# ANÁLISIS AIRBNB MADRID 2012-2016



Integrantes
Cristina Porta
Erica Martínez
Fátima Ramírez
Josselyn Jumpa
Susana Camacho

# Contexto

En el año 2019 AirBnB crea un nuevo proyecto para aumentar su cuota de mercado en capitales europeas.

En la división del sur de Europa, el equipo de analistas **Infinite loop** realiza un estudio exhaustivo de las propiedades listadas ubicadas en la ciudad de Madrid.

El objetivo del proyecto es entender:

- **¿Cómo influyen las reseñas de las viviendas en el precio final de las mismas?**

# Suposición inicial

## Tableau

Nuestra **hipótesis** antes de empezar con el estudio era la siguiente:

Las **reviews influyen directamente en el precio del alquiler, a mejor puntuación de reseñas mayor precio**.

### ¿Cuáles han demostrado ser válidas y cuáles no? ¿Por qué?

Según hemos podido comprobar, **el número de reviews no es determinante para el precio establecido en los apartamentos.** Los apartamentos son el tipo de propiedad predominante en Madrid.

Hemos observado que **la media de la puntuación en general es muy elevada** (por encima de 8,40 puntos sobre 10) para las distintas tipologías de alquiler de apartamentos (entire apartment, private room, shared room.

Sólo hemos encontrado **correlación entre el precio y la puntuación para el alquiler de apartamentos enteros.**

Para las otras dos tipologías de alquiler (Private room y Shared room) no hay una correlación tan visible, presumiblemente debido a la **poca muestra de reviews** que incluye nuestro dataset, al ser opciones de alquiler menos solicitadas por los usuarios de AirBnB.

# Métricas seleccionadas

- Reviews per month
- Precio medio + cleaning fee + security deposit
- Precio medio

## ¿Han sido las correctas o no? ¿Por qué?

Considerando el objetivo de nuestro estudio, las reviews mensuales y el precio sí **son métricas adecuadas** para averiguar la correlación existente entre ellas.

De todos modos, para profundizar más en la relación entre la ocupación de una vivienda y el número de reviews recibidas **lo más apropiado hubiese sido utilizar información sobre el número de reservas.**

En nuestro análisis hemos empleado el número de reviews como equivalente a número de reservas, puesto que no disponíamos en nuestro dataset de información sobre la cantidad de reservas de cada listing.

## Teniendo en cuenta lo aprendido ¿Qué cosas se harían igual y cuales se harían de otra forma? ¿Por qué?

Sin duda, **repetiríamos la estructura** en la que hemos basado nuestro análisis **y la distribución de las tareas.**

Merece especial mención **la primera etapa** de nuestro estudio, que consistió en familiarizarnos con el dataset. De este modo, **comprendimos qué información podíamos extraer, qué queríamos estudiar, y qué datos eran relevantes.**

En cuanto a los puntos que cambiaríamos, consideramos que **la calidad de los resultados mejoraría con el uso de una o más base de datos para aportaran información sobre el número de reservas** de cada listing, la duración de los alquileres, así como información económica del negocio de AirBnB en Madrid.

# Modelo de Regresión Lineal Múltiple

Esta sección ha tenido como objetivo desarrollar un modelo de regresión múltiple que sea capaz de predecir el precio de las propiedades de Airbnb en Madrid. Como resultado del análisis exploratorio, se han utilizado y comparado dos conjuntos de datos:

*[1] Madrid_airbnb_noreview:* consists of 9395 observations and 42 columns. It is the file resulting from removing all nulls and outliers from the original data frame.

*[2] Madrid_airbnb_imputed_reviews_others:* consists of 10181 observations and 42 columns. It is the file resulting from imputing KNN to fill the nulls and remove outliers from the original data frame

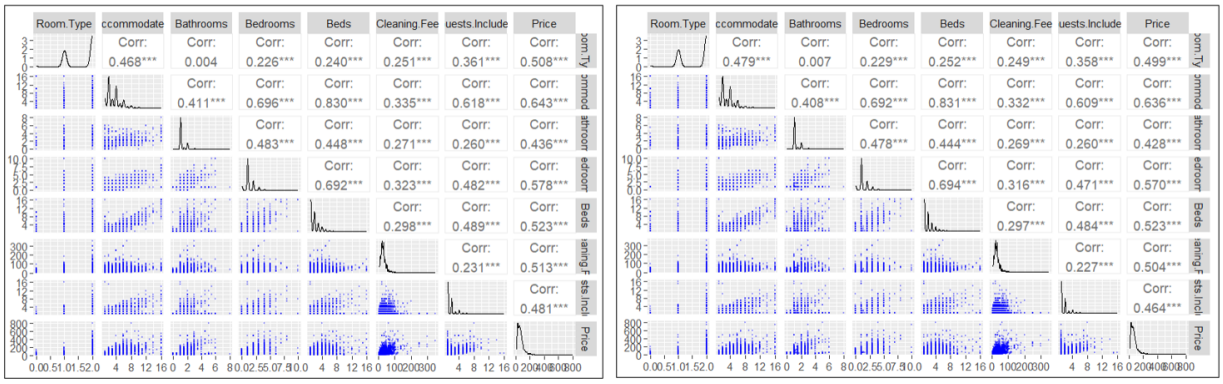
El procedimiento de análisis de datos ha consistido en tres partes que se describen a continuación.

**1. Análisis descriptivo observando la distribución y correlación de Pearson.**

- Ninguna de las variables explicativas sigue una distribución normal. La mayoría de las **distribuciones están muy sesgadas** a la derecha.

- La **relación** entre el precio y las variables independientes es **positiva**, pero existe una relación positiva particularmente fuerte con **alojamiento** (r = 0,643), **habitaciones** (r = 0,578), camas (r = 0,523) y **tarifa de limpieza** (r = 0,513).

**Figura 1. Scatterplot, distribución y coeficiente Pearson**




Dataframe no reviews

DF reviews imputed

**2. Análisis de inferencia utilizando modelos de regresión múltiple.**

La información anterior se ha utilizado como criterio para la evaluación de la regresión.

En esta sección, se dividieron los **datos en train y test** para ambos datasets y los análisis de regresión múltiple se realizaron siguiendo el enfoque de **selección *backwards*** (hacia atrás). Esto significa que el primer análisis de regresión consta de las siete variables explicativas en base a las variables más correlacionadas con el precio (*room type, accommodates, bathrooms, bedrooms, beds, cleaning fee y guest included)*. El resultado del primer modelo se comparó con el análisis del total de variables y después éstas se fueron eliminando gradualmente en función del valor de p. El mejor modelo se elige observando el R-cuadrado más alto y el Error estándar residual (RSE) más pequeño. Las Tablas 1 y 2 muestra los resultados obtenidos de los análisis de regresión.

| Model | R-squared | Adjusted $R^2$ | RSE |
|---|---|---|---|
| Model 1 | 0.62 | 0.619 | 30.88 |
| Model 2 | 0.6383 | 0.6368 | 30.17 |
| Model 3 | 0.6373 | 0.6364 | 30.18 |
| Model 4 | 0.6337 | 0.633 | 30.33 |
| **Model 5** | **0.618** | **0.6175** | **30.96** |
| Model 6 | 0.6147 | 0.6143 | 61.09 |

Tabla 1: Resumen del R cuadrado, R ajustado y RSE *(no reviews)*

| Model | R-squared | Adjusted $R^2$ | RSE |
|---|---|---|---|
| Model 1 | 0.6032 | 0.6026 | 31.63 |
| Model 2 | 0.6199 | 0.6184 | 31 |
| Model 3 | 0.6145 | 0.6139 | 31.18 |
| Model 4 | 0.6093 | 0.6088 | 31.38 |
| **Model 5** | **0.6028** | **0.6024** | **31.64** |
| Model 6 | 0.6 | 0.5996 | 31.75 |

Tabla 2: Resumen del R cuadrado, R ajustado y RSE *(reviews imputed)*

Como se mencionó anteriormente, el modelo 1 incluía las siete variables más relevantes en la correlación. El R-cuadrado ajustado muestra que los modelos 2, 3 y 4 son mejores, pero los valores de p correspondientes a las variables no eran significativos en su totalidad. Siguiendo el enfoque hacia atrás (backwards) se consideró que el modelo 5 es el mejor predictor como se muestra en la Figura 2.

**Figura 2. Regresión Lineal Múltiple**

```
summary(model5)


Call:
lm(formula = Price ~ . - Host.Response.Time - Host.Response.Rate -
    Beds - Bed.Type - Maximum.Nights - Availability.365 - Number.of.Reviews -
    Review.Scores.Accuracy - Review.Scores.Checkin - Review.Scores.Value -
    Cancellation.Policy - Minimum.Nights - Review.Scores.Cleanliness -
    Review.Scores.Communication - Host.Listings.Count - Extra.People -
    Review.Scores.Rating - Review.Scores.Location - Reviews.per.Month,
    data = airbnb.train)

Residuals:
    Min      1Q  Median      3Q     Max
-199.16  -13.99   -1.68   11.00  545.00

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -73.321115   1.922723 -38.134  < 2e-16 ***
Room.Type         32.114103   0.943488  34.038  < 2e-16 ***
Accommodates       3.740438   0.324267  11.535  < 2e-16 ***
Bathrooms         16.500068   0.816830  20.200  < 2e-16 ***
Bedrooms          10.263089   0.678752  15.121  < 2e-16 ***
Security.Deposit   0.066480   0.005248  12.667  < 2e-16 ***
Cleaning.Fee       0.637541   0.023499  27.131  < 2e-16 ***
Guests.Included    3.207081   0.425421   7.539 5.39e-14 ***
Availability.30    0.481120   0.050672   9.495  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.96 on 6567 degrees of freedom
Multiple R-squared:  0.618,     Adjusted R-squared:  0.6175
F-statistic:  1328 on 8 and 6567 DF,  p-value: < 2.2e-16
```

Dataframe no reviews

```
summary(model11)


Call:
lm(formula = Price ~ . - Host.Response.Time - Host.Response.Rate -
    Beds - Bed.Type - Maximum.Nights - Availability.365 - Number.of.Reviews -
    Review.Scores.Accuracy - Review.Scores.Checkin - Review.Scores.Value -
    Cancellation.Policy - Minimum.Nights - Review.Scores.Cleanliness -
    Review.Scores.Communication - Host.Listings.Count - Extra.People -
    Review.Scores.Rating - Review.Scores.Location - Reviews.per.Month -
    Reviews.per.Month, data = airbnb2.train)

Residuals:
    Min      1Q  Median      3Q     Max
-196.94  -14.31   -1.71   10.93  478.41

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -69.730577   1.856997 -37.550  < 2e-16 ***
Room.Type         30.855880   0.910280  33.897  < 2e-16 ***
Accommodates       4.142097   0.311070  13.316  < 2e-16 ***
Bathrooms         15.051778   0.815929  18.447  < 2e-16 ***
Bedrooms          10.768474   0.651580  16.527  < 2e-16 ***
Security.Deposit   0.061057   0.005093  11.988  < 2e-16 ***
Cleaning.Fee       0.625388   0.022167  28.212  < 2e-16 ***
Guests.Included    2.917326   0.412426   7.074 1.65e-12 ***
Availability.30    0.493447   0.047443  10.401  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.64 on 7117 degrees of freedom
Multiple R-squared:  0.6028,    Adjusted R-squared:  0.6024
F-statistic:  1350 on 8 and 7117 DF,  p-value: < 2.2e-16
```

DF reviews imputed

En resumen, a pesar de que el **modelo 5** muestra un R ajustado menor que los anteriores, se trata de un **conjunto más sencillo y óptimo**, por lo que la reducción del valor de R se justifica por la reducción de la complejidad del modelo.

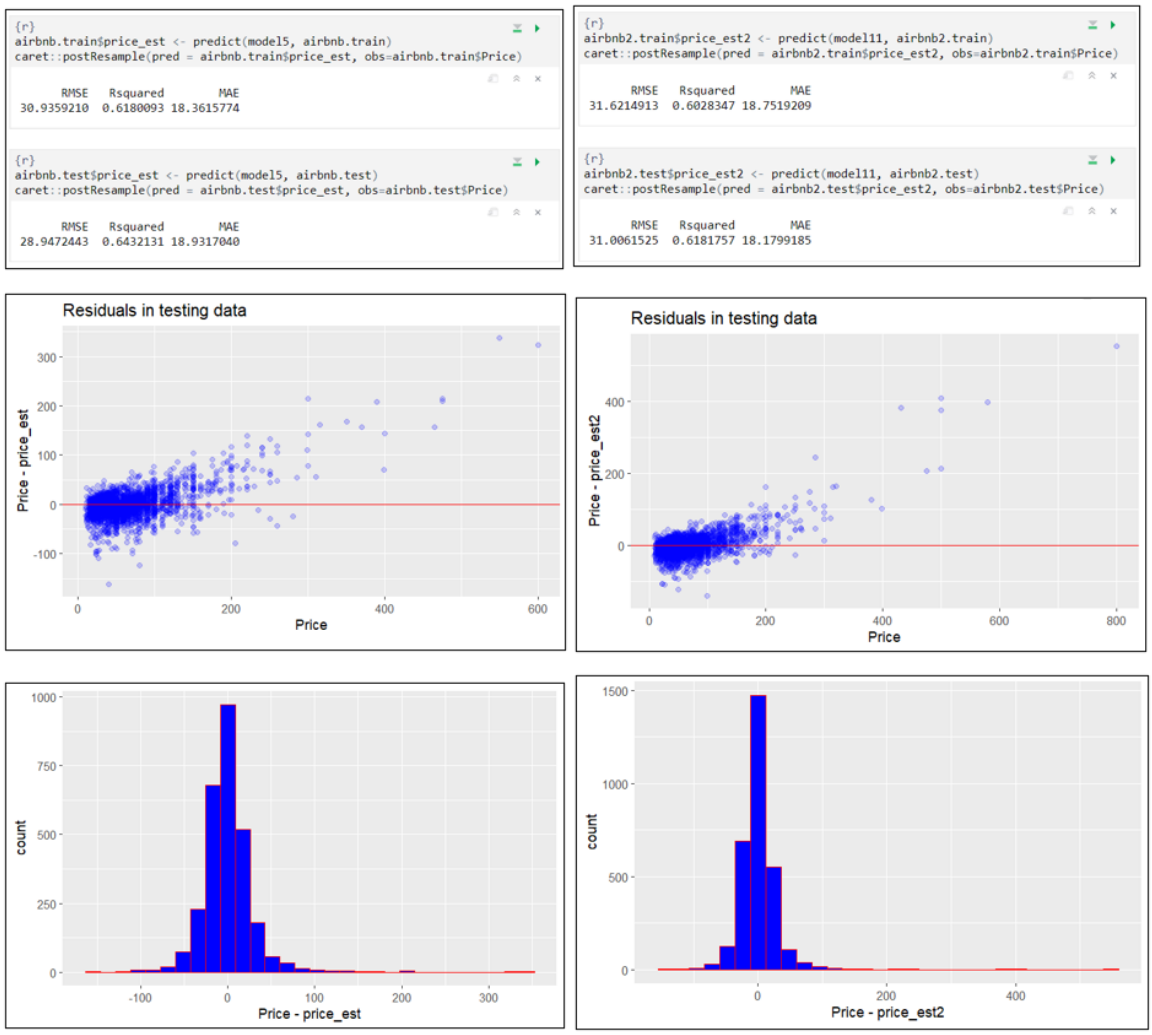
El modelo 5 para el dataframe sin reviews tiene valores más significativos que **explican casi dos tercios de la variación en el precio en las propiedades en Airbnb (R2 = 62%)**. Esto sugiere que room type, accommodates, bathrooms, bedrooms, security deposit, cleaning fee, guests included y availability 30 son útiles para predecir el precio. Por lo tanto, la ecuación de regresión es:

$$\hat{Y} = -73.3 + 32.1(\text{Room.Type}) + 3.7(\text{Accommodates}) + 16.5(\text{Bathrooms}) + 10.2(\text{Bedrooms}) + 0.06(\text{Security.Deposit}) + 0.6(\text{Cleaning.Fee}) + 3.2(\text{Guests.Included}) + 0.48(\text{Availability30})$$

## 3. Comprobación de que el mejor modelo cumple con los supuestos de regresión múltiple para generalizar la muestra a toda la población.

La Figura 3 muestra los resultados del modelo utilizando los datos de prueba. Como podemos ver, el dataset donde no se imputaron los valores NA muestra un R cuadrado más elevado.
Por otro lado, el QQ plot de los residuos muestra en ambos casos que los puntos están muy cerca de la línea recta para un precio entre 0-200€. Sin embargo, se aprecia que hay puntos que sobresalen hacia arriba y hacia la derecha, lo que significa que el modelo nos está dando un precio más alto de lo esperado. En cuanto al histograma, los residuos siguen una distribución normal con una ligera desviación a la derecha.

**Figura 3: Calidad del modelo**




Dataframe no reviews

DF reviews imputed

Como conclusión del modelo de regresión lineal, los valores de **room type, accommodates, bathrooms, bedrooms, security deposit, cleaning fee, guests included y availability30 explican una parte significativa de la variación en el precio de Airbnb en Madrid**. También es importante destacar que **existe una relación positiva entre el precio y estas variables**, lo que significa que, por ejemplo, a mayor capacidad de hospedaje, habitaciones y tarifa de limpieza cabe esperar un mayor precio de la estancia.

# Conclusiones, lessons learned.

Hemos aprendido que para analizar la actividad de una empresa es importante conocer en qué consiste su **negocio**, tener una idea general de sus **operaciones** y comprender en qué **contexto** las está llevando a cabo. Esto permite enmarcar el propósito del estudio.

**Las fuentes de datos** deben elegirse rigurosamente tanto si provienen de fuentes internas o externas a la empresa. Es importante, no sólo centrarse en hacer un análisis de los datos que genera la propia empresa de manera interna sino también explotar todos los datos que se generan en el exterior, como la opinión de los clientes, las RRSS, qué hace la competencia, eventos, etc, para poder realizar un estudio más completo.

**Familiarizarse con la información** que contienen es una de las fases más importantes del proceso. Esta etapa no debe ser precipitada, ya que es esencial para desarrollar las siguientes. Facilita la toma de decisiones en cuanto a **depuración de los datos** y definición de los **kpi's y métricas**.

Finalmente, nos gustaría puntualizar que la realización de un proyecto de análisis requiere de **tiempo** para profundizar en la información que se obtiene de los datos, relacionarla con el objetivo del estudio, y extraer conclusiones lo más precisas posibles.