Taller 1 Corte 2
Limpieza de datos con R

Hacer un análisis completo de la data suministrada (observatorio de datos para el seguimiento del objetivo de educación de calidad, caso de estudio: características de los establecimientos educativos y su influencia en los resultados de la prueba saber 11. Año 2012) SB11-20121-RGSTRO-CLFCCN-V1-0-txt-csv.csv.

Tome como referencia y apoyo el libro Ciencia de datos técnicas analíticas y aprendizaje estadístico en un enfoque práctico, disponible en la biblioteca

1. Debe realizar todo el proceso necesario de limpieza y preparación de los datos propuestos. Esta etapa incluiría la identificación de si el conjunto de datos contiene:

   Variables que no aportan información, Variables redundantes, Registros inconsistentes e incompletos (imputación), valores errores estructurales (tipográficos), entre otros.

   - Calcular y registrar la edad de cada estudiante en una nueva columna.
   - Categorizar la edad en: “joven, adulto y adulto mayor”.
   - Cambiar los tipos de documento por los nombres de cada tipo ej.: C por Cédula
   - Separe matemáticamente (sin utilizar formulas), el código DIVIPOLA de la columna ESTU_COD_RESIDE_MCPIO en código del departamento y el código del municipio en las columnas “IDDPTO” y “IDMUNICIPIO” respectivamente.

2. Quienes se destacaron más en matemáticas, si las mujeres o los hombres, teniendo en cuenta:

   - La ciudad
   - Edad de acuerdo con el tipo de documento de identidad
   - Tipo de colegio (Oficial, Privado) y caracterización del colegio (ACADEMICO, TECNICO, etc.)
   - Qué nivel de ingles
   - Nacionalidad

3. Generar las estadísticas de resumen de las siguientes columnas:

Edad_Estudiantes, PUNT_LENGUAJE, PUNT_MATEMATICAS, PUNT_C_SOCIALES, PUNT_FILOSOFIA, PUNT_BIOLOGIA, PUNT_QUIMICA, PUNT_FISICA, PUNT_INGLES

4. ¿Cuáles son las ciudades en donde las familias tienen mayores ingresos?
5. Realizar un gráfico que permita identificar y cuantificar las familias que tienen celular, internet, computador, servicio de televisión y teléfono fijo, teniendo en cuenta las ciudades principales.

Desarrollar un análisis descriptivo e interpretativo de la gráfica realizada.
De acuerdo con los resultados obtenidos, ¿Qué influencia tiene en los estudiantes los servicios representados en la gráfica, para ocupar los primeros puestos en las pruebas saber 11?

6. ¿Influye la edad en los puntajes obtenidos?, argumente su respuesta y represéntela de la forma más conveniente.

Instrucciones para la elaboración del informe

El objetivo es identificar tendencias y patrones que puedan informar decisiones estratégicas. Utilice la base de datos proporcionada para realizar esta actividad.

El informe debe estar redactado en formato Markdown y organizado en un único documento con las siguientes secciones numeradas y con encabezados descriptivos:

Elementos del Informe

1. **Introducción:**

- Proporcione un contexto y antecedentes del caso de estudio
- Describa brevemente el propósito del informe.
- Máximo 300 palabras.

2. **Análisis Descriptivo**:

- **Tratamiento de Datos:** Explique cómo se abordaron los errores, datos atípicos y datos faltantes.
- **Análisis Gráfico e Indicadores Relevantes**: Incluya gráficos e indicadores que expliquen información relevante para el caso de estudio.
- Máximo 700 palabras.

3. Discusión y Conclusiones:

- Resuma los principales hallazgos y proporcione recomendaciones basadas en el análisis.
- Identifique oportunidades de mejora.
- Máximo 400 palabras.

4. Anexos:

- Incluya gráficos y tablas del preprocesamiento de datos (errores, datos faltantes y atípicos).
- Presentación de resultados y gráficos que respalden el análisis descriptivo.
- Cada gráfico o tabla debe tener una descripción clara.
- Máximo 2 páginas.

5. Referencias:

- Liste todas las fuentes bibliográficas utilizadas en el informe.

Requisitos del Informe

- **Escritura**: El informe debe estar redactado en un lenguaje claro y conciso.
- **Formato**: Use el formato IEEE.
- **Gráficos**: Los gráficos deben medir aproximadamente 4 cm por 4 cm. Los nombres de los ejes deben estar en español. La información debe ubicarse justo después de la numeración de la figura. El tamaño y grosor de las líneas deben ser suficientes para su distinción.
- **Tablas**: Solo incluya información que se interprete en el documento principal.
- **Números**: Sea consistente con la notación de decimales y puntos de mil. Utilice el mismo número de cifras decimales en todo el documento.
- **Interpretaciones y Análisis:** Realice afirmaciones precisas y acompáñelas con cantidades exactas. En el informe descriptivo, interprete en términos del contexto del problema. En los anexos, sea lo más técnico posible e incluya la clasificación de las variables de acuerdo con su naturaleza.

Entrega y Evaluación

- La actividad debe ser entregada en el aula virtual antes de la fecha límite.

- La evaluación se basará en la claridad del análisis, la precisión de las interpretaciones, la calidad de la presentación del informe, el formato utilizado para su entrega y el cumplimiento de los requerimientos del informe.
- **Generación del Archivo Rmd:** Se debe crear un archivo R Markdown (.Rmd) que contenga el contenido requerido para la actividad. Renderización a Formato PDF: Utilizando las herramientas adecuadas, como RStudio, el archivo .Rmd debe ser renderizado en un documento PDF. Asegúrese de que el PDF resultante mantenga la integridad del contenido y la presentación adecuada de todos los elementos incluidos. Incluya el zip con el proyecto completo.