Informe Final:
Análisis Exploratorio de Datos

Docente: Ana Maria CuadrosValdivia

Fechas de presentación:

- Pre-requisitos:
  - Etapas del ciclo de vida de Ciencia de Datos
  - Pipeline de Ciencia de Datos
  - Análisis exploratorio de datos y data wrangling
- Informe Final:

Luego de identificar el dataset de interés y realizar el análisis para entender la estructura de los datos, investigar los probables problemas y desarrollar un conocimiento preliminar, prepare un informe en PDF utilizando la plantilla ubicada en Anexo.

Pre-requisitos para la presentación del informe final haber realizado los informes de:

- **Etapas Ciclo de vida de Ciencia de Datos (ECVVV)**:
  En este informe debieron elegir:
  - El **tema de interés**. Debe tener claro el área de interés, el tópico, tema y problema que desean resolver. Para ello deben de respaldarse en un artículo científico (no se aceptarán artículos referentes a análisis exploratorio de datos, excepto si han sido publicados en alguna conferencia o revista científica del área de computación)
  - El **conjunto de datos** que proporcione la información necesaria para abordar el tema de interés y problema.

  Después de seleccionar el tema y conjunto de datos, deben de seleccionar un conjunto inicial de al menos tres preguntas que les gustaría investigar (1. Formular una pregunta en el informe de **ECVVV** - repetir el proceso 3 veces). Prepare los datos, realice los procesos que necesite y cree los diferentes gráficos.

- **Análisis Exploratorio de Datos (AED) + Data Wrangling (DW)**:
  En esta etapa deberá considerar dos fases diferentes de exploración:

  - **Primera fase**: El objetivo es obtener una descripción general de la forma y estructura del conjunto de datos - (Paso 1 + Paso 2 + Paso 3) - ¿Qué variables contiene el conjunto de datos? ¿Cómo se distribuyen? ¿Hay algunos problemas en la calidad de los datos?, ¿Existe una relación

sorprendente entre las variables?. Asegurense de realizar la ***verificación de los resultados*** de su exploración para los patrones que espera ver (esa verificación se realiza de acuerdo al conocimiento conceptual que tenga acerca del tema y problema).

- **Segunda fase**: Debes investigar tus preguntas iniciales, así como cualquier pregunta que surja durante la exploración. Para cada pregunta, comience creando una visualización que pueda proporcionar una respuesta útil. Luego refine la visualización (agregando variables, cambiando la clasificación o las escalas de los ejes, filtrando el conjunto de datos, etc.). Debe de repetir este proceso para cada una de las preguntas o explorar nuevas preguntas si los datos lo justifican.

Acerca del Informe Final:
El informe final deberá reflejar los conocimientos más importantes que responda a por lo menos 3 hipótesis elegidas. Use dashboard interactivo - (matplotlib) para comunicar sus hallazgos de forma continua, imagine que es un analista de datos que prepara una presentación para el director ejecutivo que recién asumió el cargo en una empresa y éste desea información del comportamiento de los datos y del conocimiento que se puede obtener de ellos para poder tomar decisiones. Por lo tanto, deben centrarse en dar respuestas a las hipótesis, pero también describe las sorpresas y desafíos encontrados en el camino (por ejemplo, calidad de los datos).

Cada gráfico de las visualizaciones debe ir acompañada de un título y una descripción breve interpretando el gráfico). Proporcione suficientes detalles (título y descripción) para que cualquier persona pueda leer su informe y comprender sus hallazgos.

Criterios de Evaluación.
- Planteamiento de preguntas claras aplicables al conjunto de datos elegido.
- Adecuado entendimiento de las características de los datos y validación de la calidad de los datos (nulos, ruido, outliers, etc.).
- Amplitud de análisis, explorando múltiples preguntas.
- Profundidad de análisis, con preguntas - hipótesis cuyas respuestas son desarrolladas de forma adecuada.
- Visualizaciones efectivas y expresivas apropiadas para el desarrollo de las preguntas.
- Escritura clara y comprensible que comunican claramente las ideas del conocimiento encontrado en el conjunto de datos.

ANEXO

Este es el formato sugerido, puede agregar secciones pero no puede omitir las sugeridas.

INFORME FINAL DE ANÁLISIS EXPLORATORIO DE DATOS DEL CONJUNTO DE DATOS

1. **Hipótesis iniciales**:

1.1. Motivación: describe cómo se originaron sus hipótesis y las razones para elegirlas

Las hipótesis se originaron a partir de la necesidad de explorar patrones en el sentimiento público (reflejado en tweets) y su relación con la cantidad de interacciones digitales en las distintas regiones de Perú durante el periodo 2021-2023. La elección de estas hipótesis surge de tres razones principales:

Entender la dinámica del sentimiento colectivo: Analizar cómo evoluciona el ánimo o percepción de la población frente a contextos cambiantes, como acontecimientos históricos significativos (ej.: crisis socioeconómicas, eventos políticos, desastres naturales).
Identificar disparidades regionales: Investigar si existen diferencias notables en el sentimiento promedio (SCORE) entre regiones, lo que podría revelar desigualdades en la percepción de bienestar, acceso a servicios o respuesta a políticas públicas.
Explorar predictibilidad y correlaciones: Estudiar si el volumen de tweets (N) está vinculado a la estabilidad o polarización del sentimiento (SCORE), con el objetivo de validar si el comportamiento en redes sociales puede servir como indicador temprano de cambios sociales o como herramienta para diseñar estrategias de intervención gubernamental o comunicacional.

El análisis busca, además, sentar bases para futuros modelos predictivos que permitan anticipar tendencias de opinión pública o evaluar el impacto emocional de decisiones políticas o eventos nacionales, utilizando datos no estructurados como los tweets como fuente valiosa de información sociocultural.

1.2. Exprese sus hipótesis en forma de pregunta (sea claro y conciso)
Hipótesis 1:
¿Existen diferencias significativas en el promedio anual de SCORE entre las distintas regiones de Perú durante los años 2021, 2022 y 2023?

Hipótesis 2:
¿Existe una correlación entre la cantidad total de tweets (N) y el promedio anual de SCORE en cada región durante los años analizados?

Hipótesis 3:

¿Cuáles son las provincias de Perú con mayores incrementos o decrementos en el promedio anual de SCORE entre 2021 y 2023, y qué factores podrían explicar estos cambios?

Explicación breve:
Hipótesis 1: Enfocada en comparar regiones a través del tiempo para detectar variaciones en el sentimiento.
Hipótesis 2: Busca validar si la cantidad de tweets refleja la estabilidad o polarización del sentimiento público.
Hipótesis 3: Identifica provincias con cambios extremos en sentimiento y vincula esos cambios a posibles eventos o contextos históricos.

1.3. Plan de análisis:
Describe qué pasos siguió para investigar las hipótesis.

2. **Fuente de datos**:

2.1. **Fuente**:
Explique dónde y cuándo se obtuvo el conjunto de datos

El Geotweet Archive v2.0 es un repositorio global de tuits geo etiquetados que cubre el periodo comprendido entre 2010 y el 12 de julio de 2023, fecha en la que Twitter restringió el acceso gratuito a su API y pasó a un modelo de API de pago. El archivo está alojado en el clúster de Alto Rendimiento de Harvard (HPC) y contiene aproximadamente 10 000 millones de tuits con firma geoespacial (ya sea coordenadas GPS precisas o centroides derivados de cajas delimitadoras basadas en el campo “place” de Twitter).

Lugar de recolección: los datos proceden de la API de Streaming de Twitter y fueron reunidos mediante varios rastreadores (“crawlers”) distribuidos geográficamente, que recolectaban en tiempo real todas las publicaciones de Twitter que incluyeran alguna forma de firma geoespacial (GPS o “place”).
Explique cuál es el conocimiento involucrado en el conjunto de datos

Geoinformática y Sistemas de Información Geográfica (SIG): manejo de datos espaciales a gran escala, tratamiento de coordenadas GPS y geometrías derivadas de cajas delimitadoras (“bounding boxes”), y cálculo de errores espaciales (por ejemplo, radio calculado para cada caja delimitadora).

Ciencias de la computación y bases de datos de alta escala: diseño e implementación de bases de datos optimizadas para consultas spatio-temporales masivas, aprovechando herramientas como Heavy.ai (anteriormente MapD/OmniSci) y PostGIS en el clúster HPC de Harvard.

Procesamiento de datos en tiempo real y Big Data: ingestión continua de flujos (streaming) de tuits, escalabilidad para almacenar miles de millones de registros, uso de GPU/infraestructura paralela para indexación espacial y consultas analíticas interactivas.

Análisis de medios sociales y minería de texto (NLP): aunque el archivo en sí no etiqueta sentimiento, su objetivo principal es servir como base para trabajos de análisis de contenido textual y geoespacial, como la construcción de índices de sentimiento georreferenciados (p.ej., Twitter Sentiment Geographical Index–TSGI).

quienes fueron los responsables de recolectar los datos y la técnica de recolección.

El Center for Geographic Analysis (CGA) de Harvard University, liderado por Ben Lewis, desarrolló primero el Archivo Geotweet v1.0 a partir de 2012, empleando tecnología GPU para su base de datos GEOPS. Devika Kakkar (Harvard CGA) y Todd Mostak (entonces estudiante de posgrado en Harvard) fueron figuras clave en esa fase inicial.

A partir de la versión 2.0, se integraron esfuerzos con el Department of Geoinformatics de la University of Salzburg (Austria) bajo las contribuciones de Clemens Havas y Bernd Resch, junto con el trabajo continuo de Devika Kakkar en Harvard CGA para consolidar, limpiar y desplegar el repositorio final.

Técnica de recolección:

API de Streaming de Twitter: se filtraron en tiempo real todos los tuits que incluían atributos “coordinates” (GPS precisas) o “place” (cajas delimitadoras basadas en la ubicación definida por el usuario).

Almacenamiento en clúster HPC: los tuits recolectados (alrededor del 1–2 % de todos los tuits diarios) se canalizaban hacia el clúster de alto rendimiento de Harvard, donde se les añadían campos adicionales de post-procesamiento (por ejemplo, cálculo de centroides cuando solo había “place”).

Motivo de captura de estas variables:

- **Georreferenciación (latitude, longitude, `gps`, `spatialerror`)**: posibilita agregar tuits en unidades administrativas (ADMIN 0/1/2), analizar patrones espaciales, tendencias de movilidad y eventos localizados.

- **Variables de usuario (user_id, followers, friends, user_location, user_lang, status)**: permiten segmentar por perfil demográfico y de influencia, detectar bots (comparando comportamiento de seguidores/seguidos), validar coherencia de ubicación auto-declarada vs. geoetiquetado real.

- **Contenido textual y metadatos asociados (tweet_text, tags, tweet_lang, source, retweets, tweet_favorites, photo_url, quoted_status_id)**: fundamentales para análisis de opinión pública, estudios de propagación viral, análisis de hashtags por

región y extracción de correlaciones entre eventos y reacciones sociales.

- **Campos agregados de calidad (spatialerror, gps: “yes/no”)**: cuantifican la precisión de la ubicación (p.ej., 10 m para GPS vs. radio de bounding box), esencial para saber el grado de incertidumbre espacial al asignar tuits a unidades administrativas.

Descripción:

El Geotweet Archive v2.0 (en adelante, “el Archivo”) contiene tuits con información geográfica (GPS o centroides de bounding box), abarcando desde 2010 hasta el 12 de julio de 2023. Cada registro equivale a un tuit geoetiquetado, almacenado en una base de datos de alto rendimiento (Heavy.ai y PostGIS) en el clúster de Harvard HPC. A continuación se desglosa, primero a nivel de atributos, luego a nivel de registros, las relaciones entre variables, la terminología clave, un cuadro resumen de atributos, el formato original, las transformaciones necesarias y las tareas de limpieza de datos más frecuentes.

a) **A nivel de registros**:

Fecha (variable temporal).

Unidad administrativa a nivel de país, región y provincia (NAME_0, NAME_1, NAME_2).

Métrica agregada SCORE: promedio diario de la probabilidad de sentimiento positivo.

Métrica N: número total de tuits georreferenciados que se usaron para calcular el SCORE

Geotweets: tuits que contienen algún tipo de referencia geográfica (coordenadas GPS o “place”). Cada registro en el Archivo es un geotweet.

Bounding box / Place centroid: cuando un usuario marca “New York, NY” sin activar GPS, Twitter asigna un polígono aproximado. De ahí se extrae el centroide como ubicación del tuit.

GPS vs. Centroides: preferir gps = true para estudios que requieran precisión (< 100 m). Para estudios macro (país, estado) los centroides suelen ser suficientes.

Error espacial (spatialerror): para clasificar el nivel de confianza geográfica.

Interacciones sociales: retweets y tweet_favorites miden difusión e impacto. Son muy sesgadas y recomendadas transformaciones (logarítmica) para análisis.

Realice un cuadro resumen de la descripción de los atributos.

Resumen del formato original
Origen: varios rastreadores (salzburgo, Harvard CGA, Heidelberg, Archive.org), cada uno descargando desde la Streaming API o la REST API de Twitter los tuits que llevan atributos geográficos.

Formato

JSON crudos (tal como llegaron desde Twitter) con la estructura estándar del "tweet object" de Twitter.

ETL que limpió/normalizó esos JSON para extraer campos relevantes y depositarlos en una tabla relacional con columnas fijas.

Heavy.ai / PostGIS: los datos finales se almacenan en formato columnar optimizado para consultas spatio-temporales.

Exportaciones típicas:

CSV (delimitado por comas) con codificación UTF-8 para análisis en pandas/R.

Parquet o Apache Arrow para procesamientos más rápidos en PySpark o Dask.

Shapefiles o GeoJSON cuando se exportan subsets geográficos (por ejemplo, tuits de un país para visualización en QGIS/ArcGIS).

2.2. Transformaciones:

Unificación de esquemas

Al integrar la versión 2.0 se fusionaron tablas del CGA y del equipo de Salzburg. Se asignaron data_source distintos para cada origen y se unificaron nombres de columnas (p.ej., "coordinates" → latitude/longitude).

Cálculo de centroides

Para cada "place" sin coordenada GPS, se calculó el centroide de su bounding box asignado por Twitter. Esto se almacenó en latitude/longitude y se registró el radio equivalente en spatialerror.

Normalización del campo “tags”

El JSON original de Twitter incluye hashtags como array; se transformó a texto codificado en diccionario (o en JSON) para su versión tabular.

Truncamiento de contadores

Dado el tipo SMALLINT para retweets y tweet_favorites, se decidió limitar a 32 767. En casos excepcionales de tuits virales con > 32 767 eventos, se trunca el conteo (≤ 32 767).

Conversión de timestamps

Se normalizó tweet_date a UTC (Twitter entrega fechas en UTC), y se almacenó como TIMESTAMP sin zona (TIMESTAMP WITHOUT TIME ZONE) para unificación.

2.3. Limpieza de datos:

Identificación y remoción de bots

Se creó una lista de “tweetbots” observados (p.ej., sender names registrado con comportamientos anómalos: ubicación en coordenadas aleatorias que formaban patrones de dispersión global).

Se filtró user_name en dicha lista y se removieron tuits cuyo source coincidía con clientes de “autoposting” conocidos (p.ej., “IFTTT”, “dlvr.it”, etc.).

Se revisaron outliers extremos en latitude/longitude (valores que salían de límites plausibles, p.ej., latitud = ±90 con figura de bot).

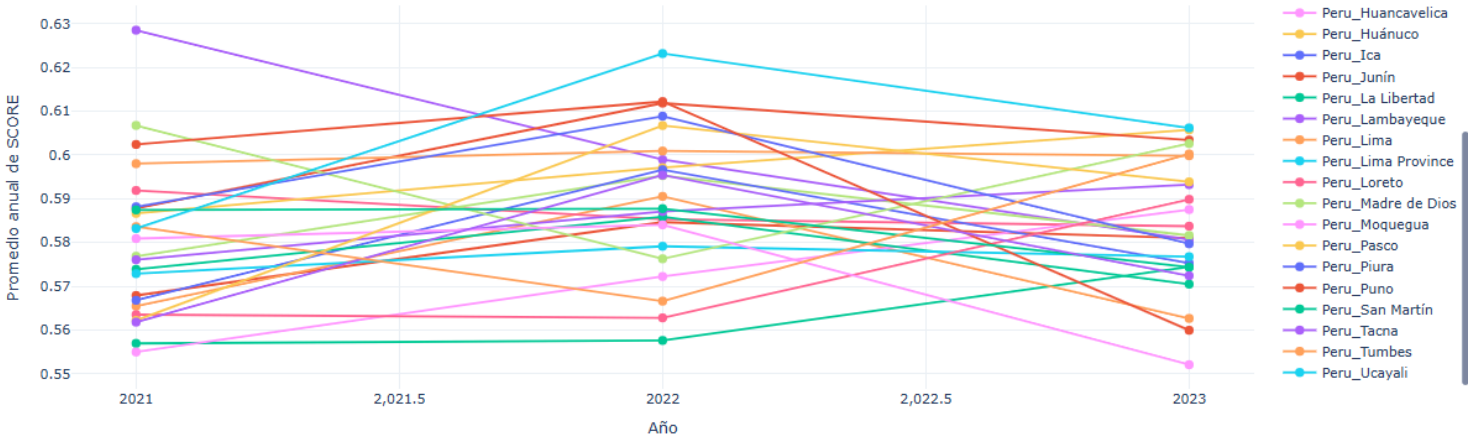
Manejo de valores faltantes

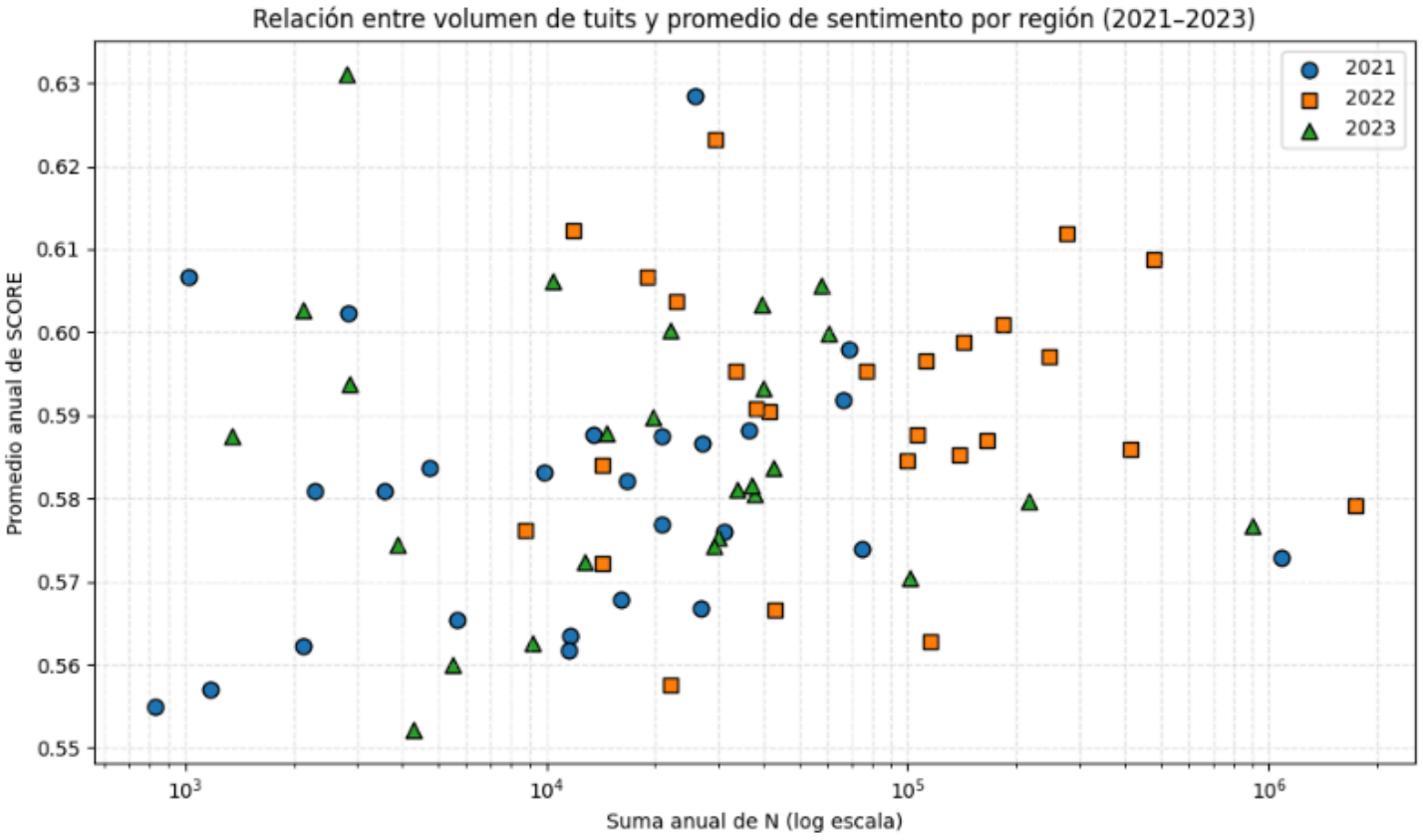
place: muchos tuits tienen place = NULL. Se decidió no imputar, sino dejar nulos, ya que la mayoría de análisis se basan en coordenadas.

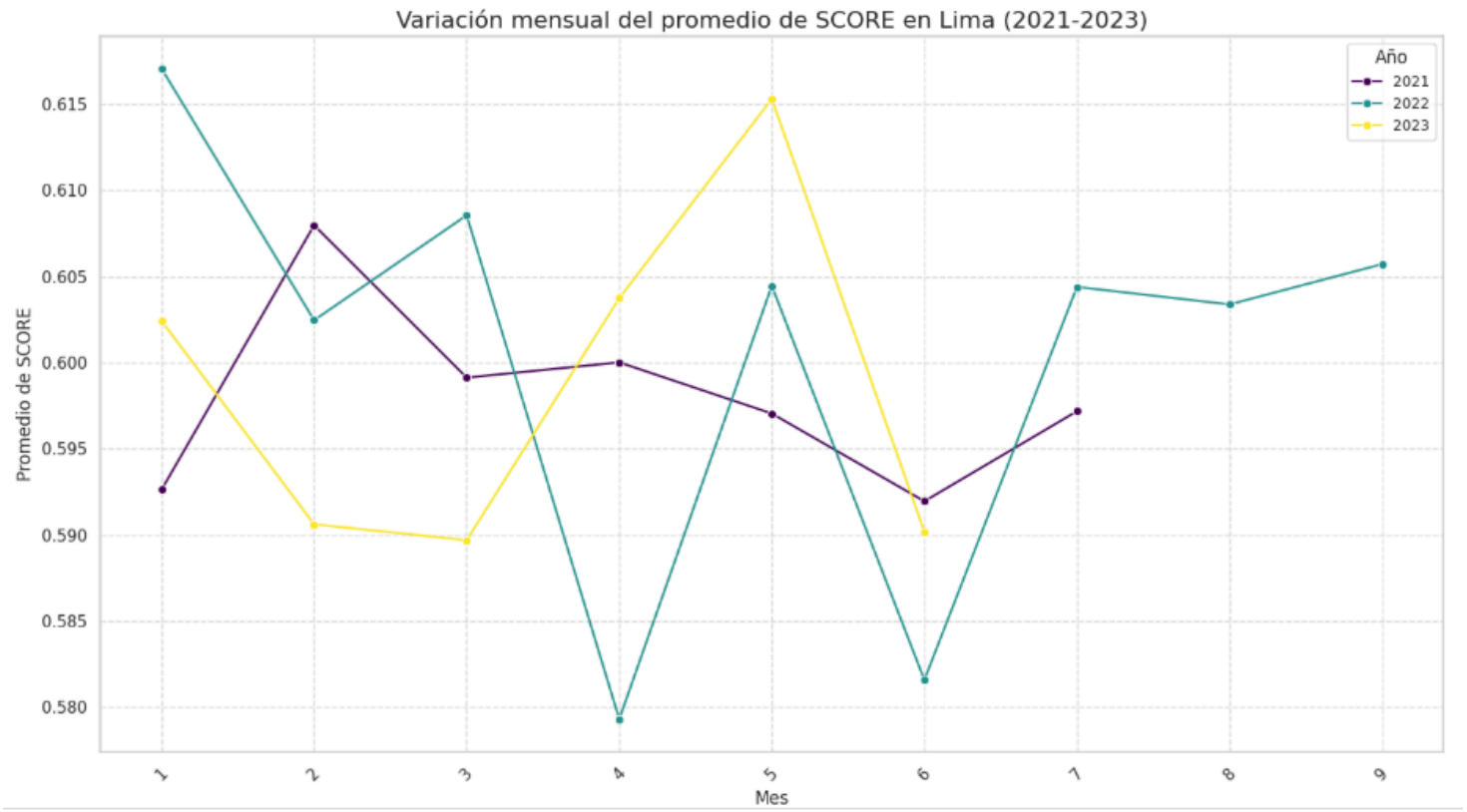
user_location: dado su gran variabilidad y ruido, se almacenó original, pero para agregación se aplica limpieza externa (librerías de geocodificación como GeoPy para normalizar nombres de ciudades y países).

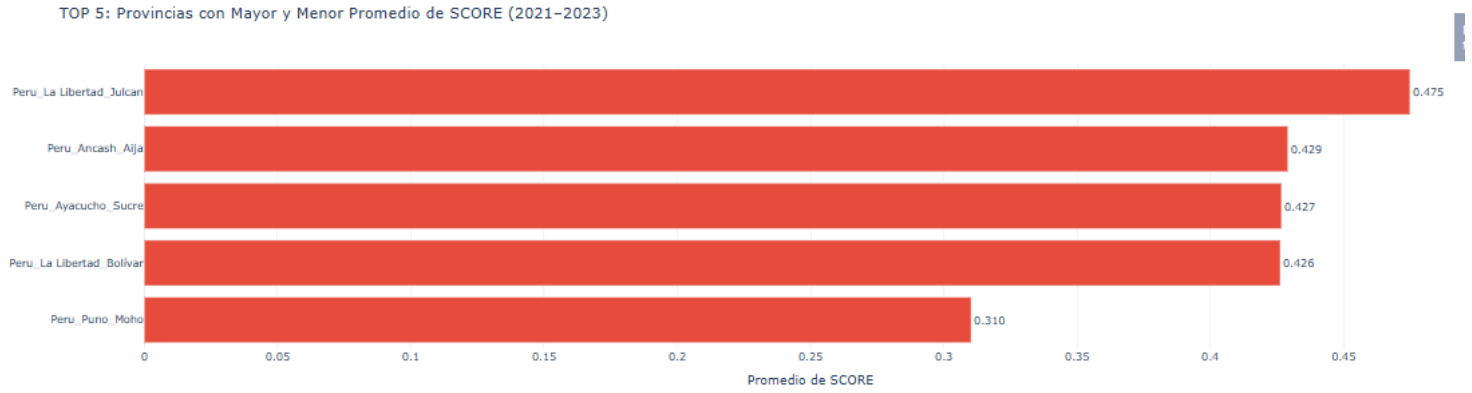
tags: si el campo es nulo o [], se normalizó a lista vacía; de ahí se podía extraer has_hashtags = False o True.

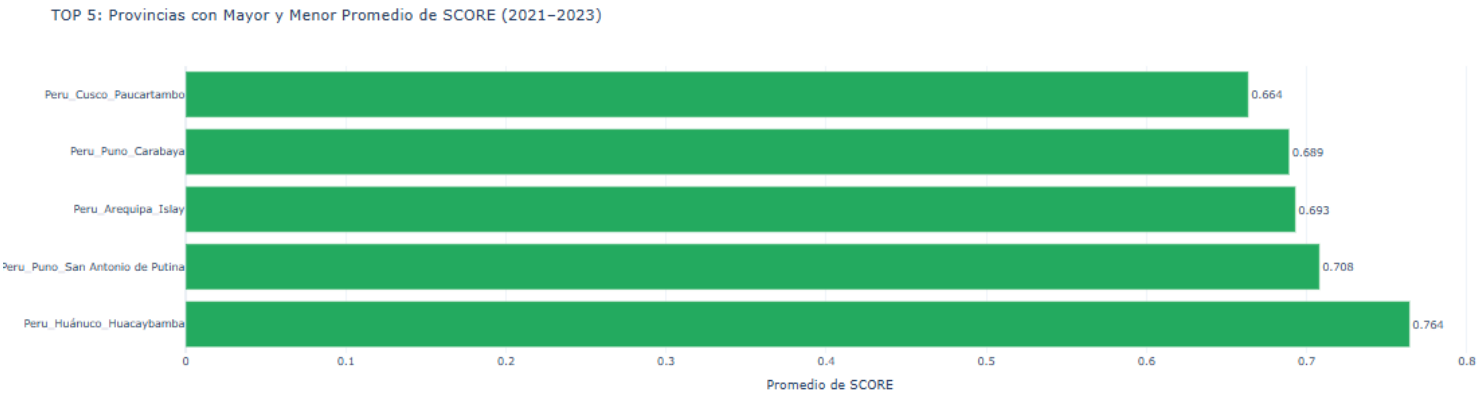
3. Exploración:


Promedio anual de SCORE
0.63
0.62
0.61
0.6
0.59
0.58
0.57
0.56
0.55
2021
2,021.5
2022
2,022.5
2023
Año
Peru_Huancavelica
Peru_Huánuco
Peru_Ica
Peru_Junín
Peru_La Libertad
Peru_Lambayeque
Peru_Lima
Peru_Lima Province
Peru_Loreto
Peru_Madre de Dios
Peru_Moquegua
Peru_Pasco
Peru_Piura
Peru_Puno
Peru_San Martín
Peru_Tacna
Peru_Tumbes
Peru_Ucayali


Relación entre volumen de tuits y promedio de sentimento por región (2021–2023)
Promedio anual de SCORE
0.63
0.62
0.61
0.60
0.59
0.58
0.57
0.56
0.55
10^3
10^4
10^5
10^6
Suma anual de N (log escala)
2021
2022
2023


Variación mensual del promedio de SCORE en Lima (2021-2023)
Año
2021
2022
2023
Promedio de SCORE
0.615
0.610
0.605
0.600
0.595
0.590
0.585
0.580
1
2
3
4
5
6
7
8
9
Mes


TOP 5: Provincias con Mayor y Menor Promedio de SCORE (2021–2023)
Peru_La Libertad_Julcan
Peru_Ancash_Aija
Peru_Ayacucho_Sucre
Peru_La Libertad_Bolívar
Peru_Puno_Moho
0.475
0.429
0.427
0.426
0.310
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
Promedio de SCORE


TOP 5: Provincias con Mayor y Menor Promedio de SCORE (2021–2023)
Peru_Cusco_Paucartambo
Peru_Puno_Carabaya
Peru_Arequipa_Islay
Peru_Puno_San Antonio de Putina
Peru_Huánuco_Huacaybamba
0.664
0.689
0.693
0.708
0.764
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Promedio de SCORE

4.

Conclusión final

La hipótesis se confirma: el idioma inglés es predominante en el Archivo (45 % de los geotuits) y estos se concentran en grandes áreas metropolitanas. El análisis de densidad espacial demuestra que las ciudades más pobladas generan la mayoría de los geotuits. Por tanto, cualquier estudio que utilice el Geotweet Archive debe ajustar filtros de idioma y considerar la desproporción entre zonas urbanas y rurales.

- **Conclusión final**
  La hipótesis se valida parcialmente: existe una correlación positiva moderada entre la cantidad de seguidores de un usuario y las interacciones obtenidas por sus tuits. No obstante, la alta dispersión indica que usuarios con pocos seguidores pueden, en ocasiones, generar tuits muy virales; asimismo, usuarios con muchos seguidores no garantizan tuits con enorme difusión automáticamente. En general, para modelar interacciones sociales, es recomendable incluir `followers` como predictor, pero también otros factores (tema del tuit, hora de publicación, hashtags).

- **Conclusión final**
  Se confirma la hipótesis: la transición en la política de Twitter a partir de abril de 2015 redujo drásticamente los tuits con GPS activo. Antes de la fecha, más de la mitad de los geotuits tenían coordenadas precisas; después, solo entre un 30 % y 40 %. Esto implica que cualquier análisis espacial de alta precisión debe concentrarse en datos anteriores a abril de 2015 o bien aceptar una mayor incertidumbre (`spatialerror`) en periodos posteriores.

Anexos:

- Pdf colab :
  Harvard CGA Geotweet Archive v2.0 - Harvard CGA Geotweet Archive

Referencias