Coursera Capstone Project - Battle of Neighborhoods

Nikolas Anic

August 4, 2020

Introduction - Business Problem

In this project, I will follow the footsteps of a successful Japanese franchise chain, which opened up its gates in Toronto, Canada, and has since expanded throughout the its origin country as Ill as the USA. Although established there, the owners plan on expanding business operations to Europe. As a former Austrian, one of the owners proposes to open up the first franchise location in Vienna, the capital of Austria.
In order to assess the profitability of said strategy, I am required to perform an analysis of districts in Vienna to reach the most promising geographic location for our venue. Especially, I am interested in a neighborhood that:

1. Is economically viable
2. Has a strong amusement area and is preferably located in a tourist area
3. Is frequently visited
4. Has not already existing strong Japanese food scene
5. Has AirBnB available and fairly even rental prices

In order to make the correct assessment, I will use a geo-location approach and analyze each district of Vienna according to our pre-defined characteristics and requirements.

Data usage

Data gathering process

For this project, I will use three different types of data repositories. On the one hand, I will require geolocation data from FourSquare that enables us to retrieve locations as Ill as types of venues for the coordinates of all 23 districts of Vienna. Also, I will segment the districts based on demographic, economic and sociog-raphic characteristics. The types of characteristics I are looking for include proxies and indicators for real estate developments, tourist characteristics, general economic status of inhabitants as Ill as infrastructure and traffic situation. Lastly, I will require some useful proxies for local and tourism demand by looking at the availability and distribution of Google searches from Google analytics as Ill as AirBnB offerings throughout the 23 districts of Vienna.

In total, I will retrieve data from three different sources available for a period of at least one year up to a decade.

Open data from public and private sources

I retrieved the following information from public and private sources with regard to socio-economic and cultural characteristics:

1. Rental prices per sqm
2. Growth of housing in the last decade
3. Two factors of AirBnB availability (if AirBnB is available at all (more than 5 offers) and if it is commonly used (more than 50 offers)
4. Gross income median
5. Google searches via a real estate platform, grouped into five bins
6. An indicator if the area is considered a tourist area, defined by the city of Vienna, tourist department
7. A survey response for the frequent availability of public transport

The logic behind choosing the above indicators is that they serve as established proxies to assess the extent to which the prerequisites formed in the first section are met for each observed area. As an example is household growth and median household income per district a valid indicator to measure economic viability as well as household demand structures and tourist indicators serve as good proxies to assess tourism prevalence.
The link with the entire csv file can be found on my GitHub account. The repositories from which I retrieved my data are cross-referenced and last updated on the 28th of July, 2020.

FourSquare API

In order to retrieve the venues per district we will use the FourSquare API. This API retrieves a dataset of existing venues within a given location. For each venue, variables defining geographic location, venue category, venue type and name as well as information to the attractiveness of the location are given.

Calling the API will retrieve 609 unique locations throughout all 23 districts of Vienna. This dataset will serve as the basis for our exploratory analysis and visualization practice. The exact code of the API can be found in my notebook.

QGis Application

As it is my aim to perform a clustering strategy in which I will assign a numeric value to each neighborhood based on the characteristics previously introduced, I am required to retrieve geo- and topographic information of Vienna through a GeoJson file. To do so, I access the data catalogue of geographic borders from the city of Vienna and download the Feature SHP file. This file contains the coordinates of each district and neighborhood of Vienna. Then, I apply the repsective layers into a program called QGis, a Geographic information system

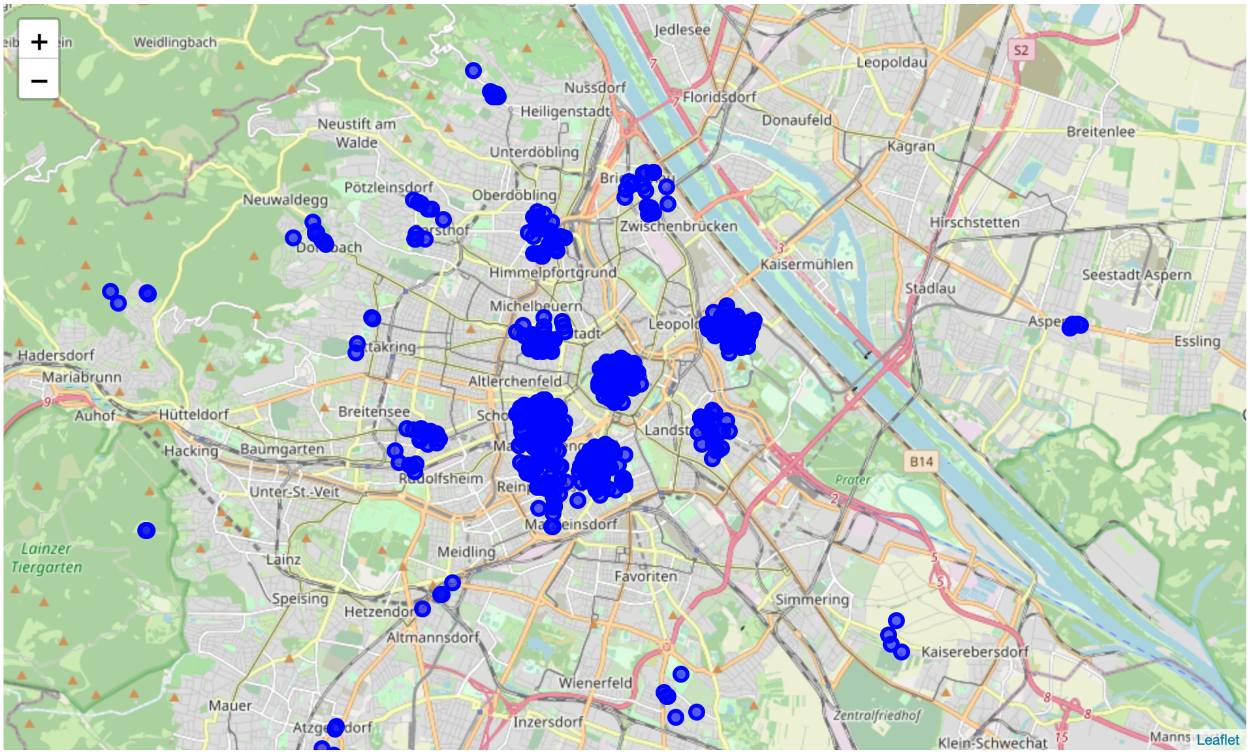

Figure 1: Vienna venues per neighborhood district via FourSquare API

for visualization purposes. The tool is able to transform the SHP file into a district map. Saving the file as GeoJson format allows me to access it through my Jupyter Notebook and use it to create a choropleth chart based on the factorial values I assign to the program. The exact approach was found on a Medium Articleby A. Gordon, which is linked in the text.

Data cleaning and wrangling process

After downloading the venue data from the FourSquare API, I create a function that displays a dataframe consisting of each venue for the respective neighborhood. In order to ease indication, I assign each neighborhood a distinctive group index, which can later be used for exploratory analysis. Further, I visualized the exact locations of each venue in Vienna, which is made available in Figure 1.
Next, I include the socio-economic indicators for each venue, based on its respective neighborhood location, and remove unnecessary or doubled columns.

Lastly, I create and add an dummy variable indicating whether the FourSquare API retrieved Asian cuisine within a specific neighborhood. Doing so, I assign five types of restaurant into this category:

1. Asian Restaurants
2. Japanese Restaurants

3. Vietnamese Restaurants
4. Thai Restaurants
5. Sushi Restaurants

I then create a column called "Asian Cuisine Available", which takes for each row the value of one if one of these restaurant categories is operating within a neighborhood, and zero otherwise. This process will help me to make a crucial difference based on the availability and prevalence of Asian cuisine.

Taken all steps together, I created a dataframe which consists geographic and categorical identification characteristics per venue as well as socio-economic indicators for each neighborhood the venue is operating in as well as a culinary dummy variable. I will only use this type of data to proceed in the methodology section. Especially, I will use the following type of information:

1. Neighborhood name
2. Neighborhood coordinates in long- and latitude
3. Postal Code
4. Grouping Index
5. Rental prices per sqm
6. airBnB availability either above 5 (available) or 50 (highly used)
7. Gross Income
8. Google searches for rental flat ranked
9. Tourist Area inidicator
10. Good Public Transport indicator percent
11. Asian cuisine available

Methodology

In the project, I will assess the profitability of Vienna's neighborhoods with regard to our Japanese franchise restaurant. As indicated in the Business Problem Section, we focus on three requirements to make a verdict.

First, the neighborhood or district cannot be saturated within the food or beverage market. This implies that we are required to find a region which either does not offer what the company is trying to introduce or that, although supply is given, demand for the product is still available. As we are unable to measure the latter (for now) we focus on the first.

Second, the neighborhood or district must be frequently visited. This implies that the region should be located in an amusement area which is frequented preferably by both, the local consumers as well as tourists. This can be measured by analyzing the density of restaurants, bars and other venues as well as tourist attractions.

Third, the price level of our offering must suit the average income for the respective region. Especially, we cannot introduce an offering with prices highly above the paying ability of the societal environment. Although this is especially hard to measure, a potential solution may lie in the analysis of average rental prices, if available. Also, the availability of services such as uber or airbnb may lead to a better understanding of the respective socio-economic status of the individual regions.

To satisfy these requirements, we will follow the subsequent steps.

Analysis of the FourSquare data

In a first step, I will look more closely at the FourSquare data to establish a distribution of venues per neighborhood. By following simple data analysis commands in my Jupyter Lab, we can assign dummy variables to each venue category and sum up the amount of categories per neighborhood. This allows me to get an overview of the exact number of venues per category as well as neighborhood. An extract of the dataframe can be found in Figure 2. Note that the Group Index [GrpIdx] indicates the respective neighborhood in an alphabetical order, where Alsersgrund is indicated 1 and Wieden 23.
Further, I add again the socio-economic characteristics and geographic indicators per neighborhood and retrieve a dataset consisting of 23 rows and 968 columns.

Distribution and prevalece of venues per neighborhood

Now that I know which venues are available in which area, I am interested in finding the distribution pattern for venues per neighborhood. Especially, I would like to know which venues are most prevalent based on the FourSquare output. Patterns that may be interesting here include combinations of culinary, cultural as well as gastronomic venues, as they may indicate a prevalence of frequent tourism. On the other hand, locations with solely cultural but non-gastronomic venues bear the potential that demand for culinary expenditures is still not fully met in the respective area. The exact output of the data can be found in Figure 3. As an example, let us look at the highlighted row in blue for the neighborhood of Florisdorf. As we can see, this neighborhood offers a wide range of culinary programs. Next to gastronomic venues, such as Pubs and Fast Food Restaurants, the area, apparently, is known for its vine cultivation. This conception may be regarded as potentially positive or negative for a Japanese-style restaurant chain. On the one hand, the prevalence of pubs and vine culture is likely to indicate an increased demand for culinary offerings within the district. Further, the fact that solely English or American-styled (Fast) Food is offered may tell that, although supply of beverage based culinary is available, the food scene is less established.

	GrpIdx	Afghan Restaurant	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	...	Tram Station	Trattoria/Osteria	Turkish Restaurant	Vegeta / Ve Restau
0	1	0	0	0	0	0	0	0	0	0	...	1	0	0	
1	2	0	0	0	0	0	0	0	0	0	...	1	0	0	
2	3	0	0	0	0	0	0	0	0	0	...	0	0	0	
3	4	0	1	0	0	0	0	0	0	0	...	0	0	0	
4	5	0	0	0	0	0	0	0	0	1	...	0	0	0	
5	6	0	0	0	0	0	0	0	0	0	...	0	0	0	
6	7	0	0	0	0	0	0	0	0	1	...	1	0	0	
7	8	0	0	0	0	0	0	0	0	0	...	0	0	0	
8	9	0	0	1	0	0	1	0	1	0	...	0	1	0	
9	10	0	0	0	0	0	0	0	1	0	...	0	0	0	
10	11	0	0	0	0	0	0	0	1	0	...	0	1	0	
11	12	0	1	0	0	0	0	0	0	0	...	0	1	0	
12	13	0	0	0	0	0	0	0	1	0	...	0	0	0	
13	14	0	0	0	0	0	0	1	0	0	...	0	0	0	
14	15	0	0	0	1	1	0	0	1	0	...	0	0	0	
15	16	0	0	0	0	0	0	0	0	0	...	0	0	0	
16	17	1	1	0	1	0	0	0	1	0	...	0	0	1	
17	18	0	0	0	0	0	0	0	0	0	...	0	0	0	
18	19	0	0	0	0	0	0	0	0	0	...	0	0	0	
19	20	0	0	0	0	0	0	0	1	0	...	0	0	0	
20	21	0	0	0	0	0	0	0	0	0	...	0	0	0	
21	22	0	0	0	0	0	0	0	0	0	...	2	0	0	
22	23	0	0	0	0	0	0	0	7	0	...	0	0	0	

Figure 2: Extract FourSquare summary

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Alsergrund	Supermarket	Café	Pharmacy	Pub	BBQ Joint
1	Brigittenau	Bus Stop	Plaza	Supermarket	Bakery	Tram Station
2	Doebling	Austrian Restaurant	Wine Bar	Restaurant	Food	Park
3	Donaustadt	American Restaurant	Indian Restaurant	Food & Drink Shop	Restaurant	Bus Stop
4	Favoriten	Shopping Mall	Italian Restaurant	Smoke Shop	Athletics & Sports	Grocery Store
5	Floridsdorf	Gastropub	Wine Shop	Wine Bar	Vineyard	Fast Food Restaurant
6	Hernals	Italian Restaurant	Athletics & Sports	Construction & Landscaping	Bakery	Austrian Restaurant
7	Hietzing	Scenic Lookout	Yoga Studio	Electronics Store	Food Court	Food & Drink Shop
8	Innere Stadt	Restaurant	Café	Plaza	Austrian Restaurant	Italian Restaurant
9	Josefstadt	Café	Hotel	Supermarket	Bar	Restaurant
10	Landstrasse	Café	Hotel	Italian Restaurant	Gourmet Shop	Greek Restaurant
11	Leopoldstadt	Theme Park Ride / Attraction	Restaurant	Café	Hotel	Museum
12	Liesing	Supermarket	Plaza	Asian Restaurant	Grocery Store	Austrian Restaurant
13	Margareten	Supermarket	Indian Restaurant	Bar	Hotel	Italian Restaurant
14	Mariahilf	Clothing Store	Café	Bar	Restaurant	Ice Cream Shop
15	Meidling	Furniture / Home Store	Pizza Place	Gym / Fitness Center	Light Rail Station	Storage Facility
16	Neubau	Café	Austrian Restaurant	Clothing Store	Coffee Shop	Ice Cream Shop
17	Ottakring	Italian Restaurant	Austrian Restaurant	Wine Bar	Food Court	French Restaurant
18	Penzing	Austrian Restaurant	Hotel	Trail	Mountain	Diner
19	Rudolfsheim-Fuenfhaus	Hotel	Supermarket	Austrian Restaurant	Chinese Restaurant	Jazz Club
20	Simmering	Park	Movie Theater	Playground	Farmers Market	Electronics Store
21	Waehring	Tram Station	Supermarket	Italian Restaurant	Pharmacy	Café
22	Wieden	Asian Restaurant	Café	Restaurant	Hotel	Austrian Restaurant

Figure 3: Most common types of venues per neighborhood

On the other hand, one could interpret that the lack of restaurants is based on the notion that many vine-related venues already offer some type of food or are even vine-based restaurants. Also, it may be open to debate whether such a restaurant class suits into the existing categories, as one may assume that wine-based culinary endeavours request a more exclusive form of restaurant than a franchise-based chain. As a consequence, the assessment whether the respective distribution of venues in Florisdorf poses an advantageous or disadvantageous background for a Japanese restaurant requires further interpretations.

Clustering Approach

Although it is hard to exactly pin down which venue combination may serve as best possible background, one can attempt to find patterns within the data that can segment each neighborhood into a certain class. Based on this distribution, a first selection of certain neighborhoods can be achieved.

In order to do so, I will use a k-means clustered approach. This approach is based on unsupervised analysis and learning methods in which several characteristics will form a pattern structure in the dataset. This pattern structure will then be used to define clusters based on the optimal distance of each combination structure to the mean values (e.g. the smallest error term value). I will perform the analysis with normalized values such that margins remain within similar sizes. In the end, I will obtain six clusters based on the characteristics described above. The distribution of the patterns is as follows:

Cluster	Amount
0	7
1	3
2	4
3	3
4	2
5	4

In a last step, I visualize these patterns with the Folium map as well as my geo coordinates I created earlier with the QGis application. The resulting image can be found in Figure 4.

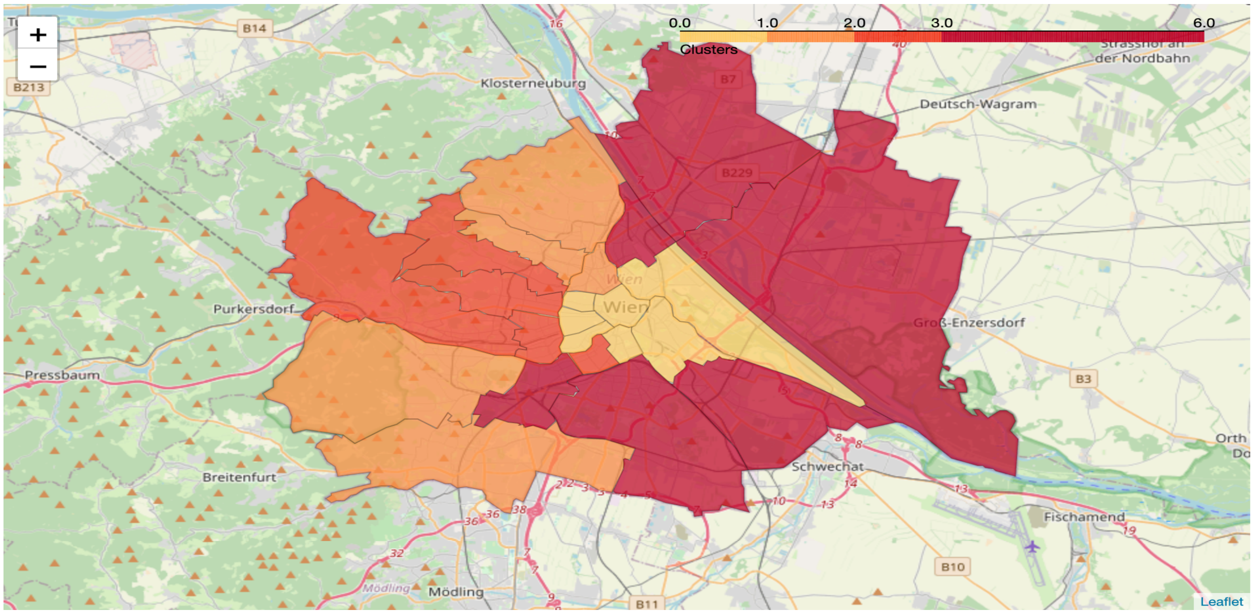

Figure 4: Clustered neighborhoods in Vienna

As we can see, from an socio-economic and cultural perspective, Vienna can roughly be clustered into five segments. If we compare the list-based data with the visualization, we can see that the center, the North-west as well as the South-West, the Central-West and the South and East show the same clustering. Socio-economically and culturally this makes sense, as the central region is both strong economically as well as culturally, the North-West and South-West areas are remote, but have a strong economic and financial background (so-called upperclass boroughs), Western Vienna (central) is pre-dominantly established as middle-class neighborhood with a cultural heritage, and finally Southern and Eastern Vienna have some parts inhabited pre-dominantly by foreign population and are either not culturally viable or are lacking certain socio-economic characteristics, which would both account for a differentiation from the other clusters.
As have our final dataset including the clusters we received with a k-means approach, we can start the selection process an analysis based on our ideas.

Selection of Neighborhoods Results

By assembling and combining all information obtained from every resource, I start selecting the neighborhoods according to the pre-defined combinations of characteristics. I will down-scale according to satisfaction levels for each prerequisite.

I start with the fact that I'd prefer a much frequented area, preferably in a neighborhood that benefits from tourism. As a consequence, I can initially filter according to the tourist indicator. This can also be seen from the clustered visualization. Furthermore, I can combine this approach with the indicators for AirBnB availability, which I retrieved earlier. I may want to filter according to either five or 50 available offerings.

Next, I want to open up the restaurant in an area which is not currently saturated when it comes to Asian food and, as a consequence, competition is less fierce.

Third, I need to position ourselves within a given price range. As defined earlier, I want to offer a Japanese form of Vapiano, implying that price ranges must remain in an affordable, but not cheap range. As I require several quality standards and likely profit from the exclusivity as well as tourist attractions, I can potentially charge a slightly higher price compared to the surrounding venues. Further assuming that the taste for exotic food is more pronounced within trendier neighborhoods primarily inhabited by young professionals and families, I may require that Gross Income should not be in the lower third as well as Rental prices should be above the mean value.

Fourth, I can assess the remaining neighborhoods according to the most common venues. If, for example, a neighborhood already offers a wide range of restaurants, but from a different kind, then this may indicate that, as I already filtered according to the "trendiness of a neighborhood", a commonly known and much appreciated food scene is established in which a new form of taste is likely to be welcomed. On the other hand, the owners could also profit in neighborhoods which would not see restaurants as most common venues, as they could obtain a first-mover advantage and, potentially, lower rental prices.

Lastly, under the condition that I did not find a sufficiently small number of neighborhoods to choose from, one may filter according to google searches of the respective area, since a larger search frequency may indicate that people are more interested in the given location and, as a consequence, are more likely to become aware of the restaurant once advertising commences on specific platforms.

As expected, the range of available districts fulfilling all prerequisites shrinks dramatically to two locations: Donaustadt and Leopoldstadt. If someone already is familiar with the socio-economic trends of the city, this result should come as no surprise. On the one hand, Leopoldstadt is one of the most central districts of Vienna. Home to numerous attractions and historical sites (such as the Prater or residences of the last Viennaise monarch, Franz Ferdinand) the district has been mainly populized by Turkish and Balcan immigrants since the late 1960s. Previously among the poorer cohorts of the city, it has gained large popularity from international institutions and corporations due to its vast, vacant land and affordable ground prices. As a consequence, throughout the last decade, many large companies built their headquarters and offices within Leopoldstadt and three of the largest Viennaise universities built their campuses there. In essence, these trends attracted many employees and students, which increased demand in housing space and different venues, marking a steep trend in gentrification. On the other hand, Donaustadt is the youngest district of the city. Located at the Eastern end, on the opposite side of the Donau, the district mainly consists of new buildings and development areas at a fairly affordable price. Mainly, city planners envisioned an affordable district with all potential infrastructures and amenities required by families and young professionals. Still, its artistic reputation as well as its reputation as venue for exhibitions and trade fairs give the district a somewhat touristic mark. A visualization of both districts can be found in Figure 5.

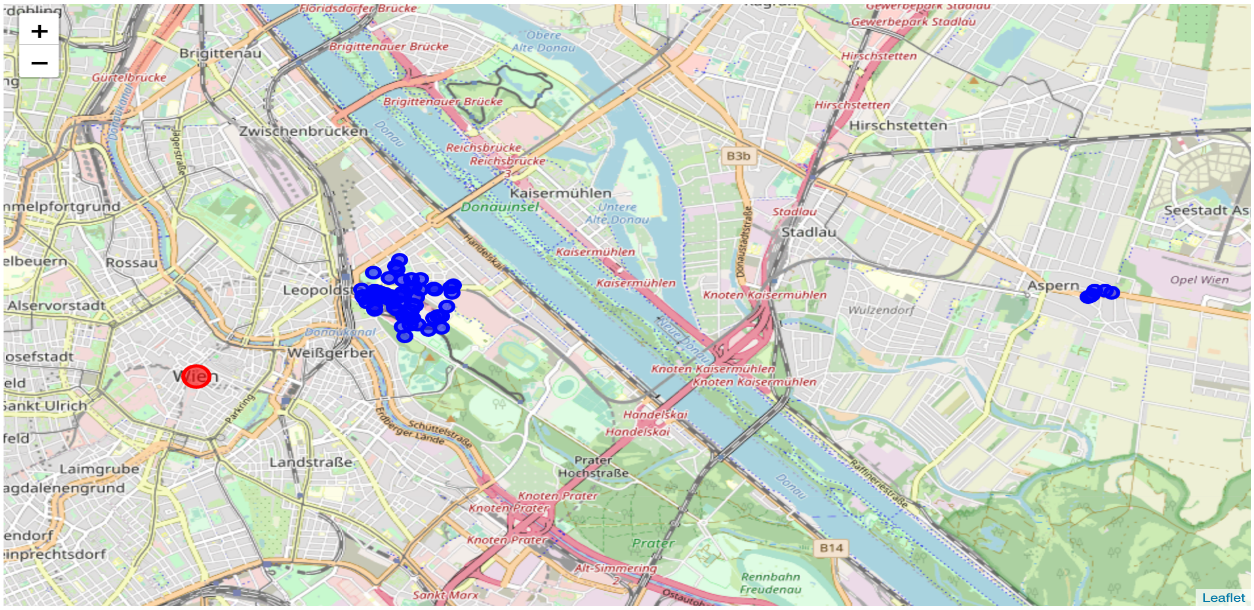

Figure 5: Leopoldstadt and Donaustadt

As we can see, Donaustadt is still a work in progress. Although established to certain parts, it still does not offer fairly many reviews and venues on FourSquare, another potential indicator that the area is not that frequently visited by tourists. Further, we observed in the clustered plot that Donaustadt is in group 4, together with the likes of Southern Vienna. As these neighborhoods are commonly known as socio-economically burdened and pre-dominantly inhabited by people with Islamic-Arabic background, a cultural scene may already exist whose demand for the proposed restaurant type may not yet be established. On the other hand, we can see that Leopoldstadt is both more centrally located as well as better equipped when it comes to venue locations. As we can see in Figure 3, the most common venues within this area are Amusement parks and attractions, Museums, Cafés, Hotels and Restaurants. Furthermore, rental growth increased at 12 percent throughout the last decade and the neighborhood is within the highest Google search category for real estate related topics. Moreover, the recent rise in facilities for public-educational as well as private-economic purposes indicates that the area is highly frequented by thousand of people every day coming from an educated background, being at least partly of young age and, potentially, having a greater interest in foreign cultures and international cuisine. In essence, Leopoldstadt offers an interesting socio-economic, cultural as well as infrastructure-oriented pattern of which we, as a Japanese restaurant chain, are likely to profit.

Therefore, let's look at the area more closely in Figure 6 and, for an interactive imagery, in my GitHub notebook. As we can see, the area consists of an exhibition site, university, three underground railway stations, parks and a wide range of attraction sites. This concludes our analysis. We have found one distinctive area of a neighborhood deemed profitable to open up our Japanese-style restaurant. Within this area, the responsible team can locate respective sites through real estate platforms to obtain the optimal location for the restaurant. However,

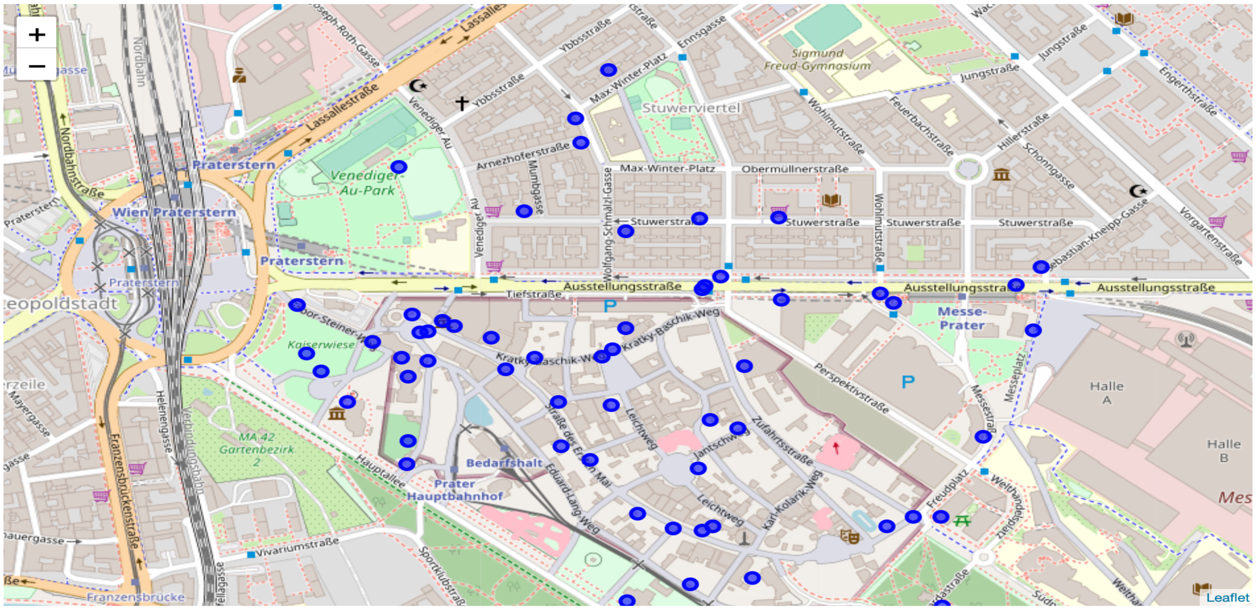

Figure 6: Leopoldstadt neighborhood

although FourSquare was able to deliver good results, it must be stressed that the definitive analysis also has to include personal research for restaurants within the area, as the platform potentially did not cover all venues.

Conclusion and Discussion

The analysis revealed that Vienna has a wide range of cultural and gastronomic venues throughout the entire city. Although important, an exact analysis cannot solely be based on prevalence of existing restaurants within a given area, but is also based on socio-economic and cultural backgrounds of the respective neighborhood. In order to satisfy both restraints, a coherent analysis requires the inclusion of both factors. We satisfied this condition by including indicators obtained from public as well as private analytics sites.

Working with the dataset, we first obtained a list of existing venues based on each neighborhood (in our case: district). We cleaned the dataset and included the additional characteristics. After, we identified existing venues per neighborhood and defined both existence and prevalence of venues within a neighborhood. In addition to geographic location, we then defined clusters based on socio-economic characteristics to portray the differences betweeen neighborhoods deemed to have an impact on demand for the chain's services. This delivered important implications to the socio-economic distributions within the city itself. Then, we combined said distribution with the addditional characteristics as well as the prevalence indiactors of Asian cuisine and restaurant density and narrowed-down the offer to two sites, namely Donaustadt and Leopoldstadt.

Further using analytics data from Google searches and AirBnB offerings, we defined a given location range in which prevalence of cultural sites is existing, removing Donaustadt from our analysis. This step left us with Leopoldstadt, which appears to suit the requirements of the stakeholders in a best possible way.

Please bear in mind that, although this area meets the respective requirements at best, it cannot be guaranteed that the pre-selected choices indeed form a valid basis for analysis. It was our aim to cluster according to certain socio-economic, touristic and gastronomic considerations. These considerations may, however be based on incorrect assumptions. Consequenlty, recommended areas should be considered only as a starting point for more detailed analysis.

In a next step, the district of Leopoldstadt should be analyzed in a more nuanced fashion, by taking in the stakeholder's opinion based on economic considerations (e.g. rental prices, advertising prices, preferred location with infrastructural needs) as well as enhanced methods of location attractiveness through further research.