

Práctica Nro. 5

Análisis de Datos y Visualización

Publicación: 27/10/2025

Finalización: 03/11/2025

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.

Kaggle (kaggle.com/datasets)

- Descripción: Ofrece acceso a miles de datasets públicos aportados por usuarios y organizaciones. Es ideal para análisis exploratorios, modelado predictivo y competencias de machine learning, con herramientas integradas para visualización y colaboración.
- Tipos de datos: Datos tabulares (CSV, JSON), imágenes (para visión computacional), texto (para procesamiento de lenguaje natural), series temporales (ventas, finanzas) y datasets especializados.

Datos Argentina (datos.gob.ar)

- Descripción: Portal oficial de datos abiertos del gobierno argentino, que centraliza información pública de diversas instituciones nacionales. Está diseñado para promover la transparencia y facilitar análisis en políticas públicas, economía y desarrollo social.
- Tipos de datos: Datos socioeconómicos (censos, empleo, inflación), estadísticas de salud (vacunación, enfermedades), datos geográficos (mapas, infraestructura), información ambiental (clima, recursos naturales) y registros gubernamentales (presupuestos, licitaciones).

GitHub (github.com/datasets)

- Descripción: Alberga repositorios de datos abiertos bajo proyectos como "awesome-public-datasets" o colecciones específicas de datasets. Incluye datos relacionados con código, ciberseguridad, redes y tendencias tecnológicas, y es un recurso dinámico para la comunidad de desarrolladores.
- Tipos de datos: Datos de código fuente (repositorios, commits), datasets de ciberseguridad (logs de ataques, vulnerabilidades), métricas de uso de software (APIs, tendencias de lenguajes de programación), datos de redes sociales extraídos de plataformas (como X), y datasets para análisis de grafos o minería de datos en entornos tecnológicos.

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. Ejemplo: El presupuesto anual de un municipio publicado en un PDF en el sitio web oficial del gobierno.

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: Un dataset en formato CSV con los horarios y rutas de transporte público de la Ciudad de Buenos Aires, disponible en el portal Datos Argentina (datos.gob.ar).

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: observar cómo varían los niveles de contaminantes a lo largo del tiempo (por día, mes o año).
- Análisis geográfico: comparar los niveles de contaminación entre distintas ciudades o regiones del país.
- Identificación de tendencias y patrones: detectar picos o descensos de contaminación asociados a estaciones del año, tráfico o eventos específicos.
- Evaluación de políticas ambientales: medir si disminuye la contaminación tras la implementación de regulaciones o restricciones.
- Visualización de datos: mediante mapas de calor, gráficos de líneas y barras para comunicar de forma clara los resultados a los ciudadanos.

5. ¿Cuáles son las condiciones de la licencia creative commons?

- Atribución (BY): se debe dar crédito al autor o fuente original.
- No comercial (NC): el material no puede ser usado con fines comerciales (solo aplica si el autor lo especifica).
- Sin obras derivadas (ND): no se permite modificar la obra (solo si el autor aplica esta restricción).
- Compartir igual (SA): si se transforma o remezcla, debe compartirse bajo la misma licencia.

6. ¿Qué significa tener datos IA-ready?

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. Según la clase, esto significa que cumplen con ciertos requisitos:

- Completitud y calidad: los datos no deben tener errores ni valores faltantes.
- Estructura adecuada: deben estar en formatos homogéneos, con metadatos claros y consistentes.
- Contexto suficiente: deben incluir información contextual que permita comprender su origen y uso.
- Ausencia de sesgos: los datos deben representar de manera justa la realidad.
- Licencias claras: deben especificar las condiciones de uso, especialmente para fines de IA.

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

Principios:

- Findable (Encontrables): los datos deben tener metadatos descriptivos y un identificador único.
- Accessible (Accesibles): deben poder descargarse o consultarse de forma abierta y estandarizada.
- Interoperable (Interoperables): deben usar formatos y vocabularios comunes para poder combinarse con otros datos.
- Reusable (Reutilizables): deben incluir licencias y documentación que faciliten su reutilización.
- Responsible (Responsables): nuevo principio añadido que promueve un uso ético y seguro de los datos en IA.

Requisitos:

- Etiquetado exhaustivo.

- Documentación completa.
- Homogeneidad de estándares y metadatos.
- Cobertura suficiente para evitar sesgos.
- Licencias que regulen su uso en IA.

Objetivo: Evitar que los sistemas de IA se construyan sobre datos incompletos, sesgados o sin control ético.

PARTE II. Visualización de Datos

Responda las siguientes preguntas sobre visualización de datos:

1. Explique que es una medida y una dimensión

Las medidas son datos numéricos cuantitativos. Las dimensiones son datos cualitativos, como un nombre o una fecha. Las dimensiones agrupan o clasifican, mientras que las medidas cuantifican.

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

Los *campos continuos* pueden contener un número infinito de valores. Puede tratarse de un rango de valores, como las ventas dentro de un determinado intervalo de fechas o cantidades.

Los *campos discretos* contienen un número finito de valores, como País, Provincia o Nombre de cliente.

El discreto se cuenta, el continuo se mide.

3. ¿Por qué es importante preparar los datos?

- Garantiza la precisión: evita errores, duplicados y valores faltantes.
- Facilita el análisis: los datos limpios permiten crear gráficos correctos y comparables.
- Mejora la comunicación visual: los gráficos reflejan correctamente los patrones y tendencias reales.
- Evita sesgos o malas interpretaciones: un dato mal cargado puede alterar completamente el mensaje visual.

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?

Se podría usar un gráfico de barras agrupadas o de columnas múltiples. Permiten comparar valores numéricos entre categorías (regiones) y observar variaciones a lo largo del tiempo.

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 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.

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?

Usaría un gráfico de dispersión, este tipo de gráfico sirve para mostrar la correlación entre dos variables numéricas. Permite observar si hay relación entre el precio del producto (eje X) y la puntuación de los clientes (eje Y).

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?

Usaría un gráfico de barras o torta. Si se busca comparar cantidades exactas, conviene usar barras. Si se desea mostrar porcentajes o proporciones del total, se usa circular.

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?

Utilizaría el gráfico de línea ya que este muestra las relaciones de los cambios en los datos en un período de tiempo, facilitando la identificación de tendencias. Se podría ver el aumento, disminución o estabilidad de las temperaturas a lo largo de los años. En el eje X se colocará la fecha y en el eje Y el valor promedio de temperatura en cada fecha.

- b) ¿Cuáles tablas son relevantes para presentar el análisis?

TemperaturasPromedio (#registroProm, fecha, valorProm)

Para el análisis es relevante esa tabla ya que:

- Con la tabla TemperaturasPromedio obtenemos la fecha (fecha) y el valor promedio (valorProm) de temperatura en dicha fecha.

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).

f. Dado el siguiente esquema de la base de datos:

CategoriaEspecie (#cat_especie, cat_nombre)

EspecieIdentificada (#especie, #cat_especie, nombre_especie, cantidad)

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

Utilizaría un gráfico circular (torta) ya que se podría mostrar cómo esas cinco categorías se comparan en porcentaje entre ellas y el total. Como son cinco las categorías se entendería el gráfico.

- d) De todas las tablas propuestas, indicar cuál o cuáles son relevantes para presentar el análisis

CategoriaEspecie (#cat_especie, cat_nombre)

EspecieIdentificada (#especie, #cat_especie, nombre_especie, cantidad)

Las dos tablas (CategoriaEspecie y EspecieIdentificada) son relevantes para el análisis, ya que la primera identifica las categorías (cat_nombre) y la segunda proporciona los datos numéricos (cantidad) necesarios para representar la proporción en el gráfico circular. Ambas se relacionan mediante el campo cat_especie, que permite vincular cada especie con su categoría correspondiente.

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.

Para el primer punto utilizaría un gráfico de barras ya que permite comparar qué productos se venden más por categoría en la sucursal. En el eje X se representarían las categorías de productos, y en el eje Y la cantidad vendida correspondiente a cada categoría.

Para el segundo punto utilizaría un gráfico de líneas ya que muestra las relaciones de los cambios en los datos en un período de tiempo, en este caso la evolución mensual de los ingresos, viendo los ascensos y descensos en las ventas. En el eje X se representarían los meses, y en el eje Y el total de ingresos correspondiente a cada mes.

Tablas importantes para el primer punto:

Producto (id_producto, nombre_producto, desc_producto, precio_unit, categoria)

Item_Venta (id_venta, id_producto, cant)

Las tablas Producto e Item_Venta son relevantes para el análisis, ya que la primera (Producto) identifica las categorías de los productos (atributo categoria), mientras que la segunda (Item_Venta) proporciona los datos numéricos (atributo cant) necesarios para calcular la cantidad total de productos vendidos por cada categoría. Ambas se relacionan mediante el campo id_producto, que permite vincular cada ítem vendido con su categoría correspondiente.

Tablas importantes para el segundo punto:

Sucursal (id_sucursal, ubicación, cant_empleados)

Venta (id_venta, fecha_venta, id_sucursal, monto_total)

Las tablas Venta y Sucursal son relevantes para el análisis, ya que la primera (Venta) contiene los datos económicos de cada operación, como el monto_total y la fecha_venta, necesarios para calcular los ingresos mensuales, mientras que la segunda (Sucursal) permite identificar la sucursal específica (atributo id_sucursal) sobre la cual se realizará el análisis. Ambas se relacionan a través del campo id_sucursal, que vincula cada venta con la sucursal donde se realizó.

PARTE III: Graficando con Tableau

En esta sección, utilizarás [Tableau Public](#) para explorar y visualizar un datasets sobre el uso de dispositivos móviles y comportamiento del usuario. El archivo con dicho dataset estará adjunto a esta práctica, y tu objetivo será importar este en Tableau y generar gráficos que representen 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.

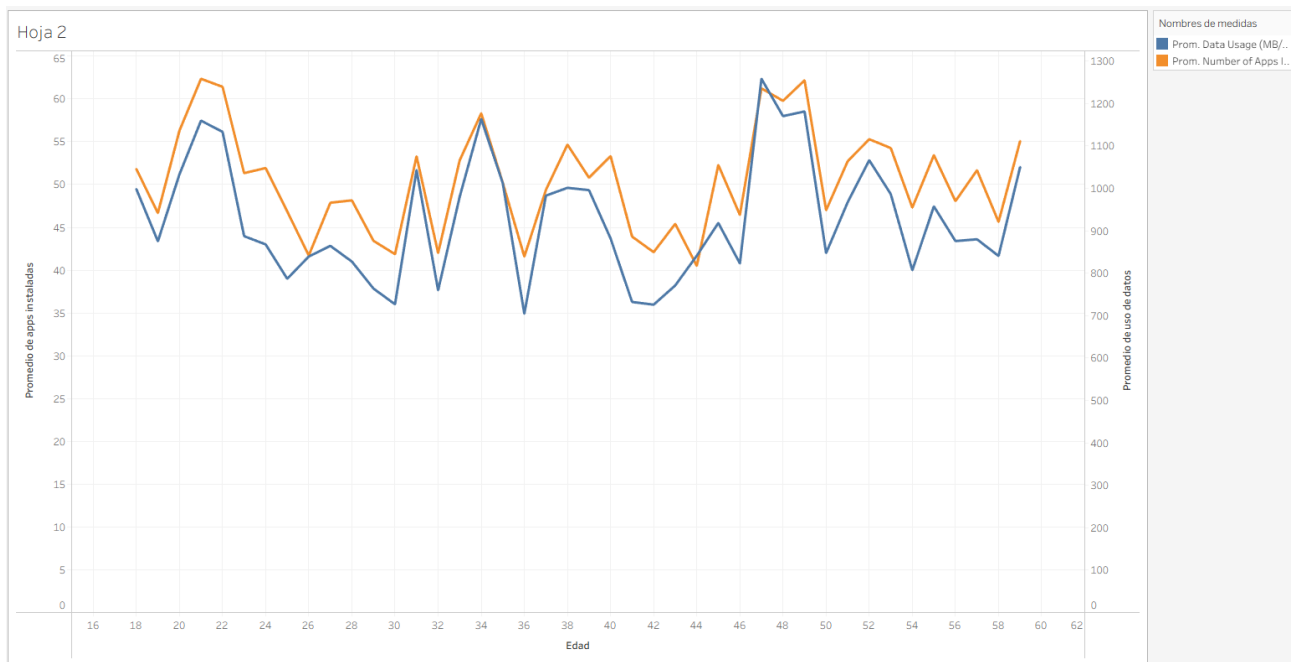
Ejercicios:

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.

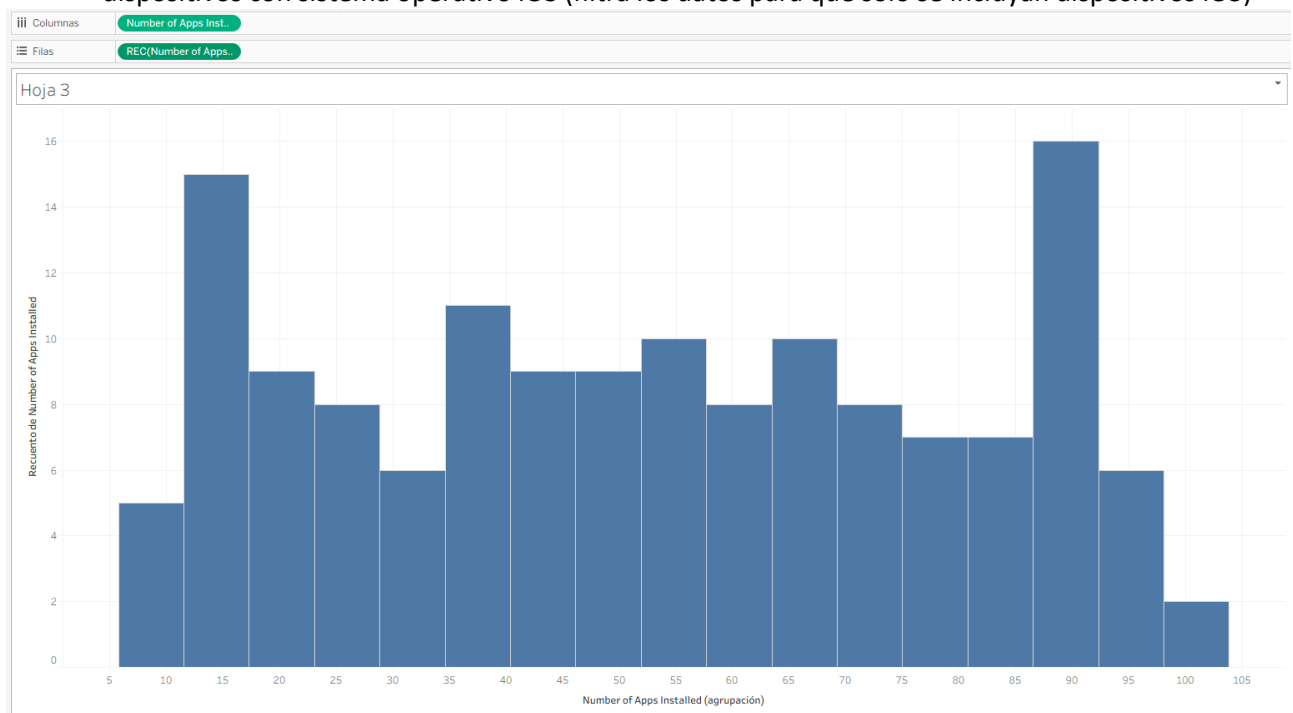
Usuarios para cada modelo según sistema operativo



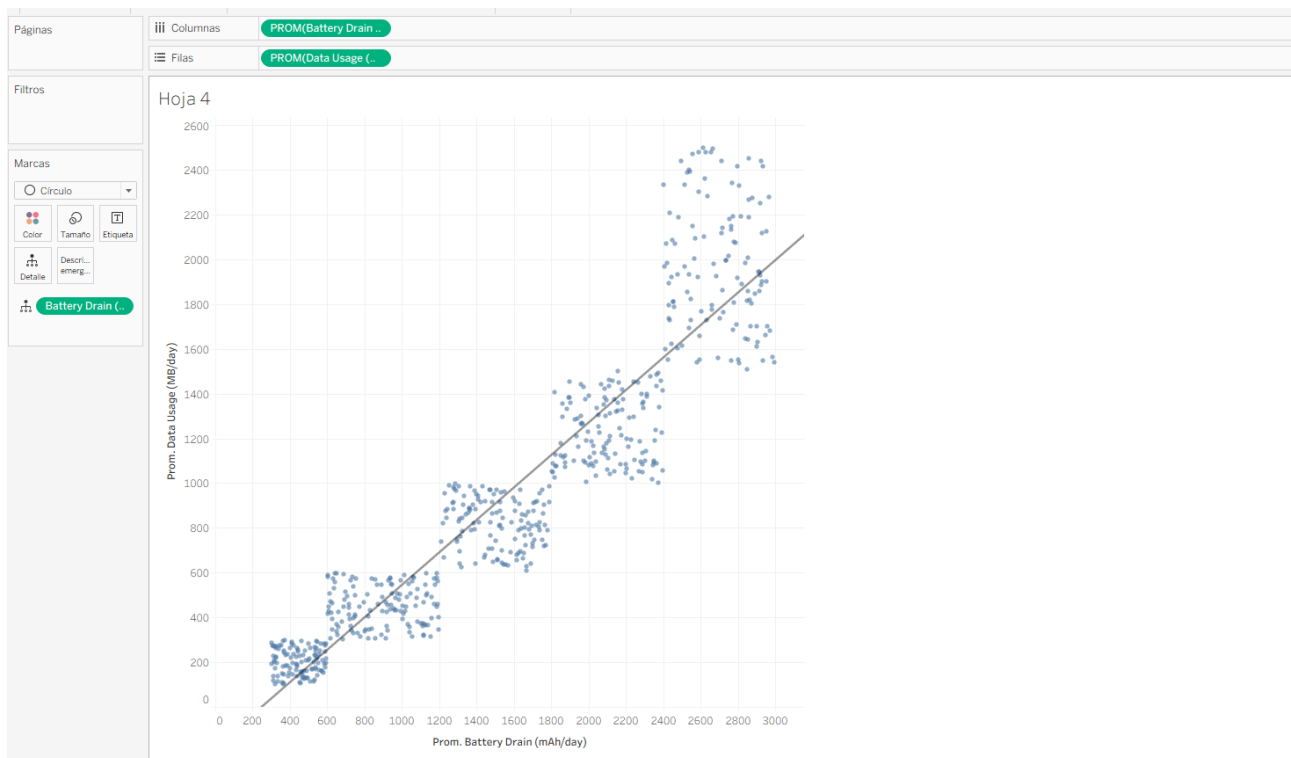
2. Muestra una línea que refleje el promedio de aplicaciones instaladas y otra línea para el promedio de datos consumidos, ambos valores en función de la edad del usuario.



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.



Hoja 4

