Practical Challenge

Data Processing

Exploratory Analysis

Before processing the data, I performed an exploratory analysis to identify and understand the database.

To solve this practical challenge, the following tools were used: Excel, Power BI, MySQL WorkBench 8.0 CE.

Data Processing: Sales Table

The following adjustments were made to process the data:

1. **Type Conversion**:
   - Columns `purchase_id` and `product_id` were converted to text.
   - Column `purchase_date` was converted to date.
   - Columns `product_price`, `service_tax`, and `comission_value` were converted to fixed decimal values:
     - Corrected the formatting of "1,000.00" to "1.000,00".
     - Reformatted values like "100000" to "1.000,00" by assuming the last two digits represent cents.
2. **Binary Columns**:
   Interpreted as "true" or "false":
   - `has_coupon`
   - `refund`
   - `cancelled`
   - `chargeback`
3. **Discount Column**:
   - Values were multiplied by 10. The following interpretation was assumed:
     - `0` = no discount.
     - `1` = 10% discount.
     - `2` = 20% discount.
     - `3` = 30% discount.

Data Processing: Producers Table

The following adjustments were made to process the data:

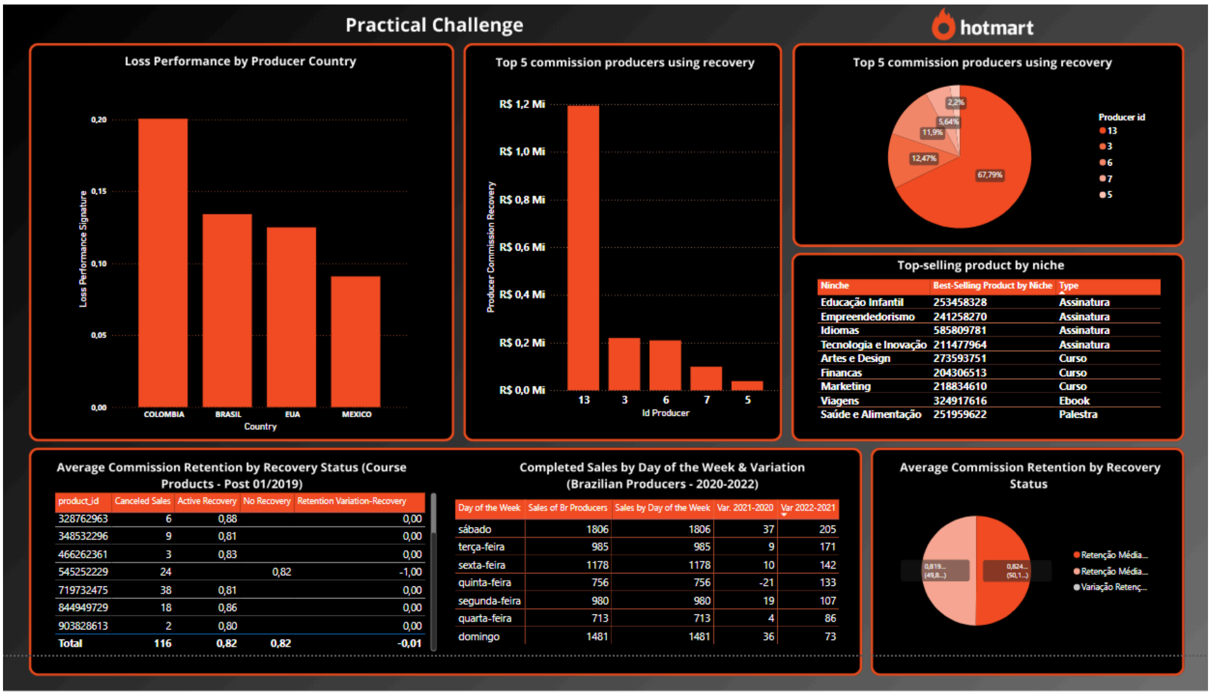
1. **Type Conversion**:
    - Column `producer_id` was converted to text.
    - Column `registry_date` was converted to date.
    - Column `country` was converted to text.

Data Processing: Products Table

The following adjustments were made to process the data:

1. **Type Conversion**:
    - Columns `product_id` and `producer_id` were converted to text.
    - Column `registry_date` was converted to date.
    - Binary columns were interpreted as "true" or "false":
        - `recovery_active`
        - `member_area_active`
2. **Other Columns**:
    - Column `niche` was converted to text.
    - Column `type` was converted to text.
    - Column `deletion_date` was converted to date.
    - Column `base_price` was converted to fixed decimal values:
        - Corrected the formatting of "1,000.00" to "1.000,00".

The following dashboard file will be attached along with the PDF of the answers.



Top-selling product by niche

Ninche	Best-Selling Product by Niche	Type
Educação Infantil	253458328	Assinatura
Empreendedorismo	241258270	Assinatura
Idiomas	585809781	Assinatura
Tecnologia e Inovação	211477964	Assinatura
Artes e Design	273593751	Curso
Financas	204306513	Curso
Marketing	218834610	Curso
Viagens	324917616	Ebook
Saúde e Alimentação	251959622	Palestra

Average Commission Retention by Recovery Status (Course Products - Post 01/2019)

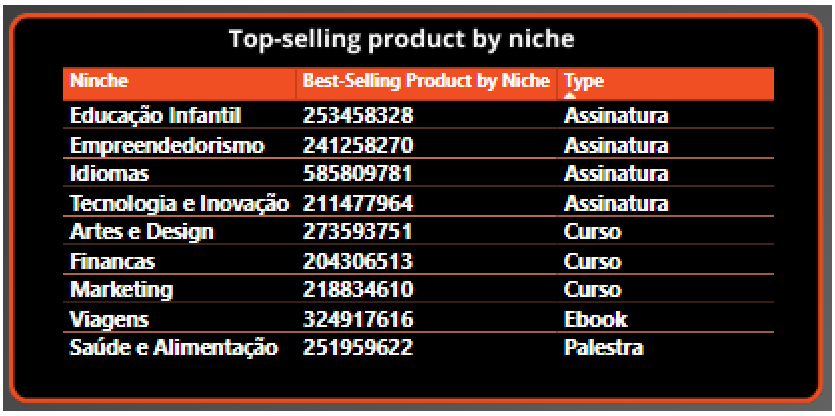
product_id	Canceled Sales	Active Recovery	No Recovery	Retention Variation-Recovery
328762963	6	0,88		0,00
348532296	9	0,81		0,00
466262361	3	0,83		0,00
545252229	24		0,82	-1,00
719732475	38	0,81		0,00
844949729	18	0,86		0,00
903828613	2	0,80		0,00
Total	**116**	**0,82**	**0,82**	**-0,01**

Completed Sales by Day of the Week & Variation (Brazilian Producers - 2020-2022)

Day of the Week	Sales of Br Producers	Sales by Day of the Week	Var. 2021-2020	Var 2022-2021
sábado	1806	1806	37	205
terça-feira	985	985	9	171
sexta-feira	1178	1178	10	142
quinta-feira	756	756	-21	133
segunda-feira	980	980	19	107
quarta-feira	713	713	4	86
domingo	1481	1481	36	73

Questions

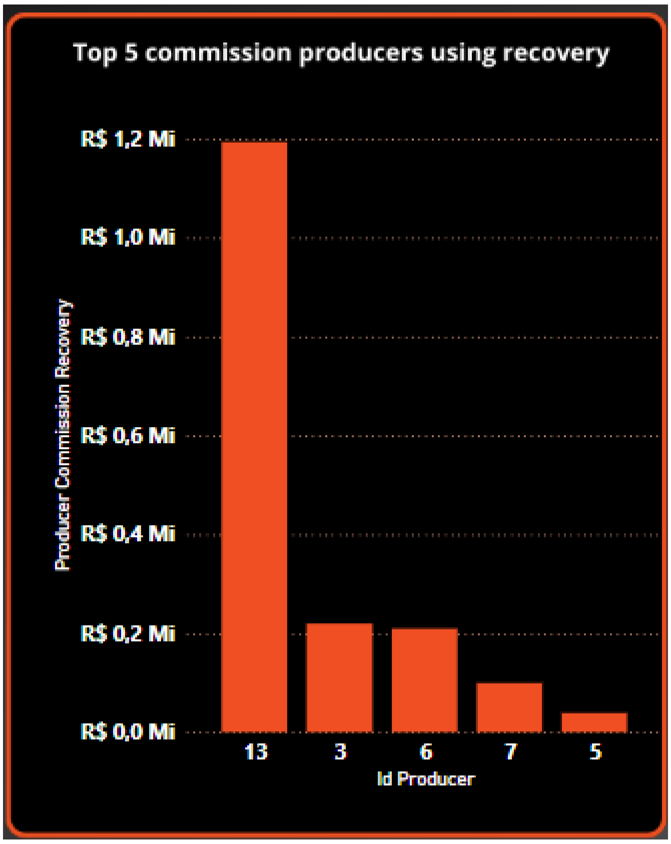
1. The top product with most valid purchases (without refund, cancellation or chargeback), in each niche with deactivated membership area and activated recovery.

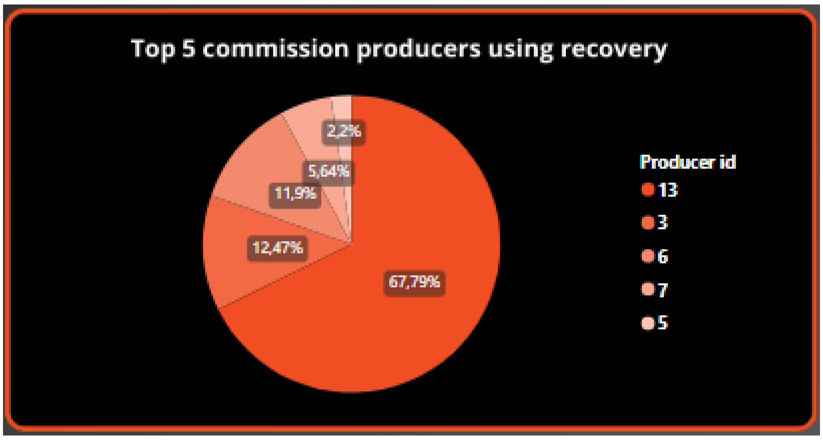


Top-selling product by niche

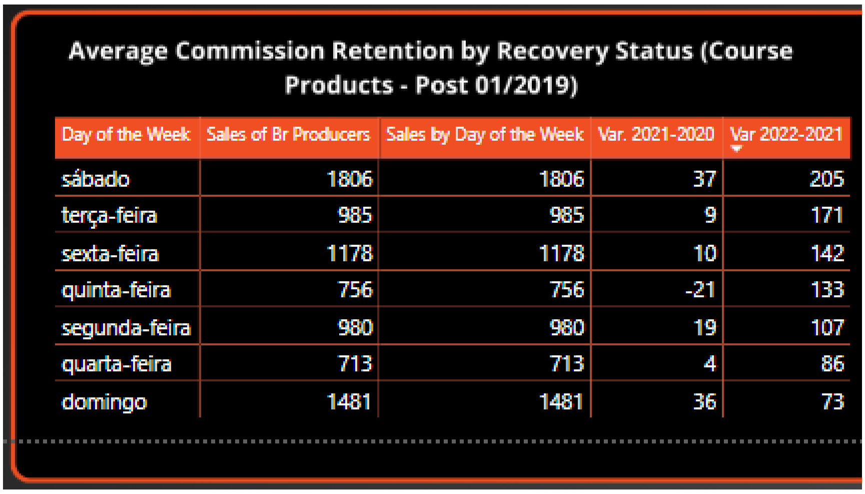
Ninche	Best-Selling Product by Niche	Type
Educação Infantil	253458328	Assinatura
Empreendedorismo	241258270	Assinatura
Idiomas	585809781	Assinatura
Tecnologia e Inovação	211477964	Assinatura
Artes e Design	273593751	Curso
Financas	204306513	Curso
Marketing	218834610	Curso
Viagens	324917616	Ebook
Saúde e Alimentação	251959622	Palestra

2. The top 5 producers who joined Hotmart from 2019 onwards and achieved the highest commission using recovery.





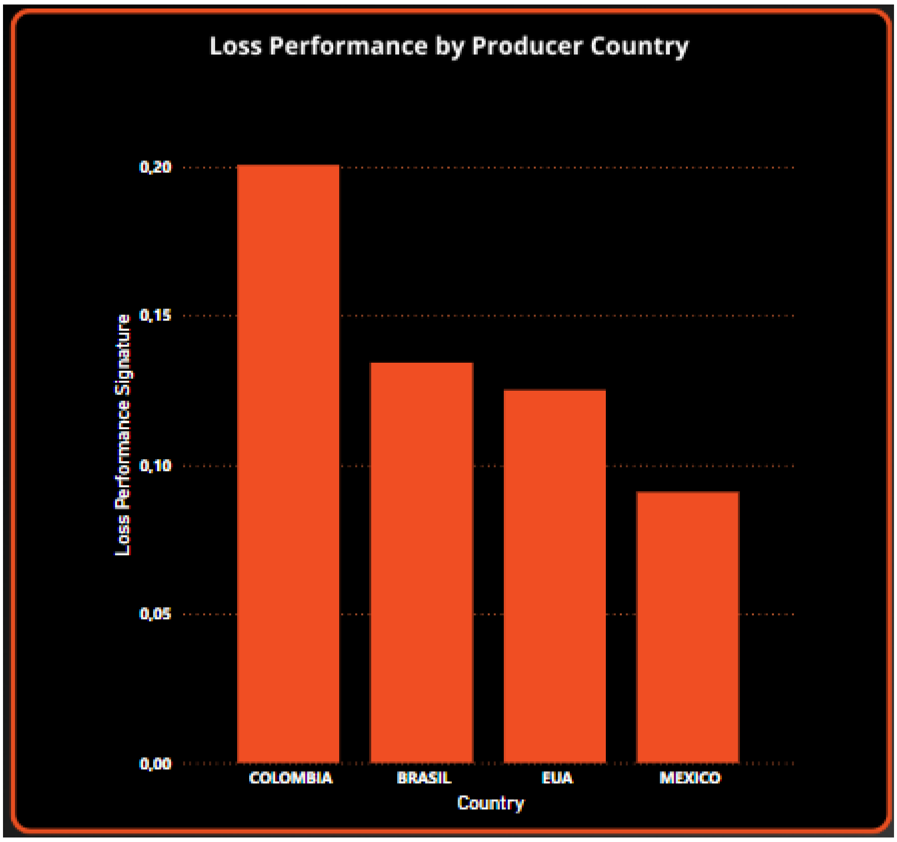
3. List the number of completed sales from Brazilian producers in each weekday for the years 2020, 2021 and 2022 and the variation between the weekday sales from 2021/2020 and 2022/2021.



Average Commission Retention by Recovery Status (Course Products - Post 01/2019)

Day of the Week	Sales of Br Producers	Sales by Day of the Week	Var. 2021-2020	Var 2022-2021
sábado	1806	1806	37	205
terça-feira	985	985	9	171
sexta-feira	1178	1178	10	142
quinta-feira	756	756	-21	133
segunda-feira	980	980	19	107
quarta-feira	713	713	4	86
domingo	1481	1481	36	73

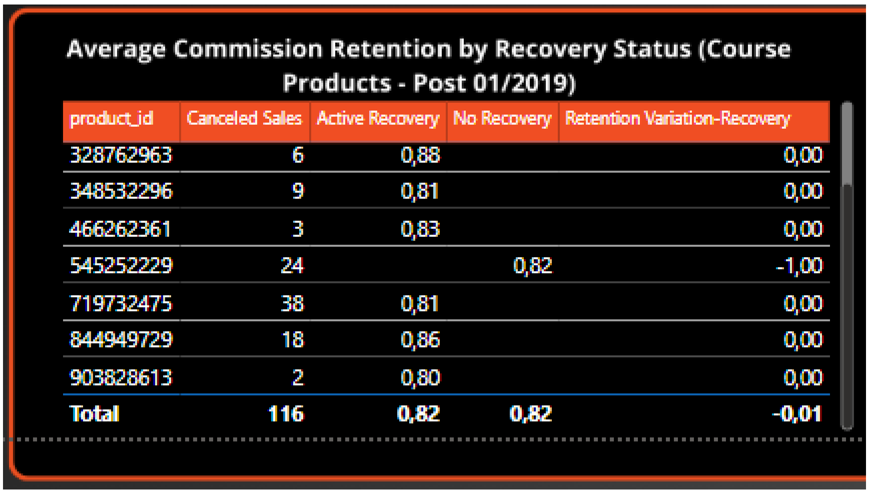
4. List each country's loss performance for each producer country, considering only products from type 'Assinatura'.
   loss_performance = (cancelations+refunds)/total_sales



Loss Performance by Producer Country

5. Looking at the products registered after 01/2019, with at least one cancellation and from the product type 'Curso', calculate the average retention performance for commission, of all products, with recovery active and without recovery active. Is there any difference for product performance considering products with the recovery tool activated?

average_commission_retention = commission_received/total_comission



Average Commission Retention by Recovery Status (Course Products - Post 01/2019)

product_id	Canceled Sales	Active Recovery	No Recovery	Retention Variation-Recovery
328762963	6	0,88		0,00
348532296	9	0,81		0,00
466262361	3	0,83		0,00
545252229	24		0,82	-1,00
719732475	38	0,81		0,00
844949729	18	0,86		0,00
903828613	2	0,80		0,00
Total	**116**	**0,82**	**0,82**	**-0,01**

Based on the analysis:

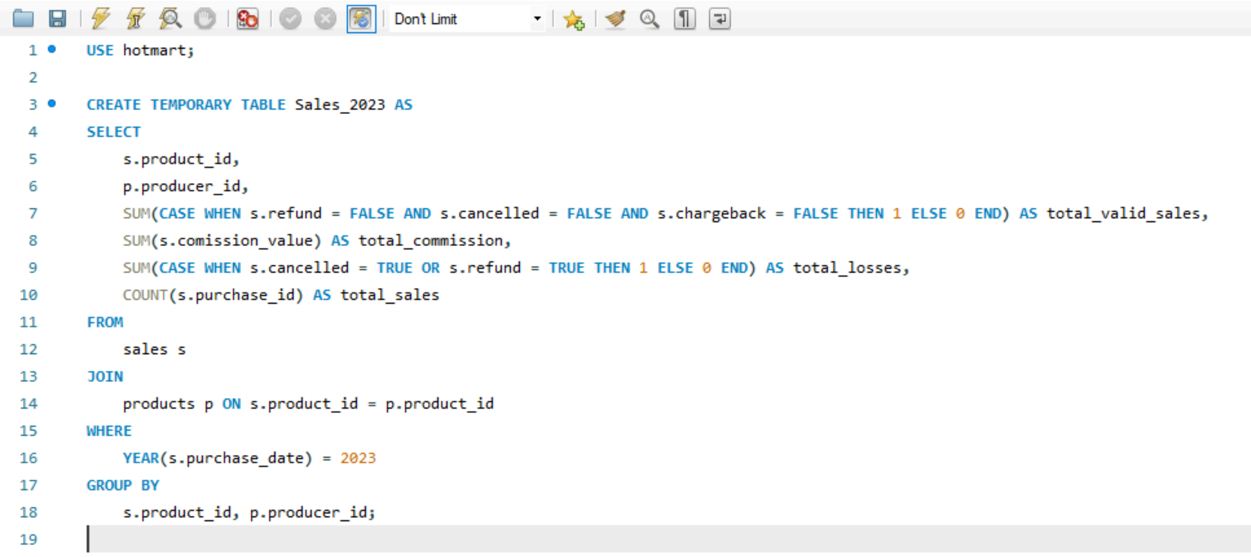
- **Products with active recovery**: Average commission retention = 0.82
- **Products without active recovery**: Average commission retention = 0.82

Conclusion: There is no difference in the average commission retention between products with active recovery and those without active recovery, as both show an average retention of 0.82.

Additionally, the **retention variation recovery** is close to zero, at 0.01.

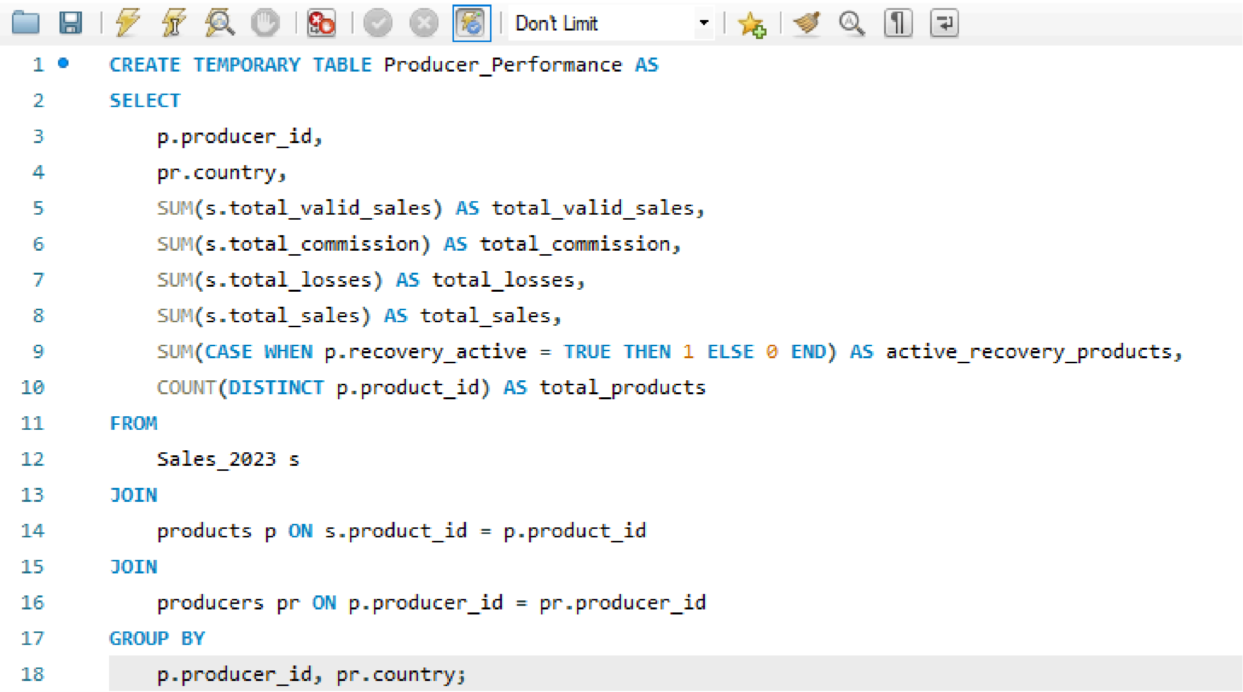
6. If you need to create a ranking of the top creators of 2023, which variables you consider crucial for ranking them? You can also create variables from the data. You must explain your reasoning and your choice of variables and show how this reflect in your SQL code.

   Step 1: Filter valid sales from 2023 Creates a temporary table with information about valid sales, cancellations, and commissions for each product.



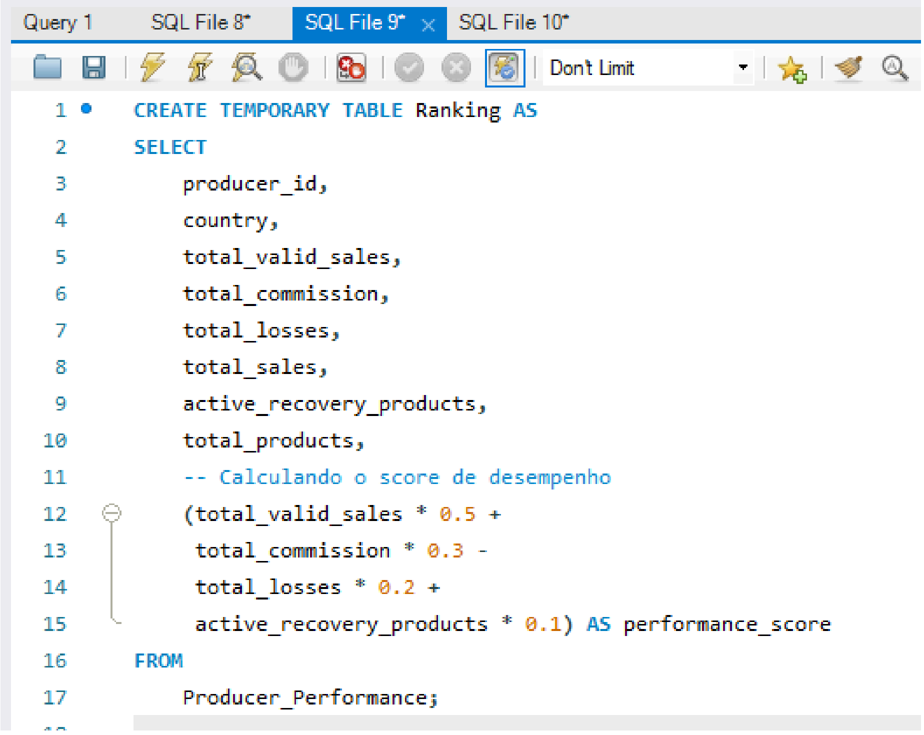
```sql
USE hotmart;

CREATE TEMPORARY TABLE Sales_2023 AS
SELECT
    s.product_id,
    p.producer_id,
    SUM(CASE WHEN s.refund = FALSE AND s.cancelled = FALSE AND s.chargeback = FALSE THEN 1 ELSE 0 END) AS total_valid_sales,
    SUM(s.comission_value) AS total_commission,
    SUM(CASE WHEN s.cancelled = TRUE OR s.refund = TRUE THEN 1 ELSE 0 END) AS total_losses,
    COUNT(s.purchase_id) AS total_sales
FROM
    sales s
JOIN
    products p ON s.product_id = p.product_id
WHERE
    YEAR(s.purchase_date) = 2023
GROUP BY
    s.product_id, p.producer_id;

```

Step 2: Consolidate producer performance Group data by producer to calculate overall performance.



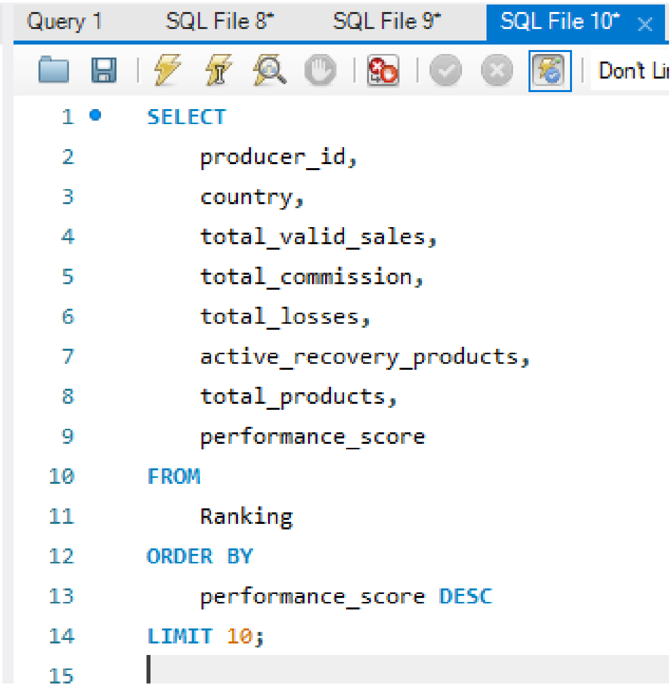
```sql
CREATE TEMPORARY TABLE Producer_Performance AS
SELECT
    p.producer_id,
    pr.country,
    SUM(s.total_valid_sales) AS total_valid_sales,
    SUM(s.total_commission) AS total_commission,
    SUM(s.total_losses) AS total_losses,
    SUM(s.total_sales) AS total_sales,
    SUM(CASE WHEN p.recovery_active = TRUE THEN 1 ELSE 0 END) AS active_recovery_products,
    COUNT(DISTINCT p.product_id) AS total_products
FROM
    Sales_2023 s
JOIN
    products p ON s.product_id = p.product_id
JOIN
    producers pr ON p.producer_id = pr.producer_id
GROUP BY
    p.producer_id, pr.country;
```

Step 3: Calculate the ranking based on the score Add a performance score to rank the producers.



```sql
CREATE TEMPORARY TABLE Ranking AS
SELECT
    producer_id,
    country,
    total_valid_sales,
    total_commission,
    total_losses,
    total_sales,
    active_recovery_products,
    total_products,
    -- Calculando o score de desempenho
    (total_valid_sales * 0.5 +
     total_commission * 0.3 -
     total_losses * 0.2 +
     active_recovery_products * 0.1) AS performance_score
FROM
    Producer_Performance;
```

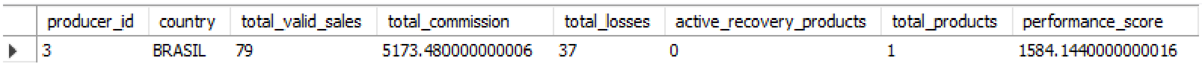
4- Step 4: Display the final ranking Returns the top 10 producers for 2023, based on performance.



```sql
SELECT
    producer_id,
    country,
    total_valid_sales,
    total_commission,
    total_losses,
    active_recovery_products,
    total_products,
    performance_score
FROM
    Ranking
ORDER BY
    performance_score DESC
LIMIT 10;
```

Summary of Steps

1. **Filter valid sales for 2023:**
   The first step is to filter the sales data for 2023, ensuring that only valid sales are included. This means excluding sales that were canceled, refunded, or had chargebacks. We create a dataset containing key metrics such as valid sales, cancellations, and commissions for the year 2023.
2. **Consolidate performance by producer:**
   In this step, we aggregate the relevant metrics by producer and country. This involves grouping the data by each producer and summarizing the total sales and total commission earned by each producer within the year 2023. This consolidation allows us to assess the overall performance of each producer in terms of sales and commissions.
3. **Calculate the ranking:**
   After consolidating the performance data, we calculate a performance score for each producer. This score is typically based on their total sales, commissions, and other performance metrics. It helps rank the producers in terms of their overall success for the given period.
4. **Display top producers:**
   Finally, we rank and display the top-performing producers based on their calculated performance scores. These top producers are displayed in descending order, showcasing the producers who achieved the highest performance, making it easy to identify the leaders in the dataset.



	producer_id	country	total_valid_sales	total_commission	total_losses	active_recovery_products	total_products	performance_score
▶	3	BRASIL	79	5173.480000000006	37	0	1	1584.144000000016

written codes

Filter valid sales for 2023:

```
USE hotmart;

CREATE TEMPORARY TABLE Sales_2023 AS
SELECT
   s.product_id,
   p.producer_id,
   SUM(CASE WHEN s.refund = FALSE AND s.cancelled = FALSE AND
s.chargeback = FALSE THEN 1 ELSE 0 END) AS total_valid_sales,
   SUM(s.comission_value) AS total_commission,
   SUM(CASE WHEN s.cancelled = TRUE OR s.refund = TRUE THEN 1 ELSE 0
END) AS total_losses,
   COUNT(s.purchase_id) AS total_sales
FROM
   sales s
JOIN
   products p ON s.product_id = p.product_id
```

```
WHERE
   YEAR(s.purchase_date) = 2023
GROUP BY
   s.product_id, p.producer_id;
```

Consolidate performance by producer:

```
CREATE TEMPORARY TABLE Producer_Performance AS
SELECT
   p.producer_id,
   pr.country,
   SUM(s.total_valid_sales) AS total_valid_sales,
   SUM(s.total_commission) AS total_commission,
   SUM(s.total_losses) AS total_losses,
   SUM(s.total_sales) AS total_sales,
   SUM(CASE WHEN p.recovery_active = TRUE THEN 1 ELSE 0 END) AS
active_recovery_products,
   COUNT(DISTINCT p.product_id) AS total_products
FROM
   Sales_2023 s
JOIN
   products p ON s.product_id = p.product_id
JOIN
   producers pr ON p.producer_id = pr.producer_id
GROUP BY
   p.producer_id, pr.country;
```

Calculate the ranking:

```
CREATE TEMPORARY TABLE Ranking AS
SELECT
   producer_id,
   country,
   total_valid_sales,
   total_commission,
   total_losses,
   total_sales,
   active_recovery_products,
   total_products,
   -- Calculando o score de desempenho
   (total_valid_sales * 0.5 +
    total_commission * 0.3 -
    total_losses * 0.2 +
    active_recovery_products * 0.1) AS performance_score
FROM
   Producer_Performance;
```

Display top producers:

```
SELECT
    producer_id,
    country,
    total_valid_sales,
    total_commission,
    total_losses,
    active_recovery_products,
    total_products,
    performance_score
FROM
    Ranking
ORDER BY
    performance_score DESC
LIMIT 10;
```