Where to go out on vacation?

Alexander Hirsch

December 23, 2020

Introduction

Business Problem

Assume you are on vacation in a foreign city, where you have never been before. You are lost by all possibilities – all the different restaurants, bars and clubs and cannot decide where to go. Further you actually like strolling through the streets and go into a restaurant, bar or club that looks most appealing to you in that very moment. However, in order to do that, you need some kind of hint which region in the city offers the best restaurants, bars and clubs. These businesses are often spread over the whole city, however there are some hotspots of the best businesses. We want to identify these hotspots!

Target Audience

The target audience for this problem is the Generation Y. These young people have plenty of opportunities. In fact, there are so many, they are overwhelmed by the abundance and cannot decide where to go. Contrarily these people often act in the "heat of the moment", when they find some visually appealing places while strolling by. These people seek for "guidance" in this strolling experience.

Data

Source of the data

The data will be fetched from the yelp developer API. This is due to the fact, that I was experiencing some verification problems with my foursquare account. Therefore, I was looking into alternatives that offer similar features. The data is being fetched via REST calls in python. By specifying a region (e.g. "Stuttgart Downtown") and a type of business (e.g. "restaurant") you then get a list of businesses in that region with lots of meta information. In this analysis I focus on the area of "Stuttgart (Germany)" and these types of businesses:

- Restaurants
- Bars
- Clubs

Data Format

The data is being stored in a pandas dataframe, which is essentially a pythonic type of table format. Each row is an individual business. Each column represents meta information (e.g. rating, review count, pricing etc.) for this business.

Data Cleaning

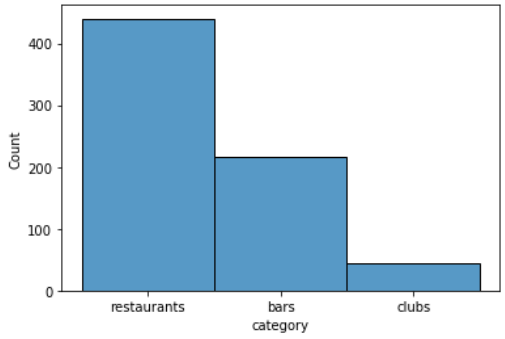
The yelp developer API is only retrieving a max. number of 50 businesses at a time. However, this is a very small data basis for further analysis. To circumvent this problem, I was retrieving those 50 businesses for each borough of Stuttgart individually. Afterwards I was merging those individual datasets into one single dataset. However, this is introducing duplicate businesses in the dataframe. Therefore, I was filtering out these duplicates by the unique "business ID".

Stuttgart has 23 boroughs. For these boroughs I was retrieving 1150 restaurants, 1065 bars and 781 clubs. In total these are 2996 potential businesses. After removing the duplicates, I was left with 2295 actual businesses.

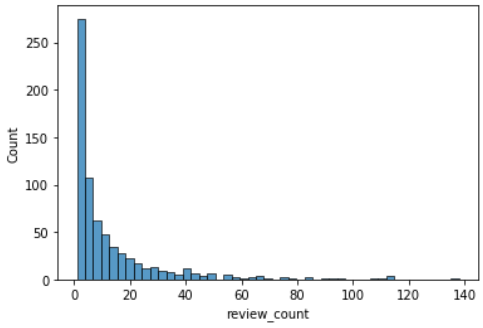
Exploratory Data Analysis (EDA)

To explore the data and find hidden correlations I was looking at many different histogram and scatter plots of the meta information in the data. These are the key findings of this analysis:

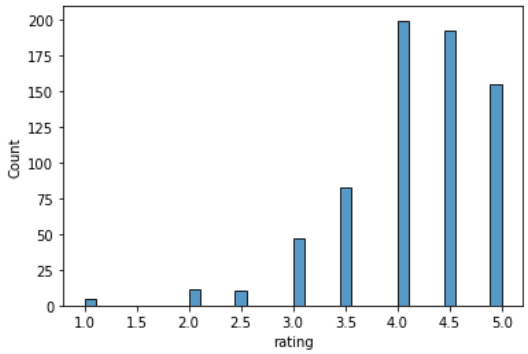
1. There are plenty of restaurants, less bars and comparatively few clubs:



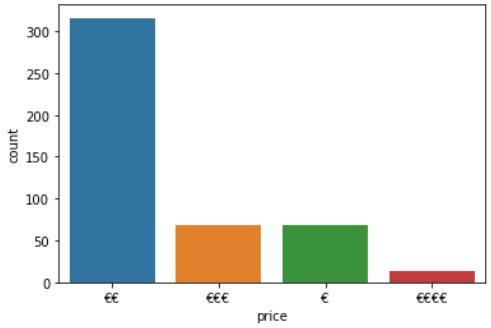
2. Most of the businesses have actually few reviews (10 or less):



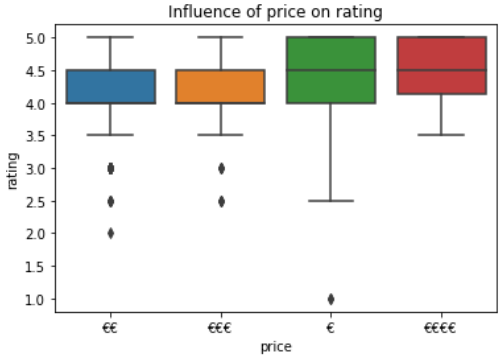
3. Most of the restaurants have high ratings (3.5 or above):



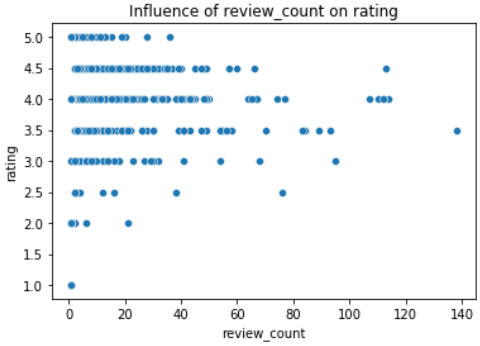
4. Most of the restaurants are in the lower middle price range (i.e. “€€”):



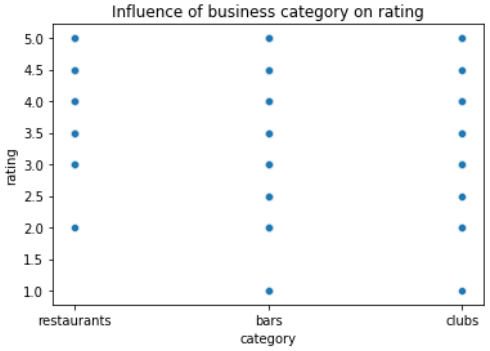
5. There is only a small influence of the pricing on the rating. Higher priced restaurants on average have slightly higher ratings and have fewer low ratings. However, these differences are not significant:



6. There are very few bad ratings (i.e. below 2.0). Most ratings are in the range 3-5



7. The business category has no significant influence on the ratings:



Methodology

In the EDA I found that the most interesting features for further analysis are Review Count and Rating.

I was deciding on the following step-by-step process in order to find the city hotspots:

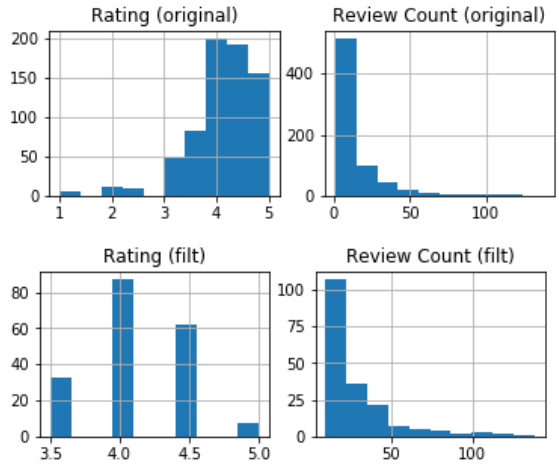
1. **Filter businesses**, which have a rating below 3.5 and a review count below 10. These businesses are assumed to be not interesting for our target audience.
2. **Plot** those **businesses** on a map for visual inspection
3. **Combine** the number of reviews and the average rating for each business **into** a single feature ("**business score**"). The weighting will be 50:50 (i.e. average combination). However before doing so, normalize those features into the interval [0,1], so they can be combined meaningfully.
4. Use kMeans **Clustering** in order to identify geographical clusters of businesses (i.e. dense areas of businesses)
5. **Combine** (i.e. average over) all "business scores" within each cluster, in order to come up with a **cluster score**, to find the hotspot clusters
6. **Plot heatmap** (acc. to cluster scores). However, in order to do so, it is reasonable to first normalize these cluster scores into the interval [0,1] so they can be plotted easier with a colormap.

Results

In this section I will go through the results which have been obtained by following the step by step process in the Methodology section.

1. Filter out uninteresting businesses

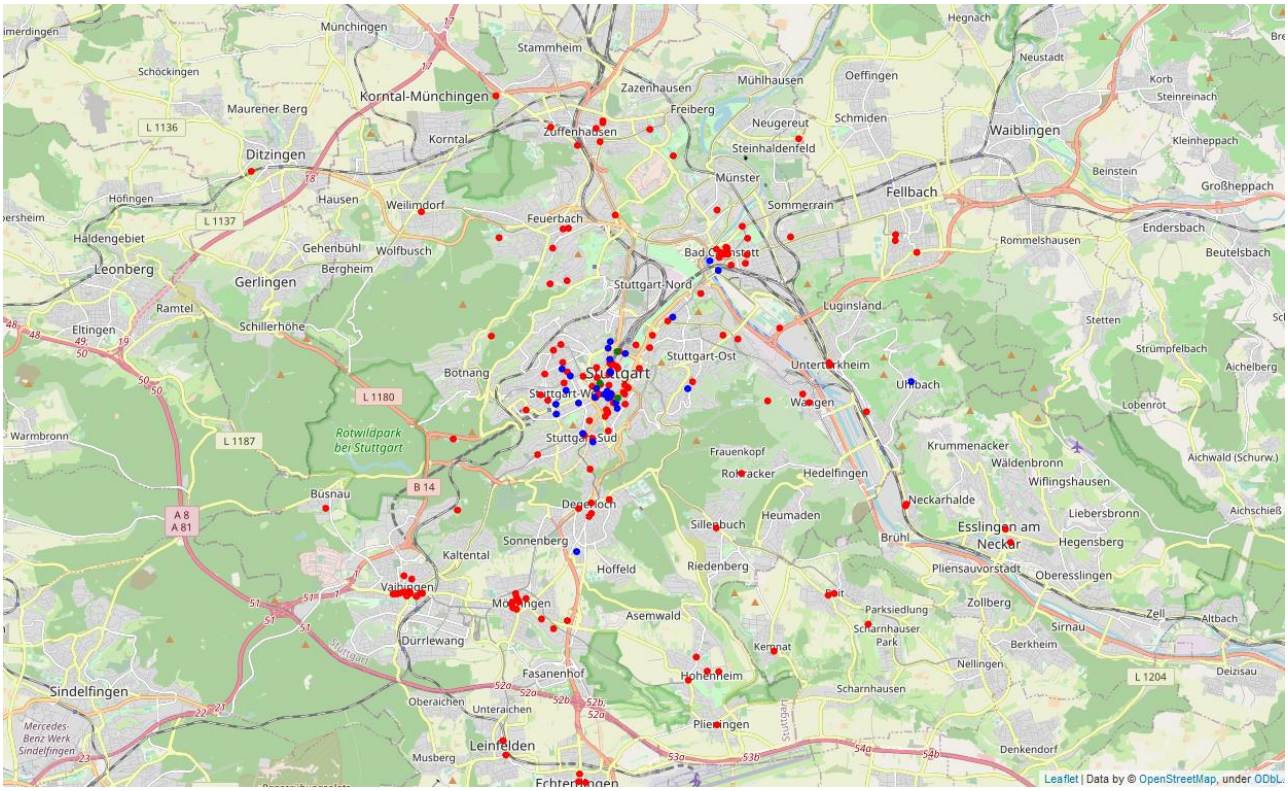
By filtering out uninteresting businesses, I reduced the number of business by 512. Therefore, the histogram of ratings and review counts changes accordingly:



2. *Plot interesting businesses on a map*

Plotting the businesses on a map, we find that most businesses are located in the city center of Stuttgart. I have used folium in order to create this map. The color encoding is:

- Red for Restaurants
- Blue for Bars
- Green for Clubs



When you hover over a business marker, you get additional info on this business:

3. *Calculate "business score" from "review count" and "rating" for each business*

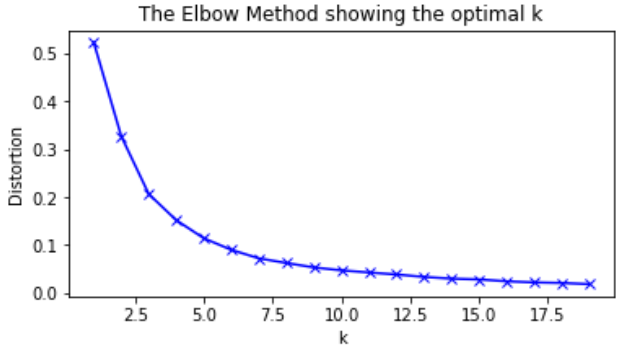
The business score is being calculated by first normalizing the review count and the rating using the sklean MinMaxScaler. Afterwards these normalized features are being averaged (i.e. 50:50 weighting) in order to form the "business score".

	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

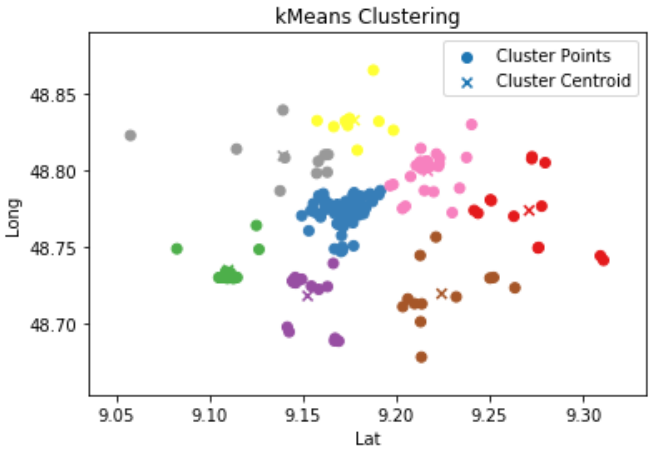
4. *Geographically cluster businesses*

The clusters are being calculated using kMeans Clustering.

Using the elbow method, we found that a reasonable number of clusters is 8:



The output of the clustering, using 8 cluster, looks like this:



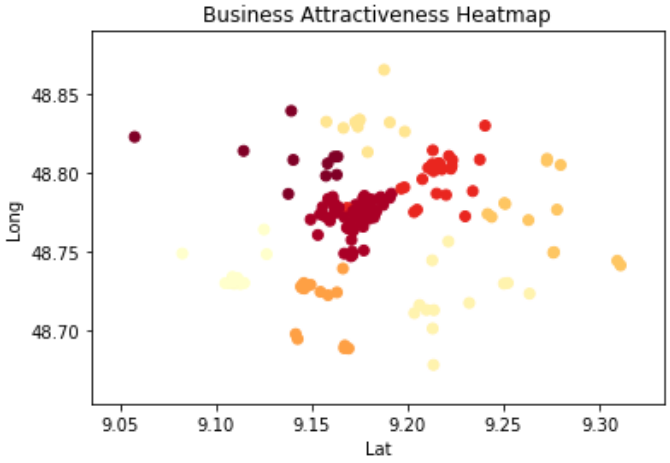
5. *Calculate "cluster score" from "business scores" within cluster*

The cluster score is being calculated by averaging the business scores of all businesses within a cluster. These cluster scores are then again being normalized into range [0,1].

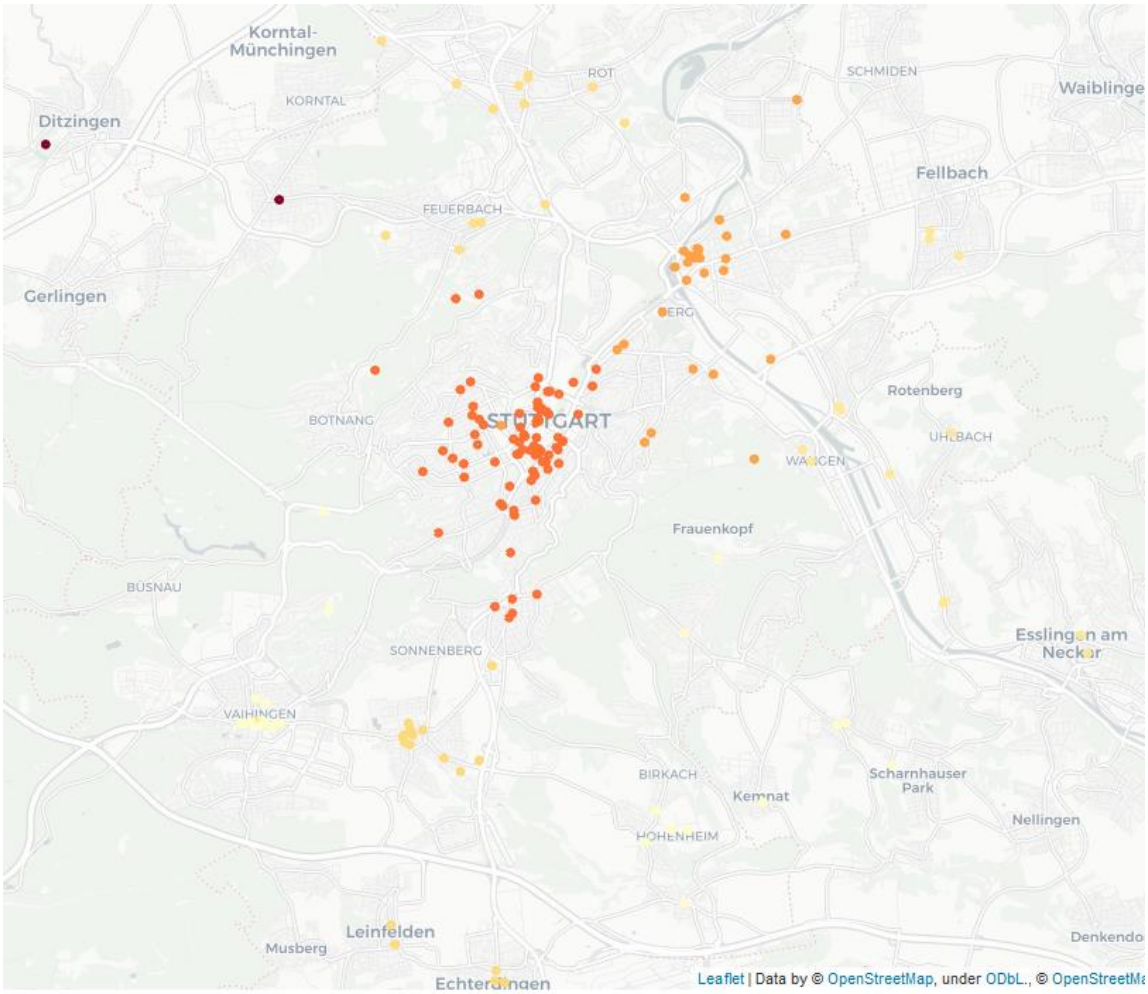
	id	name	review_count	rating	category	lat	long	review_count_normalized	rating_normalized	business_score	kmeans_cluster	cluster_score_normalized
0	qngSwQ3PmyYxmYRByOcccw	Gaststätte Schlesinger	28	4.5	restaurants	48.779510	9.172870	0.140625	0.666667	0.403646	1	0.910794
1	itdqzog_6HLeQEFQo_PBrA	Carls Brauhaus	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. *Plot clusters on map with heatmap encoding acc. to cluster scores.*

The resulting cluster scores can then be used in a colormap encoding for the kMeans cluster plot. Here all businesses within a cluster are being assigned the normalized cluster score respectively, so it is easier to visually inspect these clusters and identify the hotspots.



In a second step I projected this heatmap onto the folium map. The folium map is being greyed out a little, so it is easier to focus on the markers. It is evident, that the best place to go out is in Stuttgart Downtown, if you want to have a big number of great restaurants, bars and clubs in walking distance close to each other.



However, interestingly on the map we find two hidden gems northwest from Stuttgart, which might be worth exploring some other day. These are in Ditzingen (restaurant ”Da Michele”) and Korntal (restaurant “Baran Kebab”).

Conclusion

In this analysis, I have showed a methodology on how to find geographical hotspots of well-known and highly rated restaurants, bars and clubs in any city. I have applied this methodology to the city of Stuttgart (Germany). The results clearly show that the best place to stroll through the streets is Stuttgart Downtown, however there are some hidden gems worth exploring northwest of the city.

The analysis must be taken with some grain of salt due to several reasons. First yelp is not a very popular platform for business ratings in Germany, therefore the number of reviews is not very high for most businesses. Further there is some bias in the data due to limited number of users. Moreover, the technical limitations of the yelp developer API does not allow to retrieve all the businesses that are out there. Instead it limits the results to 50 businesses per query. Even though I found a workaround to easen this problem (querying all boroughs individually), it is still unclear how many more businesses are out there, that I could not retrieve.

Lastly, it is worthin noting that this analysis can be applied for every city in the world.