

PROYECTO FINAL BIG DATA BOOTCAMP GLOVO-MUJERES EN TECH KEEPCODING

Encuentra tu Air B&B en Madrid.

Grupo 9



Idea General

Realizar una herramienta que responde a las necesidades de búsqueda de un usuario de AirBnb que desea realizar una reserva en Madrid.

Data Set

Para el desarrollo de esta idea ha sido utilizada la base de datos de Airbnb - Listings publicada en el portal web de Open Data Soft.

Suposiciones Iniciales

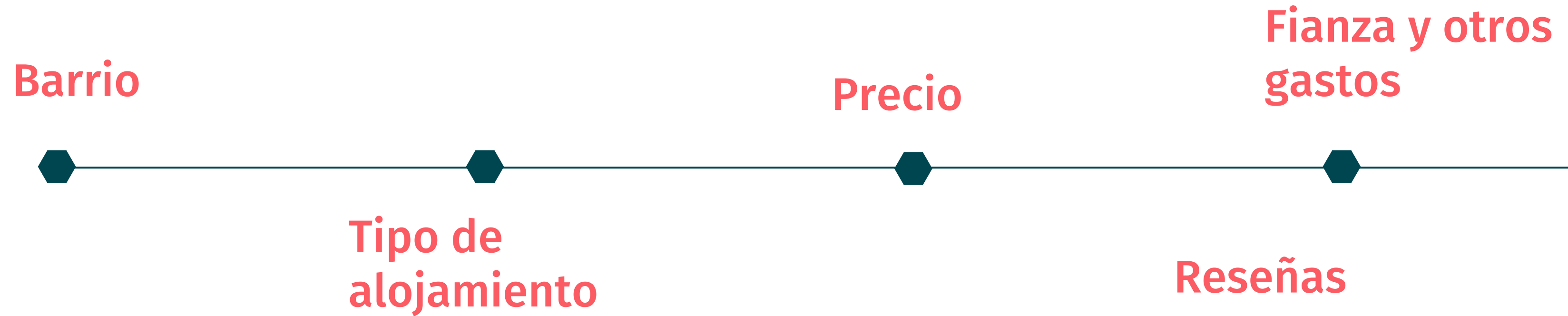
Correctas

- Tenemos una BBDD de la cual deben ser extraídos los datos. Para lograr esto, debemos tomar en consideración el tipo de alojamiento, su ubicación y su precio.

Incorrectas

- Debemos tomar en cuenta los detalles acerca del perfil del Host para que el usuario pueda filtrar por estos datos. Esta suposición resultó ser incorrecta en la práctica, ya que estos datos no aportan información vinculante en cuanto al tipo de alojamiento y precio.

Métricas



Arquitectura y validación de datos

Se creó un diagrama de Entidad-Relación con Drawio.
Luego se creó el script en DBeaver, donde luego también aparece el diagrama de ER.



Diagrama con Drawio

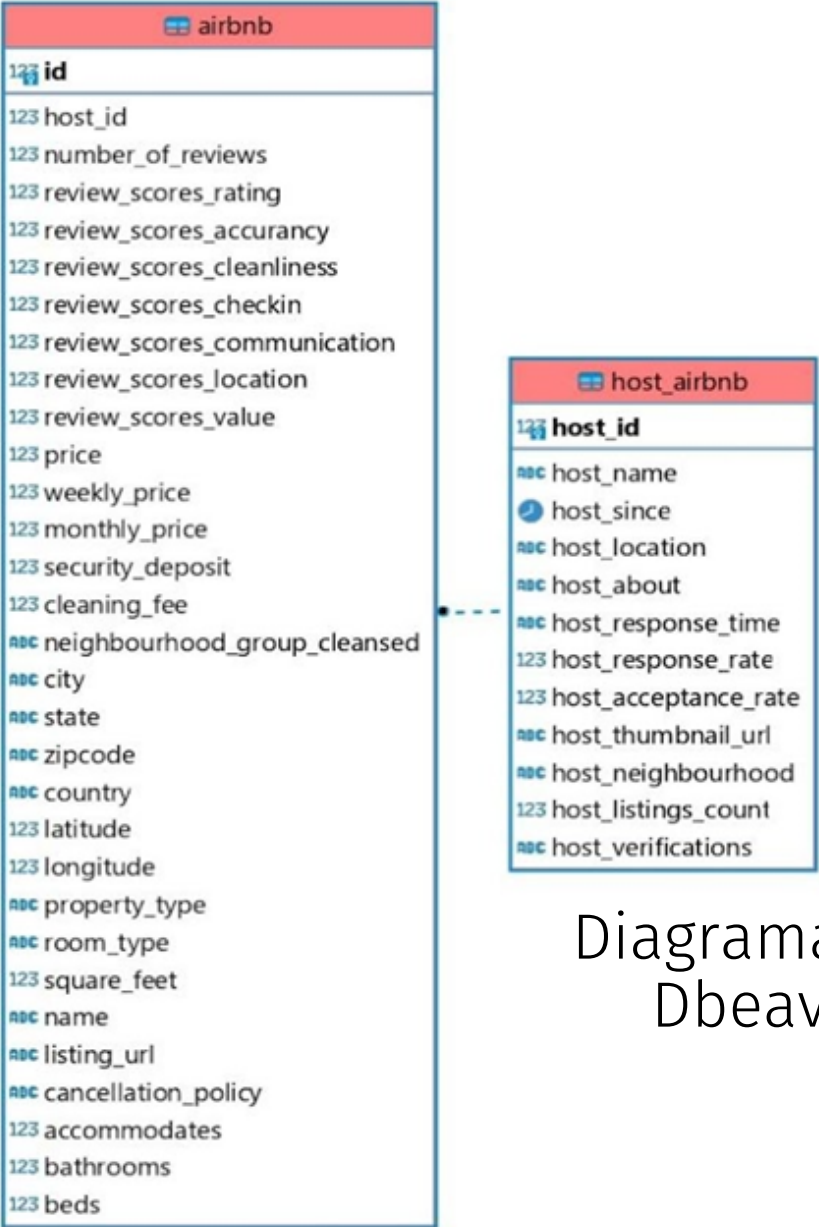


Diagrama con Dbeaver

SCRIPT

Creamos un esquema.

Añadimos la estructura.

Añadimos las PK y FK en las tablas correspondientes.

Análisis exploratorio

Tenemos un dataset con: Hacemos un df.shape para ver su forma y vemos que tiene 14780 filas y 89 columnas.

Hacemos un df.describe general para ver nuestros datos:

	ID	Host ID	Latitude	Longitude	...	Review Scores Checkin	Review Scores Communication	Review Scores Location	Review Scores Va
lue									
count	1.478000e+04	1.478000e+04	14780.000000	14780.000000	...	11443.000000	11460.000000	11440.000000	11439.000
000									
mean	1.028089e+07	3.608080e+07	40.497626	-3.858041	...	9.621778	9.647033	9.532168	9.218
201									
std	5.564829e+06	3.425360e+07	4.641387	14.123146	...	0.802736	0.767116	0.774527	0.950
578									
min	1.862800e+04	1.745300e+04	-37.851182	-123.131344	...	2.000000	2.000000	2.000000	2.000
000									
25%	5.554732e+06	6.787360e+06	40.409726	-3.707604	...	9.000000	9.000000	9.000000	9.000
000									
50%	1.133492e+07	2.464875e+07	40.419466	-3.700785	...	10.000000	10.000000	10.000000	9.000
000									
75%	1.532631e+07	5.432919e+07	40.430916	-3.684057	...	10.000000	10.000000	10.000000	10.000
000									
max	1.910969e+07	1.247534e+08	55.966912	153.371427	...	10.000000	10.000000	10.000000	10.000
000									

Tras la limpieza de las columnas, nos quedamos con un Dataset de 13306 filas y 45 columnas.

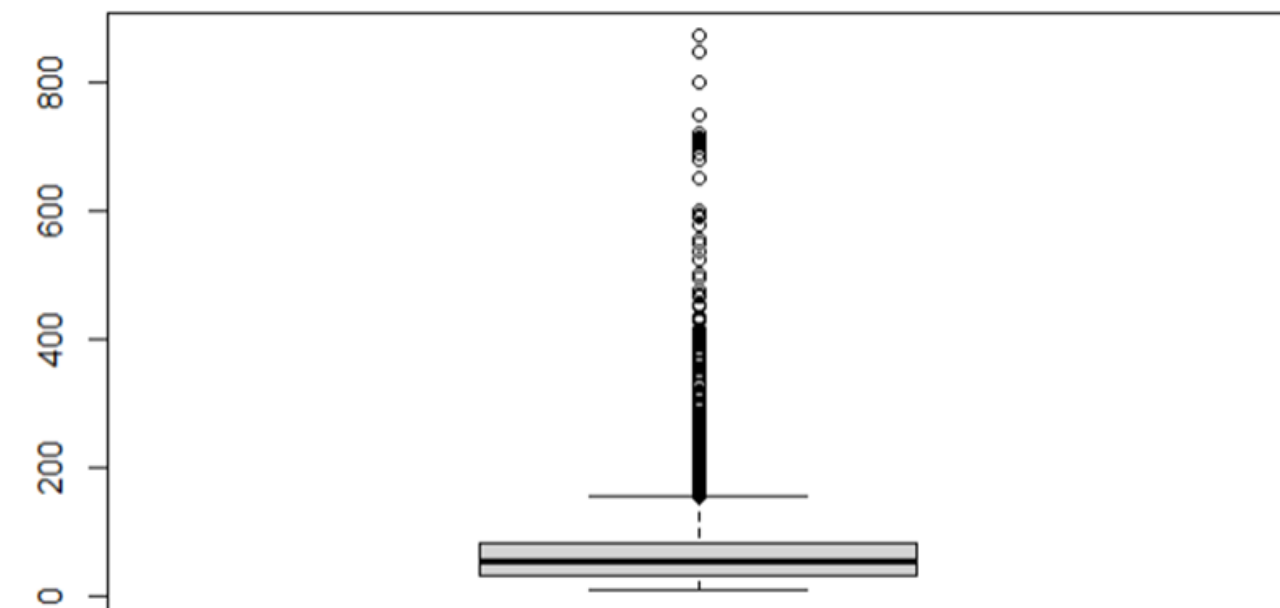
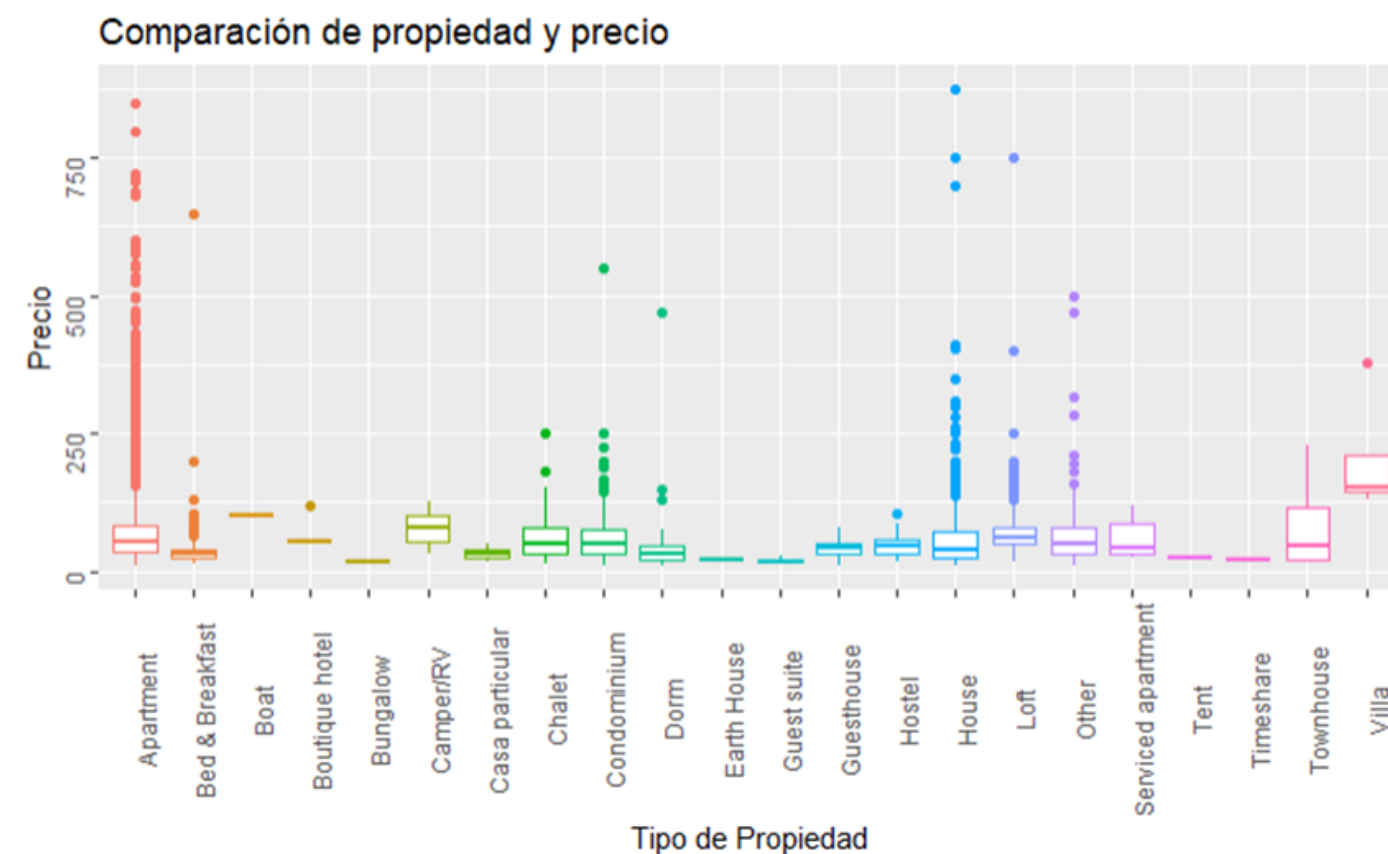
Análisis exploratorio

Empezamos haciendo un boxplot del precio, ya que consideramos que es la forma más visual de ver los valores que se alejan mucho de la media, o son muy diferentes al resto.

Obtenemos lo siguiente:

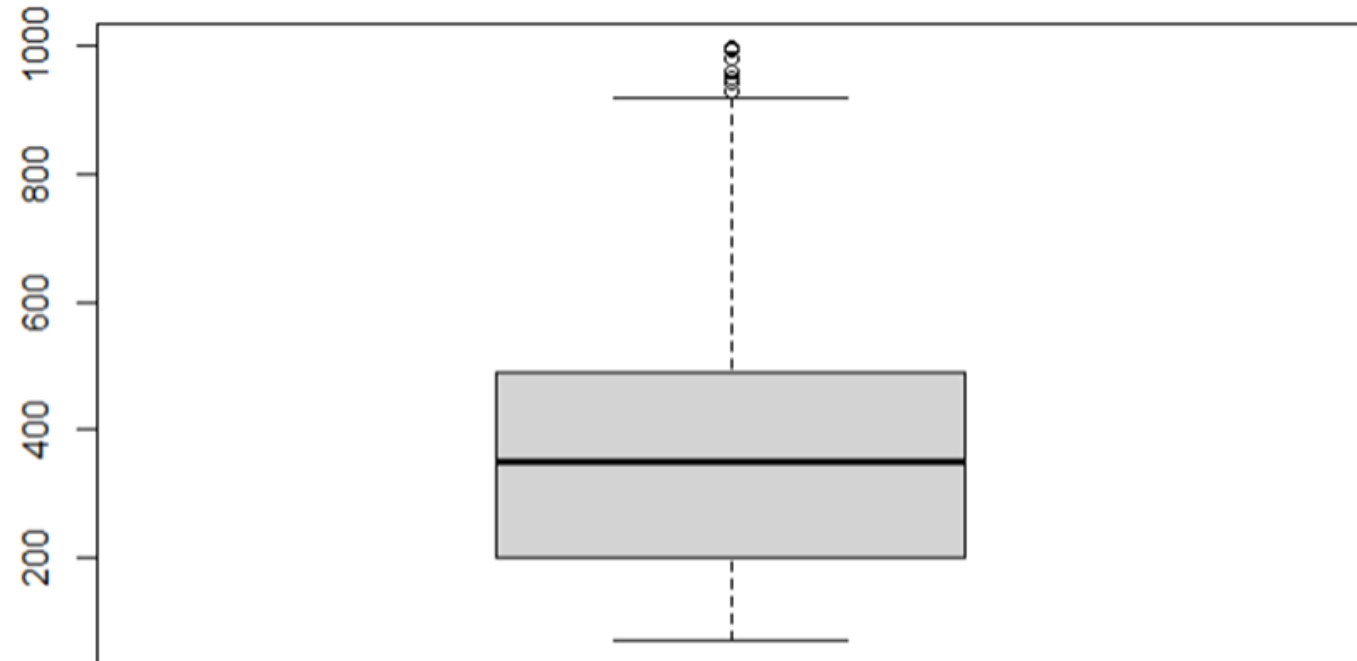
Este boxplot no nos da suficiente información para tomar decisiones acerca de los datos. Por ello hacemos otro boxplot en base al tipo de propiedad:

```
```{r}
boxplot(airbnbmadrid$Price)
```
```



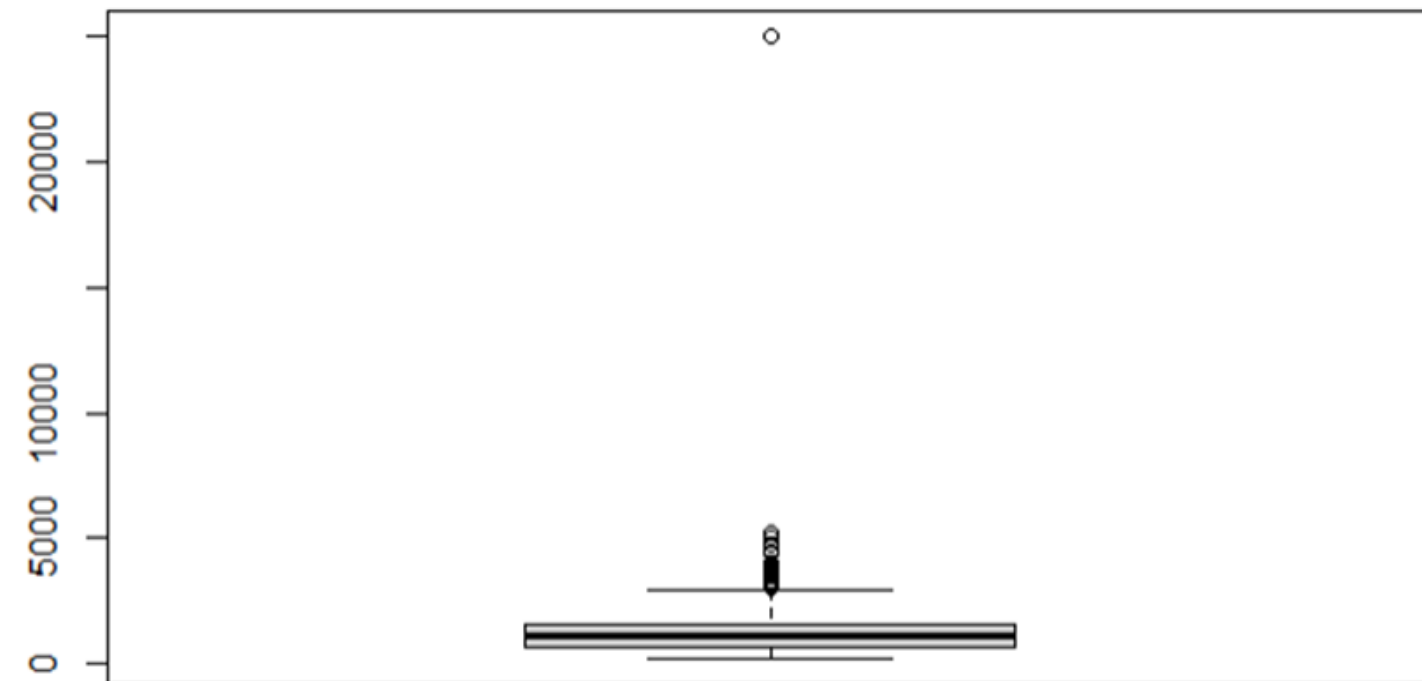
Análisis exploratorio

Ahora pasamos a ver el boxplot del Weekly Price. Decidimos basarnos en este porque contiene los datos reales del precio semanal establecidos por los dueños:



En cuanto al boxplot del Monthly Price:

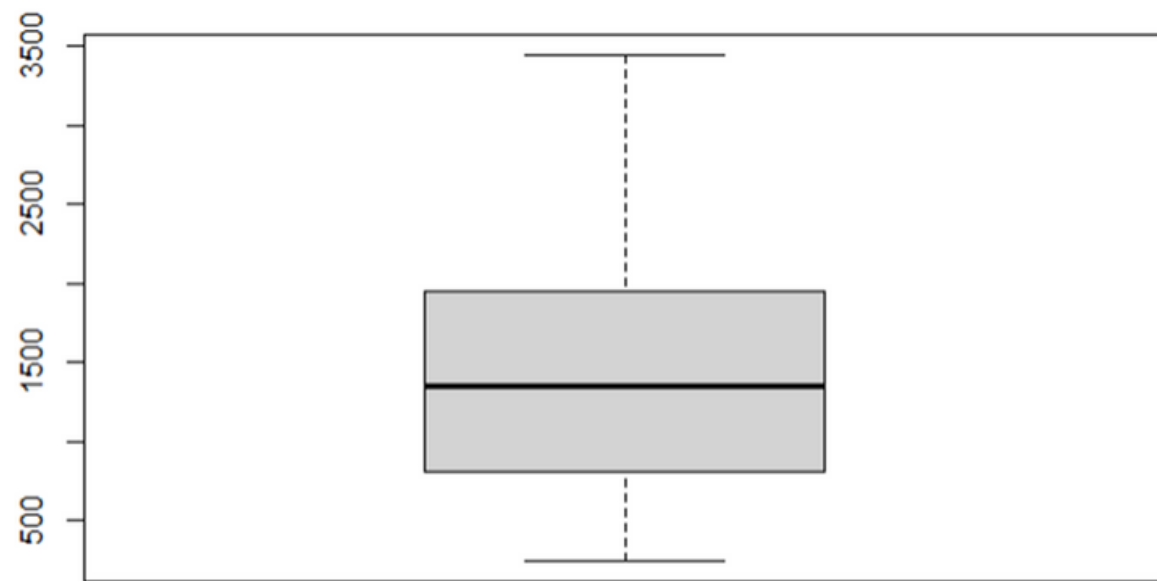
Tenemos al menos un valor que se aleja demasiado, por lo que procedemos a hacer una limpieza de ese outlier. Para poder realizar previsiones mas acertadas.



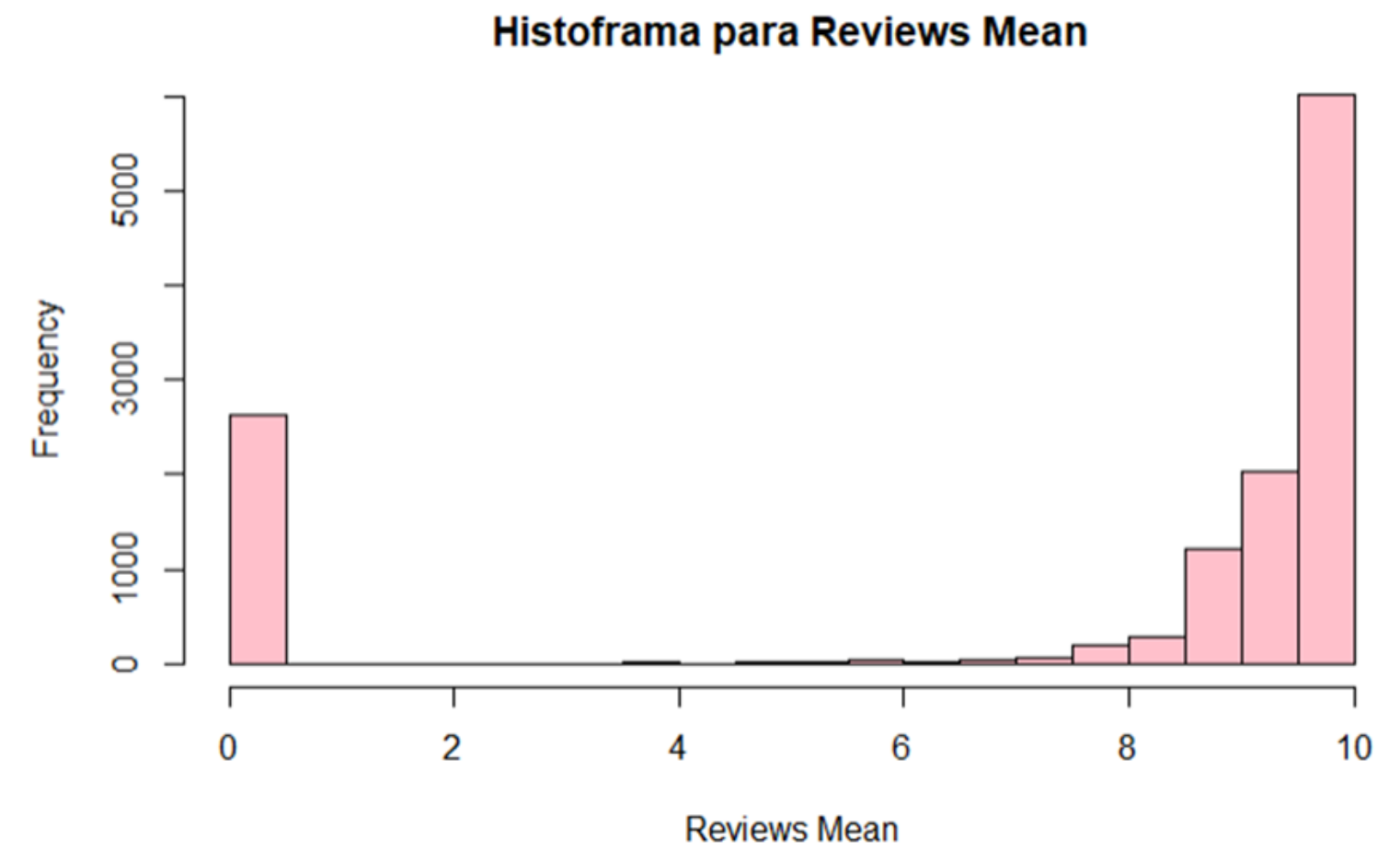
Análisis exploratorio

Decidimos hacer la limpieza sobre la columna nueva que hemos creado con los precios calculados para el mes y los precios establecidos por los dueños, ya que no tiene valores NA y es la que usaremos finalmente para la visualización y predicción.

Finalmente obtenemos este boxplot del Monthly Price New:



Por último, revisamos la distribución de la media de las reviews con un histograma.





VISUALIZACIÓN DE LAS MÉTRICAS

Visualización de los datos en Power BI

Origen de datos

El dashboard se alimenta del archivo excel: [airbnbmadrid_selected_excelfinal_v1.xlsx](#)

Filtros

Barrio

Todas

Tipo de alojamiento

Todas

Cantidad de camas

Todas

Precio noche

9 €

230 €

KPI's globales

| | | | |
|--------------------------|-------------------|------------------|------------------------|
| 12.539 mil | 12.535 mil | 7.5 | 56 € |
| Alojamientos encontrados | Hosts verificados | Puntuación media | Precio medio por noche |

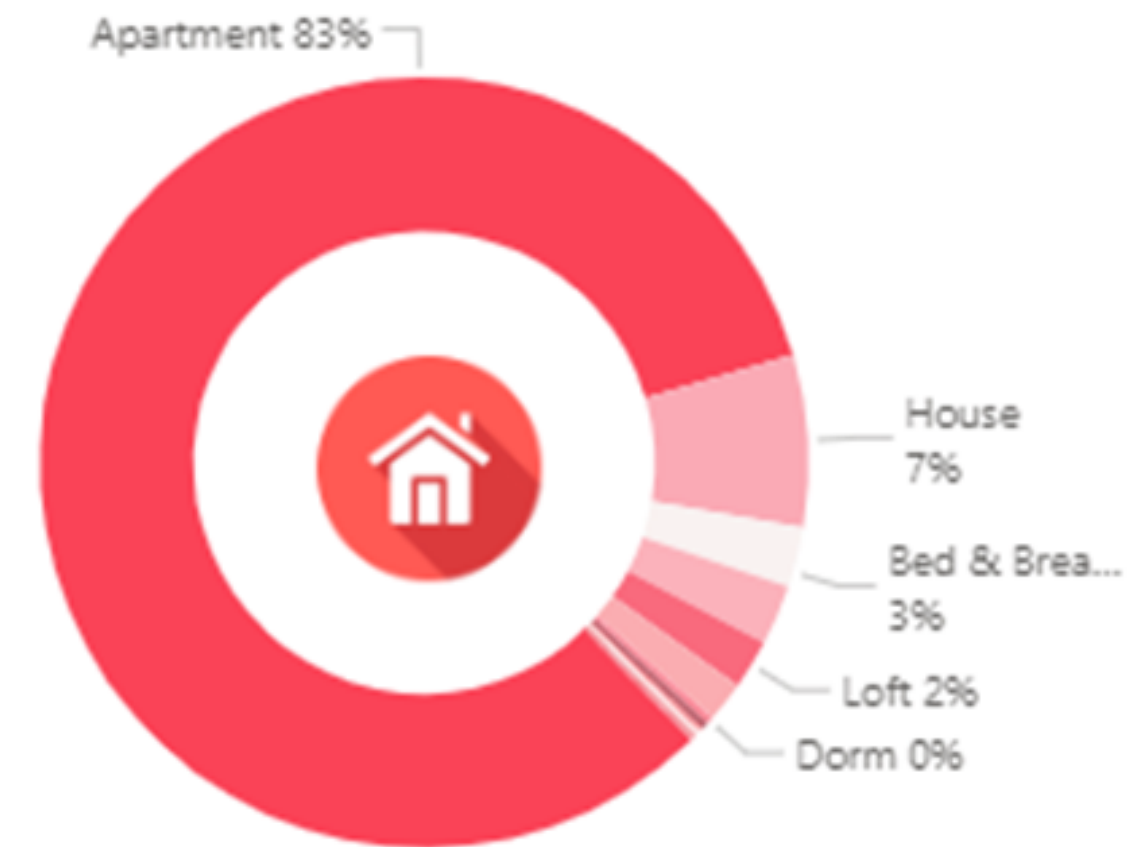
Visualizaciones

Top 10 barrios según promedio de puntuación

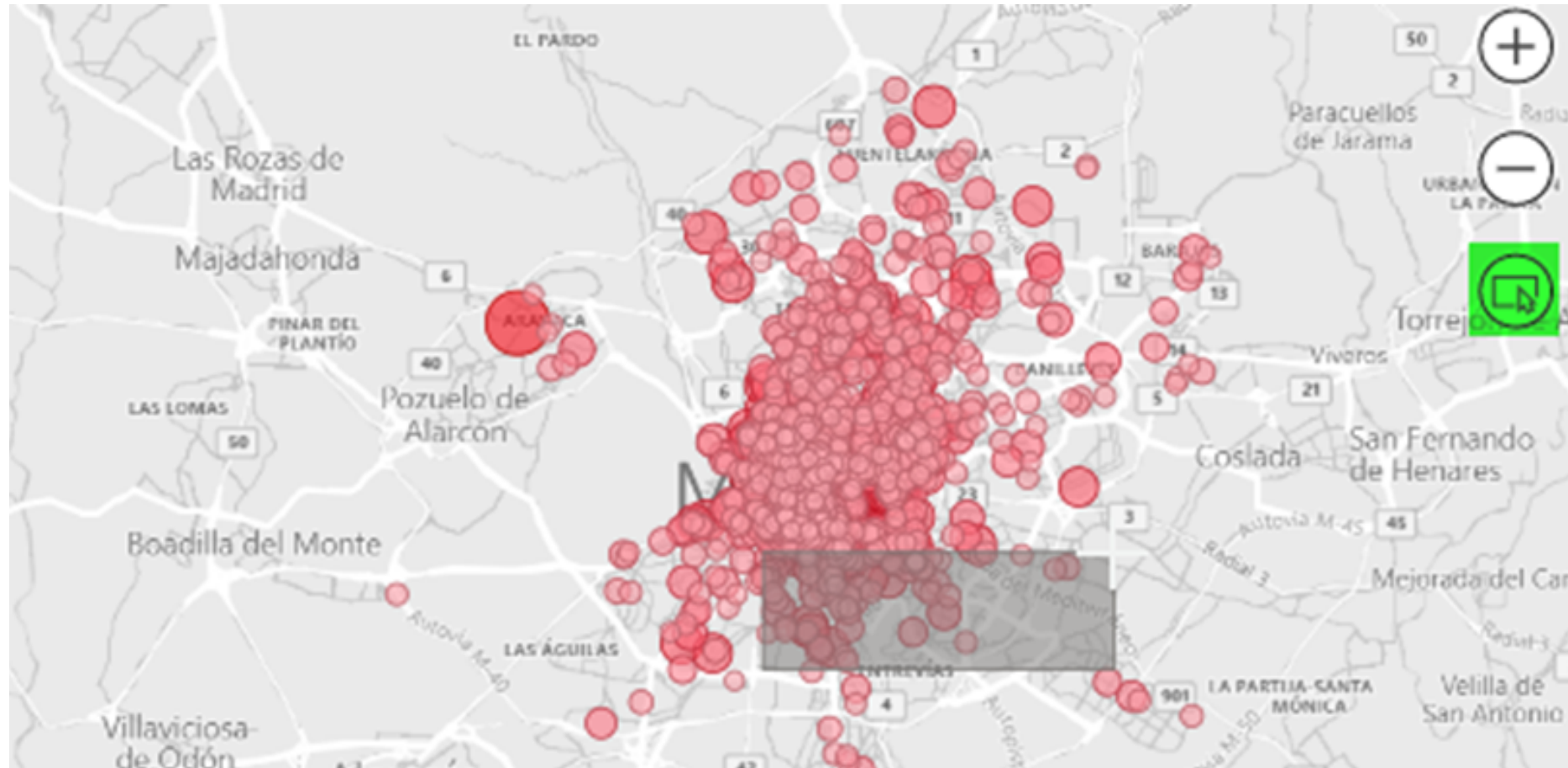


Tipos de propiedades por recuento ID

Tipo de propiedad



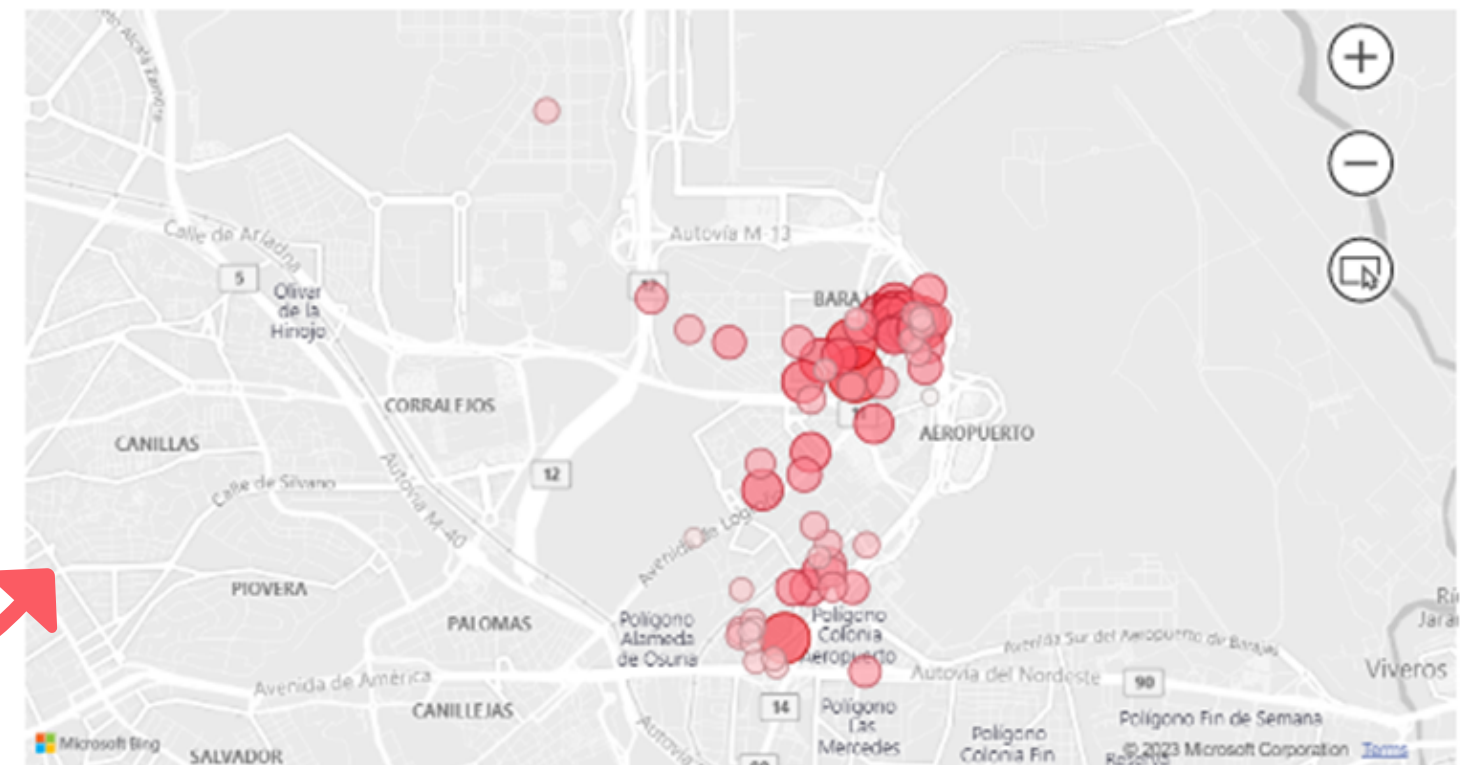
Visualizaciones



Mapa de Madrid basado en los campos: ID, Zipcode, Price, latitud y longitud.

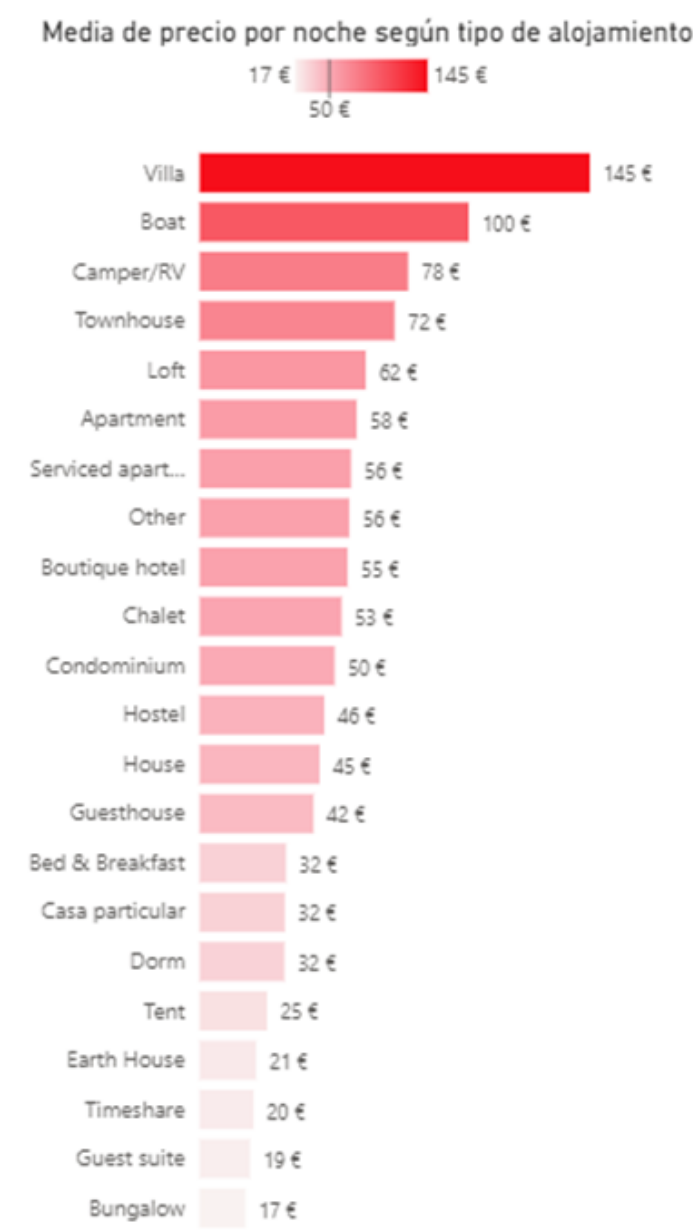
El tamaño de la burbuja indica el precio por noche

Ejemplo con filtros aplicados.
Barrio: Barajas
Precio maximo: 80€ la noche

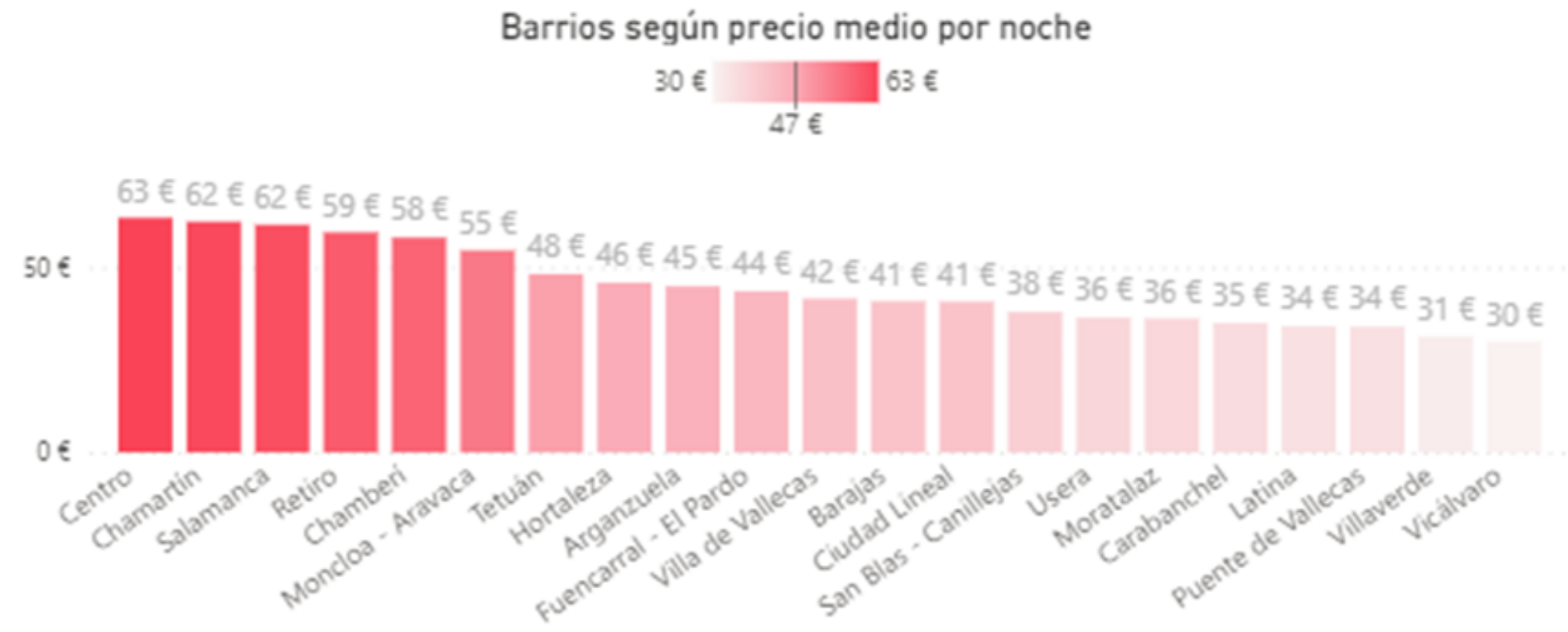


Visualizaciones

Precio medio por noche según el tipo de alojamiento



Precio medio por noche según la ubicación (barrio)



Visualizaciones

Tabla con detalles de precio y referencia del Host

| Precio por noche | Precio por una semana | Precio semanal final | Precio por un mes | Precio mensual final | Gastos de limpieza | Fianza | Host Name | Host URL |
|------------------|-----------------------|----------------------|-------------------|----------------------|--------------------|--------|----------------------------------|---|
| 70 € | 490 € | 490 € | 2,100 € | 2,100 € | 24 € | 150 € | Violeta | https://www.airbnb.com/users/show/99999180 |
| 42 € | 294 € | 294 € | 1,260 € | 1,260 € | 30 € | 100 € | Elisa | https://www.airbnb.com/users/show/99998670 |
| 32 € | 224 € | 224 € | 960 € | 960 € | 20 € | 0 € | Amanda | https://www.airbnb.com/users/show/99994045 |
| 45 € | 315 € | 315 € | 1,350 € | 1,350 € | 0 € | 0 € | Amanda | https://www.airbnb.com/users/show/99994045 |
| 80 € | 560 € | 560 € | 2,400 € | 2,400 € | 0 € | 350 € | Arizonica
Espacios Para Vivir | https://www.airbnb.com/users/show/99991901 |
| 100 € | 700 € | 700 € | 3,000 € | 3,000 € | 0 € | 350 € | Arizonica
Espacios Para Vivir | https://www.airbnb.com/users/show/99991901 |

Ejemplo de caso con descuento por tiempo de alquiler

| Precio por noche | Precio por una semana | Precio semanal final | Precio por un mes | Precio mensual final | Gastos de limpieza | Fianza | Host Name | Host URL |
|------------------|-----------------------|----------------------|-------------------|----------------------|--------------------|--------|-----------|---|
| 12 € | 84 € | 70 € | 360 € | 280 € | 10 € | 100 € | Miguel | https://www.airbnb.com/users/show/17371009 |

Visualizaciones

Valoraciones y su desglose



DEMO



Pre-procesamiento y modelado de datos

```
New names: [1] "ID"                  "Listing Url"
Cleansed"
[5] "City"                "State"
[9] "Country"             "Latitude"
[13] "Room Type"           "Price"
[17] "Security Deposit"    "Cleaning Fee"
[21] "Review Scores Accuracy" "Review Scores Cleanliness"
[25] "Review Scores Location" "Review Scores Value"
[29] "Host URL"            "Host Name"
[33] "Extras"              "Weekly Price Calculated"
[37] "Monthly Price New"   "Monthly Price Discounted"
[41] "Accommodates"       "Bathrooms"
```

```
      "Host ID...3"      "Neighbourhood Group"

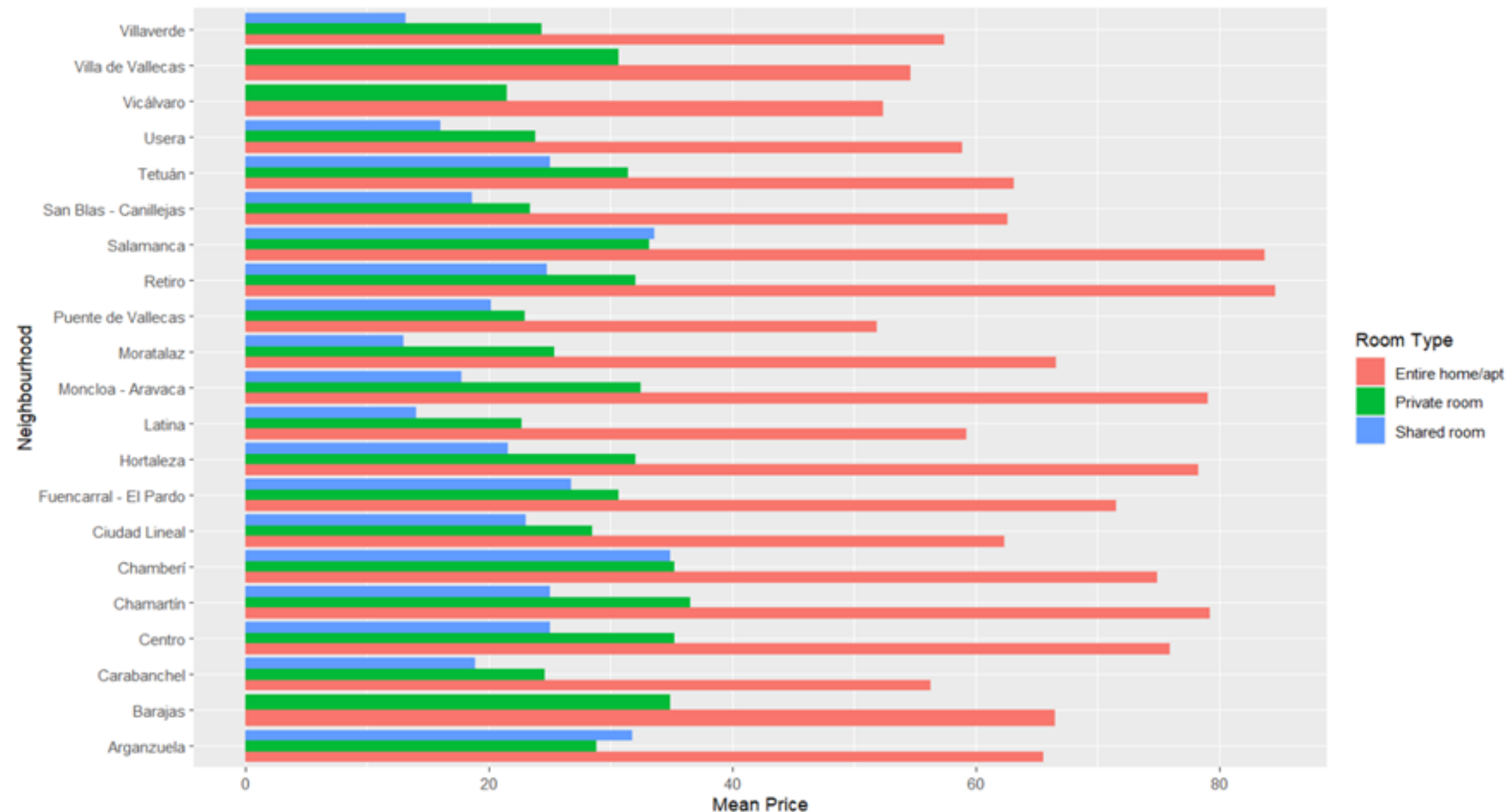
"Zipcode"                "Country Code"
"Longitude"              "Property Type"
"Weekly Price"           "Monthly Price"
"Number of Reviews"      "Review Scores Rating"
"Review Scores Checkin"  "Review Scores Communication"
"Cancellation Policy"    "Host ID...28"
"Host Since"             "Host Verifications"
"Weekly Price New"       "Monthly Price Calculated"
"Weekly Price Discounted" "Reviews Mean"
"Bedrooms"              "Beds"
```

| Var1 | Freq |
|-----------------------|-------|
| <fctr> | <int> |
| Centro | 6377 |
| Chamberí | 885 |
| Arganzuela | 785 |
| Salamanca | 760 |
| Tetuán | 444 |
| Retiro | 413 |
| Moncloa - Aravaca | 402 |
| Latina | 378 |
| Carabanchel | 351 |
| Chamartín | 340 |
| Ciudad Lineal | 302 |
| Puente de Vallecas | 217 |
| Hortaleza | 173 |
| Fuencarral - El Pardo | 147 |
| Usera | 144 |
| San Blas - Canillejas | 116 |
| Villaverde | 80 |
| Barajas | 74 |
| Moratalaz | 74 |
| Villa de Vallecas | 44 |
| Vicálvaro | 33 |

21 rows

Pre-procesamiento y modelado de datos

En la siguiente gráfica se pueden observar los valores promedios de barrio según su tipo de habitación.



Evaluar variables a incluir con toda la data y observamos cuáles son las que aportan mayor valor y mejoran el coeficiente de determinación ajustado.

```
Call:
lm(formula = log(Price) ~ Accommodates + Bedrooms + Beds + `Room Type` +
  `Neighbourhood Group Cleansed` + `Reviews Mean` + `Cancellation Policy`,
  data = df_airbnb)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|----------|----------|---------|---------|
| -1.86896 | -0.22289 | -0.01177 | 0.21931 | 1.81852 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|--|-----------|------------|---------|--------------|
| (Intercept) | 3.842562 | 0.016717 | 229.863 | < 2e-16 *** |
| Accommodates | 0.075809 | 0.003444 | 22.011 | < 2e-16 *** |
| Bedrooms | 0.117509 | 0.005705 | 20.598 | < 2e-16 *** |
| Beds | -0.034604 | 0.004119 | -8.402 | < 2e-16 *** |
| Room Type Private room | -0.674654 | 0.007958 | -84.776 | < 2e-16 *** |
| Room Type Shared room | -1.093774 | 0.026796 | -40.819 | < 2e-16 *** |
| Neighbourhood Group Cleansed Barajas | 0.104502 | 0.041533 | 2.516 | 0.011878 * |
| Neighbourhood Group Cleansed Carabanchel | -0.207922 | 0.021947 | -9.474 | < 2e-16 *** |
| Neighbourhood Group Cleansed Centro | 0.173077 | 0.013112 | 13.200 | < 2e-16 *** |
| Neighbourhood Group Cleansed Chamartín | 0.195331 | 0.022192 | 8.802 | < 2e-16 *** |
| Neighbourhood Group Cleansed Chamberí | 0.144105 | 0.016781 | 8.588 | < 2e-16 *** |
| Neighbourhood Group Cleansed Ciudad Lineal | -0.052532 | 0.023122 | -2.272 | 0.023107 * |
| Neighbourhood Group Cleansed Fuencarral - El Pardo | 0.032713 | 0.030715 | 1.065 | 0.286875 |
| Neighbourhood Group Cleansed Hortaleza | 0.094471 | 0.028691 | 3.293 | 0.000995 *** |
| Neighbourhood Group Cleansed Latina | -0.235242 | 0.021399 | -10.993 | < 2e-16 *** |
| Neighbourhood Group Cleansed Moncloa - Aravaca | 0.077999 | 0.021001 | 3.714 | 0.000205 *** |
| Neighbourhood Group Cleansed Moratalaz | -0.131404 | 0.041528 | -3.164 | 0.001559 ** |
| Neighbourhood Group Cleansed Puente de Vallecas | -0.235441 | 0.026189 | -8.990 | < 2e-16 *** |
| Neighbourhood Group Cleansed Retiro | 0.150211 | 0.020762 | 7.235 | 4.93e-13 *** |
| Neighbourhood Group Cleansed Salamanca | 0.181344 | 0.017419 | 10.411 | < 2e-16 *** |
| Neighbourhood Group Cleansed San Blas - Canillejas | -0.143656 | 0.033995 | -4.226 | 2.40e-05 *** |
| Neighbourhood Group Cleansed Tetuán | 0.017722 | 0.020284 | 0.874 | 0.382314 |
| Neighbourhood Group Cleansed Usera | -0.184004 | 0.030946 | -5.946 | 2.82e-09 *** |
| Neighbourhood Group Cleansed Vicálvaro | -0.276906 | 0.060683 | -4.563 | 5.09e-06 *** |
| Neighbourhood Group Cleansed Villa de Vallecas | -0.053437 | 0.052867 | -1.011 | 0.312148 |
| Neighbourhood Group Cleansed Villaverde | -0.273116 | 0.040308 | -6.776 | 1.29e-11 *** |
| Reviews Mean | -0.009525 | 0.000816 | -11.673 | < 2e-16 *** |
| Cancellation Policy moderate | -0.030850 | 0.008028 | -3.843 | 0.000122 *** |
| Cancellation Policy strict | -0.020970 | 0.007745 | -2.708 | 0.006787 ** |
| Cancellation Policy super_strict_30 | 0.092149 | 0.241399 | 0.382 | 0.702670 |
| Cancellation Policy super_strict_60 | 0.634502 | 0.171041 | 3.710 | 0.000208 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3411 on 12508 degrees of freedom
Multiple R-squared: 0.6616, Adjusted R-squared: 0.6608
F-statistic: 815.2 on 30 and 12508 DF, p-value: < 2.2e-16

Pre-procesamiento y modelado de datos

Una vez identificado las variables predictoras, seleccionaremos el barrio “Centro”, para ello creamos un subset.

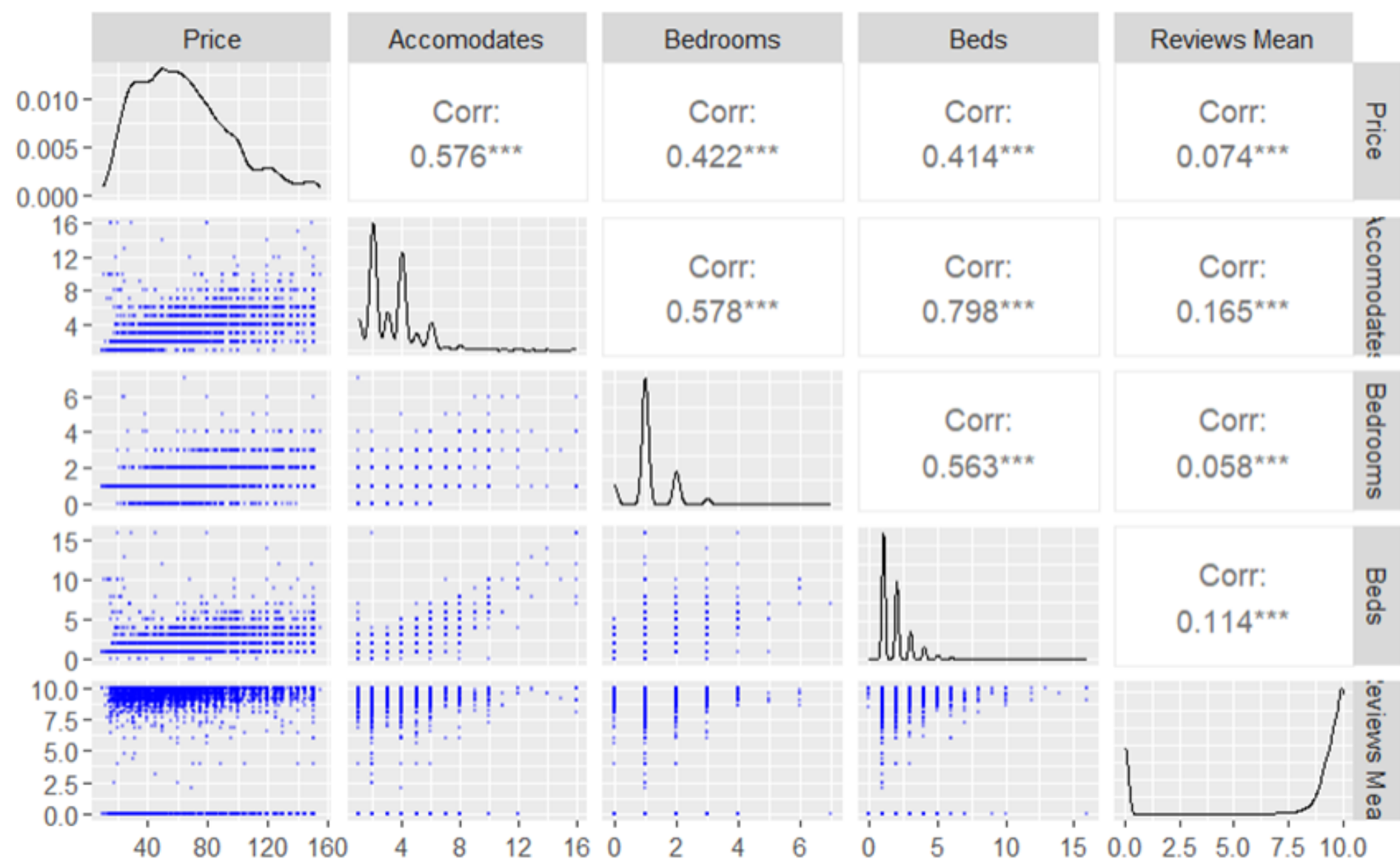
| ID
<dbl> | Listing Url
<chr> | Host ID...3
<dbl> | Neighbourhood Group Cleansed
<chr> | City
<chr> | State
<chr> | Zipcode
<chr> | Country Code
<chr> | |
|-------------|---------------------------------------|----------------------|---------------------------------------|---------------|---------------------|------------------|-----------------------|--|
| 15141125 | https://www.airbnb.com/rooms/15141125 | 96019257 | Centro | Madrid | Comunidad de Madrid | 28005 | ES | |
| 9470166 | https://www.airbnb.com/rooms/9470166 | 9885245 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |
| 17444981 | https://www.airbnb.com/rooms/17444981 | 118059488 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |
| 3284565 | https://www.airbnb.com/rooms/3284565 | 1892467 | Centro | Madrid | Community of Madrid | 28012 | ES | |
| 499911 | https://www.airbnb.com/rooms/499911 | 2467212 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |
| 1346747 | https://www.airbnb.com/rooms/1346747 | 7306349 | Centro | Madrid | Community of Madrid | 28005 | ES | |
| 3097553 | https://www.airbnb.com/rooms/3097553 | 15327748 | Centro | Madrid | Community of Madrid | 28012 | ES | |
| 13440784 | https://www.airbnb.com/rooms/13440784 | 76707968 | Centro | Madrid | Comunidad de Madrid | 28005 | ES | |
| 7818234 | https://www.airbnb.com/rooms/7818234 | 5239042 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |
| 1386096 | https://www.airbnb.com/rooms/1386096 | 6643556 | Centro | Madrid | Community of Madrid | 28012 | ES | |
| 14180827 | https://www.airbnb.com/rooms/14180827 | 18942409 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |
| 8011473 | https://www.airbnb.com/rooms/8011473 | 8831188 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |
| 13221821 | https://www.airbnb.com/rooms/13221821 | 74180884 | Centro | Madrid | Comunidad de Madrid | NA | ES | |
| 1942585 | https://www.airbnb.com/rooms/1942585 | 1528801 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |
| 9460773 | https://www.airbnb.com/rooms/9460773 | 15208964 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |
| 16311700 | https://www.airbnb.com/rooms/16311700 | 19725037 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |
| 5399733 | https://www.airbnb.com/rooms/5399733 | 3272228 | Centro | Madrid | Comunidad de Madrid | 28005 | ES | |
| 14773153 | https://www.airbnb.com/rooms/14773153 | 36963267 | Centro | Madrid | Comunidad de Madrid | 28005 | ES | |
| 1666184 | https://www.airbnb.com/rooms/1666184 | 8824421 | Centro | Madrid | Community of Madrid | 28005 | ES | |
| 17986327 | https://www.airbnb.com/rooms/17986327 | 15258781 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |
| 7796518 | https://www.airbnb.com/rooms/7796518 | 39840488 | Centro | Madrid | NA | 28013 | ES | |
| 12809312 | https://www.airbnb.com/rooms/12809312 | 2009482 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |
| 4196358 | https://www.airbnb.com/rooms/4196358 | 11488818 | Centro | Madrid | Comunidad de Madrid | 28013 | ES | |
| 13183219 | https://www.airbnb.com/rooms/13183219 | 73741764 | Centro | Madrid | Comunidad de Madrid | 28005 | ES | |
| 15508358 | https://www.airbnb.com/rooms/15508358 | 99632258 | Centro | Madrid | Comunidad de Madrid | 28013 | ES | |
| 11050879 | https://www.airbnb.com/rooms/11050879 | 57354901 | Centro | Madrid | Comunidad de Madrid | 28005 | ES | |
| 5547854 | https://www.airbnb.com/rooms/5547854 | 9193333 | Centro | Madrid | Comunidad de Madrid | 28012 | ES | |

1-27 of 6,377 rows | 1-8 of 44 columns

Previous 1 2 3 4 5 6 ... 38 Next

Pre-procesamiento y modelado de datos

Evaluamos nuevamente la tabla de correlación donde notamos que la variable Bathrooms no tiene correlación con la variable Price, por lo tanto, lo dejamos de considerar.



```
Call:
lm(formula = Price ~ Accomodates + Bedrooms + Beds + `Room Type`,
    data = train.airbnb)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|---------|--------|-------|---------|
| -86.182 | -12.999 | -2.917 | 9.445 | 114.415 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------------------|----------|------------|---------|--------------|
| (Intercept) | 48.1616 | 0.8777 | 54.875 | < 2e-16 *** |
| Accomodates | 4.6678 | 0.3484 | 13.397 | < 2e-16 *** |
| Bedrooms | 11.3484 | 0.5794 | 19.587 | < 2e-16 *** |
| Beds | -1.7274 | 0.4272 | -4.044 | 5.35e-05 *** |
| `Room Type`Private room | -31.5334 | 0.8021 | -39.315 | < 2e-16 *** |
| `Room Type`Shared room | -49.2710 | 2.8620 | -17.216 | < 2e-16 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

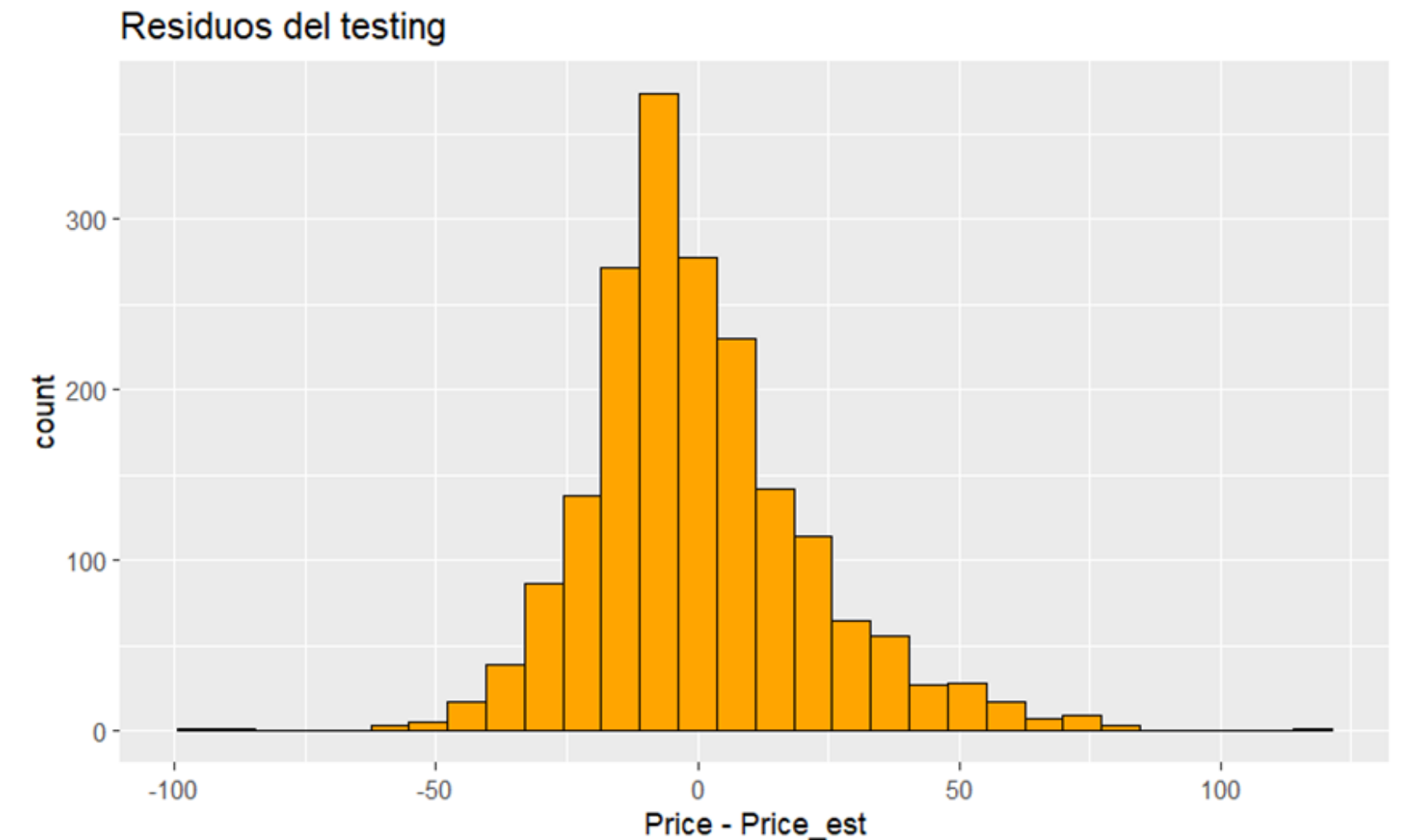
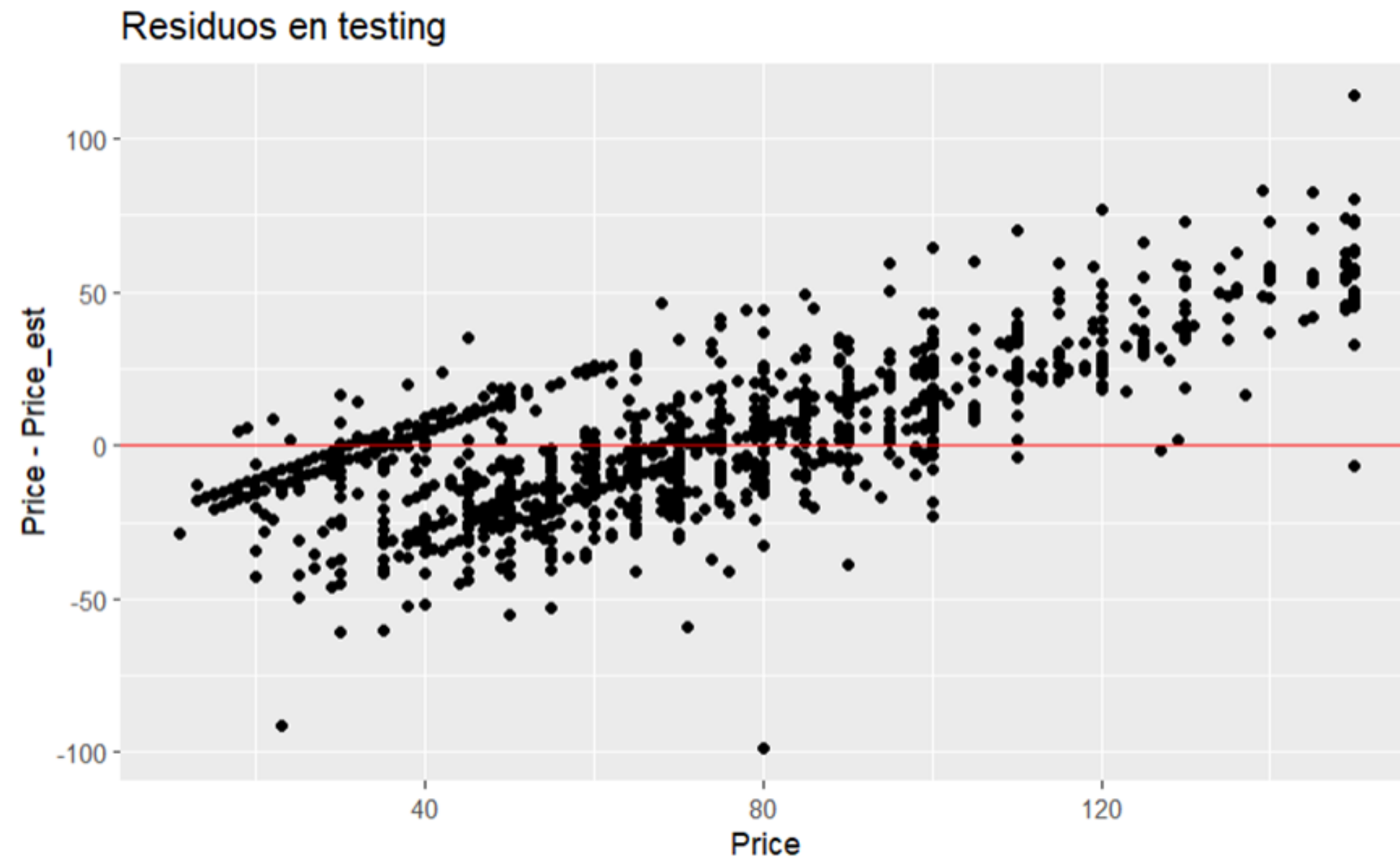
Residual standard error: 20.53 on 4457 degrees of freedom
 Multiple R-squared: 0.5435, Adjusted R-squared: 0.543
 F-statistic: 1061 on 5 and 4457 DF, p-value: < 2.2e-16

| | Accomodates | Bedrooms |
|-------------------------|-------------|-----------|
| (Intercept) | 48.161588 | 11.348429 |
| `Room Type`Private room | -31.533418 | |
| `Room Type`Shared room | -49.271012 | |

Evaluamos el modelo de regresión lineal con las variables a incluir de la data de entrenamiento.

Pre-procesamiento y modelado de datos


Luego de evaluar diferentes variables se ha mejorado el valor del coeficiente de determinación, aún así no llega a tener un valor para que el modelo sea aceptable, pasamos a evaluar el residuo o error del modelo el cual se asemeja a la campana de Gauss.



```
train.airbnb$Price_est <- predict(model_ev, train.airbnb)
caret::postResample(pred=train.airbnb$Price_est, obs = train.airbnb$Price)
```

| RMSE | Rsquared | MAE |
|------------|-----------|------------|
| 20.5176342 | 0.5435058 | 15.4028822 |

¿Que se haría igual y que se haría diferente?



Tomaríamos esta misma base de datos para trabajar, ya que tiene mucha información que se puede extraer para trabajar y sacar nuevas conclusiones sobre posibles búsquedas.

Para ampliar el desarrollo de esta herramienta, haríamos nuevos cálculos basados en los precios de alojamiento de otras ciudades y países, para ampliar las posibilidades de búsqueda a lugares fuera de Madrid.

Conclusiones y lessons learned

A través de los datos, a día de hoy es posible analizar y predecir el comportamiento que un usuario tendrá en la red, conocer qué piensan los clientes y usuarios sobre una marca o un producto, y cuáles son sus necesidades reales sobre la adquisición de productos o servicios.

El desarrollo de este proyecto nos ha dejado como enseñanza, la gran importancia que tiene aprovechar los datos y utilizarlos para identificar nuevas oportunidades. Eso, a su vez, conduce a movimientos de negocios más inteligentes, operaciones más eficientes, mayores ganancias y al alcance de grandes objetivos para el desarrollo de herramientas que aporten algo que resulte útil para la sociedad.

¡MUCHAS
GRACIAS!