Battle of the Neighborhoods

Where should the Data Scientist go?

Rodrigo Link

The problem

- After the completion of a certain Data Science course, an young and brave Brazilian data scientist is flooded with job offers from all over the world. After thoughtful consideration, he realizes he really likes where he currently lives and would like to have the same kind of venues around him wherever the job is.
- From the job offers, he selects a few cities that seem interesting and sets out to compare his current city's attractions to the neighborhoods of each new possible location, trying to identify similarities.

Data available

- Input 1 – current location [Copacabana, Rio de Janeiro, BR]
- Input 2 – list of cities where there are interesting job offers
- List of neighborhoods – webscraped from various sources
- List of venues – obtained from Foursquare API

What we want

- First Approach
  - Top 3 most similar neighborhoods from each city
  - New attractions unavailable at current location

- Second Approach
  - Clustering technique
  - What neighborhoods are in the same cluster as the current city

List of neighborhoods

- Rio de Janeiro, Brazil – 1 current location
- Porto Alegre, Brazil – 48 neighborhoods
  - https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Porto_Alegre
- Wellington, New Zealand – 12 neighborhoods
  - https://wellington.govt.nz/your-council/elections/wellington-city-wards/maps-by-ward-community-board-and-suburb
- Stockholm, Sweden – 107 neighborhoods
  - https://en.wikipedia.org/wiki/Category:Districts_of_Stockholm
- Vancouver, Canada – 22 neighborhoods
  - https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Vancouver

Webscraping results

	Country	City	Neighborhood	Latitude	Longitude
0	Brazil	Rio de Janeiro	Copacabana	-22.971964	-43.184343
1	Brazil	Porto Alegre	Aberta dos Morros	-30.160022	-51.197486
2	Brazil	Porto Alegre	Agronomia	-30.069267	-51.149217
3	Brazil	Porto Alegre	Anchieta	-29.975652	-51.174903
4	Brazil	Porto Alegre	Arquipélago	-29.992760	-51.226618
...	...	...	...	...	...
234	Canada	Vancouver	Strathcona	49.277693	-123.088539
235	Canada	Vancouver	Sunset	49.219093	-123.091665
236	Canada	Vancouver	Victoria-Fraserview	49.218979	-123.063816
237	Canada	Vancouver	West End	49.284131	-123.131795
238	Canada	Vancouver	West Point Grey	49.268102	-123.202643

239 rows × 5 columns

Foursquare API

- Using the Foursquare API, we are able to get the top 100 attractions within a 1km radius from the center of each neighborhood.

	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rio de Janeiro	Copacabana	-22.971964	-43.184343	Praia de Copacabana	-22.972441	-43.183436	Beach
1	Rio de Janeiro	Copacabana	-22.971964	-43.184343	Windsor California Hotel	-22.972704	-43.185707	Hotel
2	Rio de Janeiro	Copacabana	-22.971964	-43.184343	Hotel Sesc Copacabana	-22.973265	-43.187299	Hotel
3	Rio de Janeiro	Copacabana	-22.971964	-43.184343	Bibi Sucos	-22.972092	-43.186564	Juice Bar
4	Rio de Janeiro	Copacabana	-22.971964	-43.184343	Bar & Champanheria Copacabana	-22.974220	-43.186296	Beach Bar

Proportion of venues

- By counting the number of venues of each kind and divinding by the number of venues at each neighborhood, we can get the proportion of each type. This is going to used to compare the different neighborhoods.

	City	Neighborhood	Acai House	Accessories Store	Airport	Airport Lounge	Airport Service	Airport Terminal	Airport Tram	American Restaurant	...	Water Park	Waterfront	Wine Bar	Wine Shop	Wings Joint
0	Porto Alegre	Aberta dos Morros	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
1	Porto Alegre	Agronomia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
2	Porto Alegre	Anchieta	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
3	Porto Alegre	Arquipélago	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
4	Porto Alegre	Auxiliadora	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.01	0.0

5 rows × 423 columns

First Approach - Correlation

- We calculated the correlation between the data from each neighborhood with that from the current location. It is a measure of how similar they are.
  - Top 3 most similar neighborhoods from each city
  - New attractions unavailable at current location

First Approach – Porto Alegre

1. The first neighborhood from Porto Alegre, Cidade Baixa, has a correlation of 0.47.
The 3 new types of venues are 'Arts & Crafts store', 'Beer store' and 'Bookstore'.

2. The second neighborhood, Praia de Belas, has a correlation of 0.41.
The 3 new types of venues are 'Bar', 'Buffet' and 'Salad Place'.

3. The third neighborhood, Floresta, has a correlation of 0.40.
The 3 new types of venues are 'Bar', 'Beer bar' and 'Buffet'.

First Approach – Wellington

1. The first neighborhood from Wellington, Horokiwi, has a correlation of 0.63.
The 3 new types of venues are 'Bar', 'Hotel' and 'River'.

2. The second neighborhood, Oriental Bay, has a correlation of 0.46. The 3 new types of venues are 'Café', 'Chinese restaurant' and 'Bar'.

3. The third neighborhood, Pipitea, has a correlation of 0.43. The 3 new types of venues are 'Café', 'Bar' and 'Vietnamese Restaurant'.

First Approach – Stockholm

1. The first neighborhood from Stockholm, Riddarholmen, has a correlation of 0.52.
The 3 new types of venues are 'Scandinavian Restaurant', 'Café' and 'Bar'.

2. The second neighborhood, Gamla stan , has a correlation of 0.51.
The 3 new types of venues are 'Theater', 'Bookstore' and 'Wine bar'.

3. The third neighborhood, Älvsjö, has a correlation of 0.48.
The 3 new types of venues are 'Hotel', 'Supermarket' and 'Café'.

First Approach – Vancouver

1. The first neighborhood from Vancouver, Downtown, has a correlation of 0.50.
The 3 new types of venues are 'Seafood restaurant' and a tie for several others.

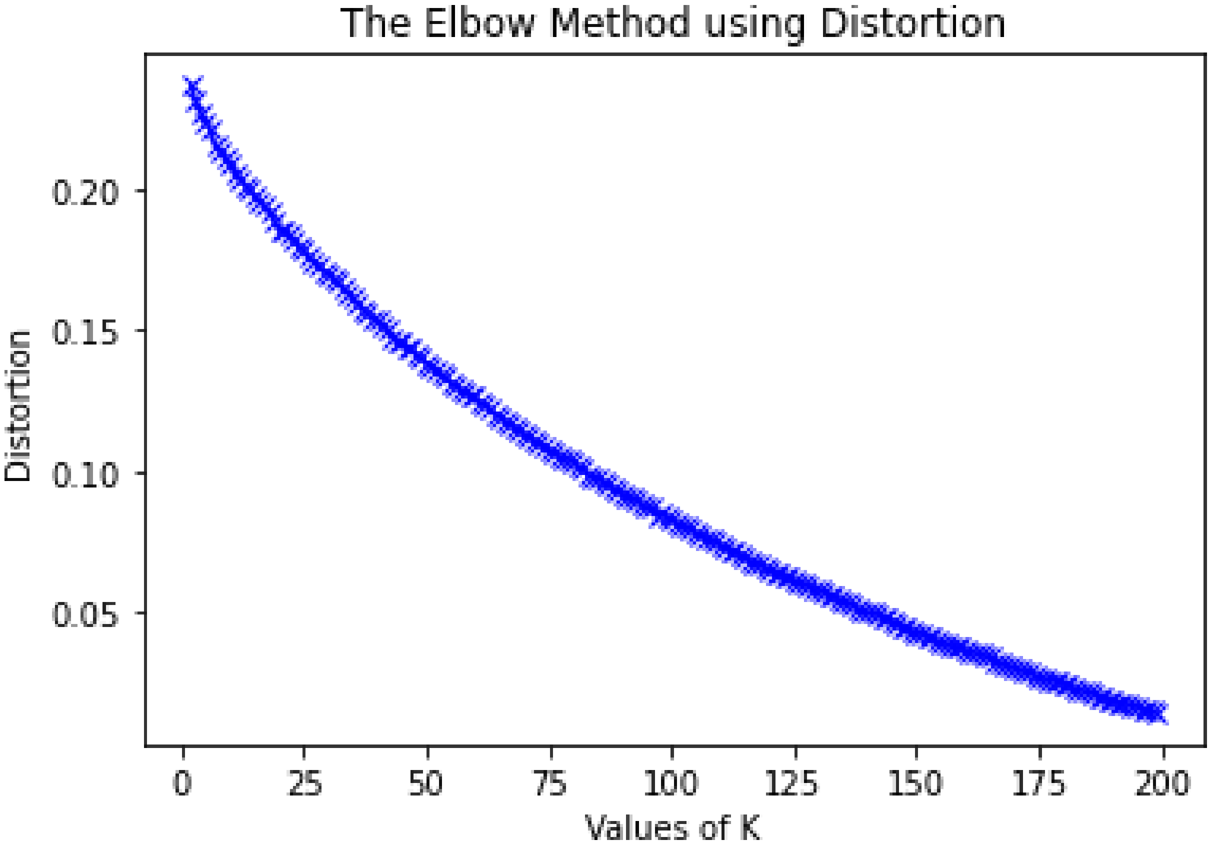
2. The second neighborhood, West End, has a correlation of 0.44.
The 3 new types of venues are 'Dessert shop', 'Bookstore' and a tie for several others.

3. The third neighborhood, Fairview, has a correlation of 0.23.
The 3 new types of venues are 'Breafast spot', 'Theater' and 'Bookstore'.

Second Approach - Clustering

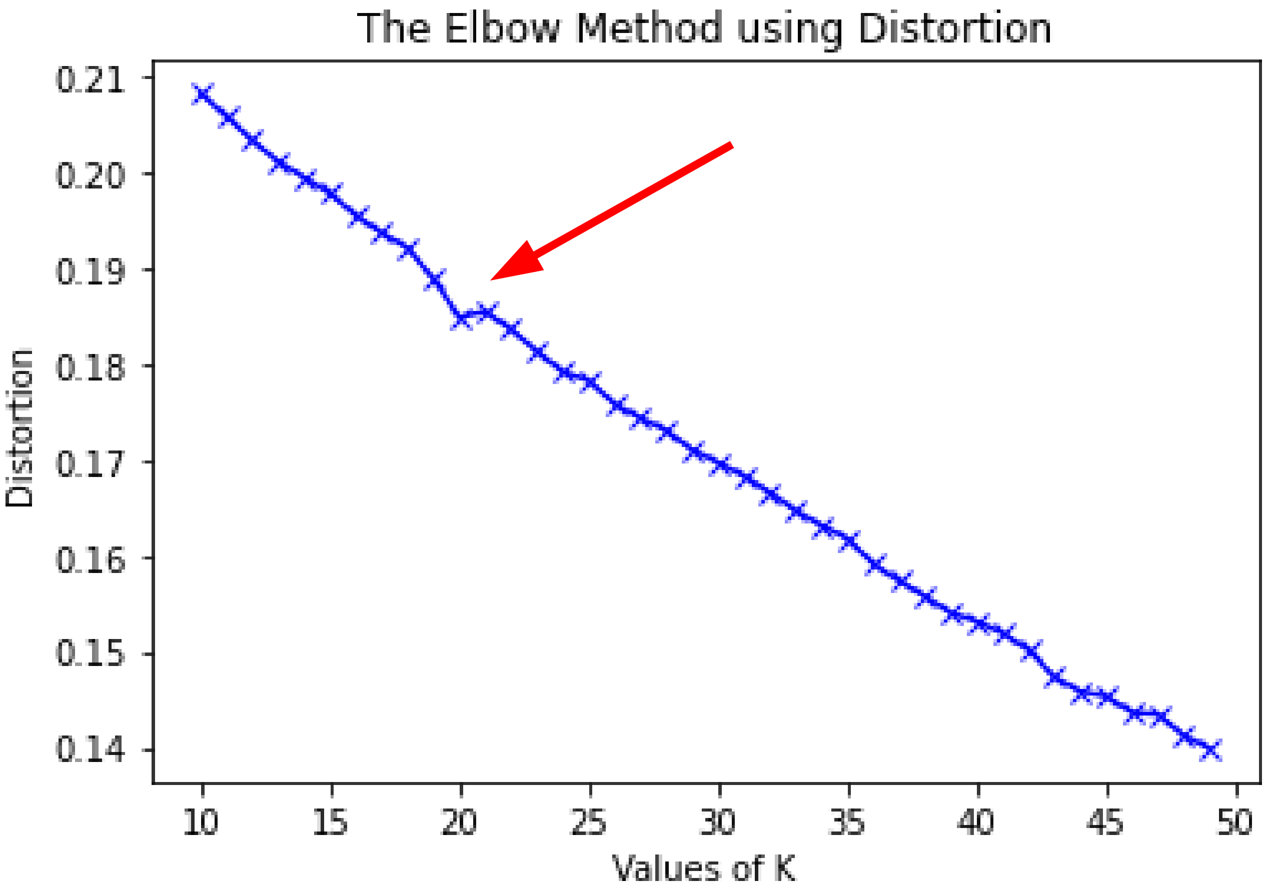
- From the data for each neighborhood, we create *k* groups to try to capture some common features.
- The first thing to do is try to figure out the best number of groups by trying a range of possibilities and measuring the sum of the distance of each point from the closest cluster center.

Ideal number of clusters



No elbow in sight!
Let’s look closer.

Ideal number of clusters



For lack of better options, **k=20**

Where is the current location?

- To find the neighborhoods more alike to the current location, let’s find out in what cluster it is in.

```
print('The current location is in cluster: ',all_grouped.iloc[85,-1])
```

```
The current location is in cluster:  4
```

Cluster 4

132	Stockholm	Riddarholmen	0.524632
106	Stockholm	Gamla stan	0.513101
155	Vancouver	Downtown	0.501553
215	Wellington	Oriental Bay	0.455498
174	Vancouver	West End	0.437037
218	Wellington	Pipitea	0.427678
129	Stockholm	Norrmalm	0.426942
208	Wellington	Mount Victoria	0.420843
145	Stockholm	Södermalm	0.402839
131	Stockholm	Reimersholme	0.402465
124	Stockholm	Långholmen	0.401361
227	Wellington	Thorndon	0.380576
15	Porto Alegre	Cascata	0.374438
226	Wellington	Te Aro	0.344184
122	Stockholm	Liljeholmen	0.341546
153	Stockholm	Östermalm	0.334611
230	Wellington	Wellington Central	0.332335
207	Wellington	Mount Cook	0.314744
120	Stockholm	Kungsholmen	0.304544
117	Stockholm	Johanneshov	0.289366
107	Stockholm	Gröndal	0.287672
92	Stockholm	Birkastad	0.263217
140	Stockholm	Stadshagen	0.261198
176	Wellington	Aro Valley	0.254607
136	Stockholm	Skeppsholmen	0.232921
157	Vancouver	Fairview	0.232422
228	Wellington	Vogeltown	0.226819
223	Wellington	Strathmore Park	0.218640
165	Vancouver	Mount Pleasant	0.199320
125	Stockholm	Marieberg	0.198011
163	Vancouver	Kitsilano	0.195735
109	Stockholm	Gärdet	0.194435
158	Vancouver	Grandview-Woodland	0.188046
168	Vancouver	Riley Park	0.186867
213	Wellington	Northland	0.164207
171	Vancouver	Strathcona	0.152995
147	Stockholm	Vasastan	0.151268
118	Stockholm	Kristineberg	0.136491
154	Vancouver	Arbutus Ridge	0.121455
170	Vancouver	South Cambie	0.105678
164	Vancouver	Marpole	0.094507
156	Vancouver	Dunbar-Southlands	0.092564
204	Wellington	Miramar	0.080751
175	Vancouver	West Point Grey	0.062817
166	Vancouver	Oakridge	0.048274
159	Vancouver	Hastings-Sunrise	0.045237
149	Stockholm	Vårberg	0.042092
97	Stockholm	Djurgården	0.034932
172	Vancouver	Sunset	0.033749
169	Vancouver	Shaughnessy	0.031581
197	Wellington	Kilbirnie	0.022443
160	Vancouver	Kensington-Cedar Cottage	-0.003920
231	Wellington	Wilton	-0.034435

Summary of the two approaches

- Riddarholmen, in Stockholm, Sweden, had the highest correlation in the its city and has the highest correlation inside cluster 4.
- Gamla Stan, in Stockholm, Sweden, had the second highest correlation in the its city and has the second highest correlation inside cluster 4.
- Downtown, in Vancouver, Canada, had the highest correlation in the its city and the third highest correlation inside cluster 4.
- Oriental Bay, in Wellington, New Zealand, had the second highest correlation in its city and the fourth highest correlation inside cluster 4.
- Cascata, in Porto Alegre, Brazil, is not in the top 3 for its city and is 13th highest correlation inside cluster 4.

Conclusion

- All things equal, the order of choice for the Data Scientist should be:

  1. Stockholm (either Riddarholmen or Gamla Stan);

  2. Vancouver (Downtown)

  3. Wellington (Oriental Bay)