City Explorer - Munich

1. Introduction

When visiting a new city, many people wish to simply stroll around to explore the shops, cafes, bars, and parks in their surroundings. By design, common city guides - digital and analog – do not cater to people's wish to explore their environment. Instead, they mostly provide readers with a more or less extensive check list of the most prominent sights of a city. The web-based application „City Explorer" is the city guide for people who wish to explore their surroundings independently. Users can select from five distinct venue categories:

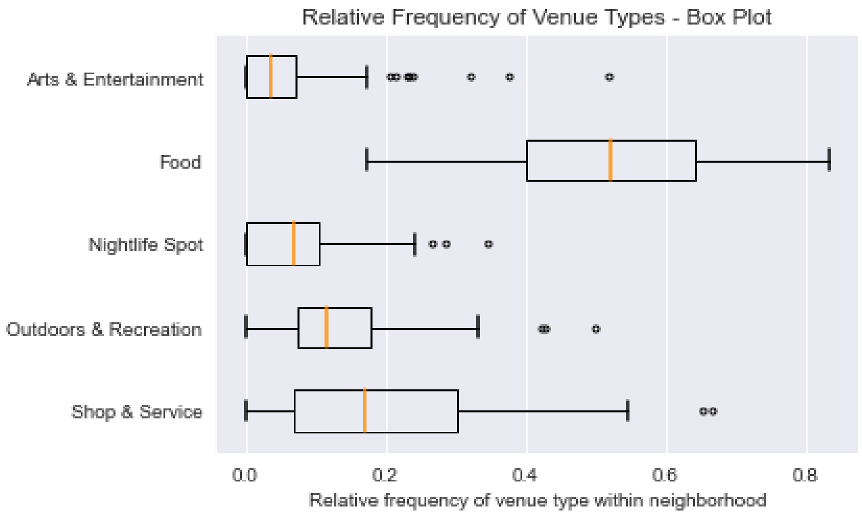
- Arts & Entertainment (e.g. art galleries, museums, theaters, cinemas, stadiums)
- Food (e.g. restaurants, pubs, cafés)
- Nightlife (e.g. bars, nightclubs)
- Outdoors & Recreation (e.g. parks, forests)
- Shops & Service (e.g. clothing stores, bookstores, flower shop)

Based on the type of venues that a user is most interested in, „City Explorer" recommends neighborhoods that the user will enjoy exploring. Moreover, users can indicate the neighborhoods that they enjoyed the most to get recommendations on further neighborhoods that they might want to explore.

2. Data

In the present case study we implement a first raw version of the „City Explorer" for the city of Munich. The analysis combines geodata from OpenStreetMap and extensive venue data provided by Foursquare. First, we access the OpenStreeMap database using the Overpass API and retrieve the geometric shapes of Munich's 108 neighborhoods. Subsequently, we use the Foursquare API to identify the type and number of venues located within the (simplified) boundaries of each neighborhood.

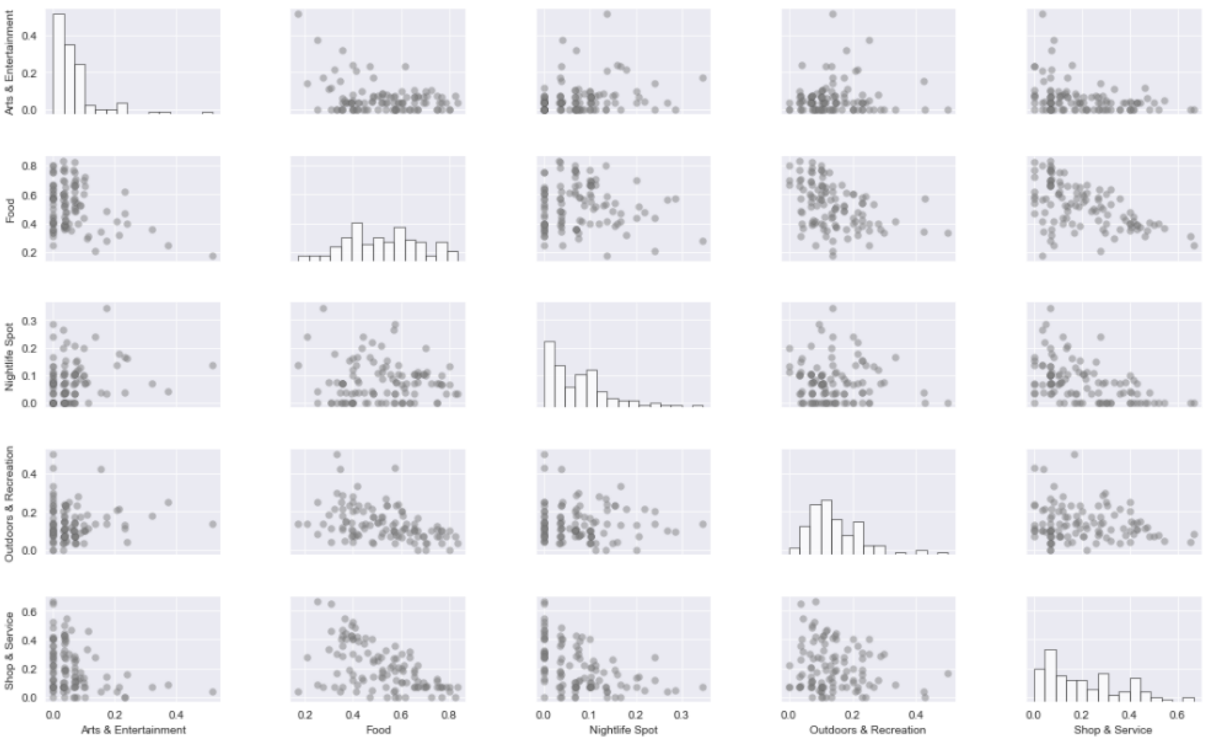
The resulting venue data set is summarized in the following two graphs. The boxplot summarizes the relative frequencies of the different venue types within neighborhoods.



For the absolute majority of neighborhoods, "Food" venues are the most common venue type.

Shop & Service" venues tend to be the second most common venue type, followed by venues of the category "Outdoors & Recreation". "Arts & Entertainment" and "Nightlife Spot" tend to be the least frequent venue types within neighborhoods

Relative Frequency of Venue Types - Pairwise Plots



The pairplots show the pairwise distributions of the relative frequencies of the different venue types. Upon visual inspection the data does not show any clearly separated clusters. The majority of graphs show at most one area of particularly high density and observations outside high density areas do not show any grouping patterns. For the vast majority of neighborhoods, the categories "Arts & Entertainment" and "Nightlife Spots" account for less than 10% of all venues. In these dimensions we see a high observation density below 10% and only a few losely scattered observations above 10%. The highest degree of dispersion can be observed in the venue categories "Food" and "Shop & Service". As a result, observation clouds tend to be stretched along these dimensions.

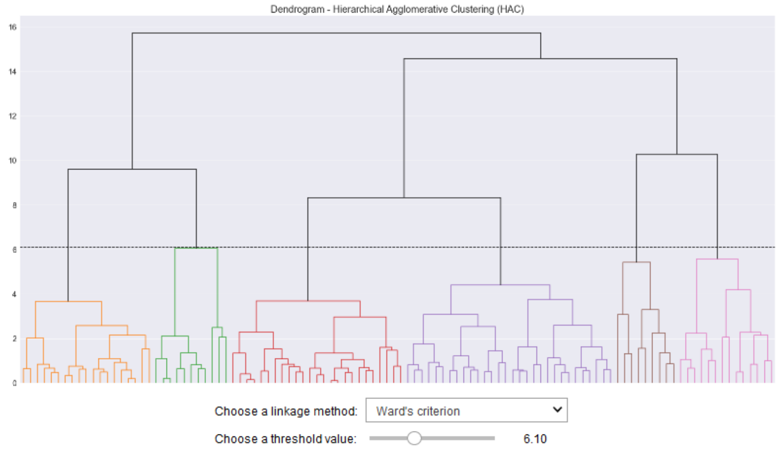
In summary, the data does not show clearly separated clusters. Thus, k-means might not be the most suitable clustering algorithm, as it tends to show poorer performance when clusters are non-circular/non-convex. In our analysis we will therefore also consider hierarchical agglomerative clustering.

3. Method

In our cluster analysis we consider K-means clustering as well as hierarchical agglomerative clustering (HAC). To narrow down the set of reasonable specifications (algorithm, number of clusters) we first collect a series of assessment metrics. Subsequently, we summarize the characteristics of the specification that appears most suitable for our purpose.

For HAC we consider three potential linkage methods: 'Ward's criterion', 'complete linkage', and 'average linkage'. The following figures present the corresponding dendograms for the considered linkage methods.

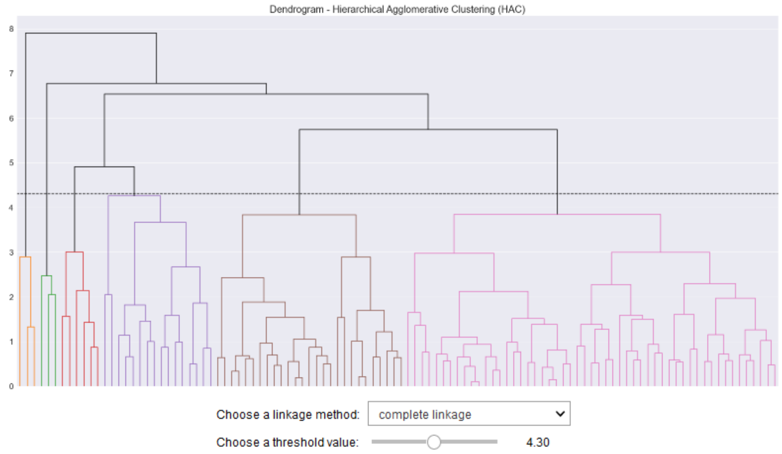
Ward's linkage criterion:



Choose a linkage method: Ward's criterion

Choose a threshold value: 6.10

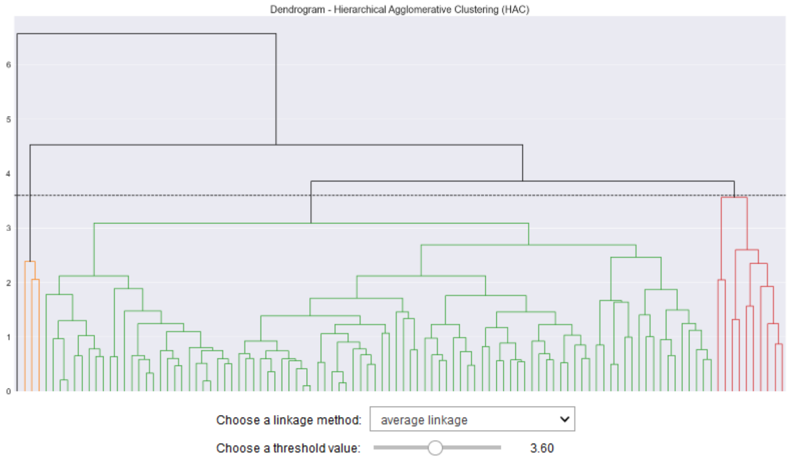
Complete linkage:



Choose a linkage method: complete linkage

Choose a threshold value: 4.30

Average linkage:

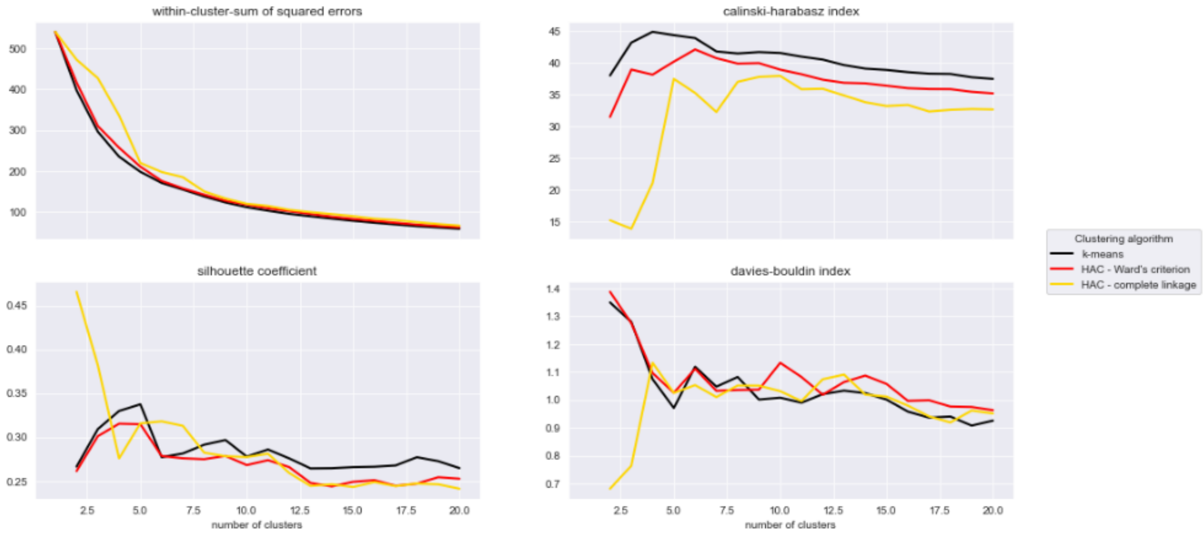


Choose a linkage method: average linkage

Choose a threshold value: 3.60

Of the three considered linkage methods, linkage according to Ward's criterion appears to be the most suitable method for our purposes. Starting from a relatively low threshold value (3.10 or higher, i.e. 15 or less clusters) it yields comparatively evenly sized clusters. 'Average linkage' does not appear suitable for our purposes. For a wide range of threshold values, the average linkage method produces singular clusters, while the majority of neighborhoods are grouped in the same cluster. As a result, the clusters generated by this method frequently are too narrow or too wide for the recommendation of similar neighborhoods. Therefore, we exclude 'average linkage' in the collection of assessment metrics.

To identify the natural number of clusters in our venue data, we make use of multiple assessment metrics depicted in the following figure:



The top left plot shows the **within-cluster-sums of squared error (WCSSE)** for a broad range of cluster numbers. Not surprisingly, the k-means algorithm tends to yield the lowest WCSSE. Yet, it is closely tracked by the clusters generated using HAC with Ward's linkage criterion. For both clustering approaches, the plotted line shows a smooth decline in WCSSE without any prominent elbow-point. For HAC with complete linkage, the plot shows multiple potential elbow points.

The top right figure plots the **Calinski-Harabasz Index (CHI)**. Higher CHI values indicate a clearer cluster segmentation. According to the CHI, therefore, the optimal numbers of clusters appear to be 4 clusters when using k-means, 6 clusters for HAC with Ward's linkage criterion, and 10 clusters for HAC with complete linkage.

The bottom left graph shows the average **silhouette scores** depending on the number of clusters. Again, a higher average silhouette score indicates a clearer distinction between clusters. For k-means and HAC with Ward's linkage criterion this indicator is maximized at 5 clusters. For HAC with complete linkage the silhouette score is maximized at 2 clusters.

Finally, the bottom right figure plots the **Davies-Bouldin Index (DBI)**. Contrary to the previous indicators, cluster segmentation is optimized when the DBI is at its minimum. For k-means and HAC with Ward's linkage criterion the first local minimum is located at k=5. Yet, within the considered range the DBI for both clustering approaches is minimized at excessively large cluster numbers. For HAC with complete linkage the DBI is minimized at 2 clusters.

Overall, the k-means algorithm and HAC with Ward's linkage criterion appear to yield the more well-behaved cluster segmentations. For both approaches, all metrics yield similar implications on the relationship between number of clusters and clustering quality. For HAC with complete linkage the relationship between number of clusters and clustering quality appears less clear, especially as CHI and silhouette scores/DBI yield completely contradicting implications.

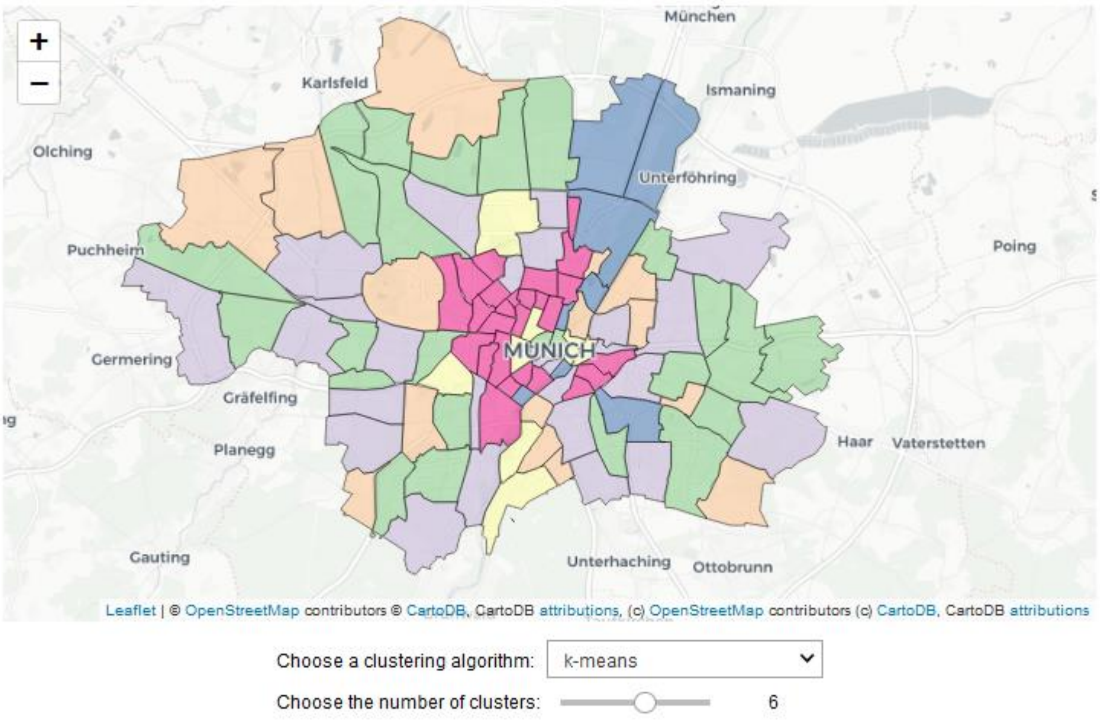
Therefore, will exclude HAC with complete linkage from our further analysis. Instead, we focus on k-means and HAC with Ward's linkage criterion as potentially fruitful clustering approaches. Given the findings of the considered assessment metrics, our analysis will moreover place particular focus on cluster numbers around 5.

4. Results

The most suitable neighborhood segmentation is obtained with the following specification:

- Clustering algorithm: k-means
- Number of clusters: 6

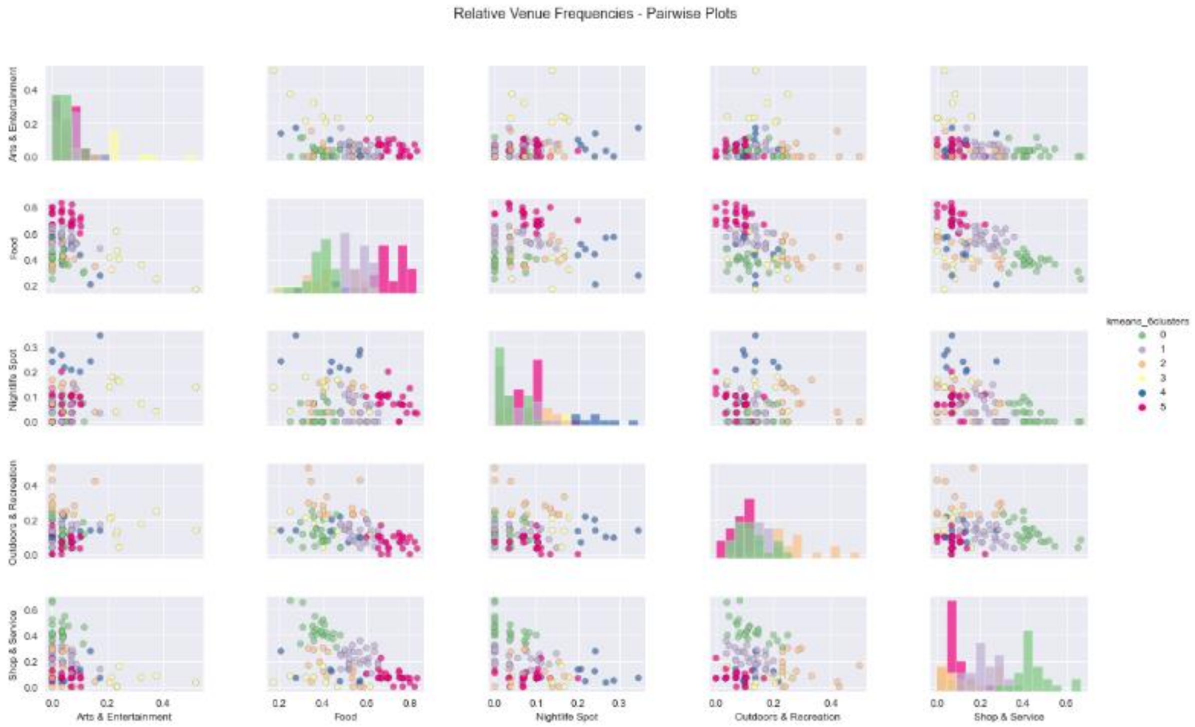
The assessment metrics presented in section 3 implied that the venue data has five natural clusters. Yet, for our purpose we identify six as our optimal number of clusters as it yields a clearer distinction between inner-city neighborhoods (pink) and similar outer-city neighborhoods (purple).


+
−
München
Karlsfeld
Ismaning
Olching
Unterföhring
Puchheim
Poing
MUNICH
Germering
Gräfelfing
Planegg
Haar
Vaterstetten
Gauting
Unterhaching
Ottobrunn
Leaflet | © OpenStreetMap contributors © CartoDB, CartoDB attributions, (c) OpenStreetMap contributors (c) CartoDB, CartoDB attributions

Choose a clustering algorithm: k-means

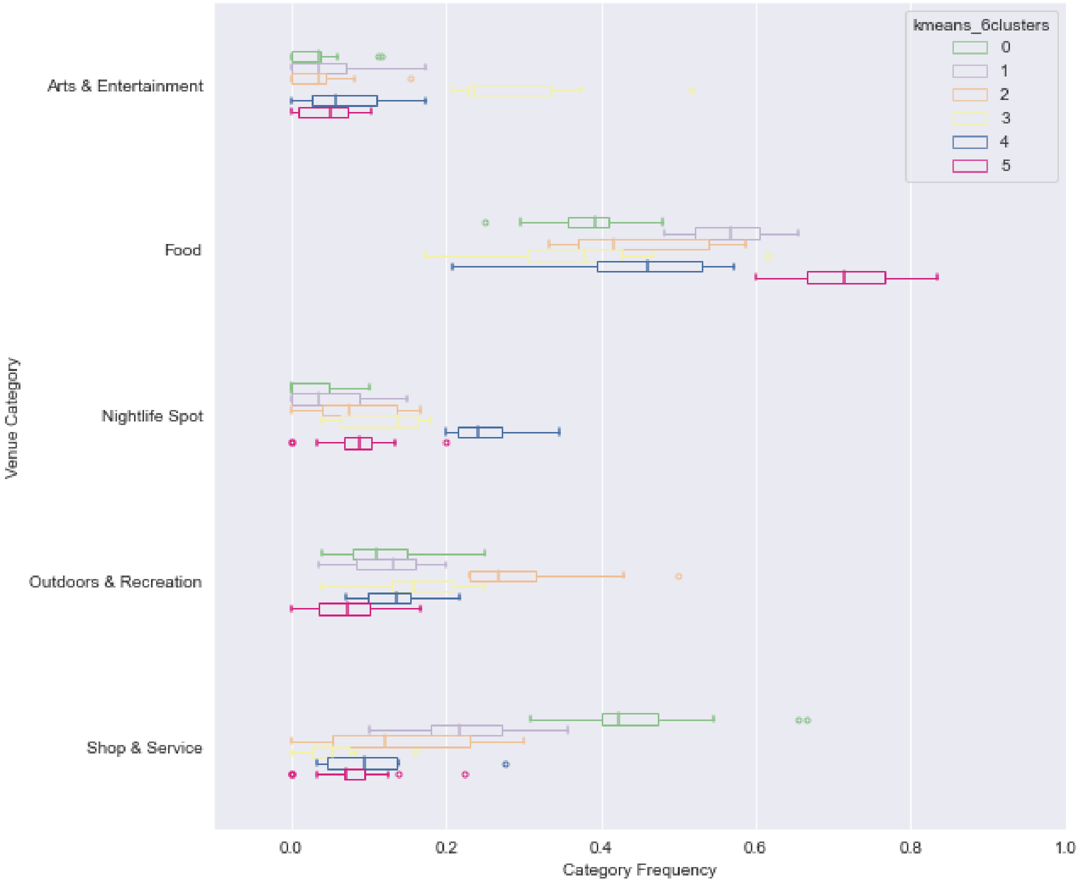
Choose the number of clusters: 6

The following pairplots show the pairwise distributions of the relative frequencies of venue types within the respective clusters. As can be seen, our specification results in six clearly distinguishable clusters. With the exception of the purple cluster, each cluster clearly stands out in at least on of the five venue categories. Neighborhoods in the purple cluster tend to take a central position across all five distributions.



Choose a clustering algorithm: k-means

Choose the number of clusters: 6



This impression is furthermore supported by the box plot presented above. The green cluster clearly stands out due to a particularly high frequency of venues from the category „Shop & Service“. The orange cluster shows a particularly high share of „Outdoors & Recreation“ venues. The yellow cluster stands out in the dimension „Arts & Entertainment“. The blue cluster is especially interesting for tourists looking for „Nightlife Spots“. Finally, the red cluster has an exceptionally high frequency of „Food“ venues.

5. Discussion

As with any cluster analysis, we had to select from an almost infinite number of potential specifications (clustering algorithm, format of venue data, number of clusters). In the context of our study, the format of venue data appears to be an especially interesting dimension that might be worth exploring in fruther studies. We built clusters based on (standardized) relative frequencies of venue types. Alternatively one could also build clusters based on absolute frequencies or based on venue density (venues per sqkm).

6. Conclusion

In this study we implement a first raw version of the „City Explorer“ for the city of Munich. In our analysis we combine geodata from OpenStreetMap and extensive venue data provided by Foursquare. We make use of a k-means clustering algorithm to divide Munich neighborhoods into clusters of similar neighborhoods. The presented study yields a first starting point for the further development of the „City Explorer“ application. Potential for next steps is outlined in the discussion section.