



PROUESTA
AIRBNB
MADRID

SEABORN GIRLS

CONTENIDOS

- 
- 01 Introducción
 - 02 Planificación
 - 03 Arquitectura de datos
 - 04 Análisis Exploratorio
 - 05 Visualización de métricas
 - 06 Modelado de datos
 - 07 Conclusiones

INTRODUCCIÓN

Como parte del proyecto final del Bootcamp KeepCoding Mujeres-In-Tech Big Data, hemos analizado el dataset **Airbnb Madrid** con datos del 2017.

El objetivo principal es analizar las variables que influyen en el precio de los alojamientos anunciados en Airbnb.

Team Seaborn Girls

Nuestro equipo está conformado por:

Paola Villalba

Cintia Guerrero

Núria Orgaz

Cherry Reynoso

PLANIFICACIÓN

Enfoque de trabajo

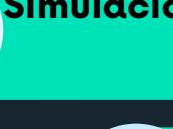
Dada la naturaleza del proyecto y del equipo, para organizarnos utilizamos dentro de las consideradas metodología ágil, la **metodología Kanban**.

El flujo de tareas se distribuyó dentro de un tablero en Trello con tres listas en un "To do", "Doing" y "Done".

The screenshot shows a Trello board titled "Seaborn Girls". The board has three main columns: "To Do", "Doing", and "Done".

- To Do:**
 - Cuarta reunión. Ensayo de presentación de resultados (Due: 12 de feb.)
 - + Añada una tarjeta
- Doing:**
 - 3.b Detección de outliers. Agrupación de elementos para eliminarlos. Imputar valores nulos.
 - Tercera reunión. (Due: 7 de feb.) CG NO
 - 4. Visualización de las métricas (Due: 9 de feb. - 14 de feb.)
 - 5. Pre-procesamiento y Modelado. Regresión Lineal (Due: 9 de feb. - 14 de feb.)
 - 6. Informe (Due: 14 de feb.)
 - + Añada una tarjeta
- Done:**
 - Primera reunión. Acercamiento como equipo. (Due: 28 de ene.)
 - Organización y repartición de tareas.
 - 1.1 Familiarización con el dataset
 - 1.2 Revisar columnas de Host listings para determinar si tienen información relevante. (Due: 31 de ene.)
 - 1.3 Buscar cómo implementar el datawarehouse y ETL (Due: 31 de ene.)
 - Segunda reunión de equipo. (Due: 31 de ene.)
 - Segunda reunión de Equipo. Tutoría en que se aclaró el punto dos "Datawarehouse y Etl"
 - + Añada una tarjeta

CALENDARIO PROYECTO SEA BORN

SÁBADO 28 DE ENERO	PRIMERA SEMANA DE FEBRERO	SEGUNDA SEMANA DE FEBRERO	TERCERA SEMANA DE FEBRERO
	<p> Planificación 28/01 - 14/02</p>		
	<p> Familiarización con el dataset 28/01 - 31/01</p>		
	<p> Validación de datos 31/01- 02/02</p>	<p> Análisis exploratorio 04/02 - 08/02</p>	
		<p> Visualización de Métricas / Modelado 09/02 - 13/02</p>	
		<p> Reporte: suposiciones iniciales. 12/02 - 12/02</p>	
		<p> Finalización de reporte 12/02 - 14/02</p>	
		<p> Simulación de presentación 14/02 - 14/02</p>	
			<p> Presentación 15/02 - 15/02</p>

ARQUITECTURA Y VALIDACIÓN DE DATOS

1

Primera depuración del Dataset, analizando la importancia de cada una de las columnas para valorar si van a ser útiles a futuro en el resto de los Hitos del proyecto.

2

Utilización de SQL con la herramienta Postgree para corroborar que columnas contienen datos vacíos o NAs

3

Tras este primer filtro , pasamos el csv modificado para un análisis mas profundo en el siguiente punto.

3. ANÁLISIS EXPLORATORIO

Importar la matriz de datos y comprobar filas y columnas

Comprobar nombre de columnas: ID... → ID

Revisión y ajuste del tipo de variable

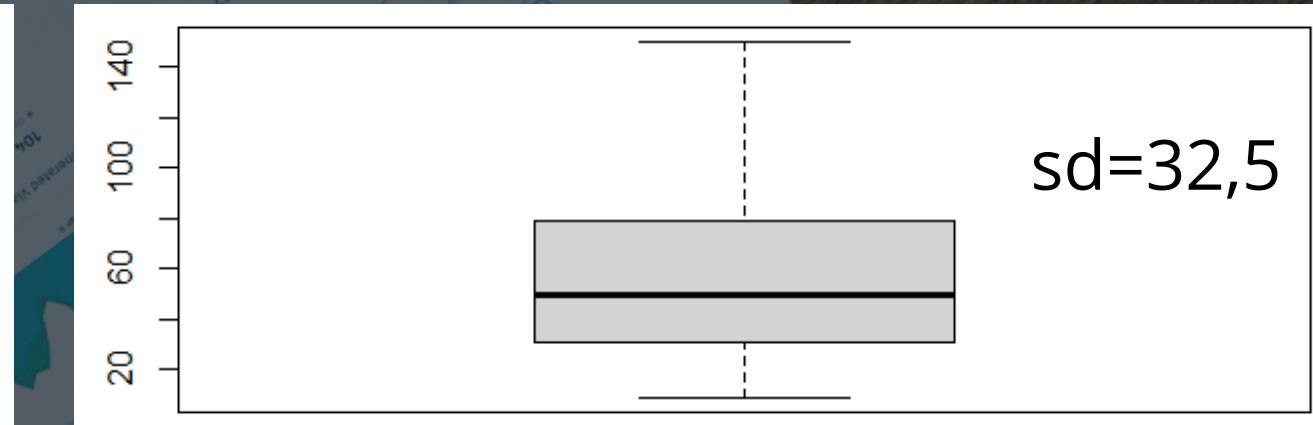
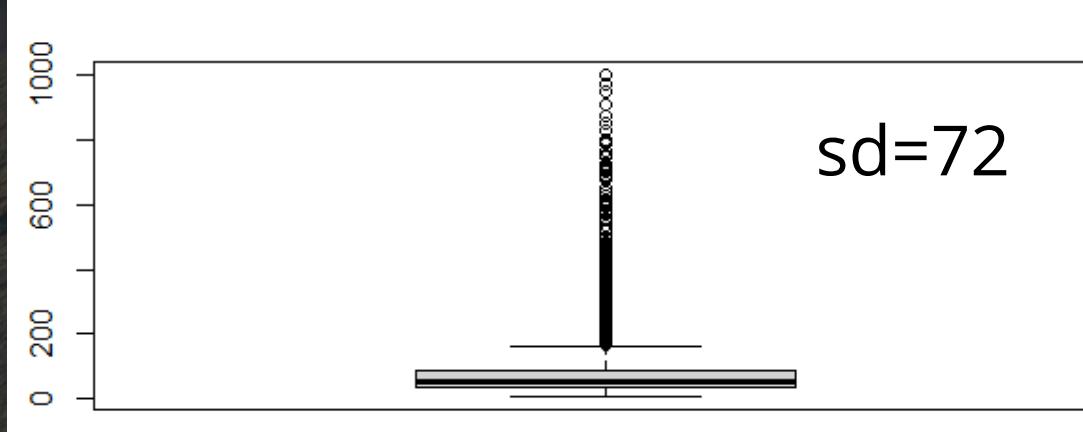
Variable ID → Comprobar que no hay duplicados

Variables numéricas:

- Ver si hay NA's para imputarlos o eliminarlos (ordenada la variable previamente)
- Summary: min, max, cuartiles, mediana y media
- Desviación estandar para ver si hay outliers
- Gràfico de cajas para ver la distribución

Variable dependiente Price:

14.780 → 13.683 filas



Se han generado dos nuevas variables booleanas de las variables Cleaning fee y Security Deposit

3. ANÁLISIS EXPLORATORIO

Variable tipo fecha

- Comprobar NA's (2) y eliminarlos 13.683 → 13.681 filas
- Crear una nueva variable que contiene solo el año

Variables factor

- Ver si hay NA's para imputarlos o eliminarlos (ordenada la variable previamente)
- Tablas de contingencia
- Análisis de las categorías, normalización y supresión de categorías vacías
- Agrupar categorías si es necesario
- Transformar en nuevas variables numéricas

Country

City

→ Country_Rec

→ City_Rec

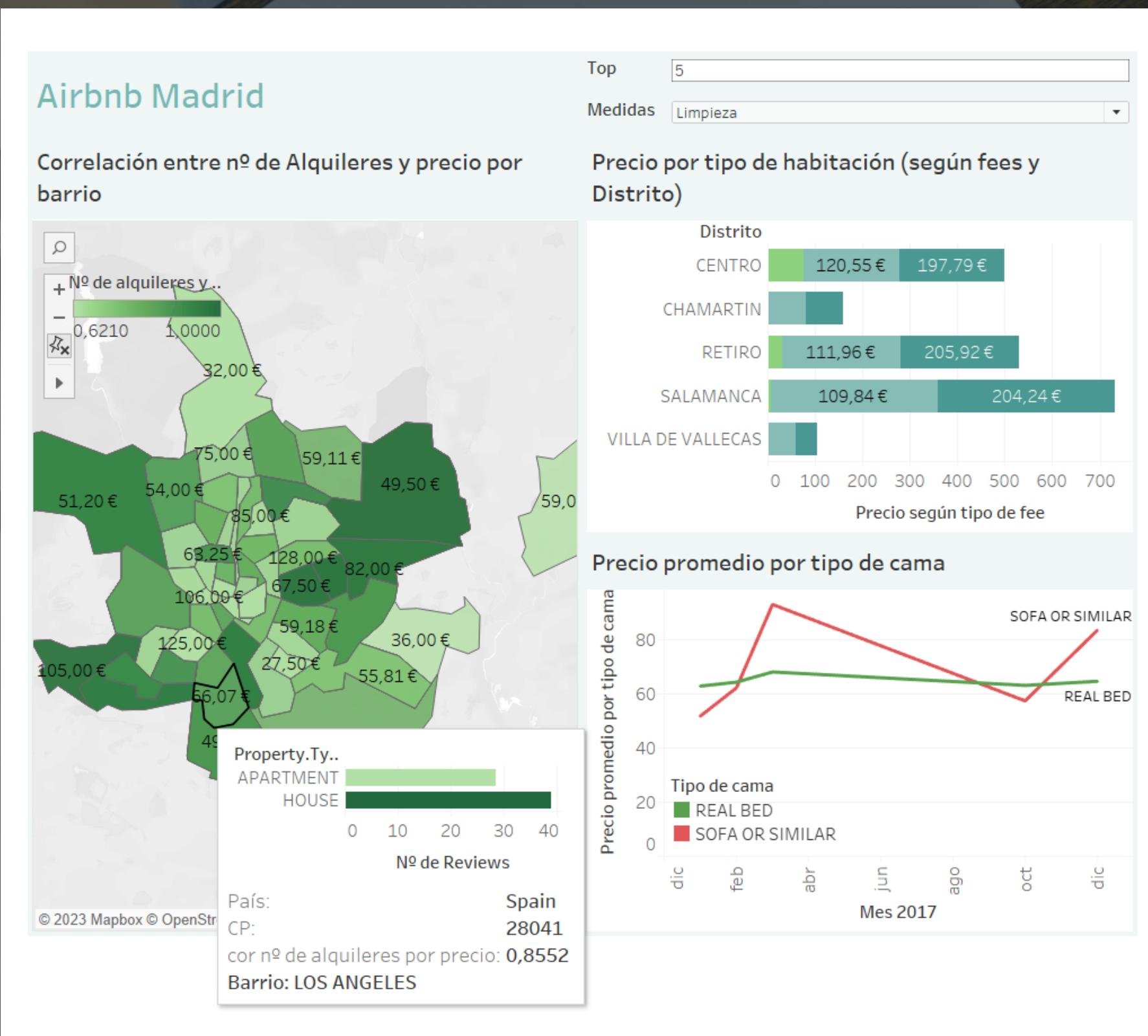
13.681

13.057

→ 13.057 filas

→ 12.375 filas

VISUALIZACIÓN DE LAS MÉTRICAS

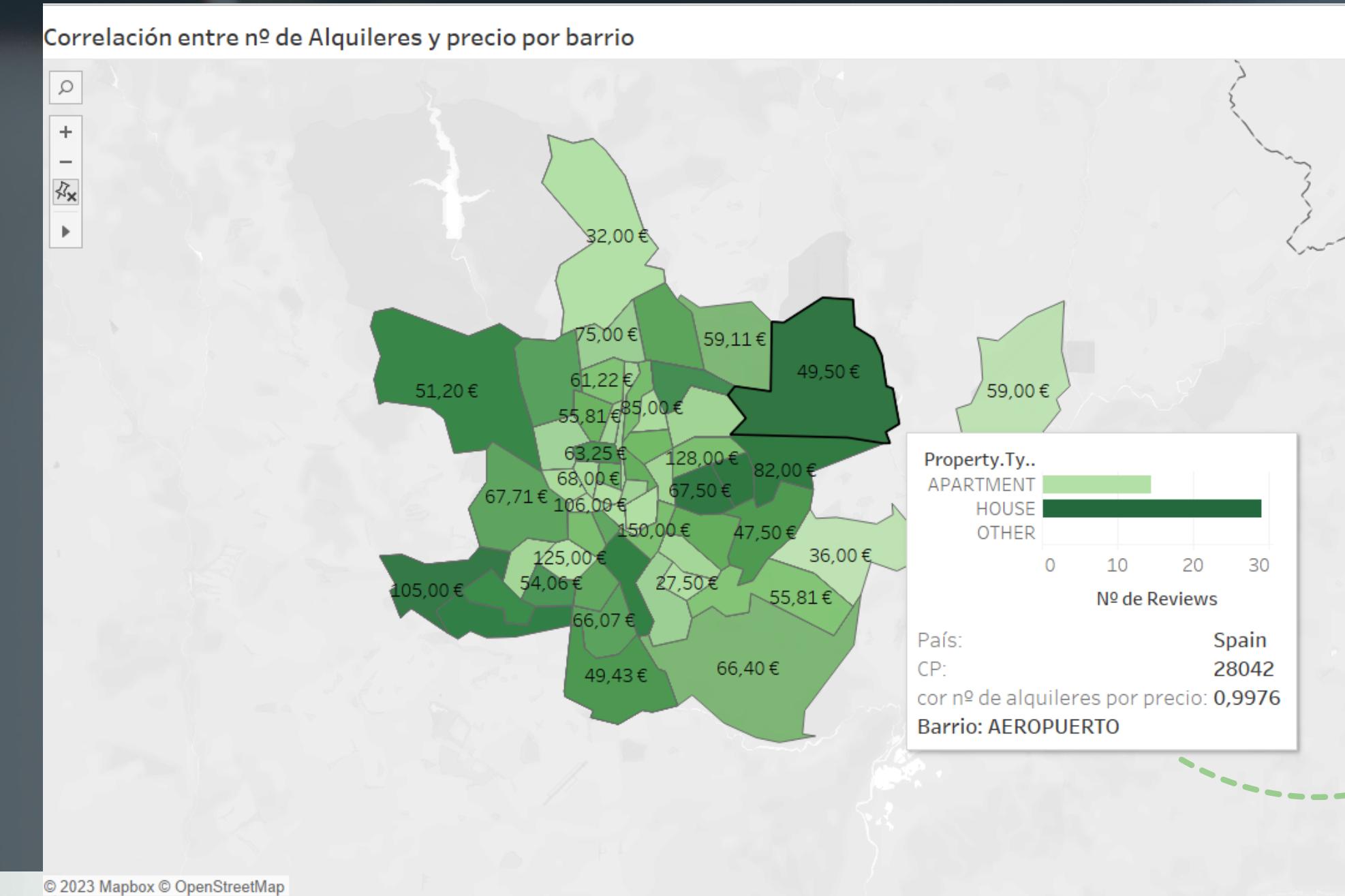


Variables con mayor correlación:

- Host Since
- Accomodates,
- Number of Reviews,
- Beds
- Calculated Host Listing account.

- Relación entre el precio y el número de alquileres por Host.
- Relación del número de reviews con el número de alojamientos.
- Distritos con los precios más altos incluyendo gastos extras.
- Evolución del precio por tipo de cama

MAPA .



- Relación entre el precio y el número de alquileres por Host por distrito.

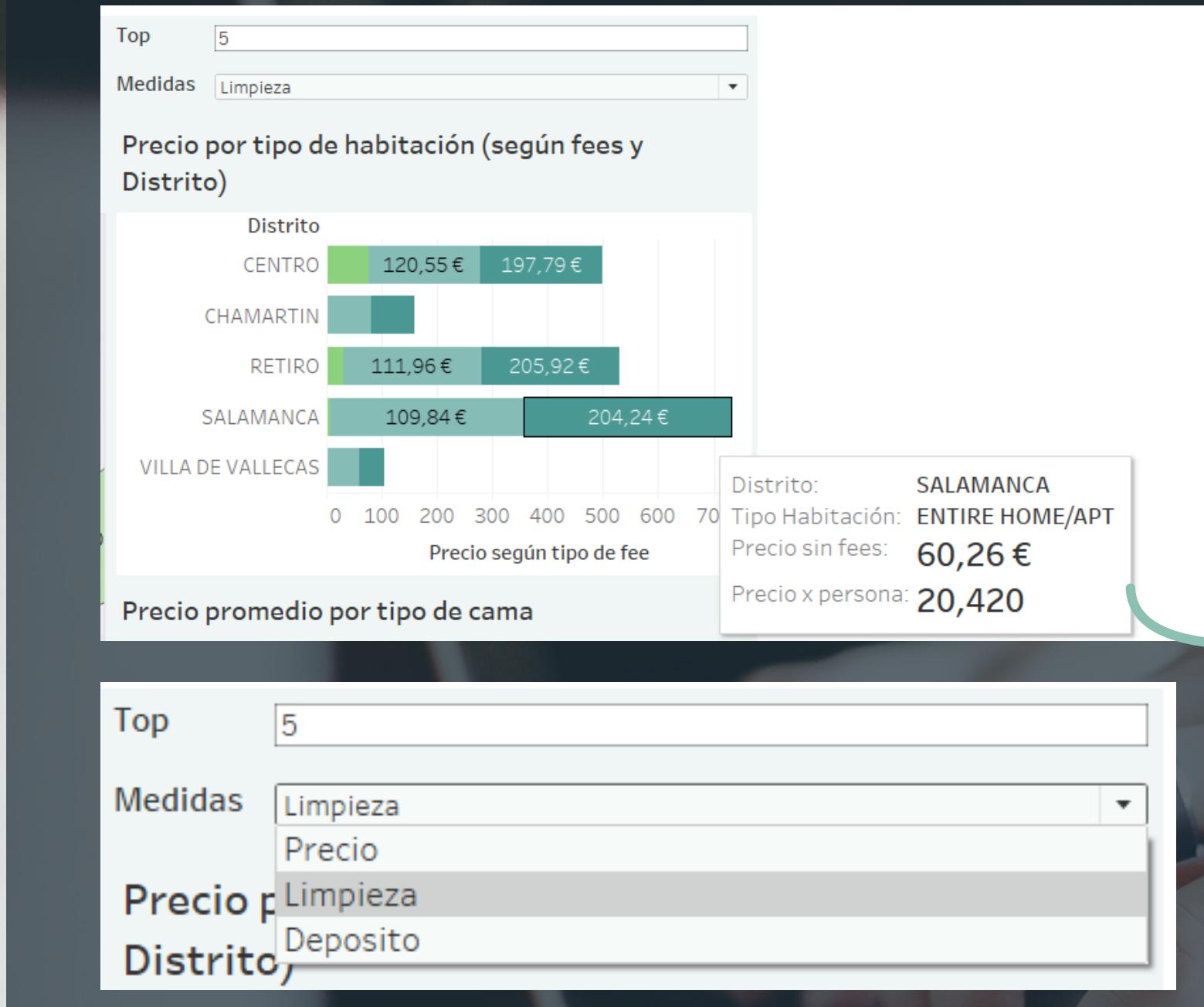
Los host con mayor número de alojamientos conocerán mejor el mercado, y por ende puede que determinen la homogeneidad del precio.

```
(CORR([Price], [Host Listing Count]))
```

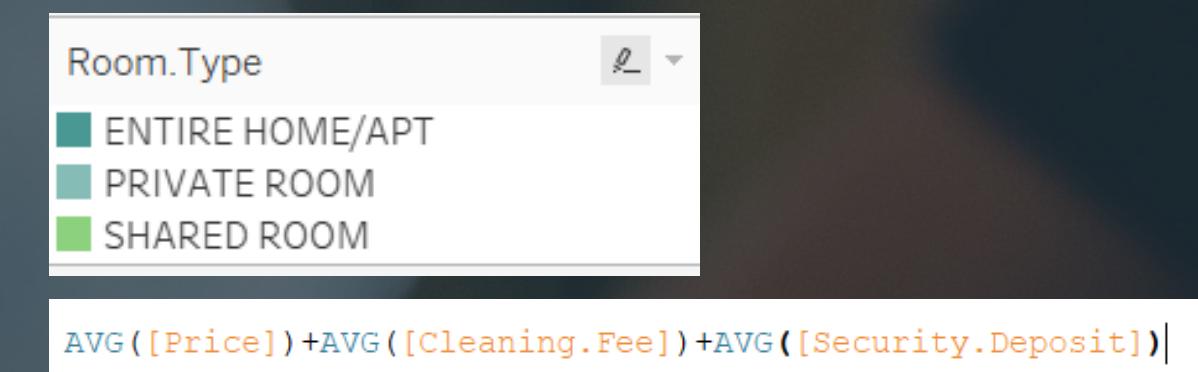
- Relación con número de reviews y host rating.

El rating del host aumenta: a mayor número de alojamientos, mayor número reviews y score. Esto se traduce en un mayor número de veces alquilado.

GRÁFICA DE BARRAS .



- Distritos con los precios más altos incluyendo gastos extras (limpieza, depósito) según el tipo de habitación



- Diferencia de precio sin gastos extras y precio por persona
Se contrastó la diferencia el precio total (gastos extras) y el precio por persona.

$\text{AVG}([\text{Price}]) / \text{AVG}([\text{Accommodates}])$

PREPROCESAMIENTO Y MODELADO

1

SELECCIÓN DE LAS VARIABLES

Qué variables son útiles y cuáles no para predecir el precio

2

TRANSFORMACIÓN DE VARIABLES

Transformar las variables categóricas en dummies

3

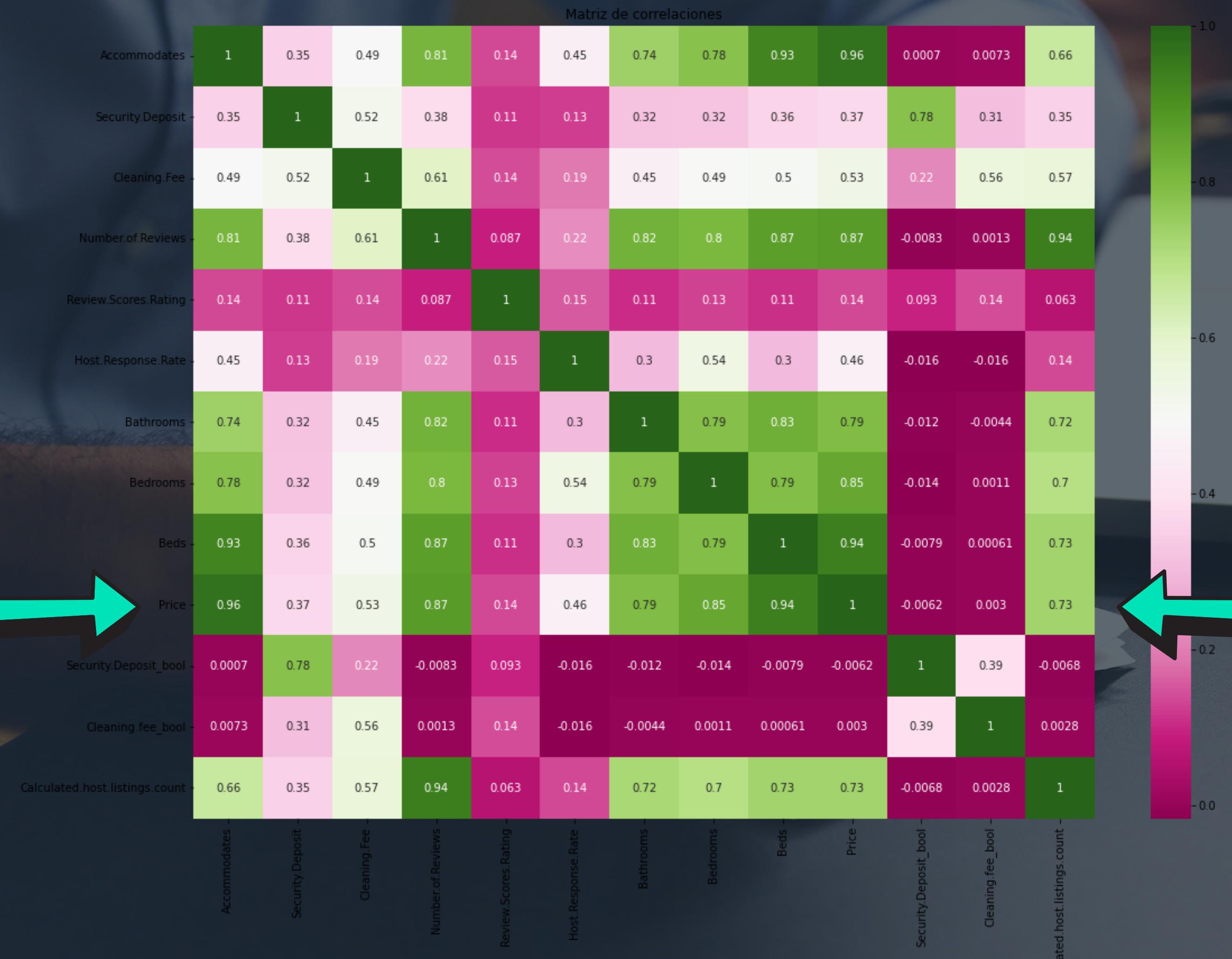
DIVISIÓN TEST Y TRAIN

El 70% del dataset se usa para entrenar el modelo y el 30% para comprobarlo

4

COMPROBAR MODELO

Se usan las métricas R^2 y MSE (error cuadrático medio)



MODELO DE REGRESIÓN LINEAL

	R ²	MSE
Regresión Lineal	0,967	33,82
Regresión Lineal con GridSearch	-34,498	33,82

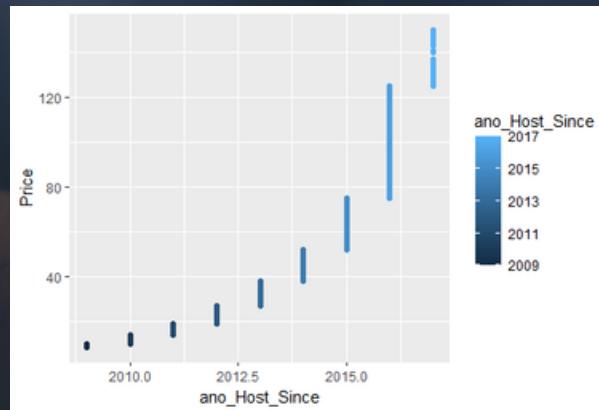
COMPARACIÓN DE MODELOS

	R ²	MSE
Regresión Lineal	0,967	33,82
Regresión Lineal con GridSearch	-34,498	33,82
Random Forest	0,9997	0,29

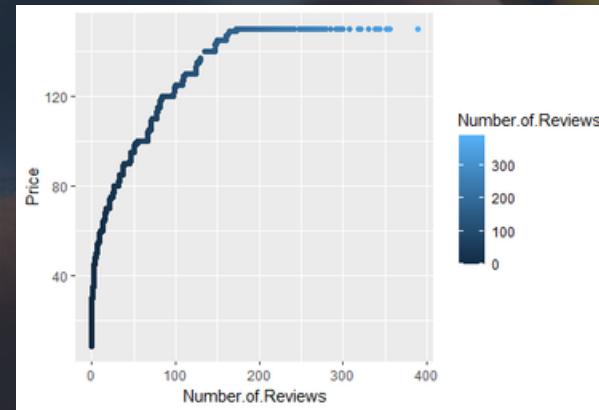
CONCLUSIONES

Análisis de correlaciones

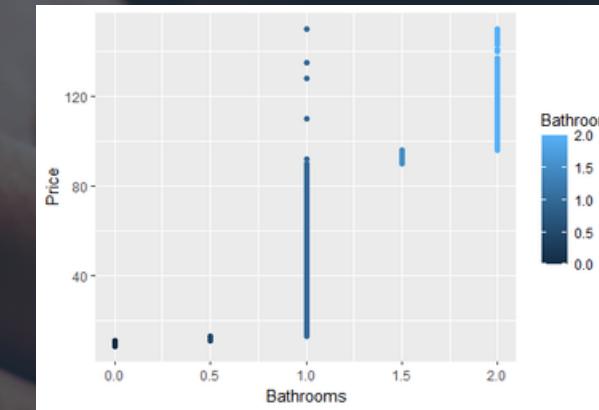
- Año Host Since (0,92)



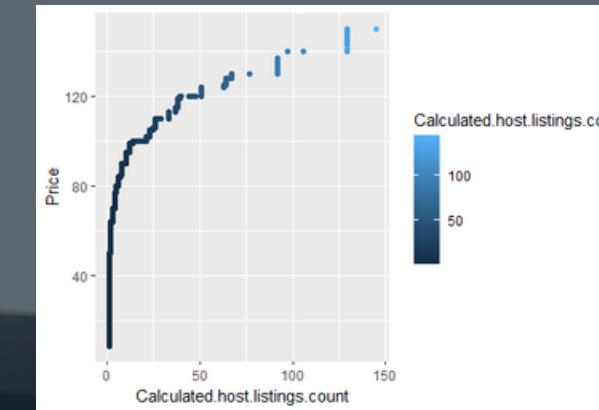
- Number of reviews (0,87)



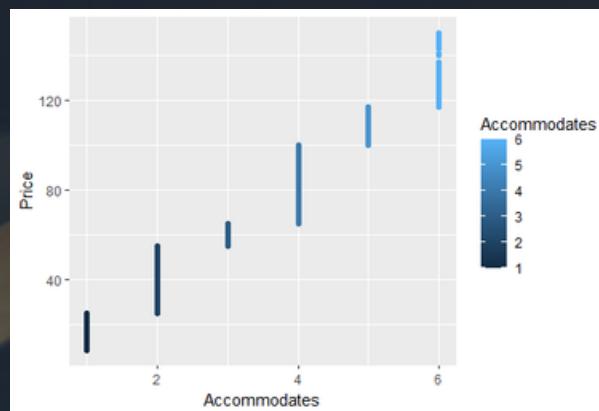
- Bedrooms (0,84)



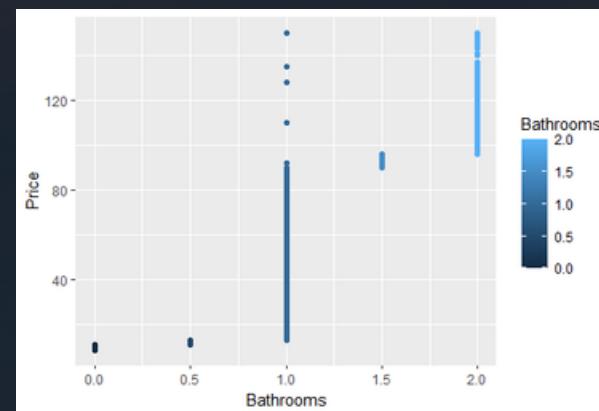
- Calculated.host.listings.count (0,73)



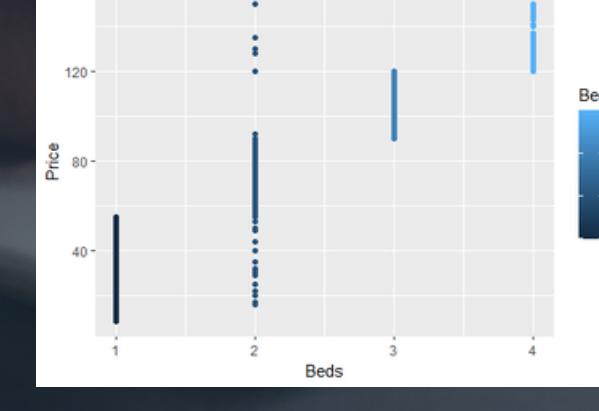
- Accomodates (0,96)



- Bathrooms (0,79)



- Beds (0,93)



LESSONS LEARNED

- 01 Volver a usar metodología ágil. Nos ha ayudado a organizarnos mejor.
- 02 Mejora técnicamente el código. Dado el límite de tiempo, no se podía depurar tan bien.
- 03 Mejor un gráfico más sencillo que "atiborrado"
- 04 En el futuro hacer un sentiment analysis usando NLP con las variables con texto





SEABORN GIRLS
GRACIAS