Where to go out on vacation?

IBM Data Science Capstone Project

Alexander Hirsch

December 23, 2020

Business Problem & Target Group

Business Problem

- Identify hotspots of restaurants, bars, clubs in a foreign city

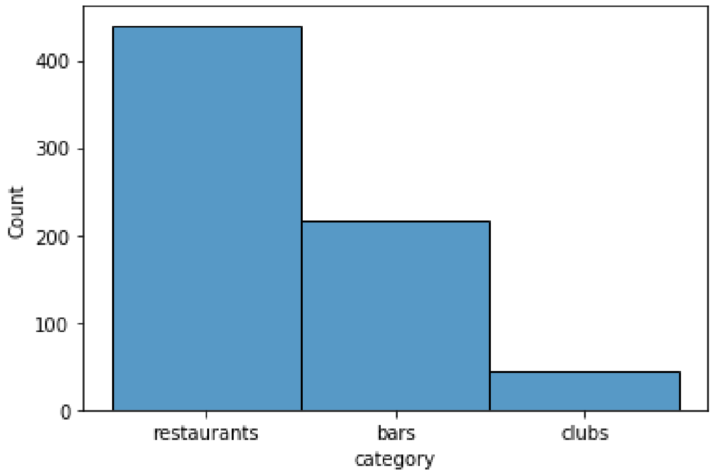
Target Group

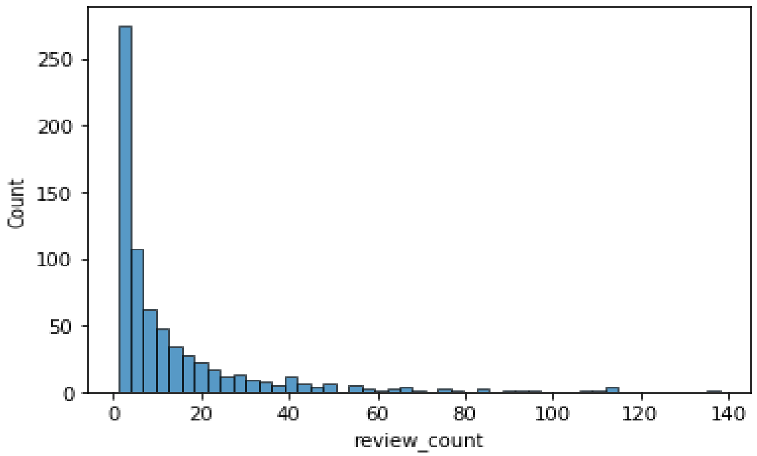
- Generation Y, with plenty of opportunities that however still need guidance

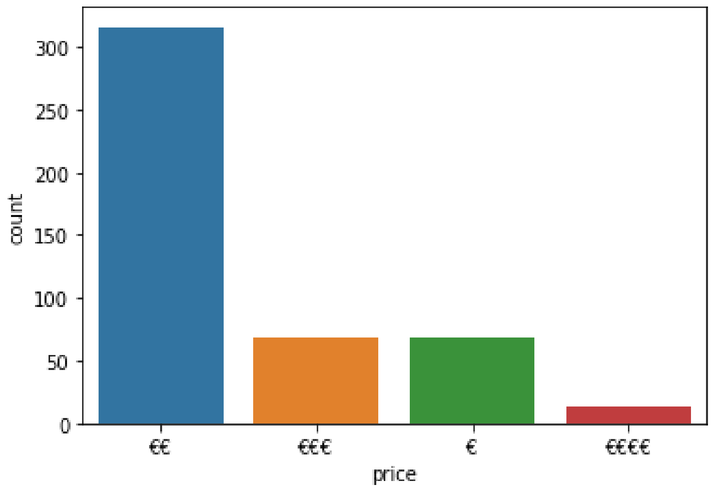
Data

- Yelp developer API: Fetching business data via REST call
- Data Format:
  - Tabular
  - Meta info on businesses (e.g. review count, rating, pricing, …)
- Yelp API Limitations: Max. 50 businesses per query
  - Workaround: Query each borough in a city individually and aggregate results
- Example: “Stuttgart (Germany)”

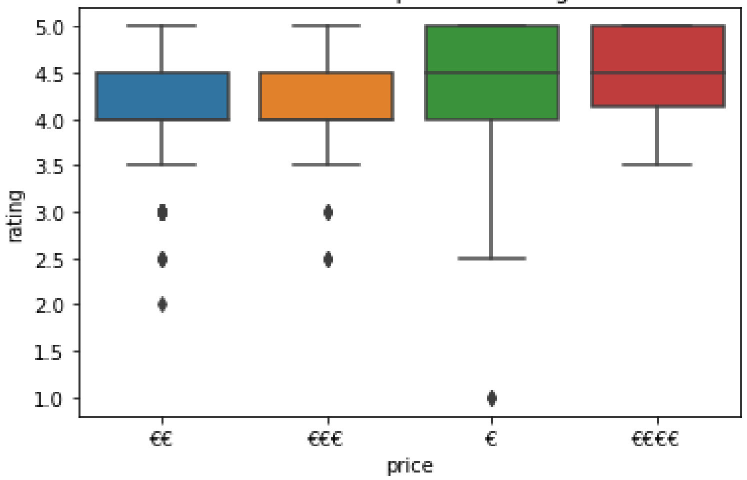
Exploratory Data Analysis (EDA)



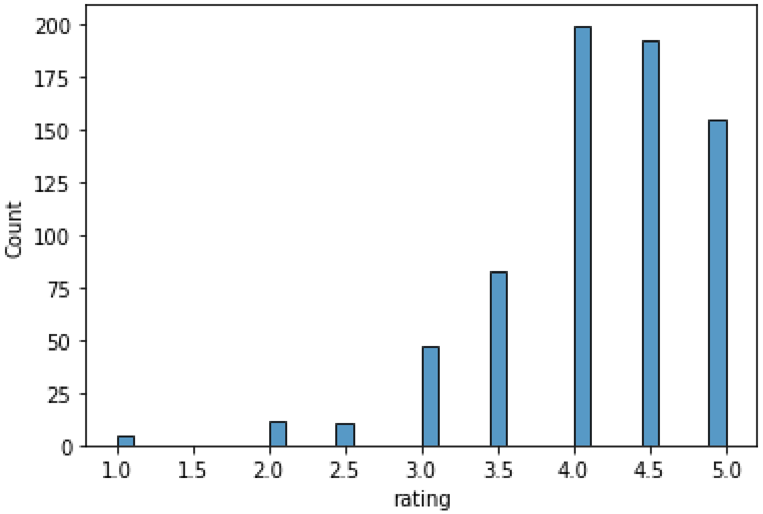


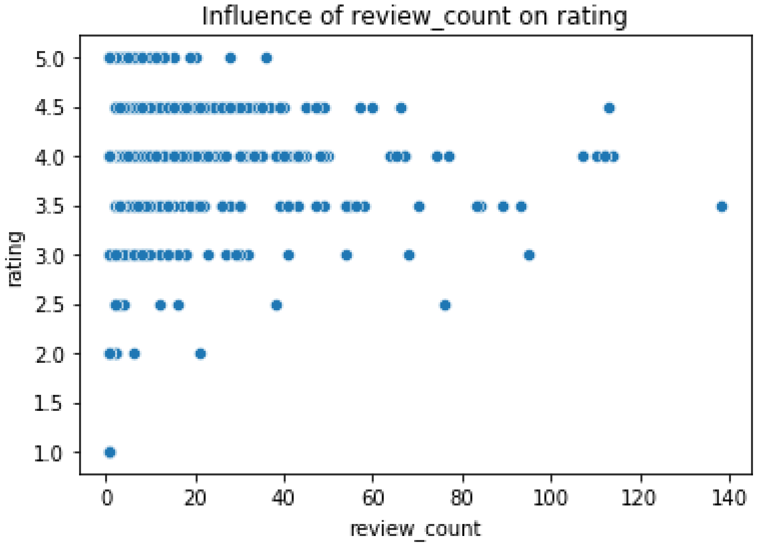


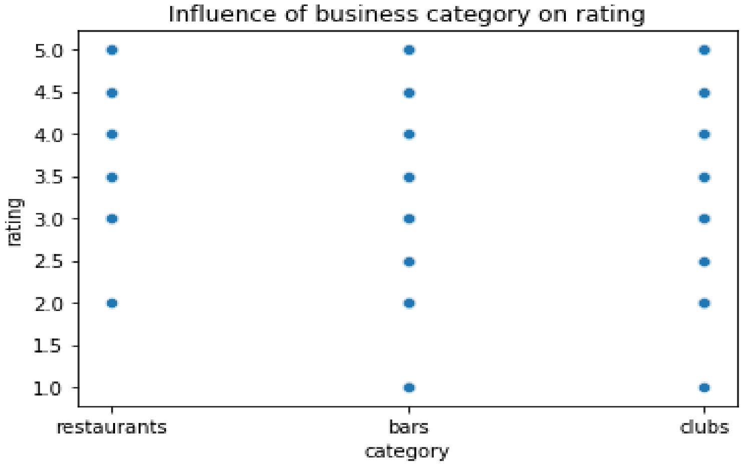
Influence of price on rating



Exploratory Data Analysis (EDA)



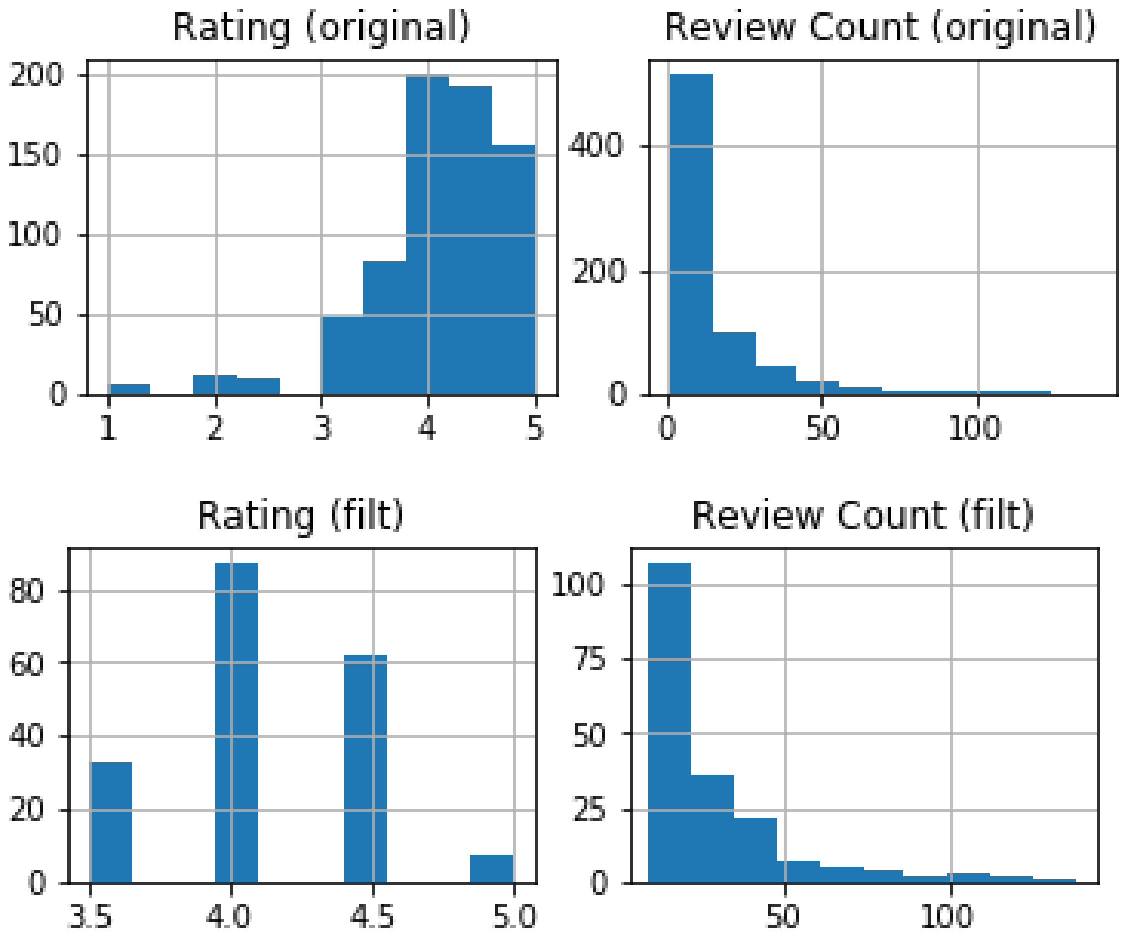




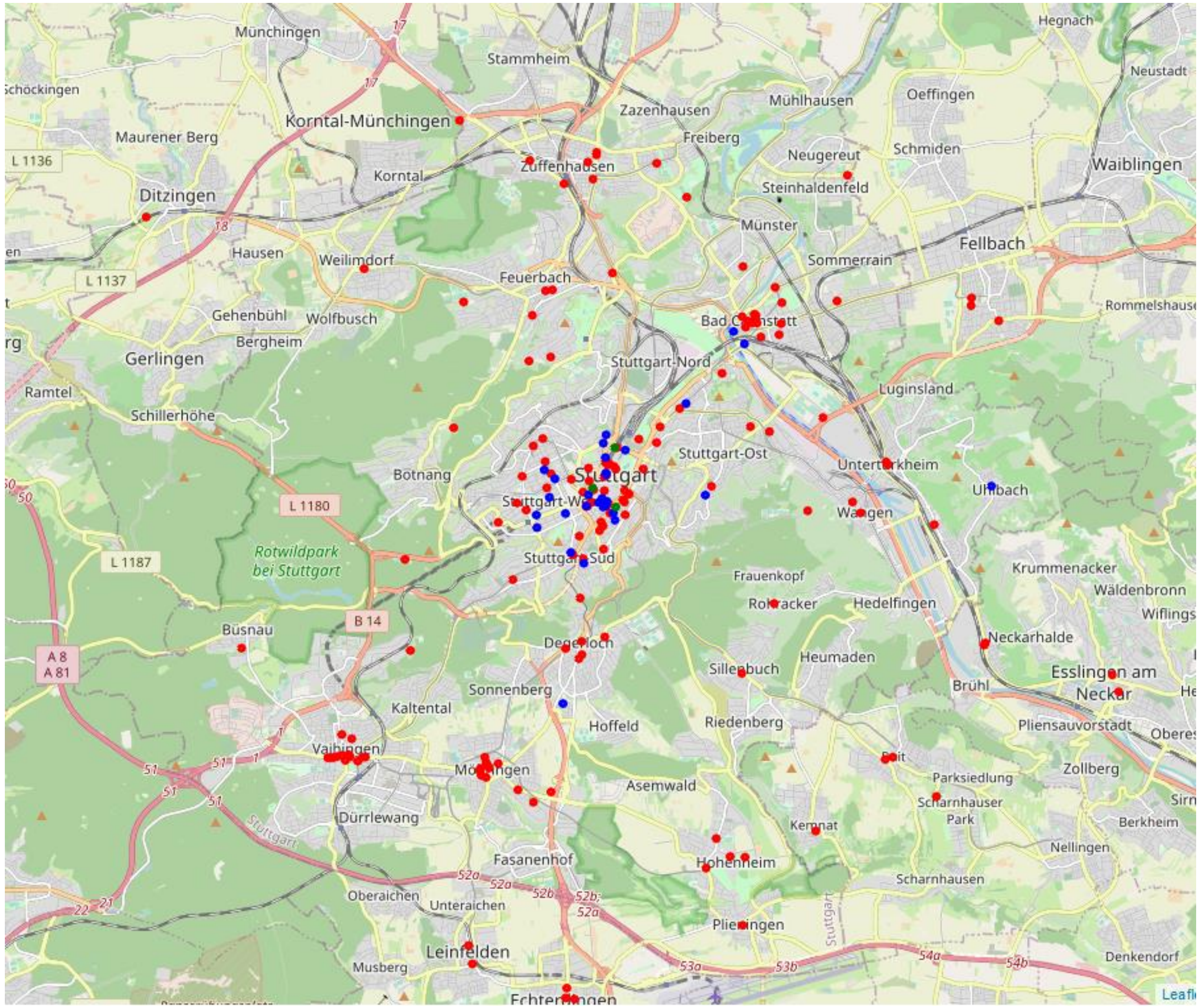
Methodology

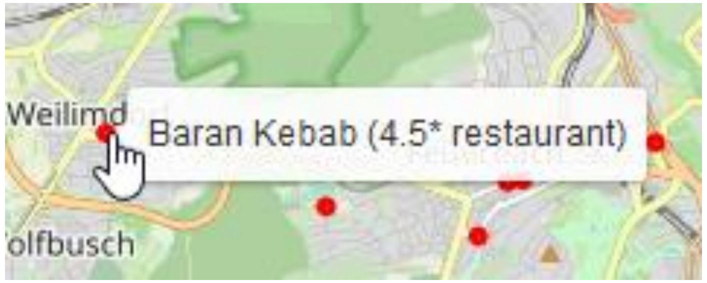
1) Filter businesses (with rating below 3.5 or review count below 10)
2) Visually inspect businesses on map
3) Aggregate review count and rating into business score
4) Use kMeans Clustering to identify geographical clusters of businesses
5) Combine business score into cluster score per cluster
6) Plot heatmap of clusters

1) Filter out uninteresting businesses



2) Visually inspect businesses

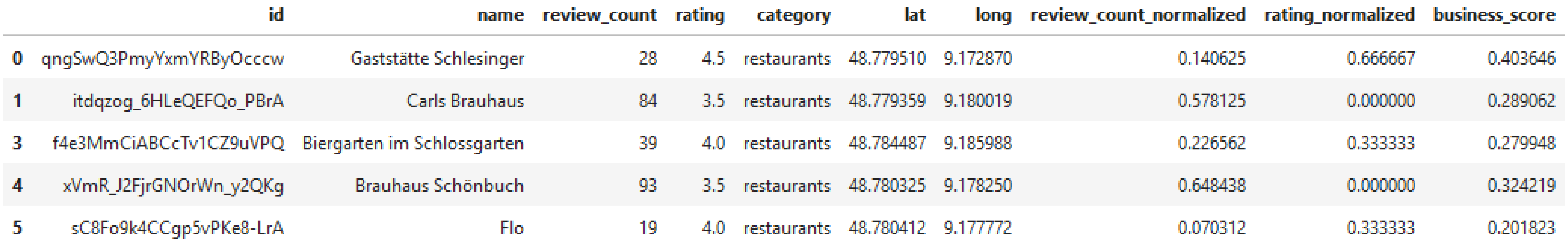




Baran Kebab (4.5* restaurant)

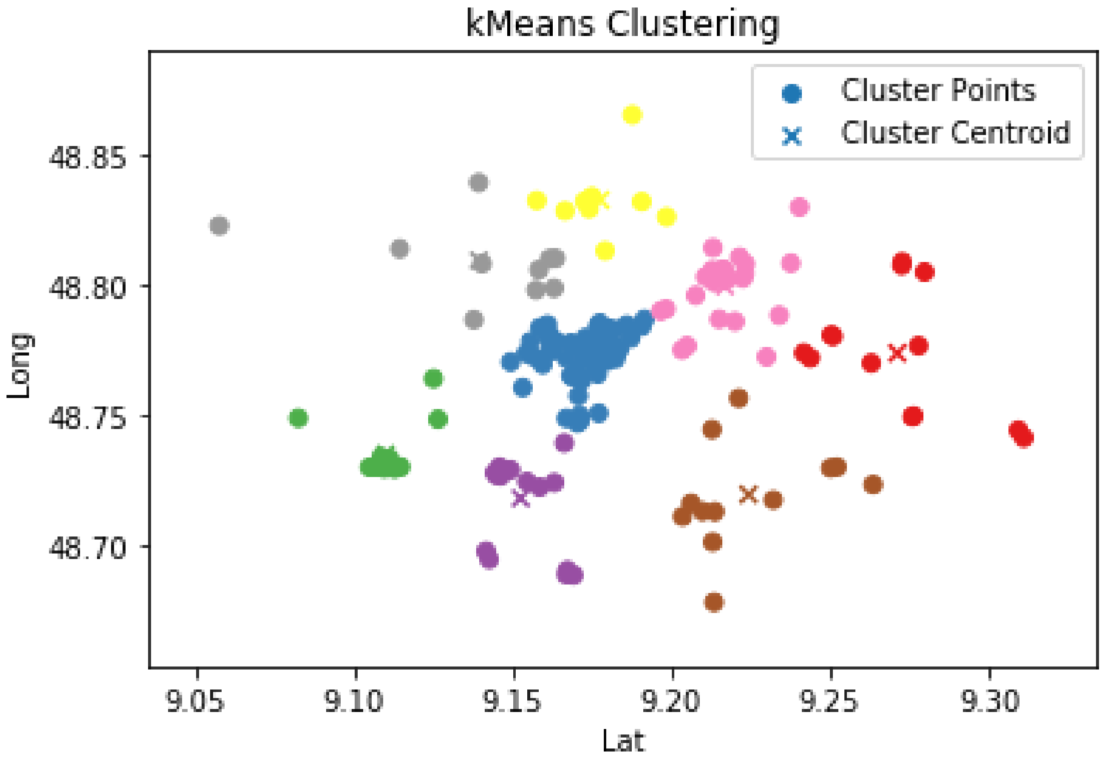
3) Calculate business score

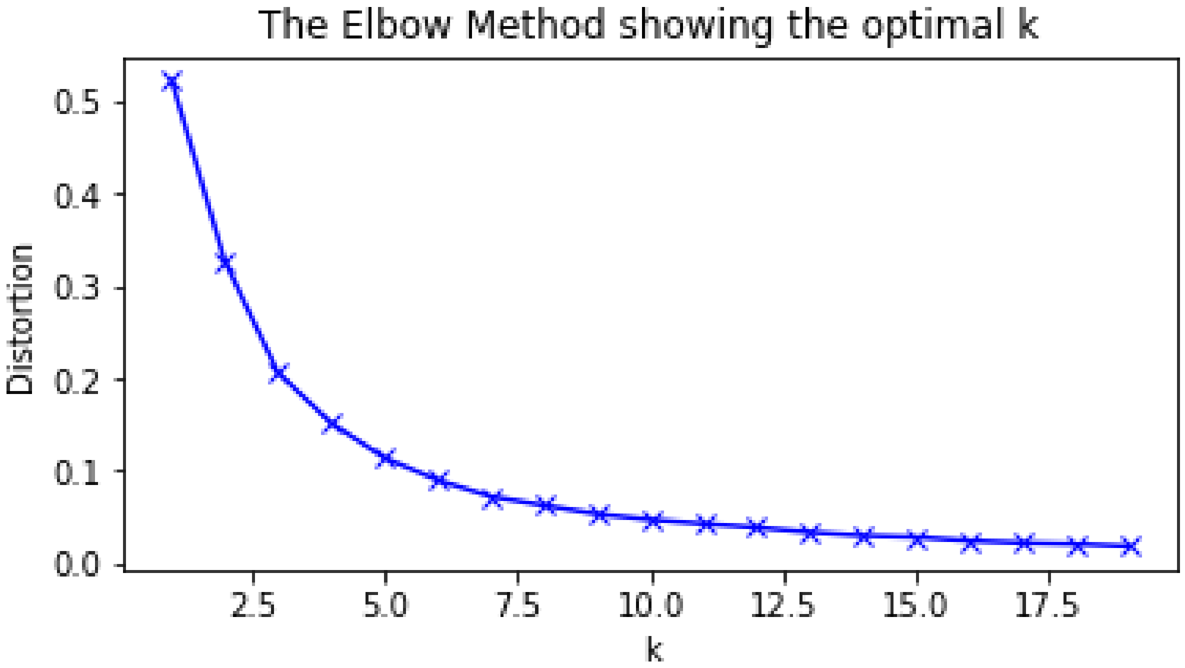
- Normalize Review Count & Rating per business into range [0,1]
- Average Review Count & Rating



	id	name	review_count	rating	category	lat	long	review_count_normalized	rating_normalized	business_score
0	qngSwQ3PmyYxmYRByOcccw	Gaststätte Schlesinger	28	4.5	restaurants	48.779510	9.172870	0.140625	0.666667	0.403646
1	itdqzog_6HLeQEFQo_PBrA	Carls Brauhaus	84	3.5	restaurants	48.779359	9.180019	0.578125	0.000000	0.289062
3	f4e3MmCiABCcTv1CZ9uVPQ	Biergarten im Schlossgarten	39	4.0	restaurants	48.784487	9.185988	0.226562	0.333333	0.279948
4	xVmR_J2FjrGNOrWn_y2QKg	Brauhaus Schönbuch	93	3.5	restaurants	48.780325	9.178250	0.648438	0.000000	0.324219
5	sC8Fo9k4CCgp5vPKe8-LrA	Flo	19	4.0	restaurants	48.780412	9.177772	0.070312	0.333333	0.201823

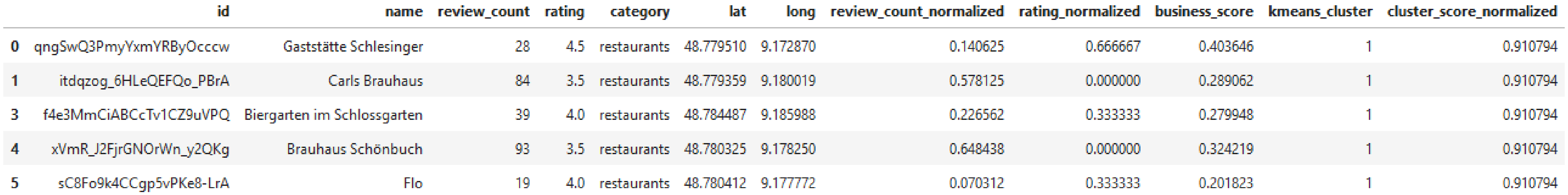
4) kMeans Clustering





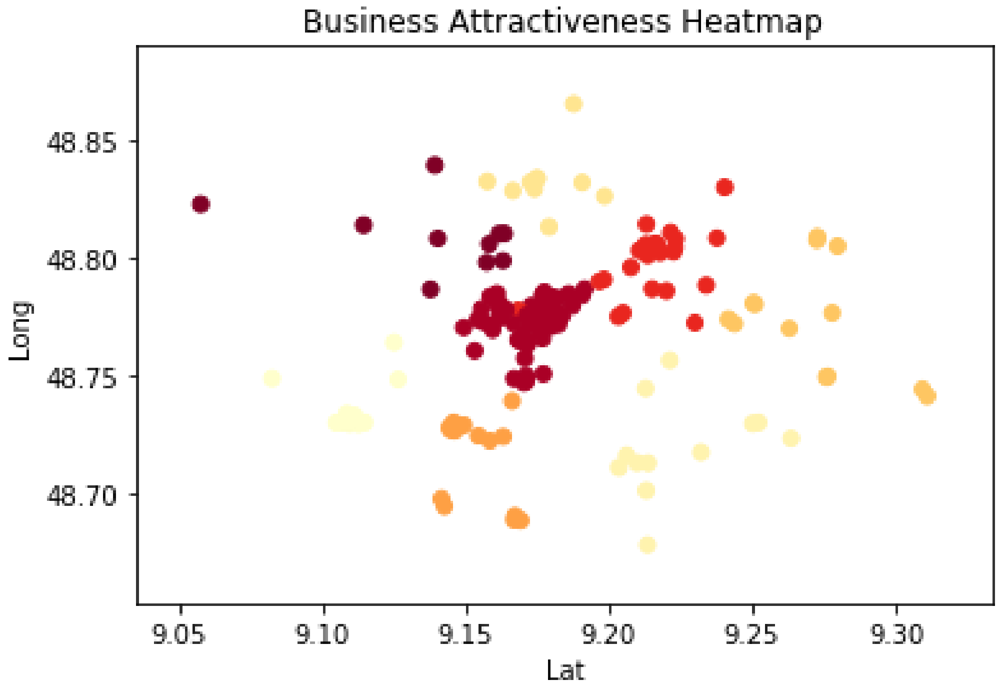
5) Calculate Cluster Score

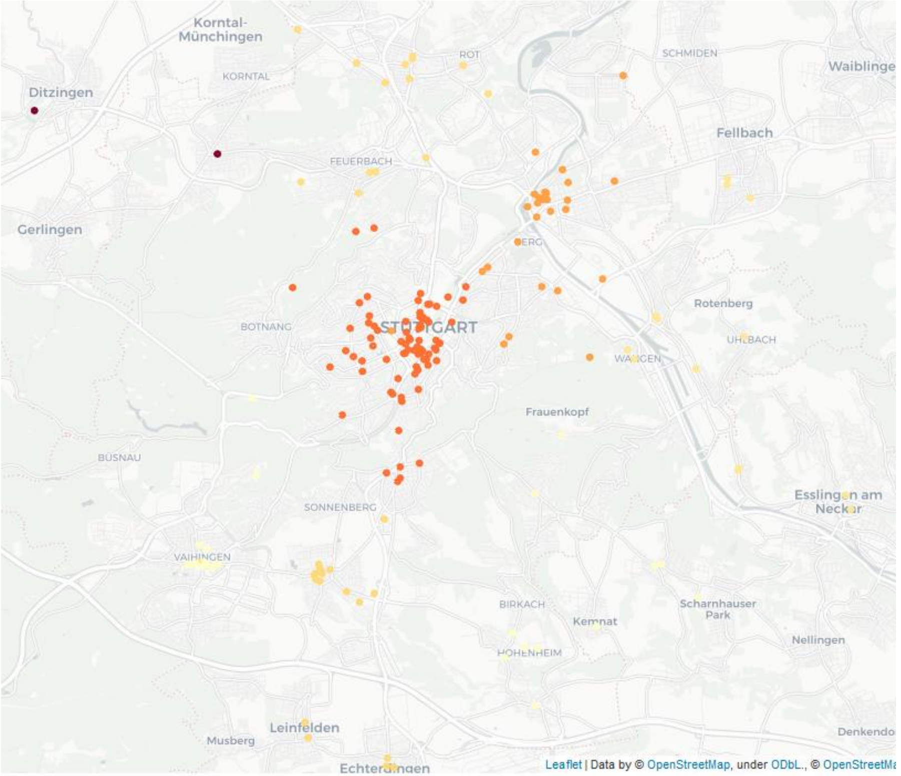
- Average business scores of all businesses within each cluster
- Normalize scores into range [0,1]



	id	name	review_count	rating	category	lat	long	review_count_normalized	rating_normalized	business_score	kmeans_cluster	cluster_score_normalized
0	qngSwQ3PmyYxmYRByOcccw	Gaststätte Schlesinger	28	4.5	restaurants	48.779510	9.172870	0.140625	0.666667	0.403646	1	0.910794
1	itdqzog_6HLeQEFQo_PBrA	Carls Brauhaus	84	3.5	restaurants	48.779359	9.180019	0.578125	0.000000	0.289062	1	0.910794
3	f4e3MmCiABCcTv1CZ9uVPQ	Biergarten im Schlossgarten	39	4.0	restaurants	48.784487	9.185988	0.226562	0.333333	0.279948	1	0.910794
4	xVmR_J2FjrGNOrWn_y2QKg	Brauhaus Schönbuch	93	3.5	restaurants	48.780325	9.178250	0.648438	0.000000	0.324219	1	0.910794
5	sC8Fo9k4CCgp5vPKe8-LrA	Flo	19	4.0	restaurants	48.780412	9.177772	0.070312	0.333333	0.201823	1	0.910794

6) Cluster Heatmap





Conclusion

Results

- Hotspot identified: Stuttgart downtown
- Hidden gems (restaurants) northwest of the city

Limitations

- Yelp is not very popular in Germany
- Query limitations (max. 50 entries); Workaround: Query boroughs individually → However, this is still lowering the coverage of the data