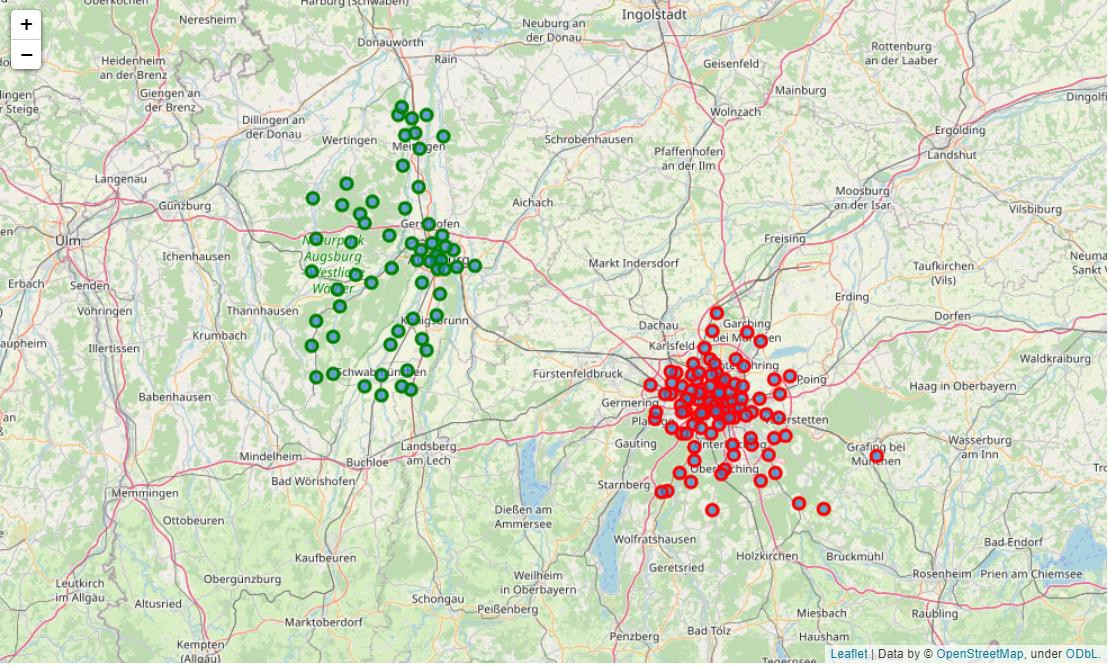
Applied Data Science Capstone Project

# “Comparison of two neighboring cities: Augsburg and Munich, Germany March 06, 2021

1. **Introduction**

One of the biggest dilemmas people face when choosing a place to live is whether to opt for a large city with a significant amount of opportunities but higher cost of living or choose a smaller location. This study aims at comparing two German cities that are relatively close located to each other but significantly differ in size. The goal is to identify similarities and dissimilarities between two cities based on clustered data.



# Data

The data for this study was obtained from the web resource [www.geonames.org](http://www.geonames.org/) where the data is ordered in form of tables and contains the location description in the format of country, region, district, city(location), postal code and coordinates (latitude and longitude). Since the administrative division of regions/ districts varies among different countries, the table has column names “admin1”,”admin2”,”admin3” and so on, which stand there as placeholders. Therefore before we could work with the data we needed to change the column names into the actual terms used for location description in the observed country (Germany).

Upon examining the data it was noticed that any distinct description of districts/ neighbourhoods within both cities is missing, which leads to certain areas having different coordinates and postal codes but having the same name. Since the goal of the study is to focus on similarities/difference between two cities and not city districts, it was decided not to go deeper into these details but simply allocate indexes to these names in order to make difference between them.

Another important thing, which had to be configured is using of the correct encoding of symbols, i.e. UTF-8. German alphabet contains special symbols, therefore certain names were read by the database incorrectly.

Since we are comparing two cities, two separate data sets were obtained for both cities, identically processed and merged for further study.

During the second stage of the study data about venues located within 500 meters from each identified location was collected using Foursquare API and sorted by the most visited places.

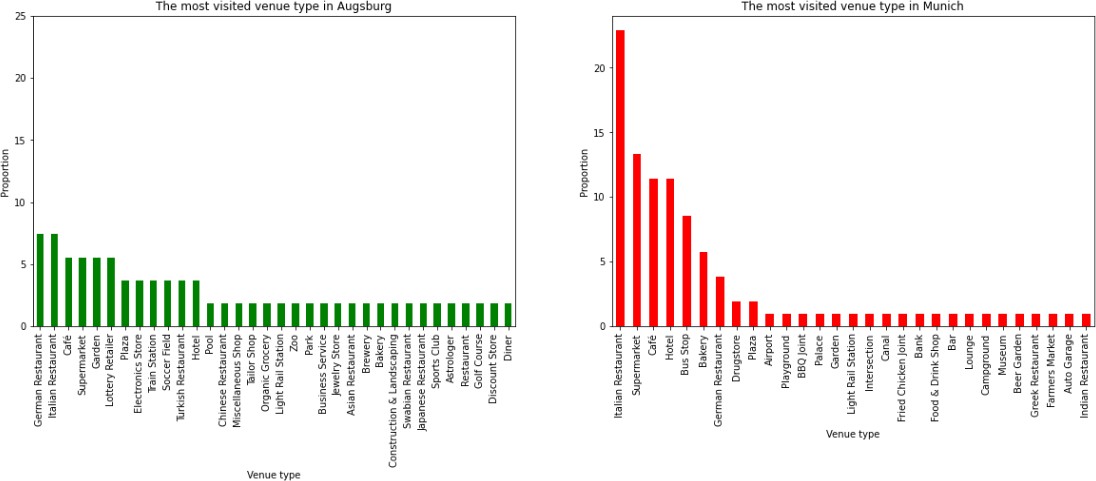
# Methodology

In order to provide a thorough comparison of both cities we created several charts based on different characteristics:

* + First we observed the most visited venues for each location in both regions with a bar chart.
  + Then we created a histogram representing the distribution of the density of places identified within 500 meters from each location, and compared both histograms for every city. We also analyzed the distribution with a box plot.
  + At the end we used the clustering method in order to identify patterns behind these locations.

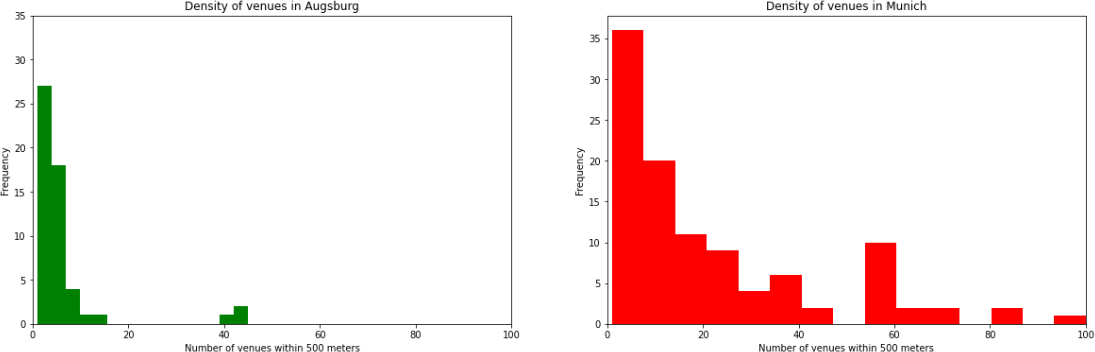
# Results

First we decided to compare the chart of most visited venues for each location in both regions. In order to compensate for obviously smaller amount of observations in Augsburg, as a smaller city, we compared both statistics in percentage.



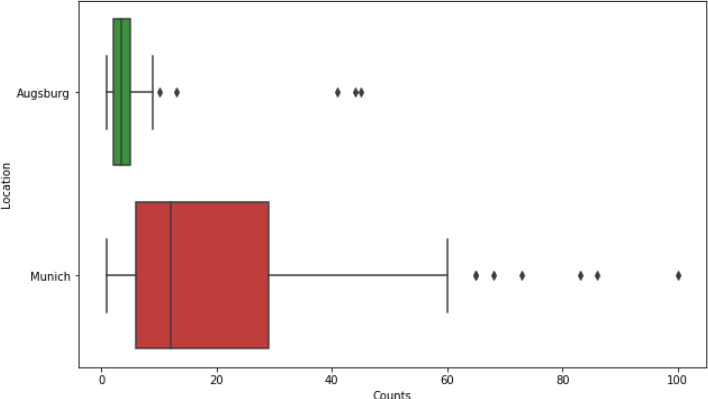
As the plot suggests the results from Munich have a clearer identified preference, while the results for Augsburg are more flat, and the preference is not that well defined.

Another point that I decided to investigate for both cities is the density of locations, which was calculated as the number of venues within a certain number of meters from every location. It was decided to choose 500 meters as the default radius for both cities in order to make the comparison more representative.



As result we can see that despite an obviously bigger amount of observations present in the Munich data, we can see that the mode for both cities lies closer to the minimum. The mode for Augsburg is 1 and for Munich is 4. Both distribution are right skewed and on the right side we see several outlier with significantly larger number of surrounding venues. In both cases such locations are found within the center of the city.

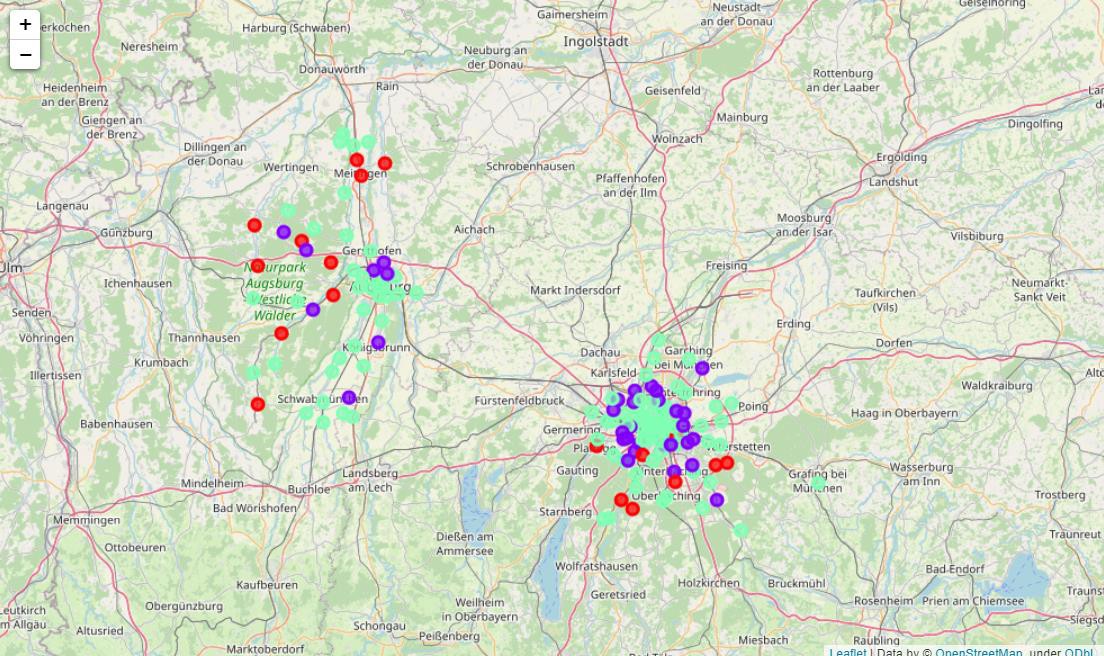
For a more clear understanding we also created a box plot, which clearly shows the distribution for both cities:



As the final step of our study we sorted venues detected in locations across both cities, extracted 10 most visited venues for each location and assigned them a corresponding ranking. Upon that we used the clustering method to identify hidden patterns between all the locations depending on their preferred places of visit.

Upon multiple tries it was decided to set the number of clusters to 3, since smaller amount of clusters is not representative enough, while with a larger amount of cluster locations are not that well distributed, due to the fact that the majority of locations in Augsburg have a small number of venues identified.

The result is presented below:



The first cluster is Purple. It shows locations where the most popular venues are supermarkets, drugstores and bus stops. All these venues are part of daily life and are mostly visited by people near to their place of residence. In the map we can clearly see that most of them are concentrated closer to the suburbs of both cities.

The second cluster is Red. It is represented by venues like German restaurants, fast food, playgrounds, markets. Similarly to the previous cluster all these venues are part of daily activities and we can also observe them closer to the suburbs.

The third cluster is Turquoise. It is represented by hotels, cafes, bars, Italian restaurants, bakeries. Most of these venues are associated with leisure or tourism and are concentrated at the center of the city.

# Discussion

Although the results represent my expectations about the amount of entries for venues present for either city, it appears that people tend to leave reviews for places mostly in touristic locations. So as result I observe certain lack of information about venues in Augsburg, especially in those further from the city center.

# Conclusion

Based on the conducted analysis we can see that both cities, despite their size difference, show similar patterns in inhabitants behavior in terms of visited places: with venues related to leisure time or tourism being concentrated in the center, while venues related to daily activities are scattered closer to the suburbs.

At the same time Munich showed a significantly higher density of venues location around each particular location.