COURSERA CAPSTONE PROJECT

CLUSTERING GERMAN UNIVERSITIES BASED ON NIGHTLIFE AND FITNESS VENUES

INTRODUCTION/BUSINESS PROBLEM

Problem definition

- In this project I want to look at universities in Germany. There are some 100 universities in Germany and it can be a challenging task to choose the right one. There are many reports and lists out there that assess universities based on the quality of their teaching and their reputation. This is of course one useful approach.
- But students do not just want to study all day long, they also want to have fun in the evening be it by going to a bar or excercising in a gym.
- Therefore, I want to look at nightlife and fitness venues in the vicinity of each university. Based on that I want to calculate a metric that ranks universities by *venues or gyms per student*. In addition I will use K-means clustering in order to define different clusters of universities.
- All the above will be visualized in maps.

INTRODUCTION/BUSINESS PROBLEM

Audience

- My main audience is students that are trying to find the right university. In addition to information that is already available in university rankings they could use my analysis to find an institute that has a high nightlife factor.

DATA

Universities (Wikipedia)

- A list of universities in Germany is available on Wikipedia here: [Liste der Hochschulen in Deutschland](https://de.wikipedia.org/wiki/Liste_der_Hochschulen_in_Deutschland "Liste der Hochschulen in Deutschland")
- The first column gives as the name of the university.
- It contains information about the type ("Form") of institution. We will only look at universities ("Uni")
- It also contains information on the number of students ("Studierende") which will be used in our analysis.

Geo data (Google)

- In order to show the universities in a map of Germany we need geo information (longitude, latitude) that is not part of the wiki page. We will use the Geocoder API from Google Maps to obtain these coordinates

Location data (Foursquare)

- The information about nightlife and fitness venues will be obtained from Foursquare. We will focus on a radius of 1000-2000m around the university.
- The venues that we are interested in will be retrieved by using the corresponding category IDs:
  - Nightlife | 4d4b7105d754a06376d81259
  - Fitness | 4bf58dd8d48988d175941735

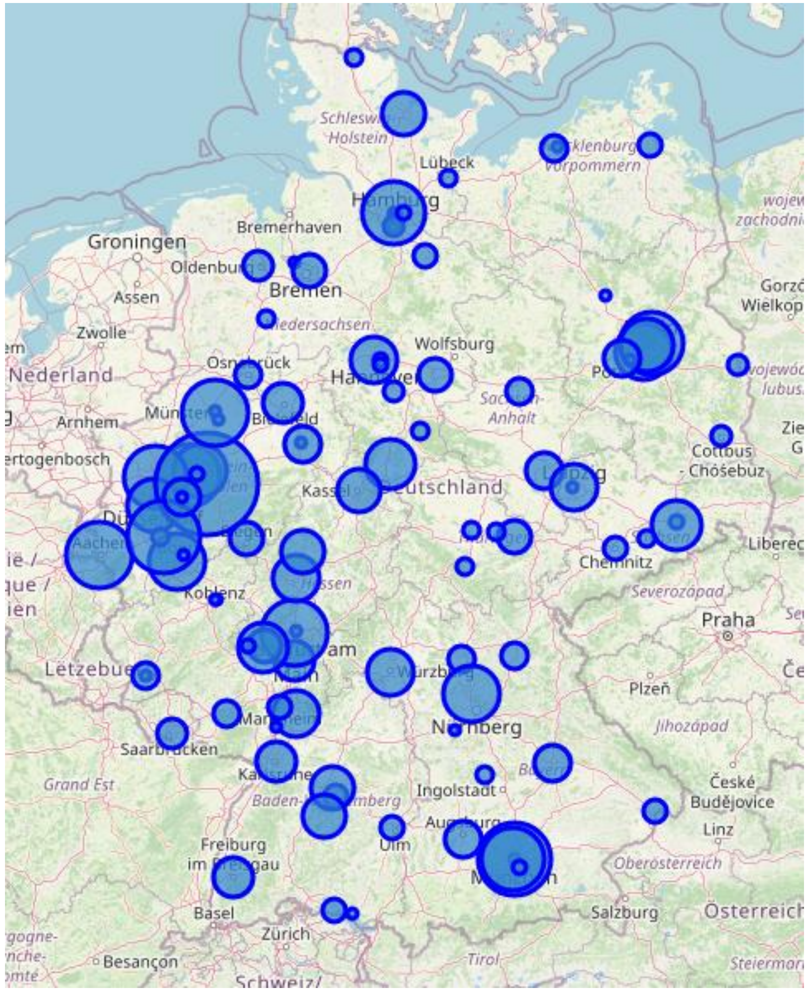
METHODOLOGY

- The _number of students_ (retrieved from Wikipedia) and the _number of venues_ per university (from Foursquare) are used for further analysis. With the objective in mind to identify universities with a high "nightlife factor" (i.e. many venues for few students) a KPI is calculated that puts the two numbers in relation (_Venues per 1000 students_)
- The new KPI is visualized on a map of Germany. The bubble size represents a high number of venues per student.
- As the next step the k-means clustering is used to cluster the universities into 5 clusters based on _number of students_ and _number of venues_

RESULTS

Geographic representation

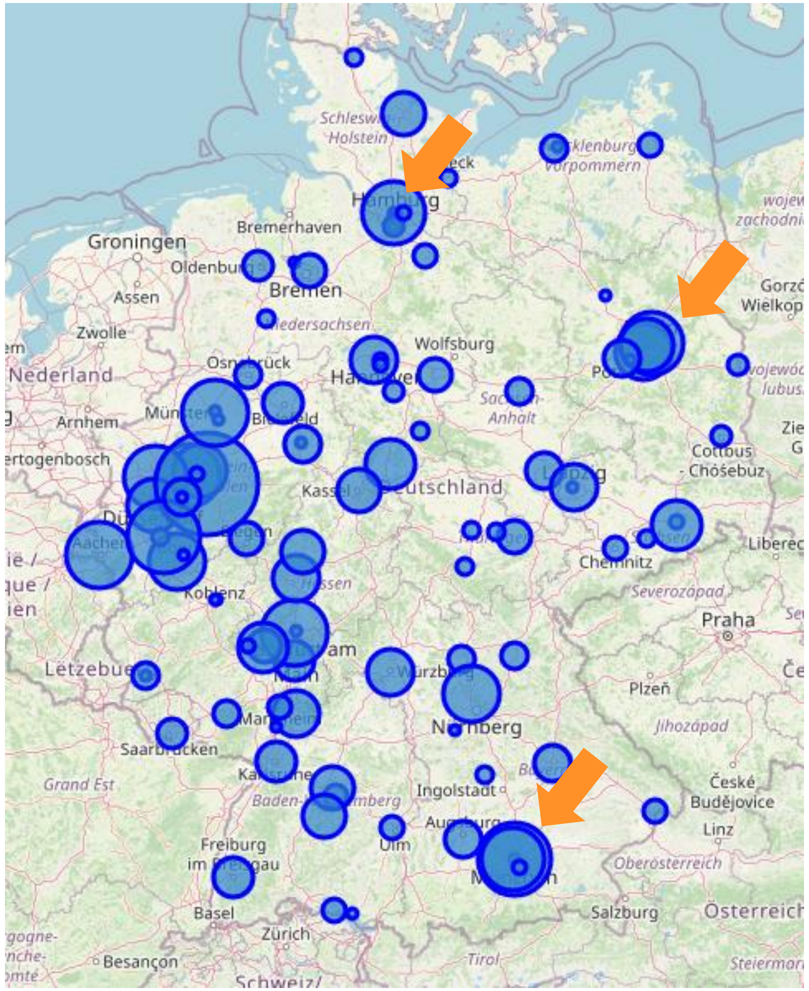
- The map visualization shows a rather even distribution of universities across Germany. There are more universities in the more crowded areas around Munich, Hamburg, Berlin, Cologne and Stuttgart. In Eastern Germany there are relatively less universities.



RESULTS

Size of universities

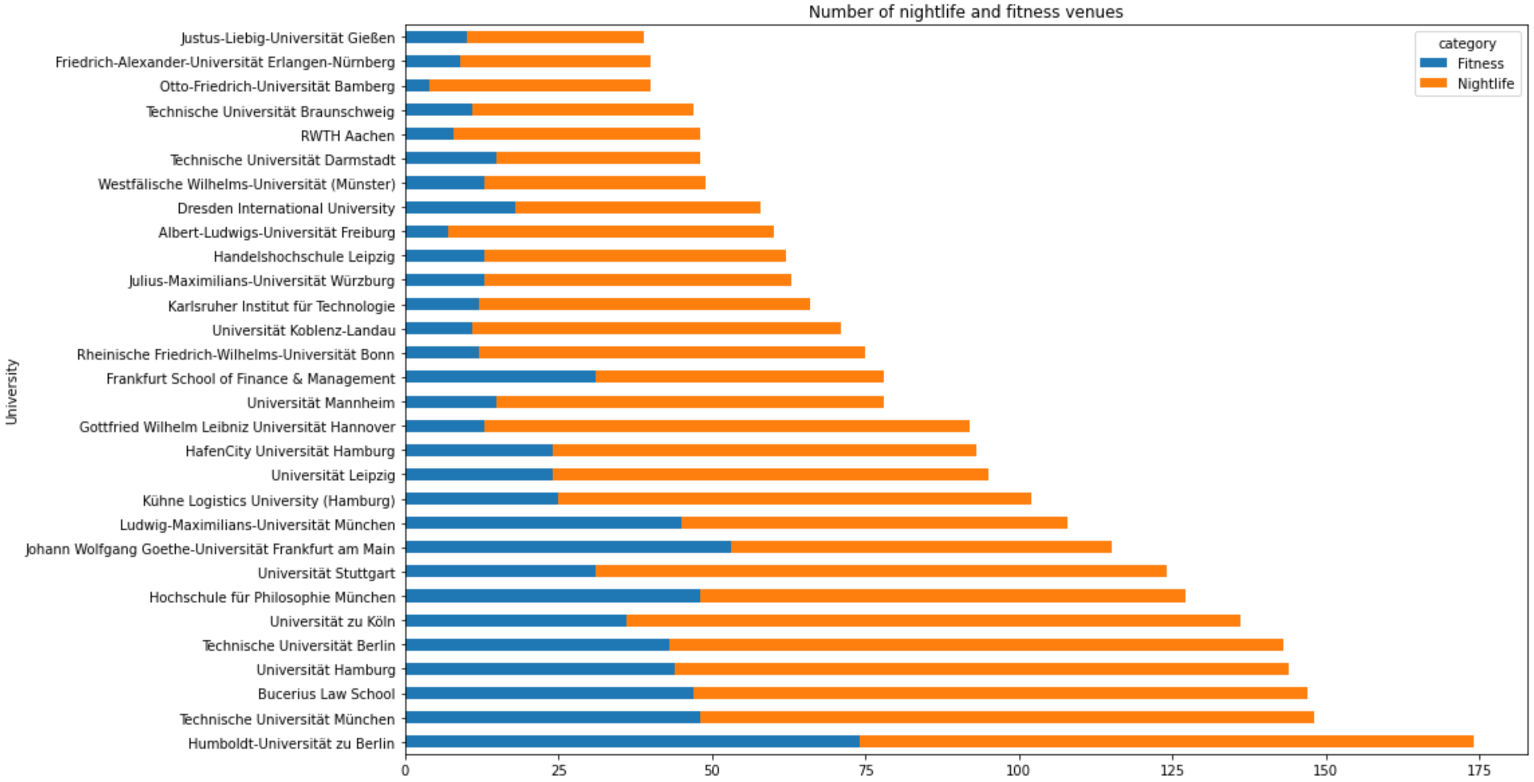
- As one would expect larger cities have larger universities (Berlin, Hamburg, Munich)



RESULTS

Nightlife and Fitness venues

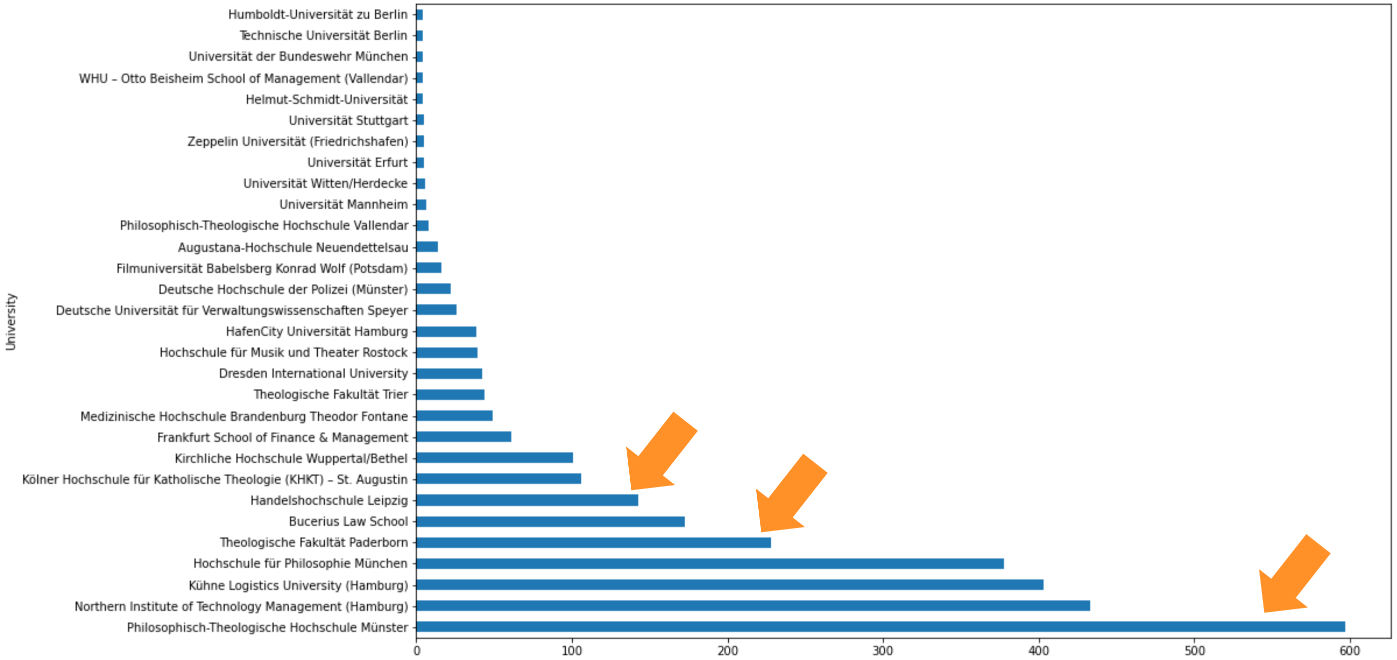
- In absolute terms the large cities/universities have the highest number of venues.



RESULTS

Nightlife and Fitness venues

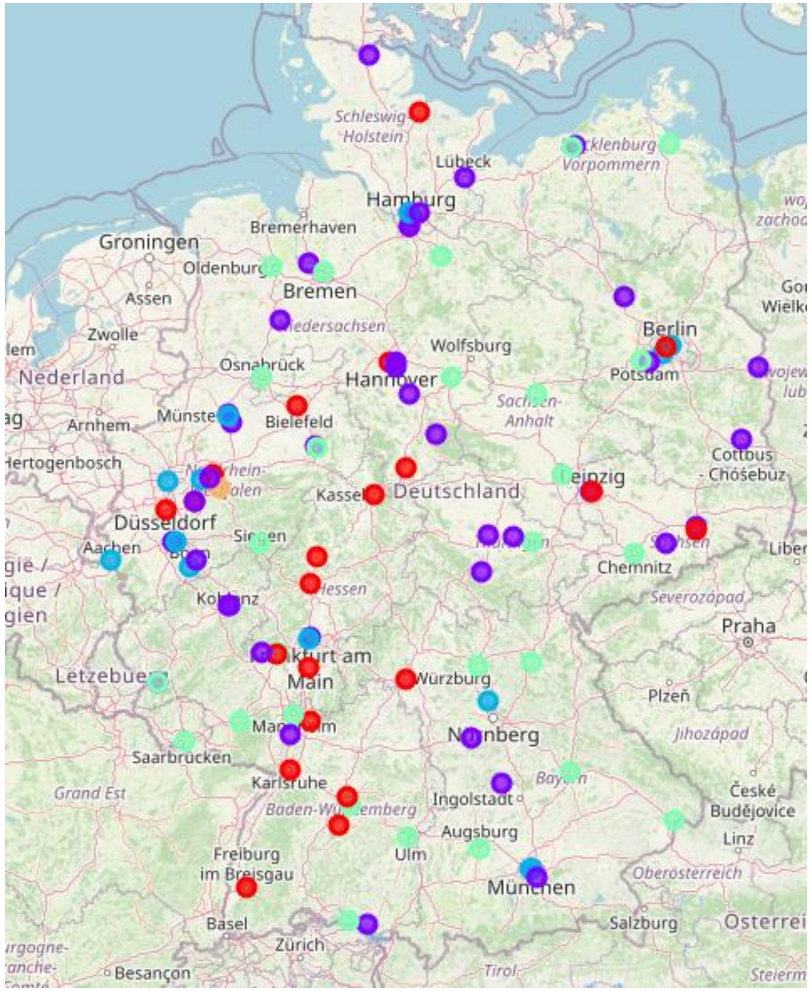
- In absolute terms the large cities/universities have the highest number of venues. However, relative to the number of students there are interesting results that show a high number of venues for fewer students.
- The top 3 among those are Münster, Paderborn and Leipzig.
- We can conclude that larger cities have an attractive nightlife and offering for fitness. However, that has to be shared with many other students.
- There are smaller universities/cities that are equally attractive in terms of nightlife per student.



RESULTS

Clustering

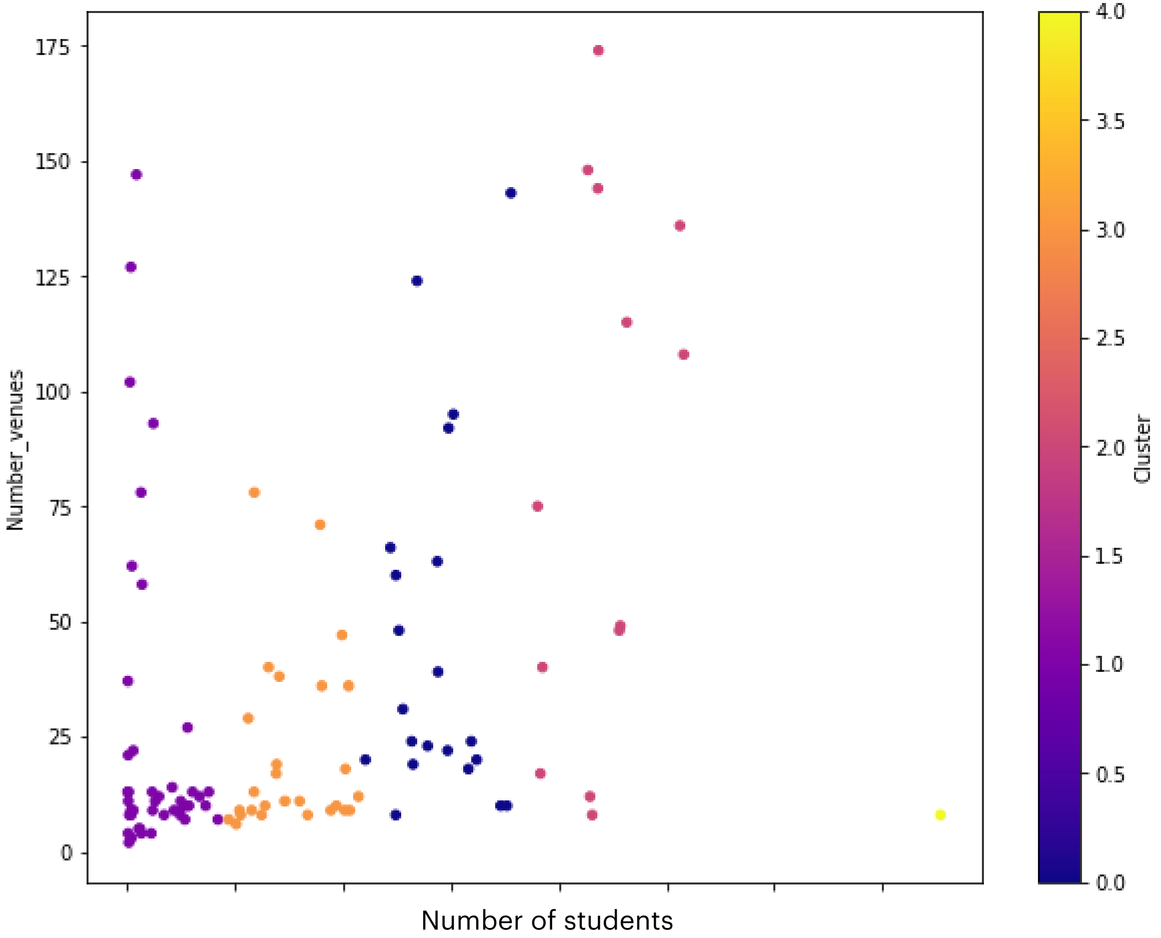
- Applying k-means to the dataset of _number of students_ and _number of venues_ resulted in 5 clusters.
- Further analysis in a scatterplot revealed that _number of students_ had the main (if not only) impact on building the clusters.



RESULTS

Clustering

- Further analysis in a scatterplot revealed that _number of students_ had the main (if not only) impact on building the clusters.



DISCUSSION/CONCLUSION

While some interesting insights could be generated the clustering has not yet led to any meaningful result. Going forward it could be interesting to include additional criteria describing the cities/universities, e.g. the relation of students to total population of a city or a price index for living.