

**Project Memory**

Team Members: Carmen Varón & Carlos Naranjo

Updated: June 8th 2018

**Airbnb CompSet Viewer**

This document is a brief summary of the Final Project for the 2017-2017 Master in Data Science at KSchool (Madrid)

Document Scheme

[1. Intro and challenge definition 3](#_Toc516166749)

[2. Project Overview & Road Map Summary 3](#_Toc516166750)

[Scope 3](#_Toc516166751)

[Road Map 3](#_Toc516166752)

[Step 0 - Getting Data 4](#_Toc516166753)

[Step 1 - Understanding Data 4](#_Toc516166754)

[Step 2.1 - Data Modeling - Establishing Success Metrics & Long Term Rental Income 4](#_Toc516166755)

[Step 2.2 - Competitors and key features 4](#_Toc516166756)

[Step 3 - Developing a visual interface for the analysis 5](#_Toc516166757)

[3. Methodology: Tools, Data & Modeling 5](#_Toc516166758)

[Step 0 - Getting Data 5](#_Toc516166759)

[Step 1 - Understanding Data 5](#_Toc516166760)

[Step 2.1 - Data Modeling - Establishing Success Metrics & Long Term Rental Income 6](#_Toc516166761)

[Income 2017 Calculation 6](#_Toc516166762)

[Long Term Rental Calculation 7](#_Toc516166763)

[Step 2.2 - Competitors and key features 7](#_Toc516166764)

[Step 3 - Developing a visual interface for the analysis 7](#_Toc516166765)

[4. Summary & Conclusions 7](#_Toc516166766)

[5. Visual Interface: Frontend Operation Manual - Airbnb CompSet Viewer 8](#_Toc516166767)

[Menu 8](#_Toc516166768)

[Intro 8](#_Toc516166769)

[Step 1 - Selection of CompSet 9](#_Toc516166770)

[Step 2 - Check CompSet Metrics 10](#_Toc516166771)

[6. Append 11](#_Toc516166772)

[Data Model 11](#_Toc516166773)

[Listing File: Fieds explanation 11](#_Toc516166774)

[Review File: Fieds explanation 13](#_Toc516166775)

[Calendar File: Fieds explanation 13](#_Toc516166776)

[Calendar File: Long\_Term Rental Price 13](#_Toc516166777)

[K-Neighbours: Fieds explanation 14](#_Toc516166778)

[2017 Performance: Fieds explanation 14](#_Toc516166779)

# Intro and challenge definition

We are currently on dates of historical records on tourism. At this juncture, two disruptive innovations has changed the tourism market:

* Irruption of the Internet (with a radical change in distribution models)
* Incipient explosion of the collaborative economy (room & home sharing)

This project will focus on the collaborative economy at the Hotel Industry. The war is served at the internet, with millions of new players acting as hotel managers with their own property. These new players are non-hoteliers, so their capabilities to price their properties are reduced.

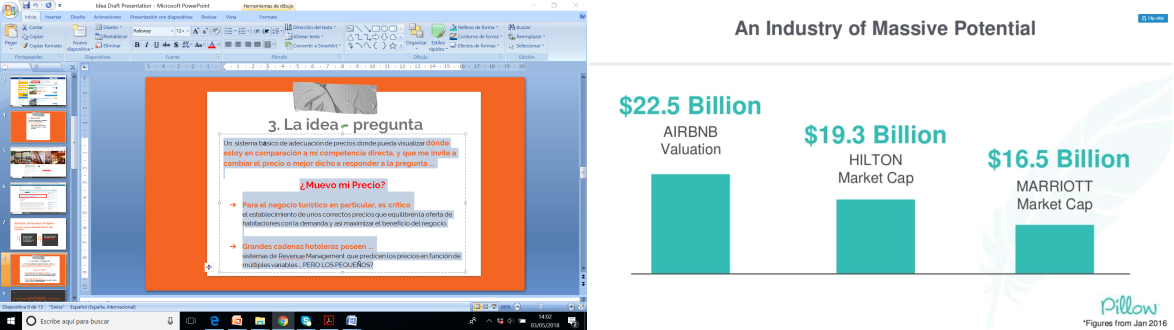


Figure 1 - Top Palyers Hospitality by number of rooms (Jan 201) - Source: Pillow

Revenue Management Systems are based on both, internal and external data, and each day these systems become more and more complex. So this project will focus on the Competitors Analysis (\*), as it is the first step to establish a pricing strategy.



Figure 2 - Revenue Management at Hospitality Industry Scheme - Source: Project Team

As Airbnb comprehends thousands of listings, it becomes more difficult to establish who are your competitors and estimate their performance metrics. This is the opportunity and challenge that this project will try to go solve.

# Project Overview & Road Map Summary

Project Goal

This Project will develop atool for property owners.This tool will help them to find theirCompSet (Top 10 competitors) and present a comparative dashboard of this group, based on their main metrics.

## Scope

The scope will be the Airbnb site properties, considered one of the biggest players at the pitch. And more specifically, Madrid as the selected city for the analysis, with over than 16K listed properties at the site.

## Road Map

The data for this project will be obtained from the Inside [Airbnb Initiative](insideairbnb.com). In their words: “Inside Airbnb is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world.”

### Step 0 - Getting Data

The typical set of data comprehends Listings, Reviews, Calendar and Neighbourhood. This data set is available for many destinations around the world. So the first decision was selecting Madrid as the destination target, and download its set of files.

### Step 1 - Understanding Data

The files kit has a large number of fields, some of them scrapped straight for Airbnb website, and some others calculated transforming the scrapped data. So it was crucial understanding Airbnb operation model for hosts, and its implications on these datasets.



Figure 3 - Project Road Map - Source: Project Team

*Step 2 - Data Modeling*

### Step 2.1 - Data Modeling - Establishing Success Metrics & Long Term Rental Income

What is a good listing for a owner? That is the first question that this team tried to answer, so we could contrast good performers and their features vs poor performers. Hosts decide open their homes to the collaborative economy, for obvious economic reasons. So, the question is easy: ‘How much money do a listing make?’. In other words: ‘What is the estimated income for a single listing’?. Scrapped datasets do not have Income Information (that information resides on hosts private area), so it has to be estimated.

Is it better for a listing vacation rental (Airbnb, HomeAway, …) or classical long term rental?. Property owners, has to decide between classical long term rental and vacation rental. For that reason the project will make a comparison for each listing between long term rental and vacation rental Income. Thus, any property owner can make a comparison of income at both models.

### Step 2.2 - Competitors and key features

*Which listings of all that at Airbnb website are a single listing competitors?* Obviously in a huge amount of listings, it is crucial to have visibility. Visibility is based on Airbnb internal algorithm, but some characteristics are known (location, capacity, number of rooms, price …). This project will try to search for a single listing, a cluster formed of similar listings with a comparison proposal.

What features are more critical to success at Airbnb? Once we have an income estimation, this project will try to find statistical relations of any feature and a listing’s performance. With the final goal of giving a single recommended price, some algorithms will be runned.

### Step 3 - Developing a visual interface for the analysis

As commented before, our target user is a listing owner that wants to know how to improve its performance. So an easy tool will be developed for that propose.

# Methodology: Tools, Data & Modeling

### Step 0 - Getting Data

Once we decided insideairbnb.com as the airbnb data provider, the first step was discovering on the internet a bunch of data science projects based on this data. Lots of different approaches were analyzed but none fit on our vision neither offered interesting insights, so we decided to start from scratch.

### Step 1 - Understanding Data

We had the data, but we had to do a business and process analysis to understanding Airbnb operation. As reported by Ajay Deep (Airbnb Researcher for OnDemand Economy), Airbnb operation can be summarized in 5 steps as follows:

1. Hosts list out their property details on Airbnb along with other factors like pricing, amenities provided etc.
2. Airbnb sends a professional photographer (if available) to the property location in order to take high quality photographs.
3. Travellers search for a property in the city where they wish to stay and browse available options according to price, amenities etc.
4. Booking is made through Airbnb where traveller pays the amount mentioned by host and some additional money as transaction charges.
5. Host approves the booking. Traveller stays there and finally Airbnb pays the amount to the host after deducting their commission.

The host and the traveler can rate each other and can write reviews based on the experience.

Business Model:

* Commission from Property Owners (Hosts)
* Airbnb charges flat 10% commission from hosts upon every booking done through the platform.
* Transaction fee from Travellers (Guests)
* Airbnb charges 3% of the booking amount as transaction charges from travellers upon every confirmed booking.

This brief summary took us to the conclusion that reservations are the key to understand what means a good or bad performance at Airbnb, but we did not have that information. Thus we made tons of paper drafts to understand how the data was related to Airbnb Operation and our project goal.

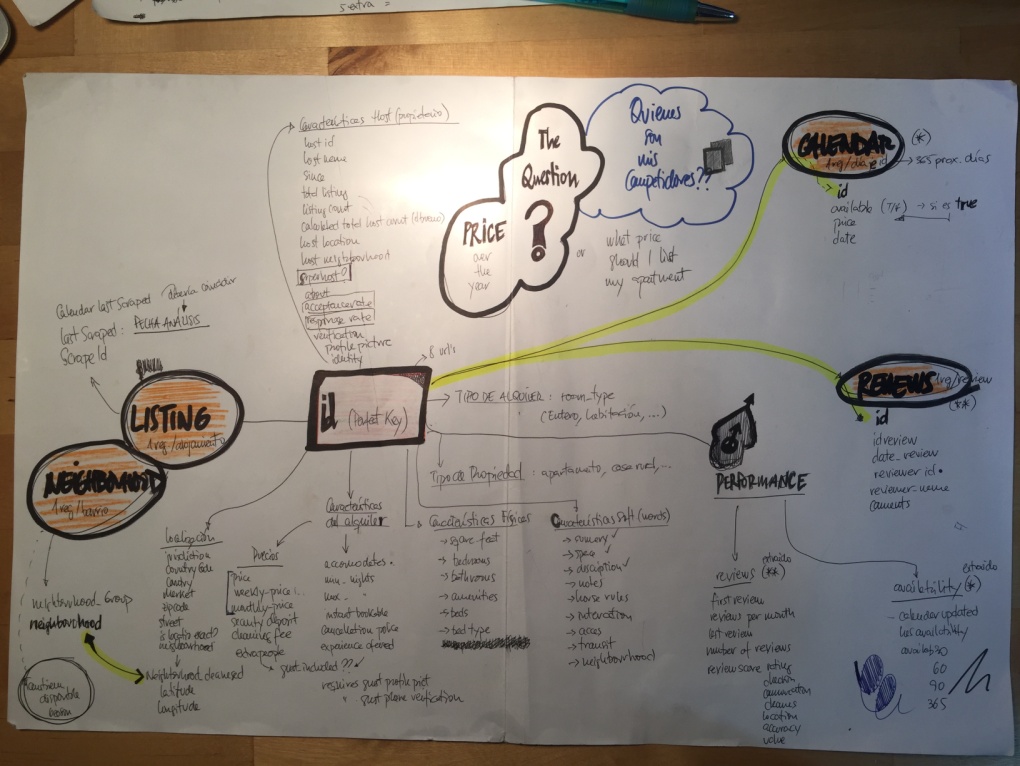


Figure 4 - Project First Draft Data Model - Source: Project Team

*Step 2 - Data Modeling*

The data model (fields, type and meaning) are explained append at this document, but this section will show the way data was transformed for the purpose of developing a final user tool, which answered our goal. Data modeling methodology can be observed at the general scheme presented below:



Figure 5 - Data modeling scheme - Source: Project Team

The data first approach was made in different tools: Python, R, Qlikview and QlikSense. But we always come to the same point: data does not show good or bad performance. So we had to build that.

### Step 2.1 - Data Modeling - Establishing Success Metrics & Long Term Rental Income

## Income 2017 Calculation

As mentioned before, Income is the metric that will allow compare properties. And we need to make some considerations:

* The cycle of demand is year based, so the calculation has to be for a whole cycle
* There are 3 type of business at Airbnb: entire house, room or shared room. So the three of them has to be analyzed separately. Our focus is the Entire House type of property.
* Income is the simple roomnights x price calculation. So we had to make some hypothesis, as we do not have that information. We have reviews, minimum nights and price:
  + Year for calculation: 2017 (last year)
  + Roomnights = (Reviews made at Airbnb + Reviews not reported by guests) *per* average duration of trip (days)
  + The price to consider will be the standard price reported at the listing
  + Average duration / booking = 4.2 days (Source: Airbnb Comunidad de Madrid 2016 Report)
  + Percentage of bookings with a review = 70% (Source: Reported by Airbnb CEO)
  + Listings puiblished before 2017 (Calculated at Qlikview as all those with a Listing Number lower than: max({1} if (year(first\_review)<year(last\_scraped)-1,id))
  + Big numbers at roomnights are avoid, limiting maximum occupancy at 80% (255 days/year) (Source: Team estimation of great occupation rate)

Once we have the model, we proceeded to contrast it with the few official Airbnb Data for Madrid. Results were pretty close to our calculations. So we decided to use then as an good estimation.

* Typical Income by host: 4.400€ (Airbnb) / 4.320 $ (Project)
* Typical nights hosted per listing: 70 (Airbnb) / 78 (Project)



Figure 6 - Airbnb official report 2017 Comunidad de Madrid - Source: Airbnb

## Long Term Rental Calculation

[El Mundo and Alquiler Seguro](http://www.elmundo.es/grafico/economia/2015/07/29/55ae303d268e3e344d8b457a.html), reported price per district in Madrid (3Q) for long term rental, based on district and number of rooms. This information is available in our data model, where 10.252 out of 10.338 Entire-House Listings has a Long Term Rental Income estimated.

### Step 2.2 - Competitors and key features

At the beginning of the project we had two main goals that we thought Data science algorithms could help us to achieve.

The first one was to find the competitors of each listing: the owner of one house could see its competitors characteristics as well as their price, reviews and scores so they can get an idea of its weaknesses and strengths, and adapt according to it (although the main characteristics such as number of rooms or location are immutable, others as price, minimum number of nights, photos and some amenities, could be changed). All Python scripts used for looking into this goal can be found in repository folder “Clustering trials”.

Our initial idea was to split the data into clusters, so all the members of a cluster would be competitors. To see if there was light in this, we tried to plot all listings reducing its dimensionality with PCA and TSNE, but there were no obvious clusters, and we experienced the curse of dimensionality, so we decided to reduce the number of variables to use (especially amenities, as we had 1 boolean feature by amenity, resulting in around 150 features).

We wanted to try some different algorithms, but we had the problem that some of the features were categorical, and most algorithms are based on classical distance metrics as Euclidean (whichare not ideal metrics forboolean features), so we finally decided to use a version of the k-means algorithm that is designed to be used with a combination of numerical and categorical features: K-prototypes. The main difference between these 2 algorithms is just the metric that it uses (this metric is explained deeper in the notebook “2.KNeighbors\_byType”).

Different number of k were tried and we tested the ones that seemed to be better with the “elbow method”. Centroids seemed to be stable, but the Silhouette coefficient was really low, telling us that the clusters were not sufficiently separated, we could see it when looking at the clusters too.

We tried a different approach, instead of splitting in clusters, we would calculate the k-nearest neighbours of each listings, using as metric the same one that the k-prototypes algorithm uses(notebook “2.KNeighbors\_byType”). We tested the results and it seemed that it was able to find listings that we could have considered competitors.

Our second goal was to find the price that a house should put in order to maximize the total income in a year. Our initial approach was to make a linear model regression with total reviews in 2017 as target, and price as one of the features used in the regression, in order to have an equation with coefficients that we could use in a system of equations to maximize the total income. We tested multiple combinations of variables and we tried different algorithms (not only linear models) and we couldn´t get any with a decent R2 and MSE.

### Step 3 - Developing a visual interface for the analysis

All these data modeling, makes sense at the interactive dashboard ‘Airbnb CompSet Viewer’. It has been developed using Qlik Sense, because of its flexibility to add extensions like action buttons or tables with url images. Over than 15 different user interfaces where developed, always in the search of usability and value for users.

# Summary & Conclusions

It is an internet race, so every decision is based on the information reported at Airbnb website. As an Internet race, success will be related with the position of your proposal among the rest, and the attractiveness of your product. This project will not cover the position at the Aribnb listing search result, neither the attractiveness of the property pictures, critical to understand success. This should be next step to cover.

Our findings (or the lack of them), is that there is no Airbnb, nor Airbnb Spain, nor Airbnb Madrid, there is a unique reality: similar behavior is just observed at neighborhood level.

This project goal is not analyzing Airbnb impact at a destination level, but offering lots of answers and clues to a Airbnb property owner:

* How much money do I make
* Do I make more money than renting my property on a Long-Term basis
* Do I make more money than similar properties
* Is my price a good one? Should I raise it or decrease it?
* What price should I list for summer?
* …

# Visual Interface: Frontend Operation Manual - Airbnb CompSet Viewer

Airbnb CompSet Viewer is a simple two steps tool. Watch [this](https://youtu.be/uUsJcLAzekE) video to see how it works:

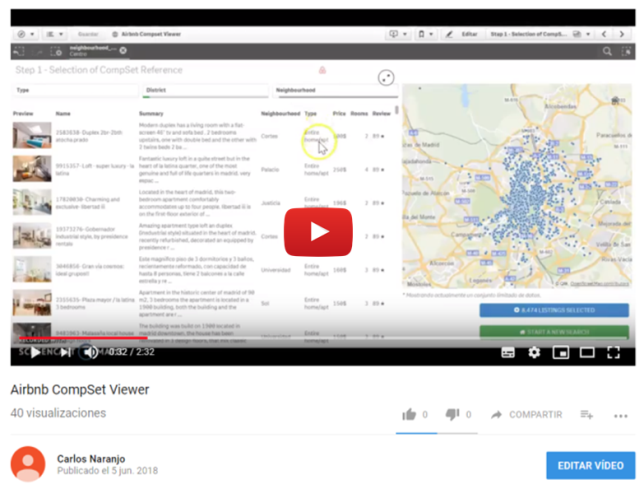
[](https://youtu.be/uUsJcLAzekE)

Figure 7 - A demo video can be played at youtube (Airbnb CompSet Viewer) - Source: Project Team

## Menu

It is the first view of the app. To start select the Intro dashboard:

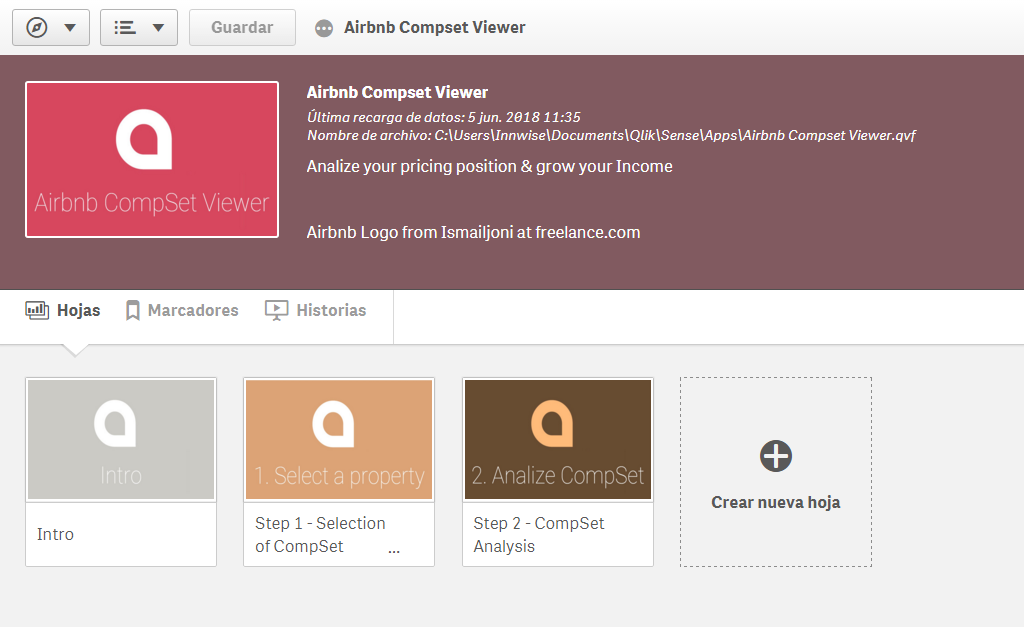


Figure 8 - Airbnb CompSet Viewer home - Source: Project Team

## Intro

At the intro you can have a explanation of what is this tool, its basics and the Start Button. You can star just pressing START TO GO TO STEP 1.

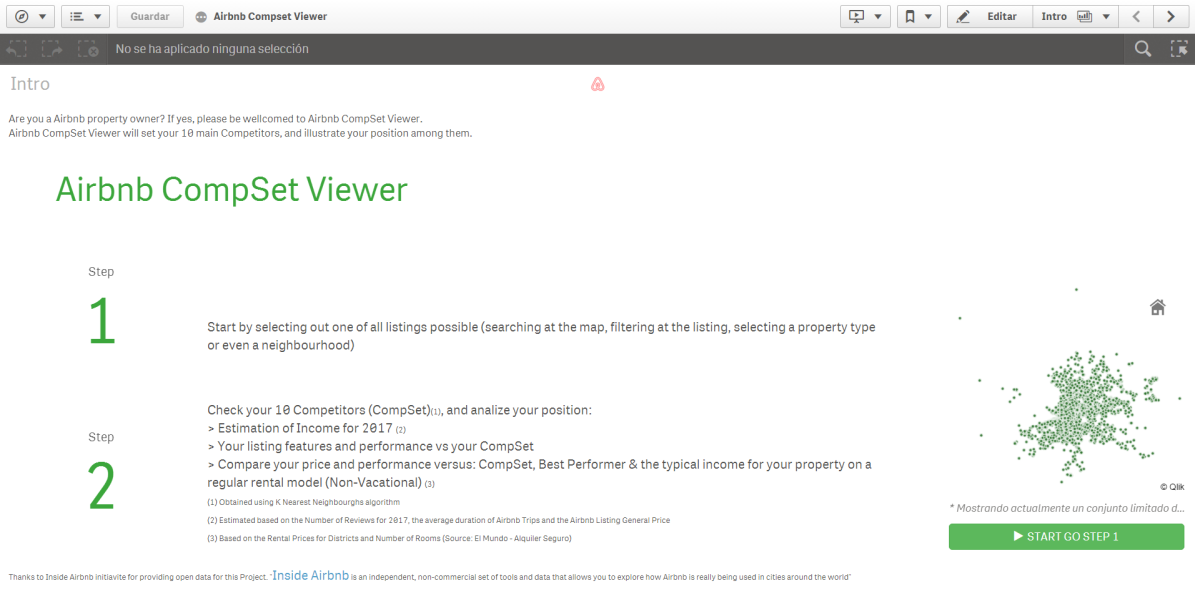


Figure 9 - Airbnb CompSet Viewer Intro - Source: Project Team

## Step 1 - Selection of CompSet

At this dashboard you have to select your single listing to compare, through a list or a map, where you can refine your search.

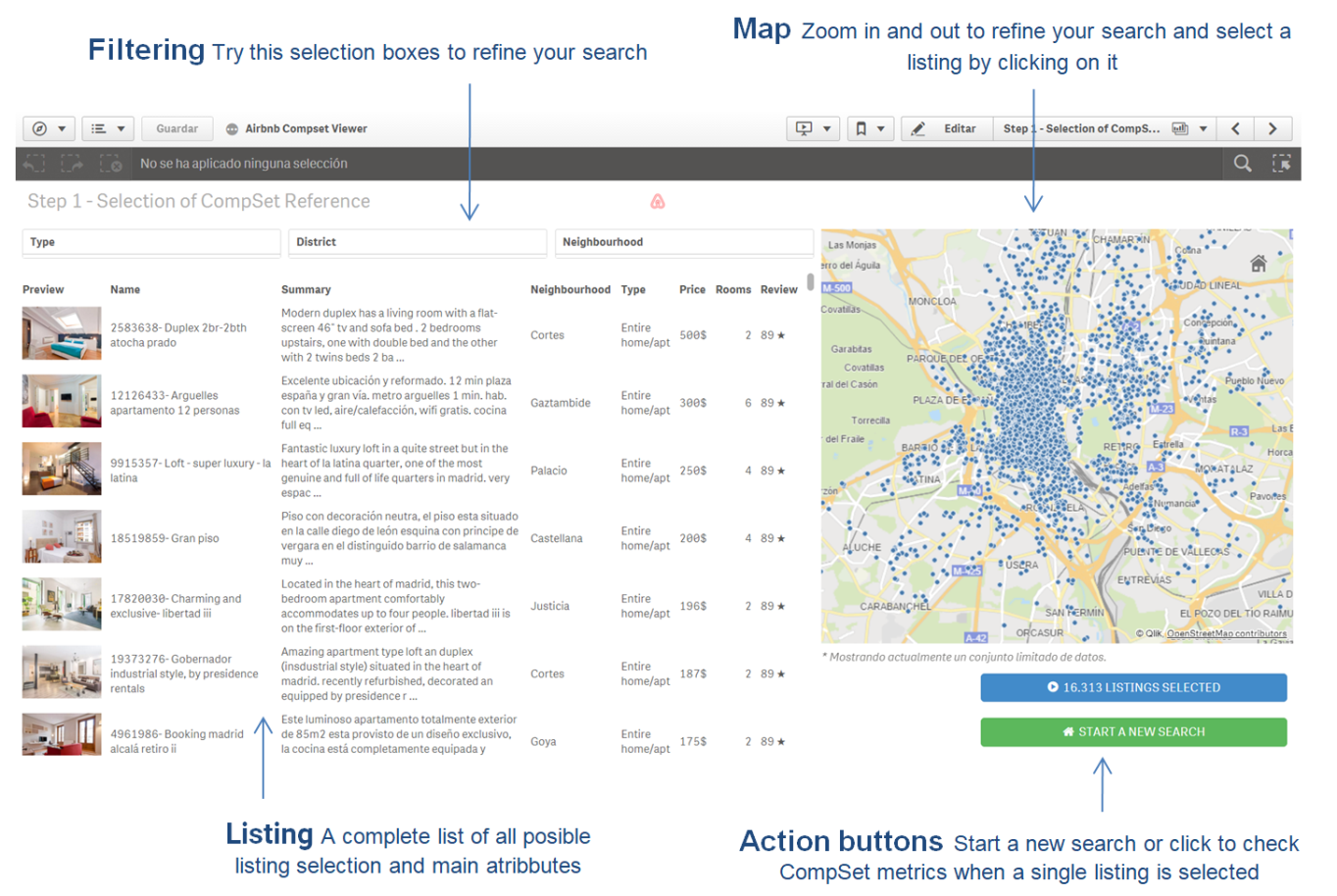


Figure 10 - Airbnb CompSet Viewer Step 1 - Source: Project Team

## Step 2 - Check CompSet Metrics

At this dashboard you have to select your single listing to compare, through a list or a map, where you can refine your search.

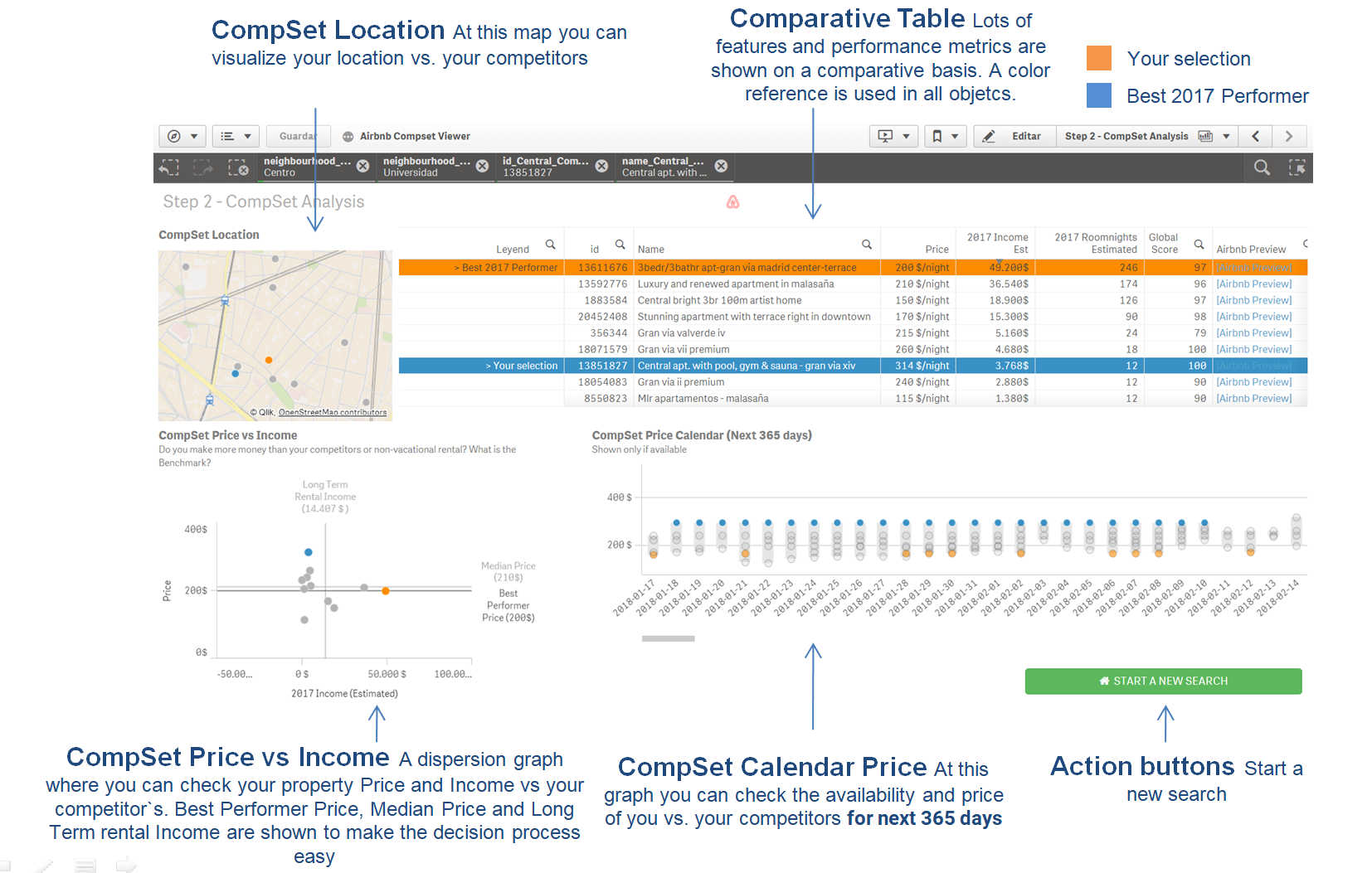
****

Figure 11 - Airbnb CompSet Viewer Step 2 - Source: Project Team

# Append

## Data Model

The main data that for this project has been downloaded from Insideairbnb.com. It will be a snapshot of date Compiled at 17 January, 2018, for Madrid with this files kit:

|  |  |  |
| --- | --- | --- |
| File Name | Used | Description |
| [listings.csv.gz](http://data.insideairbnb.com/spain/comunidad-de-madrid/madrid/2018-01-17/data/listings.csv.gz) | Yes | Detailed Listings data for Madrid |
| [calendar.csv.gz](http://data.insideairbnb.com/spain/comunidad-de-madrid/madrid/2018-01-17/data/calendar.csv.gz) | Yes | Detailed Calendar Data for listings in Madrid |
| [reviews.csv.gz](http://data.insideairbnb.com/spain/comunidad-de-madrid/madrid/2018-01-17/data/reviews.csv.gz) | Yes | Detailed Review Data for listings in Madrid |
| [listings.csv](http://data.insideairbnb.com/spain/comunidad-de-madrid/madrid/2018-01-17/visualisations/listings.csv) | No | Summary information and metrics for listings in Madrid |
| [reviews.csv](http://data.insideairbnb.com/spain/comunidad-de-madrid/madrid/2018-01-17/visualisations/reviews.csv) | No | Summary Review data and Listing ID |
| [neighbourhoods.csv](http://data.insideairbnb.com/spain/comunidad-de-madrid/madrid/2018-01-17/visualisations/neighbourhoods.csv) | No | Neighbourhood list for geo filter. |
| [neighbourhoods.geojson](http://data.insideairbnb.com/spain/comunidad-de-madrid/madrid/2018-01-17/visualisations/neighbourhoods.geojson) | No | GeoJSON file of neighbourhoods of the city. |

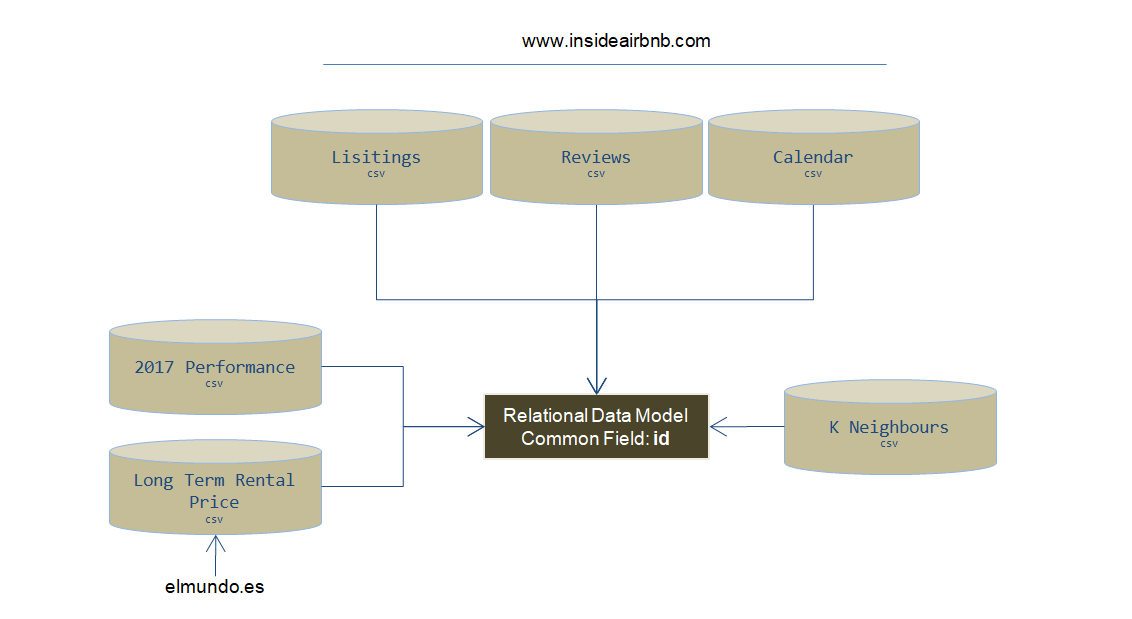


Figure 11 - Data Model - Source: Project Team

Other file is used for this project has been obtained from [El Mundo and Alquiler Seguro](http://www.elmundo.es/grafico/economia/2015/07/29/55ae303d268e3e344d8b457a.html): Entire-House Long Term Rental Income estimated.

## Listing File: Fieds explanation

Each file is a Listing

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Field** | **Extracted** | **Format** |
| Identifier - Lisiting | **id** | Scrapped | Number |
| Listing Description | **name** | Scrapped | Text |
| **summary** | Scrapped | Text |
| **space** | Scrapped | Text |
| **description** | Scrapped | Text |
| **neighborhood\_overview** | Scrapped | Text |
| **notes** | Scrapped | Text |
| **transit** | Scrapped | Text |
| **[access]** | Scrapped | Text |
| **interaction** | Scrapped | Text |
| **house\_rules** | Scrapped | Text |
| Location | **is\_location\_exact** | Calculated | Boolean |
| **latitude** | Scrapped | Number |
| **longitude** | Scrapped | Number |
| **zipcode** | Calculated | Number |
| **street** | Calculated | Text |
| **neighbourhood** | Calculated | Text Category |
| **neighbourhood\_cleansed** | Calculated | Text Category |
| **city** | Scrapped | Text Category |
| **state** | Scrapped | Text Category |
| **market** | Calculated | Text Category |
| **smart\_location** | Calculated | Text Category |
| **country\_code** | Calculated | Text Category |
| **country** | Scrapped | Text Category |
| **neighbourhood\_group\_cleansed** | Calculated | Text Category |
| Performance - Demand | **availability\_30** | Calculated | Number |
| **availability\_60** | Calculated | Number |
| **availability\_90** | Calculated | Number |
| **availability\_365** | Calculated | Number |
| **first\_review** | Scrapped | Date |
| **last\_review** | Scrapped | Date |
| **number\_of\_reviews** | Scrapped | Number |
| **reviews\_per\_month** | Calculated | Number |
| Performance - Quality | **review\_scores\_accuracy** | Scrapped | Number |
| **review\_scores\_cleanliness** | Scrapped | Number |
| **review\_scores\_checkin** | Scrapped | Number |
| **review\_scores\_communication** | Scrapped | Number |
| **review\_scores\_location** | Scrapped | Number |
| **review\_scores\_value** | Scrapped | Number |
| **review\_scores\_rating** | Scrapped | Number |
| PhisicListingFeatures | **amenities** | Scrapped | List of Text Categories |
| **square\_feet** | Scrapped | Number |
| **bathrooms** | Scrapped | Text Category |
| **bedrooms** | Scrapped | Text Category |
| **beds** | Scrapped | Text Category |
| **property\_type** | Scrapped | Text Category |
| **room\_type** | Scrapped | Text Category |
| **bed\_type** | Scrapped | Text Category |
| Pricing | **price** | Scrapped | Number |
| **weekly\_price** | Scrapped | Number |
| **monthly\_price** | Scrapped | Number |
| **security\_deposit** | Scrapped | Number |
| **cleaning\_fee** | Scrapped | Number |
| **extra\_people** | Scrapped | Number |
| RentalCharasteristics | **jurisdiction\_names** | Unknown | Unknown |
| **instant\_bookable** | Scrapped | Boolean |
| **is\_business\_travel\_ready** | Scrapped | Boolean |
| **requires\_license** | Scrapped | Boolean |
| **require\_guest\_profile\_picture** | Scrapped | Boolean |
| **require\_guest\_phone\_verification** | Scrapped | Boolean |
| **guests\_included** | Unknown | Number |
| **minimum\_nights** | Scrapped | Number |
| **maximum\_nights** | Scrapped | Number |
| **license** | Scrapped | Text |
| **cancellation\_policy** | Scrapped | Text Category |
| **accommodates** | Scrapped | Text Category |
| RentingFeatures | **experiences\_offered** | Scrapped | Text |
| ScrappingInfo | **last\_scraped** | Calculated | Date |
| **calendar\_last\_scraped** | Calculated | Date |
| **scrape\_id** | Calculated | Number |
| url | **listing\_url** | Scrapped | Text |
| **thumbnail\_url** | Scrapped | Text |
| **medium\_url** | Scrapped | Text |
| **picture\_url** | Scrapped | Text |
| **xl\_picture\_url** | Scrapped | Text |
| Calendar Info | **has\_availability** | Calculated | Boolean |
| **calendar\_updated** | Scrapped | Date |
| Host Info | **host\_is\_superhost** | Scrapped | Boolean |
| **host\_has\_profile\_pic** | Scrapped | Boolean |
| **host\_identity\_verified** | Scrapped | Boolean |
| **host\_since** | Scrapped | Date |
| **host\_verifications** | Scrapped | List of Text Categories |
| **host\_id** | Scrapped | Number |
| **host\_listings\_count** | Scrapped | Number |
| **calculated\_host\_listings\_count** | Calculated | Number |
| **host\_response\_rate** | Scrapped | Number |
| **host\_acceptance\_rate** | Scrapped | Number |
| **host\_total\_listings\_count** | Scrapped | Number |
| **host\_url** | Scrapped | Text |
| **host\_name** | Scrapped | Text |
| **host\_about** | Scrapped | Text |
| **host\_thumbnail\_url** | Scrapped | Text |
| **host\_picture\_url** | Scrapped | Text |
| **host\_location** | Scrapped | Text Category |
| **host\_neighbourhood** | Scrapped | Text Category |
| **host\_response\_time** | Scrapped | Text Category |

## Review File: Fieds explanation

Each file is a Comment

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Field** | **Extracted** | **Format** |
| Identifier Lisiting | **listing\_id** | Scrapped | Number |
| Review  info | **Id\_Review** | Scrapped | Number |
| **date** | Scrapped | Date |
| **reviewer\_id** | Scrapped | Number |
| **reviewer\_name** | Scrapped | Text |
| **comment** | Scrapped | Text |

## Calendar File: Fieds explanation

Each file is a Calendar date for next 365 days, for each listing

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Field** | **Extracted** | **Format** |
| Identifier Lisiting | **listing\_id** | Scrapped | Number |
| Pricing and availability info | **available** | Scrapped | Boolean |
| **date** | Scrapped | Date |
| **price\_Calendar** | Scrapped | Number |

## Calendar File: Long\_Term Rental Price

Each file is a District & Number of rooms Combination

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Field** | **Extracted** | **Format** |
| District | **Lugar** | Scrapped | Text |
| Number of rooms | **Dormitorios** | Scrapped | Number |
| Monthly Rent Price | **Alquiler** | Scrapped | Number |

## K-Neighbours: Fieds explanation

Each file is a id\_Central\_Comparison and an id combination for the 10 Kneighbours. Total number of fields 10 \* Number of Listings.

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Field** | **Extracted** | **Format** |
| Identifier Lisiting to Compare | **id\_Central\_Comparison** | Calculated | Number |
| Identifier Lisiting of KNeighbours | **id** | Calculated | Number |
| Number of Neighbour | **NeighbourNum** | Calculated | Number |

## 2017 Performance: Fieds explanation

Each file represents a Listing Property

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Field** | **Extracted** | **Format** |
| Identifier - Lisiting | **id** | Calculated | Number |
|  | **Published\_Before\_2017** | Calculated | Text |
| 2017 performance | **Active\_Last\_90\_days** | Calculated | Boolean |
| **2017\_Total\_Number\_Bookings\_Est** | Calculated | Number |
| **2017\_Avg\_Nights\_per\_Booking\_Est** | Calculated | Number |
| **2017\_Total\_Number\_Reviews** | Calculated | Number |
| **2017\_Total\_Income\_Est** | Calculated | Number |
| **2017\_Total\_Income\_Est/Room** | Calculated | Number |
| **2017\_Total\_Number\_Roomnigths\_Est** | Calculated | Number |
| **2017\_Occupation\_Rate\_Est** | Calculated | Number |
| Next 365 days performance | **Next\_30\_days\_Occupation\_at\_the\_date** | Calculated | Number |
| **Next\_60\_days\_Occupation\_at\_the\_date** | Calculated | Number |
| **Next\_90\_days\_Occupation\_at\_the\_date** | Calculated | Number |
| **Next\_365\_days\_Occupation\_at\_the\_date** | Calculated | Number |