Coursera Capstone Project: Applied Data Science

Viktoriia Ilina

March 5th, 2021

Introduction/Business problem

Prague is centuries of spires, centuries of magic and centuries of intrigue. The city's labyrinth of cobbled lanes and hidden, winding courtyards is a treasure trove of delight any aimless wanderer would love to explore. Actually, according to Euromonitor's annual survey 2019 Prague is the sixth most visited city in Europe after London, Paris , Istanbul, Antalya and Rome[1]. Given its relatively small size and populace compared to those giants, this is an impressive achievement. Art, culture and history play a large part in this popularity – as well as the excellent travel deals – but so too does its cuisine.

However, contemporary Czech cuisine is considered heavy and very filling, with mealscentered on meats and starches. As a result some people may prefer other cuisines for health- related, religious, cultural or moral reasons. Besides, local residents may be looking to taste something new.

Thus, the present paper aims to give a simple recommendation: in which district of the city will you find a large number or even concentration of which types of restaurants? Where to eat Mediterranean food, where to find Vietnamese restaurants, where to get Sushi? The target audience is both foreign tourists and local residents.

Description of the data

Required data has been gathered from two sources: https://foursquare.com/ and https://www.praha.eu/.

Foursquare is a location technology platform offering business solutions and consumer products through a deep understanding of location[2]. This platform lets users search forrestaurants, nightlife spots, shops and other places in a location. This paper considers followingfoursquare data about restaurants in Prague: the restaurant name, ID, location and category offood.

Praha.eu is the official tourist website for Prague. This portal reveals detailed information about 22 administrative districts and 57 municipal parts of Prague. To simplify the further analysis, the data for each district was combined by the author into one table using Excel.

This gathered data will be used for showing the district density of restaurants.

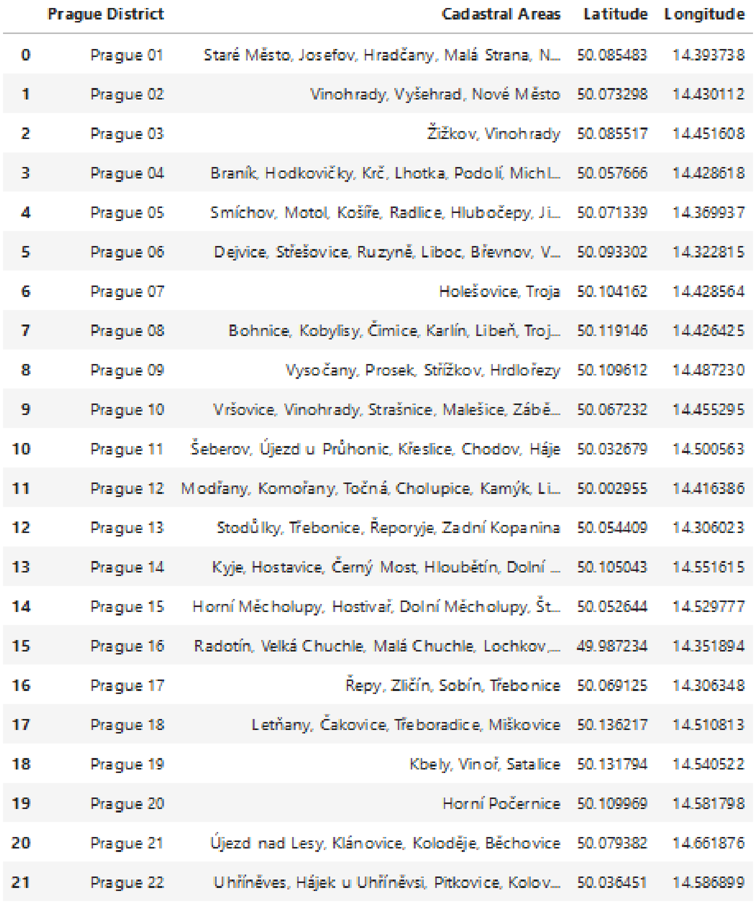
Methodology

First we need to install and import all required libraries and packages, such as Pandas, Numpy, Folium, Sklearn, Seaborn, Yellowbrick and so on.

[1] https://go.euromonitor.com/white-paper-travel-2019-100-cities.html

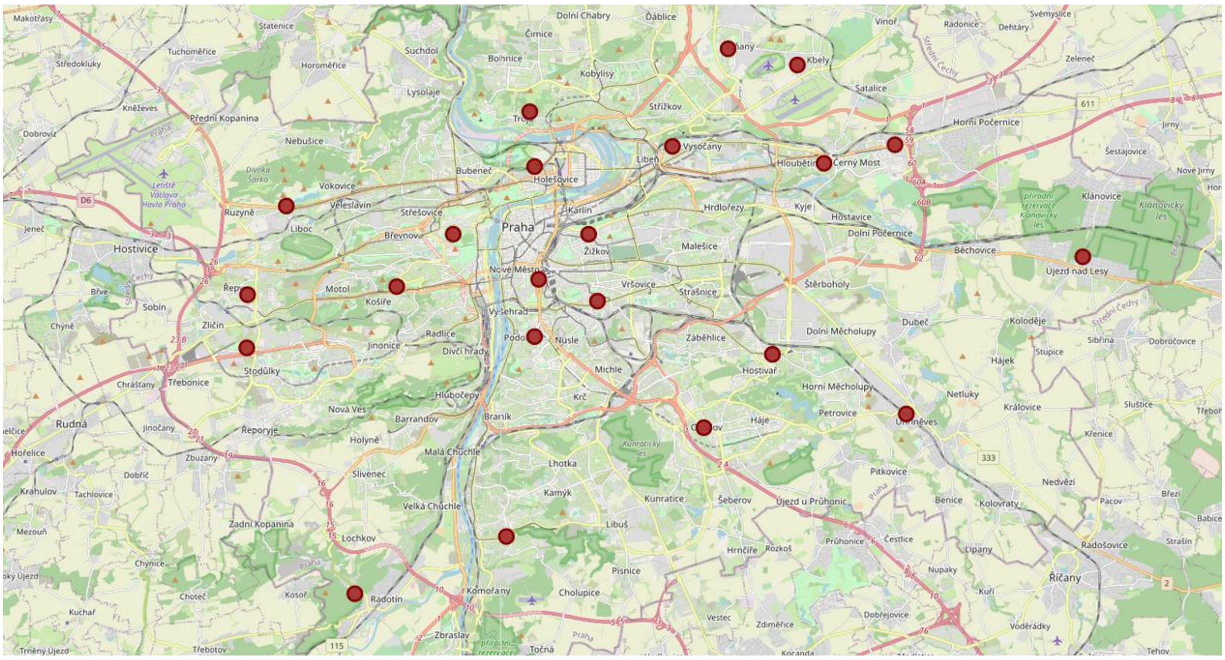
[2] https://www.crunchbase.com/organization/foursquare

Then we import excel file that contains districts data set using Pandas:



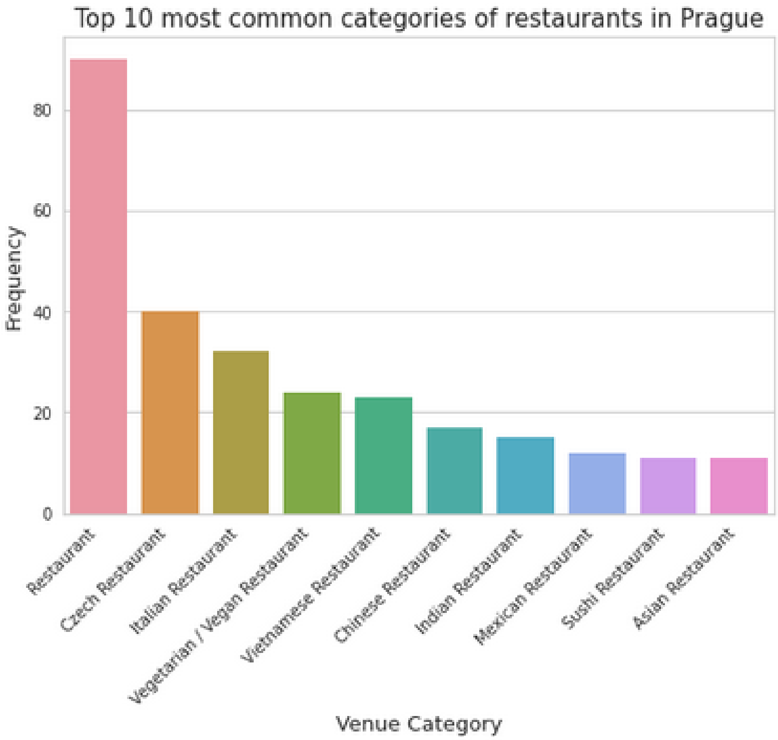
	Prague District	Cadastral Areas	Latitude	Longitude
0	Prague 01	Staré Město, Josefov, Hradčany, Malá Strana, N...	50.085483	14.393738
1	Prague 02	Vinohrady, Vyšehrad, Nové Město	50.073298	14.430112
2	Prague 03	Žižkov, Vinohrady	50.085517	14.451608
3	Prague 04	Braník, Hodkovičky, Krč, Lhotka, Podolí, Michl...	50.057666	14.428618
4	Prague 05	Smíchov, Motol, Košíře, Radlice, Hlubočepy, Ji...	50.071339	14.369937
5	Prague 06	Dejvice, Střešovice, Ruzyně, Liboc, Břevnov, V...	50.093302	14.322815
6	Prague 07	Holešovice, Troja	50.104162	14.428564
7	Prague 08	Bohnice, Kobylisy, Čimice, Karlín, Libeň, Troj...	50.119146	14.426425
8	Prague 09	Vysočany, Prosek, Střížkov, Hrdlořezy	50.109612	14.487230
9	Prague 10	Vršovice, Vinohrady, Strašnice, Malešice, Zábě...	50.067232	14.455295
10	Prague 11	Šeberov, Újezd u Průhonic, Křeslice, Chodov, Háje	50.032679	14.500563
11	Prague 12	Modřany, Komořany, Točná, Cholupice, Kamýk, Li...	50.002955	14.416386
12	Prague 13	Stodůlky, Třebonice, Řeporyje, Zadní Kopanina	50.054409	14.306023
13	Prague 14	Kyje, Hostavice, Černý Most, Hloubětín, Dolní ...	50.105043	14.551615
14	Prague 15	Horní Měcholupy, Hostivař, Dolní Měcholupy, Št...	50.052644	14.529777
15	Prague 16	Radotín, Velká Chuchle, Malá Chuchle, Lochkov,...	49.987234	14.351894
16	Prague 17	Řepy, Zličín, Sobín, Třebonice	50.069125	14.306348
17	Prague 18	Letňany, Čakovice, Třeboradice, Miškovice	50.136217	14.510813
18	Prague 19	Kbely, Vinoř, Satalice	50.131794	14.540522
19	Prague 20	Horní Počernice	50.109969	14.581798
20	Prague 21	Újezd nad Lesy, Klánovice, Koloděje, Běchovice	50.079382	14.661876
21	Prague 22	Uhříněves, Hájek u Uhříněvsi, Pitkovice, Kolov...	50.036451	14.586899

Using the folium library we visualize geographic details of Prague and its 22 city districts:



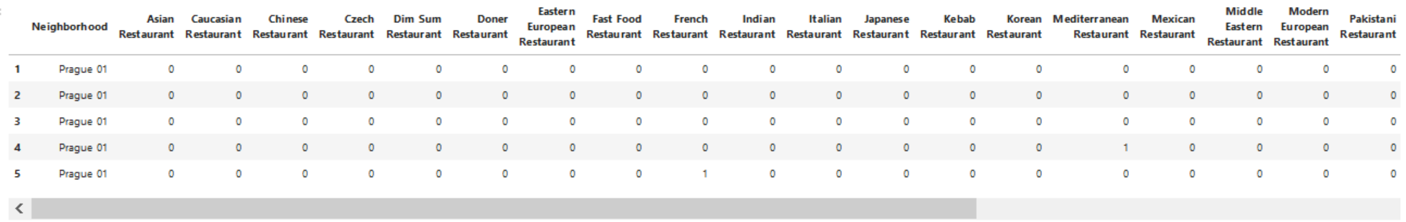
! Significant problem! : In case if you use Jupyter Notebook, Folium doesn't display special characters in tooltips (Czech, Danish, Polish and so on).

Now, foursquare data comes into play. We retrieve the foursquare data for all venues on foursquare with a distance of less than 3000 meters from each center of each city district. As a result, we have a list of 2025 venues all over Prague, including 335 restaurants come from 29 unique restaurant categories. Using Seaborn/Matplotlib packages, we visualize top 10 most common categories of restaurants in Prague:



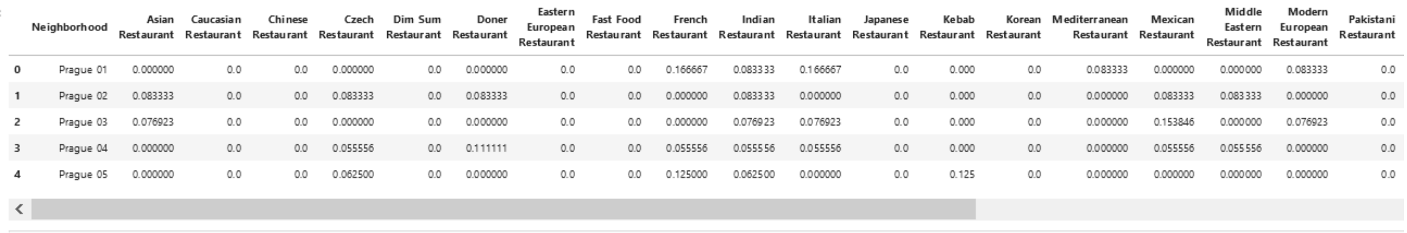
Our story just took a completely unexpected turn, the most frequent type of restaurants in Prague is restaurant with fusion cuisine!

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction[3]. For the K-means clustering algorithm, all unique items under venue category are one-hot encoded:



	Neighborhood	Asian Restaurant	Caucasian Restaurant	Chinese Restaurant	Czech Restaurant	Dim Sum Restaurant	Doner Restaurant	Eastern European Restaurant	Fast Food Restaurant	French Restaurant	Indian Restaurant	Italian Restaurant	Japanese Restaurant	Kebab Restaurant	Korean Restaurant	Mediterranean Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Modern European Restaurant	Pakistani Restaurant
1	Prague 01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Prague 01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Prague 01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Prague 01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
5	Prague 01	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

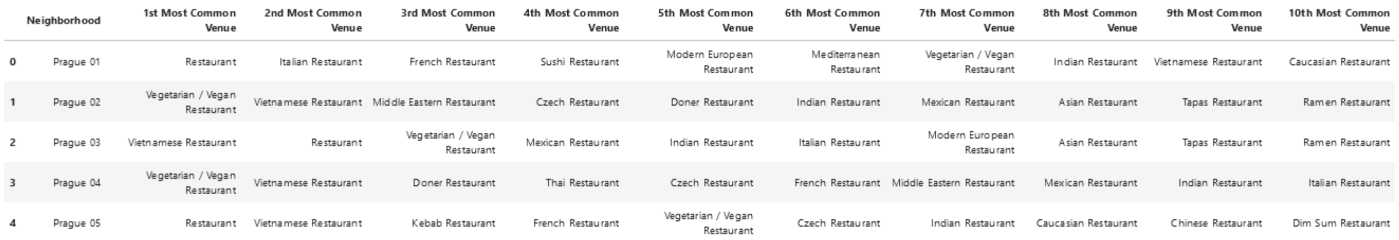
Then, we use functions groupby() and mean() to show the frequency of each category of restaurants in each city district.



	Neighborhood	Asian Restaurant	Caucasian Restaurant	Chinese Restaurant	Czech Restaurant	Dim Sum Restaurant	Doner Restaurant	Eastern European Restaurant	Fast Food Restaurant	French Restaurant	Indian Restaurant	Italian Restaurant	Japanese Restaurant	Kebab Restaurant	Korean Restaurant	Mediterranean Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Modern European Restaurant	Pakistani Restaurant
0	Prague 01	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.166667	0.083333	0.166667	0.0	0.000	0.0	0.083333	0.000000	0.000000	0.083333	0.0
1	Prague 02	0.083333	0.0	0.0	0.083333	0.0	0.083333	0.0	0.0	0.000000	0.083333	0.000000	0.0	0.000	0.0	0.000000	0.083333	0.083333	0.000000	0.0
2	Prague 03	0.076923	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.000000	0.076923	0.076923	0.0	0.000	0.0	0.000000	0.153846	0.000000	0.076923	0.0
3	Prague 04	0.000000	0.0	0.0	0.055556	0.0	0.111111	0.0	0.0	0.055556	0.055556	0.055556	0.0	0.000	0.0	0.000000	0.055556	0.055556	0.000000	0.0
4	Prague 05	0.000000	0.0	0.0	0.062500	0.0	0.000000	0.0	0.0	0.125000	0.062500	0.000000	0.0	0.125	0.0	0.000000	0.000000	0.000000	0.000000	0.0

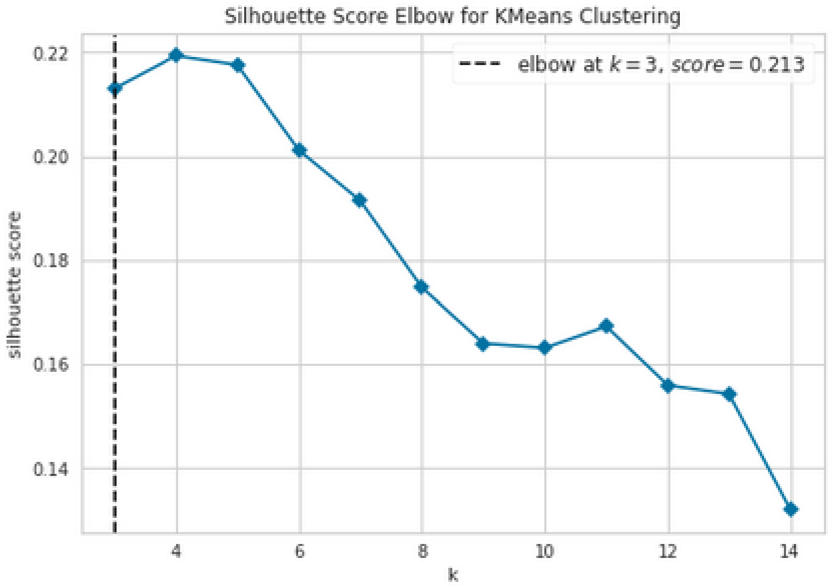
[3] https://medium.com/hackernoon/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f

After that, we create data frame with the most common restaurant venue types for each city district:



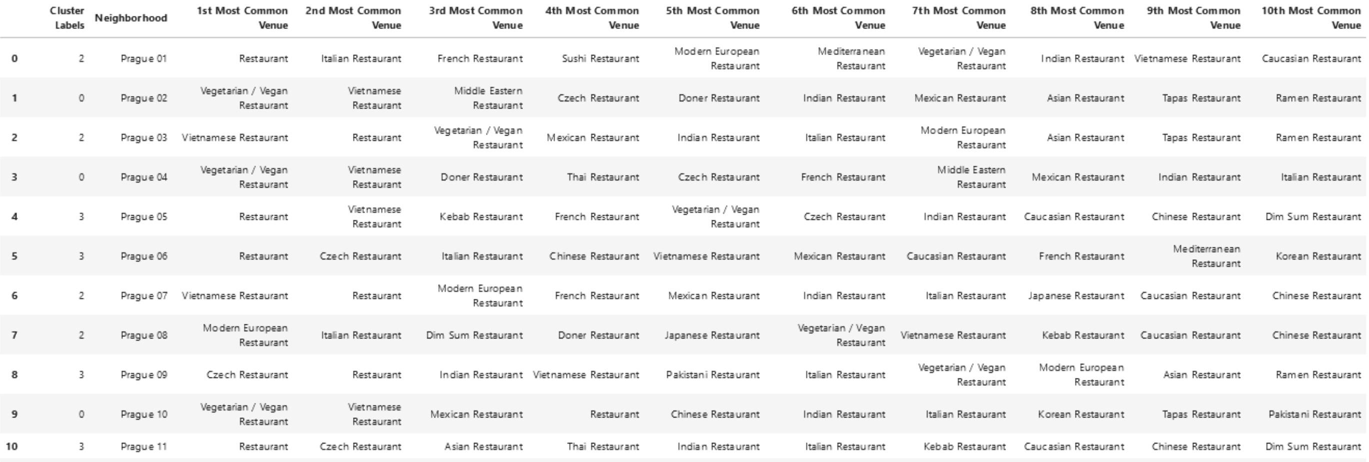
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Prague 01	Restaurant	Italian Restaurant	French Restaurant	Sushi Restaurant	Modern European Restaurant	Mediterranean Restaurant	Vegetarian / Vegan Restaurant	Indian Restaurant	Vietnamese Restaurant	Caucasian Restaurant
1	Prague 02	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Middle Eastern Restaurant	Czech Restaurant	Doner Restaurant	Indian Restaurant	Mexican Restaurant	Asian Restaurant	Tapas Restaurant	Ramen Restaurant
2	Prague 03	Vietnamese Restaurant	Restaurant	Vegetarian / Vegan Restaurant	Mexican Restaurant	Indian Restaurant	Italian Restaurant	Modern European Restaurant	Asian Restaurant	Tapas Restaurant	Ramen Restaurant
3	Prague 04	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Doner Restaurant	Thai Restaurant	Czech Restaurant	French Restaurant	Middle Eastern Restaurant	Mexican Restaurant	Indian Restaurant	Italian Restaurant
4	Prague 05	Restaurant	Vietnamese Restaurant	Kebab Restaurant	French Restaurant	Vegetarian / Vegan Restaurant	Czech Restaurant	Indian Restaurant	Caucasian Restaurant	Chinese Restaurant	Dim Sum Restaurant

Finally, we try to cluster these 22 districts based on the venue categories running an unsupervised machine learning algorithm, in particular K-means clustering algorithm, using Sklearn library. There is no golden rule of thumb for determining the appropriate value of K (if not by performing a hyperparameters' search)[4]. It largely depends on the kind of data points on which clustering is being applied. In this paper we choose to apply the ellbow method, utilizing Yellowbrick library.



As we can see on the graph, the optimal value of K is 4 (for K = 4, we get the highest average silhouette coefficient).

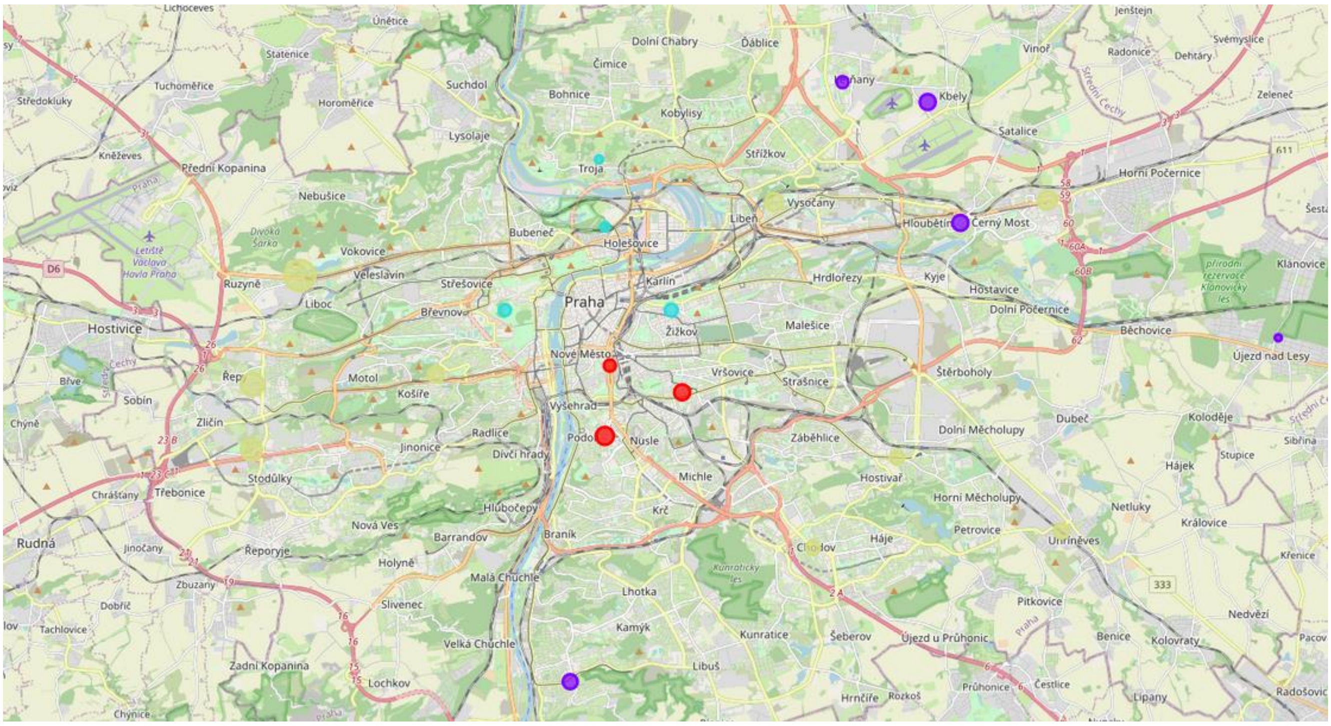
Running K-means clustering algorithm:



	Cluster Labels	Neighborhood	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	2	Prague 01	Restaurant	Italian Restaurant	French Restaurant	Sushi Restaurant	Modern European Restaurant	Mediterranean Restaurant	Vegetarian / Vegan Restaurant	Indian Restaurant	Vietnamese Restaurant	Caucasian Restaurant
1	0	Prague 02	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Middle Eastern Restaurant	Czech Restaurant	Doner Restaurant	Indian Restaurant	Mexican Restaurant	Asian Restaurant	Tapas Restaurant	Ramen Restaurant
2	2	Prague 03	Vietnamese Restaurant	Restaurant	Vegetarian / Vegan Restaurant	Mexican Restaurant	Indian Restaurant	Italian Restaurant	Modern European Restaurant	Asian Restaurant	Tapas Restaurant	Ramen Restaurant
3	0	Prague 04	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Doner Restaurant	Thai Restaurant	Czech Restaurant	French Restaurant	Middle Eastern Restaurant	Mexican Restaurant	Indian Restaurant	Italian Restaurant
4	3	Prague 05	Restaurant	Vietnamese Restaurant	Kebab Restaurant	French Restaurant	Vegetarian / Vegan Restaurant	Czech Restaurant	Indian Restaurant	Caucasian Restaurant	Chinese Restaurant	Dim Sum Restaurant
5	3	Prague 06	Restaurant	Czech Restaurant	Italian Restaurant	Chinese Restaurant	Vietnamese Restaurant	Mexican Restaurant	Caucasian Restaurant	French Restaurant	Mediterranean Restaurant	Korean Restaurant
6	2	Prague 07	Vietnamese Restaurant	Restaurant	Modern European Restaurant	French Restaurant	Mexican Restaurant	Indian Restaurant	Italian Restaurant	Japanese Restaurant	Caucasian Restaurant	Chinese Restaurant
7	2	Prague 08	Modern European Restaurant	Italian Restaurant	Dim Sum Restaurant	Doner Restaurant	Japanese Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Kebab Restaurant	Caucasian Restaurant	Chinese Restaurant
8	3	Prague 09	Czech Restaurant	Restaurant	Indian Restaurant	Vietnamese Restaurant	Pakistani Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant	Modern European Restaurant	Asian Restaurant	Ramen Restaurant
9	0	Prague 10	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Mexican Restaurant	Restaurant	Chinese Restaurant	Indian Restaurant	Italian Restaurant	Korean Restaurant	Tapas Restaurant	Pakistani Restaurant
10	3	Prague 11	Restaurant	Czech Restaurant	Asian Restaurant	Thai Restaurant	Indian Restaurant	Italian Restaurant	Kebab Restaurant	Caucasian Restaurant	Chinese Restaurant	Dim Sum Restaurant

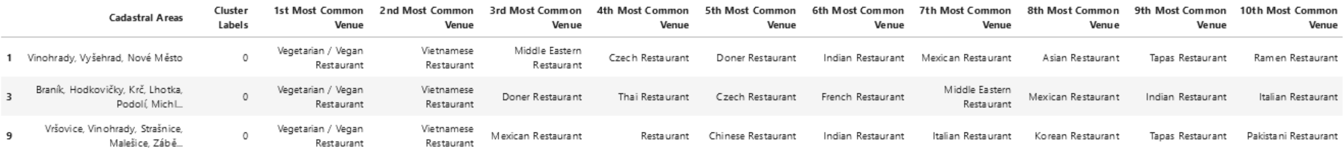
[4] https://blog.floydhub.com/introduction-to-k-means-clustering-in-python-with-scikit-learn/

Thus, we have four different cluster labels from 0 to 3 for our data set and can represent them using Folium:



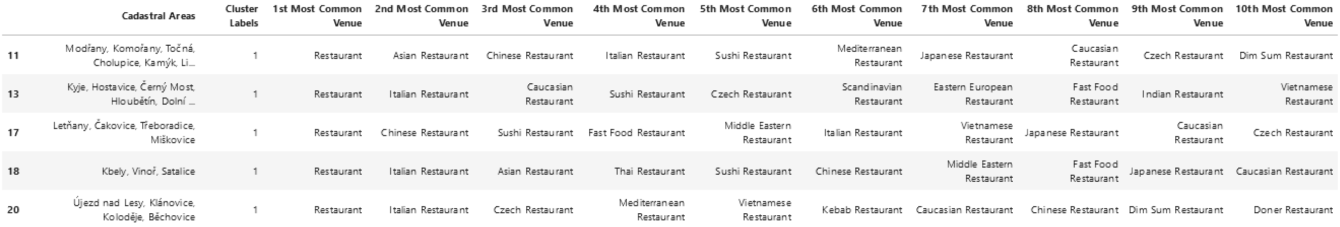
Now, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, we can then assign a name to each cluster.

The first cluster – Vegetarian/Vegan and Vietnamese restaurants



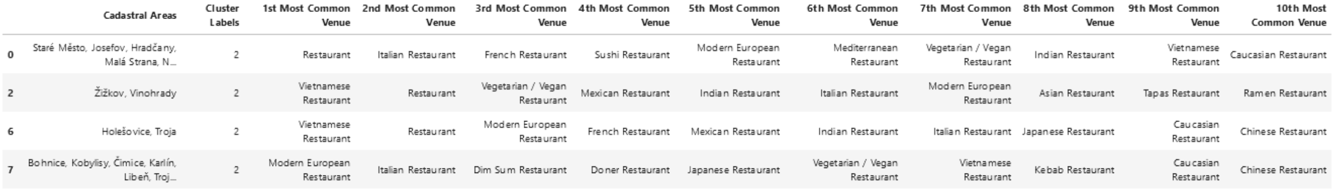
	Cadastral Areas	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Vinohrady, Vyšehrad, Nové Město	0	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Middle Eastern Restaurant	Czech Restaurant	Doner Restaurant	Indian Restaurant	Mexican Restaurant	Asian Restaurant	Tapas Restaurant	Ramen Restaurant
3	Braník, Hodkovičky, Krč, Lhotka, Podolí, Michl...	0	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Doner Restaurant	Thai Restaurant	Czech Restaurant	French Restaurant	Middle Eastern Restaurant	Mexican Restaurant	Indian Restaurant	Italian Restaurant
9	Vršovice, Vinohrady, Strašnice, Malešice, Zábě...	0	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Mexican Restaurant	Restaurant	Chinese Restaurant	Indian Restaurant	Italian Restaurant	Korean Restaurant	Tapas Restaurant	Pakistani Restaurant

The second cluster – Fusion, Italian and Asian restaurants



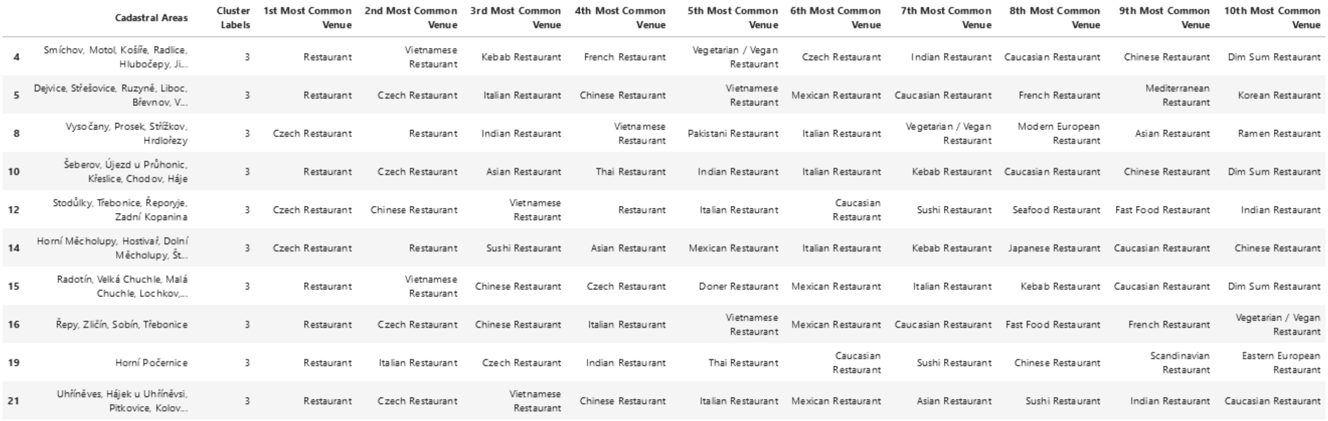
	Cadastral Areas	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
11	Modřany, Komořany, Točná, Cholupice, Kamýk, Li...	1	Restaurant	Asian Restaurant	Chinese Restaurant	Italian Restaurant	Sushi Restaurant	Mediterranean Restaurant	Japanese Restaurant	Caucasian Restaurant	Czech Restaurant	Dim Sum Restaurant
13	Kyje, Hostavice, Černý Most, Hloubětín, Dolní ...	1	Restaurant	Italian Restaurant	Caucasian Restaurant	Sushi Restaurant	Czech Restaurant	Scandinavian Restaurant	Eastern European Restaurant	Fast Food Restaurant	Indian Restaurant	Vietnamese Restaurant
17	Letňany, Čakovice, Třeboradice, Miškovice	1	Restaurant	Chinese Restaurant	Sushi Restaurant	Fast Food Restaurant	Middle Eastern Restaurant	Italian Restaurant	Vietnamese Restaurant	Japanese Restaurant	Caucasian Restaurant	Czech Restaurant
18	Kbely, Vinoř, Satalice	1	Restaurant	Italian Restaurant	Asian Restaurant	Thai Restaurant	Sushi Restaurant	Chinese Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Japanese Restaurant	Caucasian Restaurant
20	Újezd nad Lesy, Klánovice, Koloděje, Běchovice	1	Restaurant	Italian Restaurant	Czech Restaurant	Mediterranean Restaurant	Vietnamese Restaurant	Kebab Restaurant	Caucasian Restaurant	Chinese Restaurant	Dim Sum Restaurant	Doner Restaurant

The third cluster – Fusion, European and Vietnamese restaurants



	Cadastral Areas	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	10th Most Common Venue
0	Staré Město, Josefov, Hradčany, Malá Strana, N...	2	Restaurant	Italian Restaurant	French Restaurant	Sushi Restaurant	Modern European Restaurant	Mediterranean Restaurant	Vegetarian / Vegan Restaurant	Indian Restaurant	Vietnamese Restaurant	Caucasian Restaurant
2	Žižkov, Vinohrady	2	Vietnamese Restaurant	Restaurant	Vegetarian / Vegan Restaurant	Mexican Restaurant	Indian Restaurant	Italian Restaurant	Modern European Restaurant	Asian Restaurant	Tapas Restaurant	Ramen Restaurant
6	Holešovice, Troja	2	Vietnamese Restaurant	Restaurant	Modern European Restaurant	French Restaurant	Mexican Restaurant	Indian Restaurant	Italian Restaurant	Japanese Restaurant	Caucasian Restaurant	Chinese Restaurant
7	Bohnice, Kobylisy, Čimice, Karlín, Libeň, Troj...	2	Modern European Restaurant	Italian Restaurant	Dim Sum Restaurant	Doner Restaurant	Japanese Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Kebab Restaurant	Caucasian Restaurant	Chinese Restaurant

The fourth cluster – Fusion and Czech restaurants



	Cadastral Areas	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	10th Most Common Venue
4	Smíchov, Motol, Košíře, Radlice, Hlubočepy, Ji...	3	Restaurant	Vietnamese Restaurant	Kebab Restaurant	French Restaurant	Vegetarian / Vegan Restaurant	Czech Restaurant	Indian Restaurant	Caucasian Restaurant	Chinese Restaurant	Dim Sum Restaurant
5	Dejvice, Střešovice, Ruzyně, Liboc, Břevnov, V...	3	Restaurant	Czech Restaurant	Italian Restaurant	Chinese Restaurant	Vietnamese Restaurant	Mexican Restaurant	Caucasian Restaurant	French Restaurant	Mediterranean Restaurant	Korean Restaurant
8	Vysočany, Prosek, Střížkov, Hrdlořezy	3	Czech Restaurant	Restaurant	Indian Restaurant	Vietnamese Restaurant	Pakistani Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant	Modern European Restaurant	Asian Restaurant	Ramen Restaurant
10	Šeberov, Újezd u Průhonic, Křeslice, Chodov, Háje	3	Restaurant	Czech Restaurant	Asian Restaurant	Thai Restaurant	Indian Restaurant	Italian Restaurant	Kebab Restaurant	Caucasian Restaurant	Chinese Restaurant	Dim Sum Restaurant
12	Stodůlky, Třebonice, Řeporyje, Zadní Kopanina	3	Czech Restaurant	Chinese Restaurant	Vietnamese Restaurant	Restaurant	Italian Restaurant	Caucasian Restaurant	Sushi Restaurant	Seafood Restaurant	Fast Food Restaurant	Indian Restaurant
14	Horní Měcholupy, Hostivař, Dolní Měcholupy, Št...	3	Czech Restaurant	Restaurant	Sushi Restaurant	Asian Restaurant	Mexican Restaurant	Italian Restaurant	Kebab Restaurant	Japanese Restaurant	Caucasian Restaurant	Chinese Restaurant
15	Radotín, Velká Chuchle, Malá Chuchle, Lochkov,...	3	Restaurant	Vietnamese Restaurant	Chinese Restaurant	Czech Restaurant	Doner Restaurant	Mexican Restaurant	Italian Restaurant	Kebab Restaurant	Caucasian Restaurant	Dim Sum Restaurant
16	Řepy, Zličín, Sobín, Třebonice	3	Restaurant	Czech Restaurant	Chinese Restaurant	Italian Restaurant	Vietnamese Restaurant	Mexican Restaurant	Caucasian Restaurant	Fast Food Restaurant	French Restaurant	Vegetarian / Vegan Restaurant
19	Horní Počernice	3	Restaurant	Italian Restaurant	Czech Restaurant	Indian Restaurant	Thai Restaurant	Caucasian Restaurant	Sushi Restaurant	Chinese Restaurant	Scandinavian Restaurant	Eastern European Restaurant
21	Uhříněves, Hájek u Uhříněvsi, Pitkovice, Kolov...	3	Restaurant	Czech Restaurant	Vietnamese Restaurant	Chinese Restaurant	Italian Restaurant	Mexican Restaurant	Asian Restaurant	Sushi Restaurant	Indian Restaurant	Caucasian Restaurant

Results/Discussion

A compressed overview of Prague's restaurants introduces multiple useful application for travellers as well as businesses looking for new opportunities. Let's summarize our findings:

- There are 335 restaurants come from 29 unique restaurant categories in Prague.
- The most frequent type of restaurants in Prague is restaurant with fusion cuisine!
- Prague 6 and Prague 17 have maximum number of restaurants (31 and 23 accordingly).
- Prague 8 and Prague 21 have the least number of restaurants (8 and 7 respectively).
- The Prague districts are divided into 4 clusters.

The clustering is completely based on the most common venues obtained from Foursquare data.

Since the scale of the current study does not imply gathering excessively vast and detailed data sets, so certain parameters got omitted and thus the analysis ignores various other factors, such as the location's remoteness form transport stations, price ranges, and Michelin-starred restaurants, etc. The analysis then, targets at helping travelers get a quick outlook at the distributions of restaurants across 22 Prague's districts, sorted by their categories.

Furthermore, this results also could potentially vary if we use some other clustering techniques like Expectation–Maximization Clustering using Gaussian Mixture Models or Density-Based Spatial Clustering of Applications with Noise.

Conclusion

Data is a nowadays' key to finding solutions to various life situations – in regular life and unexpected occurrences as well. As for the dissected example, data made possible clustering the surrounding in Prague in terms of common food services across 22 districts of the city. The implications can come out useful, for instance, for travelers trying to pick the one district that fits their requirements or preferences the most.