AIRBNB MADRID 2024

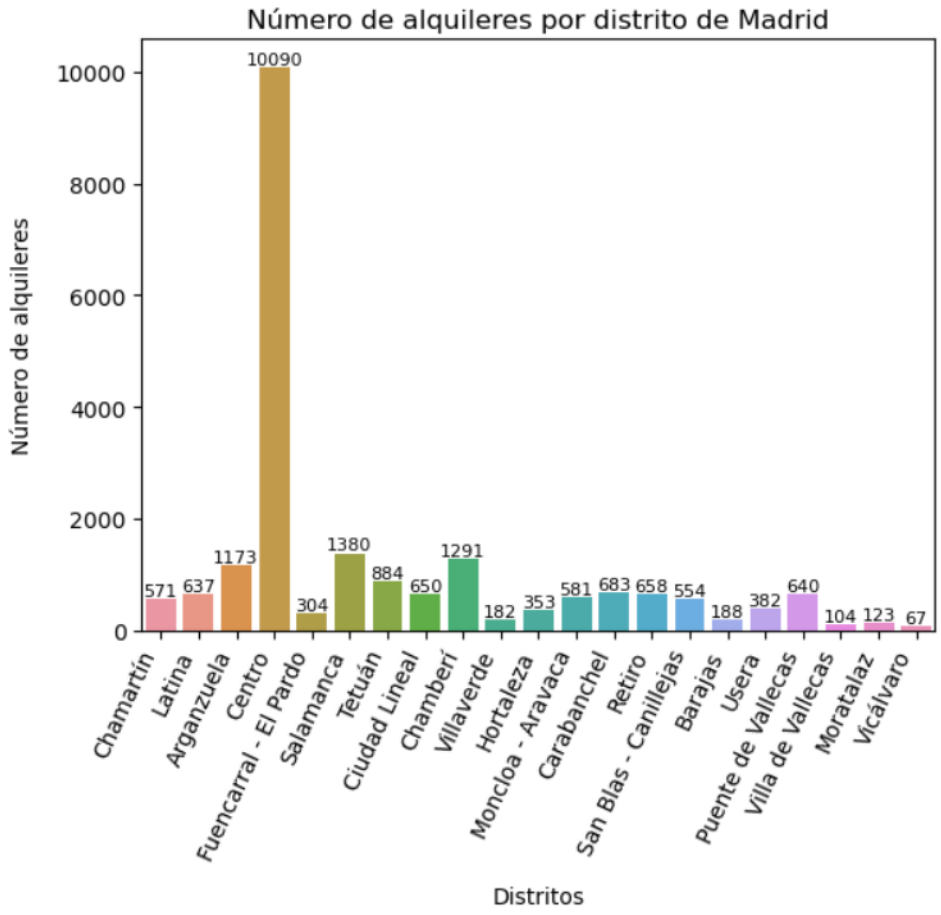
22 FEBRERO

Prueba Data Scientist
Creado por: Antonio José Toro Valderas

Análisis exploratorio

A continuación se muestran los resultados del análisis exploratorio de datos de los alquileres en Airbnb de la ciudad de Madrid.

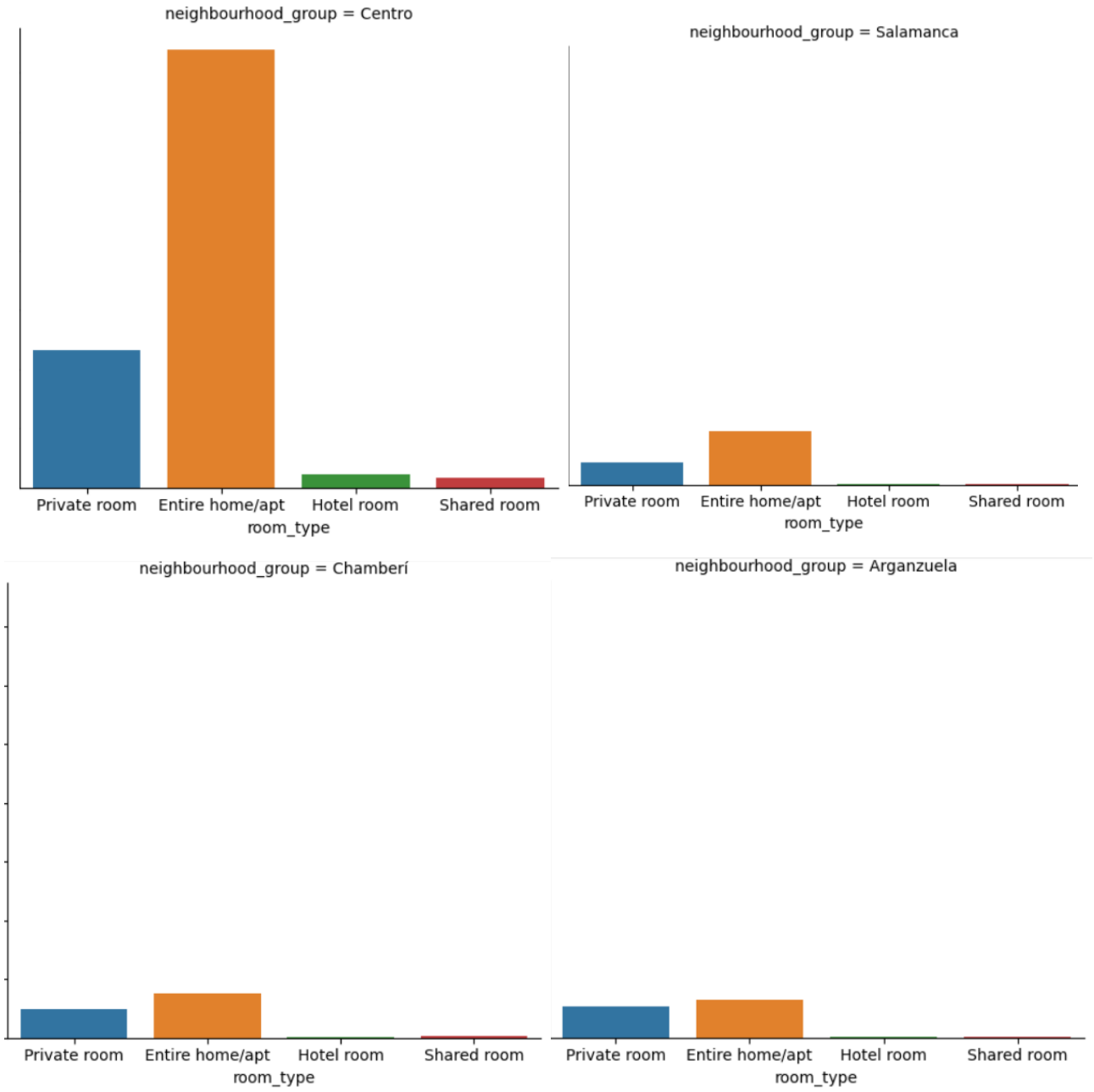
Para empezar, se muestra el número de alquileres por distritos en la ciudad de Madrid:



Sin duda alguna, el distrito del Centro es el rey de los alquileres, ya que existen más de 10000 alquileres registrados en esta zona. Después, la distancia entre el número de alquileres entre distritos se reduce. Salamanca, Chamberí y Arganzuela son los siguientes distritos donde más alquileres ofertados hay (con más de 1000 en cada uno de ellos).

Por el contrario, en Vicálvaro, Villa de Vallecas o Moratalaz son los distritos donde menos alquileres hay (menos de 150 alquileres en todos estos distritos).

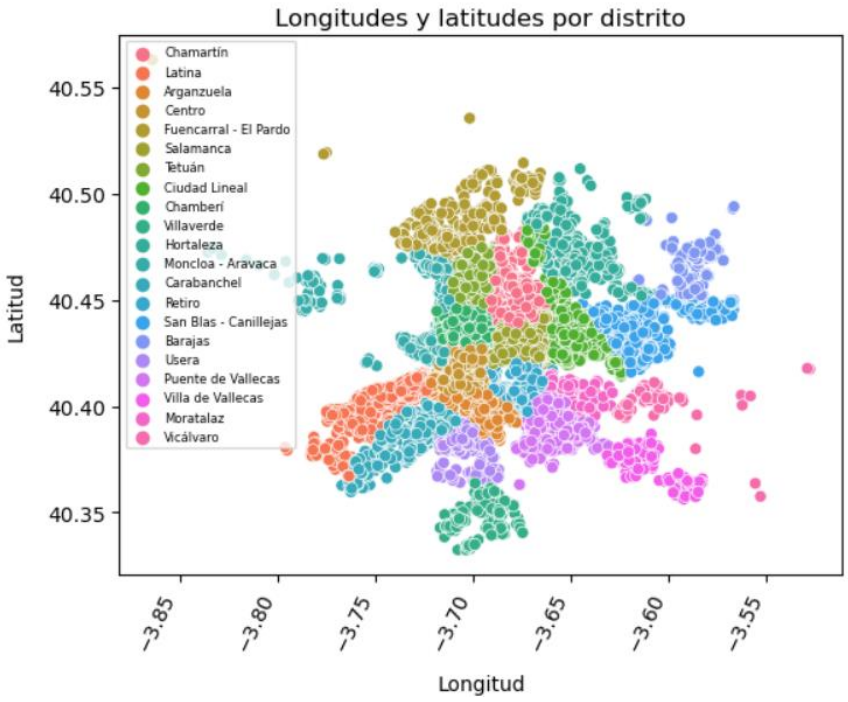
Se examina en varios de estos distritos que tipos de habitaciones se ofertan más en cada uno de ellos:



Como se puede apreciar, lo que más se ofrece son casas/apartamentos completos y habitaciones privadas. Y dentro de estas dos categorías, la mayor oferta es de casas/apartamentos completos.

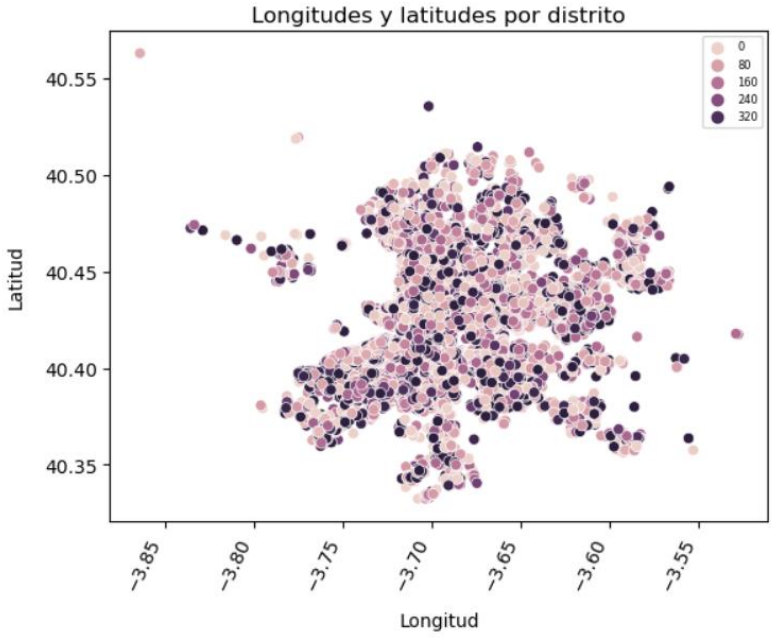
Por otra parte, las habitaciones de hotel y las habitaciones compartidas tienen pocas ofertas en comparación con las dos categorías anteriores. Esto tiene sentido, ya que la gente quiere intimidad y espacio para disfrutar de sus estancias y no compartirlas con otras personas desconocidas.

Una vez analizado el tipo de habitación por distritos, se realiza un gráfico de dispersión de la longitud y la latitud de los alquileres agrupándolos nuevamente por los distritos:



Efectivamente, longitudes y latitudes cercanas de alquileres pertenecen al mismo distrito, ya que todos los puntos de la gráfica de un mismo distrito se encuentran claramente muy cerca, agrupados, y por tanto, no están muy dispersos a lo largo del gráfico.

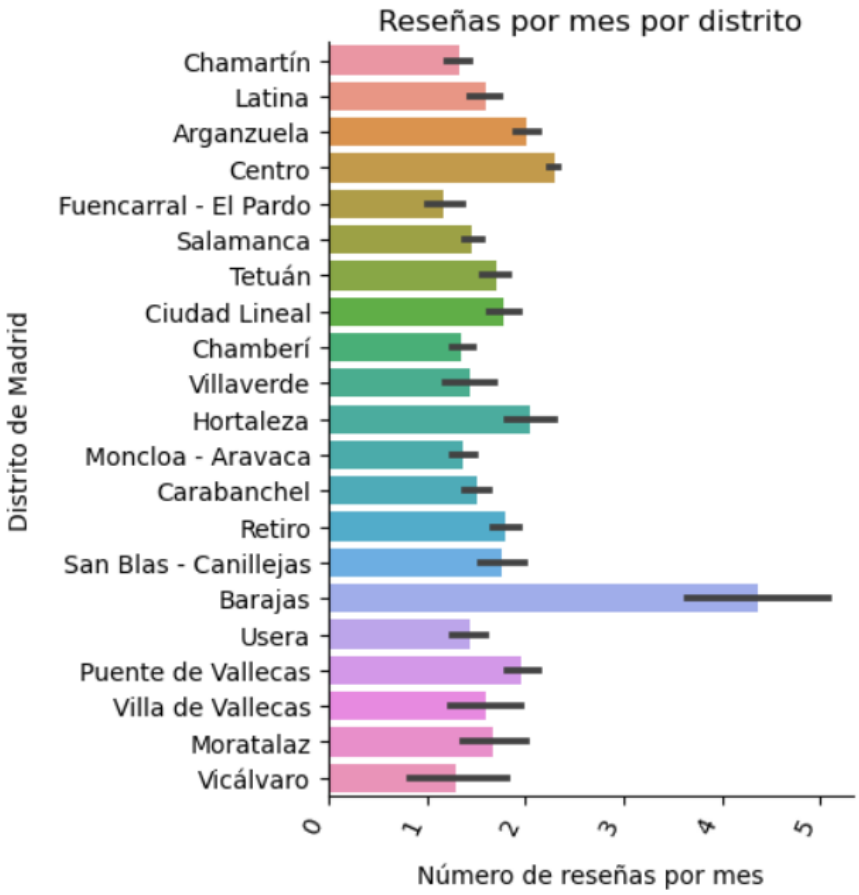
De forma análoga, se puede realizar otro gráfico de dispersión de las latitudes y longitudes de los alquileres, pero esta vez agrupándolos en función de los días de disponibilidad de alquiler al año:



Se aprecia que no hay correlación entre las variables de longitud y latitud y la disponibilidad anual. Los alquileres se reparten a lo largo de todas las longitudes y latitudes independientemente de los días de

disponibilidad de alquiler que tienen al año, por lo que hay alquileres sin y con mucha disponibilidad por todas las longitudes y latitudes de la ciudad de Madrid.

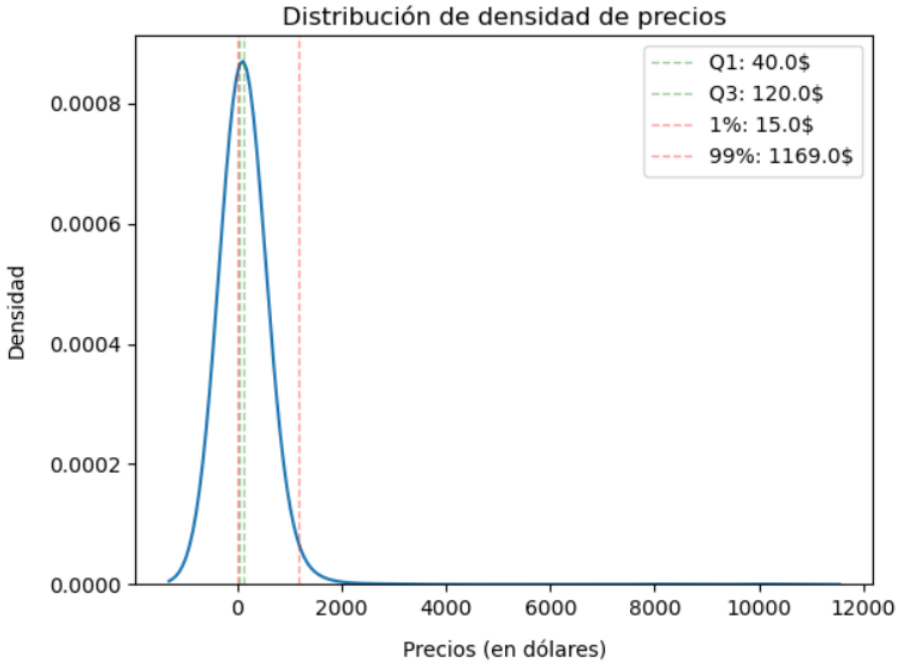
Respecto a las reseñas recibidas en la web, se muestra a continuación el número de reseñas de media por mes en cada distrito:



En el distrito de Barajas es dónde más reseñas por mes de media se realizan con bastante diferencia (más de 4 reseñas por mes) a pesar de que se ha visto anteriormente que es uno de los distritos donde menos alquileres ofertados hay. Esto indica que los huéspedes que se hospedan en este distrito es muy probable que escriban alguna reseña. Les siguen en este sentido distritos como Centro (el que más oferta de alquileres tiene) y Arganzuela, con más de 2 reseñas de media por mes.

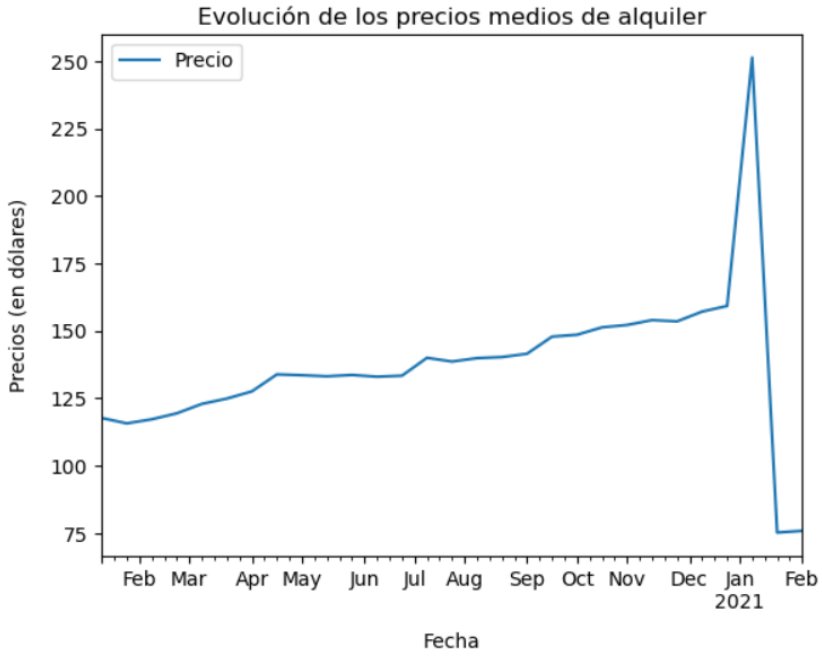
En cambio, en el lado negativo, se tienen distritos como Vicálvaro y Fuencarral - El Pardo, con menos de dos reseñas por mes, por lo que los huéspedes de estos distritos es poco probable que dejen alguna reseña en la web.

Tras esta breve introducción de la situación general del mercado de alquileres de Madrid, se analizan los precios. Para empezar, se muestra una distribución de densidad de los importes de los alquileres en Madrid:



En esta gráfica se aprecia que la mayoría de los precios está en torno al rango de los 40-120 dólares por día (primer y tercer cuartil). A precios menores o mayores a este rango, la densidad va disminuyendo, ya que la cantidad de alquileres a precios más bajos que Q1 o a precios más alto que Q3 es menor.

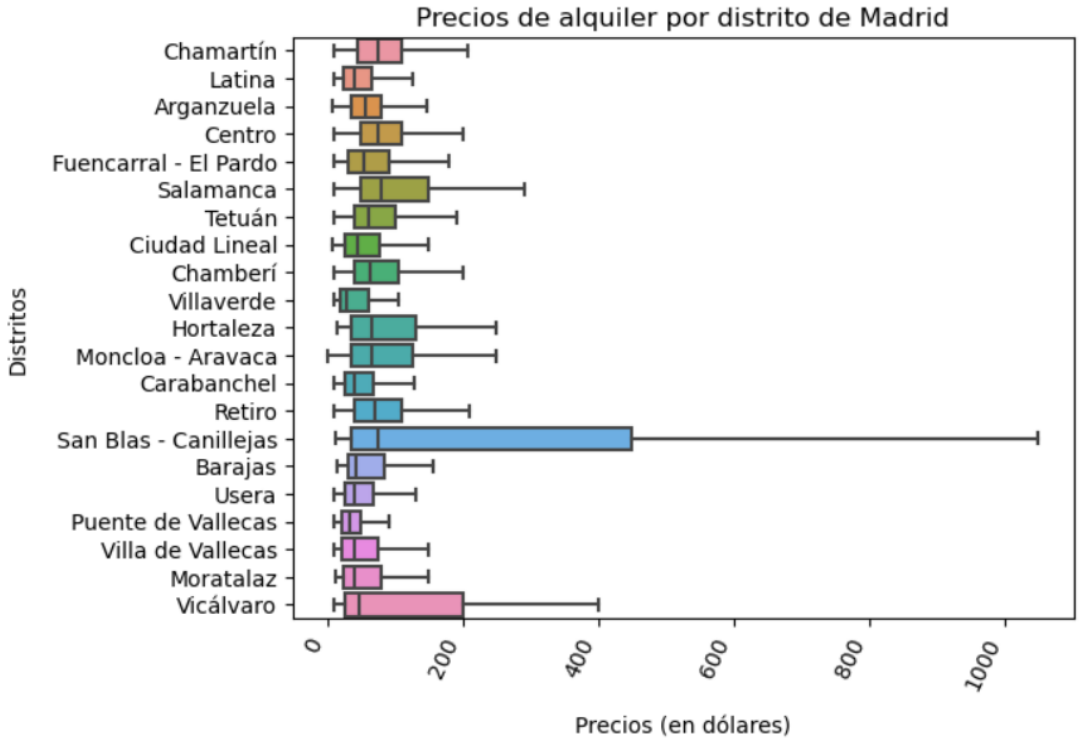
Aparte de esto, también se ha creado una gráfica que muestra la evolución temporal anual de los precios:



Como se puede apreciar, existe una tendencia general ascendente de los precios a medida que van pasando los meses del año. De forma más concreta se puede ver que de enero a febrero los precios disminuyen un poco, probablemente por el fin de las fiestas navideñas. A partir de aquí los precios van

aumentando poco a poco hasta llegar a abril, donde hay un pequeño repunte hasta mayo. Desde mayo a julio los precios medios se mantienen prácticamente y de nuevo en julio hay un pequeño repunte de los precios, quizá debido a las vacaciones de verano. Posteriormente, en septiembre los precios se mantienen constante y suben paulatinamente hasta mediados de diciembre, momento en el que se produce la mayor subida del año. Una vez pasada la Navidad y Reyes, los precios bajan de nuevo a su mínimo.

Después de ver la evolución temporal, también se puede ver los precios divididos por cada uno de los distritos de Madrid mediante un diagrama de cajas:

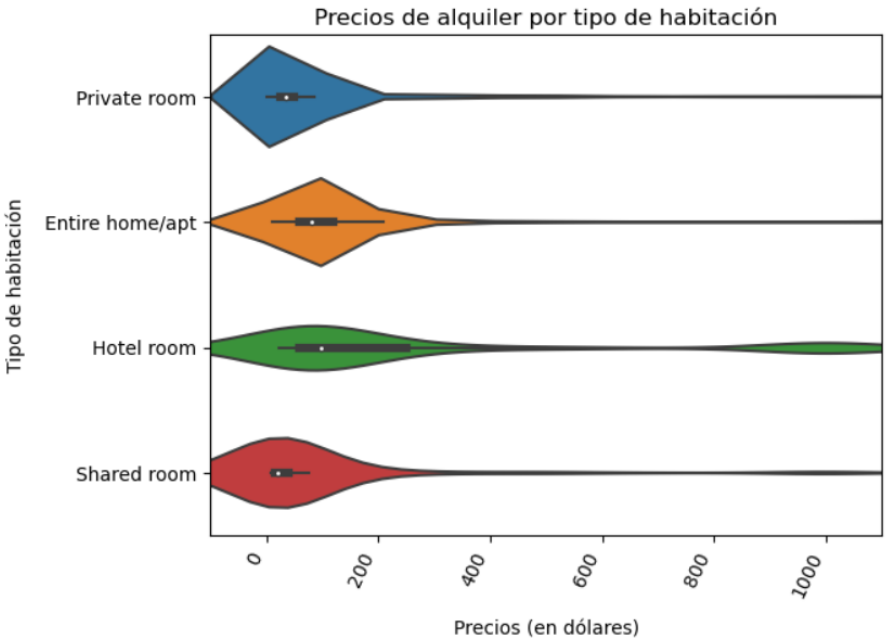


Como se puede ver en el diagrama, las medianas más altas están en distritos como los de San Blas - Canillejas, Salamanca o Chamartín, por lo que en estos lugares el precio normalmente será más alto respecto a los demás distritos. En cambio, los más bajos están en distritos como Carabanchel, Villaverde o Puente de Vallecas.

Por otro lado, hay que destacar los rangos intercuartílicos del propio distrito de San Blas – Canillejas, de Vicálvaro y de Salamanca. Son rangos muy grandes, lo que hace indicar que la variabilidad de precios en esos distritos es muy alta. Por el contrario, rangos intercuartílicos pequeños como el de Puente de Vallecas o Usera hacen indicar que la variedad de precios no es muy alta por esos emplazamientos.

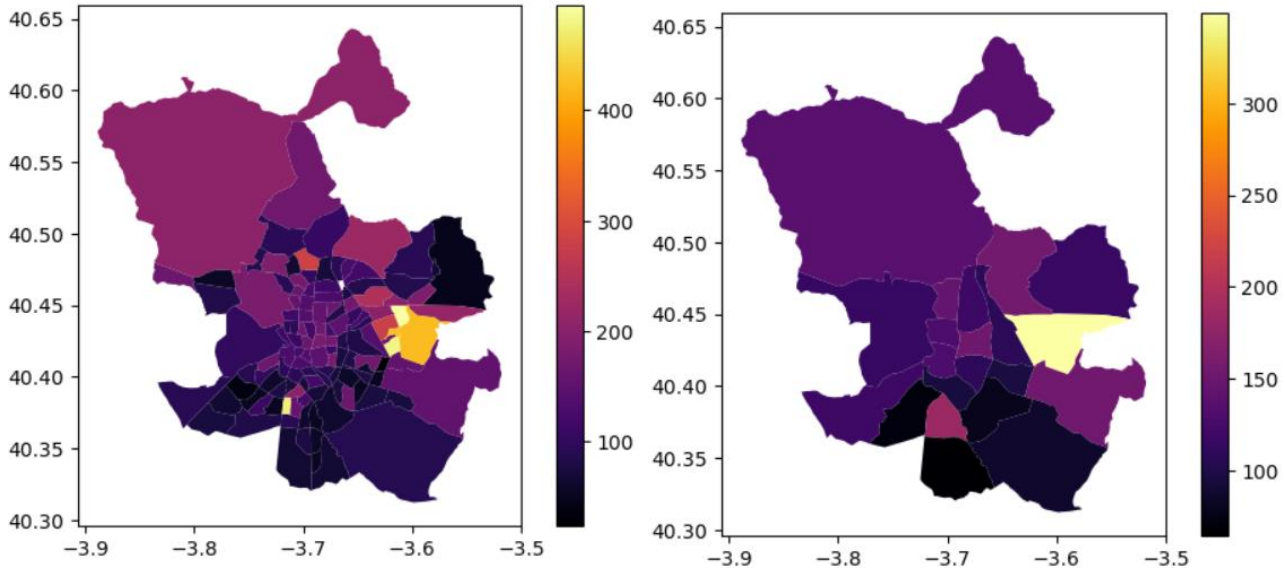
Por último, destacar que en San Blas – Canillejas y Vicálvaro los precios tienen una asimetría muy importante hacia la derecha, por lo que la mayoría de precios estarán entre la mediana y el tercer cuartil, es decir, que los precios tenderán a ser más caros que a ser baratos.

Una vez examinados los precios por distrito, se examinan por tipo de habitación con un diagrama de violines:



Este gráfico indica que las medianas más altas están en apartamentos/casas completas y en habitaciones de hotel, por lo que en estos alquileres el precio normalmente será más alto respecto a una habitación privada, y, sobre todo, respecto a una habitación compartida, que es la más baja.
Destacar el rango intercuartílico de las habitaciones de hotel. Es muy grandes, por lo que la variabilidad de precios en este tipo de alojamiento será muy elevada.
Además, tiene una asimetría contundente hacia la derecha, por lo que la mayoría de precios tenderán a ser caros.

Para terminar este análisis de los precios, se ha realizado un mapa coroplético con los precios medios del alquiler por barrios y por distritos de la ciudad de Madrid:



La información mostrada aquí informa que los precios de alquiler medios más caros están en barrios como Canillejas, Arcos o Zofío, dónde se llegan a alcanzar los 400 dólares. Por otro lado, los precios medios más baratos se encuentran en barrios como Horcajo, Aluche o Vista Alegre, lugares dónde el alquiler medio está por debajo de los 100 dólares.

En cuanto al mapa de los distritos, el precio medio más alto está en el distrito de San Blas - Canilleja, que está en torno a los 340 dólares. Por el contrario, los precios más bajos se encuentran en distritos como Villaverde o Carabanchel, dónde los alquileres están por debajo de los 100 dólares.
Como se puede comprobar, la información obtenida en este mapa se parece bastante a la obtenida en el diagrama de cajas realizado anteriormente en los distritos de Madrid.

Predicción del precio de alquileres

En esta parte del informe se va a informar del modelo de machine learning obtenido y la interpretación de las métricas conseguidas.

Tras entrenar varios modelos, el modelo elegido ha sido el del árbol de decisión de regresión:

- Este modelo ha conseguido un coeficiente de determinación ($\boldsymbol{R^2}$) de 0.73 en el conjunto de test. Este coeficiente determina la calidad del modelo para replicar resultados y la proporción de variación de los resultados que puede explicarse con él. Cuánto más cercano a la unidad sea el valor de este coeficiente, mejor será el modelo y mayor será la variabilidad explicada para la variable de respuesta, que es el precio del alquiler diario.
  Por tanto, con la métrica obtenida se ha conseguido explicar la variación para la variable de respuesta en un 73%.

- Este modelo también ha conseguido un error absoluto medio (MAE) de 10.21 dólares en el conjunto de test. Este error absoluto medio, por su parte, mide la media de las diferencias absolutas entre los valores reales y los predichos por el modelo, por lo que es muy útil para minimizar el error general del modelo.
  Por lo tanto, esto quiere decir que, de media, en todo el conjunto de test, hay un error de tan solo 10 dólares a la hora de predecir el precio diario de alquiler de un inmueble (si por ejemplo la vivienda tiene un precio de 80 dólares, el modelo, como error medio, predice que la vivienda vale 70 u 90 dólares).

Cabe destacar que se ha entrenado otro modelo (XGBoost), con mejor métrica de $\boldsymbol{R^2}$, pero al mismo tiempo con peor métrica de MAE. Como el objetivo final del cliente es predecir a que precios puede poner a alquilar sus inmuebles, se ha priorizado un modelo con el mínimo valor de MAE posible en el conjunto de test, ya que es la métrica que le va a ser más útil.

Es por eso que no se ha elegido este modelo que se está comentando, sino el árbol de decisión que se ha comentado en los párrafos anteriores.

Productivización

Para productivizar el modelo, éste se podría desplegar a un *endpoint* de AWS Sagemaker. Para empezar, primero se archiva y se comprime el modelo a formato “tar.gz”. Una vez esto es realizado, se sube el modelo comprimido a un *bucket* de S3.

Una vez el archivo esté subido, se puede inicializar dentro de AWS un SKLearnModel, que es una clase de Scikit-Learn de Sagemaker para poder desplegar un modelo a un endpoint. En esta clase se define entre otras cosas la localización del modelo (que está subido en el bucket de S3), un rol IAM para Sagemaker y la ruta del archivo de Python que se ejecutará como *entrypoint* (o punto de partida) para hospedar el modelo.

Una vez definido esto, el modelo se puede desplegar a un endpoint especificando el tipo de instancia EC2 elegida.