Airbnb Recommender System

Salvador Robles

February 2021

INTRODUCTION

Why?

Covid-19 has revolutionized the way we work, giving many people the option of working remotely 100% of time.

As a result of this, people can choose to live in a different city and country, without experiencing any negative impact in their job.

Airbnb is a digital platform with more than 2 million properties, distributed across 192 countries and 33,000 cities.

This prototype will help the user to identify cities and residences that are similar to their own, facilitating a home swap counterpart.

Audience

Any individual who may be interested in swapping homes with someone living in another city and would like to find hosts living in similar cities and residences as theirs.

Current Airbnb hosts as well as potential Airbnb hosts will benefit from this recommender, as they will be paired with users based on similarities between their cities and residences.

Data Source

For the Airbnb residences, we have worked with Airbnb-listings dataset, which provides detailed information of 500k residences from the Airbnb platform. We have focused on 15 cities from Europe.

For metrics about cities, we have worked with data from Eurostat (perception surveys by locals of their cities) and Numbeo (certain city indexes such as cost of living and purchasing power of individuals).

DATA

Original Dataset

Information about residences and hosts

From https://public.opendatasoft.com/explore/dataset/airbnb-listings we downloaded information of 500k **residences** in the world, with 89 features (columns) per residence, such as:

- *Type of property*
- *Type of room*
- *Type of bed*
- *Number of beds*
- *Price*
- *Number of reviews*
- *Rating given by users*

Information about city features – perception surveys

From https://ec.europa.eu/eurostat/web/cities/data/database we downloaded information of **perception surveys by locals**. Perception surveys are obtained by city and ask individuals to state *how satisfied or dissatisfied they are with certain aspects of their city*, such *as:*

- *City amenities* (e.g., green places, sport amenities, etc.)
- *City pollution* (e.g., air quality, noise levels, etc.)
- *City hostility* to certain groups (e.g., gays and lesbian, elderly, etc.)

Information about city features – indexes and metrics

From https://www.numbeo.com/ we downloaded information of *several city indexes and metrics* that are calculated internally in a consistent way, which allows to compare different cities against each other. Some of the metrics Numbeo provides are:

- *Cost of living (compares cost of living across cities)*
- *Local purchasing power (compares purchasing power of average salary across different cities)*
- *Groceries index (compares prices of groceries across cities)*
- *Restaurants Index (compares prices of restaurants across cities)*

As a result of our research work and data wrangling steps performed, we generated a cleaned dataset of around 237k residences in Europe, distributed across 15 cities.

METHOD

Find similar cities

There are 15 European cities available in the recommender system, distributed across 13 countries.

To find similar cities to the user's city, we leveraged on perception surveys scores, such as *city amenities and pollution levels*, as well as economic metrics, such as *cost of living and purchasing power*.

We applied **k-means clustering algorithm** to identify the user's most similar cities, considered their selected city features.

Find similar homes

To find user's most similar residences in other cities, we leveraged on several residences' features, such as *rent price, type of property, type of room, number of bedrooms or reviews* by previous guests.

We applied **k-prototypes clustering algorithm**, as it can handle both categorical and numerical values, assigning similarity scores to residences based on the number of matching features between users (for categorical features) as well as distance between points (for numerical features).

DATA WRANGLING

City Features Data

- **Eurostat city perception polls not standardized:**
  - **Problem:** answers to polls were "very satisfied-rather satisfied-rather unsatisfied-very unsatisfied" together with the % of voters per category, we needed some sort of score per feature.
  - **Solution:** score was calculated (see documents), then scores were scaled on a range 0-1.
- **Indexes and metrics not standardized:**
  - **Problem:** Numbeo's indexes had base of (100 = New York City), and our cities were in EU, also the values were too high in comparison with the scaled 0-1 perception poll scores, which is an issue for k-means algorithm clustering.
  - **Solution:** we re-based the indexes (100= average of our 15 EU cities) and scaled them to 0-1 range.

Residences Data Issues

- **Names of cities duplicated or in different languages**
  - **Problem:** some cities were named in different ways or languages
  - **Solution:** selected per country the top 15-30 cities with more residences and converged duplicated values, the rest of cities were ignored for having marginal number of residences.
- **Some important residence features missing**
  - **Problem 1:** residences had number of bathrooms, beds and bedrooms in some cases.
  - **Solution 1:** imputed the missing rows based on several consistent criteria, leveraging on the number of people allowed in the house and type of room.

  - **Problem 2:** weekly and monthly prices were missing, there were several currencies in the dataset too.
  - **Solution 2:** weekly and monthly prices waere calculated based on daily price when missing. Also, we converted all currencies to euros based on current exchange rates.

  - **Problem 3:** Host information was stacked in a column and split by commas, containing interesting categorical data in wrong format.
  - **Solution 3:** expanded columns and reframed of host characteristics, in one column per characteristic and binary 1 or 0 for values.

  - **Problem 4:** Many houses had missing rating scores
  - **Solution 4:** Filled ratings with zero when number of reviews was zero or missing, this is a correct approach because hosts with no residences will later be clustered together, which is reasonable.

Conclusion

As a result of the steps performed in data wrangling, we were able to keep around **84% of the Europe residences initially identified.**

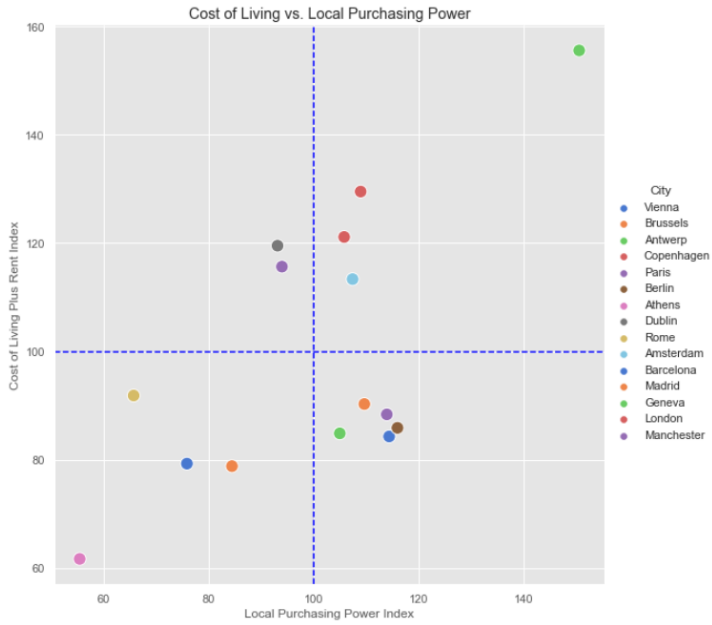
EDA – City Features

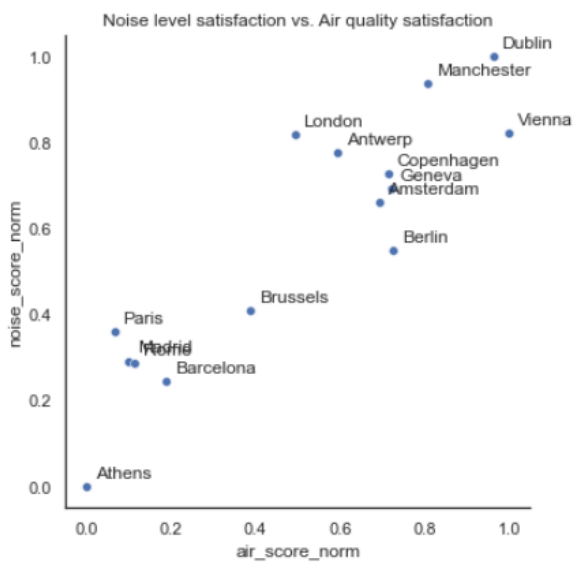
Are there significant differences between cities in terms of cost of living and purchasing power; or in how satisfied are locals with their cities?

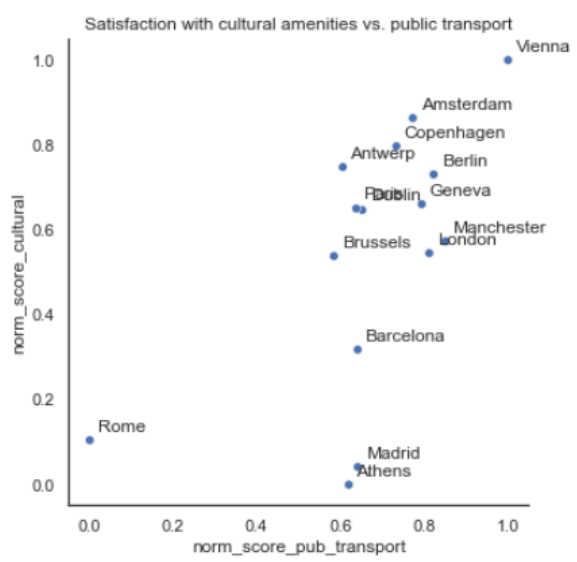
In relation with cost of living and purchase power we observed that there were significant differences between cities.

A **cost-of-living** value of 120 means than the city's cost of living is 20% higher than the average.

A **purchase power** of 120 means that with the average salary of the city, residents there could buy 20% more goods and services in the average city than where they live.







In relation with other city features, we also observed significant differences across cities.

For example, top left figure shows the satisfaction of locals with the **noise levels and air quality** of their city, values closer to 1 indicate higher satisfaction.

The bottom left figure shows the satisfaction of locals with the **cultural amenities** (museums, theaters) and **public transport** of their city.

In summary, it was concluded that **from different perspectives** (economic, environmental, structural and socio-cultural), **cities were significantly different** between each other.

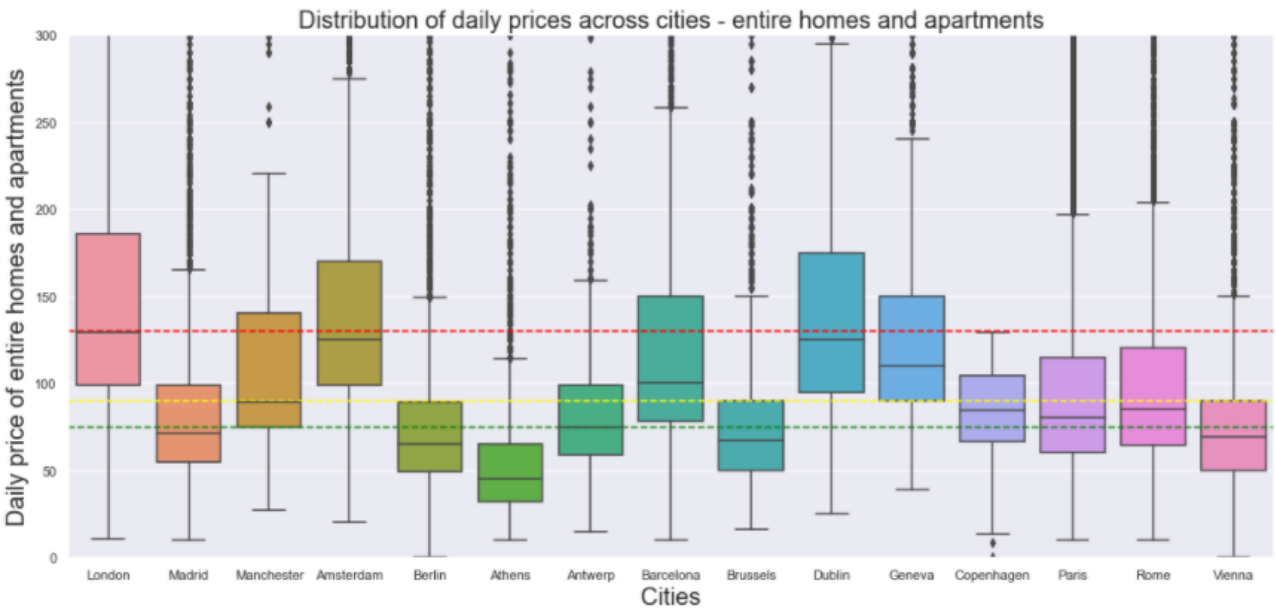
Notwithstanding of this, there is enough visual evidence to state that there are some clusters in the sample.

EDA – Home Features

What are the most frequent types of residences offered? Are residences' prices very different across cities? Are rating similar across cities?

In relation with what are the most frequent type of residences offered, we identified that **entire homes and apartments** as well as **private rooms, both with 1 or 2 beds,** were the most typical residences across the dataset.

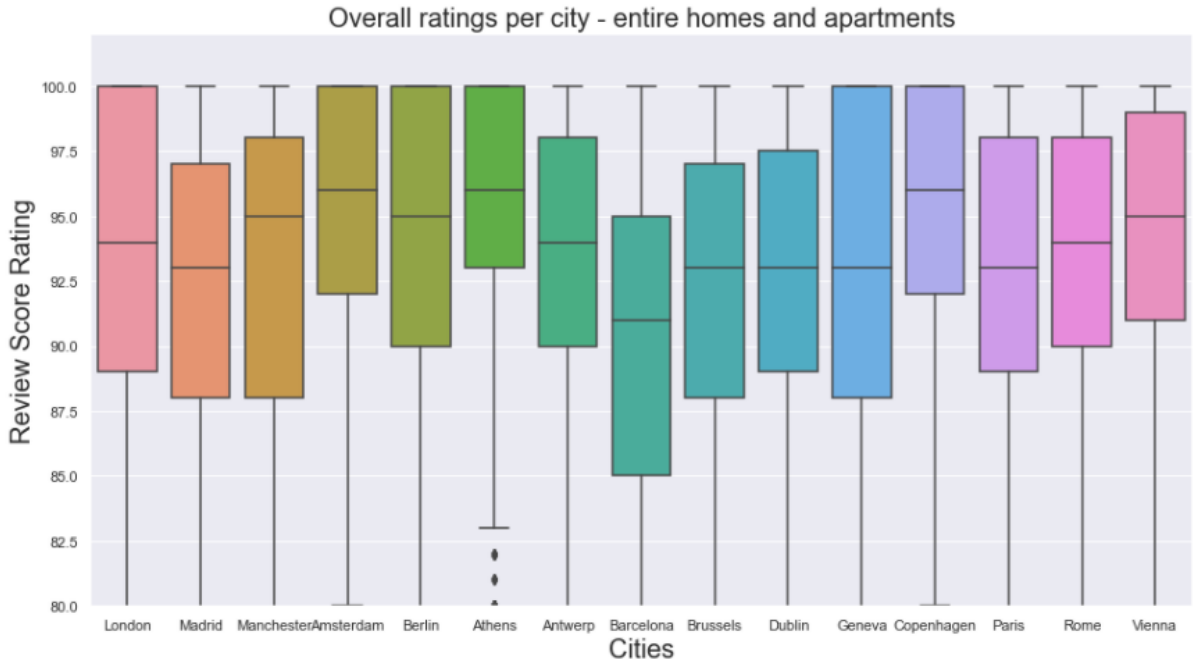
In relation with prices across cities, we observed **significant price differences** between cities. The figure below shows daily price distribution across cities.



The green, yellow and red lines could be a preliminary visual indicator of up to 3 possible clusters, such as low rent, mid rent and high rent cities.

Are ratings similar between cities?

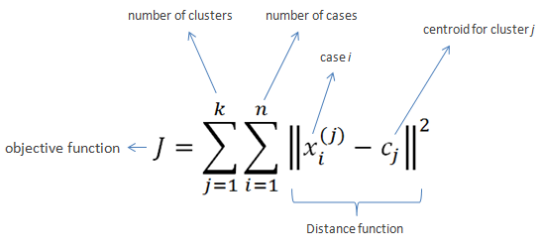
In relation with **rating scores**, we observed that they were **not significantly different** from city to city. Most residences are rated very high, with scores above 90-95% or even 100% for most of the residences. Barcelona was the exception of this, but even this city, being the lowest rated, has rating scores for 75% of houses above 85%.

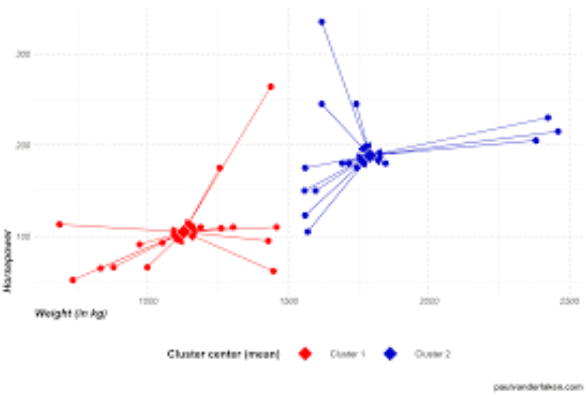


ALGORITMS & ML

To find similar cities to the user's city, we relied on k-means algorithm.

The k-means algorithm scales well with large datasets (for future project upgrades) and guarantees convergence of classes. To apply k-means, it is necessary to define the number of clusters "k" to generate. This was important for us as we wanted to keep control of the number of clusters created and how many cities existed in each cluster).





k-means struggles with categorical variables. Therefore, we preprocessed inputs to be in a scale of (0,1) for city features.

Also, centroids can be dragged by outliers and a cluster containing only the outlier may be created. This is something we faced, as Geneva was an outlier, in these cases, less clusters would be created to avoid isolation at the expense of accuracy.

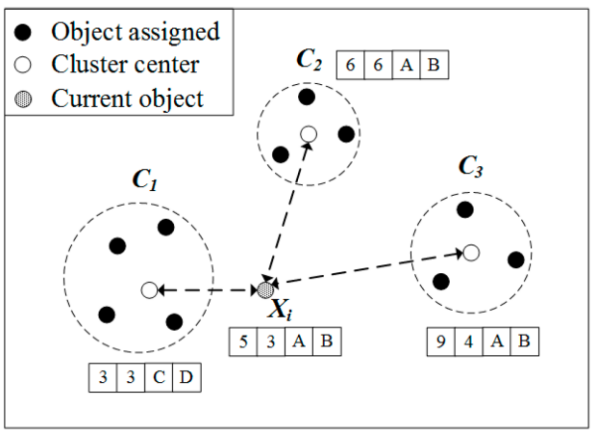
To find similar residences as the users' residence, we relied in K-prototypes algorithm

K-prototypes handles well both numeric variables (e.g., Price, n° of beds) - through k-means - and categorical variables (e.g., Property type, Room Type) -through k-modes. The rationale is that it computes the differences between numeric datapoints while computing similarity scores for categorical datapoints.

As with k-means, it is necessary to define the k-number of clusters. We used the cost function minimization for the purposes of choosing the optimal k-number of clusters.

$$E = \sum_{l=1}^{k}(E_l^r + E_l^c) = \sum_{l=1}^{k} E_l^r + \sum_{l=1}^{k} E_l^c = E^r + E^c$$

Minimize the total cost "E" which is the sum of the distances to the numeric and categorical parts of the centroid (prototype)



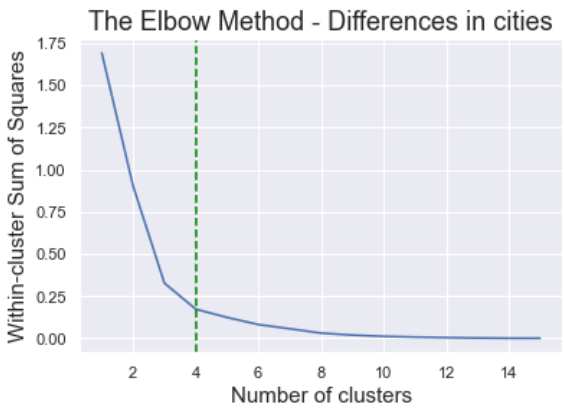
K-prototypes is a solid algorithm for problems with mixed datatypes as it minimizes distances for numerical data and matching dissimilarity function for categorical data.

ACCURACY

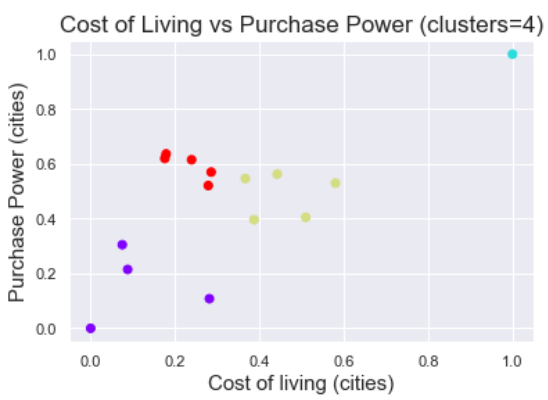
To measure quality of clusters, we applied a classifier on the clusters obtained, and calculated the F1 score as an indicator of the singularity of the groups generated.

The trade-off between number of clusters and their interpretation

To select the optimal k number of clusters, we computed wcss for k-means or the cost function for k-prototypes, that measure differences between data points and their cluster centroids.



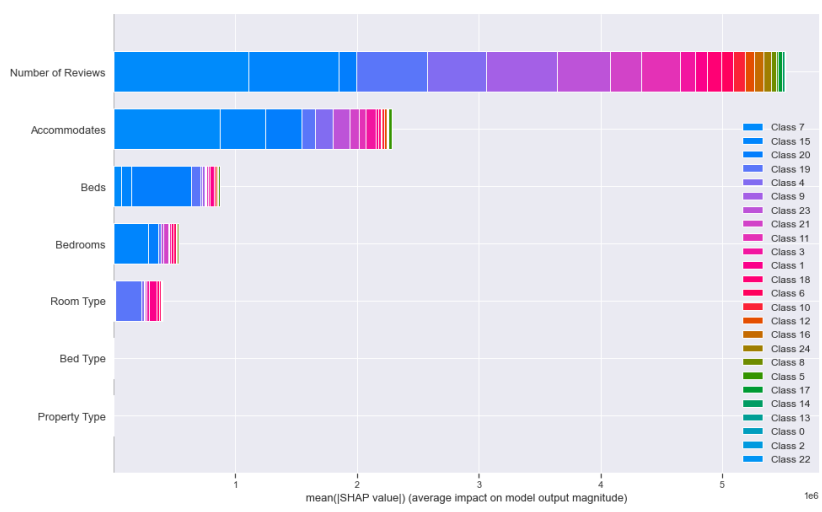
Fewer number of clusters generate bigger clusters, with larger differences between data points. More number of clusters generate smaller clusters, with lower differences between data points (but losing interpretability of clusters in the process).



The best number of clusters is subjective and must be assessed individually for every problem.

Predicting our clusters based on classifier

To test the clusters quality, we decided to follow the approach of fitting a classifier to the residence features and calculating the F1 score (0,44). This result was not very promising, but for the prototype we relied only on a subset of the available residences' features, while setting a high value of 25 for the number of clusters.



In relation with most important features to cluster, we identified through SHAP values that *number of reviews* and *accommodates* were very useful for the model to cluster. However, *Bed type* and *Property type* were not even used by the model to generate the clusters.

In conclusion, trying out different k-number of clusters together with utilizing different and/or more residences' features will be necessary to properly assess the quality of the clusters generated.

OUTPUT

City	Name	Description	Room Type	total_price_week
London	Lovely & warm flat 12 mins from central London	Clapham North has amazing restaurants, bars and open space in the form if Clapham Common. The tube is about 60 seconds walk away. That takes you to central London in 10 minutes. Flat is neat and spacious with outside space, wifi, tv etc Anywhere Can offer travel and tourist tips Great restaurants, bars and an amazing park nearby. Super quick to central London! Clapham North tube station is 1 minute away and Clapham High Street overground station is 2 minutes away	Entire home/apt	500.0
Paris	Joli trois-pièces avec vue	Ce trois-pièces, lumineux, calme et spacieux vous accueille tout près de la Place d'Italie et vous offre une magnifique sur tout Paris. A quelques minutes des métros Olympiades, Nationale, Place d'Italie. Nombreux commerces de qualité à proximité. Grand trois-pièces de 70m2, situé au 15è étage d'une tour particulièrement calme du 13ème arrondissement, cet appartement comporte une grande entrée, une cuisine/pièce-à-vivre, un salon, une chambre parentale (grand lit de 180cm composé de deux lits/sommiers de 90cm, suffisamment de place pour installer un lit bébé), une chambre pour un ou deux enfants, une salle de bain, des toilettes séparées. Notre quartier, situé entre la Place d'Italie, la Butte aux Cailles et le quartier des Gobelins, est à la fois populaire, calme et vivant. En quelques minutes de marche, vous aurez accès à de très bons commerces de bouche (excellents boulangers, fabuleux caviste et fromager par exemple) à de délicieux restaurants asiatiques, italiens, grecques ou tou	Entire home/apt	500.0
Paris	Appartement Paris Centre (9ème)	Bel appartement du centre ville parisien (place Clichy) situé au 2e étage avec ascenseur. Haut de plafond, l'appartement est un F2-3 situé dans un quartier charismatique : Grands boulevards, Montmartre et Pigalle sont accessibles à pied. L'appartement est un F2-3 : il dispose d'une chambre indépendante (lit double), d'une autre chambre (lit double) en mezzanine donnant sur le salon, lui-même ouvert sur la cuisine avec bar. WC et salle de bain séparés. Wifi, linge de lit et nécessaire de cuisine sont inclus (plaques à induction, micro-ondes, four, réfrigérateur...). Laverie automatique située à une rue. Toutes les pièces seront accessibles aux voyageurs. Nous ne serons pas sur place pendant la période où nous souhaitons mettre à disposition notre appartement, nous serons alors joignables par téléphone en cas de besoin. Pouvons nous déplacer si problème. Quartier dynamique et riche, la place de Clichy est un lieu très accessible en tout point : à pied vous vous situez à 15 minutes des G	Entire home/apt	500.0
Paris	Quiet & Bright flat	Well located in Paris intra-muros (2 min' from Olympiades Line 14 - 8' minutes from Place d'Italie, Lines 5, 6, 7). The 74 m² flat is in a 70's building, at the 26th floor (crazy view!). It has recently been refurbished.	Entire home/apt	500.0
London	Lovely 2-floor 2 bedroom garden flat in Islington	Split level basement flat in large Victorian town house, with excellent transport links. The flat consists of large double bedroom leading out to a garden area, second single bedroom, separate WC and shower, open plan living/dining/kitchen area. This is a great space with two large bedrooms. The living area works well for dining, relaxing and entertaining. There is also a decent size patio/garden area, great for eating out when it gets a bit warmer. Some things to consider when making your booking. A couple of guests have said that the ceilings in the lower floor are a bit lower than the pictures, but also that they are perfectly reasonable. Could get a bit tricky if you are 6'5 (1.95m) plus maybe. Some guests have also pointed out that they could hear noise from the house above from time to time, as they floorboards aren't soundproof. Again none have said that this necessarily impacted their stay. Also please note that the shower and toilet are separate - have a look at the p	Entire home/apt	500.0

City	Name	Listing Url	Host URL
London	Lovely & warm flat 12 mins from central London	https://www.airbnb.com/rooms/11570020	https://www.airbnb.com/users/show/27348595
Paris	Joli trois-pièces avec vue	https://www.airbnb.com/rooms/2504482	https://www.airbnb.com/users/show/7035807
Paris	Appartement Paris Centre (9ème)	https://www.airbnb.com/rooms/6987332	https://www.airbnb.com/users/show/23116772
Paris	Quiet & Bright flat	https://www.airbnb.com/rooms/5555235	https://www.airbnb.com/users/show/24411952
London	Lovely 2-floor 2 bedroom garden flat in Islington	https://www.airbnb.com/rooms/4927689	https://www.airbnb.com/users/show/4891639

FUTURE IMPROVEMENTS

- Add more residences to the dataset from same or new cities.
- Add more residences' features to the model such as description or neighborhood.
- Add more cities' features to the model such as population number and age-groups
- Automate in some way the selection of k-best number of clusters.
- Give more freedom of choice to the user in terms of what residence features to consider