

New York City Airbnb

Diplomado en ciencia de datos

Erick Ivann De La Cruz Montes De Oca



airbnb





Contenido

Objetivo

Análisis exploratorio

Dataset

Outliers y missings

Calidad de datos

**Reducción de
dimensiones**



Objetivo

El objetivo de esta iniciativa es analizar la relación que existe entre la ubicación de un inmueble , el tipo de cuarto o vivienda y otras características con respecto al precio por el que se alquila a través de la plataforma AIRBNB a fin de poder predecir el precio en el que se podría rentar un inmueble nuevo en base a ciertas características



Dataset



Nuestro conjunto de datos cuenta con 16 columnas y 48895 registros



El conjunto de datos tiene lugar a lo largo de toda la ciudad de NY con presencia en 130 barrios distintos



Una de las variables de este conjunto de datos es el **precio** que es nuestra variable objetivo



Calidad de datos

Duplicidad

0%

No se encontraron registros duplicados.

Completitud

12

Solo 2 de las 16 variables no cumplian con el minimo de porcentaje de completitud.

Normalización

2

Se normalizaron 2 de las 16 variables que contiene el dataset

Al revisar la calidad de los datos se hicieron validaciones de duplicidad, completitud y consistencia de la informacion asi como limpieza de las variables de tipo texto y normalización de variables categóricas.



Vincent

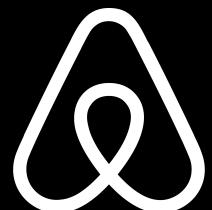
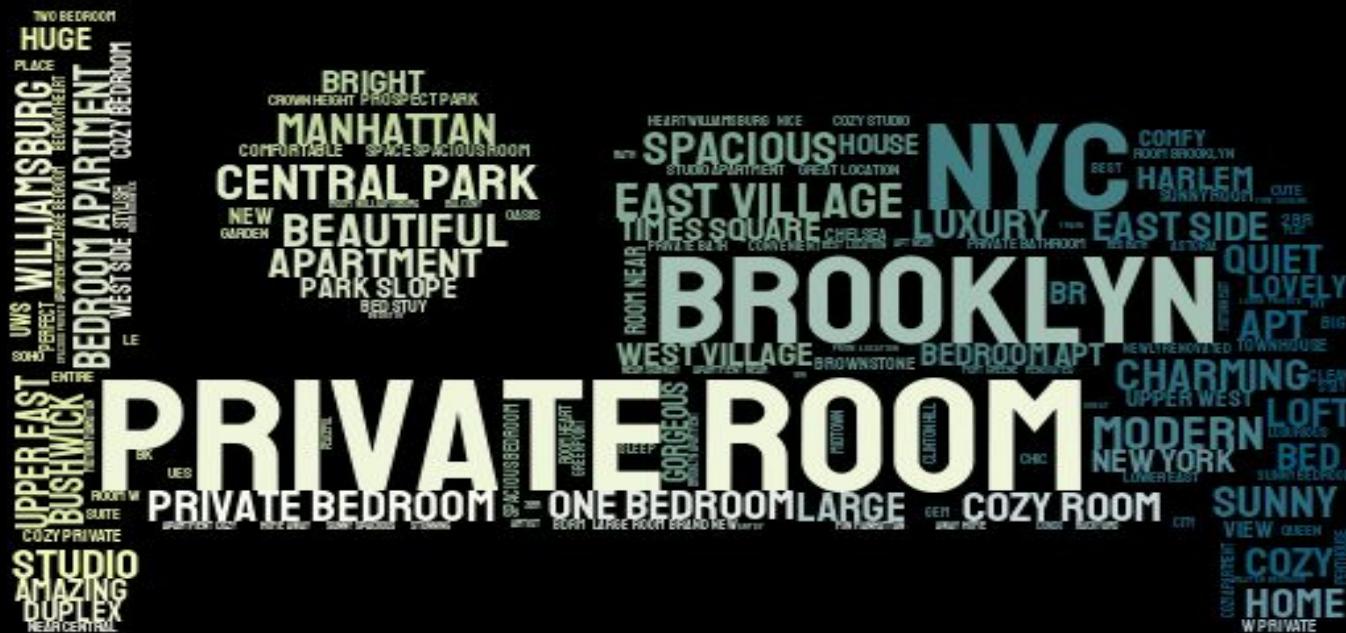
Análisis exploratorio





Nube de palabras

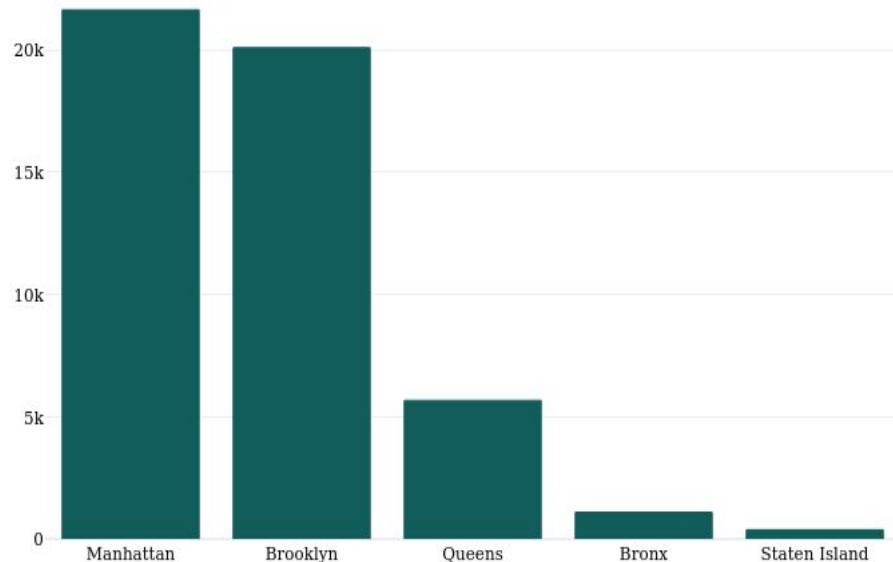
Se realizó un análisis de las palabras con mayor frecuencia dentro de la variable t_name





Distritos de NY

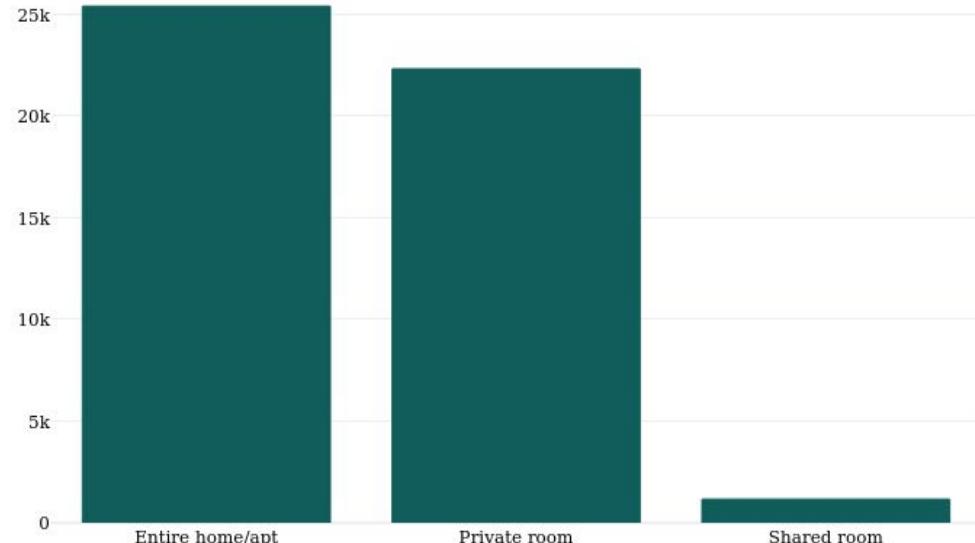
En Manhattan se concentra el mayor número de airbnb teniendo en segundo lugar a Brooklyn que en conjunto abarcan más del 80% de toda la información





Tipos de airbnb

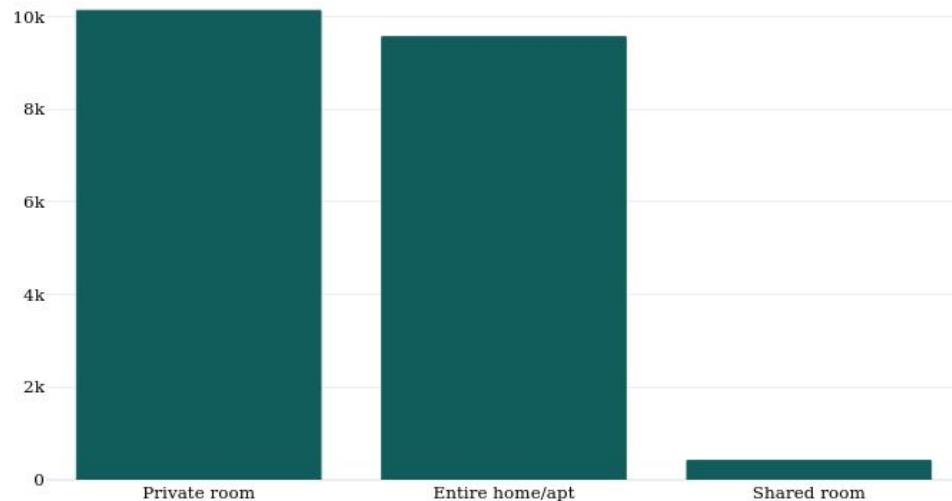
Dentro de los tipos de habitaciones o departamentos que se ofrecen en la plataforma para la ciudad de NY los departamentos o casas completas cubren casi el 50% de lo registros mientras que en segundo lugar son rentas de solo 1 habitación y sólo poco más del 2% son habitaciones compartidas.



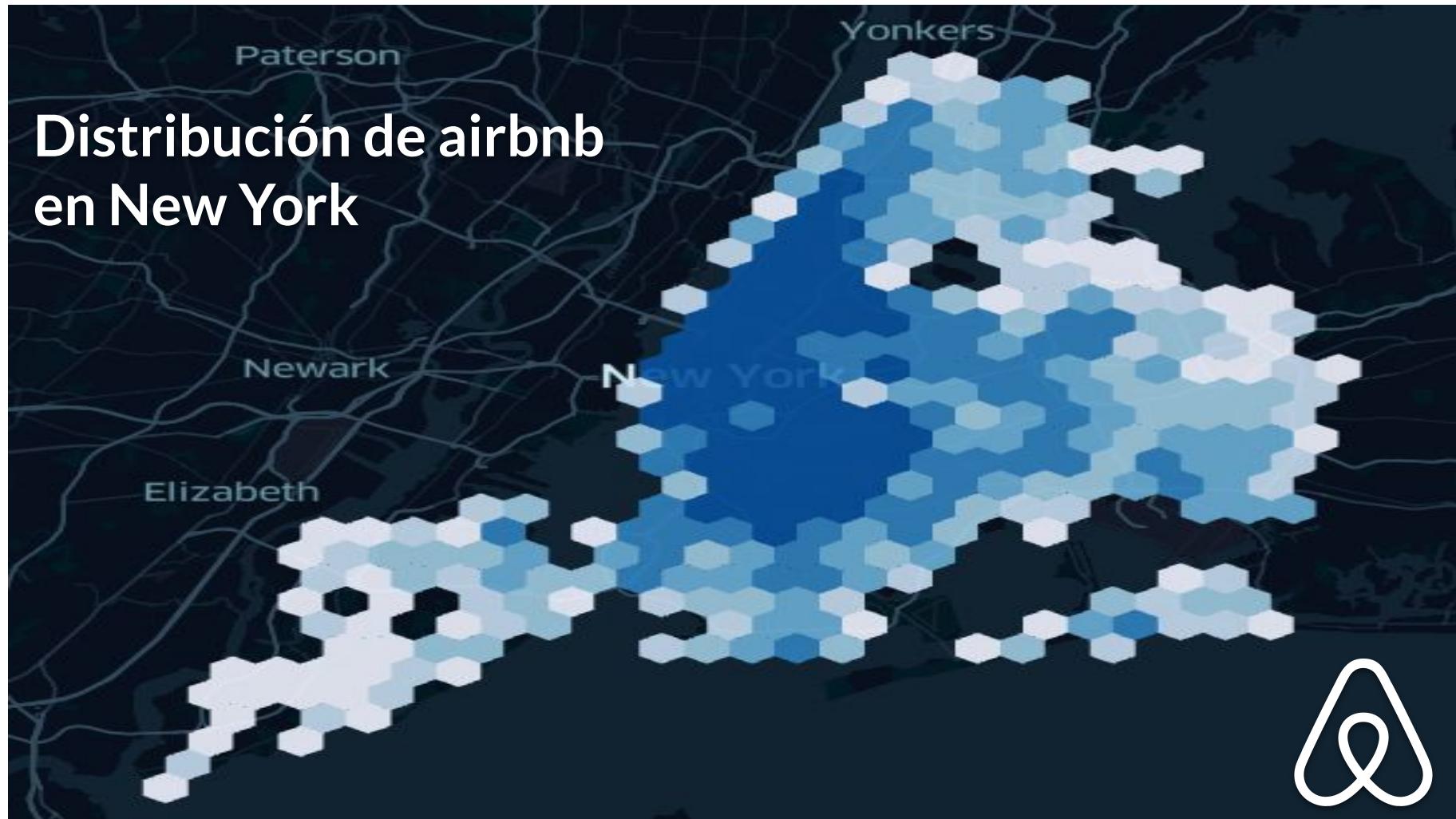


Tipos de airbnb en Brooklyn

La distribución de room_type se mantiene igual en todos los distritos con excepción de Brooklyn, donde aquí predominan los cuartos privados sobre las casa o departamentos completos



Distribución de airbnb en New York





Outliers y missings

Se removieron poco más de 6,000 registros de los casi 50,000 registros iniciales por métodos univariados.

Se trató de imputar la variable `host_name` creando un catálogo a partir de `host_id`, pero se encontró con que dichos registros eran valores únicos por lo que no se logró hacer match con el catálogo creado

	columna	total	completitud
0	<code>t_host_name</code>	21	99.957051
1	<code>t_name</code>	16	99.967277
2	<code>id</code>	0	100.000000
3	<code>host_id</code>	0	100.000000
4	<code>v_neighbourhood_group</code>	0	100.000000
5	<code>v_neighbourhood</code>	0	100.000000
6	<code>c_latitude</code>	0	100.000000
7	<code>c_longitude</code>	0	100.000000
8	<code>v_room_type</code>	0	100.000000
9	<code>tgt_price</code>	0	100.000000
10	<code>c_minimum_nights</code>	0	100.000000
11	<code>c_number_of_reviews</code>	0	100.000000
12	<code>c_calculated_host_listings_count</code>	0	100.000000
13	<code>c_availability_365</code>	0	100.000000



Reducción de dimensiones

Dado que se tienen pocas características solo se realizo un filtro de correlación donde se eliminaron 2 características que no alcanzaban el umbral mínimo de 0.05 de correlación con nuestra variable objetivo quedando así 7 variables que serán el insumo para realizar nuestra predicción.

