**Índice**

# **Introducción**

- Qué es? Para qué? Objetivos, motivación

Actualmente, es cada vez más común la economía colaborativa, que consiste en el uso de medios digitales para permitir la interacción entre un prestador de servicios y un usuario final. Su objetivo principal es satisfacer una necesidad. Algunos de sus principales ejemplos son AirBNB, Uber, Car2Go o Glovo.

# **Contexto**

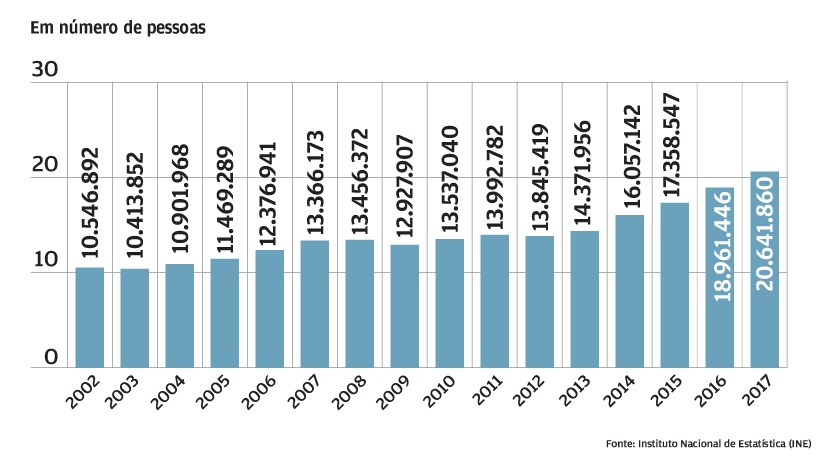
El fenómeno de la **economía colaborativa** *(sharing economy)* ha crecido en los últimos años y deriva de *“desarrollos tecnológicos que han simplificado la posibilidad de compartir bienes y servicios tangibles e intangibles a través de la disponibilidad de varios sistemas de información en Internet”* (Hamari, Sjöklint, & Ukkonen, 2015). Este concepto está estrechamente relacionado con el **consumo colaborativo** *(collaborative consumption)*, que consiste en obtener o dar acceso a bienes y servicios a través de comunidades online (Hamari, Sjöklint, & Ukkonen, 2015). En el caso de los servicios, la etapa posterior a la transacción es importante para captación de valor por parte de los proveedores. De esta forma, variables como *reviews* bidireccionales o respuestas a los usuarios ganan cada vez más importancia en páginas web de economía colaborativa (Zervas, Proserpio, & Byers, 2017).

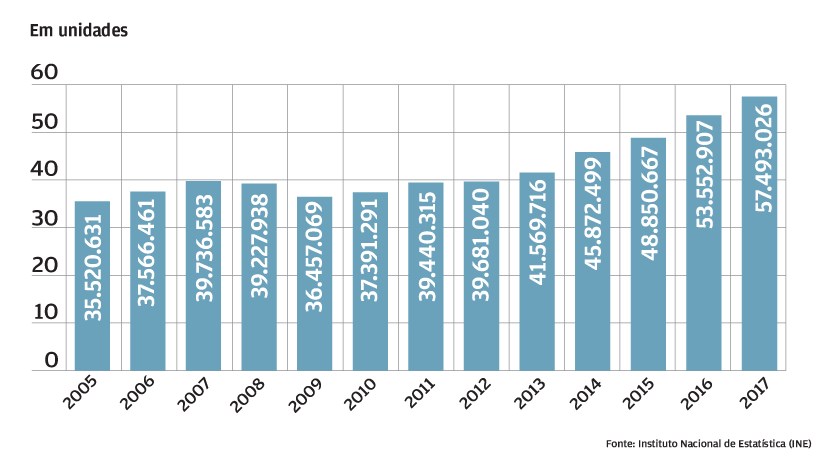
Los principales ejemplos de economía colaborativa son las plataformas Airbnb, para alquiler de alojamiento turístico y Uber, una plataforma de movilidad. En ambos casos, han empezado como startups de Silicon Valley que han tenido un rápido crecimiento y actualmente son multinacionales que operan en ciudades por todo el mundo y se estima su valoración en miles de millones de dólares (Salter, 2012; Hartmans, 2017).

Airbnb

Anti-airbnb

Turismo: Madrid y Lisboa





https://www.jornaldenegocios.pt/empresas/turismo---lazer/detalhe/a-evolucao-do-turismo-em-cinco-graficos

# **Modelo de datos**

El modelo de datos consiste en la extracción, transformación y almacenamiento de los datos en una base de datos.

* 1. **Extracción**

Airbnb no comparte sus datos de forma pública. Tiene una API de acceso limitado para ayudar a los anfitriones a gestionar sus reservas y anuncios y crear flujos de trabajo, pero no permite acceder a información sobre propiedades disponibles por región, propietarios o alquileres. Sin embargo, dentro del movimiento anti Airbnb, ha surgido el proyecto *Inside Airbnb,* creado por el activista australiano Murray Cox, para evaluar el impacto del uso de la plataforma en su país natal. Algunos ayuntamientos del país han incluso utilizado estos datos en vez de los oficiales de la plataforma para analizar el impacto local. El activista dice que el gobierno de Australia ha permitido que la empresa se implantara en la ciudad para pasar una imagen imagen liberal y adaptada a la nueva realidad, pero sin evaluar su impacto en las rentas y en la infraestructura. De esta forma, el proyecto ha sido creado con el objetivo de entender hasta qué punto se está aplicando el concepto de economía compartida, debido a un posible uso de alojamientos que serían de habitación primaria para alquiler exclusivamente turístico, impactando así la vivienda y las comunidades. Airbnb ha criticado al proyecto por información incorrecta y engañosa, pero numerosos estudios, incluidos gobiernos locales, se han basado en estos datos para estimar el impacto de Airbnb en las ciudades (Han, 2018).

Los datos de *Inside Airbnb* se han obtenido a través de web scrapping de la página web pública de Airbnb para un determinado momento en el tiempo. Estos datos han sido verificados, limpios, analizados y agregados por el equipo de *Inside Airbnb.* No se comparte información privada. Toda la información está disponible de forma pública en la página de Airbnb (Inside Airbnb, 2019).

De esta forma, los datos que sirven de base al trabajo se han extraído de esta página para las ciudades de Madrid y Lisboa para el período de XXX. Para cada ciudad y período, se ha extraído el archivo *listings.csv.gz* que contiene los datos sin agrupar, correspondientes a una fecha en específico, con periodicidad mensual. Se han considerado también los archivos *neighbourhoods.geojson* y *neighbourhoods.csv* que permiten geolocalizar los apartamentos en el mapa.

Estos archivos se han importado en un notebook de Jupyter y han sido transformados en Python.

* 1. **Transformación**

Una vez importados los archivos, se ha confirmado la uniformización en el nombre de columnas, todas en minúsculas y sin espacios.

En una primera etapa, se definen las variables no relevantes para el análisis del total de las 106 iniciales. Las variables se han eliminado según los siguientes criterios:

* Variables con campos de texto sin relevancia, no predeterminados por Airbnb y, por lo tanto, sin un formato establecido. Según este criterio, se han eliminado 6 variables: *host\_name, summary, space, neighbourhood\_overview, notes, transit, access, interaction, house\_rules*
* Variables redundantes, cuya información se puede obtener con otras variables. Según este criterio, se han eliminado 11 variables: *scrape\_id, calendar\_last\_scraped, jurisdiction\_names, country, country\_code, street, market, smart\_location, state, zipcode, city*
* Variables que corresponden a url’s no relevantes para el análisis. Según este criterio, se han eliminado 6 variables: *thumbnail\_url, picture\_url, medium\_url, xl\_picture\_url, host\_thumbnail\_url, host\_picture\_url*
* Variables cuyo método de cálculo se desconoce o cuyo significado no se comprende. Según este criterio, se han eliminado 9 variables: *is\_location\_exact, maximum\_maximum\_nights, maximum\_minimum\_nights, mínimum\_minimum\_nights, minimum\_maximum\_nights, is\_business\_travel\_ready, hsot\_verifications, host\_identity\_verified*
* Variables que hacen referencia al método de scrapping o que están condicionadas por el momento concreto en que se hizo. Según este criterio, se han eliminado 4 variables: *maximum\_nights, minimum\_nights, calendar\_updated, has\_availability*
* Variables cuya información no es necesaria para el análisis. Según este criterio, se han eliminado 2 variables: *require\_guest\_profile\_picture, require\_guest\_phone\_verification*.

Después de este proceso, sigue habiendo un total de 68 variables. Para un análisis fiable y comparable entre ciudades y períodos de tiempo, eliminamos las variables con un porcentaje de nulos superior al 30%, ya que no serán tan útiles con tanta información perdida. De esta forma, se han eliminado YY variables más:

Teniendo ya el número final de variables, se convierten las que quedan para un análisis más eficiente:

* Se cambia de formato texto a formato fecha las siguientes variables: *last\_scraped, host\_since, first\_review, last\_review*
* Se reemplazan valores de *true* y *false* por 1 y 0 en las siguientes variables: *host\_is\_superhost, requires\_license, instant\_bookable*
* Se cambia a categórica la variable *cancellation\_policy* según las políticas de cancelación de Airbnb (anexo): *flexible, moderate, strict\_14\_with\_grace\_period, super\_strict\_30, super\_strict\_60*
* Se crea una nueva columna con la ciudad a la que se refieren los datos (Lisboa, Madrid)

De esta forma, la base de datos final tiene un total de XX variables.

* 1. **Almacenamiento**

Una vez teniendo el total de variables, se definen las tablas a crear, agrupando las variables por categorías.

|  |  |  |
| --- | --- | --- |
| Variable | tipo | descripción |
| *id* |  | Id de la propiedad |
| *listing\_url* |  | url de la propiedad |
| last\_scrapped |  | Fecha de scrapping de los datos |
| name |  | Nombre del anfitrión |
| description |  | Descripción de la propiedad |
| host\_id |  | Id del anfitrión |
| host\_since |  | Fecha de inicio como anfitrión |
| host\_is\_superhost |  | Anfitrión es o no superhost (anexo) |
| host\_neighbourhood |  | Localización del anfitrión |
| host\_listings\_count |  | Total de propiedades que tiene el huésped en Airbnb |
| host\_total\_listings\_count |  | Total de propiedades que tiene el huésped en Airbnb |
| neighbourhood\_cleansed |  |  |
| neighbourhood\_group\_cleansed |  |  |
| city |  | Ciudad en análisis |
| latitude |  | Latitud de la propiedad (de 0 a 150m de la localización real) |
| longitude |  | Longitud de la propiedad (de 0 a 150m de la localización real) |
| property\_type |  | Tipo de propiedad |
| room\_type |  | Tipo de habitación |
| accommodates |  | Número máximo de huéspedes |
| bathrooms |  | Número de baños (puede ser 0,5 si no incluye ducha) |
| bedrooms |  | Número de habitaciones |
| beds |  | Número de camas |
| bed\_type |  | Tipo de cama |
| amenities |  | Comodidades de la propiedad (predefinidas por Airbnb) |
| price |  | Precio |
| cleaning\_fee |  | Tasa de limpieza |
| guests\_included |  | Número máximo de huéspedes para el precio fijo |
| extra\_people |  | Precio de cada persona extra a guests\_included |
| minimum\_nights\_avg\_ntm |  | ¿? |
| maximum\_nights\_avg\_ntm |  | ¿? |
| availability\_30 |  |  |
| availability\_60 |  |  |
| availability\_90 |  |  |
| availability\_365 |  |  |
| number\_of\_reviews |  | Número total de reviews |
| number\_of\_reviews\_ltm |  | Número de reviews en los últimos 12 meses |
| first\_review |  | Fecha de la primera review |
| last\_review |  | Fecha de la última review |
| review\_scores\_rating |  |  |
| review\_scores\_accuracy |  |  |
| review\_scores\_cleanliness |  |  |
| review\_scores\_checkin |  |  |
| review\_scores\_communication |  |  |
| review\_scores\_location |  |  |
| review\_scores\_value |  |  |
| requires\_license |  |  |
| instant\_bookable |  |  |
| cancellation\_policy |  |  |
| calculated\_host\_listings\_count |  |  |
| calculated\_host\_listings\_count\_entire\_homes |  |  |
| calculated\_host\_listings\_count\_private\_rooms |  |  |
| calculated\_host\_listings\_count\_shared\_rooms |  |  |
| reviews\_per\_month |  |  |

Explicar las variables: precio fijo y variable, reviews y Sistema de reviews, superhost, listings\_count,.

Una limitación de los datos extraídos de *Inside Airbnb* tiene que ver con la dificultad en estimar la tasa de ocupación de las propiedades listadas y los ingresos del anfitrión, para poder calcular la rentabilidad del alquiler. Sin embargo, la propia página sugiere un método de cálculo que usa las reviews como proxy de reservas estimadas. Este se denomina “Modelo de San Francisco”, por basarse en dos modelos creados para la ciudad para cuantificar el impacto de Airbnb a nivel de políticas públicas y de planeamiento urbano en las viviendas. El primer modelo, creado por Alex Marqusee, estima que las reviews representan el 72% de los alquileres, aunque no se considera una fuente fiable. El segundo modelo, del *Budget and Legislative Analyst’s Office,* indica una representatividad del 30,5%, basándose en la comparación de datos públicos con los datos del reporte de Airbnb de 2014. Sin embargo, la página recomienda el uso de un modelo intermedio, que asume una tasa del 50%.

Usando los datos de la media de la estancia y multiplicándolos por los alquileres estimados, se puede obtener la tasa de ocupación.

* An **average length of stay** is configured for each city, and this, multiplied by the**estimated bookings** for each listing over a period gives the **occupancy rate**
  + Where statements have been made about the average length of stay of Airbnb guests for a city, this was used.
  + For example, [Airbnb reported 5.5 nights](http://blog.airbnb.com/economic-impact-airbnb/#san-francisco) as the average length of stay for guests using Airbnb in San Francisco.
  + Where no public statements were made about average stays, a value of **3 nights per booking** was used.
  + If a listing has a **higher minimum nights** value than the average length of stay, the minimum nights value was used instead.
* The **occupancy rate** was **capped at 70%** - a relatively high, but reasonable number for a highly occupied "hotel".
  + This controls for situations where an Airbnb host might change their minimum nights during the high season, without the review data having a chance to catch up; or for a listing with a very high review rate.
  + It also ensures that the occupancy model remains conservative.
* **Number of nights** booked or available per year for the **high availability** and **frequently rented** metrics and filters were generally aligned with a city's short term rental laws designed to **protect residential housing**.
* Location information for listings are anonymized by Airbnb.
  + In practice, this means the location for a listing on the map, or in the data will be from 0-450 feet (150 metres) of the actual address.
  + Listings in the same building are anonymized by Airbnb individually, and therefore may appear "scattered" in the area surrounding the actual address.
* Listings can be deleted in the Airbnb platform. The data presented here is a snapshot of listings available at a particular time. Other snapshots of data from previous dates are available for analysis by request, and a future activity may be to analyse the characteristics of deleted Airbnb listings.
* The Airbnb calendar for a listing does not differentiate between a booked night vs an unavailable night, therefore these bookings have been counted as "unavailable". This serves to understate the Availability metric because popular listings will be "booked" rather than being "blacked out" by a host.
* Some hosts might not keep their calendar updated, or have it highly available even though they live in the entire home/apartment.
  + To ensure you are only seeing listings that are both "highly available" and being booked frequently, use the "Only highly available" filter with the "Only recent and frequently booked" filter to only select listings that have only been rented recently (reviewed in the last 6 months), and are being rented regularly (number of nights per year greater than the threshold for that city).
* Some reviews may be "spam" allowed by Airbnb. Analysis suggests that spam reviews are small and do not affect the statistics.
* Neighbourhood names for each listing are compiled by comparing the listing's geographic coordinates with a city's definition of neighbourhoods. Airbnb neighbourhood names are not used because of their inaccuracies.
* Extracción (fuente de datos): airbnb get the data, web scrapping
* Transformación (tratamiento y validación de datos)
* Almacenamiento

Visualización

* Análisis descriptivo
* Análisis predictivo