Informe de Calidad de los Datos

En este informe podremos encontrar las características de los datos analizados en los datasets provenientes de los reviews de la plataforma Google:

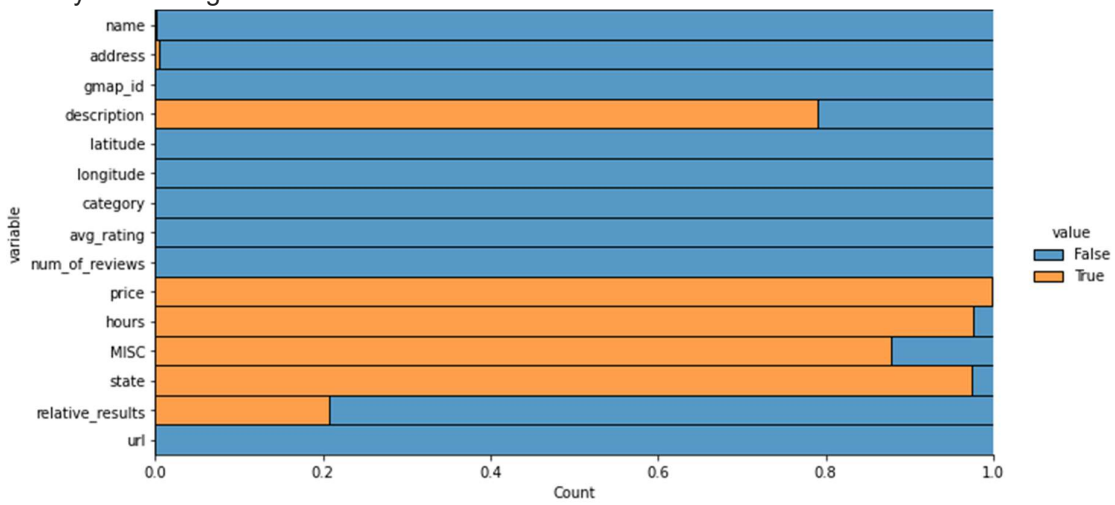
METADATA

El dataset posee 142.350 datos , contenidos en 9.490 filas y 15 columnas. De esas columnas, solo 4 contienen datos de tipo numérico y 11 de tipo object.

```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   name             9489 non-null   object
 1   address          9450 non-null   object
 2   gmap_id          9490 non-null   object
 3   description      1989 non-null   object
 4   latitude         9490 non-null   float64
 5   longitude        9490 non-null   float64
 6   category         9490 non-null   object
 7   avg_rating       9490 non-null   float64
 8   num_of_reviews   9490 non-null   int64
 9   price            17 non-null     object
 10  hours            233 non-null    object
 11  MISC             1160 non-null   object
 12  state            236 non-null    object
 13  relative_results 7528 non-null   object
 14  url              9490 non-null   object
dtypes: float64(3), int64(1), object(11)
```

Datos Nulos

Los datos faltantes en el dataset son 45.818 , representando el 32 % del total de datos , y se distribuyen de la siguiente manera:



Las columnas description, price, hours y state no se utilizarán para el análisis por tener demasiados valores nulos. En la columna MISC se esperan muchos datos nulos ya que no todos los hoteles poseen los mismos servicios.

Datos Duplicados

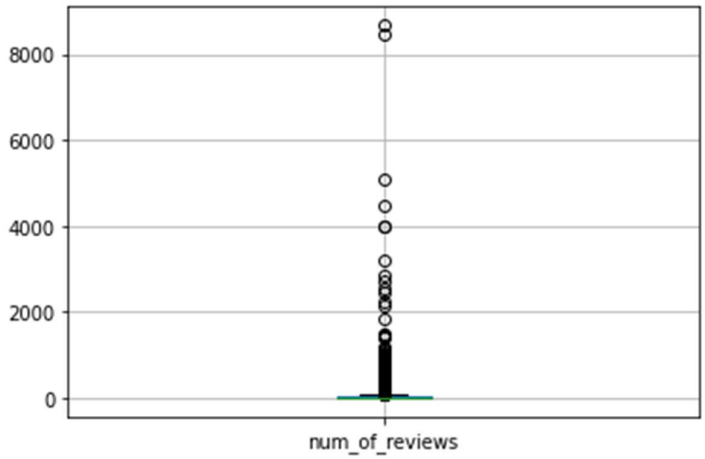
Como existe un ID unico (GmapID), realizamos el control por las columnas name, address , latitude y longitude , lo que nos arroja un total de 85 establecimientos duplicados los cuales no podremos eliminar por su posible correlación con la tabla review.

Inconsistencias en los datos

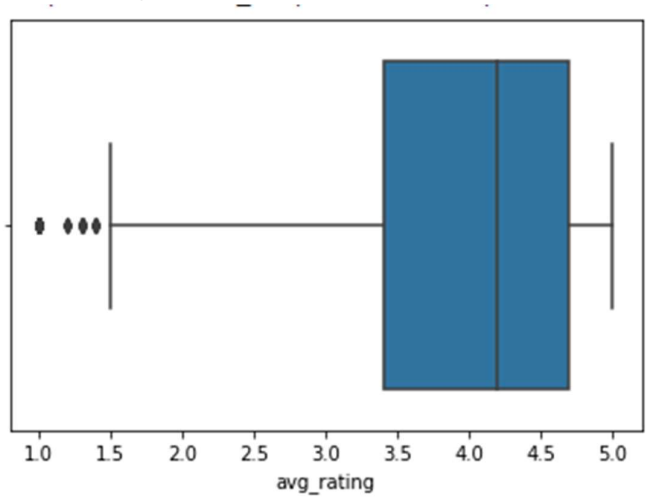
Solo un valor de Longitud se encuentra fuera de los límites del país por lo que no hay inconsistencia en latitud y longitud.

Valores Atípicos

Observamos muchos sitios con más de 550 reviews (límite de 3 desviaciones estándar) los cuales consideramos outliers ya que la media en num_of_reviews es de 38 y la desviación estándar de 183.



Si bien en la columna AVG rating no hay outliers , ya que la misma va de 1 a 5 si vemos un sesgo hacia la derecha ya que el 50% de los valores se concentra entre 3.4 y 4.7 (rango intercuartílico).



Las demás variables son categóricas , por lo que analizamos los valores únicos que asume cada columna:

VARIABLES CATEGÓRICAS	VALORES ÚNICOS
Category	1790
Hours	194
MISC	470
State (cerrado/abierto/horario)	73

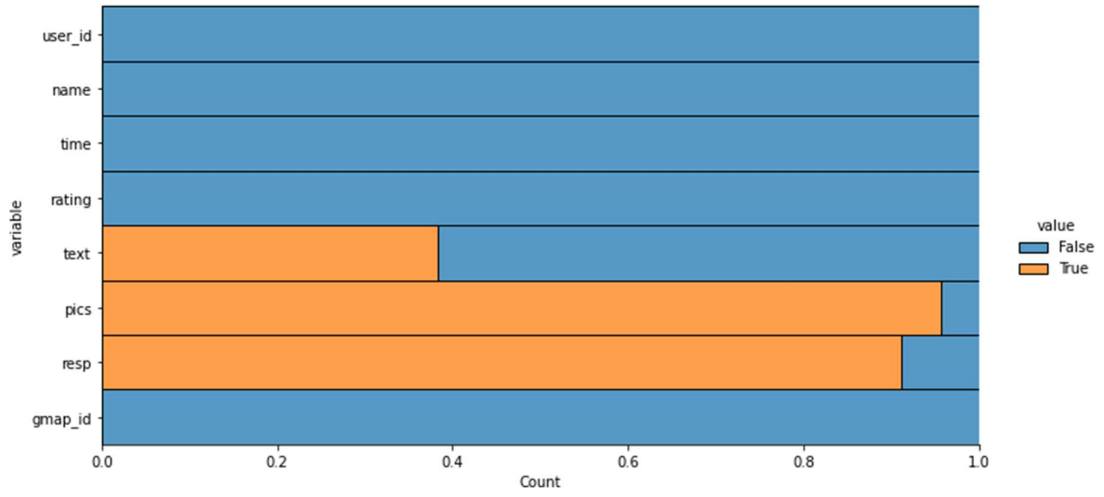
REVIEWS

El dataset posee 864.600 datos , contenidos en 108.075 filas y 8 columnas. De esas columnas, solo 3 contienen datos de tipo numérico y 5 de tipo object.

```
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   user_id  108075 non-null  float64
 1   name     108075 non-null  object
 2   time     108075 non-null  int64
 3   rating   108075 non-null  int64
 4   text     66742 non-null   object
 5   pics     4692 non-null    object
 6   resp     9556 non-null    object
 7   gmap_id  108075 non-null  object
dtypes: float64(1), int64(2), object(5)
```

Datos Nulos

Los datos faltantes en el dataset son 243.235 , representando el 28 % del total de datos , y se distribuyen de la siguiente manera:



La columna pics no se utilizará para el análisis por tener demasiados valores nulos. En las columnas text y response se realizarán transformaciones a fin de poder evaluar mejor la información que aportan.

Datos Duplicados

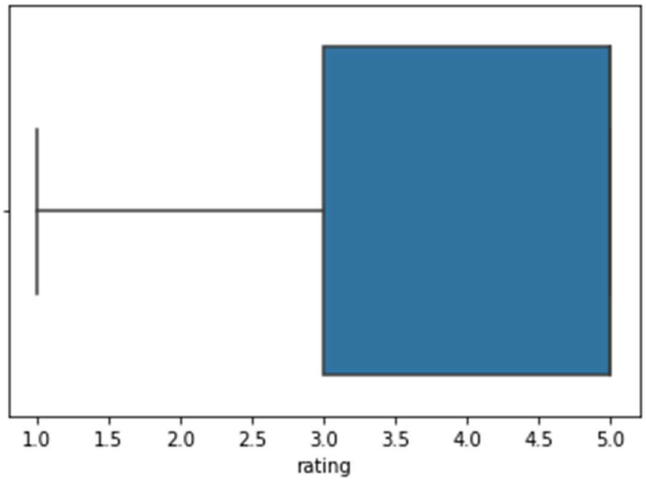
Realizamos el control por las columnas name, time, text, rating , gmap_id, pics y resp , lo que nos arroja un total de 4085 reviews duplicados los cuales se eliminan del dataset.

Tipos de datos

La columna time se encuentra en Unix Time Stamp por lo que debemos transformarla en fecha para los análisis.

Valores Atípicos

Si bien en la columna rating no hay outliers , ya que la misma va de 1 a 5 si vemos un sesgo hacia la derecha ya que el 50% de los valores se concentra entre 3 y 5 (rango intercuartílico).



Las demás variables son categóricas , por lo que analizamos los valores únicos que asume cada columna:

VARIABLES CATEGORICAS	VALORES ÚNICOS
Name	96.277
Text	57.949
Response	9160