Coursera Capstone Project: Applied Data Science

Overview

- Business Problem
- Data
- Data Preparation
- Machine Learning (Unsupervised Clustering)
- Result

Business Problem

Finding a good place for an open air restaurant:

A good place to place a open air restaurant (german Biergarten) might be a location with high throughput of people locals as well as tourists and might be marked as a place with several parks in urban area and som other outdoor takeaway stuff like cafes or icecream shops etc. An other option could be also that there are several metro stations nearby.

Data

- Districts of Stuttgart (Wikipedia)

	Number	Name	People	Size	People/Size	Districts
0	Innerer Stadtbezirk Nr. 1	Stuttgart-Mitte	23.956	3808	6.294	10
1	Innerer Stadtbezirk Nr. 2	Stuttgart-Nord	27.629	6815	4.054	11
2	Innerer Stadtbezirk Nr. 3	Stuttgart-Ost	48.730	9035	5.393	8
3	Innerer Stadtbezirk Nr. 4	Stuttgart-Süd	44.050	9586	4.594	7
4	Innerer Stadtbezirk Nr. 5	Stuttgart-West	52.668	18643	2.825	9
5	Äußerer Stadtbezirk Nr. 1	Bad Cannstatt	71.285	15713	4.537	18
6	Äußerer Stadtbezirk Nr. 2	Birkach	7.149	3089	2.383	3
7	Äußerer Stadtbezirk Nr. 3	Botnang	13.165	2135	6.166	4

Data

- Geocoding API

	Name	longitude	latitude
0	Stuttgart-Mitte	9.179800	48.775900
1	Stuttgart-Nord	9.176252	48.796661
2	Stuttgart-Ost	9.207365	48.776972
3	Stuttgart-Süd	9.132492	48.753021
4	Stuttgart-West	9.151351	48.777659
5	Bad Cannstatt	9.214680	48.804883
6	Birkach	9.203406	48.728574

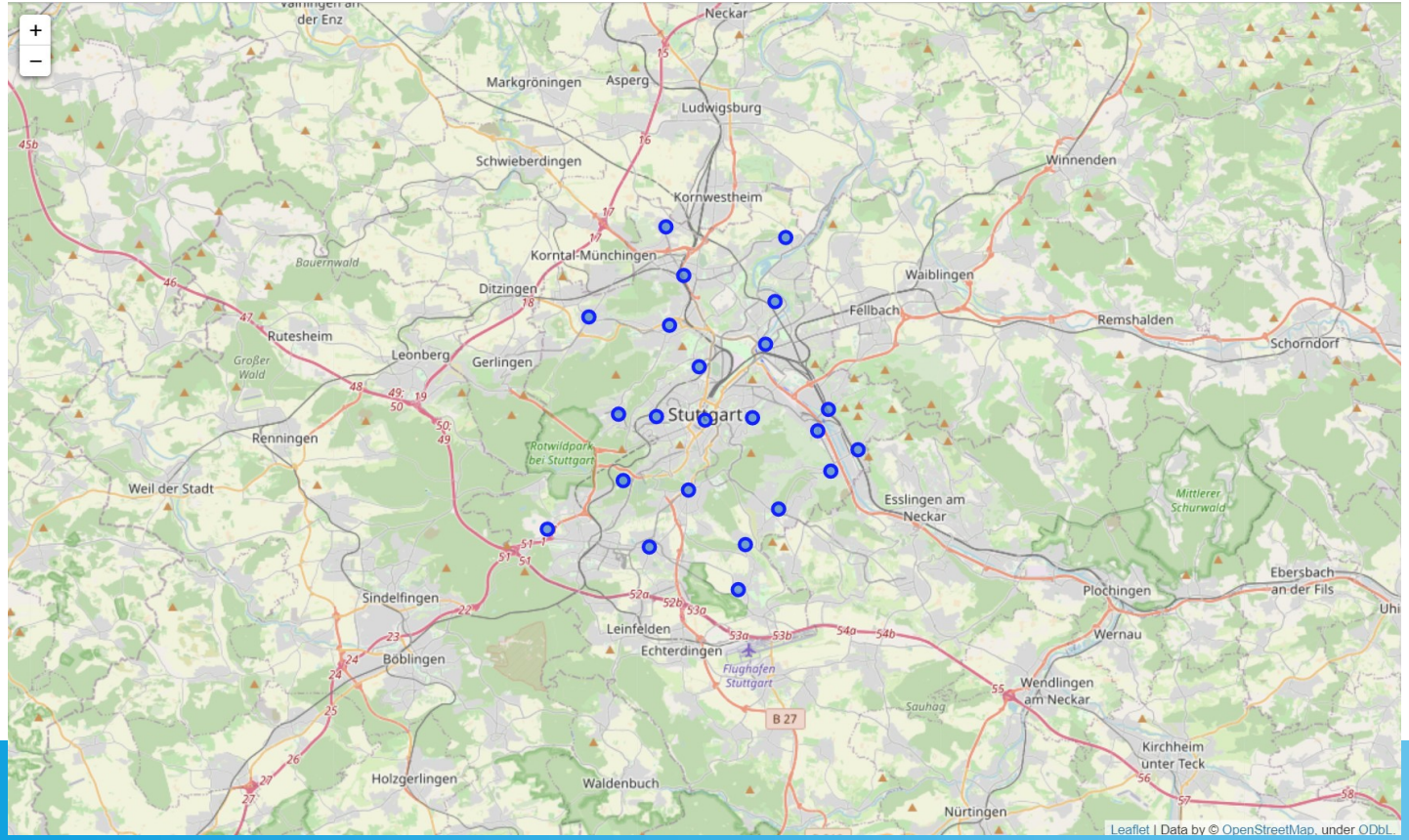
Data

- Foursquare Venues API

	Name	Venue	Latitude	Longitude	Category
0	Stuttgart-Mitte	Schlossplatz	48.778549	9.179855	Plaza
1	Stuttgart-Mitte	Old Bridge	48.774173	9.179274	Ice Cream Shop
2	Stuttgart-Mitte	Markthalle	48.776145	9.179335	Market
3	Stuttgart-Mitte	Bix Jazzclub	48.773178	9.179495	Jazz Club
4	Stuttgart-Mitte	Feinkost Böhm	48.778077	9.176317	Gourmet Shop

Data Preparation

- Visualization Stuttgart Districts

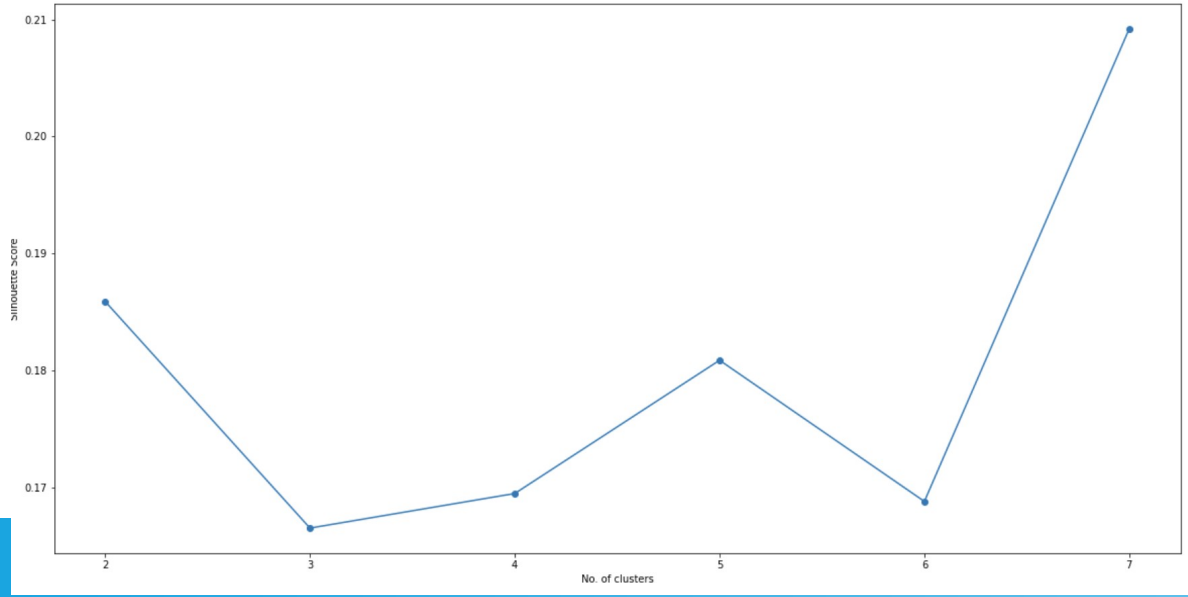


Data Preparation Steps Venues

- One Hot Encoding
  - To support machine learning algorithms there is a column created per category and the data is marked with one if venue is assigned to that category
- Top 10 most common venues
  - Grouping the one hot encoded dataframe by district and then calculating the frequency of the venue category per district after that the most common venues per district can be extracted

Data Preparation Steps Machine Learning

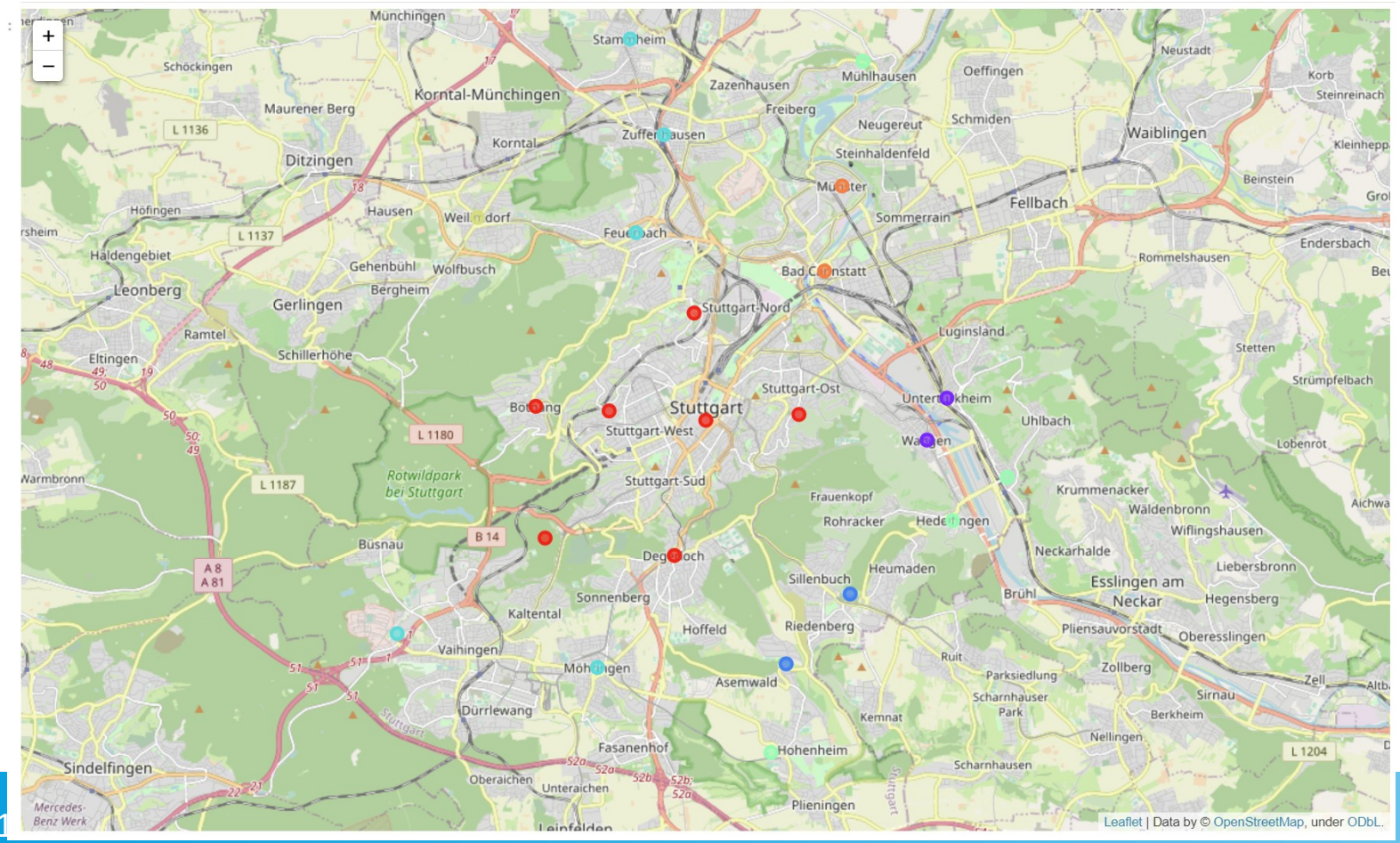
- Ideal Number of Clusters
  - To get the the ideal number of clusters the „Silhoutte Score“ was chosen. It calculats the cohesion inside the same cluster and separation to other clusters.



Clustering

- K-Means Clustering
  - Fast computation
  - Huge number of categories
  - Labeling output data with the computed K-Means Labels

Visualization of Clusters



Discussion

- Relevant Clusters
  - Families (Cluster 7)

	Cluster Labels	Name	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
0	6	Bad Cannstatt	Park	Café	Zoo Exhibit	Hotel	German Restaurant	Restaurant	Pool	Italian Restaurant	Beer Garden	Taverna
8	6	Münster	German Restaurant	Zoo Exhibit	Metro Station	Park	Greek Restaurant	Taverna	Grocery Store	Hotel	German Pop-Up Restaurant	Miscellaneous Shop

  - Tourists (Cluster 1)

	Cluster Labels	Name	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
2	0	Botnang	Italian Restaurant	Supermarket	Park	Plaza	Café	Gym / Fitness Center	Sushi Restaurant	Mountain	Cable Car	Business Service
3	0	Degerloch	Café	Bar	German Restaurant	Cocktail Bar	Hotel	French Restaurant	Ice Cream Shop	Italian Restaurant	Park	Bakery
13	0	Stuttgart-Mitte	Café	Park	Bar	Plaza	Ice Cream Shop	Italian Restaurant	Hotel	Cocktail Bar	Scenic Lookout	Coffee Shop
14	0	Stuttgart-	Café	Plaza	Park	Italian	Hotel	Bar	Restaurant	Scenic Lookout	Ice Cream Shop	Concert Hall

Discussion

- As shown in the previous slide it depends on the focus group if my focus group would be tourists and business people on their way home I would chose the City Center (Cluster 1)
- The other option if my focus group are the locals and families which bring a more consistent revenue stream but are on the other side more price sensitive I would look for alocation near one of the zoos.