Capstone Project - The Battle of Neighborhoods

Applied Data Science Capstone by IBM/Coursera

by Jan Korinek

November 18, 2019

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.**
- **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.**

List of cities

- **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.

Data

- List of districts or neighborhoods of particular city
- Coordinates assigned to each address from the lists of suburbs or districts
- Foursquare location data filtered according definition of the problem

Methodology

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

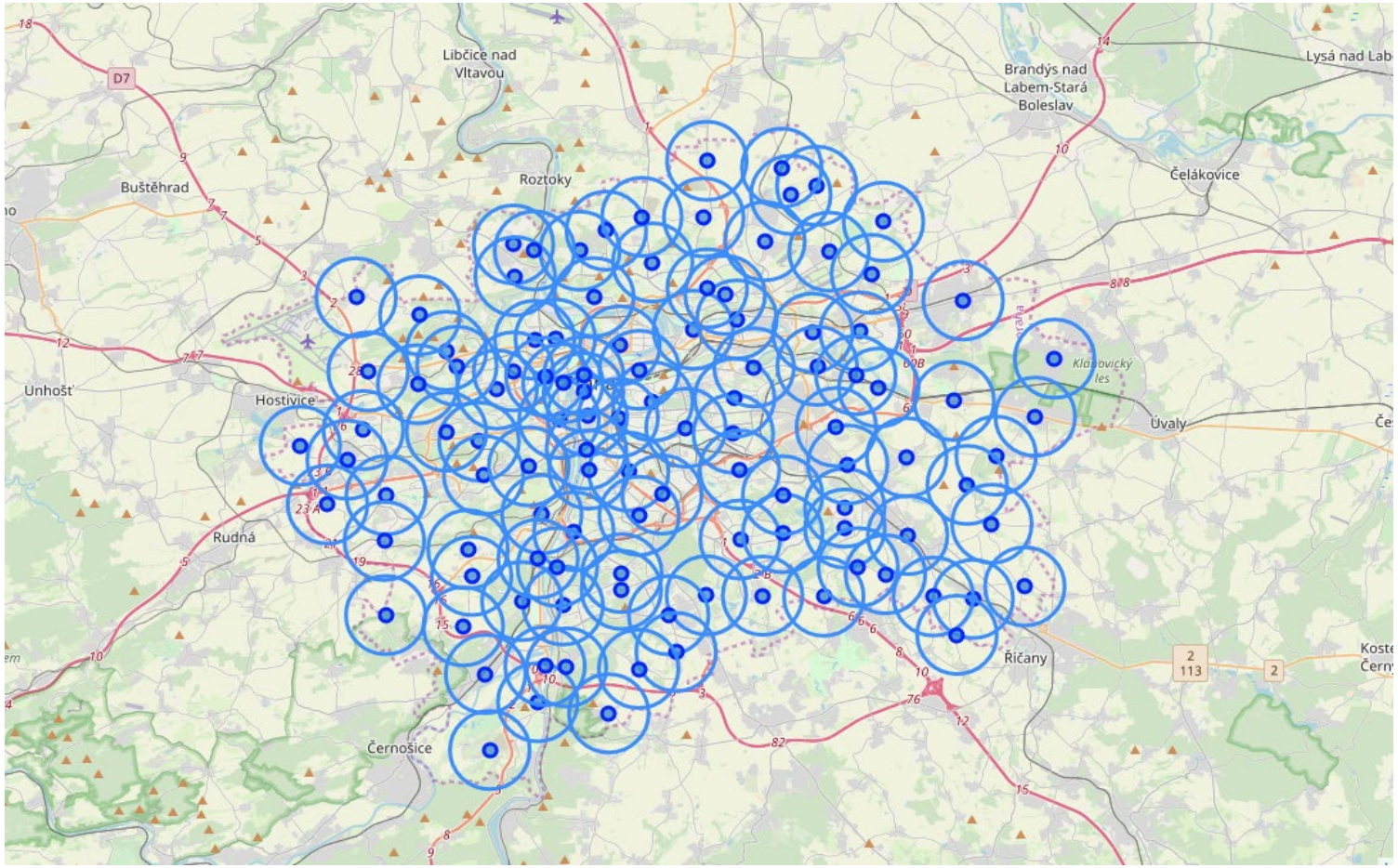
- **First part is focused on import districts/suburbs data their update in way, that they are possible to visualize on Folium map.**
- **Districts/suburbs extension about venues data via Foursquare API according defined conditions**
- **K-Means clustering**
- **Extraction of districts/suburbs from clusters which are matching the best to defined criteria.**

Analysis: Prague - Czech Republic

	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

Dataframe of the locations



Prague districts

Analysis: Prague - Czech Republic

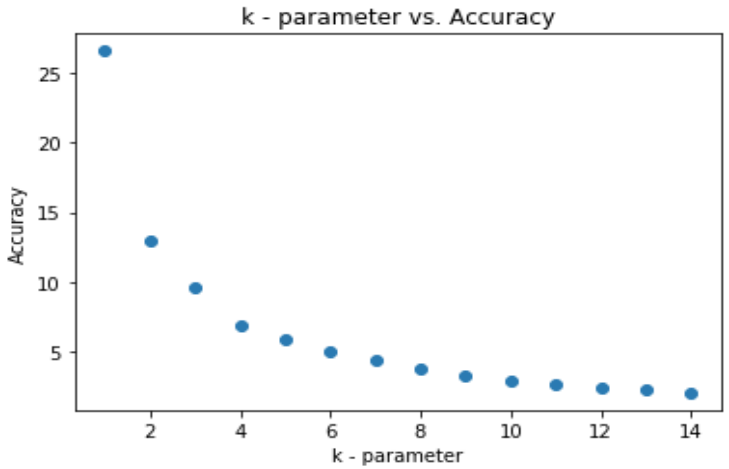
Filtered venues:

- Gym
- Fitness
- Park
- Bus
- Bus Stop
- Bus Station
- Mall
- Shopping Mall
- Shopping Plaza
- Metro
- Metro Station
- Metro Station and Building
- Train
- Train Station
- Pharmacy

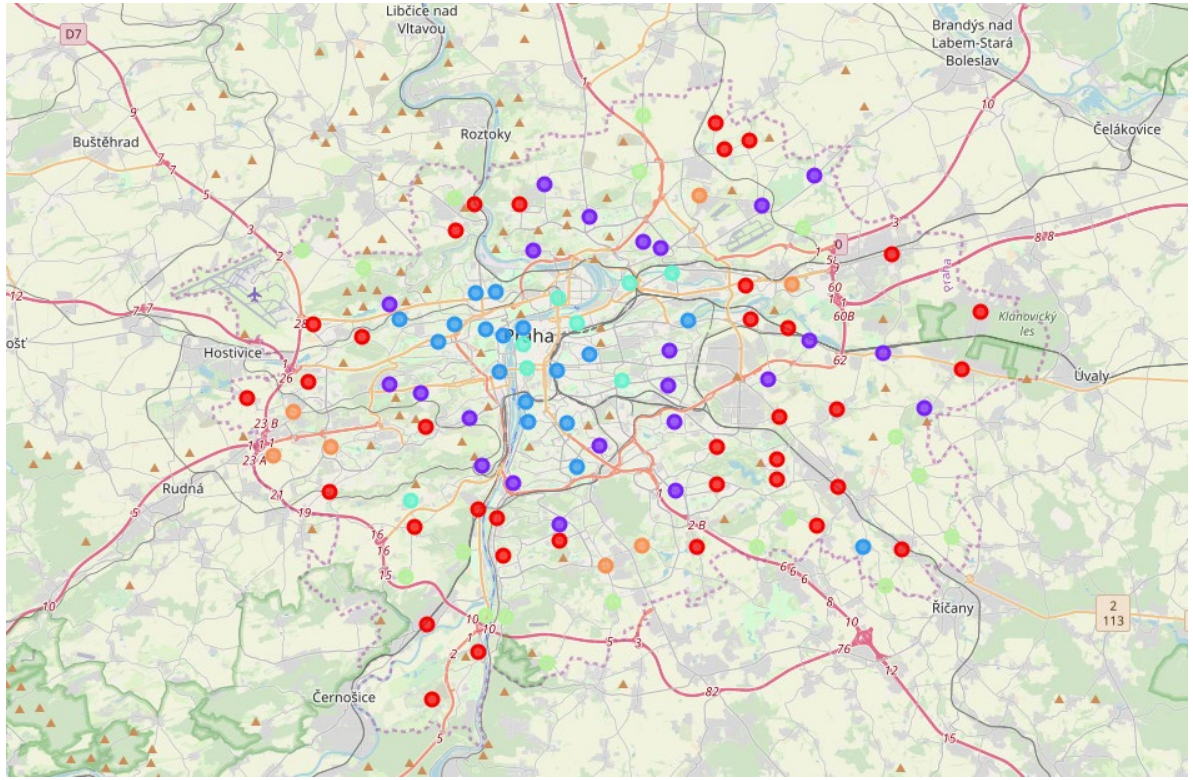
Analysis: Prague - Czech Republic

	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

Clustered dataframe



k parameters dependency



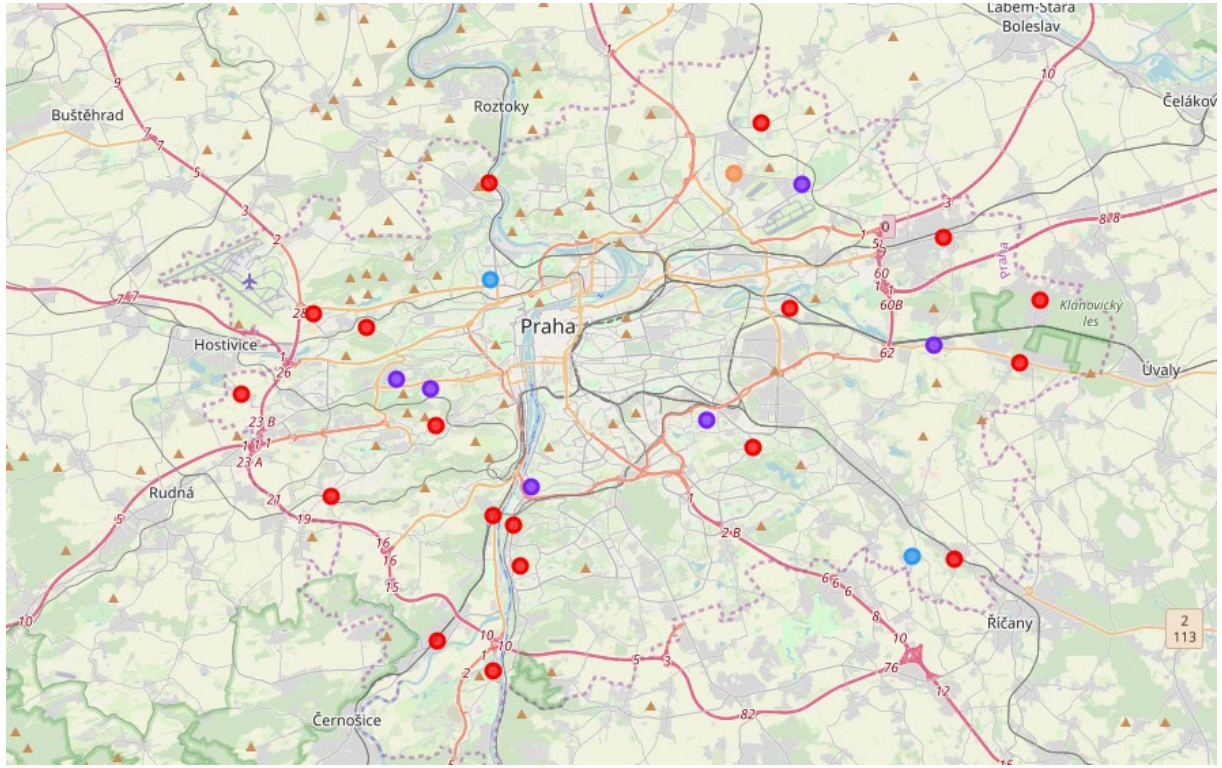
Clusters distribution

Analysis: Prague - Czech Republic

	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

Final filtered locations



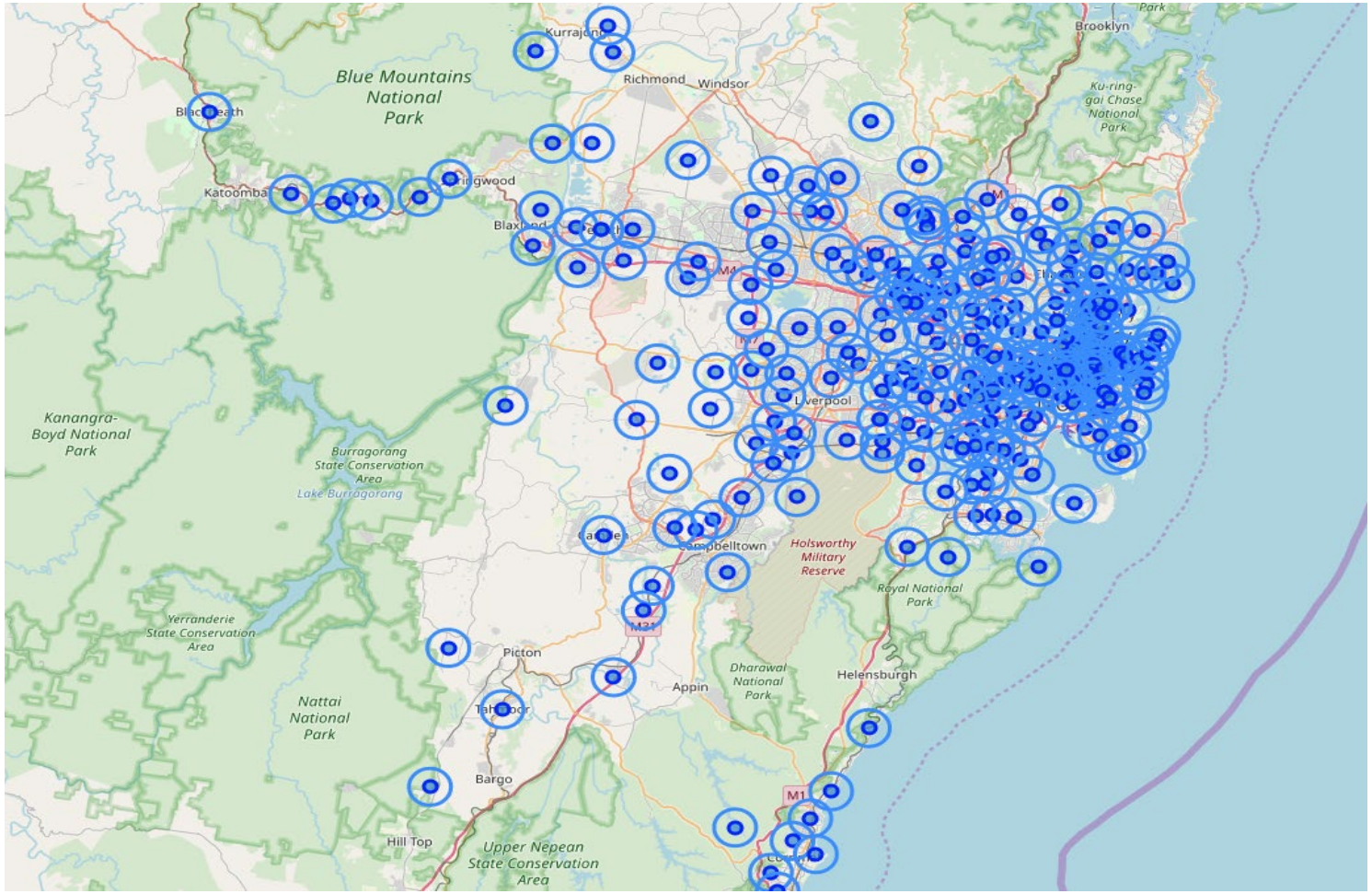
Locations distribution

Analysis: Sydney - New South Wales, Australia

	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

Dataframe of the locations

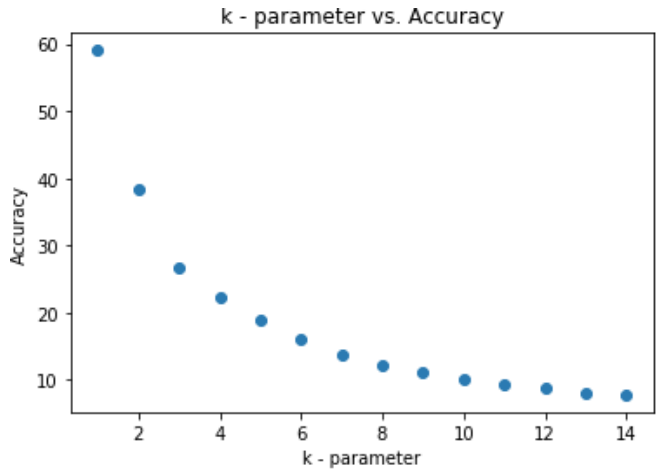


Sydney inner suburbs

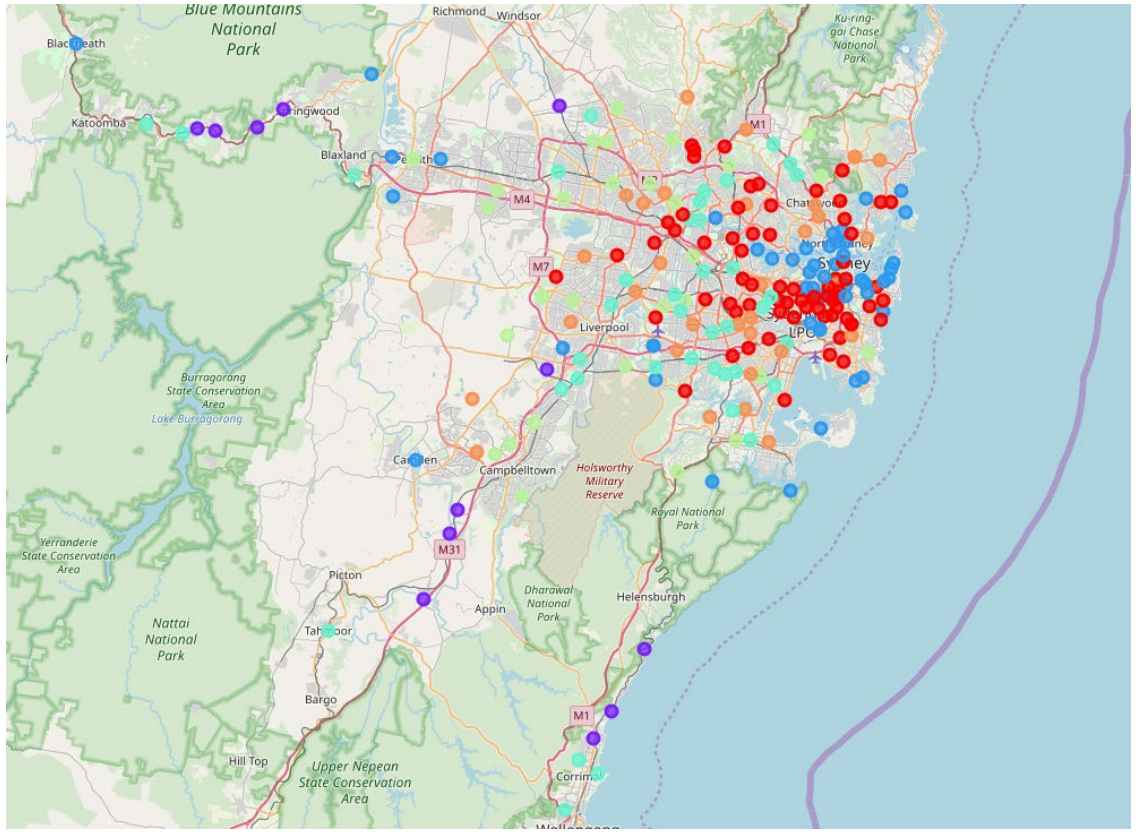
Analysis: Sydney - New South Wales, Australia

	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

Clustered dataframe



k parameters dependency



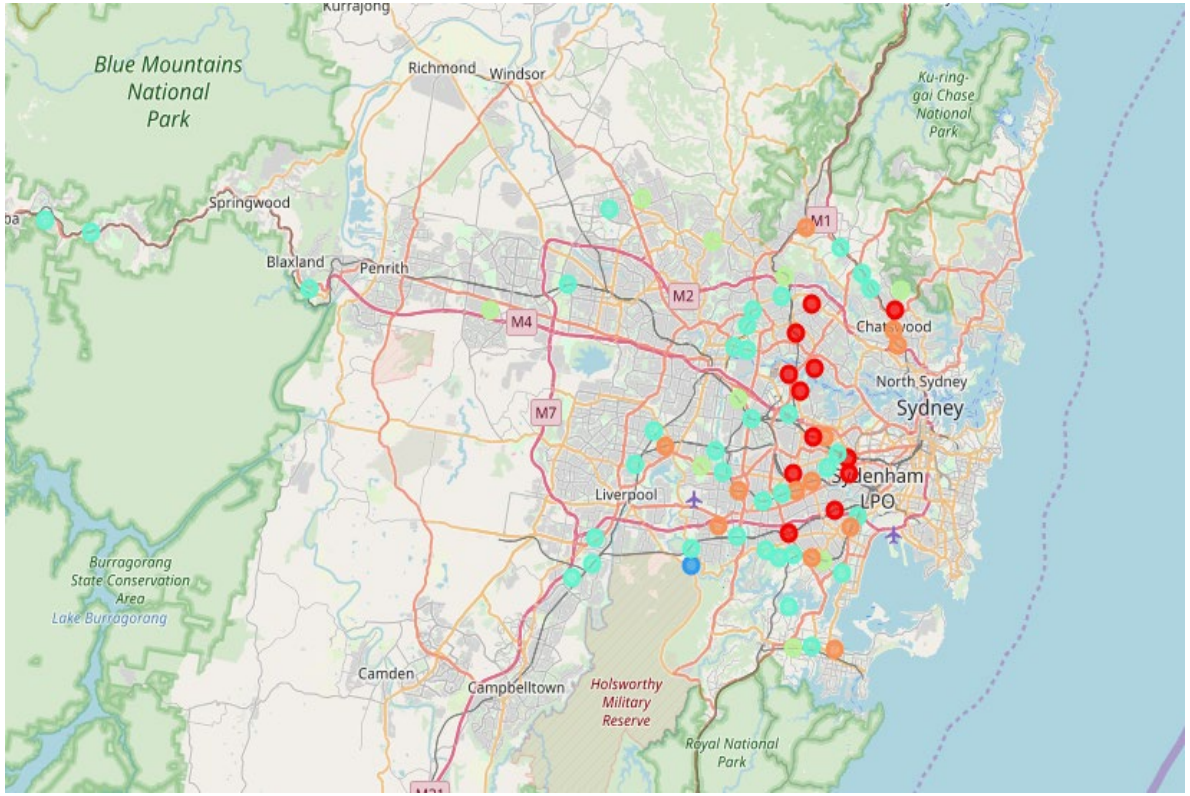
Clusters distribution

Analysis: Sydney - New South Wales, Australia

	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

Final filtered locations



Locations distribution

Results and Discussion

Performed analyses gives resulting number of locations selected according defined conditions.

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 – 24%.**
- **Visualization shows that the clusters has more or less circular shape which separates location of historical center and approaching up to outskirts.**

Sydney:

- **271 inner suburbs has been analyzed and 76 has been classified to meet the criteria, that's 28%.**
- **No clear pattern which could reflect clusters distribution,**
- **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.**

Conclusion

- Analyses done for several locations in two different cities based on specified criteria.
- In Prague it's possible to find 27 locations for further exploration. In Sydney this number is significantly higher - 76 locations. Ratios are however quite close 24% vs. 28%. This means that in Sydney are slightly higher chances to find those specified places.
- Ratios can 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%.
- 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.
- Solid base for further extension by individual factors which could reduce final number of suitable locations or find new one.