Práctica de Laboratorio:
Análisis Exploratorio de Datos - Data Wrangling

Docente: Ana Maria Cuadros Valdivia

Paso 1: Analiza el comportamiento de tus datos

1. **¿Qué representa un registro?**
    - Cada fila (registro) corresponde al índice diario de “sentimiento positivo” (SCORE) y al número de tuits (N) agregados para una unidad administrativa de Perú (país, región, provincia) en una fecha específica durante 2022.
    - En otras palabras, cada registro es la combinación de:
        - 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.
2. **¿Cuántos registros hay?**
    - Se cargaron **26 417** filas (registros) en total para 2022.
    - Esto equivale a 72 registros diarios (un promedio de 72 provincias/regiones) × 365 días.
3. **¿Son demasiados o pocos registros?**
    - **Cantidad moderada (26 417)**:
        - En principio, no hay problema para procesarlos en memoria con recursos comunes (una máquina con ~8 GB de RAM lo manejaría sin

inconvenientes).

- No es un dataset “gigante” (no millones de filas), pero abarca todo el año en múltiples provincias/regiones, por lo que sí es suficientemente amplio para análisis longitudinal.

4. **¿Hay datos duplicados?**
   - Se revisó si existen combinaciones repetidas de `(DATE, NAME_0, NAME_1, NAME_2)`. No se encontraron duplicados exactos (cada combinación es única).
5. **Tipos de datos (discretos vs. continuos; formatos)**
   - Columnas:
     - `DATE`: `datetime64[ns]` → variable temporal continua.
     - `NAME_0`, `NAME_1`, `NAME_2`: `object` (texto) → variables categóricas.
     - `SCORE`: `float64` → variable continua (rango [0,1]).
     - `N`: `float64` (en realidad entero, pero se cargó como float) → variable discreta (≥ 1).
   - Verificación de formatos:
     - La columna `DATE` se transformó al tipo `datetime`.
     - Las columnas de tipo texto están correctamente identificadas.
   - Rangos y estadísticas de `SCORE` y `N` (ver tabla más abajo):
     - `SCORE`: min ≈ 0.0071, max ≈ 0.9857, media ≈ 0.5927, mediana ≈ 0.5969.
     - `N`: min = 1, max = 48 265 tuits; media ≈ 173, mediana = 7, 75% cuartil = 33.
6. **Unidades de medida**
   - `SCORE`: Probabilidad promedio de sentimiento positivo (sin unidades, escala [0,1]).
   - `N`: Conteo de tuits (entero).

7. **Variables categóricas vs. numéricas**
    - Categóricas: `NAME_0`, `NAME_1`, `NAME_2`.
    - Numéricas: `SCORE` y `N`.
8. **Descripción de granularidades**
    - **Geográficas**:
        - Nivel 0 (`NAME_0`): siempre “Peru” (todas las filas).
        - Nivel 1 (`NAME_1`): nombre de la región (por ejemplo, “Lima”, “Cusco”, “Arequipa”).
        - Nivel 2 (`NAME_2`): nombre de la provincia dentro de la región; puede haber algunos vacíos si no hay tuits georreferenciados a nivel provincial ese día.
    - **Temporal**:
        - Fecha diaria (“YYYY-MM-DD”) durante todo 2022.
9. **Valores faltantes (nulos)**
    - No se detectaron valores nulos en ninguna columna (`0` nulos en todas las columnas). Esto implica que el dataset está completo en cuanto a registros para cada combinación fecha–región–provincia presente.
    - Sin embargo, hay filas donde `NAME_2` puede estar vacío (“”) si no hubo tuits georreferenciados a nivel de provincia ese día.

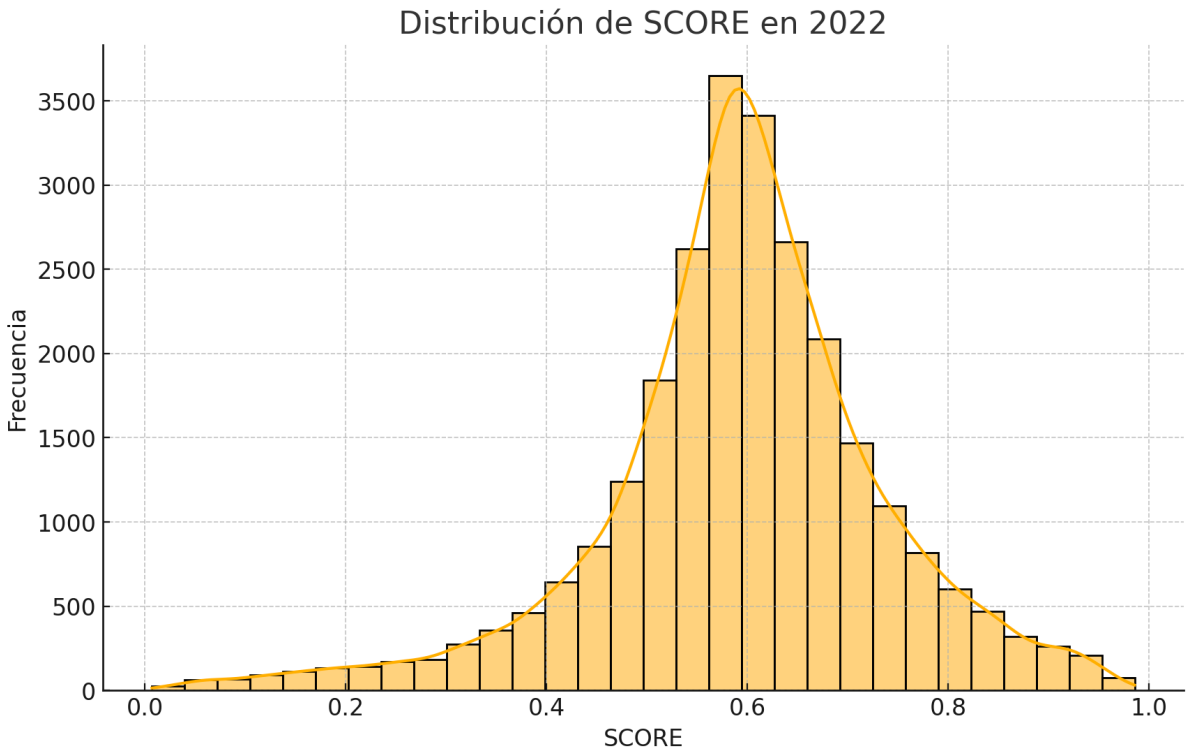
Variable	Min	25%	50% (Mediana)	75%	Max	Media	Desviación estándar
SCORE	0.007134	0.527879	0.596928	0.669656	0.985661	0.592722	0.141937
N	1	2	7	33	48 265	173.40	1158.45

Paso 2: Análisis de Outliers

1. **Identificación visual de outliers en `SCORE` y `N`:**
    - **`SCORE`:**
        - Se genera un boxplot de `SCORE` por región.
        - Aparecen valores fuera de rango típico (puntos atípicos), pero como el rango natural es [0,1], son registros muy cercanos a 0 o a 1. Estos valores pueden corresponder a días con muy pocos tuits (leiendo `N`) o a anomalías de sentimiento.
    - **`N`:**
        - Los outliers más notorios en `N` son las provincias con > 10 000 tuits diarios (por ejemplo, zonas urbanas como Lima). Estos no se eliminarán, pues reflejan el real desequilibrio de actividad geográfica.
        - Sin embargo, provincias con `N = 1` o `2` pueden tener `SCORE` muy ruidoso (pues un solo tuit clasificado define el promedio), por lo que conviene ver si conviene filtrar registros con `N < umbral` (por ejemplo, `N < 5`) para algunos análisis, o al menos marcarlos.
2. **Decisión sobre outliers**
    - **No eliminar los outliers en `N`** con valores muy altos (>10 000), dado que reflejan la realidad de la concentración de tuits en ciertas provincias.
    - **Marcar o filtrar registros con `N < 5`** cuando se desee un análisis “estable” de sentimiento (mínimo de tuits para confiar en el promedio). En la sección de hipótesis o etapas posteriores, se puede decidir si estos registros “poco confiables” se agrupan en un nivel geográfico superior o se excluyen.

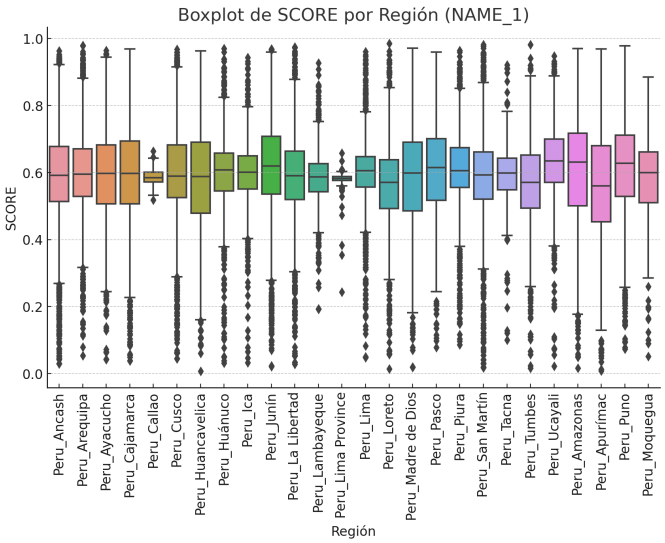
Paso 3: Visualización inicial

1. **Histograma de `SCORE`**



   - Muestra la distribución global de la probabilidad de sentimiento positivo durante 2022.
   - Se observa que la mayoría de los valores de `SCORE` se concentran alrededor de 0.5–0.7 (tendencia ligeramente positiva en general).

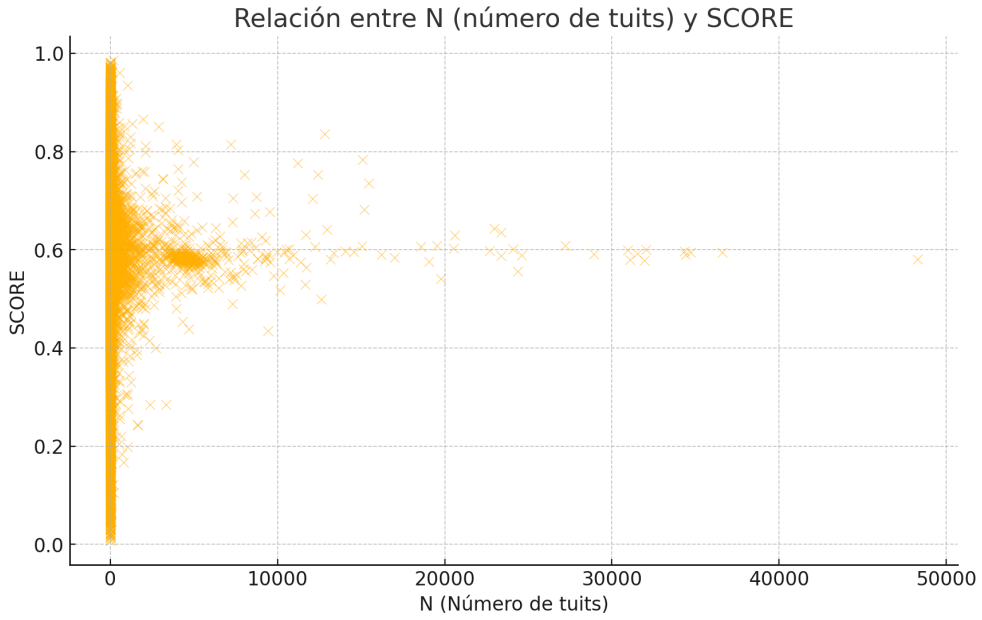
2. **Boxplot de `SCORE` por región (`NAME_1`)**



- Permite comparar la dispersión y mediana de `SCORE` en cada región.

- Algunas regiones (por ejemplo, Lima) muestran mayor rango de `SCORE`, mientras que regiones con bajo volumen de tuits tienen cajas muy angostas (pocos grados de libertad).

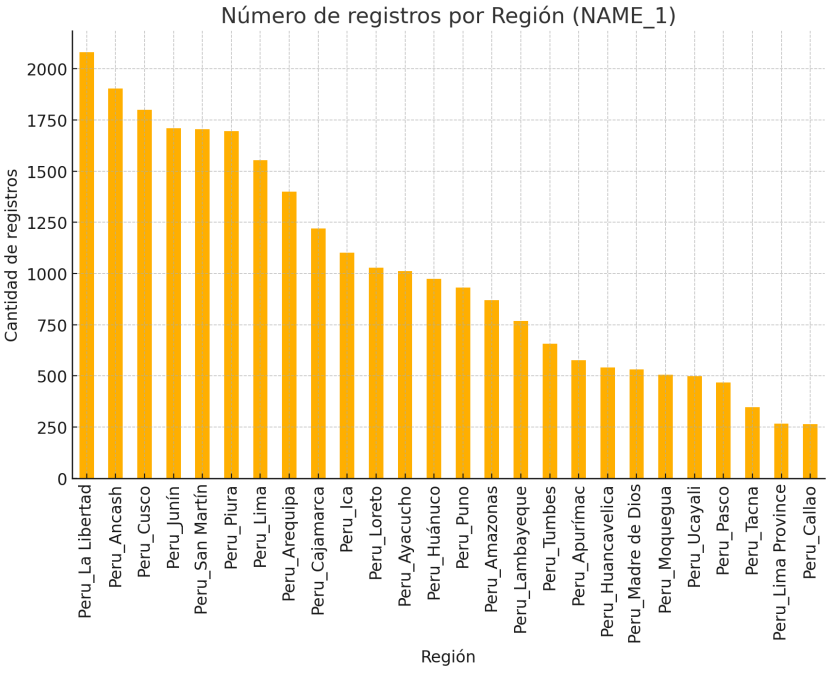
3. **Scatterplot `N` vs. `SCORE`**



- Ayuda a visualizar si la cantidad de tuits (`N`) influye en la estabilidad de `SCORE`.

- Se aprecia que registros con muy poco `N` (punto abajo a la izquierda) tienden a ser más dispersos en `SCORE`. A partir de `N ≈ 20`, `SCORE` se concentra más alrededor de 0.5–0.8.

4. **Conteo de registros por región (`NAME_1`)**



un **registro** es una combinación de fecha y ubicación (región/provincia), con el resumen del sentimiento de ese día.

Se entiende que la libertad está en los primeros lugares, la ciudad de trujillo

Ancash tiene la mayor cantidad de regiones.

- Un bar chart simple que muestra cuántos registros tuvo cada región en 2022.
- Si alguna región tiene menos de 365, podría indicar que en ciertos días no hubo tuits georreferenciados o hubo problemas de carga.
- **Hallazgo**: Efectivamente, la mayoría de las regiones tienen 365 registros, pero algunas tienen menos (p. ej., regiones muy rurales sin tuits en ciertos días). Esto implica que esos “vacíos” deben tratarse:
  - ¿Se imputan con `NaN` o con valor neutro de `SCORE`?
  - ¿Se descartan esos días para análisis longitudinal?
  - Para el objetivo de visualizar tendencias, conviene imputar o marcar esos días como “sin datos”.

Conclusiones:

- **Registro del dataset:** Índices diarios de sentimiento positivo (`SCORE`) y cantidad de tuits (`N`) por provincia/región para cada día de 2022.
- **Calidad y completitud:** No hay valores nulos, pero sí faltan registros diarios para algunas regiones (hay que imputar o marcar esos días).
- **Distribuciones y outliers**: `SCORE` se distribuye entre ~0.01 y ~0.99; `N` es muy asimétrica. las provincias con muy pocos tuits generan un SCORE ruidoso.