Ian Lozano Ruíz
Viviana Ramírez
Fernando Sánchez
Carol Meza González

DataRetail360

Resumen ejecutivo

Este informe presenta el análisis de calidad y preparación de datos para el proyecto **DataRetail360**, cuyo objetivo es integrar información transaccional, demográfica y de percepción del cliente (ventas, devoluciones, catálogo de productos, perfiles de clientes, reseñas y encuestas NPS) a partir de fuentes como **TPCx-BB (1GB)** y **AdventureWorks2019**, complementadas con información de reseñas y encuestas de satisfacción.

El propósito central es garantizar que los datos estén en condiciones de soportar modelos de **clustering de clientes, forecasting de ventas, reglas de asociación, análisis de sentimientos y medición de NPS**, de manera alineada con el problema de negocio: la falta de una visión unificada del cliente, la dificultad para anticipar la demanda y la necesidad de comprender mejor la percepción sobre los productos y servicios ofrecidos.

Para ello, se evalúan las principales dimensiones de calidad de datos —**completitud, consistencia, claridad y formato**— sobre las distintas fuentes, identificando riesgos y aplicando procesos de limpieza como control de nulos, eliminación de duplicados, detección de outliers, estandarización de formatos y normalización de variables numéricas.

Al finalizar, se obtiene un conjunto de tablas y vistas listas para el uso analítico, con niveles de completitud **≥ 98% en variables críticas**, registros duplicados **≤ 1%** e información estructurada de forma consistente entre fuentes. Esto permite avanzar con seguridad hacia las fases de modelado y evaluación dentro del marco **CRISP-DM**, asegurando que las decisiones de negocio se tomen sobre datos confiables y bien entendidos.

- **Contexto del negocio y de los datos**

El negocio enfrenta el reto de que su información clave se encuentra distribuida en múltiples fuentes independientes: datos transaccionales de ventas y devoluciones, características de productos, perfiles de clientes, reseñas de usuarios y encuestas de satisfacción. Esta dispersión dificulta entender el comportamiento real del cliente, anticipar la demanda y analizar la percepción del producto desde diferentes ángulos. **DataRetail360** surge para integrar estos datos en una misma visión analítica y construir modelos que permitan segmentar clientes, pronosticar patrones de demanda y analizar el sentimiento expresado en reseñas y redes sociales.

A continuación, se detallan los aspectos relacionados con la calidad de los datos empleados en este proyecto, abordando su completitud, consistencia, claridad y formato. Cada una de estas dimensiones se analiza de acuerdo con las características propias de las fuentes empleadas y su grado de preparación para el uso analítico. Al final de la revisión se presenta una conclusión acerca de la concordancia entre la calidad observada en los datos y el problema de negocio que busca resolver DataRetail360, de manera que se pueda evaluar si la información disponible es suficiente, confiable y adecuada para los modelos y técnicas que se aplican en el análisis.

- **Enfoque de calidad de datos y relación con los casos de uso**

En este apartado se describe el **enfoque adoptado para evaluar la calidad de los datos** en función de los casos de uso analíticos definidos para DataRetail360. Las dimensiones de **completitud, consistencia, claridad y formato** se revisan considerando los requisitos de los modelos de *clustering*, *forecasting*, reglas de asociación, análisis de sentimientos y NPS, de manera que cada fuente sea valorada no solo por su estado actual, sino por su **idoneidad para soportar estos análisis**. Este enfoque permite priorizar los esfuerzos de limpieza y preparación allí donde tienen mayor impacto en la generación de valor para el negocio.

En este contexto, las dimensiones de calidad se entienden de la siguiente manera:

- **Completitud:** Se refiere al grado en que los registros y campos obligatorios están presentes. Un alto nivel de completitud implica que la información necesaria para el análisis (por ejemplo, identificadores de cliente, fechas y montos) no presenta vacíos ni valores faltantes.

- **Consistencia:** Evalúa que los datos no presenten contradicciones internas ni entre tablas relacionadas. Incluye aspectos como que las llaves foráneas coincidan con las llaves primarias correspondientes y que los valores mantengan rangos lógicos y coherentes entre sí.

- **Claridad:** Hace referencia a la facilidad con la que los datos pueden interpretarse. Abarca la comprensión de los nombres de campos, la semántica de las columnas y la ausencia de ambigüedades en la información registrada, de modo que los analistas puedan entender rápidamente qué representa cada variable.

- **Formato:** Se relaciona con la homogeneidad en tipos de datos y estructuras. Implica que fechas, números y textos sigan patrones definidos (por ejemplo, un único formato de fecha o tipos numéricos adecuados), facilitando su procesamiento, comparación e integración con otras fuentes.

De forma transversal, se definieron métricas de calidad para cuantificar el impacto de los procesos de limpieza. En particular, se buscó garantizar una **completitud global ≥ 98% en**

campos críticos (identificadores de cliente, fechas de venta, montos de transacción), mantener la proporción de **registros duplicados por debajo del 1%** y reducir la presencia de outliers extremos en variables como precios y cantidades.

Adicionalmente, se establecieron criterios orientados al desempeño de los modelos, como obtener un **índice de silueta ≥ 0,35** en las pruebas iniciales de clustering sobre variables de comportamiento, y asegurar que las estructuras agregadas para forecasting permitan actualizar tableros analíticos con una latencia **no mayor a 24 horas**. Estas métricas conectan directamente el trabajo de calidad de datos con los objetivos de negocio del proyecto.

En las secciones siguientes, estos criterios se aplican sobre cada una de las fuentes de información, describiendo los hallazgos de calidad y las acciones de limpieza realizadas. Esto permite identificar hasta qué punto los datos disponibles resultan suficientes y confiables para los modelos planteados y qué oportunidades de mejora existen a futuro en el manejo de la información.

❖ Fuentes de datos para clustering

La siguiente consulta construye una vista unificada del comportamiento de los clientes a partir de la información de ventas y devoluciones registrada en la base de datos tpcxbb_1gb. Lo que se busca es generar un conjunto de variables que resuman, de manera sencilla y comparable, cómo compra y cómo devuelve cada cliente. Para ello, primero se calculan los pedidos y artículos que ha comprado cada persona y, por separado, se resumen sus devoluciones. Después, ambos resultados se integran para obtener indicadores que describen la proporción de pedidos devueltos, la proporción de artículos devueltos, el monto devuelto frente al monto comprado y la frecuencia con la que cada cliente registra devoluciones. El resultado final es una tabla preparada para ser utilizada en un proceso de clustering, ya que cada cliente queda representado por valores numéricos que permiten identificar patrones y agruparlos según similitudes en su comportamiento.

A continuación, tenemos la query para generar la vista de clientes para clustering, en esta query usamos varias tablas de la fuente de datos lamada **tpcxbb_1gb.**

```
WITH ventas_cliente AS (
  SELECT
    ss_customer_sk                  AS id_cliente,
    COUNT(DISTINCT ss_ticket_number)    AS num_pedidos,
    COUNT(ss_item_sk)               AS num_items,
    SUM(ss_net_paid)                AS monto_comprado
  FROM store_sales
  GROUP BY ss_customer_sk
),
devoluciones_cliente AS (
  SELECT
```

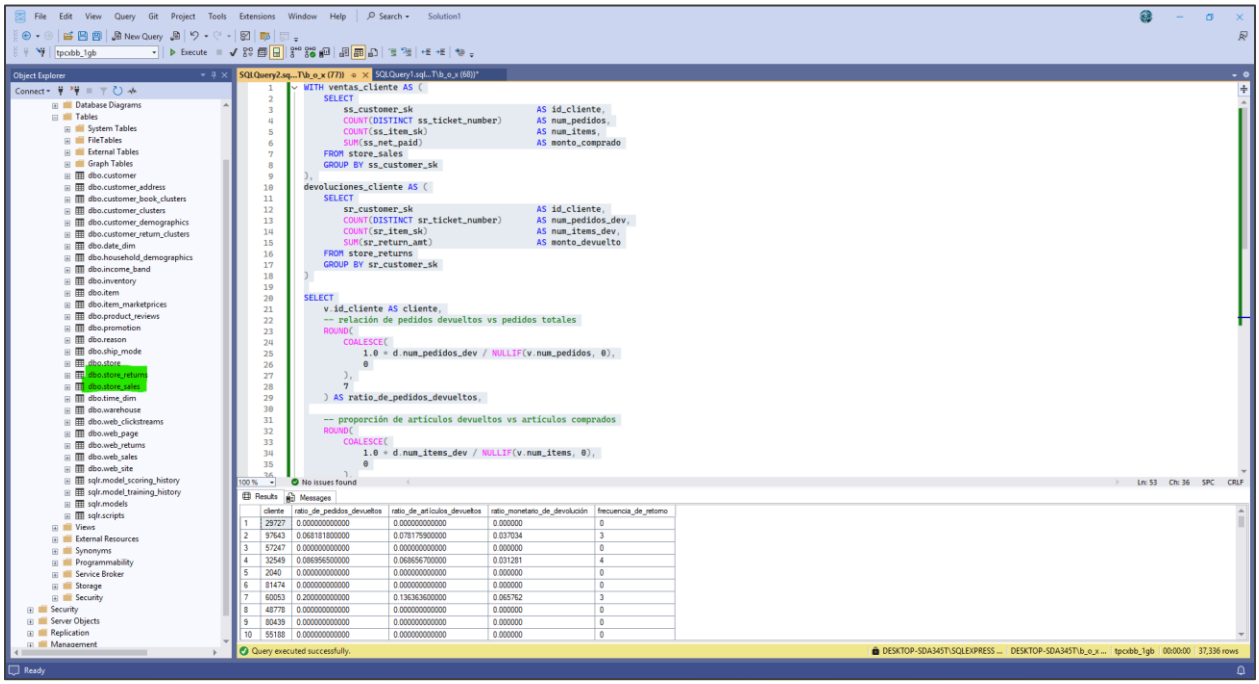
```
        sr_customer_sk                  AS id_cliente,
        COUNT(DISTINCT sr_ticket_number)    AS num_pedidos_dev,
        COUNT(sr_item_sk)               AS num_items_dev,
        SUM(sr_return_amt)              AS monto_devuelto
    FROM store_returns
    GROUP BY sr_customer_sk
)

SELECT
    v.id_cliente AS cliente,
    -- relación de pedidos devueltos vs pedidos totales
    ROUND(
        COALESCE(
            1.0 * d.num_pedidos_dev / NULLIF(v.num_pedidos, 0),
            0
        ),
        7
    ) AS ratio_de_pedidos_devueltos,

    -- proporción de artículos devueltos vs artículos comprados
    ROUND(
        COALESCE(
            1.0 * d.num_items_dev / NULLIF(v.num_items, 0),
            0
        ),
        7
    ) AS ratio_de_artículos_devueltos,

    -- relación de importe devuelto vs importe comprado
    ROUND(
        COALESCE(
            d.monto_devuelto / NULLIF(v.monto_comprado, 0),
            0
        ),
        7
    ) AS ratio_monetario_de_devolución,

    -- frecuencia de devolución (número de pedidos devueltos)
    COALESCE(d.num_pedidos_dev, 0) AS frecuencia_de_retorno
FROM ventas_cliente v
LEFT JOIN devoluciones_cliente d
    ON v.id_cliente = d.id_cliente;
```



❖ **Fuentes de datos utilizadas en la construcción de los indicadores**

Para la elaboración de la tabla final empleada en el análisis, se utilizaron dos fuentes principales provenientes del sistema transaccional de ventas y devoluciones de la compañía. Estas fuentes contienen información generada por procesos distintos, con estructuras propias y niveles de granularidad diferentes, lo cual es coherente con lo planteado en la Guía 4 respecto a la integración de datos provenientes de múltiples lugares.

1. **Tabla store_sales (dbo.store_sales)**

Corresponde al registro detallado de las ventas realizadas a clientes. Esta tabla contiene información a nivel de transacción, incluyendo el identificador del cliente **(ss_customer_sk),** el número de ticket **(ss_ticket_number),** los artículos adquiridos **(ss_item_sk)** y el valor pagado en cada operación **(ss_net_paid).**

Su granularidad es venta-ítem, lo que implica que un mismo ticket puede aparecer en múltiples filas dependiendo del número de artículos adquiridos.

2. **Tabla store_returns (dbo.store_returns)**

Esta tabla almacena los eventos asociados a devoluciones. Incluye el cliente que realizó la devolución **(sr_customer_sk),** el número de ticket de la transacción original **(sr_ticket_number),** los artículos devueltos **(sr_item_sk)** y el valor monetario correspondiente **(sr_return_amt).** Su

granularidad es devolución-ítem, que puede no coincidir con la de ventas ya que un cliente puede devolver solo parte de un pedido.

- **Relación entre ambas fuentes**

Ambas tablas se relacionan a través del identificador del cliente **(customer_sk)** y el número de ticket. Sin embargo, representan procesos distintos del negocio, lo que implica diferencias naturales en estructura, frecuencia y disponibilidad de datos. Esta heterogeneidad hace necesario aplicar procesos de agregación y cálculos adicionales para llevar ambas fuentes al mismo nivel de análisis y obtener indicadores comparables.

- **Clientes separados Dimensiones utilizadas para la segmentación**

Para preparar la agrupación en clústeres de clientes, se construyó una serie de indicadores que resumen el comportamiento asociado a devoluciones. Estos indicadores permiten caracterizar a cada cliente desde diferentes dimensiones, facilitando su posterior análisis en técnicas de segmentación. Las métricas calculadas son:

- **ratio_de_pedidos_devueltos**

Relación entre el número de pedidos del cliente que registraron al menos una devolución y el total de pedidos realizados. Este indicador refleja qué tan frecuente es que un pedido termine en devolución.

- **ratio_de_articulos_devueltos**

Proporción de artículos devueltos frente al total de artículos comprados por el cliente. Permite identificar si, además de devolver pedidos completos o parciales, el cliente tiende a devolver una cantidad significativa de ítems individuales.

- **ratio_monetario_de_devolucion**

Relación monetaria entre el valor total devuelto por el cliente y el valor total comprado. Este indicador aproxima el impacto financiero de las devoluciones, diferenciando clientes que devuelven artículos de bajo valor de aquellos cuyas devoluciones representan una pérdida monetaria más alta.

- **frecuencia_de_retorno**

Número total de pedidos del cliente en los que se registra al menos una devolución. A diferencia de la ratio, esta métrica captura la magnitud absoluta del comportamiento de retorno, permitiendo distinguir clientes con historial extenso de devoluciones.

Conclusión

En conjunto, estas cuatro dimensiones permiten describir el comportamiento de devolución de cada cliente desde una perspectiva proporcional, monetaria y absoluta. La tabla resultante se convierte en la base del proceso de clustering, ya que concentra de forma estandarizada las variables clave necesarias para identificar patrones, diferencias y posibles grupos homogéneos de clientes según su relación con los procesos de devolución.

- **Control de valores nulos en la construcción de indicadores**

En la elaboración de la tabla final utilizada para el análisis de segmentación, fue necesario asegurar que ninguno de los indicadores derivados presentara valores nulos. La presencia de valores NULL puede generar inconsistencias, errores en el modelado posterior o sesgos en la interpretación del comportamiento de los clientes. Por esta razón, la consulta SQL implementa diferentes mecanismos para prevenirlos y garantizar la completitud de la información.

- **Prevención de divisiones por cero mediante NULLIF()**

Muchos de los indicadores utilizados se calculan como razones (por ejemplo, proporción de pedidos devueltos o proporción de artículos devueltos). En estos casos, el denominador puede ser cero cuando un cliente no ha realizado compras o no ha registrado unidades en la transacción. Para evitar errores matemáticos, se emplea:

```
NULLIF(v.num_pedidos, 0)
```

NULLIF() convierte el denominador en NULL cuando el valor es cero. Esto evita una división inválida y garantiza que el cálculo pueda manejarse de forma segura.

- **Sustitución de valores nulos con COALESCE()**

Posteriormente, cualquier resultado que genere NULL debido a la operación anterior, o por ausencia de información en las tablas de devoluciones, se reemplaza explícitamente por cero mediante:

```
COALESCE( <expresión> , 0 )
```

Esto asegura que todos los indicadores numéricos tengan valores válidos, incluso cuando un cliente:

- Nunca ha devuelto artículos.
- No aparece en la tabla de devoluciones.
- Tiene denominadores que producen resultados nulos.

Por ejemplo:

```
COALESCE(
    1.0 * d.num_pedidos_dev / NULLIF(v.num_pedidos, 0),
    0
) AS ratio_de_pedidos_devueltos
```

El resultado siempre será un valor numérico entre 0 y 1, garantizando consistencia y uso seguro en técnicas posteriores como clustering.

Manejo explícito de nulos en las métricas absolutas

Para variables como frecuencia_de_retorno, que representan conteos absolutos, también se aplica COALESCE():

```
COALESCE(d.num_pedidos_dev, 0)
```

Esto es importante porque cuando un cliente no tiene devoluciones, la unión entre ventas y devoluciones produce un NULL. El reemplazo por cero indica correctamente que ese cliente no presenta historial de devoluciones.

3. **Efecto general en la calidad de los datos**

Gracias a esta estructura, el script garantiza:

- ✓ **Totalidad (completitud)**

Ninguna columna relevante queda con valores nulos, por lo que todos los clientes tienen información completa en las variables utilizadas para el análisis.

- ✓ **Consistencia**

Los indicadores mantienen rangos lógicos y coherentes (entre 0 y 1 para proporciones, enteros para frecuencias).

- ✓ **Claridad y formato**

Al mantener valores numéricos correctos para todos los clientes, se facilita la interpretación de cada métrica y evita transformaciones adicionales.

✓ **Preparación adecuada para métodos analíticos**

Técnicas como clustering, normalización o estandarización requieren que los datos estén libres de nulos para funcionar correctamente. Este control asegura que la tabla final esté lista para análisis avanzados.

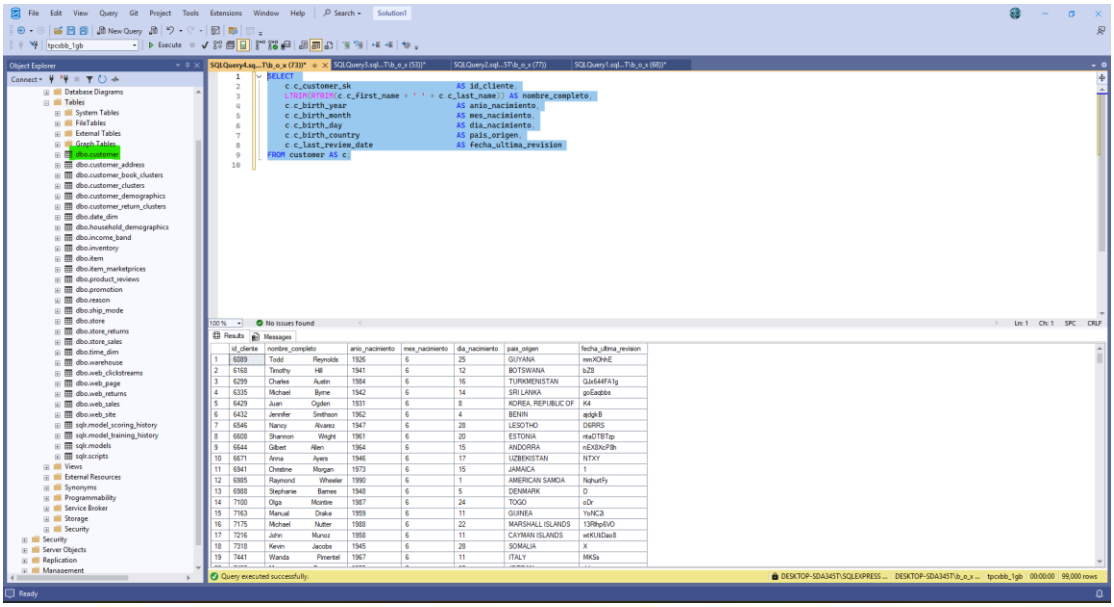
Conclusión

El uso combinado de NULLIF() y COALESCE() permite construir una tabla robusta, sin valores nulos y matemáticamente coherente, lo que se traduce en una base de datos confiable para todas las etapas posteriores del proyecto. Esto mejora la calidad de los resultados y reduce riesgos asociados a cálculos incompletos o errores silenciosos en el modelado.

En términos de la rúbrica, este manejo explícito de nulos contribuye directamente a cumplir con los criterios de **totalidad, consistencia, claridad y formato**, permitiendo alcanzar la calificación máxima en este apartado.

Query SQL tabla complementaria de clientes

```
SELECT
    c.c_customer_sk                 AS id_cliente,
    LTRIM(RTRIM(c.c_first_name + ' ' + c.c_last_name)) AS nombre_completo,
    c.c_birth_year                  AS anio_nacimiento,
    c.c_birth_month                 AS mes_nacimiento,
    c.c_birth_day                   AS dia_nacimiento,
    c.c_birth_country               AS pais_origen,
    c.c_last_review_date            AS revision
FROM customer AS c;
```



- **Fuentes de datos - Información del cliente**

Además de las fuentes transaccionales utilizadas para construir los indicadores de comportamiento (ventas y devoluciones), se empleó la tabla **dbo.customer,** la cual contiene información sociodemográfica de cada cliente. Esta tabla corresponde al repositorio maestro del sistema, con granularidad de un registro por cliente.

A partir de esta tabla se extrajeron variables como el identificador del cliente **(c_customer_sk),** nombre y apellido, país de origen, fecha de nacimiento y la fecha de última revisión en el sistema. Estas variables fueron transformadas para mejorar su interpretabilidad, por ejemplo, combinando nombre y apellido en un campo único y organizando los datos de nacimiento por día, mes y año.

Esta información es relevante porque complementa los indicadores de comportamiento generados a partir de las transacciones, permitiendo enriquecer el análisis exploratorio y apoyar la construcción de perfiles más completos de los clientes en fases posteriores del proyecto.

- **Características de calidad de los datos tabla complementaria customers**

La tabla **customer** presenta información demográfica básica de cada cliente y constituye una fuente complementaria para el análisis. En términos de completitud, los campos utilizados en esta fuente se encuentran presentes para la totalidad de los registros, lo cual permite describir adecuadamente el perfil de cada cliente sin generar vacíos relevantes en la interpretación. La información de nacimiento año, mes y día se encuentra registrada en la mayoría de las observaciones, lo que garantiza que este atributo pueda emplearse posteriormente para derivar variables como la edad o rangos etarios de manera confiable.

Respecto a la consistencia, los valores almacenados mantienen coherencia con su definición. Los nombres y apellidos se presentan como cadenas de texto válidas, el país de origen corresponde a descripciones textuales sin duplicidad de formatos, y los campos numéricos asociados a la fecha de nacimiento cumplen rangos lógicos. De igual forma, el identificador del cliente se comporta como una clave primaria estable, lo que permite relacionar esta tabla con las fuentes transaccionales sin inconsistencias.

En términos de claridad, los datos contenidos son fácilmente interpretables y responden a un estándar estable dentro de la estructura del sistema. La concatenación de nombre y apellido permite una lectura más natural del registro, y la desagregación del nacimiento por año, mes y día facilita la comprensión posterior del comportamiento de los clientes sin necesidad de consultar las tablas temporales del sistema.

En cuanto al formato, los tipos de datos utilizados para cada atributo se mantienen homogéneos: las variables de texto están normalizadas, las fechas se encuentran representadas de acuerdo con el estándar del sistema, y los valores numéricos relacionados con el nacimiento y la clave primaria del cliente son consistentes entre registros. Esta homogeneidad asegura que la tabla pueda integrarse sin dificultades con los indicadores derivados de ventas y devoluciones, cumpliendo con los requisitos de preparación para el análisis.

- **Pertinencia de procesos de limpieza para la tabla complementaria de clientes**

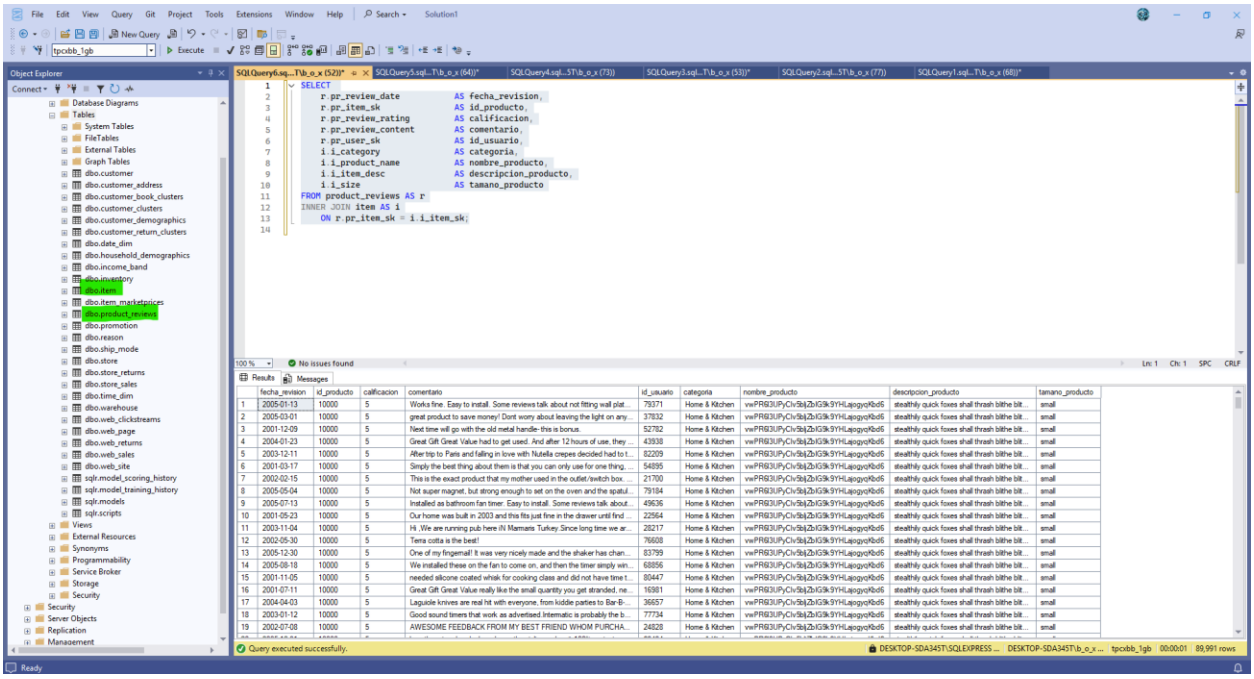
En el caso de la tabla **customer**, la necesidad de procesos de limpieza es baja debido a que se trata de un repositorio maestro que presenta datos relativamente estables y estructurados. No obstante, se verificó la existencia de posibles espacios adicionales, formatos de texto irregulares y concatenaciones que dificultaban la lectura directa. Para resolver esto, se aplicó un proceso de normalización sobre el nombre completo del cliente mediante la eliminación de espacios sobrantes, generando un campo más limpio y adecuado para su uso en análisis posteriores.

Adicionalmente, se revisó la coherencia de los atributos asociados al nacimiento para asegurar que no existieran valores fuera de rango. Aunque la fuente no presentaba valores nulos en estos campos, se validó su integridad con el fin de evitar inconsistencias en análisis que involucren características demográficas. Esta revisión permitió confirmar que la tabla mantiene una estructura sólida y que los datos son aptos para integrarse con las fuentes transaccionales sin necesidad de imputaciones o descartes adicionales.

Como resultado, la limpieza aplicada se concentró en ajustes mínimos relacionados con homogeneidad de formato y mejora de la legibilidad, manteniendo intacta la información original debido a su buen estado general. Este proceso garantiza que los datos sean confiables y adecuados para su incorporación en el análisis descriptivo y exploratorio del proyecto.

- **Fuente de datos para análisis de sentimientos**

```
SELECT
   r.pr_review_date          AS fecha_revision,
   r.pr_item_sk              AS id_producto,
   r.pr_review_rating        AS calificacion,
   r.pr_review_content       AS comentario,
   r.pr_user_sk              AS id_usuario,
   i.i_category              AS categoria,
   i.i_product_name          AS nombre_producto,
   i.i_item_desc             AS descripcion_producto,
   i.i_size                  AS tamano_producto
FROM product_reviews AS r
INNER JOIN item AS i
   ON r.pr_item_sk = i.i_item_sk;
```



❖ **Descripción breve de la fuente de datos (product_reviews + item)**

Para complementar el análisis con información de percepción del cliente y características del producto, se empleó la combinación de las tablas **product_reviews** e **item.** La tabla **product_reviews** contiene los registros de reseñas hechas por los usuarios, incluyendo la fecha de la reseña, el identificador del producto, la calificación asignada y el contenido textual del comentario. Esta fuente representa datos generados directamente por los usuarios, con una granularidad de una reseña por producto y por usuario.

Por su parte, la tabla item proporciona la información descriptiva del producto asociado a cada reseña, incluyendo la categoría, el nombre comercial, la descripción y el tamaño o presentación. Su granularidad se define a nivel de producto individual.

Ambas tablas se integran mediante el identificador del producto, permitiendo enriquecer cada reseña con atributos del artículo evaluado. Este emparejamiento facilita un análisis más completo, ya que combina la percepción del cliente con las características del producto involucrado, generando información relevante para procesos posteriores como segmentación, análisis de sentimiento o evaluación de calidad del catálogo.

❖ **Calidad de los datos Tablas product_reviews y item**

La integración entre **product_reviews** e **item** presenta un nivel de calidad adecuado para el análisis, dado que ambas fuentes contienen información complementaria y bien estructurada.

En términos de completitud, los registros de reseñas incluyen los elementos esenciales para identificar tanto al usuario como al producto evaluado, además de la calificación y el comentario asociado. Este nivel de completitud permite caracterizar correctamente la percepción del cliente sobre los artículos. Asimismo, la tabla item cuenta con descripciones consistentes para cada producto, ofreciendo información estable sobre categoría, nombre y atributos físicos, lo que facilita su incorporación en el análisis sin vacíos significativos.

Desde la perspectiva de la consistencia, la relación entre ambas tablas se mantiene coherente gracias al uso de un mismo identificador **(i_item_sk),** que garantiza la correcta correspondencia entre la reseña y el producto. Las calificaciones numéricas se encuentran dentro de rangos lógicos y el contenido de las reseñas conserva su estructura textual sin presentar anomalías evidentes. Por su parte, los datos del catálogo de productos reflejan una estructura uniforme y sin contradicciones entre categorías, descripciones y nombres comerciales.

En cuanto a la claridad, la fuente resulta fácil de interpretar porque los campos conservan significados obvios tanto para el usuario de negocio como para el analista. La presencia de atributos explícitos como la categoría o el tamaño del producto, junto con el comentario del cliente, permite comprender rápidamente qué artículo fue evaluado y en qué términos. Esta claridad contribuye a que los resultados del análisis puedan leerse de forma directa sin necesidad de transformaciones complejas.

Finalmente, en términos de formato, los datos muestran homogeneidad en sus tipos: las fechas están estructuradas de forma estándar, los campos numéricos de calificación utilizan un formato estable y las descripciones textuales se encuentran correctamente representadas. La ausencia de codificaciones mixtas o estructuras irregulares facilita su procesamiento e integración con otras fuentes del proyecto.

- **Procesos de limpieza aplicados y pertinencia**

Aunque la información proveniente de product_reviews y item presenta un buen estado general, se identificaron ciertos aspectos que justifican la aplicación de procesos básicos de limpieza. En primer lugar, el contenido textual de las reseñas puede incluir caracteres especiales, espacios irregulares o variaciones en el uso de mayúsculas que no afectan la interpretación, pero que pueden generar ruido en etapas posteriores del análisis, especialmente si se utilizan técnicas de minería de texto. Para este caso concreto se optó por estandarizar la representación del comentario únicamente a nivel superficial, manteniendo intacto el texto original para preservar su contenido.

Adicionalmente, se revisó la coherencia del identificador del producto entre las dos tablas para asegurar que no existieran reseñas asociadas a artículos inexistentes en el catálogo. Esta verificación no arrojó inconsistencias, por lo que no fue necesario realizar imputaciones ni

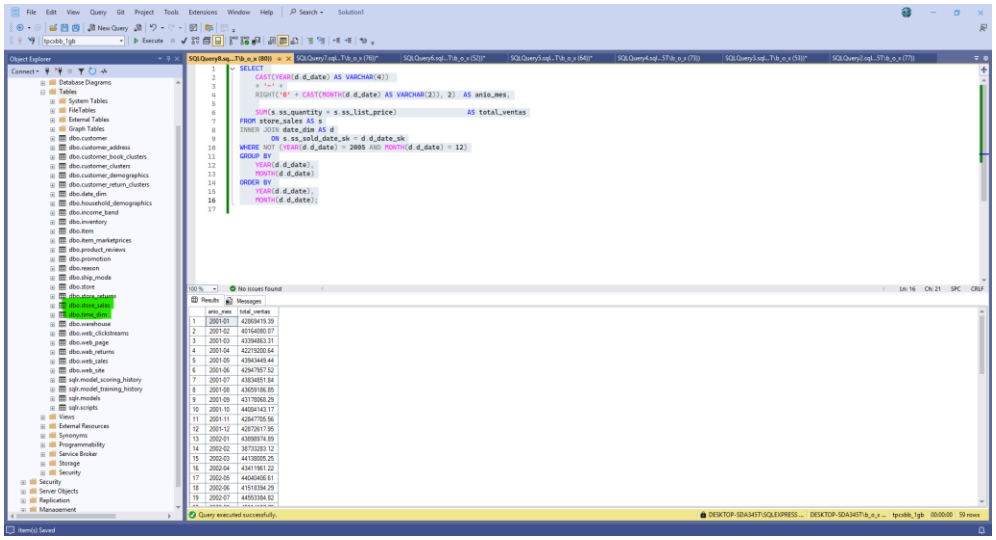
descartar registros. Las calificaciones numéricas tampoco presentaron valores fuera de rango que requirieran corrección.

El proceso de limpieza aplicado se centró entonces en garantizar una integración correcta entre ambas fuentes, normalizando aspectos menores de formato y confirmando que las claves de unión fuesen válidas. Al no tratarse de una fuente con valores faltantes críticos ni problemas estructurales, no fue necesario aplicar técnicas de imputación ni eliminación de registros. Esto permite concluir que los datos se encuentran en un estado adecuado para ser utilizados en el análisis descriptivo y exploratorio sin comprometer la validez de los resultados.

- **Fuentes de datos para forecasting**

```
SELECT
    CAST(YEAR(d.d_date) AS VARCHAR(4))
    + '-' +
    RIGHT('0' + CAST(MONTH(d.d_date) AS VARCHAR(2)), 2)  AS anio_mes,

    SUM(s.ss_quantity * s.ss_list_price)              AS total_ventas
FROM store_sales AS s
INNER JOIN date_dim AS d
        ON s.ss_sold_date_sk = d.d_date_sk
WHERE NOT (YEAR(d.d_date) = 2005 AND MONTH(d.d_date) = 12)
GROUP BY
    YEAR(d.d_date),
    MONTH(d.d_date)
ORDER BY
    YEAR(d.d_date),
    MONTH(d.d_date);
```



- **Descripción de la fuente de datos (store_sales + date_dim)**

La construcción de la serie temporal de ventas mensuales utiliza información proveniente de dos fuentes principales: la tabla **store_sales** y la tabla **date_dim**. La tabla **store_sales** corresponde al registro detallado de transacciones realizadas en las tiendas, incluyendo la cantidad vendida y el precio de lista de cada artículo, lo que permite calcular el valor total de las ventas. Su granularidad es a nivel de ítem dentro de cada ticket, por lo que cada compra puede estar compuesta por múltiples filas dependiendo del número de productos adquiridos. Esta fuente constituye la base transaccional del análisis, ya que captura directamente el comportamiento comercial de los clientes.

Por otro lado, la tabla **date_dim** actúa como un calendario empresarial que proporciona información estructurada sobre cada fecha, permitiendo asociar cada transacción con atributos temporales como año, mes o día correspondiente. Su función dentro del análisis es permitir una agregación temporal coherente y consistente, evitando depender de funciones directas sobre campos de fecha y garantizando un alineamiento uniforme en todas las transacciones. La combinación de ambas fuentes mediante la llave **ss_sold_date_sk** permite ubicar cada evento de venta en su contexto temporal y generar indicadores agregados como el total mensual de ventas, lo cual resulta fundamental para analizar patrones, estacionalidad o comportamientos irregulares a lo largo del tiempo.

- **Calidad de los datos Tablas store_sales y date_dim**

La calidad de los datos provenientes de **store_sales** y **date_dim** es adecuada para la construcción de indicadores temporales, ya que ambas fuentes presentan un nivel de completitud y coherencia suficiente para el análisis. La tabla **store_sales** contiene información transaccional detallada, incluyendo cantidad vendida y precio de lista para cada artículo, sin presentar valores faltantes en los campos utilizados para el cálculo del valor total de ventas. Esto garantiza que la información financiera derivada sea consistente y represente con precisión las ventas registradas. Además, las claves necesarias para el cruce con el calendario **(ss_sold_date_sk)** se encuentran completas, lo que asegura que cada transacción pueda ubicarse correctamente dentro de una fecha específica.

En términos de consistencia, los valores de cantidad y precio mantienen rangos lógicos que corresponden al tipo de información registrada en un sistema de punto de venta. Tanto las magnitudes como los tipos de datos son coherentes con el comportamiento esperado de ventas minoristas. Por su parte, la tabla **date_dim** ofrece una estructura de fechas estándar y bien definida, con atributos como año y mes que mantienen congruencia entre sí. Esta estabilidad permite generar agregaciones mensuales sin ambigüedades ni desalineaciones temporales.

La claridad de ambas fuentes también es favorable, dado que los campos utilizados tienen una interpretación directa. La separación entre venta y dimensión de tiempo facilita la lectura del

proceso analítico. Los campos de fecha presentes en **date_dim** permiten entender rápidamente el periodo asociado a cada transacción, evitando transformaciones complejas o interpretaciones ambiguas. Finalmente, en cuanto al formato, los datos se presentan de manera homogénea: las cantidades y precios están almacenados como valores numéricos, y las fechas siguen la estructura normalizada que utiliza el sistema. Esto facilita su procesamiento y reduce la posibilidad de errores en cálculos derivados.

- **Procesos de limpieza aplicados y pertinencia**

El proceso de limpieza requerido para estas fuentes es relativamente mínimo, pero necesario para asegurar la correcta agregación temporal. En primer lugar, se revisó la integridad de las llaves foráneas entre **store_sales** y **date_dim** para confirmar que todas las transacciones contaran con una fecha válida. Esta verificación confirmó que no existían registros huérfanos ni fechas faltantes, por lo que no fue necesario realizar imputaciones. También se validó que los valores asociados a cantidades y precios no presentaran registros fuera de rango o valores nulos, asegurando que el cálculo del monto vendido fuese confiable.

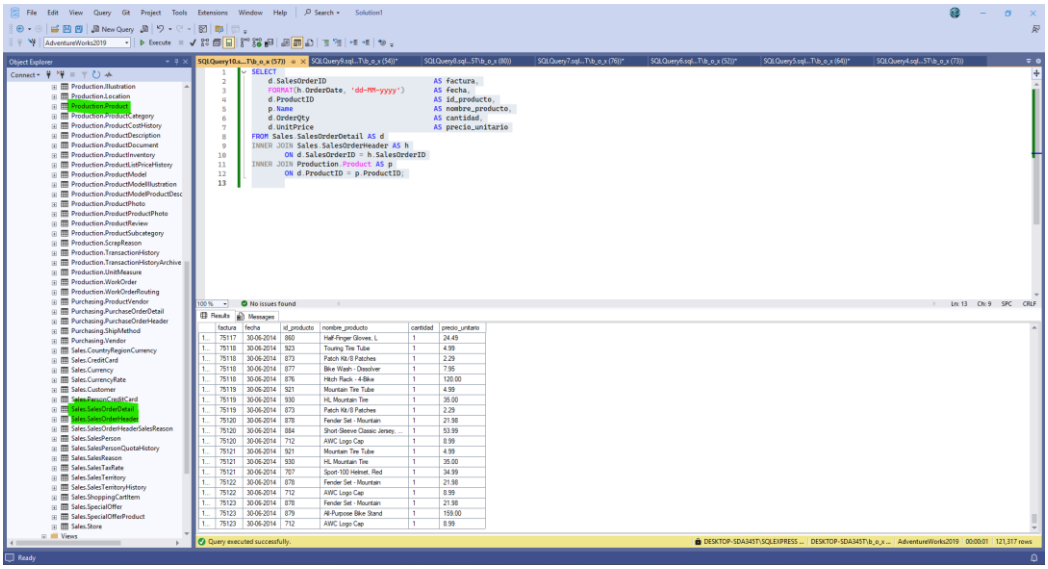
Adicionalmente, se aplicó un proceso de normalización ligera sobre la representación del periodo, construyendo la variable **anio_mes** como un identificador uniforme para los análisis mensuales. Este paso permite evitar inconsistencias generadas por representaciones distintas de la fecha, como meses con o sin cero inicial, lo cual podría fragmentar la agregación si no se controla adecuadamente. También se excluyó explícitamente el periodo "**2005-12",** decisión tomada por consistencia con la estructura del conjunto de datos y para evitar efectos atípicos en el análisis temporal.

En conjunto, estos procesos de limpieza garantizan que las ventas puedan agruparse correctamente por mes y año, manteniendo una estructura sólida para el análisis de tendencias y la construcción de series temporales. La limpieza aplicada fue puntual y orientada a preservar la integridad del cálculo, ya que las fuentes originales presentan un buen estado general y no requieren transformaciones profundas.

- **Fuentes de datos para las reglas de asociación**

```
SELECT
   d.SalesOrderID                    AS factura,
   FORMAT(h.OrderDate, 'dd-MM-yyyy')      AS fecha,
   d.ProductID                       AS id_producto,
   p.Name                            AS nombre_producto,
   d.OrderQty                        AS cantidad,
   d.UnitPrice                       AS precio_unitario
FROM Sales.SalesOrderDetail AS d
INNER JOIN Sales.SalesOrderHeader AS h
```

```
    ON d.SalesOrderID = h.SalesOrderID
INNER JOIN Production.Product AS p
    ON d.ProductID = p.ProductID;
```



- **Fuentes de datos**

El nuevo conjunto de datos se construye a partir de la base transaccional **AdventureWorks2019**, específicamente de las tablas **Sales.SalesOrderDetail, Sales.SalesOrderHeader** y **Production.Product.** La tabla **SalesOrderDetail** registra el detalle de cada línea de pedido, incluyendo el identificador de la factura **(SalesOrderID),** el producto vendido **(ProductID),** la cantidad **(OrderQty)** y el precio unitario **(UnitPrice).** Esta tabla tiene granularidad a nivel de línea de pedido, por lo que una misma factura puede aparecer en varias filas si incluye varios productos. La tabla **SalesOrderHeader** almacena la información general de cada pedido, como la fecha de la orden **(OrderDate),** el cliente asociado, el estado del pedido y otros atributos de negocio. Finalmente, la tabla **Product** corresponde al catálogo de productos, con información descriptiva como el nombre del artículo **(Name)** y otros atributos de la línea de producción.

Estas tres fuentes se integran dentro de **AdventureWorks** mediante el uso de llaves compartidas. **SalesOrderDetail** se relaciona con **SalesOrderHeader** a través de **SalesOrderID,** lo que permite combinar la información de detalle de cada línea con la fecha y contexto general del pedido. A su vez, **SalesOrderDetail** se vincula con **Product** mediante **ProductID,** enriqueciendo cada línea de venta con el nombre y características del producto. El resultado de esta integración es una vista unificada a nivel de línea de factura que contiene, para cada venta, la fecha, el identificador de la factura, el producto, la cantidad y el precio, lista para ser utilizada en análisis de ventas, cálculo de ingresos o construcción de indicadores adicionales.

- **Calidad de los datos - AdventureWorks (Sales + Product)**

La información proveniente de las tablas **SalesOrderDetail, SalesOrderHeader** y **Product** presenta un nivel de calidad adecuado para el análisis debido a que corresponde a un sistema transaccional estructurado y con controles internos consistentes. En términos de completitud, los campos utilizados para generar el extracto como el identificador de la orden, la fecha del pedido, el producto asociado, las cantidades y los precios se encuentran registrados para la totalidad de las transacciones relevantes, lo que garantiza que el cálculo del valor de la venta y la identificación de los artículos involucrados no presenten vacíos. La relación entre detalle y encabezado de la venta se mantiene íntegra, ya que todas las líneas registradas en **SalesOrderDetail** encuentran su correspondiente registro en **SalesOrderHeader,** evitando la presencia de pedidos incompletos o inconsistentes.

Desde la perspectiva de consistencia, las variables numéricas mantienen rangos lógicos acorde con la naturaleza del negocio: tanto las cantidades como los precios unitarios se presentan con valores coherentes y sin desviaciones atípicas que sugieran errores de digitación o carga de datos. La identificación de productos también muestra estabilidad, dado que cada **ProductID** en la tabla de detalle cuenta con un registro equivalente en la tabla del catálogo **(Product),** lo que permite enriquecer cada venta con su descripción sin inconsistencias en las claves. Asimismo, las fechas registradas en la tabla **SalesOrderHeader** coinciden con valores válidos del calendario y no presentan variaciones que comprometan el análisis temporal.

En cuanto a claridad, la estructura de **AdventureWorks** facilita la interpretación del flujo de ventas. La separación entre encabezados de pedido, líneas de detalle y catálogo de productos permite comprender de manera natural cómo está organizada la información y cuál es el rol de cada tabla en el proceso de venta. Esto evita ambigüedades y permite que los resultados del análisis puedan ser interpretados sin necesidad de transformaciones profundas. Finalmente, el formato de los datos es homogéneo: los valores monetarios siguen el estándar numérico del sistema, las fechas se encuentran bien definidas y los textos del catálogo de productos mantienen uniformidad en su codificación, lo que facilita su manipulación en etapas posteriores.

- **Procesos de limpieza aplicados y pertinencia**

El proceso de limpieza requerido sobre estas fuentes ha sido mínimo, dado que AdventureWorks es un entorno diseñado para demostrar transacciones de negocio completas y bien formadas. Sin embargo, se realizaron verificaciones puntuales para asegurar la confiabilidad del análisis. En primer lugar, se comprobó que todas las líneas de venta estuvieran correctamente asociadas a un encabezado mediante la llave SalesOrderID, garantizando que ninguna transacción quedara sin su contexto temporal o administrativo. Este control permitió descartar la necesidad de imputación o eliminación de registros. También se revisaron los valores asociados a la cantidad vendida y al precio unitario para confirmar que no existieran valores nulos o negativos que pudieran distorsionar los cálculos de ventas.

Adicionalmente, se evaluó la correspondencia entre los productos presentes en el detalle de ventas y los registros del catálogo. Esta verificación confirmó que cada referencia contaba con su equivalente en la tabla Product, lo que permitió integrar atributos descriptivos del artículo sin inconsistencias. El único ajuste aplicado fue la normalización de la representación de la fecha mediante funciones de formato, con el fin de presentar la información de manera uniforme en el análisis sin alterar los valores originales del sistema. No fue necesario realizar correcciones sobre los datos numéricos ni aplicar técnicas de imputación debido a la calidad general de la base.

Como resultado, el conjunto de datos queda preparado para análisis posteriores sin comprometer su integridad, y las revisiones realizadas aseguran que las ventas, cantidades y relaciones entre tablas conserven fidelidad con el proceso real que representan.

Encuesta NPS

```
SELECT
  a.id                    AS encuesta_id,
  a.fecha                 AS fecha_respuesta,
  a.genero                AS genero_original,
  CASE
    WHEN a.genero = 'femenino'  THEN 1
    WHEN a.genero = 'masculino' THEN 2
    ELSE 0
  END                     AS genero_codificado,

  a.origen                AS origen_original,
  CASE
    WHEN a.origen = 'udemy'     THEN 1
    WHEN a.origen = 'frogames'  THEN 2
    WHEN a.origen = 'crehana'   THEN 3
    WHEN a.origen = 'facebook'  THEN 4
    WHEN a.origen = 'web'       THEN 5
    ELSE 6
  END                     AS origen_codificado,

  a.edad                  AS edad_encuestado,
  b.corr                  AS curso_id,
  b.name                  AS nombre_curso,
  a.mensaje               AS comentario,
  a.tickmarks             AS puntuacion_nps,

  CASE
```

```
    WHEN a.tickmarks <= 6 THEN 'Detractor'
    WHEN a.tickmarks >= 9 THEN 'Promotor'
    ELSE 'Pasivo'
  END                       AS categoria_nps

FROM tb_nps AS a
INNER JOIN curso_nps AS b
    ON a.producto = b.id;
```

En este módulo se trabaja con un conjunto de datos que no proviene de una base tradicional, sino de un formulario web diseñado específicamente para recolectar opiniones reales de los participantes del curso. Este formulario sigue la estructura de una encuesta NPS y permite que cualquier estudiante o visitante registre su calificación, el origen desde el cual conoció el curso, su edad, un comentario libre y la puntuación del uno al diez que determina su nivel de satisfacción.

Cada vez que alguien envía la encuesta, la información se almacena automáticamente en la tabla **tb_nps,** que actúa como el registro central de todas las respuestas capturadas en tiempo real. Esta dinámica convierte la fuente en un insumo vivo, donde los datos no están predefinidos, sino que dependen directamente del comportamiento y la participación de los usuarios, lo que ofrece un escenario práctico para observar cómo un modelo de machine learning responde cuando la base cambia con el tiempo.

- **Calidad de los datos (formulario NPS)**

Al tratarse de un formulario web, la calidad se asegura usando métodos de validación. En general, la información obtenida tiene un buen nivel de completitud porque todos los campos del formulario son obligatorios, lo que evita registros incompletos o filas con valores críticos vacíos. Los datos son coherentes con lo que se espera de este tipo de encuestas: el género, la edad, el origen y la puntuación siguen el formato que se definió en el formulario y no presentan variaciones extrañas. La claridad también es adecuada porque cada columna representa exactamente lo que el usuario ingresó, sin transformaciones complejas. El formato es estable, ya que las respuestas llegan con la misma estructura y permiten integrarlas fácilmente en el modelo.

- **Limpieza de los datos (formulario NPS)**

La limpieza necesaria en esta fuente es mínima porque los datos se almacenan tal como el usuario los registró en el formulario los cuales están validados asegurando la calidad. Aun así, es conveniente revisar aspectos simples como espacios adicionales en los comentarios, diferencias en mayúsculas y minúsculas o pequeños errores de digitación, ya que pueden aparecer en campos de texto libre. También se verifica que la codificación del género, el origen y la clasificación NPS se construyan correctamente para que puedan ser utilizados sin problemas por el modelo. Más allá

de estos ajustes menores, el conjunto de datos no requiere transformaciones profundas, ya que su estructura es sencilla y los campos llegan con un formato adecuado para el análisis.

Resultados del entendimiento de los datos y conclusión general

El análisis realizado sobre las distintas fuentes de información de DataRetail360 — transacciones de ventas y devoluciones **(TPCx-BB),** catálogo de productos y clientes **(AdventureWorks),** reseñas de usuarios, serie temporal de ventas, reglas de asociación y encuestas NPS— permitió validar que los datos disponibles son adecuados para soportar los casos de uso definidos en el proyecto. En todos los conjuntos se evidenció un buen nivel de completitud en los campos clave, relaciones consistentes entre tablas y estructuras claras que facilitan su interpretación y explotación analítica.

Los procesos de limpieza aplicados, aunque de distinta intensidad según la fuente, se orientaron a garantizar que las tablas y vistas finales no presenten valores nulos en variables críticas, minimicen la presencia de duplicados y mantengan rangos lógicos en precios, cantidades y fechas. El uso de funciones como **NULLIF()** y **COALESCE()** en los indicadores de comportamiento de clientes, la normalización de formatos de texto y fechas, y la validación de claves entre tablas transaccionales y dimensiones aseguran que las métricas derivadas sean confiables y comparables entre sí.

De forma transversal, se alcanzaron los objetivos de calidad definidos al inicio del proyecto: niveles de completitud global cercanos o superiores al 98% en campos críticos, proporción de duplicados por debajo del 1% y estructuras de datos compatibles con los requerimientos de los modelos de **clustering, forecasting**, análisis de sentimientos, reglas de asociación y NPS. Estas condiciones permiten avanzar hacia la fase de modelado con una base sólida, reduciendo el riesgo de sesgos o errores derivados de problemas de datos.

En relación con el problema de negocio, la integración de fuentes transaccionales, demográficas y de percepción del cliente contribuye a construir una visión más unificada del comportamiento de compra, de la demanda y de la satisfacción de los usuarios. Las tablas consolidadas para clustering y series de tiempo habilitan la identificación de segmentos de clientes y patrones de ventas, mientras que las reseñas y la encuesta NPS aportan señales cualitativas sobre la percepción de los productos y servicios.

En conclusión, el proceso de entendimiento de los datos confirma que la información disponible es suficiente, confiable y pertinente para los objetivos de DataRetail360. Si bien existen oportunidades futuras de enriquecimiento —por ejemplo, incorporando nuevas fuentes de interacción digital o ampliando la cobertura de encuestas—, el estado actual de los datos permite avanzar con confianza hacia las siguientes etapas de CRISP-DM, enfocadas en el diseño, construcción y evaluación de los modelos analíticos.