

**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 2017-2018 Master in Data Science at KSchool (Madrid)

Document Scheme

[1. Summary 3](#_Toc516232759)

[2. Introduction and challenge definition 3](#_Toc516232760)

[3. Project Overview & Road Map Summary 4](#_Toc516232761)

[Road Map 4](#_Toc516232762)

[Step 0 - Getting Data 4](#_Toc516232763)

[Step 1 - Understanding Data 4](#_Toc516232764)

[Step 2.1-Data Modeling - Establishing Success Metrics & Long Term Rental Income 5](#_Toc516232765)

[Step 2.2-Competitors and key features 5](#_Toc516232766)

[Step 3-Developing a visual interface for the analysis 5](#_Toc516232767)

[4. Methodology: Tools, Data & Modeling 5](#_Toc516232768)

[Step 0 - Getting Data 5](#_Toc516232769)

[Step 1 - Understanding Data 6](#_Toc516232770)

[Step 2.1-Data Modeling - Establishing Success Metrics & Long Term Rental Income 7](#_Toc516232771)

[Step 2.2-Competitors and key features 8](#_Toc516232772)

[Step 3-Developing a visual interface for the analysis 8](#_Toc516232773)

[5. Visual Interface: Frontend Operation Manual - Airbnb CompSet Viewer 8](#_Toc516232774)

[Starting at the Menu Interface - Screenshot Overview 9](#_Toc516232775)

[Intro - Screenshot Overview 9](#_Toc516232776)

[Step 1 - Selection of CompSet- Screenshot Overview 10](#_Toc516232777)

[Step 2 - Check CompSet Metrics - Screenshot Overview 10](#_Toc516232778)

[6. Conclusions 11](#_Toc516232779)

[Annexes 12](#_Toc516232780)

[Data Model 12](#_Toc516232781)

[Listing File: Fieds explanation 12](#_Toc516232782)

[Review File: Fieds explanation 14](#_Toc516232783)

[Calendar File: Fieds explanation 14](#_Toc516232784)

[LongTermRentalPrice: Fields explanation 14](#_Toc516232785)

[K-Neighbours: Fieds explanation 15](#_Toc516232786)

[2017 Performance: Fieds explanation 15](#_Toc516232787)

# Summary

Airbnb site publishes thousands of listings at almost any destination worldwide. The war is served at the internet, with millions of new players acting as hotel managers with their own properties listed at Airbnb site. These new players are non-hoteliers, and their capabilities and competences to price their properties are usually limited. Nowadays property owners follow the trial and error technique for pricing.

This project will provide a visual tool that facilitates Airbnb hosts in Madrid, an analysis of their property performance versus their straight competitors. The project output will offer a tool: Airbnb CompSet Viewer. It will present lots of answers and clues to an 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?

Airbnb CompSet Viewer is an easy interface. Users have just to select a single property and then proceed to analyze a dashboard full of insights about the selected property versus its ten main competitors. These competitors are calculated through machine learning techniques, based on the similarity of listed properties at Airbnb site.

Note that our aim is not analyzing Airbnb features, neither its impact in Madrid City. Moreover, the project will not cover the position at the Airbnb listing search result, neither the attractiveness of the property pictures, both critical to understand success at this internet arena. This should be next step to cover.

# Introduction and challenge definition

Nowadays, we are living historical records on tourism. Furthermore, in recent years two disruptive innovations have changed the tourism market around the world:

* Irruption of the Internet (with a radical change in distribution models)
* Incipient explosion of the collaborative economy (room & home sharing, see figure 1)

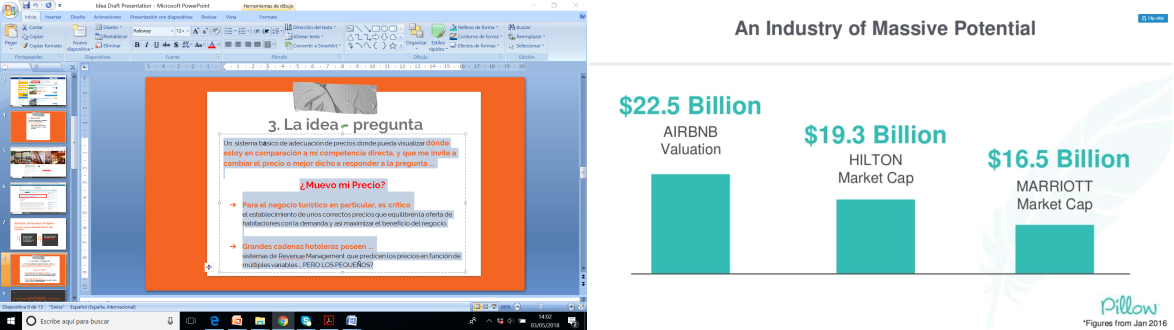


Figure 1 - Top Players Hospitality by number of rooms (Jan 2016) - Source: Pillow

One of the big challenges is being able to establish the right strategy price. Professionals manage new techniques and method to put the right price. Commonly Revenue Management Systems techniques are used (see figure 2). They are based on both internal and external data, and each day these systems become more and more complex, even for updated professionals and top rank hotels.



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

In this new scenario, the incipient explosion of the collaborative economy, and specifically the Airbnb platform, it is offering an attractive option for non-professionals in the hotel sector. People with a property, or even with an available room at home, are renting their places to tourists in exchange of, sometimes, important revenues and in parallel challenging the market. Indeed, the war is served at the internet, with millions of new players acting as hotel managers with their own properties.

These new players are non-hoteliers, hence their capabilities and competences to price their properties are usually limited. Currently, new players establish their price strategy following the trial and error method, and they do not have access to the updated techniques used by professionals at the tourist sector. Additionally, as Airbnb offers thousands of listings, it becomes more difficult to establish who are your direct competitors and estimate their performance metrics.

Consequently, there is a challenge and an opportunity that this project will try to address and hopefully, to solve. Our aim is to offer a useful and friendly tool to non-professionals new hoteliers that help them to establish a pricing strategy. Specifically the project will focus on the collaborative economy at the Hotel Industry and on the Competitors Analysis, as it is the first step to establish a pricing strategy.

# Project Overview & Road Map Summary

Project Goal and Scope

The Project will develop a tool for property owners. This tool will help them to find their CompSet (Top 10 competitors) and present a comparative dashboard of this group, based on their main metrics. The scope will be the Airbnb site properties, considered one of the biggest players at the pitch. And more specifically, Madrid area as the selected city for the analysis, with over than 16K listed properties at the site.

Airbnb in Madrid

Madrid is the capital and largest city of Spain. The city has almost 3.2 million inhabitants with a [metropolitan area](https://en.wikipedia.org/wiki/Madrid_metropolitan_area) population of approximately 6.5 million. Furthermore, in 2017 around 6.7 million international visitors arrived in the Community of Madrid, over 3.7 million more than in 2001[[1]](#footnote-1).  The Spanish capital is the third largest-city in the European Union.

Madrid is among the most important Spanish tourism hubs. As the capital city, a cosmopolitan city, a business center, seat of government, Spanish Parliament, and residence of the Spanish monarch, Madrid is also the political center of Spain. Madrid is characterized by intense cultural and artistic activity and a very lively nightlife. [Travel spending](https://www.statista.com/statistics/450363) in the Community of Madrid is higher on average than in other regions, including Catalonia, the Canary Islands and Balearic Islands.

According to Airbnb official data[[2]](#footnote-2) for 2016, 642.000 guests were accommodated at Airbnb listed properties in Madrid region, with an estimated income of 64 MM€.A recent report produced by Airbnb for 2017, shows 1.020.000 guest inbound in Madrid City with Airbnb[[3]](#footnote-3), almost twice than previous year. Therefore, Madrid City is an interesting case of study for this project goal.

## Road Map

The data for this project will be obtained from the Inside [Airbnb Initiative](file:///C:\Users\Natalia\Downloads\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 Neighborhood. This data set is available for many destinations around the world. Thus the first decision was choosing 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 from Airbnb website, and some others calculated transforming the scrapped data. So it was crucial to understand 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 an 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 to open their homes to the collaborative economy, for obvious economic reasons. So, the question is easy: ‘How much money makes a specific listing?’ In other words: “What is the estimated income for a single listing?” Scrapped datasets do not have Income Information (that information resides on host’s private area), so it had to be estimated.

Is it better for a listing vacational 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 vacational 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 the main competitors of a particular listing? 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 particular 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 between features and listing’s performance. With the final goal of giving a single recommended price, some algorithms will be run.

### 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 intend.

# 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 On Demand 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. Travelers 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 traveler pays the amount mentioned by host and some additional money as transaction charges.
5. Host approves the booking. Traveler 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 Travelers (Guests)
* Airbnb charges 3% of the booking amount as transaction charges from travelers upon every confirmed booking.

This brief summary lead us to the conclusion that reservations are the key to understand what means to have 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.

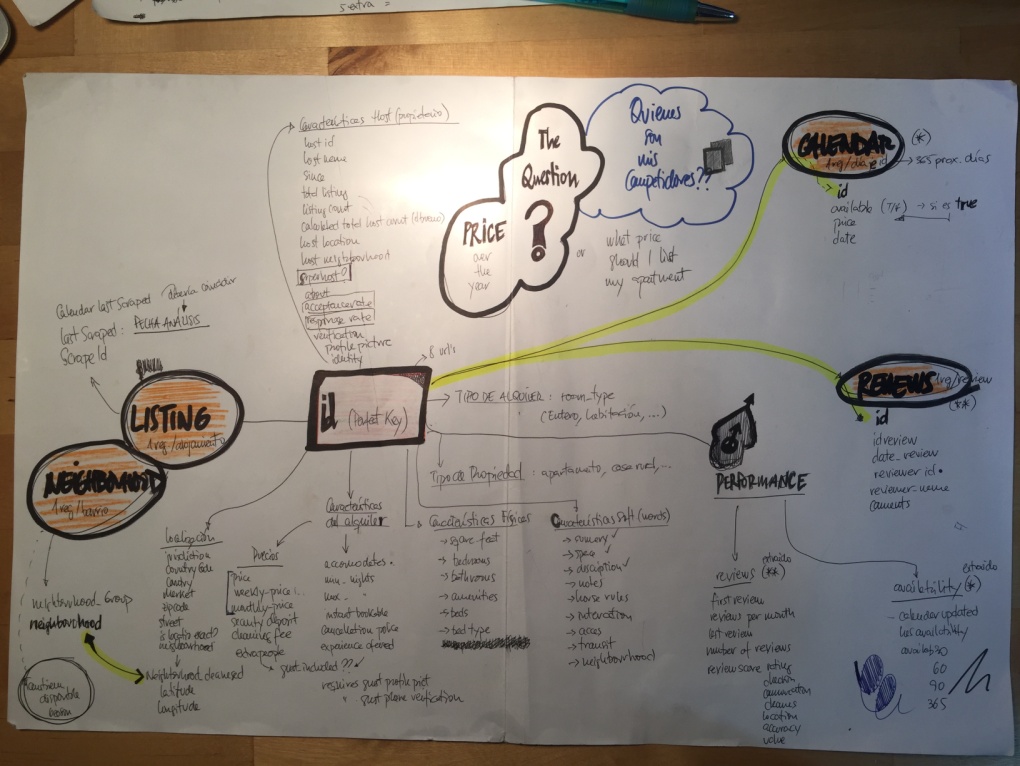


Figure 4 - Project First Draft Data Model - Source: Project Team

*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 to 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 and shared room. So the three of them have to be analyzed separately. Our focus is on the Entire House type of property.
* Income is as simple as room nights 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)
  + Room nights = (Reviews made at Airbnb + Reviews not reported by guests) x average duration of trip (days)
  + The price to consider will be the standard price reported at the listing
  + Average duration per booking = 4.2 days [[4]](#footnote-4)
  + Percentage of bookings with a review = 70%[[5]](#footnote-5)
  + Listings published before 2017 [[6]](#footnote-6)
  + Big numbers at room nights are avoided, limiting maximum occupancy at 80% (255 days/year) [[7]](#footnote-7)

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 a 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

We thought that Data Science algorithms would be key to help us in this step.

In order to find the competitors of each listing, 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 used (especially amenities, as we had 1 Boolean feature per amenity, resulting in around 150 features).

We wanted to try multiple clustering algorithms, but we had the problem that some of the features were categorical, and most algorithms are based on classical distance metrics, like Euclidean, which are not ideal metrics for Boolean features. We 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 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 plotting the clusters too. All Python Notebooks with these algorithms trials can be found in repo folder “Clustering trials”.

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 was able to find listings that we could have considered competitors if looking for them manually, so it was solving one of our problems: automatically find competitors of each listing. We used this in our tool.

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, so we could show per listing what is the maximum price they should put in order to maximize their income, as well as what that income would be. We tested multiple combinations of variables and we tried different algorithms (not only linear models) but we couldn´t get any decent R2 and MSE. We tried reducing the number of variables, joining categories in groups, constructing polynomial features, filtering outliers (we tried robust scaler and then, at a desperated point, just cleaning outliers manually, which obviously ended up in having really few records), and nothing seemed to improve the predictions. The only insights we got when studying these regressions were that the variables that seemed to impact more on the target were the location, reviews score and the people that can accommodate, but we couldn´t get anything that could improve our tool. We tried to understand why this was not working in our case, as we found a couple of other regressions done with Inside Airbnb data for other cities and it seemed to work pretty well, and we thought that it´s probably related with the number of listings we had (in the other cities they had 4 or 5 times more than us) and maybe it could mean too that Madrid is not too price sensitive yet due to not having a really high offer, or maybe it was because we were not considering the real key drivers, as photos and comments. These are just assumptions, but what we got for sure was that we couldn´t get a good regression to predict the number of reviews, so we didn´t use anything of this in the tool. All scripts used in this process can be found in repo folder “Generalized linear models”.

### 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.

# 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:

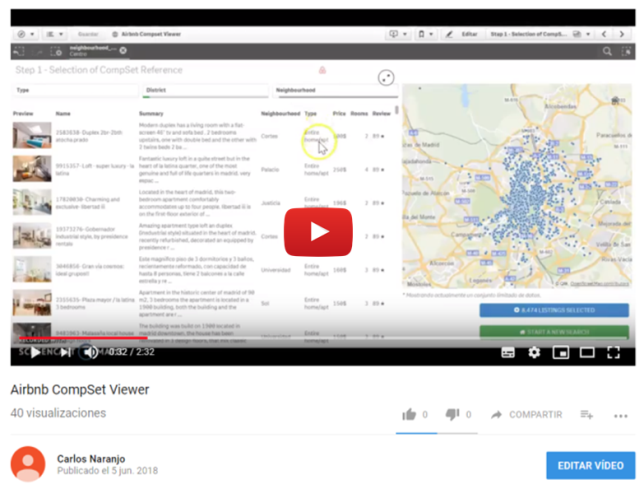
[](https://youtu.be/uUsJcLAzekE)

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

## Starting at the Menu Interface - Screenshot Overview

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

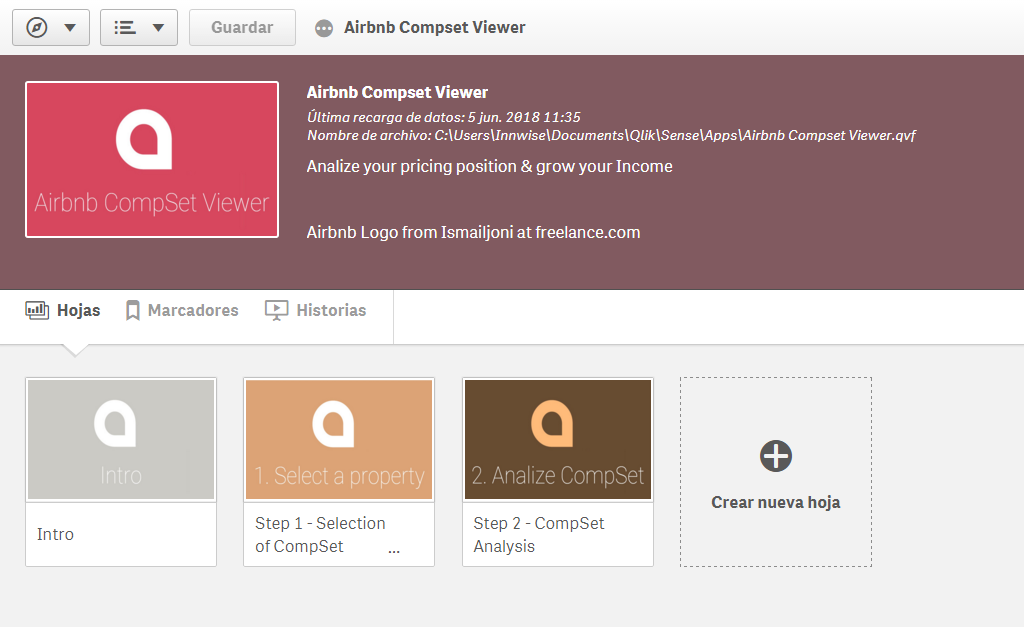


Figure 8 - Airbnb CompSet Viewer home - Source: Project Team

## Intro - Screenshot Overview

INSTRUCTIONS: At the intro you can have an explanation of what is this tool, its basics and the Start Button.

You can start just pressing START TO GO TO STEP 1.

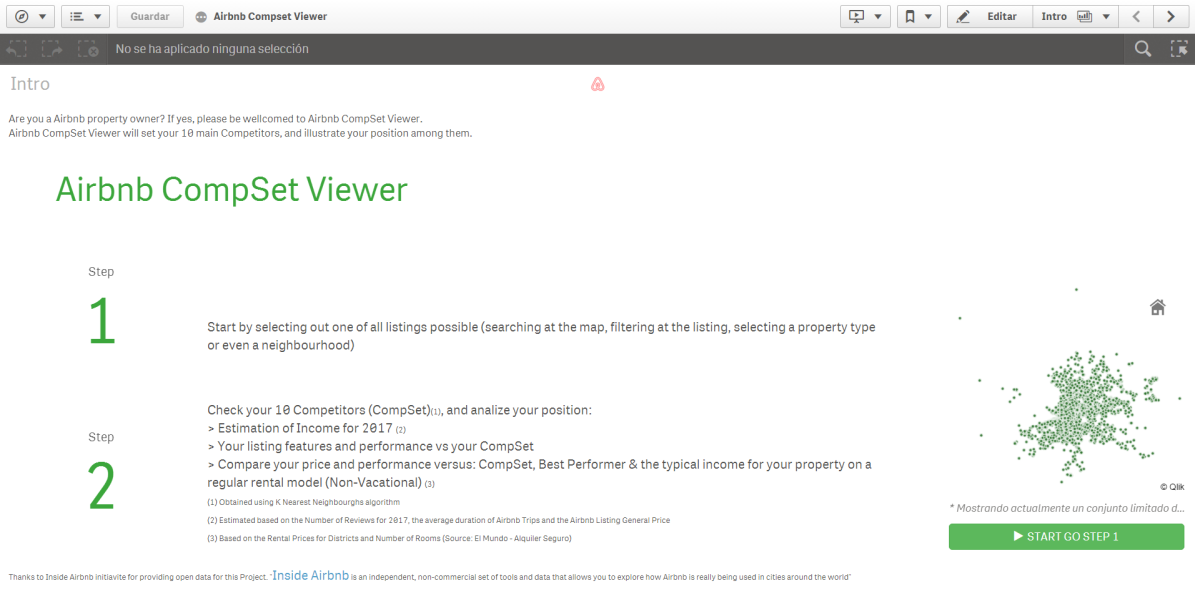


Figure 9 - Airbnb CompSet Viewer Intro - Source: Project Team

## Step 1 - Selection of CompSet- Screenshot Overview

INSTRUCTIONS: At this dashboard you have to select your single listing to compare,

through a list or a map, where you can refine your search.

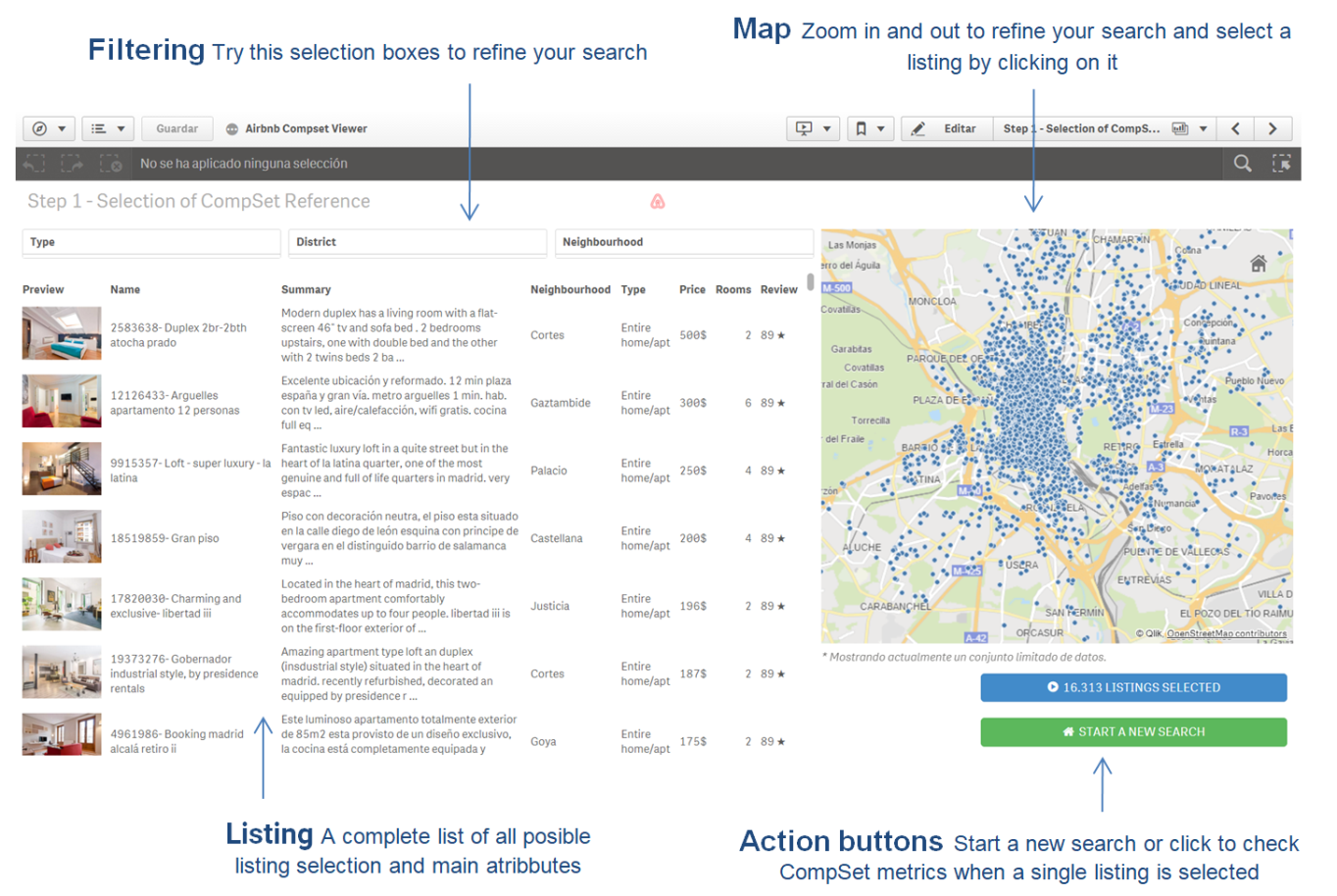


Figure 10 - Airbnb CompSet Viewer Step 1 - Source: Project Team

## Step 2 - Check CompSet Metrics - Screenshot Overview

INSTRUCTIONS: At this dashboard you have to select your single listing to compare,

through a list or a map, where you can refine your search.

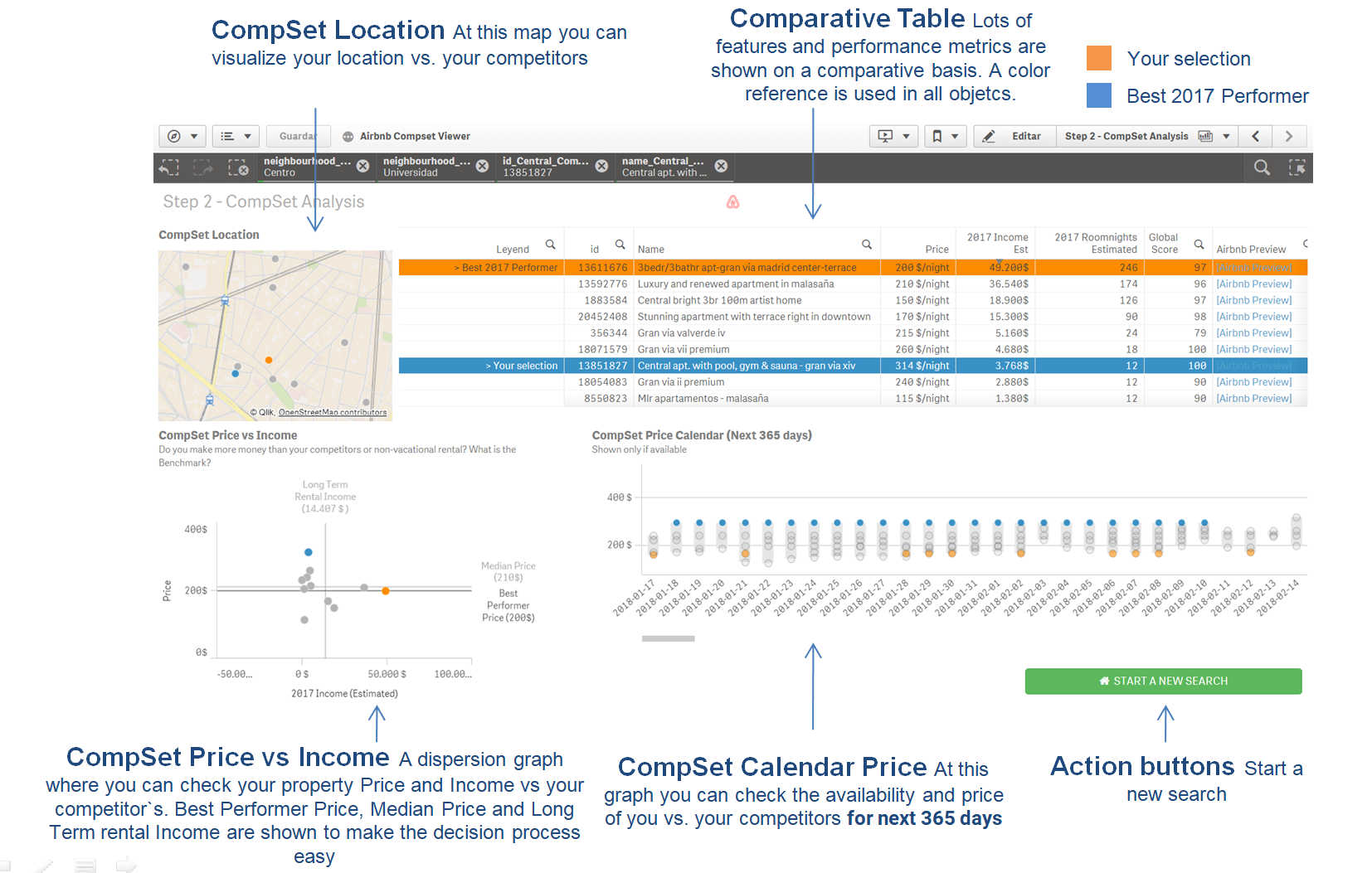
****

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

# Conclusions

Our findings (or the lack of them), is that there is no Airbnb, nor Airbnb Spain, nor AirbnbMadrid listing performance estimation, 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 an 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?
* …

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 does 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.

# Annexes

## 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 these 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. |

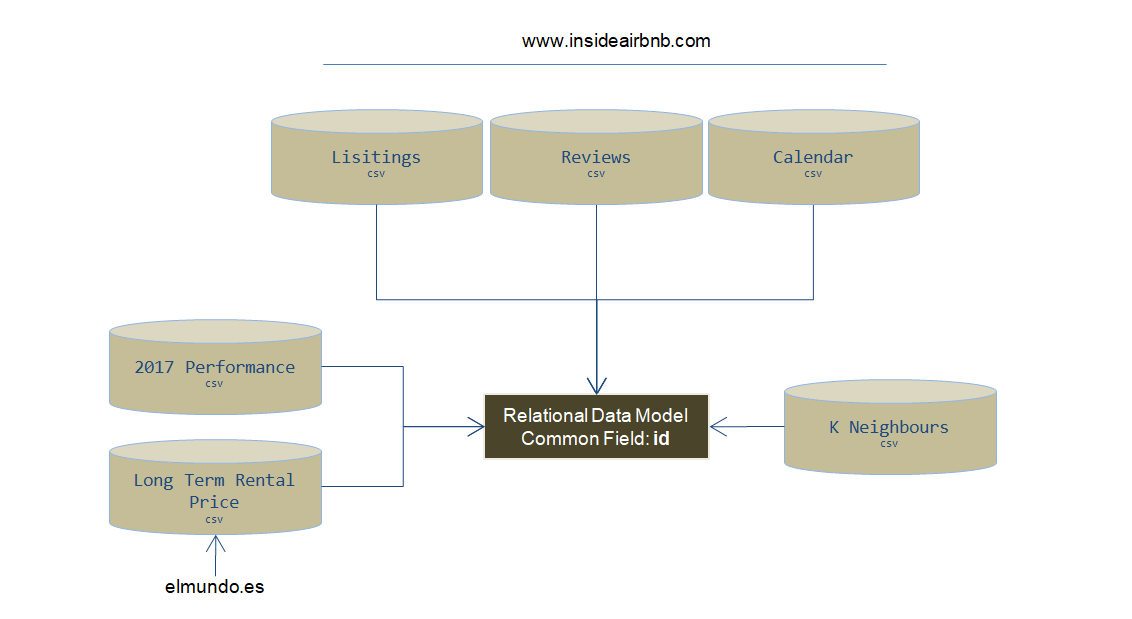


Figure 11-Data Model - Source: Project Team

Another file has been used in this project, it was 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: Fields explanation

Each record is a Listing

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Field** | **Extracted** | **Format** |
| Identifier - Lisiting | **id** | Scrapped | Number |
| ListingDescription | **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: Fields explanation

Each record 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: Fields explanation

Each record 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 |

## LongTermRentalPrice: Fields explanation

Each record 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: Fields explanation

Each record is an id\_Central\_Comparison and an id combination for the 10 K-neighbours. 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: Fields explanation

Each record 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 |

1. https://www.statista.com/statistics/449249/yearly-number-of-international-tourists-visiting-madrid-2001-2014/ [↑](#footnote-ref-1)
2. <https://static.hosteltur.com/web/uploads/2017/04/Spain-Madrid_ActivityReport-Airbnb.pdf> [↑](#footnote-ref-2)
3. Airbnb oficial report 2017 Comunidad de Madrid [↑](#footnote-ref-3)
4. Source: Airbnb Comunidad de Madrid 2016 Report [↑](#footnote-ref-4)
5. Source: Reported by Airbnb CEO [↑](#footnote-ref-5)
6. Calculated at QlikView as all those with a Listing Number lower than: max({1} if (year(first\_review)<year(last\_scraped)-1,id)) [↑](#footnote-ref-6)
7. Source: Team estimation of great occupation rate for the Hospitality Industry [↑](#footnote-ref-7)