Universidad ORT Uruguay
Facultad de Ingeniería

Predicción de Precios en Airbnb

Entregado como requisito para la obtención del título de Master en Big Data

Ana Laura Cuitiño – 289525

Andres Perelmuter - 157068

Joaquin Rama - 173098

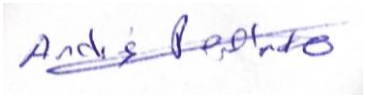
Tutor: Juan Ignacio Villalba

2024

Declaración de autoría

Nosotros, Ana Laura Cuitiño, Andres Perelmuter, y Joaquín Rama, declaramos que el trabajo que se presenta en esa obra es de nuestra propia mano. Podemos asegurar que:

- La obra fue producida en su totalidad mientras realizamos el Master en Big Data ;
- Cuando hemos consultado el trabajo publicado por otros, lo hemos atribuido con claridad;
- Cuando hemos citado obras de otros, hemos indicado las fuentes. Con excepción de estas citas, la obra es enteramente nuestra;
- En la obra, hemos acusado recibo de las ayudas recibidas;
- Cuando la obra se basa en trabajo realizado conjuntamente con otros, hemos explicado claramente qué fue contribuido por otros, y qué fue contribuido por nosotros;
- Ninguna parte de este trabajo ha sido publicada previamente a su entrega, excepto donde se han realizado las aclaraciones correspondientes.



Ana Laura Cuitiño	Andres Perelmuter	Joaquin Rama
21-03-2024	21-03-2024	21-03-2024

Agradecimientos

Al llegar a la culminación de este trabajo que ha sido nuestra tesis, quisiéramos expresar nuestro más sincero agradecimiento a todas aquellas personas que han contribuido a hacer posible el mismo.

En primer lugar, deseamos expresar nuestra profunda gratitud a nuestro tutor de tesis, Juan Ignacio Villalba, cuya guía experta, paciencia y apoyo fueron fundamentales para la realización de este proyecto. Su compromiso y pasión por la excelencia no solo nos orientaron a través de este proceso, sino que también nos inspiraron a perseguir nuestros objetivos con determinación.

Un agradecimiento especial a nuestros compañeros de Universidad ORT y amigos, por su apoyo incondicional, por las sesiones de estudio juntos y por todos los momentos de desahogo y risas que hicieron este viaje mucho más llevadero.

A nuestra familia, a la que les estamos eternamente agradecidos por su apoyo incondicional y comprensión.

Por último, pero no menos importante, queremos agradecer a todas las personas y participantes que contribuyeron a este estudio con su tiempo y conocimientos. Sin su colaboración, este trabajo no habría sido posible.

A todos , nuestro más sincero agradecimiento.

Abstract

En la economía actual, marcada por la creciente importancia de la economía compartida, la fijación y optimización de precios emergen como elementos cruciales para asegurar la rentabilidad y competitividad de las empresas, especialmente en el sector de hospedajes temporales.

El presente trabajo se enfoca en el desarrollo y aplicación de un modelo predictivo de precios en el mercado de alquileres de corta duración, puntualmente en el ámbito de Airbnb, con el objetivo de proporcionar una herramienta aplicable y comprensible en el ámbito real.

Utilizando el lenguaje de programación Python y técnicas avanzadas de modelado, este proyecto investiga varios algoritmos para identificar el más eficaz en predecir y optimizar precios basados en un amplio espectro de variables relacionadas con los alojamientos.

A través de un proceso que incluye el análisis y exploración de datos, modelado, implementación, y evaluación de modelos, se busca seleccionar el modelo óptimo basado en su rendimiento y precisión, según métricas de evaluación establecidas.

Este trabajo no solo aspira a contribuir al campo académico y práctico con un modelo predictivo de precios preciso y accesible sino también a ofrecer insights valiosos para anfitriones y gestores de propiedades en la optimización de sus estrategias de precios en plataformas de economía compartida como Airbnb.

Palabras clave

Economía Compartida; Optimización de precios; Alojamientos; Huésped; Predicción de precios; Airbnb; Río de Janeiro; Modelos de predicción.

Índice

Índice de tablas

Índice de ilustraciones

1. Introducción

Motivación del tema

La fijación de precios es un aspecto crítico en la gestión empresarial, con un impacto directo en la rentabilidad y la competitividad de las empresas en un mercado cada vez más dinámico y competitivo. En el contexto de la economía actual, donde son cada vez más los actores que actúan bajo el concepto de "economía compartida", plataformas como Airbnb han transformado la forma en que las personas buscan y ofrecen alojamiento temporal. Esta evolución ha planteado desafíos significativos en la fijación de precios, ya que los anfitriones deben tomar decisiones estratégicas para optimizar sus ingresos y la ocupación de sus propiedades, mientras que los huéspedes buscan ofertas atractivas.

La motivación detrás de esta investigación surge de la creciente importancia de establecer precios efectivos en el mercado de alquileres de corta duración de Airbnb. La plataforma ha experimentado un crecimiento exponencial en los últimos años, convirtiéndose en un actor relevante en la industria de los alquileres de alojamientos. Sin embargo, la optimización de precios en este entorno presenta desafíos interesantes debido a la diversidad de propiedades, la variabilidad de la demanda estacional, la influencia de la ubicación, la situación económica del país en cuestión, así como la de la región y la dinámica competitiva, entre algunas a destacar.

En este punto yace la intersección entre la predicción y la optimización. La predicción, en este contexto, refiere al uso de datos históricos para anticipar la demanda futura, las preferencias de los huéspedes, y las dinámicas de precios. Estas predicciones son fundamentales, ya que proporcionan una base sólida para la toma de decisiones estratégicas en la fijación de precios.

La optimización, por otro lado, se enfoca en aplicar estas predicciones para formular estrategias de precios que maximicen los objetivos de los anfitriones, como maximizar los ingresos o la ocupación. Este proceso implica ajustar los precios dinámicamente en respuesta a las variaciones anticipadas en la demanda y otras variables críticas, como la estacionalidad, eventos locales, y la competencia.

La transición de la predicción a la optimización representa una evolución de comprender qué podría suceder en el mercado, a tomar acciones concretas que alineen la oferta de alojamiento con las tendencias del mercado anticipadas.

Sin embargo, la tarea de evaluar los resultados de una estrategia de optimización de precios se encuentra imposibilitada ya que no se cuenta con los datos disponibles para poder evaluar la *performance* del modelo. La razón es que en este trabajo de investigación no se tiene una idea precisa de cuál será el porcentaje de ocupación con los precios sugeridos sin probarlos empíricamente. En el apartado Antecedentes de este trabajo se amplía esta problemática.

En este trabajo, se pone un énfasis particular en la predicción de precios alimentada con información histórica real de Airbnb. Estas predicciones no solo ayudarán a entender mejor el mercado y sus tendencias, sino que también permitirán a los anfitriones ajustar sus estrategias de precios con información real.

Este trabajo busca abordar la necesidad de estrategias de fijación de precios más efectivas y precisas en el mercado de Airbnb. La motivación radica en los beneficios potenciales para los anfitriones, quienes pueden mejorar sus ingresos y ocupación, así como para los huéspedes, que pueden encontrar opciones atractivas y asequibles.

Finalmente, cabe destacar que los autores del presente trabajo se han formado como Contadores Públicos y la optimización y/o predicción de precios es una competencia relevante que puede ser requerida en futuros roles profesionales.

1.1. Objetivos

El propósito principal de este estudio es analizar y desarrollar estrategias avanzadas de predicción de precios en la industria de alquiler de hospedajes, enfocándose en interpretar de manera efectiva dichas estrategias para potenciar la rentabilidad y competitividad de las empresas del sector. Este objetivo implica identificar y examinar exhaustivamente los factores claves que inciden en la toma de decisiones de predicción de precios. Entre estos, se incluyen aspectos como la competencia en el mercado, las características específicas de los inmuebles y su entorno, y la percepción del valor por parte de los consumidores. Además, se buscará entender cómo las tendencias del mercado, así como los patrones estacionales,

haciendo referencia a los meses del año, pueden influir en las estrategias de precios. El estudio aspira a integrar tecnologías de análisis de datos y aprendizaje automático para proporcionar una visión que abarque desde el contexto económico hasta los detalles de las preferencias de los clientes y las particularidades de los alojamientos.

1.1.1. Resultados esperados

Desarrollar modelos que sean capaces de predecir los precios de los alojamientos de Airbnb con un alto nivel de precisión, y que sean útiles para la toma de decisiones.

Adicionalmente, que estos puedan servir de base para futuras investigaciones y/o trabajos que crucen las fronteras de la predicción para llegar al campo de la optimización de precios.

1.3 Estructura del documento

Este documento se organiza en un formato estructurado para facilitar la comprensión del estudio realizado sobre la optimización de precios en el sector de hospedajes, con especial enfoque en la plataforma Airbnb. A continuación, se detalla la estructura y el contenido principal de cada capítulo:

Capítulo 1: Introducción

Este capítulo establece el contexto y la relevancia del estudio, delineando la problemática central de la optimización de precios en la economía compartida, específicamente en el mercado de alquileres de corta duración. Se discute la motivación detrás del proyecto, los objetivos perseguidos, y la importancia de desarrollar modelos predictivos de precios.

Capítulo 2: Antecedentes

Se presenta una revisión de la literatura relacionada, explorando trabajos previos en el ámbito de la predicción y optimización de precios en plataformas de alojamientos. Este capítulo subraya las metodologías utilizadas en estudios anteriores y destaca las brechas de conocimiento que el presente estudio busca llenar.

Capítulo 3: Marco Teórico

Este capítulo ofrece una base teórica sólida que respalda la investigación, abordando conceptos claves como economía compartida, modelos de predicción de precios, y el papel de Airbnb en el mercado global de alojamientos. Además, se introducen las bases teóricas de los algoritmos de aprendizaje automático aplicados en el estudio.

Capítulo 4: Consideraciones iniciales y Selección de Datos

Se detalla el proceso de selección y preparación de los conjuntos de datos utilizados para el modelado. Este capítulo describe las fuentes de datos, los criterios de inclusión y exclusión, y las técnicas de limpieza y transformación de datos aplicadas para asegurar la calidad y relevancia del análisis.

Capítulo 5: Recopilación y Análisis de Datos

Aquí se expone el análisis exploratorio de los datos, identificando patrones, tendencias y relaciones claves que informan el desarrollo de los modelos predictivos. Se discuten las variables significativas, la distribución de los precios, y otros hallazgos relevantes derivados de los datos.

Capítulo 6: Exploración de Datos

Se profundiza en el análisis exploratorio, empleando técnicas estadísticas y visuales para comprender mejor la dinámica del mercado de alojamiento y los factores que influyen en los precios. Este capítulo prepara el terreno para la fase de modelado.

Capítulo 7: Modelado

Este capítulo central del documento describe la metodología de modelado adoptada, incluyendo la selección de algoritmos, la configuración de hiper parámetros, y el proceso de entrenamiento y validación de los modelos. Se comparan los rendimientos de diferentes modelos para seleccionar el más adecuado para la predicción de precios.

Capítulo 8: Interpretabilidad y Aplicabilidad del Modelo Seleccionado

Se discute la interpretabilidad de los resultados obtenidos y cómo estos pueden ser aplicados prácticamente para la optimización de precios en Airbnb. Este capítulo destaca la relevancia práctica del modelo desarrollado, ofreciendo recomendaciones para su implementación.

Capítulo 9: Conclusiones

El documento concluye con un resumen de los hallazgos más importantes, la relevancia del estudio para el campo de la economía compartida y la optimización de precios, y las contribuciones específicas del trabajo. También se reflexiona sobre las limitaciones del estudio y se sugieren direcciones para investigaciones futuras.

Capítulo 10: Acciones Futuras

Se esbozan recomendaciones para la continuación de la investigación y la aplicación del modelo en contextos reales. Este capítulo propone pasos futuros para mejorar el modelo predictivo, explorar nuevas metodologías y expandir la aplicación del estudio a otros mercados o plataformas de la economía compartida.

2. Antecedentes

En el ámbito de la predicción de precios de alojamientos en plataformas como Airbnb, un trabajo de referencia que fue considerado en la etapa de investigación, es el proyecto documentado en un *notebook* de Jupyter titulado "*Airbnb - Price Prediction*" publicado en GitHub por un usuario de la plataforma [1]. Este estudio presenta el desarrollo de un modelo predictivo de precios utilizando un amplio conjunto de datos de listados de Airbnb de todo el mundo.

El proyecto se centra en la asignación de precios adecuados a los alojamientos listados en Airbnb. Para esto, utiliza un conjunto de datos extenso, dividido en un conjunto de entrenamiento y un conjunto de pruebas, conteniendo 84 variables predictoras. Esta gran diversidad de datos ofrece un terreno amplio para el análisis y la modelización predictiva.

Una parte fundamental del estudio es la exploración de datos, donde se examinan en detalle las múltiples categorías de variables, incluyendo información del anfitrión, características del alojamiento, y métricas de revisión. La extensa variedad de datos abarca desde descripciones textuales y variables categóricas hasta métricas numéricas y datos relacionados con fechas, proporcionando una base comprensiva para un análisis exhaustivo.

El proyecto no solo describe la construcción de un modelo base, sino que también aborda la optimización mediante ajustes de hiper parámetros. Inicia con un "modelo base", que aunque no está detallado exhaustivamente en las secciones revisadas, se menciona como el punto de partida para la modelización predictiva. Este modelo base sienta las bases para el desarrollo posterior.

Se destaca la implementación de una arquitectura de red neuronal de "*feed forward*" para el modelo predictivo. Este tipo de red neuronal, caracterizada por su flujo de información unidireccional, es frecuentemente aplicada en tareas de regresión y clasificación, lo que sugiere su adecuación para la predicción de precios en un contexto de mercado dinámico como es Airbnb.

El estudio aprovecha las capacidades de Keras, una interfaz de alto nivel para redes neuronales, operando sobre TensorFlow, una biblioteca de aprendizaje profundo más compleja y versátil.

El trabajo también muestra una sección significativa dedicada al ajuste de hiper parámetros. Esto implica una búsqueda detallada de la configuración óptima de parámetros, como el número de capas, neuronas, tasas de aprendizaje y funciones de activación, para mejorar la precisión y la capacidad predictiva del modelo.

Como conclusión se encuentra que este trabajo es de gran relevancia para cualquier investigación en la predicción de precios en Airbnb, no solo por su enfoque analítico y metodológico, sino también por su aplicación práctica en un contexto real. Ofrece un ejemplo valioso de cómo los datos pueden ser transformados en información significativa para la toma de decisiones en el mercado de alojamientos temporales.

Otro trabajo considerado en la investigación fue el estudio llevado a cabo por Dimitris Mertikas sobre la modelación de precios dinámicos para Airbnb en Ámsterdam, el cual representa un importante antecedente en la investigación sobre estrategias de optimización de precios en el sector de alquileres de corta duración. En este estudio, Mertikas se propuso explorar y desarrollar un modelo predictivo que permitiera ajustar los precios de manera dinámica, teniendo en cuenta factores clave como la demanda del mercado, la competencia existente y la disponibilidad de las propiedades.

Una de las principales motivaciones detrás de esta investigación fue la necesidad de evaluar y comparar los resultados obtenidos mediante el uso de modelos de precios dinámicos con respecto a los esquemas tradicionales de precios fijos. Esta comparación es fundamental para comprender cómo las estrategias de fijación de precios pueden influir en la rentabilidad y competitividad de los anfitriones en plataformas como Airbnb.

Para llevar a cabo su análisis, utilizó una fuente de datos confiable y ampliamente reconocida en la comunidad académica y empresarial: "Inside Airbnb". Esta plataforma proporciona acceso a una gran cantidad de información detallada sobre las propiedades disponibles en Ámsterdam, incluyendo listados, calendarios de reservas, características de las propiedades y precios históricos.

El enfoque metodológico adoptado por Mertikas se centró en la aplicación de algoritmos avanzados, como el algoritmo de agrupamiento k-means, para segmentar las propiedades en *clusters* o grupos similares. Este enfoque permitió una mejor comprensión de la diversidad y heterogeneidad del mercado de alquileres cortos en la capital holandesa, lo cual es fundamental para diseñar estrategias de fijación de precios efectivas y personalizadas.

Además del análisis de los datos, se evaluó el impacto de variables dinámicas, como eventos locales, tendencias estacionales y cambios en la demanda turística, en los precios de alquiler. Este enfoque dinámico y proactivo en la modelación de precios busca maximizar los ingresos para los anfitriones de Airbnb al adaptar los precios de manera oportuna y eficiente a las condiciones cambiantes del mercado.

El modelo implementado en este trabajo ha mostrado un buen rendimiento para el caso de estudio al considerar el efecto del precio y disponibilidad de los competidores dentro del mismo grupo de propiedades. No obstante, tal como indica el autor en la sección de evaluación, esta técnica tiene la limitación de no poder realizar una evaluación precisa de los resultados, dado que no se disponen de datos reales que permitan prever el nivel de ocupación con los precios propuestos, sin haberlos probado previamente en la práctica, algo que está más allá del alcance de nuestro proyecto. Esto dificulta la comparación entre el rendimiento "ideal" y el rendimiento real obtenido en el año anterior. Esta limitación, junto con el hecho de que se obtuvo esta referencia cerca del plazo de entrega como para implementar medidas que mitigaran el problema, desalentó la adopción de esta estrategia en el presente trabajo y confirmó la decisión a favor del enfoque de predicción [2] .

3. Marco Teórico

Este marco teórico proporciona una base conceptual y contextual para la investigación sobre la predicción de precios en alojamientos de Airbnb. Se exploran los aspectos clave de la economía compartida, el papel de Airbnb en el mercado global, y los fundamentos de los modelos de predicción de precios, ofreciendo así una comprensión integral que sustenta nuestro estudio.

Airbnb, desde su creación en 2008, ha revolucionado el mercado de alojamientos, ofreciendo una plataforma para que personas de todo el mundo alquilen sus espacios. Este modelo de negocio no sólo ha impactado la industria hotelera tradicional, sino que también ha generado nuevas dinámicas económicas y sociales, posicionando a Airbnb como un actor clave en el turismo global.

La predicción de precios es esencial en mercados dinámicos como el de los alojamientos temporales. Los precios en Airbnb son influenciados por múltiples factores: ubicación, temporada, demanda, y características específicas del alojamiento. Comprender estos factores es crucial para desarrollar modelos predictivos eficaces.

Los modelos de aprendizaje automático han ganado popularidad en la predicción de precios de alojamientos. Estos modelos se alimentan de grandes conjuntos de datos, incluyendo precios históricos, características de los alojamientos, y tendencias de mercado, para generar predicciones precisas. Estudios anteriores han demostrado el éxito de métodos como la regresión lineal, los árboles de decisión y las redes neuronales en este campo.

La predicción de precios no solo beneficia a los anfitriones en términos de maximizar sus ingresos, sino que también puede influir en las decisiones de los huéspedes y en la accesibilidad de alojamientos. Además, la fijación de precios ética y socialmente responsable es un área de creciente interés, teniendo en cuenta su impacto en las comunidades locales y la economía en general.

Los desafíos en la predicción de precios para Airbnb incluyen la adaptación a los cambios rápidos del mercado y la gestión de datos de gran volumen. Las tendencias futuras apuntan hacia la integración de tecnologías avanzadas como el aprendizaje profundo y el

análisis de *big data*, lo que promete mejorar la precisión y eficiencia de los modelos predictivos.

Este marco teórico establece las bases para una investigación detallada sobre la predicción de precios en alojamientos de Airbnb. Al abordar desde los fundamentos económicos y sociales hasta los aspectos técnicos de los modelos predictivos, este marco proporciona una perspectiva integral que es fundamental para el análisis y comprensión del fenómeno estudiado.

3.1 Economía Compartida

Los términos "economía compartida" y "consumo colaborativo" son ampliamente utilizados para describir la práctica de compartir directamente bienes y servicios entre personas, a través de plataformas en línea basadas en la comunidad.

Según expertos en consultoría, se estima que para el año 2025, la economía compartida podría generar ingresos globales de hasta $335 billones, en comparación con los $15 billones registrados en 2015. Por ejemplo, la valoración de Uber alcanza los $69 billones, superando a grandes empresas como American Airlines con $21.1 billones y Southwest Airlines con $33.6 billones.

Este modelo económico permite el intercambio de bienes y servicios entre miembros de una comunidad a través de plataformas digitales. Esta práctica ha transformado la dinámica del turismo y la hospitalidad, desafiando la forma tradicional de hacer negocios en estos sectores. Los alojamientos basados en redes *peer-to-peer* (P2P) han ganado terreno, ofreciendo a los viajeros la oportunidad de experimentar una estancia más auténtica y económica.

Las plataformas P2P han simplificado la distribución de alojamientos, siendo Airbnb una de las más destacadas al ampliar su oferta más allá de los barrios tradicionales donde se ubican los hoteles. Esto ha generado un cambio significativo en la industria hotelera, especialmente en las grandes ciudades, donde se compite directamente con el turismo de masas.

Los viajeros prefieren los alojamientos P2P debido a sus precios más bajos y a las ventajas que ofrecen, como estadías más largas y frecuentes. Estas experiencias suelen ser

percibidas como más auténticas y diferenciadas en comparación con los hospedajes tradicionales [3].

3.2 Acerca de AIRBNB

Según la plataforma Stays, quienes son especialistas en rentas por temporada, y quienes interactúan con Airbnb desde 2017, definen Airbnb como "*una plataforma que ofrece a anfitriones en todo el mundo la opción de alquilar el espacio que tengan disponible, ya sea una propiedad completa o parte de ella.*

Esta simple pero genial idea cambió definitivamente la manera de hospedarse en el mundo. Los viajeros ahora cuentan con una gran variedad de opciones para alojarse, de diferentes precios y condiciones. Además, miles de personas en el mundo pueden aprovechar esta oportunidad para generar un nuevo ingreso.

Un anfitrión de Airbnb cuenta con la opción de anunciar por medio de la plataforma cualquier espacio que tenga disponible, puede ser una habitación a compartir en una casa, apartamento, hotel o hostel; como también puede ser una propiedad completa, como una casa o cabaña" [4].

Desde Entorno Turístico, plataforma dedicada principalmente a informar el acontecer diario del medio turístico local, nacional e internacional, definen a Airbnb como *"un mercado comunitario que sirve para publicar, dar publicidad y reservar alojamiento de forma económica en más de 190 países a través de internet o desde tu smartphone. Está basado en la modalidad "Bed and Breakfast" (de donde proviene el "bnb").*

Es uno de los sistemas más exitosos de la economía colaborativa (sistema económico en el que se comparten e intercambian bienes y servicios entre particulares a través de plataformas digitales). Éste sistema permite al usuario encontrar alojamiento, con la diferencia de que no será en un hotel sino en el hogar de una persona que puede incluso estar viviendo en él. Lo interesante es que podrás alquilar desde apartamentos comunes hasta casas del árbol, iglús, geodomos, molinos, etc"[5].

Según la propia empresa, Airbnb nació en 2007 cuando dos anfitriones recibieron a tres huéspedes en su casa de San Francisco y desde entonces ha crecido hasta alcanzar más de 4 millones de anfitriones que han recibido a más de 1.500 millones de huéspedes en casi

todos los países del mundo. Cada día, los anfitriones ofrecen estadías y experiencias únicas que hacen posible que los huéspedes se conecten con las comunidades de una manera más auténtica [6].

Como resumen de lo antes mencionado, se puede decir que Airbnb se dedica a conectar a personas que buscan alojamiento con anfitriones que tienen espacio para ofrecer. Los usuarios pueden alquilar desde una habitación individual hasta una casa completa para estancias cortas. La plataforma actúa como intermediario, facilitando la transacción y proporcionando un sistema de revisión para fomentar la confianza entre anfitriones y huéspedes.

El modelo de negocio se basa en operar como una plataforma en línea que permite a los anfitriones poner a disposición sus propiedades para alquiler temporal. Los usuarios pueden explorar las opciones de alojamiento en función de diversos criterios, como ubicación, precio y tipo de propiedad. La plataforma facilita el proceso de reserva y gestiona los pagos.

3.2.1 Cómo se fija el precio en Airbnb

En este punto es importante destacar que, si bien no es la plataforma la que define los precios de los alojamientos, sino que lo hace el anfitrión, Airbnb proporciona ciertas herramientas y consejos útiles para definir el precio más adecuado.

Airbnb principalmente recomienda :

- Consultar el desglose de la tarifa para entender cómo se calcula el precio total para los huéspedes y lo que se ganará.
- Probar la función Anuncios similares, que permite comparar el precio con el de otros alojamientos similares que se han reservado en la misma zona.
- Plantearse ofrecer un descuento semanal o mensual a los huéspedes.
- Prever si cobrar una tarifa de limpieza o no [7].

Es importante destacar en este punto la conexión que existe entre el presente trabajo y la metodología de fijación de precios en Airbnb.

Como se mencionó anteriormente, es el anfitrión quien define el precio de su alojamiento, pero, ningún anfitrión fijará un precio que lo deje fuera de competencia y por eso es enriquecedor contar con modelos que predigan los precios de los alojamientos en base a características en común, lo que generará que un alojamiento sea o no similar a otro.

¿Qué son los anuncios similares?

Los anuncios similares son aquellos que muestran alojamientos que coinciden o tienen ciertos factores, o variables en este contexto, similares. Airbnb destaca: ubicación, tamaño, características, servicios, calificaciones, evaluaciones y otros anuncios que los huéspedes exploran mientras consideran reservar el tuyo [8].

3.3 Optimización de precios

La optimización de precios es un proceso que consiste en fijar el precio adecuado de un producto o servicio para maximizar los ingresos y los beneficios.

Se puede decir que la optimización de precios consiste en averiguar el precio ideal de un producto o servicio, considerando múltiples factores. Por ejemplo, hay que tener en cuenta el costo de producir un producto o prestar un servicio, el precio de los competidores y cuánto están dispuestos a pagar los clientes.

Cabe recordar que la optimización de precios no es algo que se haga una sola vez. Los precios y las condiciones del mercado pueden cambiar con el tiempo, por lo que se debe ajustar la estrategia de precios según sea necesario [9].

Actualmente, en el uso de esta técnica en la práctica de negocios se la define como una "*técnica de análisis de datos para alcanzar dos objetivos principales: comprender cómo los clientes reaccionan a diferentes estrategias de precios y encontrar los mejores precios considerando los objetivos de la empresa. Los sistemas de precios han evolucionado desde estrategias simples hasta predecir la demanda y encontrar el precio óptimo. Además de optimizar precios, entender las preferencias de los consumidores puede servir para otras aplicaciones, como la optimización del surtido de productos. Las técnicas actuales permiten considerar factores como la competencia, clima, estaciones, eventos especiales, variables macroeconómicas, costos operativos e información de almacén para determinar el precio inicial, óptimo, de descuento y promocional*" [11].

3.4 KNN

Este modelo de aprendizaje supervisado se utiliza para completar los valores faltantes en los datos seleccionados utilizando el método denominado KNN. Este algoritmo reemplaza los valores faltantes en el conjunto de datos con el valor promedio de los 'n_neighbors' vecinos más cercanos encontrados en el conjunto de entrenamiento. Para ello se debe seleccionar la medida de distancia (por ejemplo, Euclidiana) y el número de vecinos contribuyentes para cada predicción, es decir, el hiper parámetro 'k' del algoritmo KNN [15].

3.5 Grid Search

Grid Search es una técnica utilizada para encontrar los mejores hiper parámetros para un modelo de aprendizaje automático. Se realiza una búsqueda exhaustiva a través de una cuadrícula de combinaciones predefinidas de hiper parámetros y se evalúa el rendimiento del modelo utilizando validación cruzada para determinar cuáles son los mejores valores de de los mismos [12].

3.6 Cross Validation

La validación cruzada (*cross-validation*) es una técnica estadística utilizada para evaluar la habilidad de modelos predictivos al dividir el conjunto de datos en partes para entrenar y probar el modelo en diferentes iteraciones. Esto ayuda a asegurar que el modelo es capaz de realizar predicciones precisas sobre datos no vistos, mitigando problemas como el sobreajuste y proporcionando una estimación más fiable de su rendimiento en la práctica.

Una de las formas más comunes de validación cruzada es la *k-fold cross-validation*. Aquí, el conjunto de datos se divide aleatoriamente en k grupos (o "*folds*") de aproximadamente igual tamaño. El modelo se entrena k veces, cada vez usando k-1 grupos como datos de entrenamiento y el grupo restante como datos de prueba. Esto significa que cada grupo se utiliza como conjunto de prueba exactamente una vez. Al final, se calcula el rendimiento promedio del modelo sobre los k entrenamientos para obtener una medida más precisa de su eficacia.

La validación cruzada es especialmente útil cuando se trabaja con conjuntos de datos limitados, ya que permite maximizar el uso de los datos disponibles para el entrenamiento y la evaluación del modelo, proporcionando así una mejor estimación de cómo el modelo actuará en el mundo real [13].

3.7 Permutation Importance

Es una técnica de evaluación de la importancia de las características en modelos de aprendizaje automático. *Permutation Importance* es una clase en *scikit-learn* que permite calcular la importancia de las características mediante la permutación de sus valores y la observación de cómo afecta esto al rendimiento del modelo.

La idea fundamental detrás de la importancia de permutación es evaluar qué tan crucial es cada característica para el rendimiento del modelo. Se realiza de la siguiente manera:

- Se entrena un modelo utilizando el conjunto de datos original y se evalúa su rendimiento en un conjunto de prueba.
- Para cada característica, los valores de esa característica se permutan aleatoriamente en el conjunto de prueba, rompiendo la relación original entre la característica y la variable objetivo.
- Se vuelve a evaluar el rendimiento del modelo en el conjunto de prueba permutado.
- La importancia de la característica se calcula como la diferencia entre el rendimiento original y el rendimiento después de la permutación. Una gran disminución en el rendimiento indica que la característica es importante para el modelo [14].

3.8 Predicción de precios

La predicción de precios es el proceso de utilizar datos históricos, análisis estadísticos y modelos matemáticos para estimar futuros precios de bienes, servicios, activos financieros o propiedades. Este enfoque se basa en la identificación de patrones, tendencias y correlaciones en los datos pasados y actuales para hacer proyecciones sobre cómo se

comportarán los precios en el futuro. Los modelos de predicción de precios pueden incluir una amplia gama de técnicas, desde métodos estadísticos simples hasta complejos algoritmos de aprendizaje automático y de inteligencia artificial, adaptándose a la naturaleza específica del mercado o sector en cuestión.

3.9 Regresión Lineal

La regresión lineal es un método estadístico utilizado para predecir una variable dependiente continua (variable objetivo) basada en una o más variables independientes (variables predictoras). Esta técnica asume una relación lineal entre la variable dependiente y las variables independientes, lo que implica que la variable dependiente cambia de manera proporcional con los cambios en las variables independientes.

Ventajas

- Fácil de interpretar: Los coeficientes de un modelo de regresión lineal representan el cambio en la variable dependiente para un cambio de una unidad en la variable independiente, lo que facilita comprender la relación entre las variables.

- Robusto ante valores atípicos: La regresión lineal es relativamente robusta ante valores atípicos, lo que significa que se ve menos afectada por valores extremos de la variable independiente en comparación con otros métodos estadísticos.

Desventajas

- Supone linealidad: La regresión lineal asume que la relación entre la variable independiente y la variable dependiente es lineal. Esta suposición puede no ser válida para todos los conjuntos de datos.

- Puede no ser adecuado para relaciones altamente complejas: La regresión lineal puede no ser adecuada para modelar relaciones altamente complejas entre variables [15].

3.10 Árboles de decisión

Un árbol de decisión es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable de respuesta o dependiente.

Para dividir el espacio muestral en sub-regiones es preciso aplicar una serie de reglas o decisiones, para que cada sub-región contenga la mayor proporción posible de individuos de una de las poblaciones.

Si una sub-región contiene datos de diferentes clases, se subdivide en regiones más pequeñas hasta fragmentar el espacio en sub-regiones menores que integran datos de la misma clase.

Con este modelo se pueden resolver problemas tanto de regresión como de clasificación. En el presente trabajo se utilizará para predecir la variable referente al precio de los alojamientos.

Ventajas

- Facilidad en su construcción e interpretación.
- Pueden capturar relaciones no lineales sin necesidad de transformación de datos.
- Son capaces de manejar tanto variables categóricas como numéricas.
- No es necesario normalizar o escalar las características.

Desventajas

- Tienen tendencia a ajustarse demasiado a los datos de entrenamiento (sobreajuste), lo que puede llevar a predicciones inexactas con nuevos datos.
- Pequeños cambios en los datos pueden resultar en árboles de decisión muy diferentes [16].

3.11 Random Forest

Random Forest es un algoritmo de aprendizaje automático supervisado que se utiliza para problemas de clasificación y regresión. Construye múltiples árboles de decisión a partir de diferentes muestras y toma el voto mayoritario para la clasificación y el promedio para la regresión.

El algoritmo funciona mediante el principio de ensamble, donde se crea un subconjunto aleatorio de datos de muestra y se entrena un árbol de decisión independiente para cada muestra. Luego, se combina la salida de todos los árboles mediante votación por mayoría para la clasificación y promedio para la regresión.

Ventajas

- Puede usarse en problemas de clasificación y regresión.
- Resuelve el problema del sobreajuste.
- Funciona bien con datos nulos o faltantes.
- Es estable y paralelizable.
- No se ve afectado por el aumento de dimensiones.

Desventajas

- Es más complejo y requiere más tiempo de entrenamiento que los árboles de decisión.
- Requiere más tiempo de procesamiento debido a su complejidad [17].

3.12 Xgboost

XGBoost es una biblioteca de aumento de gradiente distribuido optimizada, diseñada para el entrenamiento eficiente y escalable de modelos de aprendizaje automático. Se trata de un método de aprendizaje de conjuntos que combina las predicciones de múltiples modelos débiles para generar una predicción más sólida. Con el acrónimo "*Extreme Gradient*

Boosting", se ha consolidado como uno de los algoritmos más populares y ampliamente utilizados en el campo del aprendizaje automático, gracias a su capacidad para manejar conjuntos de datos extensos y su habilidad para alcanzar un rendimiento de vanguardia en tareas como clasificación y regresión.

Ventajas

- Rendimiento: Ha demostrado su capacidad para generar resultados de alta calidad en una variedad de tareas de aprendizaje automático, destacándose especialmente en competiciones de Kaggle, donde ha sido ampliamente preferido para soluciones exitosas.

- Escalabilidad: Está diseñado para el entrenamiento eficiente y escalable de modelos de aprendizaje automático, lo que lo hace adecuado para conjuntos de datos grandes.

- Personalización: cuenta con una amplia gama de hiper parámetros que se pueden ajustar para optimizar el rendimiento, lo que lo hace altamente personalizable.

- Valores Faltantes: *XGBoost* tiene soporte incorporado para manejar valores faltantes, lo que facilita el trabajo con datos del mundo real que a menudo tienen valores faltantes.

- Interpretabilidad: Proporciona la importancia de características, lo que permite una mejor comprensión de qué variables son más importantes para hacer predicciones.

Desventajas

- Complejidad computacional: Puede demandar una cantidad significativa de recursos computacionales, especialmente durante el entrenamiento de modelos extensos, lo que podría limitar su idoneidad para sistemas con capacidades limitadas.

- Requisitos de memoria: *XGBoost* puede requerir mucha memoria,

especialmente al trabajar con conjuntos de datos grandes, lo que lo hace menos adecuado para sistemas con recursos limitados de memoria [18].

3.13 R^2

R^2 es una métrica que determina la proporción de variabilidad en los datos explicada por el modelo, brindando así una evaluación de qué tan bien se ajusta el modelo a los datos.

R^2 toma valores entre 0 y 1, siendo los más cercanos a 1, los modelos que mejor se ajustan a los datos.

La ecuación del coeficiente de determinación (R2) se expresa como:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- TSS es la suma total de cuadrados.
- RSS es la suma residual de cuadrados. Mide la cantidad de variabilidad que queda sin explicar después de realizar la regresión [13].

3.14 Error Cuadrático Medio (MSE)

MSE (*Mean Squared Error*) mide el error cuadrado promedio de las predicciones.

Para cada observación, calcula la diferencia cuadrada entre las predicciones y el objetivo y posteriormente promedia esos valores [13].

La ecuación de MSE se expresa como:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

3.15 RMSE

El RMSE (*Root Mean Squared Error*) es una medida que evalúa la diferencia entre los valores que un modelo predice y los valores que se observan realmente. Es la raíz

cuadrada del promedio de los cuadrados de las diferencias entre las predicciones y los valores reales.

RMSE es útil porque penaliza errores grandes al elevar al cuadrado las diferencias antes de promediarlas, lo que significa que grandes errores tienen un impacto mayor que los pequeños. Una RMSE baja indica un mejor ajuste del modelo a los datos. Su fórmula es la raíz cuadrada del MSE.

4. Consideraciones iniciales y Selección de datos

En la fase inicial de este proyecto, se contempló la posibilidad de trabajar con un conjunto de datos de Buenos Aires, Argentina. La elección inicial se fundamentó en la

proximidad geográfica y coincidencias socio-culturales, lo que facilita la comprensión de los datos y comportamientos. Sin embargo, durante la primera exploración básica de los datos, se enfrenta la complejidad inherente a una economía hiperinflacionaria. Esta complejidad añade un nivel de análisis significativo que, para los objetivos específicos de este trabajo, no agrega valor sustancial.

A continuación, se presenta un gráfico que ilustra de manera representativa la evolución de los datos, destacando las variaciones de precios a lo largo del tiempo. Este análisis visual proporcionará una perspectiva más clara sobre la dinámica de los datos seleccionados.

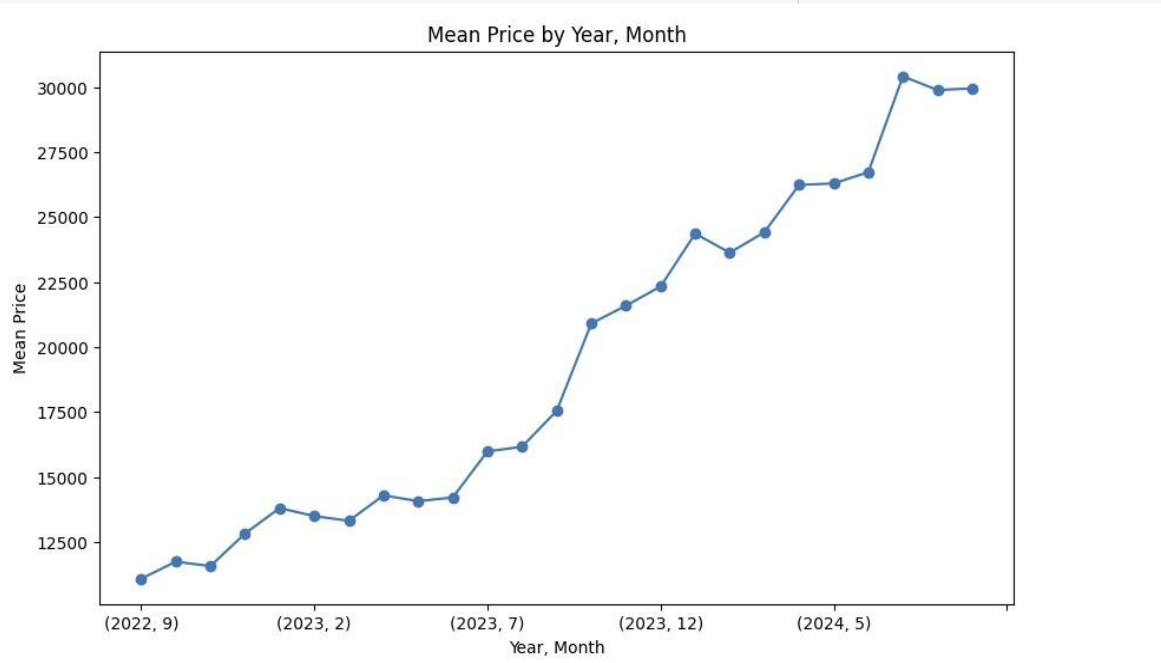


Figura 4.1: Evolución del Precio Medio , Buenos Aires

Luego de evaluar esta situación, se opta por seleccionar datos de otra ciudad como punto de partida para continuar con el desarrollo del trabajo.

5. Recopilación y Análisis de datos

Como primer punto es importante destacar que todos los conjuntos de datos utilizados en este estudio, junto con las *notebooks* de preparación y exploración de datos, así como los *scripts* de entrenamiento de modelos, están disponibles de forma pública en

un repositorio de GitHub (anexo 2). Este repositorio proporciona acceso a los datos originales, los pasos de preprocesamiento, visualización y análisis exploratorio, así como las implementaciones de los modelos de aprendizaje automático utilizados en este trabajo.

En el proceso de selección de datos para el presente trabajo, se ha optado por utilizar conjuntos de datos pertinentes a la ciudad de Río de Janeiro situada en Brasil.

Tras considerar y descartar datos de Argentina, esta elección se basa en la relevancia y riqueza de los datos disponibles para la investigación, así como en la diversidad de información que la ciudad ofrece. Al elegir Río de Janeiro como el contexto principal de estudio, se busca obtener *insights* significativos y aplicables que enriquezcan la calidad y la validez de los resultados obtenidos.

Los datos utilizados en el presente trabajo se extraen de la plataforma Inside Airbnb [19]. Esta plataforma es un proyecto que proporciona datos y promoción sobre el impacto de Airbnb en las comunidades residenciales. El fin es que se cuente con datos e información para comprender, decidir y controlar el papel del alquiler de viviendas residenciales a turistas.

5.1 Limpieza y transformación de datos

En la etapa inicial de limpieza de datos, se abordan dos conjuntos de datos principales: dflistings y dfcalendar.

dflistings:

El *dataset* dflistings contiene información detallada sobre los alojamientos de Airbnb. Este conjunto de datos es crucial para el análisis propuesto y ofrece una perspectiva amplia sobre las características y tendencias del mercado de alquileres a corto plazo.

El *dataset* consta de 31,964 filas y 75 columnas, abarcando una variedad significativa de variables relacionadas con las propiedades ofrecidas en Airbnb. Cada fila

del *dataset* representa un alojamiento único en la plataforma, y las columnas contienen datos específicos sobre estos.

Dentro del dataset, se encuentran principalmente datos como:

Identificadores y URLs: Cada listado tiene un identificador único, acompañado de su dirección *web* correspondiente en Airbnb.

Información del Anfitrión: Detalles sobre los anfitriones de las propiedades, como sus nombres, perfiles en Airbnb, y estadísticas de respuesta.

Descripciones de las Propiedades: Información textual sobre las propiedades, incluyendo descripciones generales, vistas del vecindario, y enlaces a imágenes.

Ubicación y Características de las Propiedades: Datos sobre la ubicación de las propiedades, sus características físicas, comodidades disponibles, precios, y políticas de cancelación.

Evaluaciones y Revisiones: Información sobre las calificaciones y revisiones recibidas por cada propiedad.

Con esta información se busca realizar un análisis que permita estudiar las tendencias de precios y disponibilidad en diferentes ubicaciones y períodos.

Para una comprensión más profunda de los datos, se incluye en anexos un diccionario de variables del *dataset*. Este diccionario proporciona una descripción breve de cada columna presente en el *dataset*, ofreciendo una referencia para entender la naturaleza y el alcance de los datos analizados (Anexo 1).

dfcalendar:

El segundo *dataset* mencionado se centra en la disponibilidad y precios de las propiedades. Este conjunto de datos es esencial para comprender las dinámicas de precios y patrones de disponibilidad en la plataforma.

Como características generales este *dataset* consta de 11.666.976 filas y 7 columnas, proporcionando una visión detallada de la disponibilidad y tarifas de las propiedades de Airbnb a lo largo del tiempo. Cada fila representa un registro único de un listado en una fecha específica.

Principalmente la información contenida refiere a :

- ID del Listado: Identificador único de cada propiedad listada en Airbnb.
- Fecha: La fecha específica a la que se refiere el registro.
- Disponibilidad: Indica si la propiedad está disponible o no en esa fecha.
- Precio: El precio listado para esa fecha.
- Precio Ajustado: Cualquier ajuste en el precio original para esa fecha.
- Noches Mínimas y Máximas: Restricciones sobre la cantidad mínima y máxima de noches que se pueden reservar.

De este conjunto de datos se puede extraer información de cómo varían los precios en diferentes temporadas o eventos especiales, patrones de disponibilidad para identificar períodos de alta y baja demanda, así como los cambios en los precios afectan la disponibilidad de las propiedades.

Ambos *datasets* se unen mediante un proceso denominado *'merge'* para facilitar el análisis y tener los datos en un mismo conjunto de datos. El elemento clave para esta fusión es el identificador único de los listados (‘listing_id’) presente en ambos conjuntos de datos. Utilizando este identificador como referencia, se pueden combinar las filas de ambos *datasets*, alineando la información de los listados con sus respectivos datos de disponibilidad y precio para cada fecha.

5.1.1 Limpieza de Datos

Dataset Calendar:

- Se convierte la variable ‘price’ a tipo *integer*. La columna ‘price’ originalmente contenía valores en formato de texto, incluyendo el símbolo de la moneda local y comas, lo cual no es adecuado para análisis cuantitativos. Por ende, se realizó una conversión para transformar estos valores en enteros. Se realiza esta transformación debido a que la variable ‘price’ es la variable objetivo y es fundamental contar con datos completos para la predicción. Cabe aclarar que dicha variable representa los precios de los alojamientos en moneda local, en este caso, en reales.

- Se convierte la variable ‘date’ a tipo fecha. Originalmente, esta variable estaba en un formato de cadena (*string*), lo cual no es óptimo para realizar análisis con fechas. Esto permite realizar operaciones como ordenar cronológicamente los datos, calcular intervalos de tiempo, y agrupar datos por períodos si fuera necesario.

- Se eliminan las variables ‘adjusted_price’, ‘minimum_nights’ y ‘maximum_nights’. Se observó que los valores en la columna ‘adjusted_price’ son idénticos a los de la columna ‘price’ en todas las entradas del *dataset* y en referencia a ‘minimum_nights’ y ‘maximum_nights’, esta información se encuentra presente en el *dataset* dflistings.

- Se verifica el rango de fechas que contiene el *dataset* para contextualizar el análisis y entender las tendencias dentro del período . El mismo abarca desde el 22/09/2023 hasta el 21/09/2024. Los datos “futuros”, es decir, aquellos con fechas posteriores a la cual fueron obtenidos los datos permiten visualizar el nivel de precios de cada alojamiento disponible y también observar el precio de aquellos que fueron reservados (cabe destacar que por la naturaleza del rubro, es requisito necesario contar con información a futuro ya que las reservas siempre se ejecutan a fechas posteriores). Contar con datos posteriores permiten evaluar posibles tendencias y comportamientos de los precios en base a diferentes factores, como pueden ser: cercanía de fecha de potencial reserva, fechas específicas, temporadas del año, etc.

Dataset Listings*:*

- Se realizó la conversión de las variables ‘price’ y ‘bedrooms’ a tipo *integer*. Estas conversiones se realizan con el fin de utilizar estas variables en los análisis, ya que si tienen un formato diferente no se pueden realizar operaciones y/o gráficas.

- Se eliminan variables no relevantes. Con base en el conocimiento del negocio de Airbnb, se eliminaron múltiples variables que se consideran no relevantes para la variable objetivo. Ejemplo: ‘scrape_id’, ‘name’, ‘description’, ‘host_id’, ‘license’, ‘calendar_updated’, etc.

- Se optó por eliminar todas las filas con valores nulos en la columna ‘bathrooms_text’ debido a que los valores nulos representan un porcentaje pequeño (0,087599%) , además, la presencia y el tipo de baño se consideran inicialmente como factores significativos que pueden influir en las decisiones de alquiler de los huéspedes. Por dicho motivo se decidió que mantener registros sin esta información no agrega valor.

- Las variables ‘host_response_rate’ y ‘host_acceptance_rate’ se transformaron de porcentajes a valores decimales para una mejor manipulación y análisis.

- Las columnas ‘last_scraped’, ‘first_review’ y ‘last_review’ en fueron convertidas a formato de fecha para facilitar su análisis y manipulación en caso de ser necesario.

5.1.2 Transformación de Datos

En primer lugar se aplicaron transformaciones al *dataset* dflistings con el objetivo de reducirlo y trabajar con los registros que en principio se entendieron más relevantes para el análisis. Las conclusiones sobre qué variables excluir fueron determinadas en base al conocimiento del negocio de Airbnb y a la experiencia como usuarios de la aplicación. Las variables excluidas fueron las siguientes:

-‘scrape_id’
-‘name’
-‘description’

-‘host_id’

-‘license’

-‘calendar_updated’

-‘bathrooms’

-‘neighbourhood’

-‘neighbourhood_group_cleansed’

-‘listing_url’

-‘scrape_id’

-‘source’

-‘picture_url’

-‘host_url’

-‘host_name’

-‘host_about’

-‘host_location’

-‘host_thumbnail_url’

-‘host_picture_url’

-‘host_neighbourhood’

-‘host_has_profile_pic’

-‘host_verifications’

-‘property_type’

-‘host_listings_count’

-‘host_total_listings_count’

-‘calculated_host_listings_count_entire_homes’

-‘calculated_host_listings_count_private_rooms’

-‘calculated_host_listings_count_shared_rooms’

-‘calendar_last_scraped’

-‘neighborhood_overview’

-‘minimum_minimum_nights’

-‘maximum_minimum_nights’

-‘minimum_maximum_nights’

-‘maximum_maximum_nights’

-‘minimum_nights_avg_ntm’

-‘maximum_nights_avg_ntm’

-‘review_scores_accuracy’
-‘review_scores_cleanliness’
-‘review_scores_communication’
-‘review_scores_checkin’
-‘review_scores_location’
-‘calculated_host_listings_count’

Esto permite mejorar la calidad de los datos y que el modelo presente una mejor *performance* en cuanto a la predicción, ya que no se procesa información que en principio se entiende irrelevante para la obtención del objetivo, es decir, determinar el precio del alojamiento.

Esta acción también podría haber sido realizada luego de la unificación con dfcalendar, pero se consideró más conveniente realizarlo previamente al *merge* ya que presentaba un costo computacional menor por la cantidad de registros a procesar.

La transformación del *dataset* constó de:

- Se creó una nueva variable denominada ‘amenitie_count’ en la cual se refleja la cantidad de *amenities* que posee cada alojamiento. Se consideró como aspecto importante a la hora de la definición de precios de un hospedaje la cantidad de comodidades que el mismo ofrece, por esta razón se decidió crear esta nueva variable.

- Se modificó el nombre de la variable ‘id’ por ‘listing_id’ ya que en el *dataset* dfcalendar la variable de identificación de los alojamientos se denomina de esa manera, y de esta manera será posible la unión de ambos conjuntos de datos.

- Se realizó un agrupamiento por ‘neighbourhood_cleansed’, variable perteneciente al conjunto de datos original y que representa el barrio donde se encuentra cada alojamiento, y se filtraron los barrios con una cantidad significativa de alojamientos. Posteriormente, se redujo el *dataset* para incluir solo aquellos barrios con más de 100 alojamientos.

	neighbourhood_cleansed	count
34	Copacabana	9507
8	Barra da Tijuca	3188
65	Ipanema	3074
112	Recreio dos Bandeirantes	1600
68	Jacarepaguá	1568
78	Leblon	1514
14	Botafogo	1363
122	Santa Teresa	1088
26	Centro	987
47	Flamengo	700
79	Leme	618
77	Laranjeiras	485
19	Camorim	478
135	Tijuca	435
129	São Conrado	323
23	Catete	280
76	Lagoa	265
54	Glória	245
71	Jardim Botânico	208
59	Gávea	199
145	Vidigal	183
62	Humaitá	180
9	Barra de Guaratiba	166
67	Itanhangá	164
133	Taquara	142
147	Vila Isabel	141
141	Vargem Pequena	136
75	Joá	131
49	Freguesia (Jacarepaguá)	126
58	Guaratiba	116
84	Maracanã	115
139	Urca	115
21	Campo Grande	113
140	Vargem Grande	106
116	Rio Comprido	99
38	Curicica	84
130	São Cristóvão	83
36	Cosme Velho	81
73	Jardim Guanabara	64
98	Pechincha	58
5	Anil	57
46	Estácio	56
94	Paquetá	47
108	Praça da Bandeira	47
55	Grajaú	44
2	Alto da Boa Vista	43
89	Méier	43
45	Engenho de Dentro	41
43	Engenho Novo	38
4	Andaraí	35

Tabla 5.1.2.1: Cantidad de alojamientos por barrio de dflistings

- Se crearon nuevas variables derivadas de ‘bathrooms_text’, ‘shared’ y ‘bathrooms_count’. La variable ‘shared’ es una columna binaria que indica si el baño es compartido o no. Se asigna un valor de 1 si la palabra "shared" aparece en la descripción del baño (‘bathrooms_text’), y un valor de 0 en caso contrario. Se asume que la naturaleza del baño (compartido o privado) puede influir en el precio de alquiler de la propiedad. Generalmente, se espera que los alojamientos con baños compartidos tengan precios más bajos en comparación

con aquellos que ofrecen baños privados. En cuanto a ‘bathrooms_count’ es una variable numérica que representa la cantidad de baños que posee el alojamientos, la cual se calcula extrayendo el número de la descripción del baño en la variable ‘bathrooms_text’. Es posible que la cantidad de baños en una propiedad sea un factor que afecta tanto al precio de alquiler como a la demanda de la propiedad. Al convertir esta información en una variable numérica, se facilita su inclusión en análisis cuantitativos y modelos estadísticos.

- Se utilizó *LabelEncoder* para transformar las variables categóricas ‘host_is_superhost’ y ‘room_type’ en valores numéricos. Originalmente, la variable ‘host_is_superhost’ estaba en formato de texto, con valores 't' (*true*) y 'f' (*false*), representando si un anfitrión es considerado *superhost* (buen anfitrión) o no. Para mejorar la eficiencia en el análisis y en el entrenamiento de modelos predictivos, se transformó esta variable categórica a un formato numérico binario. En el caso de la variable ‘room_type’, originalmente contenía información en formato string que especificaba el tipo de habitación, específicamente, si es compartida o no, así como en el caso anterior, se transformó la variable a un formato numérico binario.

- Se realiza mediante función *merge* de la biblioteca pandas, la unión de ambos *datasets*: dflistings y dfcalendar. Esta acción se realizó con el objetivo de contar con un solo *dataset* con la totalidad de los datos, el cual contiene 40 variables y 9.632.459 registros, siendo estos registros una representación de cada día del año para cada uno de los alojamientos que conforman el conjunto de datos (26.390). Esta acción facilita considerablemente la transformación de datos.

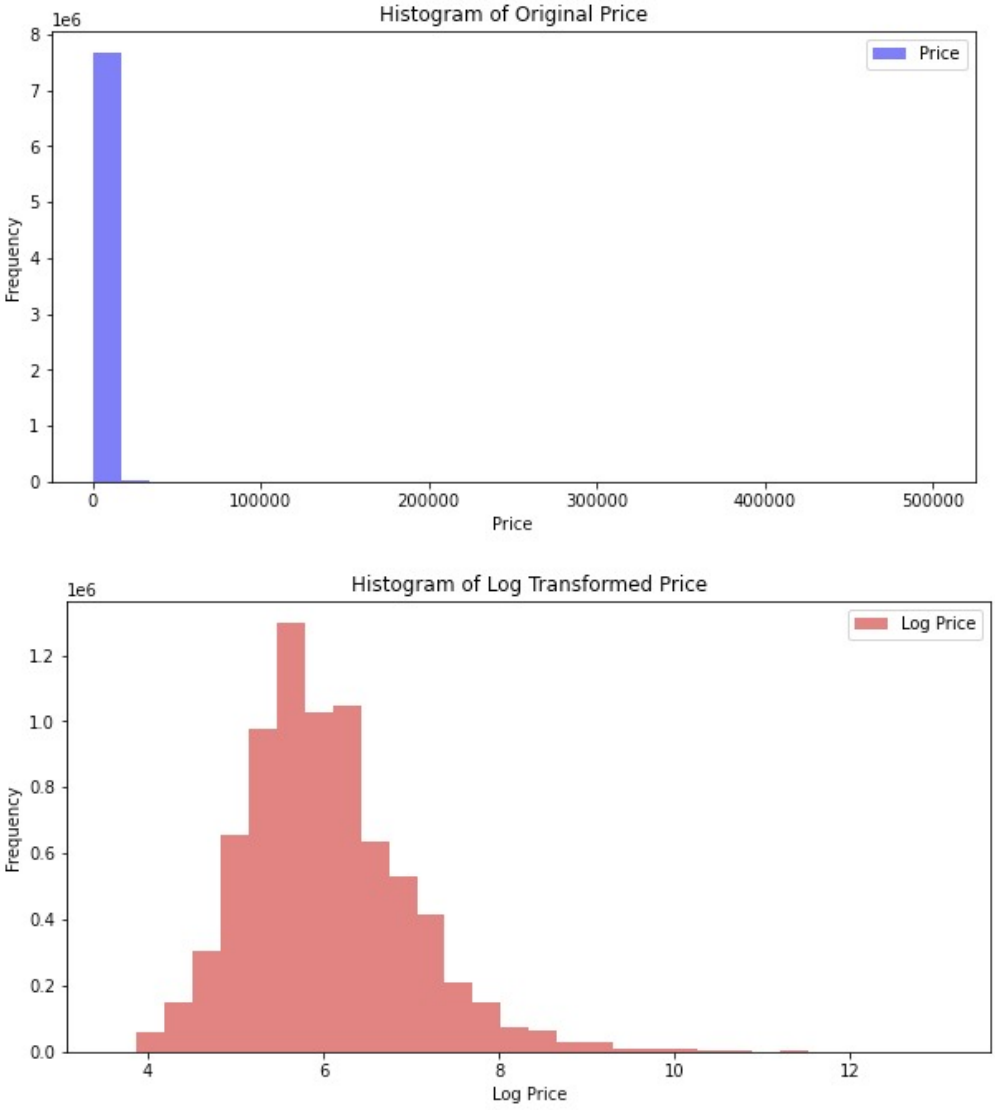
- Se dividió el *dataset* en conjuntos de entrenamiento, el cual está formado por el 80% de los registros y 39 variables (se excluyó la variable objetivo) y el conjunto de prueba, formado por el 20% restante de los registros y su respectiva variable objetivo (‘log_price’). Una vez separados los datos en estos conjuntos se realizó un tratamiento de *outliers* en los precios .

- Se calculó el logaritmo de la variable ‘price’ con el fin de reducir la varianza

de la variable y también para reducir el impacto relativo de los datos atípicos.

- La variable ‘price’ originalmente tenía una distribución que a priori parecía tener una frecuencia extremadamente alta en los valores bajos de precio con una larga cola hacia valores más altos, lo que sugiere una distribución con valores atípicos significativos en el extremo superior, lo que dificulta el análisis.

  Para justificar la decisión se adjuntan gráficos explicativos donde se observa el



comportamiento de ambas variables :

Figura 5.1.2.1: Histograma de las variables price y log_price.

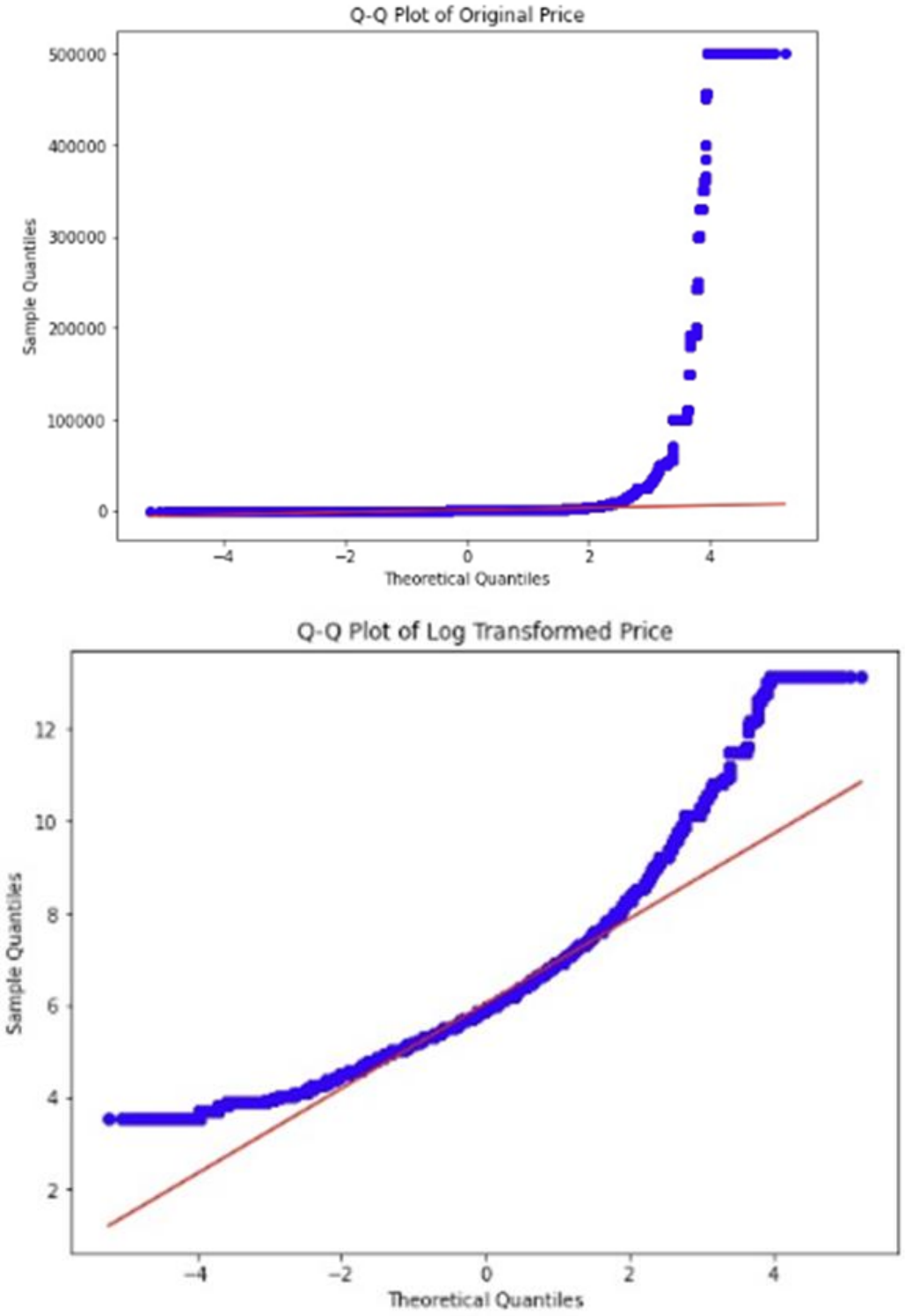


Figura 5.1.2.2: Gráfico Qqplot de las variables price y log_price

En la primera imagen se puede observar el histograma del precio original y del precio transformado con logaritmo.

La segunda imagen contiene gráficos Q-Q (cuantil-cuantil), que son útiles para evaluar si una distribución se ajusta a una distribución normal. La transformación logarítmica suele hacer que la relación entre la variable objetivo y los predictores sea más lineal, lo que puede mejorar la calidad y la interpretación de los resultados del modelo. En este gráfico se observa que la distribución del precio logarítmicamente transformado sigue más de cerca la línea roja, que representa una distribución teórica normal, en comparación con la distribución del precio original. Esto indica que la transformación logarítmica hace que la distribución del precio sea más normal, justificando así su uso para el modelado en *machine learning*.

- Se utilizó KNN para imputar valores faltantes en la columna ‘host_is_superhost_encoded’. Esto se realizó tanto en el conjunto de entrenamiento como en el de prueba. La elección de este método se basa en que el mismo no asume linealidad , lo que lo hace flexible y aplicable a una amplia variedad de situaciones de datos, incluidas aquellas donde las relaciones entre variables son complejas o no lineales.

- Se llevó a cabo un análisis sobre los valores faltantes en las variables ‘response_time’, ‘response_rate’, ‘acceptance_rate’, ‘review_scores_rating’ y ‘reviews_score_value’ en relación con el estado de *superhost* (‘host_is_superhost’), comparando las diferencias en tasas de respuesta, tasas de aceptación y calificaciones de revisión entre *superhosts* y *hosts* regulares. En base a este análisis se optó por imputar los valores nulos con la media de cada variable en relación a si el alojamiento cumplía con la condición de ser *superhost* o no.

  <u>Análisis de Tiempo de Respuesta de los Anfitriones</u>

- Se identificaron cuatro categorías principales para el tiempo de respuesta de los anfitriones en dflistings: 'within a day', 'within an hour', 'a few days or more', y 'within a few hours'.

- Se realizó un análisis visual usando *boxplots* para comparar la tasa de respuesta de los anfitriones en relación con su tiempo de respuesta. Los resultados indican que los anfitriones que responden 'within an hour' tienden a tener tasas de respuesta más altas, con una mediana cercana al 100%.

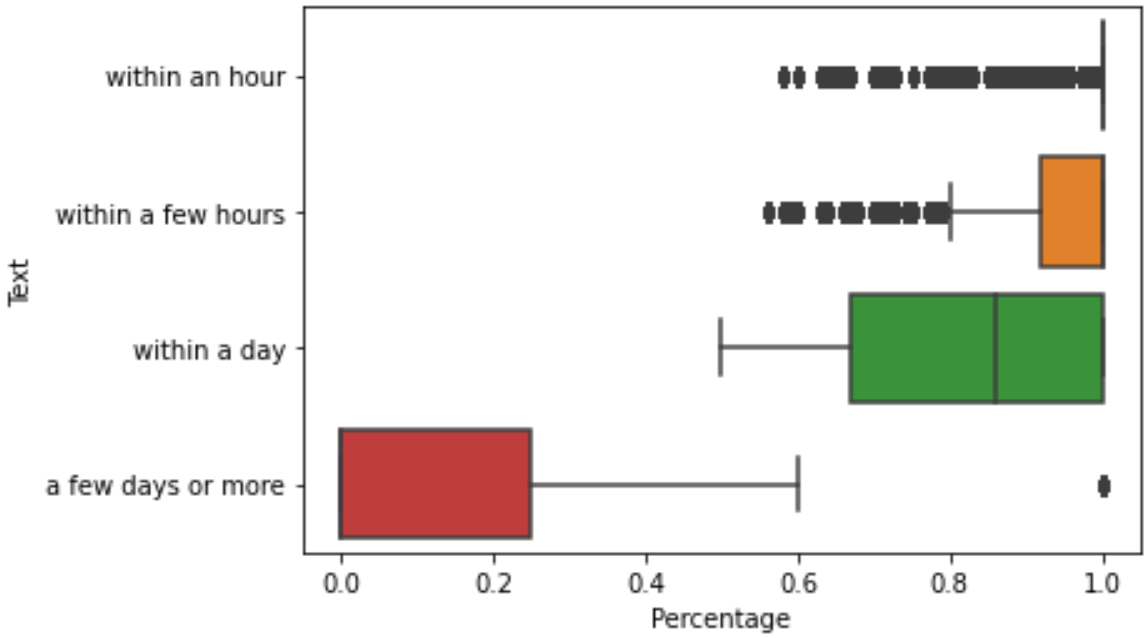


Figura 5.1.2.3: Gráfico de tiempos de respuesta

Análisis de Tasas de Respuesta y Aceptación según el status de *Superhost*

Comparación de *Superhosts* y No *Superhosts* (Entrenamiento y Pruebas):

- Se observó una diferencia notable entre Superhosts y No Superhosts en términos de tasas de respuesta y aceptación.

- Los *boxplots* y las gráficas de densidad muestran que los *Superhosts*, tanto en los conjuntos de entrenamiento como de prueba, tienden a tener tasas de respuesta y aceptación más altas y consistentes en comparación con los No *Superhosts*.

- Los *Superhosts* presentan una menor variabilidad en sus puntuaciones de revisión y tasas de aceptación, con una concentración significativa de puntuaciones altas.

- En contraste, los No *Superhosts* muestran una mayor variabilidad en estas métricas.

A continuación se presentan las gráficas mencionadas anteriormente :

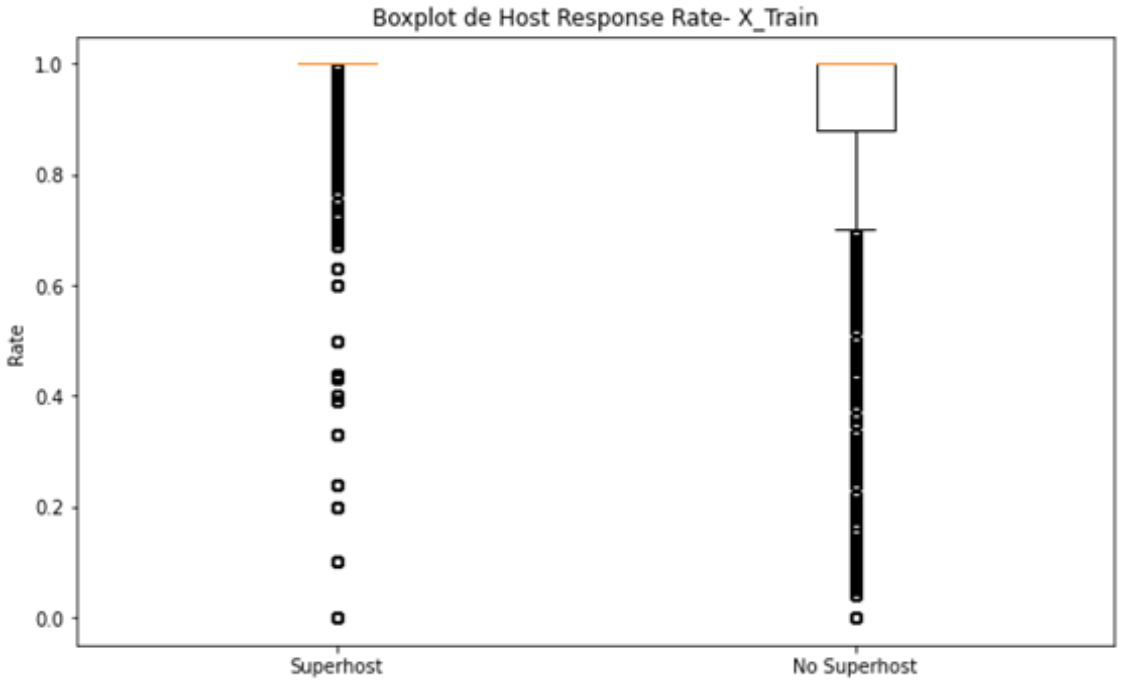


Figura 5.1.2.4: Gráfico de Boxplot de Host Response Rate- X_Train

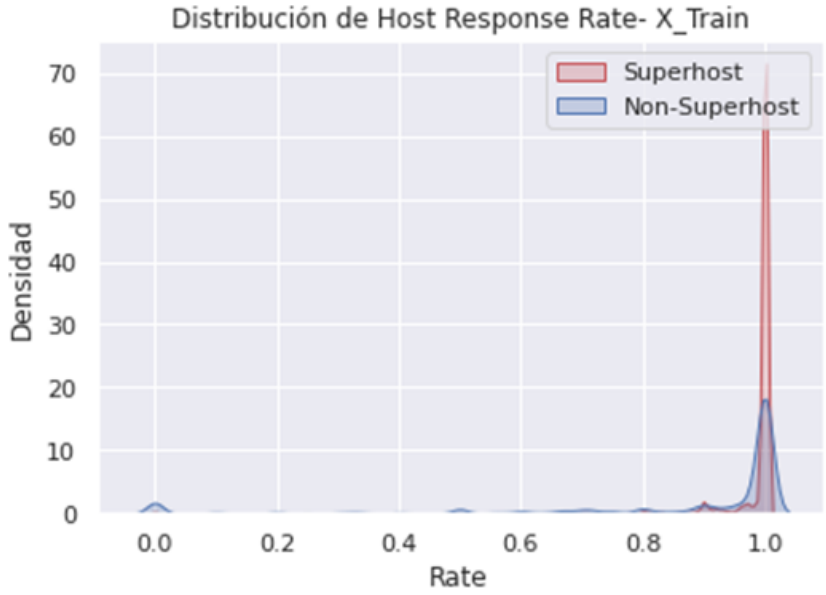


Figura 5.1.2.5: Gráfico de distribución de Host Response Rate- X_Train

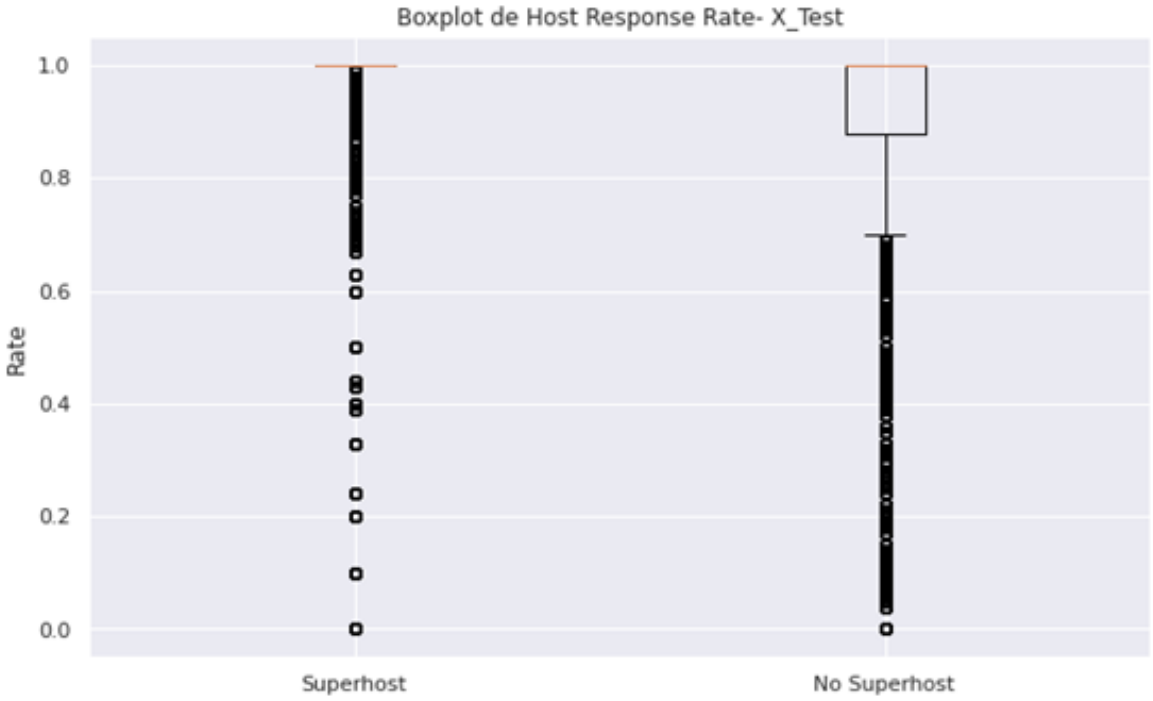


Figura 5.1.2.6: Gráfico Boxpot de Host Response Rate- X_Test

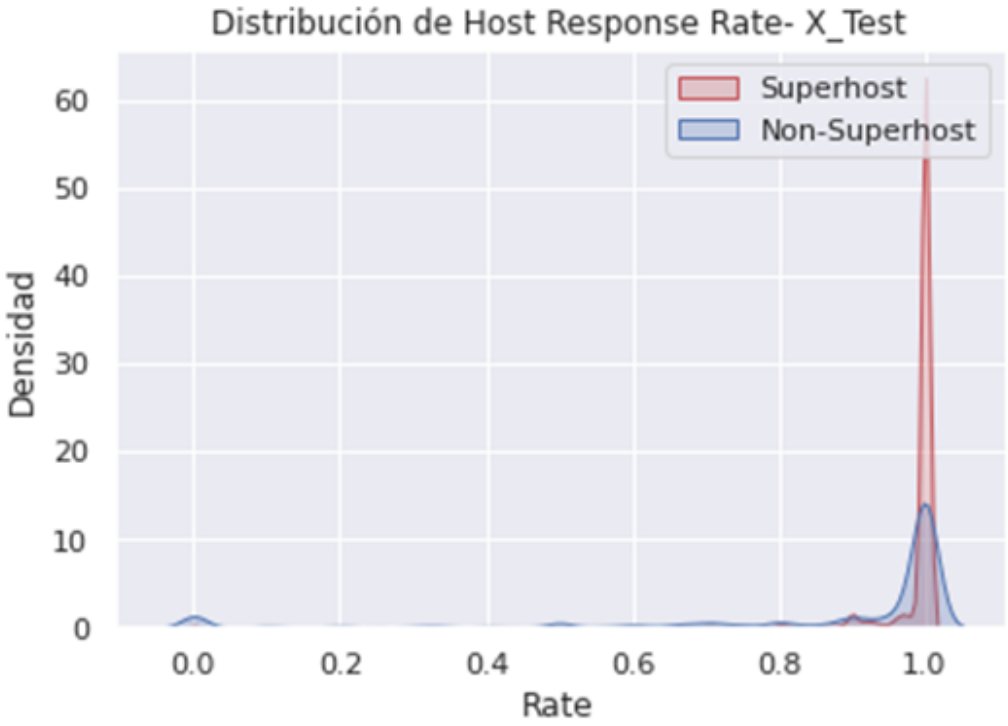


Figura 5.1.2.7: Gráfico de distribución de Host Response Rate - X_Test

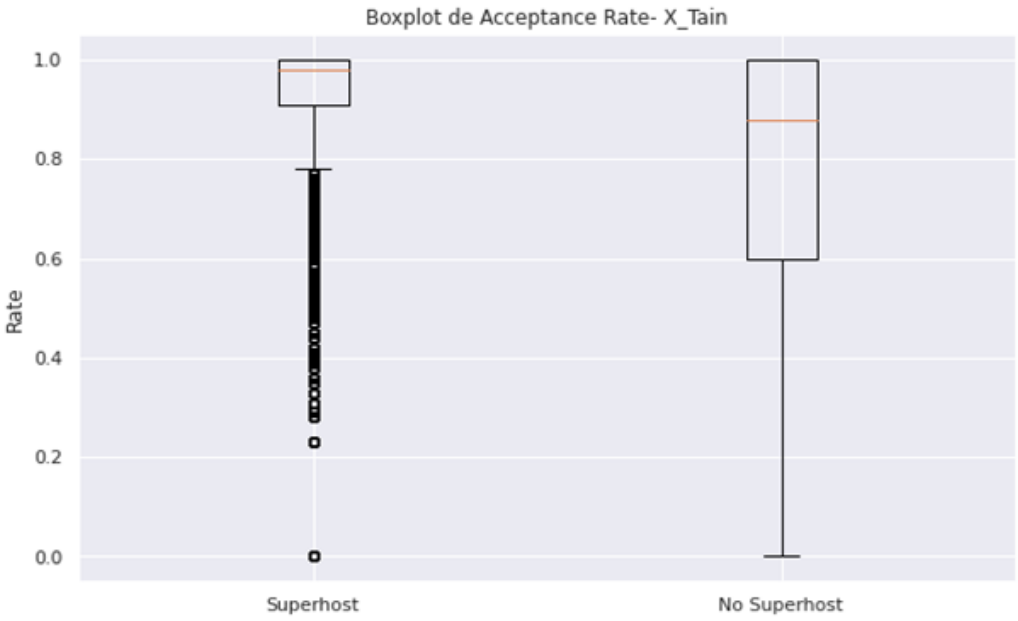


Figura 5.1.2.8: Gráfico Boxplot de Acceptance Rate - X_Train

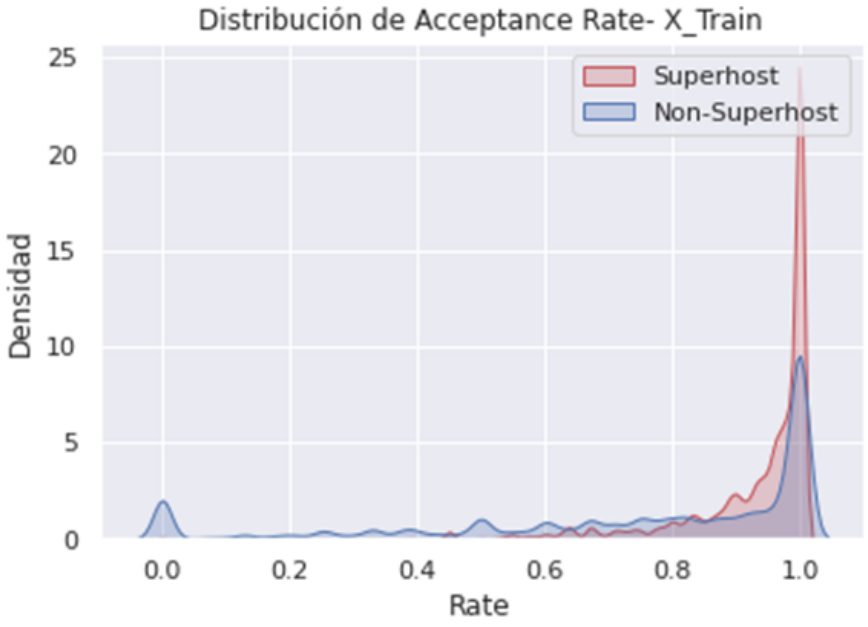


Figura 5.1.2.9: Gráfico de distribución de Acceptance Rate - X_Train

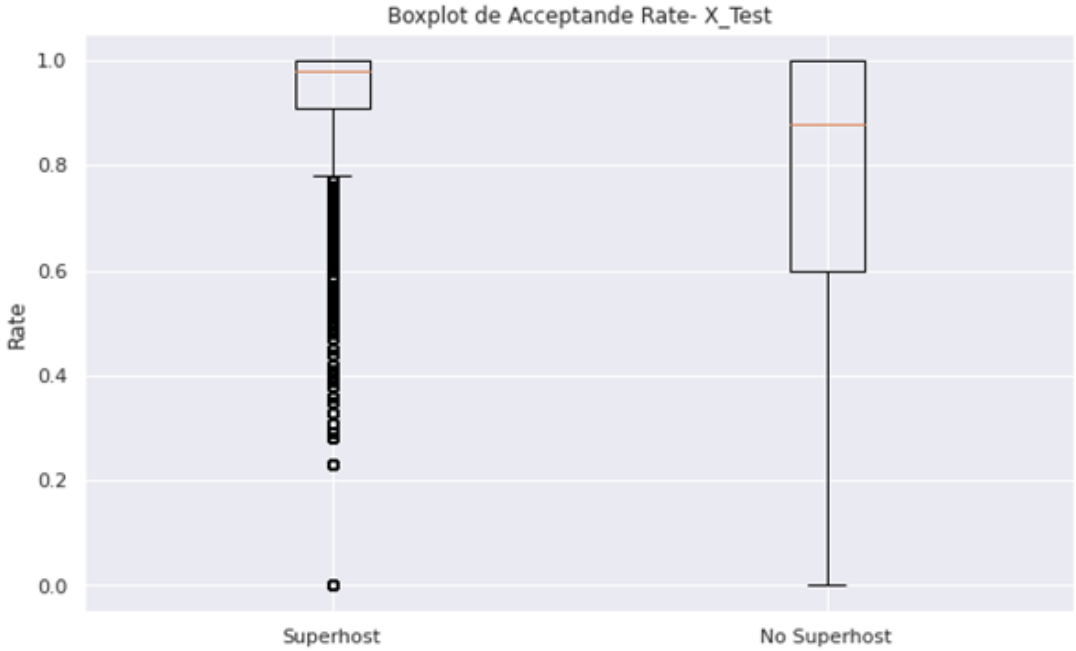


Figura 5.1.2.10: Gráfico Boxplot de Acceptance Rate - X_Test

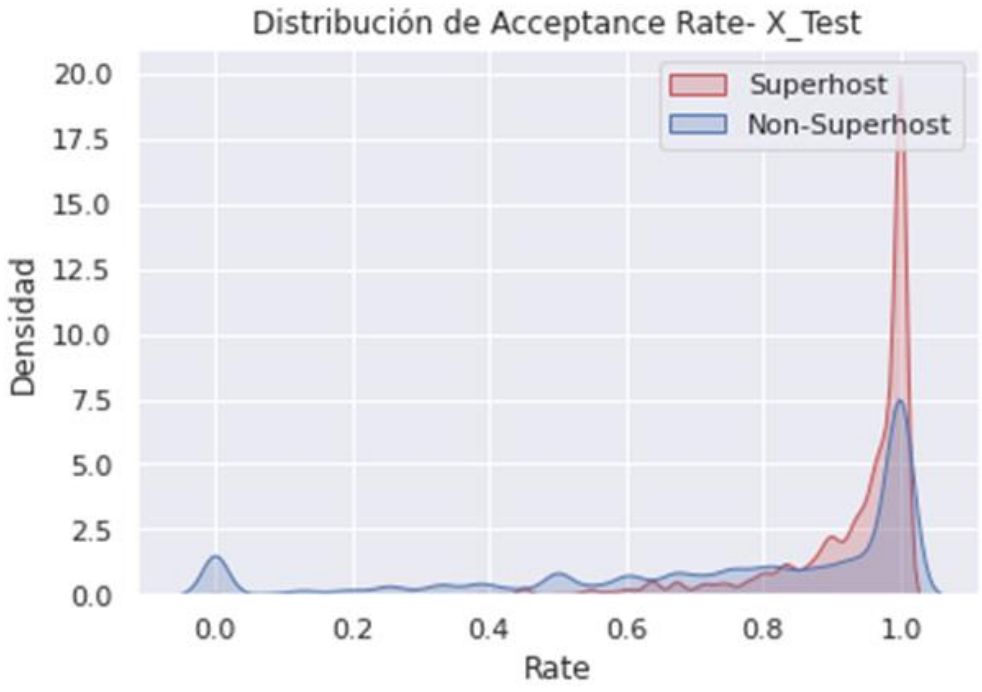


Figura 5.1.2.11: Gráfico de distribución de Acceptance Rate - X_Test

Análisis de Puntuaciones de Reseñas

Este segmento del análisis se centró en las puntuaciones de reseñas (*Review Scores Rating* y *Review Scores Value*) para los anfitriones de Airbnb, diferenciando entre *Superhosts* y No *Superhosts*.

Comparación de Puntuaciones de Reseñas entre *Superhosts* y No *Superhosts*:

- Se realizaron comparaciones utilizando *boxplots* y gráficas de densidad para ambos conjuntos de datos (entrenamiento y prueba).
- Las puntuaciones de reseñas de los *Superhosts* mostraron una mayor concentración y consistencia en puntuaciones altas, con una menor cantidad de valores atípicos.
- Por otro lado, los No *Superhosts* presentaron una mayor variabilidad en sus puntuaciones de reseñas, con medianas generalmente más bajas.

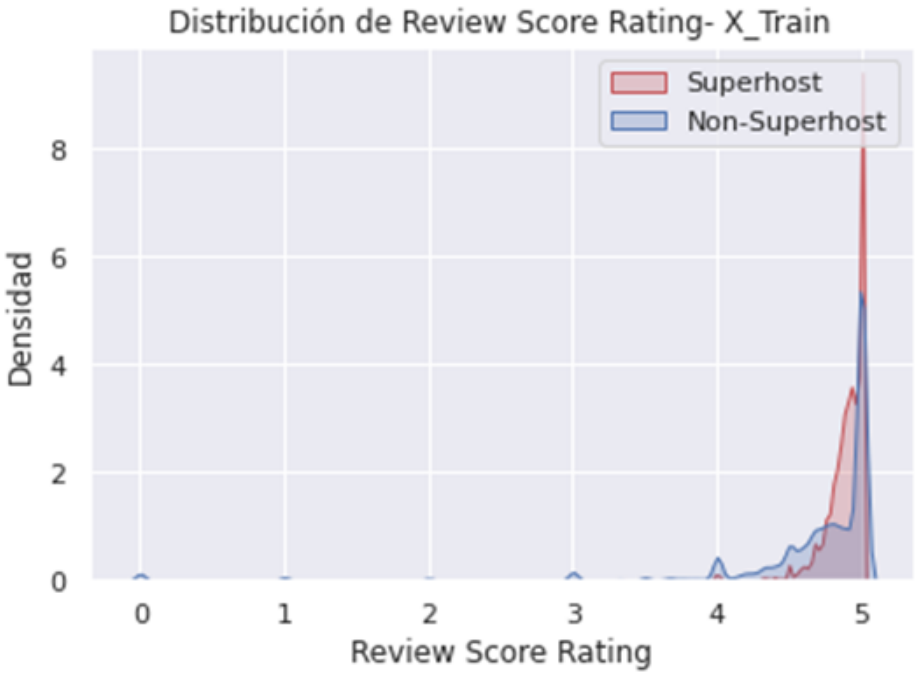
Distribución de las Puntuaciones de Reseñas:

- Las gráficas de densidad mostraron que, aunque ambos grupos tienden a recibir reseñas positivas, los *Superhosts* tienen una distribución más estrecha centrada en puntuaciones altas, mientras que los No *Superhosts* muestran una mayor dispersión.

Imputación de datos faltantes:

- Se realizó una imputación de datos faltantes en las variables de puntuaciones de reseñas y tasas de respuesta/aceptación. Esto se hizo basándose en la media de cada grupo (*Superhosts* y No *Superhosts*) y utilizando KNN para la variable ‘bedrooms’.

A continuación se presentan las gráficas mencionadas anteriormente:



Figura

5.1.2.12: Gráfico Boxplot de Review Score Rating - X_Train

Figura 5.1.2.13: Distribución de Review Score Rating - X_Train

Figura 5.1.2.14: Gráfico Boxpolot de Review Score Rating - X_Test

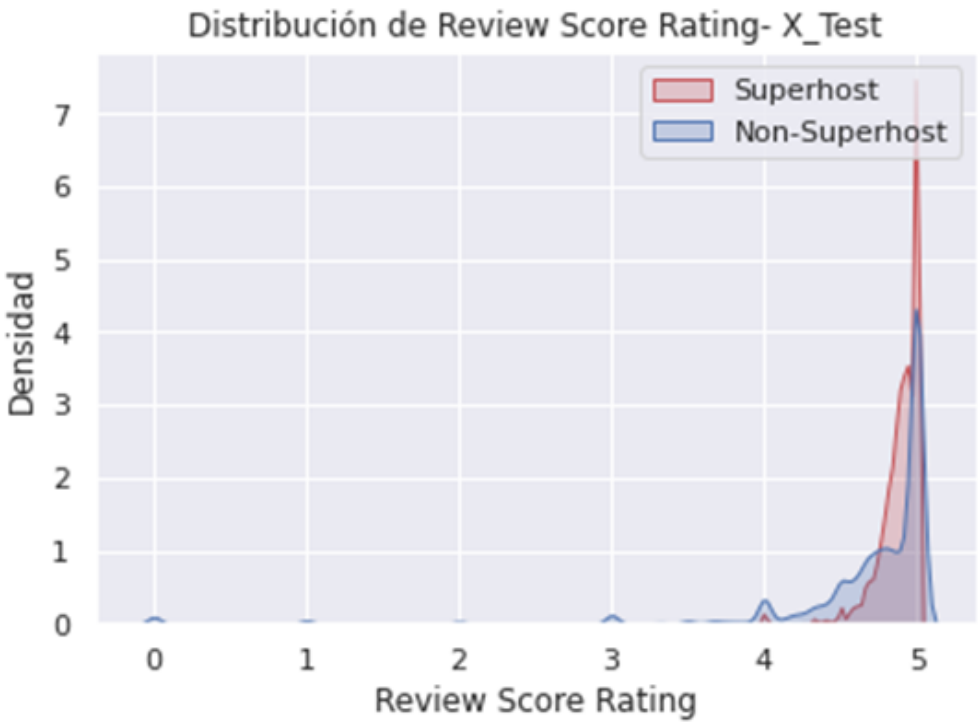


Figura 5.1.2.15: Distribución de Review Score Rating - X_Test

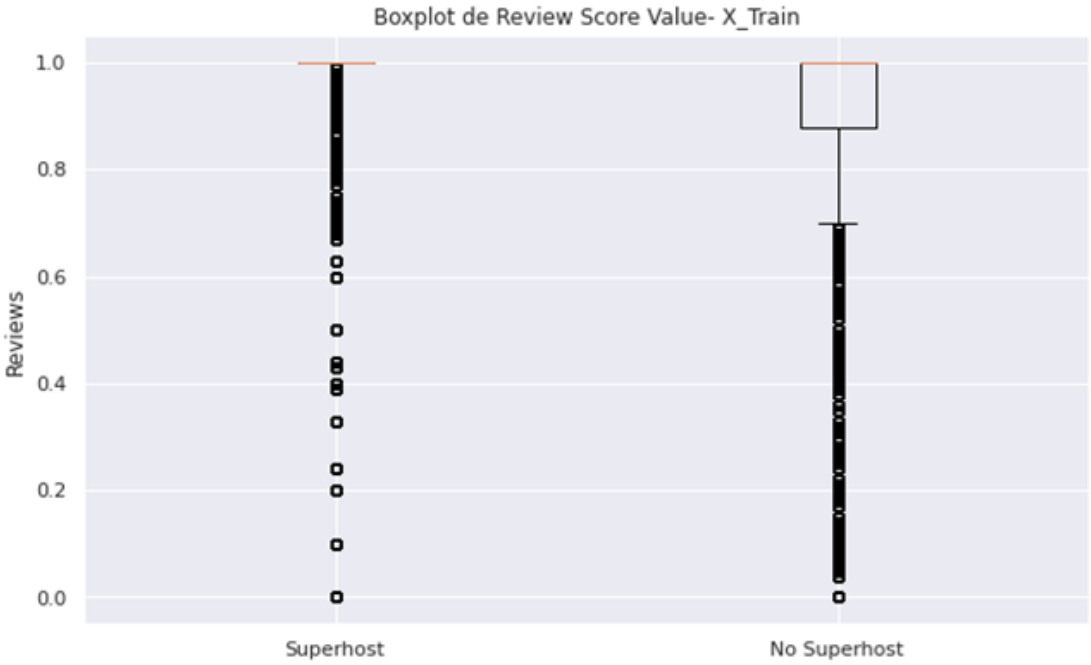


Figura 5.1.2.16: Gráfico Boxplot de Review Score Value - X_Train

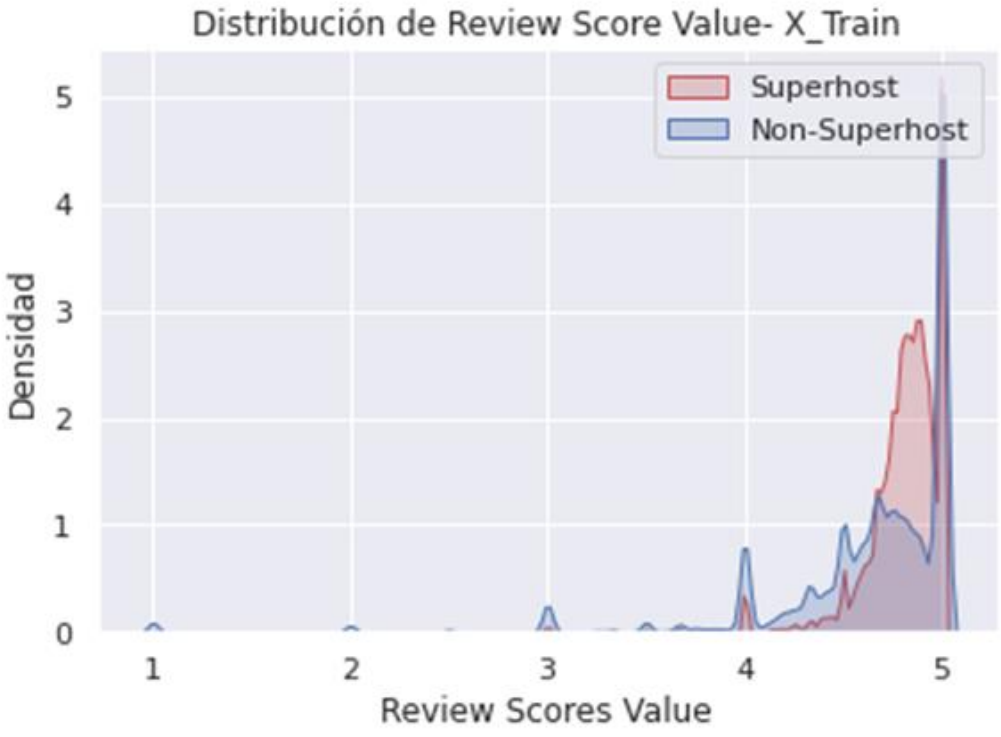


Figura 5.1.2.17: Distribución de Review Score Value - X_Train

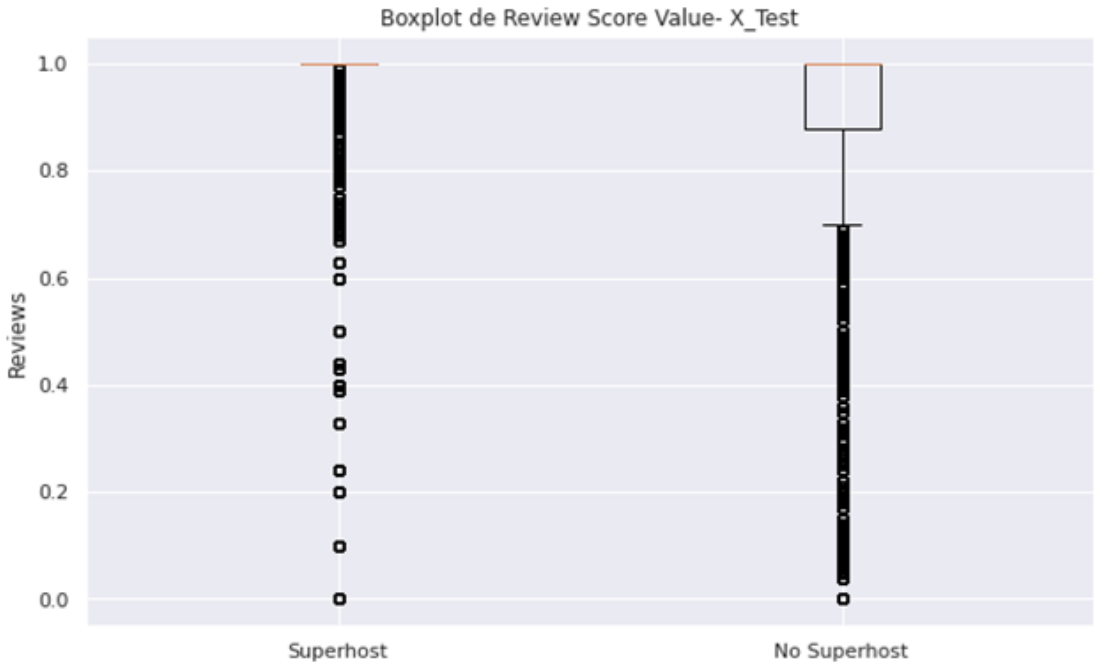


Figura 5.1.2.18: Gráfico Boxplot de Review Score Value - X_Test

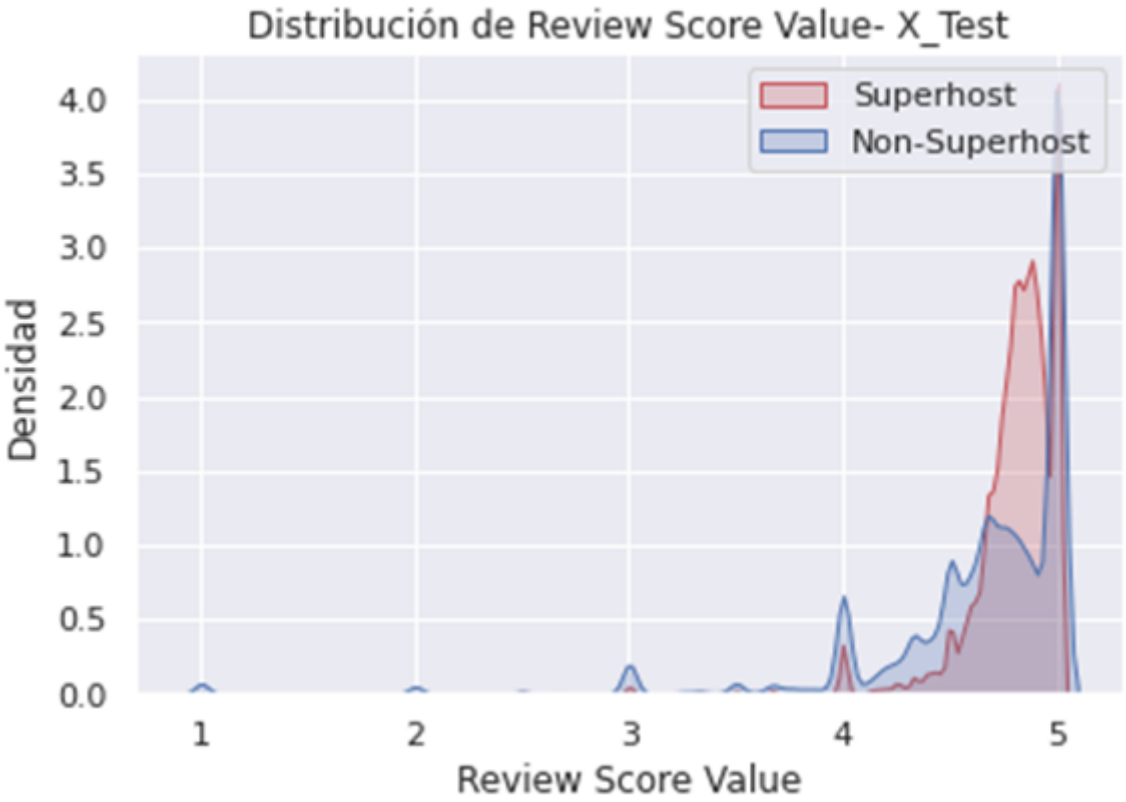


Figura 5.1.2.19: Distribución de Review Score Value - X_Test

Análisis de Reseñas Agrupadas por Mes y Año:

- Se procesaron las reseñas para obtener información sobre la primera y última reseña, el número total de reseñas y el promedio de reseñas (‘agregated_reviews’) por mes para cada anfitrión.
- Se utilizó un enfoque de agrupación por ‘listing_id’, año y mes para este análisis.

Imputación de Datos Faltantes y Conversión de Fechas:

- Se completaron los valores nulos en dflistings para ‘first_review’, ‘last_review’, y ‘reviews_per_month’ utilizando los datos correspondientes de ‘aggregated_reviews’.

- Se convirtieron las columnas ‘last_scraped’ y ‘host_since’ a formato de fecha en dflistings.

- Se creó una nueva variable antigüedad, calculando la diferencia en días entre ‘last_scraped’ y ‘host_since’ para entender la experiencia o tiempo de actividad de los anfitriones en la plataforma.

  Enriquecimiento de Características y Codificación

- Se utilizó *LabelEncoder* para transformar ‘neighbourhood_cleansed’ y ‘host_identity_verified’ en variables numéricas (‘neighbourhood_cleansed_encoded’, ‘host_identity_verified_encoded’), facilitando su uso en modelos analíticos.

- Se añadieron nuevas columnas (día, año, mes) extrayendo información detallada del día, año y mes de la columna ‘date’.

- Se eliminaron variables irrelevantes para el análisis con el fin de seguir reduciendo el *dataset*.

6. Exploración de Datos

Una vez realizada la transformación y limpieza de los diferentes conjuntos de datos, se creó un nuevo *dataset* denominado “df” el cual unifica los siguientes conjuntos de datos:

- X_train
- Y_train
- X_test
- Y_test

Sobre este nuevo *dataset* se realizó la exploración de datos con el objetivo de encontrar comportamientos y conclusiones determinantes para los entrenamientos de los diferentes modelos predictivos.

Visualización de alojamientos por ubicación y precio

El mapa muestra una distribución geográfica de precios para los diferentes alojamientos que forman parte del conjunto de datos. Los colores de los marcadores (verde, amarillo y rojo) representan diferentes rangos de precios, donde verde indica los precios más bajos, amarillo los precios medios y rojo los precios más altos.

Los hospedajes verdes parecen ser los más abundantes, lo que indica que la mayoría de las propiedades tienen precios en el tercio inferior del rango de precios. Los marcadores amarillos y rojos, que representan rangos de precios medios y altos respectivamente, están dispersos pero parecen estar más concentrados en ciertas áreas. Esto podría indicar barrios que son más caros dentro de la ciudad.

El mapa muestra una mezcla de todos los colores, principalmente aquellos correspondientes al rango inferior y medio de precios de alojamientos, lo que indica una variabilidad en los precios de las propiedades.

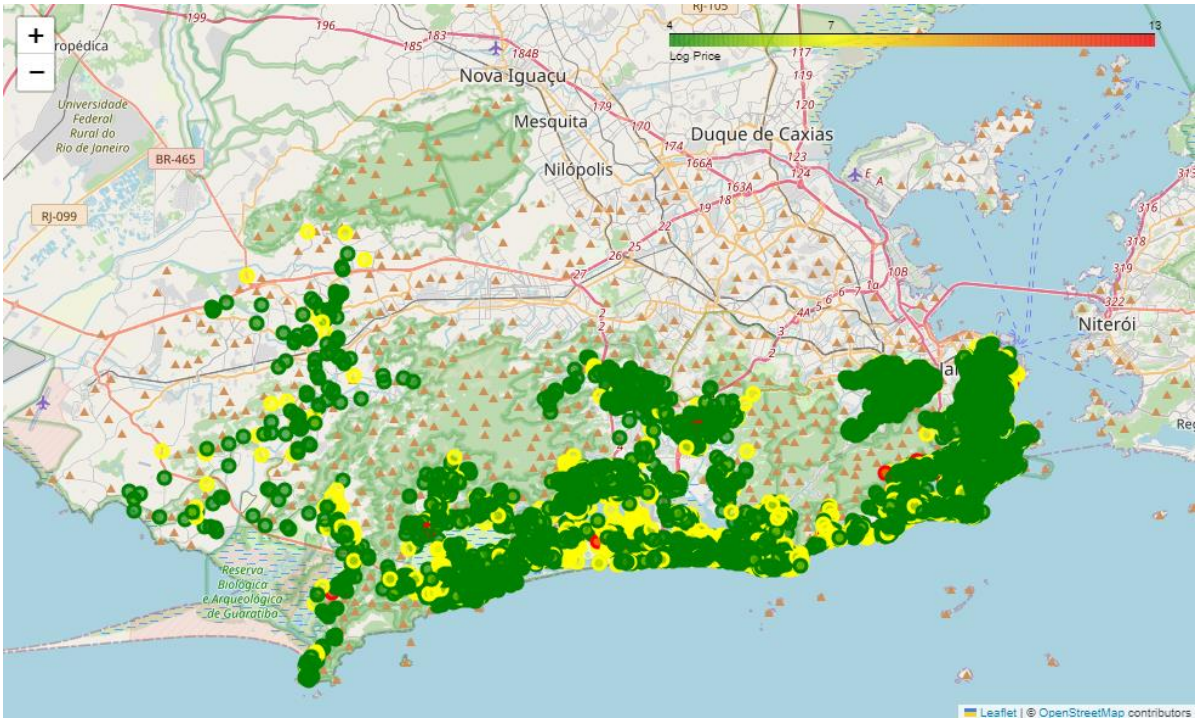


Figura 6.1: Mapa de calor basado en precios

Análisis de Distribución de Precios

Se revisa la distribución de precios con la exclusión de valores atípicos en un *boxplot*. La eliminación de estos valores proporciona una vista más concentrada de la distribución central de precios. En esta visualización ajustada, se observa que no hay puntos fuera de los 'bigotes', lo que confirma la efectiva eliminación de valores atípicos del análisis.

Esta depuración de datos permite un análisis más preciso de la estructura de precios típica. La ausencia de valores atípicos sugiere una uniformidad en la distribución de precios entre los alojamientos, con una concentración más estrecha alrededor de la mediana.

El *boxplot* actual refleja una distribución más homogénea de los precios, con los extremos superior e inferior representando los cuartiles del 75% y 25%, respectivamente. Este enfoque proporciona una comprensión más clara de la gama de precios en la que se sitúa la mayoría de los alojamientos, lo que resulta esencial para establecer estrategias de precios o para la toma de decisiones en el mercado.

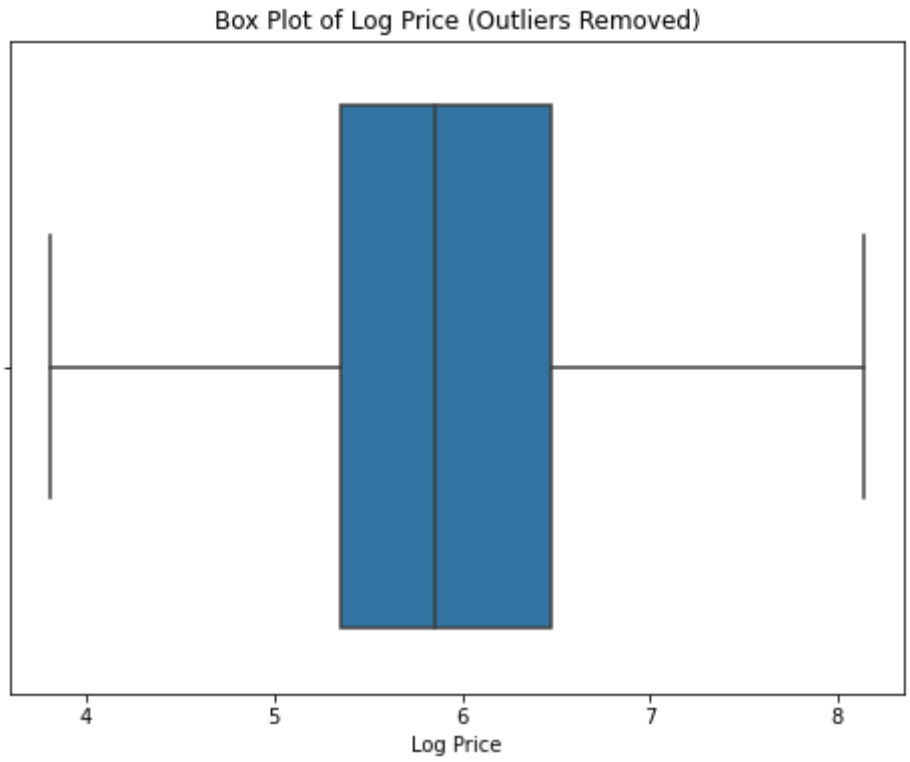


Figura 6.2: Gráfico Boxplot de Log Price

Relación entre Precio y Número de Dormitorios

Parece observarse una tendencia general de que a mayor número de dormitorios, mayor es el "Log Price". Sin embargo, esta tendencia no es uniforme y presenta una variabilidad considerable, especialmente para alojamientos con pocos dormitorios.

La mayoría de los datos se concentran en alojamientos con menos de 10 dormitorios. A medida que aumenta el número de dormitorios, la cantidad de datos disponibles disminuye.

Hay una amplia gama de "Log Price" para alojamientos con un número similar de dormitorios. Esto sugiere que hay otros factores, además del número de dormitorios, que afectan el precio de un alojamiento en Airbnb.

El gráfico muestra la relación entre el número de dormitorios y el precio de los alojamientos. Se observa que, en general, a medida que aumenta el número de dormitorios también lo hace la mediana. Hay una variabilidad notable en los precios para alojamientos con un gran número de dormitorios, lo cual se ve reflejado en los rangos intercuartílicos más amplios y la presencia de valores atípicos, especialmente en alojamientos con más de 10 dormitorios. Esto sugiere que factores adicionales podrían estar influyendo en el precio en estas categorías, aparte del número de dormitorios.

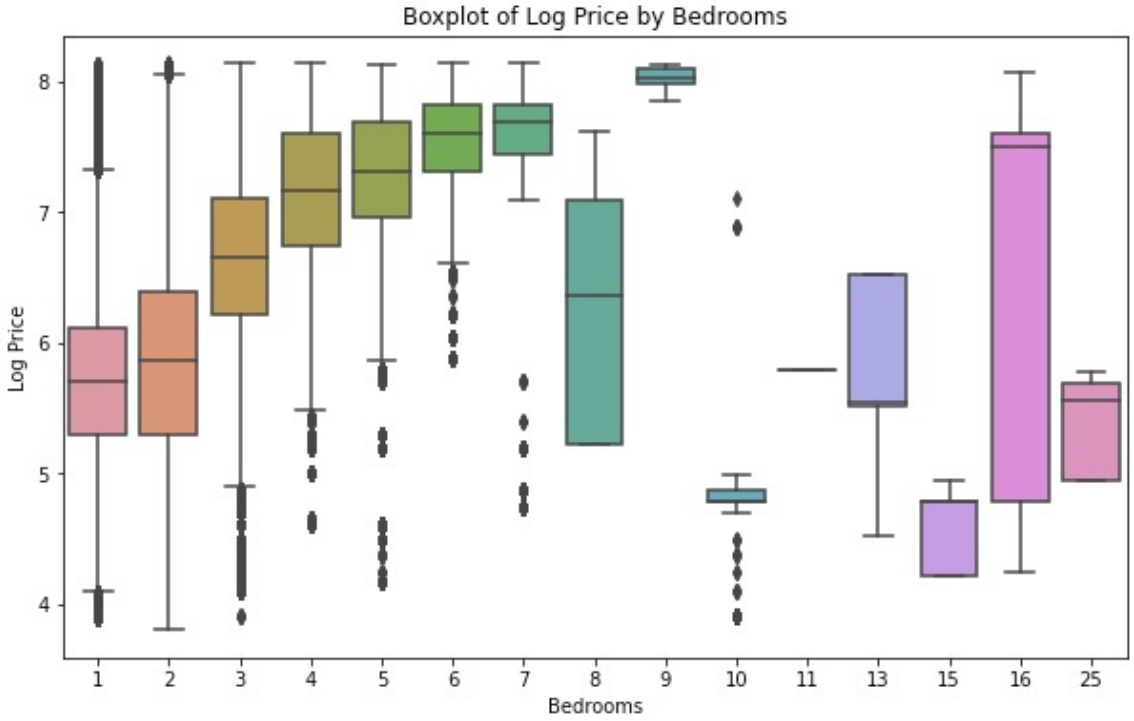


Figura 6.3: Boxplot de Log Price por cantidad de dormitorios

Correlación entre Precio y Calificación de Reseñas.

No se visualiza una relación clara y lineal entre el logaritmo del precio y las puntuaciones de las reseñas. Los precios son variados a lo largo de todas las puntuaciones de revisión, aunque hay una concentración ligeramente mayor de precios más altos en el rango de calificación más elevado.

Hay relativamente pocos puntos con calificaciones bajas (menores a 4), lo que podría sugerir que los alojamientos con calificaciones bajas son menos comunes, o que los alojamientos con malas reseñas son retirados o reciben menos reservas, y por tanto, tienen menos presencia en la muestra.

Existe una gran cantidad de alojamientos con altas calificaciones de revisión y una amplia gama de precios en la escala logarítmica. Esto podría indicar que, aunque un alojamiento tenga una buena calificación, puede variar significativamente en precio debido a otros factores como por ejemplo la ubicación, dormitorios y/o *amenities*.

En términos generales el gráfico sugiere que no hay una correlación directa y fuerte entre el logaritmo del precio y las puntuaciones de las revisiones de los alojamientos de Airbnb, pero sí hay una tendencia a tener una amplia gama de precios en alojamientos con altas puntuaciones de revisión. Se puede decir que las calificaciones se comportan prácticamente como una variable discreta para valores menores a 3, y continua a partir de 3.

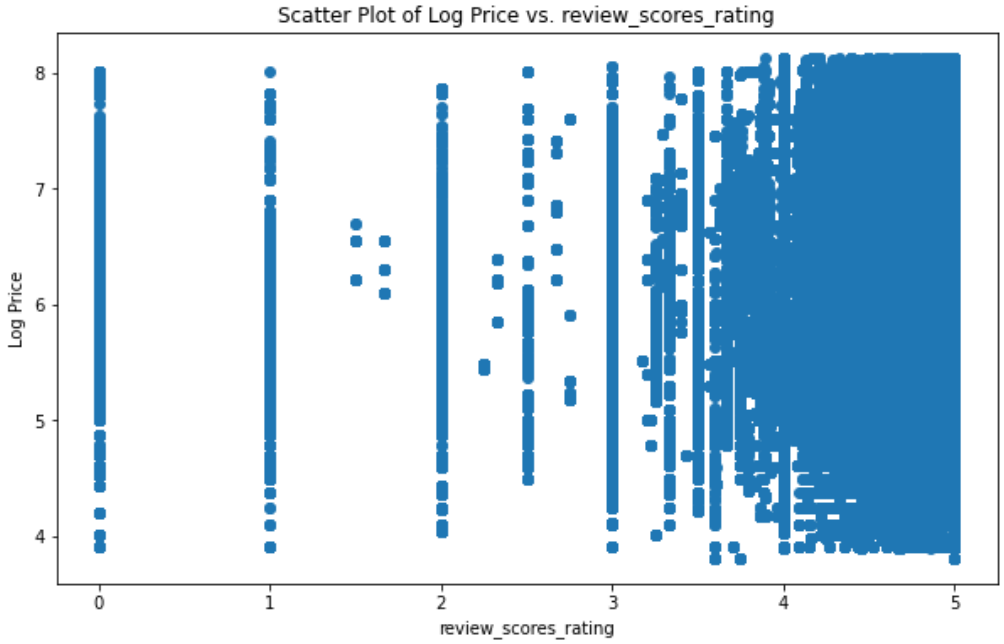


Figura 6.4: Gráfico de dispersión de Log Price vs Review scores rating

Adicionalmente se agrega un gráfico de boxplot para verificar lo antes mencionado.

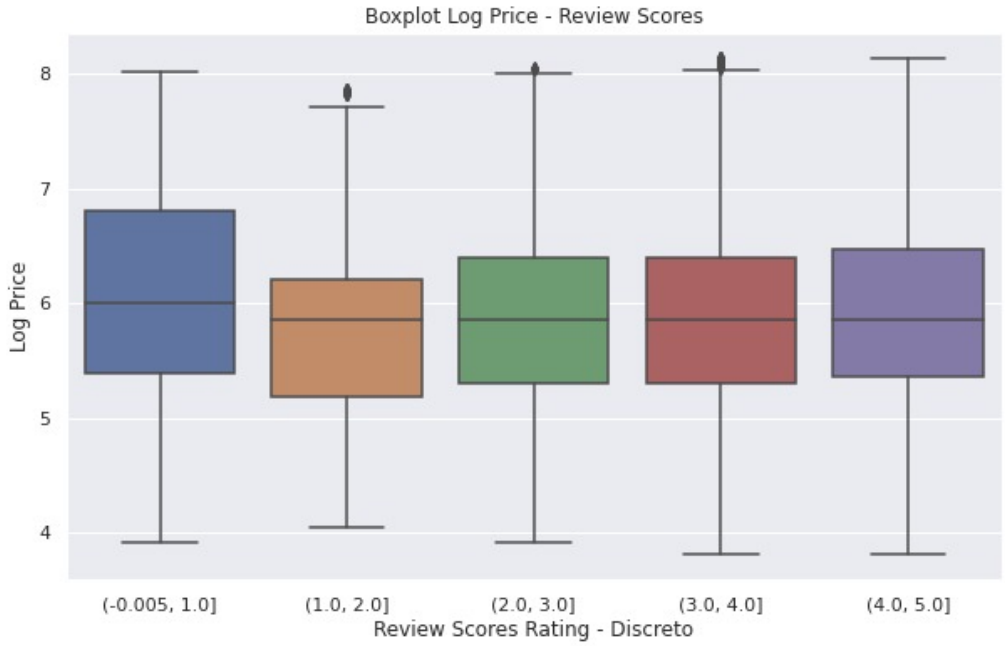


Figura 6.5: Gráfico Boxplot de Log Price - Review Scores Rating

Distribución de Precios de Alojamientos

La distribución muestra que hay varios alojamientos con precios similares, pero no un notorio valor predominante. La distribución no es simétrica y muestra una breve asimetría positiva (hacia la derecha), comportamiento típico en los precios de bienes y servicios. Este comportamiento se logra a partir de la transformación a logaritmo.

La mayoría de los alojamientos se concentran en el rango medio de precios, con menos alojamientos en los extremos más bajos y más altos de precios.

Hay menos alojamientos en el extremo superior del rango de precios, lo cual es esperable ya que generalmente hay menos propiedades de lujo o de alto valor en el mercado.

La distribución muestra que hay una dispersión significativa en los precios de los alojamientos. Esta variabilidad podría deberse a las diferentes características de los alojamientos (tamaño, ubicación, etc).

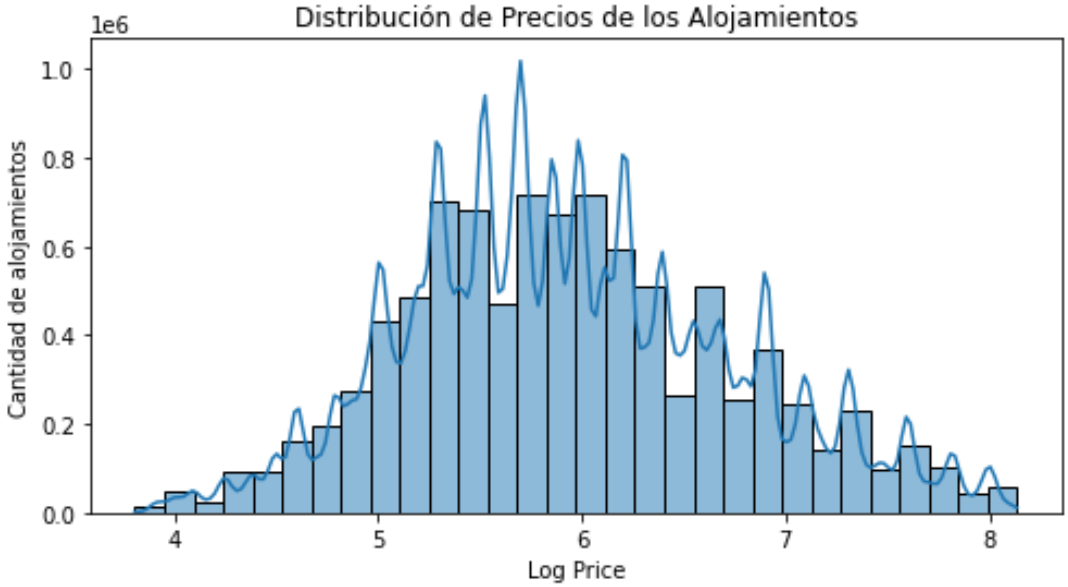


Figura 6.6: Distribución de precios de alojamientos

Distribución de Alojamientos por Tipo de Habitación

El gráfico de barras proporciona una comparación cuantitativa de los alojamientos según su clasificación por tipo de habitación. Se destaca claramente que el tipo 'Entire home/apt' (casa o apartamento completo) es la categoría más prevalente, lo que sugiere una fuerte preferencia o disponibilidad de alojamientos que ofrecen el espacio entero para los huéspedes. En contraste, el 'Private room' (habitación privada) representa una cantidad

significativamente menor, seguido por 'Shared room' (habitación compartida) y 'Hotel room' (habitación de hotel), que muestran una presencia mucho más modesta en el mercado.

Esta distribución puede reflejar las preferencias de los consumidores por privacidad y espacio.

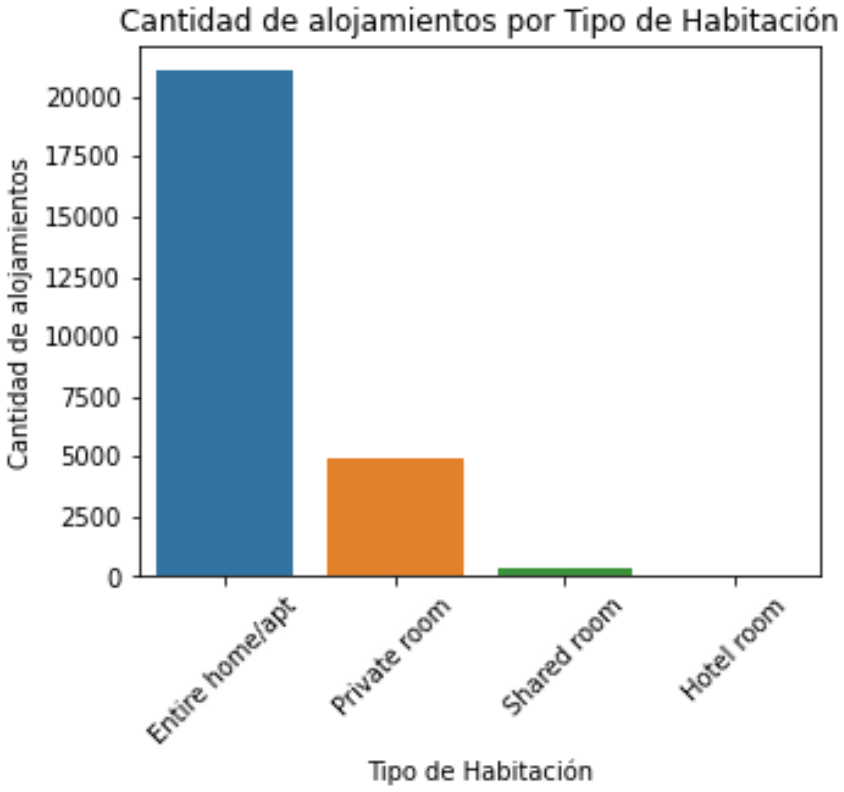


Figura 6.7: Cantidad de alojamientos por Tipo de Habitación

Concentración de Alojamientos en los Principales Barrios

Este gráfico de barras destaca los 20 barrios con la mayor concentración de alojamientos disponibles. Copacabana se distingue como el barrio con la mayor cantidad de alojamientos, seguido por Barra da Tijuca e Ipanema, lo que indica una preferencia notable por estos barrios, ya sea por parte de los anfitriones para ofrecer alojamientos o por los huéspedes al seleccionar sus estadías.

Los barrios que siguen, aunque con números menores, aún muestran una presencia significativa en el mercado, reflejando áreas potencialmente populares para el turismo. La

diversidad en la cantidad de alojamientos entre estos barrios puede deberse a una variedad de factores, incluyendo la demanda turística, la disponibilidad de inmuebles, y el perfil económico y cultural de cada barrio.

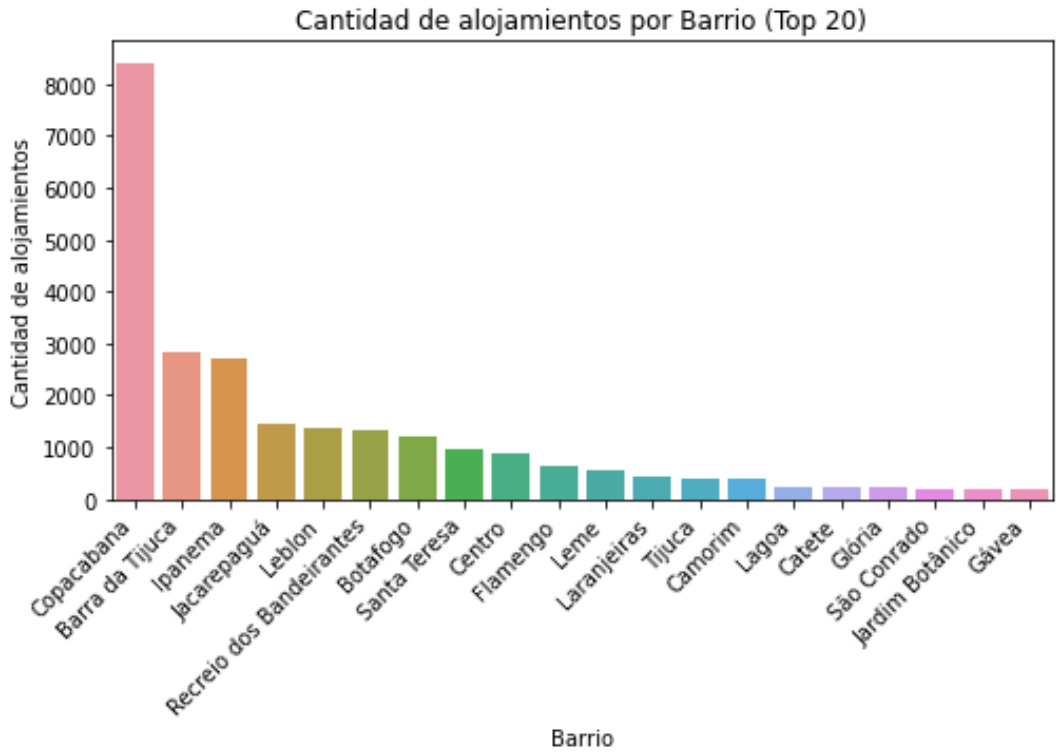


Figura 6.8: Cantidad de alojamientos por Barrio

Disponibilidad Anual de Alojamientos

El histograma muestra la cantidad de alojamientos categorizados por el número de días disponibles para alquilar durante el año. Se observa una considerable cantidad de alojamientos disponibles todo el año, reflejado por la barra alta hacia el extremo derecho del gráfico, lo que indica una amplia disponibilidad y posiblemente una preferencia por alquileres a corto plazo o propiedades dedicadas exclusivamente al alquiler vacacional.

Por otro lado, existe una cantidad significativa de alojamientos con disponibilidad muy limitada, como lo demuestra la barra alta en el extremo izquierdo. Esto podría indicar propiedades que se alquilan solo ocasionalmente, como las residencias principales durante ciertos períodos del año.

Las barras intermedias representan alojamientos con disponibilidad variable, lo que sugiere una mezcla de estrategias de alquiler por parte de los anfitriones, desde aquellos que ofrecen alquileres esporádicos hasta los que operan con una disponibilidad más regular.

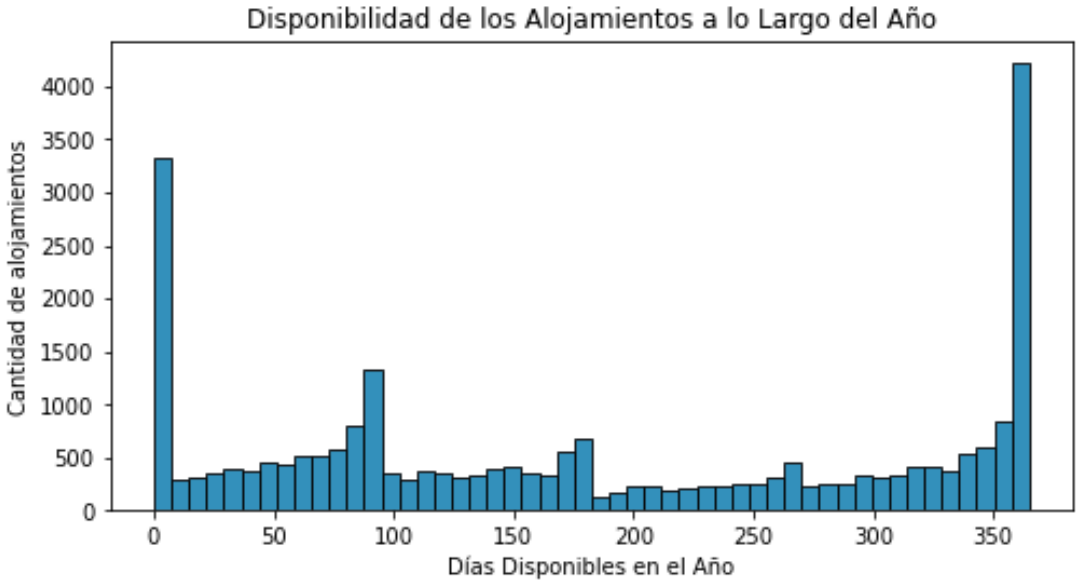


Figura 6.9: Disponibilidad de los alojamientos en el año

Análisis del Precio Medio por Tipo de Habitación

"Entire home/apt" y "Hotel room" tienen los precios promedio más altos entre los tipos de alojamiento. Este comportamiento podría deberse a que a priori se entiende que ofrecen mayor privacidad, espacio y posibles comodidades adicionales que otro tipo de alojamientos.

El precio medio para "Private room" es ligeramente más bajo que el de las casas o apartamentos completos y las habitaciones de hotel, lo que puede ser atractivo para viajeros con presupuestos más ajustados que aún desean privacidad.

Las "Shared room" tienen el precio medio más bajo, lo que es consistente con la menor privacidad y comodidades que este tipo de alojamiento suele implicar.

La similitud en los precios medios sugiere que los consumidores pueden elegir el tipo de alojamiento basado en factores distintos al precio, como la ubicación, las reseñas, las comodidades específicas y la experiencia deseada.

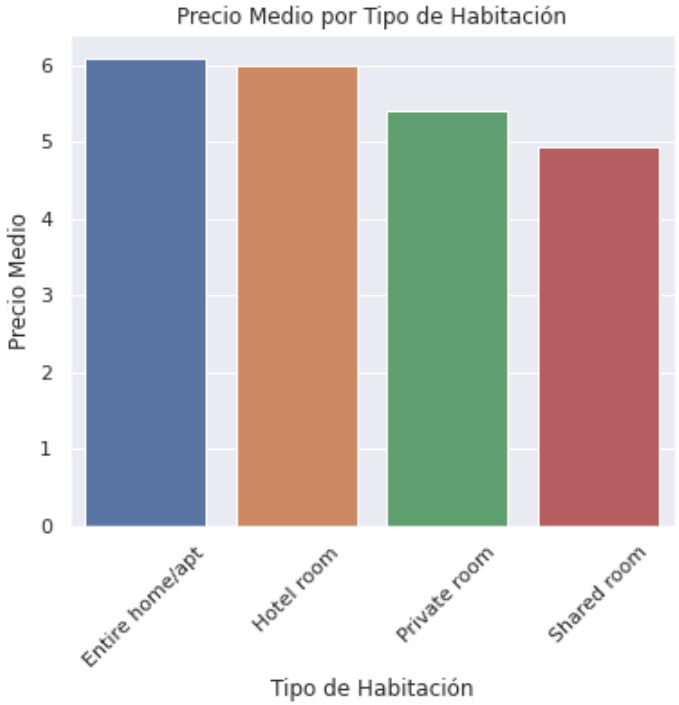
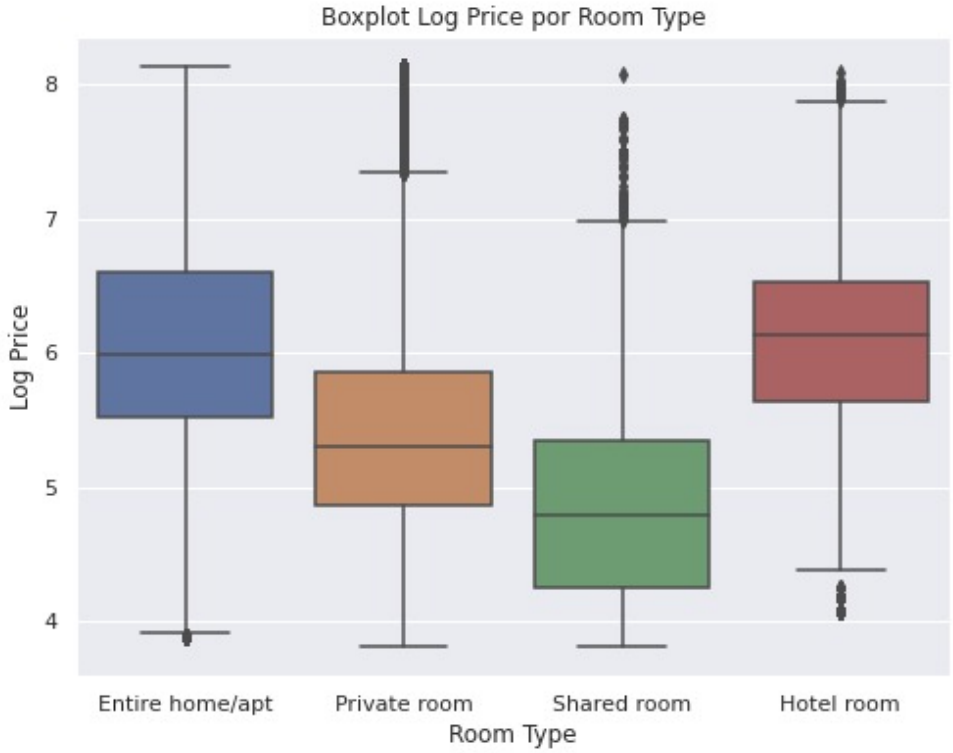


Figura 6.10: Precio medio por tipo de habitación

Figura 6.11: Gráfico Boxplot Log Price - Room Type

En el boxplot anterior se puede observar que el logaritmo del precio se compara entre diferentes tipos de habitación. Donde 'Entire home/apt' muestra una mayor mediana de precios, seguido por 'Private room' y 'Hotel room', con 'Shared room' teniendo la mediana más baja. Este patrón indica que las



propiedades que ofrecen el alojamiento completo tienden a ser más caras, mientras que las habitaciones compartidas son las más económicas. Además, se observa que los tipos 'Entire

home/apt' y 'Hotel room' tienen una variabilidad en los precios más amplia que los otros tipos de habitación, como se evidencia por la longitud de sus bigotes y la presencia de valores atípicos.

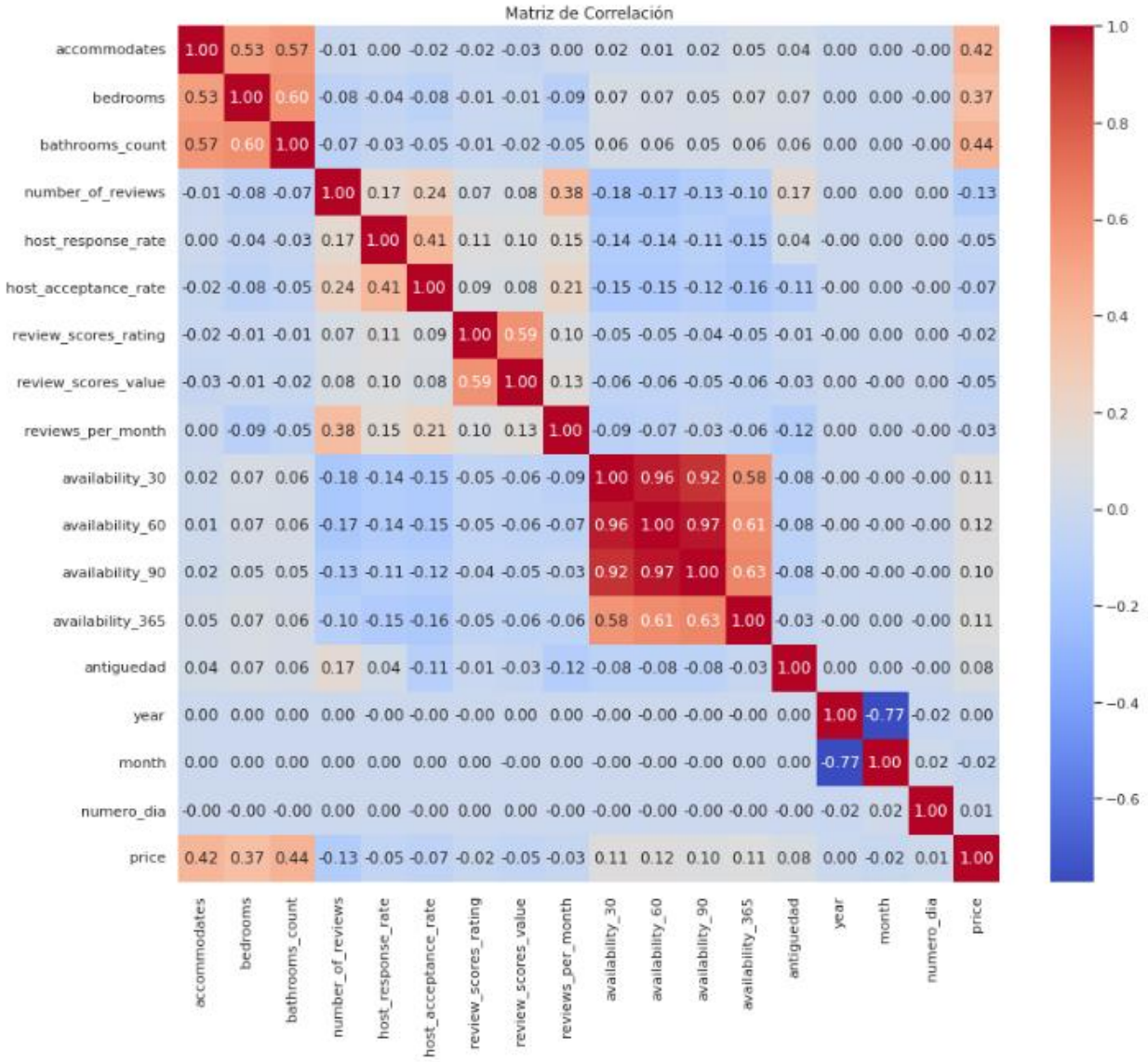
Matriz de Correlación de Características de Alojamientos

Las variables ‘accommodates’ y ‘bedrooms’ tienen una correlación positiva fuerte (0.53), lo que sugiere que a mayor número de habitaciones, mayor es la capacidad de alojamiento. ‘Accommodates’ y ‘bathrooms’ también muestran una correlación positiva similar (0.57), indicando que las propiedades con más baños tienden a hospedar a más personas.

‘Review_scores_rating’ y ‘review_scores_value’ tienen una correlación alta entre sí (0.59), comportamiento que es esperado ya que ambas son medidas de satisfacción del cliente.

‘Accommodates’ (0.42), ‘bedrooms’ (0.37) y ‘bathrooms_count’ (0.44) tienen correlaciones positivas moderadas con ‘price’. Esto indica que las propiedades con mayor capacidad o más habitaciones y baños tienden a tener precios más altos, comportamiento que también es esperado en base al conocimiento del comportamiento del mercado y también como usuarios recurrentes de alquileres temporarios de alojamientos.

Figura 6.12: Matriz de correlación



Distribución de Precios por Día de la Semana

El *boxplot* proporciona una comparación visual de la distribución de precios logarítmicos de los alojamientos por cada día de la semana. La representación gráfica no indica variaciones significativas en la mediana de los precios entre los días, lo que sugiere que el precio de los alojamientos no fluctúa considerablemente en función del día de la semana.

La similitud en el rango intercuartílico y los 'bigotes' para cada día indica que la dispersión de precios y la presencia de valores atípicos son relativamente consistentes a lo

largo de la semana. Esto puede implicar que los anfitriones mantienen una estrategia de precios uniforme, independientemente del día.

No hay diferencias significativas en las medianas de los precios a lo largo de los días de la semana. Esto sugiere que el precio promedio de los alojamientos no varía mucho de un día a otro. Además la gama de precios, indicada por la altura de las cajas, parece ser similar para todos los días. Esto implica una consistencia en la variación de precios entre los diferentes días de la semana.

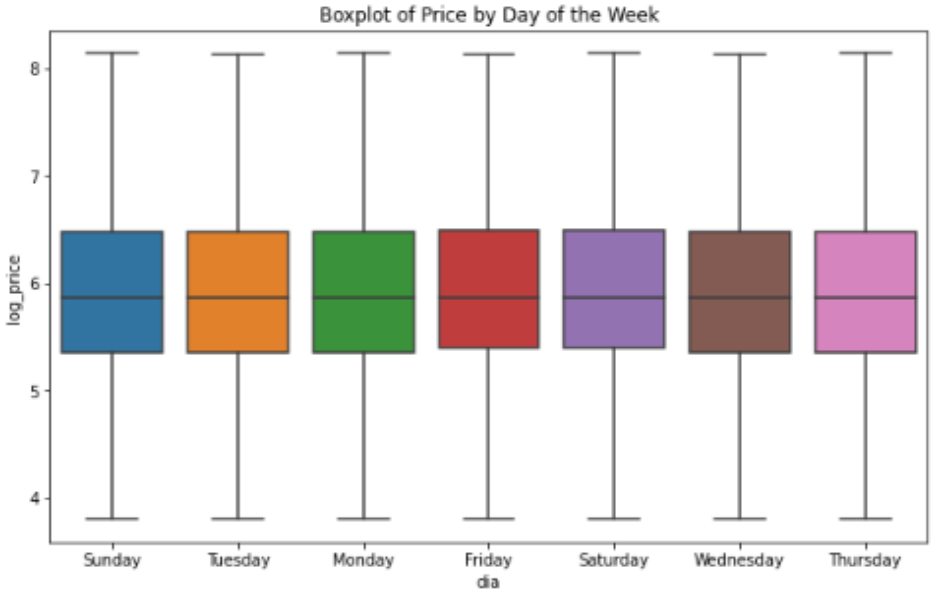


Figura 6.13: Gráfico Boxplot de Price - Day of the Week

Frecuencia de Registros de Disponibilidad por Día de la Semana

El gráfico de barras muestra la cantidad de registros de alojamientos categorizados por cada día de la semana, diferenciando entre aquellos que están disponibles ('t') y no disponibles ('f'). La distribución de los registros a lo largo de la semana no presenta variaciones significativas en cuanto a la cantidad total, lo que sugiere una consistencia en la cantidad de alojamientos listados por día.

Sin embargo, hay una ligera variación en la proporción de disponibilidad entre los días. Esto podría ser indicativo de patrones de reserva y ocupación, donde ciertos días pueden ser ligeramente más populares o tener mayor demanda que otros. La comparación entre los días muestra que la disponibilidad no fluctúa drásticamente.

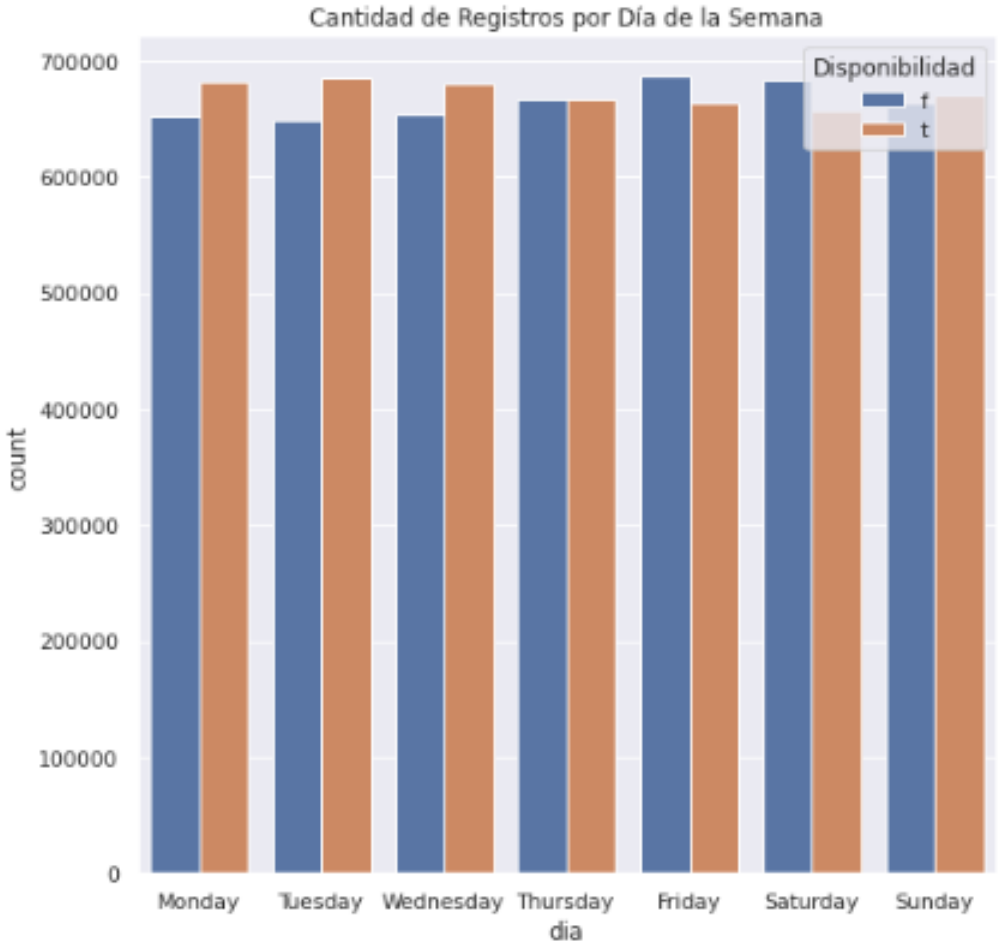
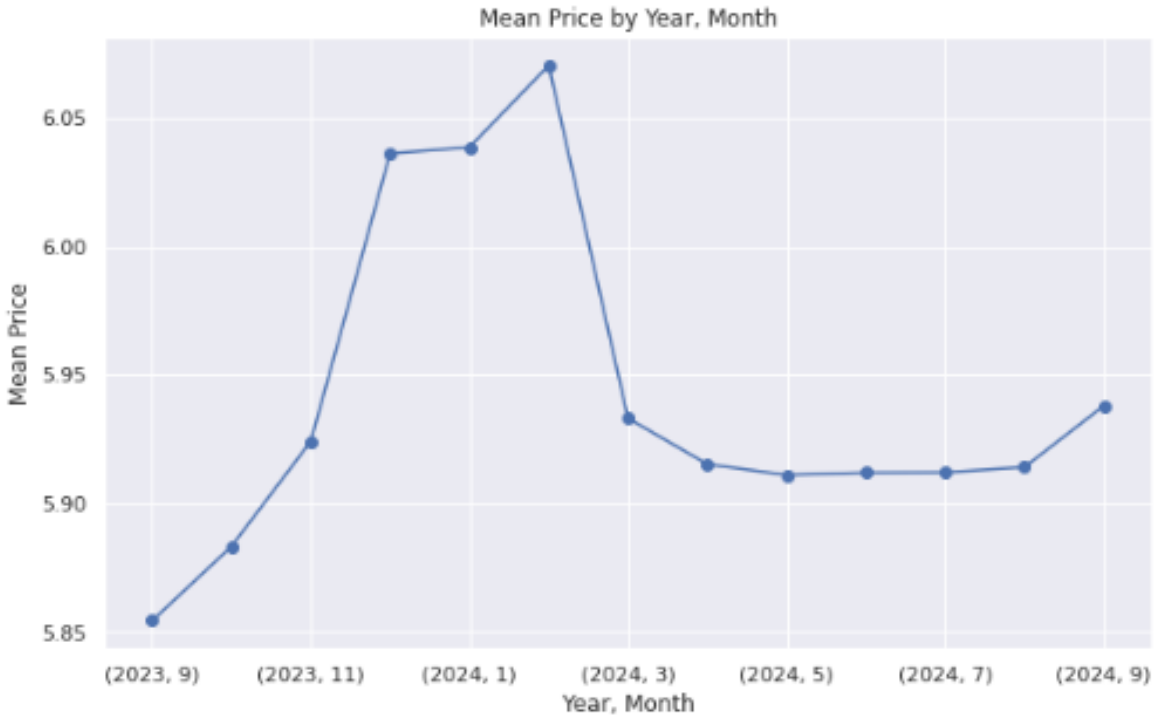


Figura 6.14: Cantidad de registros por día de la semana.



.Figura 6.15: Precio medio por año y mes.

Variación del Precio Medio por Año y Mes

El gráfico de líneas adjunto refleja la fluctuación del precio medio en función del mes para el período Septiembre 2023 a Septiembre 2024. La visualización muestra un crecimiento en los precios cuando el año va alcanzando sus últimos meses, tendencia que se mantiene hasta Febrero inclusive, para posteriormente comenzar a descender el nivel de precios hasta alcanzar un equilibrio para el resto del año.

En este período destacado se encuentra la temporada de verano donde se observa un alza en los precios que puede explicarse por la temporada del año, fechas festivas y eventos particulares de Río de Janeiro, como por ejemplo, el mundialmente famoso carnaval de Río de Janeiro.

Análisis de la Variación del Precio Promedio según Disponibilidad

El gráfico de líneas representa la variación del precio promedio de alojamientos a lo largo del tiempo, diferenciando entre aquellos que han sido reservados y los que están disponibles. Los datos, agrupados por mes y año, revelan tendencias distintivas entre las dos categorías de disponibilidad.

Se visualiza la misma tendencia que el gráfico anterior: la temporada de verano la cual coincide también con fechas festivas (navidad y fin de año), y la particularidad que presenta esta ciudad de ofrecer en esta época un evento reconocido mundialmente, muestra un comportamiento diferencial respecto al resto del año. Los precios de los alquileres disponibles son considerablemente mayor a los del resto del año, y los alquileres ya reservados acompañan esta tendencia pero en menor escala.

Este comportamiento puede deberse a que los datos fueron extraídos en la última etapa del año, por lo que al estar más próximo a dicha temporada los precios se disparan. Por otro lado, los alquileres que ya se encuentran reservados a esa fecha y que se visualizan claramente con un precio inferior, puede encontrar su explicación en que los mismos fueron confirmados con una antelación considerada y por lo tanto a precios más accesibles.

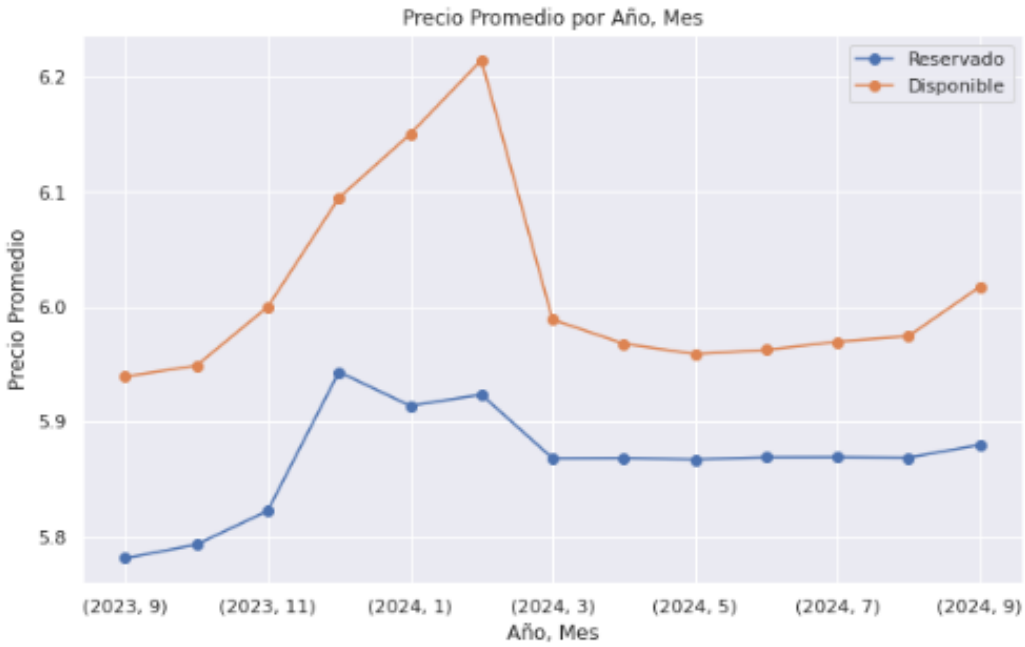


Figura 6.16: Precio medio por año y mes disponible - reservado

Análisis de la Variación del Precio según el barrio.

El gráfico representa el comportamiento de los precios de los alquileres a lo largo del período Septiembre 2023 a Septiembre 2024, según el barrio donde se encuentra ubicado el alojamiento.

La tendencia que se viene observando a lo largo del análisis también se mantiene en este caso, es decir, en la temporada de verano es cuando los precios son más elevados sea donde sea que esté ubicado el alojamiento.

Más allá de este comportamiento homogéneo en los precios, se observa que si existen diferencias en el promedio de precios de los alquileres según el barrio en el que se encuentre. Este comportamiento es razonable y predecible ya que tiene sentido que los alojamientos ubicados en lugares más atractivos, lujosos, accesibles y seguros sean más demandados y por lo tanto más costosos

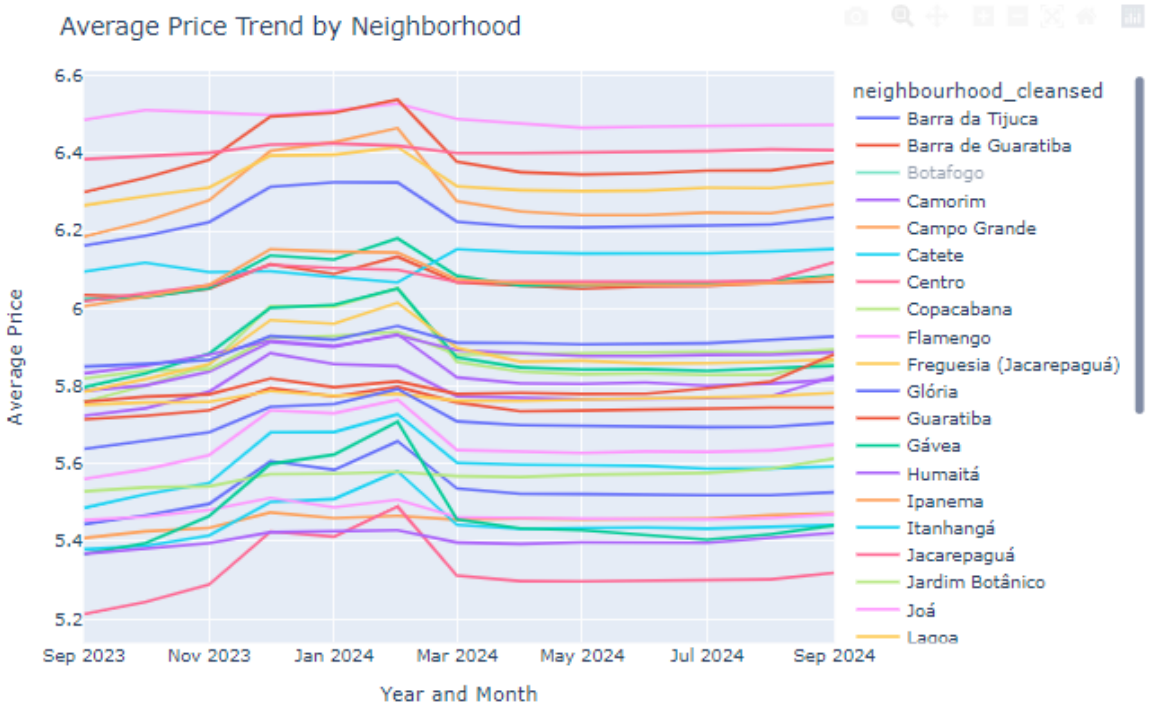


Figura 6.17: Precio promedio por barrio

Resumen

Este informe proporciona un análisis de los precios de los alojamientos en Airbnb, centrándose en varios factores influyentes. Los hallazgos claves incluyen que:

No se visualiza una única variable predominante sobre el resto que explica por sí sola o mayoritariamente la variación de los precios. A priori su fluctuación parece explicarse por la consideración de varios factores de los alojamientos: tipo de alojamiento, cantidad de huéspedes y la ubicación, por mencionar alguno de ellos.

Los precios no fluctúan significativamente según el día de la semana.

Los precios alcanzan su pico en los meses de verano pero esta tendencia disminuye hasta alcanzar un equilibrio para el resto del año.

Más allá de las características propias de un alojamiento existen otros factores que inciden directamente en el precio como lo son las fechas y eventos particulares.

Los alojamientos con precios más accesibles son reservados con preferencia sobre los de mayor precio, lo que es una tendencia lógica.

7. Modelado

Para comenzar, tal como en las *notebooks* de preparación y exploración de datos se importan las librerías necesarias y se cargan los datos generados en la *notebook* de preparación de datos. El proceso de carga de datos se realizó mediante la biblioteca Pandas para cargar los archivos CSV.

A continuación, se presentan las métricas de los distintos modelos desarrollados.

	Train		Test	
	R2	**MSE**	**R2**	**MSE**
Regresión Lineal	0,385	0,545	0,386	0,543
Árboles de Decisión	0,967	0,029	0,964	0,032
Random Forest	0,967	0,029	0,964	0,032

Gradient Boosting	0,521	0,425	0,521	0,423
Adaboost	0,318	0,604	0,318	0,603
XGBoost	0,726	0,243	0,720	0,247

Tabla 7.1 - Comparativa de Resultados

En los capítulos 7.1 y 7.2 se profundizará en su ejecución y los resultados.

Cabe destacar que los resultados de los modelos Árboles de Decisión y *Random Forest*, si bien por la representación en el cuadro parecen ser iguales los mismos presentan diferencias mínimas.

Para *Random Forest* los resultados fueron los siguientes:

Error Cuadrático Medio - Entrenamiento: 0.02911263938982635
Raíz del Error Cuadrático Medio - Entrenamiento: 0.17062426377812256
R^2 - Entrenamiento: 0.9671952549347638

Error Cuadrático Medio - Prueba: 0.03163430151885771
Raíz del Error Cuadrático Medio - Prueba: 0.17786034273794063
R^2 - Prueba: 0.9642524233550992

Para Árboles de Decisión los resultados fueron:

Error Cuadrático Medio - Entrenamiento: 0.029099970645036395
Raíz del Error Cuadrático Medio - Entrenamiento: 0.17058713505137602
R^2 - Entrenamiento: 0.9672095303475

Error Cuadrático Medio - Prueba: 0.031623533481537874
Raíz del Error Cuadrático Medio - Prueba: 0.17783006911525923
R^2 - Prueba: 0.9642645915150055

7.1 Modelo Seleccionado - Random Forest

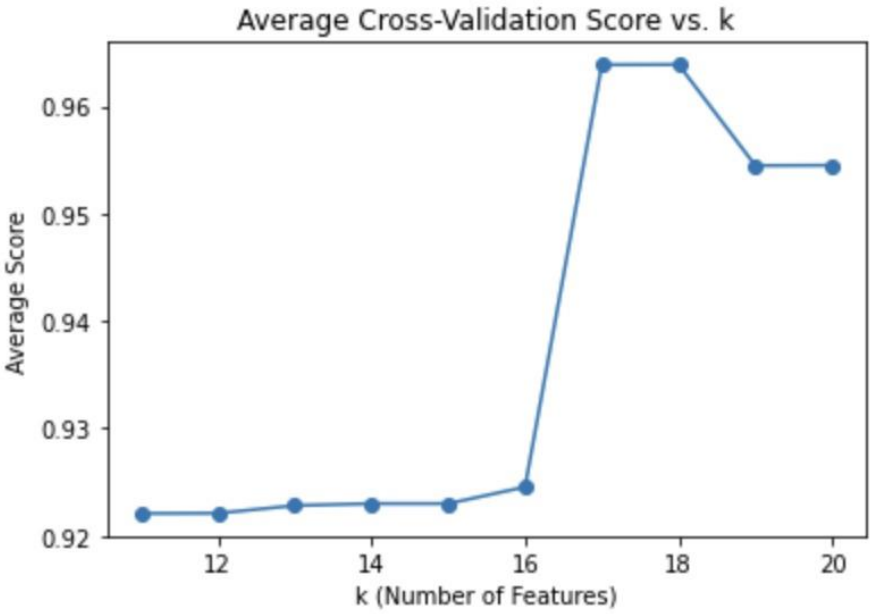


Figura 7.1.1: Selección k-best Random Forest

Se utiliza la técnica de SelectKBest, que selecciona los k mejores características del conjunto de datos de entrenamiento según la métrica de prueba *F-regression*. El rango de valores de k se establece de 11 a 20, para luego iterar sobre cada valor de k y para cada uno seleccionar las características correspondientes. Por último se entrena un modelo de *Random Forest* el cual es evaluado utilizando validación cruzada con cv = 5.

El mejor resultado se alcanzó con k igual a 18, con R2 igual a 0,9639 y las variables seleccionadas fueron las siguientes: ‘host_response_rate’, ‘host_acceptance_rate’, ‘accommodates’, ‘bathrooms_count’, ‘bedrooms’, ‘availability_30’, ‘availability_60’,

‘availability_90’, ‘availability_365’, ‘number_of_reviews’, ‘review_scores_rating’, ‘review_scores_value’, ‘reviews_per_month’, ‘shared, host_is_superhost_encoded’, ‘room_type_encoded’, ‘month’ y ‘antiguedad’.

Luego, se procedió a realizar mediante *grid search*, la búsqueda de hiper parámetros óptimos. Esta acción requirió un costo computacional demasiado alto (se generaron errores de memoria durante la ejecución de *grid search*) para los recursos con los que se cuentan, por lo que se presentó la imposibilidad de realizar esta búsqueda con las variables seleccionadas en el paso anterior.
Esta situación derivó en realizar la búsqueda de hiper parámetros con menor cantidad de variables.

Finalmente se entrena el modelo con las características e hiper parámetros seleccionados.

Se optó por fijar los hiper parámetros del modelo, estableciendo el valor de min_samples_leaf en 4 y min_samples_split en 2. El parámetro “min_samples_leaf” determina la cantidad mínima de observaciones que debe tener un nodo para que el mismo pueda dividirse, mientras que “min_samples_split” indica el número mínimo de observaciones que debe de tener cada uno de los nodos hijos para que se la división pueda ser realizada. [13]

Los resultados obtenidos en train y test son los siguientes:

Train

- Error cuadrático medio (MSE): 0.0291
- Raíz del error cuadrático medio (RMSE): 0.1706
- Coeficiente de determinación (R2): 0.9672

Test

- Error cuadrático medio (MSE): 0.0316
- Raíz del error cuadrático medio (RMSE): 0.1779
- Coeficiente de determinación (R2): 0.9643

El análisis realizado demostró que el modelo *Random Forest*, configurado con 200 estimadores, superó a los demás modelos evaluados. El hiper parámetro "n_estimators" determina la cantidad de árboles que se van a utilizar en el modelo. Este modelo arrojó un MSE de 0,0291 y un R^2 de 0,9672 en la fase de entrenamiento, así como un MSE de 0,0316 y un R^2 de 0,9642 en la fase de prueba, lo que indica una alta capacidad predictiva y una excelente generalización a datos no vistos.

Este modelo tiene un MSE bajo y un R^2 alto tanto en el conjunto de entrenamiento así como en el de prueba, lo que indica un buen equilibrio entre evitar el sobreajuste (*overfitting*) y mantener una buena capacidad de generalización.

En referencia al R^2, que es una medida de cuánta variabilidad en los datos puede ser explicada por el modelo, un R^2 de 0.9672 en el entrenamiento y 0.9643 en *test* es excepcionalmente alto, indicando que el modelo puede explicar la mayoría de la variabilidad en los precios.

En cuanto al RMSE en la fase de entrenamiento, el modelo *Random Forest* alcanzó un valor de 0.171, mientras que en la fase de prueba, el RMSE fue ligeramente superior, situándose en 0.178. Estos valores indican que, en promedio, las predicciones del modelo difieren del precio de los alojamientos en Airbnb por una cantidad relativamente pequeña. Esta diferencia marginal entre el RMSE de entrenamiento y el de prueba subraya una buena generalización del modelo a datos no vistos, reflejando una capacidad equilibrada para aprender de los datos sin caer en sobreajuste.

Una vez analizado con respecto a la variable 'log_price', se analizaron los resultados con respecto a la variable original price.

La diferencia de error en el conjunto de entrenamiento es de aproximadamente 23,8, lo que indica cuánto varían en promedio las predicciones del modelo respecto a los valores reales. Mientras tanto, el porcentaje de error en el conjunto de entrenamiento es de alrededor de 2.84%.

En cuanto al conjunto de prueba, la diferencia de error, la cual está definida como la diferencia entre la media del precio predicho y la media de la variable objetivo, es de aproximadamente 23.6. Por otro lado, el porcentaje de error (métrica definida como la

diferencia de error sobre la media de la variable objetivo) en el conjunto de prueba es de alrededor de 2.83%.

Para complementar lo realizado se visualizó la importancia de las características determinadas por el modelo, mediante la función feature_importances de *scikit learn*, proporcionando una comprensión de los factores que más influyen en la predicción de precios. Esta importancia se calcula según la contribución de cada característica al poder predictivo del modelo. Cuanto mayor sea el valor, más influyente es la característica en las predicciones del modelo.

En cuanto a estas variables, se evaluó la importancia de cada una de ellas, lo que reveló *insights* significativos sobre los factores que más influyen en los precios, ofreciendo una guía valiosa para los anfitriones en la plataforma.

La variable 'bathrooms_count' emergió como la más influyente, con una importancia relativa de aproximadamente el 23.8%, subrayando el impacto crítico que la cantidad de baños tiene en la valoración de los alojamientos por parte de los huéspedes. Esto podría reflejar las expectativas de comodidad y privacidad de los usuarios de Airbnb, especialmente en alojamientos compartidos o en destinos populares.

De cerca le sigue la antigüedad, contribuyendo con un 10.3% a la importancia del modelo. Este resultado sugiere que la duración de la oferta del alojamiento en la plataforma puede ser un indicador de calidad y fiabilidad, lo que afecta positivamente a su precio.

'Number_of_reviews' y 'reviews_per_month', con importancias del 8.5% y 8.1% respectivamente, indican que la frecuencia y la cantidad de reseñas también juegan un papel fundamental en la determinación de los precios. Estos hallazgos apuntan hacia la importancia crítica de la reputación en línea y la actividad en la plataforma en la formación de precios competitivos.

Estos hallazgos ofrecen una base empírica para que los anfitriones optimicen sus estrategias de precios. Al enfocarse en mejorar las características identificadas como más influyentes, los anfitriones pueden aumentar su competitividad y, potencialmente, su

rentabilidad en la plataforma.

	caracteristica	importancia
3	bathrooms_count	0.237529
17	antiguedad	0.103254
9	number_of_reviews	0.085231
12	reviews_per_month	0.082028
4	bedrooms	0.077573
15	room_type_encoded	0.069918
8	availability_365	0.056012
2	accommodates	0.049742
1	host_acceptance_rate	0.049449
0	host_response_rate	0.032082
11	review_scores_value	0.026788
10	review_scores_rating	0.026171
16	month	0.026134
6	availability_60	0.025813
7	availability_90	0.023520
5	availability_30	0.016330
13	shared	0.006757
14	host_is_superhost_encoded	0.005668

Figura 7.1.2: Porcentaje de importancia de variables en Random Forest

A continuación, se visualiza gráfico de la importancia de las variables al poder predictivo del modelo *Random Forest*.

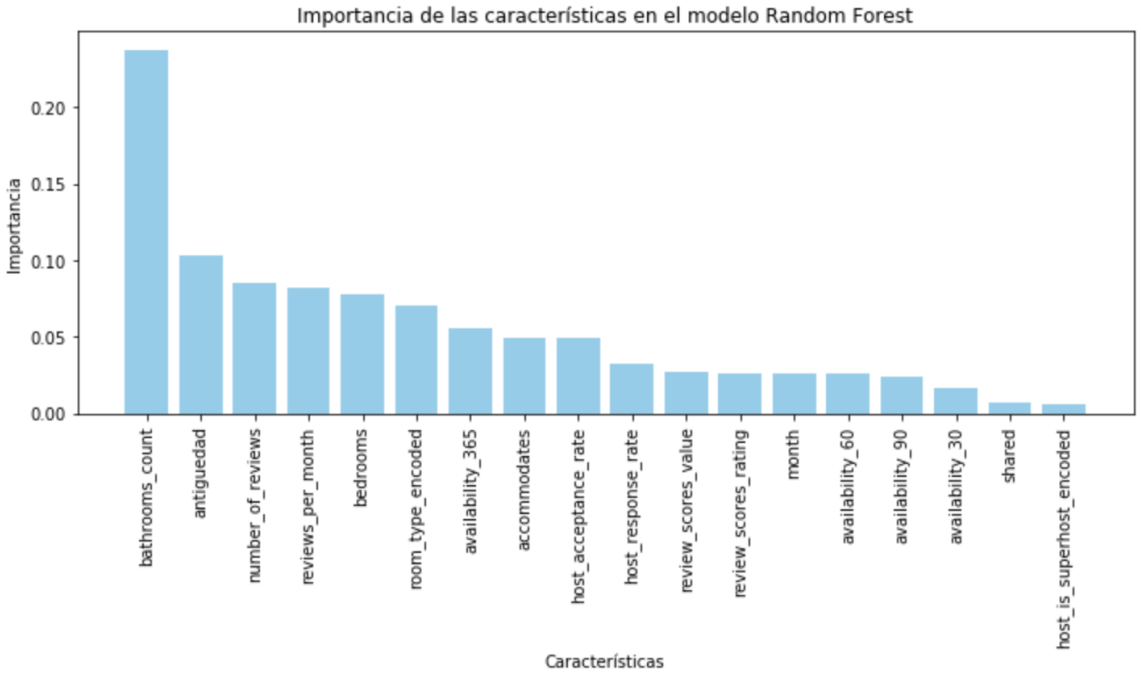


Figura 7.1.3: Gráfico de importancia de variables en Random Forest.

7.2 Otros Modelos

Se entrenaron otros modelos, entre los que se encuentran Regresión lineal, Árboles de decisión, *XG Boost*, *Adaboost* y *Gradient Boosting* bajo criterios de precisión tales como el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) y el coeficiente de determinación (R^2).

Regresión Lineal

Se llevó a cabo un análisis empleando un modelo de regresión lineal junto con la técnica de *cross-validation*. El modelo se entrenó con distintas cantidades de *features* (k), abarcando desde 2 hasta 20, utilizando cross_val_score con un valor de cv igual a 5. Para cada valor de k, se eligió un conjunto específico de características con el propósito de optimizar la capacidad predictiva del modelo.

A continuación, se presenta un gráfico que ilustra el desempeño de *cross-validation* con distintas combinaciones y cantidades de *features* utilizados como variables predictoras.

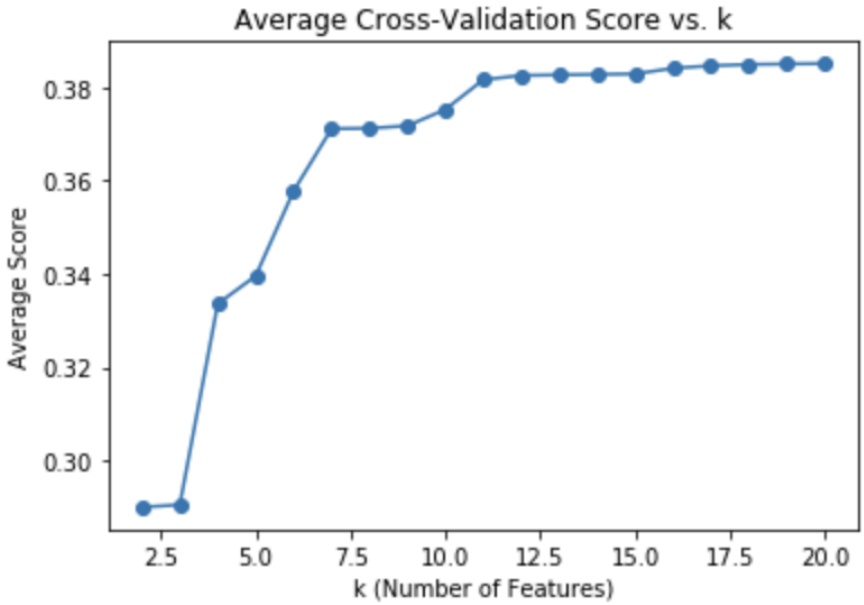


Figura 7.2.1: Selección k-best Regresión lineal

Mediante *cross-validation*, el k óptimo fue 20 con un r2 igual a 0.3851. Las variables seleccionadas mediante este proceso fueron las siguientes: ‘host_response_rate’, ‘host_acceptance_rate’, ‘accommodates’, ‘bathrooms_count’, ‘bedrooms’,

‘availability_30’, ‘availability_60’, ‘availability_90’, ‘availability_365’, ‘number_of_reviews’, ‘review_scores_rating’, ‘review_scores_value’, ‘reviews_per_month’, ‘shared’, ‘host_is_superhost_encoded’, ‘room_type_encoded’, ‘month’, ‘numero_dia’, ‘antiguedad’ y ‘neighbourhood_cleansed_encoded’.

Los resultados obtenidos en *train* y *test* son los siguientes:

Train

- Error cuadrático medio (MSE): 0.5453
- Raíz del error cuadrático medio (RMSE): 0.7384
- Coeficiente de determinación (R2): 0.3855
- Error absoluto medio (MAE): 0.5432

Test

- Error cuadrático medio (MSE): 0.5433
- Raíz del error cuadrático medio (RMSE): 0.7371
- Coeficiente de determinación (R2): 0.3861
- Error absoluto medio (MAE): 0.5425

Árboles de Decisión

Se realizó un análisis utilizando un modelo de regresión de árbol de decisión con la técnica de *cross-validation*. El modelo fue entrenado con diferentes números de características seleccionadas (k), variando desde 2 hasta 20, utilizando cross_val_score con un valor de cv igual a 5. Para cada valor de k, se seleccionó un conjunto específico de características con el objetivo de maximizar la capacidad predictiva del modelo.

A continuación, se visualiza gráfico con el rendimiento de *cross-validation* con distintas combinaciones y cantidades de factores utilizados como variables predictoras.

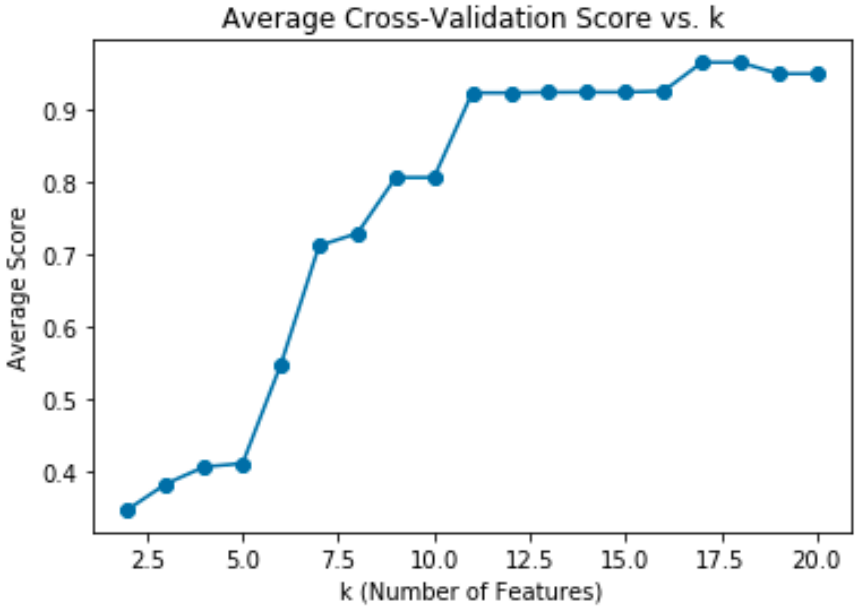


Figura 7.2.2: Selección k-best Árboles de Decisión

Mediante *cross-validation*, el k óptimo fue 18 con un r2 igual a 0.9639. Las variables seleccionadas mediante este proceso fueron las siguientes: 'host_response_rate', 'host_acceptance_rate', 'accommodates', 'bathrooms_count', 'bedrooms', 'availability_30', 'availability_60', 'availability_90', 'availability_365', 'number_of_reviews', 'review_scores_rating', 'review_scores_value', 'reviews_per_month', 'shared', 'host_is_superhost_encoded', 'room_type_encoded', 'month' y 'antiguedad'.

Se procedió a entrenar el modelo de regresión de árbol de decisión utilizando la técnica de *Grid Search* con el objetivo de encontrar los hiper parámetros óptimos. El conjunto de hiper parámetros evaluados incluyó diversas combinaciones de max_depth ([None, 10, 20, 30]), min_samples_split ([2, 5, 10]), y min_samples_leaf ([1, 2, 4]).

Los resultados obtenidos en *train* y *test* son los siguientes:

Train

- Error Cuadrático Medio (MSE): 0.0291
- Raíz del Error Cuadrático Medio (RMSE): 0.1706
- R-cuadrado (R2): 0.9672
- Error Absoluto Medio (MAE): 0.0592

Test

- Error Cuadrático Medio (MSE): 0.0316
- Raíz del Error Cuadrático Medio (RMSE): 0.1778
- R-cuadrado (R2): 0.9643
- Error Absoluto Medio (MAE): 0.0618

Estos resultados indican que el modelo entrenado con los hiper parámetros óptimos logra un rendimiento destacado tanto en el conjunto de entrenamiento como en el de pruebas.

Por último, mediante la función de *permutation importance* de *scikit learn*, se identificaron las variables más importantes para el modelo ya ejecutado.

	importances_mean	importances_std	feature
9	0.636795	0.000157	number_of_reviews
3	0.590183	0.000267	bathrooms_count
2	0.546608	0.000365	accommodates
12	0.537369	0.000368	reviews_per_month
17	0.519904	0.000244	antiguedad
1	0.485471	0.000257	host_acceptance_rate
8	0.470427	0.000220	availability_365
15	0.465451	0.000249	room_type_encoded
4	0.461541	0.000270	bedrooms
6	0.420320	0.000295	availability_60
10	0.396230	0.000192	review_scores_rating
7	0.351570	0.000298	availability_90
0	0.320585	0.000275	host_response_rate
11	0.314595	0.000291	review_scores_value
5	0.206024	0.000167	availability_30
13	0.189537	0.000410	shared
16	0.151535	0.000253	month
14	0.069240	0.000136	host_is_superhost_encoded

Tabla 7.2.1: Importancia de variables en Árboles de Decisión

XGBoost

A partir de las variables seleccionadas mediante *SelectKBest*, con la métrica de f_regression, mediante *cross-validation* con el modelo de *random forest* y *decision tree*, se entrenó el modelo de *XGBoost* utilizando la técnica de *Grid Search* con *cross-validation*, con cv igual a 5.

El conjunto de hiper parámetros evaluados incluyó diversas combinaciones de 'max_depth': [3, 5, 7], 'learning_rate': [0.1, 0.01], 'n_estimators': [100, 200] y 'min_child_weight': [1, 3].

Los hiper parámetros seleccionados fueron: learning rate: 0,1, max_depth:7, min_child_weight: 1 y n_estimators: 200.

Los resultados obtenidos en *train* y *test* son los siguientes:

Train

- Error Cuadrático Medio (MSE): 0.2432
- Raíz del Error Cuadrático Medio (RMSE): 0.4932
- R-cuadrado (R2): 0.7259
- Error Absoluto Medio (MAE): 0.3619

Test

- Error Cuadrático Medio (MSE): 0.2475
- Raíz del Error Cuadrático Medio (RMSE): 0.4975
- R-cuadrado (R2): 0.7203
- Error Absoluto Medio (MAE): 0.3644

8. Interpretabilidad y Aplicabilidad del modelo seleccionado

8.1 Introducción

En el ámbito de la predicción de precios de alojamientos en plataformas de economía compartida como Airbnb, el desarrollo y la implementación de modelos de aprendizaje automático representan una herramienta fundamental para comprender y optimizar estrategias de fijación de precios. La selección del modelo *Random Forest*, con su destacada *performance* en términos de Error Cuadrático Medio (MSE), Coeficiente de Determinación (R^2), y Raíz del Error Cuadrático Medio (RMSE), subraya no sólo su capacidad predictiva sino también su potencial aplicabilidad en escenarios reales.

Este apartado del trabajo se centra en comentar sobre la interpretabilidad y la aplicabilidad del modelo *Random Forest* elegido, destacando cómo sus características intrínsecas y los resultados obtenidos pueden traducirse en *insights* prácticos para la predicción de precios. A través del análisis de variables, hiper parámetros seleccionados, y la estructura del modelo, se busca proporcionar una comprensión de los mecanismos subyacentes que impulsan las predicciones del modelo y cómo estos pueden ser aplicados para informar decisiones estratégicas en la fijación de precios de alojamientos.

La capacidad de interpretar un modelo complejo como *Random Forest* es importante para ganar confianza en sus predicciones y para justificar las recomendaciones de precios basadas en estos. Al mismo tiempo, la aplicabilidad se refiere a cómo las predicciones y los *insights* generados pueden ser implementados efectivamente para mejorar la rentabilidad y la competitividad de los alojamientos en la plataforma. En este contexto, se explica cómo el modelo se ajusta a diferentes necesidades y escenarios de usuarios, permitiendo a los anfitriones optimizar sus estrategias de precios con base en una comprensión de los factores que afectan la demanda y los precios en el mercado de alquiler de corta duración.

8.2 Interpretabilidad del Modelo

El modelo *Random Forest* seleccionado se caracteriza por su capacidad para manejar un conjunto complejo y diverso de variables, lo que lo convierte en una herramienta poderosa para la predicción de precios en Airbnb. La incorporación de variables como el número de habitaciones, la cantidad de baños, la puntuación de las reseñas, entre otras, permite al modelo capturar la multitud de factores que pueden influir en el precio de un alojamiento.

La interpretabilidad de *Random Forest*, aunque desafiante debido a su naturaleza de ensamble, puede abordarse examinando la importancia de las características (*feature importance*). Esta métrica indica cuán valiosas son las distintas variables para las decisiones tomadas por el modelo. En este caso, variables como 'bathrooms_count' y la antigüedad (variable calculada como la diferencia en días entre 'last_scraped' y 'host_since' para entender la experiencia o tiempo de actividad de los anfitriones en la plataforma) juegan roles significativos, lo que intuitivamente coincide con las expectativas del mercado de alquileres de corta duración: los alojamientos más grandes y mejor ubicados tienden a tener precios más altos.

Además, la capacidad del modelo para ajustarse mediante hiper parámetros específicos —como el número de estimadores (árboles) y la profundidad máxima— permite una fina personalización para equilibrar complejidad y rendimiento, mejorando así la precisión sin caer en el sobreajuste.

8.3 Aplicabilidad del Modelo

Desde una perspectiva práctica, el modelo *Random Forest* ofrece ventajas significativas para los anfitriones de Airbnb buscando optimizar la fijación de precios de sus alojamientos. La alta precisión y el robusto rendimiento del modelo, reflejado en un MSE de entrenamiento de 0,0291 y un R^2 de 0,9672, proveen una base sólida para la toma de decisiones informadas.

Los anfitriones pueden aplicar las predicciones del modelo para ajustar dinámicamente sus precios en respuesta a cambios en variables clave. Por ejemplo, pueden

incrementar los precios durante períodos de alta demanda, identificados por la variable de ‘month’ o ‘availability_30’, o ajustarlos según la cantidad de habitaciones y servicios ofrecidos, para permanecer competitivos y maximizar ingresos.

Más aún, la aplicabilidad del modelo se extiende a la gestión de estrategias de *marketing* y operaciones para los alojamientos. Al entender las variables que más influyen en los precios, los anfitriones pueden focalizar mejoras en sus propiedades o destacar ciertas características en sus anuncios para atraer a un mayor número de huéspedes.

9. Conclusiones

La conclusión de este estudio subraya la eficacia del modelo de *Random Forest* en la predicción de precios para Airbnb, destacándose como una metodología robusta y flexible. Esta herramienta combina precisión analítica, habilidad para procesar múltiples variables simultáneamente, y adaptabilidad a ajustes específicos, ofreciendo así a los anfitriones una base sólida para decisiones fundamentadas en evidencia concreta. Mirando hacia adelante, el potencial de desarrollo y refinamiento de este modelo promete contribuciones valiosas al campo de la predicción de precios y al más amplio espectro de la economía colaborativa.

Al comienzo de esta investigación, se parte de ideas previas sobre la dinámica de Airbnb basadas en el conocimiento de la aplicación. Se asume que ciertas variables ejercen un impacto directo y significativo en la fijación de precios, como la localización, el tamaño del inmueble, y la estacionalidad. No obstante, el análisis desplegado reveló un panorama más complejo de lo anticipado.

Se constata que ningún factor, por sí mismo, establece de manera unilateral el precio de un alojamiento. Asimismo, variables inicialmente consideradas clave no evidenciaron la influencia pronosticada. Este descubrimiento motivó una indagación más amplia y detallada sobre posibles elementos influyentes, ampliando significativamente el espectro de variables analizadas más allá de las inicialmente señaladas como determinantes.

Aunque se identificaron incrementos tarifarios durante los meses de alta demanda —diciembre a febrero—, en línea con las expectativas para la temporada alta en Brasil, este patrón no emergió como un factor preponderante en el modelo predictivo. La búsqueda del modelo óptimo se caracterizó por un enfoque de ensayo y corrección, integrando un abanico extenso de variables para evaluar su impacto efectivo sobre los precios. Esta metodología orientó hacia una comprensión más integral y exacta de las variables que realmente influyen en la determinación de precios en Airbnb, superando suposiciones iniciales para capturar la riqueza y los detalles sutiles desvelados por el análisis de datos.

10.Acciones futuras

La evolución constante de la economía compartida y el papel crecientemente influyente de plataformas como Airbnb en los mercados de alojamiento temporal presentan tanto oportunidades como desafíos. Este dinamismo, marcado por cambios en las preferencias de los consumidores, avances tecnológicos y nuevas regulaciones, requiere un enfoque proactivo y adaptativo hacia la predicción y optimización de precios. A través de este trabajo, hemos desentrañado patrones y factores significativos que influencian la fijación de precios en Airbnb, revelando la complejidad y lo multifacético de este fenómeno.

Sin embargo, el campo de la predicción de precios en el contexto de la economía compartida está lejos de ser exhaustivo y está en constante evolución. Los resultados obtenidos y las metodologías empleadas en este estudio pavimentan el camino para futuras investigaciones y aplicaciones prácticas.

Este capítulo se dedica a esbozar un conjunto de acciones futuras recomendadas, que no solo buscan mejorar y expandir el conocimiento adquirido sino también sugerir aplicaciones prácticas de este conocimiento en el mundo real. Desde la ampliación de conjuntos de datos y la experimentación con nuevas técnicas de modelado hasta el estudio del impacto de políticas regulatorias y el desarrollo de herramientas de optimización de precios.

A continuación, se presentan una serie de recomendaciones estratégicas y áreas de investigación que se consideran claves para avanzar en el marco de la comprensión de la fijación de precios en plataformas de alojamiento temporal.

1. Ampliación del Conjunto de Datos.

Una acción futura clave es la ampliación del conjunto de datos para incluir más variables que puedan influir en la predicción de precios de Airbnb. Esto podría incluir datos más granulares sobre la ubicación de los alojamientos, como la proximidad a puntos de interés turístico o la accesibilidad a servicios de transporte público. Además, integrar variables temporales más específicas, como eventos locales o festividades, podría mejorar

la precisión de las predicciones al capturar mejor las fluctuaciones de precios relacionadas con la demanda estacional o puntual.

2. Aplicación de Nuevas Técnicas de Modelado.

Explorar y aplicar nuevas técnicas de modelado avanzado, como redes neuronales profundas o algoritmos de aprendizaje por refuerzo, podría proporcionar mejoras significativas en la capacidad predictiva de los modelos. La implementación de técnicas de aprendizaje profundo, en particular, puede ser muy prometedora dada su capacidad para manejar grandes conjuntos de datos y capturar relaciones complejas entre variables.

3. Integración de Análisis Sentimental.

El análisis sentimental de las reseñas dejadas por los huéspedes podría ofrecer insights valiosos sobre cómo las percepciones y experiencias de los usuarios influyen en los precios. Aplicar técnicas de procesamiento de lenguaje natural para analizar el sentimiento y los temas principales en las reseñas puede ayudar a identificar factores cualitativos que afectan la valoración de los alojamientos.

4. Análisis y transformación de la variable “amenities”.

La inclusión y análisis detallado de la variable "amenities" en futuras investigaciones se presenta como un paso fundamental para enriquecer el modelo predictivo de precios. Esta variable, que abarca desde WI-FI hasta aire acondicionado, refleja la diversidad y potencial impacto de las comodidades en la decisión de los huéspedes y, consecuentemente, en la predicción de precios. Se propone investigar métodos avanzados para analizar y cuantificar estas comodidades, empleando técnicas de procesamiento de lenguaje natural y aprendizaje automático. El objetivo es identificar cómo cada “amenity” contribuye a la valoración del alojamiento, permitiendo así a los anfitriones ajustar sus ofertas para mejorar su posicionamiento en el mercado y maximizar la rentabilidad. Este enfoque no solo permitirá una optimización más precisa de los precios, sino que también ofrecerá una visión más profunda sobre las preferencias y expectativas de los consumidores en el dinámico mercado de alojamiento temporal.

5. Estudio de Impacto de Políticas Regulatorias.

Realizar estudios sobre el impacto de las políticas regulatorias en los mercados locales de alquiler de corta duración podría ofrecer una comprensión más profunda de cómo las intervenciones externas afectan la dinámica de precios. Investigar la relación entre regulaciones específicas (como impuestos turísticos, límites de días de alquiler al año, etc.) y los precios de alquiler podría informar el desarrollo de políticas más efectivas para manejar el crecimiento de plataformas como Airbnb de manera que beneficie tanto a anfitriones como a comunidades locales.

6. Colaboraciones Interdisciplinarias.

Fomentar colaboraciones interdisciplinarias con expertos en turismo, economía, urbanismo, y ciencias sociales podría enriquecer la investigación sobre predicción de precios en plataformas de economía compartida. Estas colaboraciones pueden proporcionar perspectivas más amplias y profundas sobre los factores que influencian los precios de alquiler, desde tendencias económicas globales hasta dinámicas sociales y culturales locales.

7. Desarrollo de Herramientas de Optimización de Precios.

Basándose en los modelos predictivos desarrollados, una dirección futura podría ser el desarrollo de herramientas o aplicaciones de software que ayuden a los anfitriones de Airbnb a optimizar sus estrategias de precios en tiempo real. Estas herramientas podrían utilizar algoritmos predictivos para sugerir ajustes de precios basados en cambios en la demanda, competencia, y otros factores relevantes, maximizando así los ingresos y la ocupación de los alojamientos.

Estos pasos futuros podrían perfeccionar el modelo elegido y obtener predicciones más precisas sobre los precios de los alojamientos.

Una vez obtenido el mejor modelo predictivo, se trabajará en detallar la interpretabilidad y aplicabilidad del modelo y sus resultados.

11. Referencias bibliográficas

[1] BRJoaquin. “airbnb-dl”. Noviembre 2023. [Online]. Available: https://github.com/BRJoaquin/airbnb-dl

[2] Mertikas, D. "Dynamic Pricing Modelling — Airbnb Amsterdam Case Study." Octubre 2022. [Online]. Available: https://medium.com/@d.mertikas/dynamic-pricing-modelling-airbnb-amsterdam-case-study-2bd988d11e2b

[3] Santucci de Oliveira, B; Eger Bauer, J; Tomelin, C; Lisboa Sohn, A. “ECONOMÍA COMPARTIDA Un estudio sobre Airbnb”. Julio 2018. [Online]. Available: https://www.redalyc.org/journal/1807/180762492005/html/

[4] Cavalcanti, D. “Airbnb: ¿Qué es? ¿Cómo funciona? y todo lo que debes saber para incrementar tus reservas.” Agosto 2022. Available: https://stays.net/blog/es/que-es-airbnb/#:~:text=Es%20una%20plataforma%20que%20ofrecede%20hospedarse%20en%20el%20mundo

[5] Silva. C. “¿Qué es Airbnb y cómo funciona?” Noviembre 2015. [Online]. Available: https://www.entornoturistico.com/que-es-airbnb-y-como-funciona/

[6] Airbnb. “Newsroom. About Us.” [Online]. Available: https://news.airbnb.com/about-us/

[7] Airbnb. “Cómo fijar los precios de tu alojamiento”. [Online]. Available: https://es.airbnb.com/help/article/52

[8] Airbnb. “Alojamientos similares”. [Online]. Recuperado de https://es.airbnb.com/help/article/3399

[9] Ortega, C. “Optimización de precios: Qué es y guía para realizarla”. [Online]. Available: https://www.questionpro.com/blog/es/optimizacion-de-precios/

[10] Marín, G; Toscano, M; Brandes, M; Alfaro, P; Ríos, B. “How Machine Learning is reshaping. Price Optimization.” [Online]. Available: https://tryolabs.com/blog/price-optimization-machine-learning/

[11] Nerd, C. "K Nearest Neighbor (KNN) Imputer Explained.” Octubre 2023. [Online]. Available: https://medium.com/@karthikheyaa/k-nearest-neighbor-knn-imputer-explained-1c56749d0dd7

[12] The Data Schools. “Grid Search en Python”. [Online]. Available: https://thedata schools.com/python/grid-search/

[13] Gareth James, D; Trevor Hastie, R; Taylor, J. (2023). An Introduction to Statistical Learning with Python. Springer.

[14] Zohar, Y. "Explain ML Models with Permutation Importance.” [Online]. Available: https://www.aporia.com/learn/feature-importance/explain-ml-models-with-permutation-importance/

[15] GeeksforGeeks. "Linear Regression Python Implementation." Diciembre 2023. [Online]. Available: https://www.geeksforgeeks.org/linear-regression-python-implementation/?ref=lbp

[16] Amat, J. “Árboles de decisión con Python: regresión y clasificación.” Octubre 2020. [Online]. Available: https://cienciadedatos.net/documentos/py07_arboles_decision _python

[17] INESDI. “Random forest, la gran técnica de Machine Learning.” Enero 2023. [Online]. Available: https://www.inesdi.com/blog/random-forest-que-es/

[18] GeeksforGeeks. "XGBoost." Febrero 2023. [Online]. Available: https://www. geeksforgeeks.org/xgboost/

[19] Inside Airbnb. “Explore the Data”. [Online]. Available: http://insideairbnb.com/ explore/

Anexo 1

El diccionario detallado a continuación fue extraído e incluido en el presente anexo tal como se obtiene de la *web* de la cual se extrajeron los diferentes conjuntos de datos. Las variables que se detallan forman parte del *dataset* denominado “dflistings”.

Diccionario de variables

id: Identificador único de Airbnb para la lista.

listing_url: URL de publicación de alojamiento.

scrape_id: identificador único de recopilación de datos del alojamiento.

last_scraped: UTC. La fecha y hora en que esta lista fue recopilada.

source: "neighbourhood search"" o ""previous scrape"". "Neighbourhood search"" significa que el listado se encontró buscando en la ciudad, mientras que "previous scrape" significa que el listado se vio en otra extracción realizada en los últimos 65 días, y se confirmó que el listado todavía estaba disponible en el sitio de Airbnb.

name: Nombre del alojamiento.

description: Descripción detallada del alojamiento.

neighborhood_overview: Descripción del anfitrión del vecindario.

picture_url:URL a la imagen de tamaño regular alojado de Airbnb para la lista.

host_id: Identificador único de Airbnb para el *host*/usuario.

host_url: La página de Airbnb para el *host*.

host_name: Nombre del anfitrión. Por lo general, sólo el primer nombre.

host_since: La fecha en que se creó el *host*/usuario. Para los anfitriones que son invitados de Airbnb, esta podría ser la fecha en que se registraron como invitados.

host_location: La ubicación reportada por el anfitrión.

host_about: Descripción sobre el *host*.

host_response_time: tasa de tiempo de respuesta.

host_response_rate: tasa de puntaje de respuesta.

host_acceptance_rate: Esa tasa a la que un anfitrión acepta solicitudes de reserva.

host_is_superhost: identifica si es *superhost* (t) o no (f).

host_thumbnail_url: URL de imágen que representa al *host*.

host_picture_URL: URL de imágen de anfitrión.

host_neighbourhood: barrio al que pertenece el *host*.

host_listings_count: El número de alojamientos que tiene el *host* (por cálculos de Airbnb).

host_total_listings_count: El número de alojamientos que tiene el *host* (por cálculos de Airbnb).

host_verifications: define si el *host* está verificado o no.

host_has_profile_pic: define si el *host* tiene (t) o no (f) foto de perfil.

host_identity_verified: define si la identidad del *host* está verificada (t) o no (f).

neighbourhood: ciudad y país del alojamiento.

neighbourhood_cleansed: El vecindario geocodificado utilizando la latitud y la longitud contra los vecindarios según lo definido por los archivos de forma digital abiertos o públicos.

neighbourhood_group_cleansed: El grupo del vecindario geocodificado utilizando la latitud y la longitud contra los vecindarios según lo definido por los archivos de forma digital abiertos o públicos.

latitude: Utiliza la proyección del sistema geodésico mundial (WGS84) para la latitud y la longitud.

longitude: Utiliza la proyección del sistema geodésico mundial (WGS84) para la latitud y la longitud.

property_type: Tipo de propiedad seleccionada. Los hoteles y los *bed* y el desayuno son descritos como tales por sus anfitriones en este campo.

room_type: Todas las casas se agrupan en los siguientes tres tipos de habitaciones: lugar completo, habitación privada y habitación compartida.

Lugar completo

Los lugares enteros son los mejores si buscas una casa fuera de casa. Con un lugar completo, tendrás todo el espacio para ti mismo. Esto generalmente incluye un dormitorio, un baño, una cocina y una entrada separada y dedicada. Los anfitriones deben tener en cuenta en la descripción si estarán en la propiedad o no (por ejemplo: El anfitrión ocupa el primer piso de la casa) y proporcionar más detalles sobre la lista.

Cuartos privados

Las habitaciones privadas son excelentes para cuando prefiere un poco de privacidad y aún valoran una conexión local. Cuando reserve una habitación privada, tendrá su propia habitación privada para dormir y puede compartir algunos espacios con otros. Es posible que deba caminar por espacios interiores que otro anfitrión o invitado puede ocupar para llegar a su habitación.

Habitaciones compartidas

Las habitaciones compartidas son para cuando no te importa compartir un espacio con otros. Cuando reserve una habitación compartida, estará durmiendo en un espacio que se comparte con otros y compartirá todo el espacio con otras personas. Las habitaciones compartidas son populares entre los viajeros flexibles que buscan nuevos amigos y estadías económicas.

accommodates: La capacidad máxima del alojamiento.

bathrooms: El número de baños en el alojamiento.

bathrooms_text: El número de baños en el alojamiento. En el sitio *web* de Airbnb, el campo de los baños ha evolucionado de un número a una descripción textual.

bedrooms: El número de dormitorios.

beds: El número de cama (s).

amenities: Detalle de las comodidades/servicios que el alojamiento ofrece.

price: Precio diario en moneda local.

minimum_nights: Número mínimo de estadía nocturna para el alojamiento(las reglas del calendario pueden ser diferentes).

maximum_nights: Número máximo de estadía nocturna para el alojamiento (las reglas del calendario pueden ser diferentes).

minimum_minimum_nights: El valor más pequeño de la noche del calendario (mirando 365 noches en el futuro).

maximum_minimum_nights: El valor mínimo más grande de la noche del calendario (mirando 365 noches en el futuro).

minimum_maximum_nights: El valor máximo más pequeño de la noche del calendario (mirando 365 noches en el futuro).

maximum_maximum_nights: El valor máximo más grande de la noche del calendario (mirando 365 noches en el futuro).

minimum_nights_avg_ntm: el valor promedio de mínimo de la noche del calendario (mirando 365 noches en el futuro).

maximum_nights_avg_ntm: El valor promedio de máximo de la noche del calendario (mirando 365 noches en el futuro).

calendar_updated: fecha de publicación de alojamiento.

has_availability: marca si el alojamiento cuenta con disponibilidad o no [t = verdadero; f = falso].

availability_30: La disponibilidad del alojamiento x días en el futuro según lo determine el calendario. Tenga en cuenta que un alojamiento puede no estar disponible porque ha sido reservado por un invitado o bloqueado por el anfitrión.

availability_60: La disponibilidad del alojamiento x días en el futuro según lo determine el calendario. Tenga en cuenta que un alojamiento puede no estar disponible porque ha sido reservado por un invitado o bloqueado por el anfitrión.

availability_90: La disponibilidad del alojamiento x días en el futuro según lo determine el calendario. Tenga en cuenta que un alojamiento puede no estar disponible porque ha sido reservado por un invitado o bloqueado por el anfitrión.

availability_365: La disponibilidad del alojamiento x días en el futuro según lo determine el calendario. Tenga en cuenta que un alojamiento puede no estar disponible porque ha sido reservado por un invitado o bloqueado por el anfitrión.

calendar_last_scraped: Última fecha que se extrajeron los datos del alojamiento.

number_of_reviews: El número de reviews que tiene el alojamiento.

number_of_reviews_ltm: El número de *reviews* que tiene el alojamiento (en los últimos 12 meses).

number_of_reviews_l30d: El número de *reviews* que tiene el alojamiento (en los últimos 30 días).

first_review: La fecha de la primera/más antigua *review*.

last_review: La fecha de la última/nueva *review*.

review_scores_rating: Calificación del alojamiento dada por los huéspedes.

review_scores_accuracy: Métrica que define precisión de descripción del alojamiento *vs* realidad, en base a la experiencia del huésped.

review_scores_cleanliness: Calificación de la limpieza del alojamiento dada por los huéspedes.

review_scores_checkin: Calificación de ingreso al alojamiento dada por los huéspedes.

review_scores_communication: Calificación de respuesta del *host* al húesped dada por parte de este último.

review_scores_location: Calificación de ubicación del alojamiento dada por el huésped.

review_scores_value: Calificación de relación calidad/precio del alojamiento dada por el huésped.

license: La licencia/permiso/número de registro.

instant_bookable: [t = verdadero; f = falso]. Si el invitado puede reservar automáticamente el alojamiento sin que el anfitrión requiere aceptar su solicitud de reserva.

calculated_host_listings_count: El número de alojamientos que el anfitrión tiene en la geografía de la ciudad/región.

calculated_host_listings_count_entire_homes: El número de alojamientos de hogar/apto que el anfitrión tiene en la geografía de la ciudad/región.

calculated_host_listings_count_private_rooms: El número de alojamientos de habitaciones privadas que el anfitrión tiene en la geografía de la ciudad/región.

calculated_host_listings_count_shared_rooms: El número de alojamientos de habitaciones compartidas que el anfitrión tiene en la geografía de la ciudad/región.

reviews_per_month: El número de calificaciones que el alojamiento tiene durante la vida útil del alojamiento.

Anexo 2

Repositorio Github

En este anexo se proporciona la URL del repositorio de GitHub donde se encuentran todos los recursos utilizados en este trabajo, incluyendo los diferentes conjuntos de datos, *notebooks* de preparación y exploración de datos, así como los *scripts* de entrenamiento de modelos.

Repositorio de GitHub:

https://github.com/jrama947/20240321-MBD-289525-157068-173098

El repositorio incluye los siguientes archivos:

- 1 - Preparacion_de_los_Datos.ipynb
- 2 - Exploracion de Datos.ipynb
- 3 - Modelado Random Forest seleccionado.ipynb
- 4 - Otros Modelos - Regresion Lineal y Arboles de decision.ipynb
- 5 - Otros Modelos - Boosting.ipynb
- Sep23_calendar.csv.gz
- Sep23_listings.csv.gz
- Sep23_reviews.csv.gz