MEMORIA TÉCNICA

Sara Arévalo García

Andrea Brea Rodríguez

Gabriela Schneider

Claudia Soliz Campos

Contenido

Introducción 2
Definición del Dataset 2
Arquitectura y validación de los datos (Python). 2
Estudio de los datos a utilizar 4
Visualización de las métricas. 5
Pre-procesamiento y modelado. 5
Pre-procesamiento. 5
1. Modelado: Algoritmo de regresión lineal simple 6
2. Modelado: algoritmo de regresión lineal múltiple. 9
Resultados finales 13

Introducción

El objetivo de este estudio es la evaluación mediante la recta regresión de un conjunto de datos de hospedajes de *airbnb* en la ciudad de Madrid. Para ello, se va a tener en cuenta lo aprendido en el módulo de data del Bootcamp Mujeres en Tech. Para lograr este objetivo, se siguen procesos de ETL, análisis del dataset y posterior estudio en los lenguajes tanto de R como Python.

Las herramientas utilizadas son, como motor de la base de datos *elephantsql* y como cliente DBeaver. En cuanto a los entornos de trabajo, se ha utilizado *Visual Studio Code* para la parte relativa a *Python* y *RStudio* para la parte que involucra a R. En cuanto a visualización de los datos se ha optado por un Dashboard en Tableau.

Definición del Dataset

El Dataset elegido para este proyecto es "airbnb_listings_Madrid", los datos se han obtenido scrapeando la plataforma Airbnb. En la base de datos tenemos observaciones de diferentes países y ciudades, pero para este proyecto nos vamos a centrar en las observaciones que hacen referencia a Madrid.

Arquitectura y validación de los datos (Python).

Para poder extraer los datos del archivo csv se realizó un script en Python, utilizando la librería panda. Se eligió como "delimeter"= "**;**", separando así en columnas todo lo que estaba entre puntos y comas.

Para poder decidir qué transformaciones eran necesarias, se visualizó la nueva tabla en Jupyter notebook y se realizó una exploración inicial de los datos.

Procedimos a realizar la limpieza seleccionando las columnas de interés. No tuvimos en cuenta las siguientes variables:

Listing Url, Thumbnail Url, Medium Url, Picture Url, Xl Picture Url, Host Url, Host thumbnail Url, Host picture Url, para este análisis de datos no necesitamos como dato un link o url.

Name, es el título del anuncio en la plataforma y no es relevante para este análisis.

Summary, *Space*, *Description*, *Host About* y *Features* son descripciones de los inmuebles y del propietario, tiene objetivo más publicitario que informativo,

Experienced Offered, no es relevante, en la mayoría el campo está vacío o "none".

Neighborhood Overview, descripción del barrio con motivo publicitario, para el análisis ya hay unas variables como el barrio o el distrito que nos resulta más útil.

Notes, Transit, Access, Interaction, House rules, son variables que añaden información a las descripciones del anuncio, pero irrelevantes para este análisis.

Host Name, en un principio mantenemos el *Host ID*, por ello el nombre solo nos llevaría a confusión en el análisis.

Host acceptance rate, *License* y *Jurisdiction names,* son variables con información sólo para algunos inmuebles en Estados Unidos, por lo tanto no aporta información en este análisis.

Host listing count, repite información de otra variable.

Host verifications, medios de contacto verificados por el propietario, no aporta información a este análisis.

Smart location no aporta más información que la que aporta Neighborhood.

Country, no es relevante porque ya contamos con el código del país, además en este caso todos los alojamientos pertenecen al mismo país

Calendar updated, no es relevante para este análisis.

Has availability, variable completamente vacía.

Availability 30, 60, 90, 365, variables innecesarias para este análisis.

Review scores accuracy, Review scores cleanliness, Review scores checkin, Review scores communication, Review scores location, Review scores value, para todas estas variables ya tenemos la variable de puntuación principal, la cual ya tiene en cuenta el desglose de puntuaciones.

Calculated host listings count, no es una variable que aporte información relevante al análisis.

Geolocation, la información de esta variable ya está correctamente desglosada en latitud y longitud.

Se realizó un segundo filtro en la columna de “City” para quedarnos sólo con los datos que sean de “Madrid”.

Se verificó que el tipo de datos sea correcto para las columnas utilizadas en los modelos estadísticos y visualización (float, int, string). También se cuantificó la cantidad de Null por columna. (El jupyter notebook utilizado “exploracion_inicial_datos.ipynb” , se encuentra en el repositorio en GitHub)

Por último se cargó a la base de datos Postgres en Elephant, para que desde allí se pudieran tomar los datos refinados en los estudios sucesivos. El script es “airbnb_etl.py” y se encuentra en Github, en el repositorio enviado por el formulario.

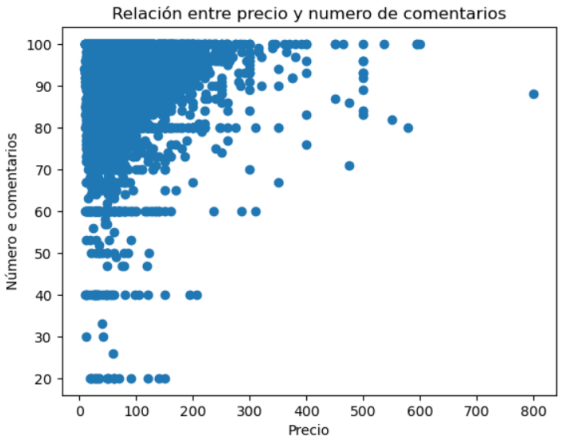
La validación de que la carga de la tabla se realizó correctamente se llevó a cabo visualmente con queries SQL con DBeaver

Estudio de los datos a utilizar.

Para llevar a cabo el estudio se van barajado varias posibilidades, que se nombran a continuación:

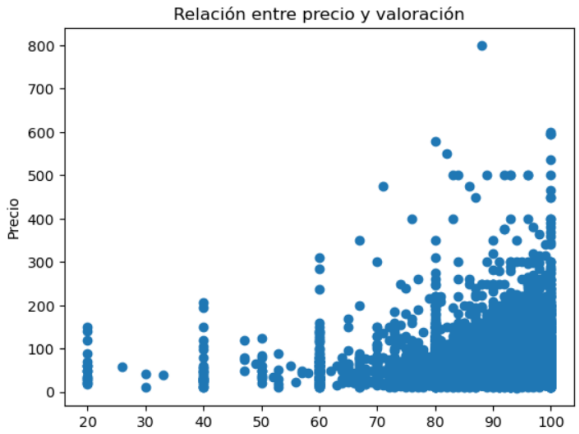
1. Relación entre número de mensajes y precio.

   Este modelo no resultó satisfactorio, la correlación de Pearson era -0.033, lo que resulta un coeficiente de determinación R^2 de 0.001. Para verificar estos datos se ha podido ver en una gráfica donde se puede ver la poca relación que tienen entre ellos.



2. Relación entre puntuación media del alojamiento y el precio

   Este modelo asemejaba algo parecido al anterior, el valor de la correlación de Pearson era de 0.055, lo que daba un coeficiente de determinación R^2 de 0.003. En este caso los valores de las varianzas eran más similares, pero con las correlaciones, se pudo ver que guardan poca relación.



Visualización de las métricas.

Para la visualización del dashboard preparado, hemos usado la herramienta de Tableau. Hemos realizado la conexión con la base de datos creada de tal forma que se vayan actualizando los datos de forma automática según se vayan modificando en la propia base de datos.

Hemos utilizado la función de campos calculados para convertir el tamaño de las propiedades en m2, ya que la medida facilitada eran pies, una medida con la que no estamos familiarizadas para poder analizar los resultados.

En el dashboard se han incluido los siguientes 3 gráficos:

- Mapa de precios: se puede ver de forma rápida información relativa a los precios medios (con diferentes tonalidades de color) y en cantidades de propiedades según el código postal (con diferentes tamaños en los círculos). Se han añadido como etiquetas los diferentes barrios y se permite la interacción con el usuario, que podrá filtrar en función de sus preferencias por tamaño de la propiedad y tipo (se pueden elegir varias opciones). También se puede obtener información adicional en forma de descripción emergente sobre los promedios de los fees de limpieza, de la fianza y el coste total. También se muestra un pequeño gráfico de barras donde se muestra qué tipo de habitación en esa zona tiene las mejores reviews.
- Pequeña muestra filtrado por el barrio “Centro” en consonancia con el análisis de regresión lineal realizado en Python, donde se pueden observar los m2 junto con los precios medios pasando el ratón por cada código postal del mapa.
- Gráfico de barras con los tipos de habitación que tienen las mejores valoraciones.

**Datos para la conexión a la base de datos: en caso de requerirlo, se pasará la contraseña por privado.*

user: phmkeaeu

password: ****

host: trumpet.db.elephantsql.com

port: 5432

dbname: phmkeaeu

Pre-procesamiento y modelado.

Pre-procesamiento.

Se ha desarrollado un estudio de algoritmo de regresión lineal simple en Python y otro múltiple en R que se describen a continuación.

1. Modelado: Algoritmo de regresión lineal simple

Para llevar a cabo el estudio en Python se han utilizado las librerías de Python, *pandas, numpym*, *matplotlib*, psycopg2, *pandas.io.sql y sklearn.*

En primer lugar, hubo que descargar el dataframe procedente de la base de datos, este proceso se lleva a cabo con la librería psycopg2 creando una conexión con *DBeaver*.

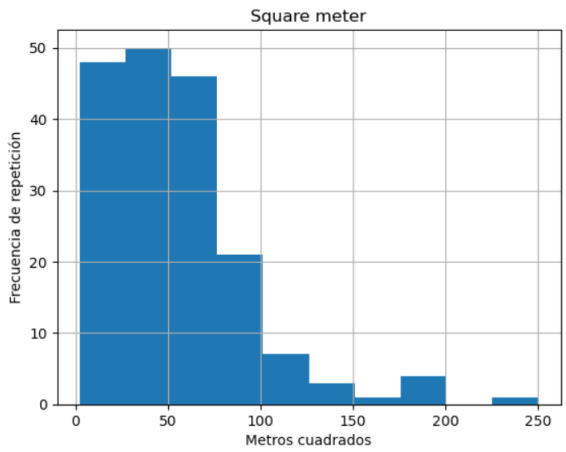
Ya con un *dataframe* con el que trabajar, hacemos un análisis de los datos obtenidos para poder representar mejor la recta. Como no es de interés en la base de datos perder valores que pueden resultar de importancia, se hace en esta etapa para intentar aproximarse a un estudio lo más real posible.

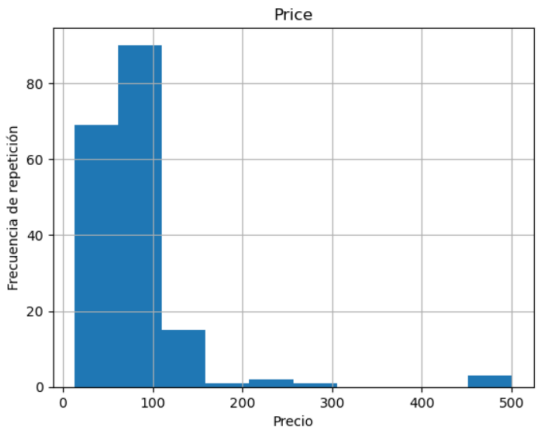
Dado que, la cantidad de muestras es muy extensa, se hace un filtrado por un distrito, en este caso el que más muestras contiene es *Centro* con 6760 muestras. Por otro lado, las estadísticas muestran varios valores a tener en cuenta. Uno de ellos es el valor mínimo de pies cuadrados que resulta 0 siendo una medida muy poco realista.

	Price	Square Feet	Neighbourhood Group Cleansed
count	13198.000000	519.000000	13207
unique	NaN	NaN	21
top	NaN	NaN	Centro
freq	NaN	NaN	6760
mean	65.924686	378.007707	NaN
std	56.008552	546.793839	NaN
min	9.000000	0.000000	NaN
25%	31.000000	0.000000	NaN
50%	52.000000	108.000000	NaN
75%	80.000000	646.000000	NaN
max	875.000000	5167.000000	NaN

Por ello, para el análisis de regresión lineal, se ha elegido como variable independiente (eje x) los metros cuadrados del hospedaje y como variable dependiente (eje y) el precio por noche de la vivienda.

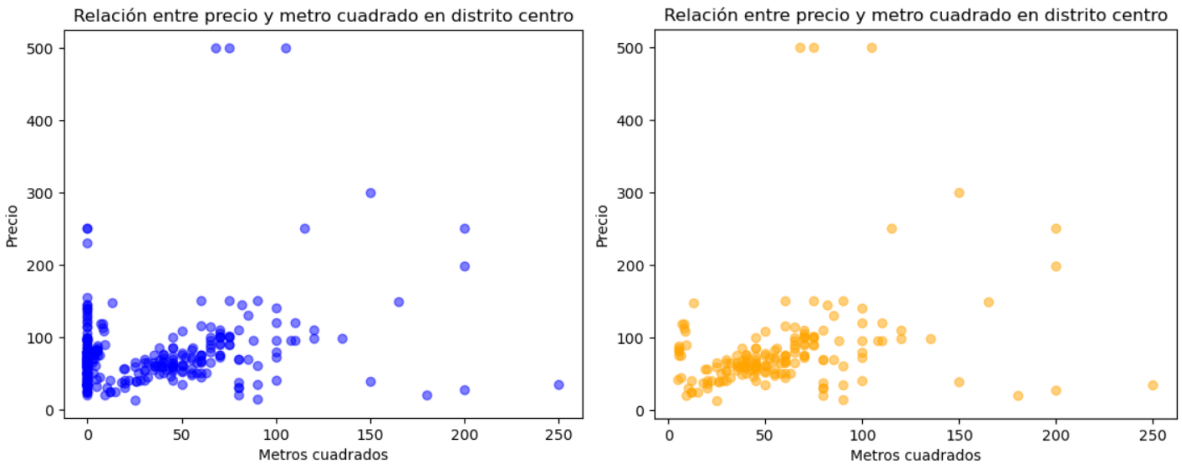
Otro parámetro a tener en cuenta a la hora de representar los datos seleccionados son los histogramas.





Puede verse en los anteriores histogramas que la mayor concentración de metros cuadrados se encuentra en hospedajes menores de 100 metros cuadrados y la mayoría de precios son también inferiores a 100 euros, siendo el precio medio 56€.

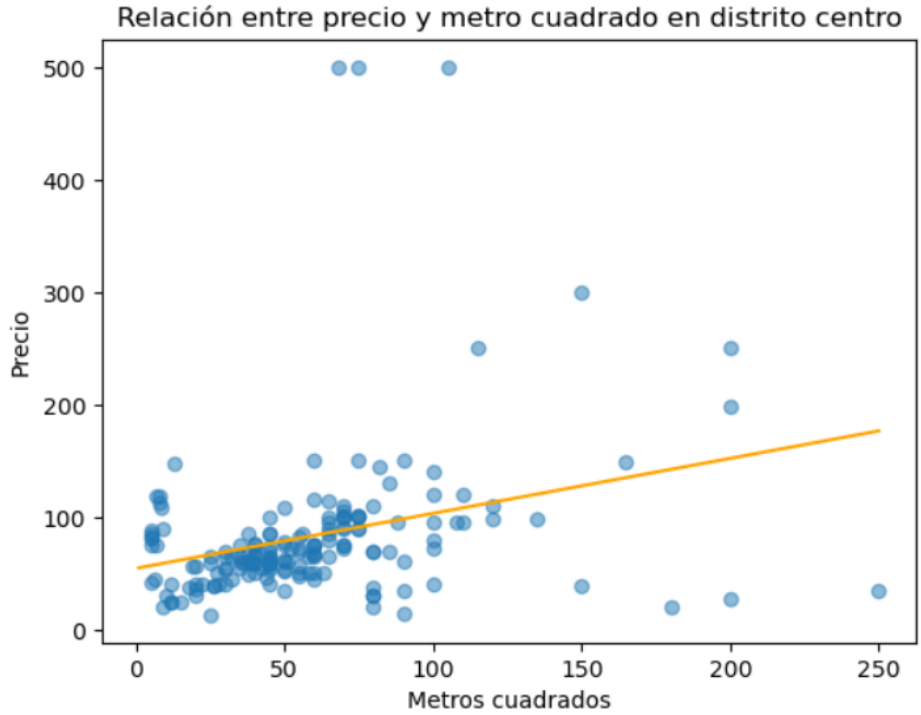
El resultado de la distribución antes y después puede verse en la gráfica inferior en color azul, una concentración importante en valores muy próximos al cero y después del filtrado una concentración de valores muy pequeños bastante menor.



Por ello, después del filtrado de datos obtenemos unas medias y valores mínimos más realistas. Vamos a suponer que todas las medidas que contienen menos de 5 metros cuadrados no son realistas. Por ello, nuestros valores estadísticos han cambiado.

	Price	Square meter
count	166.000000	166.000000
mean	83.240964	58.752849
std	69.748482	40.830819
min	13.000000	5.016722
25%	50.500000	35.024155
50%	70.000000	51.514307
75%	95.000000	74.972129
max	500.000000	250.000000

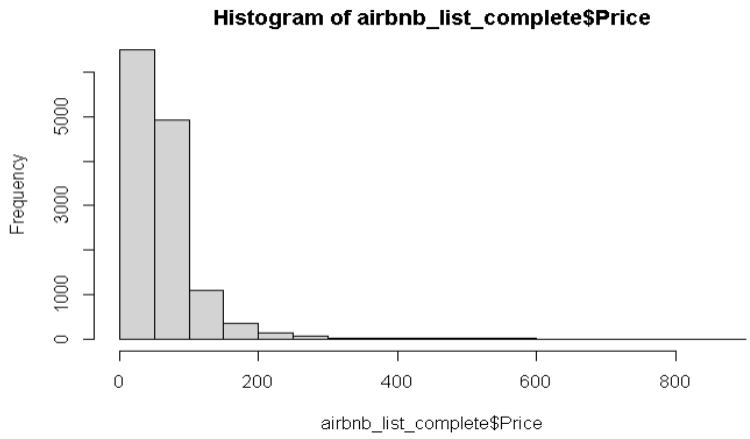
Para representar la recta de estimación se utiliza la regresión lineal por mínimos cuadrados. Todos los datos filtrados anteriormente crean dos *arrays* que configuran las muestras en las gráficas y posteriormente se intenta minimizar el error mediante la fórmula del error cuadrático medio (MCO) resultado la predicción que se muestra a continuación.


Relación entre precio y metro cuadrado en distrito centro
Precio
Metros cuadrados
0
100
200
300
400
500
0
50
100
150
200
250

2. Modelado: algoritmo de regresión lineal múltiple.

En lo referente a desarrollar un algoritmo de regresión lineal, Airbnb es una es una plataforma dedicada a la oferta de alojamiento a particulares, por ello nuestro objetivo es desarrollar un algoritmo para orientar al propietario de un precio razonable de su inmueble basado en las características de este.

Realizamos un rápido estudio de la normalidad de la variable que deseamos predecir, Price. Observamos que los precios se concentran entre los 9 y 100€.



Primero observamos que el precio de un alojamiento no va influenciado por una única característica, por ello optamos por un modelo de regresión lineal múltiple. Además, para desarrollar correctamente el algoritmo hemos dividido el dataset en dos partes, uno para train y otro para test, la decisión de que observaciones conforman el dataset de train o test es aleatoria y de esta forma reducir cualquier sesgo por nuestra parte.

Para determinar qué variables son determinantes en este análisis hemos utilizado una matriz de correlación, mediante este método descartamos *Review.Scores.Rating* por el bajo nivel de correlación.

	Price	Accommodates	Bathrooms	Bedrooms	Beds	Square.Feet	Review.Scores.Rating
Price	1.0000	0.5832	0.3412	0.5209	0.4847	0.3762	0.0734
Accommodates	0.5832	1.0000	0.3279	0.6748	0.8219	0.4390	-0.0432
Bathrooms	0.3412	0.3279	1.0000	0.4258	0.3912	0.4768	0.0148
Bedrooms	0.5209	0.6748	0.4258	1.0000	0.6788	0.4641	0.0259
Beds	0.4847	0.8219	0.3912	0.6788	1.0000	0.4099	-0.0418
Square.Feet	0.3762	0.4390	0.4768	0.4641	0.4099	1.0000	0.0118
Review.Scores.Rating	0.0734	-0.0432	0.0148	0.0259	-0.0418	0.0118	1.0000

Con las variables iniciales de tipo numérico realizamos un primer modelo (Modelo_1), el resultado fue un R^2 bajo (0,3889), pero además *Bathrooms, Beds* y *Square.Feet* no resultaban significativas en el análisis.

```
Call:
lm(formula = Price ~ Accommodates + Bathrooms + Bedrooms + Beds +
    Square.Feet, data = datos_train_1)

Residuals:
    Min      1Q  Median      3Q     Max
-155.00  -19.41   -3.19   13.60  410.42

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.762677   5.397289   3.662 0.000289 ***
Accommodates  8.030483   1.948357   4.122 4.69e-05 ***
Bathrooms    -3.463973   4.390115  -0.789 0.430620
Bedrooms     24.290750   3.702718   6.560 1.92e-10 ***
Beds         -5.317965   2.088509  -2.546 0.011313 *
Square.Feet   0.011516   0.004544   2.534 0.011695 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.21 on 352 degrees of freedom
  (8886 observations deleted due to missingness)
Multiple R-squared:  0.3889,    Adjusted R-squared:  0.3802
F-statistic: 44.81 on 5 and 352 DF,  p-value: < 2.2e-16
```

Realizamos un segundo modelo (Modelo_2) sin las variables anteriores, el resultado fue un R^2 ligeramente menor (0.3699), lo cual era esperable ya que estamos reduciendo el número de variables explicativas, pero con todas las variables significativas. Debido a esta situación nos planteamos la necesidad de incluir una variable relacionada al barrio.

```
Call:
lm(formula = Price ~ Accommodates + Bedrooms, data = datos_train_1)

Residuals:
    Min      1Q  Median      3Q     Max
-205.79  -19.97   -6.97   12.95  637.44

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.4462     0.8985   9.401   <2e-16 ***
Accommodates  11.5438     0.3039  37.983   <2e-16 ***
Bedrooms      15.4407     0.7399  20.869   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.14 on 9219 degrees of freedom
  (22 observations deleted due to missingness)
Multiple R-squared:  0.3699,    Adjusted R-squared:  0.3698
F-statistic:  2706 on 2 and 9219 DF,  p-value: < 2.2e-16
```

Optamos por incluir el nivel de renta media por persona y por hogar de cada barrio, basándonos en que un barrio con un nivel de renta elevado supone inmuebles más caros y por lo tanto el alquiler de estos es mayor que frente a un barrio con niveles de renta menores.

Los datos desglosados por barrios los obtuvimos del INE, debido a que renta media por persona y renta media por hogar tienen una alta correlación optamos por sólo utilizar para el análisis renta media por persona (correlación más elevada respecto al precio) y así evitar medir la misma característica dos veces.

	Price	Accommodates	Bathrooms	Bedrooms	Beds	Square.Feet	Renta_pers	Renta_hogar
Price	1.0000	0.5745	0.3352	0.5206	0.4862	0.3014	0.2313	0.1681
Accommodates	0.5745	1.0000	0.3114	0.6658	0.8178	0.4251	0.0771	0.0193
Bathrooms	0.3352	0.3114	1.0000	0.4054	0.3843	0.3812	0.1269	0.1138
Bedrooms	0.5206	0.6658	0.4054	1.0000	0.6801	0.4199	0.0681	0.0811
Beds	0.4862	0.8178	0.3843	0.6801	1.0000	0.4588	0.0791	0.0653
Square.Feet	0.3014	0.4251	0.3812	0.4199	0.4588	1.0000	0.0934	0.1246
Renta_pers	0.2313	0.0771	0.1269	0.0681	0.0791	0.0934	1.0000	0.9183
Renta_hogar	0.1681	0.0193	0.1138	0.0811	0.0653	0.1246	0.9183	1.0000

Incluimos la variable renta en el modelo (Modelo_3), nuestras variables explicativas son: *Accommodates, Bedrooms y Renta_pers*, el resultado fue un R2 cercano a 0.4 (0,398), lo que mostraba mejora respecto a anteriores modelos.

```
Call:
lm(formula = Price ~ Accommodates + Bedrooms + Renta_pers, data = datos_train_2)

Residuals:
    Min      1Q  Median      3Q     Max
-220.56  -18.93   -5.50   12.18  648.61

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.301e+01  2.024e+00  -16.31   <2e-16 ***
Accommodates  1.120e+01  3.071e-01   36.47   <2e-16 ***
Bedrooms      1.630e+01  7.331e-01   22.24   <2e-16 ***
Renta_pers    2.398e-03  1.064e-04   22.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42.96 on 9057 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.398,     Adjusted R-squared:  0.3978
F-statistic:  1996 on 3 and 9057 DF,  p-value: < 2.2e-16
```

Por último, nos centramos en el tipo de alojamiento, las variables *Property.Type* y *Room.Type* tienen una clara repercusión sobre el precio, un apartamento completo va tener un precio mayor frente a una habitación privada o habitación compartida.

Para medir esta variable optamos por crear una variable booleana con valor 1 para aquellos alojamientos completos y 0 para alojamientos donde lo que se ofrece es sólo una habitación.

Obtenemos el Modelo_4, un modelo con un R^2 del 0.4473, con todos los estimadores significativos.

```
Call:
lm(formula = Price ~ Accommodates + Bedrooms + Renta_pers + Entire,
    data = datos_train)

Residuals:
    Min      1Q  Median      3Q     Max
-214.59  -16.90   -3.46    9.47  659.16

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.606e+01  1.942e+00  -18.57   <2e-16 ***
Accommodates  6.384e+00  3.396e-01   18.80   <2e-16 ***
Bedrooms      1.964e+01  7.122e-01   27.58   <2e-16 ***
Renta_pers    2.180e-03  1.022e-04   21.33   <2e-16 ***
Entire        2.993e+01  1.053e+00   28.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.16 on 9056 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.4473,    Adjusted R-squared:  0.447
F-statistic:  1832 on 4 and 9056 DF,  p-value: < 2.2e-16
```

Debido al comportamiento asimétrico observado al inicio con el histograma, optamos por perfeccionar este último modelo, aplicando log a precio y renta media por persona. Obtenemos que todas las variables son significativas, un R^2 del 0.6255, y un p-valor menor que el 0.05, por lo tanto, aceptamos el modelo.

```
Call:
lm(formula = log(Price) ~ Accommodates + Bedrooms + log(Renta_pers) +
    Entire, data = datos_train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.96647 -0.24948 -0.02068  0.22612  2.76102

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.863034   0.168809  -11.04   <2e-16 ***
Accommodates     0.063329   0.003361   18.84   <2e-16 ***
Bedrooms         0.149631   0.007045   21.24   <2e-16 ***
log(Renta_pers)  0.512037   0.017382   29.46   <2e-16 ***
Entire           0.723873   0.010433   69.38   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4073 on 9056 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.6255,    Adjusted R-squared:  0.6254
F-statistic:  3782 on 4 and 9056 DF,  p-value: < 2.2e-16
```

log (Price)=-1.863+0.063329*Accom+0.146931*Bdrms+0.512037*log(renta)+0.723873*Ent

En la interpretación del modelo elegido hay que tener en cuenta que la variable endógena ahora tiene aplicado logaritmos, por lo tanto, los coeficientes representan el

- Nuestro primer coeficiente B^o, el valor constante de nuestro modelo tiene un valor de -1.863034.
- B^1 = 0.063329, es decir incrementar en un huésped la capacidad del alojamiento, supone un incremento del 6.33%.
- B^2 = 0.146931, ante un incremento de una cama en el inmueble, el precio del alojamiento se incrementa en un 15%.
- B^3 = 0.512037, se da un incremento en el precio del 0,51% por cada 1% que se incrementa la renta media por persona del barrio.
- B^4 = 0.723873, es decir, la diferencia entre contratar un alojamiento completo frente a contratar únicamente una habitación supone un incremento del 72,39% en el precio.

Por ejemplo: suponemos que somos propietarias de un apartamento y queremos alquilarlo por completo para uso turístico, tiene 3 dormitorios, cada habitación es para dos personas y un sofá cama en el salón, por ello nuestra capacidad son 8 huéspedes. Este alojamiento está situado en el barrio de Goya, que tienen asignada una renta media por persona de 22743,26€. ¿Qué precio asignamos a este apartamento?

-1.863034 + 0.063329*8 + 146819,7 *3 + 0.146931*log (22743,26) + 0.723873 = 140,48€/noche

Finalizamos evaluando la capacidad predictiva de nuestro modelo, para ello usamos la librería *Caret*, realizamos una comparación del modelo con los datos de train y con los datos test.

Los resultados son dos R^2 similares: 0.6239206 y 0.6424431 lo que nos indica que no se da un sobre ajuste, no obstante añadir que los R^2 deberían ser ligeramente más elevados para proporcionar mayor explicación sobre la variable precio.

Resultados finales

En cuanto a la obtención de parámetros para en análisis lineal simple, este no resulta un buen modelo de predicción, ya que hubo que eliminar gran cantidad de muestras porque carecían de datos relevantes quedando la muestra muy pequeña respecto al original.

Obtenemos unos predictores poco constantes, por ejemplo, la varianza alta determina que tenemos muchos cambios en nuestra muestra, y una pequeña que estos datos son muy similares.

Un valor importante a tener en cuenta en el coeficiente de correlación de Pearson en el que tenemos una correlación positiva, pero bastante baja. Por ello, el coeficiente de determinación también es muy bajo, y menor es el ajuste de nuestro modelo, o lo que es lo mismo, menos fiable.

Referente al proceso de modelado y predicción para la recta regresión múltiple llegamos a la conclusión de que el mercado de apartamentos turísticos en Madrid se encuentra en alza, sería considerable ampliar los datos relacionados con el tamaño del alojamiento, ya que el tamaño de un inmueble tiene un claro impacto sobre su precio.

Por último, destacar que nos ha sorprendido que el nivel de renta de un barrio no sea determinante en el precio de un alojamiento, esto nos lleva a considerar que la ciudad de Madrid se encuentra ante una subida general de precios, en todos sus barrios. Para un futuro análisis sería muy interesante considerar otros municipios y ciudades dentro de la Comunidad de Madrid, como por ejemplo Getafe y realizar comparaciones respecto a Madrid ciudad.