**Práctica 2: Limpieza y análisis de datos**

**Puntos para desarrollar**

**1. Descripción del dataset.** ¿Por qué es importante y qué pregunta/problema pretende responder?

Inicialmente se escoge este dataset para ver si se puede estudiar qué salario puede alcanzar una persona en base a su pertenencia a un grupo demográfico, con las características descritas en el conjunto de datos.

Las personas que no alcanzan un determinado umbral de sueldo tienen una peor calidad de vida, se encuentran con más dificultad para hacer frente a cualquier eventualidad negativa y esto provoca otras consecuencias. Determinados grupos demográficos pueden convertirse en colectivos más vulnerables de la sociedad. Factores como la raza, el sexo, el nivel de estudios… etc pueden verse con un techo salarial y verse más afectados frente a las crisis económicas.

\*\*\*\*\* preguntas a responder

**2. Integración y selección.** Integración y selección de los datos de interés a analizar.

Inicialmente, el dataset contiene las siguientes columnas

|  |  |  |  |
| --- | --- | --- | --- |
| Campo | Tipo | Descripción | Ejemplo |
| Age | Numérico | Edad de la persona | 39 |
| workclass | Texto | Clase de empleado. Campo categorizado. |  |
| fnlwgt | Numérico | Peso final dado por la oficina que ha recogido los datos y que da el número de unidades en la población objetivo en base a una fórmula no facilitada- |  |
| education | Texto | Nivel de estudios. Campo categorizado |  |
| education-num | Texto | Años dedicados a los estudios |  |
| marital-status | Texto | Situación sentimental. Campo categorizado |  |
| occupation | Texto | Trabajo actual |  |
| relationship | Texto | Relación familiar actual. Campo categorizado |  |
| race | Texto | Raza. Campo categorizado |  |
| sex | Texto | Sexo. Campo categorizado |  |
| capital-gain | Numérico | Ganancia de capital |  |
| capital-loss | Numérico | Pérdida de capital |  |
| hours-per-week | Numérico | Horas de trabajo por semana |  |
| native-country | Texto | Nacionalidad. Campo categorizado |  |
| salary | Lógico | Salario |  |

**3. Limpieza de los datos.**.

**3.1** ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Inicialmente se plantea

Sin embargo, tras la obt\*\*\* ejemplo,

**3.2.** Identificación y tratamiento de valores extremos.

Inicialmente se plantea

Sin embargo, tras la obt\*\*\* ejemplo,

**4. Análisis de los datos.**

**4.1.** Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Inicialmente se plantea

Sin embargo, tras la obt\*\*\* ejemplo,

**4.2.** Comprobación de la normalidad y homogeneidad de la varianza. Tipología y ciclo de vida de los datos Práctica 2

Inicialmente se plantea

Sin embargo, tras la obt\*\*\* ejemplo,

**4.3.** Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Inicialmente se plantea

Sin embargo, tras la obt\*\*\* ejemplo,

**5. Representación de los resultados a partir de tablas y gráficas.**

Inicialmente se plantea

Sin embargo, tras la obt\*\*\* ejemplo,

**6. Resolución del problema**. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Inicialmente se plantea

Sin embargo, tras la obt\*\*\* ejemplo,

**7. Código**: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Inicialmente se plantea

Sin embargo, tras la obt\*\*\* ejemplo,

-------------

**10. Dataset.** Publicar el dataset obtenido (\*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

El DOI correspondiente a la base de datos es:

<https://doi.org/10.5281/zenodo.5651108>

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 9. Detalle de la publicación del dataset obtenido

Tabla de contribuciones al trabajo

|  |  |
| --- | --- |
| Contribuciones | Firma |
| Investigación previa | M.P.G.R, A.C.A |
| Redacción de las respuestas | M.P.G.R, A.C.A |
| Desarrollo código | M.P.G.R, A.C.A |