PROYECTO DE ANÁLISIS DE DATOS

ENTREGA PARCIAL: ANÁLISIS EXPLORATORIO DE DATOS (EDA)

---

1. Introducción y Objetivos

Esta práctica constituye vuestro primer contacto real con un análisis exploratorio de datos completo y sistemático. El objetivo es aplicar técnicas de estadística descriptiva y visualización para comprender en profundidad la estructura, patrones y características de un conjunto de datos real.

El Análisis Exploratorio de Datos (EDA) es una fase fundamental en cualquier proyecto de ciencia de datos. Es el proceso mediante el cual "conversamos" con los datos, descubrimos sus características, identificamos patrones, detectamos anomalías y generamos hipótesis que posteriormente podrán ser validadas con técnicas de inferencia estadística.

Al finalizar esta práctica, deberéis ser capaces de:

- Explorar, limpiar y describir un conjunto de datos de forma sistemática y rigurosa
- Identificar y tratar valores atípicos, datos faltantes y posibles errores en los datos
- Crear visualizaciones efectivas que comuniquen los hallazgos de forma clara
- Analizar relaciones entre variables utilizando técnicas bivariantes y multivariantes
- Extraer insights relevantes y formular preguntas de investigación fundamentadas
- Comunicar todo el proceso de forma clara, estructurada y reproducible utilizando R/Python y Quarto

---

2. Descripción del Trabajo

Cada grupo trabajará con un conjunto de datos asignado por el profesor. A partir de él, deberéis seguir los siguientes pasos en vuestro informe:

Paso 1: Presentación y Descripción del Dataset

Esta sección es IMPRESCINDIBLE. No cumplir este requisito implica que la práctica no podrá considerarse aprobada en ningún caso.

Debéis incluir:

- Contexto y origen de los datos: ¿De dónde provienen? ¿Qué fenómeno o problema buscan describir?
- Descripción de las variables: Crear una tabla detallada con todas las variables del dataset, indicando:

  - Nombre de la variable
  - Tipo (numérica continua, numérica discreta, categórica ordinal, categórica nominal, fecha, etc.)
  - Descripción del significado
  - Unidades de medida (si aplica)
  - Rango de valores posibles o categorías
- Dimensiones del dataset: Número de observaciones y variables

Paso 2: Análisis Univariante

Realizad un análisis descriptivo completo de todas las variables del dataset:

Para variables numéricas:

- Estadísticos de centralización: media, mediana, moda
- Estadísticos de dispersión: desviación típica, rango, rango intercuartílico
- Estadísticos de forma: asimetría, curtosis
- Cuartiles y percentiles relevantes
- Visualizaciones: histogramas, diagramas de caja (boxplots), gráficos de densidad, gráficos Q-Q

Para variables categóricas:

- Tablas de frecuencia (absolutas y relativas)
- Moda y distribución de categorías
- Visualizaciones: gráficos de barras, gráficos de sectores (solo si son pocas categorías)

Durante esta fase es fundamental asegurar la calidad de los datos:

- Valores faltantes (missing values):
  - Identificad qué variables contienen valores faltantes y en qué proporción
  - Analizad el patrón de los valores faltantes (¿son aleatorios o hay un patrón?)
  - Decidid cómo tratarlos (imputación, eliminación, etc.) y justificad vuestra decisión
- Inconsistencias y errores:
  - Comprobad que los valores de las variables sean consistentes con su definición
  - Verificad posibles duplicados en los datos
  - Corregid formatos inconsistentes si es necesario

Paso 3: Análisis Bivariante y Multivariante

Explorad las relaciones entre las variables que consideréis más relevantes para comprender los datos:

Relaciones entre dos variables numéricas:

- Diagramas de dispersión (scatter plots)
- Coeficiente de correlación lineal (Pearson o Spearman según corresponda)
- Matrices de correlación con visualización (heatmap)

Relaciones entre una variable numérica y una categórica:

- Diagramas de caja por grupos
- Gráficos de violín
- Comparación de estadísticos descriptivos por grupos

Relaciones entre dos variables categóricas:

- Tablas de contingencia
- Gráficos de barras agrupadas o apiladas
- Mosaicos o heatmaps de frecuencias

Análisis multivariante (opcional pero recomendado):

- Gráficos de pares (pair plots)
- Análisis de segmentación por múltiples variables
- Visualizaciones interactivas o facetadas

Paso 4: Identificación de Patrones y Formulación de Preguntas

Basándoos en los hallazgos del análisis exploratorio, debéis:

- Identificar y describir los patrones principales encontrados en los datos
- Formular entre 3 y 5 preguntas de investigación relevantes y bien fundamentadas que hayan surgido del análisis exploratorio
    - Estas preguntas deben ser interesantes y estar respaldadas por evidencia observada en el EDA
    - Ejemplo: "¿Existe una relación entre el nivel educativo y el salario en este dataset?" o "¿Los clientes del grupo A tienen un gasto medio mayor que los del grupo B?"
    - Nota importante: En esta práctica NO debéis responder estas preguntas mediante inferencia estadística. El objetivo es plantearlas de forma fundamentada para que puedan ser respondidas posteriormente con las técnicas que veréis en clase.

Paso 5: Conclusiones del Análisis Exploratorio

Escribid una sección final donde:

- Resumáis los hallazgos principales del análisis
- Describáis las características más relevantes del dataset
- Expliquéis qué patrones, tendencias o anomalías habéis identificado
- Reflexionéis sobre las limitaciones del análisis realizado
- Comentéis qué preguntas quedan abiertas para análisis futuros (modelización)
- Mencionéis posibles sesgos en los datos o aspectos que requieran más investigación

---

3. Formato y Requisitos de Entrega

- El trabajo se realizará en grupos de 2-3 personas
- El informe debe ser un documento reproducible creado en Quarto (.qmd)
- El código utilizado para el análisis debe ser visible en el informe (echo: true)
- Se valorará positivamente:
    - Código limpio, comentado y bien estructurado
    - Uso de librerías modernas de visualización (ggplot2, plotly, etc.)
    - Visualizaciones claras y estéticamente cuidadas
    - Narrativa fluida que conecte código, visualizaciones e interpretaciones
- La entrega consistirá en un único archivo, que será el documento .html o .pdf generado por Quarto
- Aseguraos de que los nombres de todos los integrantes del grupo aparecen claramente al inicio del documento
- El documento debe incluir un índice automático para facilitar la navegación

---

4. Evaluación Detallada

La evaluación valorará tanto el rigor técnico en la aplicación de las técnicas de análisis como la claridad en la comunicación e interpretación de los resultados. Se tendrán en cuenta las siguientes dimensiones:

1. Análisis univariante y calidad del dato
2. Análisis multivariante

3. Identificación de patrones y formulación de preguntas
4. Informe, comunicación y conclusiones

Rúbrica de Calificación

Dimensión	Insuficiente (0-4)	Aprobado (5-6)	Notable (7-8)	Sobresaliente (9-10)
Análisis Univariante y Calidad de Datos	Se cometen errores técnicos graves (ej. barplot de variable continua, interpretación incorrecta de estadísticos). No se identifican o tratan valores faltantes/atípicos.	Se realiza un análisis univariante correcto de la mayoría de variables con visualizaciones adecuadas. Se identifican problemas básicos de calidad (missings, outliers) aunque el tratamiento es superficial.	Análisis univariante completo y técnicamente correcto para todas las variables. Visualizaciones apropiadas para cada tipo de variable. Se identifican y tratan adecuadamente problemas de calidad con justificación clara.	Análisis univariante exhaustivo y riguroso. Dominio completo de técnicas descriptivas y visualizaciones. Tratamiento ejemplar de problemas de calidad con justificaciones sólidas y documentadas. Demuestra profundidad en la interpretación de los estadísticos.
2. Análisis Multivariante	No se exploran relaciones entre variables o el análisis es muy limitado. Las visualizaciones son incorrectas o poco informativas.	Se exploran algunas relaciones básicas entre variables. Las visualizaciones son correctas aunque poco creativas. Se mencionan correlaciones o diferencias simples.	Se exploran múltiples relaciones relevantes con visualizaciones variadas y adecuadas. Se van más allá de las correlaciones básicas. Buen equilibrio entre análisis técnico y creatividad exploratoria.	Análisis multivariante excelente y creativo. Se descubren relaciones complejas e interesantes. Visualizaciones sofisticadas y muy informativas. Se utilizan técnicas avanzadas cuando es apropiado. Demuestra

Dimensión	Insuficiente (0-4)	Aprobado (5-6)	Notable (7-8)	Sobresaliente (9-10)
				curiosidad analítica y pensamiento crítico.
3. Identificación de Patrones y Preguntas	No se identifican patrones claros o las preguntas planteadas no están fundamentadas en el análisis. Las preguntas son triviales o imposibles de responder.	Se identifican algunos patrones básicos. Se plantean 3 preguntas relevantes fundamentadas en el EDA que podrían responderse con inferencia.	Se identifican claramente los patrones principales del dataset. Se plantean 3-5 preguntas interesantes y bien fundamentadas que surgen naturalmente del análisis exploratorio. Las preguntas son específicas y relevantes.	Se identifican patrones sutiles y no obvios. Las preguntas planteadas son excelentes, originales y demuestran una comprensión profunda de los datos. Se vinculan los hallazgos del EDA con las preguntas de forma coherente y bien argumentada.
4. Informe, Comunicación y Conclusiones	El informe es desorganizado o difícil de seguir. Las conclusiones son vagas o no se basan en el análisis. El código es caótico, no funciona, o no es reproducible.	El informe está estructurado y es legible. Las conclusiones resumen los hallazgos básicos. El código funciona y es reproducible aunque podría ser más limpio. La narrativa conecta los análisis de forma básica.	El informe es claro, bien estructurado y reproducible. Las conclusiones son coherentes, contextualizadas e incluyen reflexión sobre limitaciones. El código está bien organizado y comentado. La narrativa conecta fluidamente análisis e interpretación.	El informe es excelente: claro, profesional y muy bien narrado. Las conclusiones son profundas, incluyen reflexión crítica sobre limitaciones, sesgos y análisis futuros. El código es ejemplar y las visualizaciones son de calidad profesional. Se nota un trabajo

Dimensión	Insuficiente (0-4)	Aprobado (5-6)	Notable (7-8)	Sobresaliente (9-10)
				cuidado y pulido.

Notas Importantes

- Un trabajo que no incluya la descripción completa del dataset (Paso 1) no podrá superar el Aprobado (5 sobre 10), independientemente de la calidad de las otras secciones.
- Se valorará especialmente:
    - La capacidad de contar una historia coherente con los datos
    - La calidad y claridad de las visualizaciones
    - La profundidad del análisis más allá de lo básico
    - La honestidad al reconocer limitaciones
    - El pensamiento crítico en las interpretaciones
- Se penalizará:
    - Código que no sea reproducible o genere errores
    - Visualizaciones poco claras o mal etiquetadas
    - Interpretaciones incorrectas de los resultados
    - Plagio de código o texto de otros grupos o fuentes sin citar

---

5. Consejos y Buenas Prácticas

- Comenzad por explorar los datos sin prejuicios: Dejad que los datos "hablen" antes de buscar patrones específicos
- No os limitéis a hacer gráficos: Interpretad lo que cada visualización os está mostrando
- Usad múltiples tipos de gráficos: Diferentes visualizaciones revelan diferentes aspectos de los datos
- Cuidado con las correlaciones: Correlación no implica causalidad
- Documentad vuestro trabajo: Explicad por qué tomáis cada decisión en la limpieza y análisis
- Revisad entre compañeros: Una segunda opinión puede detectar errores o sugerir mejoras

- Consultad dudas con el profesor: Es mejor preguntar durante el proceso que descubrir errores al final

---

¡Buena suerte con vuestro análisis exploratorio! 🔍 📊