Data Science Capstone Project:
The Battle of Neighborhoods

Introduction

Moving to another city or even a different area within a city, is a difficult decision to make. Likewise, searching for a perfect city to start a new business, or simply moving to another location (city) is not an easy task.

Whether you open a new restaurant, a hotel or a bar, you need to know the existing market in the potential city of your choice. It would be beneficial to know what people are like, what kind of activities they prefer, and places they go to. There are many parameters to consider, but the location should be among the most important ones. You may feel that by being in close proximity to your competitors, you can benefit from their marketing efforts. They spent a considerable amount of resources to find their location and drive traffic to it. Alternatively, you would prefer a place with not as many similar businesses, in order to set the brand. One or the other, location data is an important asset in such decision-making processes. Putting efforts into making a good use of it could be beneficial for anyone considering starting a new business or moving it to another location.

Data

This notebook will explore Foursquare data of venues in the 10 biggest cities of Switzerland:

- Zurich
- Geneva
- Basel
- Lausanne
- Bern
- Winterthur
- Lucerne
- St. Gallen
- Lugano
- Biel/Bienne

First of all, a file containing the zip codes of the whole Switzerland is imported. The file is obtained from the location data resource AggData. For the rest of the data, a Foursquare API was used to get the information about the venues, using the imported zip codes.

Apart from the typical libraries such as *numpy* and *pandas* for data handling, *matplotlib* and *seaborn* for visualization, *requests* for HTTP GET requests, *folium* for geospatial data representation, *geocoder* for obtaining the latitude and longitude information, and *sklearn* for modeling.

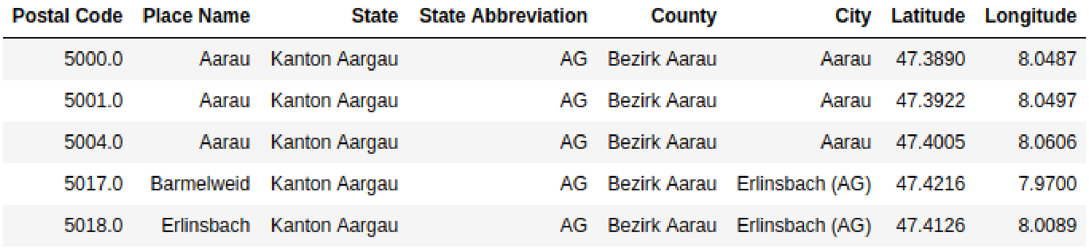
Methodology

Using the Foursquare API, the cities were explored by searching venues around each of the ZIP code's latitude and longitude points. For each ZIP code, a maximum of 200 venues has been pulled out within the radius of 1km. As some areas might overlap, it is important to remove duplicate venues, consulting the venue ID property. Venue categories were extracted from each place, and k-means clustering algorithm was used to find patterns to cluster similar city areas.

Furthermore, a specific query was sent using the API, to search for all the nearby (1km radius) venues using a keyword *restaurant.* Those venues (after removing duplicates) were then queried for details in multiple separate calls (as the free membership allows only certain number of queries per day). Once it has been done, all the data was concatenated into a single data-frame, and explored further. A typical statistical analysis was done in order to describe and explore the requested information. Furthermore, a trial modeling was performed in order to explore the possibility of predicting the restaurant rating. Few models were tested, however, the regressions were not successful enough.

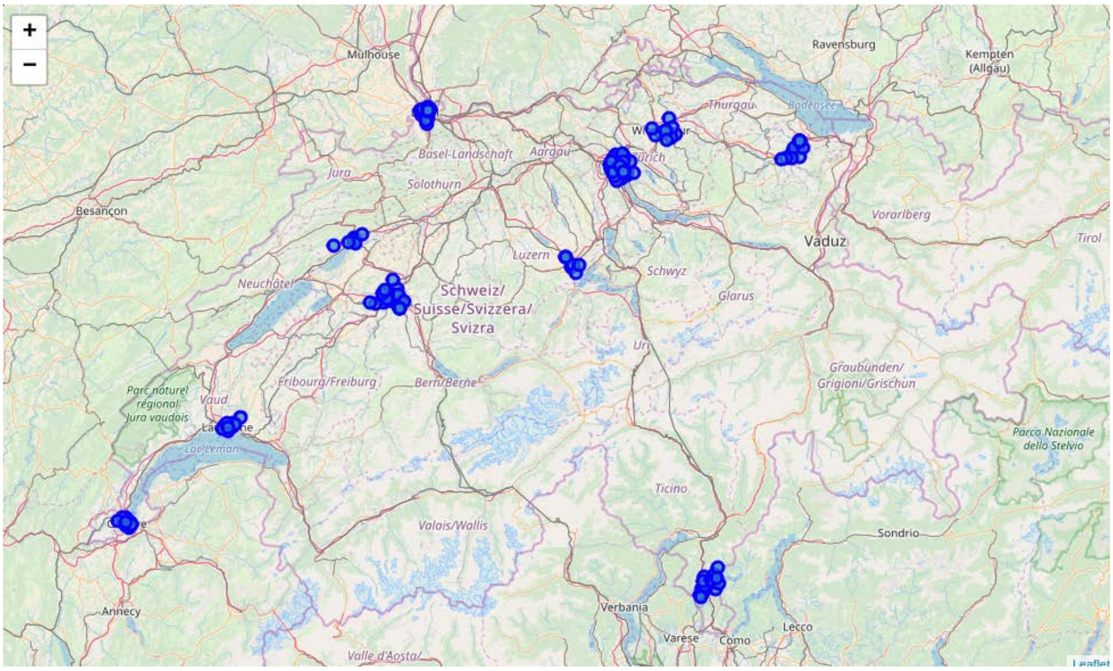
Results

The analysis starts with zip codes. After loading the file into a data-frame, certain zip codes are used only. Having the latitude and longitude already included in the imported file, there was no need for geocoder requests to obtain this information.

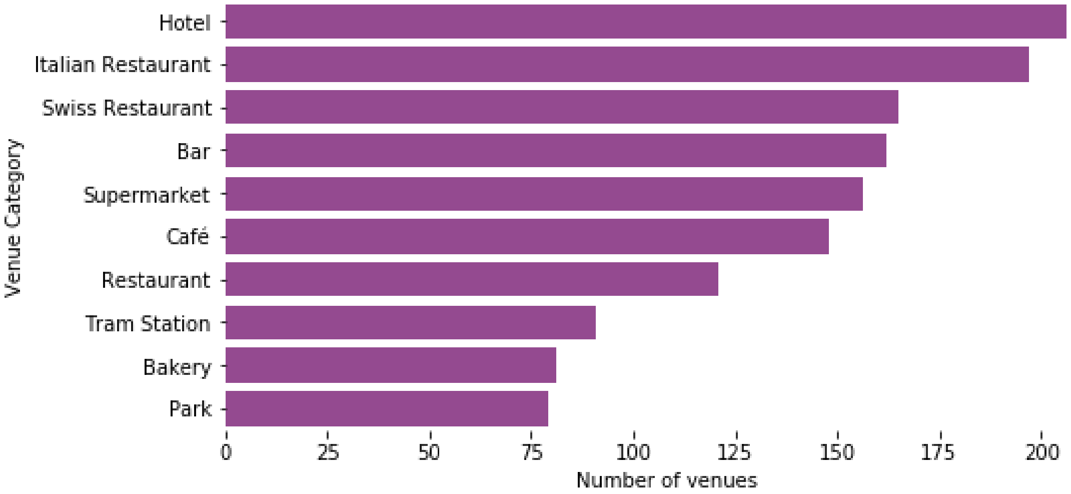


Postal Code	Place Name	State	State Abbreviation	County	City	Latitude	Longitude
5000.0	Aarau	Kanton Aargau	AG	Bezirk Aarau	Aarau	47.3890	8.0487
5001.0	Aarau	Kanton Aargau	AG	Bezirk Aarau	Aarau	47.3922	8.0497
5004.0	Aarau	Kanton Aargau	AG	Bezirk Aarau	Aarau	47.4005	8.0606
5017.0	Barmelweid	Kanton Aargau	AG	Bezirk Aarau	Erlinsbach (AG)	47.4216	7.9700
5018.0	Erlinsbach	Kanton Aargau	AG	Bezirk Aarau	Erlinsbach (AG)	47.4126	8.0089

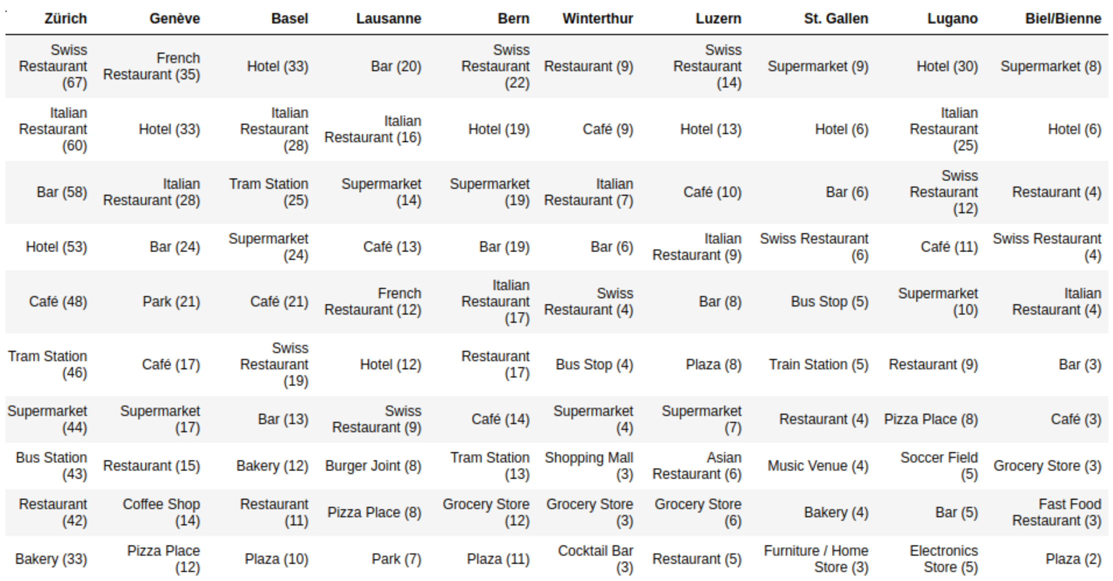
Rows containing only the selected cities are kept, while other data was dismissed. Also, certain areas had the same latitude and longitude information, and were also dismissed. Therefore, **146** areas (zip codes) were explored.



Foursquare *explore* request gave **3680** unique venues within **300** venue categories. Here are the top 10 categories, considering the number of venues belonging to each of them. The count and sorting were done for the whole data-set (across the country).

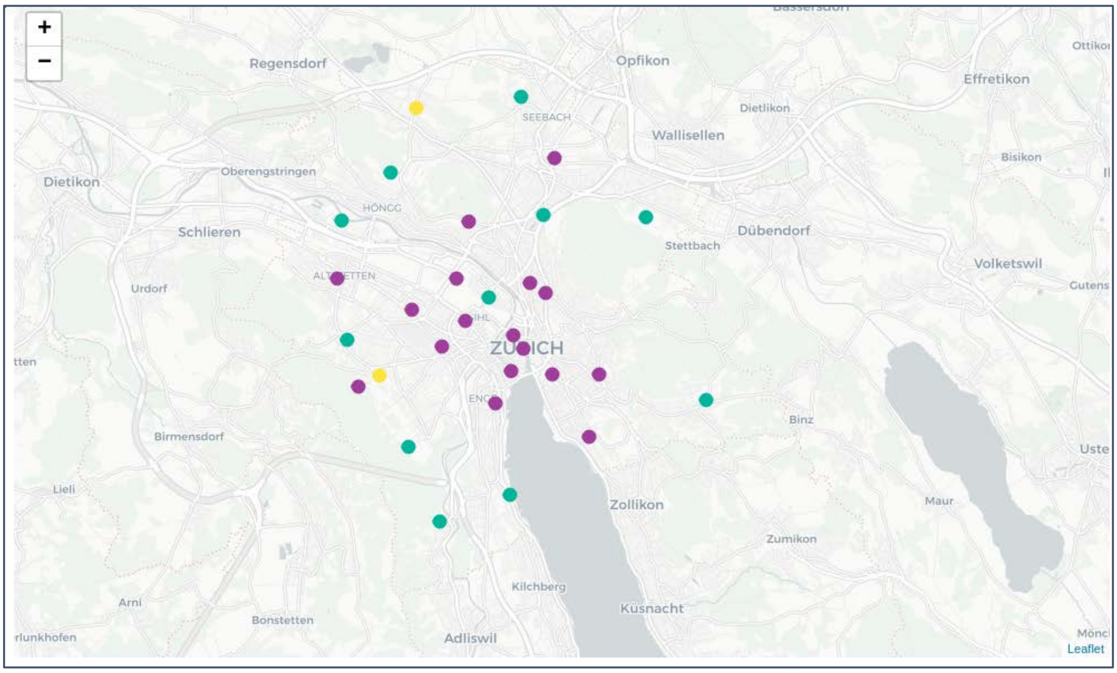


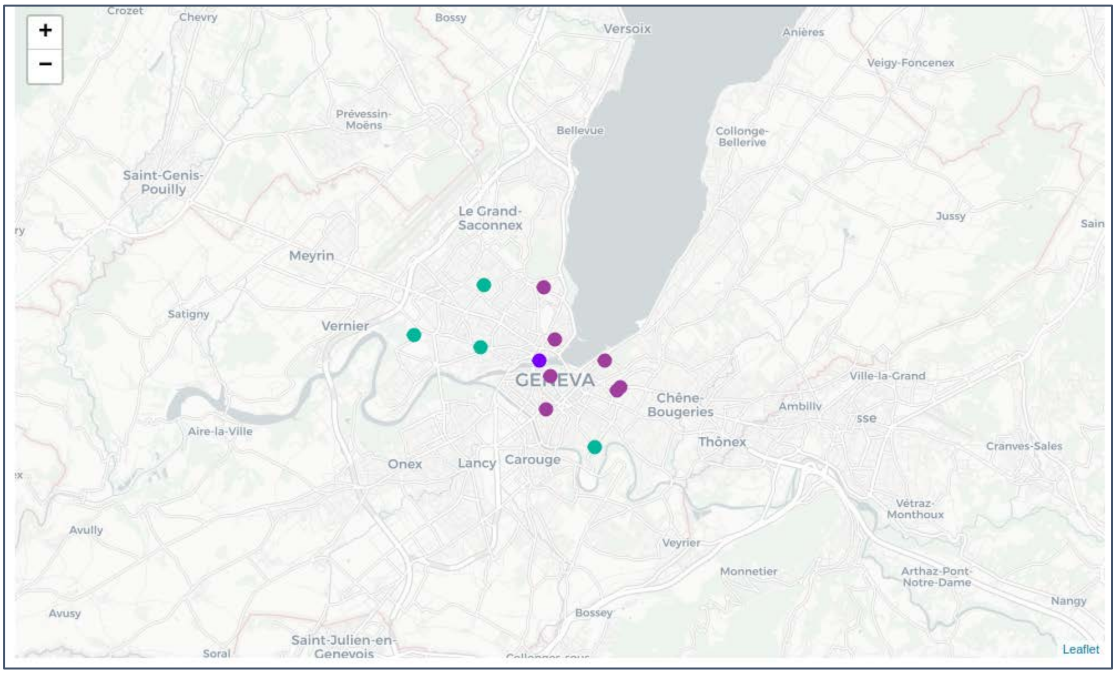
Now, it would be interesting to have a look at the top categories for each city. The table below shows the number of venues within each category, representing the top 10.

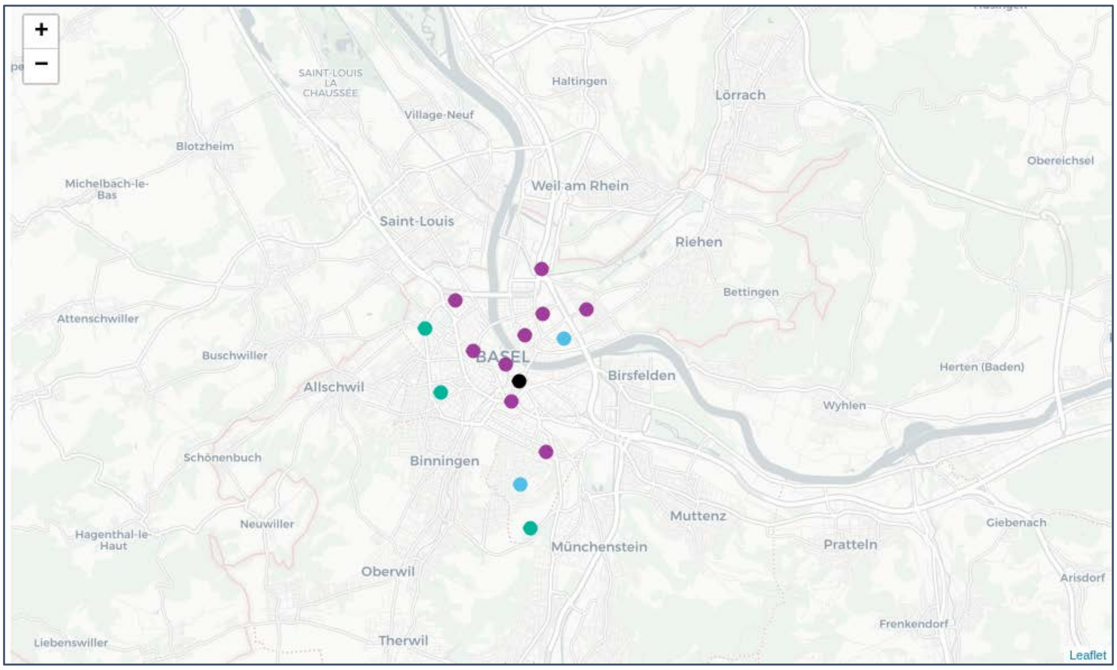


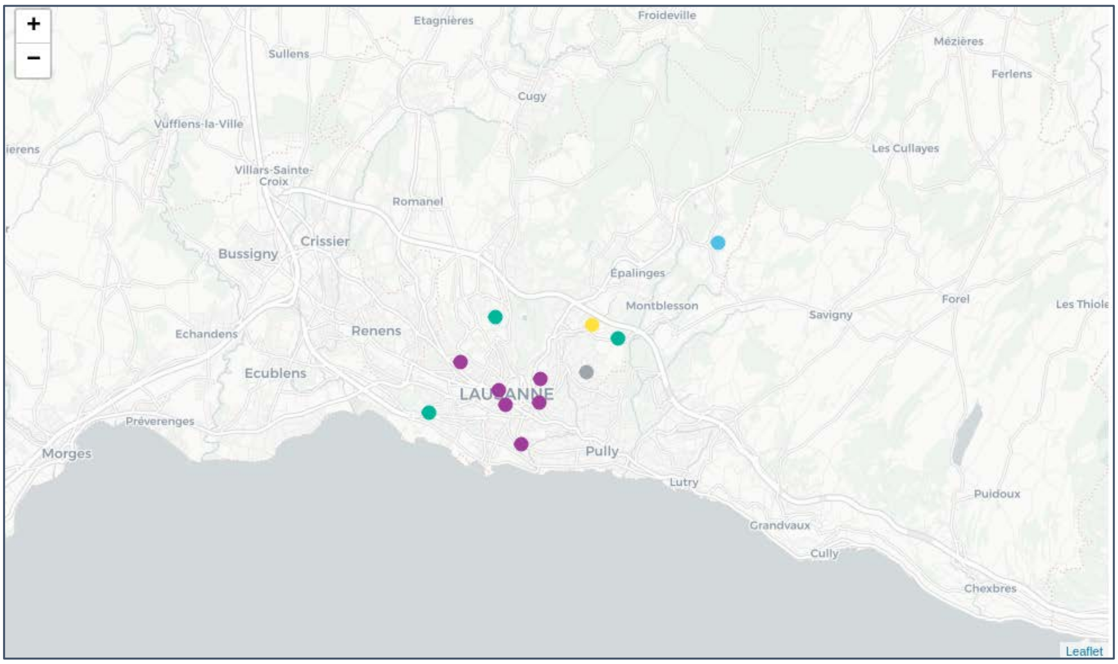
Zürich	Genève	Basel	Lausanne	Bern	Winterthur	Luzern	St. Gallen	Lugano	Biel/Bienne
Swiss Restaurant (67)	French Restaurant (35)	Hotel (33)	Bar (20)	Swiss Restaurant (22)	Restaurant (9)	Swiss Restaurant (14)	Supermarket (9)	Hotel (30)	Supermarket (8)
Italian Restaurant (60)	Hotel (33)	Italian Restaurant (28)	Italian Restaurant (16)	Hotel (19)	Café (9)	Hotel (13)	Hotel (6)	Italian Restaurant (25)	Hotel (6)
Bar (58)	Italian Restaurant (28)	Tram Station (25)	Supermarket (14)	Supermarket (19)	Italian Restaurant (7)	Café (10)	Bar (6)	Swiss Restaurant (12)	Restaurant (4)
Hotel (53)	Bar (24)	Supermarket (24)	Café (13)	Bar (19)	Bar (6)	Italian Restaurant (9)	Swiss Restaurant (6)	Café (11)	Swiss Restaurant (4)
Café (48)	Park (21)	Café (21)	French Restaurant (12)	Italian Restaurant (17)	Swiss Restaurant (4)	Bar (8)	Bus Stop (5)	Supermarket (10)	Italian Restaurant (4)
Tram Station (46)	Café (17)	Swiss Restaurant (19)	Hotel (12)	Restaurant (17)	Bus Stop (4)	Plaza (8)	Train Station (5)	Restaurant (9)	Bar (3)
Supermarket (44)	Supermarket (17)	Bar (13)	Swiss Restaurant (9)	Café (14)	Supermarket (4)	Supermarket (7)	Restaurant (4)	Pizza Place (8)	Café (3)
Bus Station (43)	Restaurant (15)	Bakery (12)	Burger Joint (8)	Tram Station (13)	Shopping Mall (3)	Asian Restaurant (6)	Music Venue (4)	Soccer Field (5)	Grocery Store (3)
Restaurant (42)	Coffee Shop (14)	Restaurant (11)	Pizza Place (8)	Grocery Store (12)	Grocery Store (3)	Grocery Store (6)	Bakery (4)	Bar (5)	Fast Food Restaurant (3)
Bakery (33)	Pizza Place (12)	Plaza (10)	Park (7)	Plaza (11)	Cocktail Bar (3)	Restaurant (5)	Furniture / Home Store (3)	Electronics Store (5)	Plaza (2)

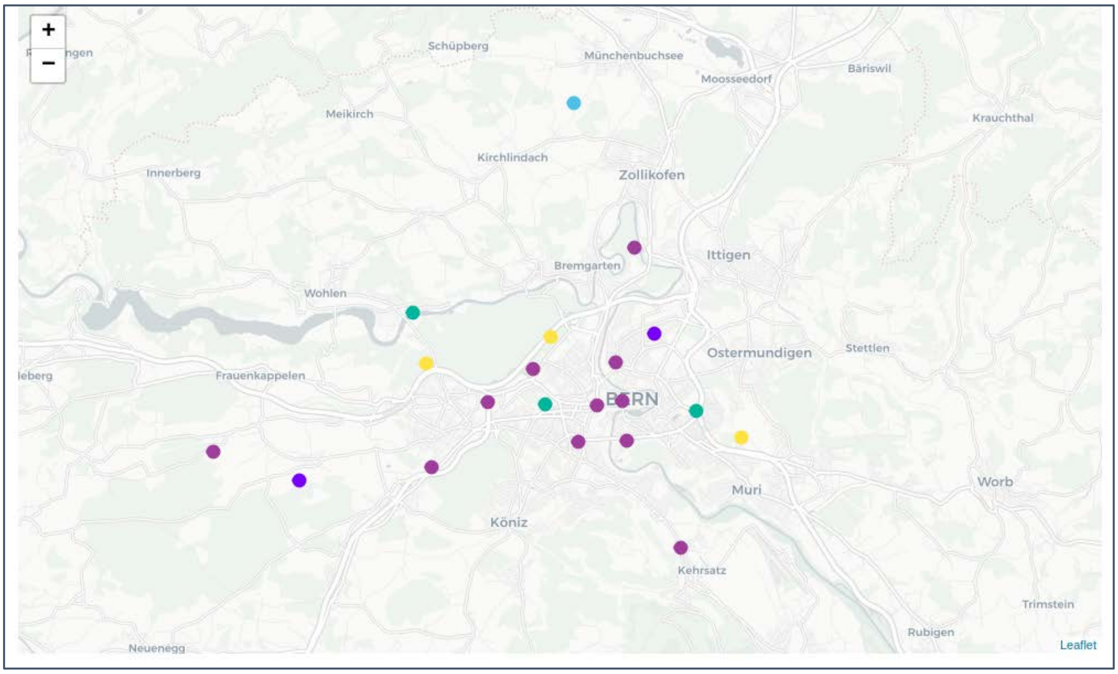
One-hot encoding was used in order to cluster the data, and a special data-frame with top categories per zip code was used in order to retrieve the most common venue categories after the labeling has been done. The total of **10 clusters** were set for the algorithm to work. The results are the following maps (in the descending order of the population of the city).

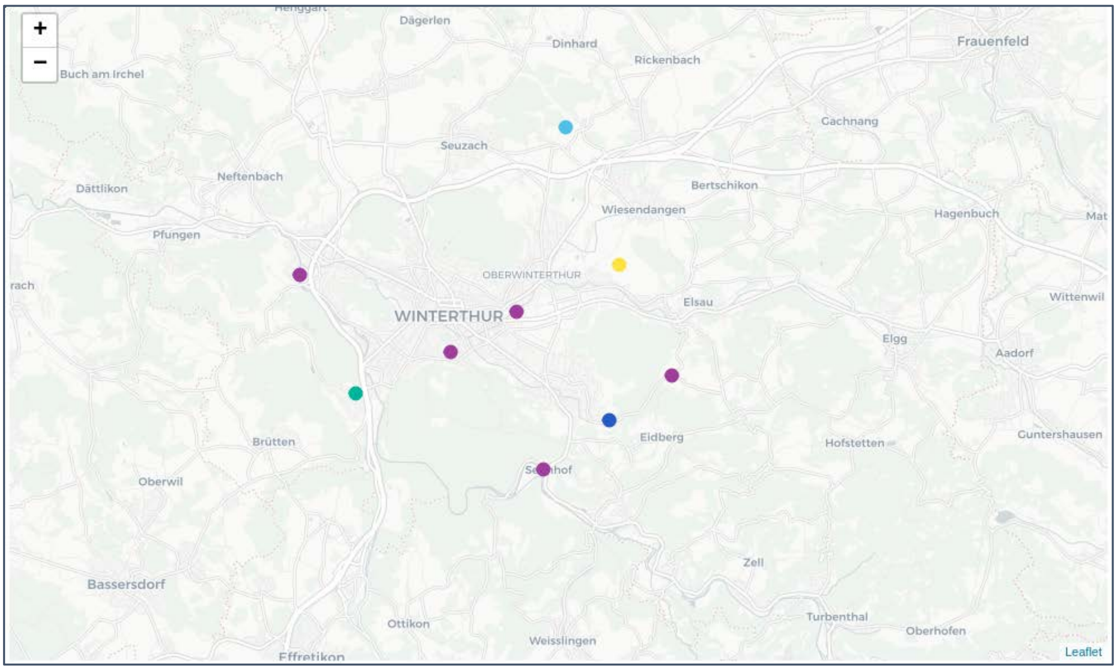


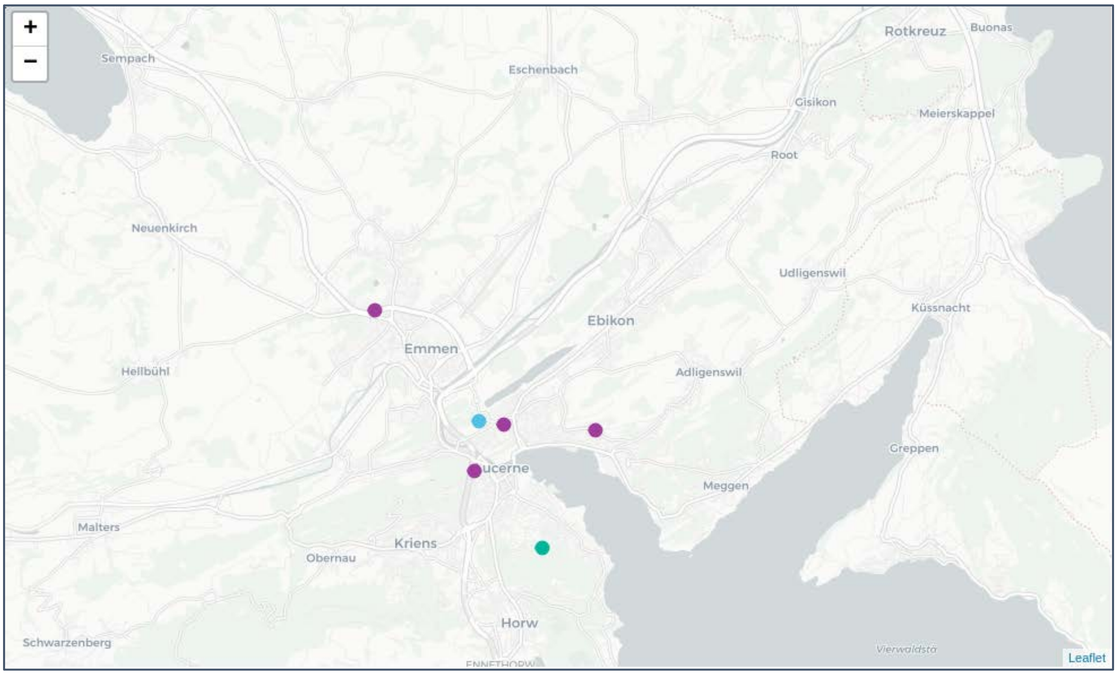

+
−
Crozet
Chevry
Bossy
Versoix
Anières
Veigy-Foncenex
Prévessin-Moëns
Bellevue
Collonge-Bellerive
Saint-Genis-Pouilly
Le Grand-Saconnex
Jussy
Meyrin
Satigny
Vernier
GENEVA
Chêne-Bougeries
Ville-la-Grand
Ambilly
Aire-la-Ville
Thônex
Cranves-Sales
Onex
Lancy
Carouge
Vétraz-Monthoux
Avully
Veyrier
Monnetier
Arthaz-Pont-Notre-Dame
Nangy
Avusy
Bossey
Saint-Julien-en-Genevois
Soral
Leaflet

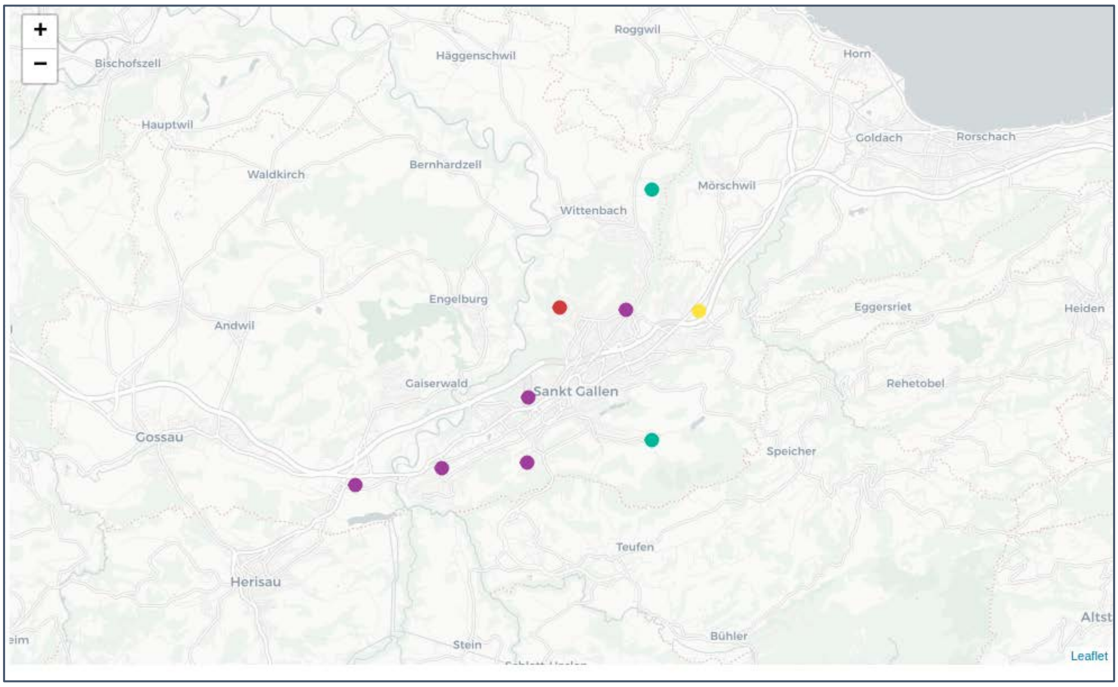

+
−
SAINT-LOUIS LA CHAUSSÉE
Haltingen
Lörrach
Village-Neuf
Blotzheim
Obereichsel
Michelbach-le-Bas
Weil am Rhein
Saint-Louis
Riehen
Bettingen
Attenschwiller
Buschwiller
BASEL
Birsfelden
Allschwil
Herten (Baden)
Wyhlen
Schönenbuch
Binningen
Muttenz
Neuwiller
Giebenach
Hagenthal-le-Haut
Münchenstein
Pratteln
Oberwil
Arisdorf
Frenkendorf
Liebenswiller
Therwil
Leaflet

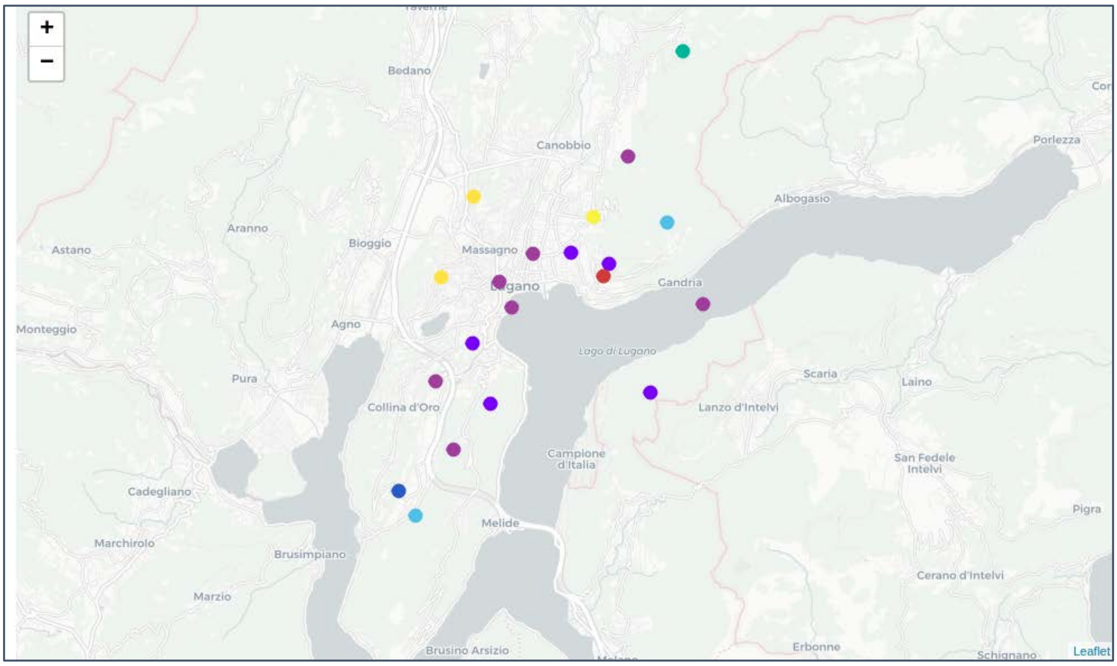

Etagnières
Froideville
Sullens
Mézières
Ferlens
Cugy
Vufflens-la-Ville
Les Cullayes
Villars-Sainte-Croix
Romanel
Crissier
Bussigny
Epalinges
Montblesson
Savigny
Forel
Les Thiole
Echandens
Renens
Ecublens
LAUSANNE
Préverenges
Morges
Pully
Lutry
Puidoux
Grandvaux
Cully
Chexbres
Leaflet

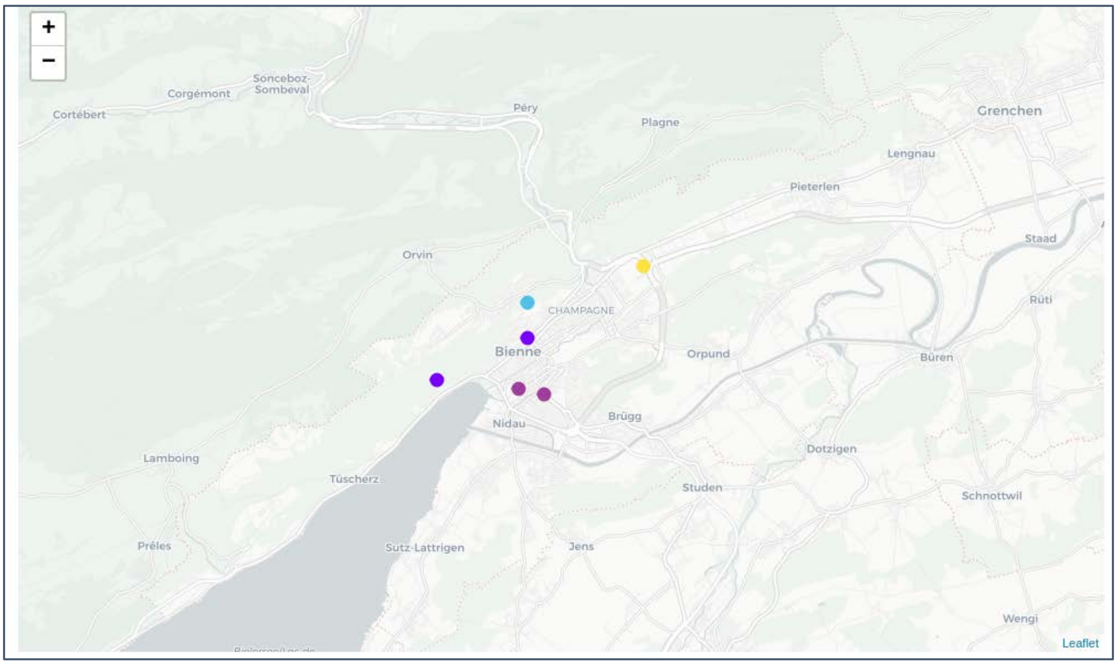

Schüpberg
Münchenbuchsee
Moosseedorf
Bäriswil
Meikirch
Krauchthal
Kirchlindach
Innerberg
Zollikofen
Bremgarten
Ittigen
Wohlen
Ostermundigen
Stettlen
Frauenkappelen
BERN
Muri
Worb
Köniz
Kehrsatz
Trimstein
Rubigen
Neuenegg
Leaflet


+
−
Hettlingen
Dägerlen
Dinhard
Rickenbach
Frauenfeld
Buch am Irchel
Gachnang
Seuzach
Neftenbach
Dättlikon
Bertschikon
Wiesendangen
Hagenbuch
Mat
Pfungen
OBERWINTERTHUR
rach
Elsau
Wittenwil
WINTERTHUR
Elgg
Aadorf
Eidberg
Hofstetten
Guntershausen
Brütten
Sennhof
Oberwil
Zell
Bassersdorf
Turbenthal
Oberhofen
Ottikon
Weisslingen
Effretikon
Leaflet


+
−
Rotkreuz
Buonas
Sempach
Eschenbach
Gisikon
Meierskappel
Root
Neuenkirch
Udligenswil
Küssnacht
Ebikon
Emmen
Hellbühl
Adligenswil
Greppen
Lucerne
Meggen
Malters
Kriens
Obernau
Horw
Schwarzenberg
ENNETHORW
Vierwaldstä
Leaflet


+
−
Roggwil
Horn
Bischofszell
Häggenschwil
Hauptwil
Goldach
Rorschach
Bernhardzell
Waldkirch
Mörschwil
Wittenbach
Engelburg
Eggersriet
Heiden
Andwil
Gaiserwald
Sankt Gallen
Rehetobel
Gossau
Speicher
Teufen
Herisau
Altst
eim
Bühler
Stein
Leaflet

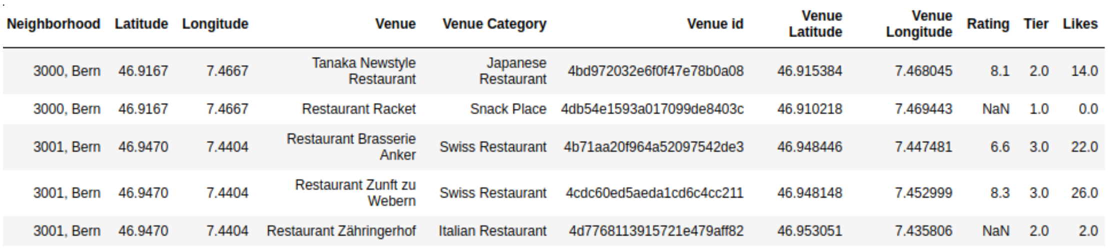

+
−
Bedano
Cor
Canobbio
Porlezza
Albogasio
Aranno
Astano
Bioggio
Massagno
Lugano
Gandria
Monteggio
Agno
Lago di Lugano
Pura
Scaria
Laino
Collina d'Oro
Lanzo d'Intelvi
Campione d'Italia
San Fedele Intelvi
Cadegliano
Pigra
Melide
Marchirolo
Brusimpiano
Cerano d'Intelvi
Marzio
Brusino Arsizio
Erbonne
Schignano
Leaflet



Searching for restaurants

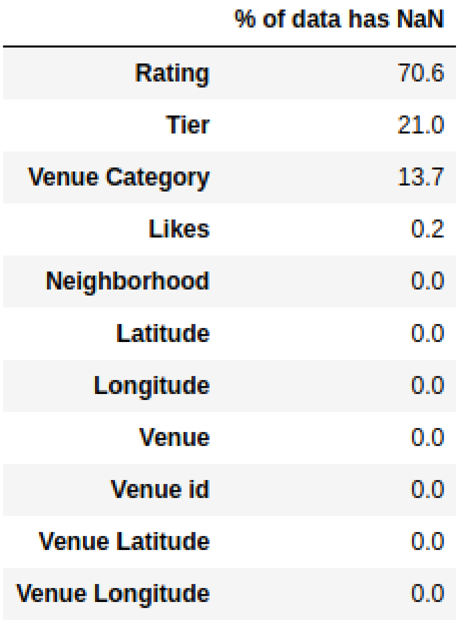
In this part of the analysis, a specific query was made. Venues with a keyword restaurant were pulled and their unique ID was used for yet another query. The last query was to obtain the details of each restaurant. Characteristics, such as number of likes, rating and price tier were extracted from the API.

The same data-frame with zip codes was used. All the points from all the cities were queried for restaurants within **1km** radius. The same code could be applied to any other venue type, such as hotels, bars, etc.



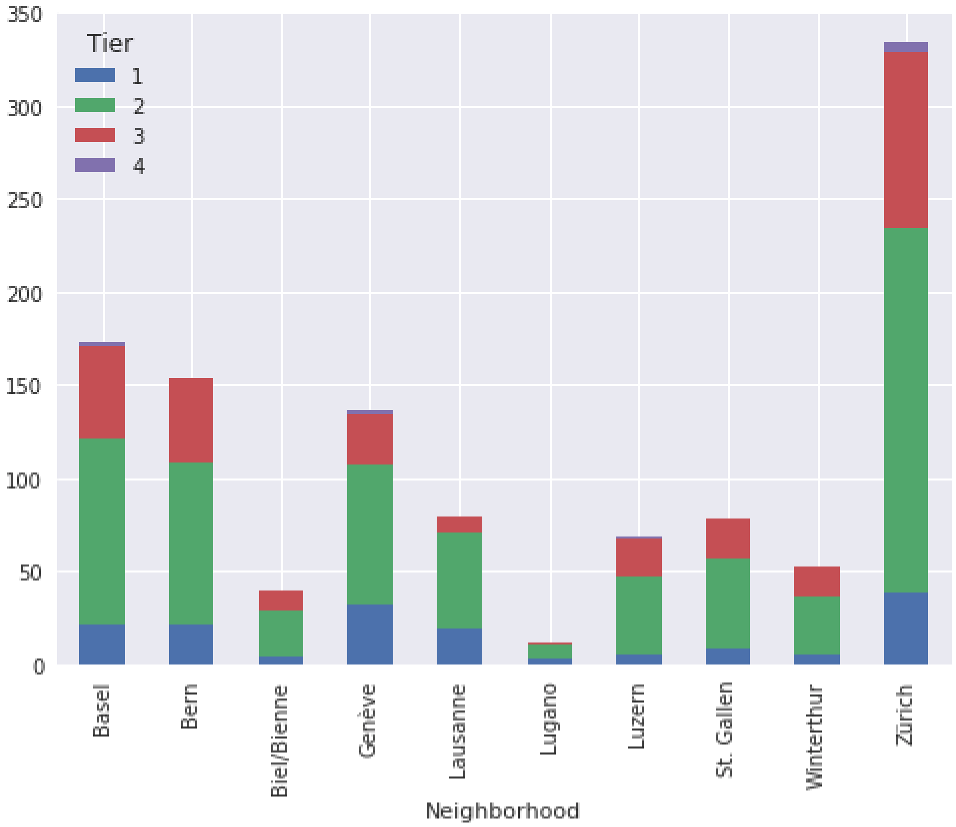
Neighborhood	Latitude	Longitude	Venue	Venue Category	Venue id	Venue Latitude	Venue Longitude	Rating	Tier	Likes
3000, Bern	46.9167	7.4667	Tanaka Newstyle Restaurant	Japanese Restaurant	4bd972032e6f0f47e78b0a08	46.915384	7.468045	8.1	2.0	14.0
3000, Bern	46.9167	7.4667	Restaurant Racket	Snack Place	4db54e1593a017099de8403c	46.910218	7.469443	NaN	1.0	0.0
3001, Bern	46.9470	7.4404	Restaurant Brasserie Anker	Swiss Restaurant	4b71aa20f964a52097542de3	46.948446	7.447481	6.6	3.0	22.0
3001, Bern	46.9470	7.4404	Restaurant Zunft zu Webern	Swiss Restaurant	4cdc60ed5aeda1cd6c4cc211	46.948148	7.452999	8.3	3.0	26.0
3001, Bern	46.9470	7.4404	Restaurant Zähringerhof	Italian Restaurant	4d7768113915721e479aff82	46.953051	7.435806	NaN	2.0	2.0

1430 unique venues that correspond to a keyword *restaurant* were obtained. The collected information was not complete, however. Certain values were missing, and the following output represents the percentage of missing values per characteristic.

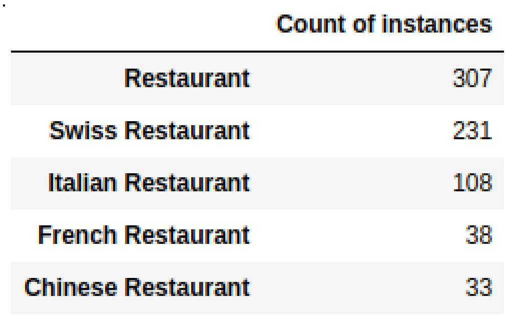


	% of data has NaN
Rating	70.6
Tier	21.0
Venue Category	13.7
Likes	0.2
Neighborhood	0.0
Latitude	0.0
Longitude	0.0
Venue	0.0
Venue id	0.0
Venue Latitude	0.0
Venue Longitude	0.0

Different cities have different number of restaurants, and within once city restaurants are classified into different number of price tiers. The following figure represents the number of restaurants per city, for each price tier.

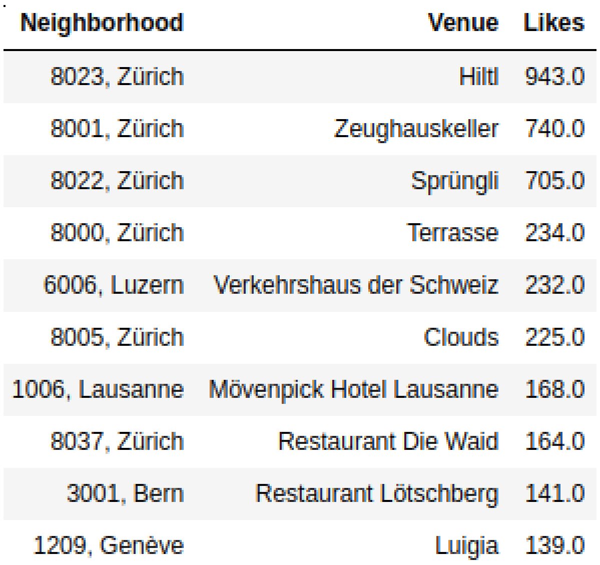


Among the characteristics for each restaurant, their co-responding category was also extracted. Here are the top 5 restaurant categories for the whole country (data from all of the cities).



	Count of instances
Restaurant	307
Swiss Restaurant	231
Italian Restaurant	108
French Restaurant	38
Chinese Restaurant	33

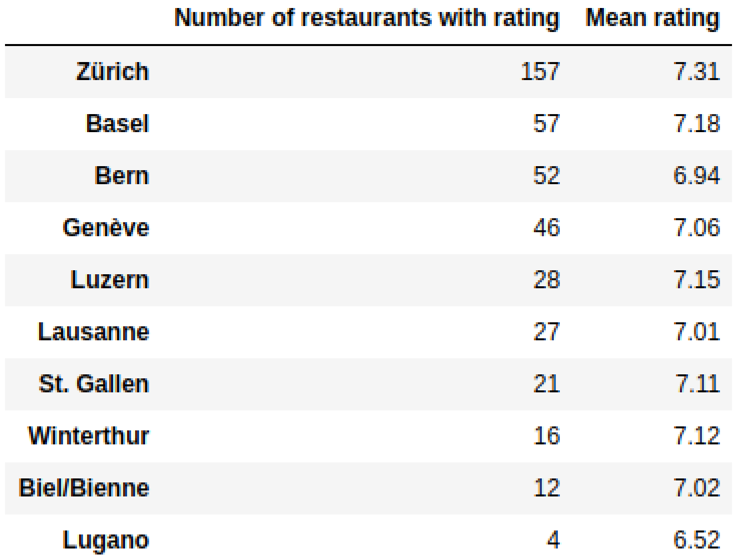
Another characteristic obtained from querying the details of each venue was the number of likes. Here are the top 10 venues (restaurants) regarding the number of likes, with co-responding zip codes and cities.



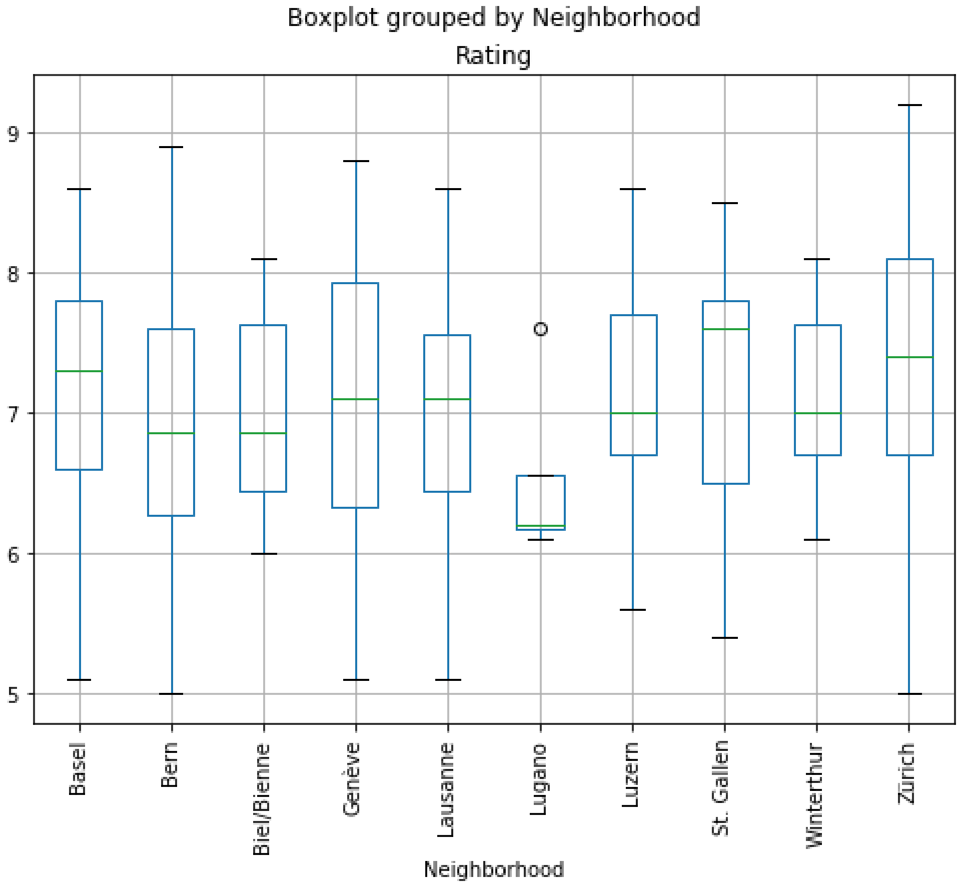
Neighborhood	Venue	Likes
8023, Zürich	Hiltl	943.0
8001, Zürich	Zeughauskeller	740.0
8022, Zürich	Sprüngli	705.0
8000, Zürich	Terrasse	234.0
6006, Luzern	Verkehrshaus der Schweiz	232.0
8005, Zürich	Clouds	225.0
1006, Lausanne	Mövenpick Hotel Lausanne	168.0
8037, Zürich	Restaurant Die Waid	164.0
3001, Bern	Restaurant Lötschberg	141.0
1209, Genève	Luigia	139.0

When it comes to restaurants, a very important feature is their rating. As seen, not all the restaurants had rating, so the following figure represents the number of venues with rating and their average. Box-plots are a convenient way of visually displaying groups of numerical data through their quartiles. The preceding figure shows a box-plot of ratings per each city (grouped zip-codes).

:



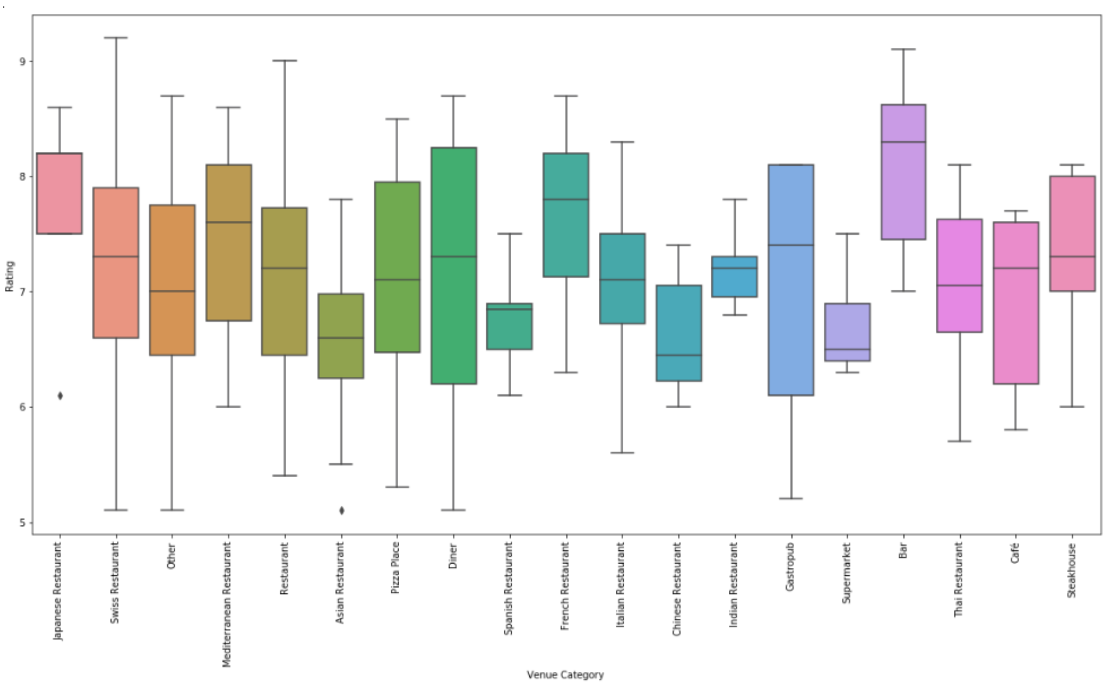
	Number of restaurants with rating	Mean rating
Zürich	157	7.31
Basel	57	7.18
Bern	52	6.94
Genève	46	7.06
Luzern	28	7.15
Lausanne	27	7.01
St. Gallen	21	7.11
Winterthur	16	7.12
Biel/Bienne	12	7.02
Lugano	4	6.52

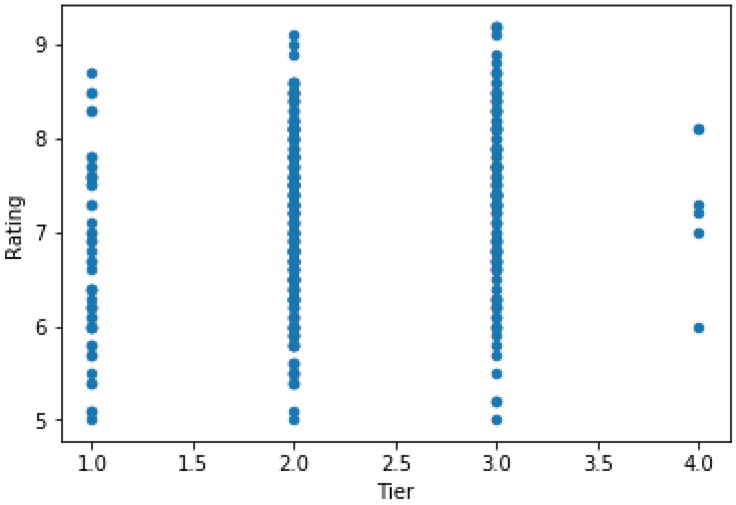


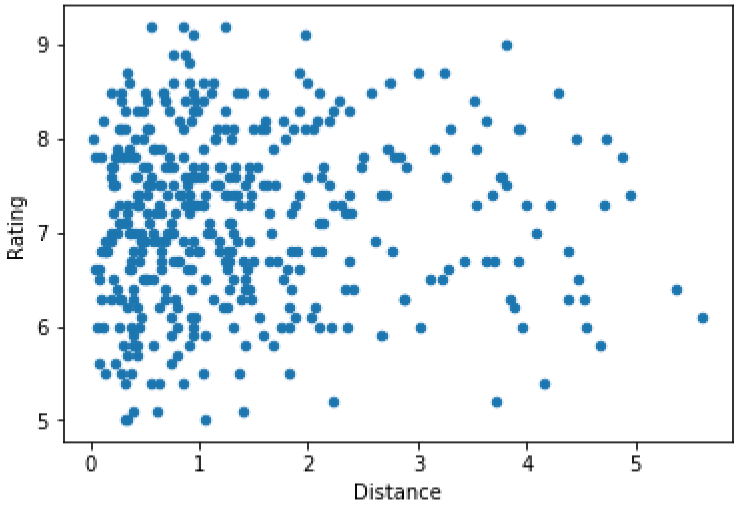
Predicting the rating

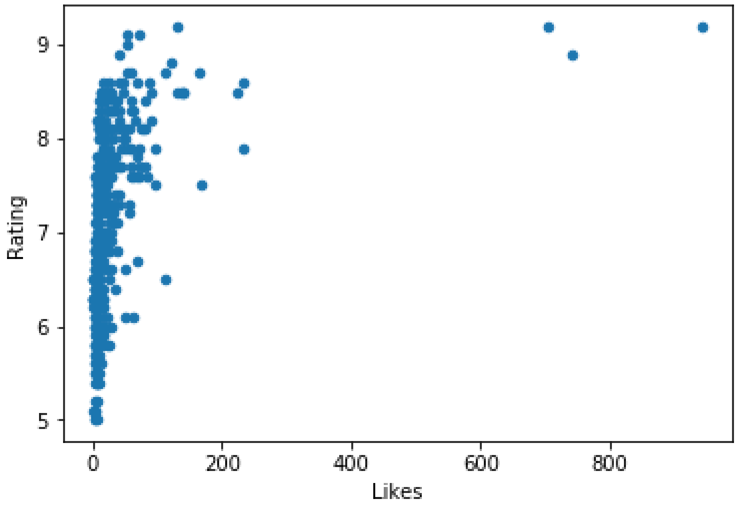
The next section of the report is a trial prediction of the restaurant rating, based on the collected features. As the free membership of the developer Foursquare account allows only limited data extraction, few features were tested in order to predict the restaurant rating. Those features include: *number of likes*, *restaurant categories*, *price tier* and *the distance from the city center*.

The following figures show how each of the features influence the label: rating. The images are in the following order: categories, tier, distance, likes.









The outcomes will be discussed in the further sections of the report.

Discussion

Looking at the most common categories of venues from the Swiss cities, we can see that the most prominent ones are **hotels**, **restaurants** and **bars**. Switzerland is a mountainous Central European country, home to numerous lakes, villages and the high peaks of the Alps. Its cities contain medieval quarters, and the country is also known for its ski resorts and hiking trails. Therefore, Swiss tourism could be a possible explanation of such data. Moreover, Foursquare is perhaps used more by people from other countries, who visit Switzerland and leave their check-in footprint, that eventually overcomes the rest of the check-ins in number. The category of restaurants is not surprising either, as Italy is in the neighborhood, and Swiss restaurants are expected to be visited often.

When we look at the most common venues per city, we see few interesting things. Biggest cities of the German-speaking part of Switzerland have a **Swiss restaurant** as the most common venue (Zurich, Bern, Luzern), while Geneva, being in the French-speaking part, has **French restaurant** as the first most common venue. **Hotels** are showing up as the most common places in Basel and Lugano, followed by **Italian restaurant** (Lugano is in the Italian-speaking part of the country). It is evident from the data that Lausanne people, myself included, prefer **bars** over other venues. Smaller cities have **supermarkets** as commonly visited places.

The KMeans algorithm aims to partition n observations into k clusters of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields. In this example, the algorithm has been used to cluster areas of the city according to venues, and to find any similarities between them.

Using 10 clusters for the analysis, it appears that zip code points can be categorized by looking at the most common venues around. The four main clusters are therefore representing the following common places:

- **Cluster 0:** Tourist points, such as hotels, cafes and other leisure points, including Italian and seafood restaurants.
- **Cluster 1:** Shopping points, such as supermarkets, shopping malls, gas stations, and some public transport and food points.
- **Cluster 3:** Traveling points, such as bus, tram and railway stations with accompanying fast food places (steakhouses, pizza, BBQ places, etc.) and some restaurants.
- **Cluster 6:** Central area places, such as bars, restaurants, cafe places and many other spots of leisure activities (boutiques, music venues, art venues, parks, etc.)

Most of the cities show a similar pattern: cluster 6 in the central areas, while cluster 3 and cluster 1 are in the periphery. This is natural, as all the leisure activities are happening in the central areas of the city, while the points of commute and shopping are situated in the periphery. Certain cities show the "touristic" side dominating, by having cluster 0 within their borders.

The two very similar cities from the French-speaking part of Switzerland: Lausanne and Geneva, have rather similar clustering. The difference is that central Geneva contains a "touristic" cluster, while Lausanne clusters suggest shopping venues are more common. A small city of Biel/Bienne, where both French and German are spoken, has a rather similar clustering. Clusters of the capital, Bern, suggest more venues with leisure activities across town, with traveling points, supermarkets and hotels around. Luzern and Basel, despite the difference in size, have similarly colored clusters. However, Basel has a very specific cluster that includes cultural and animal-loving venues. The biggest city of Switzerland, Zurich, is clustered in a typical way with leisure venues (bars, restaurants, etc.) all across the city center, and commuting and shopping points dominating the periphery. Winterthur and Sankt Gallen each have a particular cluster, that they share with Lugano (electronic stores and hotels as the most common venues, respectively). That makes Lugano, a single city from the Italian-speaking part of the country, the most diverse city in the clustering analysis. In fact, it contains all the clusters, excluding the two specific ones (Lausanne and Basel). Apart from the central leisure venues, the periphery contains clusters with traveling stops and stations. Lugano is the city with most prominent touristic clusters with hotels as the most common venue.

The case of restaurants

When it comes to the obtained data for restaurants, it is important to mention the unavailability and missing data. *Rating* had around 70% of data missing, followed by *price tier* (20%) and *category* (15%). The trial of rating prediction was performed using the remaining 30% of the data with rating.

From the data, we can see that the most common restaurants across the country (and within each city) are the ones with **price tier 2**, where 1 is least pricey, and 4 is most pricey. The top 5 categories of the collected restaurants are, unsurprisingly, the general category: **restaurant**, followed by **Swiss**, **Italian**, **French** and **Chinese** restaurants.

It is important to see where the places with the largest **number of likes** are, and possibly investigate them further. Sorting data by the number of likes across country shows the most liked places are mainly in Zurich. This city is the city with the biggest population in the whole Switzerland, and that explains the outlying number of likes. **Rating** is very similar across towns in the country, being around **7.1**, while Zurich pops up as the city with the highest rating average.

The prediction

Normalized features, mentioned in the previous sections, were used for testing several regression techniques in order to predict the rating of the restaurants. GridSearchCV was used in order to find best parameters for each regressor, and to optimize the score of the testing data-set. Linear regression in combination with polynomial features, Ridge, support vector regression, multi-layer perceptron, and k-nearest neighbors regression were tested unsuccessfully. Further investigation brought to the conclusion to use the number of likes only, considering the relationship with the outcomes (rating). After such decision, isotonic regression was trained and tested, giving an r2 score of only 0.14. Better performance (r2 = 0.24) was obtained using a custom function, based on the sigmoid curve with optimized parameters. It differs from the sigmoidal in that for higher values of the independent variable, the function does not plateau, but increases slowly (the increase in number of likes increases chances that the rating will be higher). Prediction possibilities might be exploited using much more data and different features. Nevertheless, a general overview of the numbers implicated in choosing the restaurant (number of likes, rating, tier) is sufficient for having an idea about the surrounding market, expectations and certain business decisions.

Conclusion

There is a growing popularity of smart connected devices, and one of the widely used features of mobile applications is location awareness. Mobile users take their devices everywhere and using the location services adds to their benefits by serving the right content. That makes location data is a valuable asset in researching anything regarding the best location for a new business, market and target. It entails insights around the audience movement, and if a business leader wants to maintain a competitive edge in the market today, they need to understand and leverage the location piece of their data with Location Intelligence. Benefits are apparent: identify new consumer markets, improve marketing strategies, analyzing marketing and sales activity, improve customer service, etc.

Visualization can change the way that we look at data and information. If that data contains a geospatial component, then utilizing location information can help provide a new layer of insight for certain kinds of analysis. In this project, location data was used to obtain and overview of preferences and activities across Switzerland, compare the biggest cities and investigate restaurants in the country. Very similar analysis could be used for any country, city or area, for any type of venue, not just the restaurants. Possibilities are endless.

Petar Stupar

Lausanne, Switzerland

September 15th, 2018.