

Desafío - Preparación de datos y gráficos

En este desafío validaremos nuestros conocimientos de preparación de datos y gráficos. Para lograrlo, necesitarás aplicar los contenidos vistos en clases y en la guía de estudios.

Lee todo el documento antes de comenzar el desarrollo **individual**, para asegurarte de tener el máximo de puntaje y enfocar bien los esfuerzos.

Tiempo asociado: 2 horas cronológicas

Descripción

La base de datos **world-data-2023.csv** proporciona una gran cantidad de información sobre todos los países del mundo, abarcando una amplia gama de indicadores y atributos. Incluye estadísticas demográficas, indicadores económicos, factores ambientales, métricas de atención médica, estadísticas educativas y mucho más. Con la representación de cada país, este conjunto de datos ofrece una perspectiva global completa sobre diversos aspectos de las naciones, lo que permite análisis en profundidad y comparaciones entre países. Las variables consideradas son:

- **0. Country:** Nombre del país.
- **1. Density (P/Km2):** Densidad de población medida en personas por kilómetro cuadrado.
- **2. Abbreviation:** Abreviatura o código que representa el país.
- **3. Agricultural Land (%):** Porcentaje del área de tierra utilizada para fines agrícolas.
- **4. Land Area (Km2):** Área total de tierra del país en kilómetros cuadrados.
- **5. Armed Forces Size:** Tamaño de las fuerzas armadas en el país.
- **6. Birth Rate:** Número de nacimientos por 1,000 habitantes por año.
- **7. Calling Code:** Código de llamada internacional para el país.
- **8. Capital/Major City:** Nombre de la capital o ciudad principal.
- **9. CO2 Emissions:** Emisiones de dióxido de carbono en toneladas.
- **10. CPI:** Índice de Precios al Consumidor, una medida de la inflación y el poder adquisitivo.
- **11. CPI Change (%):** Cambio porcentual en el Índice de Precios al Consumidor en comparación con el año anterior.
- **12. Currency_Code:** Código de moneda utilizado en el país.
- **13. Fertility Rate:** Número promedio de hijos nacidos de una mujer durante su vida.
- **14. Forested Area (%):** Porcentaje del área de tierra cubierta por bosques.
- **15. Gasoline_Price:** Precio de la gasolina por litro en moneda local.
- **16. GDP:** Producto Interno Bruto, el valor total de bienes y servicios producidos en el país.

- **17. Gross Primary Education Enrollment (%)**: Tasa de inscripción bruta en educación primaria.
- **18. Gross Tertiary Education Enrollment (%)**: Tasa de inscripción bruta en educación terciaria.
- **19. Infant Mortality**: Número de muertes por cada 1,000 nacidos vivos antes de cumplir un año de edad.
- **20. Largest City**: Nombre de la ciudad más grande del país.
- **21. Life Expectancy**: Número promedio de años que se espera que viva un recién nacido.
- **22. Maternal Mortality Ratio**: Número de muertes maternas por cada 100,000 nacidos vivos.
- **23. Minimum Wage**: Nivel de salario mínimo en moneda local.
- **24. Official Language**: Idioma(s) oficial(es) hablado(s) en el país.
- **25. Out of Pocket Health Expenditure (%)**: Porcentaje del gasto total en salud pagado directamente por los individuos.
- **26. Physicians per Thousand**: Número de médicos por cada mil personas.
- **27. Population**: Población total del país.
- **28. Population: Labor Force Participation (%)**: Porcentaje de la población que forma parte de la fuerza laboral.
- **29. Tax Revenue (%)**: Ingresos fiscales como porcentaje del PIB.
- **30. Total Tax Rate**: Carga tributaria total como porcentaje de las ganancias comerciales.
- **31. Unemployment Rate**: Porcentaje de la fuerza laboral que está desempleada.
- **32. Urban Population**: Porcentaje de la población que vive en áreas urbanas.
- **33. Latitude**: Coordenada de latitud de la ubicación del país.
- **34. Longitude**: Coordenada de longitud de la ubicación del país.

Vamos a realizar algunos análisis comparativos entre las variables. Para ello:

1. Carga los datos y genera un dataframe que excluya las columnas correspondientes a la abreviatura del país, código de llamada, capital, código de moneda, ciudad más grande, lenguaje oficial, latitud y longitud. Considerando estos datos, realiza una inspección inicial sobre ellos, considerando gráficos. Concluye.
2. De las variables codificadas numéricamente, ¿hay pares de ellas en la que parezca haber correlación? ¿Para cuál(es) de ellas se observa más claramente?. Elige 2 pares de variables.
3. Considera los dos pares de variables anteriores y elimina en cada caso los faltantes y atípicos. Justifica en cada caso tu procedimiento.
4. Construye un gráfico para verificar la correlación entre estas variables, ahora con los datos preparados. Compara con lo obtenido anteriormente.
5. Realiza lo mismo para comparar la cantidad de médicos por cada mil personas y la esperanza de vida. ¿Qué puedes concluir?
6. Construye un gráfico para representar la población total de los países y su producto interno bruto. Prepara los datos y escoge el gráfico adecuado (si es necesario, limpia y/o transforma los datos). Justifica.

7. Escoge otro par de datos, límpialos, prepáralos y analiza su posible relación. Justifica utilizando las herramientas vistas y gráficos adecuados.

Requerimientos

1. Explora la estructura de datos y realiza limpieza corrigiendo datos nulos o faltantes **(3 Puntos)**
2. Selecciona y transforma datos para analizarlos y/o graficarlos **(3 Puntos)**
3. Construye gráficos diversos para comparar datos, e interpreta los gráficos **(4 Puntos)**



¡Mucho éxito!

Consideraciones y recomendaciones

Aquí tienes algunas consideraciones importantes para realizar una limpieza de datos efectiva:

- **Identificar valores faltantes:** Revisa si existen valores faltantes en los datos y decide cómo tratarlos. Puedes eliminar las filas o columnas con valores faltantes si no son significativos, o completarlos
- **Eliminar outliers:** Identifica valores atípicos o extremos que puedan distorsionar el análisis y decide si eliminarlos o tratarlos de manera adecuada según el contexto del análisis. (puedes usar un box plot para ello)
- **Estandarizar formatos:** Asegúrate de que los datos estén en formatos coherentes y estandarizados para facilitar su análisis. Por ejemplo, convierte texto a números para poder realizar cálculos y gráficos.

Debes entregar tu trabajo en un archivo de Jupyter Notebook, con el desarrollo de tu trabajo y las explicaciones necesarias de tu procedimiento.