



**UNIVERSIDAD TECNOLÓGICA NACIONAL**

**FACULTAD REGIONAL CÓRDOBA  
INGENIERÍA EN SISTEMAS DE INFORMACIÓN**

## **CIENCIA DE DATOS**

### **TRABAJO PRÁCTICO INTEGRADOR – INFORME FINAL**

**Curso:** 5K1

**Grupo 8:**

- Pablo Aimar - 78996
- Francisco Capobianco - 83548
- Felipe Diaz Posee - 84056
- Mateo Toledano – 78490

**Docentes:**

- Maldonado, Calixto Alejandro
- Callejas, Marisa Del Carmen
- Vaca, Pablo Andrés

**Fecha de presentación:** 5/11/23



## Índice

Introducción.....	2
Presentación de los DataSets.....	3
Visualización y exploración de los Datos.....	5
Planteo de la Hipótesis.....	8
Extracción, transformación y carga.....	9
Aplicación de los Modelos.....	10
Conclusión.....	14
Bibliografía.....	20

## Introducción

En este trabajo, nos adentramos en un proceso integral que abarca tres etapas cruciales para comprender y aprovechar datos relacionados con alquileres en las ciudades de Buenos Aires y Nueva York a través de la plataforma Airbnb.

### Etapa 1: Investigación y Adquisición de Datos

Nuestra primera tarea se enfoca en investigar y obtener un dataset de base que sirva como fundamento para nuestro análisis. Siguiendo esta tarea hemos seleccionado dos conjuntos de datos de Airbnb, uno correspondiente a Buenos Aires y otro a Nueva York. Estos datasets contienen información sobre las propiedades y sus comodidades, sus precios, sus reseñas, etc, que se convertirán en la materia prima de nuestra investigación.

### Etapa 2: Análisis, Preparación y Exploración de Datos

En la segunda etapa de nuestro trabajo, nos sumergimos en un análisis detallado y una comprensión profunda de los datos. Nuestro objetivo primordial es evaluar la calidad de los datos, identificando posibles errores, valores faltantes e inconsistencias.

Posteriormente, determinamos estrategias adecuadas para abordar estos problemas, que pueden incluir la eliminación de datos defectuosos, la creación de nuevas columnas en base a otras existentes, etc. También seleccionamos las columnas y filas relevantes para nuestro análisis y realizamos el análisis de los datos, buscando relaciones entre variables y aplicando diferentes visualizaciones como Mapas de Calor, Matrices de Correlación, Gráficos de Torta, etc, lo que nos permitió entender y observar patrones.

### Etapa 3: Modelado y Extracción de Conocimiento

En la tercera y última etapa de nuestro proyecto, nos sumergimos en el proceso de modelado de datos según los requerimientos previamente establecidos. Utilizando el dataset objetivo y el análisis de datos realizado, seleccionamos los algoritmos adecuados para la extracción de patrones y definimos los valores de los parámetros correspondientes a cada técnica. En esta etapa utilizamos dos modelos: Random Forest y XGBoost.

Recopilamos los resultados y, finalmente, nos adentramos en la etapa de interpretación del conocimiento obtenido, analizando los resultados de los distintos modelos.

A través de estas tres etapas, nuestro trabajo tiene como objetivo proporcionar una proyección del precio de los alquileres de Airbnb en Buenos Aires y Nueva York, e interpretar los resultados del análisis de estos dos Datasets.

## Presentación de los Datasets

El dataset elegido es un web scraping de Airbnb y pertenece a los alojamientos disponibles en la aplicación correspondientes a la ciudad de Buenos Aires, Argentina, en el mes de Agosto de 2023. Originalmente 75 columnas y 26204 filas. Para el estudio, hicimos una limpieza del mismo dejando las siguientes columnas:

- last\_scraped: La fecha en la que los datos fueron extraídos por última vez del sitio web.
- host\_response\_rate: La tasa de respuesta del anfitrión, que indica qué tan rápido responde un anfitrión a las consultas de los huéspedes.
- host\_acceptance\_rate: La tasa de aceptación del anfitrión, que muestra la proporción de solicitudes de reserva que un anfitrión ha aceptado.
- host\_identity\_verified: Indica si la identidad del anfitrión ha sido verificada por la plataforma.
- \*neighbourhood\_cleansed: El nombre del barrio o vecindario en el que se encuentra la propiedad, después de haber sido limpiado.
- property\_type: El tipo de propiedad, como apartamento, casa, cabaña, etc.
- room\_type: El tipo de habitación o alojamiento que se ofrece, como habitación privada, casa/apartamento completo, habitación compartida, etc.
- accommodates: El número máximo de personas que la propiedad puede alojar.
- bathrooms\_text: El número de baños que posee la propiedad.
- bedrooms: El número de dormitorios en la propiedad.
- beds: El número total de camas disponibles en la propiedad.
- price: El precio de alquiler por noche de la propiedad.
- calendar\_last\_scraped: La fecha en la que se raspó por última vez el calendario de disponibilidad de la propiedad.
- number\_of\_reviews: El número total de reseñas que ha recibido la propiedad.
- number\_of\_reviews\_ltm: El número de reseñas que la propiedad ha recibido en el último mes.
- first\_review: La fecha de la primera reseña realizada por un huésped.
- last\_review: La fecha de la última reseña realizada por un huésped.
- review\_scores\_rating: La puntuación promedio de todas las reseñas de la propiedad, en una escala de 0 a 5.
- reviews\_per\_month: El promedio de reseñas que la propiedad recibe cada mes.
- Amenities\_Score: Una métrica o puntuación (en una escala de 1 a 5) relacionada con las comodidades o servicios que ofrece la propiedad, calculada en función de las comodidades enumeradas en el anuncio y teniendo en cuenta aquellas que son mas importantes para los huespedes.

## Captura del dataset:

<u>id</u>	<u>listing_url</u>	<u>scrape_id</u>	<u>last_scraped</u>	<u>source</u>	<u>name</u>	<u>description</u>	<u>neighborhood_picture</u>	<u>host_id</u>	<u>host_url</u>	<u>host_name</u>	<u>host_since</u>	<u>host_location</u>	<u>host_about</u>	<u>host_response_time</u>	<u>host_response_rate</u>	<u>host_acceptance_rate</u>	<u>host_is_superhost</u>	<u>host_thumbnail_url</u>	<u>host_picture_url</u>	<u>host_neighborhood</u>	<u>host_listings_url</u>	<u>host_total</u>	<u>host_verifications</u>
11508	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Condo in Bue LUXURIOUS 3 ARATE PALERRH picture https://a0.mi	42762	https://www.Candela	10/1/2009	New York, NY	within an ho	100%	79% t	https://a0.mi/https://a0.mi/Palermo	1	2	[email], phot								
14212	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in Beautiful cos Palermo is sihttps://a0.mi	8710233	https://www.Mar-a	8/3/2016	Buenos Aires SoY	within an ho	100%	100% t	https://a0.mi/https://a0.mi/muscache.com/	9	11	[email], phot								
20894	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in Recientemente renombrado https://a0.mi	1008896	https://www.Victoria	8/2/2011	Buenos Aires SoY	within an ho	100%	89% t	https://a0.mi/https://a0.mi/Palermo	15	15	[email], phot								
210045	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Condo in Bue Moderno y llo https://a0.mi	319890	https://www.RafaelOviedo	4/9/2020	Buenos Aires SoY	Althought a few days or	100%	99% f	https://a0.mi/https://a0.mi/Recoleta	25	30	[email], phot								
92228	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>The space! El barrio es https://a0.mi	494027	https://www.MirtaSusani	4/9/2011	Buenos Aires SoY uns pers N/A	within an ho	100%	0% f	https://a0.mi/https://a0.mi/Recoleta	2	2	[email], phot								
15074	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	previous	Rental unit in >B>The space!</b>> />C https://a0.mi	59538	https://www.Monica	12/7/2009	N/A	N/A	100%	https://a0.mi/https://a0.mi/muscache.com/	2	2	[email], f									
218783	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Loft in Bueno Nice apartment https://a0.mi	1613880	https://www.IvanMariano	7/8/2010	Buenos Aires Hol	within an ho	100%	100% t	https://a0.mi/https://a0.mi/Palermo	16	19	[email], phot								
219505	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Loft in Bueno Nice apartment https://a0.mi	1613880	https://www.IvanMariano	7/8/2010	Buenos Aires Hol	within an ho	100%	100% t	https://a0.mi/https://a0.mi/Palermo	10	11	[email], phot								
93174	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>equipoOur 2 appear https://a0.mi	364441	https://www.Robert	9/5/2009	New York, Un k	a few	100%	100% f	https://a0.mi/https://a0.mi/Recoleta	2	2	[email], phot								
20062	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>Estamos en la zone es cc https://a0.mi	500210	https://www.Luciano	4/11/2011	Buenos Aires Cielo Son	within an ho	100%	81%	https://a0.mi/https://a0.mi/Balvare	5	5	[email], phot								
96257	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>Very sunny https://a0.mi	75891	https://www.Sergio.Chim	1/3/2010	Buenos Aires Seg	within an ho	100%	95% t	https://a0.mi/https://a0.mi/Palermo.Hol	7	7	[email], phot								
20429	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	previous	Rental unit in >B>the space!</b>> />C https://a0.mi	580817	https://www.Juan	1/1/2010	Buenos Aires Hol	within a few days or	100%	0% f	https://a0.mi/https://a0.mi/Retiro	1	1	[email], phot								
229571	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	77065	https://www.Marcelo	4/16/2011	Buenos Aires So Marce	a few days or	100%	0% f	https://a0.mi/https://a0.mi/Construc&Ay	1	1	[email], phot								
235567	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	320283	https://www.Hortensia	12/16/2010	Buenos Aires H	within a few days or	100%	100% t	https://a0.mi/https://a0.mi/Retiro	4	4	[email], phot								
238513	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1105513	https://www.Luciano.Ol	9/3/2012	Buenos Aires A	within an ho	100%	99% t	https://a0.mi/https://a0.mi/Burzaco	8	13	[email], phot								
98460	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>IDEA PARAT'E es muy c&lt;mi https://a0.mi	1613880	https://www.Mar-a	7/8/2010	Buenos Aires Hol	within an ho	100%	100% t	https://a0.mi/https://a0.mi/muscache.com/	9	11	[email], phot								
24713	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Cristina	4/15/2011	Buenos Aires SoY	within a day	50%	99% t	https://a0.mi/https://a0.mi/Palermo	2	2	[email], phot								
98812	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Cristina	4/15/2011	Buenos Aires SoY	within an ho	100%	100% t	https://a0.mi/https://a0.mi/San Nicolis	7	12	[email], phot								
97043	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	508961	https://www.Maria	3/2/2010	Buenos Aires Hol	within a day	100%	89% t	https://a0.mi/https://a0.mi/Balvare	2	5	[email], phot								
238619	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>it's called C! The furnishes! https://a0.mi	1542049	https://www.Ceferina.D	4/16/2011	Buenos Aires Hol	a few days or	100%	67% f	https://a0.mi/https://a0.mi/San Nicolis	2	2	[email], phot								
24763	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1192725	https://www.Claudio	9/12/2011	Buenos Aires H	within a day	100%	100% t	https://a0.mi/https://a0.mi/Recoleta	1	1	[email], phot								
24764	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Carol	5/3/2010	Milano, It - We	within a day	100%	100% t	https://a0.mi/https://a0.mi/Palermo	1	1	[email], phot								
24765	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	500261	https://www.Carlos	1/2/2010	Buenos Aires M	within a day	100%	98% t	https://a0.mi/https://a0.mi/Palermo	24	31	[email], phot								
24766	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	111013	https://www.Daniel	4/19/2010	Buenos Aires Owner of thi	within a day	100%	0% f	https://a0.mi/https://a0.mi/Balvare	1	2	[email], phot								
239853	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	157968	https://www.Emmanuel	10/6/2011	Buenos Aires I am happy	within an ho	100%	100% t	https://a0.mi/https://a0.mi/Palermo	1	1	[email], phot								
99362	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Fabian	3/1/2011	Buenos Aires Hol So Y	within an ho	100%	99% t	https://a0.mi/https://a0.mi/Recoleta	26	44	[email], phot								
238244	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Fabian	3/1/2011	Buenos Aires A	within a day	100%	99% t	https://a0.mi/https://a0.mi/Recoleta	2	2	[email], phot								
100080	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Fabian	3/1/2011	Buenos Aires A	within a day	100%	99% t	https://a0.mi/https://a0.mi/Palermo	16	108	[email], phot								
104843	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	previous	Rental unit in >B>Apartment Centro de la https://a0.mi	562455	https://www.Laura	5/6/2010	Buenos Aires I'm an	a few days or	100%	83% t	https://a0.mi/https://a0.mi/Palermo	4	4	[email], phot								
243371	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Laura	4/12/2011	Buenos Aires Vivo en	within an ho	100%	100% f	https://a0.mi/https://a0.mi/Puerto Madero	4	4	[email], phot								
24990	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Laura	4/12/2011	Buenos Aires Cen	a few days or	100%	57% t	https://a0.mi/https://a0.mi/muscache.com/	2	2	[email], phot								
106475	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Laura	8/1/2011	Buenos Aires Hol	within a day	100%	75% f	https://a0.mi/https://a0.mi/Palermo	2	2	[email], phot								
245895	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	50261	https://www.Laura	4/12/2011	Buenos Aires Hoy My	within an ho	100%	98% t	https://a0.mi/https://a0.mi/Palermo	31	32	[email], phot								
24767	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Laura	4/12/2011	Buenos Aires SoY	within an ho	100%	100% t	https://a0.mi/https://a0.mi/Recoleta	1	5	[email], phot								
24768	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Laura	4/12/2011	Buenos Aires Hol	within a day	100%	100% t	https://a0.mi/https://a0.mi/Recoleta	2	2	[email], phot								
250527	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>Moderno y llo es https://a0.mi	1311901	https://www.Lucas	10/19/2011	Buenos Aires H	This is lu within an ho	100%	100% f	https://a0.mi/https://a0.mi/Recoleta	3	3	[email], phot								
31514	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>The Duplex+ San Telmo is https://a0.mi	64880	https://www.Elio.Mari	12/1/2009	Buenos Aires I was	a few days or	0%	100% t	https://a0.mi/https://a0.mi/Monserrat	10	11	[email], phot								
107759	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>7 bedooms & 4 bathooms https://a0.mi	555693	https://www.Victoria	5/3/2011	Montevideo, Soy Vitoria, within a day.	within a day	100%	77% f	https://a0.mi/https://a0.mi/Balvare	2	2	[email], phot								
31920	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	previous	Rental unit in >B>the space!</b>> />C https://a0.mi	137378	https://www.Nancy	6/7/2010	N/A	N/A	100%	f	https://a0.mi/https://a0.mi/Belgrano	1	1	[email], f								
140888	<a href="https://www.airbnb.com/rooms/2038413">https://www.airbnb.com/rooms/2038413</a>	6/29/2023	city scrape	Rental unit in >B>the space!</b>> />C https://a0.mi	1613880	https://www.Maria	1/4/2011	Buenos Aires Thanks for	within an ho	100%	f	https://a0.mi/https://a0.mi/Retiro	4	4	[email], Retir								

Fuente: [Inside Airbnb: Get the Data](https://www.insideairbnb.com/get-the-data.html)

Nota: Cabe aclarar que se utilizó un dataset de estructura idéntica, pero correspondiente a la ciudad de Nueva York, EEUU, con el fin de validar la hipótesis.

## Justificación de la elección del dataset:

- **Relevancia:** Elegimos este dataset debido a su relevancia en el contexto actual. El mercado de alquiler a corto plazo en Buenos Aires ha experimentado un rápido crecimiento, y entender los patrones y tendencias en Airbnb puede ser útil para los anfitriones, los turistas y las autoridades locales.
- **Disponibilidad:** El dataset de Airbnb en Buenos Aires es accesible y se actualiza regularmente, lo que facilita la obtención de datos actualizados y confiables.
- **Interés académico:** Nos interesa profundizar en el análisis de datos y aprender más sobre cómo los factores como la ubicación, las características de las propiedades y las revisiones de los huéspedes pueden influir en los precios y la satisfacción.

## Visualización y Exploración de Datos

### **Contexto Inicial:**

En el transcurso del Trabajo Práctico Integrador de Ciencia de Datos, investigamos a través de los datos de Airbnb, específicamente en las ciudades de Buenos Aires y Nueva York. La fase inicial se centró en la visualización y exploración de datos para desentrañar patrones y peculiaridades que podrían influir en la predicción de precios de alquiler.

### **Herramientas Visuales:**

#### **Mapas Interactivos:**

Optamos por utilizar mapas interactivos para mapear la distribución geográfica de los alquileres en ambas ciudades. Esto no solo proporcionó una visión visualmente impactante, sino que también reveló tendencias espaciales y posibles disparidades.

#### **Gráficos de Tendencias Temporales:**

La aplicación de gráficos temporales nos permitió identificar patrones estacionales y variaciones en la demanda de alquileres a lo largo del tiempo. Estos gráficos desempeñaron un papel crucial en entender las dinámicas cambiantes del mercado.

### **Descubrimientos Visuales:**

#### **1. Disparidades Geográficas:**

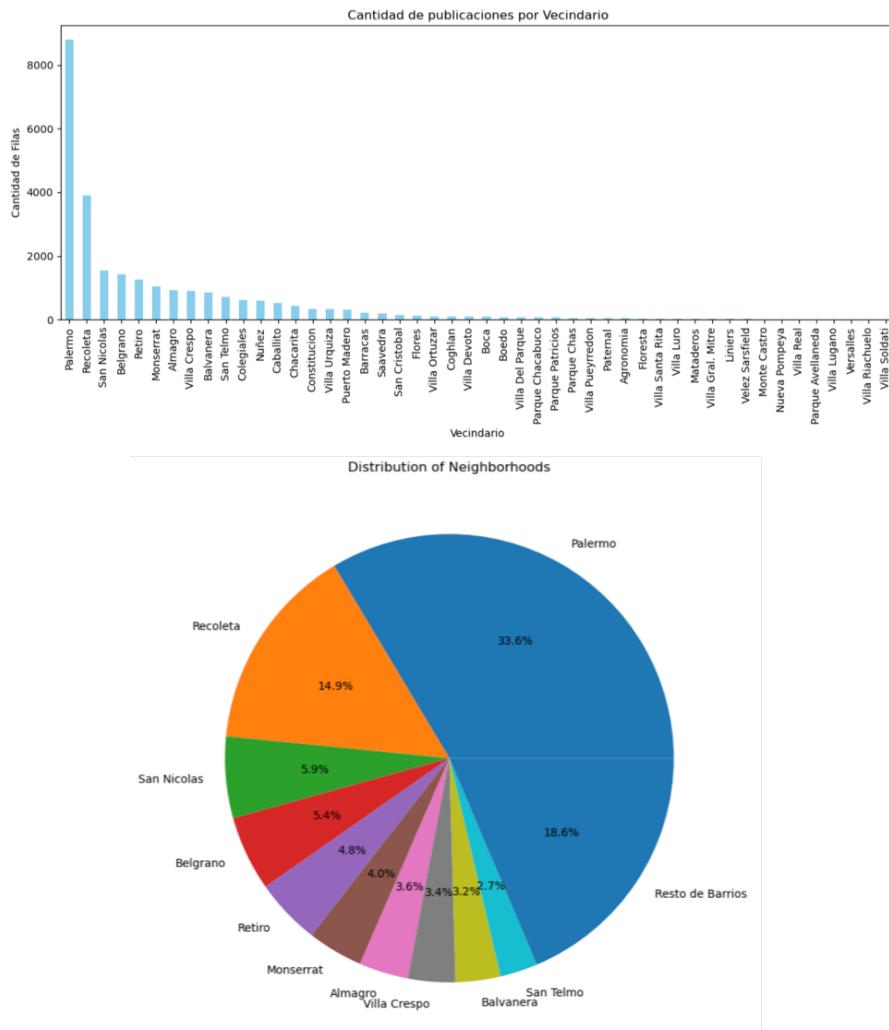
- Observamos concentraciones significativas de alquileres en ciertas áreas de ambas ciudades, destacando la importancia de la ubicación en los precios.

#### **2. Patrones Temporales:**

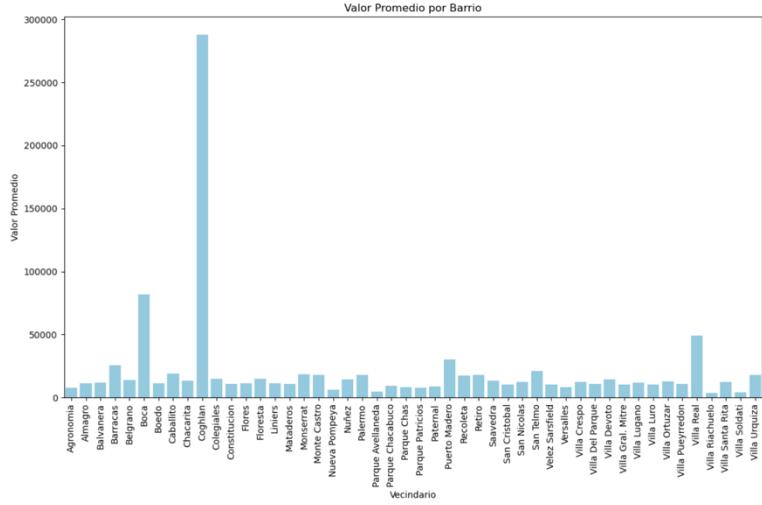
- Identificamos estacionalidades en la demanda de alquileres, con picos durante ciertos meses. Esto sugiere la necesidad de considerar factores temporales en nuestras predicciones.

### Algunos de los gráficos utilizados:

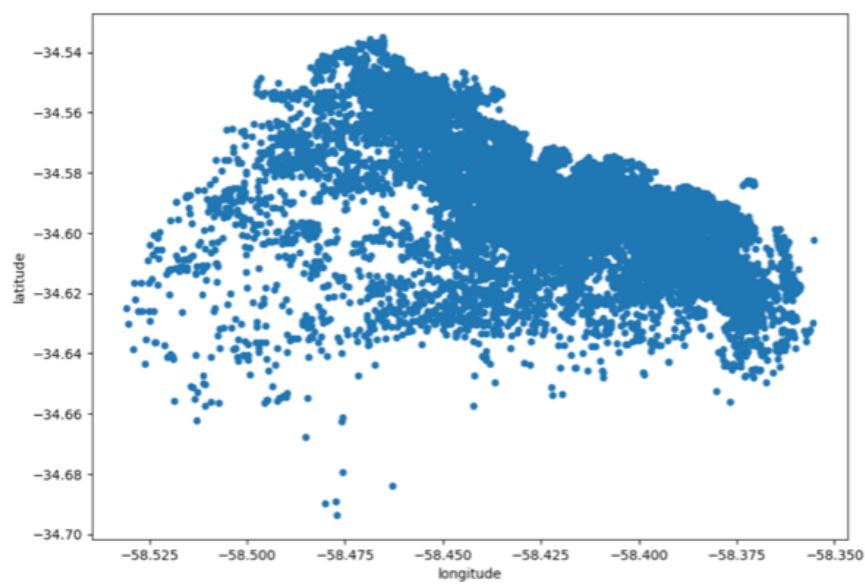
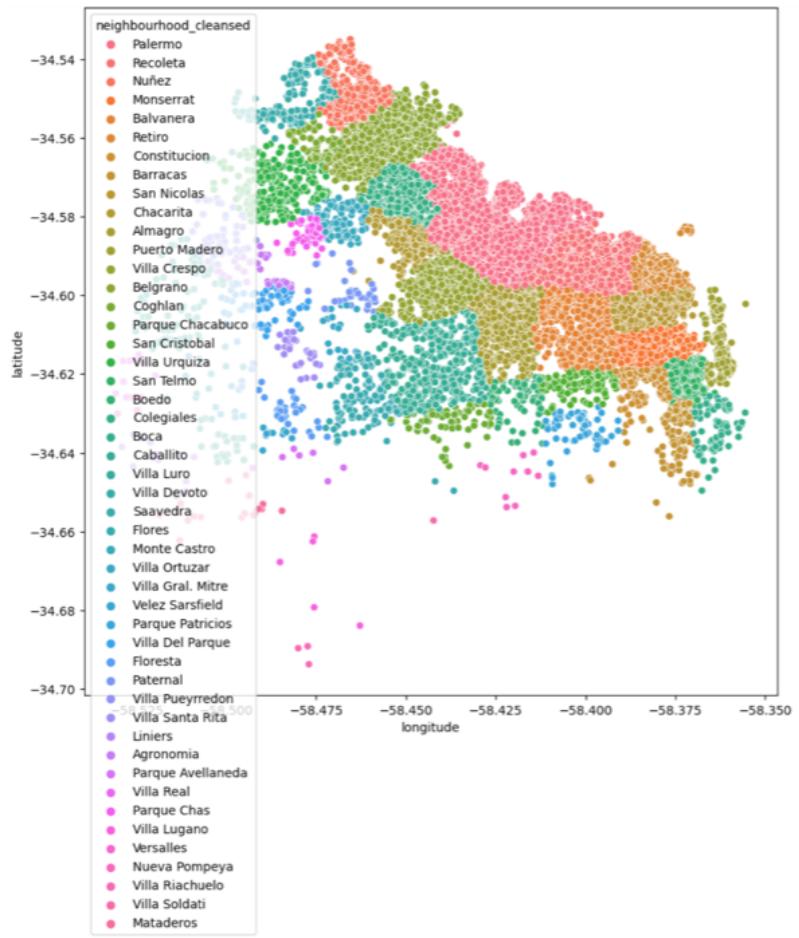
Cantidad de publicaciones por barrio:



Valor promedio por cada barrio:



Mapa de Bs As con la distribución de publicaciones por barrio hecho con la latitud y longitud:



## **Planteo de la Hipótesis**

### **Hipótesis de Predicción de Precios en Buenos Aires:**

Dada la dificultad de predecir precios de alquiler en Buenos Aires debido a la inflación, planteamos la hipótesis de que los modelos entrenados con datos de Nueva York podrían proporcionar una referencia más sólida. Nuestra estrategia fue aplicar los mismos modelos a un dataset de Nueva York para verificar su desempeño.

### **Razonamiento:**

#### **1. Estabilidad Económica en Nueva York:**

- La estabilidad económica en Nueva York proporciona un terreno más consistente para entrenar modelos, minimizando la influencia de factores económicos volátiles.

#### **2. Inflación en Buenos Aires:**

- La inflación en Buenos Aires introduce variabilidad en los precios de alquiler, dificultando la captura de patrones claros. Suponemos que esto podría afectar negativamente la capacidad de generalización de nuestros modelos.

### **Resultados y Reflexiones:**

Los modelos aplicados a Nueva York demostraron un rendimiento sólido. Sin embargo, la disparidad de precios en Buenos Aires debido a la situación inflacionaria sigue siendo un desafío. Esto destaca la complejidad de prever los precios de alquiler en un entorno económico variable y subraya la importancia de considerar múltiples factores en futuros enfoques predictivos.

## Extracción, transformación y carga de datos

El proceso ETL se centró en optimizar el conjunto de datos original, que constaba de 75 columnas, para seleccionar solo las 30 columnas más relevantes para la predicción, descartando columnas inutilizables como la descripción del departamento o el nombre del hospedador. La primera fase del proceso ETL implicó la extracción de datos del conjunto de datos original de Airbnb. Este conjunto de datos contenía información diversa sobre propiedades de alquiler, incluyendo detalles sobre las comodidades, ubicación y características de las propiedades.

La transformación de datos fue una parte crucial del proceso ETL. A continuación, se detallan algunas de las transformaciones clave que se llevaron a cabo:

1. **Manejo de Valores Nulos:** El conjunto de datos original no contenía valores nulos, lo que simplificó este aspecto del proceso. No fue necesario realizar imputaciones de datos faltantes.
2. **Creación de la Columna "amenities\_score":** Se identificó un grupo de columnas relacionadas con las comodidades ofrecidas en las propiedades de alquiler. Estas columnas incluían "Free street parking", "Elevator", "Essentials", "Hot water", y otras. Para simplificar el análisis, se consolidaron estas comodidades en una sola columna llamada "amenities\_score". Esta columna asignó una puntuación de 1 a 5 estrellas en función de la proporción de comodidades disponibles en cada propiedad.
3. **Agrupación de Barrios en Zonas:** Se creó una nueva columna que agrupaba los barrios en cinco zonas principales: Noreste, Sureste, Noroeste, Suroeste y Otro. Esta agrupación facilitó la consideración de la ubicación en el análisis de precios.
4. **Codificación One-Hot de la Columna "room\_type":** La columna categórica "room\_type" se transformó utilizando codificación one-hot. Esta técnica creó tres nuevas columnas: "Hotel room", "Private room" y "Shared room". Cada columna indicó la presencia o ausencia de un tipo de habitación específico.
5. **Transformación Yeo-Johnson:** Para otras variables categóricas, se aplicó la transformación Yeo-Johnson. Esta transformación es una técnica que ajusta la distribución de datos para que se asemeje más a una distribución normal. Ayudó a mejorar la calidad de los datos y facilitó su uso en modelos de predicción.

Una vez que se completaron todas las transformaciones, los datos se dividieron en conjuntos de entrenamiento y prueba en una proporción del 70% y 30%, respectivamente. Esta división permitió preparar los datos para su uso en modelos de predicción, asegurando que se dispusiera de un conjunto de prueba independiente para evaluar la precisión de los modelos.

## Aplicación de los modelos

### **Explicación de los modelos:**

**Regresión Lineal Múltiple** → La regresión lineal múltiple es una técnica estadística que se utiliza para entender y modelar la relación entre una variable dependiente (la que se desea predecir) y múltiples variables independientes (predictoras). En esencia, busca encontrar una relación lineal que mejor se ajuste a los datos.

En lugar de una sola variable independiente, como en la regresión lineal simple, se trabajan con múltiples variables independientes en la regresión lineal múltiple. El objetivo es determinar cómo estas variables independientes influyen en la variable dependiente y en qué medida.

Además, utilizamos dos modelos basados en “Árbol de Decisión”: “Random Forest” y “XGBOOST”. Ambos son algoritmos supervisados utilizados tanto para clasificación como para regresión.

**Random Forest** → es un algoritmo que trae la solución a uno de los más grandes problemas de los árboles de decisión, el sesgo de selección de características.

El sesgo de selección de características es una tendencia que presentan los árboles de decisión a seleccionar siempre las mismas características dominantes (que suelen ser las que proporcionan una ganancia de información alta en el inicio del modelo) cada vez que se hace la división de un nodo.

Para solucionar este problema, lo que propone Random Forest es crear muchos Árboles de Decisión que tengan distintos subespacios de características aleatorios. Esto permite que en algunos de los árboles creados se descarten las características dominantes y puedan ser seleccionadas otras características que antes eran discriminadas y que, posiblemente, podían dar una contribución a la predicción.

Lo siguiente que hace Random Forest es hacer que cada árbol generado clasifique a los vectores de entrenamiento para luego realizar una votación (ponderada a partir del grado de pertenencia que asignó cada árbol) para determinar la pertenencia del vector de entrada a alguno de los conjuntos de salida.

**XGBOOST** → es otro modelo basado en Árbol de Decisión que trae consigo mejoras respecto al Random Forest.

En Random Forest cada árbol creado se entrena de forma independiente. En cambio, en XGBOOST los árboles se entranan de forma secuencial intentando corregir los errores cometidos por el árbol anterior. Para esto, el árbol que está siendo entrenado en un momento dado, le presta especial atención a los vectores que el árbol anterior peor

clasificó. Para lograr esto cuenta con dos reguladores de coeficientes de las características, L1 y L2.

- L1 → analiza los coeficientes de acuerdo al error absoluto que aporten y para valores altos los penaliza haciéndolos tender a cero. Se puede regular con el hiperparámetro "alpha".
- L2 → analiza los coeficientes de acuerdo al error cuadrático que aporten y para valores altos los penaliza haciéndolos tender a valores bajos (pero no nulos). Se puede regular con el hiperparámetro "lambda".

Las regularizaciones hechas por L1 y L2 penaliza las particularidades del subconjunto de datos de entrenamiento, haciendo que estas no sobreajusten el modelo. Otra mejora que tiene respecto a Random Forest es que, al guiar el entrenamiento de un árbol a partir de información provista por los árboles anteriores, el entrenamiento del modelo converge más rápido y su clasificación suele ser más precisa.

Ventajas de XGBOOST respecto a Random Forest:

1. **XGBoost inmediatamente poda el árbol con una puntuación llamada "Puntuación de similitud" antes de entrar en los propósitos de modelado reales.** Considera la "ganancia" de un nodo como la diferencia entre la puntuación de similitud del nodo y la puntuación de similitud de los niños. Si se considera que la ganancia de un nodo es mínima, entonces simplemente deja de construir el árbol a una mayor profundidad, lo que puede superar en gran medida el desafío del sobreajuste. Mientras tanto, el bosque aleatorio probablemente podría sobreajustar los datos si a la mayoría de los árboles del bosque se les proporcionan muestras similares. Si los árboles han crecido completamente, el modelo colapsará una vez que se introduzcan los datos de prueba. Por lo tanto, se presta especial atención a distribuir todas las unidades elementales de la muestra con una participación aproximadamente igual a todos los árboles.
2. **XGBoost es una buena opción para conjuntos de datos desequilibrados, pero no podemos confiar en los bosques aleatorios en este tipo de casos.** En aplicaciones como falsificación o detección de fraude, es casi seguro que las clases estarán desequilibradas y el número de transacciones auténticas será enorme en comparación con las transacciones no auténticas. En XGBoost, cuando el modelo no logra predecir la anomalía por primera vez, le otorga más preferencias y ponderación en las próximas iteraciones, aumentando así su capacidad para predecir la clase con baja participación; pero no podemos asegurar que random forest trate el desequilibrio de clases con un proceso adecuado.
3. **Una de las diferencias más importantes entre XG Boost y Random forest es que XGBoost siempre le da más importancia al espacio funcional a la hora de reducir**

**el coste de un modelo mientras que Random Forest intenta dar más preferencias a los hiperparámetros para optimizar el modelo.** Un pequeño cambio en el hiperparámetro afectará a casi todos los árboles del bosque, lo que puede alterar la predicción. Además, este no es un buen enfoque cuando esperamos datos de prueba con tantas variaciones en tiempo real con una mentalidad predefinida de hiperparámetros para todo el bosque, pero los hiperparámetros de impulso XG se aplican a un solo árbol al principio, lo que se espera que ajustarse de manera eficiente cuando progresan las iteraciones. Además, XGBoost solo necesita una cantidad muy baja de hiperparámetros iniciales (parámetro de contracción, profundidad del árbol, cantidad de árboles) en comparación con el bosque aleatorio.

4. **Cuando el modelo se encuentra con una variable categórica** con un número diferente de clases, existe la posibilidad de que el bosque aleatorio dé más preferencias a la clase con más participación.
5. **XGBoost puede ser más preferible en situaciones** como la regresión de Poisson, la regresión de rango, etc. Esto se debe a que los árboles se derivan optimizando una función objetivo.

Ventajas de Random Forest respecto a XGBOOST:

1. Los bosques aleatorios son **más fáciles de ajustar** que los algoritmos de impulso.
2. Los bosques aleatorios **se adaptan más fácilmente a la computación distribuida** que los algoritmos Boosting.
3. Es casi seguro que los bosques aleatorios no se sobreajustarán si **los datos se preprocesan y limpian cuidadosamente**, a menos que se proporcionen repetidamente muestras similares a la mayoría de los árboles.

**Resultados obtenidos:**

**Regresión lineal multiple:**

Mean Squared Error: 0.1586896017261502

R2 Score: 0.38058984859526923

Mean Absolute Error: 0.11110662201500138

**Random Forest:**

RMSE en los datos de prueba: 0.1409805843451212

Coeficiente de determinación ( $R^2$ ) en los datos de prueba: 0.511122566259977

Coeficiente de determinación ( $R^2$ ) en los datos de entrenamiento: 0.9241367076369992

**XGBOOST:**

RMSE en los datos de prueba (XGBoost): 0.14156940294767362

Coeficiente de determinación ( $R^2$ ) en los datos de prueba (XGBoost): 0.507030353618454

Coeficiente de determinación ( $R^2$ ) en los datos de entrenamiento (XGBoost): 0.7626993630023169

## Conclusión

A partir del desempeño de los modelos aplicados respectos al dataset empezaron a surgir hipótesis nuevas sobre las supuestas causas que limitan el rendimiento de los modelos, inicialmente se supuso que se debía a cuestiones referidas a los procesos previos al empleo de los modelos, por este motivo se procedió inicialmente a probar diferentes alternativas correspondientes a la implementación de estos.

EJ:( Probar features diferentes, aplicar técnicas de imputación variadas, eliminación de valores atípicos, ir probar codificaciones distintas, normalización de valores )

Como resultado de seguir esto alcanzamos un incremento en el desempeño, sin embargo seguía encontrándose en niveles no tan altos. Fue en este punto donde se nos planteó la idea de que el rendimiento del modelo podría estar ligado a factores externos al desarrollo del trabajo EJ: (**La vivienda se ha convertido en un activo mega especulativo**)

Nuestro dataset de análisis primero que nada apunta a la predicción de algo que maneja demasiada volatilidad como es el valor de un alquiler (**encima temporalio**), ya que este mismo no sigue una regla que fije dicha variable, no hay nada que reglamente que bajo ciertas características un alquiler debe de tomar un valor determinado, al punto que uno puede encontrarse con casos en los cuales dos departamentos de un mismo edificio e iguales características manejan precios muy diferentes, esto explica el bajo nivel de correlación entre las variables

host_response_rate	-0.031616
host_acceptance_rate	-0.026599
host_total_listings_count	0.033613
host_identity_verified	-0.032885
latitude	0.005681
longitude	-0.000893
accommodates	0.039326
bathrooms_text	0.037506
bedrooms	0.037177
beds	0.028179
price	1.000000
minimum_nights	0.009498
maximum_nights	0.004619
minimum_minimum_nights	0.008120
maximum_minimum_nights	0.008479
number_of_reviews	-0.011743
number_of_reviews_ltm	-0.019078
review_scores_rating	-0.022242
calculated_host_listings_count	-0.002582
reviews_per_month	-0.023042
★ Amenities_Score	0.012032
Name: price, dtype: float64	

## El contexto macroeconómico del lugar de análisis

Por qué Airbnb puede ser un gran problema para los porteños que no son propietarios

20/12/19



### Airbnb y la crisis de los alquileres

Por el encarecimiento de las propiedades, varias ciudades del mundo buscan regular la modalidad de contratos temporarios, que en Argentina representan una porción cada vez mayor.

Iván Hirsch [Seguir](#)



La oferta de alquileres temporales creció 65% en un año.

### 2023, un año difícil

Hay cuatro elementos que condicionarán la economía del año próximo: el contexto internacional desfavorable; la sequía; la escasez de dólares, y el proceso electoral.



La información brindada por el dataset es de agosto de 2023 en Buenos Aires Argentina, a partir de esto es de gran importancia analizar el contexto macroeconómico para sacar conclusiones por encima de las que pueda brindarnos la parte técnica del trabajo realizado, si ya de por si el mercado inmobiliario resulta ser super especulativo habría que añadirle sus condiciones de análisis ligadas a una macroeconomía super inestable, Argentina en este contexto se encontraba atravesando una altísima inflación, devaluaciones recientes, regulaciones sobre el mercado entre otras cosas. Este conjunto de factores al poder influir muy negativamente en las ganancias termina alejando a la especulación por parte del mercado.

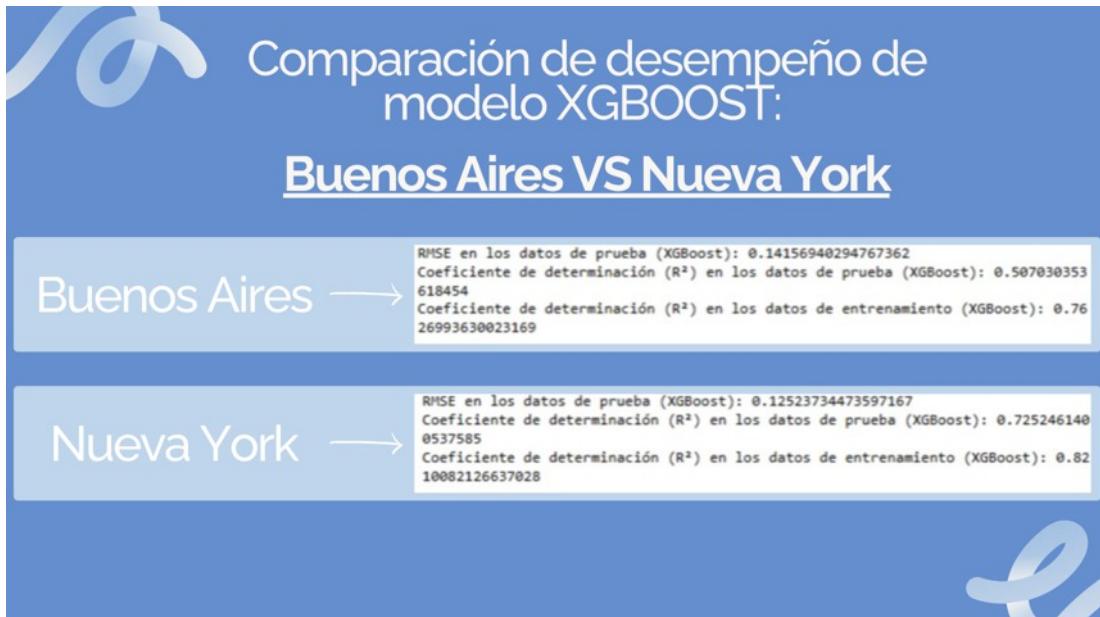
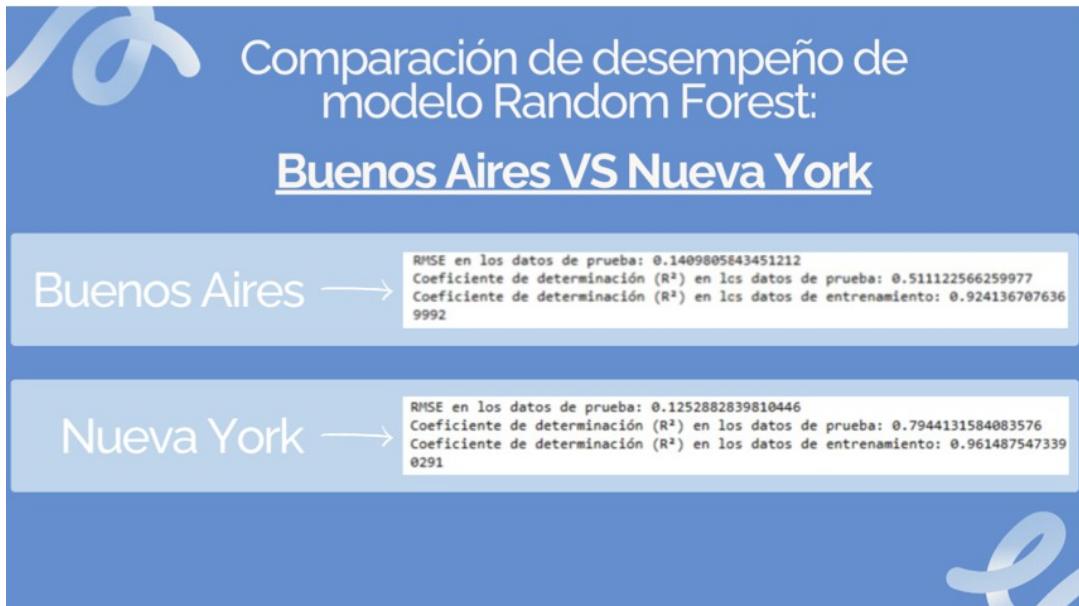
Para dar respuesta a esto recurrimos a la idea de realizar el mismo procedimiento de análisis sobre un dataset de las mismas características pero sobre una economía estable, procedimos a hacer uso de los modelo en un dataset de nueva york y las mejoras en los modelos resultaron evidentes.

Podemos observar mayores niveles de correlación entre las variables:

host_total_listings_count	0.087345
neighbourhood_cleansed	0.069787
latitude	0.137661
longitude	0.032312
property_type	-0.267863
accommodates	0.484316
bathrooms_text	0.296186
bedrooms	0.427834
beds	0.303148
price	1.000000
minimum_nights	-0.012600
maximum_nights	0.024049
minimum_minimum_nights	-0.021697
maximum_minimum_nights	-0.005767
number_of_reviews	-0.057640
review_scores_rating	-0.058996
calculated_host_listings_count	0.071469
reviews_per_month	-0.088324
★ Amenities_Score	0.122643
neighbourhood_zone	-0.169751
Hotel room	-0.027246
Private room	-0.271532
Shared room	-0.133996

Name: price, dtype: float64

### Comparación desempeño de modelos:



Si bien podemos observar una importante mejoría, aún sigue manejando niveles alejados a una precisión casi perfecta, este fenómeno puede explicarse en función al primer punto resaltado “**La vivienda se ha convertido en un commodity**”, a tal punto que el mismo lugar de ejemplo que tomamos en estas mismas fechas impuso normativas de regulación al respecto.

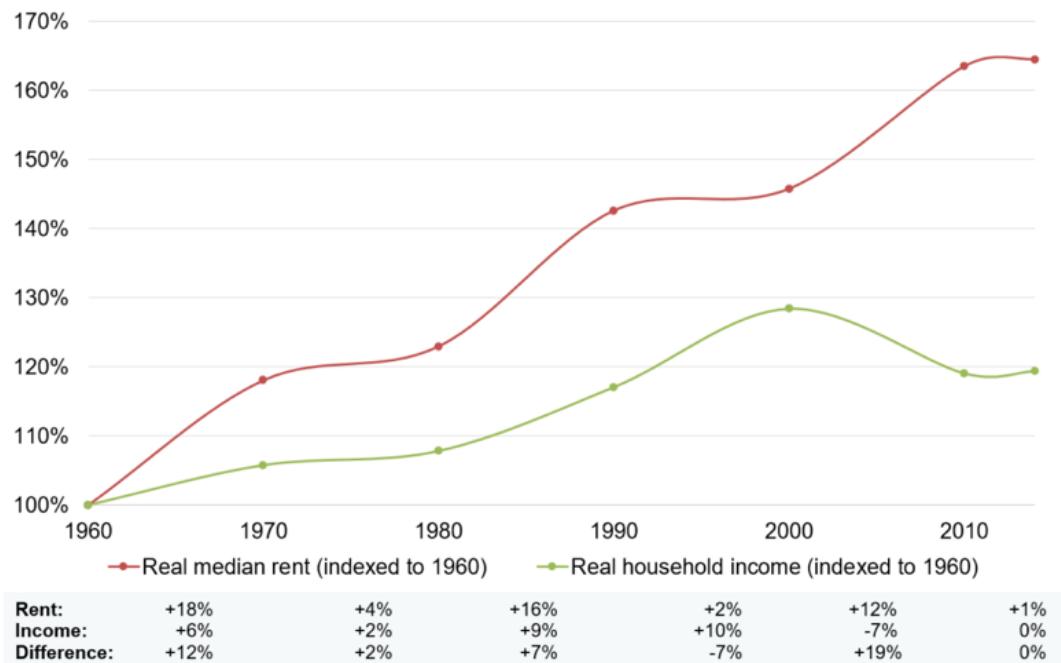
## Normativa impuesta Airbnb

Golpe de Nueva York a Airbnb y alquileres temporarios: el proyecto local para regularlos y la ley que no se cumple

- El Senado ya discute una iniciativa para ponerles límites.
- Qué dice la ley porteña que se aprobó en 2019.



Median rents vs. median household income, 1960 - 2014

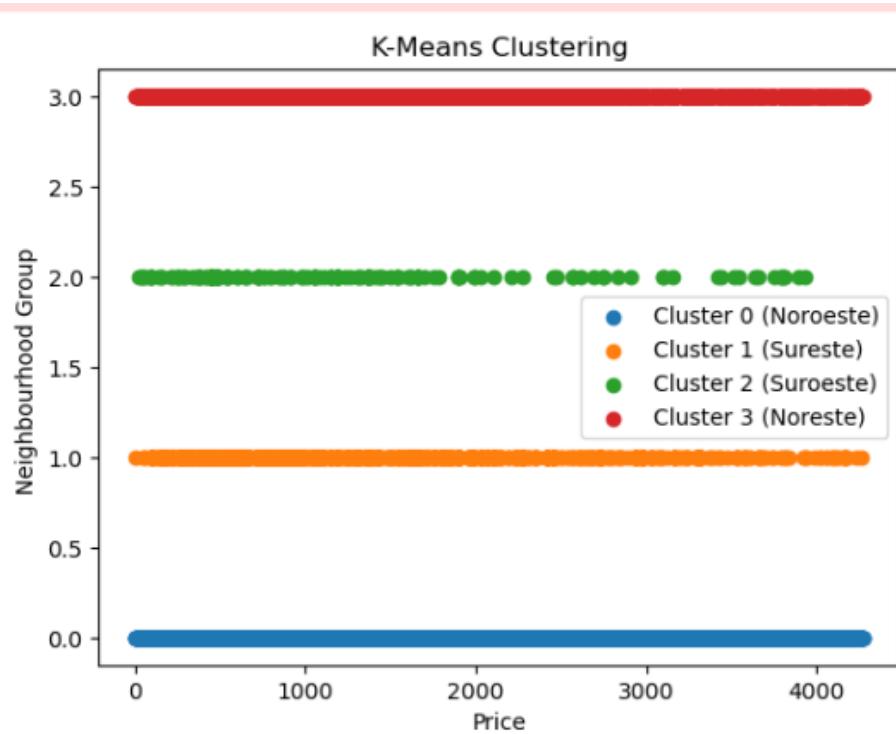
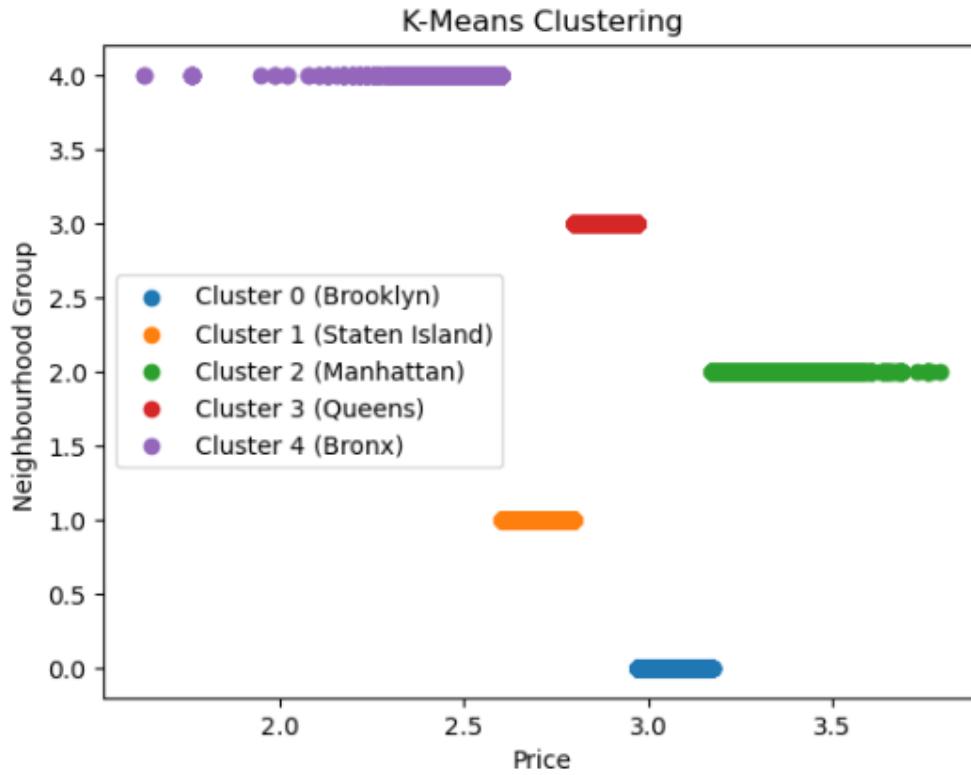


Note: Chart only includes data for 1960, 1970, 1980, 1990, 2000, 2010, and 2014. Numbers may not sum due to rounding  
Source: 1960-2000 Decennial Censuses and 2010-14 American Community Surveys

apartment

Dejando de lado el punto de análisis del rendimiento de los modelos, pudimos responder en cierta manera a la hipótesis referida a la existencia de una correlación representativa entre el nivel de precios según el barrio.

Para esto recurrimos a la creación de una nueva columna llamada zonas de barrios, en las cuales dividimos en 4( sur, norte, este y oeste).

**Buenos Aires:****Nueva York:**

Para concluir nuestro trabajo destacamos la importancia de la limpieza y preparación de datos, ya que la calidad de los datos es fundamental para lograr predicciones precisas. Hemos explorado y aplicado técnicas de ingeniería de características para identificar las variables más relevantes que influyen en los precios de alquiler. Además, hemos utilizado algoritmos de aprendizaje automático y técnicas de evaluación de modelos para crear un modelo de predicción robusto.

Durante el proceso, también hemos enfrentado desafíos, como la gestión de valores atípicos, la selección de variables adecuadas y la optimización de hiper parámetros. Sin embargo, estos desafíos nos han brindado la oportunidad de mejorar nuestras habilidades y adoptar un enfoque más riguroso en nuestro análisis.

En última instancia, hemos obtenido un modelo que puede proporcionar estimaciones razonablemente precisas de los precios de alquiler en función de las características específicas de las propiedades y del mercado local. Esto podría ser de gran utilidad tanto para los propietarios que desean establecer precios competitivos como para los inquilinos que buscan alojamiento asequible.

Nuestro trabajo no solo ha mejorado nuestras habilidades técnicas, sino que también nos ha permitido comprender la importancia de la colaboración y la comunicación en proyectos de análisis de datos. Cada miembro del equipo ha aportado su experiencia y conocimientos únicos, lo que ha enriquecido nuestro enfoque y resultados.

Destacamos también la importancia de la comprensión contextual de los datos con los que trabajaremos, hay cuestiones fundamentales que no pueden cuantificarse y por lo tanto debemos analizar desde una perspectiva más social para hacer del proceso de análisis de datos más eficiente.

Luego de realizado este trabajo práctico integrador de Ciencia de Datos sobre Airbnb podemos concluir que ha sido una experiencia valiosa y enriquecedora. Hemos adquirido habilidades técnicas, desarrollado un modelo de predicción útil y fortalecido nuestra capacidad para trabajar en equipo. Este trabajo es un testimonio de nuestro compromiso con la excelencia en el campo de las Ciencias de Datos y demuestra cómo podemos aplicar nuestras habilidades para abordar desafíos del mundo real con éxito.

## Bibliografía

- McKinney, W. (2023). Python para análisis de datos: Manipulación de datos con pandas, NumPy y Jupyter. O'Reilly.
- Bonacorso, G. (2018). Machine Learning Algorithms, Second Edition: Popular algorithms for data science and machine learning. Packt Publishing.
- <http://insideairbnb.com/get-the-data/>
- <https://www.lanacion.com.ar/opinion/dilema-urbano-airbnb-vs-crisis-habitacional-mas-regulacion-o-mejor-planificacion-nid09042023/>
- <https://www.redaccion.com.ar/los-alquileres-por-airbnb-se-salieron-de-control-como-es-la-situacion-en-buenos-aires-en-comparacion-a-otras-ciudades/>
- <https://es.wired.com/articulos/la-prohibicion-de-airbnb-en-nueva-york-esta-generando-un-mercado-negro>
- [https://www.clarin.com/internacional/golpe-nueva-york-airbnb-ciudad-pone-lmites-alquileres-temporarios\\_0\\_jXkFNNdOQj.html](https://www.clarin.com/internacional/golpe-nueva-york-airbnb-ciudad-pone-lmites-alquileres-temporarios_0_jXkFNNdOQj.html)
- <https://www.kranio.io/blog/serie-data-science-extraer-y-transformar-los-datos-2-de-3>
- <https://www.codificandobits.com/blog/analisis-exploratorio-de-datos/>
- <https://scikit-learn.org/stable/index.html>
- <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- <https://medium.com/geekculture/xgboost-versus-random-forest-898e42870f30>
- <https://xgboost.readthedocs.io/en/stable/>