FACULTAD DE INGENIERIA
Universidad de Buenos Aires

75.06/95.58 Organización de Datos
Primer Cuatrimestre de 2020
Trabajo Práctico 1

Grupo:35
Data Hunters

Apellido/s	Nombre/s	Padrón	E-mail
Giampietri	Mauro Gabriel	101186	mgiampietri@fi.uba.ar
Brocca	Pablo Martín	104256	pbrocca@fi.uba.ar
Inneo Veiga	Sebastian Bento	100998	sinneo@fi.uba.ar

Índice

Introducción

En este informe analizaremos los datos provistos por el archivo train.csv, este archivo contiene datos sobre ciertos tweets que pueden o no ser acerca de un accidente y consta de los siguientes datos (nota: se muestran solo los primeros 10):

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1
5	8	NaN	NaN	#RockyFire Update => California Hwy. 20 closed...	1
6	10	NaN	NaN	#flood #disaster Heavy rain causes flash flood...	1
7	13	NaN	NaN	I'm on top of the hill and I can see a fire in...	1
8	14	NaN	NaN	There's an emergency evacuation happening now ...	1
9	15	NaN	NaN	I'm afraid that the tornado is coming to our a...	1

El significado de las columnas es el siguiente:
id: identificación del usuario que escribió el tweet.
keyword: palabra clave por la cual podría considerarse un tweet acerca de algún desastre.
location: lugar desde el cual se escribió el tweet.
text: el tweet.
target: la veracidad del tweet (1 para verdadero y 0 para falso).

Link al repositorio de GitHub.

Objetivo

El objetivo de este informe es tratar de llegar a alguna relación entre los distintos datos de las columnas con el fin de predecir qué tweets son verdaderos y cuáles no a partir de la siguiente pregunta: ¿cómo se relacionan las columnas respecto a su veracidad?

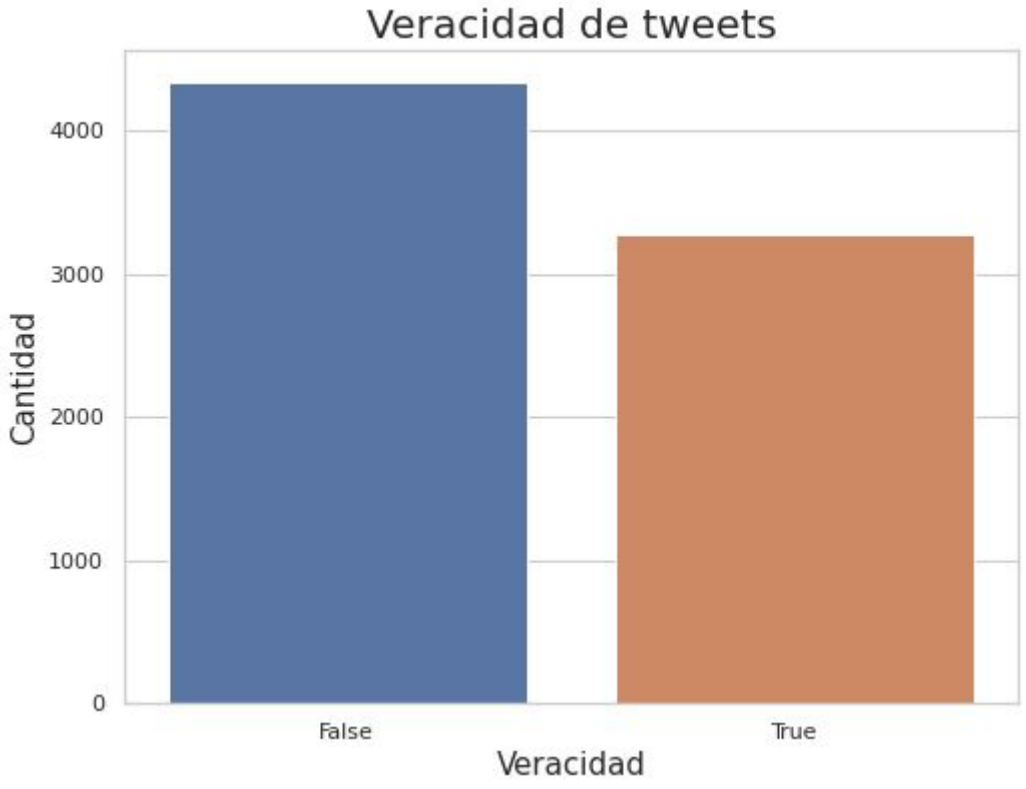
Análisis general

Al hacer un análisis general del archivo obtenemos lo siguiente:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7613 entries, 0 to 7612
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   id         7613 non-null   int64
 1   keyword    7552 non-null   object
 2   location   5080 non-null   object
 3   text       7613 non-null   object
 4   target     7613 non-null   int64
dtypes: int64(2), object(3)
memory usage: 297.5+ KB
```

Esto nos permite afirmar que hay 7613 tweets de los cuales todos tienen algún valor no nulo en sus columnas id, text y target sin embargo solo 7552 de ellos tienen keyword y 5080 tienen location.

A continuación analizamos los valores de target:



Gracias a esto sabemos que hay 4342 tweets falsos y 3271 tweets verdaderos

Con respecto al contenido de los tweets obtuvimos lo siguiente:

```
count                                                7613
unique                                               7503
top     11-Year-Old Boy Charged With Manslaughter of T...
freq                                                   10
Name: text, dtype: object
```

Lo que nos permite decir que hay 7503 tweets entre los cuales algunos de ellos se repiten. El más repetido es un tweet acerca de un niño de 11 años acusado de asesinato en masa el cual se repite 10 veces y luego de un análisis más profundo sobre ese tweet podemos confirmar que todos estos tweets se refieren al mismo hecho y no son sobre hechos distintos ya que las ubicaciones de estos tweets son en zonas de la india y todos ellos son verdaderos.

Al ver las keywords obtenemos esto:

```
count       7552
unique       221
top     fatalities
freq          45
Name: keyword, dtype: object
```

Por lo cual afirmamos que solo hay 221 tipos de accidentes siendo el más frecuente fatalities con 45 tweets.

Al analizar los tweets con esta keyword vemos que hay 26 reales y 19 falsos todos ellos dispersos alrededor del mundo.

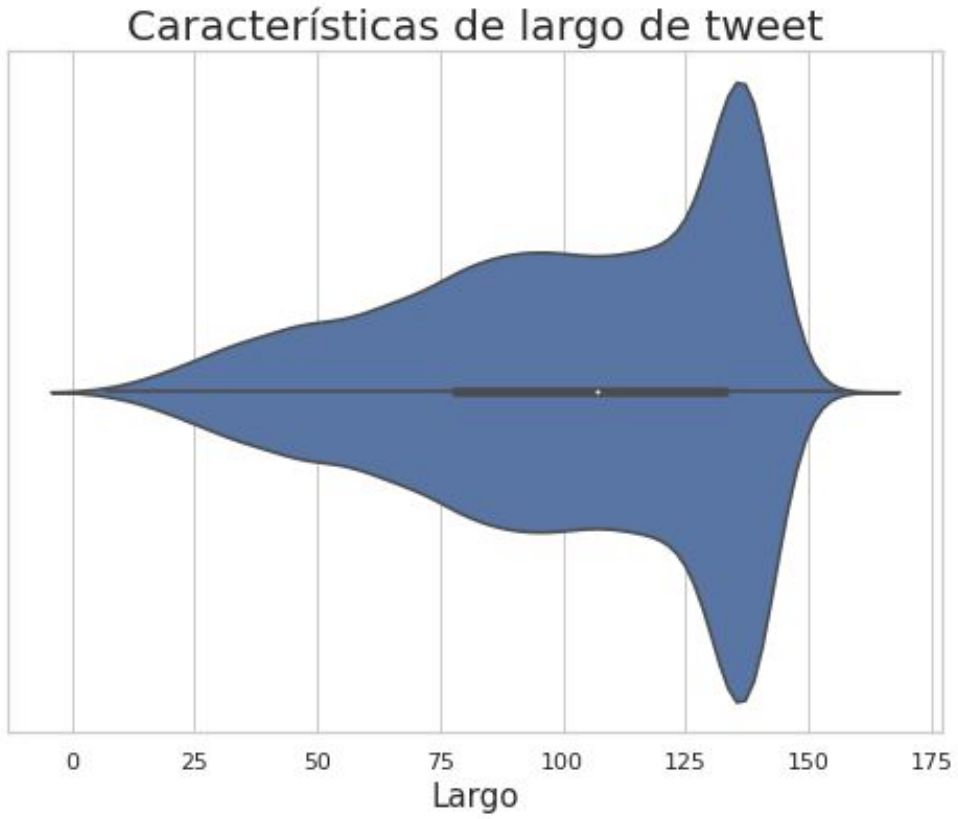
Correlaciones

Ahora nos concentramos en ver las relaciones entre los elementos del archivo

Largo de texto (tweet)

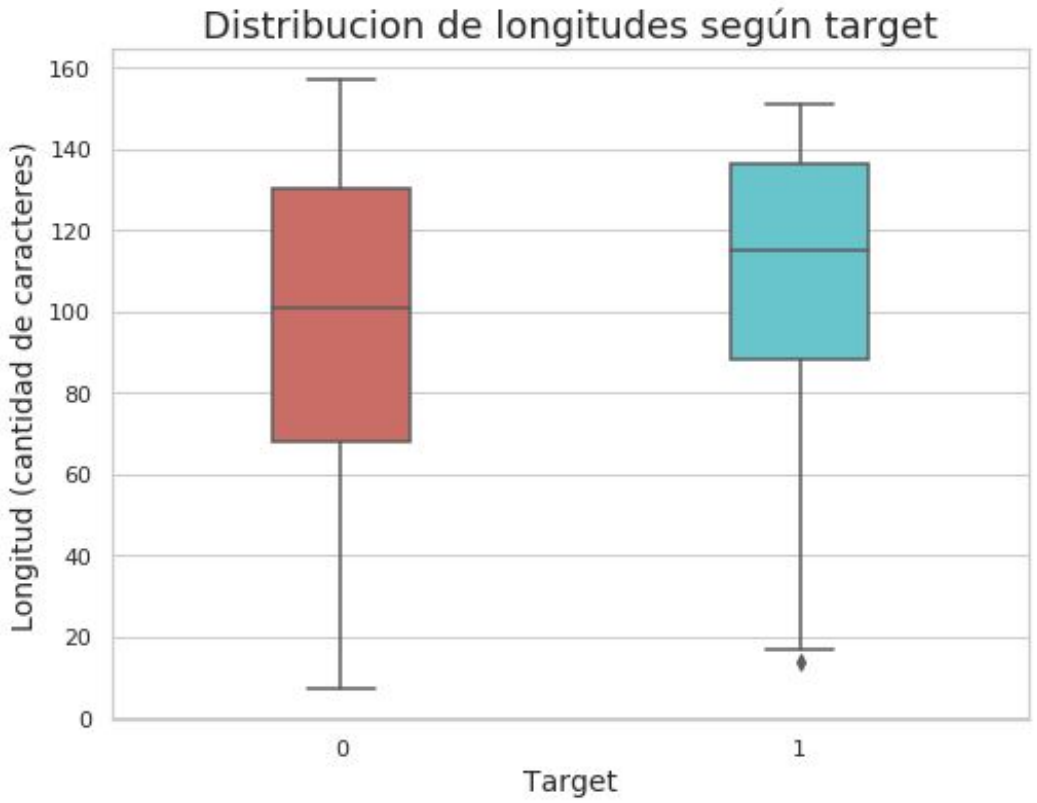
Primero vemos si hay alguna relación entre el tamaño de los tweets y su veracidad, para esto incorporamos la columna len_text la cual indica cuántos caracteres hay por texto.

Al analizar esta nueva columna obtenemos los siguientes datos:



Esto nos dice que el promedio de longitud es de 101,037436 caracteres, el mínimo es de 7 caracteres y el máximo es de 157.

Para cada valor del target, analizamos estos datos en un gráfico de cajas:



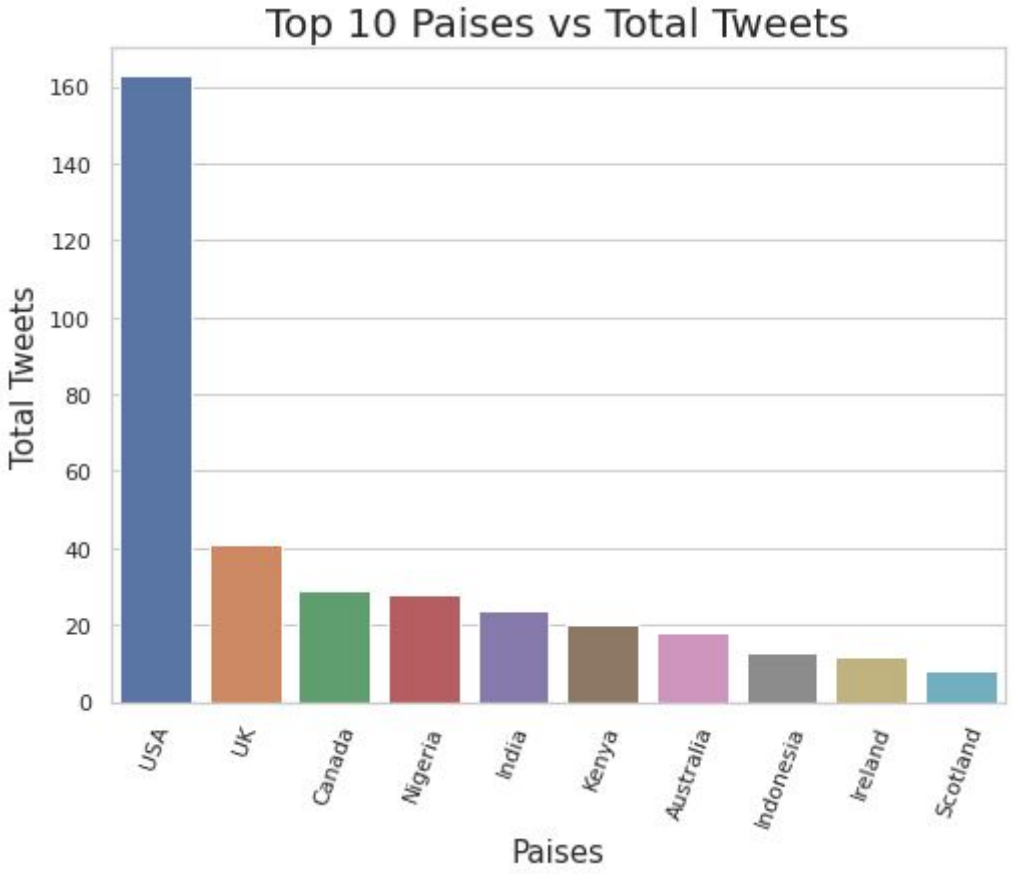
Acá podemos observar cómo el largo se relaciona con la veracidad del tweet. Los verdaderos tienen un promedio de 108 caracteres con máximo de 151 caracteres y un mínimo de 14 caracteres. Por el lado de los falsos un promedio de 95 caracteres con máximo de 157 caracteres y un mínimo de 7 caracteres. Teniendo esto en cuenta pareciera ser que la gran mayoría de tweets verídicos tienden a estar entre los 88 y 136 caracteres mientras que los falsos entre 68 y 130 caracteres lo cual hace pensar que el largo no está directamente relacionado a si es verdadero o no el tweet.

Top 10 Location vs Total tweets

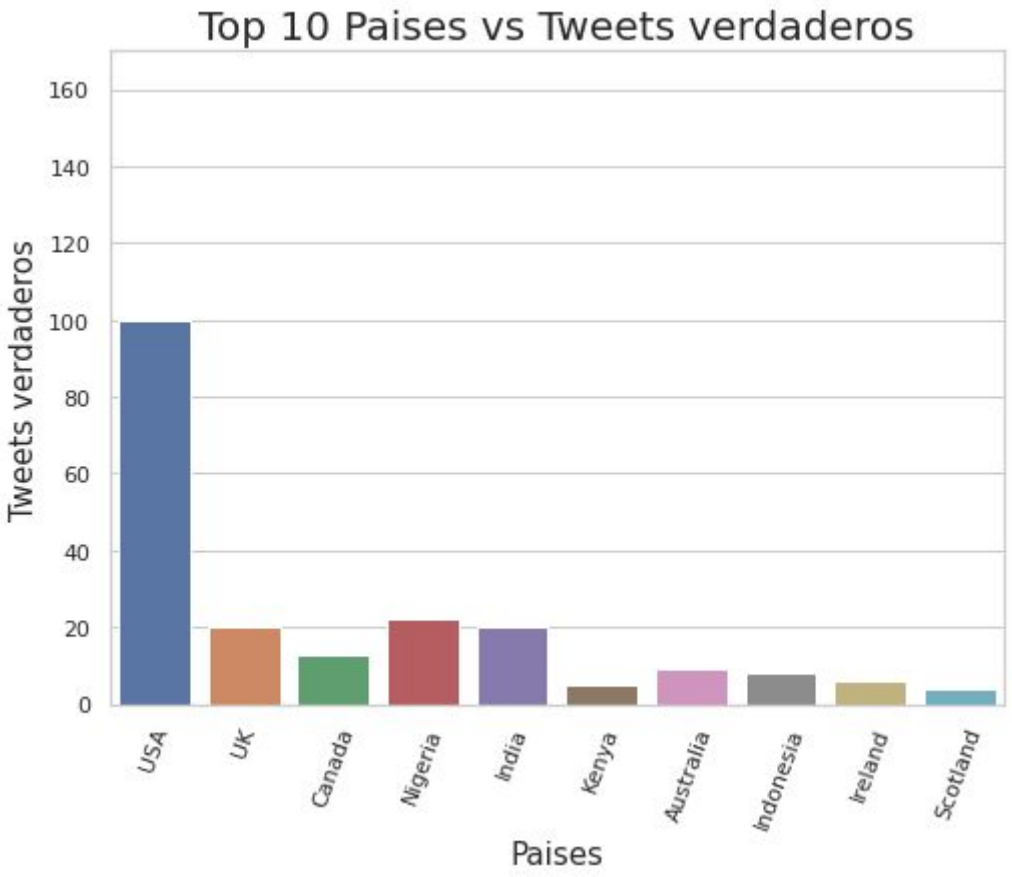
A continuación analizaremos el top 10 países y top 10 ciudades tomando en cuenta aquellos que tengan al menos 8 Tweets en total. Cabe aclarar que en el caso de los países se toma sólo aquellos tweets en los que se indica sólo el país en cuestión y no se toman en cuenta las ciudades para obtener el total ni la cantidad de tweets verdaderos. Por lo tanto la distinción entre países y ciudades se hace para una comparación entre los mismos y no contra todos.

Países

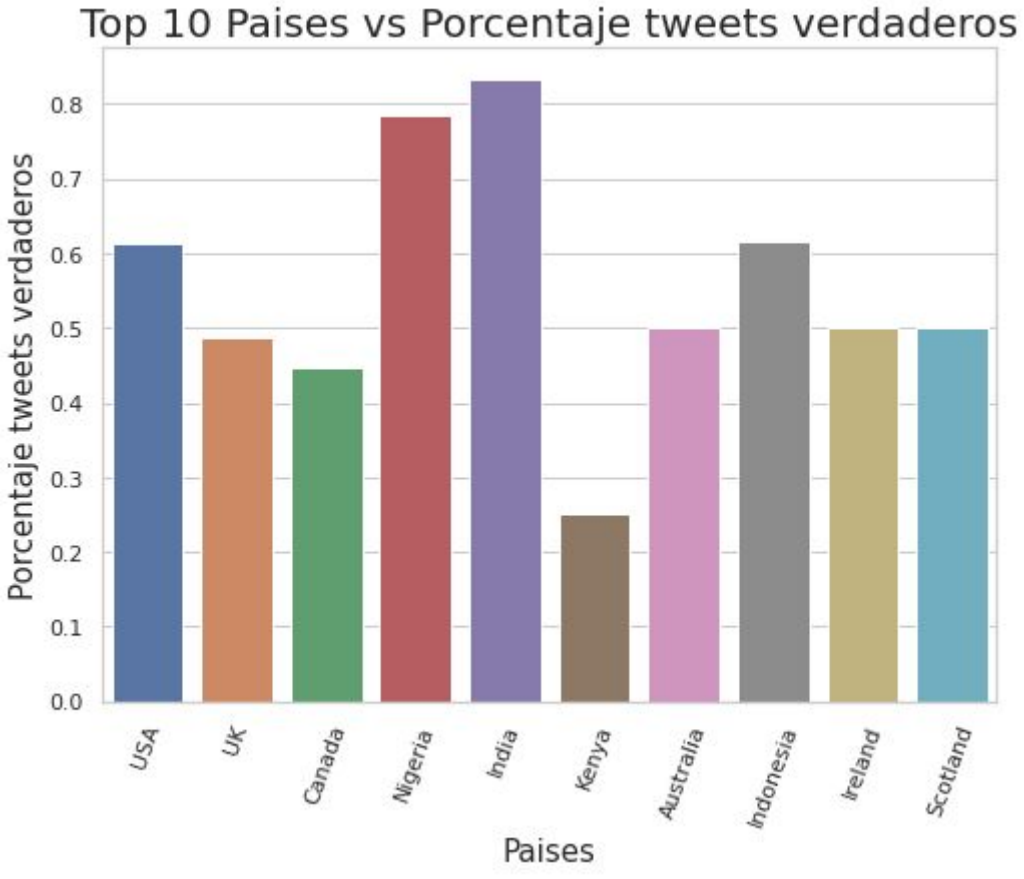
Empezando por los países (ordenados por los que tienen más cantidad de tweets) obtenemos un gráfico de barras del estilo:



Hay una clara mayoría en USA comparado con otros países que están un poco más parejos y no tienen tanta diferencia pero si vemos los que son verdaderos:



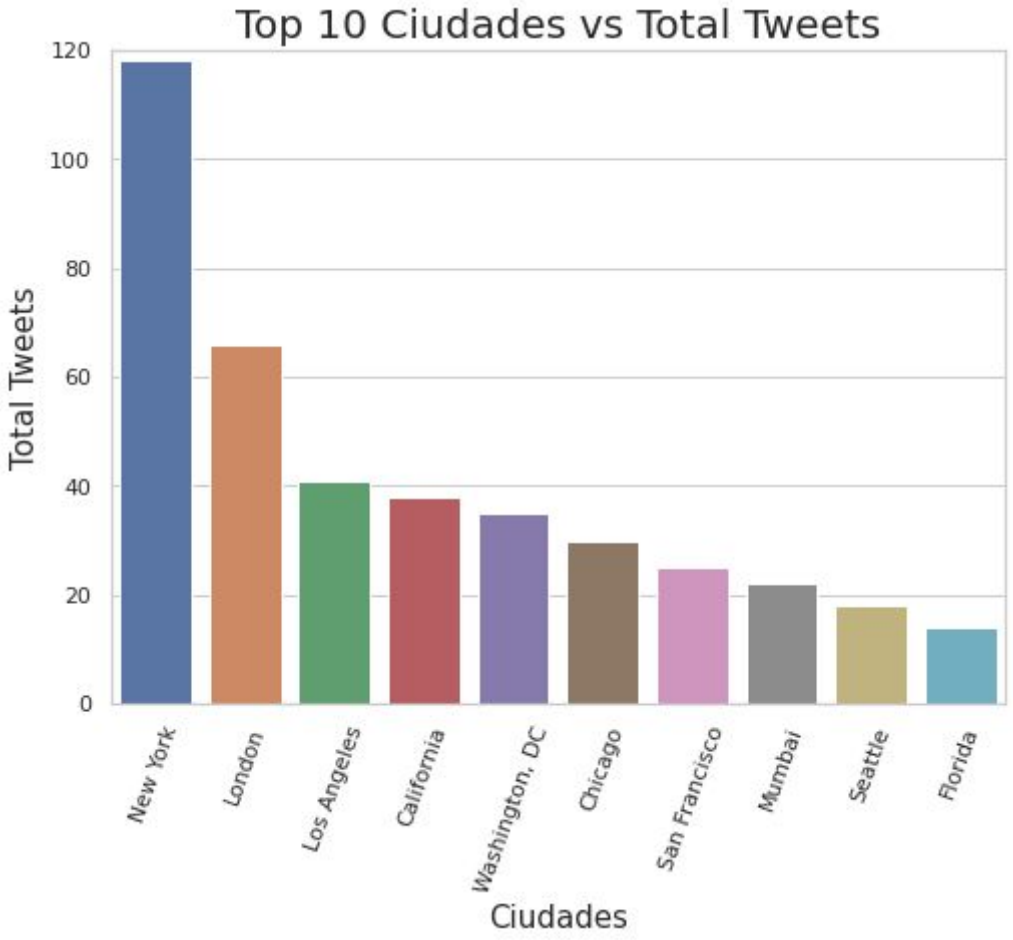
También se puede observar como sigue liderando USA y la diferencia parece más marcada entre los otros países. Se puede ver cómo los países que le siguen a USA son Nigeria e India. Sin embargo, en proporción, se podría decir que estos últimos tienen más tweets verdaderos como se aprecia en el siguiente gráfico:



Esto podría interpretarse como que, a pesar de que estos países tienen menos tweets, tienden a tener más veracidad. Sin embargo no pasaría lo mismo con el país que les sigue en cantidad de tweets totales que es Kenya el cual tiene la menor proporción de tweets verdaderos.

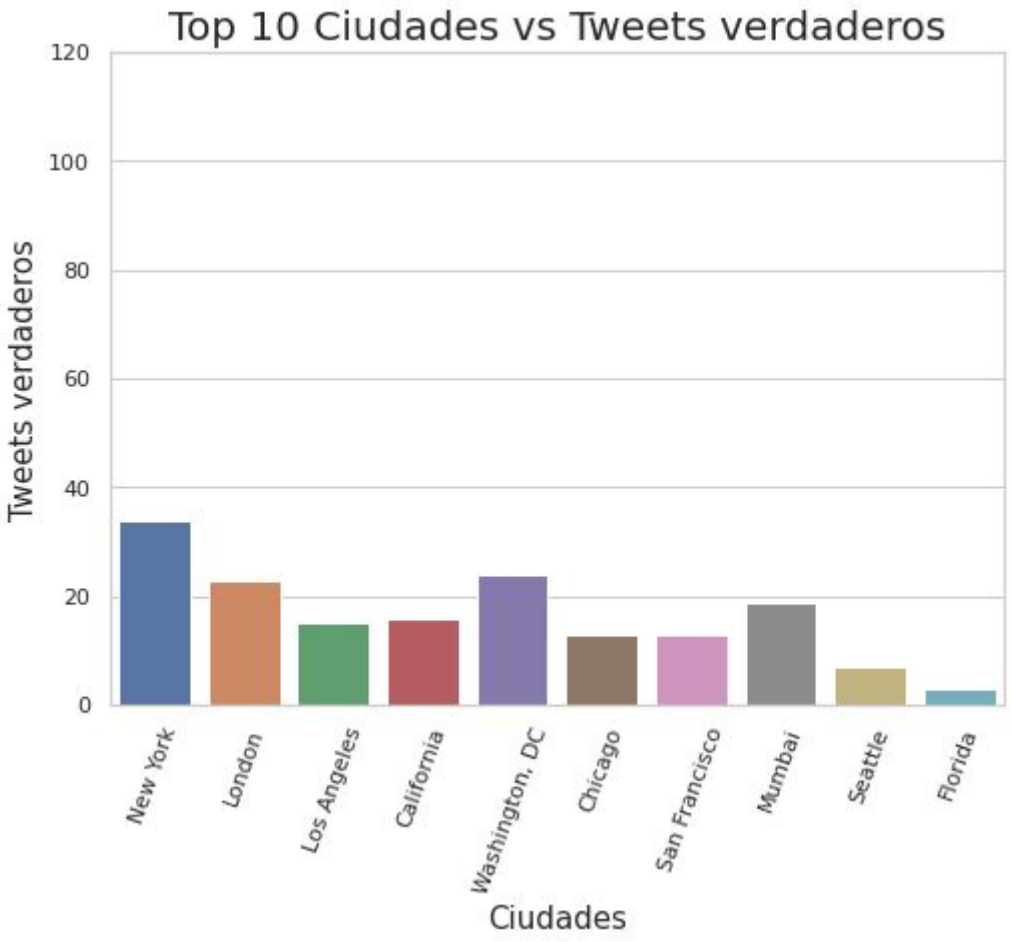
Ciudades

Prosigamos a analizar lo que sucede con las ciudades:



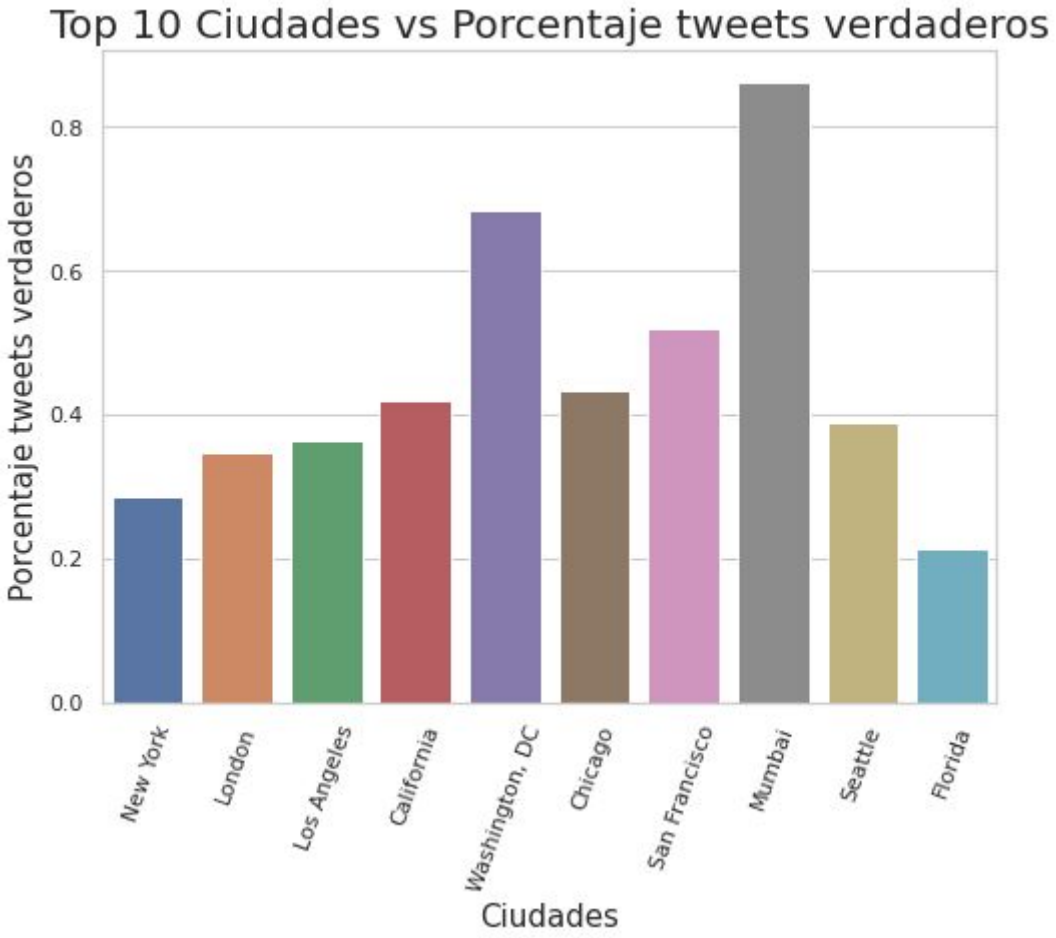
Como podemos ver la mayoría son ciudades de USA que era casualmente el país con más tweets. La segunda ciudad, London, también comparte esta relación con el gráfico anterior. Nuevamente vemos como el primer puesto lleva una diferencia marcada con las otras ciudades que tienen una menor diferencia en el total entre ellas.

Analicemos cuántos de los tweets totales son verdaderos:



En este caso pasa todo lo contrario al gráfico de los países, la diferencia de tweets verdaderos entre las ciudades es mucho menor.

Analicémoslo en proporción:



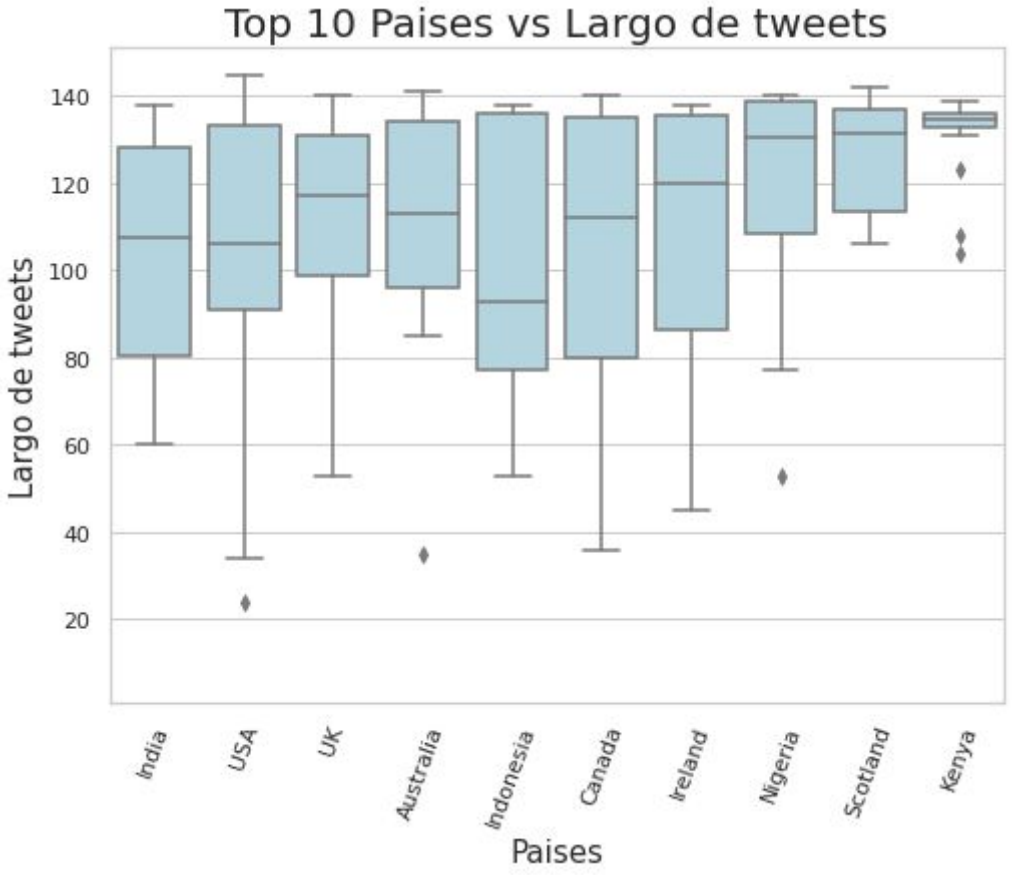
Acá podemos observar como Mumbai es la ciudad con mayor cantidad de tweet verdaderos seguido por Washington, DC.

De estos gráficos podemos concluir que tweets provenientes de India, Nigeria y Mumbai son en su mayoría verdaderos, al contrario de lo que pasa con los provenientes de Kenya y Florida. Mayor cantidad de tweets no significa que tengan mayor veracidad como es el caso de New York. En cuanto a los demás países y ciudades habrá que analizarlos con otra variable para ver si tienen alguna tendencia o no.

Top 10 Location vs Largo tweets

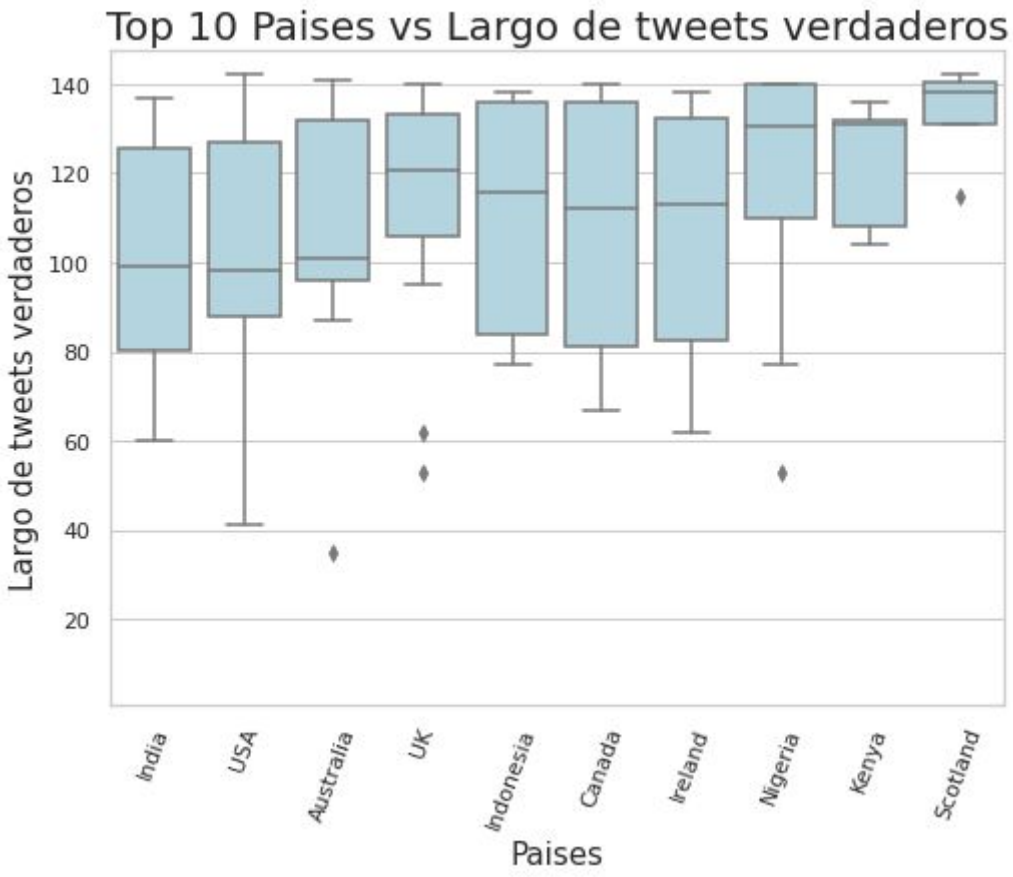
Países

Veamos cómo se relacionan los distintos largos de los tweets (verdaderos o no) con sus respectivos países:



La mayoría de los tweets presentan un largo entre 80 y 140 caracteres aproximadamente, hay muy pocos casos fuera de ese rango. El país que más resalta es Kenya, cuya mayoría de tweets tienen casi el mismo largo salvo por 3 excepciones.

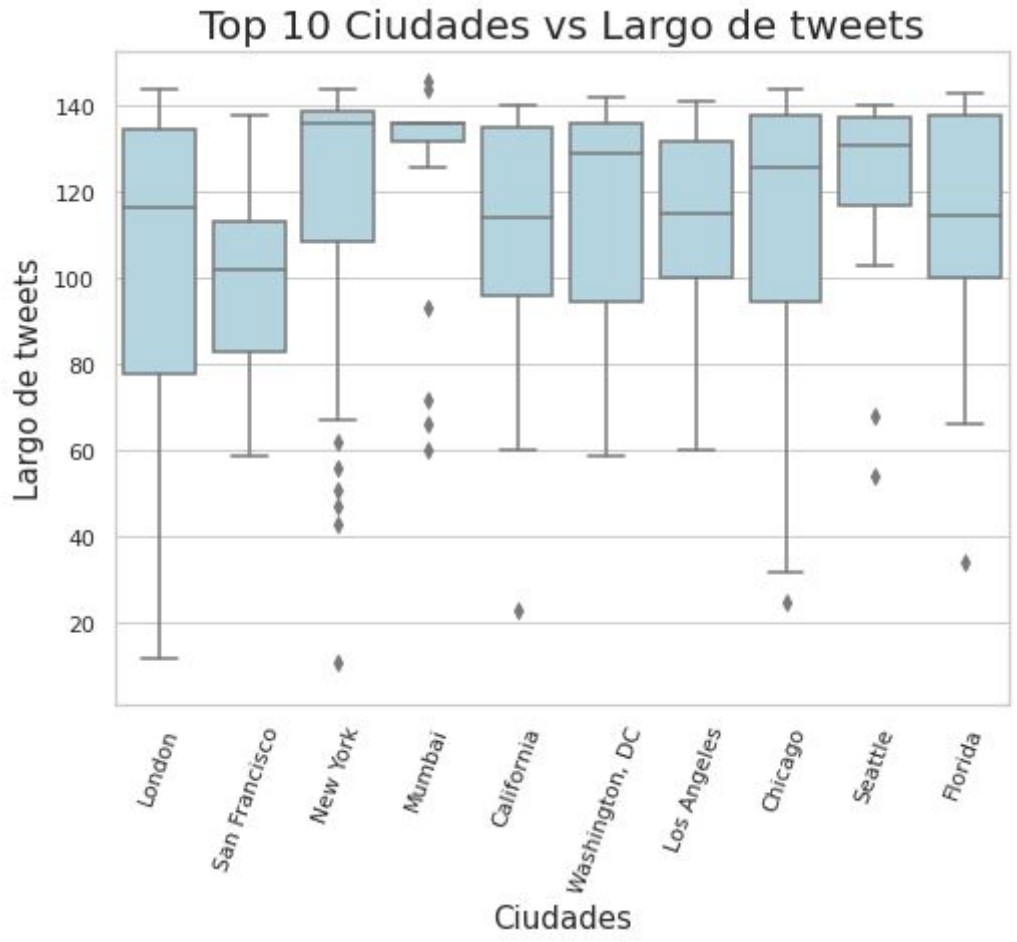
Prosigamos analizando ahora únicamente el largo de aquellos tweets que son verdaderos:



En general la media tendió a subir, pero los máximos y mínimos bajaron un poco en el caso de tweets verdaderos, por lo que será importante tener en cuenta aquellos tweets que estén por encima de los 80 rondando los 100 a 120 caracteres en el caso que provengan de países. Un caso destacable es el de Kenya que antes tenía 3 tweets aislados que por lo visto eran parte de los verídicos pero como habíamos visto en los gráficos anteriores su porcentaje de tweets verdaderos era muy bajo a pesar de tener tweets dentro del rango de largo de los verdaderos.

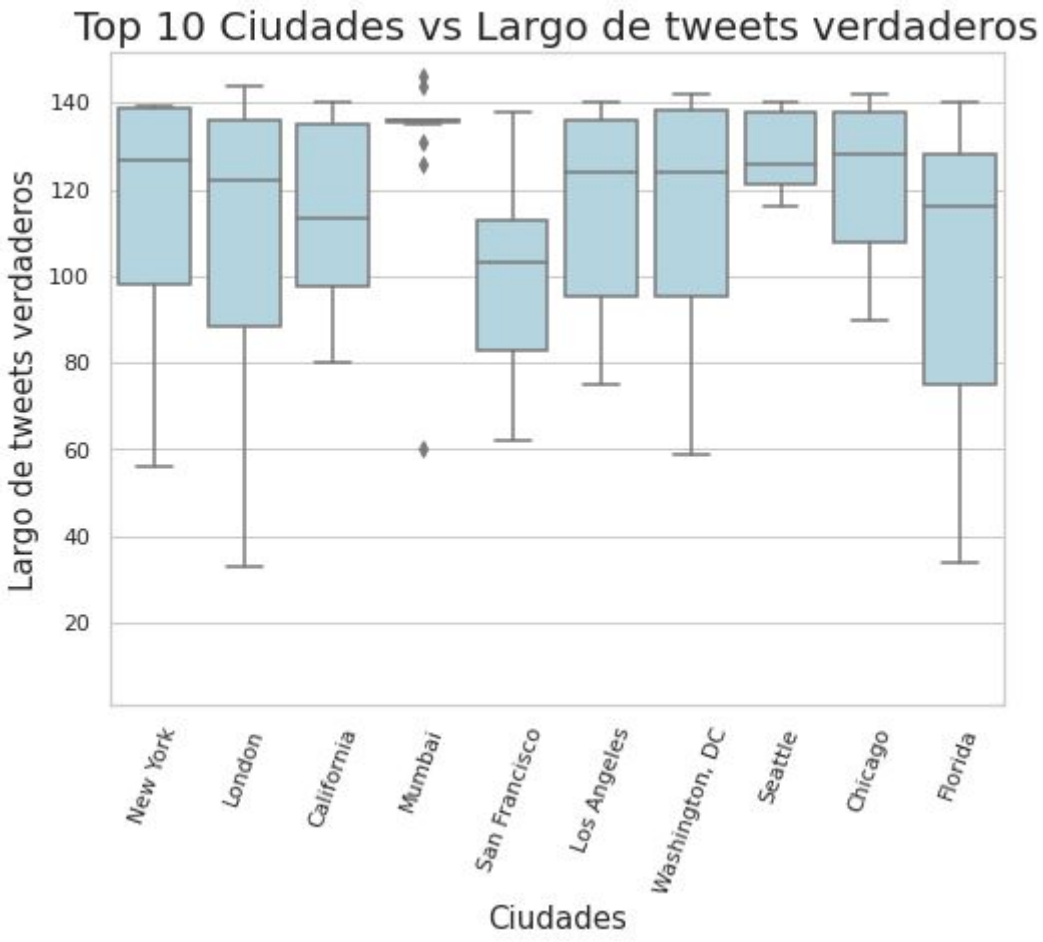
Ciudades

Veamos qué sucede en el caso de las ciudades:



Acá vemos como la cantidad de casos aislados es mayor principalmente en New York y Mumbai. El largo promedio de los tweets, comparado con los países, es un poco más alto, tienden más a estar entre 100 y 140 aproximadamente y tienen mínimos más bajos.

Veamos cuál es el largo de los que son verdaderos entonces:

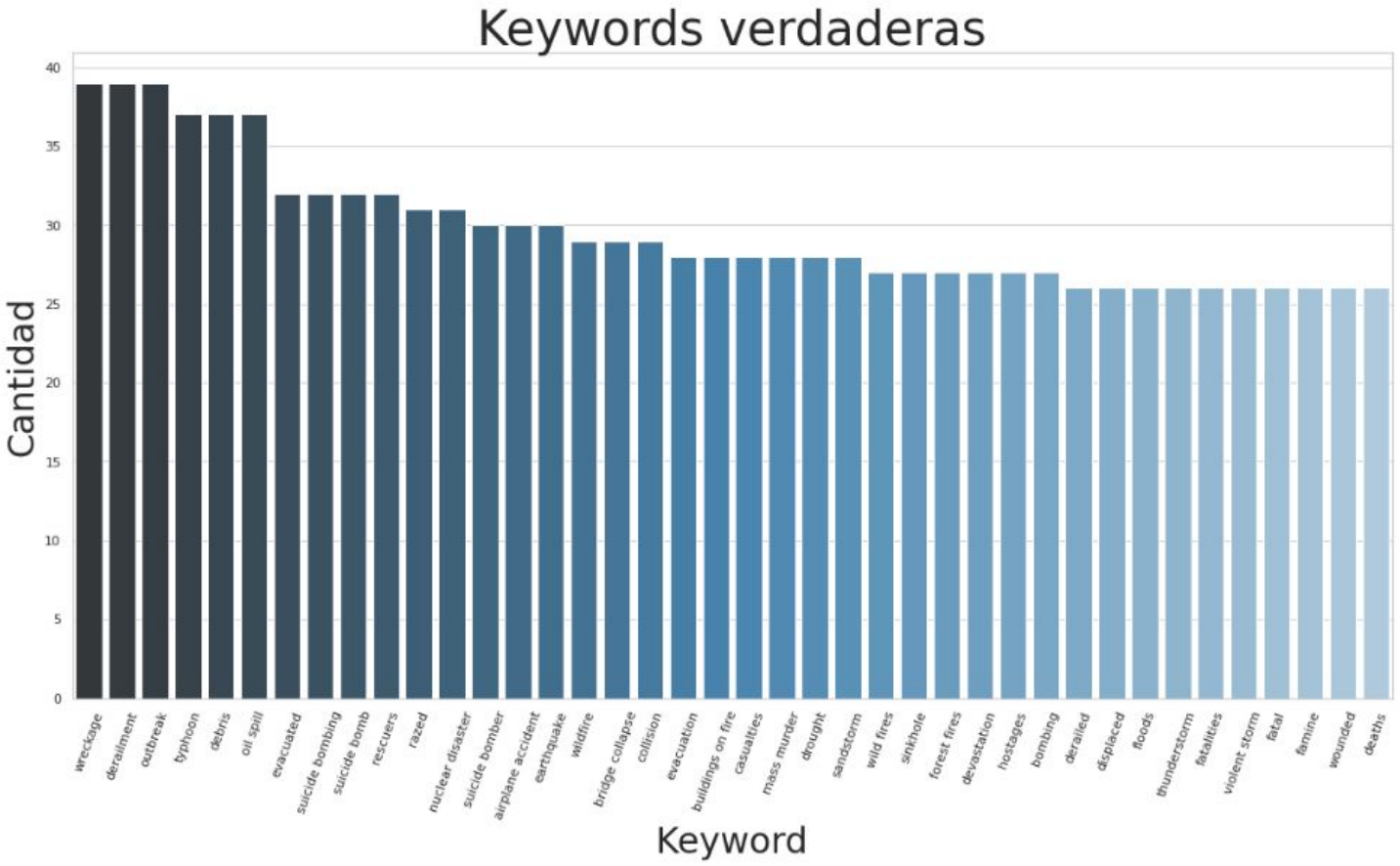


En general no hay una tendencia de la media, en algunos casos subió como con Florida, en otro bajo como con New York y otros quedó igual como Mumbai. En cuanto a los mínimos estos subieron salvo el caso de Florida y New York. Los máximos se mantuvieron estables y el rango de verdaderos se concentra entre los 100 y los 140 caracteres, muy cercano a lo que pasaba con los países, tal parece que ese es el rango más confiable para que un tweet sea verídico. Un caso especial es el de Mumbai que, recordemos, era el que más porcentaje de tweets verdaderos tenía pero por lo que vemos en el gráfico están dentro del rango antes mencionado a excepción de un único caso aislado con 60 caracteres.

De esta forma ya podemos darnos una idea de cuál va a ser el largo de los tweets que sean verdaderos sumado a los demás análisis para identificarlos más fácilmente.

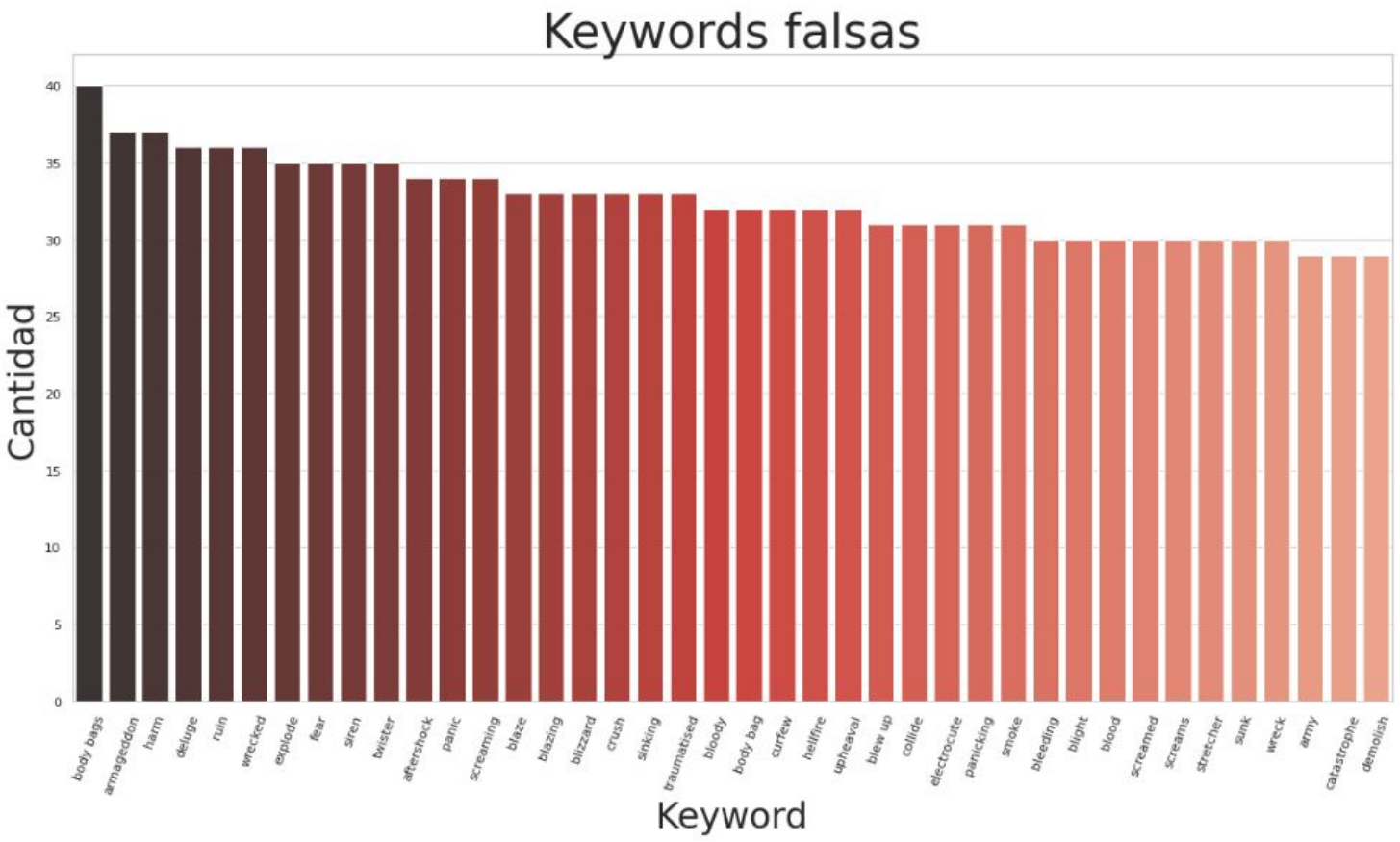
Keyword vs target

Al analizar los tweets verdaderos según su keyword obtenemos lo siguiente:



Lo que nos dice que los tweets acerca de desastres como outbreak, wreckage y derailment son los que tienen más cantidad de tweets verdaderos con 39 tweets.

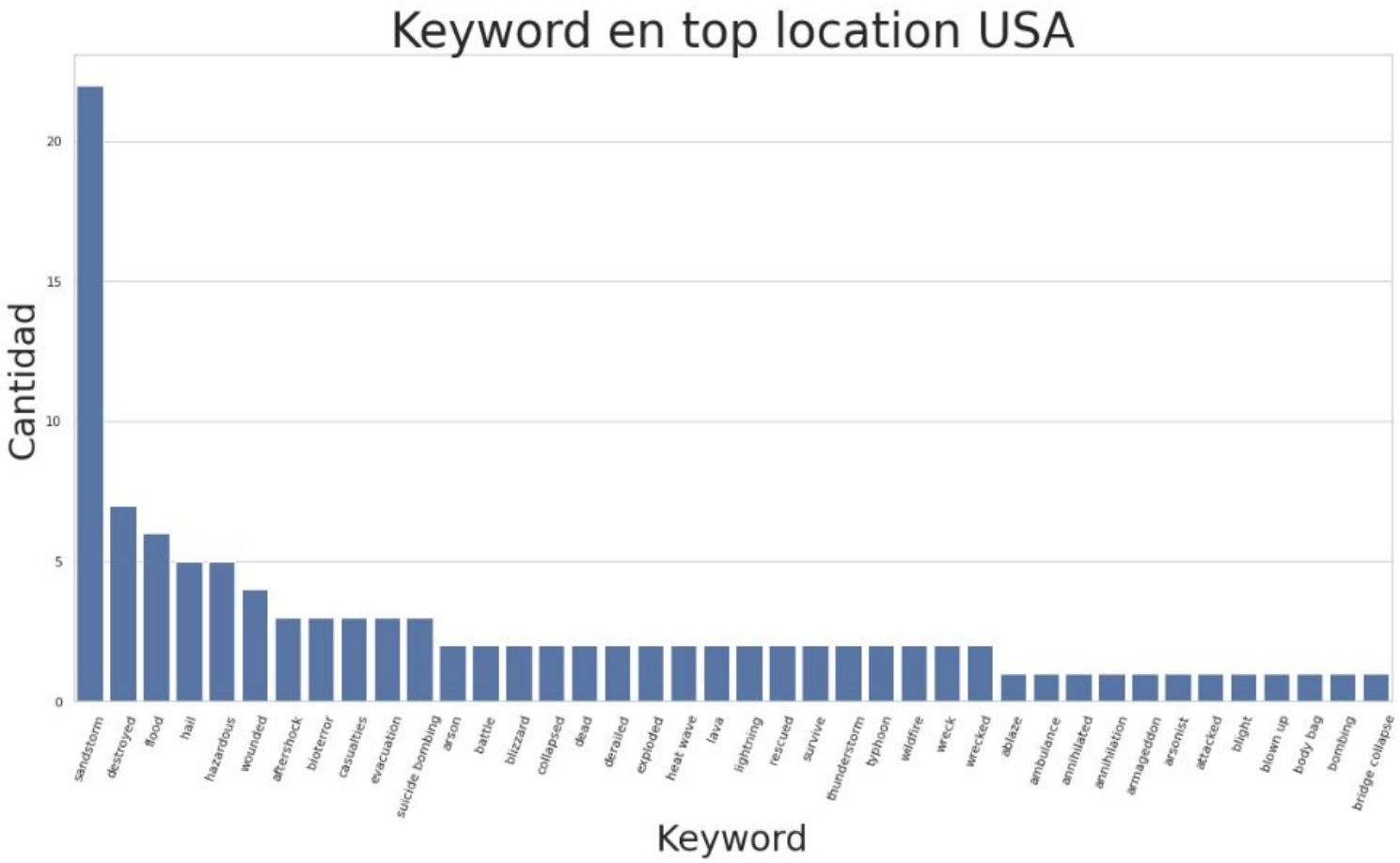
Luego al ver los falsos:



Encontramos que los tweets acerca de body bags son los que contienen más cantidad de tweets falsos con 40 tweets.

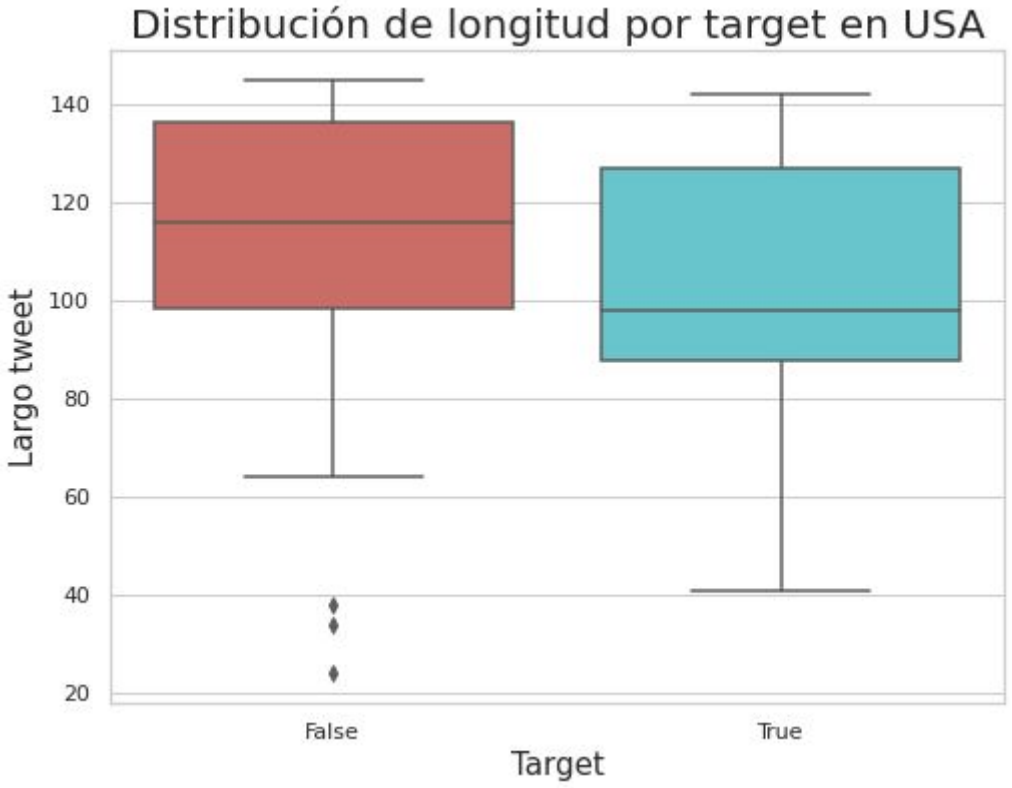
Top location

Al analizar los tweets de la ubicación con mayor cantidad de tweets (USA) obtuvimos lo siguiente:



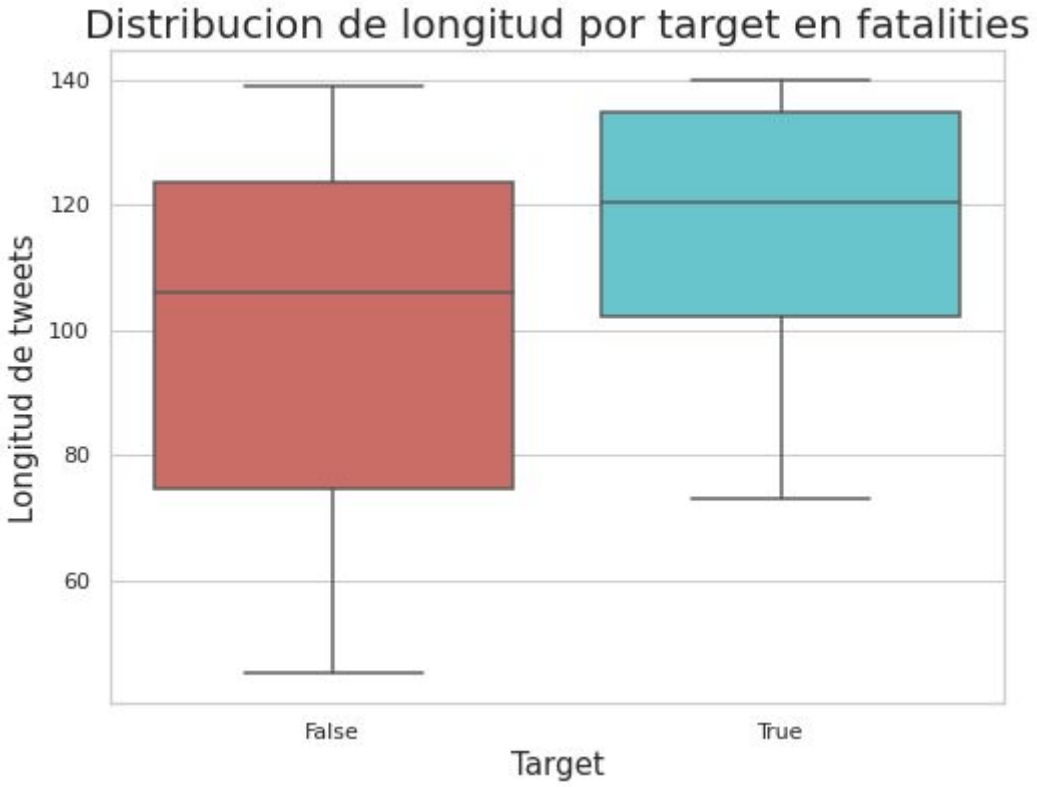
Esto nos permite afirmar que la mayoría de los tweets en esta ubicación son acerca de tormentas de arena (sandstorm).

En cuanto al largo de los tweets en comparación a si son verdaderos o falsos:



En comparación a cómo se comportan los tweets totales aquí vemos que los tweets falsos tienden a ser más largos que los verdaderos, el promedio de los verdaderos es más bajo que el general y el de los falsos subió.

Ahora veamos qué sucede con la keyword más usada en USA:



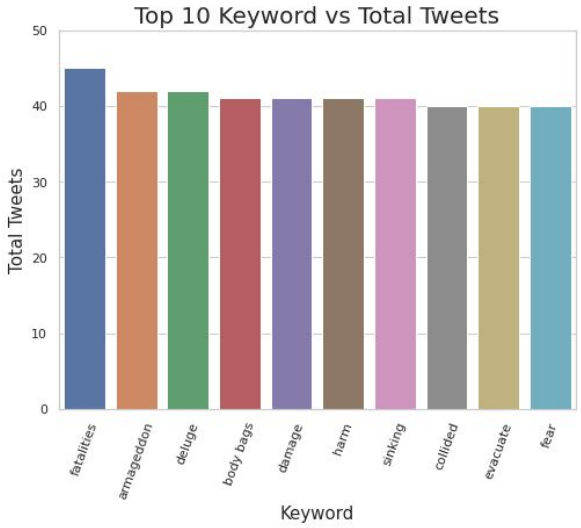
Nuevamente observamos que los promedios se mueven, comparados al gráfico anterior, los tweets con keyword fatalities verdaderos rondan los 120 caracteres mientras que los falsos está por debajo.

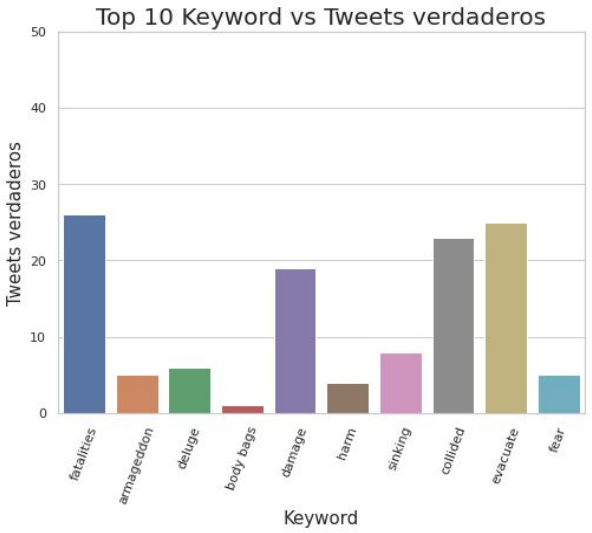
Top 10 Keyword

A continuación analizaremos las 10 keywords con más ocurrencias, en la columna sum se muestran la cantidad de tweets verdaderos con esta keyword y en count la cantidad de tweets totales.

	sum	count
keyword		
fatalities	26	45
armageddon	5	42
deluge	6	42
body bags	1	41
damage	19	41
harm	4	41
sinking	8	41
collided	23	40
evacuate	25	40
fear	5	40

A continuación deseamos ver si hay alguna relación entre los tweets reales de estas keywords y la cantidad



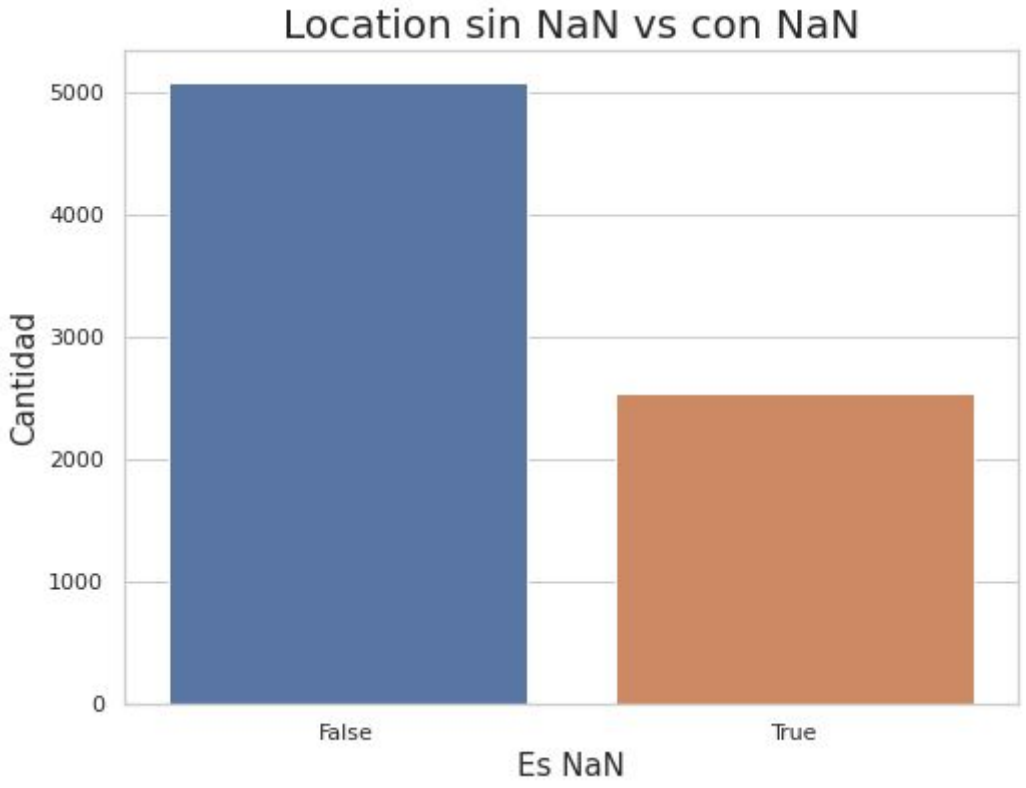


Con estos gráficos podemos ver que, salvo en casos como fatalities, collided y evacuate, hay más tweets falsos que verdaderos. Lo opuesto que con las

ubicaciones (cabe destacar que ninguna de estas top 10 se encuentra entre las keywords con más tweets verdaderos).

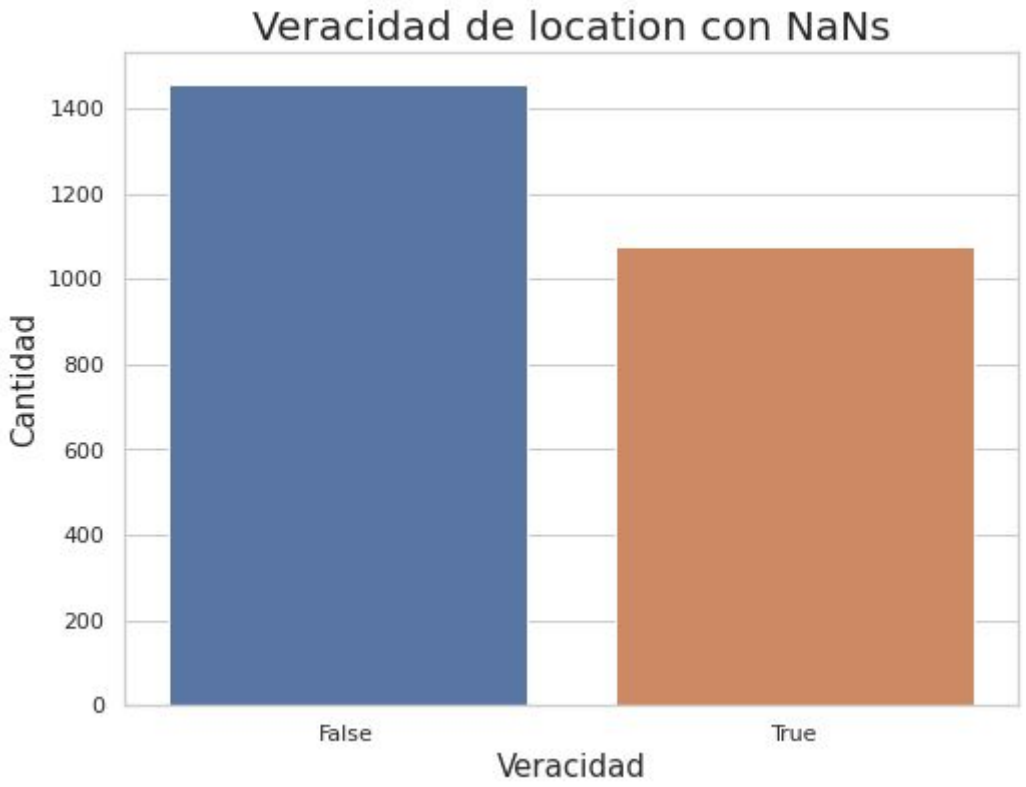
Locations NaN

Ahora queremos ver cuántos de nuestros tweets tienen como location un valor nulo o NaN . Al analizarlo obtenemos esto:



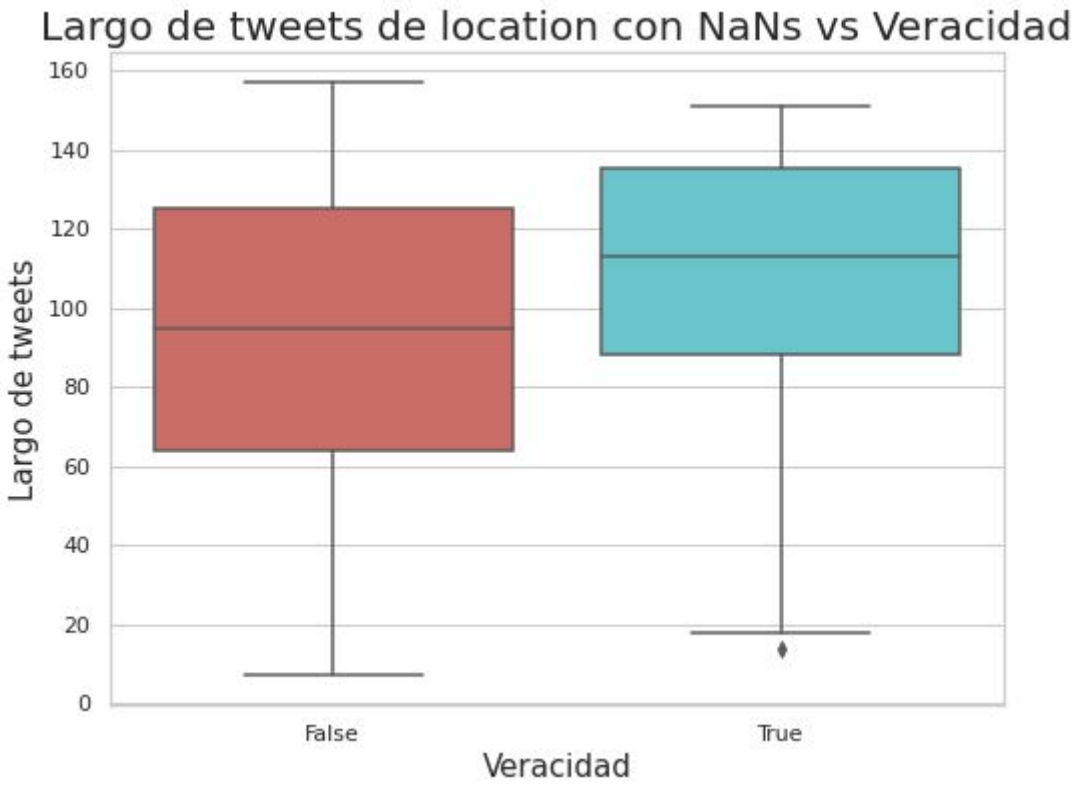
Lo que nos permite afirmar que hay 2533 tweets cuyo valor de location es NaN, lo cual representa aproximadamente un 33% de todos los datos.

Luego queremos ver cuántos de ellos son verdaderos y cuántos no:



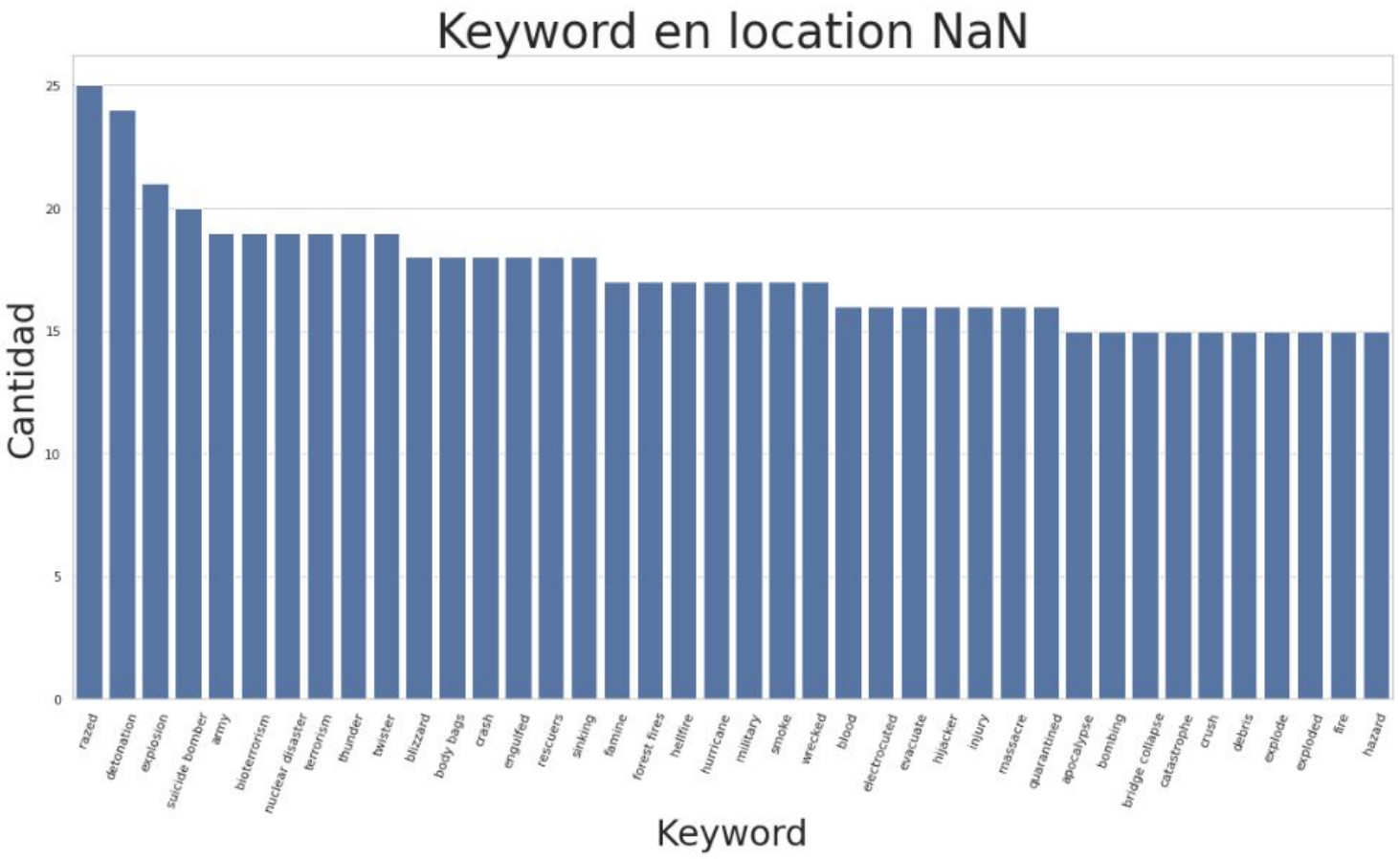
Con esto vemos que hay 1458 falsos y 1075 verdaderos haciendo que tengamos que tener en cuenta casi la mitad de los tweets con location NaN, ya que van a representar casi un séptimo de la cantidad total de tweets.

Veamos las propiedades de los largos de los tweets tanto verdaderos como falsos:



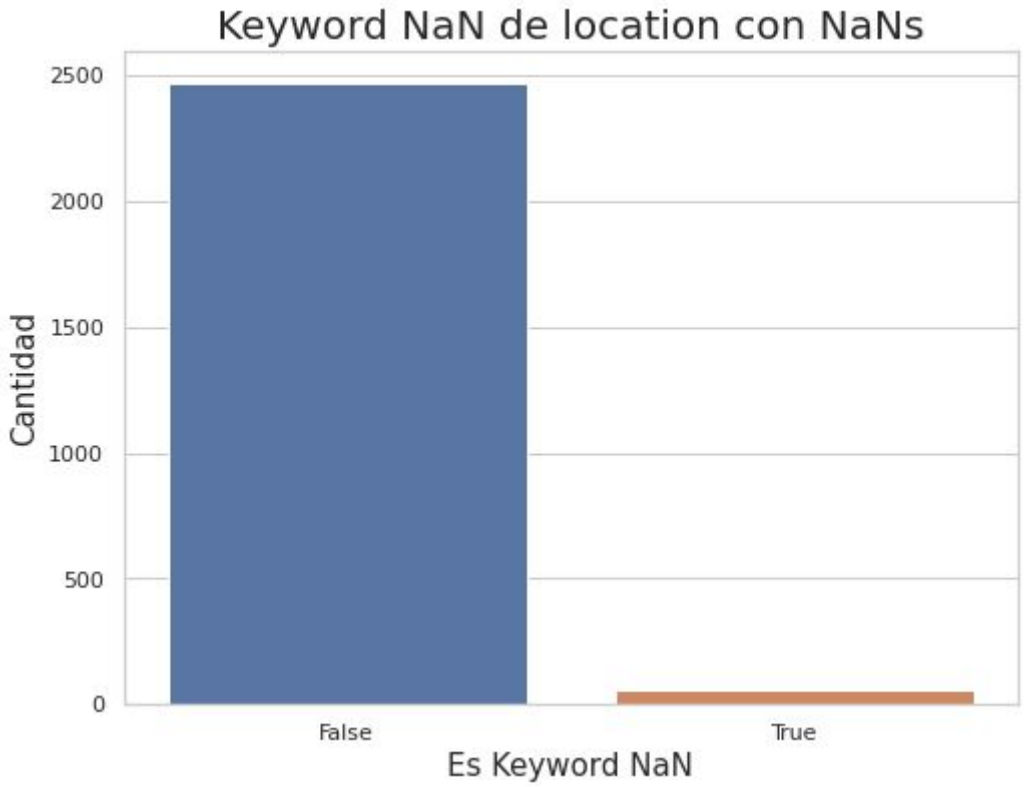
Podemos ver que los tweets verdaderos tienden a ser un poco más largos que los falsos, pero a su vez los tweets falsos tienen largos en ambos extremos (muy cortos y muy largos) que si bien no son la mayoría hay que tenerlos en cuenta para no confundirlos con los verdaderos.

Ahora queremos ver sobre qué tipos de desastres son estos tweets



Como vemos hay tanto keywords verdaderas como falsas vistas en los gráficos anteriores. Algo destacable es que las keyword con en los primeros puestos (por cantidad) pertenecen al grupo de las keyword con target verdadero.

Y finalmente veremos cuántos de ellos tienen keywords NaN:



Esto nos indica que hay 61 de estos tweets que tienen un NaN en su keyword (nótese que este es el número total de tweets sin keyword por lo que podemos afirmar que si un tweet no tiene keyword tampoco tiene location).

Conclusión

El análisis de este set de datos nos permite extraer ideas generales sobre cada uno de los factores que influyen sobre si un tweet refiere a un desastre natural verdadero o no. Entre estas ideas destacamos el hecho de que las palabras clave más repetidas no son compartidas entre cada valor del target, así como la variación de las longitudes de los tweets entre cada uno. A la luz de la cantidad de tweets analizados, estas relaciones sirven de guía para comenzar a predecir la veracidad sin contar con dicho valor como dato.