Capstone Project - The Battle of Neighborhoods

Applied Data Science Capstone by IBM/Coursera

by Jan Korinek

November 18, 2019

1. Table of contents

2. Introduction

In this final capstone project I will try to find several locations in two different cities based on specified criteria.

This analysis is suitable for the situation, when person needs to move from city A to city B and parallelly wants to keep its living habits. Main goal is then to find districts, suburbs or boroughs in each city with mutually similar characteristics. Some of the cities presented in this analysis are intentionally chosen from different geographical environments and hasn't the same size of population. Intention is to find out if those aspects has significant impact on amount of resulting locations. Analyzed cities are following:

- **Prague** - capital of Czech Republic with population of 1,3 mil. (2019) located in central Europe and founded in 8th century AD.
- **Sydney** - city of Australia with population of 4,6 mil. (2011) located in New South Wales and founded in 18th century AD.

Conditions which all places has to meet are defined based on individual preference. Idea is to maintain healthy lifestyle, minimize time necessary for routine activities like traveling to job or shopping and dedicate it more to fitness activities in gyms or parks or spend more time on cultural events.

Based on that, targeted location should be covered by dense net of public transport. It should include shopping malls not far located and also parks and gyms approachable within several minutes of walking.

With help of Foursquare platform geolocation data and data science approaches, we will get as a result list of appropriate locations for each city with mutual comparison.

3. Data

Inputs used in this type of study falls into three categories.

As a first step, it’s necessary to obtain list of districts or neighborhoods of particular city. Goal is to get at least 100 locations for analysis across the city as input. In case of Prague, list of Prague districts is sufficient. List of inner suburbs is used in case of Sydney.

Next type of data are coordinates assigned to each address from the lists of suburbs or districts by GeoPy. This is used as input for function localizing venues at Foursquare. Last category is then Foursquare location data filtered according definition of the problem like public transport stations, shopping plazas, pharmacies and parks. Venues located in each district or suburb are then used for further processing and evaluation.

4. Methodology

Methodology scheme is very simple and for analyzing each city it's applied in same the way.

First part is focused on import districts/suburbs data their update in way, that they are possible to visualize on Folium map. This leads to second step where city districts/suburbs are extended about venues data via Foursquare API according defined conditions and prepared for clustering process. Third part is focused on usage of K-Means algorithm, therefore finding of the best k parameter is done and then all venues are clustered. Finally, last step is focusing on extraction of districts/suburbs from clusters which are matching the best to defined criteria.

For detailed commentary on methodology, please move forward to analysis section below. There you can find further explanation of each step done in analysis and factors which leads to that. This is done completely for Prague city analysis. Due to fact that overall methodology is always the same, for Sydney are commented only steps which differentiate from method applied in case of Prague.

5. Analysis

5.1 Prague - Czech Republic

Dataset Import and Modification

As a first step, we need to create dataset containing names of the Prague districts. This is done by importing of the excel file into pandas dataframe. In early stages of the script creation it was discovered that Nominatim module converts full district addresses with less error than name of the district only, therefore columns of "City" and "Country" were added into .xlsx file before import into dataframe.

Dataframe contains 112 districts. Now, we extend Prague districts it about the new column "Address" by merging data from previous three columns into the one.

Dataset Extension by Coordinates

At this moment, we can convert addresses into coordinates with help of Nominatim module and extend the dataframe. We also clean dataset about unnecessary columns like "City", "Country", "Address" and "Coordinates" and create new dataset.

Final dataframe is ready and it's possible to see on picture below.

	Prague District	Latitude	Longitude
0	Stodůlky	50.048307	14.312404
1	Žižkov	50.081054	14.454917
2	Chodov	50.032843	14.501643
3	Vinohrady	50.075359	14.436394
4	Vršovice	50.071885	14.472665
...	...	...	...
107	Lahovice	49.988587	14.397336
108	Nedvězí u Říčan	50.016467	14.653807
109	Lipany	49.999546	14.617539
110	Malá Chuchle	50.026136	14.393634
111	Zadní Kopanina	50.006452	14.312206

112 rows × 3 columns

Fig. 1 Dataframe of the locations

Prague Districts Map Creation

For better understanding of the districts location across Prague and later distribution of the clusters, we will create a map with help of Folium library. In order to define an instance of the geocoder, we need to define a user_agent. We will name our agent prg_explorer.

Now, let's create a map of Prague with districts superimposed on top. Map also shows intended radius size, which will be later used in Foursquare API to locate venues in defined distance from the center of the district. For optimal coverage even the larger districts, the radius size was determined at 1500m.

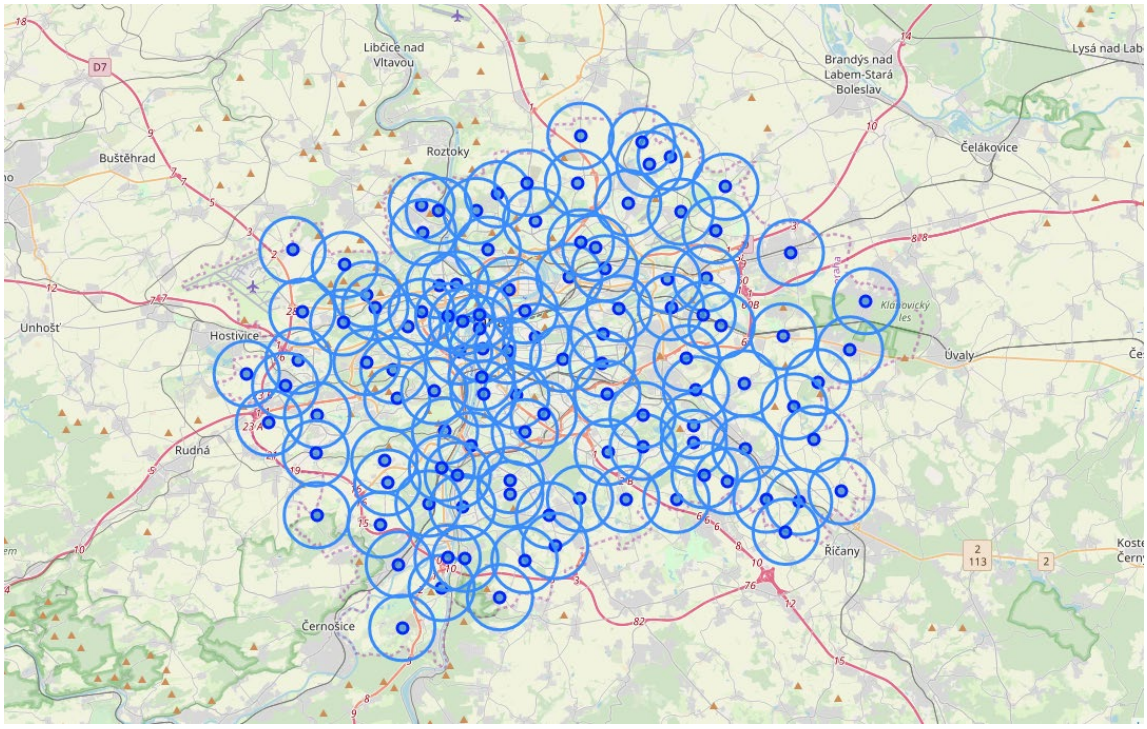

Fig. 2 Prague districts

Venues Extraction via Foursquare API

At this moment, it is necessary to create function which can extract all venues within radius of 1500m from district center location. We need to extract all possible venues, therefore limit of number of venues returned by Foursquare API is set to 1000. Whole extraction process is repeated for all defined districts within the city.

Based on this function, it’s possible to create new dataframe containing all returned venues in city districts. Each venue is assigned to corresponding district with its own name, latitude, longitude and category. This information will be important for further processing. Whole new dataframe is created by merging cleaned dataframe of Prague districts and venues returned by function.

Function returned 6092 venues in total in 359 unique categories. Now, let's filter these categories and select only those relevant for definition of our problem and put them into the new dataframe. Based on individual preference of keeping current lifestyle in any of intended cities, we will select venues related to public transport, shopping, fitness, parks and pharmacies. Full list of venue categories is possible to see below:

- Gym
- Fitness
- Park
- Bus
- Bus Stop

- Bus Station
- Mall
- Shopping Mall
- Shopping Plaza
- Metro
- Metro Station
- Metro Station and Building
- Train
- Train Station
- Pharmacy

After filtering is done, 693 venues from 6092 match our criteria.

Data Preparation for Clustering

In this section, we're going to prepare data of selected venues for K-Means clustering process. Therefore One-hot encoding and re-grouping operations of the dataset will be done.

Another step is to perform series of merging operations to get more space efficient dataframe. In case of Prague, it's possible to merge together "Bus" and "Bus Station" columns under one category called "Bus". Same process can be also done with "Shopping Mall" and "Shopping Plaza" merged under the "Shopping".

K-Means Clustering

	District	Gym	Metro Station	Park	Pharmacy	Train Station	Bus	Shopping
0	Benice	0.000000	0.0	0.500000	0.000	0.5	0.00	0.0
1	Bohnice	0.000000	0.0	0.125000	0.125	0.0	0.75	0.0
2	Braník	0.000000	0.0	0.400000	0.000	0.3	0.30	0.0
3	Bubeneč	0.333333	0.0	0.666667	0.000	0.0	0.00	0.0
4	Běchovice	0.000000	0.0	0.300000	0.000	0.2	0.50	0.0

Fig. 3 Clustered dataframe

At this point our dataframe has structure showed on picture above and itis ready to be used in clustering process. Algorithm of K-Means will be used for this purpose since is the most commonly used. Also, by time of analysis preparation, there wasn't known any reason which would implied usage of different approach. Results accuracy of clustering process is significantly dependent on which value of k parameter is used. Therefore, accuracy vs. k parameter characteristics will be printed.

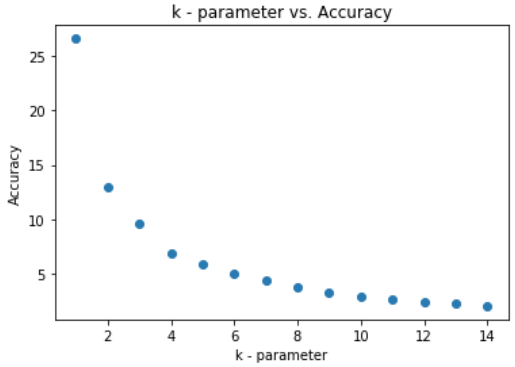

Fig. 4 k parameters dependency

To choose optimal k parameter, it's recommended to select such a value where k characteristic sharply shifts (elbow point). From diagram above, it's possible to see that elbow point lies where k parameter is 5 or 6. Based on trials where clustered data was processed with k=5 and k=6 it was concluded, that for further interpretation is more suitable to continue with k=6. Cluster labels will be then merged with PragueDistricts_cl dataframe to obtain coordinates and cluster labels for each particular city district.

Now, let's visualize Prague districts with assigned color for each cluster. This step will be beneficial later for final results evaluation.

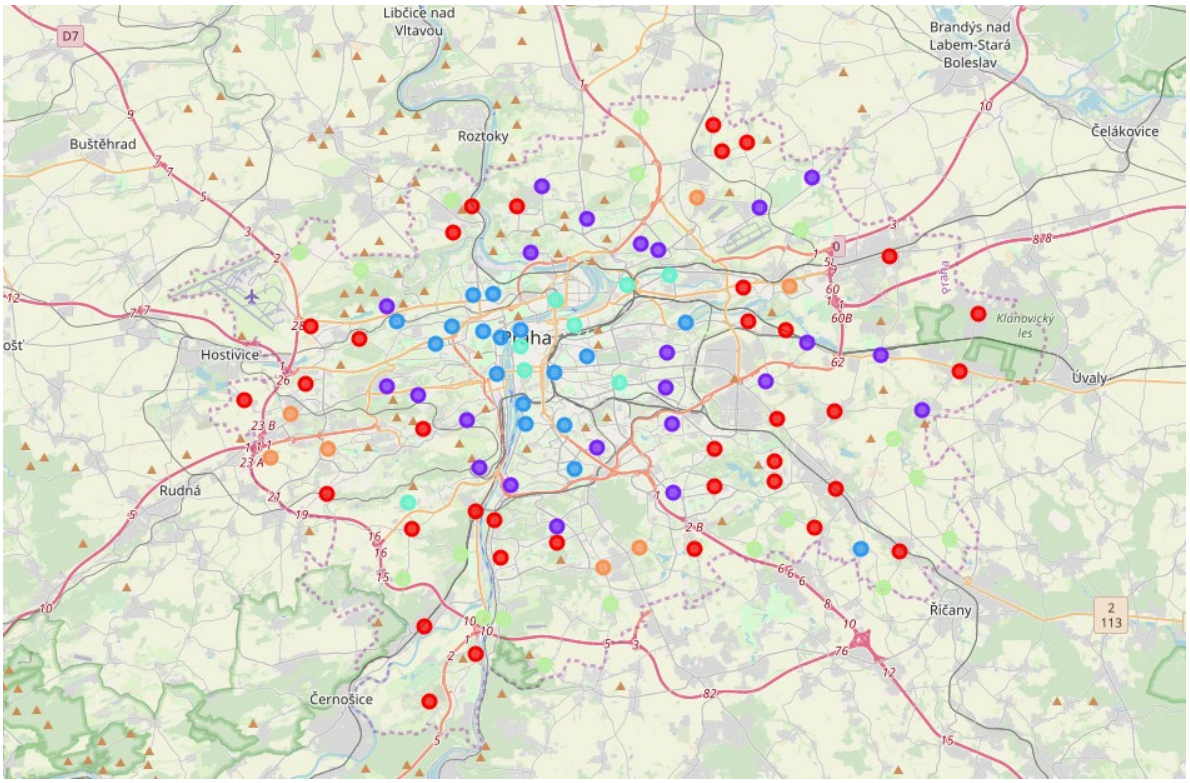

Fig. 5 Clusters distribution

Clusters Examination and Final Selection of the Locations

Now, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster.

Final selection and filtering of the locations is done based on combination of observing cluster map, prague_index dataframe where data are grouped according cluster labels and lists of locations in clusters generated in this section. Those data are then placed against selection criteria defined at the beginning of the analysis.

When we have a look at prague_index dataframe, it's possible to notice that some of the clusters doesn't contains one or more venue categories at all. This would jeopardize fulfillment of defined conditions. In case of Prague this is the case of cluster 3 and cluster 4. Therefore they're excluded from selection process.

Another selection criteria is related to mobility and availability of public transport. Preference is to select locations, where is possible to travel ideally by both bus and train. If this combination isn't possible to achieve, preferred transport is on train and therefore all locations without train station availability are excluded.

Final outcome locations and their distribution on map is possible to see on figures below.

	Prague District	Latitude	Longitude	Cluster Labels	Gym	Metro Station	Park	Pharmacy	Train Station	Bus	Shopping
7	Záběhlice	50.057282	14.501349	1	0.166667	0.0	0.250000	0.083333	0.083333	0.416667	0.000000
10	Modřany	50.009806	14.406989	0	0.100000	0.0	0.000000	0.100000	0.100000	0.700000	0.000000
19	Dejvice	50.102556	14.391797	2	0.250000	0.0	0.625000	0.000000	0.125000	0.000000	0.000000
24	Letňany	50.136969	14.514886	5	0.142857	0.0	0.142857	0.000000	0.142857	0.428571	0.142857
25	Braník	50.035728	14.412717	1	0.000000	0.0	0.400000	0.000000	0.300000	0.300000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...
77	Běchovice	50.081210	14.616026	1	0.000000	0.0	0.300000	0.000000	0.200000	0.500000	0.000000
96	Sedlec 160 00	50.133703	14.391723	0	0.117647	0.0	0.058824	0.000000	0.058824	0.705882	0.058824
101	Benice	50.012960	14.604874	2	0.000000	0.0	0.500000	0.000000	0.500000	0.000000	0.000000
103	Sobín	50.065626	14.266559	0	0.000000	0.0	0.000000	0.000000	0.333333	0.666667	0.000000
110	Malá Chuchle	50.026136	14.393634	0	0.000000	0.0	0.250000	0.083333	0.083333	0.583333	0.000000

27 rows × 11 columns

Fig. 6 Final filtered locations

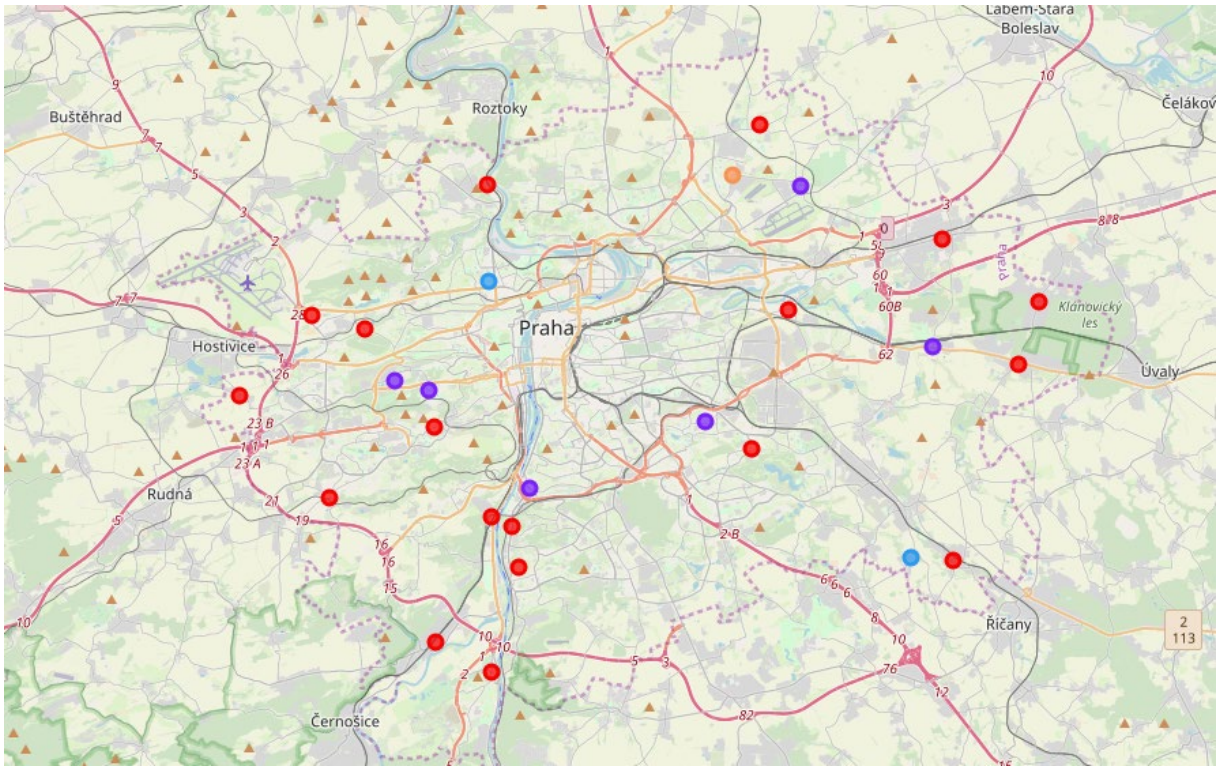

Fig. 7 Locations distribution

5.2 Sydney - New South Wales, Australia

Dataset Import and Modification

Instability issues has appeared when using Nominatim module. Therefore, load into pandas dataframes was divided into several .xlsx files where there was no more than 70 inner suburbs names.

Dataset Extension by Coordinates

At this moment, we can convert addresses into coordinates with help of Nominatim module and extend the dataframe. Final dataset is possible to see on picture below.

	Sydney Suburb	Latitude	Longitude
0	Australia Square	-33.864946	151.207793
1	Grosvenor Place	-33.741295	151.034051
3	Queen Victoria Building	-33.871435	151.206669
4	Eastern Suburbs	-33.870260	151.270226
5	Haymarket	-33.881441	151.204452
...	...	...	...
311	Wentworth Falls	-33.715639	150.369759
312	Lawson	-33.719448	150.431562
313	Bullaburra	-33.724860	150.413798
314	Blackheath	-33.633889	150.284722
315	The Ponds	-33.706667	150.909167

271 rows × 3 columns

Fig. 8 Dataframe of the locations

Sydney Inner Suburbs Map Creation

User_agent function is called syd_explorer.

In the step of the map creation, radius size was extended from 1500m to 2000m for optimal coverage of the larger districts.

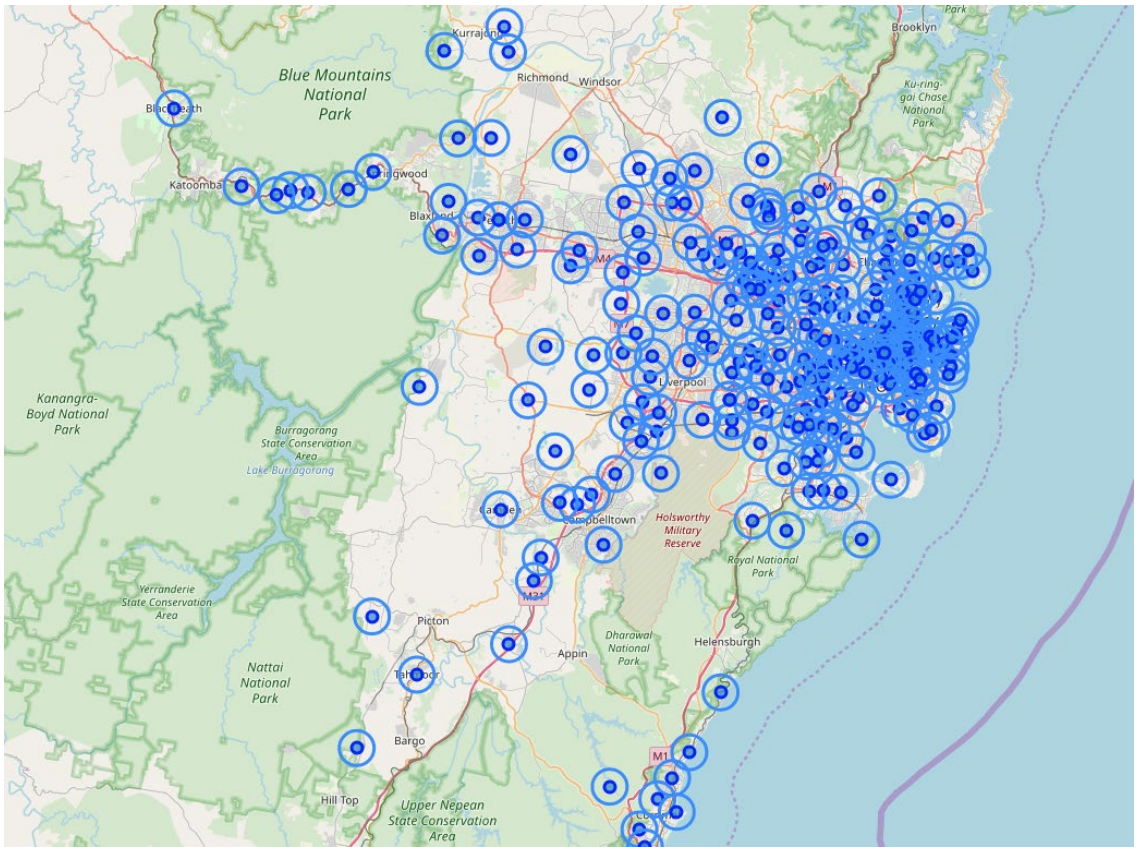

Fig. 9 Syndney inner suburbs locations

Venues Extraction via Foursquare API

We have 13478 venues within in 341 unique categories. After filtering 1127 venues match our criteria.

Data Preparation for Clustering

See Prague section of data preparation for clustering.

K-Means Clustering

	District	Gym	Park	Pharmacy	Train Station	Bus	Shopping
0	Abbotsbury	0.200000	0.600000	0.0	0.0	0.0	0.200000
1	Abbotsford	0.142857	0.857143	0.0	0.0	0.0	0.000000
2	Acacia Gardens	0.200000	0.200000	0.0	0.0	0.2	0.400000
3	Airds	0.000000	0.000000	0.0	0.0	0.0	1.000000
4	Alexandria	0.000000	0.666667	0.0	0.0	0.0	0.333333

Fig. 10 Clustered dataframe

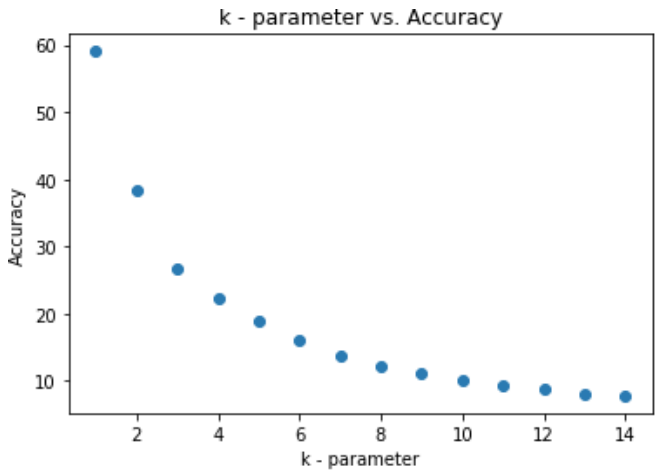

Fig. 11 k parameters dependency

Based on above characteristic, as optimal value of the k was selected at k=6.

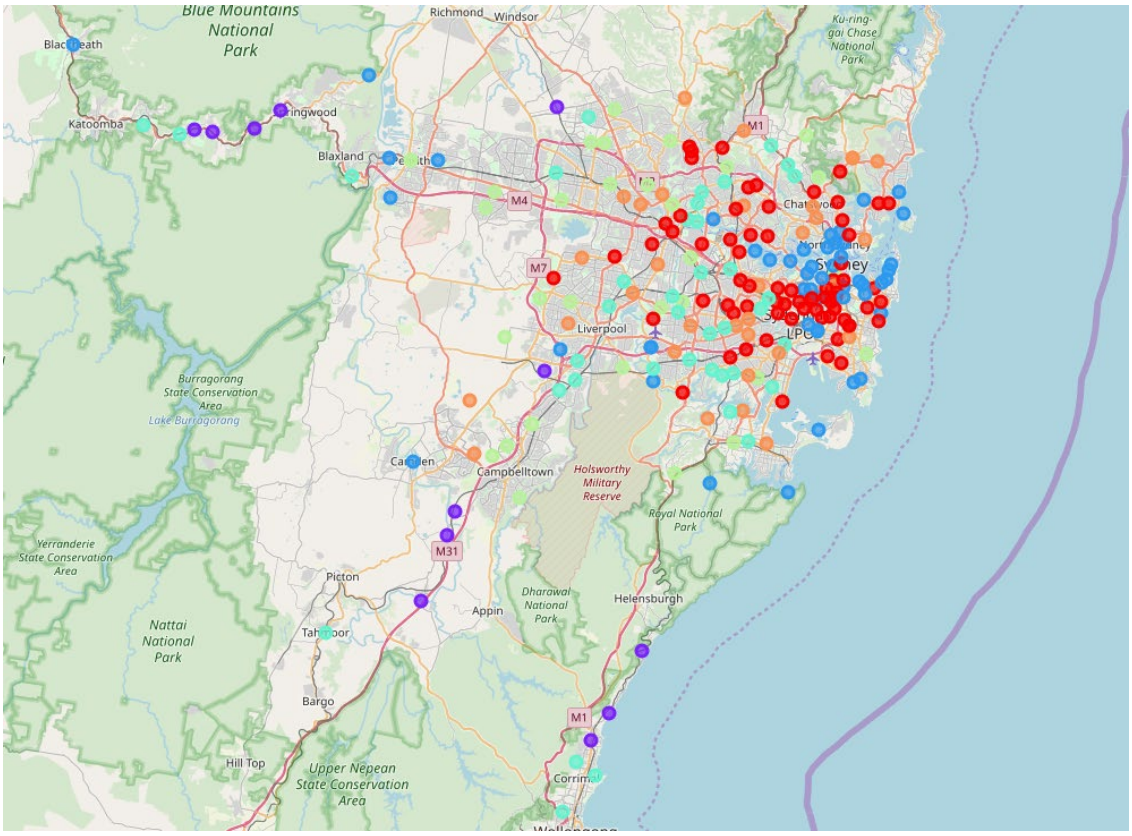

Fig. 12 Clusters distribution

Clusters Examination and Final Selection of the Locations

After clustering of locations of the Sydney It was find out that cluster 1 doesn't meet defined conditions in almost any category and due to that was excluded from further processing. As in the previous cases, the condition of preference for using public transport by train was applied here and therefore all locations without train station availability were excluded too.

Final outcome locations and their distribution on map is possible to see on figures below.

	Sydney Suburb	Latitude	Longitude	Cluster Labels	Gym	Park	Pharmacy	Train Station	Bus	Shopping
46	Rydalmere	-33.810017	151.029281	3	0.222222	0.222222	0.0	0.444444	0.111111	0.000000
114	Chatswood	-31.872494	147.500293	5	0.428571	0.142857	0.0	0.142857	0.000000	0.285714
120	Artarmon	-33.808955	151.185309	5	0.444444	0.111111	0.0	0.111111	0.000000	0.333333
123	Chatswood	-33.797481	151.180939	5	0.428571	0.142857	0.0	0.142857	0.000000	0.285714
125	Roseville	-33.782646	151.182726	0	0.250000	0.333333	0.0	0.166667	0.000000	0.250000
...	...	...	...	...	...	...	...	...	...	...
302	Doonside	-33.763689	150.869183	3	0.000000	0.400000	0.2	0.200000	0.000000	0.200000
305	Glenbrook	-33.767066	150.622492	3	0.000000	0.500000	0.0	0.500000	0.000000	0.000000
311	Wentworth Falls	-33.715639	150.369759	3	0.000000	0.500000	0.0	0.500000	0.000000	0.000000
313	Bullaburra	-33.724860	150.413798	3	0.000000	0.333333	0.0	0.666667	0.000000	0.000000
315	The Ponds	-33.706667	150.909167	3	0.000000	0.333333	0.0	0.333333	0.000000	0.333333

76 rows × 10 columns

Fig. 13 Final filtered locations

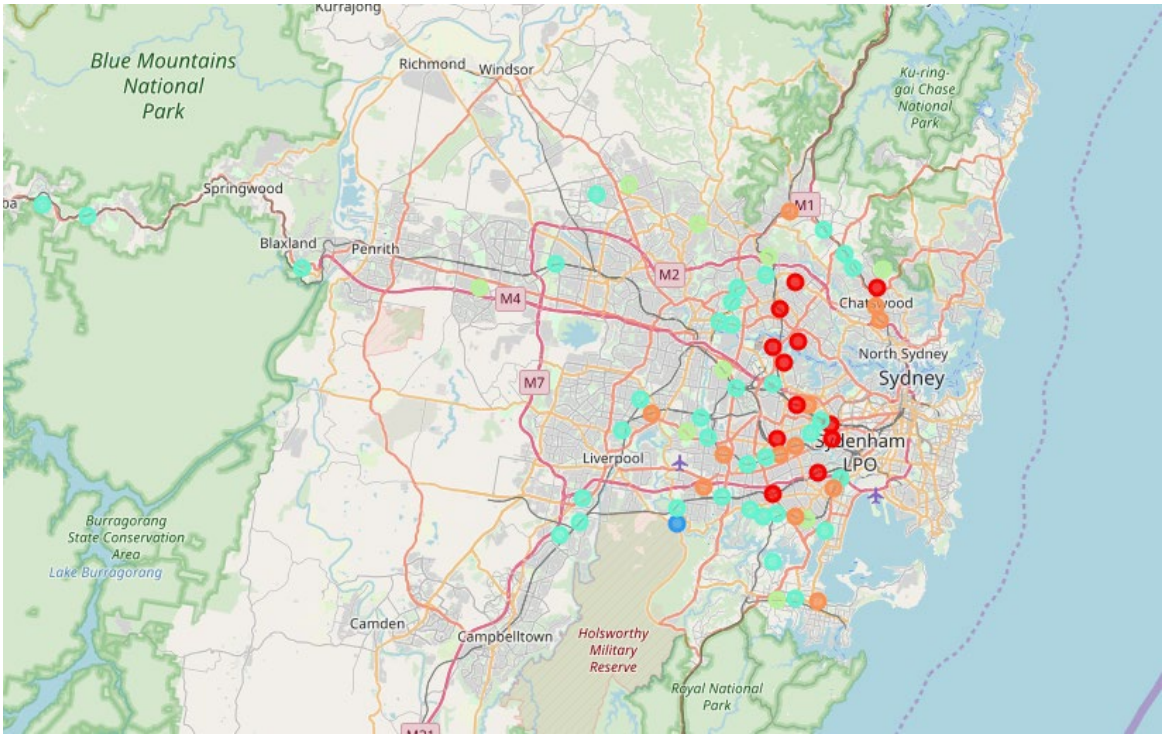

Fig. 14 Locations distribution

6. Results and Discussion

Analyses performed in section 5 gives resulting number of locations selected according defined conditions.

In case of Prague, 112 districts were used as an input and after finishing of selection process, 27 locations has been evaluated with highest potential to meet all criteria. That's 24% of all analyzed districts. Visualization shows that the clusters has more or less circular shape which separates location of historical center and approaching up to outskirts. Most of the finally selected districts are impacted by this trend and therefore lies in the circular pattern too. It is also possible to see that most of them are located near outskirts and are shaped according to train corridors.

Sydney distribution of the clusters is different from Prague. There is no clear pattern which could reflect clusters distribution, however from the visualization of the finally selected inner suburbs is possible to claim, that most of them are located parallelly to coast but keeping certain distance from it. There is also significant distance from the outskirts from inland side. Minority of inner suburb remains at the outskirt shaped clearly according to train corridors. In total, 271 inner suburbs has been analyzed and 76 has been classified to meet the criteria, that's 28%.

7. Conclusion

Purpose of this project was to find similarly specified districts, suburbs or neighborhoods across two different cities. This might come handy when person needs to move from one to another city and wants to keep current lifestyle. Conditions were setup to maintain healthy lifestyle, minimize time necessary for routine activities like traveling to job or shopping and dedicate it more to fitness activities in gyms or parks or spend more time on cultural events.

According to specified conditions, in Prague it's possible to find 27 locations for further exploration. In Sydney this number is significantly higher - 76 locations. This is due to large loaded

dataset containing inner suburbs. Ratios are however quite close 24% vs. 28%. This means that in Sydney are slightly higher chances to find those specified places. Situation will change if we decide to exclude locations without bus stops instead of locations without train station. After this, filtering return 87 location within Prague and 41 locations in Sydney. Ratios will be then 78% vs. 15%. This means that bus stops network is way denser in Prague than in Sydney.

Is it also worth to mention that resulting number of locations may vary according to precision of description of the venues obtained from Foursquare API. Some of them might be classified incorrectly like shopping malls and therefore there are unintentionally excluded.

Above presented result can be considered as starting point for deeper analysis of each location within clusters. With regard to that might appear new selection criteria like job location, traffic exploitation of the routes or location of the school for own kids. There might appear many other individual factors which could significantly reduce final number of suitable locations or find new one. This is however out of scope of this project. Despite this analysis create solid base for further extension in described way.