Capstone – project:

AUSTRIAN CITIES

„Analysis of Foursquare Profiles“

28. JUNI

Datahouse.at
author: Albert

DataHouse.at

Austrian Cities

1 Introduction and description of the business problem

According to the Austrian Business Agency(ABA) foreign visitors to Austria spent a total of EUR 17.4 billion in 2016. Total tourism revenue amounted to EUR 40 billion. The impact of tourism and leisure on the Austrian GDP compared to other sectors amounted to 15.4% in 2018 according to statista/WTTC. Austria holds the 5th place in this global ranking ahead of other European countries like Spain or Italy.

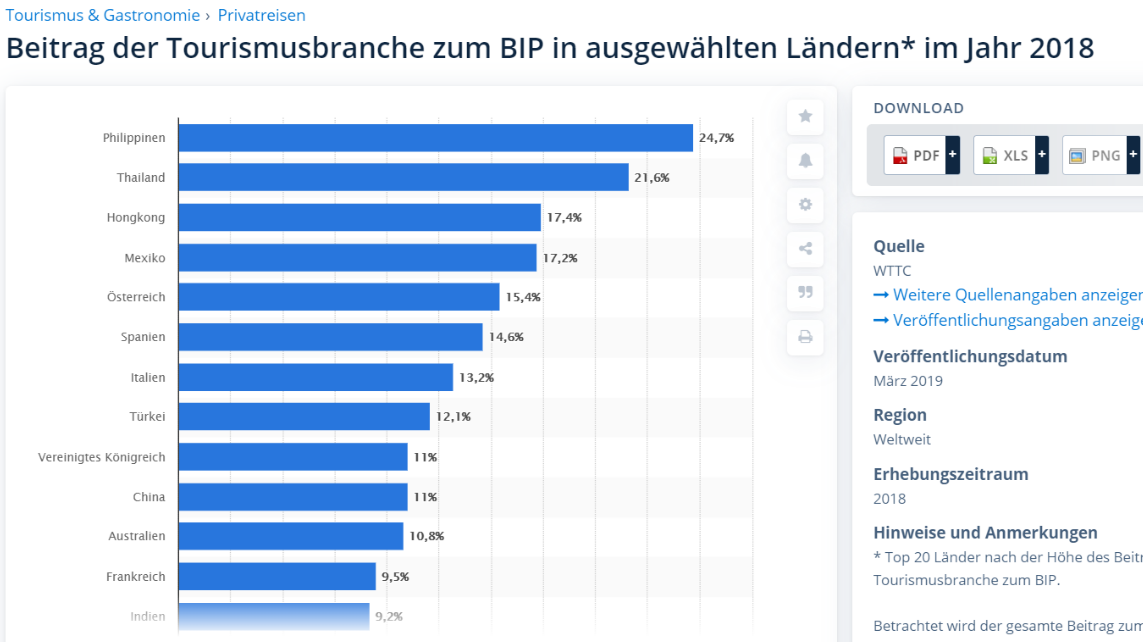


Figure 1: https://de.statista.com/statistik/daten/studie/289171/umfrage/beitrag-der-spanischen-tourismusbranche-zum-bip-im-vergleich/#professional

In Austria tourism is an important economic factor for the whole country but due to the COVID 19 crisis the whole sector is experiencing a temporary downturn. The high internationality of Austrian tourism in an EU comparison is a major part of this current crisis. Most recently, foreign visitors accounted for about 70 % of all overnight stays.

Although Austria has a great reputation as a safe travel destination the country can't benefit from this asset right now as many people are not allowed to leave their home countries due to the global pandemic covid crisis.

The two major selling propositions for vacation and recreation in Austria are:

- outstanding cultural offerings (Amadeus, Sound of Music, Vienna State Opera, …)
- untouched natural environment (the alps for hiking, lakes and rivers, …)

As there is currently a major trend towards domestic vacation this favors the countryside while cities offering diverse cultural highlights might not be able to keep up in the short term.
The goal of this analysis is therefore to compare the different venues (i.e. offerings) of Austrian cities in order to identify different clusters. In a next step the analysis intends to build high-level "citiy profiles" which can be communicated easily to appropriate target groups. At the same time this allows to design and communicate more comprehensive offerings that allow for a "real customer journey".

2 Data and how data sources will be used to solve the problem

After a short research it turned out that there are hardly any API`s available for retrieving coordinates of Austrian cities.
It seems straightforward to use the geopy() package to address this problem. A table of Austrian cities is available on a Wikipedia website.
As the cities are relatively stable it seems not useful to use a scraper for this special task.
Instead of this the basic list has been formatted as a .csv file and is also hosted on github in order for everyone to get this data:

de.wikipedia.org/wiki/Liste_der_Städte_in_Österreich#FN_1

Stadt	Bezirk	Bundesland	Markt seit	Stadt seit	VZ 2001	RZ 2011	POPREG 2019
Allentsteig	Zwettl	Niederösterreich		1380	2.163	1.998	1.808
Altheim	Braunau am Inn	Oberösterreich	1581	2003	4.875	4.784	4.933
Althofen	St. Veit an der Glan	Kärnten	ca. 1230	1993	4.732	4.656	4.678
Amstetten	Amstetten	Niederösterreich	1276	1897	22.595	22.847	23.727
Ansfelden	Linz-Land	Oberösterreich		1988	14.789	15.716	16.654
Attnang-Puchheim	Vöcklabruck	Oberösterreich	1955	1990	8.756	8.862	9.042
Bad Aussee	Liezen	Steiermark	1285 (?)	1994	5.086	4.879	4.862
Bad Hall	Steyr-Land	Oberösterreich	1287	2001	4.752	4.773	5.375
Bad Ischl	Gmunden	Oberösterreich	1466	1940	14.081	13.876	14.126
Bad Leonfelden	Urfahr-Umgebung	Oberösterreich	1356	2001	3.847	4.086	4.245
Bad Radkersburg	Südoststeiermark	Steiermark	1265/67	1299	1.599	1.332	3.154

Figure 2: https://de.wikipedia.org/wiki/Liste_der_St%C3%A4dte_in_%C3%96sterreich#FN_1

The **raw data for the Austrian** cities is available via:
https://raw.githubusercontent.com/schwingi2/Coursera_Capstone/master/cities_at.csv

In a next step the coordinates will be added to the dataset by using the geopy package and geolocate.
An initial assumption that turned out to be unsustainable was the interpretation of the retrieved geopy latitude and longitude values as "center" of the cities. This issue could be solved by adapting the search string of the geolocater and is discussed in the "methodology section"

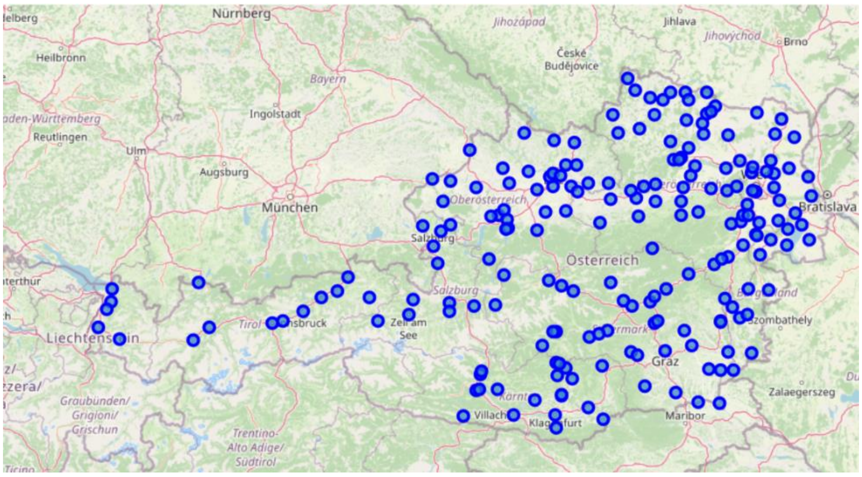
The **adjusted raw data for the Austrian cities** including the appropriate **coordinates** is available via:
https://raw.githubusercontent.com/schwingi2/Coursera_Capstone/master/my_city_coordinates01.csv

The **venues data** is retrieved via the **Foursquare API**
https://raw.githubusercontent.com/schwingi2/Coursera_Capstone/master/my_venues_all.csv

As the free API is limited in terms of API-calls the data was divided into several parts, written as file and finally merged into a complete file.

3 Methodology: exploratory data analysis, statistics, machine learning

The coordinates were used to identify venues within the foursquare API and determine the “density” of locations per city.
In a first step the cities were plotted onto a map. In a next step an analysis of the number (i.e. density) of venues per city was carried out.



This should lead to “high”,”medium” and “low” densitiy cities (large/medium/small) that shouldn`t be compared or clustered identically.

This step requires some additional analysis of the data, but it seems to be quite important as in Austria we have only a few larger cities with a high number of expected venues. That’s why it wouldn`t be appropriate to compare the percentage of pubs in

```
city_densities1.head()
```

[144]:

	Neighborhood	Total_Venues
0	Allentsteig	4
1	Altheim	6
2	Althofen	5
3	Amstetten	25
4	Ansfelden	38

```
[145]: city_densities1.describe()
```

[145]:

	Total_Venues
count	171.000000
mean	28.812865
std	28.816051
min	2.000000
25%	9.000000
50%	17.000000
75%	38.000000
max	100.000000

Vienna (which might be 100+) to the percentage of pubs in e.g. Litschau which might be incidentially one pub that makes up more than 25% of overall restaurants in this small town.
A description of the "density" DataFrame showed a mean of approximately 29 venues per city and a 50% quartile of 17. This deviation indicates that a division into different density categories is also justified by looking at the numbers.

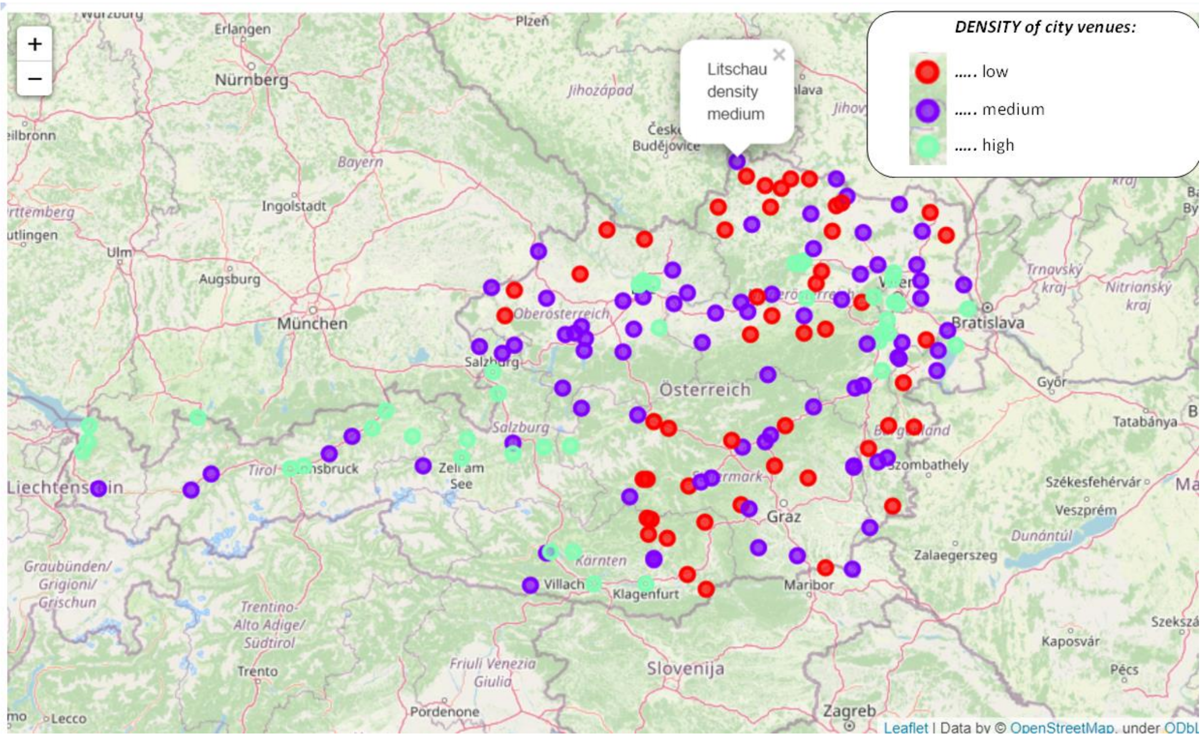
According to the density levels the cities were subdivided into "low", "medium" and "high" density regarding the venues. The "radius" parameter is of great importance for doing this as the "ultimate" center of any town might retrieve many venues and result in a high density whereas this should change with increasing the radius. Hence it was initially set to 7,000 meters.

The density levels were set within the following lines of code:

```
#build three categories: low (25% <=9) medium (75% <=38) high (> 75%) restricted
by max=100 due to API limitations

city_densities1['density']='None'
city_densities1.loc[city_densities1.Total_Venues<=9,'density']="low"
city_densities1.loc[city_densities1.Total_Venues>38,'density']="high"
city_densities1.loc[city_densities1.density=='None','density']="medium"
```

The resulting categories were plotted in folium:



The analysis of venues per city category revealed a clear difference in absolute number of venues. The “high-density” category comprises nearly twice as many venues as the “medium-density” category although the “high-density” cities are only 41 compared to 80 “medium-density” cities.

```
The number of venues is:
  313 ........... for low density cities compared to:
 1579 ........... for medium density cities and
 3035 ........... for high density cities


The number of cities is:
 50 ........... for low density cities compared to:
 80 ........... for medium density cities and
 41 ........... for high density cities
```

For further analysis a KNN approach is used to analyze each density category separately.

The encoding is done with onehot.encoding. Based on this data the DataFrames are grouped by “city” and the basis for KNN is the relative share of each specific venue category in comparison to the total.

If these numbers are summed up across all cities, we get the following numbers for the top venues. As these are the absolute figures and not ratios they should be interpreted merely with regard to the importance of a venue category within a single density level. This is important as we have a different number of cities within each category.

```
**********  LOW DENSITY CITIES Top Venues Scores **********


              Top_Venue_Scores_low
Restaurant                 5.41706
Supermarket                4.54008
Hotel                      4.16667
Train Station              2.82976
Café                       2.73929


**********  MEDIUM DENSITY CITIES Top Venues Scores **********


              Top_Venue_Scores_medium
Hotel                         7.24864
Supermarket                   7.18812
Restaurant                    6.25747
Café                          3.78236
Train Station                 3.35658


**********  HIGH DENSITY CITIES Top Venues Scores **********

                    Top_Venue_Scores_high
Hotel                             4.43448
Restaurant                        3.01256
Supermarket                        2.3515
Café                              2.23123
Austrian Restaurant                1.6498
```

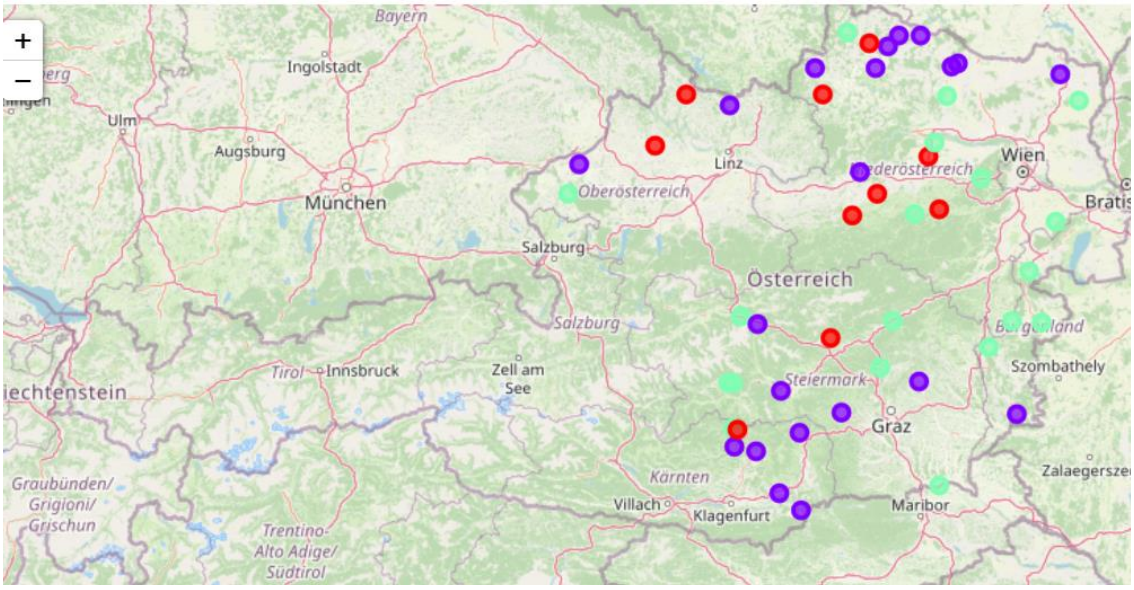
4 Results and intermediate findings

After adding the Cluster Labels to each “density category” as shown in the table below, the clusters are analyzed via a folium map and an “in-depth analysis” for each cluster label.

	Cluster Labels	city	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	l
0	0	Amstetten	Supermarket	Fast Food Restaurant	Gastropub	Grocery Store	Asian Restaurant	Restaurant	Cocktail Bar	Sauna / Steam Room	Pizza Place	Event Service	48.1230
1	0	Ansfelden	Supermarket	Furniture / Home Store	Restaurant	Shopping Mall	Fast Food Restaurant	Asian Restaurant	Park	Hotel	Clothing Store	Buffet	48.2119
2	0	Attnang-Puchheim	Train Station	Supermarket	Multiplex	Museum	Mountain	Furniture / Home Store	Shopping Mall	Gas Station	Electronics Store	Seafood Restaurant	48.0126
3	1	Bad Aussee	Hotel	Restaurant	Austrian Restaurant	Supermarket	Lake	Mountain	Gastropub	Diner	Pizza Place	Resort	47.5990
4	1	Bad Hall	Hotel	Austrian Restaurant	Construction & Landscaping	Pool	Café	Spa	Supermarket	Gas Station	Gastropub	Restaurant	48.0351

Figure 3: example for medium-density cities

“low-density” cities



<u>Cluster “RED”: food, restaurants and supermarkets</u>

Cluster RED is situated mostly in upper and lower Austria. It is dominated by food, restaurants and supermarkets. As we don't have per definition - many observations of venues for low density cities we can`t expect a very reliable result due to a lack of data.

The “**in depth analysis” for Cluster RED** is performed by creating an overview table of all Cluster RED low density cities including the most frequent venues. The description of the Cluster is chosen by the Author and could be subject to text analysis methods in a next step.

	city	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	lat	
7	Groß Gerungs	Supermarket	Restaurant	Construction & Landscaping	Food	Grocery Store	Eastern European Restaurant	Fast Food Restaurant	Food & Drink Shop	Furniture / Home Store	Gas Station	48.573109	1
11	Hainfeld	Supermarket	Train Station	Pizza Place	Plaza	Hostel	Gastropub	Drugstore	Hotel	Eastern European Restaurant	Fast Food Restaurant	48.034881	1
13	Herzogenburg	Plaza	Pharmacy	Train Station	Café	Supermarket	German Restaurant	Drugstore	Eastern European Restaurant	Fast Food Restaurant	Food	48.285907	1
20	Mank	Supermarket	Gastropub	Cosmetics Shop	Pizza Place	Austrian Restaurant	Hostel	Home Service	Hotel Pool	Eastern European Restaurant	Fast Food Restaurant	48.111069	1
27	Peuerbach	Supermarket	Clothing Store	Multiplex	Gym / Fitness Center	Eastern European Restaurant	Fast Food Restaurant	Food	Food & Drink Shop	Furniture / Home Store	Gas Station	48.337974	1
34	Rohrbach-Berg	Supermarket	Plaza	Performing Arts Venue	Paper / Office Supplies Store	Outdoors & Recreation	Gastropub	Eastern European Restaurant	Fast Food Restaurant	Food	Food & Drink Shop	48.572372	1
36	Scheibbs	Supermarket	Italian Restaurant	Restaurant	Construction & Landscaping	Business Service	Dessert Shop	Gastropub	Eastern European Restaurant	Hotel Pool	Hotel	48.003677	1
43	Trofaiach	Supermarket	Grocery Store	Lake	Seafood Restaurant	Drugstore	Eastern European Restaurant	Fast Food Restaurant	Food	Food & Drink Shop	Furniture / Home Store	47.420494	1
	Waidhofen an					Shopping		Eastern	Fast Food		Food &		

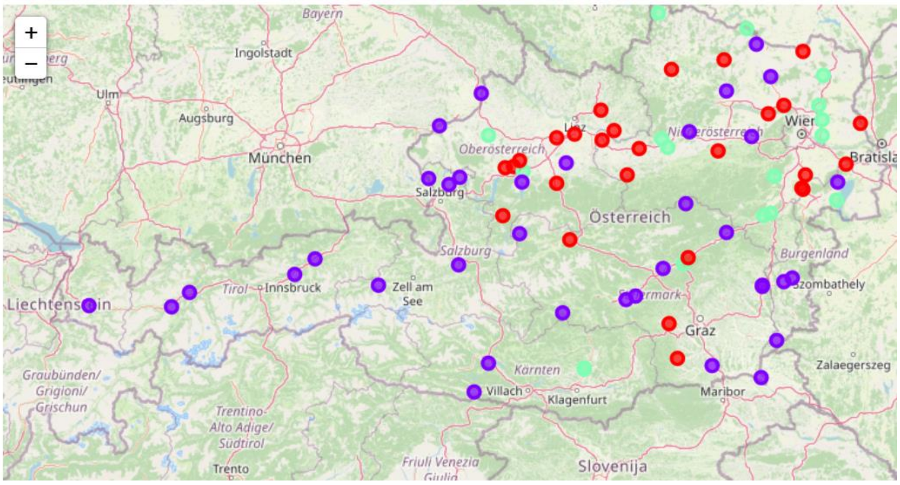
***Cluster “BLUE”:** Hotel-Motel-Cafe, recreation & sports*

This Cluster includes additional categories, especially concerning sports, recreation sometimes theaters or historic sites. These additional venues seem to be accompanied by a higher density of Hotels or Bed & Breakfast

***Cluster “GREEN”:** Train & Travel, Restaurants, Bars & FastFood*

This Cluster is predominant in Burgenland and therefore seems to be geographical cluster at the same time.

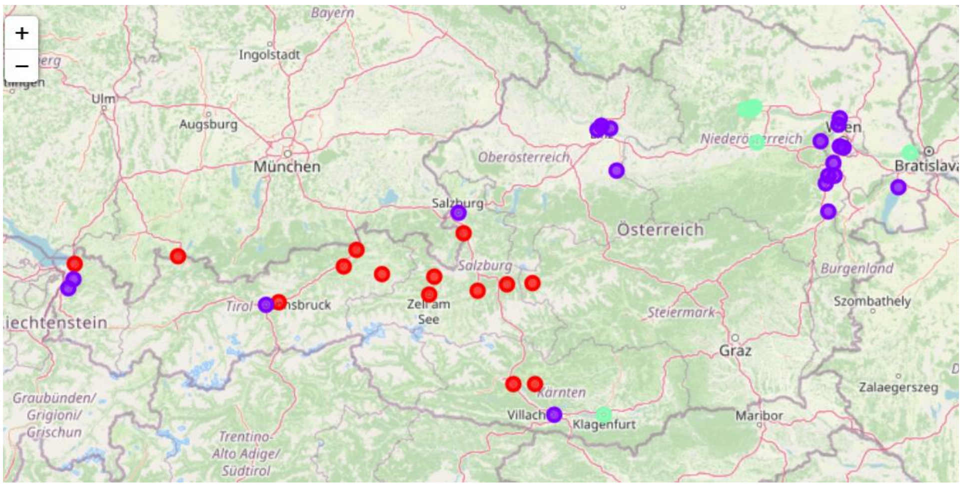
“medium density” cities



***CLUSTER “RED”**: Supermarket, Train Station and Restaurants*
***CLUSTER “BLUE”**: Hotels, Pubs& Cafes and recreation*
***CLUSTER “GREEN”**: Restaurants, Fast Food, Gastro & Shopping*

In this Cluster it seems that some special cases do not fit into this profile. E.g. Litschau with a Golf Course as #1, Border crossing and some other quite unique venues doesn't fit well into this cluster. Rust is very similar despite different venues compared to Litschau it also offers many venues that are unique in this Cluster like ferries, beach or a theme park.

“high density” cities



***CLUSTER “RED”:** Hotel, Sports & Recreation Cluster*
***CLUSTER “BLUE”:** Hotel, Food & Restaurant*

within the high - density cities we experience many clusters where hotels are in the first place. The importance of recreation opportunities, sports and leisure seems less significant

CLUSTER "GREEN": *The unexpected Danube and Lake Cluster*

This is a very small cluster compared to all other clusters. It seems to be a mixture of different venues including historic sites and castles as potential differentiator.

5 Discussion of observations and recommendations

A first interesting observation that does not depend on the knn clustering algorithm is due to the division of cities into density clusters.

Some cities in the low-density category wouldn`t be expected here. This is almost certainly due to missing data of venues in the foursquare API.

For example we can see Waidhofen an der Thaya or Heidenreichstein in this low-density section whereas Litschau is in the medium - section.

It seems that in terms of Marketing Litschau as the smaller city did a much better job here.

A second observation is the existence of geo- clusters. For example we have an accumulation of low density "green-cluster" cities in Burgenland and an accumulation of "blue-cluster" cities in the north of lower Austria. These blue-cluster cities could try to improve their profile by cooperating within their "blue Hotel, sports and recreation" cluster. Especially for smaller cities this could make sense to improve awareness on e.g. social media channels.

On the other hand we can also see geo-clusters for high-density cities around Vienna (blue cluster) and around St. Pölten and Krems (green cluster). As these cities will almost certainly have a professional tourist management and strategies in place it wouldn´t be appropriate to draw immediate conclusions from this "high level" analysis of their city centers. It could be however a starting point for further discussions and analyses that could be performed on the level of districts or neighbourhoods within the cities.

6 Conclusion

On the one hand it's quite difficult to give general advice to cities based on a high-level cluster analysis. On the other hand this analysis might reveal interesting facts about one or another city and may be the starting point for further discussions and initiatives that could be more fact-based.
Almost certainly I would conclude that the availability of appropriate data – e.g. via improving the cities profiles on foursquare could reveal even more facts and improve the reliability of analyses.
I couldn't find any Austrian data providers (social media, APIs, etc.) although lot`s of data should be available in different Austrian chambers and federal institutions.

The current foursquare partner initiatives show the increasing importance of this platform as companies like Samsung, Uber or Airbnb are incorporating it into their products and services.

"Using Foursquare's database of 105 million places, Samsung allows users to geo-tag their photos with detailed info about a destination, while also serving up contextually relevant locations for those who are searching for a place to eat, drink or explore. "

At the same time foursquare is an important source for travel advice. This should be an incentive for Austrian cities to -at least -participate and sharpen their profile on foursquare.