

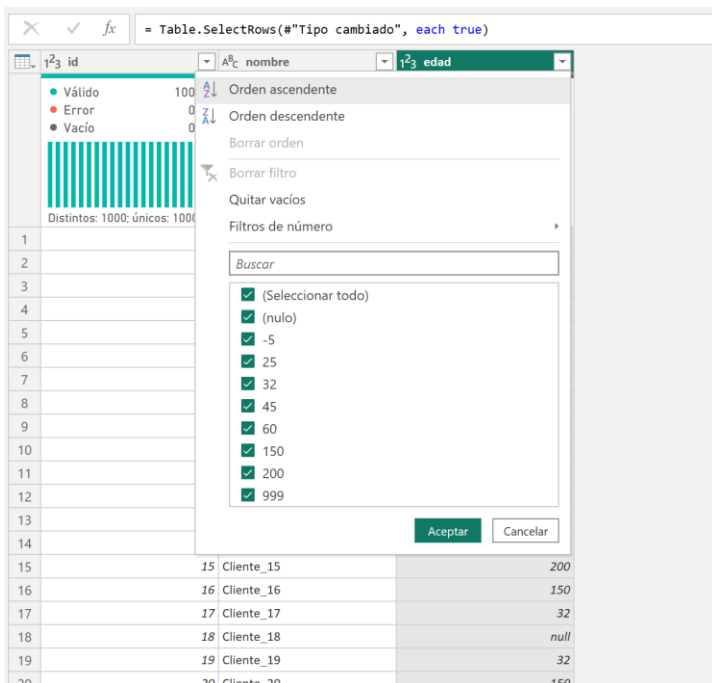
PRUEBA TÉCNICA DATA ANALYST

AIRAM KATIZA COBO SOLIS

Análisis de Datos

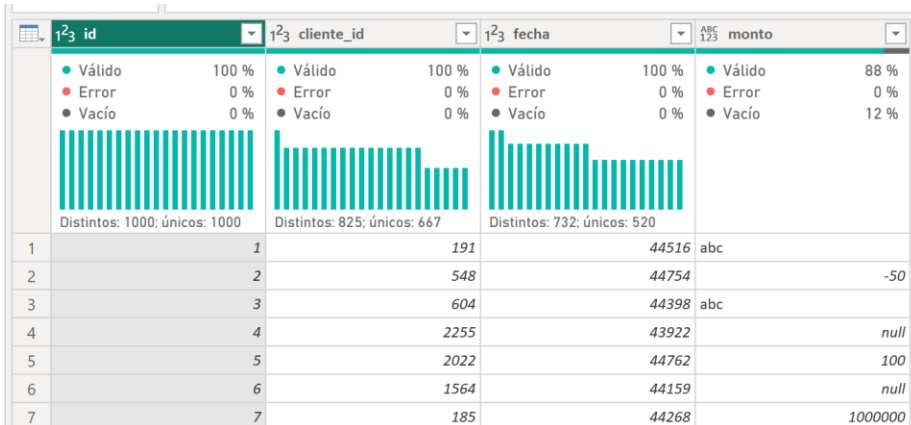
1. Se realiza una copia del dataset para realizar los procesos de limpieza y dejar una copia de los datos crudos
2. Se realiza un análisis exploratorio de los datos en POWER BI y Python: Dada la cantidad de datos se procede a trabajar con Excel y powerBI directamente.

LA TABLA DE CLIENTES TIENE:



Datos nulos en la columna edad o valores numéricos que no están de acuerdo con la edad de un usuario, por ejemplo (-5,150, 200).

LA TABLA DE COMPRAS:



id	cliente_id	fecha	monto
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			

Presenta fechas sin formato de datetime, y la columna monto tiene montos nulos, negativos y valores tipo string que no van asociados a la columna.

LA TABLA DEPARTAMENTOS

id	nombre
11	Ventas

La columna id presenta datos nulos, y los nombres de departamentos presentan datos nulos o valores tipo numérico que no concuerdan con el atributo.




TABLA EMPLEADOS

id	nombre	salario	departamento_id
1	Empleado_1	3000	61
2	Empleado_2	2000	38
3	Empleado_3	1000	22
4	Empleado_4	dosmil	69
5	Empleado_5	3000	23
6	Empleado_6	3000	28
7	Empleado_7	3000	36
8	Empleado_8	dosmil	40
9	Empleado_9	999999	44
10	Empleado_10	dosmil	35
11	Empleado_11	3000	15
12	Empleado_12	999999	64
13	Empleado_13	dosmil	64
14	Empleado_14	null	42
15	Empleado_15	-100	4

La tabla presenta valores de salario nulos, valores de salario como string y en departamento_id también presenta valores nulos.

MUESTRA LAS PRIMERAS 5 FILAS DE LOS DATOS CARGADOS Y LIMPIOS.

TABLA CLIENTES

   = Table.TransformColumnTypes("#Filas filtradas",{{"id", type text}})

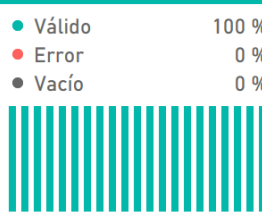
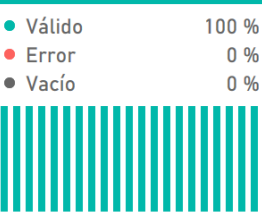
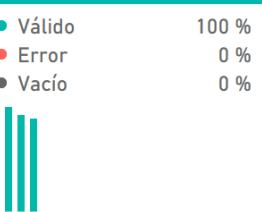
	A ^B _C id	A ^B _C nombre	1 ² ₃ edad
	 <p>Válido 100 % Error 0 % Vacío 0 %</p> <p>Distintos: 841; únicos: 841</p>	 <p>Válido 100 % Error 0 % Vacío 0 %</p> <p>Distintos: 841; únicos: 841</p>	 <p>Válido 100 % Error 0 % Vacío 0 %</p> <p>Distintos: 4; únicos: 0</p>
1	1	Cliente_1	32
2	8	Cliente_8	32
3	9	Cliente_9	32
4	10	Cliente_10	45
5	12	Cliente_12	45

TABLA COMPRAS

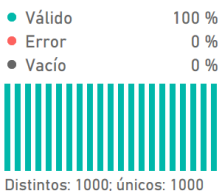
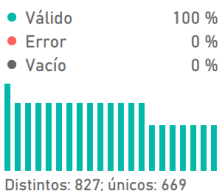
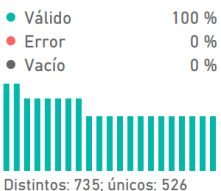
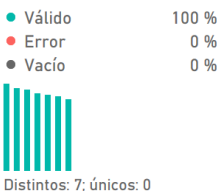
	A ^B _C id	A ^B _C cliente_id	fecha	1 ² ₃ monto de compra
	 <p>Válido 100 % Error 0 % Vacío 0 %</p> <p>Distintos: 1000; únicos: 1000</p>	 <p>Válido 100 % Error 0 % Vacío 0 %</p> <p>Distintos: 827; únicos: 669</p>	 <p>Válido 100 % Error 0 % Vacío 0 %</p> <p>Distintos: 735; únicos: 526</p>	 <p>Válido 100 % Error 0 % Vacío 0 %</p> <p>Distintos: 7; únicos: 0</p>
1	2	548	12/07/2022	-50
2	4	2255	1/04/2020	0
3	5	2022	20/07/2022	100
4	6	1564	24/11/2020	0
5	7	185	13/03/2021	1000000

TABLA DEPARTAMENTOS

Orígenes de datos | Parámetros | Consulta | Ai

`= Table.SelectRows("#Tipo cambiado", each`

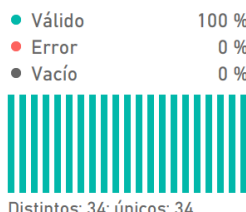
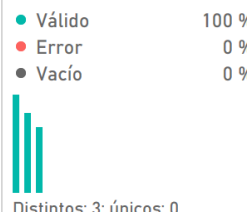



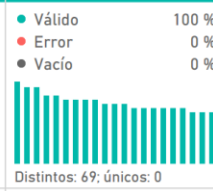
1 ² id	A ^B C nombre
	
1	Ventas
2	IT
3	IT
4	IT
5	RRHH
6	RRHH

TABLA EMPLEADOS

`= Table.SelectRows("#Tipo cambiado1", each ([departamento_id] <> null))`

A ^B C 1 ² id	1 ² nombre	1 ² salario	A ^B C departamento_id
			
1	Empleado_1	3000	61
2	Empleado_2	2000	38
3	Empleado_3	1000	22
4	Empleado_4	2000	69
5	Empleado_5	3000	23
6	Empleado_6	3000	28

Se limpiaron los datos que no correspondían con los valores de la columna, datos nulos o con otro formato, se reemplazaron valores como por ejemplo en la tabla empleados el salario estaba dosmil tipo string y se reemplazo por el valor numérico de 2000.

Se analizaron las columnas numéricas y se verificaron que los datos fueran acordes a lo necesitado para el análisis por ejemplo en edad se quitaron edades mayores a los 100 años.

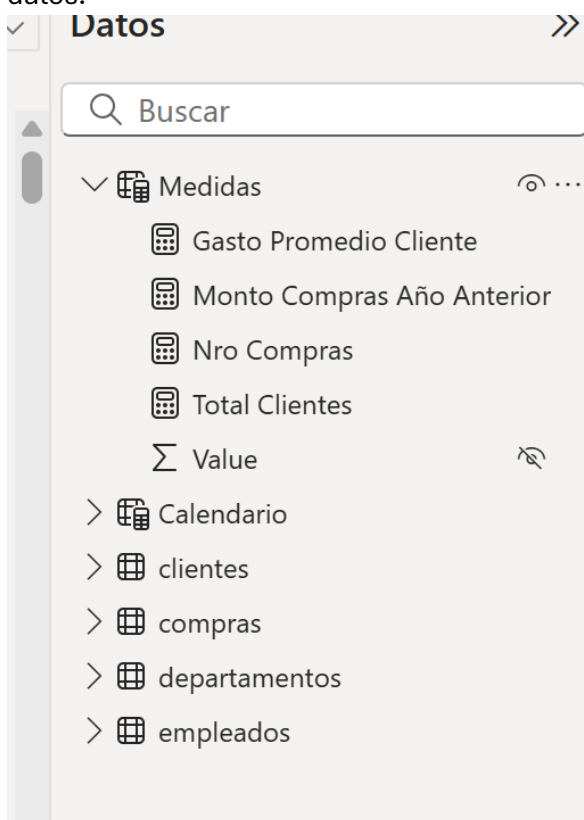
En la tabla de compras se modificaron los montos por cliente reemplazando datos null con cero y se dejaron los valores negativos, aunque puede que sea mejor eliminarlos porque no representan un dato real, los valores de cero en cada cliente se pueden representar posibles clientes que han empezado a tener un registro dentro de la empresa, pueden ser clientes potenciales o algunas compras canceladas.

Los valores negativos en montos de compras pueden ser analizados con el fin de entender el por qué los clientes han realizado la devolución de algún producto o pedido reembolso, dado que no se tiene mucha información sigue adelante con esos datos, **aclorando que se debe revisar el área de negocio y su significado para poder eliminarlos, se pueden marcar como valores atípicos y ya dependiendo del contexto tomar la decisión de eliminarlos o dejarlos.**

Se observé también que los id de los departamentos de empleados no corresponden con el valor id por lo que se realizó un análisis con el fin de darle un nombre de departamento a cada id para realizar después del dashboard.

id	nombre	salario	departamento_id	DepartamentoNombre
2	Empleado_2	2000	38	Otro
4	Empleado_4	2000	69	Otro
8	Empleado_8	2000	40	Ventas
10	Empleado_10	2000	35	Ventas
13	Empleado_13	2000	64	Otro
17	Empleado_17	2000	62	Otro
23	Empleado_23	2000	53	Otro
24	Empleado_24	2000	50	Otro
26	Empleado_26	2000	25	Ventas
27	Empleado_27	2000	32	Otro
31	Empleado_31	2000	18	RRHH
38	Empleado_38	2000	49	Ventas
42	Empleado_42	2000	45	RRHH

Esto permitió darle nombre a cada id de departamento y poder realizar un análisis posterior de los datos.



Se crearon algunas medidas y la tabla calendario para poder analizar mejor los datos y poder crear o proponer una idea y planear los análisis que se realizarán más adelante.



IDENTIFICA UNA COLUMNA NUMÉRICA RELEVANTE (EJ. "MONTO DE COMPRA").

TABLA CLIENTES

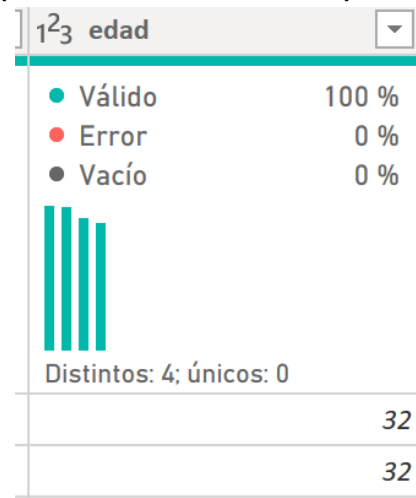


TABLA COMPRAS

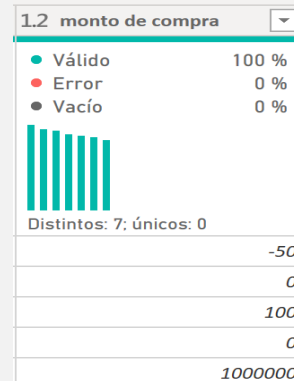
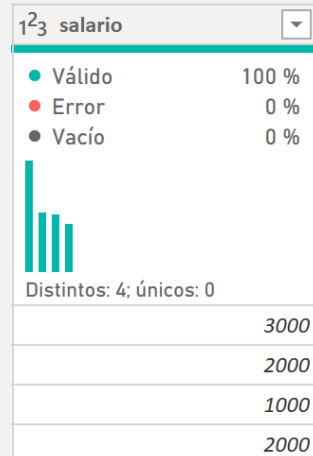


TABLA DEPARTAMENTOS

No presenta una columna numérica relevante

TABLA EMPLEADOS



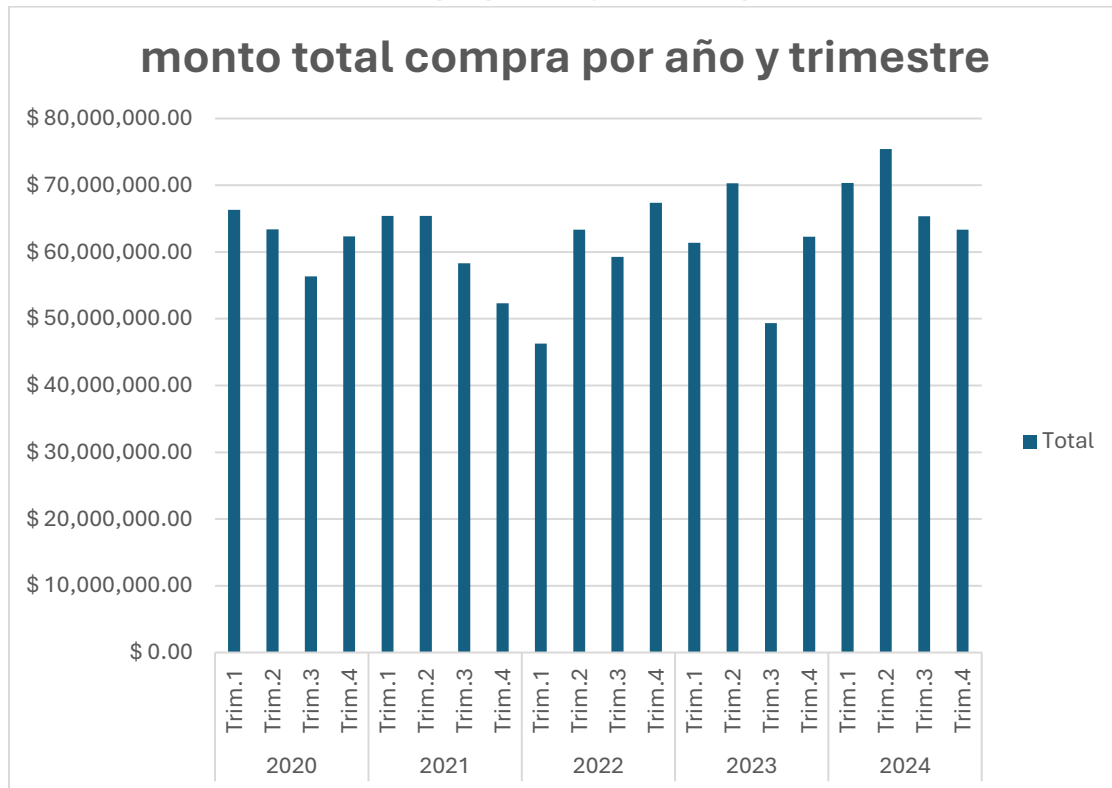


CALCULA Y PRESENTA LA MEDIA Y MEDIANA DE DICHA COLUMNA.

TABLA	MEDIA	MEDIANA
CLIENTES (EDAD/AÑOS)	40	32
COMPRAS (MONTO DE COMPRA /DÓLARES)	142653.9	200.0
DEPARTAMENTOS	No aplica	No aplica
EMPLEADOS (SALARIO /DÓLARES)	171308.8	2000.0

VISUALIZACIÓN:

Grafico de barras de monto total compra por año y trimestre para clientes.



En el gráfico se muestra la variación del monto total de compras por cliente entre los años 2020 y 2024. La información está dividida por trimestres, lo cual permite observar con mayor detalle la evolución de las compras a lo largo de cada año.

En el grafico se observa que el monto total oscila entre los 40 millones y los 70 millones durante el periodo analizado. Se evidencia un crecimiento moderado hacia los últimos años, destacando que el trimestre 2 de 2024 registra el nivel más alto de compras.

Uno de los comportamientos más relevantes es la caída sostenida entre el trimestre 3 de 2021 y el trimestre 1 de 2022, donde durante tres periodos consecutivos se redujo el volumen de compras. Esto podría estar asociado a factores externos como crisis económicas, cambios en la dinámica del mercado o variaciones en las preferencias de consumo.

Posteriormente, se presenta otra caída importante en el trimestre 3 de 2023, aunque el comportamiento posterior muestra una recuperación con tendencia al alza, lo que indica que los clientes retomaron niveles de compra más elevados en los siguientes periodos.

Finalmente, aunque el gráfico refleja altibajos naturales en la dinámica de compra, la tendencia general en los últimos años es positiva, lo cual puede interpretarse como un valor positivo para el mercado y para las ventas en los próximos años, recomendaría con estos datos realizar una serie de tiempo y analizar más a fondo este comportamiento con datos modelados.



DOUBLE V
PARTNERS



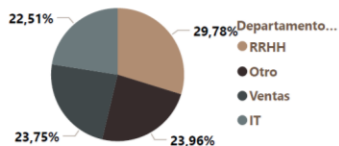
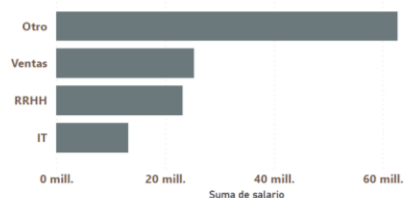
NYX

A Double V Partners company

DASHBOARD (EN POWER BI):

EMPLEADOS

Suma de salario por DepartamentoNombre



725

Número de empleados

171,31 mil

Promedio de salario

1000

Mín. de salario

1 mill.

Máx. de salario

CLIENTES

\$1,24 mil M

Suma de monto de compra

\$142,6 mil

Promedio de monto de compra

841

Total Clientes

8723

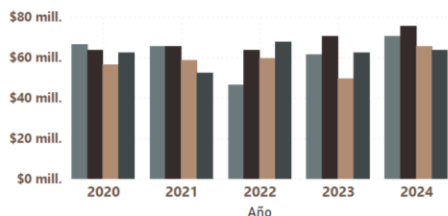
Nro Compras

39,65

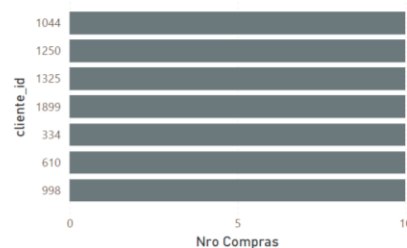
Promedio de edad

Suma de monto de compra por Año y Trimestre

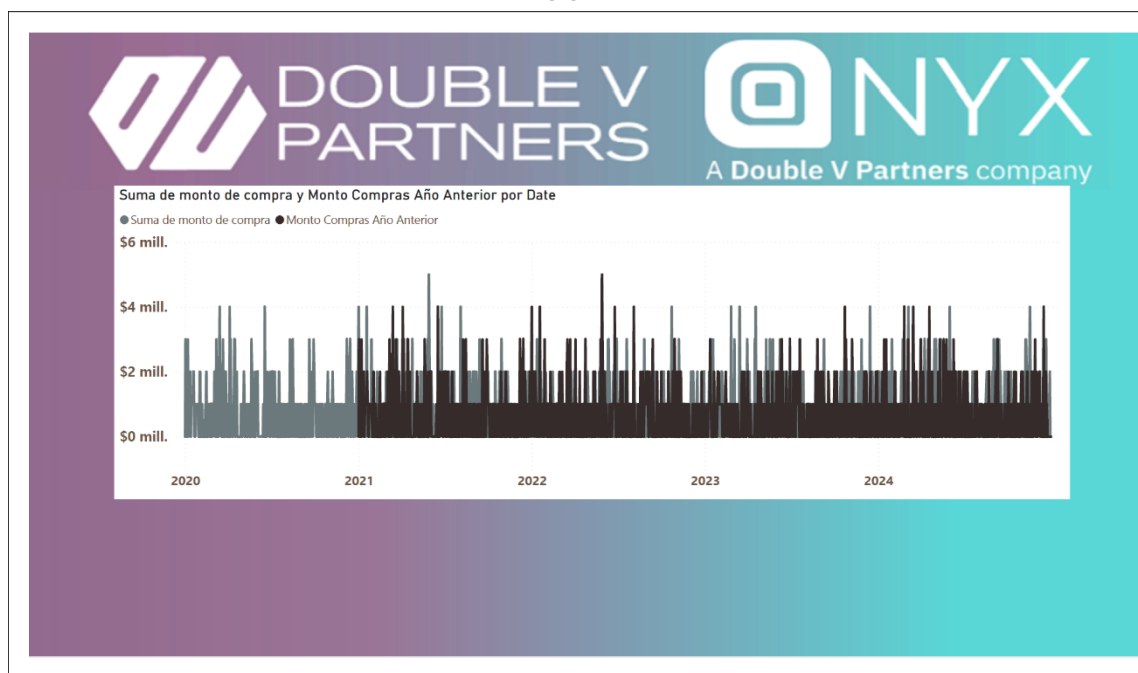
Trimestre: T1 T2 T3 T4



2020
2021
2022
2023
2024



TOOLTIP



Contexto

El dataset suministrado tenía cuatro tablas, clientes, compras, empleados, departamento. Se realizó el proceso de carga y transformación de los datos, usando buenas prácticas, se analizaron los valores de cada columna y se dejaron los datos que se consideraron viables para este análisis.

Resumen

Para el análisis de los datos se dividió en dos bloques principales, primero los empleados y luego los clientes.

En la parte superior del dashboard tenemos el análisis de los datos para los empleados, donde encontramos métricas clave como salarios (mínimo, máximo, promedio y distribución por departamento), número de empleados y participación de cada área.

En la parte inferior destinada para clientes van a encontrar un resumen de los datos como compras totales y promedio, número de clientes, compras realizadas, edad promedio y evolución de compras por año y trimestre.

Puntos de interés

Empleados:

- El salario promedio es de 171308.8, con un rango amplio entre 1.000 y 1 millón.
- El área de RRHH concentra la mayor parte de la masa salarial (29,78%), seguida por "Otro" y Ventas.
- La mayor proporción de gasto salarial no está en áreas directamente productivas (Ventas o IT), sino en otros departamentos.

Clientes:

- El monto total de compras es de \$1,24 mil M, con un promedio de \$142,6 mil por cliente.
- Hay 841 clientes que generaron 8.723 compras, lo que da un promedio cercano a 10 compras por cliente.
- La edad promedio de los clientes es 40 años, lo que sugiere un mercado de adultos en etapa económicamente activa.
- Las compras han tenido un comportamiento estable con leve crecimiento hacia 2023 y 2024, destacando picos en ciertos trimestres.

Conclusiones

1. Se puede observar que la esta compañía invierte la mayor parte de su salario a departamentos como RRHH y otras dependencias. Esto se debe analizar un poco más a fondo y detallar cuales son esos otros departamentos, ya que con la información suministrada no se puede concluir más a fondo está sobredimensión de gastos administrativos.
2. Algunos salarios están sobre dimensionados y al no tener bien esa división por departamentos no se pude monitorear si existe una desigualdad salarial o si algunos puestos están mal pagos.
3. En cuanto a los clientes a pesar de que algunos trimestres entre el trimestre 3 de 2021 y el trimestre 1 de 2022, las compras muestran una buena solidez y estabilidad a través de los años. Esto demuestra que la empresa tiene clientes solidos y mantiene sus ventas constantes a pesar de algunos trimestres bajos.

Recomendaciones

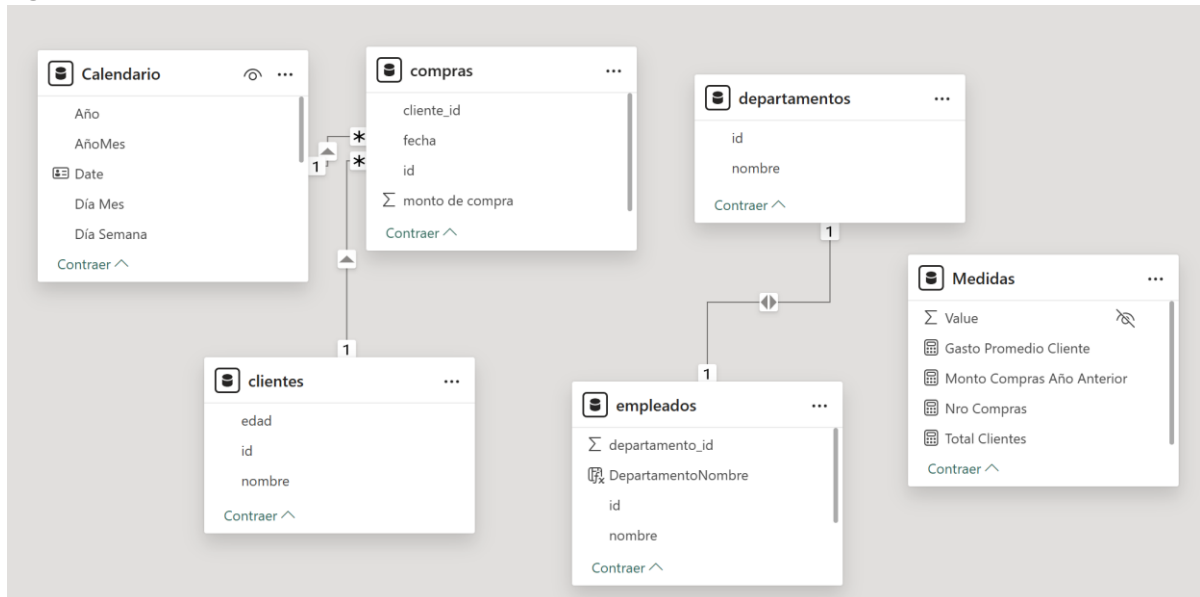
- Mejorar o complementar los datos de la sección empleados para poder determinar bien si los salarios presentan una estructura salarial desigual.
- Realizar algún modelo que me permita revisar esa variación de los datos en el tiempo, entrenar modelos de Machine Learning para predecir ese comportamiento en los meses futuros. **(Se realizó un modelo preliminar para mirar esto se dejó en el repositorio con**

los datos, es un modelo preliminar, dado que los datos no son reales fue complicado tener mejores resultados)

- Crear bases de datos normalizadas con validaciones de integridad (restricciones de edad, rangos de salarios, formatos de fechas, unicidad en IDs de clientes).
- Implementar un proceso de ETL automatizado que limpie, transforme y cargue la información en el sistema de análisis, reduciendo errores manuales.
- Establecer alertas automáticas para detectar outliers (ejemplo: salarios fuera de rango, edades irreales como <15 o >100 años, montos de compra negativos).

CONSULTAS AVANZADAS(SQL). ASUME QUE TIENES UNA BASE DE DATOS CON LAS SIGUIENTES TABLAS:

Diagrama de la base de datos



RESPONDE LAS SIGUIENTES CONSULTAS:

1. Muestra el total de compras realizadas por cada cliente, incluyendo sólo aquellos que han realizado al menos 3 compras.

```
1 • SELECT c.id AS cliente_id,
2       c.nombre AS cliente,
3       COUNT(co.id) AS total_compras,
4       SUM(co.monto_de_compra) AS monto_total
5 FROM clientes c
6 JOIN compras co ON c.id = co.cliente_id
7 GROUP BY c.id, c.nombre
8 HAVING COUNT(co.id) >= 3
9 ORDER BY total_compras DESC;
```

2. Salarios por departamento: Muestra el salario de cada empleado junto con el salario promedio de su departamento.

```
SELECT e.id AS empleado_id,
       e.nombre AS empleado,
       e.salario AS salario,
       d.nombre AS departamento,
       AVG(e.salario) OVER (PARTITION BY d.id) AS promedio_salario
FROM empleados e
JOIN departamentos d ON e.departamento_id = d.id
ORDER BY d.nombre, e.salario DESC
```

3. Top clientes: Devuelve los 5 clientes que más han gastado, mostrando su nombre y el total gastado.

```
SELECT c.id AS cliente_id,  
       c.nombre AS cliente,  
       SUM(co.monto_de_compra) AS monto_total  
FROM clientes c  
JOIN compras co ON c.id = co.cliente_id  
GROUP BY c.id, c.nombre  
ORDER BY monto_total DESC  
LIMIT 5;
```

4. Ventas por mes: Muestra el total de ventas por mes durante el último año. La consulta debe devolver el mes (formato YYYY-MM) y el total de ventas.

```
SELECT TO_CHAR(co.fecha, 'YYYY-MM') AS mes,  
       SUM(co.monto_de_compra) AS total_ventas  
FROM compras co  
WHERE DATE_PART('YEAR', co.fecha) = 2024  
GROUP BY TO_CHAR(co.fecha, 'YYYY-MM')  
ORDER BY mes;
```

5. Compras por grupo de edad:

Compara el total de compras de clientes según su grupo etario:

- Menores de 30
- Entre 30 y 50
- Mayores de 50
- Devuelve el grupo de edad y el total de compras correspondiente

```
SELECT  
  CASE  
    WHEN c.edad < 30 THEN 'Menores de 30'  
    WHEN c.edad BETWEEN 30 AND 50 THEN 'Entre 30 y 50'  
    WHEN c.edad > 50 THEN 'Mayores de 50'  
  END AS grupo_edad,  
  SUM(co.monto_de_ompra) AS total_compras  
FROM clientes c  
JOIN compras co ON c.id = co.cliente_id  
GROUP BY grupo_edad  
ORDER BY grupo_edad;
```