Capstone project - The Battle of Neighbourhoods

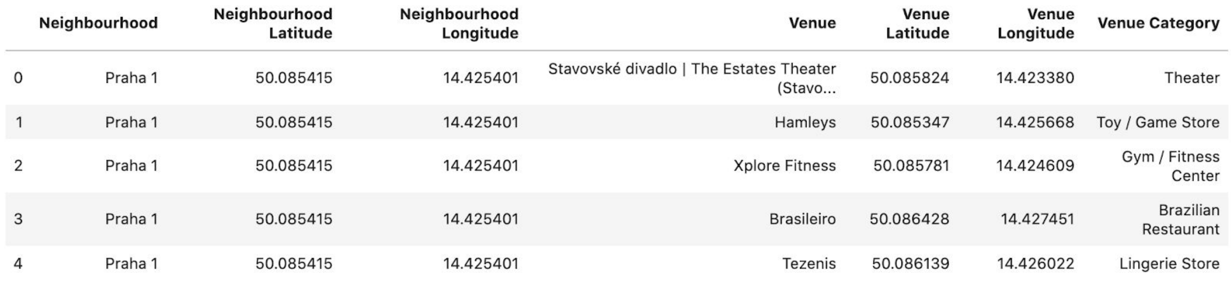
Prague Start-up searching for location of their new office

Introduction

A Prague start-up company is growing very quickly. Its current offices in Prague 6 do not have sufficient capacity for all the new employees. The company is thus searching for new offices. It has a very specific corporate culture and believes that the current location was a perfect fit for its employees. The founders believe that its location in a young gentrifying neighborhood attracts the best young workers who share the same mindset with the founders. Therefore while searching for a new location for their new office, they want to compare Prague neighbourhoods to test how similar or dissimilar they actually are. Prague is divided into districts labeled with numbers. This distribution will be key for the founders to find out where they want to move their new headquarters.

Data

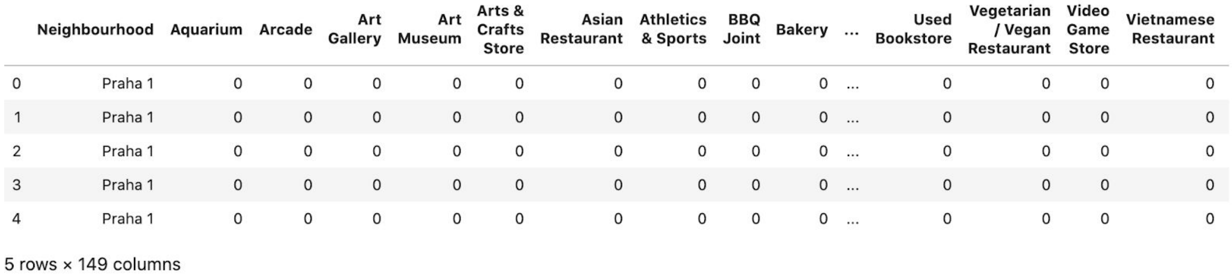
In order to find out which neighbourhoods fit the corporate culture of this start-up, the founders need Foursquare location data. The data will be accessed via Foursquare API. As mentioned above, the founders will use Prague's numerical division of districts for their decision on where to locate their new office. Therefore, data about Postal codes of these districts will be needed. These data will be accessed thanks to a csv file from Prague's info websites. Together, our dataset will thus contain information about different venues in each Prague's district. With this data, we should be able to cluster each neighbourhood as the founders request and find the perfect matches for them. Below you will find the first five rows of our dataset.



	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Praha 1	50.085415	14.425401	Stavovské divadlo | The Estates Theater (Stavo...	50.085824	14.423380	Theater
1	Praha 1	50.085415	14.425401	Hamleys	50.085347	14.425668	Toy / Game Store
2	Praha 1	50.085415	14.425401	Xplore Fitness	50.085781	14.424609	Gym / Fitness Center
3	Praha 1	50.085415	14.425401	Brasileiro	50.086428	14.427451	Brazilian Restaurant
4	Praha 1	50.085415	14.425401	Tezenis	50.086139	14.426022	Lingerie Store

Methodology

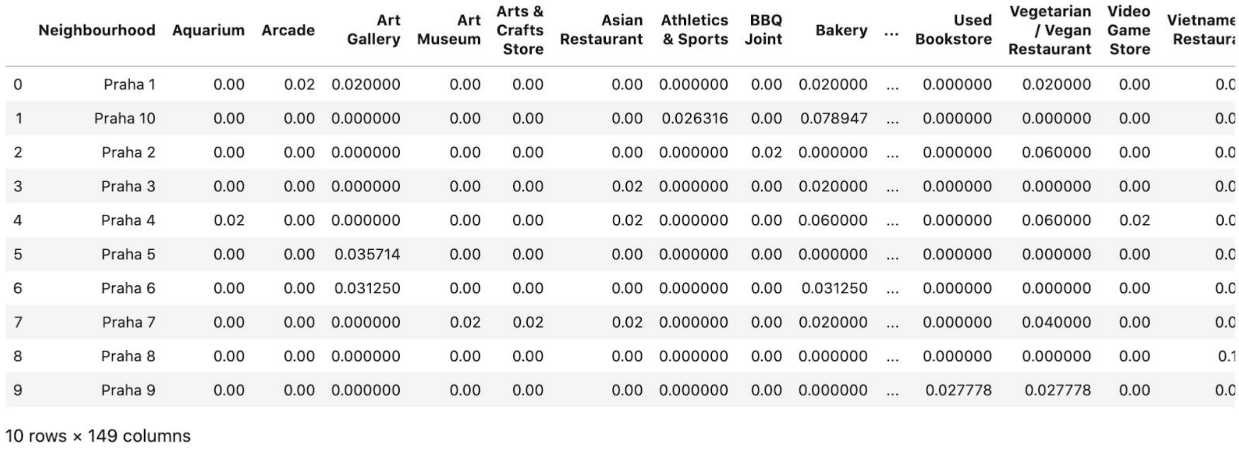
For this Capstone project the Machine learning technique - K-means clustering was used. This classification method allows processing data for finding similarities and differences in data. In our case, finding the most similar and dissimilar neighbourhoods in Prague it fits perfectly. Firstly, after uploading Foursquare API and Prague district data as seen in the dataset above, we need to obtain dummy variables for each venue category, in order to have a dataset with integers rather than strings. For each venue the function assigns 0 if it does not match the venue category and 1 if it does match. A preview of the resulting dataset can be seen below.



	Neighbourhood	Aquarium	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	...	Used Bookstore	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant
0	Praha 1	0	0	0	0	0	0	0	0	0	...	0	0	0	0
1	Praha 1	0	0	0	0	0	0	0	0	0	...	0	0	0	0
2	Praha 1	0	0	0	0	0	0	0	0	0	...	0	0	0	0
3	Praha 1	0	0	0	0	0	0	0	0	0	...	0	0	0	0
4	Praha 1	0	0	0	0	0	0	0	0	0	...	0	0	0	0

5 rows × 149 columns

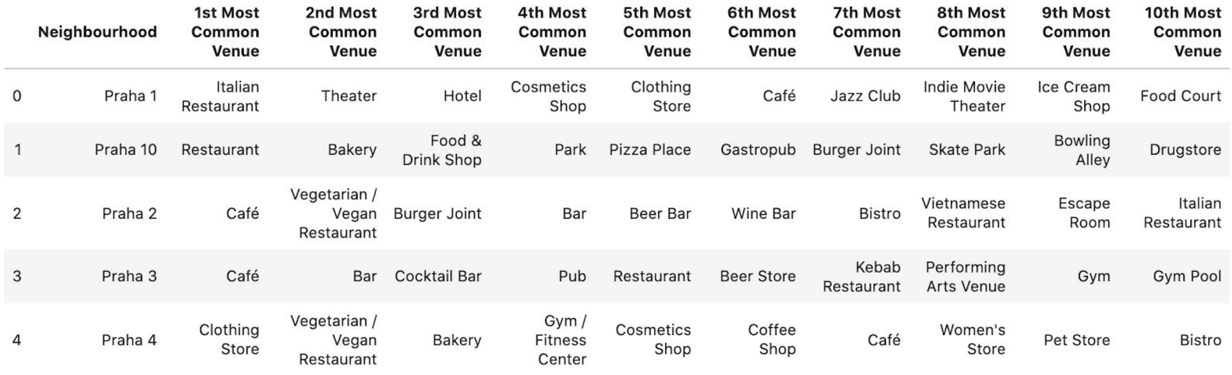
Thanks to this transformation, we can obtain the occurrence of each venue category in each neighbourhood. This can be done grouping by each neighbourhood and the mean values of each venue category which therefore shows the frequency of each venue category in the dataset (table below).



	Neighbourhood	Aquarium	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	...	Used Bookstore	Vegetarian / Vegan Restaurant	Video Game Store	Vietname Restaura
0	Praha 1	0.00	0.02	0.020000	0.00	0.00	0.00	0.000000	0.00	0.020000	...	0.000000	0.020000	0.00	0.0
1	Praha 10	0.00	0.00	0.000000	0.00	0.00	0.00	0.026316	0.00	0.078947	...	0.000000	0.000000	0.00	0.0
2	Praha 2	0.00	0.00	0.000000	0.00	0.00	0.00	0.000000	0.02	0.000000	...	0.000000	0.060000	0.00	0.0
3	Praha 3	0.00	0.00	0.000000	0.00	0.00	0.02	0.000000	0.00	0.020000	...	0.000000	0.000000	0.00	0.0
4	Praha 4	0.02	0.00	0.000000	0.00	0.00	0.02	0.000000	0.00	0.060000	...	0.000000	0.060000	0.02	0.0
5	Praha 5	0.00	0.00	0.035714	0.00	0.00	0.00	0.000000	0.00	0.000000	...	0.000000	0.000000	0.00	0.0
6	Praha 6	0.00	0.00	0.031250	0.00	0.00	0.00	0.000000	0.00	0.031250	...	0.000000	0.000000	0.00	0.0
7	Praha 7	0.00	0.00	0.000000	0.02	0.02	0.02	0.000000	0.00	0.020000	...	0.000000	0.040000	0.00	0.0
8	Praha 8	0.00	0.00	0.000000	0.00	0.00	0.00	0.000000	0.00	0.000000	...	0.000000	0.000000	0.00	0.1
9	Praha 9	0.00	0.00	0.000000	0.00	0.00	0.00	0.000000	0.00	0.000000	...	0.027778	0.027778	0.00	0.0

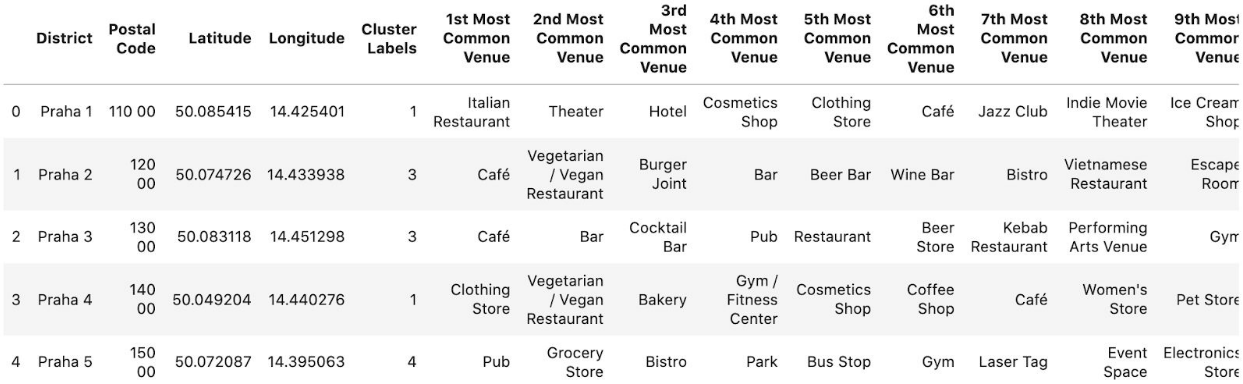
10 rows × 149 columns

The next step is to sort the data and get the 10 most common venues in each neighbourhood. Again, this can be seen in the table below.

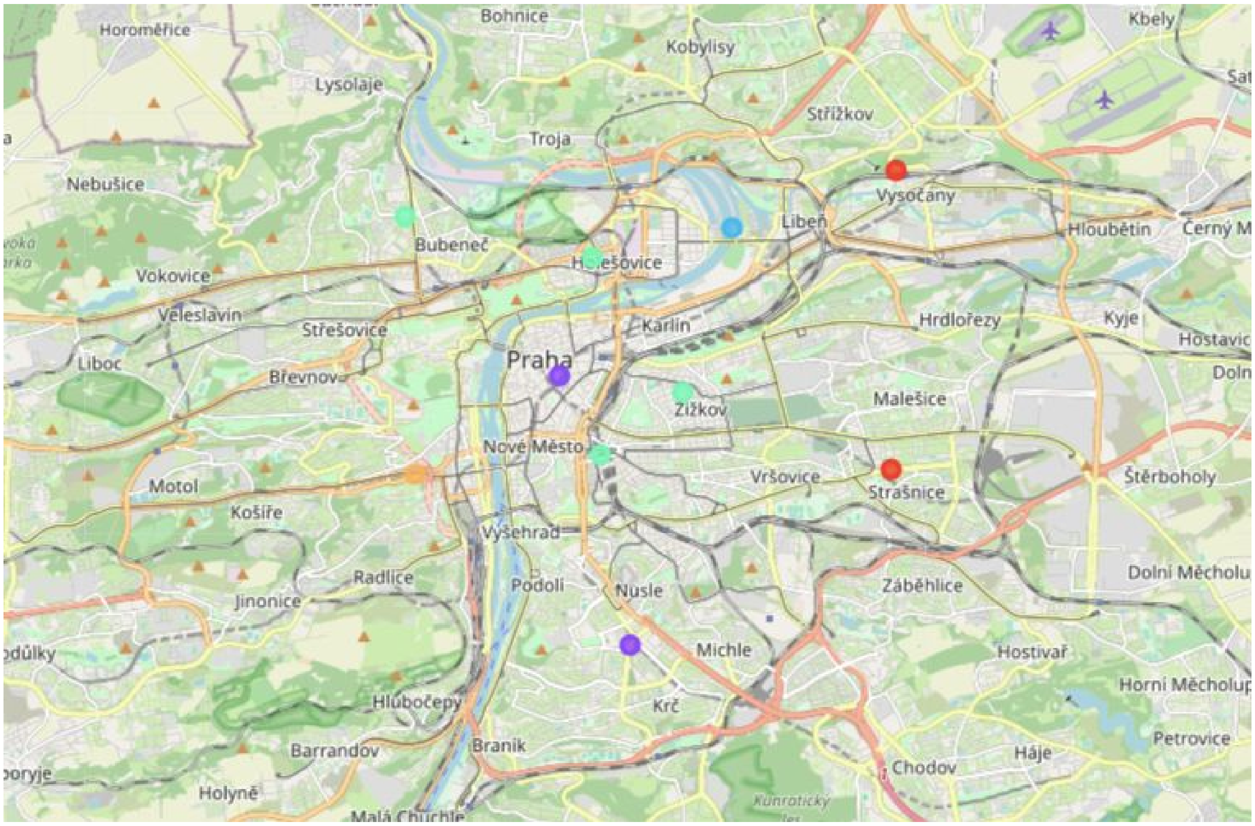


	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Praha 1	Italian Restaurant	Theater	Hotel	Cosmetics Shop	Clothing Store	Café	Jazz Club	Indie Movie Theater	Ice Cream Shop	Food Court
1	Praha 10	Restaurant	Bakery	Food & Drink Shop	Park	Pizza Place	Gastropub	Burger Joint	Skate Park	Bowling Alley	Drugstore
2	Praha 2	Café	Vegetarian / Vegan Restaurant	Burger Joint	Bar	Beer Bar	Wine Bar	Bistro	Vietnamese Restaurant	Escape Room	Italian Restaurant
3	Praha 3	Café	Bar	Cocktail Bar	Pub	Restaurant	Beer Store	Kebab Restaurant	Performing Arts Venue	Gym	Gym Pool
4	Praha 4	Clothing Store	Vegetarian / Vegan Restaurant	Bakery	Gym / Fitness Center	Cosmetics Shop	Coffee Shop	Café	Women's Store	Pet Store	Bistro

Everything is thus set for clustering. We will choose to divide the Prague neighbourhoods into 5 clusters. Below is the dataset with Cluster labels and a map with each neighbourhood by color of its cluster is attached.



	District	Postal Code	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Praha 1	110 00	50.085415	14.425401	1	Italian Restaurant	Theater	Hotel	Cosmetics Shop	Clothing Store	Café	Jazz Club	Indie Movie Theater	Ice Cream Shop
1	Praha 2	120 00	50.074726	14.433938	3	Café	Vegetarian / Vegan Restaurant	Burger Joint	Bar	Beer Bar	Wine Bar	Bistro	Vietnamese Restaurant	Escape Room
2	Praha 3	130 00	50.083118	14.451298	3	Café	Bar	Cocktail Bar	Pub	Restaurant	Beer Store	Kebab Restaurant	Performing Arts Venue	Gym
3	Praha 4	140 00	50.049204	14.440276	1	Clothing Store	Vegetarian / Vegan Restaurant	Bakery	Gym / Fitness Center	Cosmetics Shop	Coffee Shop	Café	Women's Store	Pet Store
4	Praha 5	150 00	50.072087	14.395063	4	Pub	Grocery Store	Bistro	Park	Bus Stop	Gym	Laser Tag	Event Space	Electronics Store



Results & Discussion

We have obtained the requested clusters of Prague neighbourhoods. As can be seen the clusters are not evenly distributed:

Cluster 0: **Prague 10, Prague 9**

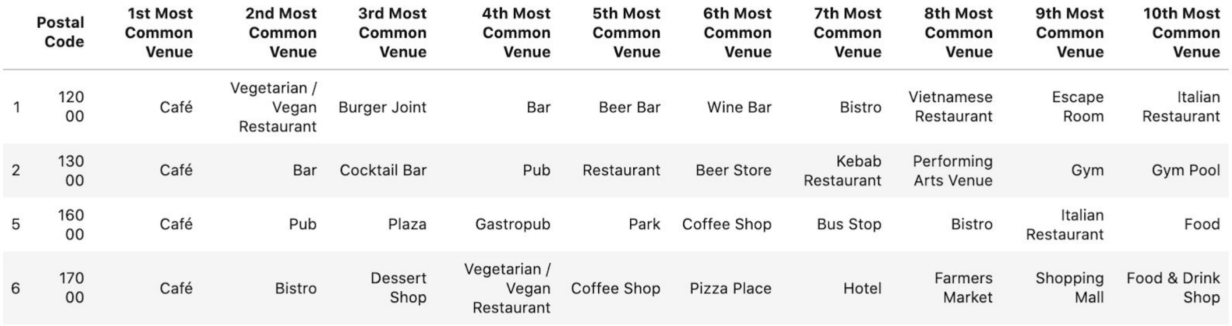
Cluster 1: **Prague 1, Prague 4**

Cluster 2: **Prague 8**

Cluster 3: **Prague 2, Prague 3**, *Prague 6*, **Prague 7**

Cluster 4: **Prague 5**

The findings are interesting as the most central neighbourhood is more like Prague 4 and not like Prague 2, 3, 6 and 7 which are the most gentrified and young neighbourhoods. These are thus the neighbourhoods where our start-up should move if they want to choose a neighbourhood which is similar to their current location - Prague 6. As can be seen below these neighbourhoods have bars, pubs and bistros as their most common venues. Thus moving to one of these neighbourhoods the company should be able to attract similar employees as they did until now to fit their corporate culture.



	Postal Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	120 00	Café	Vegetarian / Vegan Restaurant	Burger Joint	Bar	Beer Bar	Wine Bar	Bistro	Vietnamese Restaurant	Escape Room	Italian Restaurant
2	130 00	Café	Bar	Cocktail Bar	Pub	Restaurant	Beer Store	Kebab Restaurant	Performing Arts Venue	Gym	Gym Pool
5	160 00	Café	Pub	Plaza	Gastropub	Park	Coffee Shop	Bus Stop	Bistro	Italian Restaurant	Food
6	170 00	Café	Bistro	Dessert Shop	Vegetarian / Vegan Restaurant	Coffee Shop	Pizza Place	Hotel	Farmers Market	Shopping Mall	Food & Drink Shop

While examining the most similar and dissimilar neighbourhoods in Prague we have been using the Foursquare API data. With other datasets, that can be more specific, or even containing different information, we might get a bit different results. However, for the needs of our capstone project we have chosen this method and this dataset that despite its simplicity has led to interesting results.

Conclusion

We have obtained a dataset containing data about Prague venues. Thanks to k-means clustering we have been able to help a Prague start-up company to find a new location for their office. We recommend them to move to either Prague 2, 3 or 7. Because these neighbourhoods are the most similar to their current location - Prague 6. Thanks to the fact that the company has a specific corporate culture, this should help them to attract similar employees as for the current location of the office.