Coursera specialization capstone project
"IBM Data Science Professional Certificate"

"The Battle of Neighborhoods"

Leandro Oliveira Bortot

Ribeirão Preto
2019

1. Introduction

Problem definition:

→ Shortening the process of finding a new house or apartment when moving.
- Lengthy
- Costly

→ In this notebook we propose a way of shortening this process by using data about the most popular venues of each neighborhood of a given city to find their "venue profile" and match it with the client’s profile.

Target audience

→ Real estate companies.

→ Web form to determine what is the "venue profile" that each client prefers

→ Suggest houses located in neighborhoods that matches the client's preferences.

→ Company: Increases the chances of closing deals before its competitors

→ Client: Saves time and money

2. Data

Web scraping with BeautifulSoup to obtain a list of all neighborhoods

https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Ribeir%C3%A3o_Preto

Google Maps API for geocoding the neighborhood names into their coordinates.

→POST request that returns a JSON file:

https://maps.googleapis.com/maps/api/geocode/json
?key=YOUR_API_KEY
&address=Centro,+Ribeirão+Preto,+São Paulo,+Brazil

Foursquare API to get all venues in each neighborhood

→POST request that returns a JSON file:

https://api.foursquare.com/v2/venues/search
?client_id=YOUR_ID&client_secret=YOUR_SECRET&v=VERSION
&ll=-21.1704008,-47.8103238
&radius=1000
&limit=9999
&intent=browse

3. Methodology

1. Web Scraping

→ BeautifulSoup library

2. Geocoding

→Google Maps API

3. Data acquisition about venues

→Foursquare's API

4. Feature reduction

→ Only most common venue categories

5. Machine learning

→k-Means clustering

- Simple clustering algorithm
- It was capable of meeting the proposed objective

→ We will consider the city of Ribeirão Preto, one of the largest of the State of São Paulo, Brazil.

4. Results

→234 neighborhood names were scraped from Wikipedia

https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Ribeirão_Preto

Lista de bairros de Ribeirão Preto

Origem: Wikipédia, a enciclopédia livre.

Esta página ou secção **não c**
as corretamente no texto ou
—*Encontre fontes:* Google (notíci

Esta é uma **lista de bairros do município brasileiro de Ribeirão**

=== Zona Norte

Índice [esconder]

1 Campos Elíseos
2 Vila Gertrudes
3 Vila Carvalho
4 Vila Albertina
5 Vila Augusta
6 Vila Esperança
7 José Sampaio
8 Jardim Procópio
9 Parque das Figueiras
10 Jardim Alexandre Balbo

```
['Campos Elíseos',
 'Vila Gertrudes',
 'Vila Carvalho',
 'Vila Albertina',
 'Vila Augusta',
 'Vila Esperança',
 'José Sampaio',
 'Jardim Procópio',
 'Parque das Figueiras',
 'Jardim Alexandre Balbo']
```

4. Results

→ All of them were geocoded with the Google Maps API

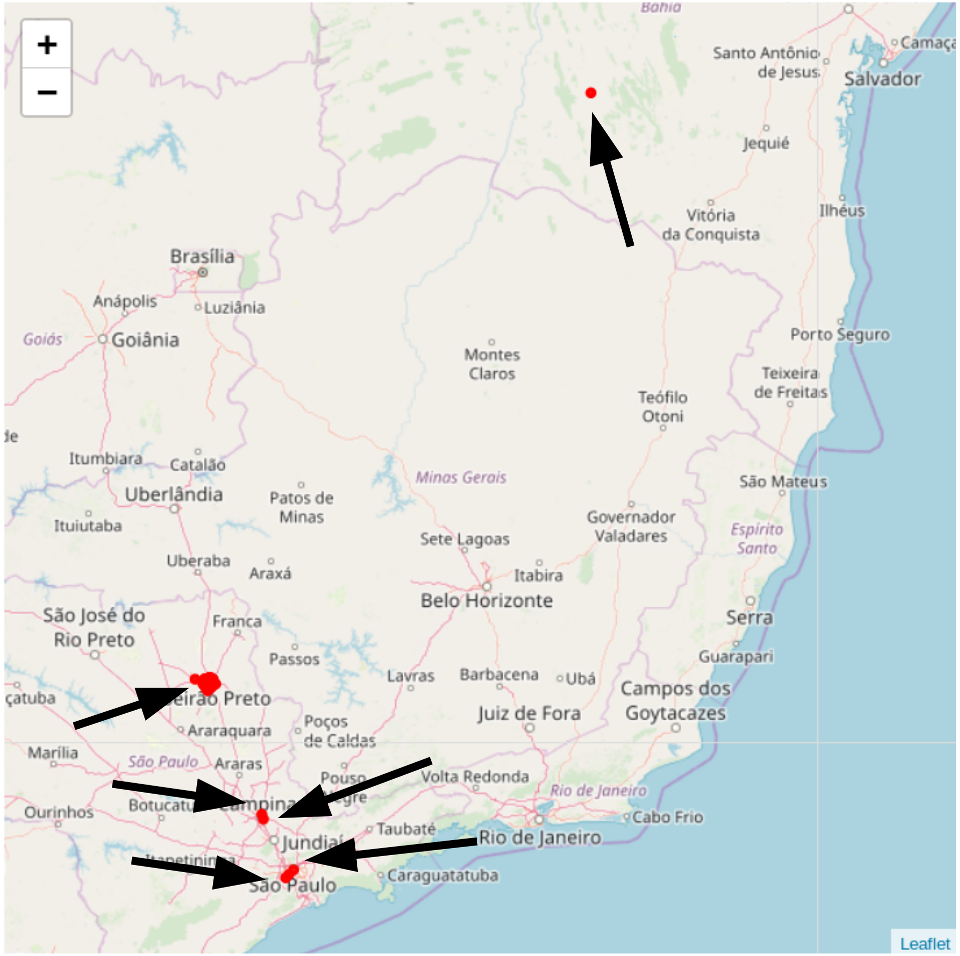
	Latitude	Longitude
Campos Elíseos	-21.162380	-47.798293
Vila Gertrudes	-23.621491	-46.696494
Vila Carvalho	-21.142448	-47.792124
Vila Albertina	-21.150803	-47.815570
Vila Augusta	-21.140195	-47.821237
Vila Esperança	-21.154167	-47.770728
José Sampaio	-21.138542	-47.826679
Jardim Procópio	-21.133577	-47.834086
Parque das Figueiras	-21.130052	-47.838307
Jardim Alexandre Balbo	-21.129665	-47.829148

4. Results

→However, 6 of them were wrong (black arrows)

→Those 6 neighborhoods doesn’t exist anymore, so Google Maps pointed at the same neighborhood names in other cities.

→ Those 6 neighborhood were dropped from the dataframe



4. Results

→ All venues around each neighborhood obtained with the Foursquare API a

→ Coordinates dataframe was merged with the venues dataframe

→ Dropped venues for which the category was missing

→ Dropped categories that were present in less than 0.5% of the total venues

→ Applied the one-hot encode approach to the “Venue category” column

→ Calculated the frequency of each venue category for each neighborhood

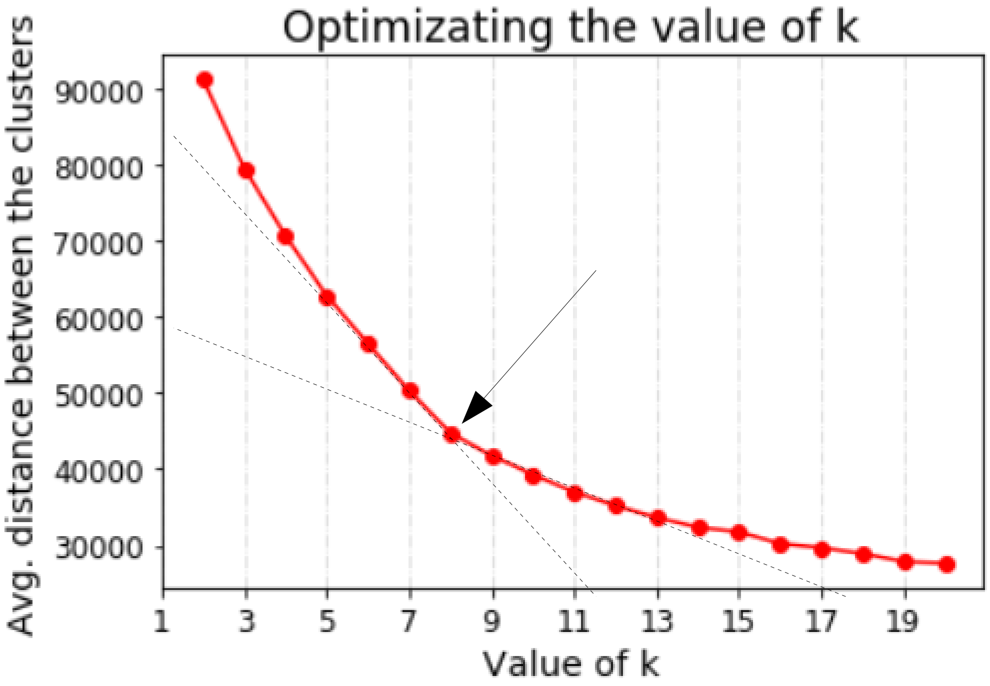
	Venue Name	Venue Category	Latitude	Longitude	Neighborhood
0	Salgaderia 2 irmãos	Snack Place	-21.161945	-47.798283	Campos Elíseos
1	Esquina Do Sr. Olívio Salgaderia	Snack Place	-21.161947	-47.797994	Campos Elíseos
2	Pinguim Frios	MISSING	-21.163100	-47.797900	Campos Elíseos
3	Alex Cabeleireiro	Salon / Barbershop	-21.162427	-47.797245	Campos Elíseos
4	AGN CONSTRUÇÕES LTDA - RIBEIRÃO PRETO	Office	-21.163195	-47.798478	Campos Elíseos
5	Congregação Crista no Brasil (central)	Non-Profit	-21.161597	-47.799066	Campos Elíseos
6	xapuri	Brazilian Restaurant	-21.162982	-47.797731	Campos Elíseos
7	Clinica Integral	Dentist's Office	-21.163759	-47.798519	Campos Elíseos
8	Aurora Festas	General Entertainment	-21.163899	-47.797780	Campos Elíseos
9	Posto Beta News	Gas Station	-21.163554	-47.797596	Campos Elíseos

	Venue Name	Venue Category	Latitude	Longitude	Neighborhood
30080	Antonia Store	Women's Store	-21.266308	-47.816792	Recanto das Flores
30081	Loja ANTÔNIA	Women's Store	-21.266278	-47.816931	Recanto das Flores
30082	Ponto De Encontro	Restaurant	-21.265534	-47.818270	Recanto das Flores
30083	Cervejaria Walfänger	Brewery	-21.262733	-47.819446	Recanto das Flores
30084	Quitanda César	Deli / Bodega	-21.259098	-47.814017	Recanto das Flores
30085	Alpha Instrumentos	Health & Beauty Service	-21.262761	-47.817134	Recanto das Flores
30086	Comunidade Nova Geração	Non-Profit	-21.289214	-47.812672	Recanto das Flores
30087	Feira de Bonfim	Farmers Market	-21.258972	-47.814057	Recanto das Flores
30088	Recanto Caipira	Comfort Food Restaurant	-21.268431	-47.815307	Recanto das Flores
30089	Sítio Manga Rosa	Garden Center	-21.288823	-47.812413	Recanto das Flores

4. Results

→ k-Means clustering

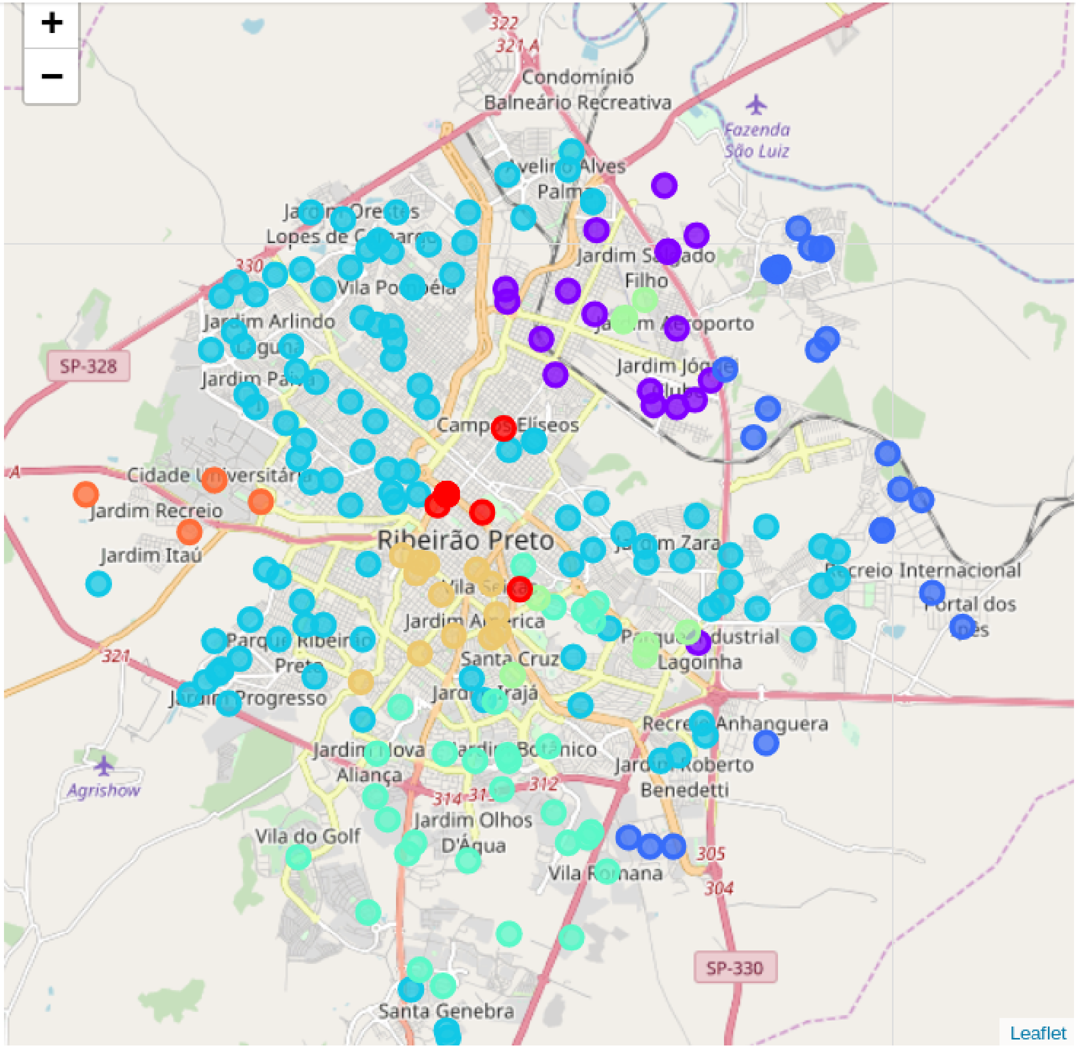
→ Slight elbow point at *k*=8



4. Results

→ Bigger cluster = 50% of neighborhoods
→ Smaller cluster = 2% of neighborhoods

Cluster ID	Count	Percentage
2	112	50.2
3	30	13.5
1	24	10.8
0	18	8.1
7	13	5.8
5	13	5.8
4	9	4.0
6	4	1.8



5. Discussion

→ Three most common venue categories of each cluster reveal differences between them
→ Allow us to imagine which "client profile" fits the best with each cluster

ID	Three most common venue categories	Good for clients that ...
0	Factory, Office, Entertainment	will work in factories and want to live close to work.
1	Entertainment, Residential, Events	prefer a residential area with lots of entertainment options.
2	Beauty, Residential, Entertainment	prefer a quiet residential area.
3	Residential, Office, Entertainment	work in offices and want a good equilibrium between work and leisure.
4	Office, Buildings, Co-working space	work in offices and want to be immersed in the work environment.
5	Doctors, Office, Beauty	need constant medical attention and want to live close to several doctor's offices.
6	College, Hospital, Student Center	are students.
7	Automotive, Office, Shops	want to start a new business in the automobile sector.

6. Conclusion

→ Simple analysis

→ City of Ribeirão Preto (SP, Brazil):

- 8 clusters of neighborhoods
- Different venues profiles
- Mapped to different "client profiles"

→ This can be helpful for a real estate company when suggesting new houses to its clients.

→ This analysis can be improved in several ways to become more useful.

- User feedback is essential

→ If the cluster that best matches the client's "venue profile" has too many neighborhoods:

- Perform new clustering considering only the elements of the chosen cluster
- Allows the selection of neighborhoods based on smaller differences
- This can be done several times until there's only one cluster left as the best match