Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Análsis de la oferta inmobiliaria

Airbnb CDMX

Proyecto Final

P R E S E N T A :

Hernández Castellanos David

México, CDMX, 2021

«Empezar una compañía es más arte que ciencia porque es un territorio desconocido. En lugar de querer resolver todos los problemas del mundo, trata de solucionar la situación que te sea más personal. Idealmente, si eres una persona normal y resuelves la situación que te incomoda a ti, habrás encontrado la respuesta para millones de personas.»

Brian Chesky, CEO de Airbnb

Índice general

1 Introducción

Con más de 8.5 millones de habitantes, la Ciudad de México representa un mercado atractivo para cualquier industria. Una ciudad tan turística y con tanto movimiento económico simboliza un atractivo importante para la industria hotelera; pero con el paso de los años el negocio de hospedaje ha ido evolucionando hasta lo que conocemos actualmente como la plataforma de hospedaje más grande en el mundo. Con un concepto que invita a sentirte en casa en cualquier parte del planeta donde puedas hospedarte, Airbnb ha conseguido establecerse en un mercado tan competitivo como lo representa la Ciudad de México, desde 2011 cuando inicio operaciones en la capital del país.

El éxito de la plataforma no es coincidencia y existen factores que determinan la demanda o la irregularidad en los negocios, el claro ejemplo se encuentra en la industria hotelera que se ha comenzado a ver superada por la gran cantidad de alojamientos con los que cuenta la plataforma y la diversidad de precios que la integran. La limitada capacidad del sector hotelero de añadir inmuebles a su oferta ha representado un reto para detener el crecimiento acelerado que ha tenido airbnb en los años recientes. Lo único que ha parado este crecimiento, desafortunadamente es la pandemia con la que nos enfrentamos en la actualidad, por ello, el análisis lo realizo tiempo antes del confinamiento en la Ciudad de México, de tal forma busco conocer el estado inicial y el comportamiento sobre un ambiente no estresado dejando para otro análisis el estrés de los datos y las propiedades que predominan en un ambiente tan conflictivo como lo representa el impacto de la enfermedad COVID-19.

Mi objetivo inicial es descubrir la características que definen un precio adecuado para una propiedad de acuerdo a sus características, gracias a esta información definir un estándar de calidad que permita incrementar el valor de las propiedas y el promedio de ventas por noche. Posteriormente me gustaría realizar una segmetación de las propiedades para definir las relaciones que se pueden establecer entre sus características para poder realizar agrupaciones, por ejemplo, las propiedades 'departamentos con 3 habitaciones con precio $ 1,000 por noche, ubicadas en condesa y reserva inmediata' tienen mayor numero de reservaciones, con esta información se puede definir un patrón de inversión que permita aumentar el número de rentas de ciertas propiedades o descubrir nuevas propiedades potenciales.

1.1 Dataset

Utilicé para el análisis una muestra de las propiedades publicadas en febrero del 2020 en la plataforma Airbnb gracias a la página Inside Airbnb. Se comenzó con un total de 21,663 registros y 106 variables que describen a los anfitriones, sus propiedades y contenidos.

1.1.1 Diccionario de datos - Etiquetado de variables

Se presentan las variables con las que se trabajaron y el etiquetado de variables correspondiente

Variable	Etiqueta	Descripción
id'	Llave	id de la propiedad anunciada
listing_url'	Llave	url de la propiedad
scrape_id'	Llave	id del corte de los datos
thumbnail_url'	Llave	url de vista previa
medium_url'	Llave	url de vista previa extendido
picture_url'	Llave	url de imagenes
xl_picture_url'	Llave	url de imagen extendida
host_id'	Llave	id del host
host_url'	Llave	url perfil del host
host_thumbnail_url'	Llave	url de vista previa perfil del host
host_picture_url'	Llave	url foto de host
license'	Llave	Licencia propiedad
'host_listings_count'	Continua	Publicaciones realizadas por el host
host_total_listings_count'	Continua	Total de publicaciones realizadas por el host
latitude'	Continua	Latitud
longitude'	Continua	Longitud
bathrooms'	Continua	Numero de baños
square_feet'	Continua	Tamaño de la propiedad en pies
price'	Continua	Precio en pesos por noche
weekly_price'	Continua	Precio por semana
monthly_price'	Continua	Precio por mes
security_deposit'	Continua	Deposito de seguridad
cleaning_fee'	Continua	Cuota de limpieza
extra_people'	Continua	Precio por persona extra
minimum_nights_avg_ntm'	Continua	Promedio de noches minimas
maximum_nights_avg_ntm'	Continua	Promedio de noches maximas
'host_response_rate'	Discreta	Rate sobre respuesta del host
host_acceptance_rate'	Discreta	Rate sobre tiempo de aprobacion del host
host_is_superhost'	Discreta	Clasificacion de superhost
host_has_profile_pic'	Discreta	Verificacion de foto de perfil host
host_identity_verified'	Discreta	Host verificado
zipcode'	Discreta	codigo postal
is_location_exact'	Discreta	Verificacion localizacion exacta
accommodates'	Discreta	Numero de huespedes
bedrooms'	Discreta	Numero de habitaciones
beds'	Discreta	Numero de camas
guests_included'	Discreta	Huespedes incluidos por precio
minimum_nights'	Discreta	Noches minimas
maximum_nights'	Discreta	Noches maximas
minimum_minimum_nights'	Discreta	Minimo de noches minimas
maximum_minimum_nights'	Discreta	Maximo de noches minimas
minimum_maximum_nights'	Discreta	Minimo de noches maximas
maximum_maximum_nights'	Discreta	Maximo de noches maximas
has_availability'	Discreta	Verificacion de disponibilidad
availability_30'	Discreta	Dias disponibles en los siguientes 30 dias
availability_60'	Discreta	Dias disponibles en los siguientes 60 dias
availability_90'	Discreta	Dias disponibles en los siguientes 90 dias
availability_365'	Discreta	Dias disponibles en los siguientes 365 dias
number_of_reviews'	Discreta	Numero de criticas
number_of_reviews_ltm'	Discreta	Numero de criticas en el ultimo mes

Figura 1.1: Variables.

Variable	Etiqueta	Descripción
reviews_per_month'	Discreta	Criticas por mes
review_scores_rating'	Discreta	rate de criticas para la propiedad
review_scores_accuracy'	Discreta	Rate sobre la veracidad de los datos
review_scores_cleanliness'	Discreta	Rate sobre la limpieza
review_scores_checkin'	Discreta	Rate sobre el checkin
review_scores_communication'	Discreta	Rate sobre la comunicacion con el host
review_scores_location'	Discreta	Rate sobre la ubicacion
review_scores_value'	Discreta	Rate sobre la calidad
requires_license'	Discreta	Requisito de licencia
instant_bookable'	Discreta	Verificacion reserva instantanea
is_business_travel_ready'	Discreta	Preparada para viajes de negocios
require_guest_profile_picture'	Discreta	Solicita foto de huesped
require_guest_phone_verification'	Discreta	Solicita telefono de huesped
calculated_host_listings_count'	Discreta	Conteo de publicaciones del host
calculated_host_listings_count_entire_homes'	Discreta	Conteo de publicaciones de casas enteras del host
calculated_host_listings_count_private_rooms'	Discreta	Conteo de publicaciones de habitaciones privadas del host
calculated_host_listings_count_shared_rooms'	Discreta	Conteo de publicaciones de habitaciones compartidas del host
neighbourhood_cleansed'	Discreta	Alcaldia
calendar_updated'	Discreta	Ultima actualizacion del calendario
property_type'	Discreta	tipo de propiedad
cancellation_policy'	Discreta	Politica de cancelacion
room_type'	Discreta	Tipo de habitacion
bed_type'	Discreta	tipodecama
'last_scraped'	Fecha	Fecha en la que se obtuvieron los datos
host_since'	Fecha	Fecha de registro del host
calendar_last_scraped'	Fecha	Fecha de ultimo corte de informacion
first_review'	Fecha	Fecha de primer review
last_review'	Fecha	Fecha de ultimo review
'name'	Texto	Titulo del anuncio
summary'	Texto	Resumen de la propiedad
space'	Texto	Resumen del espacio
description'	Texto	Descripcion de la propiedad
experiences_offered'	Texto	Experiencias ofrecidas por la propiedad
neighborhood_overview'	Texto	Resumen del vecindario
notes'	Texto	Notas del host
transit'	Texto	Accesibilidad de la ubicacion del inmueble
access'	Texto	Zonas disponibles en la casa
interaction'	Texto	Formas de comunicacion con host
house_rules'	Texto	Reglas de la propiedad
host_name'	Texto	nombre del host
host_location'	Texto	Localización del host
host_about'	Texto	Descripcion del host
host_response_time'	Texto	Tiempo de respuesta del host
host_neighbourhood'	Texto	vecindario host
host_verifications'	Texto	documentos de verificacion del host
street'	Texto	calle de la propiedad
neighbourhood'	Texto	Colonia
neighbourhood_group_cleansed'	Texto	Grupo de Alcaldias
city'	Texto	Ciudad
state'	Texto	Estado

Figura 1.2: Variables.

Variable	Etiqueta	Descripción
market'	Texto	Mercado de propiedades ubicacion
smart_location'	Texto	Ubicacion simplificada
country_code'	Texto	Codigo de pais
country'	Texto	Pais
amenities'	Texto	amenidades
jurisdiction_names'	Texto	claves localizacion

Figura 1.3: Variables.

Se retiran las variables de llaves ya que no nos aportan información relevante para el análisis de los datos, ya que representan las llaves o id que le brindan identidad a los registros, para las propiedades y para los anfitriones.

2 Calidad de Datos

2.1 Duplicados

Se registraon 0 % duplicados en la tabla, por lo que no se retira ninguna fila en esta sección. Se realizó el filtro por 'id' y no se encontraron duplicados.

2.2 Completitud

Para esta sección se registraron 26 variables que presentan menos del 80 % de completitud en sus datos. Junto con las variables llaves que se retiraron en esta sección, conservamos un total de 68 variables. Se presenta un gráfico con todas las variables que presentan valores faltantes.

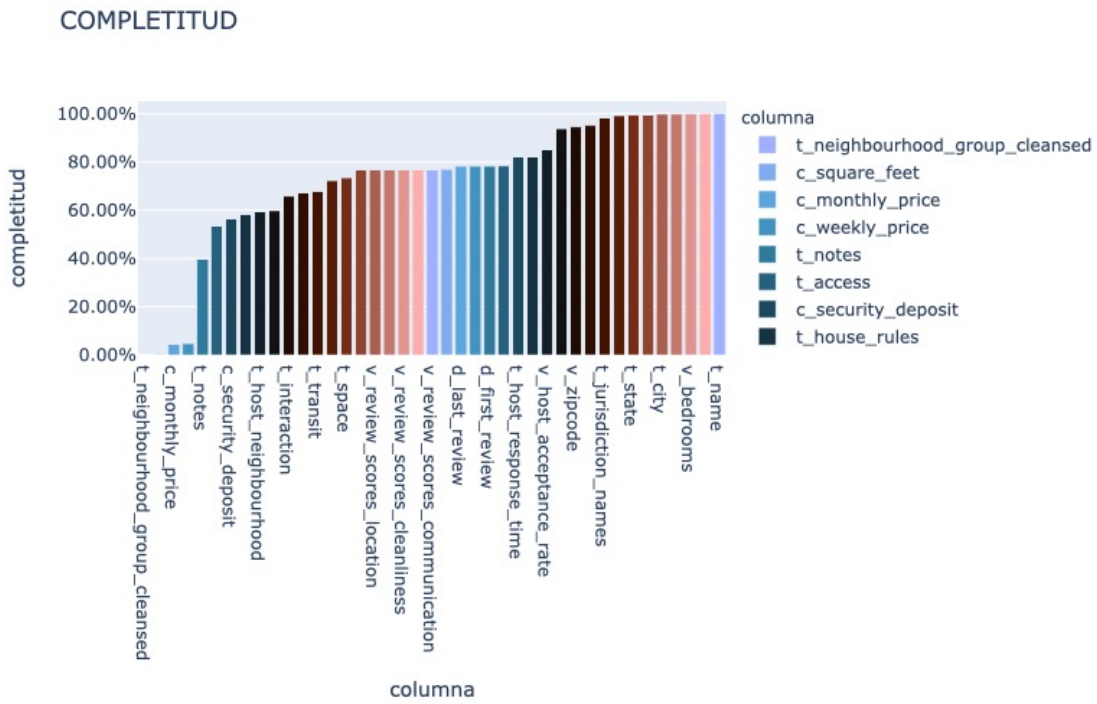

Figura 2.1: Completitud 1.

Después de retirar las variables mencionadas se presentan 15 variables con datos faltantes, las cuáles se muestran a continuación. En otro módulo serán imputados dichos valores.

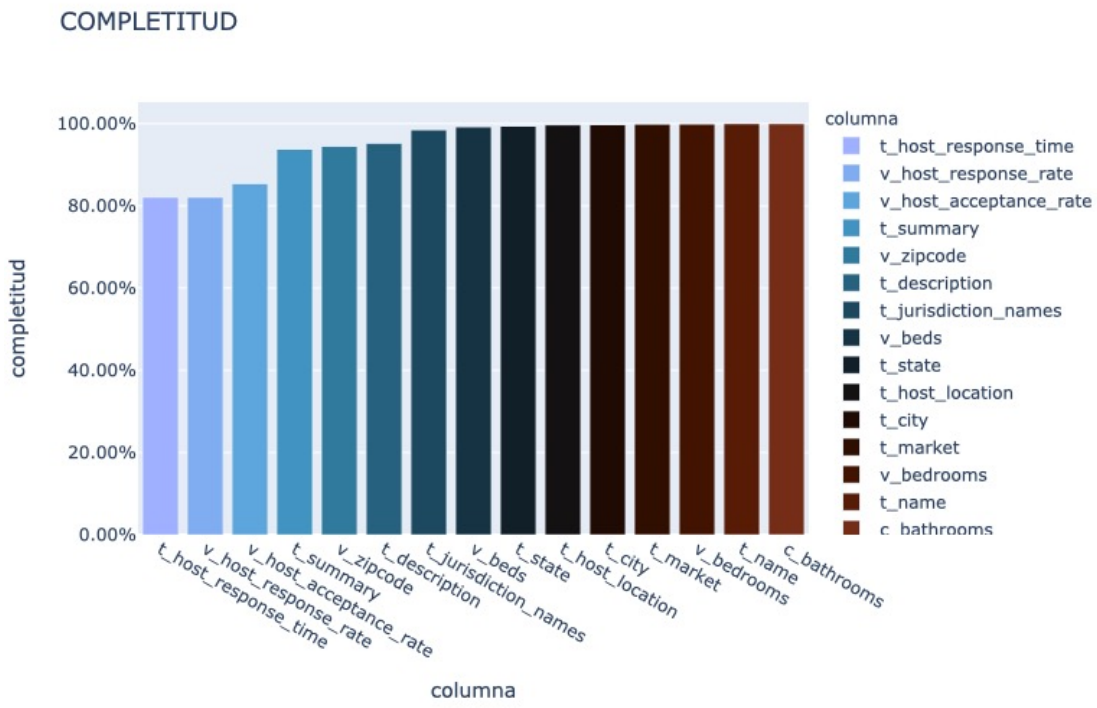

Figura 2.2: Completitud 2.

2.3 Consistencia

En esta sección se busca que las variables tengan el tipo de dato correcto, de acuerdo a su clasificación o etiqueta. Se busca retirar los valores incosistentes según los valores de la variable y se retiran las variables unitarias, ya que no aportan información adicional a la tabla.

Para la variable objetivo se retiraron los caracteres especiales, se colocó el tipo de dato correcto y se renombró la variable como 'tgt_price'.

Para las variables con fecha, se revisó que no existieran datos inconsistentes, según la fecha de inicio de operaciones de airbnb en el 2008. Se retiró la variable 'd_last_scraped' al tratarse de una variable unitaria.

Para las variables discretas se revisaron que no existieran datos inconsistentes. Para la variable 'v_accommodates' se encontraron 2 valores superiores al número máximo de huespedes que permite la plataforma que son 16, se colocó el número 16

para esos valores, para ser congruente con el número de camas y cuartos que contiene la casa. Se retiraron los caracteres especiales para el resto de variables discretas y se les coloco el tipo de dato 'int' para las que no presentaran missings y tipo float para respetar las variables que tuvieran datos faltantes.

Para las variables continuas, se revisó que la longitud y latitud corresponda a la Ciudad de México con coordenadas: "Latitud: 19.4978, Longitud: -99.1269 ", se realizó un mapa para apreciar la distribución de las propiedades en la ciudad.

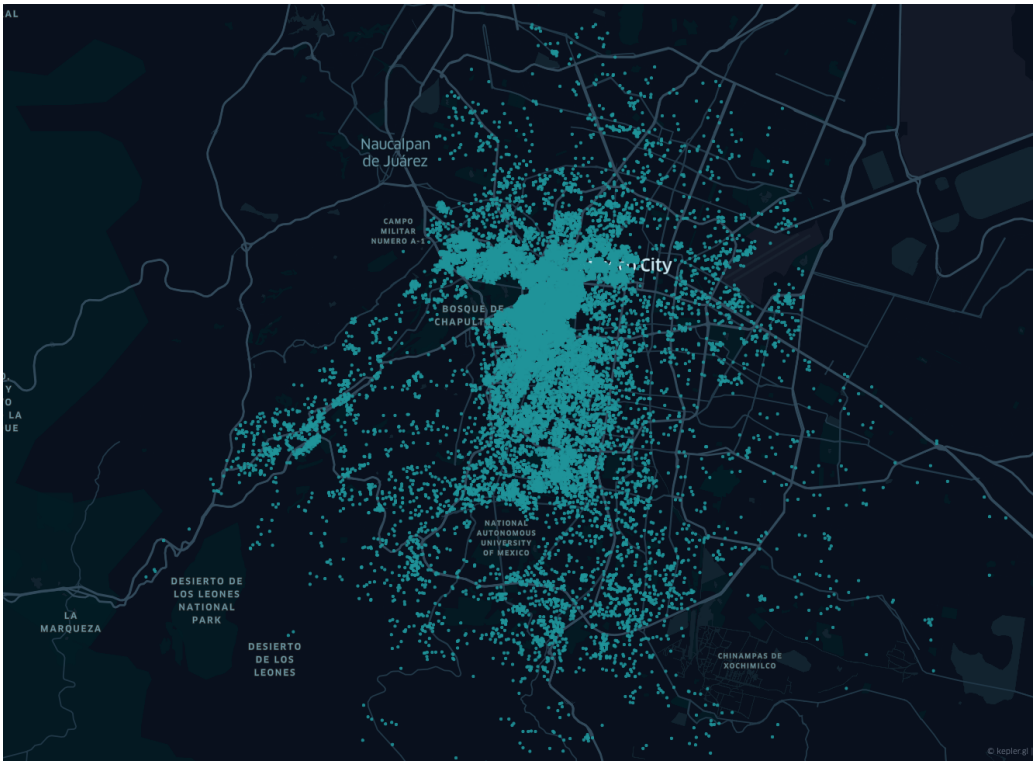

Figura 2.3: Mapa Ciudad de México.

Para el resto de las variables se retiraron los caracteres especiales y se asignó el tipo de dato correcto, corresponde a tipo float y se conservaron los valores faltantes.

Para las variables de texto se limpiaron los registros quitando los caracteres especiales y números, se retiraron los stopwords y hapaxes para las variables 't_name' y 't_summary'.

Se retiraron las siguientes variables unitarias

- 'v host has profile pic'
- 'v has availability'
- 'd calendar last scraped'

- 'v requires license'
- 'v is business travel ready'
- 'v require guest profile picture'
- 'v require guest phone verification'
- 'v calculated host listings count shared rooms'
- 't experiences offered'
- 't market'
- 't country code'
- 't city'
- 't state'
- 't country'
- 't country code'
- 'v bed type'
- 't jurisdiction names'
- 't description'

3 Análisis Exploratorio

3.1 Año antigüedad host

El primer dato importante que obtuve fue conocer el comportamiento de registro de host por año, como se puede notar en la gráfica existe un aumento exponencial en los primeros 8 años de la plataforma, posteriormente comienza a decrecer para encontrar su estabilidad entre los años 2018 y 2019.

3.2 Reviews último mes

Para conocer un poco más sobre la relación que establecen los hosts con los huespedes es interesante analizar cuantas personas colocan alguna reseña sobre el establecimiento, esto también nos puede dar una idea de la demanda de los inmuebles el último mes al tener una concentración de 32 % en 0 reseñas nos indica que quizá no existe una demanda considerable en el último mes, tiene sentido porque en ese mes comenzaba a tomar fuerza el tema de la pandemia en el país .

3.3 Amenities

Hay que considerar los elementos indispensables para poder recibir a una persona como en su propia casa, por eso es importante conocer que elementos proliferan en las propiedades que ya están funcionando en airbnb, contar con la mayor cantidad de elementos con ese valor agregado para los clientes es crucial para aumentar la probabilidad de reservación.

3.4 Política de cancelación

Un factor que puede determinar la reservación es la flexibilidad con el cliente, sabemos que pueden presentarse miles de incovenientes en una ciudad tan alocada como lo representa la Capital de México; siempre hay que considerar que la retención de clientes está ampliamente ligada a la capacidad de la entidad de solucionar los problemas a los que se enfrentan sus consumidores, generando una lealtad genuina, dificilmente reemplazable.

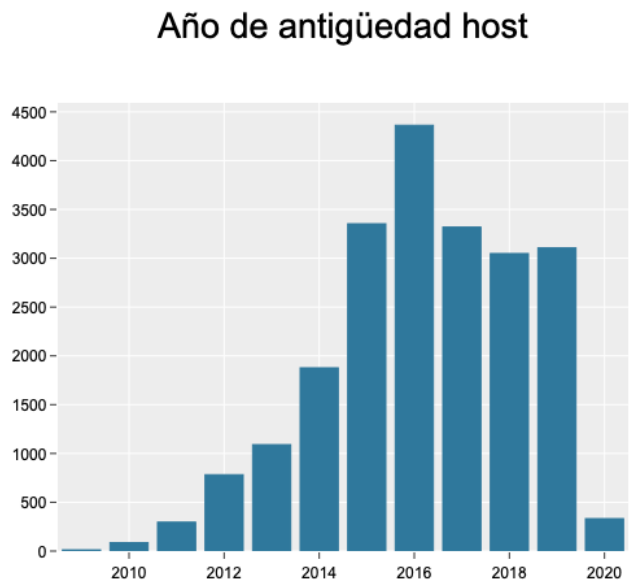

Figura 3.1: Año de ingreso host.

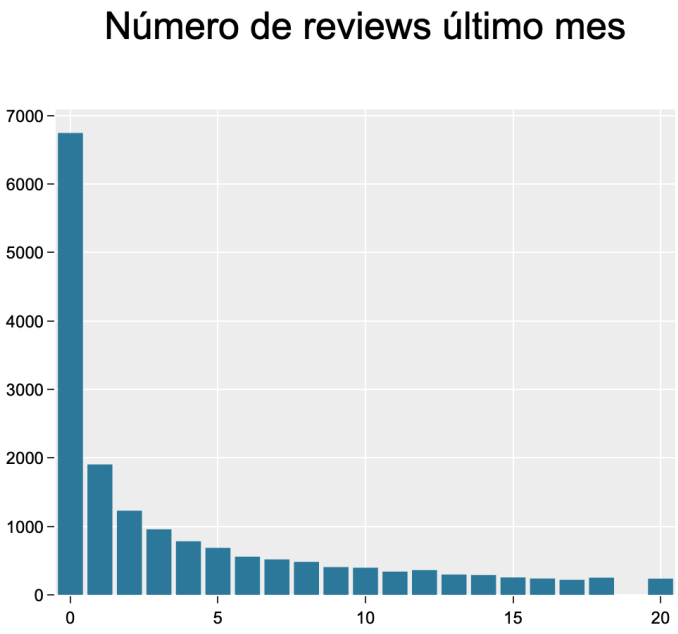

Figura 3.2: Reviews último mes.

Figura 3.3: Amenities.

Politica de cancelación

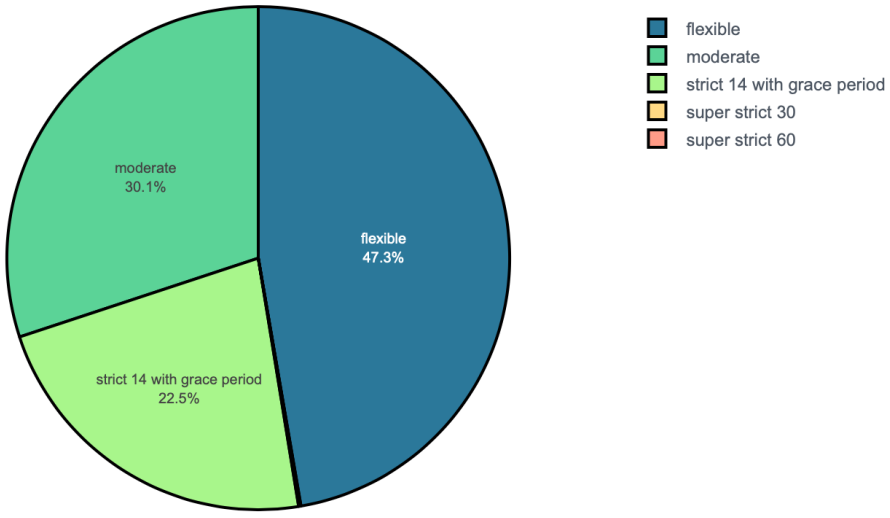

Figura 3.4: Política de cancelación.

3.5 Tipo de propiedad

Para conocer la tendencia de desarrollo inmobiliario en la ciudad, basta con observar la gran cantidad de edificios que se han construido en los años recientes, pero la tabla nos confirma dicha información con una fuerte presencia de apartamentos en el conjunto de datos con 64 % de los registros.

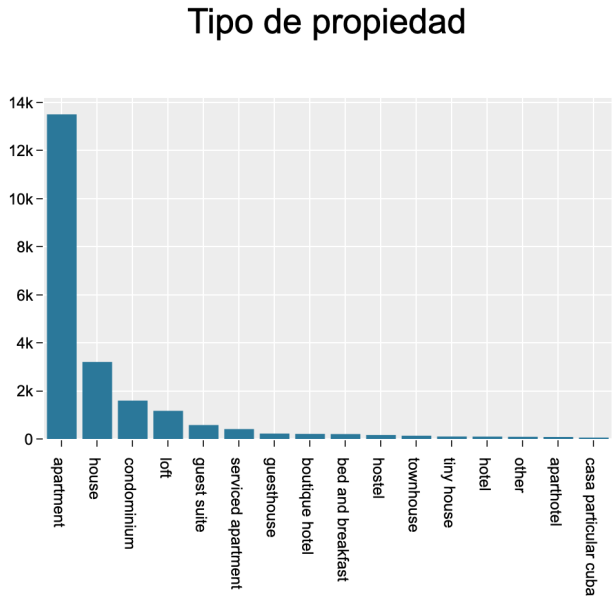

Figura 3.5: Tipo de propiedad.

3.6 Propiedades por alcaldía

Me parece interesante conocer en que parte de la ciudad se encuentra más desarrollado el mercado de alojamientos, esto nos podría permitir conocer los patrones que existen en esa zona e implementar dichas características para fomentar el crecimiento en otras áreas no solo de la ciudad, de la república mexicana con el potencial de poder generar índices similares de acuerdo con las características de los inmuebles y la posición geográfica de los mismos.

3.7 Descripción del inmueble

Si quieres ser la primer opción entre los huespedes debes usar las palabras clave de manera inteligente a tu favor, por ello es importante conocer que conceptos se repiten o tienen una mayor presencia en las ofertas de la plataforma, de esta forma se podría ubicar entre las primeras opciones de busqueda de los clientes, lo cuál adquiere mayor relevancia cuando el número de ofertas crece.

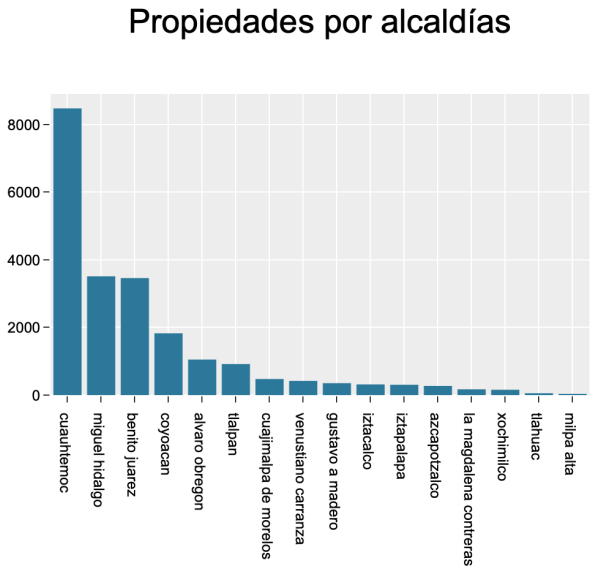

Figura 3.6: Alcaldías.

Figura 3.7: Palabras clave.

3.8 Número de huespedes

Finalmente me gustaría presentar la distribución de huespedes que reciben las propiedades, englobando todos los datos anteriores podemos visualizar un posible patrón de propiedades, con poco espacio, flexibles, ubicadas en zonas comerciales y en apartamentos, por lo que esta gráfica nos muestra la relación que representan los alojamientos para menos de 3 personas con los datos anteriores .

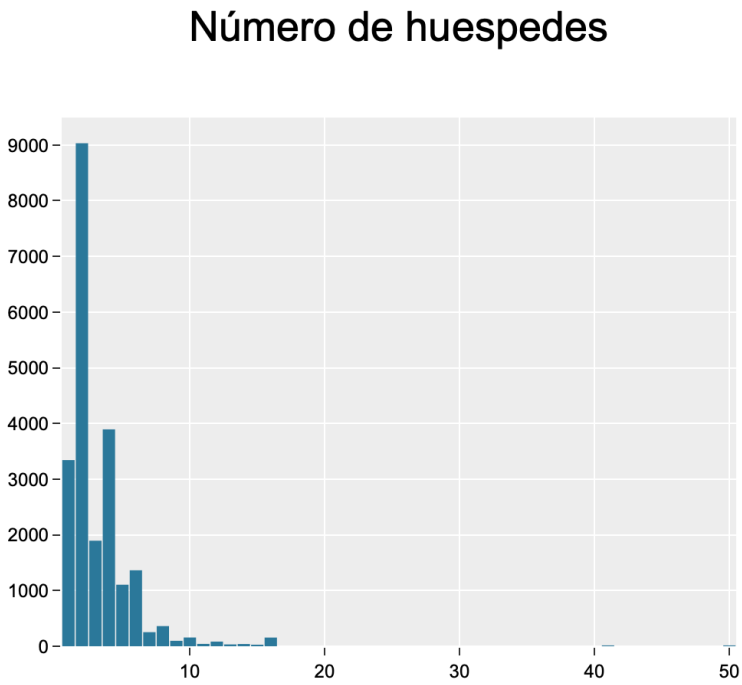

Figura 3.8: Número de huespedes.

4 Outliers

Para esta sección se realizó el tratamiento para 5 variables continuas. Se realizó una prueba de normalidad para determinar si era posible utilizar el método Z-score; la prueba indicó que las variables no tenían una distribución normal, por lo que se optó por los métodos de IQR y percentiles para determinar los valores atípicos en los registros.

Se eliminaron 2654 registros que corresponden al 12 % de los registros iniciales.

	features	n_outliers_IQR	n_outliers_Percentil	n_outliers_IQR_%	n_outliers_Percentil_%	total_outliers	%_outliers	indices
0	c_host_listings_count	2441	1046	11.34	4.86	1046	4.86	[18440, 16284, 6156, 10252, 6158, 16285, 6160,...
1	c_host_total_listings_count	2441	1046	11.34	4.86	1046	4.86	[18440, 16284, 6156, 10252, 6158, 16285, 6160,...
2	c_extra_people	870	1075	4.04	5.00	870	4.04	[12293, 18439, 4104, 10, 11, 20492, 18447, 102...
3	c_minimum_nights_avg_ntm	2410	822	11.20	3.82	822	3.82	[12288, 8194, 8202, 10251, 16405, 12310, 4122,...
4	c_maximum_nights_avg_ntm	6	1081	0.03	5.02	6	0.03	[2785, 2787, 2788, 2344, 4328, 2781]

Figura 4.1: Tabla outliers.

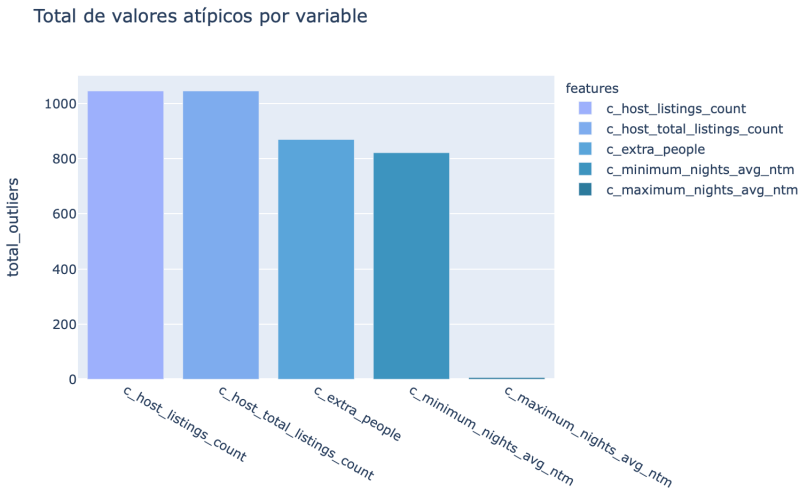

Figura 4.2: Outliers por variable.

4.1 Gráficos tratamiento

Se presentan los gráficos previos de las variables continuas antes del tratamiento de valores atípicos.

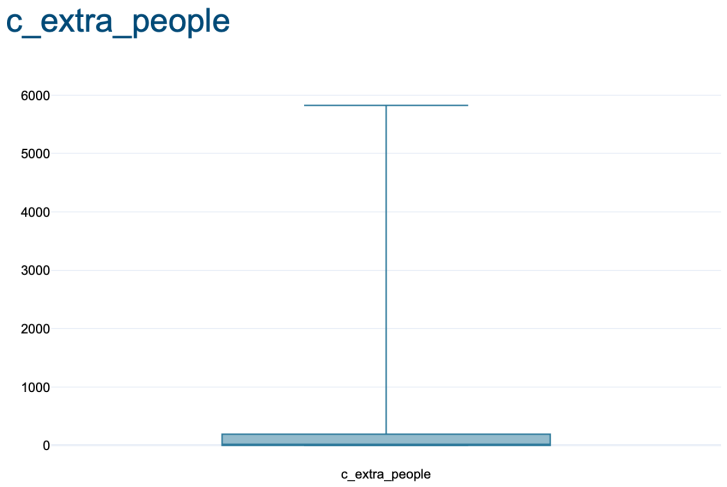

Figura 4.3: Costo por persona extra.

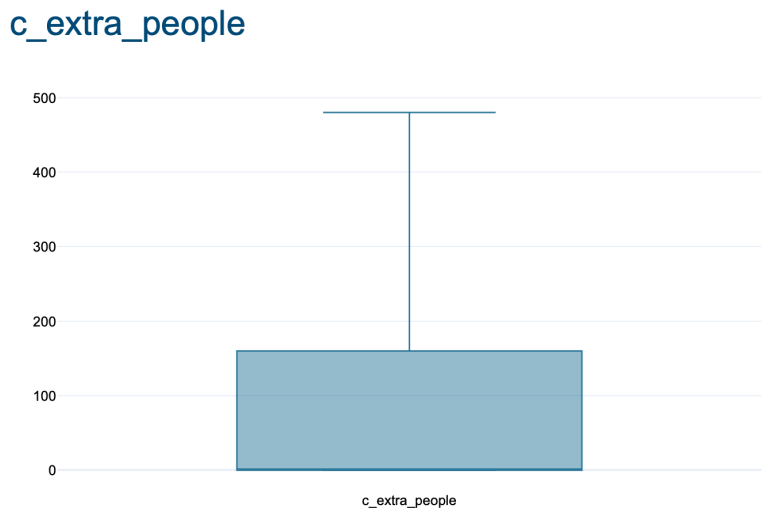

Figura 4.4: Costo por persona extra después de outliers.

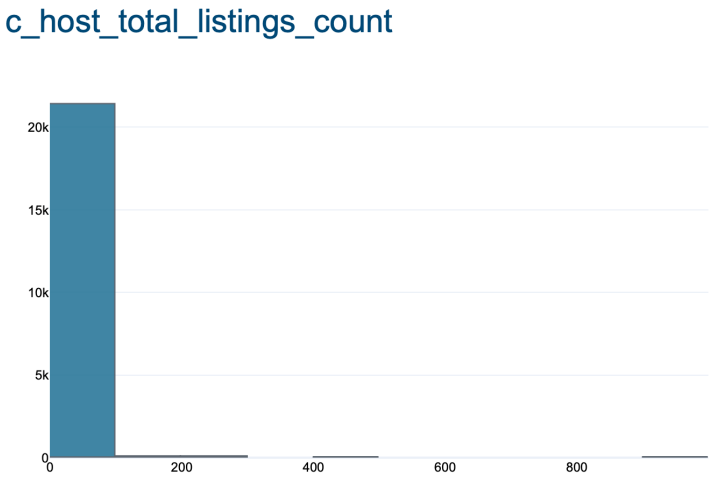

Figura 4.5: Número total de publicaciones.

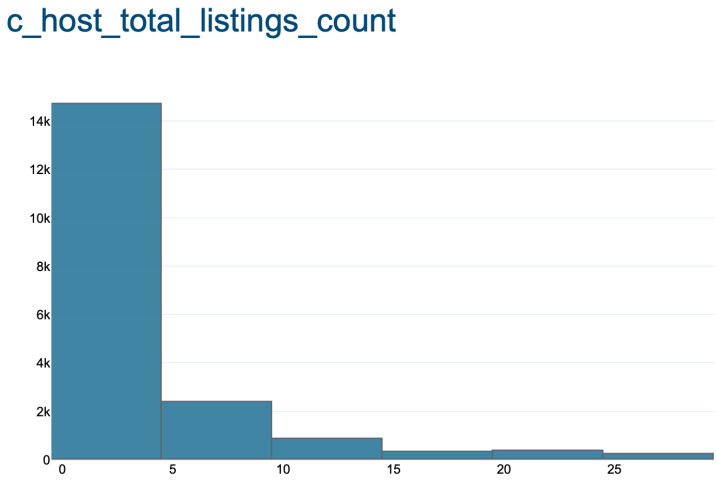

Figura 4.6: Número total de publicaciones después de outliers.

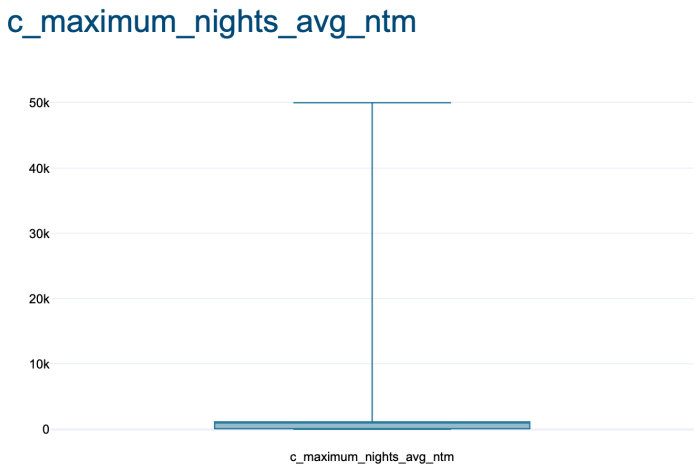

Figura 4.7: Máximo de noches promedio.

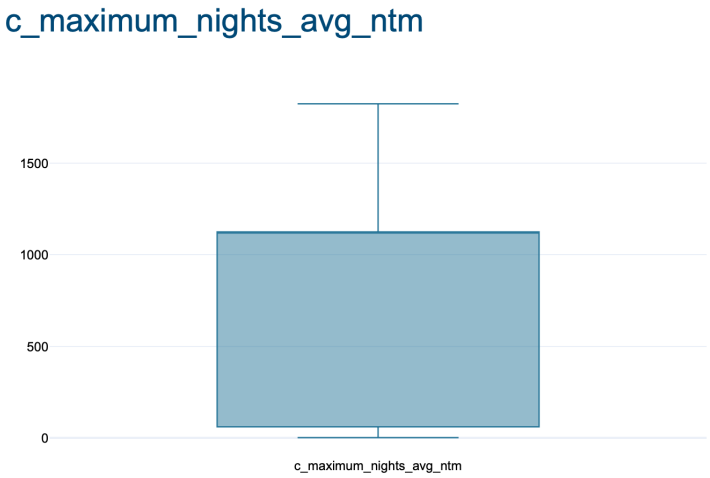

Figura 4.8: Máximo de noches promedio después de outliers.

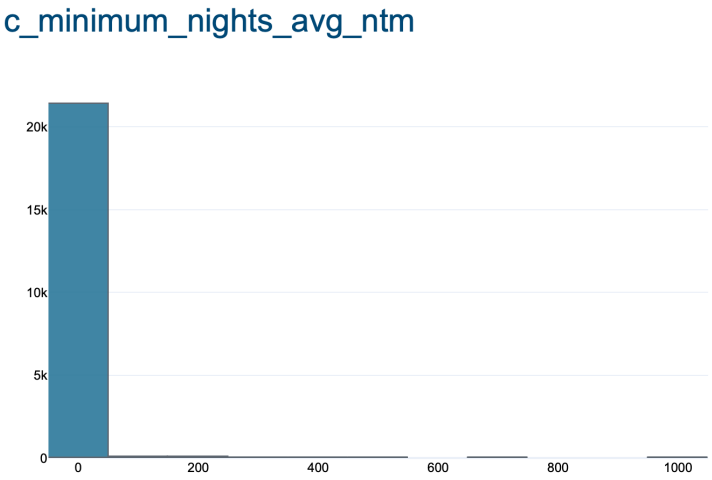

Figura 4.9: Mínimo de noches promedio.

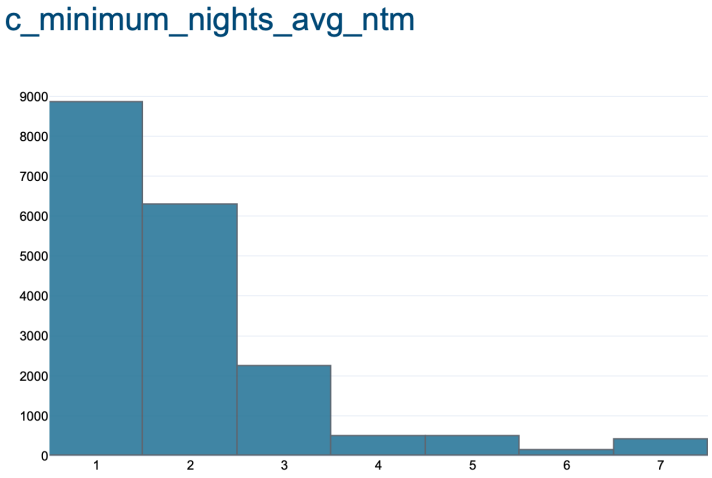

Figura 4.10: Mínimo de noches promedio después de outliers.

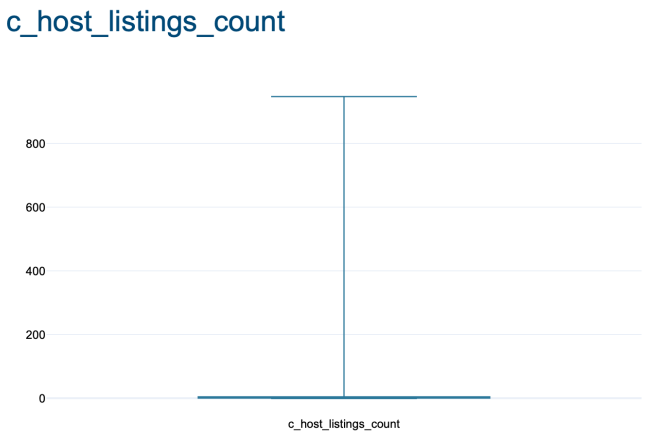

Figura 4.11: Publicaciones activas.

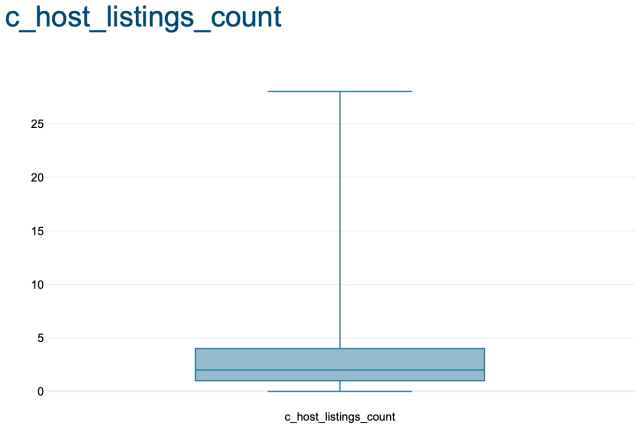

Figura 4.12: Publicaciones activas después de outliers.

5 Missings

5.1 Eliminación por baja varianza

- 'v host response rate' 18.0 % missings
- 'v host acceptance rate' 18.0 % missings
- 't host response time' 14.31 % missings

Se retiraron de la tabla debido a que tenían muy baja varianza y un elevado número de missings, al momento de imputar y comprobar que se respetaban las proporciones no se obtenía un estadístico que nos permitiera no rechazar dicha proporción, por lo que al ser cercanos a valores unitarios y no aportar información relevante, decidí que lo mejor sería retirar estas variables del dataset.

5.2 Variables discretas

- 'v zipcode' 5.5 % missings
- 'v beds' 0.7 % missings
- 'v bedrooms' 0.1 % missings

Se imputó por moda y se respetó la proporción original después de la imputación, por lo que se procedió a mantener la imputación por moda para las tres variables.

5.3 Variable continua

- 'c bathrooms' 0.03 % missings

Después de realizar la prueba de Kolmogorov-Smirnov sobre la distribución de dos muestras, se obtuvo que el mejor método era por mediana por lo que se procedió con la imputación bajo este método.

Hay que recordar que para realizar este procedimiento se partió la tabla de datos de manera estratificada con una proporción 70-30 para entrenamiento y test, respectivamente.

6 Ingeniería de Variables

6.1 Dummies

Después de normalizar y transformar según corresponda, se obtuvieron las siguientes variables dummies:

- 'v host identity verified'
- 'v host is superhost'
- 'v instant bookable'
- 'v is location exact
- 'v property type'
- 'v calendar updated'
- 'v neighbourhood cleansed'
- 'v room type'
- 'v cancellation policy'

6.2 Año registro host

Se generó la variable para conocer la antigüedad en años del host en la plataforma, podría estar relacionado con el refinamiento del precio para obtener una mayor demanda.

6.3 Noches reservadas en los siguientes 30 días

Un dato fundamental para conocer la demanda que tienen las propiedades y el número de reservaciones futuras que podría ser predecido en un futuro modelo con la tabla analítica.

7 Reducción de dimensiones

7.1 Relación de valor perdido

Para este rubro no se encontraron variables que presentaran valores faltantes después de los tratamientos.

7.2 Filtro de baja varianza

Gracias a los filtros realizados durante todo el proceso no se eliminó ninguna variable por este filtro.

7.3 Filtro alta correlación

Se eliminan por este filtro las siguientes variables, el criterio de selección corresponde a las variables que presentan mayor información para los siguientes 30 días.

- 'c host listings count'
- 'c maximum nights avg ntm'
- 'v maximum maximum nights'
- 'v minimum maximum nights'
- 'c minimum nights avg ntm'
- 'c minimum nights avg ntm'
- 'v maximum minimum nights'
- 'v minimum minimum nights'

- 'v availability 60'
- 'v availability 90'
- 'v availability 365'
- 'v number of reviews'
- 'v calendar updated mas de un mes'
- 'v calendar updated menos de 3 semanas'

7.4 Correlación con la variable objetivo

Se eliminan las siguientes variables por este filtro:

- v neighbourhood cleansed iztapalapa
- v minimum nights
- v neighbourhood cleansed gustavo a madero
- v neighbourhood cleansed iztacalco
- v cancellation policy moderate
- v host identity verified
- v neighbourhood cleansed azcapotzalco
- v neighbourhood cleansed venustiano carranza
- v room type hotel room
- v neighbourhood cleansed cuajimalpa de morelos
- c extra people
- v zipcode
- v property type condominio
- v cancellation policy super strict 30
- v neighbourhood cleansed xochimilco
- v neighbourhood cleansed tlahuac

- v is location exact
- v neighbourhood cleansed la magdalena contreras
- v cancellation policy super strict 60
- v neighbourhood cleansed milpa alta
- v availability 30
- v property type otros
- v instant bookable

7.5 Multicolinealidad

A pesar de que se obtuvieron 12 variables con un VIF mayor a 5 decidí no retirarlas de la tabla debido a su cercanía con el 5 y no ser tan estricto para este punto por la información que aportan para describir a la variable objetivo. Como nota hay que considerar que para las variables un coeficiente de regresión no se estima adecuadamente.

7.6 PCA

Realicé el proceso para componentes principales con 20 componentes y escalando los datos para eliminar las posibles influencias. Obtuve que con 13 componentes se representa el 80 % de la varianza por lo que se tiene un porcentaje considerable con una reducción del 48 % de las variables.

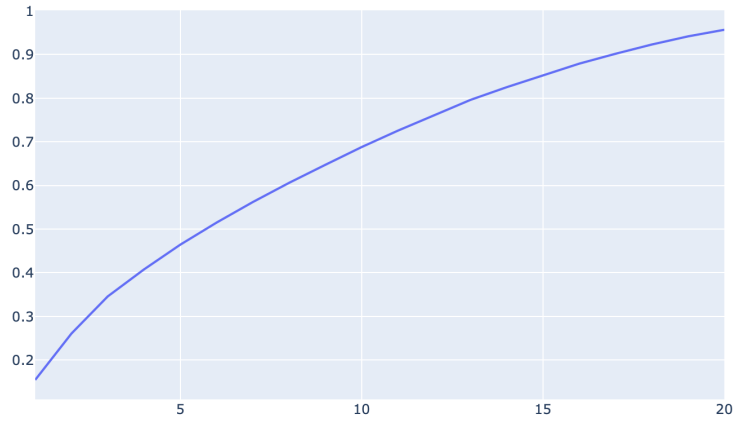

Figura 7.1: PCA 20 componentes.

8 Modelos

Después de un exhaustivo análisis de reducción de dimensiones, obtuve la mejor tabla analítica posible de acuerdo a mi selección de variables, para esta entrega realicé algunas modificaciones buscando un mejor desempeño de mi conjunto de datos para la sección de modelos.

El primer punto, que me parece de los más importantes fue reducir o segmentar la información que tenía, ya que mantenía un alto nivel de información, pero que se encontraba muy dispersa al tratarse de grupos de propiedades con características particulares, donde entran algunas clasificaciones como el lujo o un mejor nivel de propiedades que destacaban por encima de las demás con precios fuera de los rangos conocidos. Es por eso que opte por segmentar los datos que tengo en el rango de precios donde se encontraba la mayor parte de la informacion acumulada. Esto para describir de una mejor forma los precios de los alojamientos y que se tuviera una mejor aproximación en los resultados.

Otra segmentación que realicé fue por tipo de propiedad, ya que la mayor parte de mi información hacia referencia a propiedades de tipo apartamento, por lo que decidí enfocarme en este tipo de propiedades, para conocer mejor las cualidades que deben mantener para tener un mejor precio. Es evidente que este análisis se puede repetir para todos los tipos de propiedad, pero es necesario tener información suficiente en el conjunto de datos, para poder realizar un análisis adecuado de dichas propiedades, en este caso no se tiene la información necesaria para expandir el análisis a otras propiedades.

8.1 Modelos y métricas

Decidí realizar una división de dataset con 80 % para el conjunto de entrenamiento y 20 % para en conjunto de prueba.

Probé un total de 5 modelos de regresión con el apoyo de GridSerch para optimizar la busqueda de los hiperparametros que me permitieran obtener los mejores resultados para el modelo. Utilicé una función que me devuelve una tabla con el resumen de las métricas de los modelos que fueron seleccionados y un dataframe con la información necesaria para realizar un análisis de estabilidad de los modelos. La información obtenida es la siguiente:

	modelo	datos	r2	r2_adj	mae	mse	rmse
0	Regresion Lineal	train	0.514128	0.512780	216.982413	81026.162151	284.650948
1	Regresion Lineal	test	0.513364	0.507919	215.189044	80194.067113	283.185570
2	Lasso	train	0.514106	0.512759	217.022752	81029.691855	284.657148
3	Lasso	test	0.513166	0.507718	215.250091	80226.773945	283.243312
4	Ridge	train	0.514126	0.512778	217.003191	81026.433675	284.651425
5	Ridge	test	0.513286	0.507841	215.219859	80206.854221	283.208146
6	ElasticNet	train	0.490225	0.488811	225.250120	85012.271112	291.568639
7	ElasticNet	test	0.478245	0.472408	224.037066	85981.364427	293.225791
8	AdaBoost	train	0.466317	0.464837	239.696509	88999.277089	298.327466
9	AdaBoost	test	0.461131	0.455102	238.245983	88801.673722	297.996097

Figura 8.1: Métricas modelos de regresión.

Después de aplicar crossvalidation, tratar de realizar toda la optimización posible con los recursos disponibles, obtuve que el modelo con el mejor desempeño es el modelo de Regresión Lineal Múltiple. Esto de acuerdo en primer lugar a que tiene la mejor R2 de los modelos obtenidos. Es normal que obtenga una R2 tan baja debido a la variación tan importante de las propiedades en la Ciudad de México, en mi opinión para obtener mejores resultados se pueden aplicar o iniciar relaciones con información de otras aplicaciones, por ejemplo crear un métricas de información desde las avenidas principales de las propiedades, ya que esto incrementa el valor de los alojamientos; también hay que considerar que Airbnb en la Ciudad de México aún no es el gigante que es en otros países, conforme pasen los años es posible que se agreguen más propiedades y pueda describirse de mejor forma el comportamiento y estabilidad de los precios de los alojamientos. Esta información podemos verla claramente representada en las otras métricas donde la distancia con los datos reales es relativamente grande, debido a que aún no existe una estabilidad en los precios que pueda representar una relación clara entre la calidad o los contenidos de las propiedades y el precio. Una relación natural que podría definir el precio, son las calificaciones asignadas por los huespedes, pero desafortunadamente aún no hay una participación importante por parte de los huespedes que permita obtener información destacada de estos medios. Como un primer acercamiento me parece que el modelo conserva su estabilidad de manera adecuada con respecto a su contraparte de prueba donde no se observa una pérdida significante de la información descrita por parte del entrenamiento. Por ello el modelo no presenta un sobreajuste de los datos. En comparación y dando sentido a la selección del modelo, notamos que las variantes de la regresión lineal, a pesar de presentar similitudes, mantienen ligeramente, resultados peores en comparación. Analizando el modelo de AdaBoost, presenta resultados similares, pero por debajo por los obtenidos por los otros modelos, aunque de todos en general es importante destacar que no es notorio un sobreajuste en ellos.

8.1 Análisis de estabilidad

Una vez seleccionado el modelo de Regresión Lineal Múltiple como el mejor modelo, busqué comparar de manera visual el comportamiento de las predicciones con los resultados reales de la tabla. Un acercamiento una muestra de la tabla, me permitió observar como al modelo le cuesta definir una aproximación certera con precios en un rango medio, ya que podría presentarse similitudes entre alojamientos pero con precios muy distintos, esto como comento anteriormente, puede ser a que no existe una estabilidad tangible en los precios de la aplicación que determine el comportamiento del mercado, se dará eventualmente con el aumento de la oferta y la comparación inteligente por parte de los anfitriones de la competencia, por eso nuestra labor es tan importante antes de realizar modelos, para poder realizar una correcta segmentación de la información orientada a un objetivo en particular. Para este caso yo trate de realizar un mapeo general y obtener resultados a partir de ese objetivo, claramente analizar de manera particular me permitiría obtener muchos mejores resultados y un modelo con un mejor desempeño, por ejemplo, analizando todos los departamentos en polanco para 2 huéspedes; pero desde mi punto de vista ese es el proposito de esta práctica analizar la información desde un punto incial y presentar cual sería la forma correcta de realizar un análisis de la información. Baso mis argumentos en los resultados obtenidos al segmentar la información, donde mis métricas fueron mejorando cada vez que yo tomaba datos más detallados de las propiedades y me enfocaba un mercado definido dentro de la aplicación. Es ahí donde yo noto un valor de la información y del proceso mismo del análisis de los datos. Por cuestiones de resumen y no desviar el valor de los datos obtenidos, presento la información con el mayor impacto posible y no con todas las combinaciones que realicé para encontrar la que tuviera mayor impacto. Aunque al seguir el proceso desde el análisis de datos, pude notar gracias a esa sección que el camino lógico sería continuar con el tipo de propiedad más destacado en los datos.

Figura 8.2: Train vs Pred.

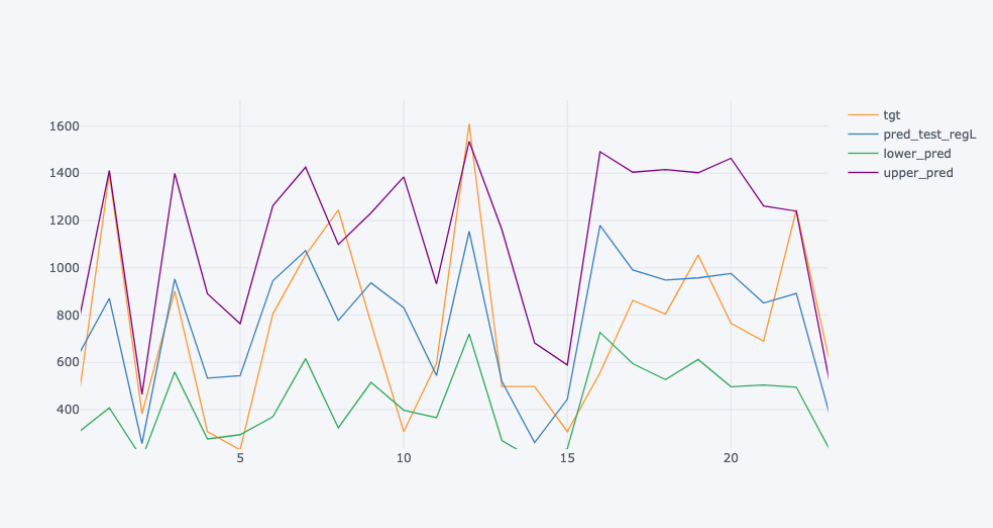

Figura 8.3: Test vs pred.

9 Conclusiones

Gracias a todo el proceso que seguí para realizar el análsis de los datos pude notar lo que realmente es valioso de la información y como obtener datos de alto impacto para la toma de decisiones. Para los pasos posteriores que yo prospecto para este proyecto es realizar una segmentación detallada de la información donde exista información suficiente para obtener modelos eficientes, de acuerdo a la información aquí presentada un posible candidato con una eficiencia correcta sería los apartamentos en la Alcaldía con mayor presencia de propiedades, siguiendo el análisis previo un aumento en la oferta permite una estabilida de los precios de acuerdo a la demanda de los mismos. También la colaboración con otras fuentes de información como lo son google maps o el aumento de las reseñas y calificaciones de los usuarios puede permitir un mejor desempeño de los modelos y potenciar el resto de las variables del conjunto de datos.

Realmente es un tema que me apasionado y que considero que puede ser llevado a un servicio de intermediación para presentarle a los anfitriones como potenciar sus propiedades a través de la manipulación del precio o sus contenidos, claro al estar limitados por la posición de sus propiedades. Otro camino a seguir sería analizar posibles propiedades con un alto potencial de destacar en la aplicación de acuerdo a su posicioón geográfica, es por eso la insistencia en la conectividad con otras fuentes de información y obtener más datos sobre lo que rodea al gigante digital en todo el mundo que es Airbnb.

BIBLIOGRAFÍA

Recuperado de : 'http://insideairbnb.com/get-the-data.html'