

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.

Descripción:

Classroom IA School F5 P6 TAREA II: EDA Alexandra Zambrano • Ayer (Última modificación: Ayer)

[https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRzZ-DS/view?usp=classroom\\_web&authuser=0](https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRzZ-DS/view?usp=classroom_web&authuser=0)

<https://classroom.google.com/c/ODMwODM3MDA1ODQ4/a/ODQyNDY3NDc4OTQz/details>

Fecha de creación:

20260203100800

Versión:

20260203131300

Página 01 de 7

¡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 quieras 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).

Descripción:

Classroom IA School F5 P6 TAREA II: EDA Alexandra Zambrano • Ayer (Última modificación: Ayer)

[https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRz-DS/view?usp=classroom\\_web&authuser=0](https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRz-DS/view?usp=classroom_web&authuser=0)

<https://classroom.google.com/c/ODMwODM3MDA1ODQ4/a/ODQyNDY3NDc4OTQz/details>

## 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.

Descripción:

Classroom IA School F5 P6 TAREA II: EDA Alexandra Zambrano • Ayer (Última modificación: Ayer)

[https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRzZ-DS/view?usp=classroom\\_web&authuser=0](https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRzZ-DS/view?usp=classroom_web&authuser=0)

<https://classroom.google.com/c/ODMwODM3MDA1ODQ4/a/ODQyNDY3NDc4OTQz/details>

Fecha de creación:

20260203100800

Versión:

20260203131300

Página 03 de 7

- **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?

Descripción:

Classroom IA School F5 P6 TAREA II: EDA Alexandra Zambrano • Ayer (Última modificación: Ayer)  
[https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRzZ-DS/view?usp=classroom\\_web&authuser=0](https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRzZ-DS/view?usp=classroom_web&authuser=0)  
<https://classroom.google.com/c/ODMwODM3MDA1ODQ4/a/ODQyNDY3NDc4OTQz/details>

Hola Gemini.

¿Puedes visitar:

[https://github.com/Factoria-F5-madrid/ai\\_project\\_EDA./blob/main/README.md](https://github.com/Factoria-F5-madrid/ai_project_EDA/blob/main/README.md)?.

Me gustaría que generases código de Python

(usando Requests; Pandas; Matplotlib, Pyplot, Seaborn, Bokeh)

que aplique estos pasos a un dataset (pueden encontrarse ejemplos en:

<https://covidtracking.com/data/download>

).

Querría poder crear un repositorio en GitHub para mostrar el trabajo,

con al menos un Jupyter Notebook integrado en vscode,

que use los servidores de Google para pruebas/ejecución

con el análisis que se ha llevado a cabo claramente explicado,

ordenado adecuadamente en ramas y mensajes de commit limpios.

Un informe ejecutivo de las conclusiones obtenidas.

ejemplo:

<https://github.com/Factoria-F5-madrid/project-ai-data-analyst>

Briefing: PROYECTO DATA ANALYST

#### 🔍 Planteamiento

Somos el departamento de datos de una StartUp de reciente fundación que pretende desarrollar producto de software propio en el campo inmobiliario. Nuestro principal reclamo como empresa y de cara a los inversores es que somos una empresa moderna completamente data driven que incorporamos tecnologías modernas como la IA en nuestros productos. Recientemente hemos sido contratados por la empresa AirBnB, quien nos ha cedido algunos datos de algunas de las ciudades más importantes para que extraigamos toda la información útil que seamos capaces de encontrar en ellos. De alguna manera es una prueba para contratarnos para futuros proyectos, por lo que cuanto más encontremos, mejor (explicaciones a outliers, confirmación de hipótesis, patrones extraños, etc) La clave es usar todas las armas de nuestro arsenal para impresionar al cliente.

#### 🎯 Objetivos del Proyecto

Analizar un dataset.

Preprocesar los datos.

Generar visualizaciones avanzadas y creación dashboards.

Generación de insights.

#### 📦 Condiciones de Entrega

Para la fecha de entrega, los equipos deberán presentar:

Repositorio en GitHub con el código fuente documentado.

Un panel de visualización de datos en PowerBI/Tableau el cual debe mostrar la información que consideráis importante para la gente de negocio.

Notebook en el que habéis hecho el análisis Debe estar todo comentado usando las celdas de markdown, con todo el proceso y las conclusiones obtenidas. Pensad que de este análisis saldrá lo que mostráis en el panel de visualización.

Demo en vivo mostrando el funcionamiento del panel de visualización/dashboard.

Presentación técnica, explicando los objetivos, desarrollo y tecnologías utilizadas.

Tablero Kanban con la gestión del proyecto (Trello, Jira, etc.).

#### Descripción:

Classroom IA School F5 P6 TAREA II: EDA Alexandra Zambrano • Ayer (Última modificación: Ayer)

[https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRzZ-DS/view?usp=classroom\\_web&authuser=0](https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRzZ-DS/view?usp=classroom_web&authuser=0)

<https://classroom.google.com/c/ODMwODM3MDA1ODQ4/a/ODQyNDY3NDc4OTQz/details>

**⚙️ Tecnologías Recomendadas**

Control de versiones: Git / GitHub

Entorno de ejecución: Docker

Lenguaje principal: Python

Librerías útiles: Pandas, Numpy, Matplotlib, Seaborn, Plotly

Visualización de Datos: Streamlit, Dash, PowerBI, Tableau

Gestión del proyecto: Trello, Jira, Github

**💡 Datos**

AirBnB ciudades CSV

<https://drive.google.com/drive/folders/17sYr63LjEX30-3-KjXlaPP-bRwEmMqpf>**Listado de datos:**

#	Modificado	Tamaño	Comprimido	Suma de control	Nombre
#					
2019-09-10 00:32	1738173	536554	09900810	primer_proyecto_datos\tokyo_airbnb.csv	# Archivo C:\Users\Coder\Proyectos\EDA\Datos\primer_proyecto_datos-20260203T100006Z-3-001.20260203110038M.zip
2019-10-27 14:05	7077973	2360257	8CC9B852	primer_proyecto_datos\NY_airbnb.csv	
2019-11-17 17:40	5504518	2109580	AD748547	primer_proyecto_datos\sydney_airbnb.csv	
2021-05-22 07:09	2801783	917236	4E01497F	primer_proyecto_datos\madrid_airbnb.csv	
2021-08-26 14:00	2393568	836538	07CAD94C	primer_proyecto_datos\milan_airbnb.csv	
2022-08-16 15:53	11578155	3995593	B618D848	primer_proyecto_datos\london_airbnb.csv	
2024-10-08 01:49	1779197	1767681	C1C061FE	primer_proyecto_datos\milan_airbnb.xlsx	
2024-10-11 06:24	8230750	8180990	0681ADBB	primer_proyecto_datos\london_airbnb(1).xlsx	
2024-10-24 04:16	8262897	8210286	478ADBD2	primer_proyecto_datos\london_airbnb.xlsx	
2026-01-15 00:17	1972142	1961027	E952CECC	primer_proyecto_datos\madrid_airbnb.xlsx	
#					
# Total		Tamaño	Comprimido	Ficheros	
#		51339156	30875742	10	

**💡 Niveles de Entrega****● Nivel Esencial:**

- Análisis exploratorio de datos (EDA) completo con visualizaciones clave que muestren la distribución de las variables, patrones generales, detección de outliers,etc.

- Documentación del análisis en el notebook, bien comentada y estructurada en celdas de markdown, con conclusiones claras.

- Kanban board mostrando la planificación y organización del proyecto.

- Repositorio Git con ramas bien organizadas y commits limpios y descriptivos.

- Documentación del código y un README en GitHub.

**● Nivel Medio:**

- Panel de visualización de datos funcional en PowerBI/Tableau, que presente la información más relevante.

- Uso de visualizaciones avanzadas con Plotly o Seaborn, incluyendo gráficos interactivos.

- Segmentación del análisis en función de distintas ciudades o grupos de usuarios, mostrando diferencias clave entre ellos.

**● Nivel Avanzado:**

- Implementación de un sistema de filtros interactivos en el panel, para que el usuario pueda explorar los datos de forma dinámica.

- Versión dockerizada del panel de visualización que permita fácil despliegue y portabilidad.

- Hipótesis verificadas extraídas de los datos, acompañadas de análisis estadísticos adecuados (p. ej. test de hipótesis o correlaciones) que justifiquen las conclusiones presentadas.

**Descripción:**

Classroom IA School F5 P6 TAREA II: EDA Alexandra Zambrano • Ayer (Última modificación: Ayer)  
[https://drive.google.com/file/d/112ViwnTJZ-OsEG8L3Cq7fenXPLRz-DS/view?usp=classroom\\_web&authuser=0](https://drive.google.com/file/d/112ViwnTJZ-OsEG8L3Cq7fenXPLRz-DS/view?usp=classroom_web&authuser=0)  
<https://classroom.google.com/c/ODMwODM3MDA1ODQ4/a/ODQyNDY3NDc4OTQz/details>

**● Nivel Experto:**

- Análisis predictivo o de clustering, utilizando modelos clásicos de machine learning (p. ej. regresión lineal, clustering k-means) para identificar patrones en los datos.
- Integración de datos adicionales (p. ej., datos demográficos o económicos de las ciudades) para enriquecer el análisis.
- Despliegue en un servidor accesible públicamente.

**Evaluación**

Se considerarán los siguientes criterios:

Competencia: Aplicar algoritmos de aprendizaje automático según el problema, identificando y resolviendo problemas clásicos de inteligencia artificial:

- Seleccionar las variables que son útiles y las que no lo son.
- Aplicar modelos de clustering.
- Uso de técnicas avanzadas de visualización como heatmaps, pairplots, y 3D visualizations.
- Introducción a estadística para Machine Learning.

Competencia: Modificar los parámetros y componentes de la inteligencia artificial para mejorar su rendimiento:

- Visualización de resultados y métricas.

Más detalles en: [roadmap-mad-ai-p4.coderf5.es](http://roadmap-mad-ai-p4.coderf5.es)  
<https://roadmap-mad-ai-p4.coderf5.es/>

¿Puedes tener en cuenta los siguientes requisitos?:

Uso y gestión de formato .csv  
 Limpieza y preprocesado de datos  
 Visualización de datos (seaborn, matplotlib, plotly)  
 Análisis exploratorio detallado (EDA)

Me gustaría poder ejecutar dicho código y que estuviera comentado inline siguiendo los pasos que me has indicado en los 10 puntos que me has indicado.

Esto es un proyecto complejo a desarrollar en múltiples iteraciones de consultas. En caso de que sea necesario refinar la consulta no dudes en pedírmelo.

Creo que no podré subir todos los datos de una única vez.

Por favor, Gemini, ¿puedes guiarme para que vayamos resolviendo juntos el proyecto?. Ten en cuenta que soy novel en esto. Hay una semana desde hoy. Vamos a intentar resolver sólo el Nivel Esencial.

En primer lugar, creo que sería necesario hacer un esquema de cómo abordarlo.

Necesitaría comprobar que es lo que tengo instalado (vscode y extensiones, Git Bash, Python, etc). Necesitaría que me guiaras en los pasos para crear .git, .venv, dependencias, .gitignore, programas de bases de datos asociados, docker, conectar con el github de Factoria F5 desde vscode y crear el correspondiente repositorio (puede ser desde la interfaz web).

Gracias Gemini.

## Descripción:

Classroom IA School F5 P6 TAREA II: EDA Alexandra Zambrano • Ayer (Última modificación: Ayer)  
[https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRzZ-DS/view?usp=classroom\\_web&authuser=0](https://drive.google.com/file/d/1i2ViwnTJZ-OsEG8L3Cq7fenXPLRzZ-DS/view?usp=classroom_web&authuser=0)  
<https://classroom.google.com/c/ODMwODM3MDA1ODQ4/a/ODQyNDY3NDc4OTQz/details>