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.

2.2 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	40.425904	-3.681647	https://www.airbnb.com/rooms/4941335	Comodo estudio en la calle Goya	1.506046
12510055	Salamanca	Apartment	Entire home/apt	TV,Internet,Wireless Internet,Air conditioning,...	70.0	40.428352	-3.687155	https://www.airbnb.com/rooms/12510055	Studio & Terrace - Barrio Salamanca	1.567554
15332216	Salamanca	Apartment	Entire home/apt	TV,Internet,Wireless Internet,Air conditioning,...	70.0	40.426671	-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	1.456292
3962279	Salamanca	Apartment	Entire home/apt	TV,Internet,Wireless Internet,Air conditioning,...	60.0	40.426550	-3.670036	https://www.airbnb.com/rooms/3962279	Beautiful apartment in the center	1.956402

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)
```

2.3 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

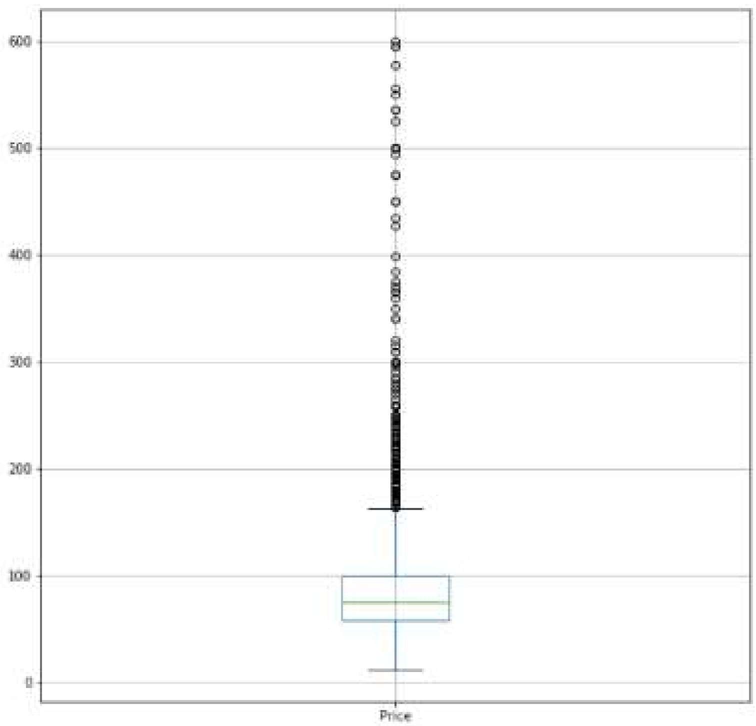
When it comes to numerical variables, we will focus mainly on the price.

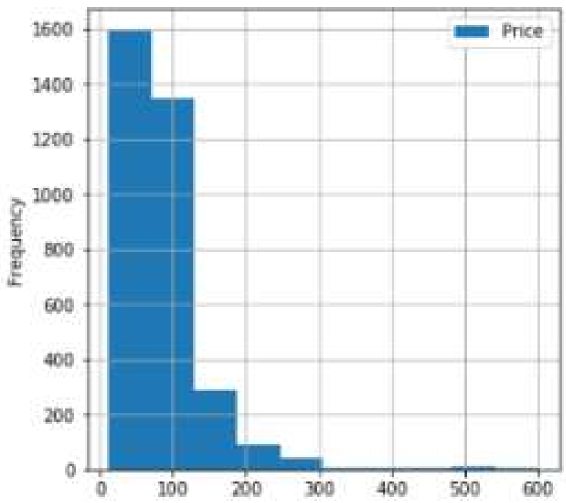
First, it checks for outliers

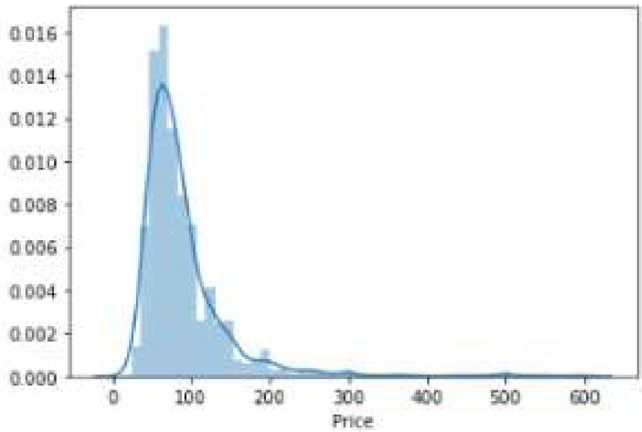
```
print(df['Price'].describe())
df.plot(kind='box',y='Price',grid=True,figsize=(10 ,10))

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.

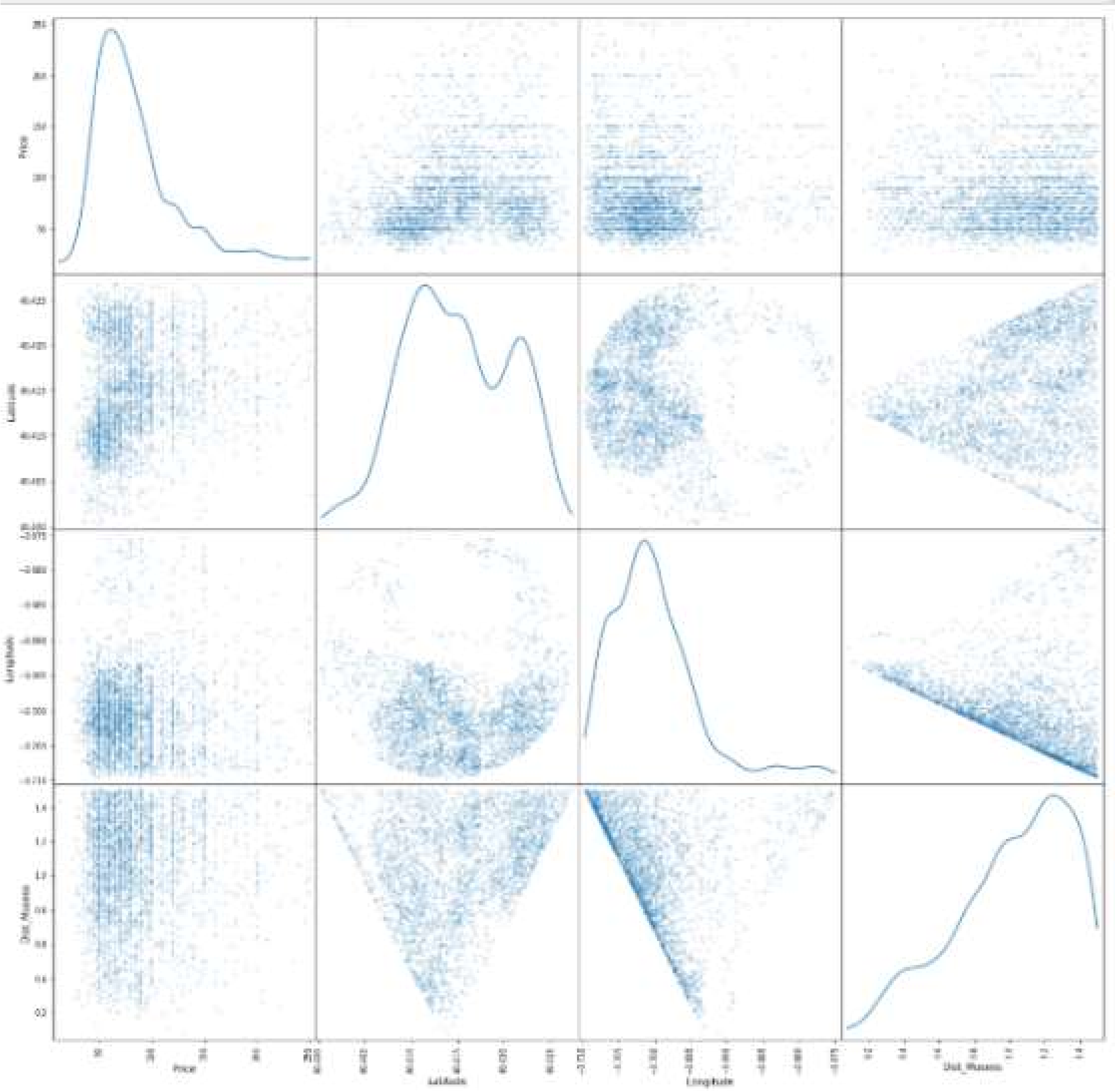
We see that outliers are considered approximately from 180. But it is seen that until approximately 250 there is a large group of values.

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

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 propertie.

	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