INTRODUCCIÓN

Análisis exploratorio del set de datos “train.csv” que fue provisto por twitter. En una primera instancia vamos a recolectar información de los datos, y es lo que se muestra a continuación.

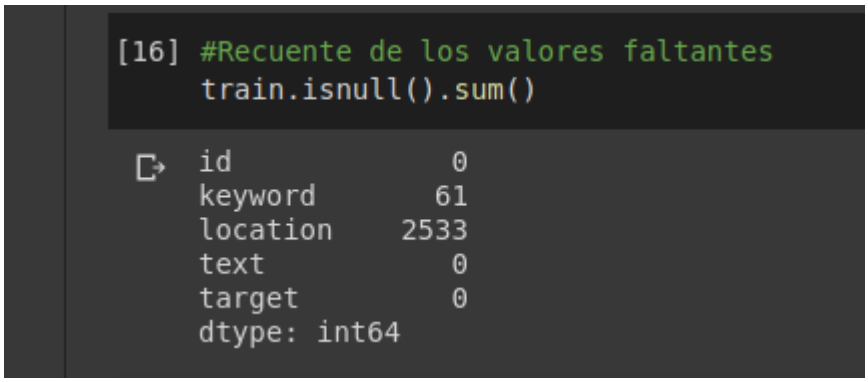
- Id : Es un identificador que proporciona un número único de reconocimiento para cada tuit.
- Text : Este es un campo que contiene el texto escrito en el tuit.
- location – Nos informa sobre la ubicación de donde fue enviado un tuit, y esta misma podría encontrarse o no.
- keyword - Una keyword o palabra clave es el término o conjunto de términos que utilizan los usuarios cuando buscan en los buscadores.
- target – Este set de datos, nos indica si el tuit se trata de algún desastre real o no, esta diferenciación de produce por medio de los valores 1 o 0.

El objetivo de este primer informe es analizar los datos descritos anteriormente para observarlos de manera general, y sacar características de su comportamiento que nos puedan servir más adelante.

1. ¿ Qué dimensiones poseé el data frame?

Los datos están organizados por filas y columnas, tenemos 7613 files y 5 columnas.

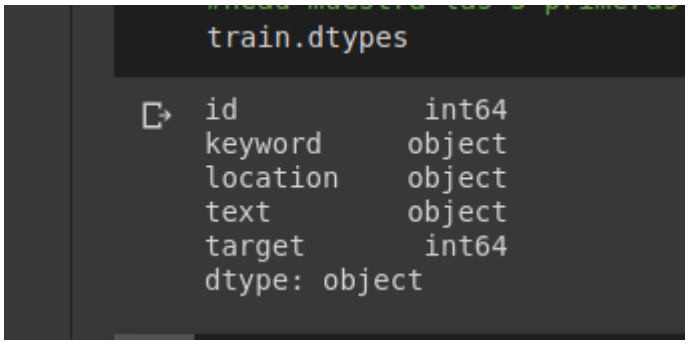
2.¿Todos los datos están completos?



Tenemos 61 palabras claves faltantes, 61 tuits no van a tener un acceso directo y rápido.

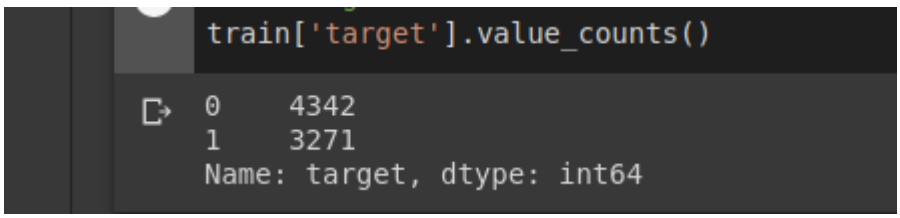
También nos encontramos con 2533 tuits sin ubicación, lo cual dificultará su averiguar su origen.

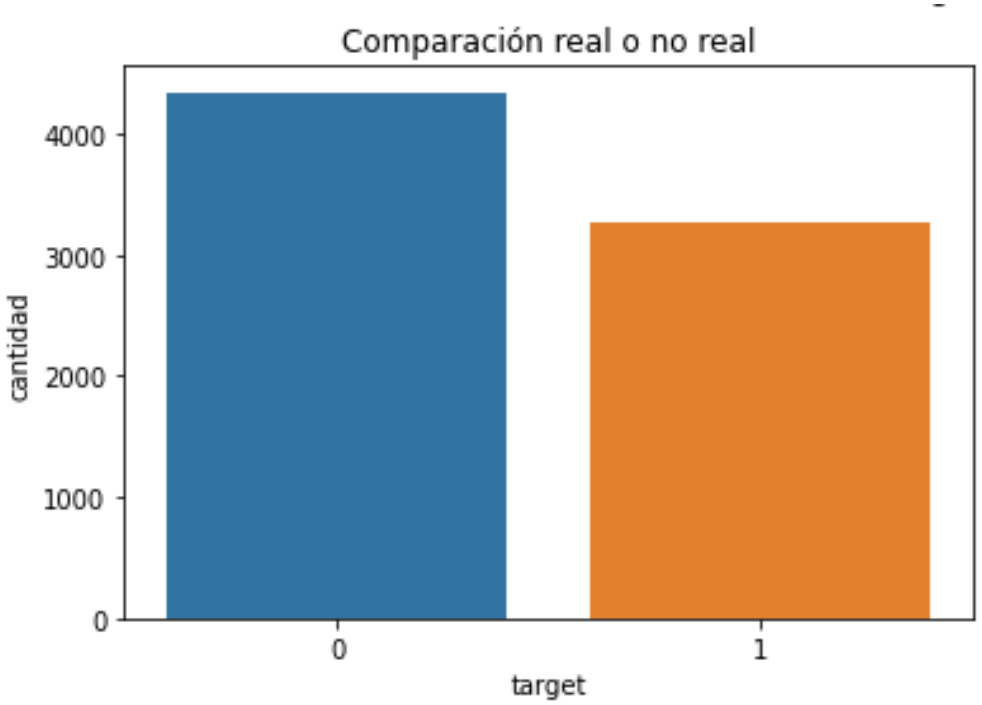
3.¿Qué tipos de datos encontramos en cada columna?



Trabajaremos con tipos de datos: Objects e Integer.

4. Averigüemos la cantidad de “desastres reales y no reales”





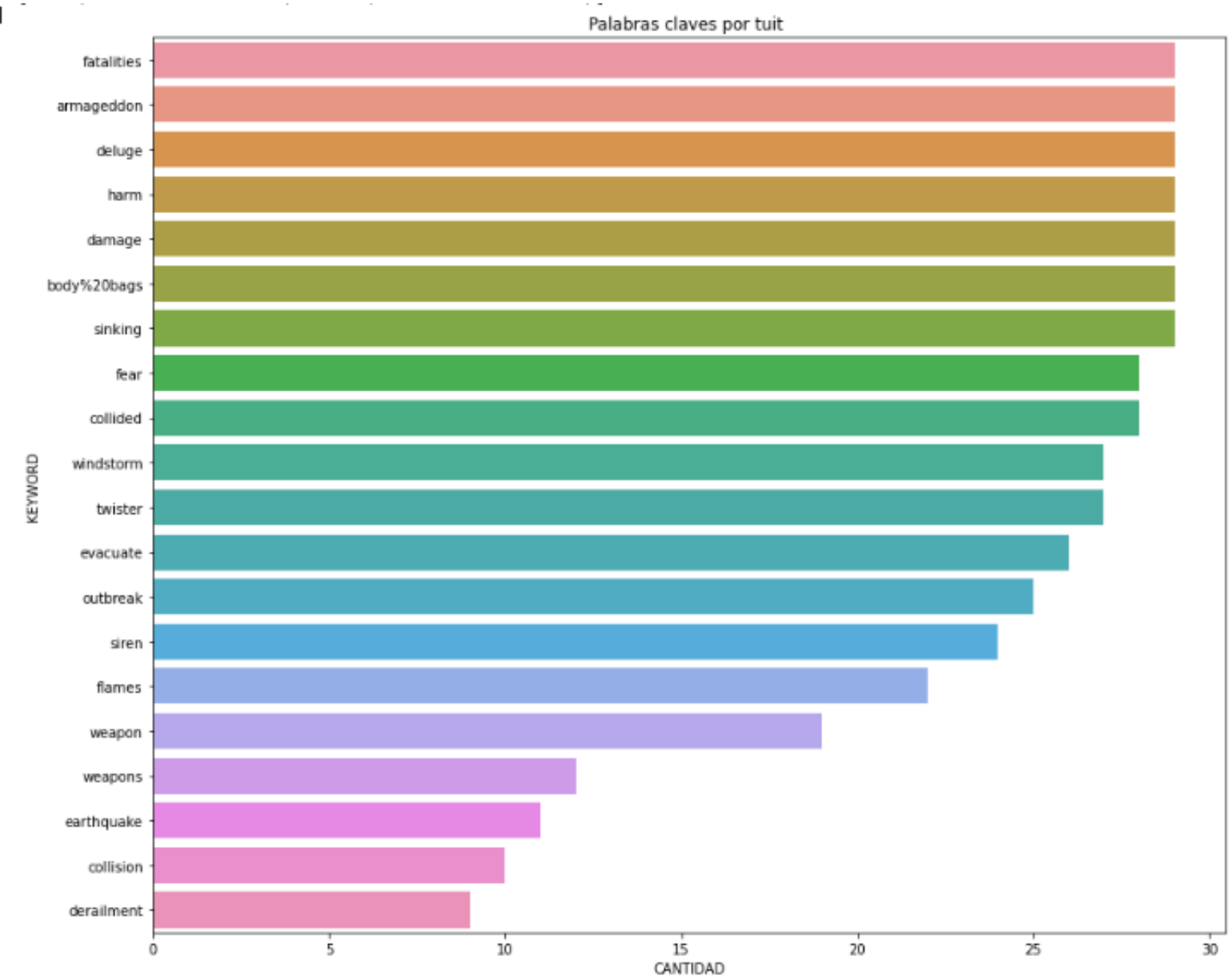
5. Enlistamos la cantidad de keyword

target

```
[54] s=train['keyword'].value_counts()
     s
```

```
fatalities              45
armageddon              42
deluge                  42
harm                    41
damage                  41
                        ..
forest%20fire           19
epicentre               12
threat                  11
inundation              10
radiation%20emergency    9
Name: keyword, Length: 221, dtype: int64
```

Mostramos las palabras que mas que mas se repiten

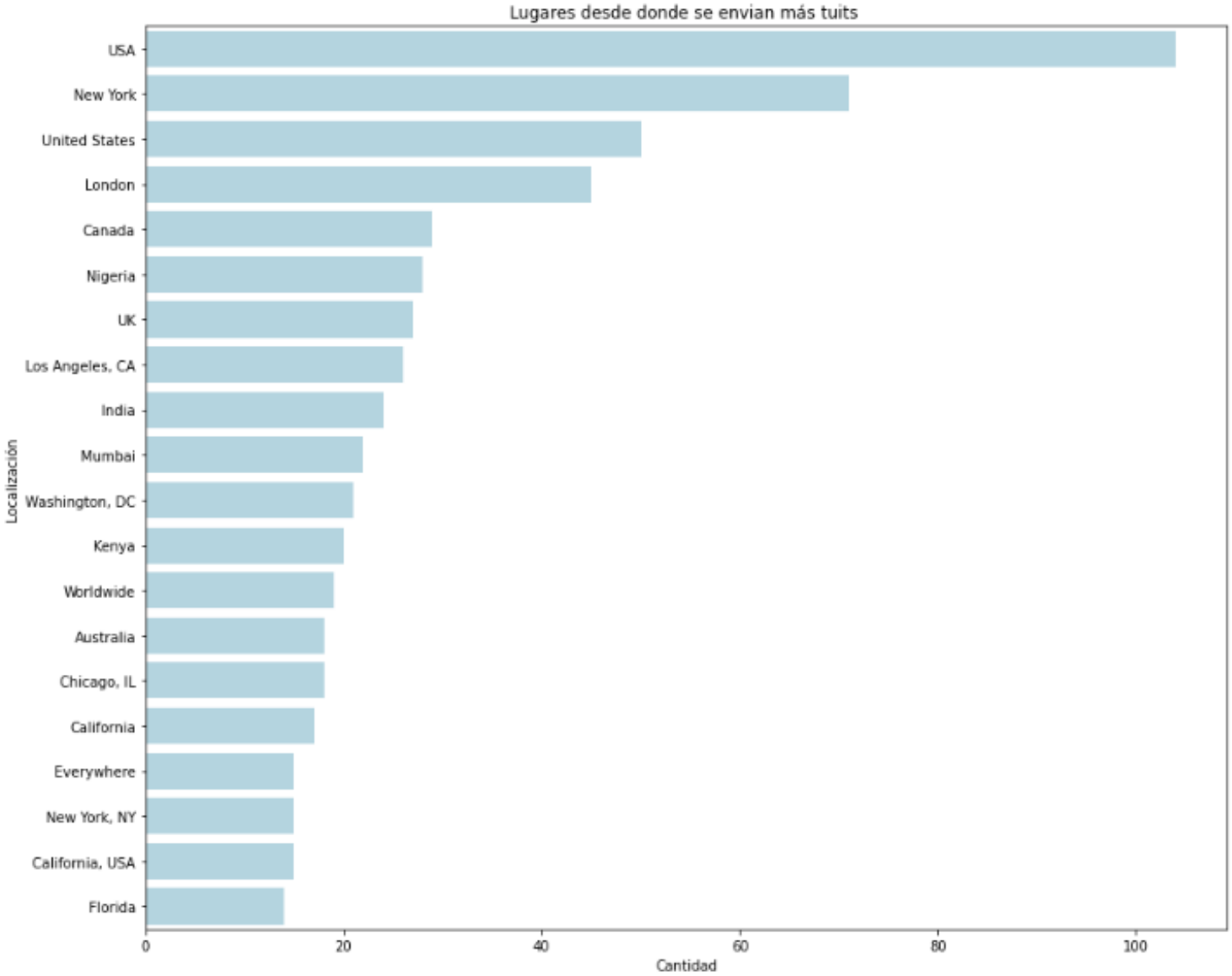


6. Cantidad de palabras que cuentan con una localización

```
[ ] print(train['location'].nunique())

    3341
```

El siguiente gráfico muestra las 20 ubicaciones más comunes de donde vienen la mayor cantidad tuits.



7. ahora veamos las 20 ubicaciones menos comunes