PIPELINE PROYECTO

Alumno: Jayan Cáceres Cuba
Curso: Tópicos en Ciencia de Datos

1 Tema/ Area de interes:
Twitter(tweets) Análisis visual

2 Contexto Corto:
Promover el bienestar es una de las metas clave de los Objetivos de Desarrollo Sostenible de las Naciones Unidas. Muchos gobiernos de todo el mundo están incorporando indicadores de bienestar subjetivo (SWB, por sus siglas en inglés) para complementar las métricas objetivas y económicas tradicionales. Nuestro Índice Geográfico de Sentimiento de Twitter (TSGI) puede proporcionar un monitor de alta granularidad del bienestar en todo el mundo.

Este conjunto de datos es un esfuerzo conjunto del Laboratorio de Urbanización Sostenible del MIT y el Centro de Análisis Geográfico de Harvard.

3 Análisis de la Dataset:
La dataset se divide en 3 archivos asociados a cada país y uno asociado al mundo

0. fecha , N , Score ,asociado al mundo

	DATE	N	SCORE
1	2017-01-01	483831	0.6823056014352119
2	2017-01-02	310587	0.6538674122741777
3	2017-01-03	362766	0.6425331066389904
4	2017-01-04	373639	0.6418016552849141
5	2017-01-05	355892	0.6413148010210963
6	2017-01-06	318655	0.6452813849586544
7	2017-01-07	702575	0.6549700258378109
8	2017-01-08	717768	0.6536327779449628
9	2017-01-09	629269	0.6427291661308597
10	2017-01-10	545323	0.6440940638208915
11	2017-01-11	559057	0.6454347286180121
12	2017-01-12	574718	0.645267756221312

1. fecha , nombre del país , N , score

	DATE	NAME_0	N	SCORE
1	2021-01-01	Afghanistan	271	0.6859137601476013
2	2021-01-01	Albania	200	0.6335534200000001
3	2021-01-01	Algeria	1385	0.6642562953068591
4	2021-01-01	American Samoa	2	0.592435
5	2021-01-01	Angola	395	0.5742634784810127
6	2021-01-01	Argentina	67323	0.6170747723957637
7	2021-01-01	Australia	19519	0.6295564949024027
8	2021-01-01	Austria	2764	0.6487709641823445
9	2021-01-01	Azerbaijan	1571	0.5834570464672184
10	2021-01-01	Bangladesh	2516	0.6994925612082671
11	2021-01-01	Belarus	1247	0.5870970232558138
12	2021-01-01	Belgium	5979	0.6302301941796287

2. fecha , nombre del país , provincia,N ,score

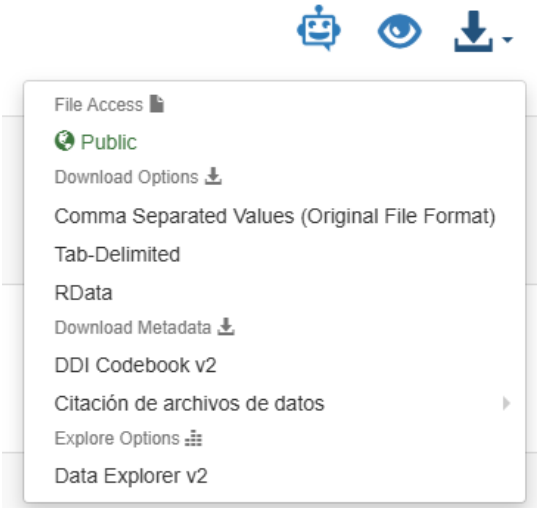
```
,DATE,NAME_0,NAME_1,N,SCORE
0,2012-01-01,Algeria,Algeria_Alger,1,0.88081
1,2012-01-01,Angola,Angola_Luanda,1,0.615514
2,2012-01-01,Argentina,Argentina_Buenos Aires,22,0.6396650909090908
3,2012-01-01,Argentina,Argentina_Ciudad de Buenos Aires,5,0.7094976000000001
4,2012-01-01,Argentina,Argentina_Corrientes,1,0.6325970000000001
5,2012-01-01,Argentina,Argentina_Córdoba,1,0.6038180000000001
6,2012-01-01,Argentina,Argentina_Jujuy,1,0.3746219999999999
7,2012-01-01,Argentina,Argentina_La Pampa,1,0.843993
8,2012-01-01,Argentina,Argentina_Mendoza,3,0.3078976666666667
9,2012-01-01,Argentina,Argentina_Salta,1,0.662137
10,2012-01-01,Argentina,Argentina_San Luis,1,0.8133670000000001
11,2012-01-01,Australia,Australia_Australian Capital Territory,4,0.47158449999999996
12,2012-01-01,Australia,Australia_New South Wales,35,0.5970145714285714
```

3. fecha , nombre del país , provincia , región ,N , score

```
DATE;NAME_0;NAME_1;Column1;SCORE;N
01/01/2023;Peru;Peru_Ancash;Peru_Ancash_Huaylas;0.647061;36
01/01/2023;Peru;Peru_Ancash;Peru_Ancash_Santa;0.6308600065288357;919
01/01/2023;Peru;Peru_Apurímac;Peru_Apurímac_Abancay;0.95817925;16
01/01/2023;Peru;Peru_Arequipa;Peru_Arequipa_Arequipa;0.665501626506024;83
01/01/2023;Peru;Peru_Ayacucho;Peru_Ayacucho_Huamanga;0.57068025;412
01/01/2023;Peru;Peru_Cajamarca;Peru_Cajamarca_Cajamarca;0.7029850272373542;257
01/01/2023;Peru;Peru_Cajamarca;Peru_Cajamarca_Chota;0.3622751538461539;26
01/01/2023;Peru;Peru_Cajamarca;Peru_Cajamarca_San Marcos;0.393291;19
01/01/2023;Peru;Peru_Callao;Peru_Callao_Callao;0.6634403191489361;235
01/01/2023;Peru;Peru_Cusco;Peru_Cusco_Cusco;0.6413751803607215;499
01/01/2023;Peru;Peru_Cusco;Peru_Cusco_Paruro;0.6423052844036699;109
01/01/2023;Peru;Peru_Huánuco;Peru_Huánuco_Puerto Inca;0.693452952275329;13976
01/01/2023;Peru;Peru_Ica;Peru_Ica_Chincha;0.689761279661017;118
```

Los archivos que contienen dataset de manera global pesan unos cuantos kbs.
Los archivos que contiene dataset asociado al país , pesan entre 10mb y 30 mb .
Los archivos que contienen dataset asociados al pais , provincia , pesan entre 50bm y 100mb
Los archivos que contienen dataset asociados al país, provincias y regiones pesan entre 300mb y 800 mb.

Delimitación de los archivos :
te permite descargarlo de varias maneras , la mejor manera es separado por comas para poder trabajar con un .csv

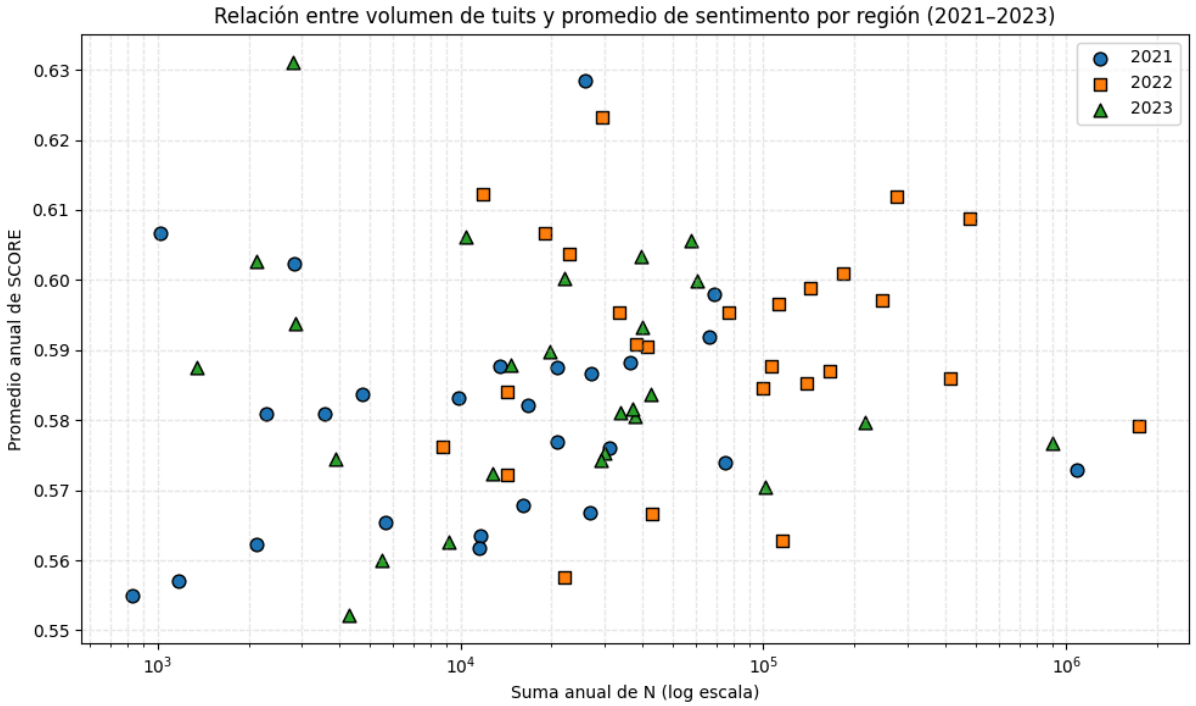


Problemas con la dataset:

Con respecto a la dataset ,se han identificado varios problemas , unos mencionados en el paper y otros con los que me encontré .

a. **Meses inconclusos de algunos años.**
   Esto se debe a que Twitter dejó de permitir el acceso gratuito a su API, pasando el acceso a la API a un modelo de pago a partir del 12 de julio de 2023.
b. **Ubicación opcional en las cuentas de Twitter**
   Antes de abril de 2015, la opción predeterminada para la captura de coordenadas GPS estaba activada para los usuarios de Twitter. Después de esta fecha, los usuarios han tenido que optar por compartir su ubicación precisa. Esta es una de las razones de la gran disminución en el volumen de geotweets después de esta fecha.
c. **Boots automáticos**
   una serie de tweet-bots automatizados que generan tweets con coordenadas (aparentemente) falsificadas aleatoriamente. Estos bots parecen representar no más que un pequeño porcentaje de los geotweets recolectados.

Descubrimientos al analizar la data



En un primer análisis que realiza con respecto a los años más recientes que me ofrecía la dataset .
tuve que utilizar una escala logarítmica para que se permita ver bien tanto los valores pequeños como los grandes, ya que en algunas provincias la cantidad de tweets era muy baja y en otras como la capital eran demasiadas.
Las figuras geométricas más a la derecha indican una mayor cantidad de tweets ,y las de la izquierda una menor cantidad.
Se puede apreciar una mayor cantidad por parte del 2022 con respecto al 2023 porque el 2023 no tiene todos los meses completos .