Optimal locations to open a Chinese restaurant in Stuttgart, Germany
- *Strategic decision making using machine learning*

Author: Cao
July 01, 2020

Table of contents

1. Introduction/Business Problem

The success of opening a new restaurant depends on several factors: location, menu, customers, competition and so on. A restaurant's location is as crucial to its success as great food and service. Therefore, it is of utmost importance to determine the location in a strategic way in order to maximize business profits.

To the audiences who are interested in opening a Chinese restaurant in Stuttgart ([1]) Germany, the data analysis in this project will tell them in which regions of Stuttgart to look for the optimal location. The restaurant concept and cuisine type define the demographics and psychographics of the ideal customers. A Chinese restaurant in western cities provides exotic tastes for the local people without traveling to China. The age and density of the population in the area determine the attraction of the restaurant and therefore would be considered with the competition by leveraging the Foursquare location data to solve the business problem.

Fellow entrepreneurs who plan to open their first Chinese restaurant or to expand their franchised restaurants in Stuttgart would be very interested in the insights of this report for competitive advantages and business values. Those who seeking to establish an asian restaurant may also be interested.

2. Data

2.1. Data sources

As described in the business problem, the following aspects will be combined and considered for the decision making. Namely,

- Stuttgart regions and their relevant population information including density, density of age group between 20 and 65, average age;
- geographical coordinate of each region;
- number of existing restaurants in each region which are grouped into three categories: Chinese restaurant, asian restaurant and other restaurant for the competition study.

The population information in Stuttgart area from year 2001 to 2019 can be found on the Stuttgart website. The data in Excel format are provided by Statistikatlas Stuttgart and are downloaded as 'Statistikatlas_Stuttgart_Datentabelle_Stadtbezirke.xlsx' for the further use. Geographical coordinate of each region will be obtained using the Geocoder package for the address near the region center. The Foursquare location API ([2]) is used to search the existing restaurants in the vicinity of each region.

2.2. Data cleaning

Stuttgart has altogether 23 regions. The names of the regions are imported from the Excel table. One of the key points to locate a restaurant is a good reachability of the place either by cars or by public transportations. Specially for a Chinese restaurant, a certain percentage of the customers will be Chinese overseas students who may not have cars. Therefore, the town hall of each region which usually locates in the center and is easily reachable will be defined as the address for Geocoder to get the latitude and longitude coordinates for the Foursquare searching later on.

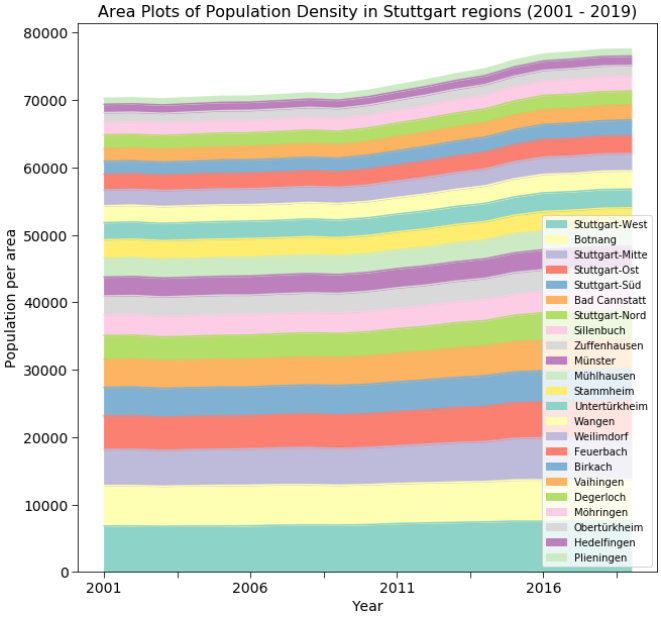
Remark: For some regions, Geocoder fails to get the correct coordinates by only giving 'town hall' in the address. Hence, a specific address near the region center is explicitly given.

Population and its density of each region from year 2001 to 2019 are imported and the region area is calculated. The region area is further used to obtain the population density of the age group between 20 and 65 which can better describe the potentiality of customers in each region. The average age is also imported from the data source as one of the features.

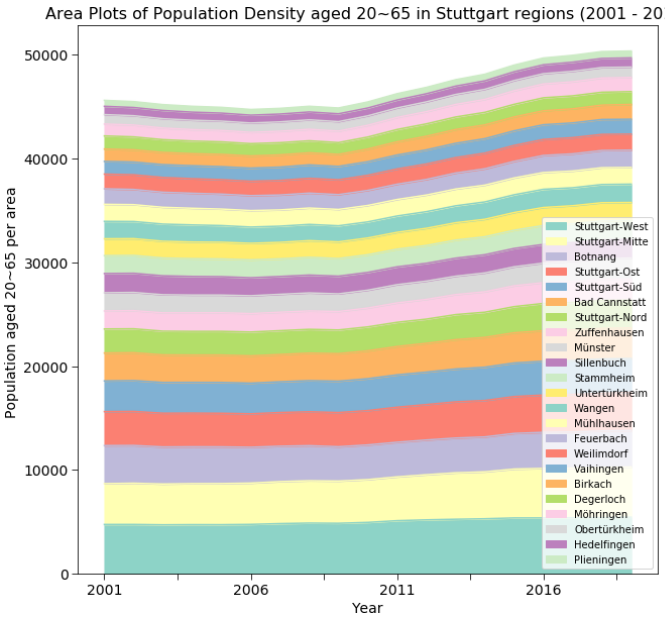
3. Methodology

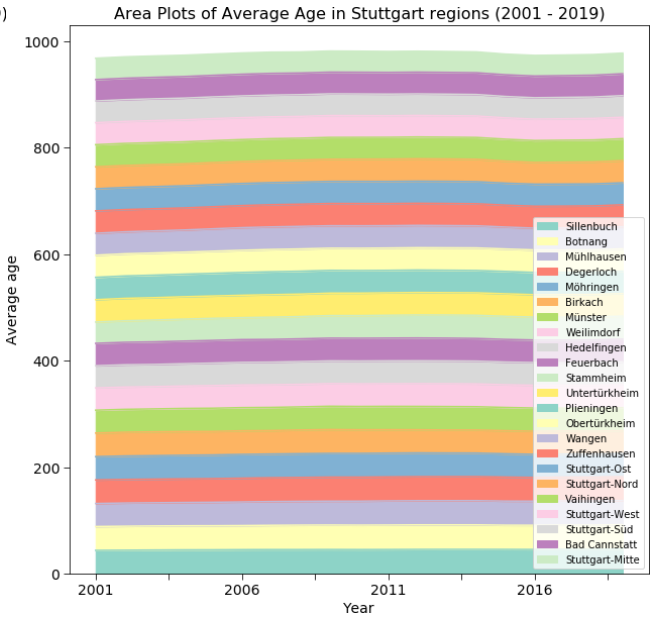
3.1. Exploratory data analysis

3.1.1. Feature changes over time



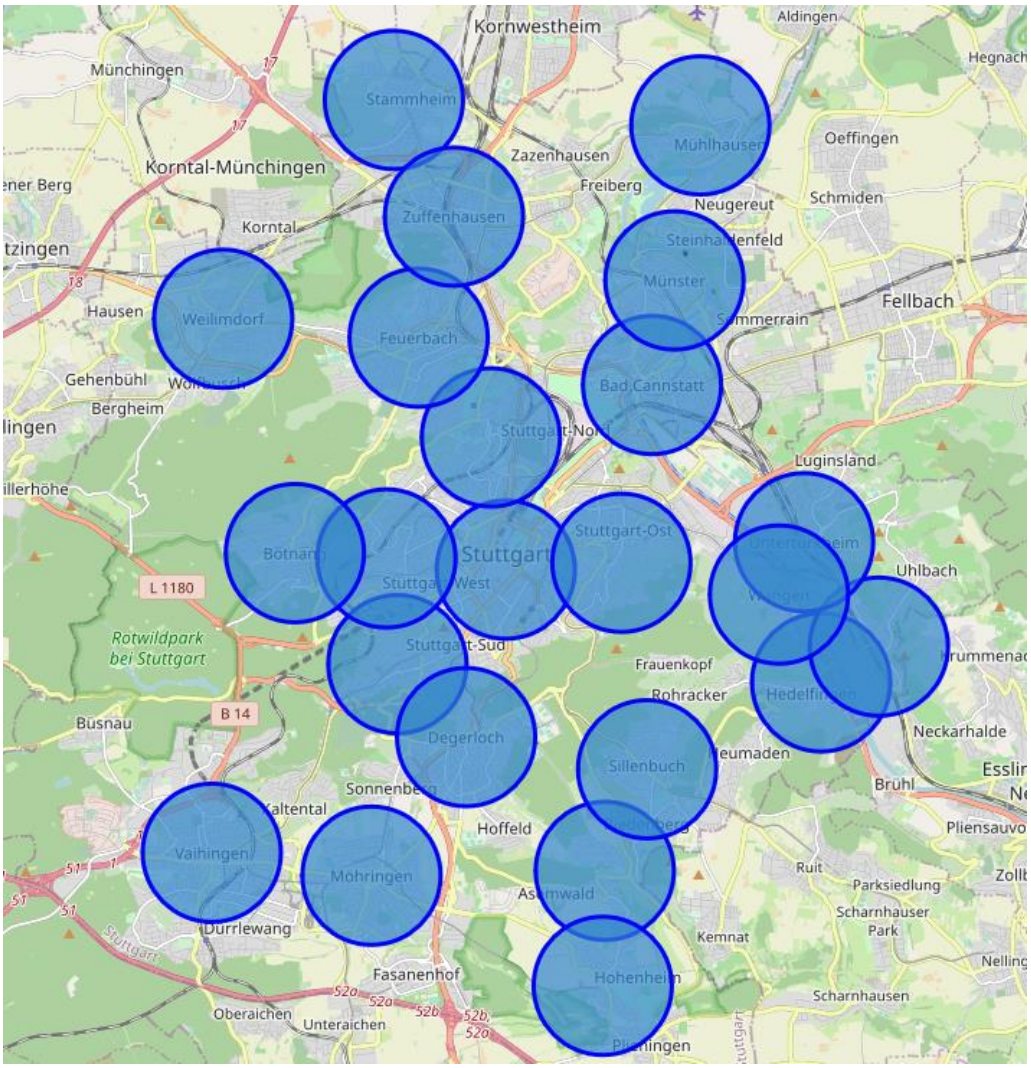
A quick visualization for population density, population density aged 20~65 and average age from year 2001 to 2019 is shown using stacked area plots. As we can see from the three area plots, the feature changes of each region are slow over time, especially the change of the average age. Hence, only the data of the latest year 2019 for each region will be merged into the feature dataframe for the machine learning modeling.





3.1.2. Folium map

The Geocoder package and Folium library are used to get the latitude and longitude values of each region center and to geographically visualize the distribution of each region in Stuttgart with markers (see map below). The radius of markers is set to 1 km which will be later on also applied to the Foursquare API as the searching radius for existing restaurants. The markers with the actual searching radius can give an intuitional illustration for the coverage of the searching. The reason for choosing 1 km radius is that this distance is reachable on foot around 10~12 minutes and there are no big overlapping between every two neighbor regions.



3.1.3. Existing restaurants using Foursquare API

As mentioned in section 3.1.2., the searching radius using Foursquare API is set to 1 km. The searching results are grouped into three categories in terms of the category names in Foursquare for the analysis afterwards: Chinese restaurant, asian restaurant and other restaurant. The idea behind the grouping is the similarities, dissimilarities and the competition levels to the Chinese restaurants. The cuisine and style of the asian restaurants are quite similar among each other in comparison to the western, South American or African restaurants. The existing Chinese restaurants are of course the direct competitors and the other asian restaurants are also quite competitive due to the exotic similarities. The grouping gives a quick and more clear exploration of the number distribution.

3.2. Unsupervised machine learning model with clustering

Since we know little to the outcomes that are to be expected for the optimal location of a new opening Chinese restaurant, **clustering** would be a proper machine learning technique to discover demographic structure and restaurant distribution in Stuttgart area. There are many models for clustering out there. I

select the **K-Means** which is vastly used for clustering in many data science applications, especially useful to quickly discover insights from unlabeled data as for our case. By segmenting and clustering the regions, a recommendation for optimal location candidates will be given.

Essentially, determining the number of clusters in a data set, or K as in the K-Means algorithm, is a frequent problem in data clustering. The correct choice of K is often ambiguous because it is very dependent on the shape and scale of the distribution of points in a dataset. Here, I use the elbow method with the help of **KElbowVisualizer** to select the optimal number of clusters by fitting the model with a range of values from 1 to 10 for K. The scoring parameter **metric** is by default set to **distortion**, also as attribute **inertia**, which computes the sum of squared distances of samples to their closest cluster center.

3.2.1. Pre-processing

The existing restaurants of each region searched from Foursquare are collected in the dataframe column 'Restaurant Category' which is a categorical variable. The K-Means algorithm cannot operate on such variables directly because Euclidean distance function is not really meaningful for discrete variables. It requires all input variables and output variables to be numeric. Here, I use **One-hot encoding** to convert the categorical data to numerical data. After encoding, the number of existing restaurants for the three cuisines (Chinese, asian and other) is obtained for each region. These features combined with population density and average age (see table below) are prepared for the machine learning model.

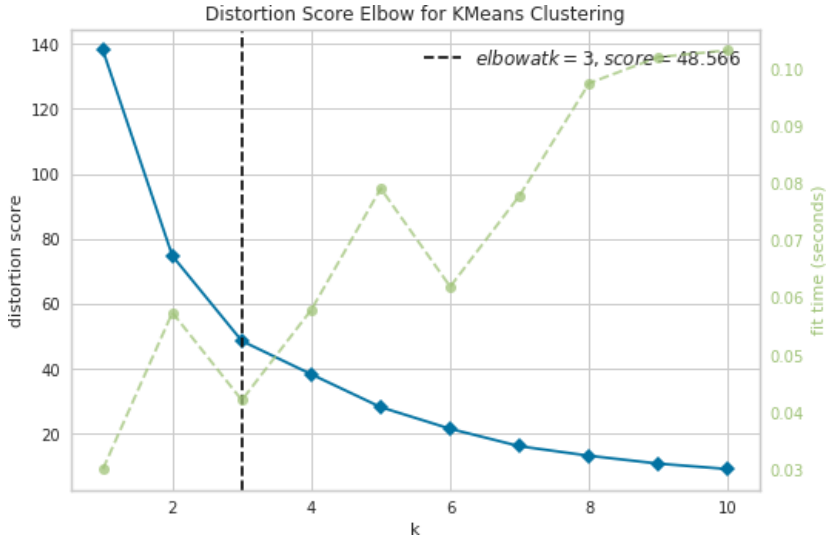
	Population Density	Population Density Age 20~65	Average Age	Asian Restaurant	Chinese Restaurant	Other Restaurant
0	6340	4832.840980	38.7	5	3	34
1	4068	2657.910629	41.6	0	0	5
2	5416	3675.605387	41.3	1	1	9
3	4613	3257.480792	40.6	2	0	11
4	7640	5462.230138	40.3	2	1	9

3.2.2. Normalization

Altogether six features for the clustering have three different magnitudes and distributions for density, age and number. In order to interpret theses features equally, we use **StandardScalar()** to normalize the dataset over the standard deviation using a statistical method.

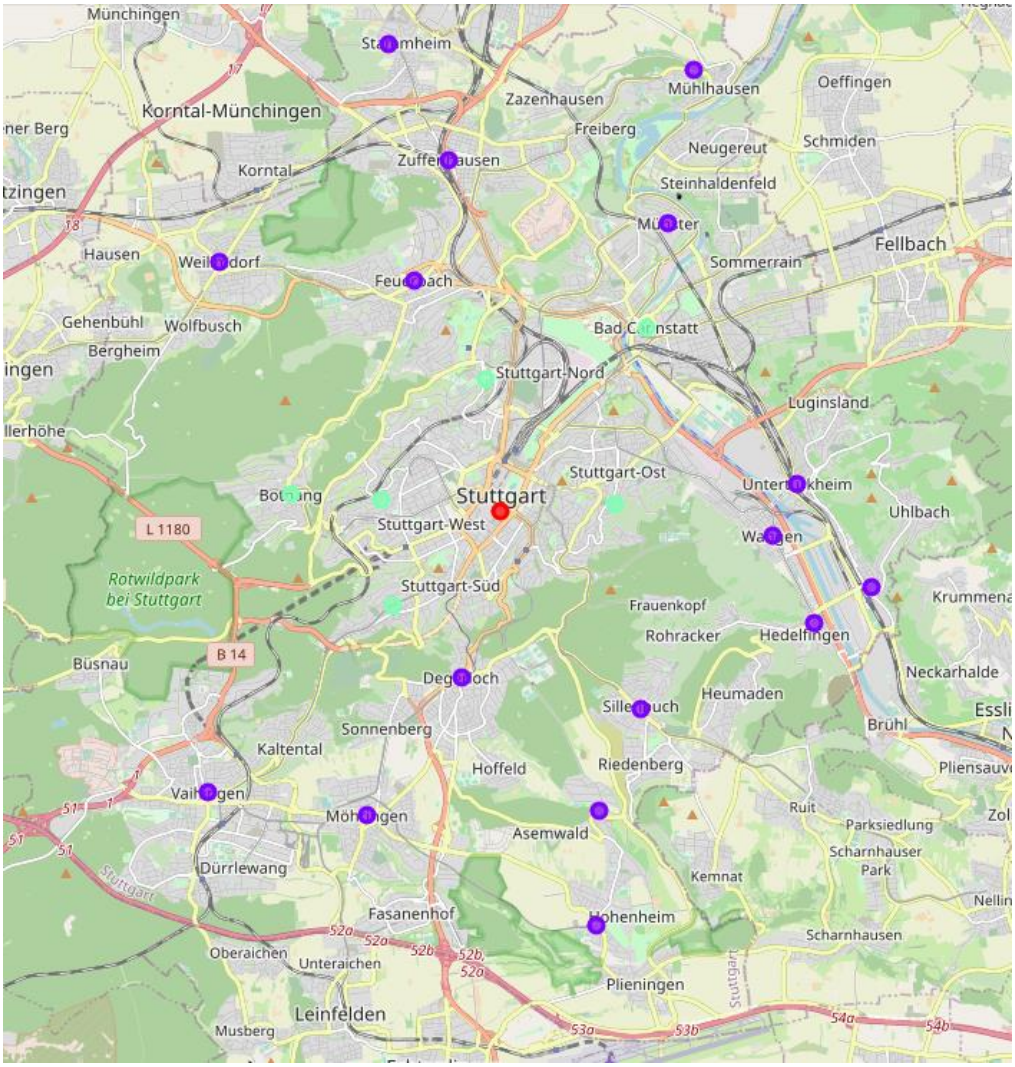
3.2.3. Modeling

As mentioned before, I first use **KElbowVisualizer** to determine the optimal number of clusters. The scoring parameter **distortion** as a blue solid line and the amount of time to train the clustering model per **K** as a dashed green line are displayed in the figure 'Distortion Score Elbow for KMeans Clustering' (see below). The optimal cluster number is marked as a dashed black line, in our case, 3 which is used to fit the model and get cluster labels.



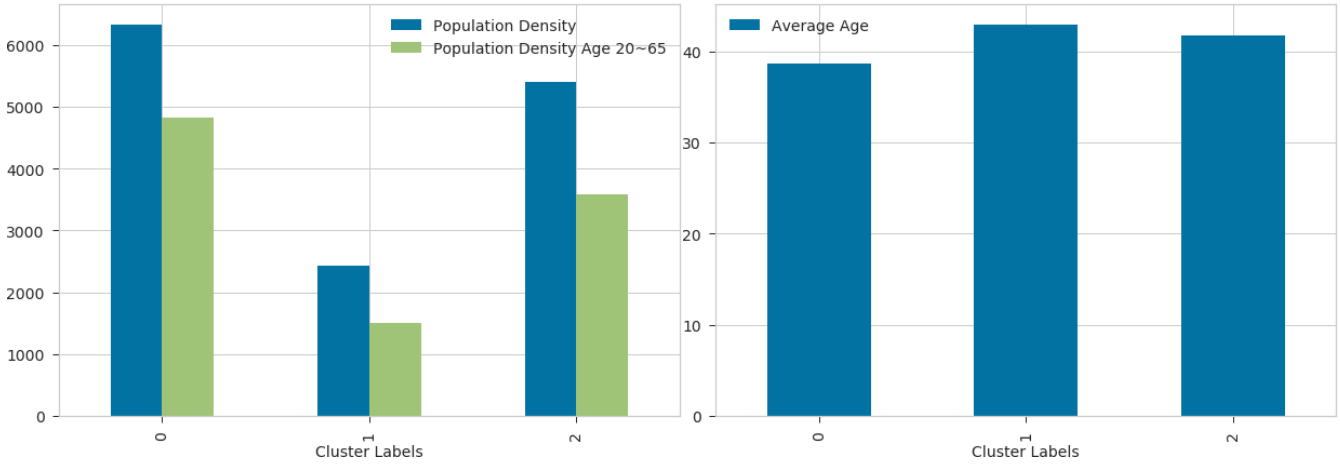
4. Results

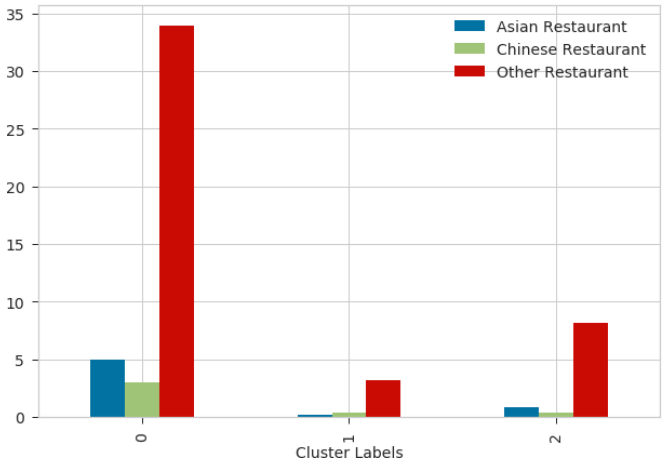
After the optimal clustering, let's merge the cluster labels into the feature dataframe, sort the values by column 'Cluster Labels' and visualize the resulting clusters in the map (see right). The colorful markers show a radial distribution of the clusters. The region **Stuttgart-Mitte** alone constitutes **Cluster 0** (red) which is the center of the cluster distribution. The other four downtown regions surrounding the center, namely, **Stuttgart-Ost, Stuttgart-Süd, Stuttgart-West** and **Stuttgart Nord**, plus another two nearby regions **Botnang** and **Bad Cannstatt** form **Cluster 2** (green). The rest further outside the circle belongs to **Cluster 1** (purple).



Now I group the features in each cluster and calculate the mean values to get insights into the cluster properties. The mean values of six features are compared in the bar charts below for three clusters and are described as follows:

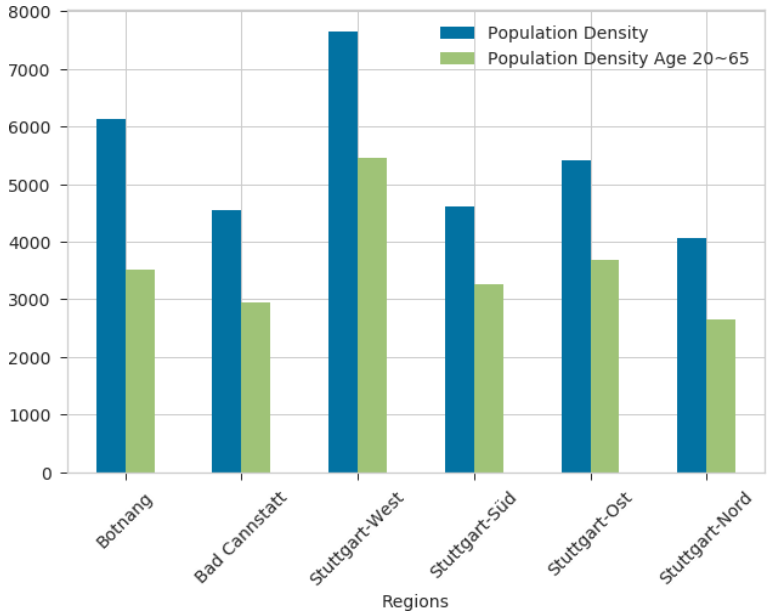
- **Cluster 0:**
  - high population density (also age 20~65)
  - average age: 38.7
  - lower middle number of Chinese & asian restaurants
  - very high number of other restaurants
- **Cluster 1:**
  - low population density (also age 20~65)
  - average age: 43
  - very low number of Chinese & asian restaurants
  - low number of other restaurants
- **Cluster 2:**
  - upper middle population density (also age 20~65)
  - average age: 41.8
  - low number of Chinese & asian restaurants
  - lower middle number of other restaurants

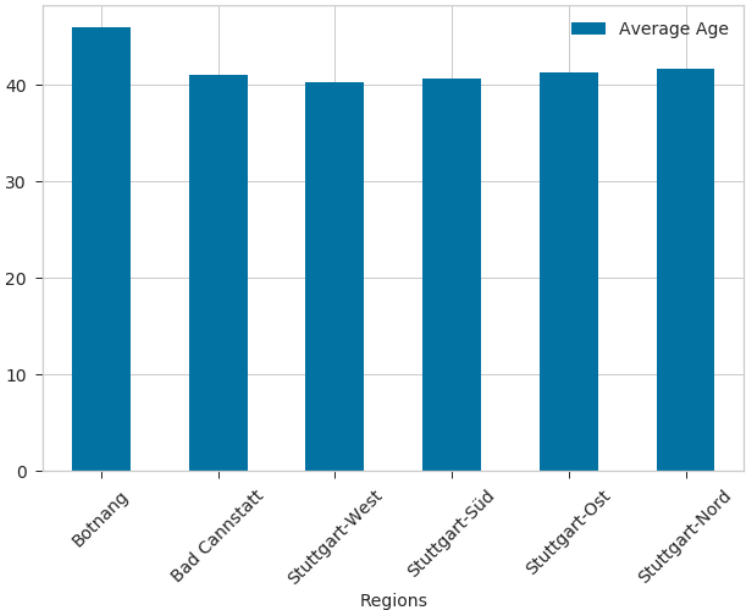




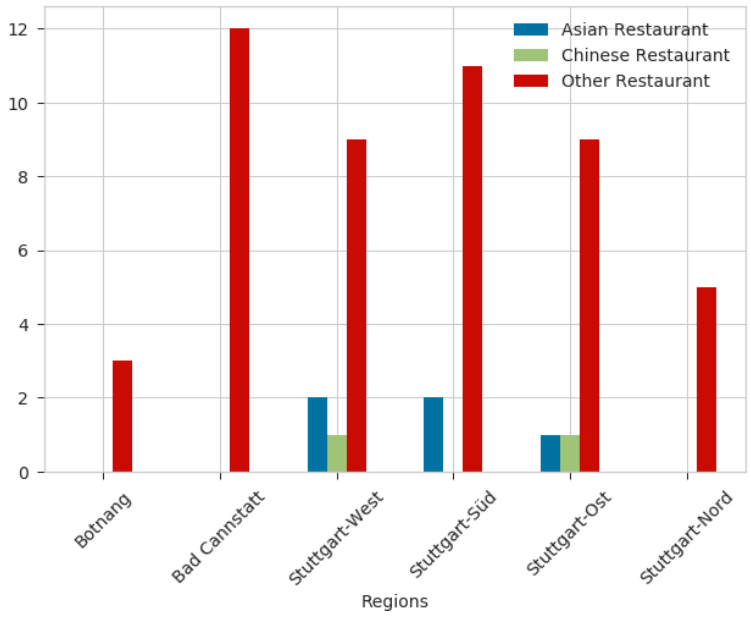
Let's look more deeply into Cluster 2 with bar charts. The six features for each region in Cluster 2 are illustrated and compared. As before, the potential aspects of selecting the optimal locations are listed for each region:

- **Botnang:**
  - upper middle population density (age 20~65)
  - average age: 45.9
  - no Chinese restaurants
  - no asian restaurants
  - low number of other restaurants
- **Bad Cannstatt:**
  - middle population density (age 20~65)
  - average age: 41
  - no Chinese restaurants
  - no asian restaurants
  - high number of other restaurants
- **Stuttgart-West:**
  - high population density (age 20~65)
  - average age: 40.3
  - low number of Chinese restaurants
  - lower middle number of asian restaurants
  - upper middle number of other restaurants
- **Stuttgart-Süd:**
  - middle population density (age 20~65)
  - average age: 40.6
  - no Chinese restaurants
  - lower middle number of asian restaurants
  - high number of other restaurants
- **Stuttgart-Ost:**
  - upper middle population density (age 20~65)
  - average age: 41.3
  - low number of Chinese restaurants
  - low number of asian restaurants
  - upper middle number of other restaurants
- **Stuttgart-Nord:**
  - lower middle population density (age 20~65)
  - average age: 41.6
  - no Chinese restaurants
  - no asian restaurants





- lower middle number of other restaurants



5. Discussion

Remember that the size of the searching circle for restaurants is the same for each region. This implies that the existing number of restaurants also represents the restaurant's concentration. The clustering analysis shows that highest concentration of restaurants is detected in Cluster 0, the very center area of Stuttgart. It is crowded with competitors of various cuisines including many Chinese/asian restaurants. In contrast, Cluster 2 has low ~ lower middle concentration of restaurants which is comparable with Cluster 1 from the competitor point of view but fairly closer to the center area.

If we only focus on the customers from the region itself, the population density especially the age group 20~65 plays an important role. The reason to target the people between 20 and 65 years old is that they are more likely to accept or like the asian cuisine than the other age group. Children until teenage may still be influenced by their parents and the very old generation may on one hand seldom go to restaurants and on the other hand be less international for other kitchens. The results show that the average population density (also age 20~65) of Cluster 2 has a similar level as Cluster 0 but considerably higher than Cluster 1. Of course, customers are not limited to the region where a restaurant locates. The neighborhoods near the very center are more popular for shopping, relaxing, friends' gatherings, tourists and so on. Therefore, whether the customers from the region itself or from all surroundings, Cluster 2 is surely a potential candidate.

At last, the average age of Cluster 0 is 38.7 which is 4.3 years younger than Cluster 1 and 3.1 years younger than Cluster 2. As mentioned before, a region with more younger adults may better activate the restaurant business. For this reason, Cluster 2 also stands on a positive position.

Overall, **Cluster 2** appears to be the most promising group of regions to the audiences for opening a new Chinese or an asian restaurant. More specifically and similar to the analysis for clusters, the list of the

analysis results for the six regions in Cluster 2 would recommend **Stuttgart-Nord** as the optimal location.

The recommended cluster or concrete region is based on the available data we could find on the Internet. Some other relevant features like income, rental, amount of Chinese/asian migrants and even the business profits of existing restaurants will definitely contribute to a more detailed and accurate clustering model and guidance. In the future, the data set can be further expanded with these features whenever it is reachable for the model improvement and a second level of clustering for the optimal 1st-level cluster candidate can also be conducted when necessary.

6. Conclusion

The aim of this study is to identify an optimal location for opening a new Chinese restaurant in the city Stuttgart, Germany. Based on the available data, the population density including for the age group 20~65, average age and the calculated restaurant density from Foursquare data among the most important features are determined for the strategic analysis. One prerequisite of the recommendation is defining a circle around each region center as the search target.

The unsupervised machine learning model 'K-Means' is implemented to cluster the regions for recognizing the major zone of interest which contains potential regions. I use the elbow method to identify the optimum K value as 3. The three resulted clusters can be well distinguished among each other and a clear choice is made for Cluster 2. Folium map, stacked area plots and bar charts are combined to illustrate the analysis results.

Final decision-making of a specific optimal location will be done by the audiences based on the recommendation of this study and their own liking of the regions. This study can also be applied to other interchangeable scenarios such as opening a restaurant of different cuisines.

7. References

[1]. (2020). Retrieved from Stuttgart: https://en.wikipedia.org/wiki/Stuttgart

[2]. (2020). Retrieved from Foursquare API: https://developer.foursquare.com/