

tp5

Primera Parte: PARTE I. Datos abiertos

1. Investigar y mencionar 3 sitios de datos abiertos donde se puedan obtener datasets relevantes para diferentes análisis a realizar. Describa el tipo de datos que se puede encontrar en cada uno.
 2. ¿Cuál es la diferencia entre datos públicos y datos abiertos? Proporciona un ejemplo de cada tipo.
 3. Mencione 3 tipos de licencias que pueden tener los datos abiertos, describiendo diferencias entre ellas.
 4. Suponga que tiene acceso a un dataset abierto sobre los niveles de contaminación del aire en diferentes ciudades del país. ¿Qué tipos de análisis podría realizar para obtener información útil?
 5. ¿Cuáles son las condiciones de la licencia creative commons?
 6. ¿Qué significa tener datos IA-ready?
 7. ¿Cuáles son los principios FAIR-R y sus requisitos?
-

1. Tres sitios de datos abiertos y qué tipo de datos ofrecen

a) Datos.gob.ar

Es el portal oficial de datos abiertos del Gobierno de Argentina.

👉 Ofrece datasets sobre transporte público, educación, salud, economía, ambiente, seguridad, entre otros.

Por ejemplo, podés descargar datos sobre accidentes viales o calidad del aire por ciudad.

b) Data.gov (Estados Unidos)

Portal del gobierno de EE. UU. con miles de conjuntos de datos públicos.

👉 Contiene información sobre clima, agricultura, energía, estadísticas laborales, y más.

Ideal para análisis comparativos internacionales o estudios de tendencias globales.

c) Kaggle Datasets

Plataforma colaborativa de ciencia de datos.

👉 Incluye datasets subidos por usuarios e instituciones, sobre temas como machine learning, deportes, salud, economía o redes sociales.

Muchos vienen listos para entrenar modelos o hacer análisis exploratorios.

2. Diferencia entre datos públicos y datos abiertos

- **Datos públicos:**

Son datos a los que el público puede acceder, pero no necesariamente se pueden reutilizar libremente.

Ejemplo: un informe PDF del INDEC sobre inflación — es público, pero no “abierto” porque no está en formato reutilizable ni con licencia de uso.

- **Datos abiertos:**

Son datos accesibles, reutilizables y redistribuibles libremente, con una licencia clara que lo permita y en formatos legibles por máquina (como CSV o JSON).

Ejemplo: un dataset de transporte urbano publicado en formato CSV bajo licencia abierta.

3. Tres tipos de licencias de datos abiertos

a) Creative Commons CC-BY

Permite copiar, modificar y redistribuir los datos, siempre que se mencione la fuente.

- Es muy abierta, solo exige atribución.

b) CC-BY-SA (ShareAlike)

Permite uso y modificación, pero cualquier obra derivada debe mantenerse bajo la misma licencia.

- Promueve compartir con las mismas condiciones.

c) CC0 (Dominio Público)

Renuncia total de derechos.

- Los datos pueden usarse sin atribución ni restricciones.
-

4. Análisis posibles con un dataset sobre contaminación del aire

Podrías hacer varios tipos de análisis útiles, por ejemplo:

- **Análisis temporal:** ver cómo varía la contaminación a lo largo del tiempo.
 - **Análisis geográfico:** comparar niveles de contaminación entre provincias o ciudades.
 - **Correlaciones:** estudiar la relación entre contaminación y factores como población o tráfico.
 - **Predicciones:** aplicar modelos de machine learning para estimar niveles futuros de contaminación.
 - **Visualizaciones:** mapas de calor o gráficos de tendencia para comunicar resultados fácilmente.
-

5. Condiciones de la licencia Creative Commons

Las licencias **Creative Commons (CC)** permiten combinar diferentes condiciones:

- **BY (Atribución):** se debe citar al autor.

- **SA (ShareAlike):** las obras derivadas deben mantenerse con la misma licencia.
- **NC (No Comercial):** prohíbe el uso con fines comerciales.
- **ND (No Derivatives):** no permite crear obras derivadas.

Ejemplo: **CC-BY-NC-SA** → se puede compartir y modificar con atribución, pero no para fines comerciales, y con la misma licencia.

6. Qué significa tener datos “IA-ready”

Significa que los datos están preparados para ser utilizados en inteligencia artificial o machine learning.

👉 En otras palabras, están **limpios, estructurados, etiquetados y en formatos estándar** (como CSV, JSON o Parquet).

También deben tener **metadatos claros** y cumplir principios de calidad, coherencia y actualización.

7. Principios FAIR-R y sus requisitos

Los **principios FAIR-R** buscan que los datos sean:

1. **Findable (Encontrables):** tengan identificadores únicos y metadatos claros.
 2. **Accessible (Accesibles):** puedan obtenerse fácilmente, con protocolos abiertos.
 3. **Interoperable (Interoperables):** se integren con otros sistemas o formatos estándar.
 4. **Reusable (Reutilizables):** incluyan licencias claras y documentación completa.
 5. **Responsible (Responsables):** se gestionen de forma ética, segura y con respeto a la privacidad.
-
-

PARTE II. Visualización de Datos

1. Explique qué es una medida y una dimensión

- **Medida:** es un valor numérico que se puede calcular, sumar o promediar. Representa **cantidades** (por ejemplo, ventas, precios, ingresos).
👉 Ejemplo: el monto total de ventas en una sucursal.
 - **Dimensión:** es una categoría o calidad que permite agrupar, filtrar o segmentar las medidas.
👉 Ejemplo: la región, el producto o el mes de la venta.
-

2. Explique la diferencia entre un dato discreto y un dato continuo

- **Dato discreto:** solo puede tomar **valores específicos** y contables (números enteros).
👉 Ejemplo: cantidad de clientes o número de autos vendidos.
 - **Dato continuo:** puede asumir **cualquier valor dentro de un rango**, incluyendo decimales.
👉 Ejemplo: temperatura, peso o tiempo de respuesta.
-

3. ¿Por qué es importante preparar los datos?

Preparar los datos es esencial para garantizar que el análisis sea **preciso y confiable**.

Permite:

- Corregir errores, valores nulos o duplicados.
- Unificar formatos (fechas, unidades, nombres).
- Estandarizar categorías y normalizar valores.
- Aumentar la calidad y la coherencia de la visualización final.

En resumen, sin una buena preparación, los gráficos pueden ser engañosos o incorrectos.

Selección de tipos de gráficos y justificación

a. Comparación de suscripciones anuales por región geográfica

Pregunta: ¿Qué tipo de gráfico utilizaría para comparar las ventas entre las regiones durante los 5 años?

Respuesta:

Usaría un **gráfico de líneas múltiples** o un **gráfico de barras agrupadas**.

- El **gráfico de líneas** permite ver la evolución de las suscripciones a lo largo del tiempo en cada región.
- Si el objetivo es comparar más claramente entre regiones año a año, el **gráfico de barras agrupadas** facilita esa comparación directa.

👉 **Justificación:** ambos muestran bien los cambios en el tiempo, pero si querés resaltar tendencias, las líneas son más efectivas.

b. Análisis de la distribución de las edades de clientes

Pregunta: ¿Qué tipo de gráfico utilizará para representar la distribución de las edades de los clientes?

Respuesta:

Usaría un **histograma**.

👉 **Justificación:** permite ver cómo se distribuyen las edades agrupadas en intervalos (por ejemplo, 18–25, 26–35, etc.), mostrando dónde se concentra la mayoría de los clientes.

c. Relación entre el precio y la puntuación otorgada por el cliente

Pregunta: ¿Qué tipo de gráfico usaría para analizar si existe una relación entre el precio del producto y la puntuación del cliente?

Respuesta:

Usaría un **gráfico de dispersión (scatter plot)**.

👉 **Justificación:** permite visualizar la relación entre dos variables numéricas (precio y puntuación), y detectar si existe correlación positiva, negativa o ninguna.

d. Análisis de los préstamos de libros por género

Pregunta: ¿Qué gráfico es adecuado para visualizar la proporción de préstamos de cada género?

Respuesta:

Usaría un **gráfico de torta (pie chart)** o un **gráfico de barras simples**.

👉 **Justificación:** ambos muestran proporciones. El gráfico de torta es útil si se quiere resaltar qué género predomina, mientras que el de barras facilita comparar las cantidades exactas.

e. Análisis de la temperatura promedio del mar a lo largo del tiempo

Pregunta (a): ¿Qué tipo de gráfico utilizará para mostrar los cambios de la temperatura promedio a lo largo del tiempo?

Respuesta:

Usaría un **gráfico de líneas**.

👉 **Justificación:** es ideal para representar series temporales, mostrando claramente cómo varía la temperatura promedio a lo largo de los 7 años.

Pregunta (b): ¿Cuáles tablas son relevantes para presentar el análisis?

Respuesta:

Las tablas **TemperaturasPromedio** y **TipoGradoTemp** son las más relevantes.

- **TemperaturasPromedio** contiene la **fecha y valor promedio**, necesarios para la gráfica temporal.
 - **TipoGradoTemp** puede ayudar a indicar el tipo de unidad (Celsius, Fahrenheit, etc.) en la visualización.
-

f. Visualización de la proporción de especies por categoría

Pregunta (c): ¿Qué gráfico utilizaría si se quiere visualizar la proporción de especies de cada categoría?

Respuesta:

Usaría un **gráfico de torta** o un **gráfico de barras horizontales**.

👉 **Justificación:** ambos permiten ver fácilmente qué categorías (Moluscos, Peces, etc.) tienen mayor cantidad de especies identificadas.

Pregunta (d): ¿Cuál o cuáles tablas son relevantes para presentar el análisis?

Respuesta:

Las tablas **CategoríaEspecie** y **EspecieIdentificada**.

- **CategoríaEspecie** aporta el nombre de la categoría.
 - **EspecieIdentificada** brinda la cantidad de especies por categoría, necesaria para el gráfico.
-

g. Visualización de las estadísticas de una cadena de supermercados

Pregunta: ¿Qué tipo de gráfico podría utilizar y cuáles tablas son relevantes para el análisis?

Respuesta:

Primer caso: *Cantidad de productos vendidos por categoría en todas las sucursales*

- Usaría un **gráfico de barras agrupadas** (categorías vs. cantidad).

👉 **Justificación:** muestra claramente qué tipo de producto se vende más.

Segundo caso: *Total de ingresos de la sucursal número 10 mes a mes*

- Usaría un **gráfico de líneas** para mostrar la evolución mensual de ingresos.
- 👉 **Justificación:** permite ver fácilmente si hubo aumentos o caídas mes a mes.

Tablas relevantes:

- **Venta**: aporta el monto total y la fecha de cada venta.

- **Item_Venta** : permite conocer cuántos productos se vendieron.
 - **Producto** : indica el precio unitario y categoría.
 - **Sucursal** : identifica la sucursal y su ubicación.
-

PARTE III: Graficando con Tableau

