CLUSTERING FOR HOUSE HUNTING IN BERLIN

CAPSTONE PROJECT - THE BATTLE OF NEIGHBOURHOODS

THIAGO DE CARVALHO - APRIL/2021

Starting Point

The problem situation is the complete lack of knowledge about the destination city of a couple that has just moved to Berlin. Some important starting point parameters based on the couple's preferences

The objective is to identify, based on the couple's preferences, the priority geographic narrow areas to start house hunting.

Couple’s preferences

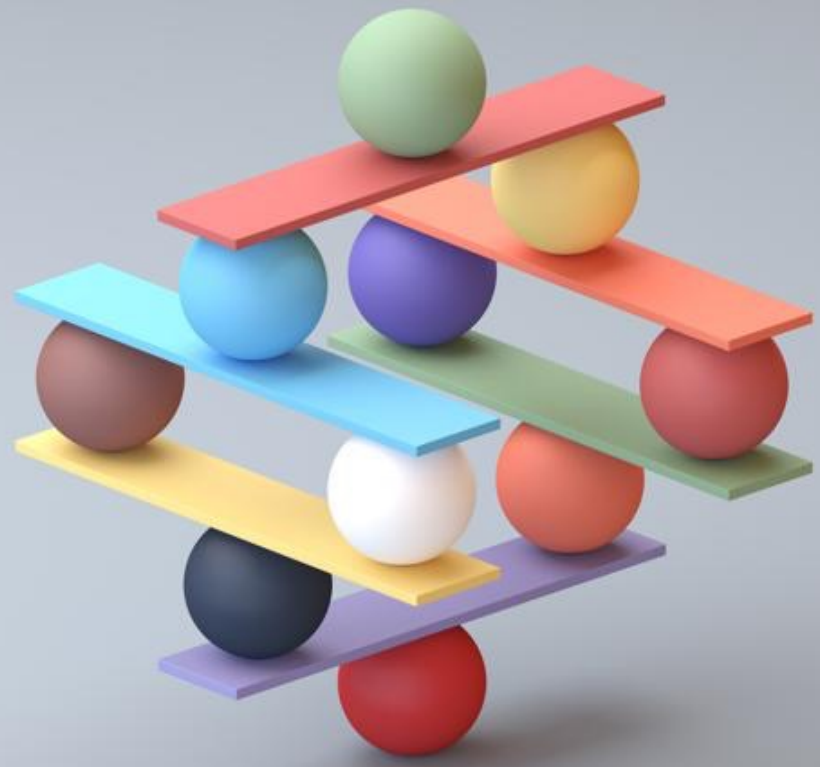
- Use of public transportation.
- Sports practitioners indoors and outdoors.
- They value social life, cooking at home, occasionally in restaurants.
- They want to live in neighborhoods that are considered safe.

Datasets

- Berlin 2020 venues by category geodata
  - https://developer.foursquare.com/
  - (extracted from foursquare API)
- 2019 Berlin Crimes by neighbourhood
  - https://www.kaggle.com/danilzyryanov/crime-in-berlin-2012-2019/download
- Berlin's neighbourhood geospatial data
  - http://data.insideairbnb.com/germany/be/berlin/2021-02-20/visualisations/neighbourhoods.geojson
- Average rental prices by Berliners districts.
  - https://www.statista.com/statistics/800765/rent-expenditure-apartments-berlin-germany-by-district/

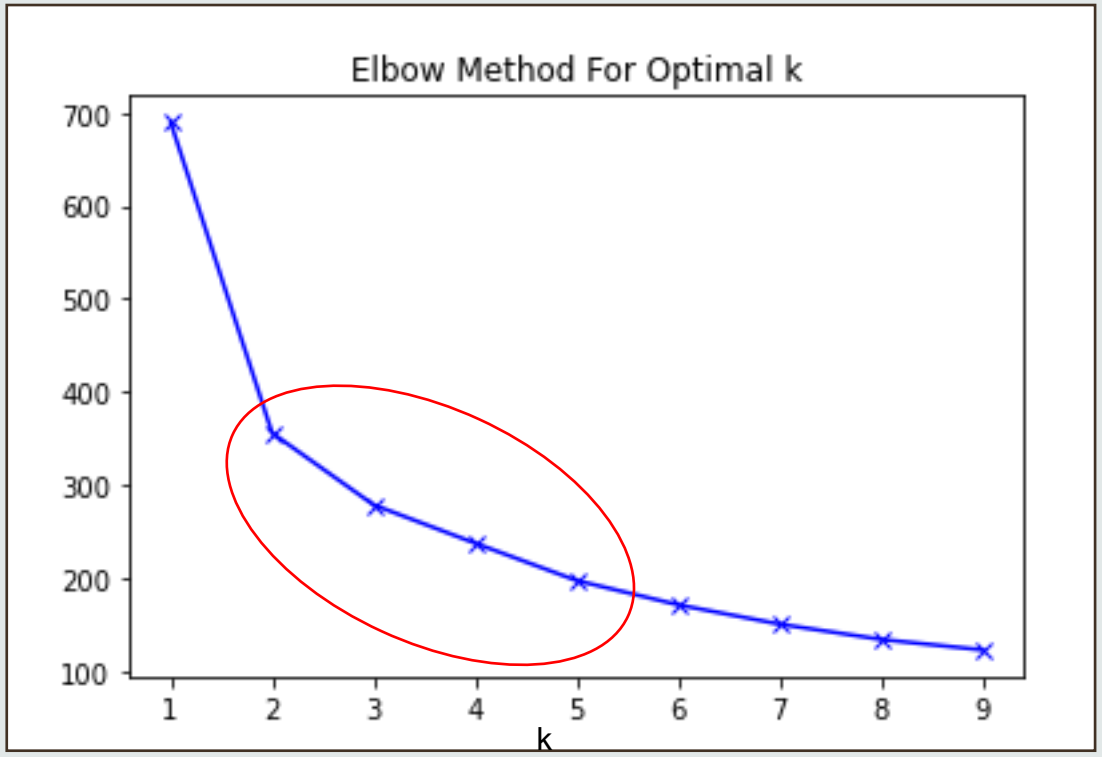
Data Analysis Stages

1. Data Acquisition and cleaning.
2. Get the coordinates of the centroids of each neighbourhood
3. Use Foursquare API, passing the central coordinates to get the venues information of each neighbourhood
4. Set the priority venues categories (couple's perspective)
5. Merge geodatabase, with priority venues categories, crimes and rental average prices per neigbourhood



Finding prospects neighbourhoods with k-means clustering (1/3)

Elbow Method

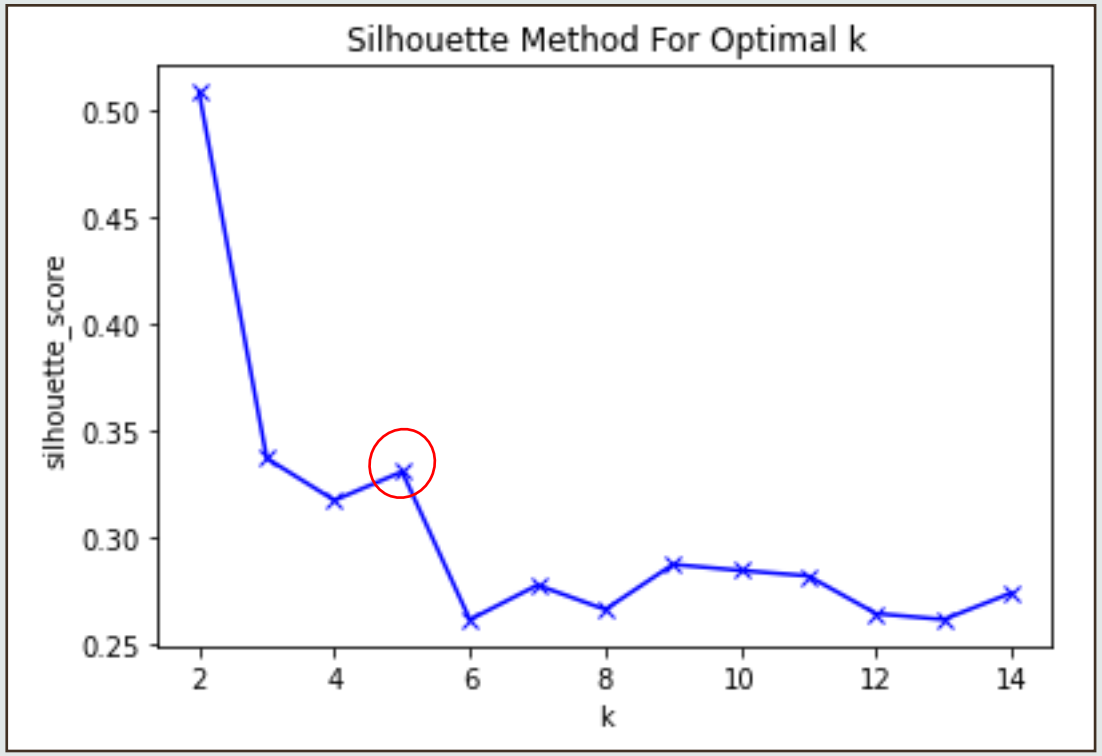


Considering the dataset fitted, **the optimal number of clusters should be between 2 to 5 clusters** according to the elbow method.

Observing the line there was not a relevant specific point, despite the k=2 that does not seem to be convenient to understand almost 140 neighbourhoods of a so diverse city as Berlin. In this case is convenient to check other metric, to decide about the number of clusters to pick.

Finding prospects neighbourhoods with k-means clustering (2/3)

Silhouette_score



The Silhouette score graph showed the optimal, again despite the k=2, **that clustering in 5 groups bring more separate groups** with slightly difference for 3 that also seemed small as k=2.

So, for the reasons described above and considering the results of the 2 metrics, **the decision for the number of clusters was set to k=5 clusters.**

Cluster Decision

The bar graph shows each cluster profile considering their mean in each variables of the model.

Observing closely the cluster that were not chosen:

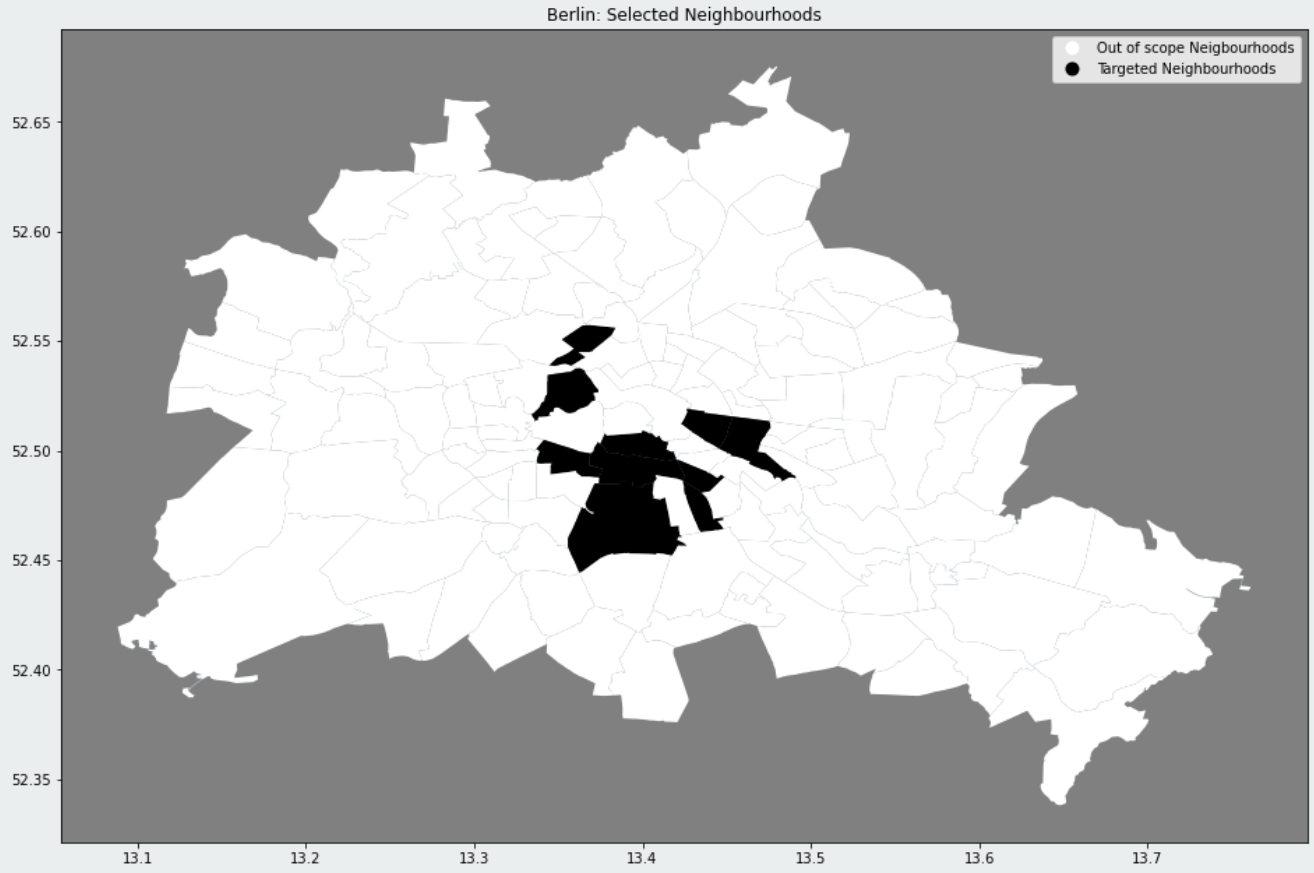
- Cluster 4 – They are the less safe neighourhoods.
- Clusters 0,1 and 3- Although being the safer neighourhoods, they are not meeting significatively all dimensions the couple's need.

The chosen one:

- Cluster 2 - Despite its moderate safety, it has really representative averages on the 4 priority venues categories.

The couple's set needs are considerably found in Berlin's downtown surrounds.

- As result of the k-means clustering stage the searching zone was reduced from 138 neighbourhoods to 10. **A reduction of 93% in the number of searching areas.**
- In order to geographically see the targeted neigbourhoods the clustering preliminary results, a map was plotted →

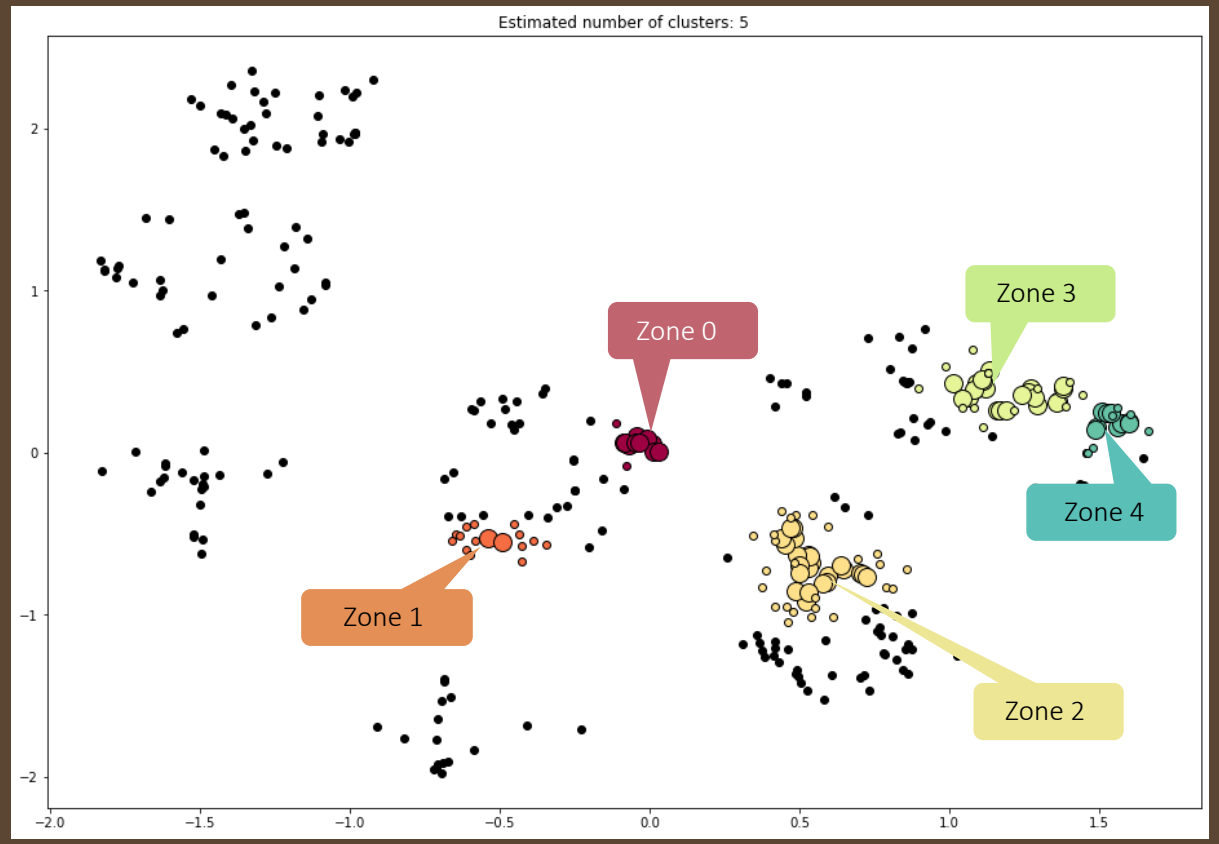


Finding narrower zones to house hunting with DBSCAN

DBSCAN was set with two hyperparameters to assure finding narrower areas within the targeted neighbourhoods.

As result DBSCAN found 5 zones with high density of venues that meets the couple's interests.

Theses zones concetrate 40% of the venues of the 10 neighbourhoods regardless of the neighbourhood that the venues belonging to.



Finding more convenient zones regarding average m² price

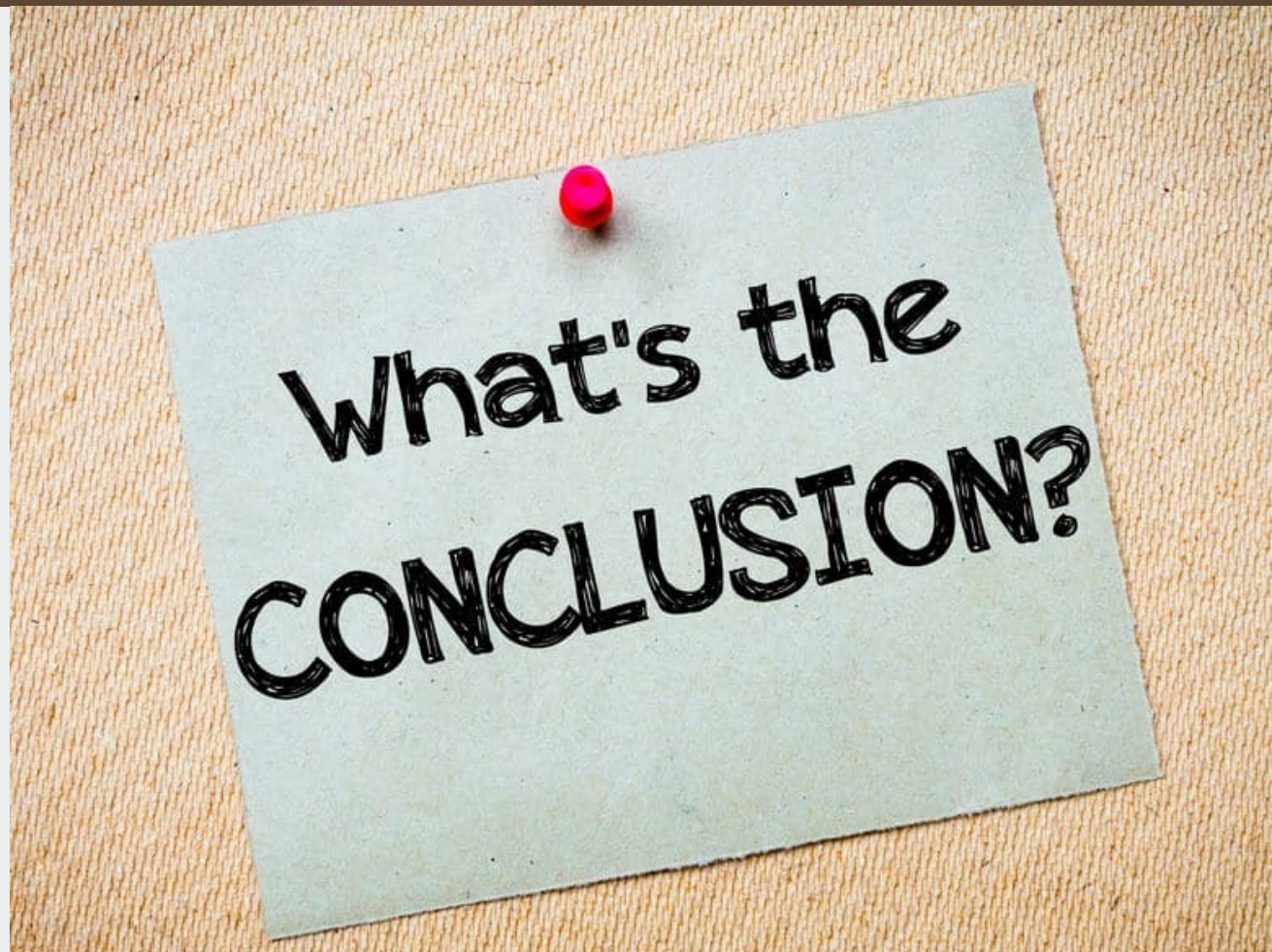
- The Berlin's map shows the neighbourhood boundaries with the average rental price per squared meter and the five blue circles of the zones.
- By the map its possible say that, considering the lower prices of average square meters, the targeted zones starting point for the house hunting is zone 2.

B4 - Conclusion

It is important to point out that this exercise is intended to meet the couple's needs (only and exclusively).

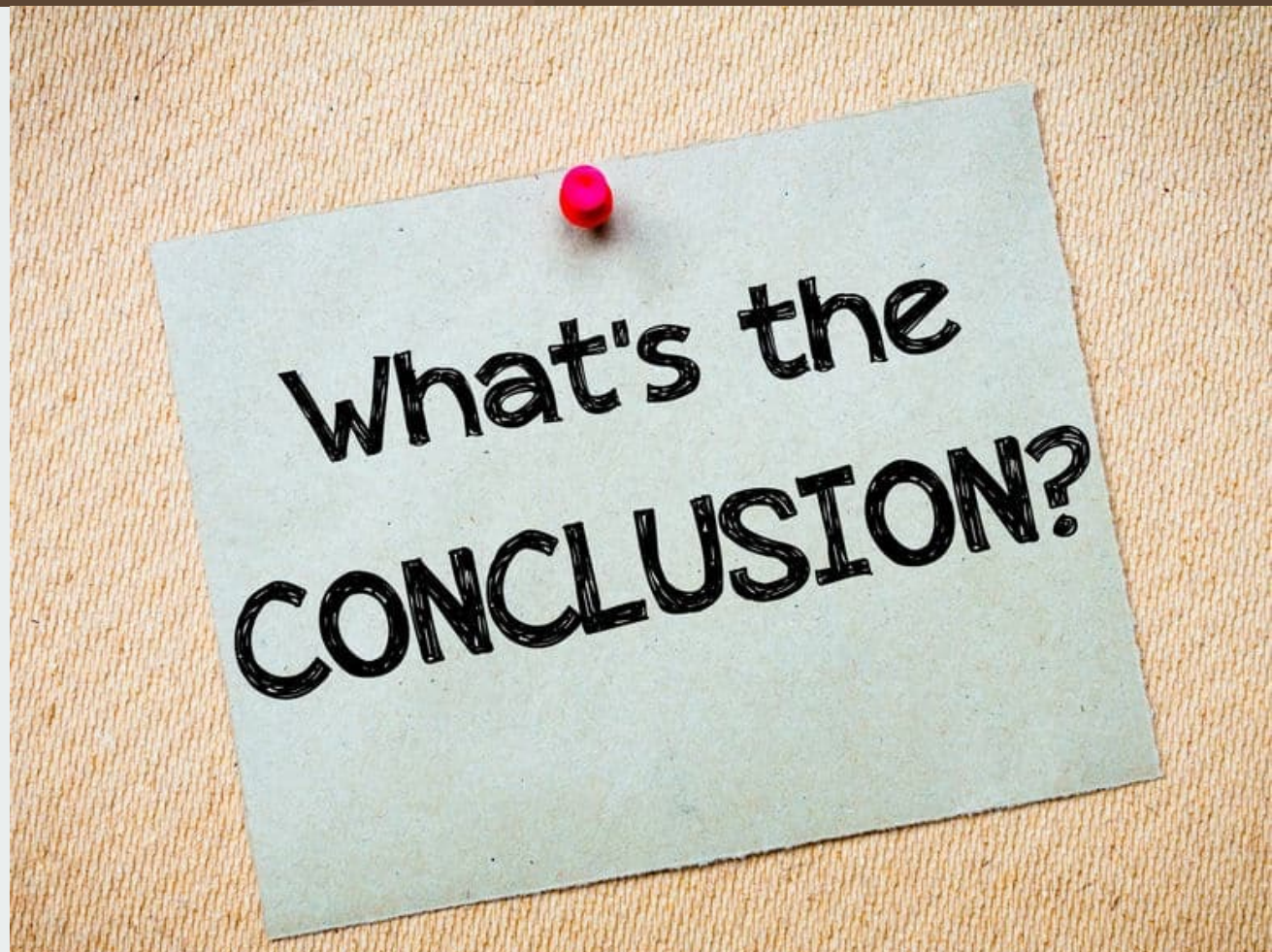
In no way does the analysis serve as a comparison of the attractiveness of Berlin's neighborhoods.

The actual data presented here show that Germany's capital is an attractive and safe city to live.



Conclusion

- Considering the results presented above in the analysis the couple's apartment search, clustering methods applied to geodata were satisfactory. The search area for a flat to rent was reduced from 138 possible to 5 mainly desirable/priority areas for the couple.
- Taking into consideration the average price per square meter of the properties for rent in these neighborhoods, the search should start in the zone marked on the south-center of Berlin.



Project evolution

As an end user needs, would be necessary apply some Web scrapping techniques for searching data from announced flats.

As insights for future evolutions of this project, thinking about develop a front-end application as a commercial plugin for current house hunting websites:

- Would be useful developing a front end in order to allow anyone to easily set up the priority venues parameters.
- Using machine learning algorithms to calculate by a specific location the adherence to the set parameters.
- Addition of variables such price and all related characteristics of the flats (size, number of rooms, etc).