Practica 5 - Visualización de datos

Parte 1

Inciso 1

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.

Respuesta

Menciono 3 sitios de datos abiertos donde se pueden obtener datasets relevantes:

1. **Datos Argentina** (`https://datos.gob.ar`):
   - **Tipo de datos:** Datasets del Sector Público Nacional de Argentina. Se pueden encontrar datos sobre **Estadísticas Educativas**, **Cultura**, **Agroganadería**, **Ciencia y Tecnología**, y **Economía y Finanzas**.
2. **Portal Europeo de Datos** (`https://data.europa.eu/en`):
   - **Tipo de datos:** Conjuntos de datos, catálogos y publicaciones de las instituciones de la Unión Europea y de los Estados miembros. Abarca temas muy diversos como política, economía, geografía y resultados de investigación.
3. **Wikidata** (`https://www.wikidata.org/?uselang=es`):
   - **Tipo de datos:** Base de conocimiento secundaria, libre y abierta que almacena **datos estructurados** para ser utilizados en proyectos de Wikimedia y otros sitios. Su contenido está disponible bajo la licencia Creative Commons Dedicación al Dominio Público (CC0).

Inciso 2

2. ¿Cuál es la diferencia entre datos públicos y datos abiertos? Proporciona un ejemplo de cada tipo.

Respuesta

Datos públicos	Datos abiertos
Un dato público es cualquier dato generado en el ámbito gubernamental, o que se encuentra bajo su guarda	Los datos abiertos son aquellos de origen público o no a los que cualquier persona puede acceder, usar y compartir libremente. Sólo deben atribuirse y compartirse con la misma

Datos públicos	Datos abiertos
	licencia con la que fueron publicados.

Tipo de Dato	Ejemplo Concreto
Dato Público	Un **informe anual escaneado** sobre el presupuesto nacional, publicado en la web oficial como un **archivo PDF no seleccionable ni editable**.
Dato Abierto	Un **Dataset de las estadísticas de tráfico** de una ciudad, publicado como un **archivo CSV** y con licencia **Creative Commons Reconocimiento (CC-BY)**.

Inciso 3

Mencione 3 tipos de licencias que pueden tener los datos abiertos, describiendo diferencias entre ellas.

Respuesta

Licencia	Descripción	Diferencia Clave
ODC Public Domain Dedication and License (PDDL)	Permite difundir, reutilizar o adaptar los datos **sin restricción alguna**.	Es la menos restrictiva. Equivale a dominio público.
ODC Attribution License (ODC-BY)	Exige la **referencia a la autoría o fuente** de los datos (atribución) para la reutilización.	Requiere únicamente la **atribución** al autor o fuente.
ODC Open Database License (ODbL)	Permite la reutilización siempre que se **reconozca la autoría** y, si se crea una base de datos derivada, esta se comparta bajo la **misma licencia** (Compartir Igual).	Requiere **atribución y "Compartir Igual"** (licenciamiento recíproco).

Inciso 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?

Respuesta

Con un dataset abierto sobre niveles de contaminación del aire por ciudad, se podrían realizar los siguientes análisis útiles:

- **Análisis Descriptivo y Comparativo:**
  - Calcular el **promedio y variabilidad** de los contaminantes (PM2.5, O3, NO2) por ciudad, mes o año.
  - **Clasificar y comparar** las ciudades según sus niveles de contaminación para identificar las más y menos afectadas.
- **Análisis de Series de Tiempo:**
  - Identificar **tendencias** a largo plazo (si la contaminación sube o baja) y **patrones estacionales** (ej., más contaminación en invierno o verano).
- **Análisis de Correlación:**
  - Investigar la relación entre los niveles de contaminantes y otros factores (ej., densidad de población, tráfico vehicular o condiciones meteorológicas) para comprender las **causas subyacentes**.
- **Modelos Predictivos:**
  - Desarrollar modelos de *machine learning* para **predecir** los niveles de contaminación con anticipación, lo que permite a las autoridades tomar medidas preventivas.

Inciso 5

¿Cuáles son las condiciones de la licencia creative commons?

Respuesta

Condiciones:

- Atribución: Exige reconocer la autoría de la obra.
- No comercial: Prohíbe el uso de la obra con fines de lucro directos o indirectos.
- Sin obras derivadas: La autorización para explotar la obra no incluye la posibilidad de crear una obra derivada
- Compartir igual: Si se genera una obra derivada, el resultado debe distribuirse bajo una licencia **igual o similar** a la que regula la obra original.

Inciso 6

¿Qué significa tener datos IA-ready?

Respuesta

Tener datos IA-ready significa que estos datos cumplen una serie de requisitos técnicos, estructurales y de calidad que optimizan su aprovechamiento por parte de los algoritmos de inteligencia artificial. Esto incluye múltiples aspectos como:

- completitud de los datos
- ausencia de errores e inconsistencias
- uso de formatos adecuados, metadatos y estructuras homogéneas
- proporcionar el contexto necesario para poder verificar que estén alineados con el uso que la IA les dará.

Inciso 7

¿Cuáles son los principios FAIR-R y sus requisitos?

Respuesta

- Encontrables
- Accesibles
- Interoperables
- Reutilizables
- Preparados para la IA

Requisitos:

- etiquetado exhaustivo
- documentación completa del origen de los datos
- homogeneidad de estándares y metadatos
- cobertura suficiente que evite sesgos
- licencias que regulen claramente su uso en IA.

Parte 2

Preguntas

1. Explique que es una medida y una dimensión
2. Explique la diferencia entre un dato discreto y un dato continuo
3. ¿Por qué es importante preparar los datos?

Respuesta

1. Medida: Las medidas son datos numéricos cuantitativos.
   Dimension: Las dimensiones son datos cualitativos, como un nombre o una fecha.
2. Continuos: Pueden contener un numero infinito de valores. Puede tratarse de un rango, como las ventas de determinado intervalo de fechas o cantidades
   Discretos: Contienen numero finito de valores, como país, provincia o nombre de cliente.
3. La preparación de datos (o **limpieza y transformación de datos**) es un paso crucial porque garantiza que los datos sean de **alta calidad** y estén en un **formato adecuado** antes de cualquier análisis, modelado o visualización. **Los resultados de un análisis**

son tan buenos como los datos que se utilizan. Preparar los datos transforma la "materia prima" defectuosa en información confiable y procesable.

Escenarios

Dadas las situaciones que se presentan a continuación, decidir qué tipo de gráfico utilizaría para visualizar la información de manera clara y efectiva. Justifique su elección indicando por
qué ese tipo de gráfico es el más adecuado para cada caso.

Escenario A

Comparación de suscripciones anuales por región geográfica
Se cuenta con un conjunto de datos de las suscripciones anuales de una empresa de telefonía
celular en distintas regiones (Norte, Sur, Este, Oeste) durante los últimos 5 años. ¿Qué tipo de
gráfico utilizaría para comparar las ventas entre las regiones durante los 5 años?

Respuesta

Utilizaría un gráfico de lineas ya que muestra las relaciones de los cambios en los datos en un período de tiempo, facilitando la identificación de tendencias. Donde cada linea podría ser tranquilamente una region en este caso

Escenario B

Análisis de la distribución de las edades de clientes
Se tiene un dataset con datos de los clientes de una tienda virtual, entre ellos, la edad. El objetivo es entender cómo se distribuyen las edades de los clientes. ¿Qué tipo de gráfico utilizará para representar la distribución de las ciudades de los clientes?

Respuesta

Utilizaría el grafico de barras ya que nos permiten comparar valores numéricos como números
enteros y porcentajes. Este tipo de gráfico es ideal para **comparar las frecuencias absolutas o relativas** de categorías discretas. Cada barra representaría una ciudad, y su altura o longitud indicaría cuántos clientes provienen de allí.

Escenario C

Relación entre el precio y la puntuación otorgada por el cliente
Se tiene información sobre el precio de diferentes servicios ofrecidos y la calificación otorgada
por los clientes. ¿Qué tipo de gráfico usaría para analizar si existe una relación entre el

precio
del producto y la puntuación marcada por el cliente?

Respuesta

Utilizaría gráfico de Dispersión porque es la herramienta fundamental para visualizar la **relación, correlación o asociación** entre **dos variables numéricas** (o una numérica y una ordinal, como una puntuación).

En este caso, tendrías el **Precio** en el eje X y la **Puntuación del Cliente** en el eje Y.

Escenario D

Análisis de los préstamos de libros por género
Se cuenta con el registro de los préstamos de una biblioteca escolar. Entre los datos de cada
uno de estos se tiene el género del libro (narrativa, poesía, cuento, novela y biografía). ¿Qué
gráfico es adecuado para visualizar la proporción de préstamos de cada género?

Respuesta

Eligiría el **Gráfico de Barras** porque es superior al Gráfico de Torta para la **comparación de magnitudes** o proporciones entre 5 categorías/géneros (**Narrativa, Poesía, Cuento, Novela y Biografía**). Aunque el gráfico de torta no se elige principalmente por si se tiene mayor cantidad de géneros de libros

Escenario E

Se han almacenado datos de registro de la temperatura promedio del mar, a partir de mediciones diarias frente a la costa de Las Toninas, un kilómetro mar adentro. Esta medición
se viene realizando durante los últimos 7 años para estudios del comportamiento de la fauna
del lugar.
Dado el siguiente esquema de la base de datos:
TipoGradoTemp (**#tipoGradoTemp**, descripcionTipoGrado)
TemperaturaRegistrada (**#registroTemp,** fecha, hora, valorTemp, #tipoGradoTemp)
TemperaturasPromedio (**#registroProm**, fecha, valorProm)

a) ¿Qué tipo de gráfico utilizará para mostrar los cambios de la temperatura promedio a lo largo del tiempo?
b) ¿Cuáles tablas son relevantes para presentar el análisis?

Respuesta

Eligiría un gráfico de líneas porque muestran relaciones de los cambios de datos dentro de un periodo de tiempo

Con la tabla temperaturas promedio sería suficiente
Contiene directamente los campos necesarios para el gráfico de líneas:

- `fecha` (Eje X o tiempo)
- `valorProm` (Eje Y o valor numérico de la temperatura promedio)

Escenario F

Además, el dataset muestra la cantidad de especies que se identificaron en el mar en las 5 distintas categorías (Moluscos, Artrópodos, Cnidarios, Peces y Mamíferos marinos).
Dado el siguiente esquema de la base de datos:
CategoriaEspecie (`#cat_especie` , cat_nombre)
EspecieIdentfificada (`#especie` **,** #cat_especie, nombre_especie, cantidad)

c) ¿Qué gráfico utilizaría si se quiere visualizar la proporción de especies de cada categoría?
d) De todas las tablas propuestas, indicar cuál o cuáles son relevantes para presentar el análisis

Respuesta

Se puede utilizar un gráfico de barras, ya que sirve para comparar valores numericos. Cada barra representará una de las 5 categorías, y su longitud indicará la **cantidad** de especies identificadas, permitiendo una comparación clara y precisa de qué categoría domina en el total.

Deberían utilizar ambas tablas ya que para ayudar a la comprensión, se requiere el nombre de la especie

Escenario G

Visualización de las estadísticas de una cadena de supermercados
Una cadena de supermercados quiere contar con estadísticas de las ventas por sucursal y por
tipo de producto a lo largo del último año (mes a mes). Se quiere identificar las sucursales con
mayores ventas y los tipos de productos que generan más ingresos.
Esta cadena de supermercados quiere visualizar:

- La cantidad de productos vendidos de cada categoría para todas las sucursales, para conocer qué tipo de producto es el que más se vende.
- El total ingresos de la sucursal número 10, mes a mes, durante los últimos 12 meses, para determinar si hubo o no un incremento de los ingresos. Para ello dispone de una base de datos con el siguiente esquema:

Venta (**id_venta**, fecha_venta, id_sucursal, monto_total)
Item_Venta (**id_venta, id_producto**, cant)
Sucursal (**id_sucursal**, ubicación, cant_empleados)
Producto (**id_producto,** nombre_producto, desc_producto ,precio_unit, categoria)
Cliente (**id_cliente**, nombre_cliente, apellido_cliente, tipo_cliente)
Determine qué tipo de gráfico podría utilizar y justifique su elección.
Con los esquemas proporcionados, elegir cuáles -con sus atributos- son relevantes para presentar el análisis visual propuesto anteriormente

Respuesta

Para la visualización por categoría, utilizaría un grafico de barras. Me parece ideal para contar las cantidad de cada uno de los productos. Para este grafico es importante la tabla item venta, los atributos cant junto con id_producto

Para los ingresos de la sucursal numero 10, utilizaria un gráfico de lineas. facilitando la identificación de las tendencias a lo largo de los 12 meses que se requiere. Y para este se requiere el id de la sucursal, junto con las fechas de cada una de las ventas , y dentro de un mes, contar el monto recaudado, con la tabla de ventas alcanzaría.