Jose-Julio Ramírez y Sánchez-Escobar
Móvil: 640583015 WhatsApp: 636110348 Telegram: 640583015
fantvamp@yahoo.es Maestro Lassalle 11, 28016, Madrid. DNI: 51063210J
vscode01.jj@gmail.com vscode01.jj@hotmail.com jjramirez@isciii.es

Tarea 1 - c:\Users\Coder\Proyectos\EDA\Docs\EDA_Tarea.20260203090648M.png 90.113 2026-02-03 10.22
-a--

Tarea: Investigación y Desarrollo de un Análisis Exploratorio de Datos (EDA)

Los coders deben investigar cómo se realiza un proceso de **EDA (Exploratory Data Analysis)** aplicando Python y librerías comunes (**pandas, matplotlib, seaborn, etc.**).

Parte 1: Análisis Exploratorio de Datos

Cada estudiante debe investigar y responder las siguientes preguntas:

1. ¿Qué es el EDA y cuál es su propósito en el análisis de datos?
   Añade un ejemplo práctico de un caso en el que el EDA sea fundamental antes de aplicar modelos de Machine Learning.
2. ¿Qué tipos de datos existen (categóricos, numéricos, ordinales, etc.) y cómo se tratan en un EDA?
3. ¿Qué técnicas básicas se usan en un EDA?
   (Ejemplo: estadísticas descriptivas, histogramas, boxplots, correlaciones, detección de valores atípicos).
4. ¿Cuál es la diferencia entre análisis univariado, bivariado y multivariado?
5. ¿Qué es la limpieza de datos y qué tareas suelen incluirse en este paso?
   (Ejemplo: manejo de valores nulos, duplicados, outliers, normalización de datos).
6. ¿Qué papel juegan las librerías pandas, matplotlib y seaborn en un EDA?
7. Explica brevemente el flujo típico de un EDA desde la carga del dataset hasta las conclusiones.
8. ¿Qué es una matriz de correlación y para qué sirve en el EDA?
9. ¿Qué son los outliers y qué métodos existen para detectarlos y tratarlos en un análisis exploratorio?
10. ¿Qué es hipótesis testing y para qué sirve en el EDA?

Entrega

1. Subir un PDF al classroom con tus respuestas.

Jose-Julio Ramírez y Sánchez-Escobar
Móvil: 640583015 WhatsApp: 636110348 Telegram: 640583015
fantvamp@yahoo.es Maestro Lassalle 11, 28016, Madrid. DNI: 51063210J
vscode01.jj@gmail.com vscode01.jj@hotmail.com jjramirez@isciii.es

¡Hola! Veo que tienes una tarea de investigación sobre **Análisis Exploratorio de Datos (EDA)**. Es un tema fundamental en la ciencia de datos.

A continuación, he preparado un borrador estructurado con las respuestas a las 10 preguntas que plantea tu documento, diseñado para que puedas utilizarlo como base para tu entrega en PDF.

Parte 1: Respuestas a la Investigación sobre EDA

1. ¿Qué es el EDA y cuál es su propósito?

El **Exploratory Data Analysis (EDA)** es una fase crítica en el análisis de datos que consiste en "interrogar" al conjunto de datos antes de hacer suposiciones o aplicar modelos complejos. Su objetivo es entender la estructura de los datos, detectar anomalías, probar hipótesis iniciales y verificar su calidad.

- **Propósito:** Asegurar que los datos estén limpios y sean comprensibles para elegir las técnicas de modelado adecuadas.
- **Ejemplo Práctico:** Imagina que quieres predecir el precio de casas usando Machine Learning. Sin un EDA, podrías no notar que la columna "Tamaño del terreno" tiene valores de "0" (lo cual es imposible) o que está en diferentes unidades (metros vs. pies). Si entrenas el modelo con estos errores, las predicciones serán basura (*Garbage In, Garbage Out*).

2. Tipos de datos y su tratamiento en EDA

Los datos se clasifican generalmente en dos grandes grupos, y cada uno requiere gráficos y estadísticas diferentes:

- **Numéricos (Cuantitativos):**
  - *Continuos:* Pueden tomar cualquier valor (ej. altura, temperatura). Se analizan con histogramas y medias.
  - *Discretos:* Valores enteros contables (ej. número de hijos).
- **Categóricos (Cualitativos):**
  - *Nominales:* Sin orden (ej. color de ojos, país). Se analizan con gráficos de barras y modas.
  - *Ordinales:* Con orden lógico (ej. nivel educativo: bajo, medio, alto).

3. Técnicas básicas usadas en un EDA

Las técnicas se dividen en gráficas y no gráficas:

- **Estadísticas descriptivas:** Uso de .describe() en pandas para ver la media, desviación estándar, mínimos y máximos.
- **Visualización de distribución:** Histogramas y gráficos de densidad.
- **Análisis de relaciones:** Diagramas de dispersión (scatter plots) y mapas de calor.
- **Detección de valores atípicos:** Boxplots (diagramas de caja).

Jose-Julio Ramírez y Sánchez-Escobar
Móvil: 640583015 WhatsApp: 636110348 Telegram: 640583015
fantvamp@yahoo.es Maestro Lassalle 11, 28016, Madrid. DNI: 51063210J
vscode01.jj@gmail.com vscode01.jj@hotmail.com jjramirez@isciii.es

4. Diferencia entre análisis univariado, bivariado y multivariado

La diferencia radica en la cantidad de variables que se analizan simultáneamente:

- **Univariado:** Analiza una sola variable a la vez. *Objetivo:* Entender la distribución central y dispersión (ej. histograma de edades).
- **Bivariado:** Analiza la relación entre dos variables. *Objetivo:* Encontrar correlaciones (ej. scatterplot de Edad vs. Salario).
- **Multivariado:** Analiza tres o más variables. *Objetivo:* Entender interacciones complejas (ej. pairplot o gráficos 3D).

5. ¿Qué es la limpieza de datos?

Es el proceso de detectar y corregir (o eliminar) registros corruptos o inexactos. Las tareas comunes incluyen:

- **Imputación:** Rellenar valores nulos (media, mediana o valor fijo).
- **Eliminación de duplicados:** Borrar filas repetidas.
- **Manejo de Outliers:** Decidir si eliminar o ajustar valores extremos.
- **Normalización/Estandarización:** Poner diferentes variables en la misma escala (ej. 0 a 1).

6. Papel de las librerías en Python

- **Pandas:** Es la herramienta principal para la manipulación de datos (DataFrames). Se usa para cargar, filtrar, limpiar y transformar los datos.
- **Matplotlib:** Es la base de la visualización en Python. Permite crear gráficos estáticos, pero requiere mucho código para personalizarlos.
- **Seaborn:** Construida sobre Matplotlib, ofrece una interfaz de alto nivel para dibujar gráficos estadísticos atractivos e informativos con menos código.

7. Flujo típico de un EDA

1. **Carga de datos:** Importar el dataset (ej. `pd.read_csv`).
2. **Inspección inicial:** Ver las primeras filas (`.head()`), tipos de datos (`.info()`) y dimensiones (`.shape`).
3. **Limpieza:** Tratar nulos, duplicados y errores de tipo.
4. **Análisis Univariado:** Ver la distribución de cada variable individualmente.
5. **Análisis Bivariado/Multivariado:** Buscar relaciones entre variables.
6. **Conclusiones:** Resumir hallazgos para la siguiente fase (modelado).

8. Matriz de correlación

Es una tabla que muestra los coeficientes de correlación entre variables numéricas.

Jose-Julio Ramírez y Sánchez-Escobar
Móvil: 640583015 WhatsApp: 636110348 Telegram: 640583015
fantvamp@yahoo.es Maestro Lassalle 11, 28016, Madrid. DNI: 51063210J
vscode01.jj@gmail.com vscode01.jj@hotmail.com jjramirez@isciii.es

- **Uso:** Sirve para identificar qué variables están fuertemente relacionadas (positiva o negativamente).
- **Importancia:** Ayuda a evitar la *multicolinealidad* (variables redundantes) en modelos de regresión. Se visualiza comúnmente con un Heatmap.

9. Outliers: Detección y tratamiento

Los **outliers** (valores atípicos) son datos que se desvían drásticamente del resto de la muestra.

- **Detección:**
  - *Rango Intercuartílico (IQR):* Datos por debajo de $Q1 - 1.5 \times IQR$ o por encima de $Q3 + 1.5 \times IQR$.
  - *Z-Score:* Datos que están a más de 3 desviaciones estándar de la media.
- **Tratamiento:** Se pueden eliminar (si son errores), transformar (logaritmo) o capar (asignarles un valor máximo/mínimo permitido).

10. Hipótesis testing en el EDA

Es un método estadístico para tomar decisiones sobre los datos basándose en evidencia experimental.

- **Uso en EDA:** Sirve para confirmar si una diferencia observada en los gráficos es estadísticamente significativa o solo producto del azar.
- *Ejemplo:* ¿Es el salario promedio de los hombres significativamente diferente al de las mujeres en este dataset? (Usando un T-test).

Siguiente paso

¿Te gustaría que generase un **script de Python de ejemplo** (usando pandas y seaborn) que aplique estos pasos a un dataset ficticio para que puedas incluir capturas de código en tu PDF?