# AirBnb Data. Data Mining practice statement 21_22, GIN2. Provisional English version

Group and name of each practice

2021-2022

## Contents



Title:

Authors:


1. Last name, First name
2. Last name, First name
3. Last name, First name
4. Last Name, First Name
5. Last name, First name


# 1 Workshop evaluable in groups AirBnb data

- Here you have the link to this data from AirBnb
- Generate a new project.
- Download the AirBnb data to a folder/directory called `AirBnb` and inside `AirBnb` create a folder/directory called `data`.
- You can (have to) use the workshop help for this data.

## 1.1 Instructions

- Hand out in practice groups.
- Can be done with R or Python.
- Rmd/Notebook must be delivered together with its html/pdf output.
- Maximum length: equivalent to 10 pages in pdf.
- Care must be taken in presentation, spelling and writing.
- Deadline and place of delivery please consult the moodle space of the course.

## 1.2 Context of the data

The website Inside Airbnb http://insideairbnb.com/ contains information about the data on vacation rental apartments or residences in various locations around the world.

The collected data are spread over various regions, provinces, departments, counties... of the world. The data are Open Source and we can use them (see license About Inside Airbnb http://insideairbnb.com/about.html.)

In summary access and data dictionaries and other utilities are accessible from the home page or at the following links:

**Data Resources**

- Get the data
- View Data Dictionary
- Read Data Policies including aligning data availability to the mission, Community Guidelines and policies on Archived and New Data
- Make a Data Request for Archived Data or Data for a new region

Attention!!! The latter service is chargeable for data older than one year.

### 1.2.1 Access to the data

In the link Get the data we can download for each city the following files:

| File Name | Description |
|---|---|
| listings.csv.gz | Detailed Listings data for Name City. |
| calendar.csv.gz | Detailed Calendar Data for listings in Name City. |
| reviews.csv.gz | Detailed Review Data for listings in Name City. |
| listings.csv | Summary information and metrics for listings Name City (good for visualisations). |
| reviews.csv | Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing) N/A Name City. |
| neighbourhoods.csv | Neighbourhood list for geo filter. Sourced from city or open source GIS files N/A Name City. |
| neighborhoods.geojson | GeoJSON file of neighborhoods of the city. |

Define a `data` folder and inside a folder by zone `Mallorca`, `Valencia`, `Barcelona`, etc. Download the data, you can use the program `download_city_inside_airbnb.R` that you can find in the root of the practice github.

### 1.2.2 Specification of the data tables

To understand each data table we need to access the Data Dictionary.

We have to understand what variables we are going to load and the type of data. As there are data of all types we have to go with special attention:

- To long integer ids that can be confused with numeric variables: we have to read them as character strings.
- To the numeric variables that can contain special characters: symbol of $, symbol of €, the symbol of %, separators of thousands...
- Variables that are lists; for example housing extras: wifi, TV, swimming pool,.... .
- Other special types of variables: latitude, longitude, text, etc....

As the problem is a data problem without a clear structure, each group will have to study the zones:

- Mallorca
- Valencia
- Barcelona
- Several more cities until completing (together with the three previous ones) the number of members of the group.

#### 1.2.3 Bibliography and additional software

- Dynamic graphics with plotly: https://plotly.com/r/animations/
- MAPS of Spain: https://www.cienciadedatos.net/documentos/58_mapas_con_r.html fixed

### 1.3 Question 1: Context of the problem and data model (25%)

Load `listing.csv`, `calendar.csv` and `reviews.csv` file for each city. You have to study them and decide which type of data and which variables to load. It is necessary to explain the transformations that you make to manipulate the data; for example 50$ I transform it to 50, "2020-01-30" I read it in type `date`.... . 2. Define a **data model** with all tables. For example join all the listings of your cities in one table, adding a variable specifying the zone: Mallorca, Valencia, Barcelona, CityX, CityY.... 3. Save the data model in `.csv` or `Robj` files for the second part of the practice. 4. Write a report explaining the three previous sections.

### 1.4 Question 2: Exploratory data analysis (EDA). (50%)

In the following questions apply everything we have seen about documentation in EDA: Title of graphs, labels of axes, coloring with information, legends, well presented tables (knitr)....

Calculate the frequency of the number of reviews per apartment. I.e. how many apartments have 1 review, 2 reviews, 3 reviews and so on. Does the frequency of the number of reviews per vacation apartment and the range of reviews follow a "power law" (potential relationship)? 1. From the number of reviews per area, neighborhood, day of the week and per month, perform a comprehensive descriptive and graphical analysis. The correct choice of statistics and corresponding graphs will be positively valued. 1. From each city select the 5 zones/neighborhoods with more vacation apartments. From these zones and for each city compare the average prices (for the whole period), the number of rooms and the number of beds for each apartment. 1. For each city calculate and compare (analytically and graphically) the time series of average, maximum and minimum prices per day, week and month (for the whole period). For each city graphically represent a map showing the number of apartments and their average price during the Christmas period (December 23rd to January 2nd), by latitude and longitude or grouped by squares. Select the most informative graphical representation.

### 1.5 Question 3: Final presentation. (25%)

Final presentation:

- 15 minute presentation types transparencies: ioslide, powerpoint.... (remember that ioslides with Rmd have output to pdf, html and ppt).
- All members of the group must present and the order of presentation will be drawn by lot.
- Two notes may be given, one for the overall presentation and one for the individual presentation.
- Extra question: interactive panel +20%.

You can get extra points by making an interactive panel with any of the libraries or programs seen in the master: shiny, graphana, Power BI, tableau.