Modelización y análisis exploratorio para visualización

Tema 4.
Bloque 2: Preprocesamiento y Modelización de Datos para Visualización

unir LA UNIVERSIDAD EN INTERNET

Profesor: Luis F. Ruiz

¿Qué veremos hoy?

01 **Introducción y objetivos.**

02 **Modelización descriptiva y predictiva.**

03 **Análisis exploratorio.**

04 **Agrupamiento, segmentación y reducción de dimensionalidad.**

05 **Aplicación en entornos BI.**

4.1. El desafío del análisis de datos en entornos profesionales

El Reto Analítico

En los entornos profesionales nos enfrentamos constantemente a preguntas críticas que requieren respuestas fundamentadas: ¿Qué factores impulsan realmente nuestras ventas? ¿Qué segmentos de clientes comparten necesidades comunes? ¿Cómo podemos predecir comportamientos futuros?

La respuesta a estos interrogantes se encuentra en la combinación estratégica de la modelización de datos y el análisis visual exploratorio, dos pilares fundamentales del análisis moderno.

Nuestro Propósito

Transformar grandes volúmenes de información en visualizaciones comprensibles que nos permitan validar hipótesis, identificar patrones ocultos y comunicar resultados con claridad y rigor profesional.

Este tema te capacitará para aplicar técnicas exploratorias de segmentación visual, combinando rigor estadístico con representación gráfica efectiva para identificar patrones complejos en tus datos.

- ”Ver el modelo”: que cada técnica vaya acompañada de una visualización interpretative.
- Entender conceptos, no solo nombres de técnicas.

Objetivos de aprendizaje

01

Fundamentos de Modelización

Comprender los principios de la modelización predictiva y descriptiva, orientada específicamente a la creación de visualizaciones efectivas que comuniquen insights de forma clara.

02

Análisis Exploratorio Avanzado

Aplicar técnicas de análisis exploratorio para identificar patrones significativos, relaciones entre variables y segmentos relevantes en conjuntos de datos complejos.

03

Técnicas de Agrupamiento

Implementar procesos de clustering y reducción de dimensionalidad (PCA, t-SNE) directamente en entornos de Business Intelligence para simplificar datos complejos.

04

Visualizaciones Interpretativas

Integrar visualizaciones que faciliten la interpretación de resultados y apoyen la toma de decisiones estratégicas basadas en evidencia empírica.

4.2. Modelización Descriptiva: explicando el comportamiento de los datos

La modelización descriptiva busca explicar el comportamiento actual de los datos y determinar las relaciones existentes entre variables. Es fundamental para comprender qué está ocurriendo y por qué antes de intentar predecir el futuro.

Análisis de Correlación

Objetivo: Desvelar si variables como la satisfacción del cliente están ligadas al tiempo de entrega o a la calidad de la posventa.

Visualización: Diagramas de dispersión que muestran la fuerza y el sentido de la relación entre variables mediante la distribución de puntos.

Regresión Lineal

Objetivo: Estimar el impacto cuantitativo de una o varias variables independientes sobre una variable dependiente.

Visualización: Diagramas de dispersión enriquecidos con líneas de tendencia y bandas de confianza. Power BI facilita modelos exploratorios rápidos.

Análisis de Varianza (ANOVA)

Objetivo: Evaluar si existen diferencias estadísticamente significativas entre grupos distintos, como segmentos de clientes.
Visualización: Gráficos de cajas o diagramas de violín que ilustran distribuciones y variabilidad en cada grupo.

4.2. Modelización Predictiva: anticipando comportamientos futuros

La modelización predictiva utiliza datos históricos para estimar comportamientos futuros, permitiendo a las organizaciones anticiparse a tendencias y tomar decisiones proactivas.

Regresión Logística

Predice variables categóricas binarias, como la probabilidad de abandono de un cliente o la conversión de un lead. Se visualiza mediante gráficos de barras que muestran probabilidades estimadas o mapas de calor que identifican zonas de riesgo.

Árboles de Decisión

Ofrecen una representación jerárquica y transparente de las reglas que conducen a una predicción. Son especialmente apreciados en entornos de negocio porque los diagramas de árbol permiten seguir el razonamiento del modelo paso a paso, comunicando la lógica predictiva incluso a audiencias no técnicas.

- Lineal ➔ valores continuos (p.ej. ventas).
- Logística ➔ categorías vía probabilidad (p.ej. abandono SÍ/NO).

4.3. Análisis Exploratorio: descubriendo lo inesperado

El Análisis Exploratorio de Datos (EDA) no busca confirmar hipótesis preestablecidas, sino descubrir patrones y anomalías que pueden generar nuevas preguntas de negocio y abrir líneas de investigación previamente no consideradas.

Iteración continua

El proceso no es lineal; requiere reformular preguntas conforme aparecen hallazgos. Cada descubrimiento puede abrir nuevas vías de exploración.

Flexibilidad metodológica

Adaptar herramientas y técnicas a la naturaleza cambiante de los datos. No existe una única forma correcta de explorar.

Enfoque visual primero

La representación gráfica es el núcleo del análisis exploratorio. Ver los datos es comprenderlos mejor.

FASES

Evaluación de Calidad

Detectar valores ausentes, duplicados o inconsistentes mediante histogramas y gráficos de dispersión iniciales.

Reducción de Complejidad

Agrupar categorías o normalizar escalas para facilitar comparaciones y la identificación de relaciones significativas.

~~Confirmar hipótesis previas~~ vs **Descubrir patrones inesperados**

4.3. Visualización para la Exploración de Datos

La visualización traduce los resultados del análisis exploratorio en argumentos claros y comprensibles para diferentes audiencias, desde analistas técnicos hasta directivos que toman decisiones estratégicas.

Diagramas de Dispersión

Esenciales para evaluar correlaciones, detectar concentraciones de datos y resaltar valores atípicos (outliers) que pueden indicar oportunidades o problemas.

Matrices de Correlación

Muestran cómo se relacionan múltiples variables simultáneamente. Codifican la fuerza y el sentido mediante colores, facilitando la comprensión a perfiles no técnicos.

Exploración Dinámica

Segmentaciones y filtros permiten observar cómo cambian las visualizaciones en tiempo real, estimulando hipótesis.

Power BI

Utiliza segmentaciones dinámicas y filtros interactivos para observar cómo cambian las visualizaciones en tiempo real, permitiendo la exploración guiada por la curiosidad y estimulando la generación continua de nuevas hipótesis de análisis.

Tableau

Su función *Show Me* sugiere automáticamente representaciones adecuadas según el tipo de dato seleccionado, agilizando significativamente la navegación inicial y fomentando la exploración intuitiva incluso para usuarios menos experimentados.

4.4. Clustering: descubriendo Estructuras Latentes en los datos

El clustering o agrupamiento agrupa observaciones con características similares para identificar estructuras latentes que no son evidentes a simple vista, revelando segmentos naturales en los datos.

K-Means

Técnica que asigna cada punto de datos al grupo más cercano a su centroide. Es especialmente útil para segmentar clientes o productos con comportamientos homogéneos.
Visualización: Diagramas de dispersión donde cada clúster se diferencia mediante un color distinto, facilitando la interpretación inmediata de los segmentos identificados.

- Agrupa observaciones según su distancia al **centroide** del clúster.
- Muy usado para segmentar clientes, productos o regiones.
- *Visualización*: scatter plot donde cada clúster se pinta con un **color diferente**; se pueden mostrar también los centroides.
- *Ejemplo*: segmentación de clientes por ingresos y gasto mensual

Validación de Segmentos

Es fundamental validar los resultados del clustering con métricas de calidad objetivas, como el **Coeficiente de Silueta**, que mide tanto la cohesión interna de los grupos como la separación entre ellos. Un buen clustering muestra grupos bien definidos y claramente separados.

Clustering: Descubriendo Estructuras Latentes en los datos

Las métricas agregadas no muestran toda la historia; necesitamos descubrir **grupos con comportamientos similares**.

•Construye un árbol de relaciones (dendrograma) que muestra cómo se van fusionando grupos.

•Ayuda a decidir el número óptimo de segmentos.

•Visualización: dendrograma con el “corte” a cierta altura define los clústeres.

Clustering Jerárquico

Construye un árbol de relaciones (dendrograma) mostrando cómo se agrupan los elementos de forma progresiva. Es particularmente útil para determinar el número óptimo de segmentos.

Ventaja: Permite visualizar todo el proceso de agrupamiento en un único gráfico, desde elementos individuales hasta grupos completos.

Coeficiente de Silueta

Es una métrica clave para ver la Calidad del clúster. Mide la cohesión interna y separación entre grupos.

4.4. Reducción de Dimensionalidad: simplificando la complejidad

La reducción de dimensionalidad permite simplificar conjuntos de datos con muchas variables (alta dimensionalidad) proyectándolos en un espacio de dos o tres dimensiones, preservando la información más relevante y facilitando su visualización efectiva.

PCA: Análisis de Componentes Principales

Naturaleza: Técnica **lineal** que transforma variables originales en nuevos componentes no correlacionados que capturan la máxima varianza de los datos.

Objetivo: Reducir dimensiones manteniendo la mayor cantidad de información posible.

Uso en Visualización: Representación efectiva de tendencias y agrupaciones en gráficos 2D o 3D, facilitando la comprensión de estructuras complejas.

Una matriz de calor que muestre la contribución de las variables originales a los componentes principales facilita enormemente la interpretación de qué factores explican la mayor parte de la varianza.

t-SNE: Preservando Relaciones Locales

Naturaleza: Técnica **no lineal** que proyecta datos de alta dimensión preservando las relaciones de proximidad local entre observaciones.
Objetivo: Mantener los puntos similares cerca y los diferentes alejados en el espacio reducido.

Uso en Visualización: Especialmente eficaz para visualizar clústeres naturales en problemas de alta complejidad, como análisis de comportamiento digital o segmentación avanzada de clientes.
Es ideal cuando las relaciones entre datos son no lineales y PCA no captura adecuadamente la estructura subyacente.

- **PCA** es la técnica adecuada para “proyectar datos multivariantes en dos dimensiones manteniendo la varianza”.
- **t-SNE** es “no lineal” y “preserva relaciones locales”.

4.5. Aplicación Práctica en entornos de BI

Los entornos de BI modernos han democratizado estas técnicas avanzadas, haciéndolas accesibles mediante interfaces visuales intuitivas que no requieren conocimientos profundos de programación.

Power BI: Clustering Automático

Permite realizar clustering automático directamente en gráficos de dispersión. La herramienta calcula y colorea los segmentos, permitiendo al usuario definir el número de clústeres o dejar que el algoritmo lo decida automáticamente.
Soporta la creación de líneas de tendencia (regresiones lineales) en gráficos para validar hipótesis rápidamente, incluso mostrando intervalos de confianza.

Tableau: Clustering Intuitivo

Ofrece clustering simplemente arrastrando la funcionalidad sobre la hoja de trabajo. Tableau calcula automáticamente los grupos y genera un panel de resumen con estadísticas clave de cada segmento identificado.

Permite añadir tendencias (lineales, cuadráticas) y modelos de predicción desde el menú Análisis, facilitando la exploración predictiva básica.

> **Ejemplo Práctico:** Segmentación de clientes en clústeres según ingresos y gasto mensual, visualizando los centroides de cada grupo en un gráfico de dispersión interactivo. Cada segmento revela patrones de comportamiento específicos que pueden guiar estrategias de marketing personalizadas.

RESUMEN Tema 4

- **Modelización descriptiva**: explica **relaciones y patrones históricos** (correlación, regresión lineal, ANOVA) usando visualizaciones como dispersión, cajas, líneas de tendencia.
- **Modelización predictiva**: anticipa resultados futuros (regresión logística, árboles de decisión) y se comunica con gráficos de probabilidad y diagramas de árbol.
- **EDA visual**: orientado a **descubrir** patrones, no a confirmar hipótesis. Basado en iteración, flexibilidad y enfoque visual (dispersión, cajas, matrices de correlación).
- **Clustering y segmentación**: k-means y clustering jerárquico para identificar grupos; calidad evaluada con **coeficiente de silueta**. Se interpretan con scatterplots coloreados y dendrogramas.
- **Reducción de dimensionalidad**:
  - **PCA** (lineal, maximiza varianza; adecuada para proyectar datos multivariantes en 2D).
  - **t-SNE** (no lineal, preserva relaciones locales).
  - Se complementan con matrices de calor de cargas de variables.
- **BI y automatización visual**: Power BI y Tableau ofrecen clustering automático, líneas de tendencia y sugerencias de visualización (Show Me).

Checklist Tema 4

- ✓ **Modelización descriptiva** ➔ explicar relaciones y patrones históricos.
- ✓ **PCA** ➔ técnica adecuada para proyectar datos multivariantes manteniendo varianza.
- ✓ **Árbol de decisión** ➔ facilita la interpretación de reglas lógicas y condiciones de decisión.
- ✓ **t-SNE** ➔ método no lineal que preserva relaciones locales, ideal para visualizar clústeres complejos.
- ✓ **Clustering en Power BI** ➔ función de agrupación automática con colores diferenciados en un gráfico.
- ✓ **EDA** ➔ no se centra en confirmar hipótesis previas.
- ✓ **Calidad de clustering** ➔ coeficiente de silueta.
- ✓ **Visualización de PCA** ➔ matriz de calor de cargas de variables en componentes.
- ✓ **Regresión logística** ➔ predice variables categóricas mediante probabilidad.
- ✓ **Tableau – Show Me** ➔ sugiere automáticamente tipos de visualización.

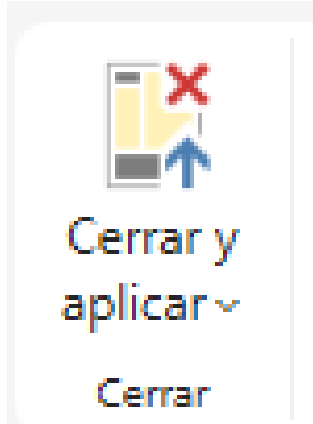
Ejemplo con Power BI

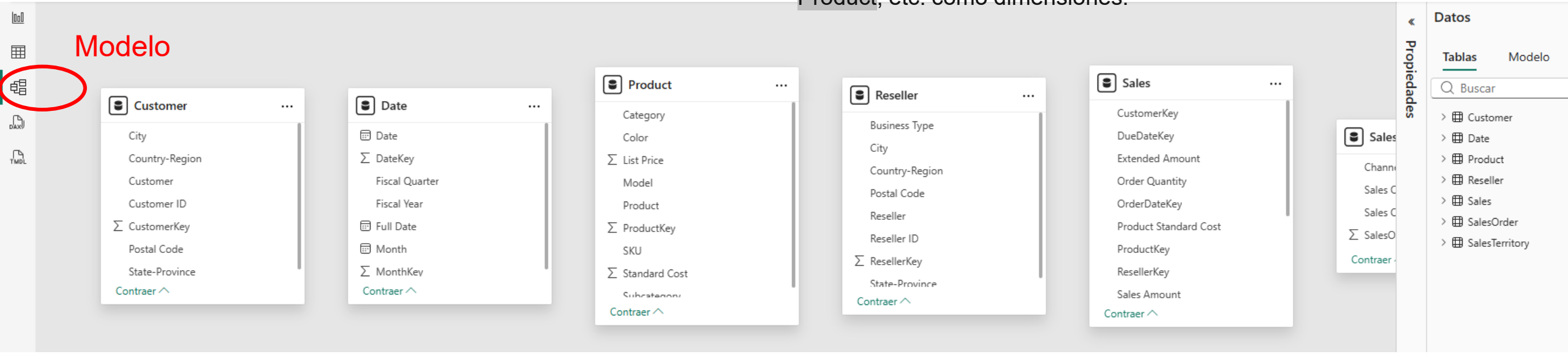
¿Qué vamos a hacer? (¿Qué queremos ver?)

- **Visual**: Gráfico de dispersión (scatter).
- Cada punto: un cliente (Customer).
- Eje X: **ventas totales del cliente**.
- Eje Y: **nº de pedidos del cliente**.
- Clústeres: creados con **"Automatically find clusters / Encontrar clústeres automáticamente"** en el scatter (Power BI usa k-means).

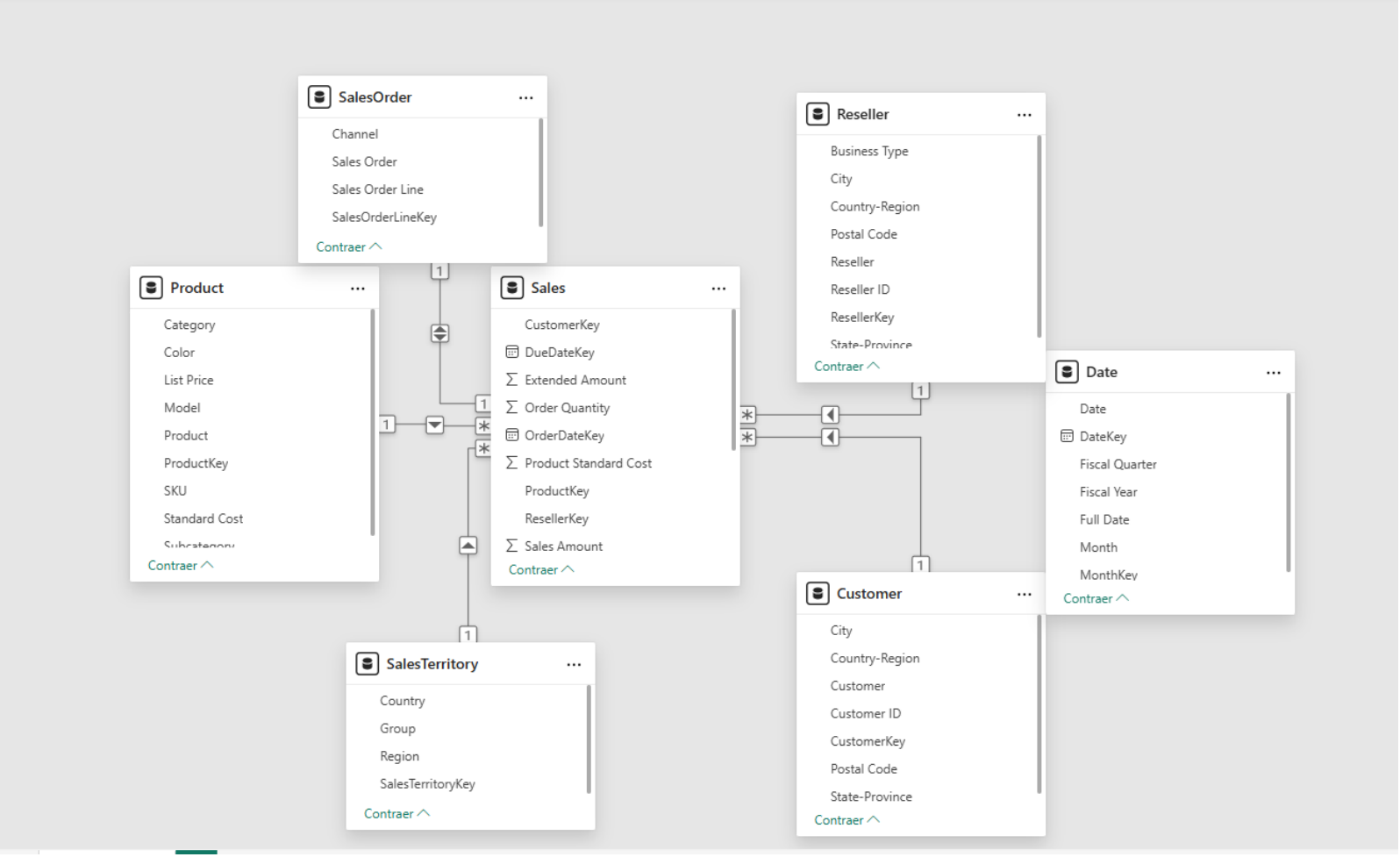
Ejemplo con Power BI

1. En el **navegador**:
   - Marcar las tablas: Customer, Date, Product, Reseller, Sales, SalesTerritory (todas salvo las hojas _data).
   - Pulsar **Transformar datos / Transform Data**.
2. En Power Query:
   - Verificar que las columnas clave (CustomerKey, SalesOrderLineKey, SalesAmount, etc.) tienen el tipo correcto (Número entero / Decimal).
3. Pulsar **Cerrar y aplicar (Close & Apply)**.
4. Power BI montará el **modelo en estrella** con Sales en el centro y Customer, Date, Product, etc. como dimensiones.





Modelo en estrella



Crear las medidas necesarias

Crear 3 medidas:

- Total Sales – ventas totales
- Order Count – nº de líneas de pedido (o pedidos)
- Avg Sales per Order – valor medio por línea/pedido

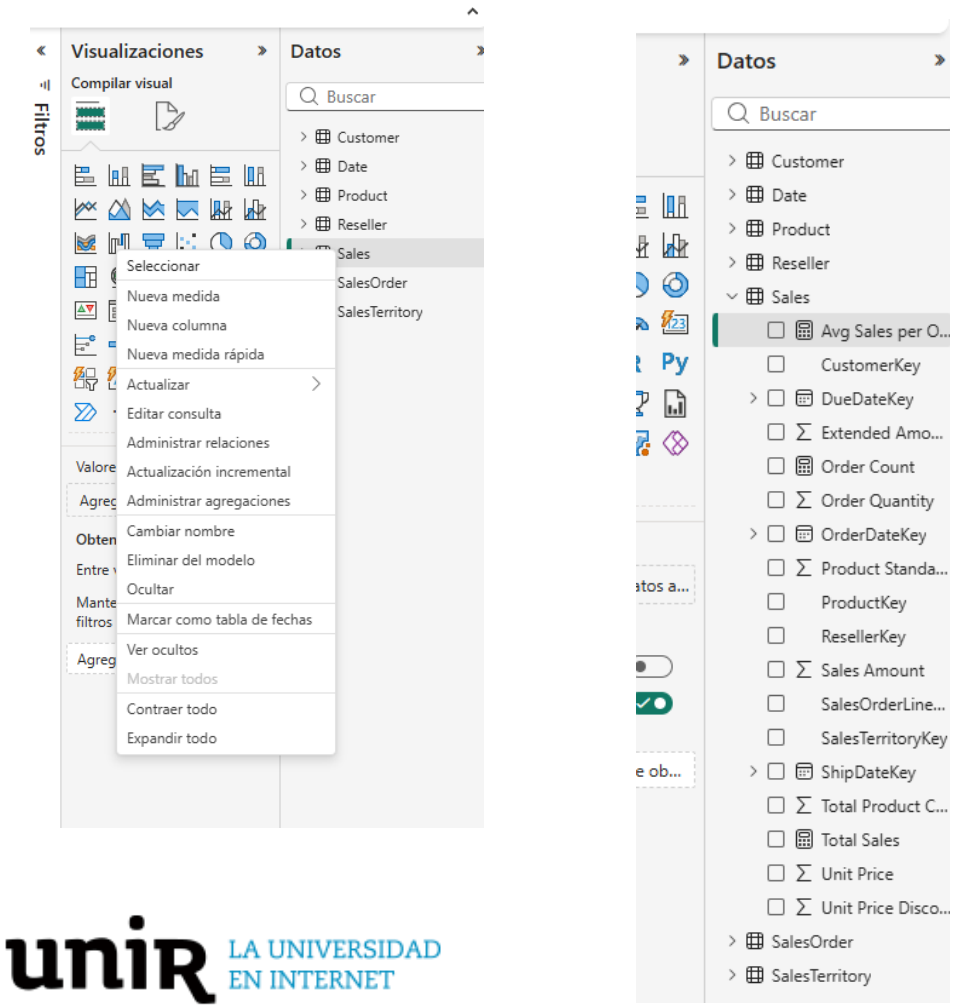


Tabla **Sales**:

1. En el panel **Campos/Fields**, clic dcho. sobre la tabla Sales

   ➔ **Nueva medida / New measure**.

2. Escribir (DAX):

```
Total Sales = SUM ( Sales[Sales Amount] )
```

3. Otra medida (DAX):

```
Order Count = DISTINCTCOUNT ( Sales[SalesOrderLineKey] )
```

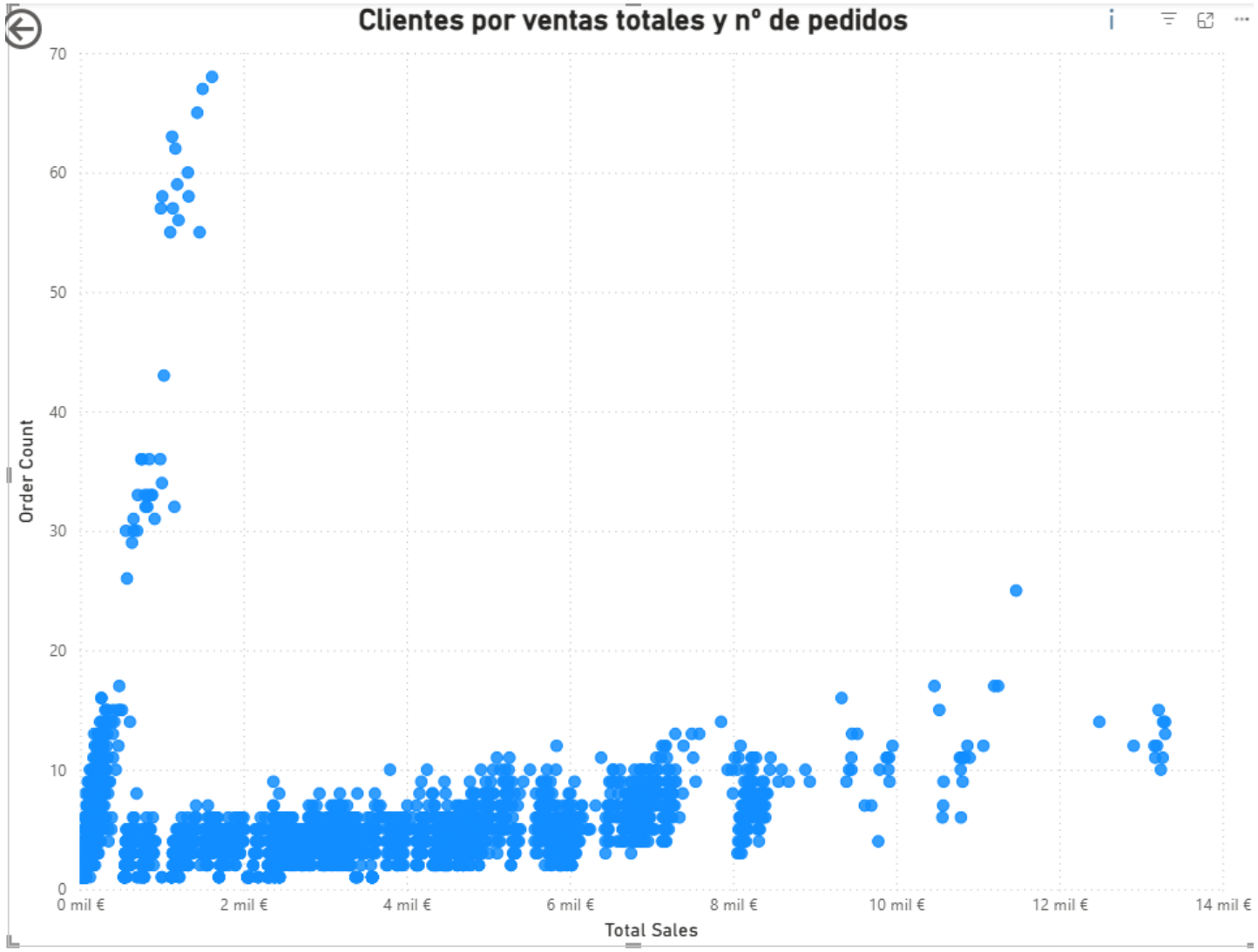
ó

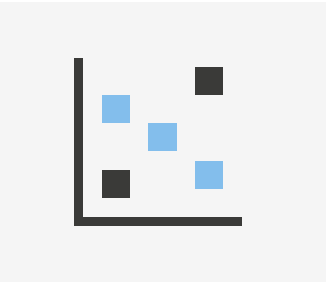
```
Order Count = DISTINCTCOUNT ( Sales[SalesOrderNumber] )
```

4. Valor medio por pedido:

```
Avg Sales per Order =
DIVIDE ( [Total Sales], [Order Count] )
```

Construir el scatter de clientes





Valores

CustomerKey

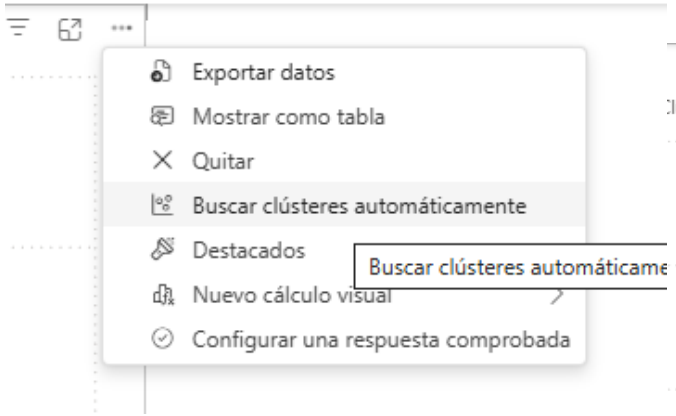
Eje X

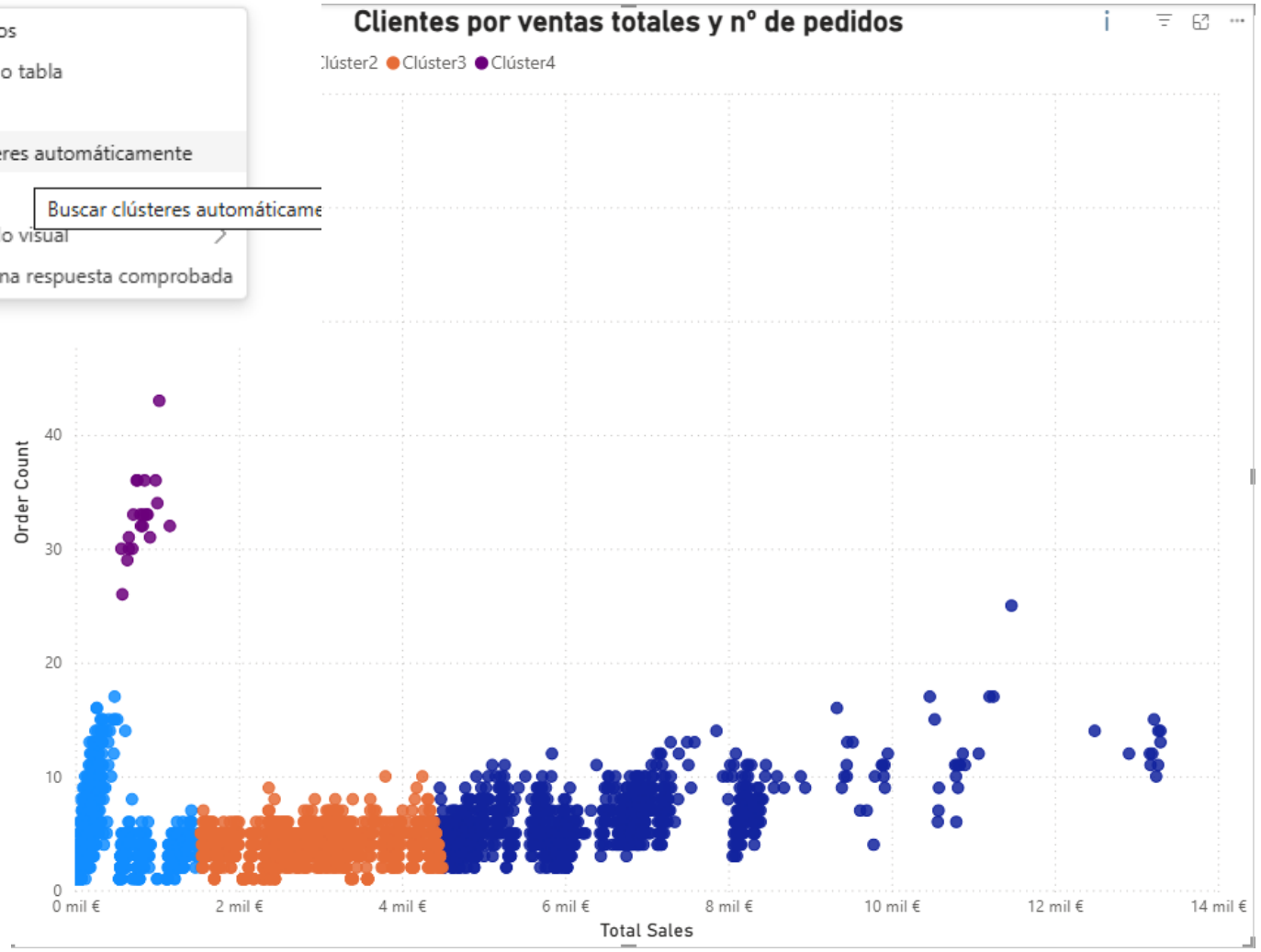
Total Sales

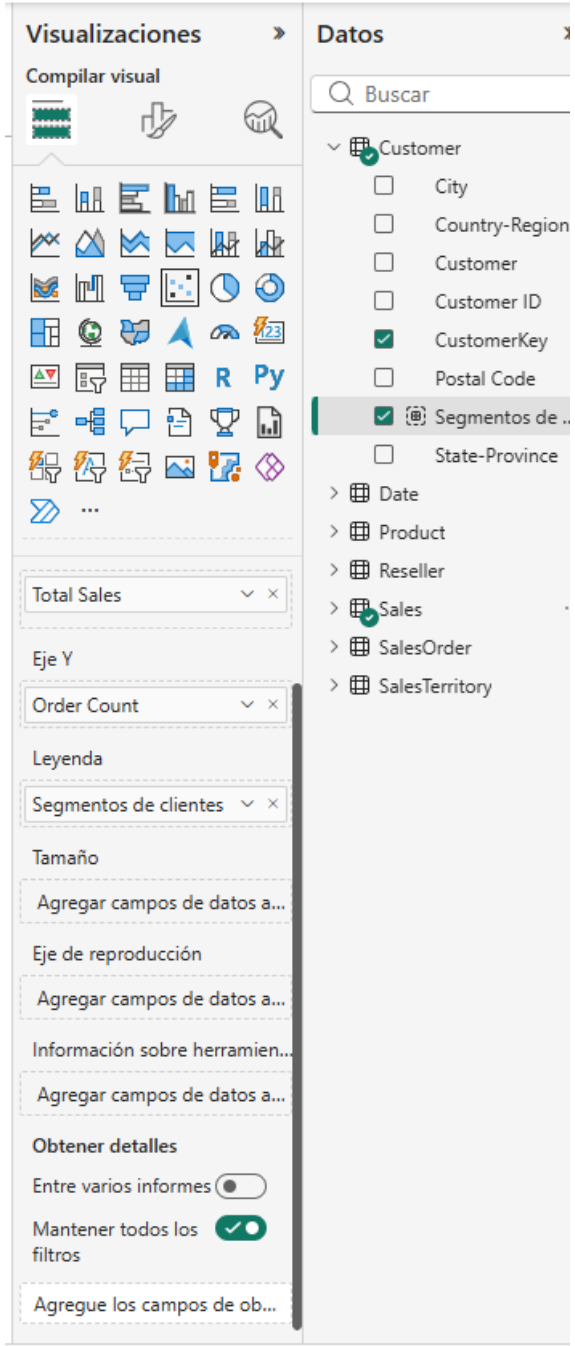
Eje Y

Order Count

Aplicar clustering automático en el scatter







Pasos del ejemplo

1. Cargamos el dataset.
2. Creamos medidas en Sales:
   - Total Sales = SUM(Sales[SalesAmount])
   - Order Count = DISTINCTCOUNT(Sales[SalesOrderLineKey])
3. Crear un **scatter**:
   - X = [Total Sales]
   - Y = [Order Count]
   - Details = Customer[CustomerKey]
   - Tooltips = nombre de cliente, país, etc.
4. En el scatter → … → **Automatically find clusters** → Number of clusters = 4.
5. Añadir **segmentadores** por año y país.
6. Contar la historia de los 4 segmentos de clientes.

INTERPRETACIÓN

Cada punto es un cliente de AdventureWorks.

Eje Horizontal: ventas totales

Eje Vertical: número de pedidos que nos ha hecho.

➔ Nube de puntos ¿Segmentos claros?

Power BI encuentra automáticamente clústeres = grupos de clientes con comportamiento similar.

Quiero 4 clústeres:

- Clúster 1: clientes con muchas ventas y muchos pedidos → *nuestros clientes estratégicos*.
- Clúster 2: clientes con baja venta y pocos pedidos → *clientes de bajo valor*.
- Clúster 3: clientes con pocas ventas pero muchos pedidos pequeños → quizá compran cosas baratas o hacen muchos pedidos mínimos.
- Clúster 4: clientes con poca frecuencia pero pedidos muy altos → clientes de “golpe fuerte”, interesantes para acciones concretas.

Si filtro por año o por región…¿Cambia el tamaño de los clústeres? ¿Aparecen más clientes ‘estratégicos’ en cierto país?

Muchas gracias por vuestra atención

www.unir.net