Where should I live? An early analysis about life quality between two big EU cities: Rome and Berlin

Capstone Project - The Battle of Neighborhoods (Week 1)

Vincenzo Scoca

July 9, 2021

Contents

1 Introduction

Many reports and rankings about city life quality are currently available on books, internet and other sources.
However, those reports do not always provide a clear evidence of how these scores are affected by the actual city structure. Essentially, they are computed by means of interviews and/or polls submitted to the citizens without actually mapping those results on the city infrastructures in order to detect the main features that led to those scores.
More in details, those studies do not carry out a further analysis on the distribution of the essential services and venues over the city area to explain the polls results.
For example, the Teleport Company [1] developed a great reporting service providing fully fledged reports about cities quality life, computed over many and very detailed indexes, such as cost of living, housing, safety, healthcare, education, taxation, tolerance and so on and so forth. However, to the best of my knowledge, the ranking methods do not take into account the any info about the infrastructure of the cities studied.
For example, let us consider the Education Score assigned to the city of Rome. According to what has been reported on their page ([2]) the city of Rome has an education score of 4 out of 10. This score is then explained by means of a set of indexes based on the results of some academic tests carried out in one or more city universities.
However, this analysis does not clearly reflects the actual satisfaction of the education service provided by in Rome. Indeed, for a more exhausting research, we should take into account not only a quality score based on the academic relevance of the universities, but also some complementary information as the distribution of any grade schools and universities among the city neighbourhoods. In this way, we can evaluate the quality of educational services, based not only on the academic quality of the service itself, but also on the easiness to reach the school. Indeed, having some few high quality schools eventually placed in one or just few neighbourhoods, should not lead to a high ranking for educational services.
Moreover, this further study about the distribution of services among neighbourhoods, will also support new and former citizens throughout the *"where should I live?"* decision making process. They will have a more comprehensive view of service quality, based not only on generic quality scores, but also on their actual location in the area.
Therefore, the aim of this work is to provide and early analysis about the distribution of ten main service categories, among the neighbourhoods of two main EU cities as Rome and Berlin, to extend the quality analysis provided by the Teleport Company.
Moreover, for each city we compared and grouped all the similar neighbourhoods in terms of venue categories distribution within the neighbourhood itself. The purpose was to check what is the distribution of similar districts in the city area.

2 Dataset

The main purpose of this work is to analyse the distribution of venues of different categories among the neighbourhoods of two main cities as Rome and Berlin.
Therefore, two datasets were needed:

- **Venue Dataset:** containing data related to the venue location and category;
- **Neighbourhoods Dataset:** containing all the relevant info about neighbourhoods zip-codes and coordinates;

2.1 Venue-Dataset

Fer each city we retrieved 50 venues for each of the following categories:

- Art;

[1] https://teleport.org/
[2] `https://teleport.org/cities/rome/`

- Food;
- Nightlife;
- University;
- Events;
- Outdoors Recreation;
- Professional Other Places;
- Residence;
- Shop Service;
- Travel Transport;

We used the Foursquare API [3] to collect the raw set of data the has been refined removing the useless columns not for the analysis purpose. Indeed, only the following columns has been held in the clean version of the dataset:

- "id"
- "name"
- "categories"
- "location.postalCode"

The "id" and the "name" column have been held only for the element description, whereas the "categories" and the "location.postalCode" provided the data needed to carry out the study about the distribution of the category venues among the different neighbourhoods.

2.2 Neighbourhoods Dataset

To find the exact location of each neighbourhood, we retrieved two datasets providing the latitude and the longitude of each postal code for both Rome and Berlin:

- **Rome:** `http://download.geonames.org/export/zip/`
- **Berlin:** `https://gist.githubusercontent.com/iteufel/af379872bbc3bf5261e2fd09b681ff7e/raw/205e9a5f1a9f2fb70b58bf9022cab3cf6cc13a19/zipcodes.germany.sql`

In this way we could easily represent on our map each neighbourhood.

3 Methodology

The raw data collected as in Sec.2.1 has been then cleaned has reported in Fig.1 In order to analyse the venue category distribution among the districts we needed only partial information about the venues and all the useless features have been removed from the dataframe.
Afterwards, a further data modelling operation has been carried out to obtain a suitable data representation for the evaluation of the category distribution. Basically, the data has been converted in their related binary representation by means of the **one-hot encoding**. Thus, each venue has been represented by its postal-code and its related binary encoding as in Fig.2.
Finally, the categories distribution among the cities districts has been carried out. All the venues have been grouped by their postal-code summing up the number of venues for each category located in each district. This analysis will show us how the categories are distributed among all the districts, to understand if all of them provide access to the every category.

[3] `https://developer.foursquare.com/docs/api-reference/venues/search/`

	id	name	categories	location.postalCode
0	4c1df65ffcf8c9b6fb8bac0b	Basilica di Santa Maria Maggiore	Art	00184
1	4adcdac6f964a520285321e3	Piazza Navona	Art	00186
2	4adcdac7f964a520735321e3	Piazza San Pietro	Art	00120
3	4adcdac6f964a520105321e3	Basilica di San Pietro (Basilica Sancti Petri)	Art	00120
4	4adcdac6f964a5202b5321e3	Trevi-fontein (Fontana di Trevi)	Art	00187

Figure 1: An excerpt of Rome cleaned data venues

	location.postalCode	Art	Events	Food	Nightlife	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport	University
0	00184	1	0	0	0	0	0	0	0	0	0
1	00186	1	0	0	0	0	0	0	0	0	0
2	00120	1	0	0	0	0	0	0	0	0	0
3	00120	1	0	0	0	0	0	0	0	0	0
4	00187	1	0	0	0	0	0	0	0	0	0

Figure 2: An excerpt of Rome one-hot data representation

A further study has estimated the Pearson's correlation between the categories to detect eventual connections between different categories within the districts. To check if the evolution the district over the time has followed some rule about the type of venues to develop.
Finally the similarity between the neighbourhoods has been computed, by means of an unsupervised clustering method as the **K-means** clustering. The data previously collected and grouped has been normalized and the used as input for the clustering process. Thus, the optimal of cluster has been selected through the Elbow Method. The final aim of the district clustering was to detect the similar districts and to show them on the map to see how similar cluster are distributed over the cities are.

4 Results

4.1 Venue Categories

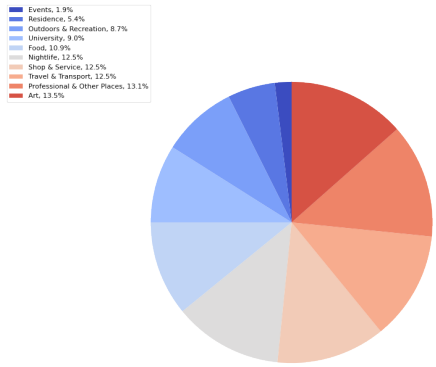
The early analysis provided an overview about the category distribution over the whole city area as depicted in Fig.3a and in Fig.3b. As reported in the graph, the overall trend is quite similar. Berlin and Rome have a very similar venue distribution over their area, the difference among the categories is very restricted (i.e. $\pm 1\%$).

4.2 Venue Categories District Distribution

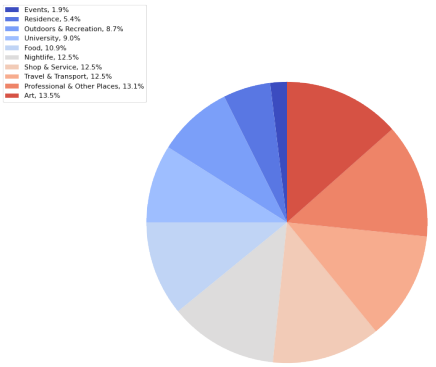
The next step of the study pointed out that the categories are not evenly distributed among the districts. As depicted in Fig.4 and Fig.5 each district has in average less than one venue for every category. Moreover, very few categories provided more that one location for each category Fig.4 5.

4.3 Venue Categories Correlation

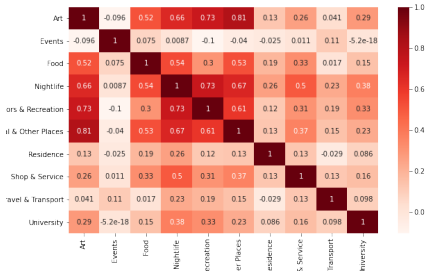
A further analysis has been carried out in order to discover the correlation between categories. As pointed out by the results in Fig.3c only few categories show a stronger correlation. The city of Berlin, instead, does not point out such correlation between the categories as reported in Fig.3d.



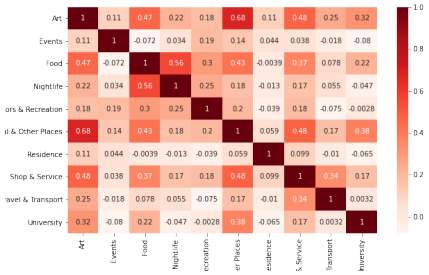
(a) Rome



(b) Berlin



(c) Rome



(d) Berlin

Figure 3: Rome and Berlin Venue Categories

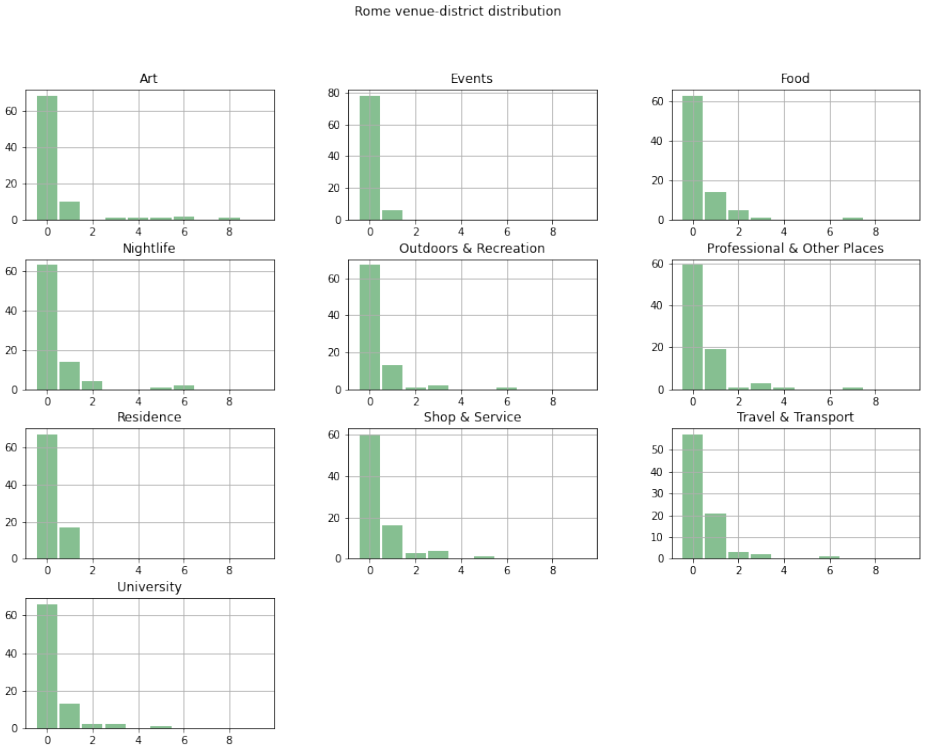


Figure 4: Rome category-district distribution

4.4 Venue Categories Clustering

For each city all the district have been grouped by means of the venue categories distribution to detect similar districts and locate them on the map to find out how they are located over the city area.
The optimal number of the K clusters has been estimated using the Elbow Method as reported in Fig.6 The optimal K is similar for both cities and it can be fixed on 5. The final clusters have then been placed on the map as in Fig.6c and Fig.6d

5 Discussion

As denoted by the results of this study, both cities are characterized by same venue categories distribution over the city area.
Indeed, the number of venues for each category over the whole city territory is similiar for both cities and also the distribution over the neighbourhoods does not reveal remarkable differences. Therefore, this study reveals a similarity between the cities that can be used as further support for the quality score computation carried out by the Teleport Company in [4] and [5] were they denoted larger differences for few indexes, especially the one related to the Culture Score.

[4] `https://teleport.org/cities/rome/`
[5] `https://teleport.org/cities/berlin/`

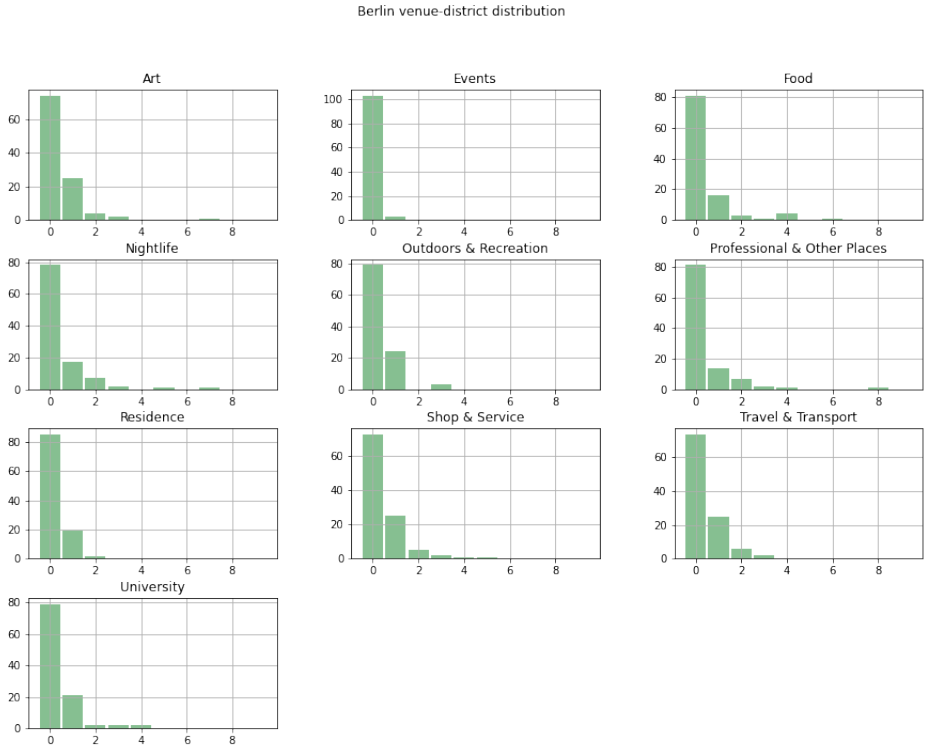
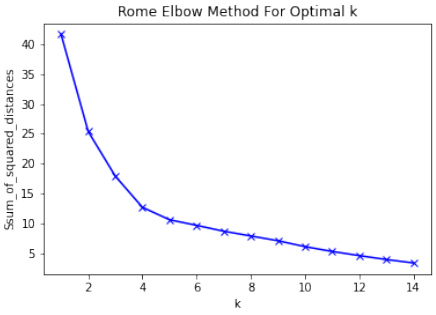
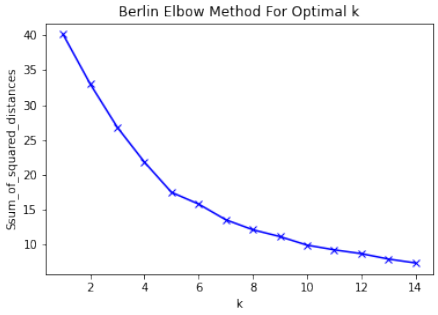


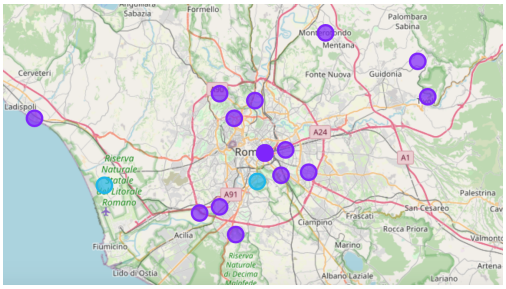
Figure 5: Berlin category-district distribution



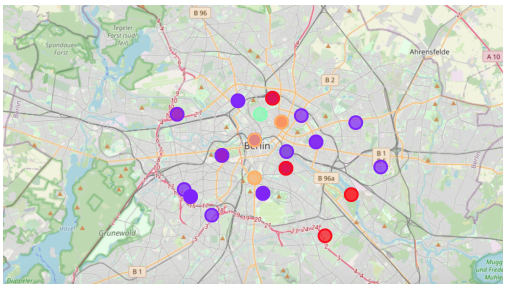
(a) Rome



(b) Berlin



(c) Rome



(d) Berlin

Figure 6: Optimal K estimation for and clusters visualization for Rome and Berlin

6 Conclusion

This work represent an early analysis about the city quality score, is not a comprehensive study. All the results must be extensively mastered with further elaborations to be considered as a real support for the city evaluation.
It only provides some insights about how to start a territorial analysis about the city services and how to use the outcome of such study.