Technical case | Straight forward questions

During the preparation of this case, I prioritized total transparency regarding the methods used. Therefore, each question is followed by the development of the answer, which includes the SQL query used and the results obtained. Additionally, I have provided a graphical visualization to make it easier for the reader to understand.

Before addressing the issues in the *case,* I consider it important to highlight the strategy adopted to resolve them.

Initially, after carefully reading the instructions, I chose to start the resolution with the technical *case.* To do this, I revisited the instructions to better understand the datasets and items that would be evaluated in this stage.

To organize the resolution process, I created a repository dedicated to the technical case and started some essential tasks. First, I established a SQL database integrating the three tables: Producer, Product and Sales. I then thoroughly examined the tables for null or empty data, row repetitions, typos, incorrect data types, and other relevant aspects.

During this stage of exploration, I identified a single significant inconsistency: the presence of niches with different names, such as ' Financas ' and 'Finança', as well as "Tecnologia e Informação and 'Tecnologia-e-Inovação".

To resolve this inconsistency, I implemented the following query:

```
query_ ="""
UPDATE Product
SET niche = CASE
WHEN niche = 'Financas' THEN 'Finanças'
WHEN niche = 'Tecnologia-e-Inovação' THEN 'Tecnologia e Inovação'
ELSE niche
END;
"""
```

After ensuring that all the data was correct and without inconsistencies, I moved on to resolving the issues proposed in the case.

Resolution of Direct Questions - Technical Case

1. **The top 10 products that sell the most in each niche with deactivated membership area and activated recovery.**

    a. **Development**

    After understanding the question about the number of sales made, I proceeded to count how many times each product, categorized by niche, appears in the table.

    Before creating the query for the answer, I examined the number of products per niche in the table.

```
query_q1_1 = """
SELECT niche, COUNT(*) AS count
FROM Product
GROUP BY niche
ORDER BY
count DESC;
"""
```

	niche	count
0	Finanças	15
1	Saúde e Alimentação	10
2	Tecnologia e Inovação	8
3	Empreendedorismo	7
4	Marketing	6
5	Educação Infantil	4
6	Viagens	2
7	Artes e Design	2
8	Idiomas	1

    When analyzing the query results, I noticed that the niches with at least 10 products are Finance (with 15 products) and Health and Food (with 10 products).

    b. **Query**

```
query_q1_2 = """
SELECT
    p.niche,
    p.product_id,
    COUNT(s.product_id) AS total_sales
FROM
    Product p
    JOIN Sales s ON p.product_id = s.product_id
WHERE
    p.recovery_active = 1
    AND p.member_area_active = 0
GROUP BY
    p.niche,
    p.product_id
```

```
ORDER BY
      p.niche,
      total_sales DESC;
"""
```

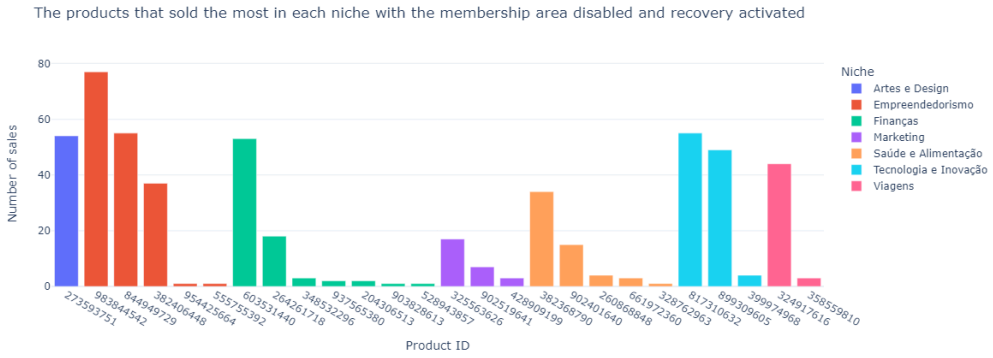
c. Query result

	niche	product_id	total_sales
0	Artes e Design	273593751	54
1	Empreendedorismo	983844542	77
2	Empreendedorismo	844949729	55
3	Empreendedorismo	382406448	37
4	Empreendedorismo	954425664	1
5	Empreendedorismo	555755392	1
6	Finanças	603531440	53
7	Finanças	264261718	18
8	Finanças	348532296	3
9	Finanças	937565380	2
10	Finanças	204306513	2
11	Finanças	903828613	1
12	Finanças	528943857	1
13	Marketing	325563626	17
14	Marketing	902519641	7
15	Marketing	428909199	3
16	Saúde e Alimentação	382368790	34
17	Saúde e Alimentação	902401640	15
18	Saúde e Alimentação	260868848	4
19	Saúde e Alimentação	661972360	3
20	Saúde e Alimentação	328762963	1
21	Tecnologia e Inovação	817310632	55
22	Tecnologia e Inovação	899309605	49
23	Tecnologia e Inovação	399974968	4
24	Viagens	324917616	44
25	Viagens	358559810	3

d. Final answer

From the results of the consultation, it was observed that it was not possible to identify the top 10 in any niche due to the total number of products per category and the established conditions.

To provide a clearer understanding of the best-selling products, I chose to create a graphical visualization.



The products that sold the most in each niche with adhesion area disabled and recovery enabled were:

- ‘Artes e Design’:
  - Product_ID = 273593751.
  - Total = 1.
- ‘Empreendedorismo’:
  - Product_ID = 983844542, 844949729, 382406448, 954425664 and 555755392.
  - Total = 5.
- ‘Finanças’:
  - Product_ID = 603531440, 264261718, 348532296, 937565380, 204306513, 903828613 and 528943857.
  - Total = 7.
- ‘Marketing’:
  - Product_ID = 325563626, 902519641 and 428909199.
  - Total = 3.
- ‘Saúde e Alimentação’:
  - Product_ID = 382368790, 902401640, 260868848, 661972360 and 328762963.
  - Total = 5.
- ‘Tecnologia e Inovação’:
  - Product_ID = 817310632, 899309605 and 399974968.
  - Total = 3.
- ‘Viagens’:
  - Product_ID = 324917616 and 358559810.
  - Total = 2.

2. **The top 10 producers who joined platform from 2020 onwards and achieved the highest sales using recovery.**

   a. **Development**

   Before preparing the query, I conducted research to determine the total number of producers present in the database, as well as how many of them joined the platform as of 2020.

   The research results revealed that the database includes a total of 12 producers, of which only 4 were registered as of 2020.

   b. **Query**

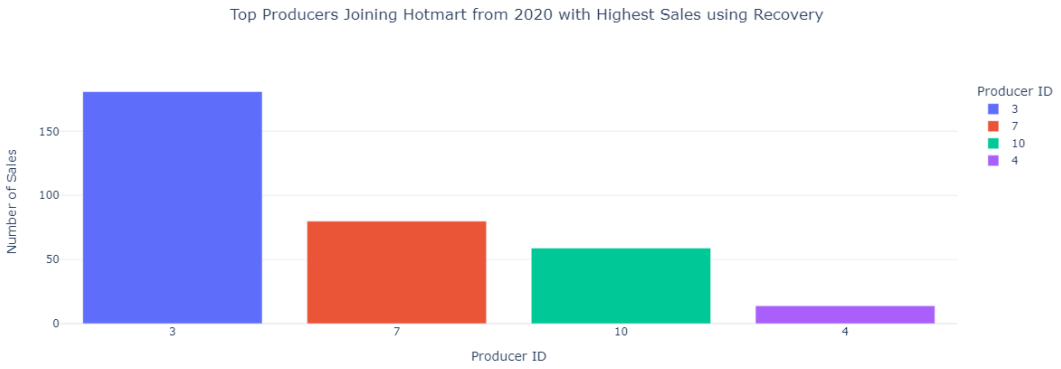
```
q query_q2 = """
SELECT
    p.producer_id,
    COUNT(s.product_id) AS total_sales
FROM
    Product p
    JOIN Sales s ON p.product_id = s.product_id
    JOIN Producer pr ON p.producer_id = pr.producer_id
WHERE
    pr.registry_date >= '2020-01-01'
    AND p.recovery_active = 1
GROUP BY
    pr.producer_id
ORDER BY
    total_sales DESC
LIMIT 10;
"""
```

   c. **Query result**

	producer_id	total_sales
0	3	181
1	7	80
2	10	59
3	4	14

d. **Final anwser**

From the results of the consultation, we can conclude that the main producers who joined the platform from 2020 onwards and achieved the highest sales using the recovery functionality are those identified 3, 7, 10 and 4.



3. **How much more a producer with the recovery feature activated is likely to sell in each niche? Consider only producers who registered from 2020 onwards.**

   a. **Development**

   After analyzing the data from the database, I started searching for information [query_q3_1] that would allow us to compare the average sales of producers with the recovery feature activated in each market niche with the average sales of producers without this feature. However, the maximum amount of information obtained was the result of the query [result_query_q3_1].

   Additionally, I considered analyzing data prior to 2020 to make a possible projection and comparison of sales with the recovery tool activated and deactivated. The results obtained are in [result_query_q3_2].

   b. **Query**

      i. **[query_q3_1]**

```
query_q3_1 = """
SELECT
    niche,
    recovery_active,
    AVG(total_sales) AS average_total_sales,
    AVG(total_sales_cancelled) AS
average_total_sales_cancelled,
    AVG(total_sales_refund) AS
average_total_sales_refund
```

```
FROM
    (
    SELECT
        pr.producer_id,
        p.niche,
        p.recovery_active,
        COUNT(s.product_id) AS total_sales,
        SUM(CASE WHEN s.cancelled = 1 THEN 1 ELSE 0
END) AS total_sales_cancelled,
        SUM(CASE WHEN s.refund = 1 THEN 1 ELSE 0 END)
AS total_sales_refund
    FROM
        Product p
    JOIN
        Sales s ON p.product_id = s.product_id
    JOIN
        Producer pr ON p.producer_id = pr.producer_id
    WHERE
        pr.registry_date >= '2020-01-01'
    GROUP BY
        pr.producer_id,
        p.niche,
        p.recovery_active
    ) AS subquery
GROUP BY
    niche,
    recovery_active;
"""
```

ii. **[query_q3_2]**

```
query_q3_2 = """
SELECT
    niche,
    recovery_active,
    ROUND(AVG(total_sales),1) AS average_total_sales,
    ROUND(AVG(total_sales_cancelled),1) AS
average_total_sales_cancelled,
    ROUND(AVG(total_sales_refund),1) AS
average_total_sales_refund
FROM
    (
    SELECT
        pr.producer_id,
```

```
        p.niche,
        p.recovery_active,
        COUNT(s.product_id) AS total_sales,
        SUM(CASE WHEN s.cancelled = 1 THEN 1 ELSE 0
END) AS total_sales_cancelled,
        SUM(CASE WHEN s.refund = 1 THEN 1 ELSE 0 END)
AS total_sales_refund
    FROM
        Product p
    JOIN
        Sales s ON p.product_id = s.product_id
    JOIN
        Producer pr ON p.producer_id = pr.producer_id
    WHERE
        pr.registry_date < '2020-01-01'
    GROUP BY
        pr.producer_id,
        p.niche,
        p.recovery_active
    ) AS subquery
GROUP BY
    niche,
    recovery_active;
"""
```

c. Query result

i. [result_query_q3_1]

	niche	recovery_active	average_total_sales	average_total_sales_cancelled	average_total_sales_refund
0	Artes e Design	1	56.0	3.0	3.0
1	Educação Infantil	1	14.0	0.0	0.0
2	Empreendedorismo	1	132.0	3.0	6.0
3	Finanças	1	3.0	0.0	0.0
4	Saúde e Alimentação	1	39.0	3.0	4.0
5	Tecnologia e Inovação	1	45.0	2.5	0.5

ii. [result_query_q3_2]

	niche	recovery_active	average_total_sales	average_total_sales_cancelled	average_total_sales_refund
0	Empreendedorismo	1	13.8	0.5	1.0
1	Finanças	0	1.0	0.0	0.0
2	Finanças	1	23.8	0.5	1.3
3	Marketing	1	128.5	5.5	2.5
4	Saúde e Alimentação	1	43.0	0.0	1.0
5	Tecnologia e Inovação	0	101.0	12.0	1.0
6	Tecnologia e Inovação	1	23.3	1.3	2.0
7	Viagens	1	47.0	1.0	4.0

d. Final answer

Based on the [result_query_q3_1] data, the sales conversion rate was calculated as a metric to evaluate the producer's success when the recovery tool is activated:

- ‘Arte e Design’:
    - Average success = 50
    - Average cancellation/refund = 6
    - Average conversion = ~89.3%
- ‘Educação Infantil’:
    - Average success = 14
    - Average cancellations/refunds = 0
    - Average conversion = 100.0%
- ‘Empreendedorismo’:
    - Average success = 132
    - Average cancellation/refund = 9
    - Average conversion = 93.2%
- ‘Finanças’:
    - Average success = 3
    - Average cancellation/refund = 0
    - Average conversion = 100.0%
- ‘Saúde e Alimentação’:
    - Average success = 39
    - Average cancellation/refund = 7
    - Average conversion = ~82.0%
- ‘Tecnologia e Inovação’:
    - Average success = 45
    - Average cancellation/refund = 3
    - Average conversion = ~93.3%

After analyzing the [result_query_q3_2] table, we observed that the 'Finance' and 'Technology and Innovation' niches offer both recovery tool options (activated/disabled). Although cancellation and refund data were included to enrich the analysis, I chose not to consider it in this context.

Based on the results obtained, we concluded that the 'Finance' niche, in years prior to 2020, presented an average result of 2,380% in the number of sales for products with the recovery tool activated, compared to products in which the tool was disabled. On the other hand, the 'Technology and Innovation' niche showed a greater number of sales for products where the recovery tool was disabled.

I decided not to consider the data above to answer the question, as I was unable to find measures that would help make this projection in 2020 data.

4. **The product niche(s) with the highest number of cancellations and refunds.**
    a. **Query**

```
query_q4 = """
SELECT
    p.niche,
    COUNT(s.product_id) AS total_sales
FROM
    Product p
    JOIN Sales s ON p.product_id = s.product_id
WHERE
    s.cancelled = 1
    OR s.refund = 1
GROUP BY
    p.niche
ORDER BY
    total_sales DESC;
"""
```

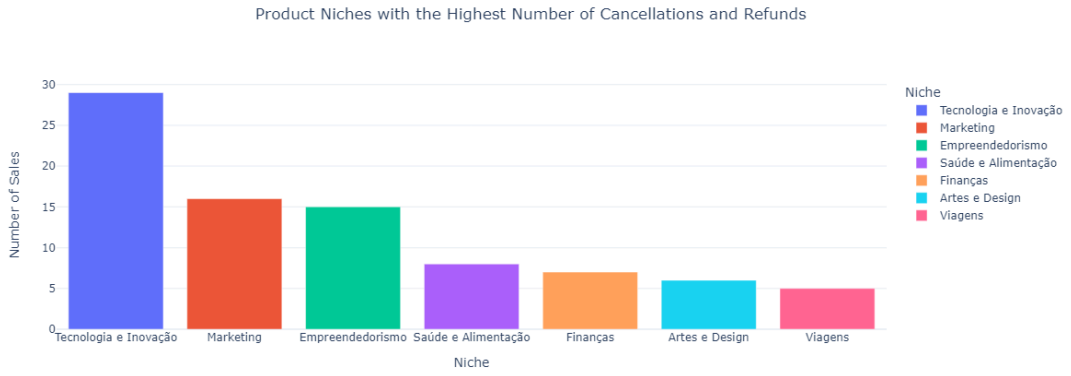
b. Query result

	niche	total_sales
0	Tecnologia e Inovação	29
1	Marketing	16
2	Empreendedorismo	15
3	Saúde e Alimentação	8
4	Finanças	7
5	Artes e Design	6
6	Viagens	5

c. Final answer

The top three product niches with the highest number of cancellations and refunds are Technology and Innovation, Marketing and Entrepreneurship. Next in order we have Health and Food, Finance, Art and Design and Travel.

For a better visualization of the data, I present the corresponding graph below.



5. Calculate the total money lost by producer due to cancellations and refunds. Is there any difference for producers considering products with the recovery tool activated?

a. Development

In this case, I planned to create a query that would provide me with data on the total amount lost due to cancellations and refunds, as well as bring these amounts with the condition of whether the recovery tool is activated or not. This way, it would be possible to easily identify the values that consider the activation of this tool.

b. **Query**

i. **[query_q5_1]**

```
query_q5_1 = """
SELECT
    SUM(CASE WHEN s.cancelled = 1 OR s.refund = 1 THEN
s.product_price ELSE 0 END) AS total_lost_money
FROM
    Sales s
    JOIN Product p ON s.product_id = p.product_id
    JOIN Producer pr ON p.producer_id = pr.producer_id
"""
```

ii. **[query_q5_2]**

```
query_q5_2 = """
SELECT
    pr.producer_id,
     SUM(CASE WHEN s.cancelled = 1 OR s.refund = 1 THEN
s.product_price ELSE 0 END) AS total_lost_money
FROM
    Sales s
    JOIN Product p ON s.product_id = p.product_id
    JOIN Producer pr ON p.producer_id = pr.producer_id
GROUP BY
    pr.producer_id
"""
```

c. **Query result**

i. **[query_q5_1]**

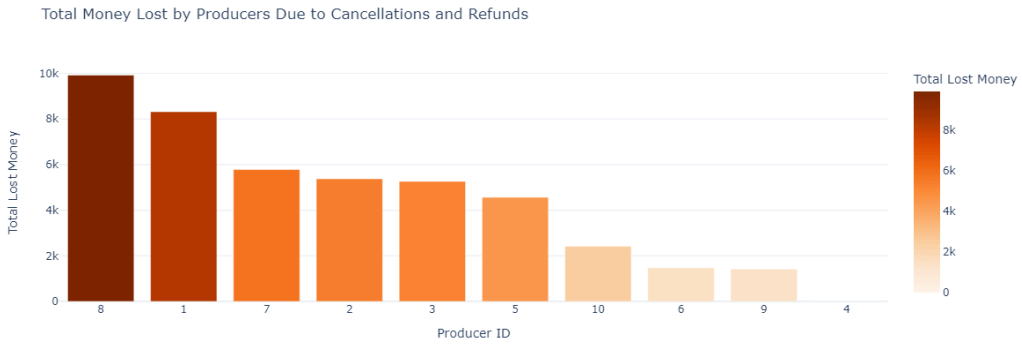
	total_lost_money
0	44541.0

ii. **[query_q5_2]**

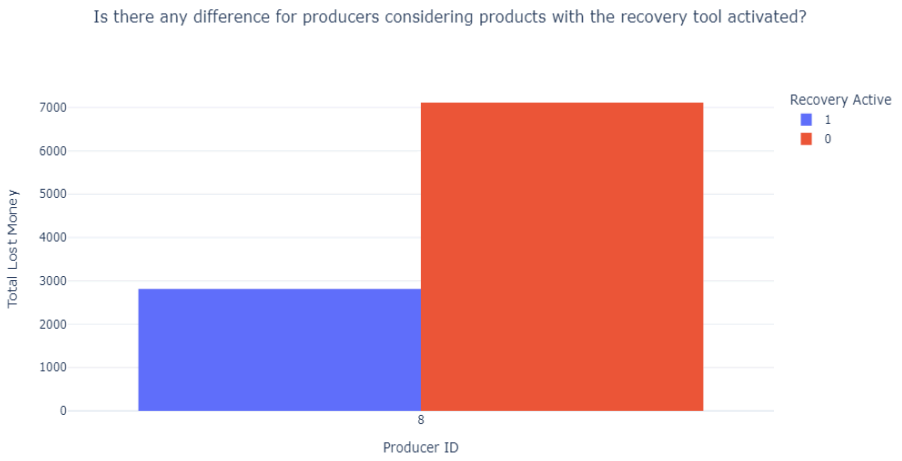
	producer_id	total_lost_money
0	1	8318.0
1	2	5376.0
2	3	5265.0
3	4	0.0
4	5	4566.0
5	6	1470.0
6	7	5782.0
7	8	9924.0
8	9	1421.0
9	10	2419.0

d. Final answer

Based on the consultation carried out, it was found that the total amount lost by producers due to cancellations and refunds was R$44,541.00. Furthermore, when analyzing the results of the second consultation, it is possible to observe the values lost by each producer.



During the consultations, it was found that producer 8 was the one that recorded the biggest losses in sales due to cancellations and refunds. Furthermore, by analyzing producer 8's results in more detail, it is possible to better understand the losses related to whether the recovery tool is activated.



Analysis of the graph above reveals a clear disparity between producers who activate the recovery tool on their products. An example is producer 8, who recorded losses of R$7,112.00 with the recovery tool disabled, compared to just R$2,812.00 when the tool was activated. This represents around 72% of the total lost with products where the recovery tool was not used.

For a more comprehensive assessment, we can examine the number of canceled purchases according to the status of the recovery tool in the case of producer 8.

Furthermore, it is noticeable that the number of canceled or refunded sales tends to be lower when the products are associated with the activation of the recovery tool.

6. **If you need to create a ranking of the top creators of 2023, which variables do 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 reflects in your SQL code.**

   a. **Development and final response**

   In my opinion, evaluating ranking of the top creators of 2023 should consider four essential variables, each playing a significant role. Below, I highlight these variables and explain their impact in detail:

- Time period: It is important to pay attention to the number of specific sales for the year 2023 for creators in this period, providing correct results.
- Total Number of Sales: This direct indicator reflects the overall performance of a creator and plays a vital role in determining their position in the rankings.
- Total Revenue: In addition to the number of sales, total revenue offers a comprehensive view of a creator's financial success during the measured period.
- Cancellation and Refund Rate: A high cancellation and refund rate can be indicative of issues such as customer dissatisfaction, product issues, or underutilization of recovery tools, negatively influencing the creator's reputation.

In addition to these fundamental variables, others can be considered in cases of unusual results or draws:

- Quantity of Products: The variety of products offered can affect the creator's attractiveness to customers and its ability to reach different market segments.
- Diversity of Product Types: Product diversity reflects the creator's ability to meet a wide range of customer needs and explore different market niches, which can positively impact their position in the rankings.
- Uptime: Creators with longer uptime on the platform generally have more opportunities to accumulate sales, which can significantly contribute to their ranking prominence.

b. Query

i. [query_q6_anwer]

```
query_q6_answer = """
SELECT
    pr.producer_id,
    SUM(CASE WHEN s.cancelled = 0 AND s.refund = 0
THEN s.product_price ELSE 0 END) AS total_revenue,
    COUNT(s.product_id) AS total_sales,
    ROUND((SUM(CASE WHEN s.cancelled = 1 OR s.refund =
1 THEN 1 ELSE 0 END) * 1.0) / COUNT(s.product_id), 2)
AS cancellation_refund_rate,
```

```
    ROUND(julianday('now') -
julianday(pr.registry_date),1) AS
activity_time_in_days,
    COUNT(DISTINCT s.product_id) AS
total_unique_products,
    COUNT(DISTINCT p.type) AS total_unique_types
FROM
    Product p
    JOIN Sales s ON p.product_id = s.product_id
    JOIN Producer pr ON p.producer_id = pr.producer_id
WHERE
    pr.registry_date >= '2023-01-01'
GROUP BY
    pr.producer_id
ORDER BY
    total_sales DESC;
"""
```

ii. **[query_q6_example]**

```
query_q6_example = """
SELECT
    pr.producer_id,
      SUM(CASE WHEN s.cancelled = 0 AND s.refund = 0
THEN s.product_price ELSE 0 END) AS total_revenue,
    COUNT(s.product_id) AS total_sales,
     ROUND((SUM(CASE WHEN s.cancelled = 1 OR s.refund =
1 THEN 1 ELSE 0 END) * 1.0) / COUNT(s.product_id), 2)
AS cancellation_refund_rate,
                          ROUND(julianday('now')      -
julianday(pr.registry_date),1)                       AS
activity_time_in_days,
                 COUNT(DISTINCT    s.product_id)    AS
total_unique_products,
    COUNT(DISTINCT p.type) AS total_unique_types
FROM
    Product p
    JOIN Sales s ON p.product_id = s.product_id
    JOIN Producer pr ON p.producer_id = pr.producer_id
GROUP BY
    pr.producer_id
ORDER BY
    total_sales DESC;
"""
```

c. Query Result

i. [result_query_q6_example]

	producer_id	total_revenue	total_sales	cancellation_refund_rate	activity_time_in_days	total_unique_products	total_unique_types
0	1	146271.0	245	0.06	2261.0	7	3
1	3	87750.0	181	0.07	1561.0	3	3
2	8	68858.0	147	0.12	1917.0	5	4
3	5	43585.0	92	0.10	1886.0	3	3
4	7	42792.0	80	0.13	1362.0	4	3
5	2	40796.0	78	0.13	2289.0	8	3
6	6	42416.0	60	0.03	1622.0	7	4
7	10	21655.0	59	0.10	1551.0	3	2
8	9	12110.0	47	0.11	2178.0	2	2
9	4	8806.0	14	0.00	1195.0	2	2