UNIVERSIDAD NACIONAL
AUTONOMA DE MEXICO

DIPLOMADO EN CIENCIA DE DATOS

Presencia de Airbnb en la CDMX

Rocío Ortega Kingston

28 de noviembre de 2020

Índice

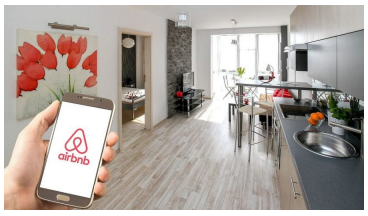
Índice de figuras

1. Introducción

El turismo es parte muy importante en las estrategias económicas de la mayoría de los países, tal es el caso del nuestro, México es reconocido por su gran potencial y riqueza de recursos naturales y culturales y para el ser humano, viajar siempre ha sido una actividad primordial y en medida en que evoluciona el turismo las empresas de este ramo buscan darle mayor confianza y conformidad a sus usuarios.

En el 2008 un grupo de 3 jóvenes encontraron una forma de diversificar y capitalizar una manera diferente de alojamiento, primero dirigido hacia quien buscara una arrendamiento común pero con el tiempo se convirtió en algo más allá de buscar una habitación en un resort todo incluido o un hotel sencillo, Airbnb es hoy una empresa multimillonaria que a través de su plataforma permite que los anfitirones puedan rentar sus propiedades disponibles para turistas quienes visitan la ciudad o personas que tienen un viaje de negocios. Lo que hace interesante a este negocio es la diversidad de alojamientos que se ofrecen en cuanto a tamaño y comodidades, hoy con esta plataforma se puede rentar una casa en tu destino preferido y viajar con toda tu familia o rentar un pequeño departamento un mes para un viaje de negocios, sin necesidad de tramites legales o avales.

Segun un artículo en *El Financiero*, Airbnb calcula que un propietario mexicano promedio puede ganar 2 mil 300 dólares adicionales al año alquilando una sola propiedad. La oprotunidad de negocio para nuestro país resulta bastante atractiva, pero ¿Cómo son los anfitriones de Airbnb en la Ciudad de México?¿Qué ofrecen sus alojamientos?¿Cuál es el precio promedio? La intensión de este análisis es encontrar respuestas a estas preguntas.



2. Datos

Para este análisis se cuenta con una muestra inicial de 2999 registros de anuncios de anfitriones de Airbnb y 36 atributos que describen al alojamiento y al anfitrión, extraída del sitio *Inside Airbnb*.

2.1. Diccionario de Datos

Antes de definir el tipo de valiables con el que se trabajó en este proyecto es necesario identificar las variables y su descripción.

VARIABLE	TIPO	DESCRIPCIÓN
host total listings count	float	Total de publicaciones de anuncios del anfitrión
bathrooms	float	Número de baños en el alojamiento
bedrooms	float	Número de recámaras en el alojamiento
beds	float	Número de camas en el alojamiento
accommodates	int	Número de huéspedes en el alojamiento
price	float	Precio del alojamento
minimum nights	int	Mínimo de noches para reservar
maximum nights	int	Máximo de noches para reservar
number of reviews	int	Número de reviews del alojamiento
id	int	Id del anuncio publicado
host is superhost	str	v = Es súper anfitrión / f = No es super anfitrión
host identity verified	str	v = Anfitrión verificado / f = Anfitrión no verificado
has availability	str	v = Tiene disponibilidad el alojamiento / f = No tiene disponibilidad el alojamiento
cancellation policy	str	Tipo de política de cancelación del alojamiento
country	str	País donde se ubica el alojamiento
state	str	Estado donde se ubica el alojamiento
city	str	Alcaldía donde se ubica el alojamiento
neighbourhood	str	Colonia donde se ubica el alojamiento
property type	str	Tipo de alojamiento
zipcode	float	Código postal de la ubicación del alojamiento
first review	DateTime	Fecha de la primera vista del anuncio
last review	DateTime	Fecha de la última vista del anuncio
name	Str	Nombre del alojamiento
description	Str	Descripción del alojamiento
amenities	Str	Comodidades en el alojamiento
house rules	Str	Reglas dentro del alojamiento

Figura 1: Diccionario de Datos

En el apéndice podrán encontrar información sobre el resto de los atributos.

2.2. Tipos de Variables

2.2.1. Variables Cualitativas

Son el tipo de variable que puede tomar cualquier valor dentro del espectro de los números reales.

En este proyecto dichas variables son: c host total listings count, c zipcode, c latitude, c longitude, c accommodates, c bathrooms, c bedrooms,c beds, c price,

c guests included, c minimum nights, c maximum nights, c number of reviews, c review scores rating, c reviews per month

2.2.2. Variables Categóricas

Estas variables sólo puede tomar un valor específico entre el conjunto de todos los valores posibles.

En este proyecto contamos con las siguientes variables categóricas: v host is superhost, v host identity verified, v has availability, v cancellation policy, v country, v state, v city, v neighbourhood, v property type, v zipcode

2.2.3. Variables Temporales

Se refiere a las variables de fecha.

d last scraped, d host since, d first review, d last review

2.2.4. Variables de Texto

Aqui estamos hablando de las variables que contienen literalmente texto.

t name, t description, t house rules, t amenities, t space, t summary

2.2.5. Variables de Indices

id id, id host id

3. Calidad de Datos

Por calidad de datos nos referimos al acondicionamiento de nuestros datos a través de técnicas y procesos que permitan realizar análisis precisos con ellos. Los datos en una organización son la información más valiosa para cualquier negocio ya que en ellos se basa el mayor porcentaje de la toma de decisiones, es por esta razón que es mandatorio contar con una base de datos consistente.

3.1. Duplicidad

Es importante revisar como primer paso si existen registros duplicados ya que de existir podrían alterar análisis posteriores.

En esta ocasión tenemos 0 % de datos duplicados en nuestra tabla.

3.2. Completitud

En esta parte de la calidad de datos buscamos tener registros completos para poder procesar correctamente la información.

A continuación veremos una tabla donde se aprecia el porcentaje de datos faltantes en la muestra.

	% de missings
t_space	0.233411
t_house_rules	0.286429
v_neighbourhood	0.157719
v_city	0.001000
v_state	0.005002
v_zipcode	0.068023
c_bathrooms	0.001667
c_bedrooms	0.000667
c_beds	0.001667
d_first_review	0.193398
d_last_review	0.193398
c_review_scores_rating	0.202734
c_reviews_per_month	0.193398

Figura 2: Completitud Previa

Observamos que tenemos variables con más del 20 % de su información faltante, esto puede ser contraproducente más adelante ya que es un porcentaje importante de información que si tratamos de imputar despues podría alterar los analisis posteriores. Dado lo anterior se decide eliminar las siguientes variables:

t house rules, t space y c review scores rating

3.3. Conformidad y Consistencia

En esta dimensión de datos se busca que estén en un formato estándar y legible para poder tener un buen procesamiento.

Para las variables continuas y variables de id se validó que todas estén en formato *float* o *int* y que además no tuvieran registros negativos.

Para las variables categoricas se validó también variable por variable que estén en el mismo formato, algunas en *float* o *int* y otras en *str*.

Para las variables de fecha se validó que estuvieran en formato "DateTime".

Y finalmente en las variables de texto se validó que todos los registros estén en formato tipo *str* además de eliminar palabras cortas como preposiciones y caractéres especiales y finalmente, que no contuvieran registros con números.

Todos los registros que no se encontraban en su formato de naturaleza fueron replazados por nulos, tal como en las siguientes variables:

Variable	Dato no válido
v_zipcode	méxico
v_neighbourhood	cuauhtemoc
v_city	mexico df

Figura 3: Calidad de Datos

3.4. Precisión

Después de revisar la conformidad y consistencia de nuestra tabla, notamos que en la variable v city la información no era precisa por lo que con la ayuda de una tabla extraída del sitio web *Datos Abiertos Ciudad de México* se logró hacer un cruce con nuestra tabla de datos a través de la columna v neighbourhood, dando como resultado lo siguiente:

CUAUHTEMOC	1292
MIGUEL HIDALGO	466
BENITO JUAREZ	405
COYOACAN	234
VENUSTIANO CARRANZA	154
LA MAGDALENA CONTRERAS	83
ALVARO OBREGON	52
AZCAPOTZALCO	27
IZTAPALAPA	21

Figura 4: Normalización de Alcaldías

Otra variable que se analizó en esta parte fue la de v state ya que en sus registros contenía datos fuera de su naturaleza totalmente, sin embargo sabemos que el estado de toda esta información es la Ciudad de México por lo que la corregimos.

3.5. Normalización

Otra cosa que notamos en la parte de conformidad y consistencia es que en algunas variables contábamos con muchas categorías, por lo que decidimos re-categorizar para facilitar el análisis posterior de las siguientes variables:

v neighbourhood y v property type

4. Análisis Exploratorio

A continuación algunas gráficas de variables antes y después de calidad de datos.

- En esta primera figura vemos el cambio gracias al proceso de conformidad y consistencia que le dimos a la variable v neighbourhood.

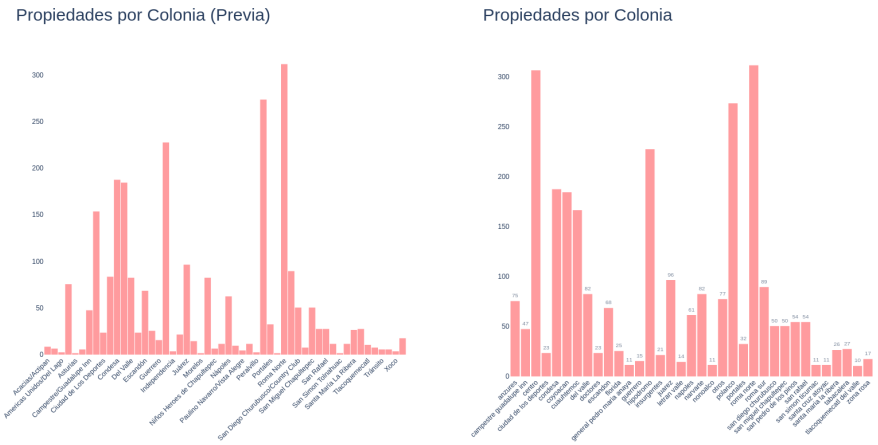


Figura 5: Análisis Exploratorio: Alojamiento por Colonia

El top 5 de colonias en donde encontramos más alojamientos es Roma Norte (11.04 %), Centro (10.86 %), Polanco (9.69 %), Hipódromo (8.06 %) y Condesa (6.64 %).

- En la siguiente apreciamos claramente el cruce con la tabla de alcaldías y colonias para precisar el dato de la variable v city.

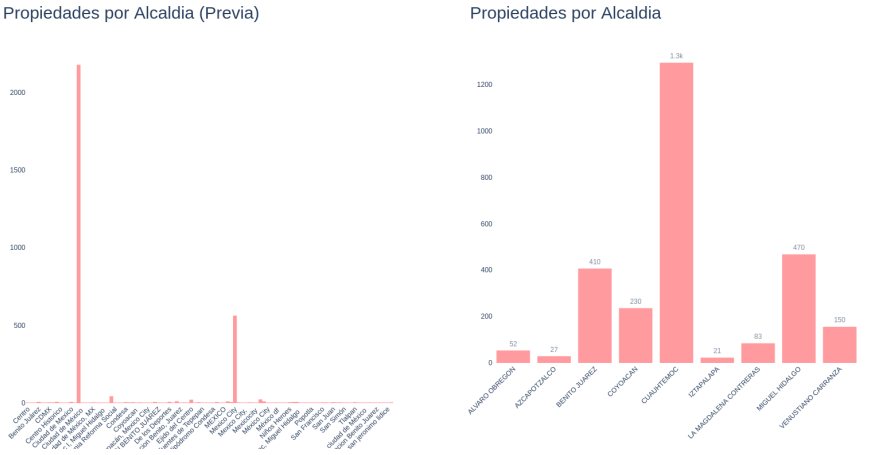


Figura 6: Análisis Exploratorio: Alojamiento por Alcaldía

Es claro que la alcaldía Cuauhtémoc lidera la lista en ubicación de los alojamientos con un 47.25 % mientras que la alcaldía Iztapalapa ocupa el último lugar con apenas un 0.76 % de alojamientos.

- En la siguiente también apreciamos claramente el cambio en la precisión de la variable v state con el 100 % de los registros en este estado.

Figura 7: Análisis Exploratorio: Alojamiento por Estado

- En estas gráficas observamos la normalización de la variable v property type para reducir las categorías, creando una categroría de otros con las menos frecuentes.

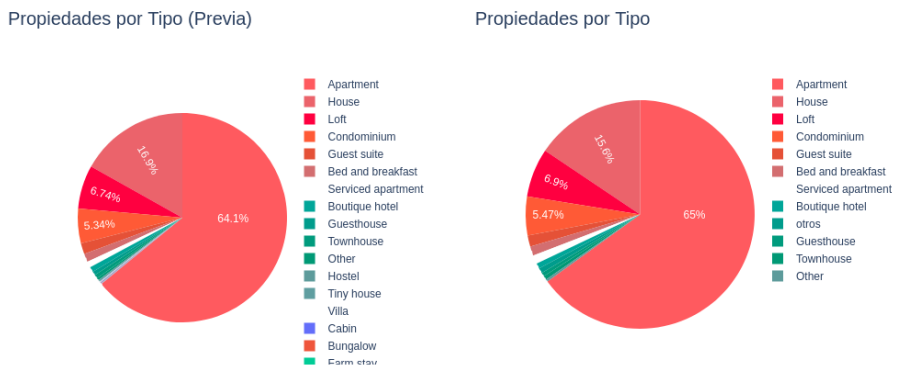


Figura 8: Análisis Exploratorio: Alojamiento por Tipo

También se observa que el 65 % de los alojamientos son departamentos seguido de un 15.6 % de casas, siendo estos dos tipos la gran mayoría de alojamientos.

- Ahora, ilustrando las variables de texto, en las siguientes nubes de palabras vemos como cambian los datos después del proceso de conformidad y consistencia.

Figura 9: Análisis Exploratorio: Descripción del Alojamiento

Lo más común al describir un alojamiento es colocar la ubicación, el tipo

de alojamiento y alguna comodidad respecto a la distancia o los servicios cercanos.

Figura 10: Análisis Exploratorio: Nombre de los Alojamientos

En cuanto al nombre de los alojamientos lo común es de igual manera colocar la ubicación y el tipo de alojamiento.

Para ver la comparación del resto de las variables consulte el apéndice.

5. Ingeniería de Variables

En esta parte básicamente se trata de crear variables a partir de las ya existentes.

5.1. Variable Descripción

Después del análisis exploratorio notamos que las variables t description y t summary se complementan ya que en ambas la información es muy similar por lo que para enriquecer el análisis se sumaron las variables para crear una sola en t description.

Figura 11: Ingeniería de Variables: Descripción del Alojamiento

5.2. Codificación a Nivel Nominal

En esta parte nos ayudamos de variables dummies para convertir variables categóricas en continuas.

Las variables transformadas fueron las siguientes: v host is superhost, v host identity verified y v has availability

Creándose las siguientes variables:

VARIABLE	TIPO	DESCRIPCIÓN
v host is superhost f	Int	0 : Si es super anfitrión / 1 : Si no es super anfitrión
v host is superhost t	Int	0 : Si no es super anfitrión / 1 : Si es super anfitrión
v host identity verified f	Int	0 : Si es anfitrión verificado / 1 : Si no es anfitrión verificado
v host identity verified t	Int	0 : Si no es anfitrión verificado / 1 : Si es anfitrión verificado
v has availability t	Int	0 : Si el alojamiento no tiene disponibilidad / 1 : Si el alojamiento tiene disponibilidad

Figura 12: Variables Dummies

5.3. Codificación a Nivel Ordinal

Esto consiste también en una forma de transformar variables categóricas en continuas, la diferencia con las nominales es que aquí debe haber un orden de importancia.

Para este análisis la variable que se utilizó fue v cancellation policy ya que en esta tabla de Airbnb contamos con 4 políticas diferentes:

- Flexible: Cancelación gratuita hasta 24 horas antes de la llegada.

- Moderada: Cancelación gratuita hasta 5 días antes de la llegada.
- Estricta: Cancelación gratuita durante 48 horas tras hacer la reserva, siempre y cuando se haga al menos 14 días antes de la llegada.
- Superestricta de 30 días: Los huéspedes pueden cancelar hasta 30 días antes de la llegada y recibir un reembolso del 50 % del precio por noche y los gastos de limpieza, aunque no de la tarifa de servicio

Por lo tanto en nuestra nueva variable la codificación quedó de la siguiente forma:

CATEGORIA	TRANSFORMACION
Flexible	0
Moderate	1
Strict 14 with grace period	2
Super Strict 30	3

Figura 13: Políticas de Cancelación

6. Outliers

En esta sección nos basamos en revisar que las variables continuas no cuenten con valores extremos, ya que esto puede entorpecer los análisis posteriores, sobretodo al imputar datos faltantes. En caso de existir valores extremos estos se eliminan de la tabla de datos.

En este análisis se utilizaron 3 métodos para revisar los datos extremos:

- IQR: Es un método matemático que a partir de los cuantiles I y III genera rangos para definir cotas que delimiten los valores de nuestras variables, es decir, los valores que se encuentren fuera de esas cotas se consideran outliers.
- Zcore: Esta técnica utiliza la desviación estandar de nuestras variables, se considerarán outliers aquellos datos que se encuentren fuera de un rango de -3 a 3 desviaciones.
- Percentiles: Aquí se revisan puntualmente los percentiles de cada variable, cuando se encuentran datos muy alejados de la multidud en los percentiles extremos, es decir datos fuera de un rango del percentil 5 y 95 se consideran outliers.

Una vez explicado lo anterior, se eliminaron los datos extremos considerando las variables que contaran con la aceptación de al menos dos de los metodos. En la siguiente tabla podemos ver los resultados:

Es decir, se eliminaron un total de 1030 registros quedándonos con el 68.67 % de la información de la tabla de datos inicial.

v_feature	c_n_rows
Inicial	3288
c_host_total_listings_count	3124
c_latitude	2980
c_longitude	2798
c_accommodates	2740
c_bathrooms	2654
c_bedrooms	2651
c_beds	2637
c_price	2557
c_guests_included	2510
c_minimum_nights	2399
c_maximum_nights	2393
c_number_of_reviews	2275
c_reviews_per_month	2258

Figura 14: Outliers

6.1. Visualzaciones

En esta sección vamos a ver gráficos del antes y después del tratamiento de Outliers.

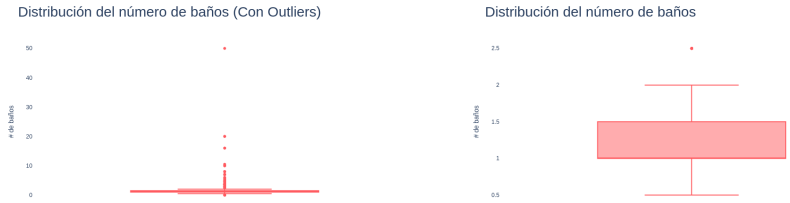


Figura 15: Outliers: Distribución del Número de Baños

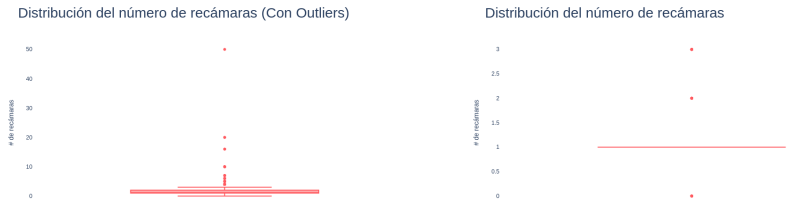


Figura 16: Outliers: Distribución del Número de Recámaras

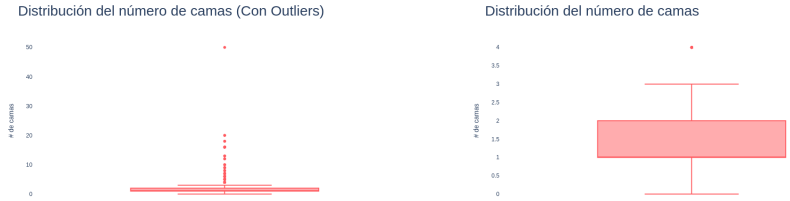


Figura 17: Outliers: Distribución del Número de Camas

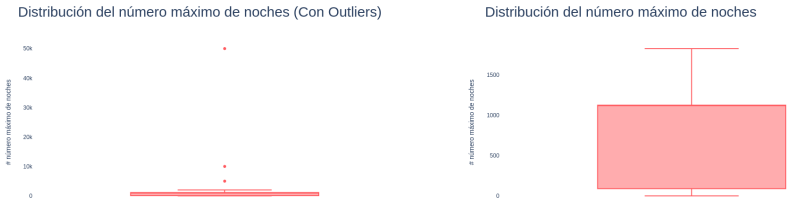


Figura 18: Outliers: Distribución del Número Máximo de Noches

Para ver el comportamiento del resto de las variables consulte el apéndice.

7. Datos Ausentes

Esta parte está dedicada 100 % a definir y tratar los datos ausentes. A continuación, todas las variables que contienen datos faltantes:

	% de missings
v_neighbourhood	0.046944
v_zipcode	0.072188
c_bathrooms	0.002657
c_bedrooms	0.000443
c_beds	0.002214
d_first_review	0.184677
d_last_review	0.184677
c_reviews_per_month	0.184677
v_city	0.075288

Figura 19: Completitud Después de Calidad de Datos

7.1. Eliminación por %

De acuerdo a la tabla anterior se decide eliminar aquellas variables que cuentan con más del 18 % de datos ausentes, ya que además no aportan información relevante al análisis y de imputar sus valores se podría alterar el mismo. Estas varibables son:

d first review y d last review

7.2. Imputación de Variables Categóricas

En esta parte se encontró la moda de las variables categóricas que cuentan con datos faltantes y se imputaron para así completar toda la información.

VARIABLE	MODA
v_neighbourhood	roma norte
v_zipcode	6700
v_city	CUAUHTEMOC

Figura 20: Moda

7.3. Imputación de Variables Continuas

En esta parte se calculó la media de las variables continuas que cuentan con datos faltantes y se imputaron para así completar toda la información.

VARIABLE	MEDIA
c_bathrooms	1.239786856
c_bedrooms	1.206025698
c_beds	1.400355082
c_reviews_per_month	0.914530147

Figura 21: Media

8. Reducción de Variables

Por último a través de diferentes técnicas, se revisó si podemos reducir la dimensión de nuestras variables para asegurar que la tabla final no contenga información duplicada o variables que se puedan predecir o explicar a través de otras.

8.1. Varianza Cero

Esta técnica consiste en revisar la varianza de cada variable, si esta es de 0 % significaría que sólo hay una categoría en esa variable y debe ser eliminada ya que no aporta relevancia en el análisis.

Para esta ocasión las variables eliminadas son las siguientes: v state, v country y v has availability t

8.2. Variables Dummies

Como ya vimos en otra sección, cuando se crean variables dummies se tiende a crear información duplicada con ellas ya que una variable dummie se podria explicar con la otra, por lo tanto siempre hay que eliminar una variable dummie por cada variable incicial transformada.

Para este caso se eliminaron las siguientes: v host is superhost f y v host identity verified f

8.3. Alta Correlación

Como última técnica utilizada tenemos Alta Correlación, la cual consiste en calcular la correlación entre variables, cuando esta es alta se puede decir que una de esas variables se puede explicar con la otra o que la información es muy similar con la misma tendencia, por tanto se decide eliminar una de esas variables.

En la siguiente imagen podemos apreciar la correlación entre nuestras variables:

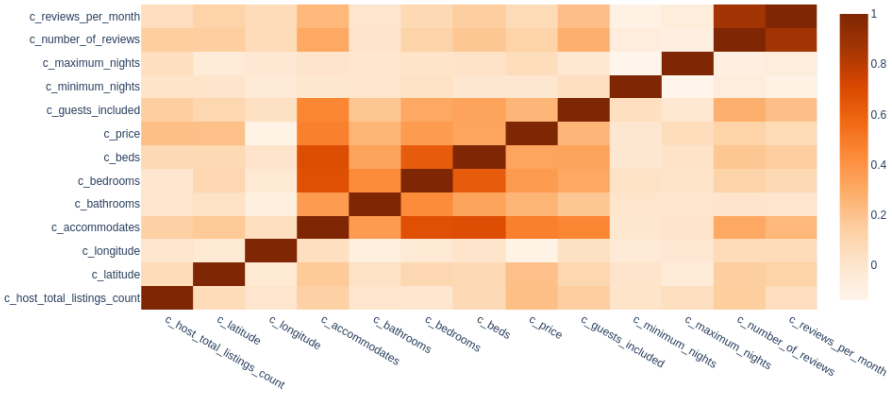


Figura 22: Correlación entre Variables

En particular, las variables c reviews per month y c number of reviews tienen una correlación mayor al 80 % por lo cual se decide eliminar una de ellas (c reviews per month)

9. Tabla Analítica Final

Finalmente nos quedamos con una tabla analítica de datos lista para algún análisis posterior que cuenta con 2258 registros y 26 atributos.

Referencias

[1] Inside Airbnb. http://insideairbnb.com/get-the-data.html

[2] El Financiero. https://www.elfinanciero.com.mx/bloomberg-businessweek/los-mexicanos-la-estan-rompiendo-en-airbnb, Julio 2017

[3] Airbnb MX. https://www.airbnb.mx/home/cancellation$_p oliciessuper - strict - 30$

[4] Datos Abiertos CDMX. https://datos.cdmx.gob.mx/explore/dataset/coloniascdmx/table/

A. Apéndice

A.1. Diccionario de Datos

A continuación la descripción del resto de variables iniciales.

VARIABLE	TIPO	DESCRIPCIÓN
latitude	float	Latitud de la ubicación del alojmiento
longitude	float	Longitud de la ubicación del alojmiento
guests included	int	Invitados incluidos en el alojamiento
reviews per month	float	Número de reviews del alojamiento por mes
last scraped	DateTime	Fecha de la última recopilación del registro
host since	DateTime	Fecha inicial del anfitrión en la plataforma
summary	Str	Resúmen del alojamiento
space	Str	Referencias sobre el alojamiento
review scores rating	float	Calificación del alojamiento
host id	int	Id del Anfitrión

Figura 23: Diccionario de Datos

A.2. Análisis Exploratorio

En ésta sección se puede observar la comparativa del comportamiento del resto de las variables antes y después de calidad de datos.

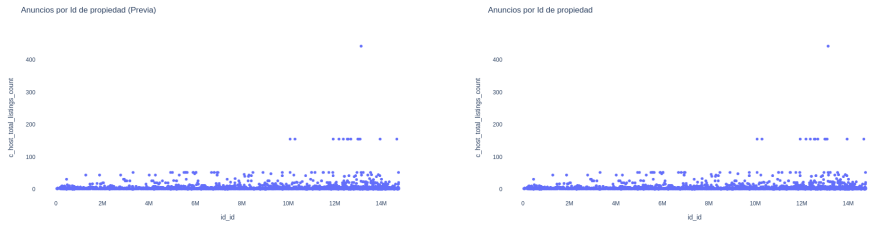


Figura 24: Análisis Exploratorio: Anuncios por Id de Propiedad

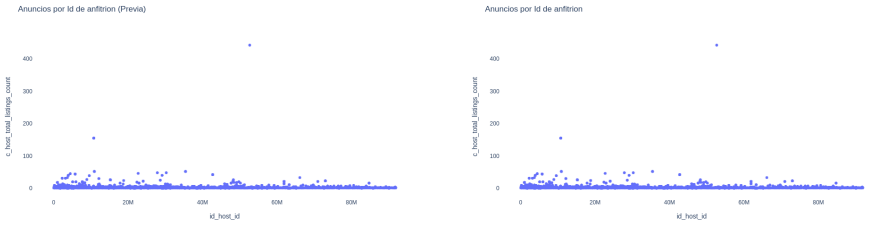


Figura 25: Análisis Exploratorio: Anuncios por Id de Anfitrión

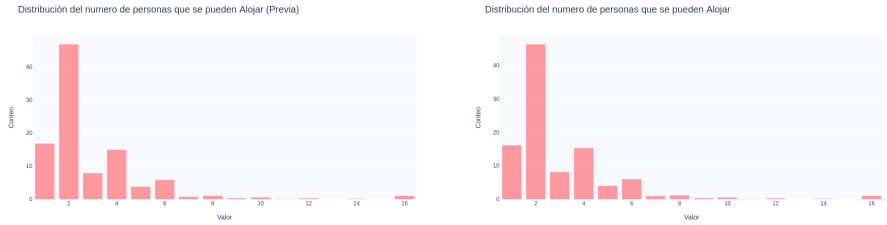


Figura 26: Análisis Exploratorio: Número de Huéspedes

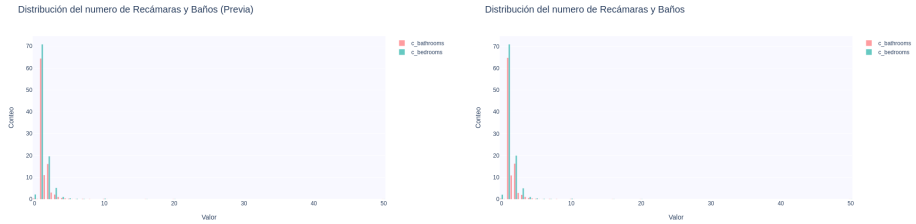


Figura 27: Análisis Exploratorio: Recámaras y Baños

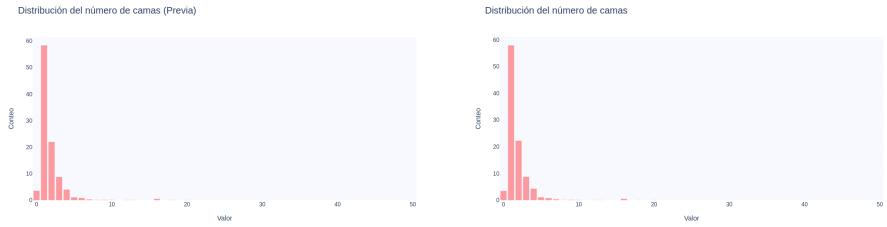


Figura 28: Análisis Exploratorio: Número de Camas

Figura 29: Análisis Exploratorio: Distribución de Precio

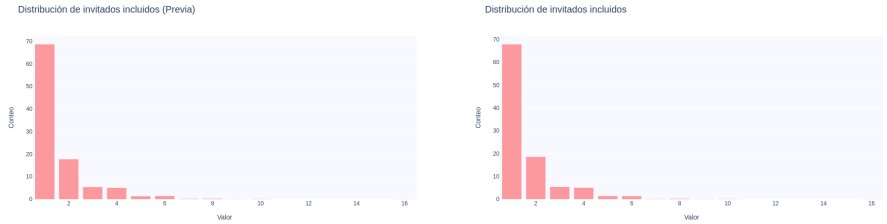


Figura 30: Análisis Exploratorio: Invitados Incluidos

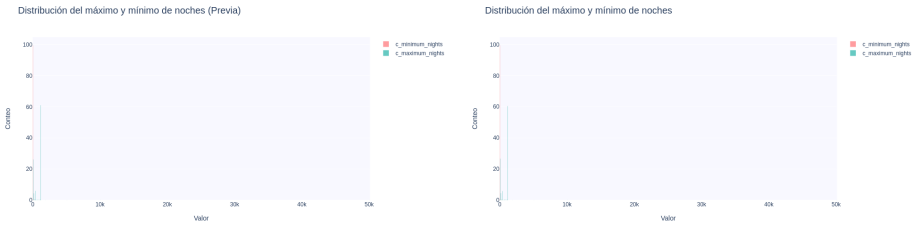


Figura 31: Análisis Exploratorio: Máximo/Mínimo de Noches

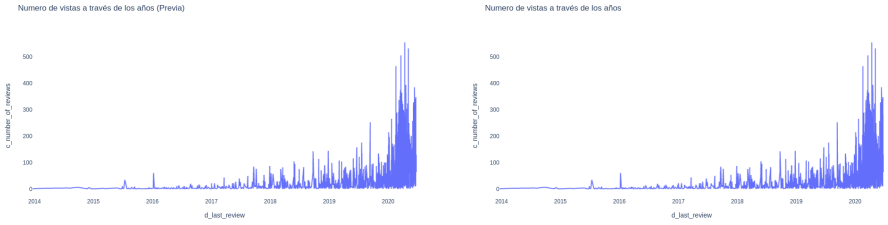


Figura 32: Análisis Exploratorio: Visitas a través de los años

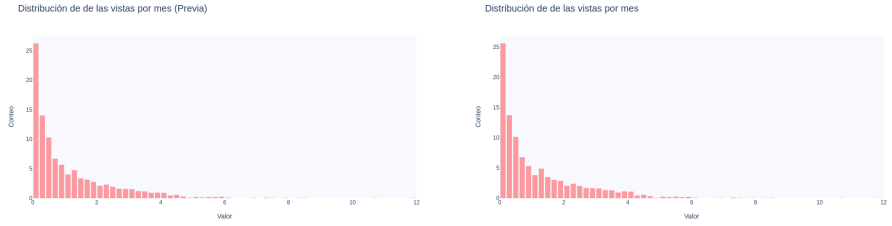


Figura 33: Análisis Exploratorio: Visitas por mes

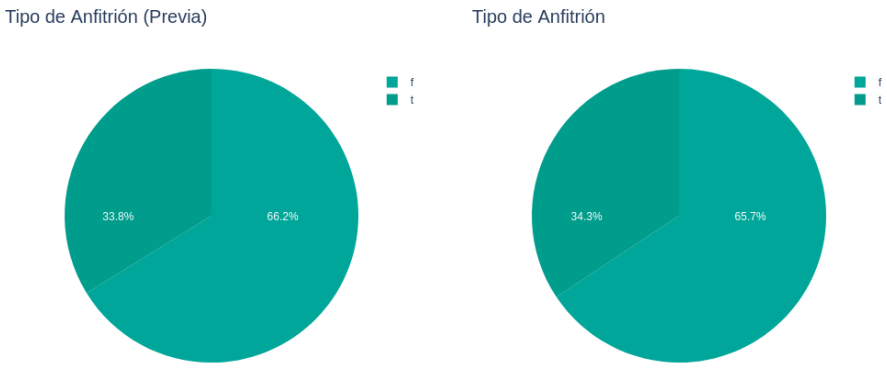


Figura 34: Análisis Exploratorio: Tipo de Anfitrión

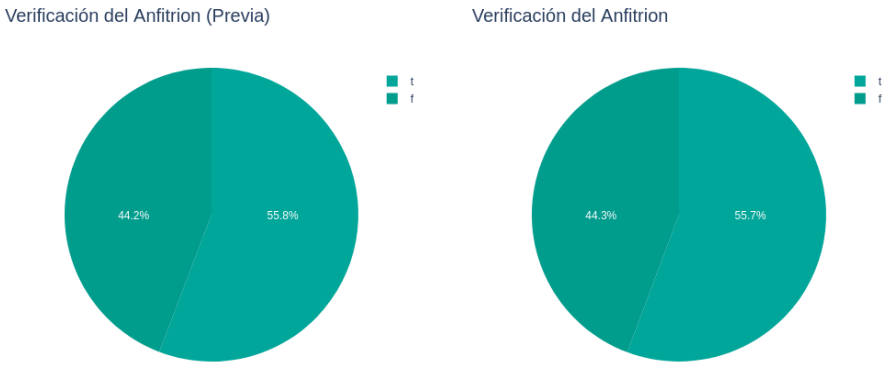


Figura 35: Análisis Exploratorio: Seguridad del Anfitrión

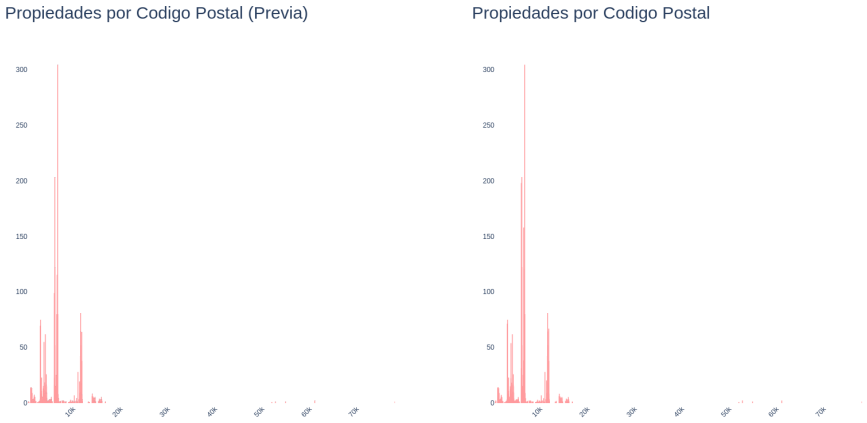


Figura 36: Análisis Exploratorio: Alojamiento por Código Postal

Figura 37: Análisis Exploratorio: Alojamiento por País

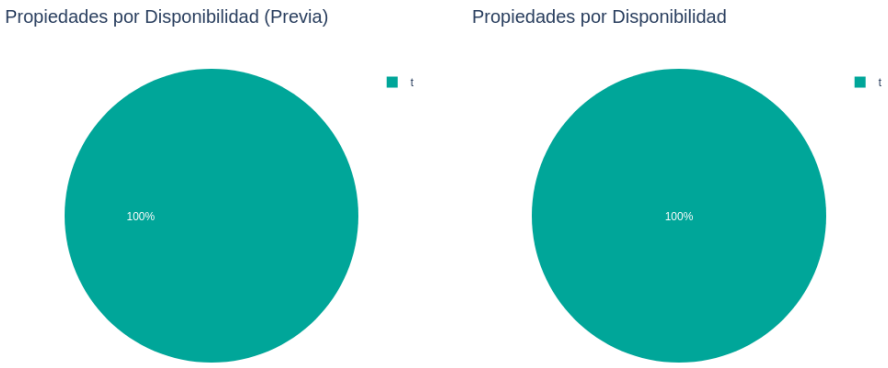


Figura 38: Análisis Exploratorio: Alojamiento por Disponibilidad

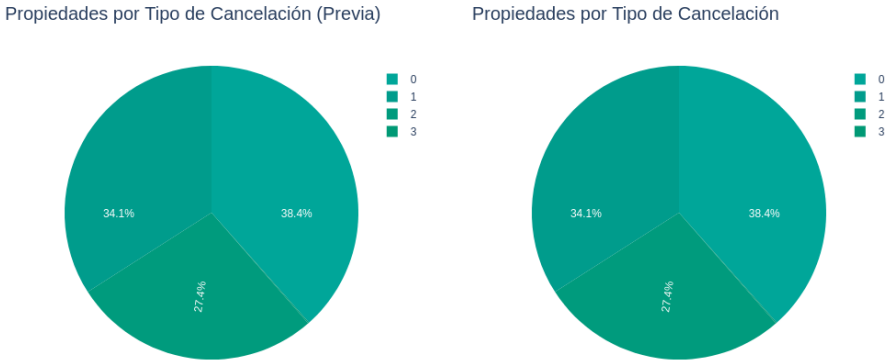


Figura 39: Análisis Exploratorio: Alojamiento por Tipo de Cancelación

Figura 40: Análisis Exploratorio: Comodidades

Figura 41: Análisis Exploratorio: Resúmen de los Alojamientos

A.3. Outliers

En ésta sección se puede observar la comparativa del comportamiento del resto de las variables antes y después del tratamiento de outliers.

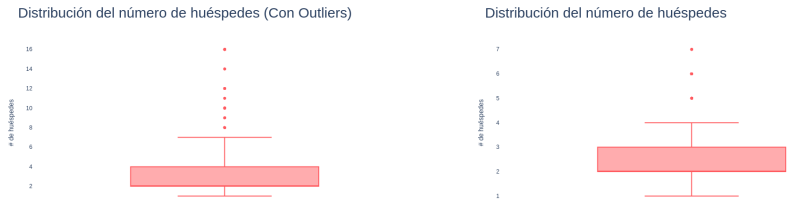


Figura 42: Outliers: Distribución del Número de Huéspedes

Figura 43: Outliers: Distribución del Número de Precio

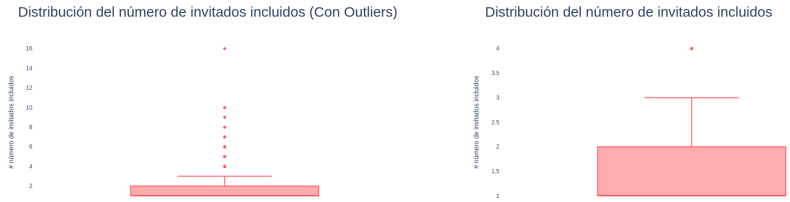


Figura 44: Outliers: Distribución del Número de Invitados Incluidos

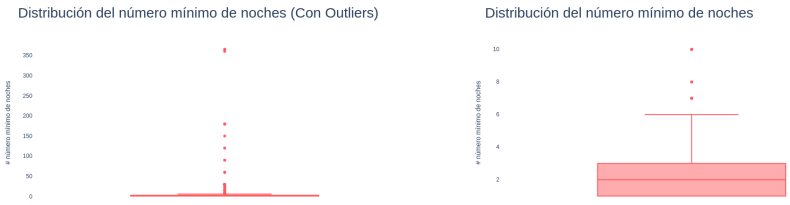


Figura 45: Outliers: Distribución del Número Mínimo de Noches

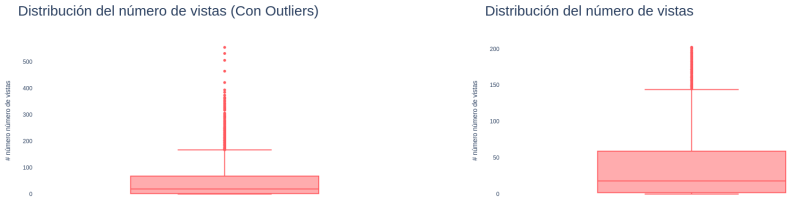


Figura 46: Outliers: Distribución del Número de Visitas

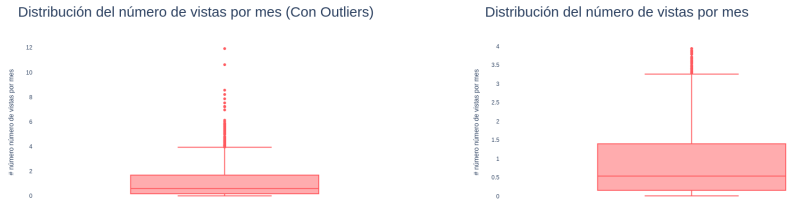


Figura 47: Outliers: Distribución del Número de Visitas por Mes