PROYECTO DATA: Análisis AIRBNB

GRUPO 15

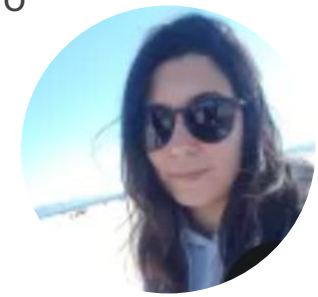
INTRODUCCIÓN

GRUPO 15

Sandra Moreno

Jessica Piñas

Carmen Santaló



Nerea Castañares

Olga Marín

OBJETIVO DEL PROYECTO

¿Qué variables influyen más en los precios de alquiler de un alojamiento en Madrid?

Partiendo como hipótesis que las siguientes variables son las que más influyen:

- La ubicación y el número de habitaciones
- Barrios con rentas más altas
- Número de reviews positivas

Metodología

FASE 1

Recopilación de datos

FASE 2

DW+ETL

FASE 3

Análisis Exploratorio

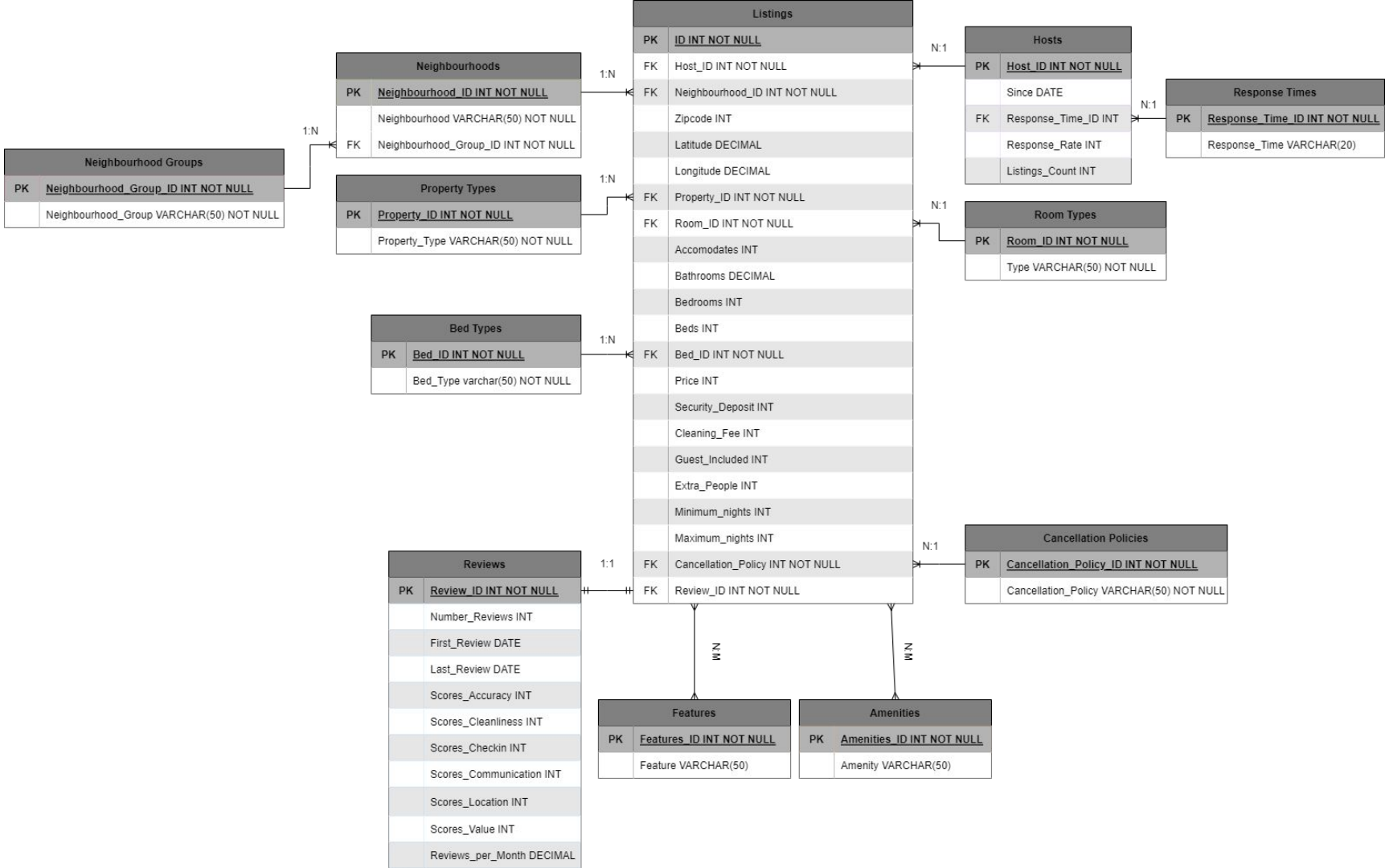
FASE 4

Visualización

FASE 5

Modelo de regresión Lineal

DATA WAREHOUSE


Listings
PK ID INT NOT NULL
FK Host_ID INT NOT NULL
FK Neighbourhood_ID INT NOT NULL
Zipcode INT
Latitude DECIMAL
Longitude DECIMAL
FK Property_ID INT NOT NULL
FK Room_ID INT NOT NULL
Accomodates INT
Bathrooms DECIMAL
Bedrooms INT
Beds INT
FK Bed_ID INT NOT NULL
Price INT
Security_Deposit INT
Cleaning_Fee INT
Guest_Included INT
Extra_People INT
Minimum_nights INT
Maximum_nights INT
FK Cancellation_Policy INT NOT NULL
FK Review_ID INT NOT NULL
Neighbourhoods
PK Neighbourhood_ID INT NOT NULL
Neighbourhood VARCHAR(50) NOT NULL
FK Neighbourhood_Group_ID INT NOT NULL
Neighbourhood Groups
PK Neighbourhood_Group_ID INT NOT NULL
Neighbourhood_Group VARCHAR(50) NOT NULL
Property Types
PK Property_ID INT NOT NULL
Property_Type VARCHAR(50) NOT NULL
Bed Types
PK Bed_ID INT NOT NULL
Bed_Type varchar(50) NOT NULL
Hosts
PK Host_ID INT NOT NULL
Since DATE
FK Response_Time_ID INT
Response_Rate INT
Listings_Count INT
Response Times
PK Response_Time_ID INT NOT NULL
Response_Time VARCHAR(20)
Room Types
PK Room_ID INT NOT NULL
Type VARCHAR(50) NOT NULL
Cancellation Policies
PK Cancellation_Policy_ID INT NOT NULL
Cancellation_Policy VARCHAR(50) NOT NULL
Reviews
PK Review_ID INT NOT NULL
Number_Reviews INT
First_Review DATE
Last_Review DATE
Scores_Accuracy INT
Scores_Cleanliness INT
Scores_Checkin INT
Scores_Communication INT
Scores_Location INT
Scores_Value INT
Reviews_per_Month DECIMAL
Features
PK Features_ID INT NOT NULL
Feature VARCHAR(50)
Amenities
PK Amenities_ID INT NOT NULL
Amenity VARCHAR(50)
1:N
N:1
1:1
N:M

ETL

Limpieza, Retos y Automatización

ETL: Limpieza

1. Filtro por ‘Comunidad de Madrid’
2. Códigos postal incorrectos
3. Nulos
4. Tipo de datos

ETL: Retos

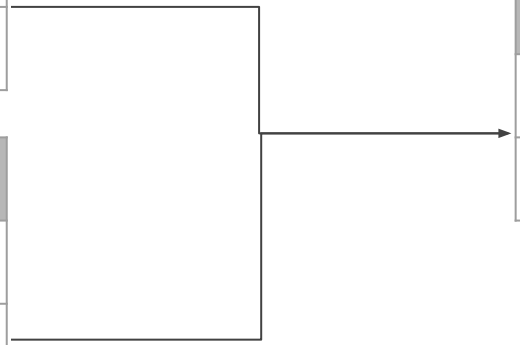
1. Normalización
2. Columnas ‘Amenities’ y ‘Features’
    a. Ex: “TV, Wireless Internet, Kitchen, Breakfast, Heating, First aid kit, Essentials, Shampoo, Hangers”
3. Dos relaciones N:N

ETL: Normalización

.loc

Index	Feature
1	Wireless Internet
2	Kitchen

Index	Feature
1	TV
2	Wireless Internet
3	Kitchen



Index	Feature
1	2
2	3

ETL: ‘Amenities’ y ‘Features’

listas

Index	Feature
1	TV, Wireless Internet, Kitchen
2	TV, Garden, Pool

Index	Feature
1	TV
2	Wireless Internet
3	Kitchen
4	Garden
5	Pool

ETL: N:N

.explode

ID	Feature
1	TV, Wireless Internet, Kitchen

Index	ID	Feature
1	1	TV
2	1	Wireless Internet
3	1	Kitchen

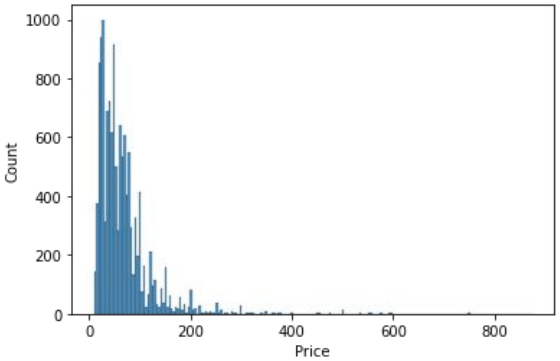
ETL: Automatización

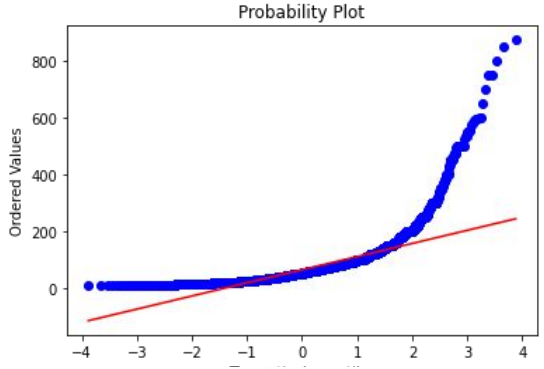
Index	Bed Type
0	Real Bed
1	Pull-out Sofa
2	Futon
3	Couch
4	Airbed

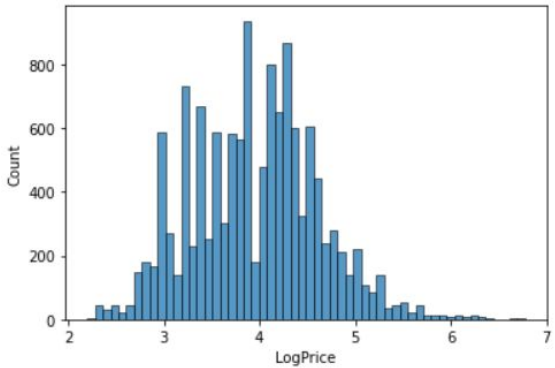
```
INSERT INTO
    grupo_15.bed_types(id, bed_type)
VALUES
    (0, 'Real Bed'),
    (1, 'Pull-out Sofa'),
    (2, 'Futon'),
    (3, 'Couch'),
    (4, 'Airbed');
```

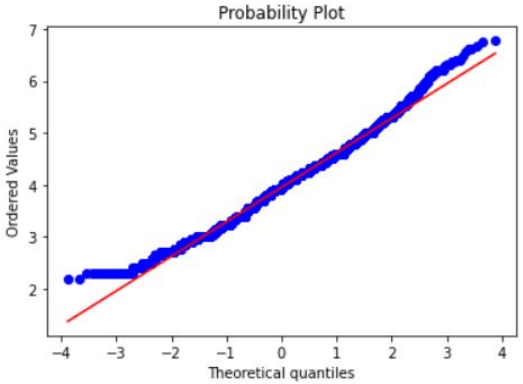
ANÁLISIS EXPLORATORIO

Análisis variable Precio

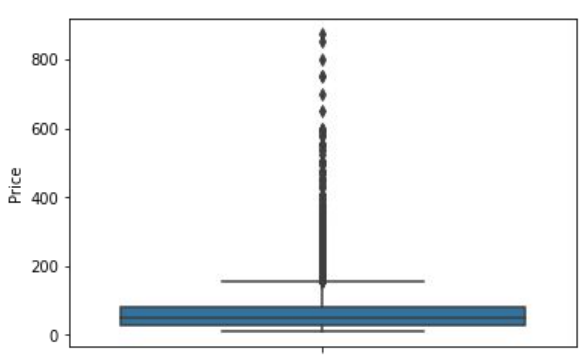


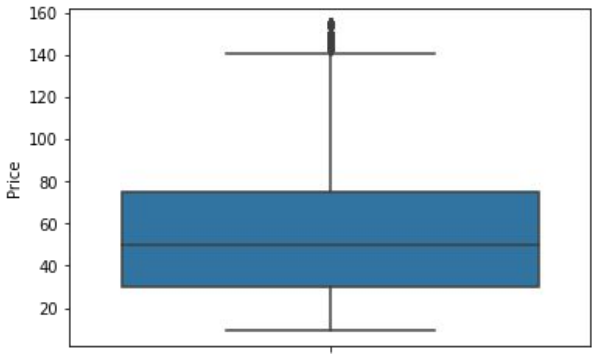






Outliers

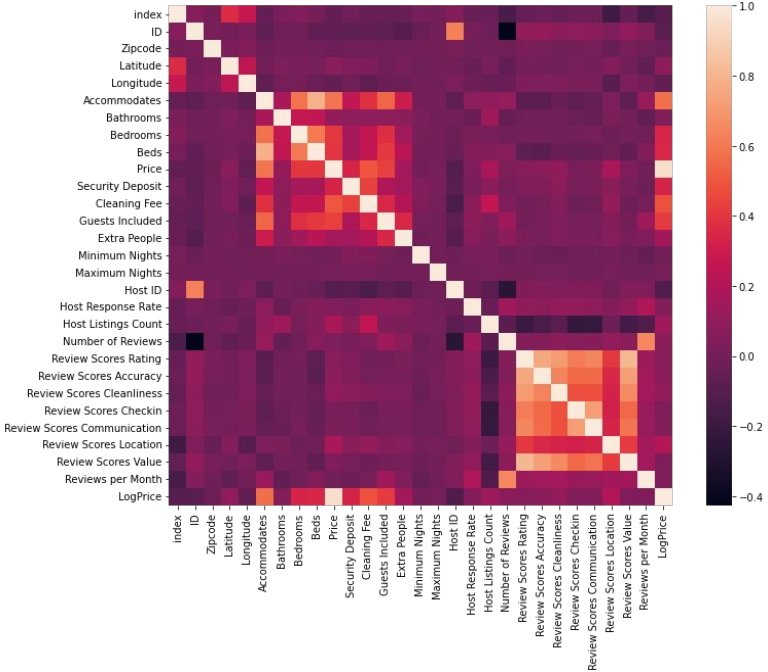




Detectamos que existe un número elevado de Outliers a través de un Boxplot.

Eliminamos aquellos datos que están a más de 1,5 veces del IQR

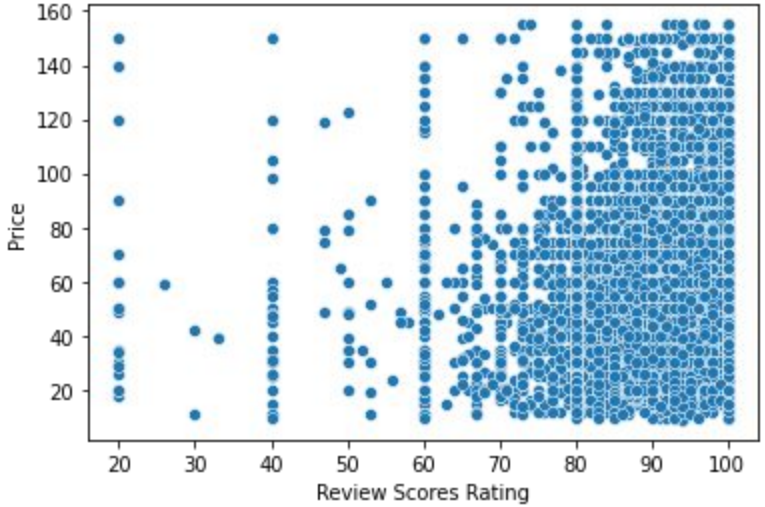
Correlaciones mapa de calor



Destaca la correlación con el precio de las variables:

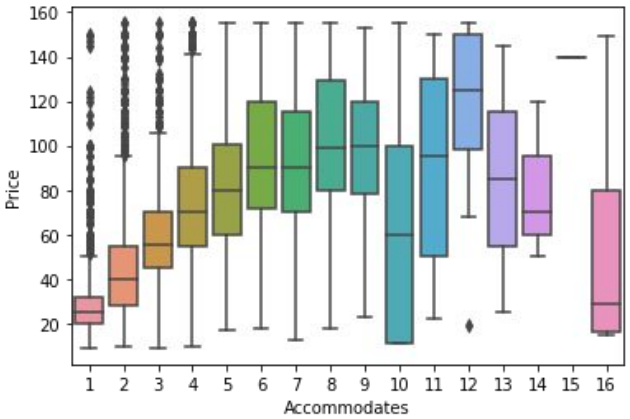
- Accommodates
- Cleaning fee
- Bedrooms

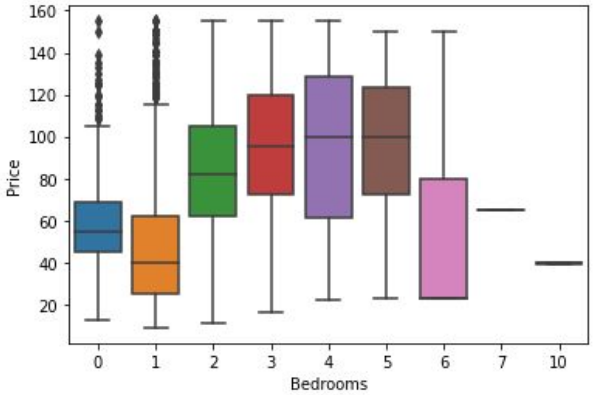
Reviews



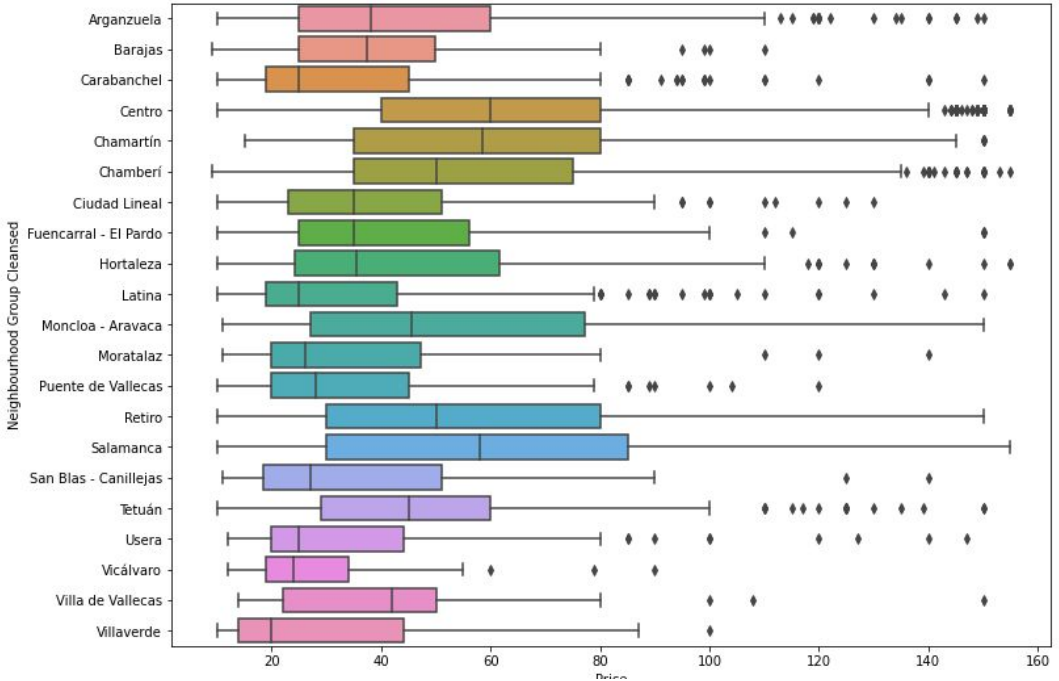
El número de reviews positivas no está necesariamente relacionado con el precio

Características del alojamiento





Ubicación del alojamiento



Barrio rentas altas

```
# Precios medios de los distintos tipos de Neighbourhood

N_Prices = data_clean.groupby(['Neighbourhood Cleansed']).Price.mean()
N_Prices = N_Prices.reset_index()
N_Prices=N_Prices.sort_values('Price',ascending=[0])
print(f'\nPrecios más Altos:\n\n{N_Prices.head(10)}\n')
print(f'Precios más Bajos:\n\n{N_Prices.tail(10)}')
```

Precios más Altos:

	Neighbourhood Cleansed	Price
98	Recoletos	81.193548
47	El Plantío	78.333333
29	Castellana	75.493151
112	Sol	74.538550
59	Hispanoamérica	73.173913
62	Jerónimos	72.986842
39	Cortes	70.349003
75	Nueva España	68.980769
60	Ibiza	68.891089
56	Goya	68.004926

Precios más Bajos:

	Neighbourhood Cleansed	Price
70	Marroquina	25.083333
84	Palomeras Sureste	24.733333
4	Aguilas	24.437500
68	Los Rosales	24.266667
91	Portazgo	22.421053
10	Ambroz	21.833333
42	Cuatro Vientos	21.333333
11	Amposta	20.400000
78	Orcasitas	18.500000
101	Rosas	17.500000

```
rentas = pd.read_csv('30677bsc.csv', sep=(';'), encoding='latin-1', decimal= ('.'))
rentas['Total'] = rentas['Total'].apply(lambda x: x.replace('.',''))
rentas['Total'] = rentas['Total'].apply(lambda x: x.replace(',','.')).astype('float').astype("Float64")
rentas = rentas.drop(['Sexo'], axis=1)
rentas = rentas.sort_values(['Total'], ascending=False)
print(f"\nRentas más Altas: \n\n{rentas['Nivel territorial: Nivel 2'].head(10)}\n")
print(f"Rentas más Bajas: \n\n{rentas['Nivel territorial: Nivel 2'].tail(10)}")
```

Rentas más Altas:

```
24                      El Viso (Madrid)
106                     Piovera (Madrid)
18                    Recoletos (Madrid)
23                   Castellana (Madrid)
28                 Nueva España (Madrid)
55     Aravaca-Plantio-Valdemarin (Madrid)
114                     Palomas (Madrid)
50                   Mirasierra (Madrid)
17           Niño Jesús-Jerónimos (Madrid)
39                      Almagro (Madrid)
Name: Nivel territorial: Nivel 2, dtype: object
```

Rentas más Bajas:

```
115     San Andrés-2 (Madrid)
129           Hellín (Madrid)
127           Ambroz (Madrid)
90         Numancia-1 (Madrid)
81         Pradolongo (Madrid)
82          Entrevías (Madrid)
83        San Diego-1 (Madrid)
135           Amposta (Madrid)
140       San Diego-2 (Madrid)
117    San Cristobal (Madrid)
Name: Nivel territorial: Nivel 2, dtype: object
```

Dashboard

Visualización de datos:

Decidimos qué KPIs queremos mostrar:

- Promedio Precio, Promedio limpieza, Promedio Fianza, Promedio Precio Total, Habitaciones

- Reviews

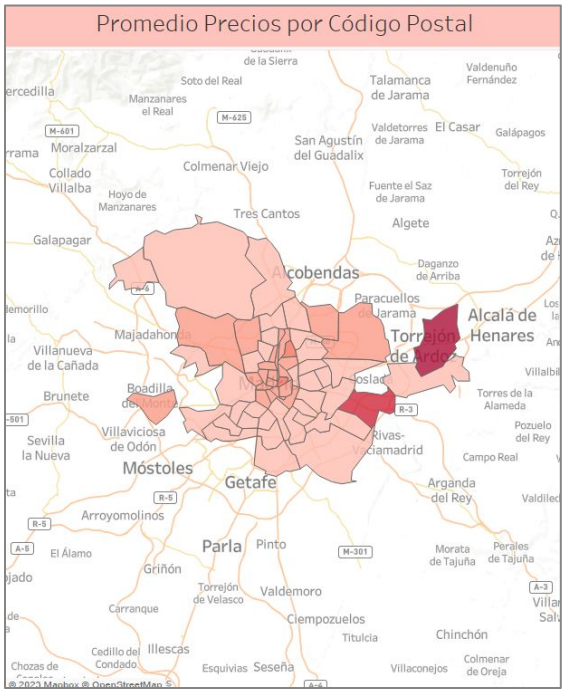
Creamos los campos calculados y parámetros necesarios

Decidimos qué gráficos queremos usar

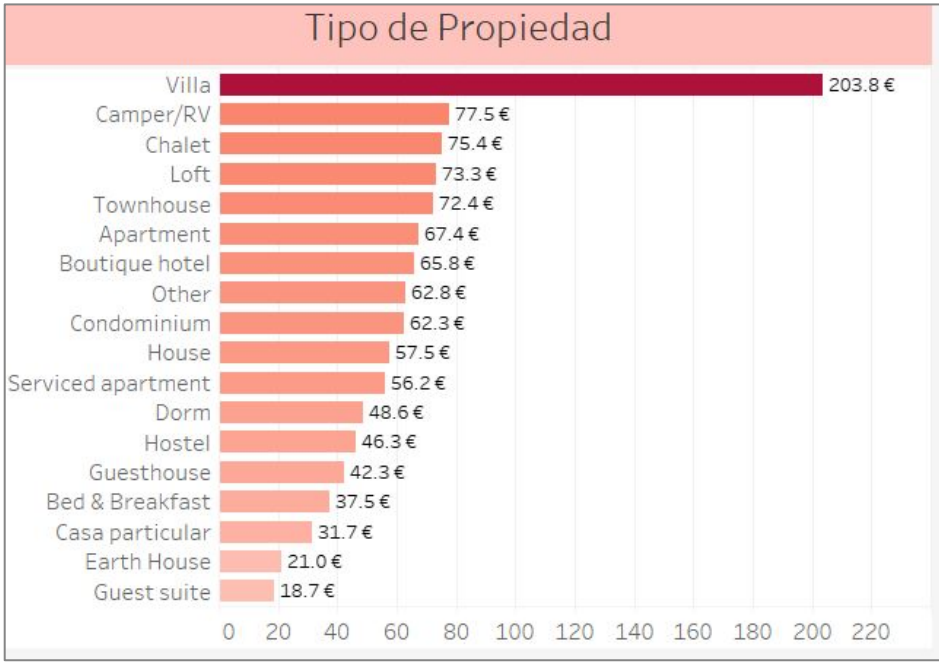
Creamos filtros

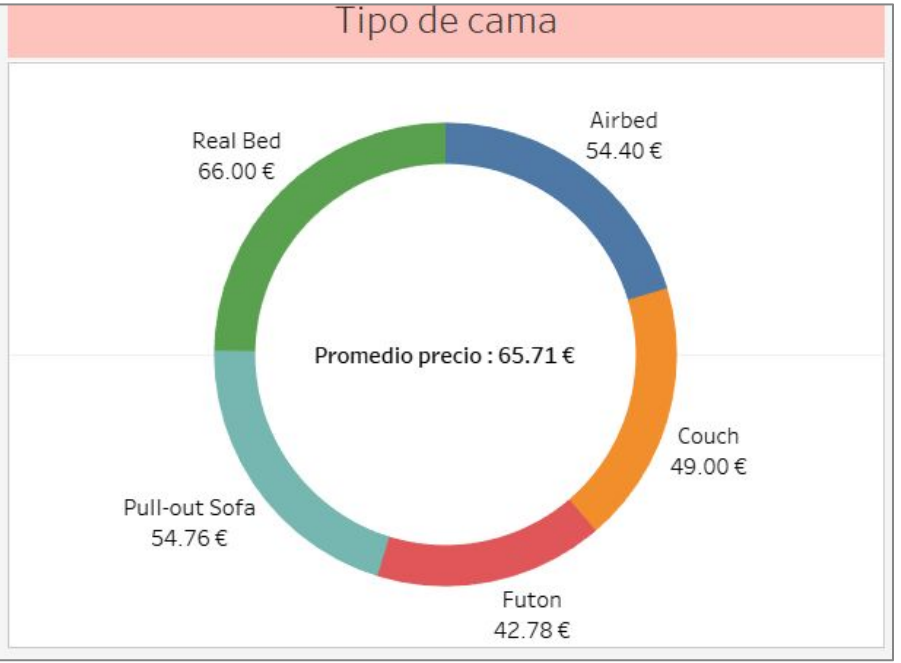
Creamos los dashboards

Visualización de datos: Precio

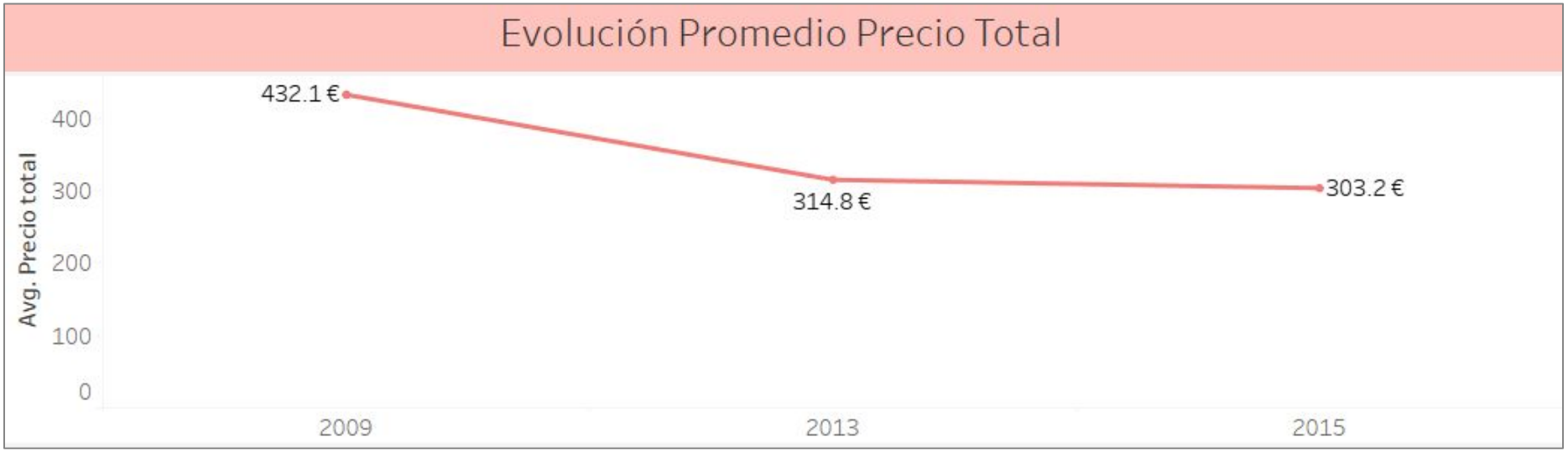


Visualización de datos: Precio

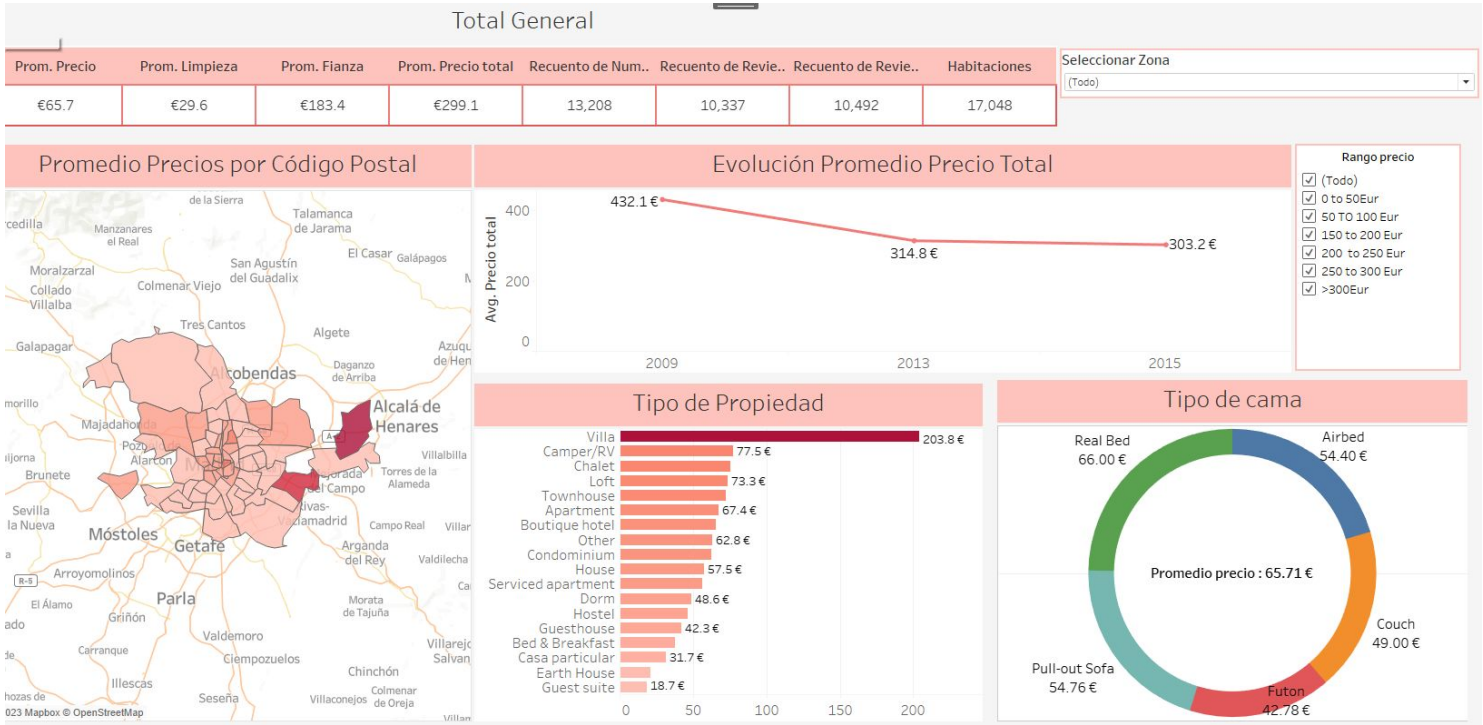




Visualización de datos: Precio

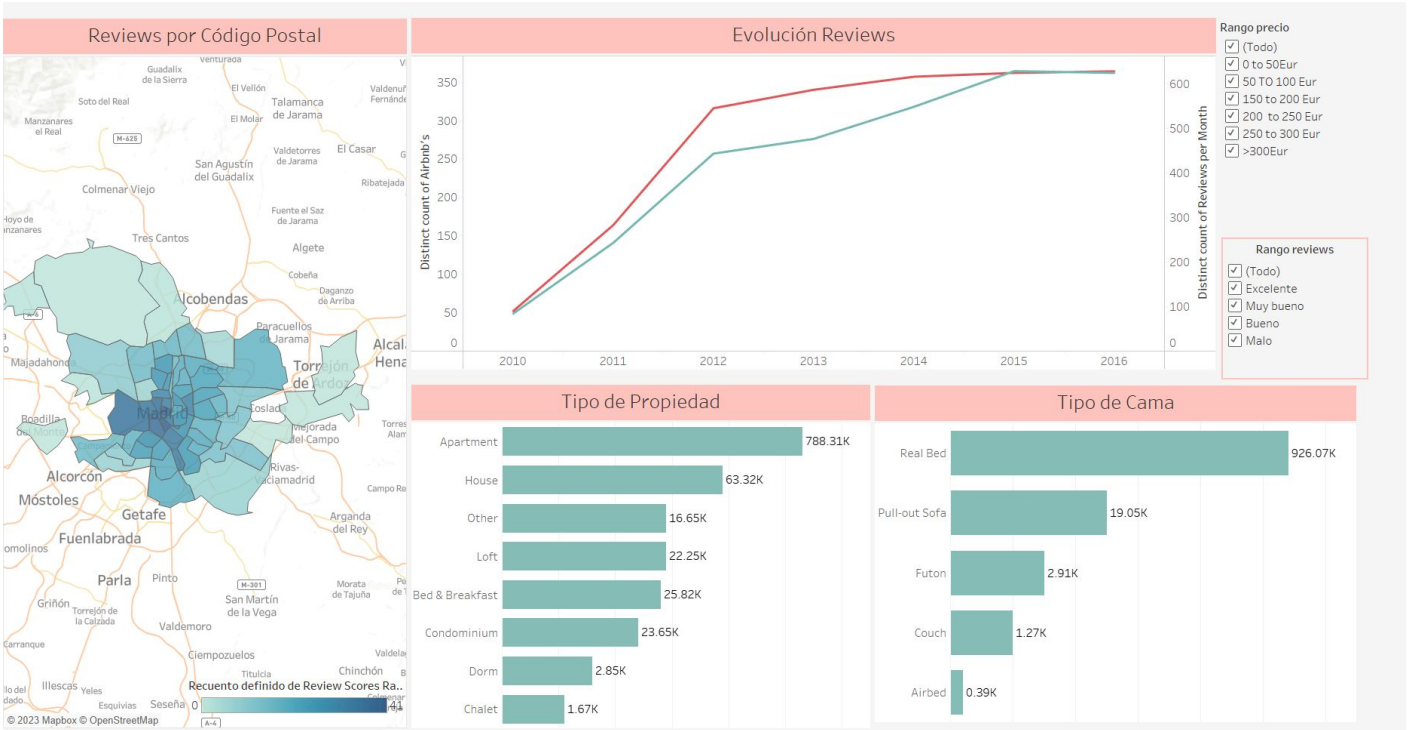


Visualización de datos: Precio



https://public.tableau.com/app/profile/jesica.pinas/viz/DashboardProyectofinalG15_/DashboardPrecio

Visualización de datos: Reviews

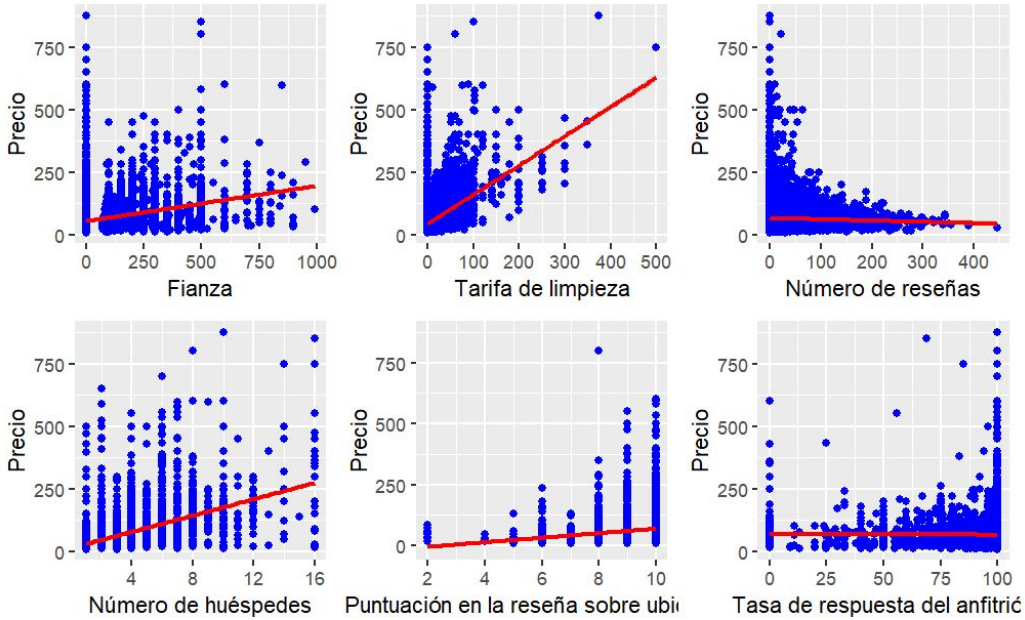


https://public.tableau.com/app/profile/jesica.pinas/viz/DashboardProyectofinalG15/DashboardReviews#8

Modelo de regresión lineal

Regresión lineal

- fianza (Security.Deposit)
- tarifa de limpieza (Cleaning.Fee)
- número de reseñas (Number.of.Reviews)
- número de huéspedes (Accommodates)
- puntuación en la reseña sobre ubicación (Review.Scores.Location)
- tasa de respuesta del anfitrión (Host.Response.Rate)



Regresión lineal

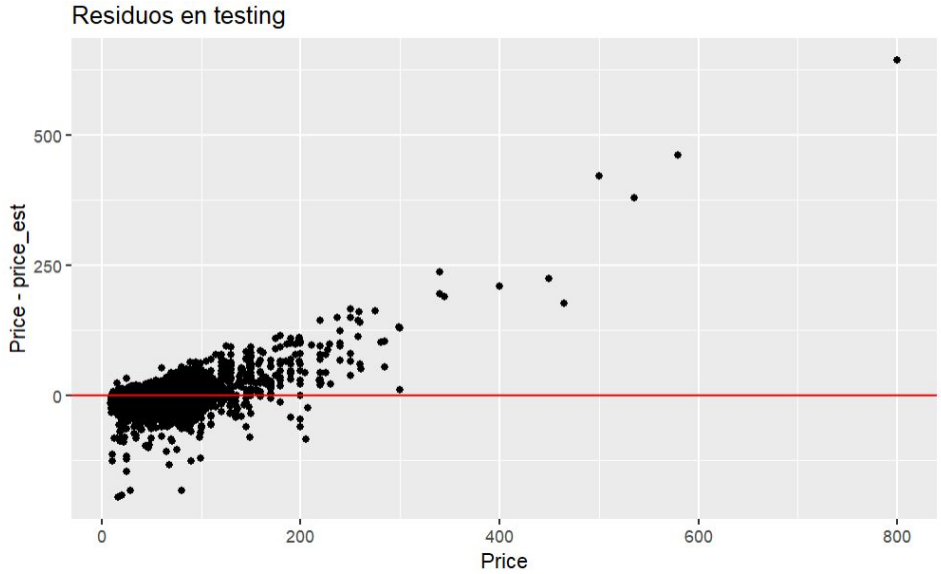
*Price = -43.3 + 0.04*Security.Deposit + 0.67*Cleaning.Fee -0.056*Number.of.Reviews + 12.4*Accommodates + 6.47*Review.Scores.Location - 0.08*Host.Response.Rate*

Training

$R^2 = 0.54$

Testing

$R^2 = 0.48$



Regresión lineal

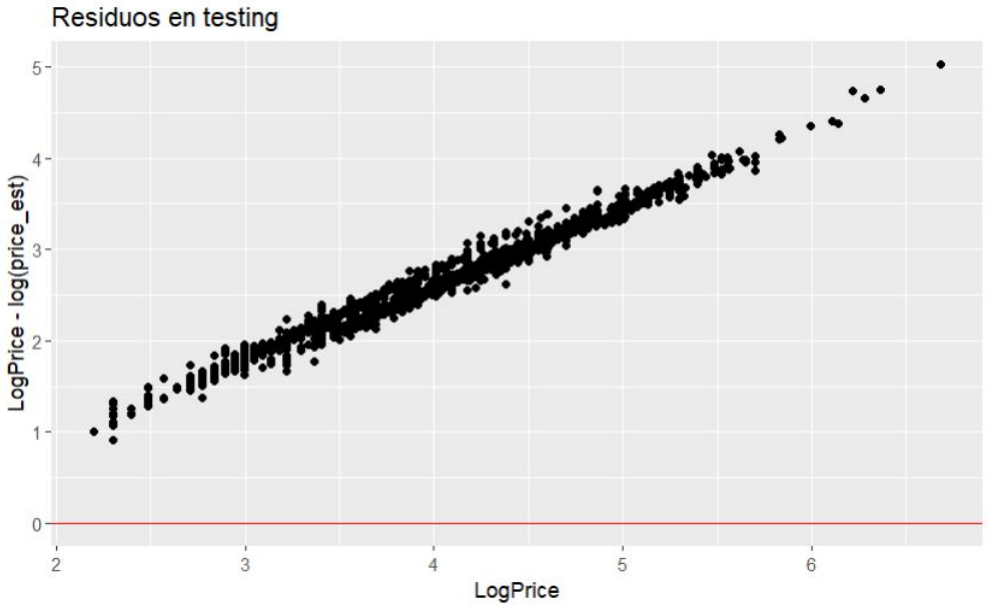
*log(Price) = 2.7 + 0.0002*Security.Deposit + 0.004*Cleaning.Fee - 0.0005*Number.of.Reviews + 0.06*Accommodates + 0.1*Bedrooms + 0.1*Bathrooms + 0.1*Review.Scores.Location - 0.0006*Host.Response.Rate*

Training

$R^2 = 0.72$

Testing

$R^2 = 0.71$



Conclusiones

Conclusiones del desarrollo del proyecto

- Desarrollar el proyecto nos ha ayudado a tener una visión holística de todo el proceso.

- Hemos podido llevar a cabo todos los pasos y entender cómo las diferentes materias del bootcamp están relacionadas con otras.

- Podríamos haber mejorado el proceso de planificación para poder documentarnos más y definir mejores objetivos.

- Entre las mejoras que nos gustaría realizar en el futuro:
  - Mejorar el rendimiento de la ETL
  - Amplia del número de variables a tener en cuenta
  - Crear modelos predictivos más complejos