# Search for accommodation in Madrid

## (Capstone Project)

by César Portales Alonso

# Table of contents

Table of contents........ 2

1. Introduction ........ 3

1.1 Background........ 3

1.2 Problem ........ 3

1.3 Interest ........ 3

2. Data acquisition and cleaning ........ 4

2.1 Data Sources........ 4

3. Methodology ........ 4

3.1 Tools used ........ 4

3.2 Data cleaning........ 5

3.3 Exploratory Data Analysis........ 6

Numerical variables........ 6

**Categorical data** ........ 9

3.4 Clustering ........ 13

**Modeling** ........ 13

3.5 Explore Neighborhoods in Madrid ........ 14

4. Results ........ 15

5. Discussion........ 16

6. Conclusion ........ 17

# 1. Introduction

## 1.1 Background

Madrid Familiar Tours (MFT) is a company dedicated to organizing guided tours of the main monuments of Madrid and its surroundings. Visits last (depending on the package you hire) of between three and five days.

MFT, as the name suggests, has a clear family vocation and offers a wide range of discounts and packages for members of the same family.

The company organizes the visits, manages the guides and accesses the main monuments, but does not provide the necessary accommodations. Therefore, to provide facilities to its customers, it has requested us to provide a list of possible accommodations with certain characteristics to make your stay as easy and pleasant as possible.

## 1.2 Problem

The accommodations to be suggested must meet a number of requirements:

- The visits will focus mainly on what is known as the "Walk of Museums", which is home to some of the world's leading museums, such as the Prado Museum, the Thyssen-Bornemisza museum and the Reina Sofía modern art center. Some packages include trips to cities near Madrid (such as Toledo or Aranjuez) with quick access from the train, which can be taken at Atocha station (which is located on the same walk as the museums). Therefore, the accommodation is required to be a **maximum distance of one and a half kilometers** from this location, which will facilitate quick access even on foot.
- Since most of MFT's clients are family units, it is intended that the accommodations are **complete apartments.**
- Of course, it will be worth providing information about the **price**, and it would be desirable to know if the accommodations have a **kitchen**, are prepared for **children**, **allow smoking**, if they have **access for the disabled**, if they have **access to the internet**, if there is **temperature control** (heating / air conditioning), or even if they **allow pets**
- It would be welcome to provide information on the main points of interest (**venues**)near the selected properties.

## 1.3 Interest

Obviously, this information is of interest to any of the potential MFT customers: individuals / families interested in knowing the main monuments of Madrid and immerse themselves in its culture and gastronomy.

# 2. Data acquisition and cleaning

## 2.1 Data Sources

For this purpose, we will use a public Airbnb file in csv format (airbnb-extract-Madrid.csv), which contains information about accommodation in Madrid.


airbnb-extract-Madrid.csv: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
ID;Neighbourhood;Property Type;Room Type;Amenities;Price;Latitude;Longitude;URL;NAME
6478626;Retiro;Apartment;Private room;TV,Wireless Internet,Kitchen,Heating,Essentials,Hair dryer,Iron;28;40.40695285;-
3.670894491;https://www.airbnb.com/rooms/6478626;Habitaci¢n con ba¤o .
4941335;Salamanca;Apartment;Entire home/apt;TV,Internet,Wireless Internet,Air conditioning,Kitchen,Elevator in
building,Heating,Family/kid friendly,Washer,First aid kit,Essentials,Shampoo,translation missing:
en.hosting_amenity_49;58;40.42590392;-3.68164746;https://www.airbnb.com/rooms/4941335;C¢modo estudio en la calle Goya
12510355;Salamanca;Apartment;Entire home/apt;TV,Internet,Wireless Internet,Air conditioning,Kitchen,Smoking
allowed,Elevator in building,Buzzer/wireless intercom,Heating,Family/kid friendly,Washer,Essentials,Shampoo,24-hour
check-in,Hangers,Hair dryer,Laptop friendly workspace;70;40.42835189;-
3.687154663;https://www.airbnb.com/rooms/12510355;Studio & Terrace - Barrio Salamanca
4756811;Salamanca;Apartment;Private room;Internet,Wireless Internet,Kitchen,Doorman,Elevator in building,Buzzer/wireless
intercom,Heating,Washer,Fire extinguisher,translation missing: en.hosting_amenity_49,translation missing:
en.hosting_amenity_50;26;40.42745443;-3.685763023;https://www.airbnb.com/rooms/4756811;Cute single room -private bathroom
15332216;Salamanca;Apartment;Entire home/apt;TV,Internet,Wireless Internet,Air conditioning,Kitchen,Doorman,Elevator in
building,Buzzer/wireless intercom,Heating,Family/kid friendly,Washer,Smoke detector,First aid kit,Safety card,Fire
extinguisher,Essentials,Shampoo,24-hour check-in,Hangers,Iron,Laptop friendly workspace;70;40.42867086;-
3.685422757;https://www.airbnb.com/rooms/15332216;Deluxe Apartment in Barrio Salamanca (next to IE)
15825062;Salamanca;Other;Entire home/apt;Internet,Wireless Internet,Air conditioning,Kitchen,Breakfast,Heating,Family/kid
friendly,Washer,Dryer,Smoke detector,Carbon monoxide detector,Essentials,Shampoo,24-hour check-in,Hangers,Hair
dryer,Iron,Laptop friendly workspace;108;40.42177819;-3.69106518;https://www.airbnb.com/rooms/15825062;BEST Location for
DESIGN lovers
3116679;Salamanca;Apartment;Entire home/apt;TV,Internet,Wireless Internet,Air conditioning,Kitchen,Doorman,Elevator in
```

The Dataset contains the following fields: Property ID, Neighborhood, Property Type, Room Type, Amenities, Latitude, Longitude, Property URL, and Property Name.

As far as points of interest near the accommodation are concerned, we will need to use the **Foursquare location** data.

# 3. Methodology

## 3.1 Tools used

The main tool to use is a Jupyter notebook, in which they will be used as Python and HTML programming languages.

The libraries that will need to be used to perform this task are:

- To perform mathematical calculations, and manage arrays and datasets:

```python
# mathematical functions, management of arrays and datasets
import math
import numpy as np

from scipy import stats
from scipy.stats import norm
import pandas as pd

import warnings
warnings.filterwarnings('ignore')
```

- To generate statistical graphs

```python
#graphics management functions

import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
%matplotlib inline

cm = plt.cm.RdBu
cm_bright = ListedColormap(['#FF0000', '#0000FF'])

import seaborn as sns
```

- To preprocessing and clustering

```python
# preprocessing and clustering functions
#from sklearn.preprocessing import LabelBinarizer
from sklearn.preprocessing import MinMaxScaler # to normailize data
from sklearn.cluster import KMeans # clustering
# NLP for extract Amenities
from sklearn.feature_extraction.text import CountVectorizer
import string
```

- To geolocate, access Foursquare data and draw maps

```python
# geolocation functions
!pip install geocoder
import geocoder # import geocoder
!pip install geopy
from geopy.geocoders import Nominatim
from geopy.distance import geodesic # To calculate distances to the Prado museum
from geopy.extra.rate_limiter import RateLimiter #necessary to pause between each search
import requests # library to handle Foursquare requests

# graph the maps
!pip install folium
import folium
geolocator = Nominatim(user_agent="mad_explorer")
```

## 3.2 Data cleaning

To accommodate our client's requests, we will select only the apartment information that is rented in full

We select only complete rented apartments

```python
print("Dataset size:", df_Completo.shape)
df0 = df_Completo.drop(df_Completo[df_Completo['Property Type'] != "Apartment"].index)
df = df0.drop(df0[df0['Room Type'] != "Entire home/apt"].index)

df.reset_index(drop=True, inplace=True)
print("Dimensions of the dataset once the data for Madrid are selected are:", df.shape)
```

```
Dataset size: (13251, 10)
Dimensions of the dataset once the data for Madrid are selected are: (7013, 10)
```

By geolocation and from latitude and longitude, the distance to the Prado museum is calculated in the "Dis_Museos" field.

| ID | Neighbourhood | Property Type | Room Type | Amenities | Price | Latitude | Longitude | URL | NAME | Dist_Museos |
|---|---|---|---|---|---|---|---|---|---|---|
| 4941335 | Salamanca | Apartment | Entire home/apt | TV,Internet,Wireless Internet,Air conditioning,... | 58.0 | 40.425604 | -3.681647 | https://www.airbnb.com/rooms/4941335 | Cómodo estudio en la calle Goya | 1.506046 |
| 12510355 | Salamanca | Apartment | Entire home/apt | TV,Internet,Wireless Internet,Air conditioning | 70.0 | 40.428352 | -3.687155 | https://www.airbnb.com/rooms/12510355 | Studio & Terrace - Barrio Salamanca | 1.567554 |
| 15332216 | Salamanca | Apartment | Entire home/apt | TV,Internet,Wireless Internet,Air conditioning,... | 70.0 | 40.428671 | -3.685423 | https://www.airbnb.com/rooms/15332216 | Deluxe Apartment in Barrio Salamanca (next to IE) | 1.743295 |
| 3116679 | Salamanca | Apartment | Entire home/apt | TV,Internet,Wireless Internet,Air conditioning,... | 260.0 | 40.425606 | -3.683705 | https://www.airbnb.com/rooms/3116679 | Elegant & central luxury 3 bedroom apartment | 1.458292 |
| 3962279 | Salamanca | Apartment | Entire home/apt | TV,Internet,Wireless Internet,Air conditioning | 60.0 | 40.426550 | -3.676096 | https://www.airbnb.com/rooms/3962279 | Beautiful apartment in the center | 1.356402 |

And accommodation that is more than a mile away is eliminated.

```python
print(df.shape)
df = df.drop(df[df['Dist_Museos']>=1.5].index)
df = df.reset_index(drop=True)
df.shape
```

```
(7013, 11)

(3404, 11)
```

## 3.3 Exploratory Data Analysis

As a preliminary step, the presence of null values is studied. It is checked that there is a priceless accommodation. Since the data is essential, and it is only a record, it is carried out to remove it.

```python
df.isnull().sum()
```

```
Neighbourhood    0
Amenities        8
Price            1
Latitude         0
Longitude        0
URL              0
NAME             0
Dist_Museos      0
dtype: int64
```

The eight records without Amenities are maintained. We can't tell if the data is missing or if the property simply doesn't provide it, so the latter option is assumed instead of deleting the records.

### Numerical variables

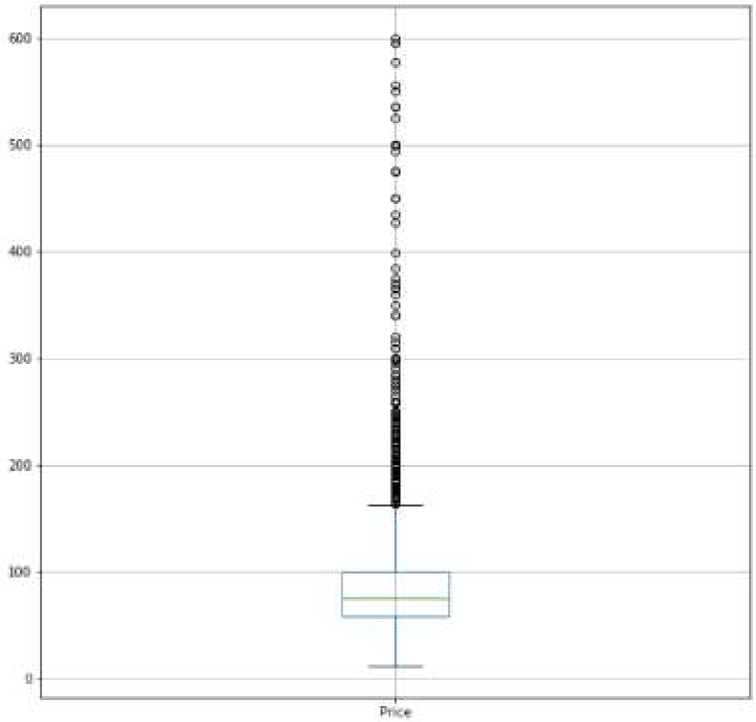
When it comes to numerical variables, we will focus mainly on the price.

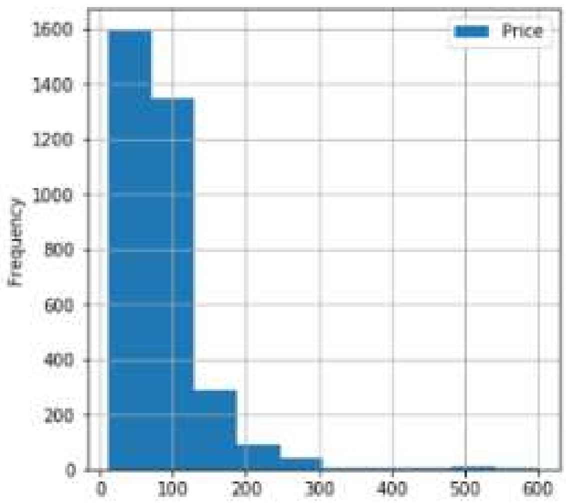
First, it checks for outliers

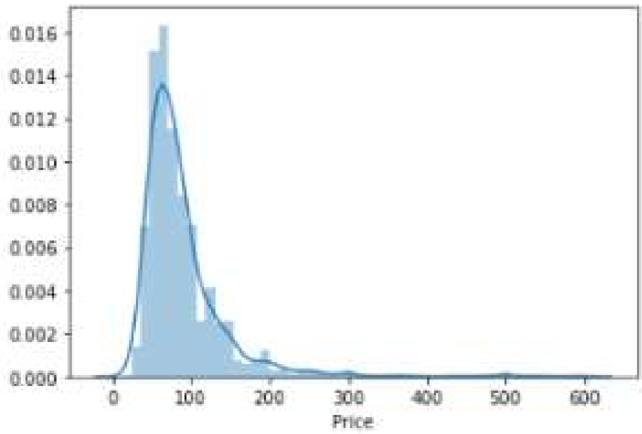
```
print(df['Price'].describe())
df.plot(kind='box',y='Price',grid=True,figsize=(10 ,10))

plt.show()
```

```
count    3403.000000
mean       89.532471
std        58.482222
min        12.000000
25%        58.000000
50%        75.000000
75%       100.000000
max       600.000000
Name: Price, dtype: float64
```

You can see that the distribution is not very balanced. It can be seen more easily graphically by boxplot or through a histogram.






At first glance it is seen that there is a standard deviation from the normal distribution, a positive symmetry and the presence of some peaks is observed.

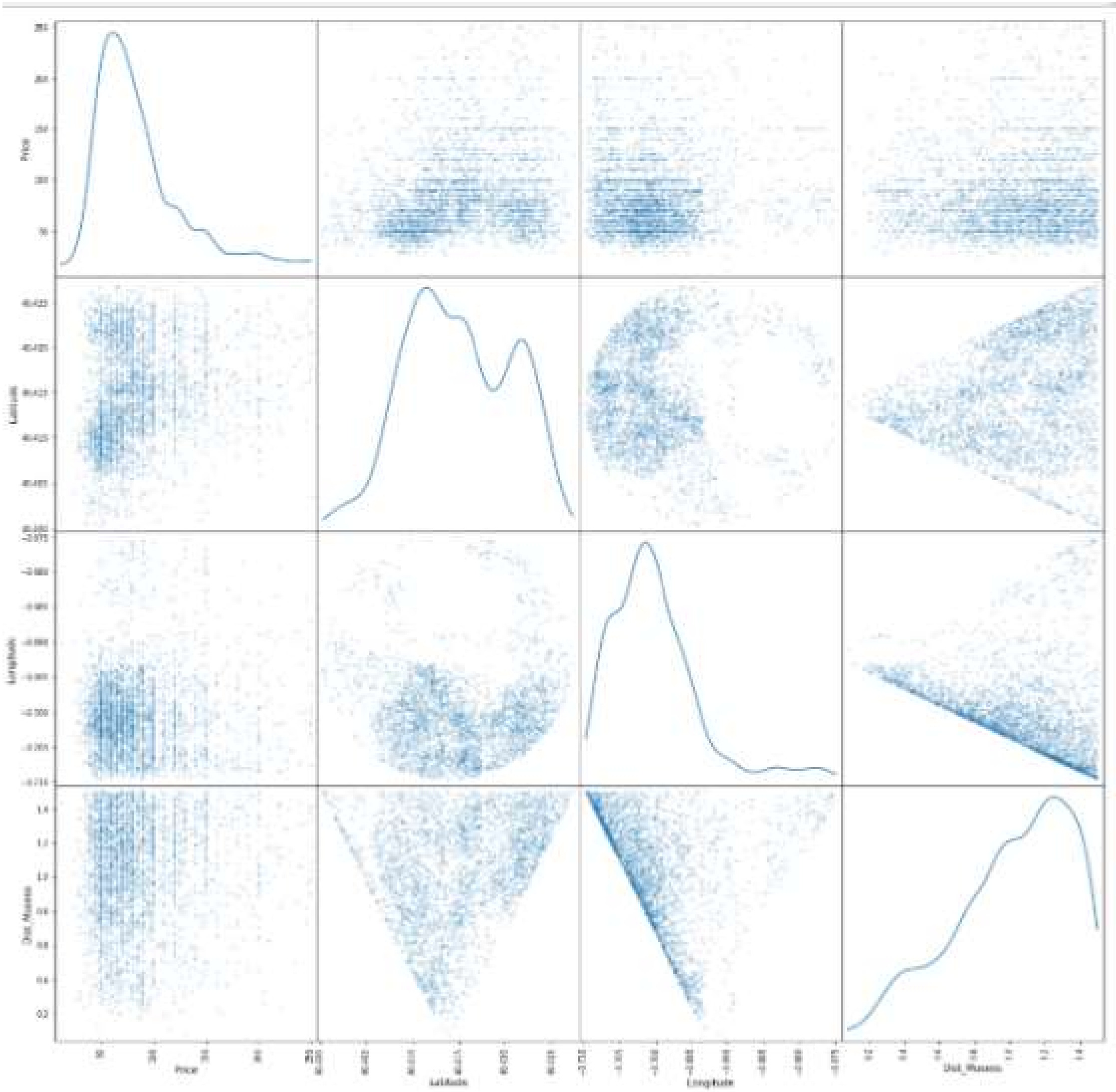
We see that outliers are considered approximately from 180. But it is seen that until approximately 250 there is a large group of values.

We check how many values are above 250 price and it is decided to work only with these values so that you don't have too distorted results.

**Study of correlations and distribution of data**

Correlations are then studied and the distribution of the data is checked. In case there is too correlated data, we would have to study whether any of the variables are removed from the problem when running our model, but you can verify that this is not the case and the correlation between the data is scarce, and the distribution of the data is adjusted to what you would expect.

| | Price |
|---|---|
| **Price** | 1 |
| **Latitude** | 0.194602 |
| **Longitude** | 0.0334467 |
| **Dist_Museos** | 0.0328302 |



## Categorical data

The ID of the accommodation, as well as the name and URL of the same is not of interest to the study that we are going to carry out (except its use as tags to display them on the map), but they are not necessary (rather the opposite) or to perform clustering or to obtain the venues.

That's why it's only worked with Neighbourhoods information and amenities.

**Working with Amenities**

The 'Amenities' field does have important information,

Amenities are saved separated by commas, and that we will study with basic NLP techniques.

| Neighbourhood | Amenities | URL |
|---:|---:|---:|
| Chamberí | Internet,Wireless Internet,Air conditioning,Ki... | https://www.airbnb.com/rooms/15261457 |
| Arganzuela | Internet,Wireless Internet,Kitchen,Elevator in... | https://www.airbnb.com/rooms/5257204 |
| Arganzuela | TV,Internet,Wireless Internet,Air conditioning... | https://www.airbnb.com/rooms/13892731 |
| Arganzuela | TV,Cable TV,Internet,Wireless Internet,Air con... | https://www.airbnb.com/rooms/1740331 |
| Centro | TV,Internet,Wireless Internet,Air conditioning... | https://www.airbnb.com/rooms/16399219 |

A very advanced study will not be required because, as we will see below, the information is very little varied (it looks like it has been selected from some type of menu). So in principle we will use sklearn's 'Feature_extraction', which is efficient and easy to use.

We will create a column in the dataset for each of the elements of the attribute, and then we will be left with only the columns that interest us, saving them with values of one or zero.

Through the library feature_extraction de sklearn , we load CountVectorizer, which will perform the entire data extraction process and create a new field in the dataset for every value it finds.

These are the values found

```
Columns generated from the values of 'Amenities'

['24HOUR CHECKIN' 'AIR CONDITIONING' 'BABY BATH'
 'BABYSITTER RECOMMENDATIONS' 'BATHTUB' 'BREAKFAST'
 'BUZZERWIRELESS INTERCOM' 'CABLE TV' 'CARBON MONOXIDE DETECTOR' 'CATS'
 'CHILDRENS BOOKS AND TOYS' 'CHILDRENS DINNERWARE' 'CRIB' 'DOGS' 'DOORMAN'
 'DOORMAN ENTRY' 'DRYER' 'ELEVATOR IN BUILDING' 'ESSENTIALS'
 'FAMILYKID FRIENDLY' 'FIRE EXTINGUISHER' 'FIRST AID KIT'
 'FREE PARKING ON PREMISES' 'FREE PARKING ON STREET' 'GAME CONSOLE' 'GYM'
 'HAIR DRYER' 'HANGERS' 'HEATING' 'HIGH CHAIR' 'HOT TUB'
 'INDOOR FIREPLACE' 'INTERNET' 'IRON' 'KEYPAD' 'KITCHEN'
 'LAPTOP FRIENDLY WORKSPACE' 'LOCK ON BEDROOM DOOR' 'LOCKBOX' 'OTHER PETS'
 'OUTLET COVERS' 'PACK N PLAYTRAVEL CRIB' 'PAID PARKING OFF PREMISES'
 'PETS ALLOWED' 'PETS LIVE ON THIS PROPERTY' 'POOL' 'PRIVATE ENTRANCE'
 'PRIVATE LIVING ROOM' 'ROOMDARKENING SHADES' 'SAFETY CARD' 'SELF CHECKIN'
 'SHAMPOO' 'SMARTLOCK' 'SMOKE DETECTOR' 'SMOKING ALLOWED' 'STAIR GATES'
 'SUITABLE FOR EVENTS' 'TABLE CORNER GUARDS'
 'TRANSLATION MISSING ENHOSTINGAMENITY49'
 'TRANSLATION MISSING ENHOSTINGAMENITY50' 'TV' 'WASHER' 'WASHER  DRYER'
 'WHEELCHAIR ACCESSIBLE' 'WINDOW GUARDS' 'WIRELESS INTERNET' 'Z']
```

And this is the Dataframe obtained

| | 24HOUR CHECKIN | AIR CONDITIONING | BABY BATH | BABYSITTER RECOMMENDATIONS | BATHTUB | BREAKFAST | BUZZERWIRELESS INTERCOM | CABLE TV | CARBON MONOXIDE DETECTOR | CATS | CHILDRENS BOOKS AND TOYS | CHILDREN DINNERWAR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | |
| 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 5 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | |
| 6 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | |
| 7 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | |
| 9 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |

We are going to eliminate the fields that the client has not asked for, and regarding what he has asked us, we will make combinations of values. For example, if we are interested in information about whether or not there is internet access, we will combine fields such as "Internet" and "Wireless Internet", or in the case of whether the home is child-friendly, we will combine information about whether it has games or books to children, if you have a high chair, etc.

*df_am['Pets'] =(df_am['CATS'].astype(bool)) | (df_am['DOGS'].astype(bool)) | (df_am['OTHER PETS'].astype(bool)) | (df_am['PETS ALLOWED'].astype(bool)) |(df_am['PETS LIVE ON THIS PROPERTY'].astype(bool))*

*df_am_to_drop =['CATS','DOGS','OTHER PETS','PETS ALLOWED','PETS LIVE ON THIS PROPERTY']*

*df_am['InternetAccess'] =(df_am['INTERNET'].astype(bool)) | (df_am['WIRELESS INTERNET'].astype(bool))*

Finalmente, nos quedamos solo con los datos que nos han solicitado:

| | ESSENTIALS | KITCHEN | SMOKING ALLOWED | WHEELCHAIR ACCESSIBLE | Pets | InternetAccess | TempControl | KidsFriendly |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| 2 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 3 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| 4 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |

**Working with Neighbourhoods**

If we analyze the distribution of values in this field, the result obtained is as follows:

```
Centro        3003
Arganzuela     131
Retiro         115
Salamanca       79
Chamberí         6
Name: Neighbourhood, dtype: int64
```

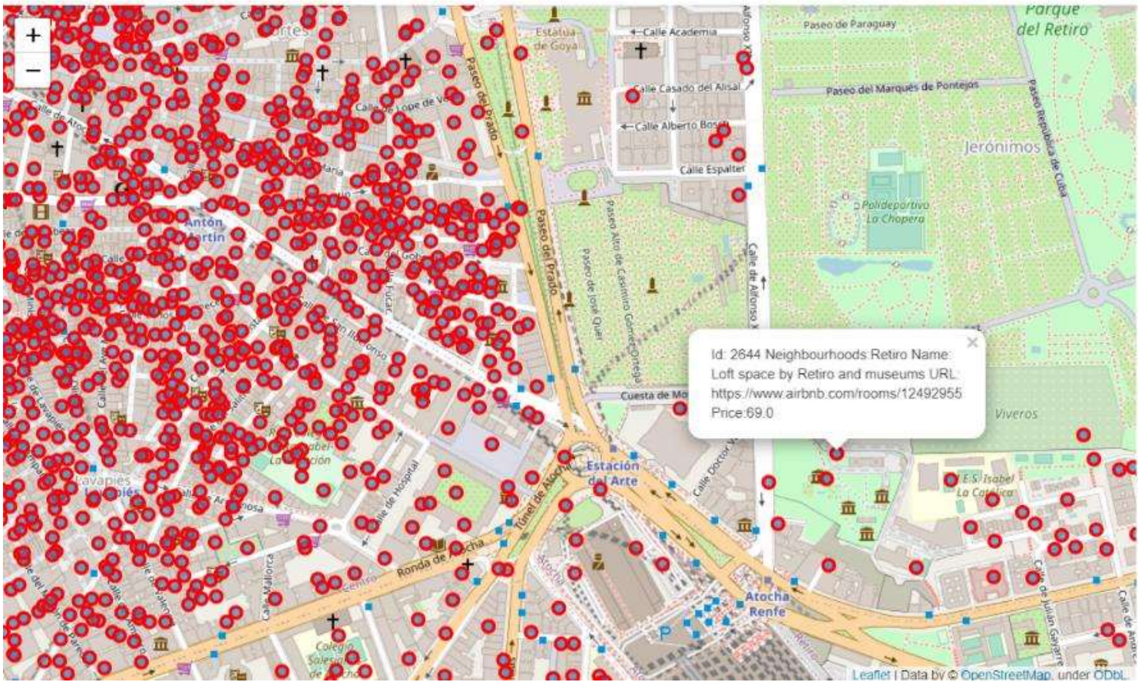
This leads us to believe that when we do clustering, the neighborhood will not influence much, as the vast majority of the accommodations will be in the Centro district.

Since the K-means algorithm does not work with categorical data, we will apply one-hot encoding to this field.

This is the result:

| | Arganzuela | Centro | Chamberí | Retiro | Salamanca |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 | 0 |
| 3 | 1 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 |
| 5 | 0 | 1 | 0 | 0 | 0 |
| 6 | 0 | 1 | 0 | 0 | 0 |
| 7 | 0 | 1 | 0 | 0 | 0 |
| 8 | 0 | 1 | 0 | 0 | 0 |
| 9 | 0 | 1 | 0 | 0 | 0 |
| 10 | 0 | 1 | 0 | 0 | 0 |
| 11 | 0 | 1 | 0 | 0 | 0 |
| 12 | 0 | 1 | 0 | 0 | 0 |

This is the map obtained from the accommodations near the Prado Museum and Atocha Train Station. By clicking on each of them we get information about the accommodation (Id, neighborhood, price and URL for more information).




To explore the points of interest of each neighborhood, we will use the library **Foursquare**. We will define Foursquare Credentials and Version and obtain ten venues for each property.

| | neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Arganzuela | Tapas Restaurant | Restaurant | Beer Garden | Bakery | Coffee Shop | Museum | Chinese Restaurant | Mediterranean Restaurant | Café | Market |
| 1 | Centro | Plaza | Hotel | Hostel | Tapas Restaurant | Gourmet Shop | Wine Bar | Ice Cream Shop | French Restaurant | Bistro | Restaurant |
| 2 | Chamberí | Tapas Restaurant | Spanish Restaurant | Bar | Restaurant | Café | Theater | Plaza | Bakery | Gastropub | Coffee Shop |
| 3 | Retiro | Spanish Restaurant | Bar | Grocery Store | Art Gallery | Tapas Restaurant | Supermarket | Burger Joint | Indian Restaurant | Museum | Mexican Restaurant |
| 4 | Salamanca | Spanish Restaurant | Restaurant | Bar | Hotel | Plaza | Burger Joint | Grocery Store | Seafood Restaurant | Gymnastics Gym | Bakery |

## 3.4 Clustering

### Modeling

We will train using a K-means algorithm to classify the information into five clusters

As a pre-step to do the analysis, the data is normalized using MinMaxScaler

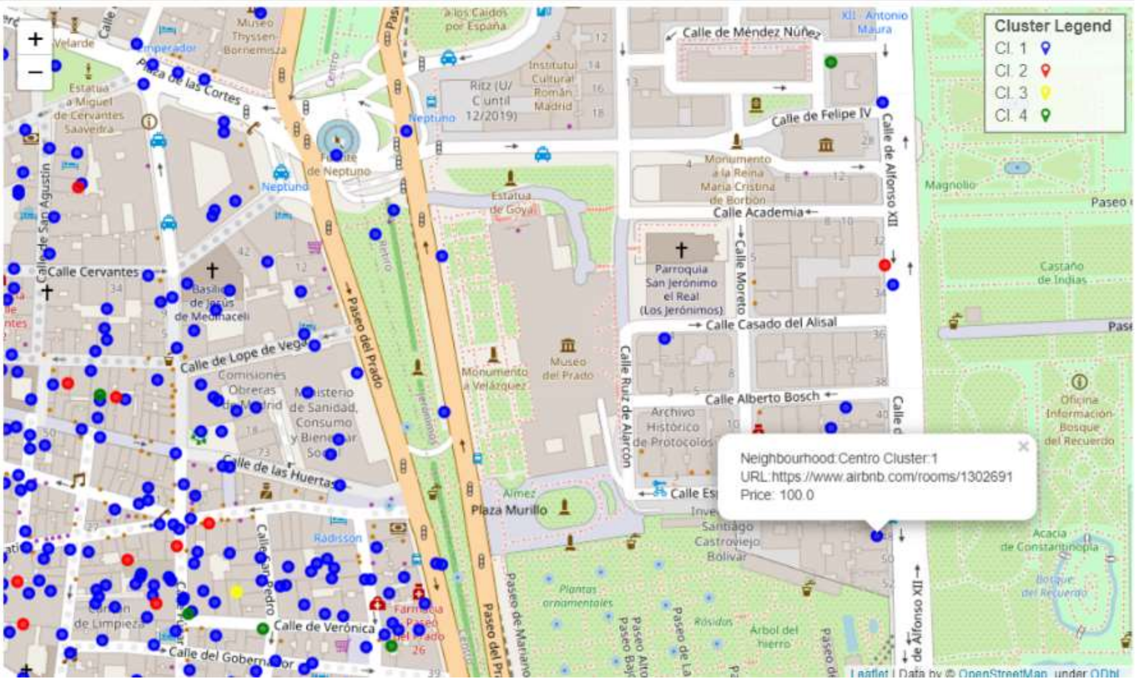
| | Price | Latitude | Longitude | Dist_Museos | Arganzuela | Centro | Chamberi | Retiro | Salamanca | ESSENTIALS | KITCHEN | SMOKING ALLOWED | WHEELCHAIR ACCESSIBLE | Pets |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.113445 | 0.959529 | 0.518384 | 0.912492 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 1 | 0.138655 | 0.220781 | 0.457714 | 0.538627 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 |
| 2 | 0.252101 | 0.108433 | 0.503794 | 0.764622 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.264706 | 0.056220 | 0.538937 | 0.876051 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 4 | 0.306723 | 0.637508 | 0.227703 | 0.596616 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 |
| 5 | 0.474790 | 0.592194 | 0.192215 | 0.629938 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 |
| 6 | 0.142857 | 0.623802 | 0.051529 | 0.927670 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 7 | 0.348739 | 0.594563 | 0.190304 | 0.635154 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 8 | 0.600840 | 0.631943 | 0.201229 | 0.641207 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 |
| 9 | 0.327731 | 0.596728 | 0.103793 | 0.808815 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 10 | 0.399160 | 0.673225 | 0.243201 | 0.606490 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 11 | 0.491597 | 0.587670 | 0.100870 | 0.810390 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 |
| 12 | 0.243697 | 0.653017 | 0.242082 | 0.585988 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 |
| 13 | 0.529412 | 0.573347 | 0.078844 | 0.849209 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 14 | 0.558824 | 0.637881 | 0.231934 | 0.589162 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 15 | 0.327731 | 0.726957 | 0.252025 | 0.661307 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 16 | 0.369748 | 0.517895 | 0.087334 | 0.818946 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 |
| 17 | 0.407563 | 0.624048 | 0.088940 | 0.853550 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 18 | 0.247899 | 0.570295 | 0.175848 | 0.651234 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 19 | 0.113445 | 0.629376 | 0.067720 | 0.898761 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |

The model is trained and the information is grouped into five clusters

| Labels | Price | Latitude | Longitude | Dist_Museos | Arganzuela | Centro | Chamberí | Retiro | Salamanca |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.304548 | 0.545882 | 0.269896 | 0.668313 | 0.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 |
| 1 | 0.293149 | 0.537250 | 0.260128 | 0.655259 | 0.000000 | 0.998239 | 0.001761 | 0.000000 | 0.000000 |
| 2 | 0.295151 | 0.560264 | 0.290779 | 0.705152 | 0.400631 | 0.000000 | 0.009464 | 0.356467 | 0.233438 |
| 3 | 0.289366 | 0.534205 | 0.263884 | 0.655839 | 0.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 |
| 4 | 0.306681 | 0.525110 | 0.271621 | 0.678320 | 0.000000 | 0.994118 | 0.001961 | 0.000000 | 0.003922 |

| ESSENTIALS | KITCHEN | SMOKING ALLOWED | WHEELCHAIR ACCESSIBLE | Pets | InternetAccess | TempControl | KidsFriendly |
|---|---|---|---|---|---|---|---|
| 0.926065 | 0.988095 | 0.000000 | 0.085213 | 0.000000 | 0.981203 | 0.979950 | 1.000000 |
| 0.889085 | 0.955986 | 0.123239 | 0.026408 | 0.000000 | 0.957746 | 0.940141 | 0.000000 |
| 0.917981 | 0.981073 | 0.138801 | 0.151420 | 0.151420 | 0.940063 | 0.984227 | 0.810726 |
| 0.945455 | 0.985455 | 0.000000 | 0.076364 | 1.000000 | 0.996364 | 0.978182 | 0.865455 |
| 0.917647 | 0.992157 | 1.000000 | 0.090196 | 0.521569 | 0.968627 | 0.949020 | 0.939216 |

We create a new map that adds information about the cluster to which each accommodation belongs. A different color is assigned to each cluster, and a legend is added to the chart (using HTML).




## 3.5 Explore Neighborhoods in Madrid

Next, we are going to start utilizing the Foursquare API to explore the neighborhoods and segment them.

A function is created to obtain points of interest within 500 meters of each host, and the number of venues that has been found in each district is counted

| neighbourhood | neighbourhood Latitude | neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| **Arganzuela** | 31 | 31 | 31 | 31 | 31 | 31 |
| **Centro** | 46 | 46 | 46 | 46 | 46 | 46 |
| **Chamberi** | 81 | 81 | 81 | 81 | 81 | 81 |
| **Retiro** | 39 | 39 | 39 | 39 | 39 | 39 |
| **Salamanca** | 73 | 73 | 73 | 73 | 73 | 73 |

Next, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category

| | neighbourhood | American Restaurant | Argentinian Restaurant | Art Gallery | Art Museum | Asian Restaurant | BBQ Joint | Bakery | Bar | Bed & Breakfast | Beer Bar | Beer Garden | Beer Store | Bistro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Arganzuela | 0.000000 | 0.000000 | 0.032258 | 0.000000 | 0.000000 | 0.000000 | 0.064516 | 0.032258 | 0.000000 | 0.000000 | 0.064516 | 0.000000 | 0.000000 |
| **1** | Centro | 0.000000 | 0.000000 | 0.000000 | 0.021739 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.021739 | 0.000000 | 0.000000 | 0.043478 |
| **2** | Chamberi | 0.012346 | 0.000000 | 0.000000 | 0.000000 | 0.012346 | 0.012346 | 0.037037 | 0.086420 | 0.000000 | 0.024691 | 0.000000 | 0.012346 | 0.000000 |
| **3** | Retiro | 0.000000 | 0.000000 | 0.051282 | 0.000000 | 0.000000 | 0.000000 | 0.025641 | 0.076923 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **4** | Salamanca | 0.000000 | 0.013699 | 0.000000 | 0.000000 | 0.013699 | 0.000000 | 0.027397 | 0.054795 | 0.013699 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |

Now let's create the new dataframe and display the top 10 venues for each neighborhood

| | neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Arganzuela | Tapas Restaurant | Restaurant | Beer Garden | Bakery | Coffee Shop | Museum | Chinese Restaurant | Mediterranean Restaurant | Café | Market |
| 1 | Centro | Plaza | Hotel | Hostel | Tapas Restaurant | Gourmet Shop | Wine Bar | Ice Cream Shop | French Restaurant | Bistro | Restaurant |
| 2 | Chamberí | Tapas Restaurant | Spanish Restaurant | Bar | Restaurant | Café | Theater | Plaza | Bakery | Gastropub | Coffee Shop |
| 3 | Retiro | Spanish Restaurant | Bar | Grocery Store | Art Gallery | Tapas Restaurant | Supermarket | Burger Joint | Indian Restaurant | Museum | Mexican Restaurant |
| 4 | Salamanca | Spanish Restaurant | Restaurant | Bar | Hotel | Plaza | Burger Joint | Grocery Store | Seafood Restaurant | Gymnastics Gym | Bakery |

## 4. Results

As we have seen above, the clusters obtained would be:

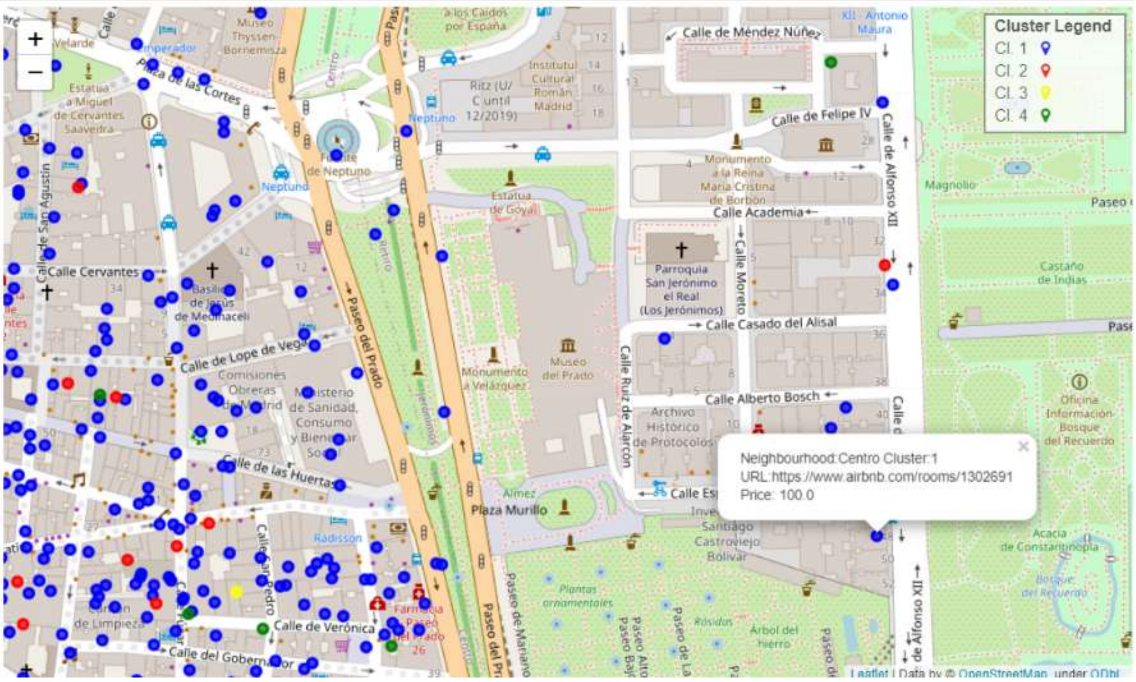
| Labels | Price | Latitude | Longitude | Dist_Museos | Arganzuela | Centro | Chamberí | Retiro | Salamanca |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.304548 | 0.545882 | 0.269896 | 0.668313 | 0.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 |
| 1 | 0.293149 | 0.537250 | 0.260128 | 0.655259 | 0.000000 | 0.998239 | 0.001761 | 0.000000 | 0.000000 |
| 2 | 0.295151 | 0.560264 | 0.290779 | 0.705152 | 0.400631 | 0.000000 | 0.009464 | 0.356467 | 0.233438 |
| 3 | 0.289366 | 0.534205 | 0.263884 | 0.655839 | 0.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 |
| 4 | 0.306681 | 0.525110 | 0.271621 | 0.678320 | 0.000000 | 0.994118 | 0.001961 | 0.000000 | 0.003922 |

| ESSENTIALS | KITCHEN | SMOKING ALLOWED | WHEELCHAIR ACCESSIBLE | Pets | InternetAccess | TempControl | KidsFriendly |
|---|---|---|---|---|---|---|---|
| 0.926065 | 0.988095 | 0.000000 | 0.085213 | 0.000000 | 0.981203 | 0.979950 | 1.000000 |
| 0.889085 | 0.955986 | 0.123239 | 0.026408 | 0.000000 | 0.957746 | 0.940141 | 0.000000 |
| 0.917981 | 0.981073 | 0.138801 | 0.151420 | 0.151420 | 0.940063 | 0.984227 | 0.810726 |
| 0.945455 | 0.985455 | 0.000000 | 0.076364 | 1.000000 | 0.996364 | 0.978182 | 0.865455 |
| 0.917647 | 0.992157 | 1.000000 | 0.090196 | 0.521569 | 0.968627 | 0.949020 | 0.939216 |

The data on venues of each neighborhood are:

| | neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Arganzuela | Tapas Restaurant | Restaurant | Beer Garden | Bakery | Coffee Shop | Museum | Chinese Restaurant | Mediterranean Restaurant | Café | Market |
| 1 | Centro | Plaza | Hotel | Hostel | Tapas Restaurant | Gourmet Shop | Wine Bar | Ice Cream Shop | French Restaurant | Bistro | Restaurant |
| 2 | Chamberí | Tapas Restaurant | Spanish Restaurant | Bar | Restaurant | Café | Theater | Plaza | Bakery | Gastropub | Coffee Shop |
| 3 | Retiro | Spanish Restaurant | Bar | Grocery Store | Art Gallery | Tapas Restaurant | Supermarket | Burger Joint | Indian Restaurant | Museum | Mexican Restaurant |
| 4 | Salamanca | Spanish Restaurant | Restaurant | Bar | Hotel | Plaza | Burger Joint | Grocery Store | Seafood Restaurant | Gymnastics Gym | Bakery |

And the map showing the ranking results:




## 5. Discussion

As for neighborhood classification there are several facts that draw attention. The first significant fact is that the price has barely influenced the ranking (the average of the results ranges from 0.28 to 0.30, which is not significant).

It was more expected that the neighborhood did not influence too much, since the vast majority of the accommodations are located in Centro.

They also don't have too much influence and you have similar data at each Essentials, Kitchen, or Temperature Control classification level, all with very similar means and very close to one.

You also get almost equal averages in the case of wheelchair accessibility, although in this case the surprise is negative, getting surprisingly low results (the highest data tells us that only 15% of the accommodation is accessible).

Where the biggest differences are observed is in the cases of Pets Allowed, Smoking Allowed and Kids Friendly.

This leads us to conclude that the classification carried out would allow us to select by the following groups:

- **Cluster 1**: In Centro Ward (or very close), pets are allowed, smoking is not allowed, and very high probability of being Kids Friendly
- **Cluster 2**: In Centro Ward, pets are not allowed, smoking is not allowed, and not kids friendly
- **Cluster 3**: In Centro Ward, pets are not allowed, smoking is not allowed, and kids friendly
- **Cluster 4**: Not In Centro Ward, pets are not allowed, smoking is not allowed, and very probably kids friendly

As far as venues are concerned, the elements that appear most often are clearly the references to gastronomic offer (Tapas Bar, Bar, Spanish Food...), which is interesting, since the cultural tourist offer is covered by our client.

## 6. Conclusion

The offer of tourist accommodation provided is of great interest, having the vast majority of essentials accommodations, kitchen, internet access at a reasonable price and with a lot of close supply when it comes to food.

Iit would be recommended to improve the wheelchair access part. It is clear that in the downtown area (with the oldest houses in Madrid) it was going to be difficult to meet this requirement (it is not therefore a problem of the dataset that has been used).

If you want to satisfy this type of tourism, it would be advisable to contact companies transporting specialized travelers, since accommodations with wheelchair access are far from the points where you are going to work.