Trabajo práctico N° 2

Informe escrito

Autoras

Clara Bernhardt

María Teresa Laszeski

Rosario Luque

Profesores

María Noelia Castillo

Ignacio Spiousas

Asignatura

Ciencia de Datos

Semestre y año de presentación

2° semestre 2024

I. Pre procesamiento de datos

La Base Airbnb NY es una base de datos que contiene información de valor sobre 48905 departamentos en la ciudad de Nueva York disponibles para ser alquilados por la plataforma de Airbnb.

El primer paso en el limpiado de la base de datos es identificar las filas duplicadas y eliminarlas. Al hacerlo, vemos que hay 10 entradas repetidas, por lo que terminamos quedándonos con 48895 filas. En segundo lugar, procedemos a eliminar las columnas que no tienen información de interés. Originalmente la base de datos tenía 16 columnas, de las cuales descartamos "id", "name", "host id", "host name", "neighbourhood" y "last review" por no ser de interés para el análisis posterior.

De esta forma, las 10 columnas que se conservan son: (1) "neighbourhood group" (Manhattan, Brooklyn, Queens, Bronx y Staten Island, class: object); (2) "latitude" y (3) "longitude" (coordenadas geográficas del Airbnb, class: float); (4) "room type (Private room, Entire home/apt y Shared room, class: object); (5) "price" (en dólares, class: float); (6) "minimum nights" (class: integer); (7) "number of reviews" (class: integer); (8) "reviews per month" (class: float): (9) "calculated host listing count (class: integer) y (10) "availability 365" (cantidad de días del año que está disponible, class: integer).

En tercer lugar, evaluamos la cantidad de *missing values* (celdas vacías) por columna. Al hacerlo, observamos que las únicas dos columnas son "price" con 15 y "reviews per month" con 10052 valores faltantes. Si bien es cierto que la imputación de datos es una herramienta clave cuando el objetivo del análisis es la predicción, dado que permite reducir el error estándar (mikeenguyen13, s.f.), al hacerlo en un primer momento, con la técnica de imputación por la mediana, observamos que esto de hecho incrementaba la variabilidad de los datos. Luego de analizar las variables en cuestión, llegamos a la conclusión de que podríamos suponer que los missing values en este caso están dados por el azar. En este sentido, suponemos que la falta de precio es un error del host en el cargado del perfil y reviews per month puede estar vacío si en un mes en particular nadie dejó una reseña, pero que la falta de datos no está relacionada con ninguna otra variable. Por lo tanto, procedimos a eliminar 10064 filas que tenían al menos un *missing value*, quedándonos con un total de 38831 datos.

En cuarto lugar, realizamos el análisis de los outliers mediante un proceso sistemático. Inicialmente, generamos un boxplot de los datos, lo cual nos llevó a eliminar todos los valores de "availability 365" menores o iguales a 0, ya que no tiene sentido publicar un departamento sin disponibilidad o con disponibilidad negativa. Posteriormente, graficamos histogramas por variable para conocer la distribución de los datos, lo que nos indicó la necesidad de aplicar una transformación logarítmica (Log+1) a ciertas variables para evitar los valores cero. Utilizamos el método de la Desviación Absoluta de la Mediana (MAD) para identificar outliers dado que es un método robusto (Rousseeuw, 1990), fijando un umbral de 3.5 y etiquetando las filas sin outliers con un 0 y las que tenían outliers con un 1. Decidimos no realizar este análisis sobre la variable "calculated host listings count" para evitar una pérdida significativa de información, dado que eliminar outliers hacía que todos los datos de la variable fueran 1. Como resultado, identificamos 2741 outliers para "minimum nights", 2 para "reviews per month", 1349 para "price" y 885 para "availability 365". Finalmente, creamos una nueva base de datos con todas las filas etiquetadas con 0, resultando en un conjunto de datos depurado con 25,249 departamentos.

En quinto lugar, procedimos a la transformación de variables categóricas a numéricas. Convertimos “neighbourhood group” asignando valores numéricos a cada distrito: Manhattan (2), Brooklyn (1), Queens (3), Bronx (0) y Staten Island (4). De manera similar, transformamos “room_type” en valores numéricos: Entire home/apt (0), Private room (1) y Shared room (2). Finalmente, como sexto y último paso, creamos una nueva variable denominada “offer group”, la cual cuantifica la cantidad de departamentos ofrecidos por zona.

II. Gráficos y visualizaciones

Una vez hecha la limpieza de datos, se generó una matriz de correlación (**Fig 1**) con las siguientes variables: ‘neighbourhood group’, ‘latitude’, ‘longitude’, ‘room type’,‘price’, ‘minimum nights’, ‘number of reviews’, ‘reviews per month’,‘calculated host listings count’, ‘availability 365’. Se incluyeron las variables que originalmente eran categóricas de forma numérica para poder interpretar mejor la matriz de correlación. A continuación, se detallan algunas observaciones claves.

En primer lugar, existe una correlación moderada entre ‘neighbourhood group’ y ‘latitud’ (0.20), y entre ‘longitud’ y ‘neighbourhood group’ (0.13). Esto indica que la ubicación geográfica está algo relacionada con el área en la que se encuentran las propiedades, aunque no es una relación fuerte.

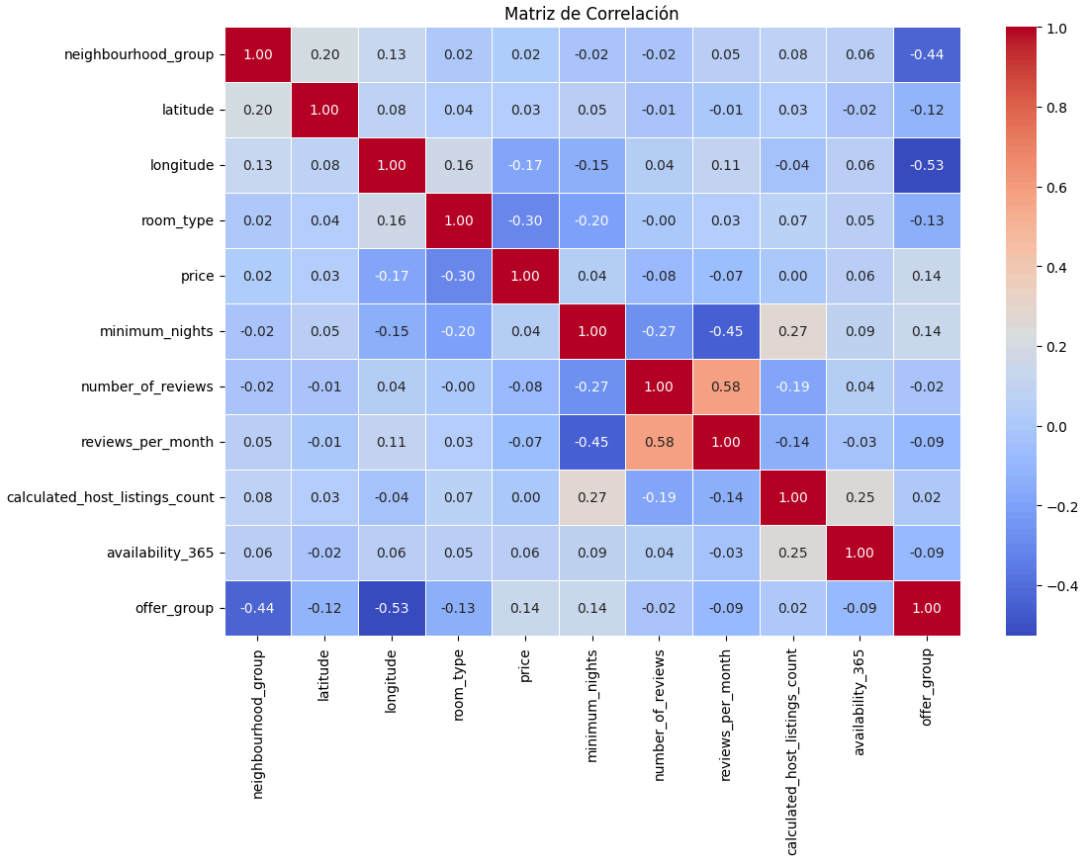

Fig 1. Matriz de correlaciones de la Base Airbnb NY.

En segundo lugar, se puede notar que hay una correlación negativa entre ‘price’ y ‘room type’ (-0.30), lo que sugiere que los tipos de habitación (Entire home/apt, Private room, Shared room) están inversamente relacionados con el precio. También hay una

correlación negativa entre ‘price’ y ‘longitude’ (ubicación longitudinal de las propiedades) (-0.17). Finalmente, no hay una correlación significativa entre ‘price’ y las demás variables, lo que indica que las demás variables no influyen directamente en el precio.

En tercer lugar, se observa una correlación positiva entre ‘number of reviews’ y ‘reviews per month’ (0.58). Tiene lógica, ya que un mayor número de reseñas mensuales tiende a estar asociado con un mayor número total de reseñas. En cuarto lugar, existe una correlación negativa entre ‘minimum nights’ y ‘number of reviews’ (-0.27) y ‘reviews per month’ (-0.45). Esto puede deberse a que los alquileres con más restricciones de noches mínimas suelen recibir menos reseñas, lo que podría indicar menor rotación de inquilinos.

Por último, la variable ‘availability 365’ tiene una correlación baja con la mayoría de las variables. Esto sugiere que la cantidad de días que los alojamientos están disponibles al año no está fuertemente relacionada con características como el precio, el número de reseñas o las noches mínimas.

A continuación se analizaron las proporciones de oferentes en los diferentes barrios (‘Neighbourhood group’) de la ciudad de Nueva York (**Fig 2)**.

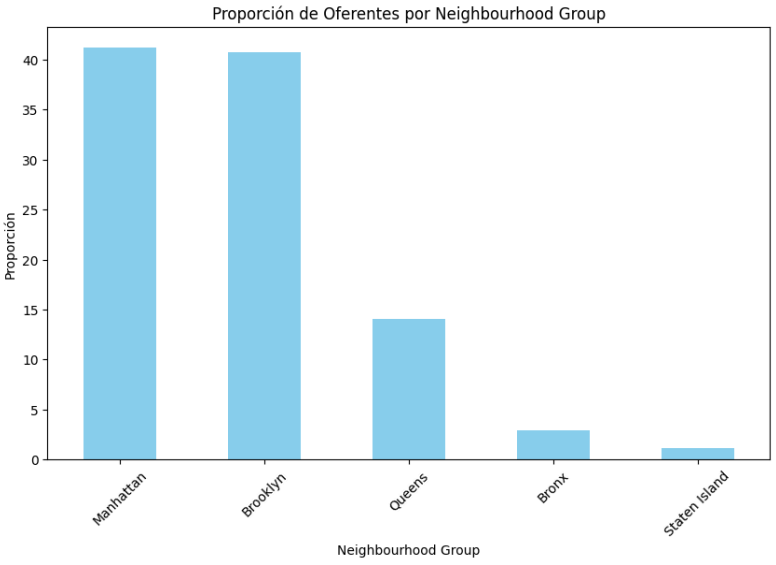

Fig 2. Histograma de la proporción de oferentes por zona.

Este análisis evidencia que los barrios de Manhattan (41.18%) y Brooklyn (40.71%) concentran la mayoría de los oferentes, acumulando entre ambos más del 80% del total. Esto se podría deber a su ubicación central y atractivos turísticos. Queens, con un 14.04%, ocupa el tercer lugar, mientras que Bronx (2.93%) y Staten Island (1.14%) cuentan con una baja participación, con menos del 3% cada uno, posiblemente debido a menor demanda o menor desarrollo turístico.

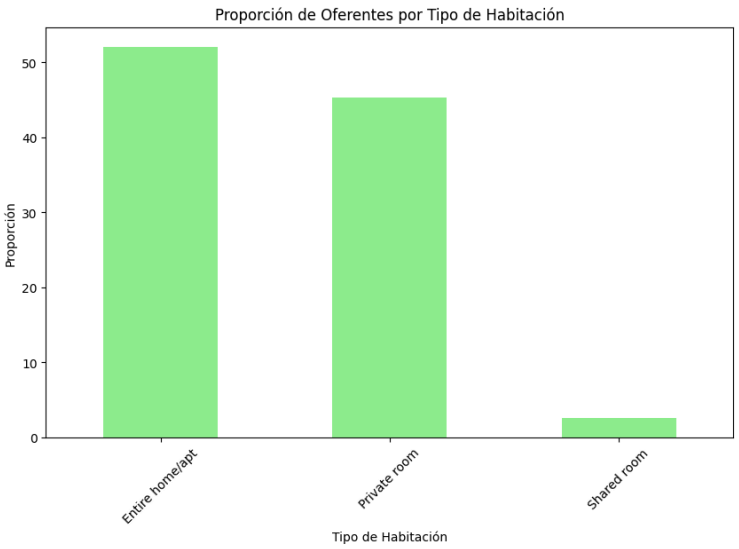

Fig 3. Histograma de la proporción de oferentes por tipo de habitación.

Las proporciones según el tipo de habitación ofrecido son las siguientes: ‘Entire home/apt’: 52.05%, ‘Private room’: 45.36% y ‘Shared room’: 2.59% (**Fig 3**). La preferencia por el alquiler de viviendas completas o departamentos puede estar relacionada

con el mayor grado de privacidad que ofrecen, comparado con las habitaciones privadas o compartidas. Las habitaciones privadas mantienen una oferta competitiva, lo que podría estar relacionado con precios más accesibles para los viajeros que buscan ahorrar. Por último, las habitaciones compartidas no son una opción popular, seguramente por su poca privacidad.

El histograma de los precios de alojamiento (**Fig 4**) muestra una clara concentración de precios hacia valores bajos. La mayoría de los alojamientos tiene un precio inferior a los $500, pero hay precios más altos (más de $4000) que llaman la atención. El gráfico muestra una asimetría positiva, lo que significa que hay pocos valores extremadamente altos que arrastran la media hacia arriba. La línea de densidad refuerza esta observación, mostrando un pico fuerte alrededor de los precios bajos y una larga cola hacia los precios más altos.

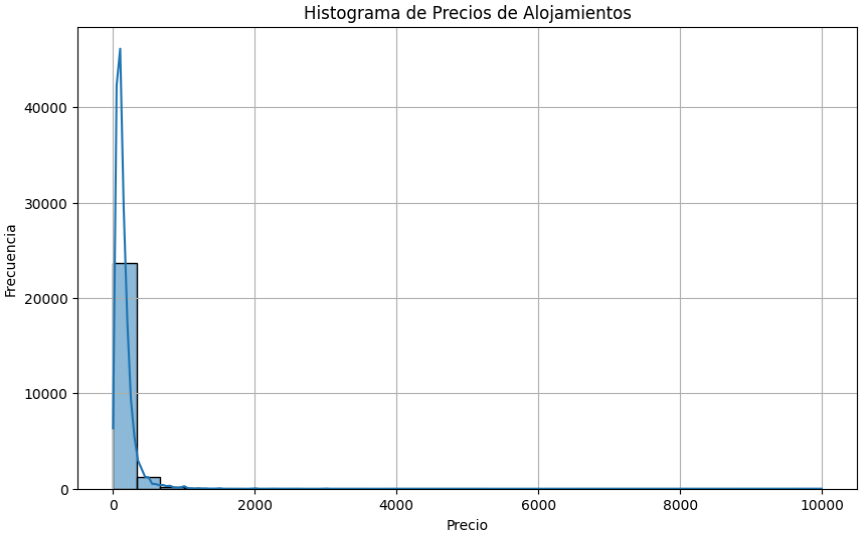

Fig 4. Histograma de los precios de los alojamientos.

El precio mínimo es de $0 (puede haber alojamientos gratuitos o errores en los datos), luego el precio máximo es de $9999 (este valor podría ser un outlier extremo) y el precio promedio es de $150.57.

La media de precios por barrio revela que Manhattan ($197.59) tiene los precios más altos, lo cual es esperable dado que es la zona más exclusiva y turística de la ciudad. Le sigue Brooklyn ($128.73), después Queens ($95.68), mientras que Bronx ($80.25) y Staten Island ($88.98) tienen los precios más bajos.

Los alojamientos de departamentos son significativamente más caros ($210.00), lo cual es razonable, ya que ofrecen más espacio y privacidad. Las habitaciones privadas tienen precios intermedios ($87.42), mientras que las habitaciones compartidas son las más asequibles ($62.48), probablemente orientadas a viajeros con presupuestos más ajustados.

Este análisis muestra que la mayoría de los viajeros optan por opciones más asequibles, pero sigue existiendo un segmento dispuesto a pagar precios mucho más altos, particularmente en áreas como Manhattan y para los departamentos.

A continuación, se realizaron dos scatter plots con dos variables de interés en cada uno. En el primer scatter plot (**Fig 5**), donde se ve una relación entre precio de los alojamientos y número de reseñas por mes, cuantas más reseñas tiene un lugar de alojamiento el precio menor es el precio. Si el lugar tiene menos reseñas, los precios son más variables, y casi siempre más altos.

En el siguiente scatter plot se ve una relación entre el precio y la disponibilidad (días al año, **Fig 6**). En la parte inferior del gráfico, hay una concentración de precios bajos con distintos niveles de disponibilidad que varían entre los 0 a 365 días del año, lo que podría

significar que no hay una relación directa entre precio bajo y disponibilidad. No se puede ver una relación clara y fuerte entre precio y disponibilidad. En cambio, se podría decir que los alojamientos con precio alto, tienden a tener más disponibilidad durante el año. Pero, cabe aclarar que son pocos los alojamientos que tienen un precio elevado.

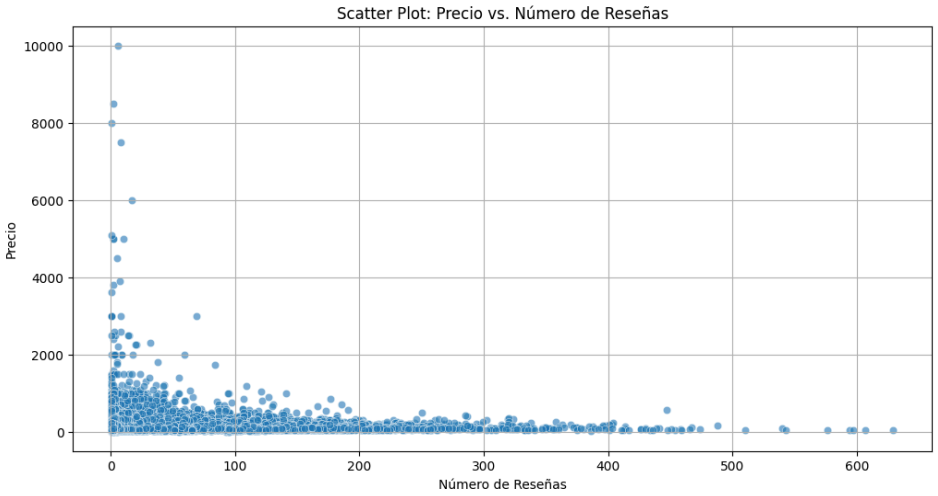

Fig 5. Scatter plot del precio y el número de reseñas.

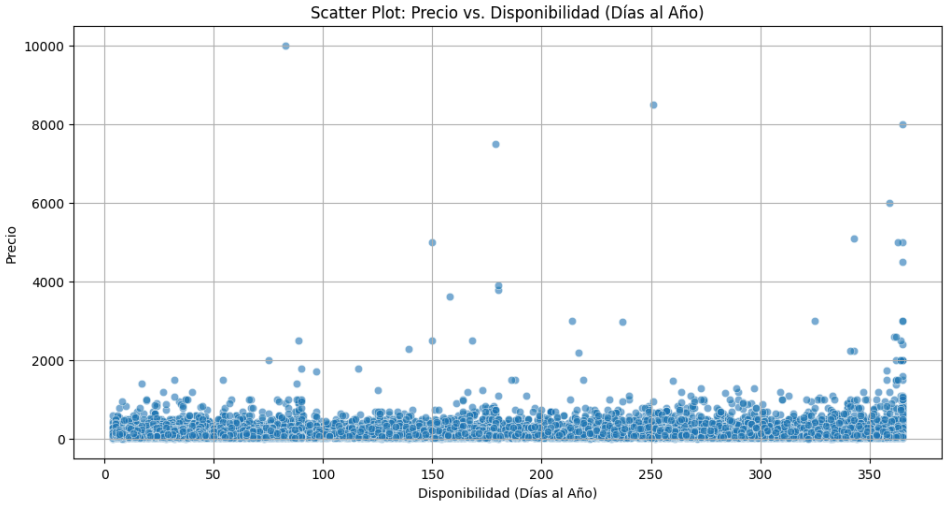

Fig 6. Scatter plot del precio y la disponibilidad.

Para realizar el análisis de componentes principales (PCA), escalamos las variables estandarizándolas para que los datos contribuyan equitativamente al análisis, independientemente de sus unidades o magnitudes originales. Después de estandarizar las variables, ejecutamos el PCA y observamos las varianzas explicadas por los componentes (**Fig 7**): El primer componente principal explica el 20.06% de la varianza total, el segundo componente principal explica el 16.38% de la varianza total y el tercer componente principal explica el 11.62% de la varianza total.

Los loadings representan el aporte de cada variable a los componentes principales. En este caso, los loadings del componente 1 son: [-0.1865939, -0.06070954, -0.36362495, -0.21796445, 0.21299138, 0.43692055, -0.38454653, -0.46067971, 0.21217665, 0.0237288, 0.37392077]. De estos valores se puede interpretar que: la variable 8 ("reviews per month") tiene una relación fuerte e inversa con el componente 1. Otros loadings significativos son los de las variables 6 ("minimum nights"), 7 ("number of reviews") y 11 ("offer group"). Las variables 6 y 11 tienen una relación fuerte y directa con el componente.

Los loadings para el componente 2 son: [0.39427018, 0.23898463, 0.34501326, 0.17936625, -0.10641674, 0.23403178, -0.3419576, -0.30599721, 0.31605774, 0.22081585,

-0.4592539]. En este caso, las variables que tienen mayor peso y relación directa en el componente 2 son: variable 1 ("neighbourhood") y variable 3 ("latitud"). Por otro lado, las

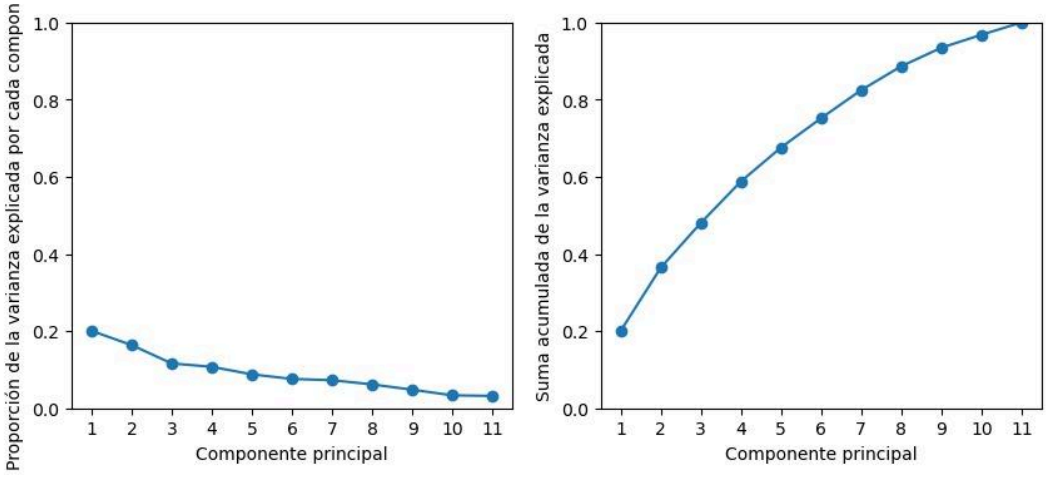

Fig 7. Gráfico de varianza explicada y varianza acumulada CP

variables con relación significativa e inversa son las variables 7 ("number of reviews") y 11 ("offer group"). Posteriormente, calculamos la norma euclídea de los loadings de cada componente y verificamos que la suma de los cuadrados de los loadings de cada componente es igual a 1.

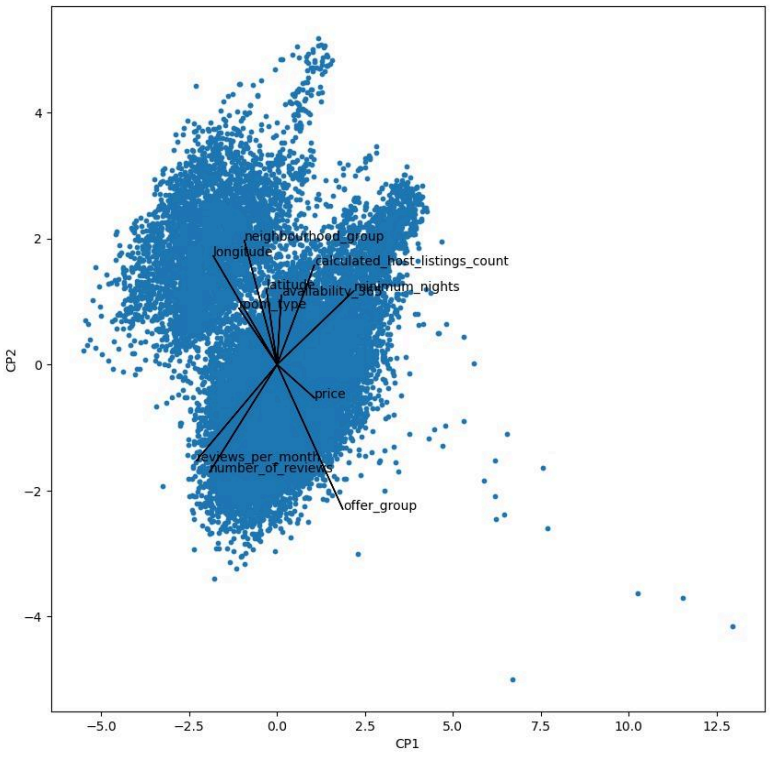

Fig 8. Bipilot de CP 1 y CP 2

En el gráfico (**Fig 8**) presentado a continuación, se puede observar que aquellos loadings mencionados anteriormente con relación inversa se encuentran en los valores negativos de sus respectivos ejes. De manera similar, los loadings con relación directa se ubican en los valores positivos de los ejes del componente 1 y 2, según corresponda.

III. Predicción

Para realizar el modelo lineal, iniciamos dividiendo la base de datos en "base de prueba (test)" y "base de entrenamiento (train)" utilizando la función “train_test_split” de la librería” sklearn.model_selection”, e implementando una semilla (201) con el parámetro

random_state. El 30% de los datos se asignaron a la base de prueba y el 70% restante a la base de entrenamiento.

Después de dividir la base, se definió la variable dependiente ("Price") en la base de entrenamiento, mientras que las variables restantes conformaron la variable independiente en formato de matriz. Además, se agregó una columna de unos (1) a la matriz de variables independientes.

Para crear el modelo de regresión lineal, utilizamos la clase LinearRegression de la librería "sklearn.linear_model". El modelo se ajustó utilizando las variables dependiente e independiente de la base de entrenamiento. A continuación, creamos la variable "y_pred" (variable predictora) utilizando la variable independiente de entrenamiento.

Se realizó el cálculo del MSE y R^2 para evaluar el ajuste del modelo para predecir el precio. El MSE obtenido es de 0.146 y el R^2 es de 0.70.

Para evaluar visualmente el rendimiento del modelo, se creó un gráfico de dispersión (**Fig 9**) comparando los precios predichos (eje Y) con los precios reales (eje X). La distribución de los puntos muestra un buen ajuste al modelo de regresión lineal, lo que confirma su capacidad para predecir precios de manera efectiva en este contexto.

Fig 9. Gráfico de dispersión de valores predichos y valores reales..

Referencias

mikeenguyen13 (s.f.). Chapter 11. Imputation (Missing Data) | A Guide on Data Analysis. Bookdown.org. https://bookdown.org/mike/data_analysis/imputation-missing-data.html#imputation-missing-data

Rousseeuw, P. J. (1990). Robust estimation and identifying outliers. Handbook of statistical methods for engineers and scientists, 16, 16-11.