# CHOOSING THE BEST CITY FOR A LIVING

LONDON / PARIS / AMSTERDAM

**02.01.2021**

## COURSERA CAPSTONE DATA SCIENCE PROJECT SELECTION OF A EUROPEAN CITY FOR A LIVING.

Diana Tolstyga

## 1. INTRODUCTION TO THE BUSINESS PROBLEM.

We are a family of three (husband, wife, four years old kid) from Toronto planning to move to Europe. Occasionally we have got three job offers: at London (UK), Amsterdam (Netherlands), and Paris (France) and need to decide which city to choose for a living. For now, each city looks equally attractive to us: expecting salary, way of living, language, climate, culture, education, medicine - all these aspects are comparable to us across these three places and does not help us to make the right decision. The only thing that helps to decide where to move - is its

downtown's infrastructure: number of interesting places (sights, restaurants, children activities, sport centers) that makes our life comfortable and full. The Foursquare data will help us to identify all these places and make the final decision by comparing the results for all three cities.
All these three cities are being very attractive for young well-educated STEM ex-pats who want to immigrate to Europe and the Data Science project is meant to help that group of people in making the right choice.

# 2. DATA

We will explore only 3-5 downtown neighborhoods for each city. The reason for such limitation is:
(a) all companies that offered a job are located at the very center of each city;
(b) we are not interested to spend too much time/money for a commute;
(c) we want to live in a very diverse area full of various interesting places and enjoy exploring all of them by walking around.
To get those neighborhoods (incl. longitude/latitude) we have to manually gather data from google maps by taking the longitude and latitude of a centroid of each neighborhood:

2.1. Paris.
Top neighborhoods are taken from the site [https://www.parisinsidersguide.com/paris-neighborhoods.html](https://www.parisinsidersguide.com/paris-neighborhoods.html) and five the nearest ones to the hypothetic office have been selected:

| | Borough | Borough\nLocal name | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | Eiffel Tower | Eiffel Tower | Eiffel Tower | 48.853684 | 2.301351 |
| 1 | The Heart | 1st arrondissement of Paris | 2nd arrondissement of Paris | 48.863698 | 2.335927 |
| 2 | Marais | Le Marais | Le Marais | 48.858677 | 2.360471 |
| 3 | Latin Quarter | Quartier latin | Quartier latin | 48.847843 | 2.353432 |
| 4 | St. Germain | Saint-Germain-des-Prés | Saint-Germain-des-Prés | 48.853890 | 2.334440 |

2.2. London:
Top boroughs are taken around Buckingham Palace which is the location of the potential office in London and have the following locations:

| | Borough | Neighbourhood\nLocal Name | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | Mayfair | Mayfair | Mayfair | 51.510314 | -0.147814 |
| 1 | Knightsbridge | Knightsbridge | Knightsbridge | 51.499878 | -0.167863 |
| 2 | Kensington | Kensington | Kensington | 51.498138 | -0.192754 |
| 3 | Covent Garden | Covent Garden | Covent Garden | 51.512428 | -0.123448 |
| 4 | Carnaby | Carnaby | Carnaby | 51.513142 | -0.138690 |

2.3. Amsterdam.
Top three boroughs are taken as the nearest ones to the hypothetic office buildings. For each borough we have selected the biggest neighborhood based on the variety of postcodes from postcode site: https://postcode.site/noord-holland/municipality/amsterdam/district/grachtengordel-west. Then google maps are used to get the exact location:

| | Borough | Borough\nLocal name | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | Canal Ring | Grachtengordel | Felix Meritisbuurt | 52.369860 | 4.885504 |
| 1 | The Jordaan | Jordaan | Elandsgrachtbuurt | 52.375380 | 4.880826 |
| 2 | Museum Quarter | Museumkwartier | Concertgebouwbuurt | 52.355504 | 4.882079 |

2.4. Venues.
To get top venues for each neighborhood of the city, we use Foursquare data which include places of interest, its location, and venue category for each selected neighborhood. Venue category will be later used to rate each category based on our personal preferences and will help to make a final decision.

# 3. METHODOLOGY

To choose the best neighborhood in each city we will use a combined rating calculated from two different approaches:

(1) First approach is using classification modeling to categorize the venues and to select the most diverse neighborhood.
For classification modeling we will use k-clustering analysis to cluster venues categories and to select the most diverse neighborhood i.e. the one that has the biggest number of venues with different clusters
The neighborhood with venues from all clusters will get the highest rating.

(2) Second approach is using our personal ranking system. To do that we manually rank each category from 1 to 10 based on our personal preferences i.e. science museums and cultural places are getting the highest score (9-10) because we like science, while smoke shops and liquor stores are getting the lowest score (1-2) because we are not smoking and drinking fans. The more places with high venue category's rating in the neighborhood, the higher its total ranking.

After the best neighborhoods are chosen for each city, we compare its total ranking and choose the city with the highest rank as a final winner.

3.1. Paris
Based on longitude and latitude of each neighborhood we request information on venues from the Foursquare:

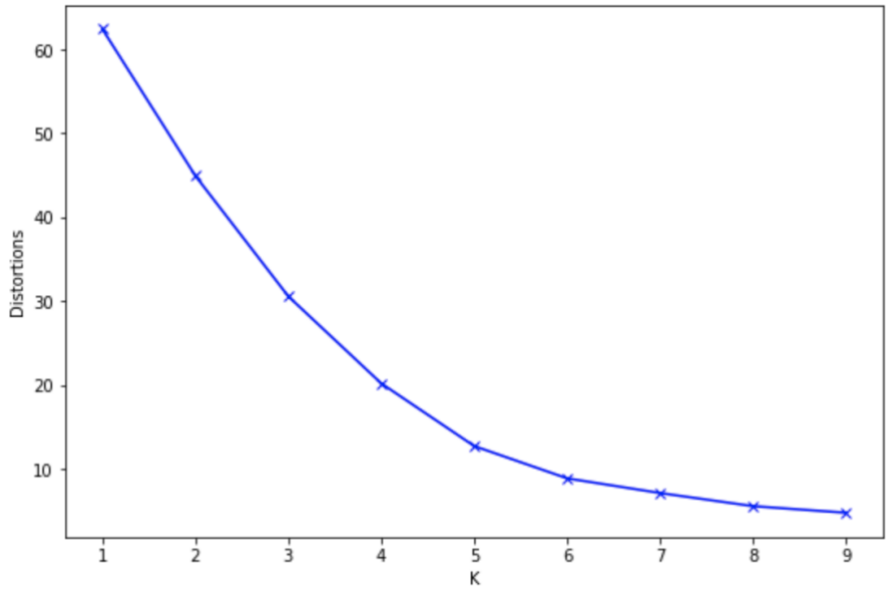
Table 4. Paris Venues and Neighborhoods with its coordinates.

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Eiffel Tower | 48.853684 | 2.301351 | École Militaire | 48.852553 | 2.303390 | Historic Site |
| 1 | Eiffel Tower | 48.853684 | 2.301351 | Parc du Champ de Mars (Jardin du Champ-de-Mars) | 48.855567 | 2.298760 | Garden |
| 2 | Eiffel Tower | 48.853684 | 2.301351 | Place Joffre | 48.853004 | 2.302698 | Plaza |
| 3 | Eiffel Tower | 48.853684 | 2.301351 | Aux Cerises | 48.853653 | 2.297019 | Tea Room |
| 4 | Eiffel Tower | 48.853684 | 2.301351 | Kozy | 48.855395 | 2.305185 | Coffee Shop |

```
Paris_venues.shape
```

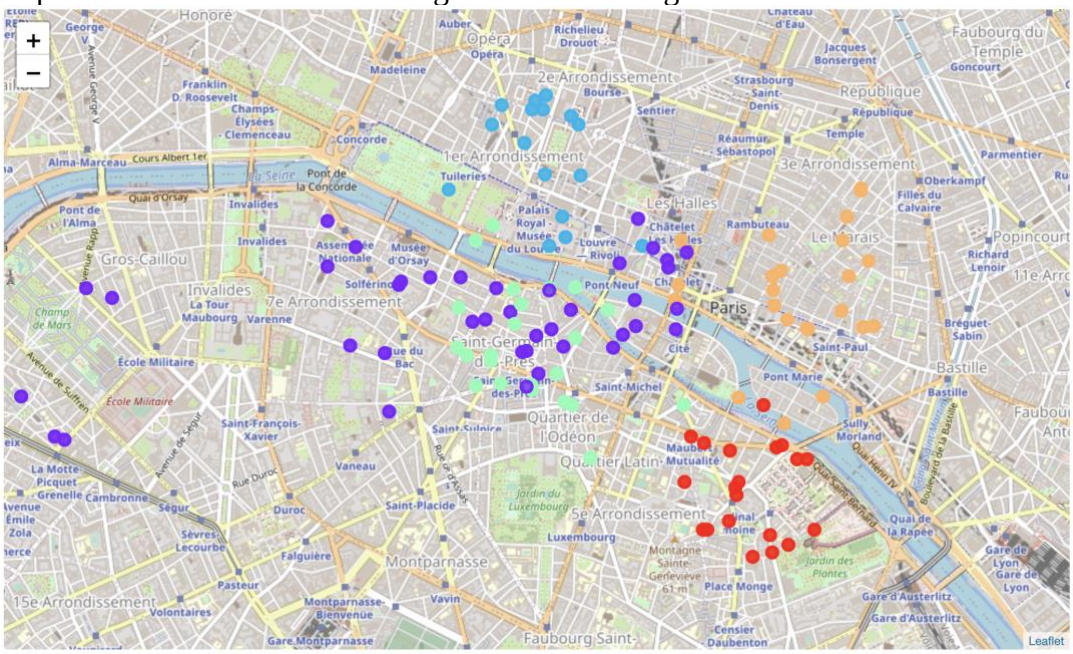
```
(460, 7)
```

The Foursquare portal provides us the information of all nearby places and its categories, together with longitude and latitude of each place. We will re-group those 460 venues by venue category to get a mean of each category in a particular neighborhood. And we will use these data for further clustering analysis as we need to categorize venue categories in some way.
To proceed the k-clustering analysis let us first choose the best k. To do that, we will define the distortions function for k in range (1,10) and select k for which distortions decrease significantly which is k=5:
Chart 1. Distortions function for k-clustering of venues in Paris neighborhoods.




Once we receive all the clusters, we can visualize the results on a map:

Map 1. Five clusters of venue categories in Paris neighborhoods.




We can now choose the best neighborhood in Paris. To do that we will use a combined rating:

1. The first rating is using a cluster analysis.

We have 5 clusters of venue categories and our preference is to live at the center of the city at the most diverse community which means that we want venues from different clusters to present in our neighbourhood. That is why the metrics which considers the variety of clusters is taken into consideration. The neighbourhood with venues from all 5 clusters will get the highest rating.

To get the first rating we make some transformations with Paris data and then rate each neighbourhood based on the following assumptions: (1) if the value for a venue in a neighbourhood column is less than 1.0, we assume that this type of venue is underrepresented in a neignbourhood and it earns zero point; (2) if the value for a venue in a neighbourhood column is more or equal to 1.0. we assume that this type of venue is presented in a neighbourhood and we keep it and its cluster label. And for each such cluster, a neighbourhood earns 10 points, i.e. if the neignbourhood has venues from all 5 clusters, it will get 50 points.

In more detail, a particular neighborhood is getting 10 points for each cluster that is essentially presented in it.[1]

[1] Essentially presented in a neighborhood means the following: we assume alpha > 1, where alpha is the sum of means of venues in particular category across all the neighborhoods. This methodology restricts from the analysis those categories, which have small amount of venues as well as the categories with a little presence in a particular cluster.

Table 5. 1-st ranking of Paris neighborhoods.

| 1nd arrondissement of Paris | Eiffel Tower | Le Marais | Quartier latin | Saint-Germain-des-Prés |
|---|---|---|---|---|
| 30 | 10 | 20 | 40 | 30 |

We see that Quartier latin has the highest rating while Eiffel Tower has the lowest one.

2. The second rating is using our personal ranking system.

To do that we first need to get the list of all venues and manually rank each category from 1 to 10. The combined list of categories for all the cities has 206 different categories which we manually rate; its sum for each neighbourhood will lead to a total rank.

Table 5. Venue Category Ranking combined with clustering results

| | Venue Category # | Venue Category | Ranking | Cluster Labels | 1nd arrondissement of Paris | Eiffel Tower | Le Marais | Quartier latin | Saint-Germain-des-Prés | Venue Latitude | Venue Longitude |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | American Restaurant | 1 | 3.0 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 | 48.854200 | 2.330697 |
| 1 | 1 | Art Gallery | 7 | 4.0 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 0.000000 | 48.859120 | 2.362594 |
| 2 | 2 | Art Museum | 8 | 2.0 | 0.600000 | 0.000000 | 0.400000 | 0.000000 | 0.000000 | 48.859936 | 2.344638 |
| 3 | 3 | Asian Restaurant | 3 | 4.0 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 0.000000 | 48.855680 | 2.362163 |
| 4 | 4 | Bakery | 5 | 1.0 | 0.142857 | 0.214286 | 0.142857 | 0.357143 | 0.142857 | 48.854395 | 2.335114 |

Function to define a rank of each neighbourhood is a product of Ranking column and the data for each neighbourhood:

We can see that the first neighborhood has the highest rank:

Table 6. 2-nd ranking of Paris Neighborhood:

| 1nd arrondissement of Paris | Eiffel Tower | Le Marais | Quartier latin | Saint-Germain-des-Prés |
|---|---|---|---|---|
| 140.9 | 57.08 | 131.23 | 119.37 | 107.4 |

After combining these two ranking we have got the following results:

Table 6. Final ranking of Paris Neighborhoods.

| 1nd arrondissement of Paris | Eiffel Tower | Le Marais | Quartier latin | Saint-Germain-des-Prés |
|---|---|---|---|---|
| 170.9 | 67.08 | 151.23 | 159.37 | 137.4 |

Where we could see that “1nd Arrondissement of Paris” neighborhood has got the highest rank.

## 3.2. London

We will now repeat the process for London neighbourhoods.

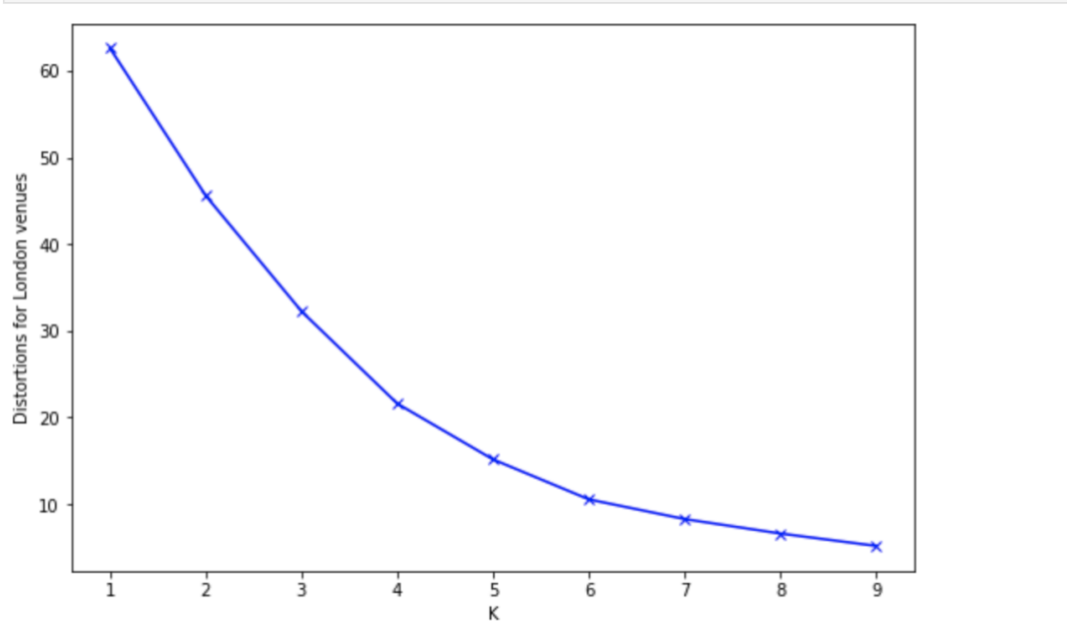
To do that, we will import the venues from Foursquare for each neighbourhood and save the results to a dataframe.

```
London_venues.head()
```

```
Mayfair
Knightsbridge
Kensington
Covent Garden
Carnaby
```

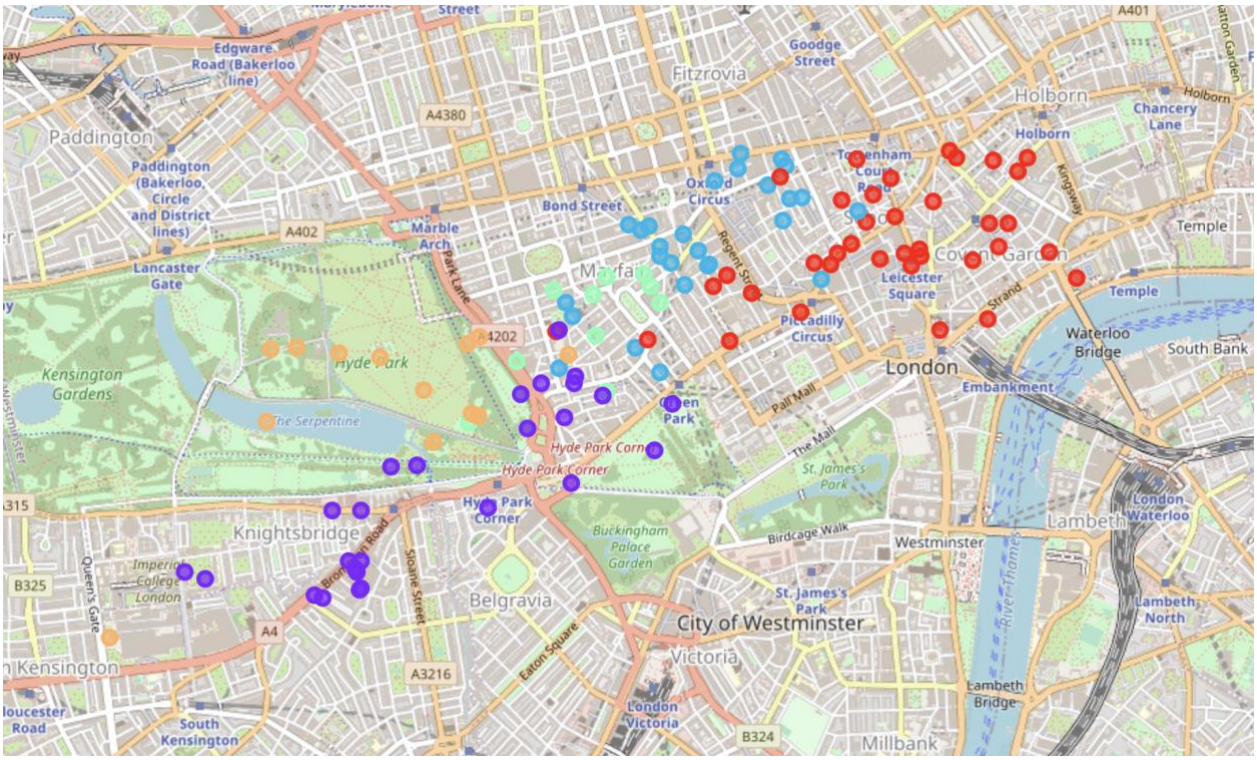
| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Mayfair | 51.510314 | -0.147814 | The Connaught | 51.510138 | -0.149498 | Hotel |
| 1 | Mayfair | 51.510314 | -0.147814 | Phillips | 51.510381 | -0.147017 | Art Gallery |
| 2 | Mayfair | 51.510314 | -0.147814 | Hedonism Wines | 51.510803 | -0.147450 | Wine Shop |
| 3 | Mayfair | 51.510314 | -0.147814 | Connaught Bar | 51.510042 | -0.149628 | Hotel Bar |
| 4 | Mayfair | 51.510314 | -0.147814 | Annabel's | 51.509352 | -0.146759 | Lounge |

To complete the first ranking we need to proceed k-clustering analysis (choosing the best k preliminary) and calculate the rank of each neighborhood based on clusters representation.
From the picture below we can see that the best k is 5, so we proceed k-clustering analysis for k=5.




| | Cluster Labels | Venue Category | Carnaby | Covent Garden | Kensington | Knightsbridge | Mayfair |
|---|---|---|---|---|---|---|---|
| 0 | 1 | American Restaurant | 0.000000 | 1.000000 | 0.0 | 0.000000 | 0.0 |
| 1 | 0 | Art Gallery | 0.400000 | 0.000000 | 0.1 | 0.100000 | 0.4 |
| 2 | 0 | Art Museum | 0.333333 | 0.333333 | 0.0 | 0.333333 | 0.0 |
| 3 | 0 | Arts & Crafts Store | 1.000000 | 0.000000 | 0.0 | 0.000000 | 0.0 |
| 4 | 3 | Asian Restaurant | 0.000000 | 0.000000 | 0.0 | 0.500000 | 0.5 |

And visualize the results on a map:




An then we will use the same two metrics methodologies to rate the neighborhoods.

1. We calculate the first ranking of a neighbourhood based on its cluster labels u sing the same methodology as we used for Paris and see that the neighbourho od “Covent Garden” has the highest score which is 40 :

| Carnaby | Covent Garden | Kensington | Knightsbridge | Mayfair |
|---|---|---|---|---|
| 30 | 40 | 10 | 30 | 30 |

2. Now we calculate the second ranking based on our personal preferences of so me neighbours. Again, we will use venues' categories list to get it:

| Carnaby | Covent Garden | Kensington | Knightsbridge | Mayfair |
|---|---|---|---|---|
| 77.05 | 100.38 | 36,63 | 48.95 | 63.99 |

We combine these two ratings to get total score for each neighborhood and choose the best one:

| Carnaby | Covent Garden | Kensington | Knightsbridge | Mayfair |
|---|---|---|---|---|
| 107.05 | 140.38 | 46.63 | 78.95 | 93.99 |

We can conclude that the best neighbourhood in London is "Covent Garden" base on our rating system.

## 3.3. Amsterdam

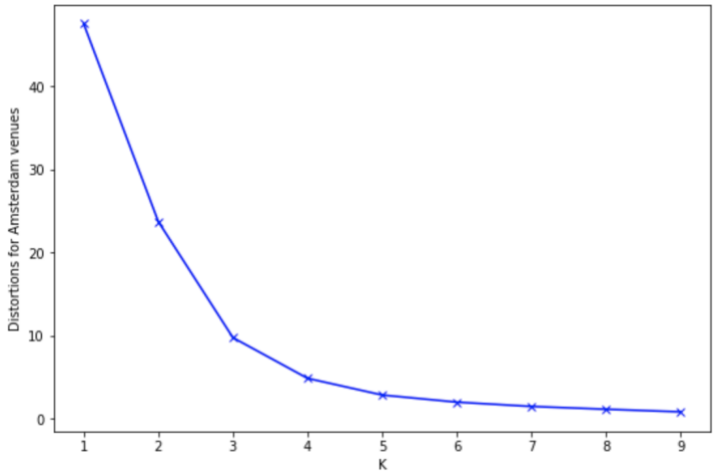
We will now reapeat the same algorithm for all selected neighbourhoods in Amsterdam.
As before, we get a test sample of venues for one neighbourhood only from Foursquare first and then replicate the process for all the remaining neighbourhoods.

```
Amsterdam_venues.head()
```

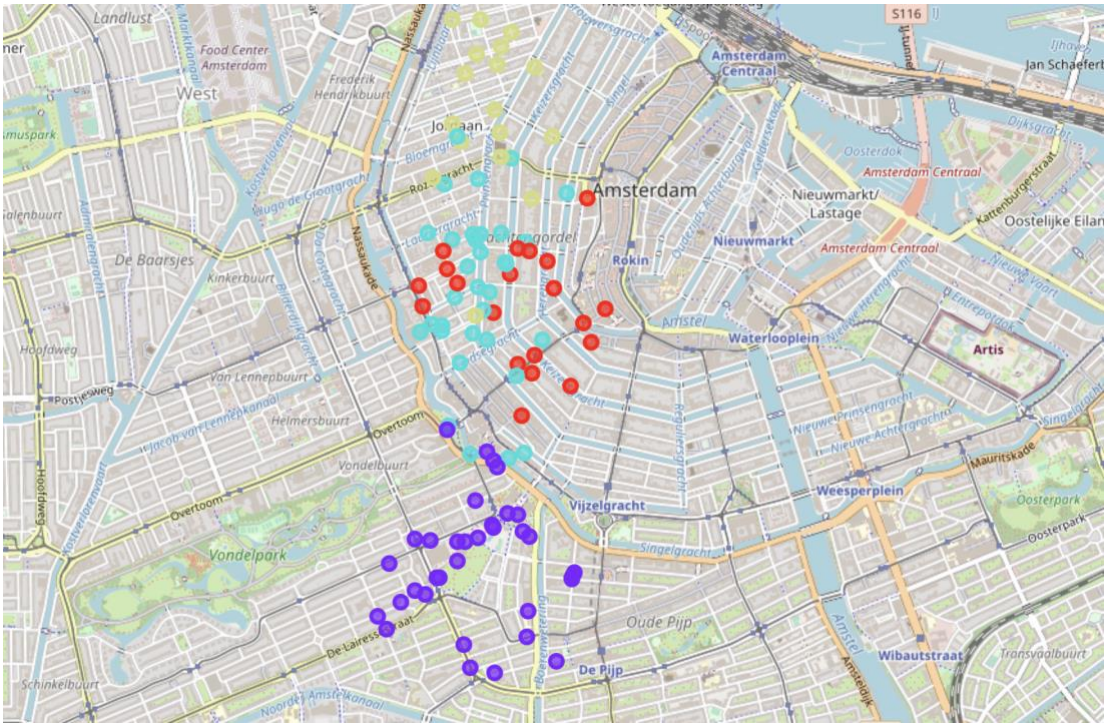
```
Felix Meritisbuurt
Elandsgrachtbuurt
Concertgebouwbuurt
```

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Felix Meritisbuurt | 52.36986 | 4.885504 | Jumbo City | 52.370369 | 4.885127 | Organic Grocery |
| 1 | Felix Meritisbuurt | 52.36986 | 4.885504 | 9 Straatjes | 52.370251 | 4.885907 | Shopping Mall |
| 2 | Felix Meritisbuurt | 52.36986 | 4.885504 | The Lebanese Sajeria | 52.368756 | 4.887505 | Lebanese Restaurant |
| 3 | Felix Meritisbuurt | 52.36986 | 4.885504 | The Dylan | 52.369273 | 4.884004 | Hotel |
| 4 | Felix Meritisbuurt | 52.36986 | 4.885504 | Urban Cacao | 52.368888 | 4.885056 | Chocolate Shop |

As before, we proceed k-clustering analysis for the best k, and get 4 clusters for Amsterdam venues.




We will visualize the results:




Now we have enough information to complete a ranking analysis using two0way approach.

1. The first ranking is based on cluster results.

We will modify the previous methodology  so we could get comparable rating for all three cities. For both Paris and London we had 5 clusters to take into consideration

and the maximum rating was equal to 50 (10 * maximum number of clusters presented in a neighbourhood). That is why we will normalize our rating values to match the results for all three cities.

| | Cluster Labels | Concertgebouwbuurt | Elandsgrachtbuurt | Felix Meritisbuurt |
|---|---|---|---|---|
| 0 | 0 | 0.783333 | 0.000000 | 19.216667 |
| 1 | 1 | 33.103030 | 0.772727 | 1.124242 |
| 2 | 2 | 7.714957 | 12.243162 | 13.041880 |
| 3 | 3 | 0.333333 | 17.666667 | 0.000000 |

| Concertgebouwbuurt | Elandsgrachtbuurt | Felix Meritisbuurt |
|---|---|---|
| 26.7 | 26.7 | 26.7 |

We can see that all neighborhoods have the same rating.

2. Let us calculate the second ranking for all Amsterdam neighbourhoods.

To do that, as before , we add our personal ranking to Amsterdam data and calculate the score for each neighbourhood as a sum of ranking across all venues categories normalized by its number:

| Concertgebouwbuurt | Elandsgrachtbuurt | Felix Meritisbuurt |
|---|---|---|
| 119.71 | 82.62 | 103.67 |

The neighbourhood 'Concertgebouwbuurt' has the highest rank with the score 119.71.

| Concertgebouwbuurt | Elandsgrachtbuurt | Felix Meritisbuurt |
|---|---|---|
| 146.41 | 109.32 | 130.37 |

Having both ranking we can now combine them to receive the total rank of Amsterda m neighbourhood and choose The neighbourhood 'Concertgebouwbuurt' has the high est total rank with total score: 146.37

## 3.4. Choosing the best city for a living.

To find the best city for a living we will just compare its total score:

| Paris | London | Amsterdam |
|---|---|---|
| 170.9 | 140.38 | 146.41 |

And conclude that **Paris** is the best city for a living.

# 4.CONCLUSIONS

In this study, we analyzed which is the best city for a living based on our own preferences. We have created two-way approach to evaluate each of three cities: London, Paris, and Amsterdam. The first approach takes into consideration the clusters which divide the city venues into different categories. The neighborhood which has the biggest number of clusters gets the highest score. The second approach takes into consideration manually elaborated venue categories rating. The sum of these two ratings help us first to identify the best neighborhood and then to select the city with the highest rating of a neighborhood.
This model can be very useful in helping young professionals who are in seek of relocation opportunities. Slight modifications of venue categories ranking could make the model applied to other people with different preferences and taking into account neighborhoods' coordinates of other cities could make globally adjust the approach.

# 5.FUTURE DIRECTIONS

We were able to take into consideration only 3-5 neighborhoods the most appropriate for our case. However, the analysis would be more comprehensive if more neighborhoods could be taken into consideration. We could also make our metrics more complicated by adding costs of commute, rent, house prices, age of buildings, GDP, local schools' rating, climate characteristics etc.
We could also complicate the way of calculation the resulted city ranking by not only taking into consideration the best neighborhood's ranking but also rankings of other neighborhoods. It might be a large dispersion in neighborhood rankings for one particular city which might also be important for evaluation.
To visualize the results, one could get json files with city neighborhoods' coordinates and use choropleth library to add some layers. We could also expend the approach to other cities.
These interactions data are obviously more difficult to extract and quantify, but if added, could bring significant improvements to the model.