EXAMEN PANDAS 23-01-2025

El dataset se ha obtenido de: https://www.kaggle.com/datasets/lumierebatalong/covid-19-variants-survival-data

Todas las preguntas tendrán los comentarios necesarios para saber qué hace cada línea de código y cómo lo hace.

Exploración y limpieza de datos

1. Muestra las primeras 15 filas del dataset para obtener una idea de su estructura.
2. Identifica cuáles son las columnas categóricas y cuáles son numéricas.
3. Verifica si hay valores nulos en el dataset y calcula su proporción por columna. El número debe aparecer con dos decimales y terminado en '%'.

Transformaciones y manipulaciones

4. Convierte las columnas `first_seq`, `last_seq`, y `censure_date` al tipo de dato `datetime`.
5. Extrae el mes y el año de la columna `first_seq` y guárdalos en dos nuevas columnas llamadas `first_seq_month` y `first_seq_year`.
6. Crea una nueva columna `fatality_rate` calculada como: `fatality_rate = total_deaths / total_cases`.
7. Filtra las filas donde `variant` sea 'S.Q677' y crea un nuevo `DataFrame` con esos datos.
8. Crea una nueva columna `active_cases` calculada como: `active_cases = total_cases - total_deaths`.

Manejo de datos faltantes

9. Identifica cuántos valores faltantes tiene la columna `growth_rate` y su proporción en el dataset.
10. Rellena los valores nulos de la columna `growth_rate` con la media de la misma.
11. Rellena los valores nulos de `growth_rate` utilizando la interpolación lineal.

Visualización y análisis exploratorio

12. Crea un histograma para analizar la distribución de la columna `mortality_rate`.
13. Genera un gráfico de barras para mostrar el promedio de `growth_rate` por variante (`variant`).
14. Crea un gráfico de dispersión para analizar la relación entre `total_cases` y `total_deaths`, diferenciando los puntos por país (`Country`).
15. Genera un boxplot para analizar la variabilidad de `duration` por variante (`variant`).

Análisis de correlaciones

16. Calcula la matriz de correlación para las columnas numéricas del dataset.
17. Representa un mapa de calor (heatmap) para visualizar las correlaciones entre las variables numéricas.

18. Identifica la variable que tiene mayor correlación con `growth_rate` y analiza su relación mediante un gráfico de dispersión.

Análisis estadístico

19. Agrupa el dataset por `Country` y calcula: el promedio de `duration` y el total de casos (`total_cases`) acumulados por país.
20. ¿Cuál es el país con la mayor tasa de mortalidad (`mortality_rate`) promedio?
21. Analiza la evolución del número de casos (`total_cases`) a lo largo del tiempo (`first_seq`) para los 5 países con más registros.

Expresiones regulares

22. Verifica si alguna de las entradas en `variant` contiene números. Si es así, extrae esos números en una nueva columna.
23. Comprueba si las fechas de `first_seq` contienen el formato esperado (YYYY-MM-DD) utilizando expresiones regulares.

Expresiones regulares

24. Evalúa el uso de memoria del dataset antes y después de optimizar todas las columnas categóricas y numéricas.
25. Genera un resumen con los cinco países con mayor tasa de mortalidad y genera un segundo resumen con los cinco países con mayor número de casos activos.