Applied Data Science Capstone Project

"Business integration trips in the Silesian region".

Introduction

I have chosen to investigate the Silesian Voivodship as this is where I was brought up. This part of Poland is situated in southern Poland. The region's population is the second-highest in Poland. In the past, the Silesian region was known mainly from its dominant heavy industry (60 mines, 18 iron and steel companies). Today it is a place for dynamically developing motor, machinery, equipment, and energy industries as well as companies operating in the business process outsourcing and logistics sectors. Silesia is now one of the top regions in terms of the number of big investors mainly because of its characteristic of an agglomeration. Foreign investors include IBM, Unilever, Rockwell, Capgemini, Deloitte, Vattenfall and ABB to name a few.[1]

I come up with an idea that might be of interest to businesses or tourist companies who might look for business integration trips for employees or business venues with entertainment opportunities.

I was contacted by a businessman who is starting his business in the Silesian region and wants to learn more about pleasant locations with entertainment opportunities. He wants to offer full packages of business venues and integration trips to companies. The package should include hotel and entertainment possibilities. He gave me some directions in terms of ideal location as well as the type of entertainment he is looking to include in his packages.

My job is to investigate selected venues popularities and their locations in all Silesian's cities. Having a big picture of interesting venues, I will choose the most promising location for the trip and next, I will find hotels/ conference facilities. Finally, I will investigate venues around the chosen hotel in more details.

[1] https://ec.europa.eu/growth/tools-databases/regional-innovation-monitor/base-profile/silesia

Data

Geometries of cities in the Silesian region were extracted from json file taken from the NYU Spatial Data Repository website[2]. Those data represent the second-level administrative divisions of Poland and contain all regions (voivodeships) and their corresponding cities in Poland. The geometries are described as multipolygons. To generate polygons and centroid location coordinates values for each city, the library Shapely[3] was used. Geometries of cities belonging to the Silesian region were used to look for venues in those regions and to visualise results on maps.

Venues of interest in the Silesian region were found using the Foursquare API[4]. Multiple searches were performed to look for all venue categories, search for selected venues and explore popular spots. The foursquare website also contains information about all venue's categories and subcategories. Initially, the following categories were chosen to find recommended venues in all cities belonging to the Silesian Voivodeship: Cave, Forest, Hill, Lake, Mountain, National Park, Nature Preserve, River, Bowling Alley, Go Kart Track, Golf Course, Golf Driving Range, Paintball Field, Reservoir, Scenic Lookout, Sculpture Garden, State / Provincial Park, Windmill and Vineyard. In the next step, the location with the highest number of venues of interest was further investigated and categories such as Recreation Center, Convention Center, Resort and Hotel were included in searches for the best hotel facility. Finally, the location around that hotel was further explored.

Methodology

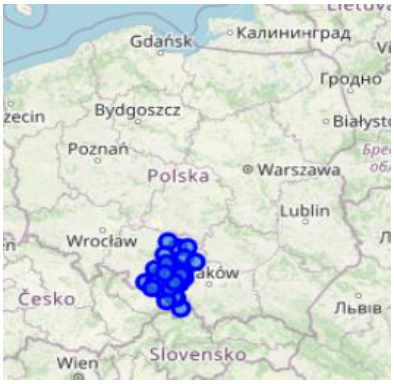
The second-level administrative divisions of Poland data contain geometry information about 16 voivodeships (376 neighborhoods in total)(a). The geojson file with those data was used as a starting point. The shapely package was utilised to generate polygons (filled areas representing neighborhoods) and centroid coordinates for each neighborhood in Poland. As we are only interested in one voivodeship (Silesian - that has 36 neighborhoods)(b), only data relevant to this area were pulled out and used in further searches. Silesian's neighborhoods include 19 cities and 17 districts hence some names might suggest the same neighborhood, but in fact, there are differences in terms of administrative rights[5].

[2] https://geo.nyu.edu/catalog/stanford-xh662zc5620
[3] https://pypi.org/project/Shapely/
[4] https://developer.foursquare.com/docs/
[5] https://pl.wikipedia.org/wiki/Wojew%C3%B3dztwo_%C5%9Bl%C4%85skie

The geometry data for all neighborhoods in the Silesian region contain information such as region (the voivodeship), Neighborhood (cities and districts), centroid location coordinates and geometries of polygons for each city.

	region	Neighborhood	centroid_x	centroid_y	coordinates
0	Silesian	Żory	18.685270	50.022560	{'type': 'MultiPolygon', 'coordinates': [[[[18...
1	Silesian	Żywiec	19.184707	49.587275	{'type': 'MultiPolygon', 'coordinates': [[[[19...
2	Silesian	Świętochłowice	18.826438	50.235390	{'type': 'MultiPolygon', 'coordinates': [[[[18...
3	Silesian	Będzin	19.116783	50.371653	{'type': 'MultiPolygon', 'coordinates': [[[[19...
4	Silesian	Bielsko-Biała	19.058871	49.791412	{'type': 'MultiPolygon', 'coordinates': [[[[19...

a) All neighborhoods in Poland



b) Map of Poland with Silesian's neighborhoods in blue

Table 1. Map of Poland and the Silesian region

The python folium library was used to visualise Silesian's neighborhoods (the map of Katowice with neighborhoods superimposed on top) using centroid coordinates for each neighborhood (Table1/b).

Foursquare API was used to explore all possible categories that can be used for this project. After consultation with the client, we have decided to find neighborhoods with the highest number of venues belonging to the following categories: Cave, Forest, Hill, Lake, Mountain, National Park, Nature Preserve, River, Bowling Alley, Go Kart Track, Golf Course, Golf Driving Range, Paintball Field, Reservoir, Scenic Lookout, Sculpture Garden, State / Provincial Park, Windmill and Vineyard. All Silesian's neighborhoods were considered in the search.

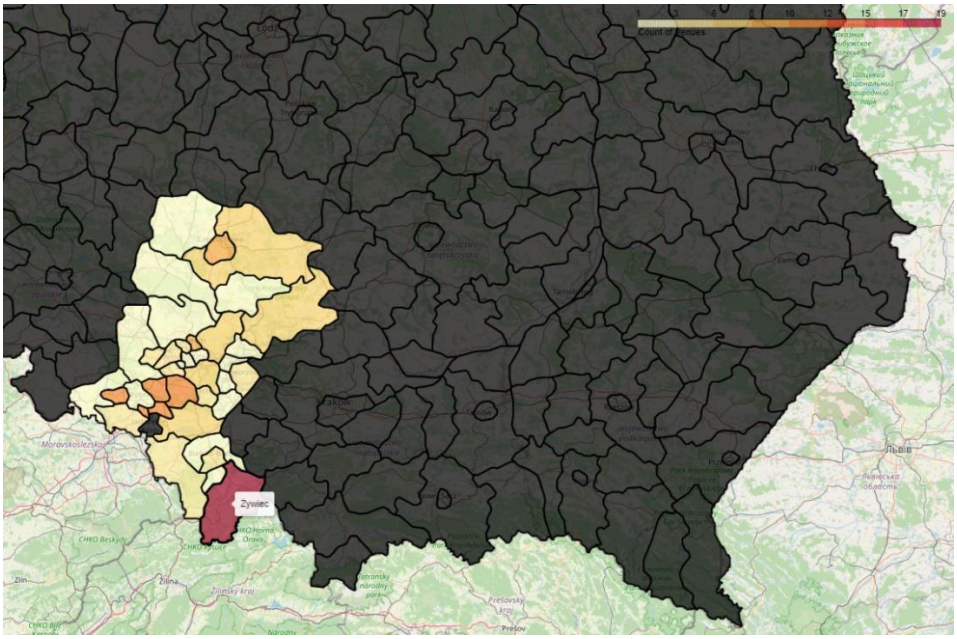
The sample of an output from the search for the most recommended venues for Silesian's neighborhoods using selected categories is shown below:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Żory	50.022560	18.685270	Staw Jesionka	50.056743	18.689521	Lake
1	Żory	50.022560	18.685270	Śmieszek	50.052871	18.717026	Lake
2	Żory	50.022560	18.685270	Żory Wygoda	50.060107	18.677101	Lake
3	Żory	50.022560	18.685270	Gichta	50.072865	18.709553	Lake
4	Żory	50.022560	18.685270	Halda Paragliding	49.997228	18.593819	Mountain
5	Żory	50.022560	18.685270	Staw Zarzyna	50.082324	18.747439	Lake
6	Żory	50.022560	18.685270	Ośrodek Baron	50.071583	18.769624	Lake
7	Żory	50.022560	18.685270	Staw Walszówka	50.085943	18.752724	Lake
8	Żory	50.022560	18.685270	Staw Jesionka	50.085954	18.757978	Lake
9	Żory	50.022560	18.685270	Las Kyndra	49.948726	18.630994	Forest
10	Żory	50.022560	18.685270	Paintball Jastrzębie	49.949497	18.617217	Paintball Field
11	Żywiec	49.587275	19.184707	Prusów	49.557950	19.148847	Mountain
12	Żywiec	49.587275	19.184707	Romanka 1366 m n.p.m.	49.559802	19.241499	Mountain

Groupby method was used to find out how many venues were returned for each neighborhood.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Bielsko	2	2	2	2	2	2
Bielsko-Biała	4	4	4	4	4	4
Bieruń-Lędziny	1	1	1	1	1	1
Bytom	5	5	5	5	5	5
Będzin	7	7	7	7	7	7
Chorzów	3	3	3	3	3	3
Cieszyn	4	4	4	4	4	4
Częstochowa	6	6	6	6	6	6
Częstochowa City	8	8	8	8	8	8

I have generated a Choropleth map that shows venue distribution within the Silesian region. The red colour shows the highest number of recommended venues found, light yellow, the lowest and grey where no venues were found (1 neighborhood in the region) or the neighborhoods that were not considered in the search, because they belong to other voivodeships.



One Hot Encoding technique was used to convert the list of categories to a vector where each column corresponds to one possible value of the feature. Every different category is shown in columns that contain either 1 or 0. 1 show that the venue was recommended, whereas 0 that it did not return venues.

	Neighborhood	Bowling Alley	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill	Hotel	Lake	Mountain	National Park	Nature Preserve	Other Great Outdoors	Paintball Field	Park	Pool	Public Art	River	Scenic Lookout	Sculpture Garden	Shopping Mall	Ski Area	Ski Chairlift	Sports Bar	Trail
0	Żory	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Żory	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Żory	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Żory	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Żory	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Groupby method was used to group rows by Neighborhood and the mean of the frequency of occurrence of each category was assessed.

	Neighborhood	Bowling Alley	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill	Hotel	Lake	Mountain	National Park	Nature Preserve	Other Great Outdoors	Paintball Field	Park	Pool	Public Art	River	Scenic Lookout	Sculpture Garden
0	Bielsko	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.0	0.00	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
1	Bielsko-Biała	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.0	0.00	0.250000	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
2	Bieruń-Lędziny	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.0	0.00	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
3	Bytom	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.0	0.00	0.200000	0.200000	0.000000	0.000000	0.200000	0.000000	0.000000	0.0	0.000000	0.000000	0.200000	0.000000
4	Będzin	0.000000	0.00	0.000000	0.000000	0.142857	0.142857	0.0	0.00	0.714286	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
5	Chorzów	0.333333	0.00	0.000000	0.000000	0.000000	0.000000	0.0	0.00	0.000000	0.333333	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.333333	0.000000	0.000000	0.000000
6	Cieszyn	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.0	0.00	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
7	Częstochowa	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.0	0.00	0.333333	0.333333	0.166667	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.166667	0.000000	0.000000
8	Częstochowa City	0.375000	0.00	0.000000	0.000000	0.125000	0.125000	0.0	0.00	0.125000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.250000	0.000000
9	Dąbrowa Górnicza	0.000000	0.00	0.333333	0.000000	0.000000	0.000000	0.0	0.00	0.666667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000

I was interested to see the summary of the top 5 venues for each neighborhood together with their mean values. An example of the output file is shown below.

```
----Częstochowa City----
            venue  freq
0   Bowling Alley  0.38
1  Scenic Lookout  0.25
2   Go Kart Track  0.12
3     Golf Course  0.12
4            Lake  0.12
5            Park  0.00
6      Sports Bar  0.00
7   Ski Chairlift  0.00
8        Ski Area  0.00
9   Shopping Mall  0.00


----Dąbrowa Górnicza----
             venue  freq
0             Lake  0.67
```

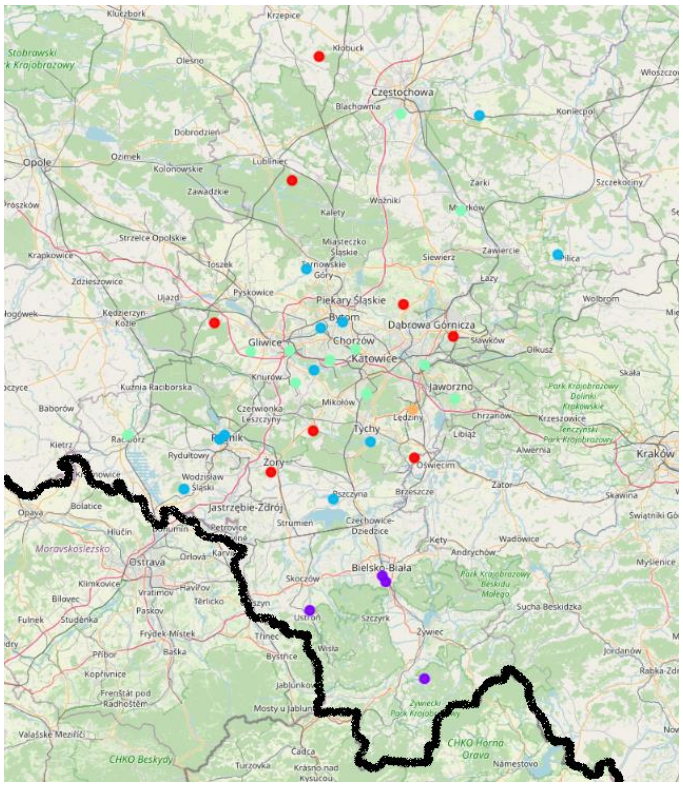
Finally, I wated to compare 10 most recommended venues with their rankings of popularity for the neighborhoods.

	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	Bielsko	Mountain	Trail	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill	Hotel
1	Bielsko-Biała	Mountain	Trail	Lake	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill
2	Bieruń-Lędziny	Lake	Trail	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill	Hotel
3	Bytom	Other Great Outdoors	Ski Chairlift	Scenic Lookout	Lake	Mountain	National Park	Church	Dive Spot	Forest	Go Kart Track
4	Będzin	Lake	Go Kart Track	Golf Course	Trail	Nature Preserve	Church	Dive Spot	Forest	Hill	Hotel
5	Chorzów	Bowling Alley	Public Art	Mountain	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill
6	Cieszyn	Mountain	Trail	Ski Area	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill
7	Częstochowa	Lake	Mountain	River	National Park	Trail	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track
8	Częstochowa City	Bowling Alley	Scenic Lookout	Go Kart Track	Golf Course	Lake	Nature Preserve	Church	Dive Spot	Forest	Hill
9	Dąbrowa Górnicza	Lake	Dive Spot	Trail	Nature Preserve	Church	Forest	Go Kart Track	Golf Course	Hill	Hotel
10	Gliwice	Lake	Trail	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill	Hotel
11	Gliwice City	Bowling Alley	Church	Scenic Lookout	Mountain	Nature Preserve	Dive Spot	Forest	Go Kart Track	Golf Course	Hill
12	Jaworzno	Bowling Alley	Dive Spot	Scenic Lookout	Pool	Lake	National Park	Church	Forest	Go Kart Track	Golf Course
13	Katowice City	Bowling Alley	Lake	Sculpture Garden	Scenic Lookout	National Park	Church	Dive Spot	Forest	Go Kart Track	Golf Course
14	Kłobuck	Lake	Trail	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill	Hotel
15	Lubliniec	Lake	Trail	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill	Hotel
16	Mikołów	Lake	Hill	Bowling Alley	Sculpture Garden	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course
17	Myszków	Bowling Alley	Scenic Lookout	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill	Hotel
18	Mysłowice	Scenic Lookout	Trail	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill	Hotel
19	Piekary Śląskie	Lake	Other Great Outdoors	Scenic Lookout	Golf Course	Public Art	Bowling Alley	Ski Chairlift	Sculpture Garden	Ski Area	Church
20	Pszczyna	Lake	Forest	Scenic Lookout	Golf Course	Trail	Nature Preserve	Church	Dive Spot	Go Kart Track	Hill
21	Racibórz	Forest	Scenic Lookout	River	Trail	National Park	Church	Dive Spot	Go Kart Track	Golf Course	Hill
22	Ruda Śląska	Hill	Bowling Alley	Mountain	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hotel
23	Rybnik	Forest	Scenic Lookout	Lake	Mountain	Shopping Mall	Go Kart Track	Park	Trail	National Park	Church
24	Rybnik City	Shopping Mall	Forest	Trail	Nature Preserve	Church	Dive Spot	Go Kart Track	Golf Course	Hill	Hotel
25	Siemianowice Śląskie	Bowling Alley	Sculpture Garden	Public Art	National Park	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill
26	Sosnowiec	Dive Spot	Go Kart Track	Scenic Lookout	Trail	Nature Preserve	Church	Forest	Golf Course	Hill	Hotel
27	Tarnowskie	Other Great Outdoors	Ski Chairlift	Lake	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill
28	Tychy City	Lake	Other Great Outdoors	Bowling Alley	Ski Area	Shopping Mall	Church	Dive Spot	Forest	Go Kart Track	Golf Course
29	Wodzisław	Golf Course	Hotel	Lake	Sports Bar	Trail	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track
30	Zabrze	Bowling Alley	Church	Scenic Lookout	Mountain	Nature Preserve	Dive Spot	Forest	Go Kart Track	Golf Course	Hill
31	Zawiercie	Lake	Mountain	Nature Preserve	Forest	Trail	Church	Dive Spot	Go Kart Track	Golf Course	Hill
32	Świętochłowice	Lake	Mountain	Trail	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill
33	Żory	Lake	Forest	Mountain	Paintball Field	Trail	Nature Preserve	Church	Dive Spot	Go Kart Track	Golf Course
34	Żywiec	Mountain	River	Scenic Lookout	Trail	National Park	Church	Dive Spot	Forest	Go Kart Track	Golf Course

To understand data, I further analysed the results and clustered neighborhoods with common venues using unsupervised learning k-mean algorithm. I run k-mans to cluster the neighbourhoods into 5 clusters. I had to clean the data as one of the neighborhoods (Jastrzebie Zdroj) did not returned venues and another issue was that the cluster labels were returned as float numbers, not integers. The cleaned data sample is shown below.

	region	Neighborhood	centroid_x	centroid_y	coordinates	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	Silesian	Żory	18.685270	50.022560	{'type': 'MultiPolygon', 'coordinates': [[[18...	2	Lake	Forest	Mountain	Paintball Field	Trail	Nature Preserve	Church	Dive Spot	Go Kart Track	Golf Course
1	Silesian	Żywiec	19.184707	49.587275	{'type': 'MultiPolygon', 'coordinates': [[[19...	3	Mountain	River	Scenic Lookout	Trail	National Park	Church	Dive Spot	Forest	Go Kart Track	Golf Course
2	Silesian	Świętochłowice	18.826438	50.235390	{'type': 'MultiPolygon', 'coordinates': [[[18...	4	Lake	Mountain	Trail	Nature Preserve	Church	Dive Spot	Forest	Go Kart Track	Golf Course	Hill

I used the folium map to visualise clusters, showing Silesian's neighborhoods grouped into 5 clusters. Clustered neighbourhoods are shown in 5 different colours, representing common venue popularities in Silesian's neighborhoods.

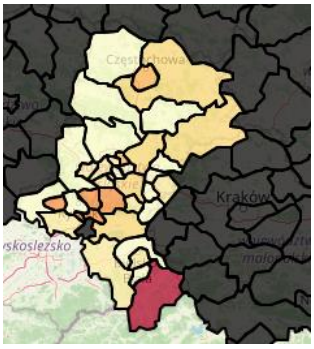


Foursquare API was also used to find hotels for the chosen location. The information about hotels was investigated in details looking at users' review comments, ratings and tips to meet the client expectation.

Finally, the hotel matching the requirements best was further explored for any nearby venues to provide complete information about the recommended spot for the integration trip.

Results

The initial search for the most recommended venues belonging to selected categories gave 168 results in all Silesia's neighborhoods. Those venues belong to 25 unique categories: Lake, Mountain, Forest, Paintball Field, Scenic Lookout, River, Go Kart Track, Golf Course, Trail, Other Great Outdoors, Ski Chairlift, Public Art ,Bowling Alley, Ski Area, National Park, Dive Spot, Church, Pool, Sculpture Garden, Hill, Shopping Mall, Park, Sports Bar, Hotel, Nature Preserve. The highest number of recommended venues was returned for Zywiec neighbourhood, which can be seen on the map in red colour. Other neighboorhoods with a high number of venues were observed in Zory, Rybnik and Mikolow (orange colour). In one neighborhood - Jastrzebie Zdroj, no venues were observed.



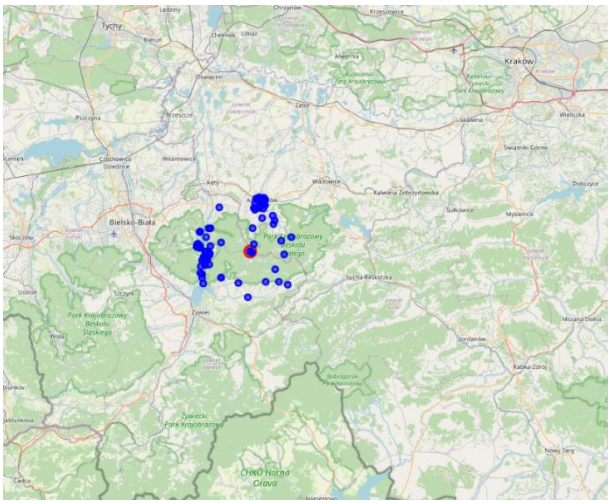
When venues were clustered it was observed that the area highlighted in red colour on the map above is also the one that the client likes most because of the highly valued venues. The common highly valued venues were observed:

- Purple cluster
    - location surroundings: mountains and trail
    - entertainment: diving, go kart, golf course
- Blue cluster
    - location surroundings: lake, forest
    - entertainment: bowling, diving
- Green cluster
    - locations surroundings: various, not consistent
    - entertainment venues: bowling
- Red cluster
    - locations surroundings: lake and natural preserve
    - entertainment: go kart, paintball
- Orange cluster
    - locations surroundings: scenic lookout, trail
    - Entertainment: go card and golf course

The client's ideal location was the one that has in close vicinity mountains or lakes/ rivers or forests and the purple cluster ideally matches that description. For that reason, Zywiec was chosen as a location to search for a hotel. When searching for hotels one of the requirements was the conference facilities. A few hotels were investigated and information such as users' review comments, ratings and tips were taken under consideration. Finally, the hotel matching the client's requirements best was further explored for any nearby venues to provide complete information about the recommended spot for the integration trip.

Based on the customer's review and rating Kocierz Hotel & SPA was chosen as a recommended place for an integration trip. The venues within 500m from the hotel were investigated in addition to previously found venues for this region. Restaurants and additional to previously found entertainment venues were listed in the clients' report.

The map shows the hotel location together with nearby venues.



Discussion

Recommendation: I would recommend the process described in Methods and Results of this work to anyone interested in finding the perfect place based on the individual's requirements. Analysis of generated data based on user's recommendations provide a useful tool and can be very insightful in making successful decisions by companies and investors.

Observations: Please note the search location coverage in this work was limited. All found hotels could be investigated more carefully. More detailed analysis of venues generated around hotels could be provided. For this project, I thought it is more important to show various possibilities that data science give us rather than concentrate on a detailed program of the integration trip. I would be willing to spend additional time, play with parameters and provide alternative spots for integration trips if that business offer is of interest to anyone.

Conclusion

Companies often choose locations outside their offices to have business meetings, training, and conferences with the aim to integrate employees. Tools like Foursquare can help decision-makers to choose the right spot for business venues. In this work, I show the process, tools and methods that can be used to design an integration trip offering.