Cátedra de Bases de Datos 1
Cursada 2024

Práctica Nro. 5

Análisis de Datos y Visualización

Publicación: 30/10/2024
Finalización: 08/11/2024

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.

Datos Argentina:

- **Descripción:** Este portal ofrece una amplia variedad de datos públicos en formatos abiertos, incluyendo datos sobre agropecuaria, pesca, forestación, educación, cultura, deportes, entre otros.
- **Tipo de datos:** Datos sobre producción agropecuaria, estadísticas educativas, indicadores culturales, y mucho más.

World Bank Open Data:

- **Descripción:** El Banco Mundial proporciona un extenso repositorio de datos sobre lo que está sucediendo en diferentes países del mundo.
- **Tipo de datos:** Datos económicos, sociales, ambientales, y geoespaciales, incluyendo microdatos, estadísticas de series temporales y datos geoespaciales.

Kaggle:

- **Descripción:** Una plataforma popular para científicos de datos y analistas donde se pueden encontrar datasets de diversas áreas y participar en competiciones de análisis de datos.
- **Tipo de datos:** Datasets para análisis de datos, machine learning, competiciones de ciencia de datos, y proyectos de investigación.

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

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 licencia con la que fueron publicados.

Ejemplo de dato público: El presupuesto asignado a un ministerio (es público pero podría no estar en formato accesible/procesable)

Ejemplo de dato abierto: Dataset en formato CSV con el desglose detallado de gastos gubernamentales, disponible para descarga y uso libre.

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

1. Creative Commons Attribution (CC BY)
    a. Permite usar y modificar los datos
    b. Requiere atribución al autor original
    c. Permite uso comercial
2. Open Database License (ODbL)
    a. Similar a CC BY pero específica para bases de datos
    b. Requiere que trabajos derivados mantengan la misma licencia
    c. Incluye protecciones específicas para bases de datos
3. CC0 (Creative Commons Zero)
    a. Renuncia a todos los derechos
    b. No requiere atribución
    c. La más permisiva de todas

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?

Análisis temporal:

- Tendencias por hora/día/mes/año
- Patrones estacionales
- Correlación con eventos específicos.

Análisis espacial:

- Mapas de calor por zonas
- Comparación entre ciudades
- Identificación de puntos críticos

PARTE II. Visualización de Datos

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.

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?

Sería lógico usar un mapa de área, dado que sirve para mostrar la relación entre el dato (suscripciones anuales) y la región geográfica específica que es analizada. Podrían usarse distintos colores para diferenciar las áreas.

b. Seguimiento del caudal de un río

Se han almacenado datos de registro del caudal promedio mensual en un punto de interés del río Salado durante los últimos 10 años dada la baja que se observa. ¿Qué tipo de gráfico usaría para mostrar la evolución del caudal a lo largo del tiempo?

Deberíamos usar un gráfico de línea, para graficar el caudal del agua en la unidad de medida pertinente comparada al paso del tiempo.

c. 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 edades de los clientes?

Usaría un gráfico de barras que podría comparar por rangos de edades. Es un gráfico óptimo para la tarea porque nos permite comparar valores numéricos uno al lado del otro.

d. 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?

Utilizaría un gráfico de dispersión para encontrar un punto en el esquema donde se relacionen las 2 variables observadas.

e. 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?

Podría ser un gráfico de barras, dado que nos permite comparar porcentajes.

También, dado que son 5 géneros, podríamos usar un gráfico de torta si sabemos que, por ejemplo, novela es el más prestado.

Si no, un gráfico de burbujas también podría servir con las mismas precondiciones que el de torta.

f. Distribución del presupuesto 2025 por departamento y actividades

Se dispone de un dataset que contiene el presupuesto asignado a los diferentes departamentos de una empresa (Marketing, Ventas, Desarrollo, Recursos Humanos), y dentro de cada departamento, las actividades específicas (publicidad, investigación, capacitación, etc.). Qué gráfico utilizará para mostrar cómo se distribuye el presupuesto total por categoría y cuánto consume cada actividad dentro de la categoría?

Un treemap serviría para este caso. Pueden dejarse los departamentos como encabezados, de tamaño respectivo al presupuesto, y las actividades dentro de los departamentos siguiendo la misma lógica. Como son 4 departamentos, es viable hacer un treemap.

g. Tendencia de ingresos mensuales de una empresa

Se registraron los ingresos mensuales de una empresa mes a mes durante los últimos 3 años. Se quiere observar la tendencia para determinar la tendencia al alza o a la baja. ¿Qué tipo de gráfico es adecuado para representar la tendencia de ingresos elegiría?

Un gráfico de línea, así comparamos el alza/baja de los ingresos con el período de tiempo estipulado en una línea de tiempo.

h. Evaluación de comentarios de un trailer de película

Se tiene un conjunto de datos con comentarios sobre el trailer de la película realizados por los seguidores de perfil fan de la misma. ¿Qué tipo de gráfico usaría para evaluar qué comentarios son más frecuentes y por lo tanto, describen al trailer de la película?

Un gráfico de burbujas. Con burbujas más grandes dejaría en claro cuáles son los comentarios más frecuentes, en contraposición con las más chicas (o sea, menos frecuentes).

i. Evolución del precio de una acción

Se dispone de un registro del precio diario de una acción durante el último año. ¿Qué tipo de gráfico elegiría para mostrar la evolución del precio de la acción observada a lo largo del tiempo?

Línea. Porque nos permitría ver la evolución creciente/decreciente como una línea de tiempo.

j. Densidad poblacional por provincia

Un conjunto de datos contiene la densidad de población (habitantes por km2) de cada provincia del país. Qué gráfico le permitirá visualizar rápidamente qué provincias tienen una mayor densidad de población, representando con diferentes tonos? Más oscuro, mayor densidad poblacional.

Un mapa de áreas. Cada provincia es un área y, en base al color, podemos representar as áreas

más y menos densamente pobladas.

k. Visualización de la proporción de ventas por categoría de producto
El dataset muestra la cantidad de clientes que compraron productos de distintas categorías (Vegano, vegetariano, Sin TACC, sin azúcar agregada, orgánico, cosmética).¿Qué gráfico es adecuado para visualizar la proporción de clientes que compraron cada categoría de producto?
Un gráfico de torta quedaría horrible porque son demasiadas categorías. Así que lo mejor sería un gráfico de barras o uno de burbujas, que nos permiten comparar los datos de estas categorías porcentualmente.

PARTE III: Graficando con Tableau

En esta sección, utilizarás Tableau Desktop para explorar y visualizar dos datasets distintos: uno sobre el uso de dispositivos móviles y comportamiento del usuario, y otro con datos de ubicación de tiros de la temporada regular de la NBA. Estos archivos estarán adjuntos a esta práctica, y tu objetivo será importar cada dataset en Tableau y generar gráficos que representan patrones y relaciones claves en los datos.

Importante: Agrega leyendas, etiquetas, títulos y otros elementos visuales a cada gráfico, según crea necesario, asegurándose de esta manera que los gráficos sean comprensibles y que proporcionen contexto suficiente.

1. **Mobile Device Usage and User Behavior Dataset**
Este dataset contiene una muestra de 700 usuarios y detalla patrones de uso de dispositivos móviles. Incluye métricas como el tiempo de uso de aplicaciones, consumo de batería y de datos móviles. Cada usuario está clasificado en una de cinco categorías de comportamiento, desde uso ligero hasta extremo. A continuación se describen las columnas más relevantes del dataset:

- User ID: Identificador único del usuario.
- Device Model: Modelo del dispositivo utilizado.
- Operating System: Sistema operativo del dispositivo (iOS o Android).
- App Usage Time: Tiempo diario en minutos dedicado a aplicaciones.
- Screen On Time: Promedio diario de tiempo de pantalla activa.
- Battery Drain: Consumo diario de batería en mAh.
- Number of Apps Installed: Total de aplicaciones instaladas en el dispositivo.
- Data Usage: Consumo diario de datos en MB.
- Age: Edad del usuario.
- Gender: Género del usuario (Masculino o Femenino).
- User Behavior Class: Clasificación de comportamiento en una escala de 1 a 5, según los patrones de uso.

Ejercicios:
1. Crea un gráfico de barras horizontales que muestre la cantidad de usuarios para cada modelo de dispositivo. Segmenta las barras por el color del sistema operativo (iOS y Android) para identificar las preferencias de uso por sistema.
2. Muestra una línea que refleje el promedio de aplicaciones instaladas y otra línea para el

promedio de datos consumidos, ambos valores en función de la edad del usuario.
3. Crea un histograma que muestre la distribución del número total de aplicaciones instaladas en dispositivos con sistema operativo iOS (filtra los datos para que solo se incluyan dispositivos iOS)
4. Representa en un gráfico de líneas la relación entre la batería gastada y los datos consumidos. Este gráfico permitirá observar cómo el consumo de datos puede influir en el uso de batería.

2. **NBA Regular Season Shot Location Data (2023)**
Este dataset contiene datos detallados sobre los tiros realizados en la temporada regular de la NBA 2023, incluyendo ubicación, distancia y resultado de cada tiro. A continuación, las columnas más importantes:

- TEAM_NAME: Nombre del equipo.
- PLAYER_NAME: Nombre del jugador.
- POSITION_GROUP y POSITION: Grupo de posición y posición del jugador.
- HOME_TEAM, AWAY_TEAM: Equipos local y visitante.
- GAME_DATE: Fecha del juego (Mes-Día-Año).
- EVENT_TYPE: Resultado del tiro (Acertado o Fallado).
- SHOT_MADE: Resultado en formato booleano (Verdadero o Falso).
- SHOT_TYPE y ACTION_TYPE: Tipo de tiro y descripción (ej. bandeja, mate, lanzamiento).
- BASIC_ZONE: Zona de la cancha desde donde se realizó el tiro.
- ZONE_NAME, ZONE_ABB: Zona de la cancha por nombre y abreviatura.
- LOC_X, LOC_Y: Coordenadas de cada tiro en la cancha.
- SHOT_DISTANCE: Distancia del tiro en pies desde el aro.

Ejercicios:

1. Crea un treemap que muestre la cantidad de tiros en cada zona de la cancha, identificando visualmente las zonas donde se realizaron más intentos.
2. Genera un mapa de la cancha utilizando las coordenadas LOC_X y LOC_Y para representar cada tiro realizado por el equipo ‘Boston Celtics’. Usa un color para los tiros acertados y otro para los tiros fallados.
3. Representa la cantidad de tiros acertados en cada una de las zonas de la cancha en un gráfico de torta. Este gráfico debe mostrar la efectividad en cada zona de forma clara.