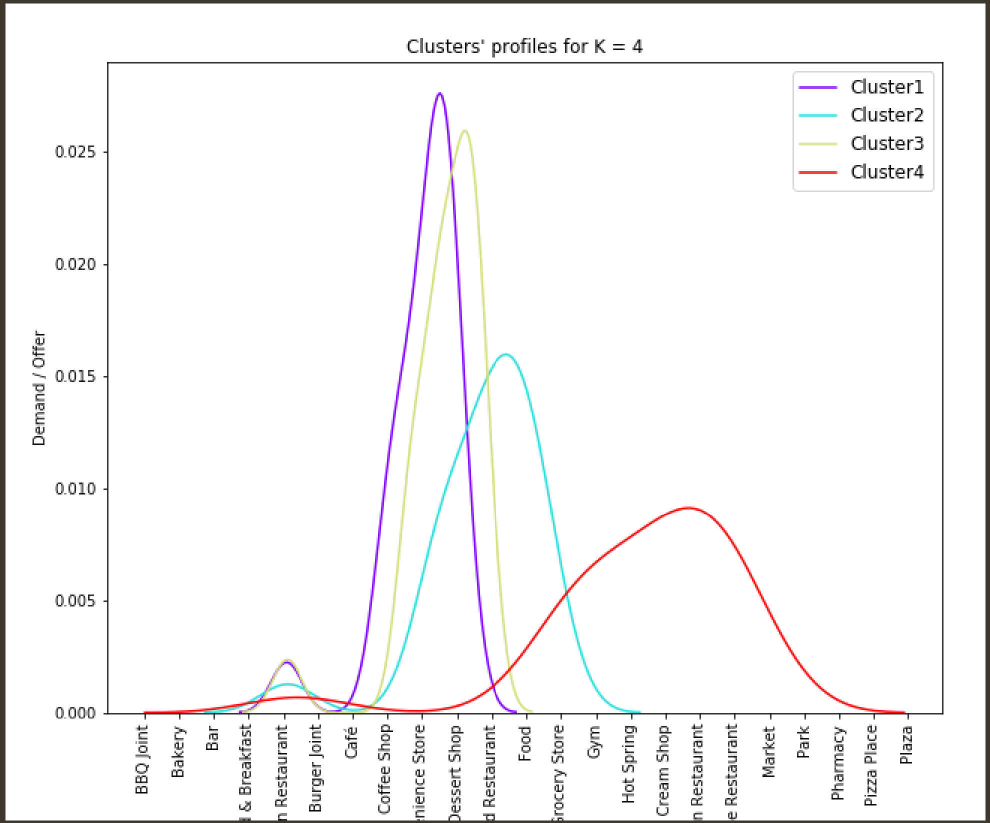
Exploring best number of clusters

- To find best K value for K-Means clustering, we tried modeling varying K from 3 to 8.

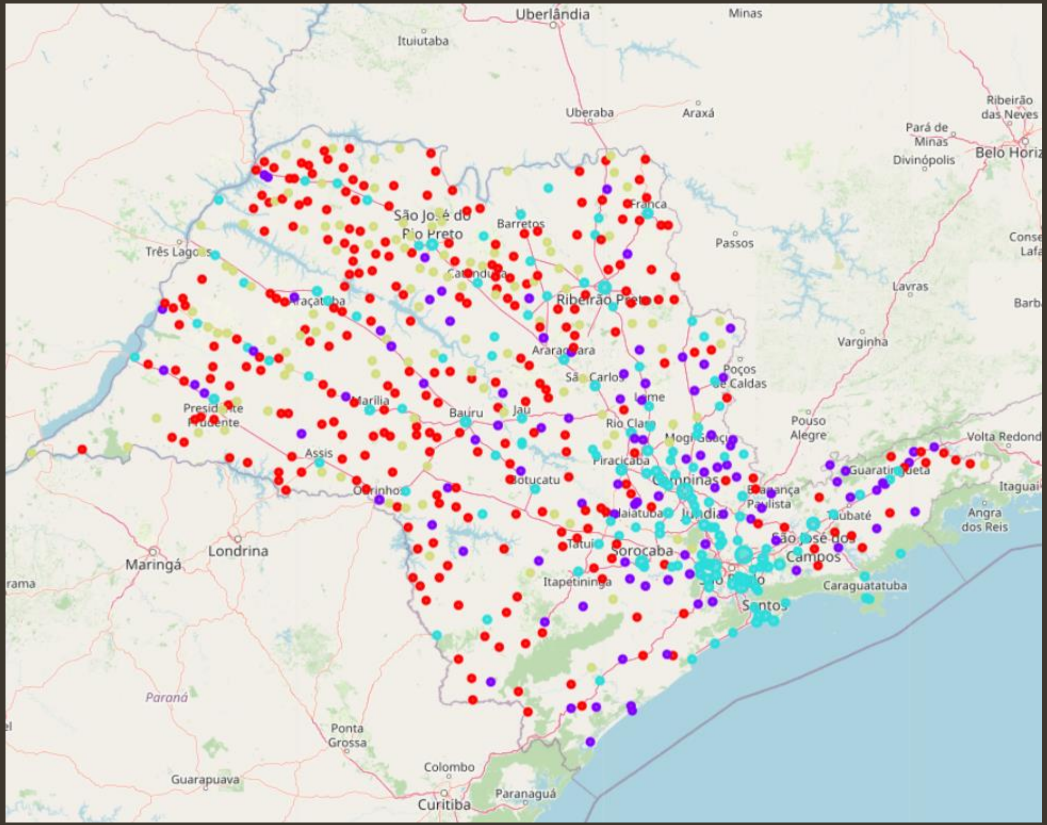


4 Clusters of cities

Distribution curves of each cluster.



SP cities clusters



Conclusion and future directions

- Useful cities segmentation and clustering to understand a state, or region, profile.
  - Good to find opportunities for entrepreneurship and in which area of business.
  - Good to find similar cities, in case a stakeholder wants to move to other town keeping same quality of life.
- Assertivity of model has room for improvement.
  - Consider quality of venues, capturing feedbacks from clientes.
  - Consider criminality.