Proyecto Final - Airbnb Data Analysis

Future Host Guide

Noelia Álvarez Miriam Sánchez Elena Candela
Carol Vílchez Sandra Rodríguez

Table of contents

Airbnb Data Analysis

Data Set

Arquitectura y validación de los datos

a. Muestreo y exploración inicial de los datos

```
[1] "El dataset tiene: 21278 filas y  17 columnas."
```

```
[1] "Las variables que tenemos son: "
```

```
 [1] "Room ID"                     "Name"
 [3] "Host ID"                     "Neighbourhood"
 [5] "Room type"                   "Room Price"
 [7] "Minimum nights"              "Number of reviews"
 [9] "Date last review"            "Number of reviews per month"
[11] "Rooms rent by the host"      "Availibility"
[13] "Updated Date"                "City"
[15] "Country"                     "Coordinates"
[17] "Location"
```

Para mayor facilidad en el manejo de los datos reemplazamos los espacios en los nombres de las columnas por guiones bajos "_" y se cambian los nombres por minúsculas.

```
colnames(data) = colnames(data) %>%  str_replace_all(' ', '_') %>% tolower()
```

Hacemos un análisis inicial de los datos, en este encontramos:

- La variable room_price tiene como valor máximo 9999, este valor se puede considerar como outlier pero usualmente se usa para codificar valores no válidos, por lo que se tendrá en cuenta para eliminar y no ser considerado como valor máximo observado.
- La variable number_of_reviews_per_month es numérica pero se ha cargado como character, tenemos que transformarla a numérica.
- Aunque las columnas room_id y host_id se cargan como variables cuantitativas, por sus características debemos transformarlas en variables de tipo texto.
- La variable minimum_nights tiene un valor máximo de 1125, tendremos que analizarlo para saber si se considera un outlier o si se puede corresponder con un valor anómalo.
- La variable coordinates es una variable que contiene la latitud y longitud en una misma cadena de texto, vamos a separar separar la variable usando la "," como separador y lo vamos a transformar a variable numérica.
- Las variables date_last_review y number_of_reviews_per_month tienen NA's, en el caso de number_of_reviews_per_month al ser cuantitativa se puede considerar imputar con un 0, ya que la interpretación de la variable no se vería afectada.

```
summary(data)
```

```
    room_id              name              host_id          neighbourhood
 Min.   :    6369   Length:21278       Min.   :     5154   Length:21278
 1st Qu.:18047743   Class :character   1st Qu.: 25506202   Class :character
 Median :28823834   Mode  :character   Median : 95416752   Mode  :character
```

```
Mean   :26951219                     Mean   :120527709
3rd Qu.:37708095                     3rd Qu.:208979694
Max.   :44274350                     Max.   :356881304

 room_type           room_price     minimum_nights     number_of_reviews
Length:21278       Min.   :   0.0   Min.   :   1.000   Min.   :  0.00
Class :character   1st Qu.:  36.0   1st Qu.:   1.000   1st Qu.:  0.00
Mode  :character   Median :  60.0   Median :   2.000   Median :  6.00
                   Mean   : 163.9   Mean   :   5.522   Mean   : 34.85
                   3rd Qu.: 100.0   3rd Qu.:   3.000   3rd Qu.: 38.00
                   Max.   :9999.0   Max.   :1125.000   Max.   :661.00

date_last_review     number_of_reviews_per_month rooms_rent_by_the_host
Min.   :2012-08-04   Length:21278                Min.   :  1.00
1st Qu.:2019-10-06   Class :character            1st Qu.:  1.00
Median :2020-02-23   Mode  :character            Median :  2.00
Mean   :2019-10-20                               Mean   : 14.26
3rd Qu.:2020-03-12                               3rd Qu.:  7.00
Max.   :2020-07-28                               Max.   :244.00
NA's   :5413
 availibility    updated_date           city              country
Min.   :  0.0   Min.   :2020-06-11   Length:21278       Length:21278
1st Qu.:  0.0   1st Qu.:2020-07-17   Class :character   Class :character
Median :132.0   Median :2020-07-17   Mode  :character   Mode  :character
Mean   :158.9   Mean   :2020-07-16
3rd Qu.:335.0   3rd Qu.:2020-07-17
Max.   :365.0   Max.   :2020-07-17

coordinates          location
Length:21278       Length:21278
Class :character   Class :character
Mode  :character   Mode  :character
```

Aplicamos los cambios que hemos observado en el análisis inicial antes de realizar el análisis de las variables cuantitativas.

```
data[,c("room_id", "host_id")] %<>%  toString()
data$number_of_reviews_per_month %<>%  as.numeric()
```

```
data$lat <- data$coordinates %>% str_split_i(', ', 1) %>% as.numeric()
data$long <- data$coordinates %>% str_split_i(', ', 2) %>%  as.numeric()
```

Analizamos cuántos valores distintos tienen las variables cualitativas:

```
data %>% select_if(where(is.character)) %>% map_dfc(n_distinct)
```

```
# A tibble: 1 x 9
  room_id  name host_id neighbourhood room_type  city country coordinates
    <int> <int>   <int>         <int>     <int> <int>   <int>       <int>
1       1 20410       1           142         4    13       8       21278
# i 1 more variable: location <int>
```

Las variables City y Country deberían tener un único valor, ya que solo nos centramos en una ciudad y en un país.

```
print('Ciudades únicas:')
```

```
[1] "Ciudades únicas:"
```

```
data$city %>% unique()
```

```
 [1] "Madrid"        "Sevilla"       "Girona"        "Barcelona"
 [5] "Lisbon"        "Florence"      "Istanbul"      "San-francisco"
 [9] "Brussels"      "London"        "Sydney"        "Mallorca"
[13] "Sicily"
```

```
print('Países únicas:')
```

```
[1] "Países únicas:"
```

```
data$country %>% unique()
```

```
[1] "Spain"          "Portugal"       "Italy"          "Turkey"
[5] "United states"  "Belgium"        "United kingdom" "Australia"
```

```
data = data[data$city=="Madrid",]
```

```
data %>% select_if(where(is.character)) %>% map_dfc(n_distinct)
```

```
# A tibble: 1 x 9
  room_id  name host_id neighbourhood room_type  city country coordinates
    <int> <int>   <int>         <int>     <int> <int>   <int>       <int>
1       1 20391       1           127         4     1       1       21255
# i 1 more variable: location <int>
```

b. Definir e implementar el Datawarehouse

```
data[is.na(data$number_of_reviews_per_month), 'number_of_reviews_per_month'] = 0
```

c. (Opcional) Ingesta de datos (ETL) y validación de que se ha cargado correctamente

Análisis Exploratorio

Hacer un estudio estadístico con R o Python, según preferencia personal, y averiguar cuales son las métricas adecuadas para el dataset. No olvidemos: a. Revisión de la calidad de los datos b. Detección outliers (rango de variables), imputación valores nulos. c. Boxplots, histogramas, etc. d. Normalización de los valores de las tablas (quitar tildes, "dobles espacios", etc.)

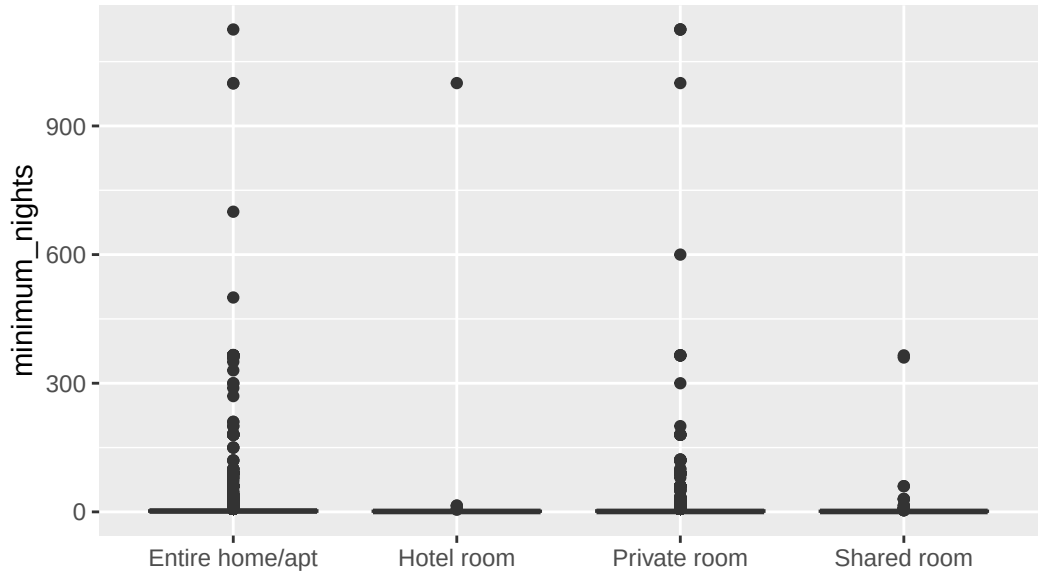
```
data %>% select_if(where(is.numeric)) %>% summary()
```

```
   room_price   minimum_nights     number_of_reviews
 Min.   :   0   Min.   :   1.000   Min.   :  0.00
 1st Qu.:  36   1st Qu.:   1.000   1st Qu.:  0.00
 Median :  60   Median :   2.000   Median :  6.00
 Mean   : 164   Mean   :   5.522   Mean   : 34.88
 3rd Qu.: 100   3rd Qu.:   3.000   3rd Qu.: 38.00
 Max.   :9999   Max.   :1125.000   Max.   :661.00
 number_of_reviews_per_month rooms_rent_by_the_host  availibility
 Min.   : 0.000              Min.   :  1.00         Min.   :  0.0
 1st Qu.: 0.000              1st Qu.:  1.00         1st Qu.:  0.0
 Median : 0.390              Median :  2.00         Median :133.0
 Mean   : 1.065              Mean   : 14.25         Mean   :158.9
 3rd Qu.: 1.540              3rd Qu.:  7.00         3rd Qu.:335.0
 Max.   :27.250              Max.   :244.00         Max.   :365.0
      lat             long
```

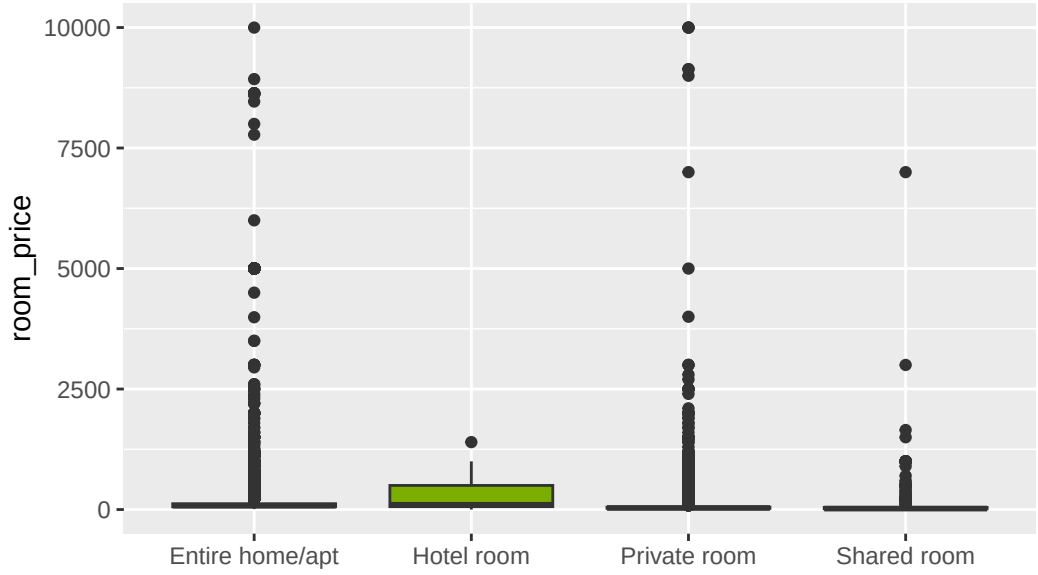
```
 Min.   :40.33   Min.   :-3.864
 1st Qu.:40.41   1st Qu.:-3.708
 Median :40.42   Median :-3.701
 Mean   :40.42   Mean   :-3.694
 3rd Qu.:40.43   3rd Qu.:-3.687
 Max.   :40.56   Max.   :-3.524
```

En las variables room_price y minimum_nights tiene un máximo muy lejano al tercer cuartil, vamos a analizar si estos valores se pueden considerar válidos o no.

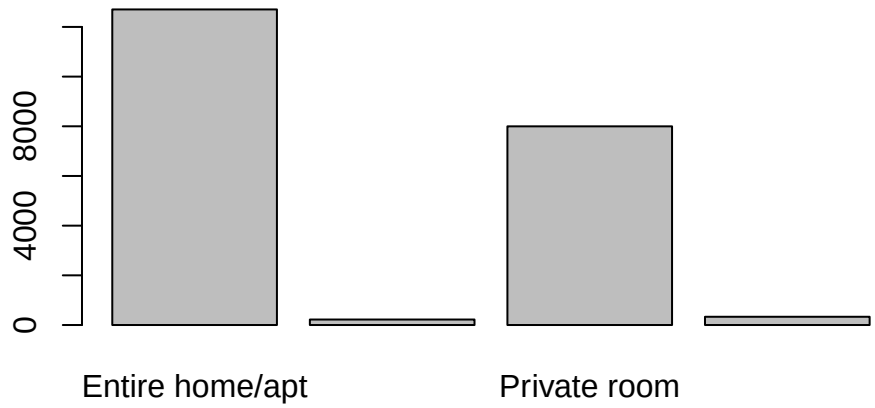
```
data %>% select ('room_type','room_price', 'minimum_nights') %>%
  ggplot( aes(x=room_type, y=minimum_nights, fill=room_type)) +
    geom_boxplot() +
 #   scale_fill_viridis(discrete = TRUE, alpha=0.6, option="A") +
  #  theme_ipsum() +
    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    ggtitle("") +
    xlab("")
```



```
data %>% select ('room_type','room_price', 'minimum_nights') %>%
  ggplot( aes(x=room_type, y=room_price, fill=room_type)) +
    geom_boxplot() +
 #   scale_fill_viridis(discrete = TRUE, alpha=0.6, option="A") +
  #  theme_ipsum() +
    theme(
      legend.position="none",
      plot.title = element_text(size=11)
    ) +
    ggtitle("") +
    xlab("")
```




```
data$room_type %>% table() %>% barplot()
```



Visualización de las métricas

A partir de los datos de Airbnb, obtén los KPIs que puedan ser de relevancia y contesta a través de un dashboard a una pregunta relevante que hagas sobre los datos. a. a. Se valorará el diseño final del dashboard. b. b. El uso de buenas prácticas. c. c. El cálculo de KPIs adecuados y el uso de campos calculados avanzados. d. d. El uso de vistas interactivas. Nota: En este ejercicio no habrá un dashboard modelo sino que se basará en valorar vuestras capacidades con el uso de la herramienta de Tableau y que podáis demostrar todo lo aprendido durante este bloque

Pre-procesamiento y Modelado

La tarea asignada es hacer un algoritmo de regresión lineal que prediga el precio de un inmueble en función de las características que elijáis

En primer lugar dividimos los datos en train y test para para poder **evaluar el rendimiento del modelo**, ya que si usamos todos los datos del dataset, el modelo estaría entrenado para valores específicos de esos datos y que no se podrían generalizar a otros conjuntos de datos, suceso llamado **overfitting**.

Al dividir el dataset, vemos el rendimiento en datos del propio conjunto que no han sido usados para su entrenamiento, por lo que nos da una visión más realista de cómo se comportará el modelo con datos nuevos.

```
data <- data[data$room_price<9999,]
set.seed(123)
idx <- sample(1:nrow(data), 0.7*nrow(data))
train_data <- data[idx,]
test_data <- data[-idx,]
```

El modelo de regresión lineal muestra cómo los precios de las habitaciones varían según el barrio, el tipo de habitación, la disponibilidad y el minimo de noches. Hay una gran parte de la variabilidad que no se explica por estas variables, lo que sugiere que se podría considerar agregar más variables independientes para mejorar la capacidad explicativa del modelo o utilizar otro tipo de algoritmo que tenga en cuenta relaciones mas complejas y no lineales entre las variables independientes y la variable dependiente.

En este caso, el término del intercept se ha eliminado, y el precio predicho se calcula directamente sumando los efectos de todas las variables predictoras. Esta aproximación asume que el precio base de una habitación es 0 cuando todas las demás variables son 0.

```
model <- lm(room_price ~ neighbourhood + room_type + availibility +
minimum_nights + 0, data = train_data)

summary(model)
```

```
Call:
lm(formula = room_price ~ neighbourhood + room_type + availibility +
    minimum_nights + 0, data = train_data)

Residuals:
   Min     1Q Median     3Q    Max
-578.8 -123.4  -72.0  -11.7 8847.9

Coefficients:
                                          Estimate Std. Error t value
neighbourhoodAbrantes                    132.59565   92.43629   1.434
neighbourhoodAcacias                     139.89287   44.72304   3.128
neighbourhoodAdelfas                      68.79104   70.61239   0.974
neighbourhoodAeropuerto                  110.63598  172.36575   0.642
neighbourhoodAguilas                     126.01836   86.59882   1.455
```

neighbourhoodAlameda de Osuna	263.60136	83.57202	3.154
neighbourhoodAlmagro	199.85219	45.04158	4.437
neighbourhoodAlmenara	123.99171	70.01863	1.771
neighbourhoodAlmendrales	200.59025	82.47622	2.432
neighbourhoodAluche	99.50312	63.54879	1.566
neighbourhoodAmbroz	214.78613	136.65312	1.572
neighbourhoodAmposta	281.63403	222.61796	1.265
neighbourhoodApostol Santiago	164.41481	132.43741	1.241
neighbourhoodArapiles	172.44048	45.44937	3.794
neighbourhoodAravaca	118.42970	99.67481	1.188
neighbourhoodArcos	537.24163	87.82190	6.117
neighbourhoodArgüelles	194.58777	37.35442	5.209
neighbourhoodAtocha	158.75365	136.34619	1.164
neighbourhoodBellas Vistas	213.76715	50.89376	4.200
neighbourhoodBerruguete	202.70248	51.30663	3.951
neighbourhoodBuenavista	134.18712	80.86655	1.659
neighbourhoodButarque	96.56889	164.41940	0.587
neighbourhoodCampamento	126.97479	128.85178	0.985
neighbourhoodCanillas	108.66756	64.33459	1.689
neighbourhoodCanillejas	573.24849	72.11802	7.949
neighbourhoodCármenes	113.65628	84.42109	1.346
neighbourhoodCasa de Campo	120.73157	65.02091	1.857
neighbourhoodCasco Histórico de Barajas	128.05865	111.44072	1.149
neighbourhoodCasco Histórico de Vallecas	150.05878	69.12245	2.171
neighbourhoodCasco Histórico de Vicálvaro	216.75541	91.28630	2.374
neighbourhoodCastellana	348.31257	47.58660	7.320
neighbourhoodCastilla	191.50518	80.50411	2.379
neighbourhoodCastillejos	249.98368	54.44213	4.592
neighbourhoodChopera	107.44298	52.91533	2.030
neighbourhoodCiudad Jardín	195.48248	63.29086	3.089
neighbourhoodCiudad Universitaria	125.02420	77.41618	1.615
neighbourhoodColina	81.01971	145.73733	0.556
neighbourhoodComillas	123.62896	65.58092	1.885
neighbourhoodConcepción	138.76694	63.72367	2.178
neighbourhoodCorralejos	146.42670	164.35941	0.891
neighbourhoodCortes	185.10335	21.85947	8.468
neighbourhoodCostillares	131.06877	92.35164	1.419
neighbourhoodCuatro Caminos	162.63745	44.38933	3.664
neighbourhoodCuatro Vientos	143.38249	222.52117	0.644
neighbourhoodDelicias	188.74689	46.96720	4.019
neighbourhoodEl Goloso	229.71940	164.31911	1.398
neighbourhoodEl Pardo	321.26911	385.09157	0.834
neighbourhoodEl Plantío	114.91702	243.70591	0.472

neighbourhoodEl Viso	368.82143	66.79563	5.522
neighbourhoodEmbajadores	140.19054	13.98047	10.028
neighbourhoodEntrevías	100.86819	93.93703	1.074
neighbourhoodEstrella	116.90765	113.76723	1.028
neighbourhoodFontarrón	112.35182	93.93994	1.196
neighbourhoodFuente del Berro	114.00114	55.36851	2.059
neighbourhoodFuentelareina	90.56941	544.56523	0.166
neighbourhoodGaztambide	128.01871	42.40071	3.019
neighbourhoodGoya	313.90879	35.82082	8.763
neighbourhoodGuindalera	557.21486	39.67873	14.043
neighbourhoodHellín	488.34022	116.57267	4.189
neighbourhoodHispanoamérica	384.48465	61.37296	6.265
neighbourhoodHorcajo	80.02136	385.14881	0.208
neighbourhoodIbiza	183.73246	47.09345	3.901
neighbourhoodImperial	245.46365	55.01353	4.462
neighbourhoodJerónimos	195.51674	62.32931	3.137
neighbourhoodJusticia	191.40256	20.44106	9.364
neighbourhoodLa Paz	293.57184	114.01388	2.575
neighbourhoodLegazpi	87.40739	89.69149	0.975
neighbourhoodLista	454.17709	44.72539	10.155
neighbourhoodLos Angeles	105.61772	107.20580	0.985
neighbourhoodLos Rosales	112.75678	79.43360	1.420
neighbourhoodLucero	131.14470	64.67688	2.028
neighbourhoodMarroquina	219.79521	119.24818	1.843
neighbourhoodMedia Legua	135.41476	157.35200	0.861
neighbourhoodMirasierra	151.87792	109.17274	1.391
neighbourhoodMoscardó	240.40553	63.49419	3.786
neighbourhoodNiño Jesús	208.41377	99.60999	2.092
neighbourhoodNueva España	218.95387	84.33028	2.596
neighbourhoodNumancia	87.35683	47.73899	1.830
neighbourhoodOpañel	97.54601	68.50745	1.424
neighbourhoodOrcasitas	131.83294	192.74270	0.684
neighbourhoodOrcasur	105.04536	125.29065	0.838
neighbourhoodPacífico	120.15050	42.54765	2.824
neighbourhoodPalacio	163.60998	18.13056	9.024
neighbourhoodPalomas	253.78938	145.75017	1.741
neighbourhoodPalomeras Bajas	230.67099	67.53338	3.416
neighbourhoodPalomeras Sureste	146.55732	101.54315	1.443
neighbourhoodPalos de Moguer	104.19846	34.18876	3.048
neighbourhoodPavones	100.10207	314.42482	0.318
neighbourhoodPeñagrande	214.47374	70.22237	3.054
neighbourhoodPilar	95.36521	84.43748	1.129
neighbourhoodPinar del Rey	170.19925	72.02346	2.363

neighbourhoodPiovera	209.72423	107.09787	1.958
neighbourhoodPortazgo	112.19345	88.76833	1.264
neighbourhoodPradolongo	374.79591	113.91037	3.290
neighbourhoodProsperidad	98.17707	51.26203	1.915
neighbourhoodPueblo Nuevo	164.17531	55.16284	2.976
neighbourhoodPuerta Bonita	162.36560	73.88368	2.198
neighbourhoodPuerta del Angel	213.68011	35.81195	5.967
neighbourhoodQuintana	140.56674	62.28541	2.257
neighbourhoodRecoletos	440.33158	36.35599	12.112
neighbourhoodRejas	212.70836	73.23989	2.904
neighbourhoodRios Rosas	176.72281	44.97291	3.930
neighbourhoodRosas	542.51120	70.52069	7.693
neighbourhoodSalvador	216.35060	119.25593	1.814
neighbourhoodSan Andrés	90.62391	98.19279	0.923
neighbourhoodSan Cristobal	123.98561	140.94326	0.880
neighbourhoodSan Diego	116.13099	47.09915	2.466
neighbourhoodSan Fermín	103.42681	132.35352	0.781
neighbourhoodSan Isidro	109.96766	52.23854	2.105
neighbourhoodSan Juan Bautista	243.70365	116.34637	2.095
neighbourhoodSan Pascual	162.00066	105.12771	1.541
neighbourhoodSanta Eugenia	127.39288	222.43740	0.573
neighbourhoodSimancas	307.23695	56.39756	5.448
neighbourhoodSol	136.45956	18.43809	7.401
neighbourhoodTimón	162.67347	81.60184	1.994
neighbourhoodTrafalgar	215.28584	33.21669	6.481
neighbourhoodUniversidad	153.75989	15.48767	9.928
neighbourhoodValdeacederas	152.73832	61.40207	2.488
neighbourhoodValdefuentes	245.24306	63.25901	3.877
neighbourhoodValdemarín	109.20879	243.73691	0.448
neighbourhoodValdezarza	143.87232	86.62971	1.661
neighbourhoodVallehermoso	250.93800	74.47249	3.370
neighbourhoodValverde	124.65503	67.99807	1.833
neighbourhoodVentas	149.26425	48.35534	3.087
neighbourhoodVinateros	80.21900	164.42448	0.488
neighbourhoodVista Alegre	102.92840	61.86959	1.664
neighbourhoodZofío	433.90943	92.35190	4.698
room_typeHotel room	84.87463	44.54045	1.906
room_typePrivate room	-108.36120	9.92710	-10.916
room_typeShared room	-67.18772	36.88499	-1.822
availibility	0.14480	0.03109	4.657
minimum_nights	-0.30104	0.15939	-1.889
	Pr(>|t|)		
neighbourhoodAbrantes	0.151464		

```
neighbourhoodAcacias                        0.001764 **
neighbourhoodAdelfas                        0.329970
neighbourhoodAeropuerto                     0.520969
neighbourhoodAguilas                        0.145636
neighbourhoodAlameda de Osuna               0.001613 **
neighbourhoodAlmagro                        9.19e-06 ***
neighbourhoodAlmenara                       0.076608 .
neighbourhoodAlmendrales                    0.015023 *
neighbourhoodAluche                         0.117423
neighbourhoodAmbroz                         0.116027
neighbourhoodAmposta                        0.205855
neighbourhoodApostol Santiago               0.214458
neighbourhoodArapiles                       0.000149 ***
neighbourhoodAravaca                        0.234789
neighbourhoodArcos                          9.75e-10 ***
neighbourhoodArgüelles                      1.92e-07 ***
neighbourhoodAtocha                         0.244304
neighbourhoodBellas Vistas                  2.68e-05 ***
neighbourhoodBerruguete                     7.83e-05 ***
neighbourhoodBuenavista                     0.097064 .
neighbourhoodButarque                       0.556989
neighbourhoodCampamento                     0.324428
neighbourhoodCanillas                       0.091221 .
neighbourhoodCanillejas                     2.02e-15 ***
neighbourhoodCármenes                       0.178226
neighbourhoodCasa de Campo                  0.063358 .
neighbourhoodCasco Histórico de Barajas     0.250526
neighbourhoodCasco Histórico de Vallecas    0.029954 *
neighbourhoodCasco Histórico de Vicálvaro   0.017588 *
neighbourhoodCastellana                     2.62e-13 ***
neighbourhoodCastilla                       0.017381 *
neighbourhoodCastillejos                    4.43e-06 ***
neighbourhoodChopera                        0.042327 *
neighbourhoodCiudad Jardín                  0.002015 **
neighbourhoodCiudad Universitaria           0.106340
neighbourhoodColina                         0.578267
neighbourhoodComillas                       0.059431 .
neighbourhoodConcepción                     0.029449 *
neighbourhoodCorralejos                     0.373001
neighbourhoodCortes                          < 2e-16 ***
neighbourhoodCostillares                    0.155851
neighbourhoodCuatro Caminos                 0.000249 ***
neighbourhoodCuatro Vientos                 0.519356
```

```
neighbourhoodDelicias                  5.88e-05 ***
neighbourhoodEl Goloso                 0.162132
neighbourhoodEl Pardo                  0.404144
neighbourhoodEl Plantío                0.637262
neighbourhoodEl Viso                   3.42e-08 ***
neighbourhoodEmbajadores                < 2e-16 ***
neighbourhoodEntrevías                 0.282937
neighbourhoodEstrella                  0.304153
neighbourhoodFontarrón                 0.231717
neighbourhoodFuente del Berro          0.039516 *
neighbourhoodFuentelareina             0.867911
neighbourhoodGaztambide                0.002538 **
neighbourhoodGoya                       < 2e-16 ***
neighbourhoodGuindalera                 < 2e-16 ***
neighbourhoodHellín                    2.82e-05 ***
neighbourhoodHispanoamérica            3.84e-10 ***
neighbourhoodHorcajo                   0.835413
neighbourhoodIbiza                     9.60e-05 ***
neighbourhoodImperial                  8.18e-06 ***
neighbourhoodJerónimos                 0.001711 **
neighbourhoodJusticia                   < 2e-16 ***
neighbourhoodLa Paz                    0.010037 *
neighbourhoodLegazpi                   0.329808
neighbourhoodLista                      < 2e-16 ***
neighbourhoodLos Angeles               0.324549
neighbourhoodLos Rosales               0.155772
neighbourhoodLucero                    0.042610 *
neighbourhoodMarroquina                0.065324 .
neighbourhoodMedia Legua               0.389481
neighbourhoodMirasierra                0.164195
neighbourhoodMoscardó                  0.000154 ***
neighbourhoodNiño Jesús                0.036429 *
neighbourhoodNueva España              0.009430 **
neighbourhoodNumancia                  0.067287 .
neighbourhoodOpañel                    0.154504
neighbourhoodOrcasitas                 0.493996
neighbourhoodOrcasur                   0.401812
neighbourhoodPacífico                  0.004751 **
neighbourhoodPalacio                    < 2e-16 ***
neighbourhoodPalomas                   0.081658 .
neighbourhoodPalomeras Bajas           0.000638 ***
neighbourhoodPalomeras Sureste         0.148957
neighbourhoodPalos de Moguer           0.002310 **
```

```
neighbourhoodPavones                    0.750212
neighbourhoodPeñagrande                 0.002261 **
neighbourhoodPilar                      0.258740
neighbourhoodPinar del Rey              0.018135 *
neighbourhoodPiovera                    0.050220 .
neighbourhoodPortazgo                   0.206289
neighbourhoodPradolongo                 0.001003 **
neighbourhoodProsperidad                0.055486 .
neighbourhoodPueblo Nuevo               0.002923 **
neighbourhoodPuerta Bonita              0.027994 *
neighbourhoodPuerta del Angel           2.48e-09 ***
neighbourhoodQuintana                   0.024034 *
neighbourhoodRecoletos                   < 2e-16 ***
neighbourhoodRejas                      0.003687 **
neighbourhoodRios Rosas                 8.55e-05 ***
neighbourhoodRosas                      1.53e-14 ***
neighbourhoodSalvador                   0.069672 .
neighbourhoodSan Andrés                 0.356065
neighbourhoodSan Cristobal              0.379045
neighbourhoodSan Diego                  0.013687 *
neighbourhoodSan Fermín                 0.434554
neighbourhoodSan Isidro                 0.035299 *
neighbourhoodSan Juan Bautista          0.036220 *
neighbourhoodSan Pascual                0.123341
neighbourhoodSanta Eugenia              0.566847
neighbourhoodSimancas                   5.18e-08 ***
neighbourhoodSol                        1.43e-13 ***
neighbourhoodTimón                      0.046225 *
neighbourhoodTrafalgar                  9.39e-11 ***
neighbourhoodUniversidad                 < 2e-16 ***
neighbourhoodValdeacederas              0.012875 *
neighbourhoodValdefuentes               0.000106 ***
neighbourhoodValdemarín                 0.654116
neighbourhoodValdezarza                 0.096780 .
neighbourhoodVallehermoso               0.000755 ***
neighbourhoodValverde                   0.066791 .
neighbourhoodVentas                     0.002027 **
neighbourhoodVinateros                  0.625644
neighbourhoodVista Alegre               0.096207 .
neighbourhoodZofío                      2.65e-06 ***
room_typeHotel room                     0.056726 .
room_typePrivate room                    < 2e-16 ***
room_typeShared room                    0.068544 .
```

```
availibility                                3.23e-06 ***
minimum_nights                              0.058954 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 544.6 on 14743 degrees of freedom
Multiple R-squared:  0.1159,    Adjusted R-squared:  0.108
F-statistic: 14.64 on 132 and 14743 DF,  p-value: < 2.2e-16
```

```r
predictions <- predict(model, newdata = test_data)
accuracy <- sqrt(mean((predictions - test_data$room_price)^2))
print(paste("RMSE del modelo:", accuracy))
```

```
[1] "RMSE del modelo: 510.73186956232"
```

```r
coefficients <- coef(model)

predict_room_price <- function(neighbourhood, room_type, availibility,
minimum_nights, coefficients) {
  # Intercepto
  intercept <- coefficients["(Intercept)"]

  # Extraer los coeficientes de las variables categóricas
  neighbourhood_coef <- coefficients[paste0("neighbourhood", neighbourhood)]
  room_type_coef <- coefficients[paste0("room_type", room_type)]

  # Coeficiente de la variable continua
  availibility_coef <- coefficients["availibility"]
  minimum_nights_coef <- coefficients["minimum_nights"]

  # Verificar si los coeficientes de las interacciones existen
  interaction_coef <- coefficients[paste0("neighbourhood", neighbourhood,
  ":room_type", room_type)]
  if (is.na(interaction_coef)) {
    interaction_coef <- 0
  }

  # Calcular la predicción
  predicted_price <- intercept + neighbourhood_coef + room_type_coef +
  availibility_coef * availibility + + minimum_nights_coef * minimum_nights +
  interaction_coef
```

```
  return(predicted_price)
}
```

```
new_data <- data.frame(neighbourhood = "Sol", room_type = "Entire home/apt",
availibility = 120,  minimum_nights = 1)

# Usar la función predict() para predecir el precio
predicted_price <- predict(model, new_data)
cat("La predicción del precio para un apartamento entero en Sol con \n
disponibilidad de 120 dias y que se reserve como minimo una noche es ",
round(predicted_price,2), "€ \n")
```

```
La predicción del precio para un apartamento entero en Sol con

disponibilidad de 120 dias y que se reserve como minimo una noche es  153.53 €
```

```
new_data <- data.frame(neighbourhood = "Sol", room_type = "Entire home/apt",
availibility = 365,  minimum_nights = 1)

# Usar la función predict() para predecir el precio
predicted_price <- predict(model, new_data)
cat("La predicción del precio para un apartamento entero en Sol con \n
disponibilidad de 365 dias y que se reserve como minimo una noche es ",
round(predicted_price,2), "€. \n
Se puede comprobar con respecto al anterior que con un precio más alto \n
hay una mayor disponibilidad, el apartamento con el precio más bajo se \n
alquila más y por lo tanto tiene menos disponibilidad.\n")
```

```
La predicción del precio para un apartamento entero en Sol con

disponibilidad de 365 dias y que se reserve como minimo una noche es  189.01 €.

Se puede comprobar con respecto al anterior que con un precio más alto

hay una mayor disponibilidad, el apartamento con el precio más bajo se

alquila más y por lo tanto tiene menos disponibilidad.
```

```
new_data <- data.frame(neighbourhood = "Sol", room_type = "Hotel room",
availibility = 365, minimum_nights = 1)

# Usar la función predict() para predecir el precio
predicted_price <- predict(model, new_data)
cat("La predicción del precio para una habitación de hotel en Sol con \n
disponibilidad de 365 dias y que se reserve como minimo una noche es ",
round(predicted_price,2), "€. \n
Si el tipo de alojamiento es una habitación de hotel, el precio de la misma \n
aumenta notablemente. \n")
```

```
La predicción del precio para una habitación de hotel en Sol con

disponibilidad de 365 dias y que se reserve como minimo una noche es  273.88 €.

Si el tipo de alojamiento es una habitación de hotel, el precio de la misma

aumenta notablemente.
```

```
new_data <- data.frame(neighbourhood = "Prosperidad",
room_type = "Entire home/apt", availibility = 365, minimum_nights = 1)

# Usar la función predict() para predecir el precio
predicted_price <- predict(model, new_data)
cat("La predicción del precio para un apartamento entero en Prosperidad con \n
disponibilidad de 365 dias y que se reserve como minimo una noche es ",
round(predicted_price,2), "€. \n
En un barrio más alejado del centro el precio es menor.\n")
```

```
La predicción del precio para un apartamento entero en Prosperidad con

disponibilidad de 365 dias y que se reserve como minimo una noche es  150.73 €.

En un barrio más alejado del centro el precio es menor.
```

```
new_data <- data.frame(neighbourhood = "Prosperidad",
room_type = "Entire home/apt", availibility = 365, minimum_nights = 20)

# Usar la función predict() para predecir el precio
predicted_price <- predict(model, new_data)
```

```r
cat("La predicción del precio para un apartamento entero en Prosperidad con \n
disponibilidad de 365 dias y que se reserve como minimo 20 noches es ",
round(predicted_price,2), "€. \n
Al aumentar el minimo numero de noches obligatorias que se debe reservar el \n
precio debe bajar para que sea más interesante. \n")
```

```
La predicción del precio para un apartamento entero en Prosperidad con

disponibilidad de 365 dias y que se reserve como minimo 20 noches es  145.01 €.

Al aumentar el minimo numero de noches obligatorias que se debe reservar el

precio debe bajar para que sea más interesante.
```

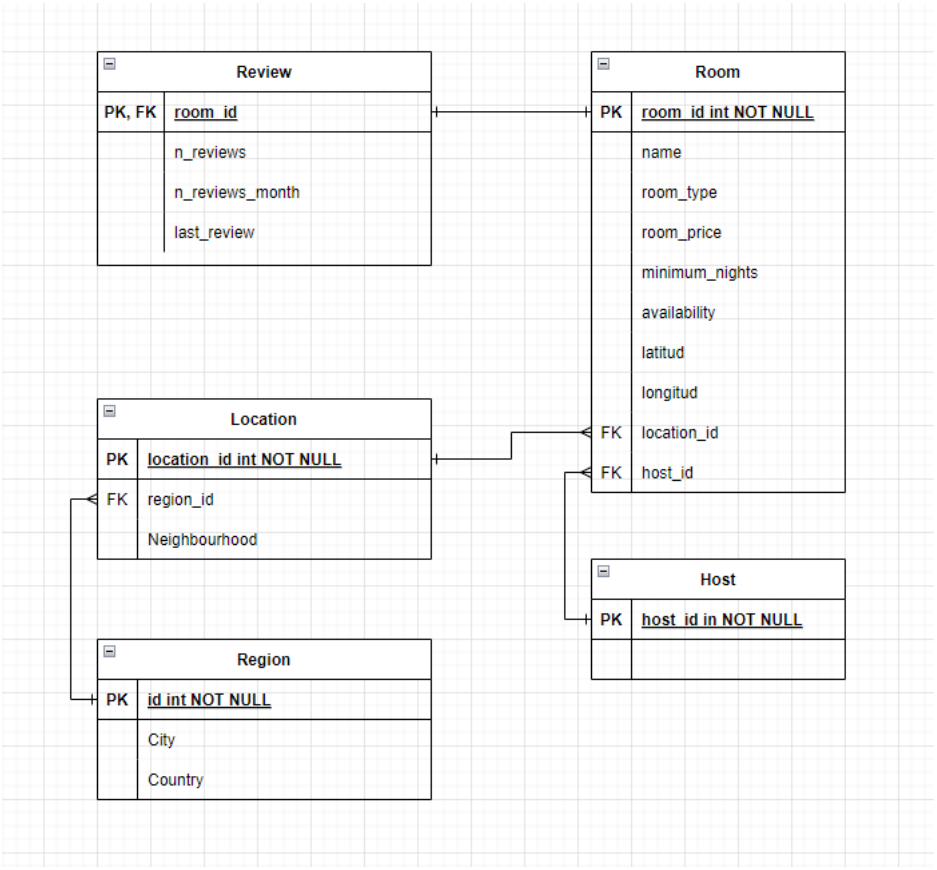
```r
new_data <- data.frame(neighbourhood = "Prosperidad", room_type = "Private room",
availibility = 365, minimum_nights = 1)

# Usar la función predict() para predecir el precio
predicted_price <- predict(model, new_data)
cat("La predicción del precio para una habitación privada en Prosperidad con \n
disponibilidad de 365 dias y que se reserve como minimo una noche es ",
round(predicted_price,2), "€. \n
Si el tipo de alojamiento es compartido aunque la habitación sea privada, el \n
precio de la misma disminuye notablemente. \n")
```

```
La predicción del precio para una habitación privada en Prosperidad con

disponibilidad de 365 dias y que se reserve como minimo una noche es  42.37 €.

Si el tipo de alojamiento es compartido aunque la habitación sea privada, el

precio de la misma disminuye notablemente.
```

Informe

En esta etapa se debe de simular la presentación de resultados en un entorno real de empresa. Suposiciones iniciales Cuales han demostrado ser válidas y cuáles no. ¿Por qué? Métricas seleccionadas: ¿han sido las correctas o no? ¿por qué?

Primeramente realizamos el diseño de las tablas que vamos a utilizar y las dividimos de la siguiente manera:

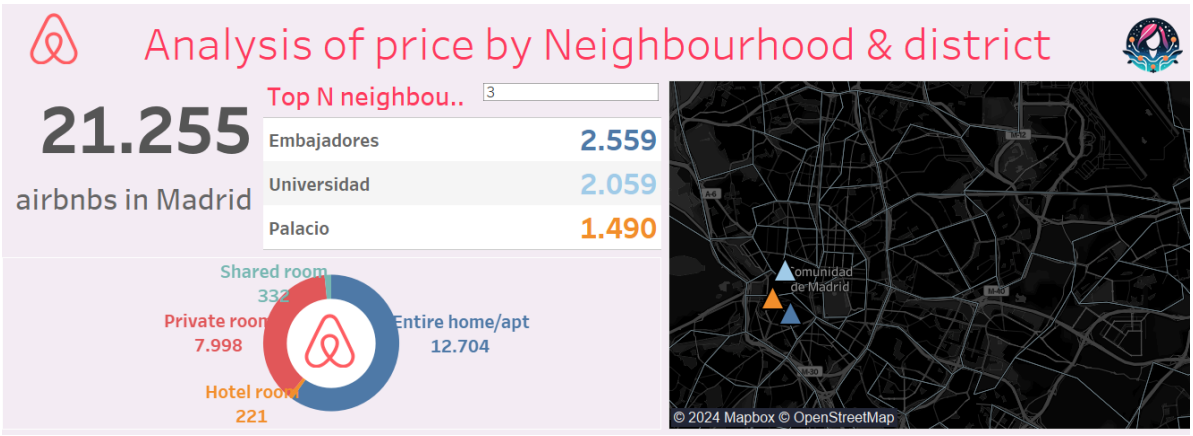


Trás analizar los datos facilitados en el dataset, nos enfocamos a responder preguntas de un futuro anfitrion que quiere conocer la rentabilidad que tendría a la hora de ofertar un alojamiento teniendo en cuenta varias metricas como la zona, tipo de alojamiento, etc…

Visualizamos en Tableau varias metricas que nos ayudan a poder aconsejar a futuros host.

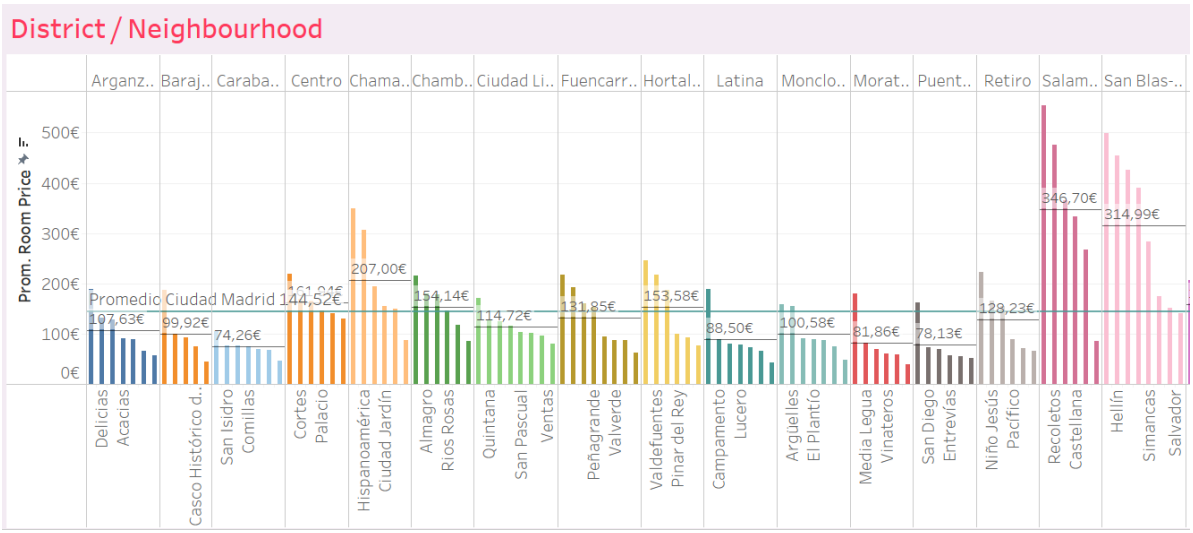
Para ello vemos primeramente la cantidad de Airbnb que hay en Madrid, 21.255 a fecha del dataset y vemos la cantidad de host que hay por barrios, siendo el barrio de Embajadores en el que más se concentran.

Tambien podemos ver que los alojamientos preferidos son el apartamento entero, seguido de la habitación privada.

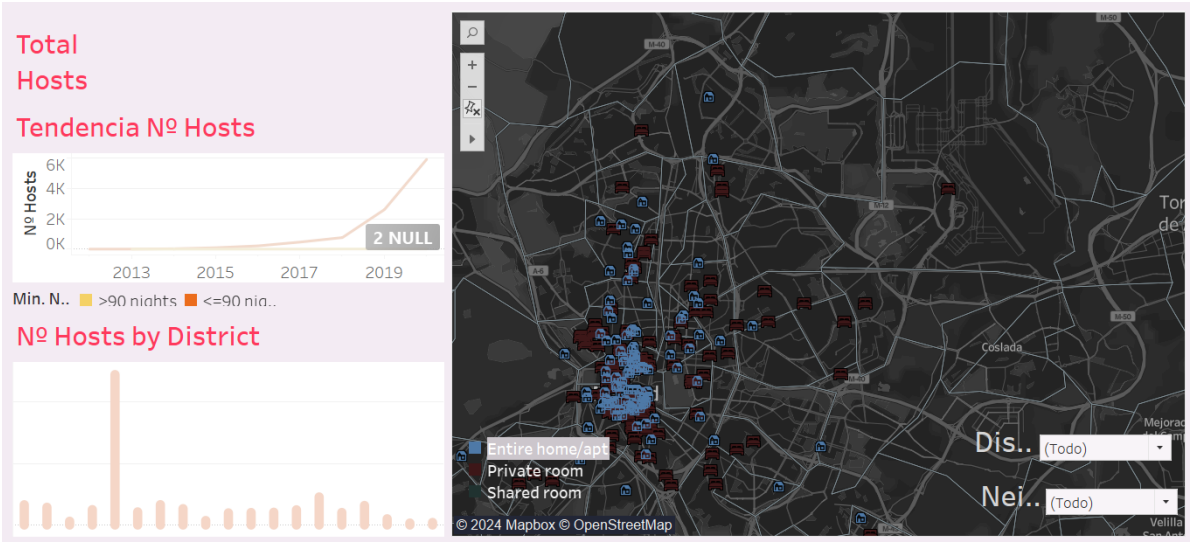


Otro punto importante es ver el promedio de precios por distrito y por barrio. Aqui vemos los barrios y distritos más caros y más baratos y cual es el promedio total de todo Madrid, que se encuentra en 144,52€ precio por noche.

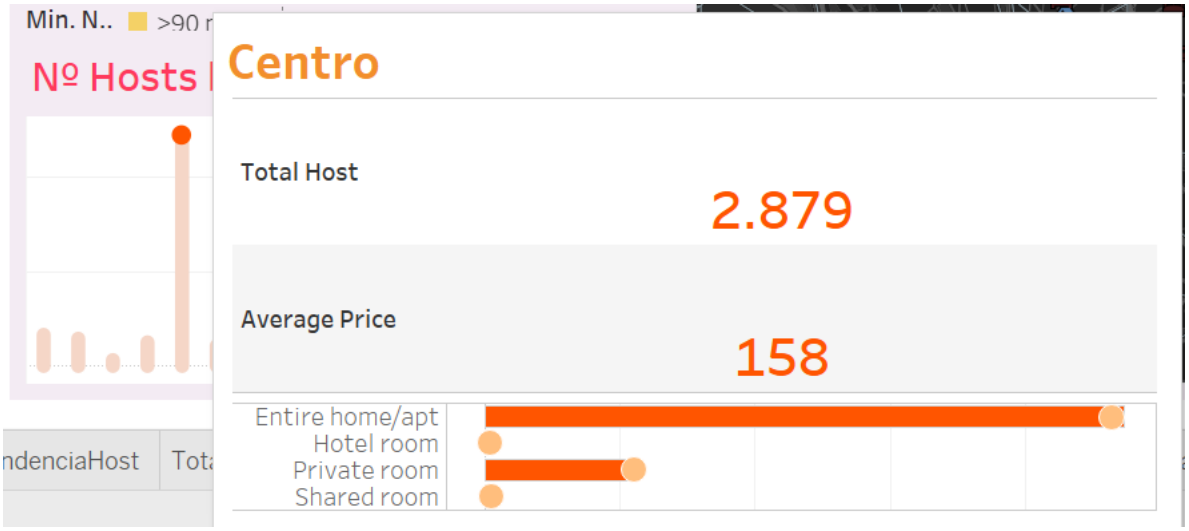
Siendo el precio medio más bajo el del barrio de Orcasur en Usera (31,96€) y el más alto el de Guindalera en el distrito de Salamanca (555,10€)



Continuamos viendo cual ha sido la tendencia de crecimiento en los ultimos años desde sus inicios

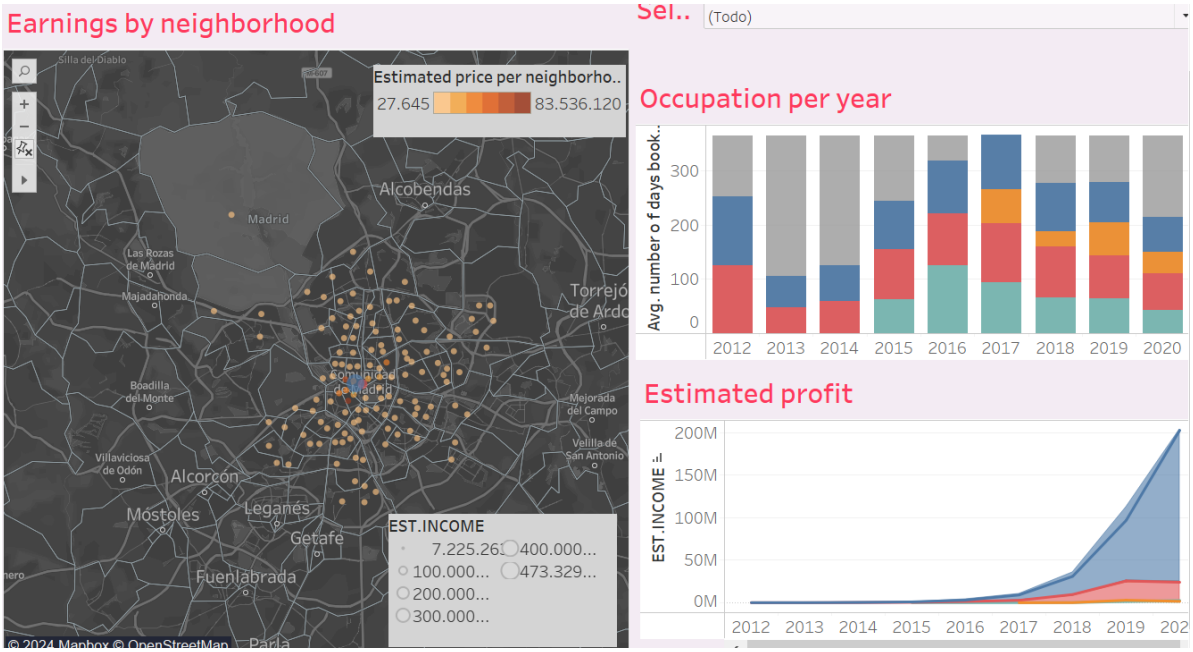


En este gráfico podemos ver la evolución desde los primeros Airbnb en Madrid en el año 2012. Existe un notable crecimiento a lo largo de los años y las zonas en las que más alojamientos hay por distrito, que se suele concentrar principalmente en el centro de la capital, debido a la amplia oferta turistica que existe.



Como se puede ver, solo el distrito centro cuenta con 2.879 host y la media de precio son 158€/noche siendo, como ya hemos mencionado, el apartamento completo el tipo de alojamiento preferido por los usuarios.

Una vez conocemos la tendencia, vamos a ver la media de ingresos por año en cada barrio



Como podemos comprobar, en los barrios más centricos es donde mayor media de ganancia existe al año, tambien porque es donde más alojamientos hay, por lo que debemos tener en cuenta este dato, ya que si se eligen estos barrios, contaremos con mayor competencia.

Por otro lado, nos aseguramos una mayor ocupación practicamente durante todo el año, ya que es la zona preferida por los turistas de todo el mundo. Como hemos comprobado, hay una notoria cantidad de alojamientos cuyos anuncios están en inglés y chino principalmente, aunque tambien hay anuncios en otros idiomas.

En la gráfica de "ocupacion al año" podemos apreciar que los primeros alojamientos que se ofrecian era el apartamento entero y habitación privada, y la tendencia fue cambiando a partir de 2015 donde apareció la habitación compartida y la habitación de hotel. No obstante, siguen siendo la habitación privada y el apartamento entero las estancias que más se alquilan, teniendo un promedio de ocupación de casi 200 dias al año.

La gráfica de "estimación de ganancias" nos muestra de un vistazo como se han incrementado las ganancias en los ultimos años.

Estos datos se pueden ver con mayor detalle en el archivo adjunto"Proyecto_final.twbx"

NORMATIVA DE NO ALQUILAR POR MAS DE 90 NOCHES

Según los datos estudiados en el dataset hay un porcentaje (6,7%) de viviendas que no cumplirian con la normativa conocida como "regla de los 90 dias"

El Plan Especial de Hospedaje (PEH), limita a 90 días la posibilidad de alquilar una vivienda con fines turísticos sin permiso y a partir de ese plazo obliga a obtener una licencia de uso terciario de hospedaje.

https://www.madrid.es/portales/munimadrid/es/Inicio/Vivienda-y-urbanismo/Urbanismo/Plan-Especial-de-regulacion-del-uso-de-servicios-terciarios-en-la-clase-de-hospedaje/?vgnextoid=b71cbc8d3c9f4610VgnVCM1000001d4a900aRCRD&vgnextchannel=2af331d3b28fe410VgnVCM1000000b205a0aRCRD

En el siguiente enlace podemos ver con más detalle la legislacion de la comunidad de Madrid sobre alquileres turisticos:

https://www.comunidad.madrid/servicios/consumo/normativa-alojamientos-turisticos

Conclusiones

- Con un precio más alto hay una mayor disponibilidad, el apartamento con el precio más bajo se alquila más y por lo tanto tiene menos disponibilidad.
- Si el tipo de alojamiento es una habitación de hotel, el precio de la misma aumenta notablemente.
- En un barrio más alejado del centro el precio es menor.
- Al aumentar el mínimo número de noches obligatorias que se debe reservar el precio debe bajar para que sea más interesante.
- Si el tipo de alojamiento es compartido aunque la habitación sea privada, el precio de la misma disminuye notablemente

Lecciones aprendidas

Aspectos Positivos y Recomendaciones para Repetir

- Exploración de Datos Exhaustiva:

Lo que funcionó: Realizar un análisis exploratorio de datos (EDA) fue crucial para entender las variables y su impacto en el precio de alquiler. Visualizar distribuciones y relaciones entre variables ayudó a identificar patrones y posibles transformaciones necesarias.

Repetir: Siempre realizar un EDA detallado al inicio de cualquier proyecto de análisis de datos para orientar mejor el enfoque del modelado.

- Preprocesamiento de Datos:

Lo que funcionó: Limpiar y preparar los datos, incluyendo el manejo de valores nulos, la conversión de tipos de datos y la codificación de variables categóricas, fue esencial para asegurar que el modelo funcionara correctamente.

Repetir: Implementar un proceso robusto de preprocesamiento de datos para garantizar la calidad y consistencia de los datos antes de aplicar cualquier técnica de modelado.

- Evaluación del Modelo con Métricas Apropiadas:

Lo que funcionó: Utilizar la RMSE (Raíz del Error Cuadrático Medio) como métrica para evaluar el desempeño del modelo proporcionó una medida clara de la precisión de las predicciones.

Repetir: Elegir y usar las métricas de evaluación correctas para el problema en cuestión. En problemas de regresión, la RMSE, MAE y R^2 son métricas fundamentales.

- Documentación y Comunicación de Resultados:

Lo que funcionó: Documentar cada paso del proceso y comunicar en el equipo ayudó a alinear expectativas y a demostrar el valor del análisis.

Repetir: Mantener una documentación detallada y comunicar los hallazgos de manera efectiva a todas las partes interesadas en el proyecto.

Áreas de mejora y nuevas estrategias

- Manejo de Datos Faltantes:

Lo que aprendimos: Algunos datos faltantes en columnas clave como date_last_review pueden afectar el rendimiento del modelo si no se manejan adecuadamente.

Mejorar: Implementar técnicas avanzadas de imputación para manejar datos faltantes, como la imputación múltiple o el uso de modelos predictivos para estimar valores faltantes.

- Validación del Modelo:

Lo que aprendimos: Validar el modelo solo con una división de los datos en entrenamiento y prueba puede no ser suficiente.

Mejorar: Utilizar validación cruzada (cross-validation) para evaluar el rendimiento del modelo de manera más robusta y asegurar que el modelo generaliza bien a datos no vistos.

- Experimentación con Diferentes Modelos:

Lo que aprendimos: Si bien el modelo lineal fue un buen punto de partida, explorar otros tipos de modelos podría haber mejorado la precisión de las predicciones.

Mejorar: Probar con diferentes algoritmos de machine learning, como Random Forest, Gradient Boosting Machines o modelos de redes neuronales, y comparar sus desempeños con el modelo lineal.