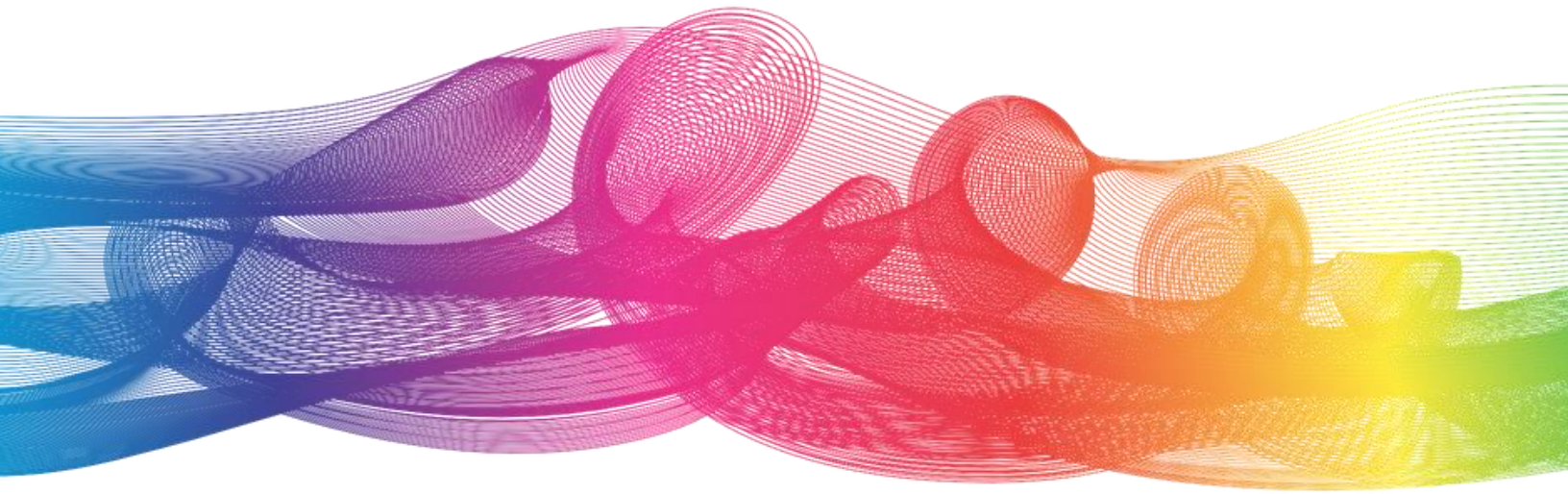


Instituto Politécnico Nacional

Escuela Superior de Cómputo

Grupo 3CV19 Data Mining

Profesor Zagal Flores Roberto Eswart

1er Parcial Practica #3: Definición del proyecto semestral, carga y exploración de datos

Alumno García González Aarón Antonio

Lunes 29 de marzo de 2021

Índice

Introducción ***.....3***

Desarrollo ***.....4***

Conclusiones ***.....17***

Introducción

Este proceso es particular para mi caso de estudio y de acuerdo con mi experiencia con procesamiento y limpieza de los datos, explico la propuesta de solución.

- Elección de tema principal
- Recopilación de todos los archivos a considerar en la muestra
- Limpieza por columna con python
- Unión de documentos con python
- Limpieza por tupla con manejador de base de datos
- Análisis en manejador de base de datos

Desarrollo

Procedimiento: Definir el alcance del proyecto semestral de datos, realizando un primer reconocimiento a la muestra de datos a elegir obtenida en el repositorio de la Ciudad de México.

1. Revise la clase que corresponda al tema “exploración básica de datos con Tableau” y el tema de “limpieza de datos”.
2. Explore las diferentes categorías de los conjuntos de datos abiertos de la Ciudad de México: https://datos.cdmx.gob.mx/.

   En la sesión de dudas comente mi interés por usar dataset de AirBnB, por lo cual la fuente de datos se encuentra en: http://insideairbnb.com/get-the-data.html y en la sección de Mexico City, Distrito Federal, Mexico.
3. Seleccione un conjunto de datos (dataset) que cumpla con las siguientes condiciones:

3.1. Tener al menos tres años de registros o tuplas. Que el dataset se pueda reducir su tamaño haciendo filtros por año (filtrando al año más reciente), en caso de que no sea posible procesar todos los registros.

3.2. El dataset debe contener al menos en la dimensión del tiempo “año” y “mes” como dimensión mínima de temporalidad.

3.3. La dimensión de espacio, al menos deben contener “delegación o alcaldía” y “coordenadas (latitud-longitud)”.

3.4. Que la cantidad de registros mínima del dataset debe ser 3 veces mayor al de incidentes viales usado en prácticas anteriores; es decir aproximadamente mayor a 90 mil registros. En caso de que el dataset en su tamaño original no pueda ser procesado, filtre los datos hasta que el dataset cumpla con este requisito. ES IMPORTANTE IMPORTAR LOS DATOS AL MANEJADOR DE SU PREFERENCIA PARA CONOCER SI ESTE REQUISITO SE CUMPLE.

3.5. Buscar una aplicación o caso de estudio de valor adicional del dataset elegido si este se complementa o se le integra información sobre el perfil de la población en la CDMX, está información será obtenida desde el sitio oficial del INEGI.

4. Explique en el reporte escrito, cómo el dataset elegido cumple cada uno de los requisitos del punto anterior.

4.1. Adicionalmente, explique cuántas dimensiones temáticas identificó en el dataset. Es importante identificar si el dataset cuenta con diccionario de datos. Por ejemplo, tipo de incidente vial, clasificación del origen del reporte, etc.

El volumen de datos elegidos es desde enero de 2020 hasta febrero de 2021, en total se encuentran 11 archivos CSV, cada uno incluye alrededor de 20K tuplas, por lo que aproximadamente se tienen 220K tuplas en total. Se encontraron 9 archivos CSV para el año 2020 y 2 archivos CSV para el 2021, cada uno de estos archivos representa un mes, por lo que hay meses que no se ven reportados. También cabe destacar que todos los archivos CSV tienen los mismos atributos o columnas.

Como comenté en la sesión de zoom, esta el caso donde la información omite algunos meses, por ejemplo: Abril, Marzo, Abril y Mayo si están los datos pero se salta hasta el mes de Agosto, es decir los meses de Junio y Julio no estarían reflejados, hay varias opciones a seguir:

1. Cuando se recopile la información complementaria por ejemplo de INEGI o del gobierno de la CDMX, contrastar o mantener el mismo rango de tiempos, es decir eliminar u omitir los meses que el data set principal de AirBnB no considera. *
2. Solo considerar los meses consecutivos que si están disponibles y eso mismo reflejarlo en la información secundaria.

Las dimensiones temáticas que incluye son la delegación, las coordenadas geográficas, el tipo de oferta, el precio, el mínimo de noches a rentar, el numero de calificaciones, la fecha de la ultima calificación o reseña, el numero de veces que se ha rentado en el mes, el total de días que se ofrece al año el servicio, así como el mes y año, hay un atributo el cual es el nombre del oferente, dado que un oferente puede ofertar mas de un airbnb pues igual podría ser llamado dimensión.

Con base a los datos recolectado parece ser que hay dos catálogos; la alcaldía y el tipo de servicio (cuarto de hotel, casa completa, habitación dentro de casa, etc.

4.2. Ponga especial atención en describir a detalle la granularidad temporal y espacial. Por ejemplo, el nivel de descripción del tiempo: día, mes, año minuto, segundo, etc.

A cada archivo que se encontró anteriormente es necesario agregar el campo mes y año correspondiente del cual se descargó, esto será necesario hacerlo previo a la integración de todos los archivos para la obtención del dataset a trabajar.

De igual manera se incluye un campo de fecha compuesto (día, mes y año), se intentará separar este dato en las unidades atómicas de información que lo conforman.

4.3. Explique a detalle el caso de estudio adicional donde el dataset elegido se pueda complementar con datos del perfil población de la CDMX. Por ejemplo, en el dataset de incidentes viales integrando el perfil poblacional, podemos conocerla relación entre la cantidad de incidentes y la cantidad de población que vive en la delegación Coyoacán.

Todos los archivos CVS contienen la dimensión “neighbourhood” que representa la alcaldía, igualmente se incluyen las coordenadas geográficas.

Yo espero que al contrastar los datos de localización del airbnb con el volumen turístico o datos del perfil de población de la CDMX, encontrar la relación entre el nivel de ingresos y el costo de airbnb, o tal vez la relación del turismo por delegación hospedada en airbnb.

4.4. Explique las razones o los motivos por las que ha elegido el dataset.

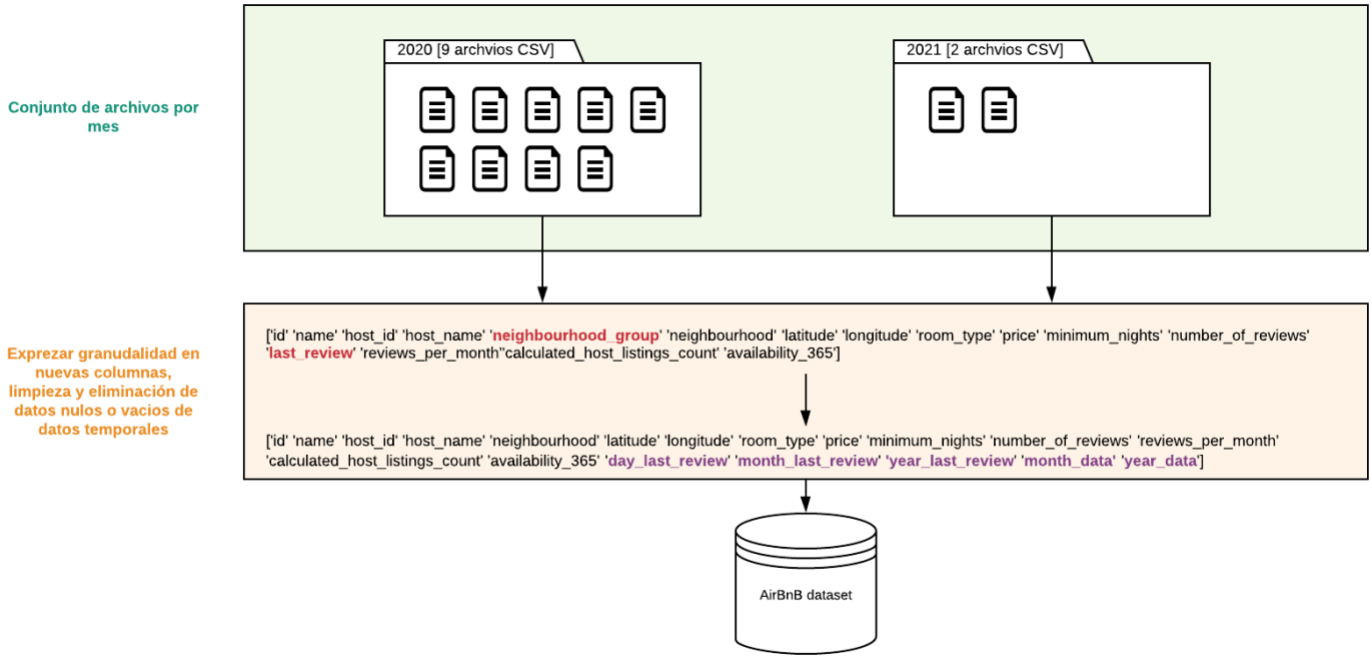
La razón principal de elegir airbnb como dataset principal se debe a que me gustaría ofertar en el futuro parte de un departamento en airbnb, que voy a rentar completo con algún arrendatario cuando me mude a la ciudad de México para comenzar a trabajar allá y si me funciona igual escalarlo a un negocio mas rentable como pasatiempo.

También busque primero los datasets que nos compartió de la ciudad de México, por alguna u otra razón no me convenció ninguno y no cumplían con los requerimientos que nos solicito para el curso, pero si voy a complementar con alguno de estos datasets.

4.5. Explique el problema que quiere resolver al explorar los datos. El alcance del proyecto: es decir explicar cual es el conocimiento que espero descubrir al estudiar el dataset.

Con base a los atributos que contiene el dataset, estoy interesado en caracterizar algunas de las dimensiones que dan mayor peso en que un airbnb sea exitoso o no lo sea, de igual manera al combinar o relacionar estos datos con otros datos por ejemplo del INEGI de la cantidad de personas que encontramos por delegación o con los datos de la ciudad de México en la sección de turismo, encontrar la relación de turismo con la renta de airbnb, específicamente utilizare el dataset de “Establecimientos de hospedaje por categoría” en el cual busco encontrar la relación e influencia que tienen los hoteles con respecto a los airbnb por delegación de la CDMX.

A manera general, describe el proceso de procesamiento y limpieza de datos realizado hasta este punto.

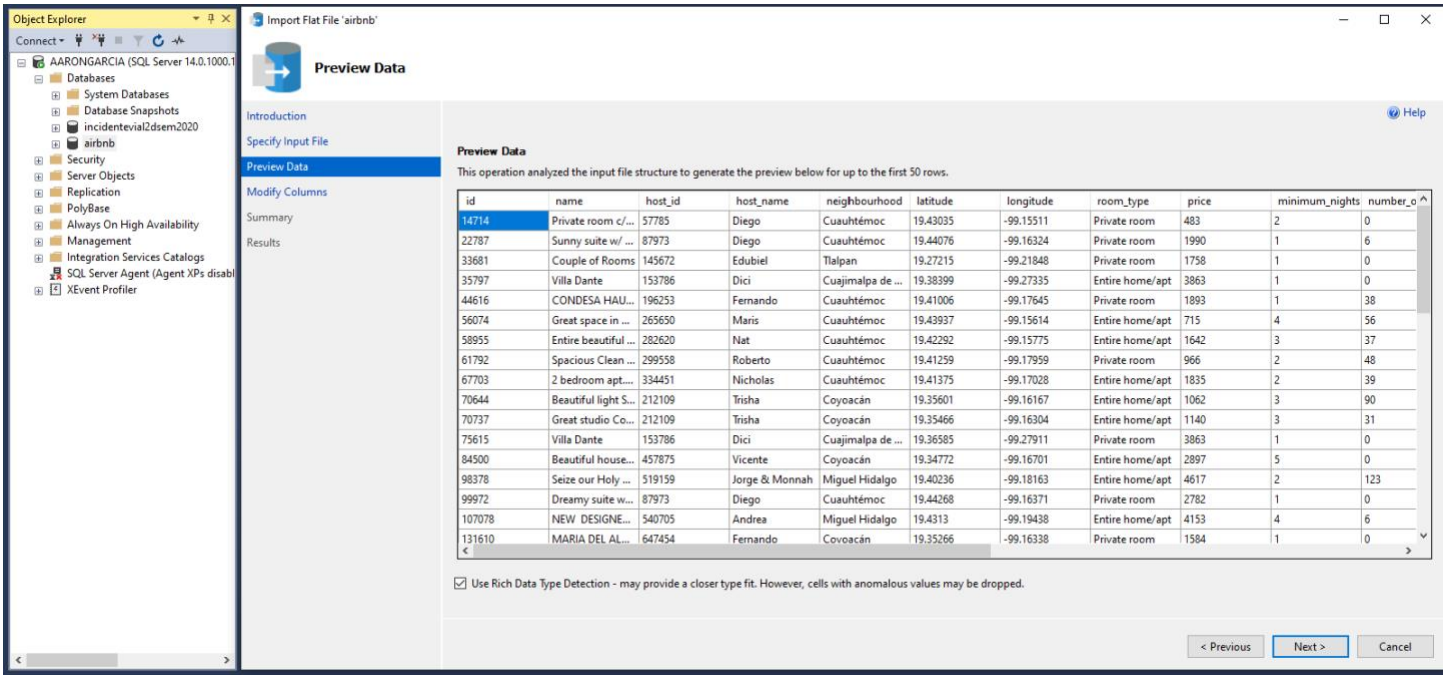


Me gustaría destacar que la ultima fecha de reseña, hay varias tuplas que están en blanco, esto se debe a que jamás han recibido una valoración, hay dos opciones, dejar en null o agregar una fecha comodín que no este dentro del rango de fechas que incluye esta dimensión, la segunda opción fue la que decidí seguir por mera presentación de datos, considero que es valido si y solo si soy yo la única persona que va a estar interactuando con esta base de datos, ya que conforme se pase el conocimiento, se podría llegar a perder este detalle.

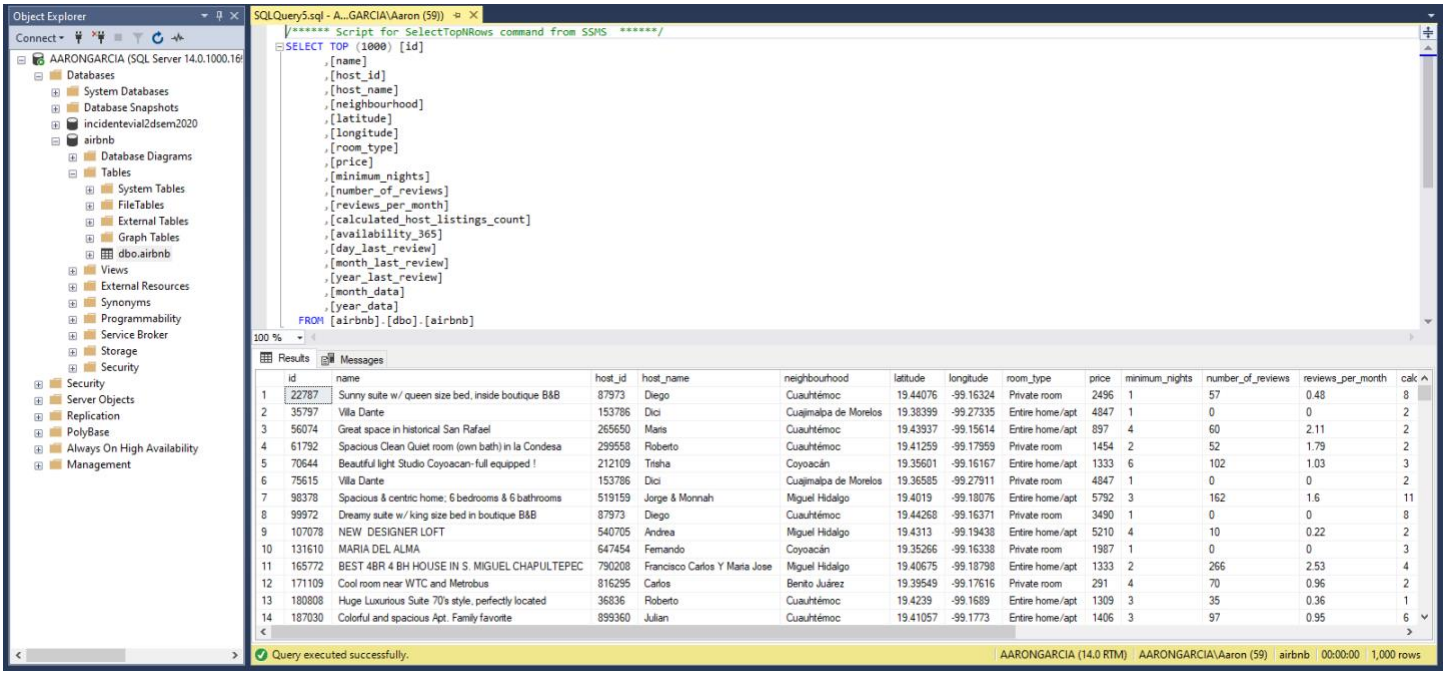
De igual manera describo el documento CSV resultante:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 228782 entries, 0 to 228781
Data columns (total 19 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   id                              228782 non-null  int64
 1   name                            228708 non-null  object
 2   host_id                         228782 non-null  int64
 3   host_name                       228756 non-null  object
 4   neighbourhood                   228782 non-null  object
 5   latitude                        228782 non-null  float64
 6   longitude                       228782 non-null  float64
 7   room_type                       228782 non-null  object
 8   price                           228782 non-null  int64
 9   minimum_nights                  228782 non-null  int64
 10  number_of_reviews               228782 non-null  int64
 11  reviews_per_month               228782 non-null  float64
 12  calculated_host_listings_count  228782 non-null  int64
 13  availability_365                228782 non-null  int64
 14  day_last_review                 228782 non-null  int64
 15  month_last_review               228782 non-null  int64
 16  year_last_review                228782 non-null  int64
 17  month_data                      228782 non-null  int64
 18  year_data                       228782 non-null  int64
dtypes: float64(3), int64(12), object(4)
memory usage: 33.2+ MB
None
```

Ahora vamos a analizarlo con sql server y limpiar las tuplas que tengas inconsistencias:



Vista previa de los datos, hasta este momento parece todo ir bien de acuerdo con el plan.



Se completo la importación de manera satisfactoria, se tienen 19 columnas con 228,781 filas de datos, en recursos de la computadora se tienen: Windows 10, 8 GB RAM e Intel Pentium por lo que se tardo aproximadamente 6 minutos en realizar la importación.

```
SQLQuery5.sql - A...GARCIA\Aaron (59))*
SELECT COUNT(*) AS 'Numero de registros'
FROM airbnb;

SELECT COUNT(DISTINCT id) AS 'Numero de vendedores diferentes'
FROM airbnb;
```

100 %

Results | Messages

	Numero de registros
1	228782

	Numero de vendedores diferentes
1	31454

Se tienen en total 228,792 registros de los cuales solo hay 31,45 vendedores o usuarios diferentes que ofertaron sus inmuebles en el ejercicio que estamos considerando.

```
SQLQuery5.sql - A...GARCIA\Aaron (59))*
SELECT COUNT(*) AS 'VALORES NULOS'
FROM airbnb
WHERE name IS NULL;

SELECT COUNT(*) AS 'VALORES NULOS STRING'
FROM airbnb
WHERE name = 'NULL';

SELECT *
FROM airbnb
WHERE name = 'NULL'
OR name IS NULL;

UPDATE airbnb
SET name = 'Desconocido'
WHERE name = 'NULL'
OR name IS NULL;

SELECT *
FROM airbnb
WHERE name = 'NULL'
OR name IS NULL;
```

100 %

Results | Messages

	VALORES NULOS
1	74

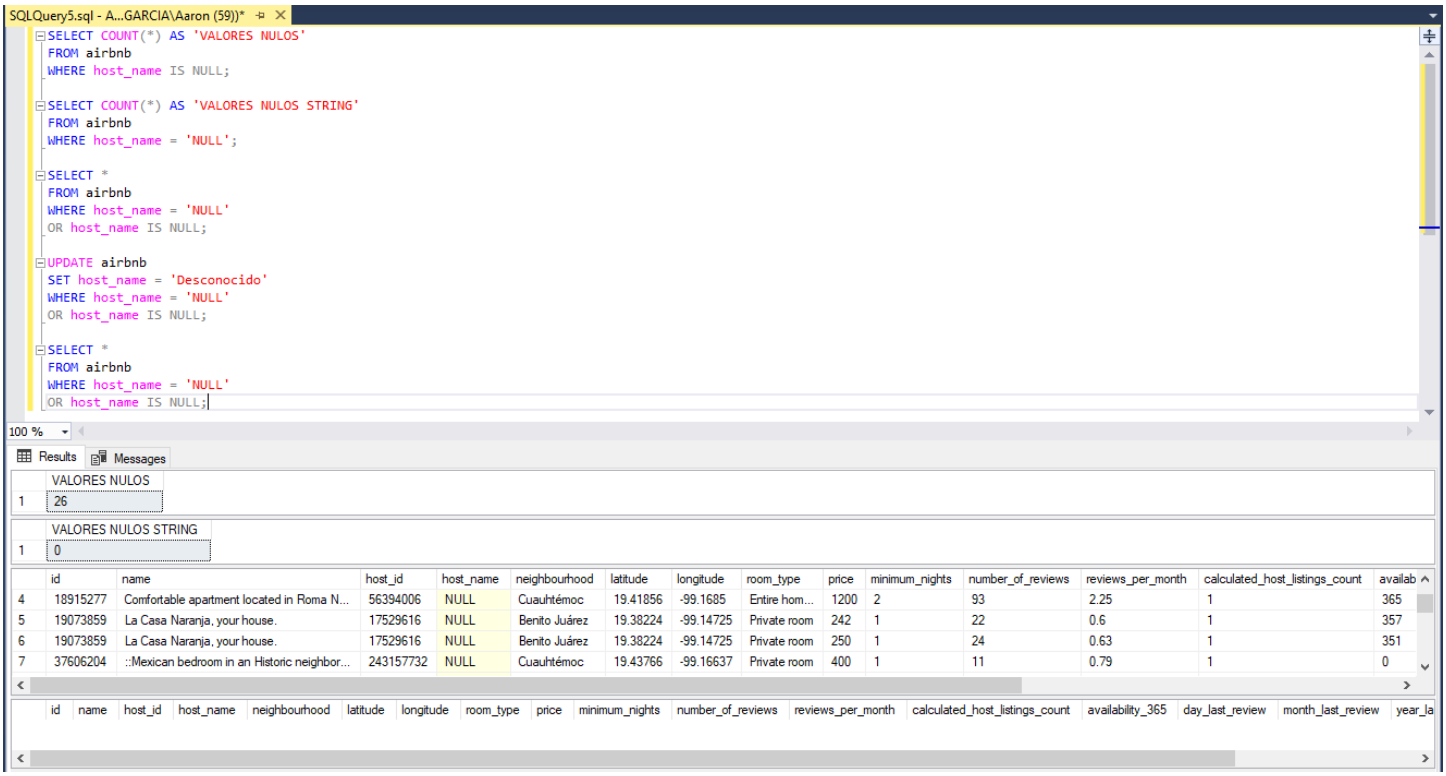
	VALORES NULOS STRING
1	0

	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	day_last_review
4	7902171	NULL	41666883	Yamel	Benito Juárez	19.38633	-99.14814	Private room	291	1	1	0.02	1	365	19
5	10765...	NULL	55677877	Humberto En...	Miguel Hidalgo	19.40031	-99.17431	Private room	24...	1	0	0	1	365	1
6	12158...	NULL	25512962	Álvaro	Benito Juárez	19.38078	-99.16407	Private room	236	4	0	0	2	0	1
7	13533...	NULL	2818226	Emiliano	Benito Juárez	19.36606	-99.16305	Private room	495	1	1	0.02	1	0	12
8	32524...	NULL	244396...	Alicia	Álvaro Obreg...	19.34433	-99.19149	Entire home/apt	460	28	2	0.18	1	179	1

	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	day_last_review	month_last_review	yea

Query executed successfully. | AARONGARCIA (14.0 RTM) | AARONGARCIA\Aaron (59) | airbnb | 00:00:02 | 76 rows

Cuando se busca en la columna 'name' por valores nulos se encontraron 76 registros con valor nulo, dado que este dato no es una dimensión para el análisis de datos, procedí en actualizar esos valores por 'Desconocido', ya que los demás campos parecen ser coherentes.



```
SELECT COUNT(*) AS 'VALORES NULOS'
 FROM airbnb
 WHERE host_name IS NULL;

SELECT COUNT(*) AS 'VALORES NULOS STRING'
 FROM airbnb
 WHERE host_name = 'NULL';

SELECT *
 FROM airbnb
 WHERE host_name = 'NULL'
 OR host_name IS NULL;

UPDATE airbnb
 SET host_name = 'Desconocido'
 WHERE host_name = 'NULL'
 OR host_name IS NULL;

SELECT *
 FROM airbnb
 WHERE host_name = 'NULL'
 OR host_name IS NULL;
```

Results | Messages

	VALORES NULOS
1	26

	VALORES NULOS STRING
1	0

	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availab
4	18915277	Comfortable apartment located in Roma N...	56394006	NULL	Cuauhtémoc	19.41856	-99.1685	Entire hom...	1200	2	93	2.25	1	365
5	19073859	La Casa Naranja, your house.	17529616	NULL	Benito Juárez	19.38224	-99.14725	Private room	242	1	22	0.6	1	357
6	19073859	La Casa Naranja, your house.	17529616	NULL	Benito Juárez	19.38224	-99.14725	Private room	250	1	24	0.63	1	351
7	37606204	::Mexican bedroom in an Historic neighbor...	243157732	NULL	Cuauhtémoc	19.43766	-99.16637	Private room	400	1	11	0.79	1	0

	id	name	host_id	host_name	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	day_last_review	month_last_review	year_la

Cuando se busca en la columna 'host_name por valores nulos se encontraron WCECECEC registros con valor nulo, dado que este dato no es una dimensión para el análisis de datos, procedí en actualizar esos valores por 'Desconocido', ya que los demás campos parecen ser coherentes.

Hasta este punto ya revisamos todas tuplas por cada atributo y se ha limpiado la base de datos, por lo que ahora consideramos una segunda versión de nuestro dataset, este será el que usaremos en Tableau.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 228782 entries, 0 to 228781
Data columns (total 19 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   id                              228782 non-null  int64
 1   name                            228708 non-null  object
 2   host_id                         228782 non-null  int64
 3   host_name                       228756 non-null  object
 4   neighbourhood                   228782 non-null  object
 5   latitude                        228782 non-null  float64
 6   longitude                       228782 non-null  float64
 7   room_type                       228782 non-null  object
 8   price                           228782 non-null  int64
 9   minimum_nights                  228782 non-null  int64
 10  number_of_reviews               228782 non-null  int64
 11  reviews_per_month               228782 non-null  float64
 12  calculated_host_listings_count  228782 non-null  int64
 13  availability_365                228782 non-null  int64
 14  day_last_review                 228782 non-null  int64
 15  month_last_review               228782 non-null  int64
 16  year_last_review                228782 non-null  int64
 17  month_data                      228782 non-null  int64
 18  year_data                       228782 non-null  int64
dtypes: float64(3), int64(12), object(4)
memory usage: 33.2+ MB
None
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 228782 entries, 0 to 228781
Data columns (total 19 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   id                              228782 non-null  int64
 1   name                            228782 non-null  object
 2   host_id                         228782 non-null  int64
 3   host_name                       228782 non-null  object
 4   neighbourhood                   228782 non-null  object
 5   latitude                        228782 non-null  float64
 6   longitude                       228782 non-null  float64
 7   room_type                       228782 non-null  object
 8   price                           228782 non-null  int64
 9   minimum_nights                  228782 non-null  int64
 10  number_of_reviews               228782 non-null  int64
 11  reviews_per_month               228782 non-null  float64
 12  calculated_host_listings_count  228782 non-null  int64
 13  availability_365                228782 non-null  int64
 14  day_last_review                 228782 non-null  int64
 15  month_last_review               228782 non-null  int64
 16  year_last_review                228782 non-null  int64
 17  month_data                      228782 non-null  int64
 18  year_data                       228782 non-null  int64
dtypes: float64(3), int64(12), object(4)
memory usage: 33.2+ MB
None
```

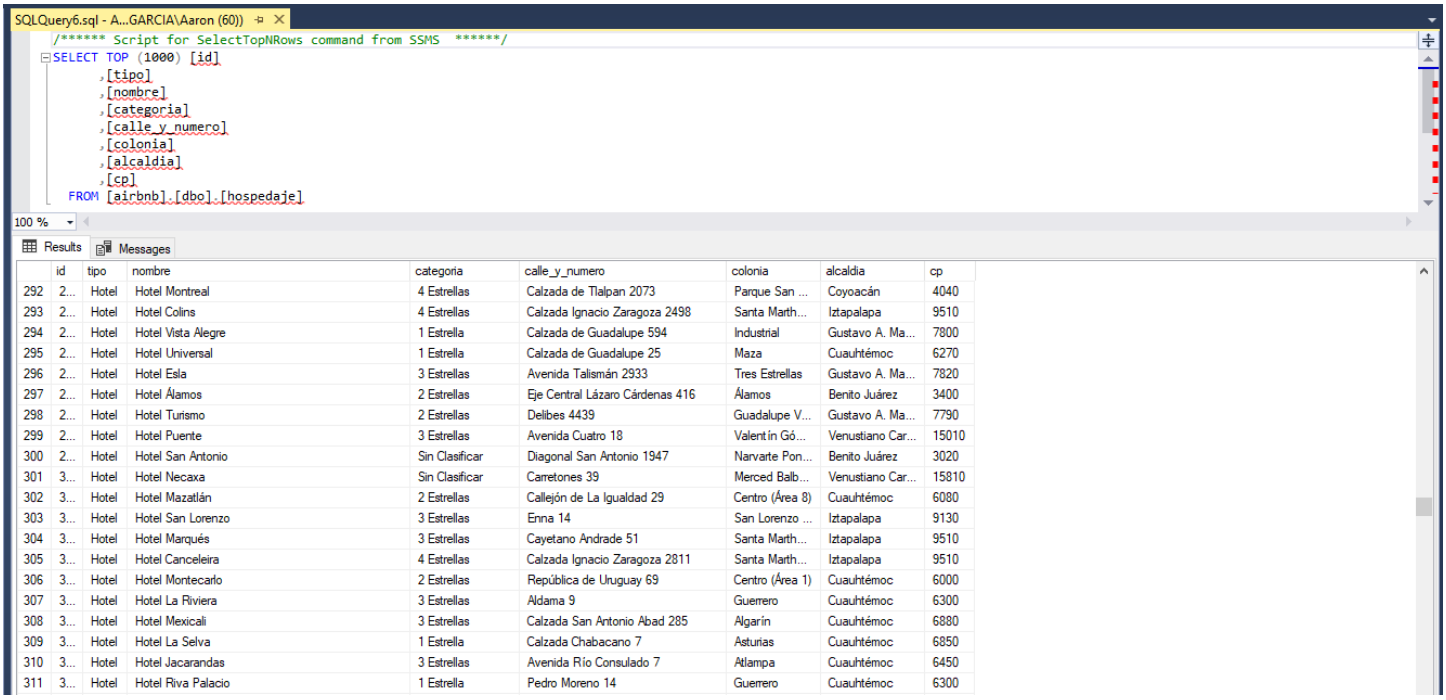
Comparativo de información rápida obtenida del dataset original respecto a la segunda versión, ahora si tenemos las 228,781 tuplas para todos los atributos.

Con respecto a los datasets, ahora muestro el dataset complementario, decidí utilizar el dataset “Establecimientos de hospedaje por categoría” de https://datos.cdmx.gob.mx/dataset/establecimientos-de-hospedaje-por-categoria.

El cual incluye el siguiente diccionario de datos:

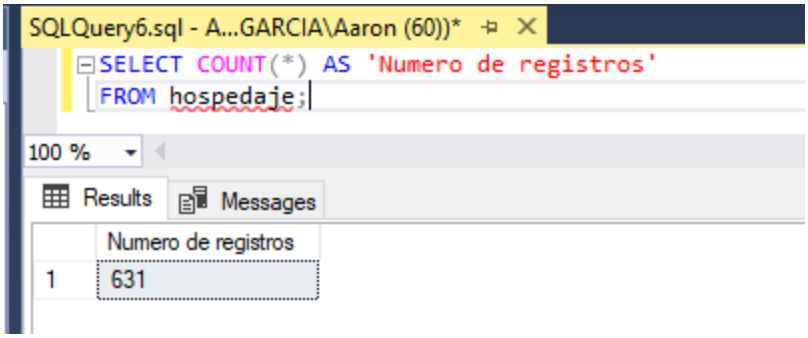
Diccionario de datos

Columna	Tipo
id	numeric
tipo	text
nombre	text
categoria	text
calle_y_numero	text
colonia	text
alcaldia	text
cp	numeric

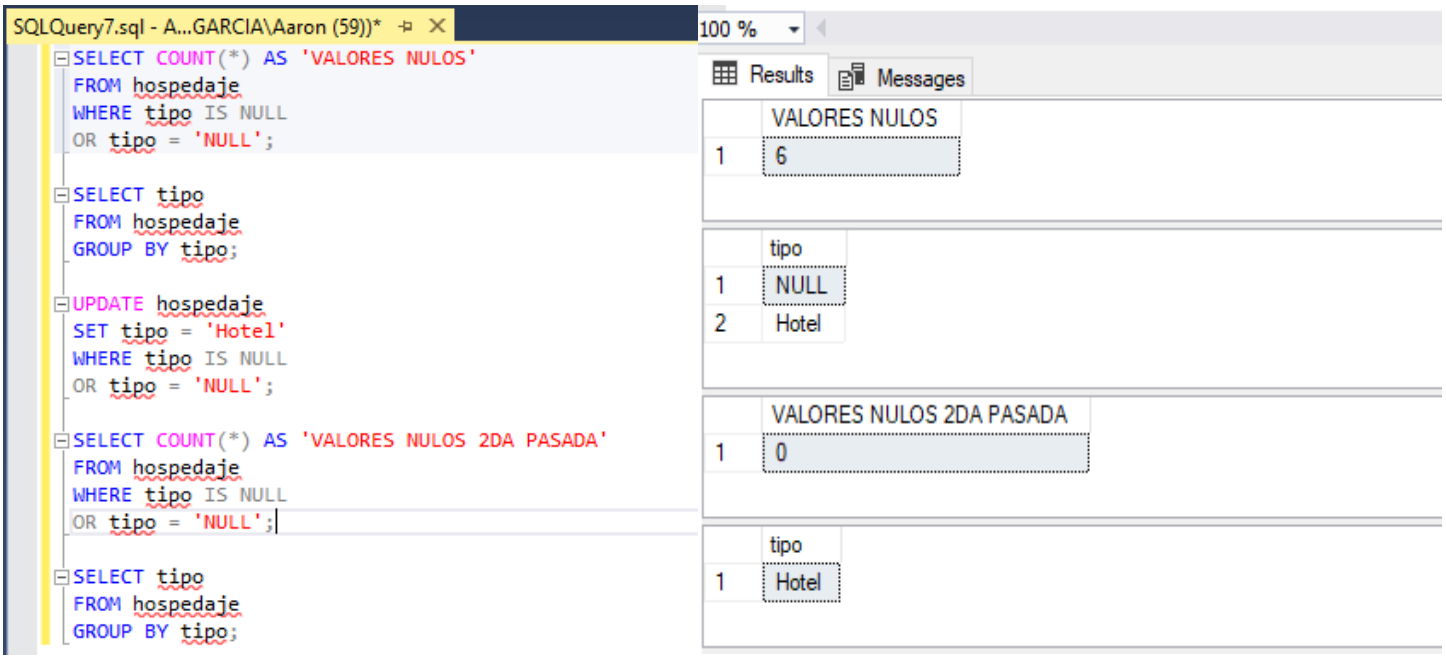


	id	tipo	nombre	categoria	calle_y_numero	colonia	alcaldia	cp
292	2...	Hotel	Hotel Montreal	4 Estrellas	Calzada de Tlalpan 2073	Parque San ...	Coyoacán	4040
293	2...	Hotel	Hotel Colins	4 Estrellas	Calzada Ignacio Zaragoza 2498	Santa Marth...	Iztapalapa	9510
294	2...	Hotel	Hotel Vista Alegre	1 Estrella	Calzada de Guadalupe 594	Industrial	Gustavo A. Ma...	7800
295	2...	Hotel	Hotel Universal	1 Estrella	Calzada de Guadalupe 25	Maza	Cuauhtémoc	6270
296	2...	Hotel	Hotel Esla	3 Estrellas	Avenida Talismán 2933	Tres Estrellas	Gustavo A. Ma...	7820
297	2...	Hotel	Hotel Álamos	2 Estrellas	Eje Central Lázaro Cárdenas 416	Álamos	Benito Juárez	3400
298	2...	Hotel	Hotel Turismo	2 Estrellas	Delibes 4439	Guadalupe V...	Gustavo A. Ma...	7790
299	2...	Hotel	Hotel Puente	3 Estrellas	Avenida Cuatro 18	Valentín Gó...	Venustiano Car...	15010
300	2...	Hotel	Hotel San Antonio	Sin Clasificar	Diagonal San Antonio 1947	Narvarte Pon...	Benito Juárez	3020
301	3...	Hotel	Hotel Necaxa	Sin Clasificar	Carretones 39	Merced Balb...	Venustiano Car...	15810
302	3...	Hotel	Hotel Mazatlán	2 Estrellas	Callejón de La Igualdad 29	Centro (Área 8)	Cuauhtémoc	6080
303	3...	Hotel	Hotel San Lorenzo	3 Estrellas	Enna 14	San Lorenzo ...	Iztapalapa	9130
304	3...	Hotel	Hotel Marqués	3 Estrellas	Cayetano Andrade 51	Santa Marth...	Iztapalapa	9510
305	3...	Hotel	Hotel Canceleira	4 Estrellas	Calzada Ignacio Zaragoza 2811	Santa Marth...	Iztapalapa	9510
306	3...	Hotel	Hotel Montecarlo	2 Estrellas	República de Uruguay 69	Centro (Área 1)	Cuauhtémoc	6000
307	3...	Hotel	Hotel La Riviera	3 Estrellas	Aldama 9	Guerrero	Cuauhtémoc	6300
308	3...	Hotel	Hotel Mexicali	3 Estrellas	Calzada San Antonio Abad 285	Algarín	Cuauhtémoc	6880
309	3...	Hotel	Hotel La Selva	1 Estrella	Calzada Chabacano 7	Asturias	Cuauhtémoc	6850
310	3...	Hotel	Hotel Jacarandas	3 Estrellas	Avenida Río Consulado 7	Atlampa	Cuauhtémoc	6450
311	3...	Hotel	Hotel Riva Palacio	1 Estrella	Pedro Moreno 14	Guerrero	Cuauhtémoc	6300

Hacemos la primera visualización do consulta de los datos.



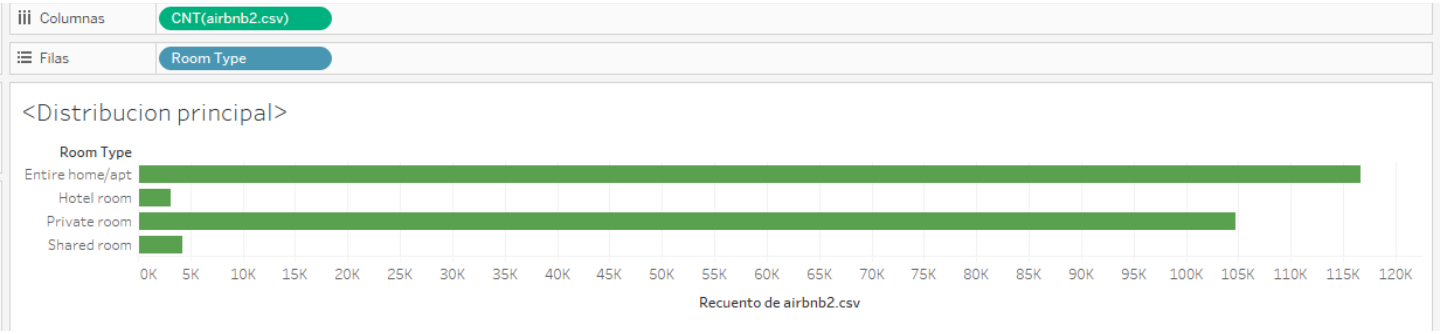
Hay un total de 631 hoteles registradoe en el dataset.



En el campo 'tipo' se actualizaron al otro valor posible que es "Hotel", esto ya que todos os registros corresponden a hoteles y al continuar revisando los demás atributos parece que todo esta bien, por lo que continuaremos ahora con el análisis en tableau.

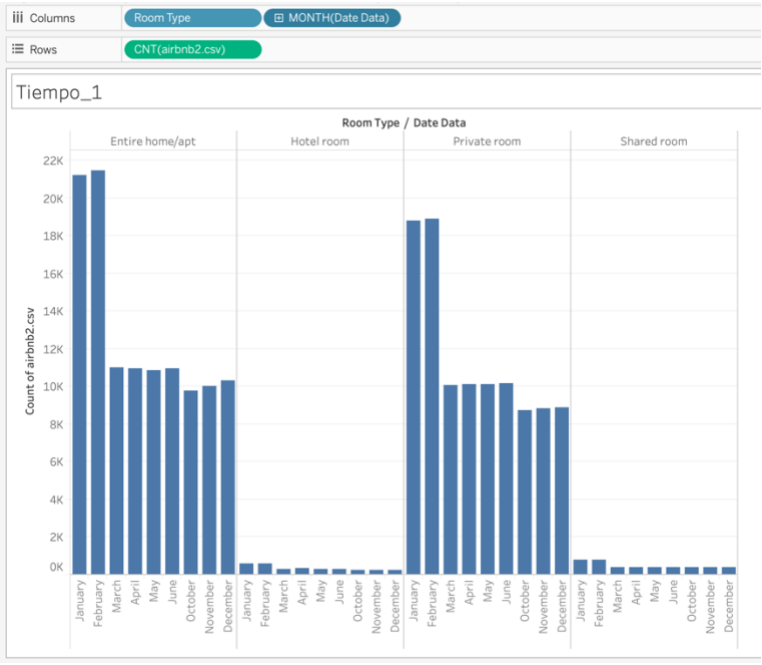
5. Realice el análisis exploratorio básico usando Tableau, contestando las siguientes preguntas generales. Responda aplicando su propio criterio, es decir filtrando la información como considere conveniente. Agregue los resultados en el reporte.

5.1. ¿Cual es la distribución de la dimensión categorial o temática (el tema del dataset) más importante (del fenómeno que es descrito por el dataset)? Ej. La distribución general de incidentes viales.

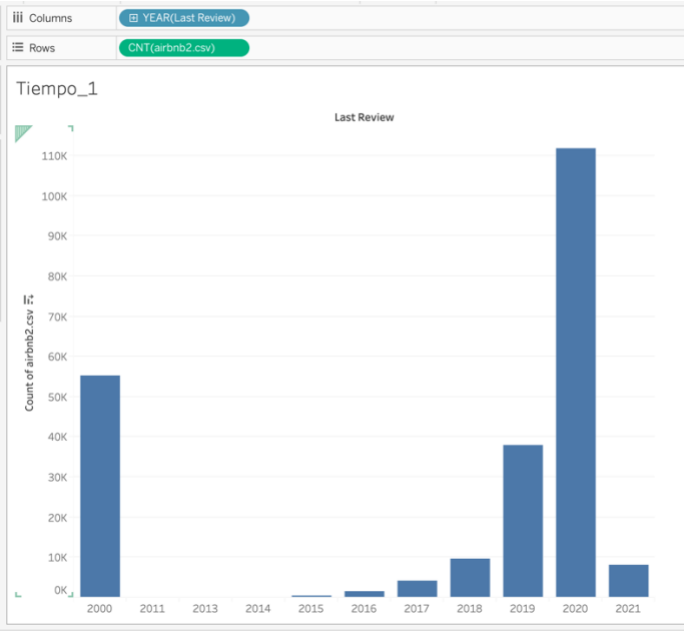


Considero que la categoría temática mas importante es el tipo de Airbnb que se oferta.

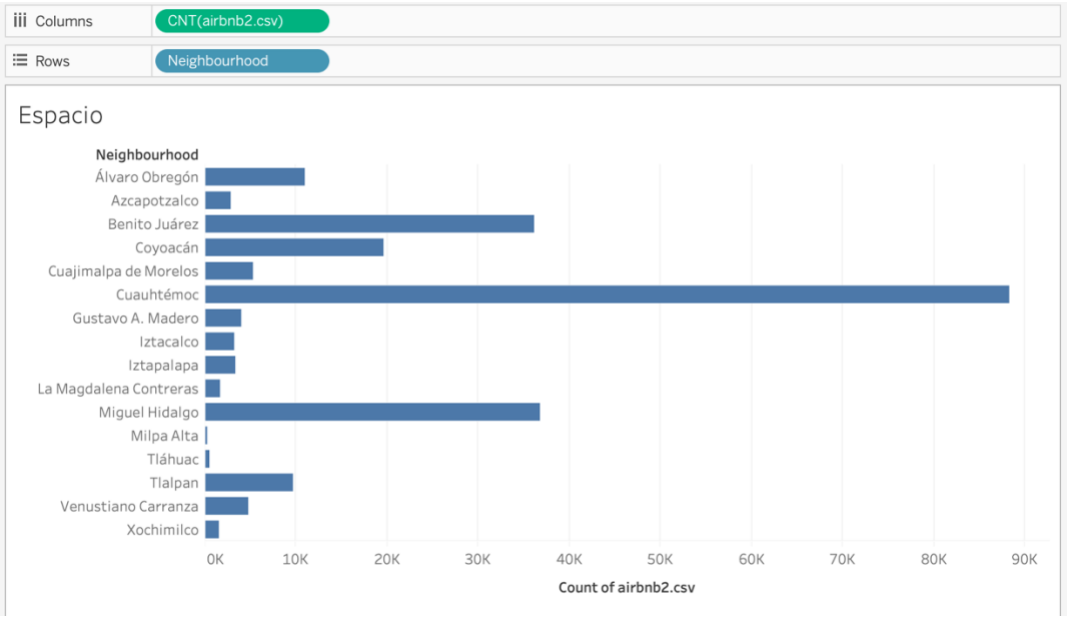
5.2. ¿Cual es la distribución del fenómeno que mide el dataset en el tiempo?, explorar la mayor cantidad de los niveles de granularidad de tiempo. Ej. La distribución anual de incidentes viales por mes.



Distribución de tipo de airbnb por mes

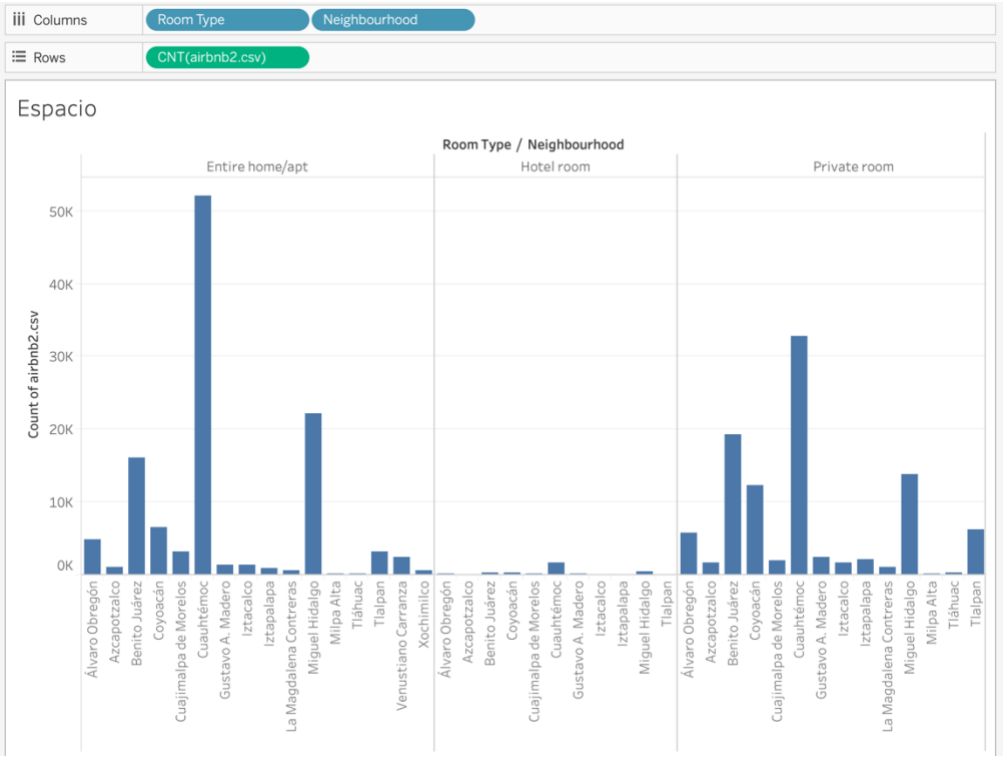


Distribución de ultima reseña por año.



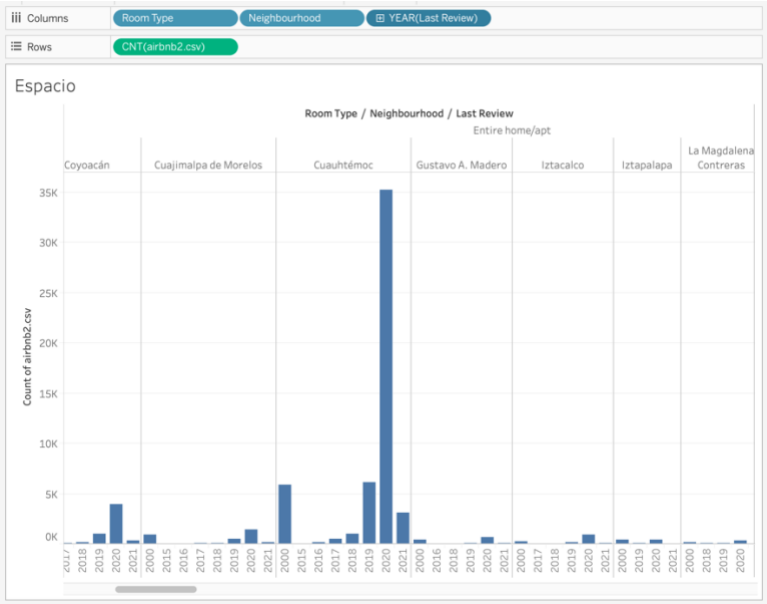
Distribución de airbnb por alcaldia de la CDMX.

5.3. ¿Cual es la distribución del fenómeno que mide el dataset en el espacio?, explorar la mayor cantidad de los niveles de granularidad. Ej. La distribución anual de incidentes viales en la delegación Coyoacán.



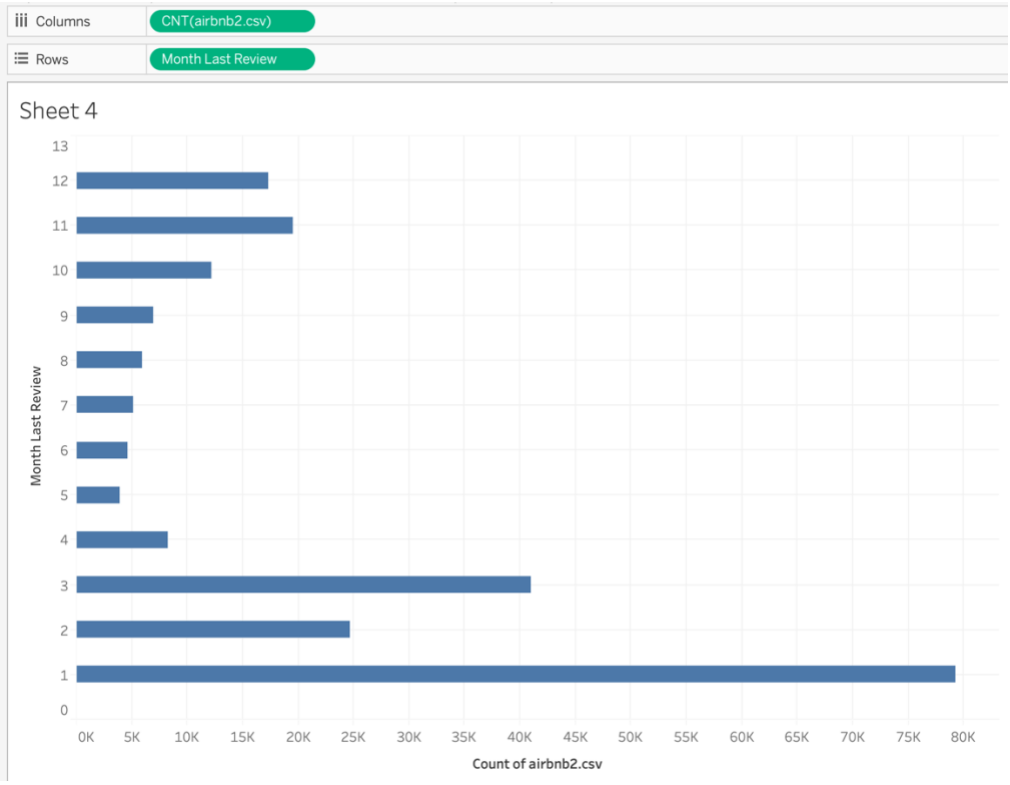
Distribución por alcaldía de tipo de airbnb ofertado.

5.4. ¿Cual es la distribución del fenómeno que mide el dataset en el tiempo y en el espacio?, explorar la mayor cantidad de los niveles de granularidad. Ej. La distribución anual de incidentes viales en la delegación Coyoacán en el 2012.

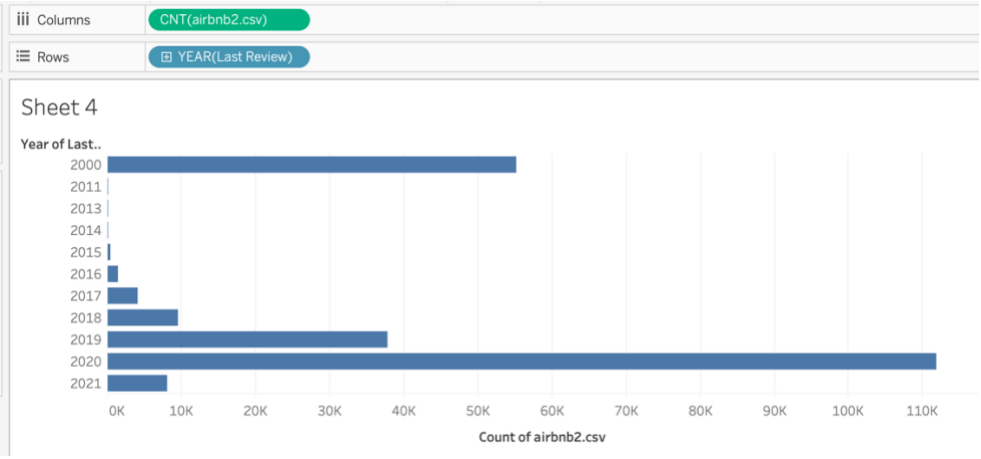


Distribución por alcaldía de tipo de airbnb ofertado y por año.

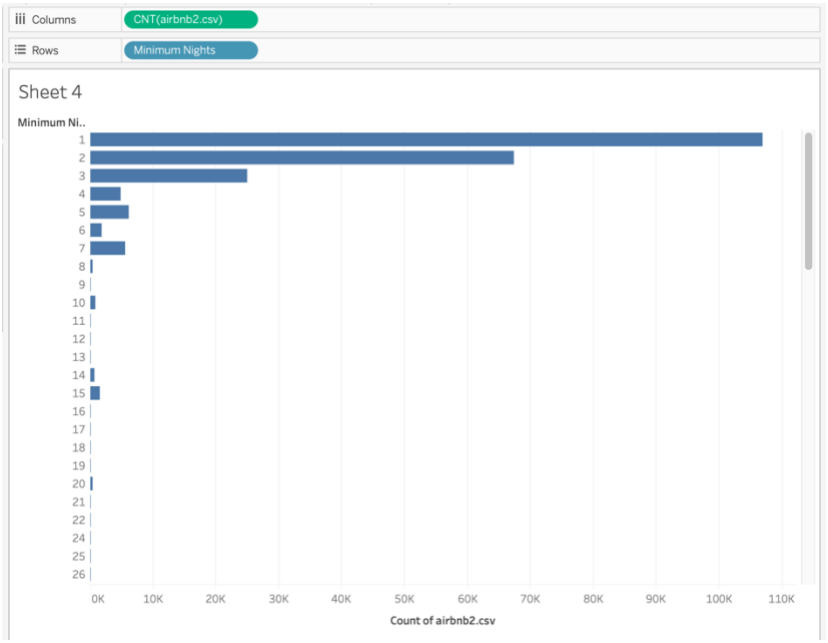
5.5. ¿Cual es la distribución de otras dimensiones temáticas (que consideren importante) del dataset? Ej. La distribución de los medios por los que son reportados los incidentes viales.



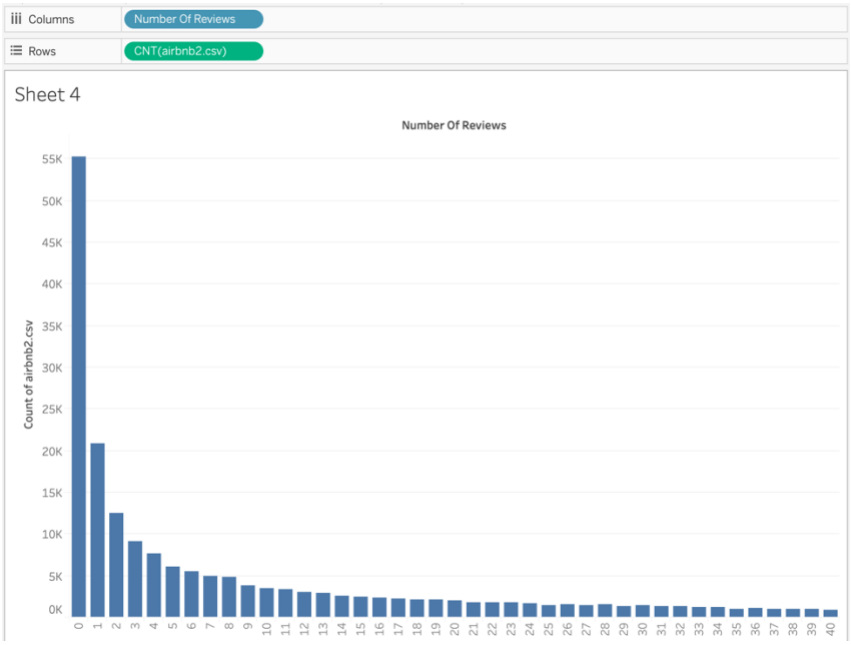
El numero de reviews que se hacen y la distribución por mes.



La distribución anual del ultimo review, donde 2000 es no tener ningún comentario.



La distribución por numero de días mínimos a ocupar el airbnb



Distribución de numero de reviews

5.6. ¿Encontró valores atípicos en el dataset o valores inconsistentes?

En la ultima fecha de review, hay airbnb que jamas han tenido comentarios, por lo que opte por dejarles fecha del 2000, o bien puedo hacer una segunda versión donde lo pongo null como estaba.

También es difícil utilizar un id único por cada registro, ya que como junte varios documentos, pero en diferentes momentos y en la misma zona geográfica, el id único de cada documento es el proveedor de airbnb, entonces hay documentos que contienen el mismo id, por lo que entonces podría ser bueno asignar ids como índice de tupla.

5.7. Verifique si las preguntas se pueden procesar con todos los registros originales del dataset o explique si el dataset fue recortado o filtrado por tiempo u otra variable.

Dado que elegi airbnb como fuente principal, encontré sus datos de manera separada por mes, por lo que yo opte por utilizar datos únicamente de 2020 y 2021, así el numero de tuplas a procesar seria aceptable para mis recursos.

Conclusiones

Ya tenia rato que no buscaba datos de manera masiva, definitivamente el acoplamiento, el usar mismas dimensiones, el verificar la fuente, los tipos de datos, la granularidad de datos, creo que fue bastante ejemplificador en mi caso que use una fuente distinta, espero encontrar información útil para mi, estoy ansioso por seguir analizando esta información, igual me gustaría aprender mas de Tableau por lo que revisare algunos tutoriales, hasta ahora no he logrado hacer una agrupación basada en rangos, por ejemplo precios de airbnb por rangos: $0 a $200, $200 - $500, $500 - $1,000, etc.