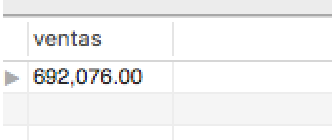
# Learndata Enterprise- Data Cleansing in SQL

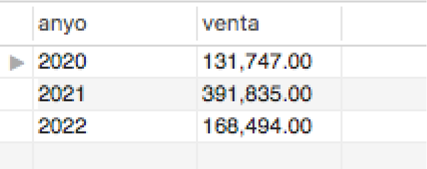
Responses & recommendations from the financial e-commerce analysis

1. **What is the total sales of the company?**



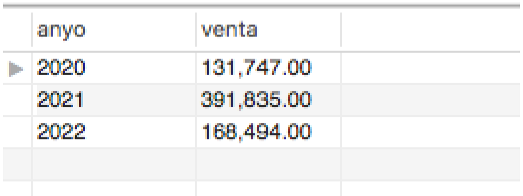
| ventas |
|---|
| 692,076.00 |

2. **What is the total sales per year?**



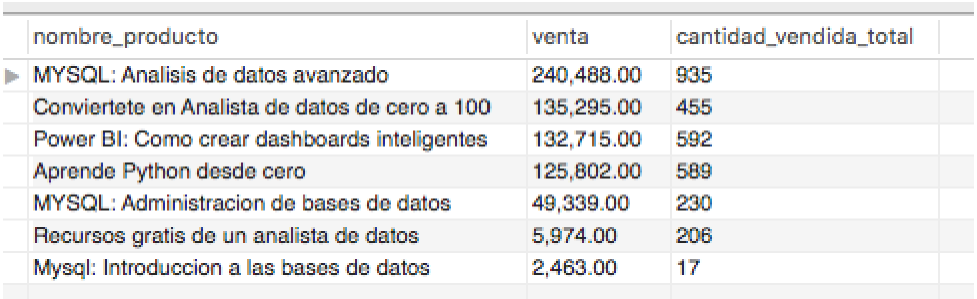
| anyo | venta |
|---|---|
| 2020 | 131,747.00 |
| 2021 | 391,835.00 |
| 2022 | 168,494.00 |

3. **What is the total sales per product?**



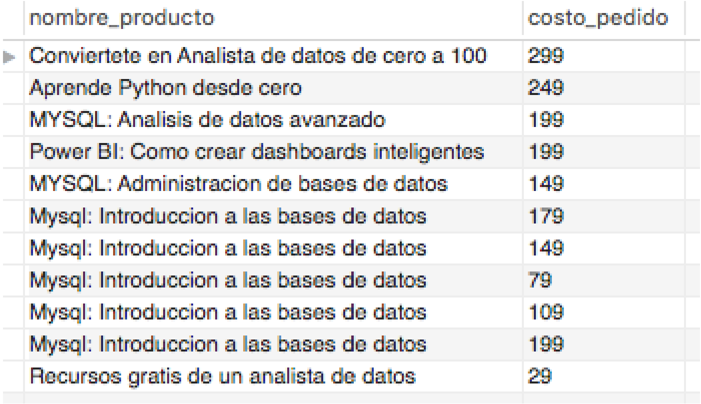
| anyo | venta |
|---|---|
| 2020 | 131,747.00 |
| 2021 | 391,835.00 |
| 2022 | 168,494.00 |

4. **What is the total sales per product and the number of orders placed?**



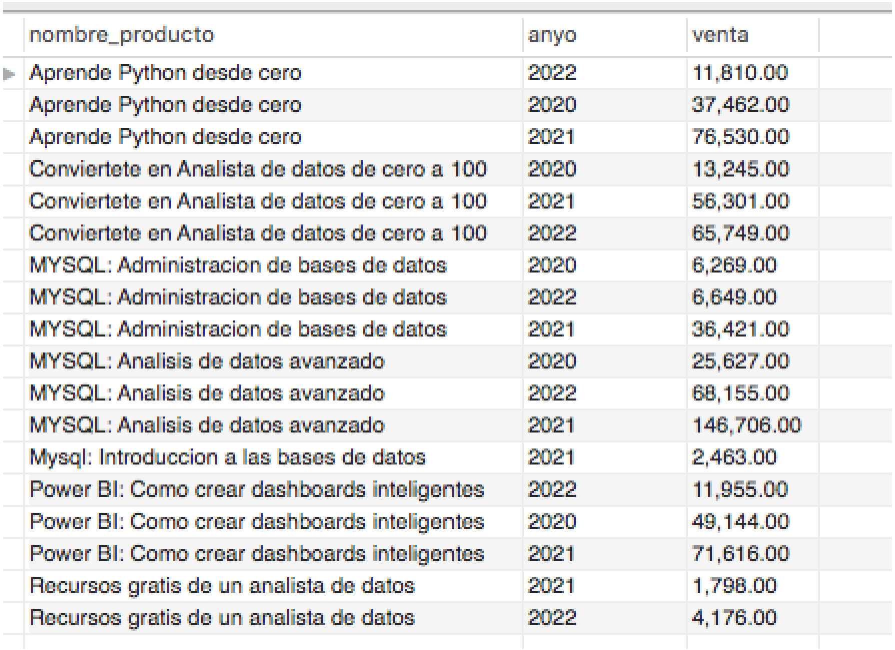
| nombre_producto | venta | cantidad_vendida_total |
|---|---|---|
| MYSQL: Analisis de datos avanzado | 240,488.00 | 935 |
| Conviertete en Analista de datos de cero a 100 | 135,295.00 | 455 |
| Power BI: Como crear dashboards inteligentes | 132,715.00 | 592 |
| Aprende Python desde cero | 125,802.00 | 589 |
| MYSQL: Administracion de bases de datos | 49,339.00 | 230 |
| Recursos gratis de un analista de datos | 5,974.00 | 206 |
| Mysql: Introduccion a las bases de datos | 2,463.00 | 17 |

5. **At what price has each product been sold? Could you get the unique value?**



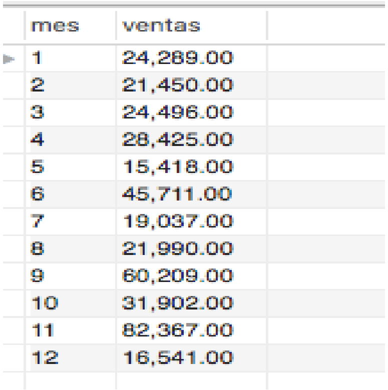
| nombre_producto | costo_pedido |
|---|---|
| Conviertete en Analista de datos de cero a 100 | 299 |
| Aprende Python desde cero | 249 |
| MYSQL: Analisis de datos avanzado | 199 |
| Power BI: Como crear dashboards inteligentes | 199 |
| MYSQL: Administracion de bases de datos | 149 |
| Mysql: Introduccion a las bases de datos | 179 |
| Mysql: Introduccion a las bases de datos | 149 |
| Mysql: Introduccion a las bases de datos | 79 |
| Mysql: Introduccion a las bases de datos | 109 |
| Mysql: Introduccion a las bases de datos | 199 |
| Recursos gratis de un analista de datos | 29 |

6. **To what could we attribute this growth of sales? Could we see the sales by product and by year?**



| nombre_producto | anyo | venta |
|---|---|---|
| Aprende Python desde cero | 2022 | 11,810.00 |
| Aprende Python desde cero | 2020 | 37,462.00 |
| Aprende Python desde cero | 2021 | 76,530.00 |
| Conviertete en Analista de datos de cero a 100 | 2020 | 13,245.00 |
| Conviertete en Analista de datos de cero a 100 | 2021 | 56,301.00 |
| Conviertete en Analista de datos de cero a 100 | 2022 | 65,749.00 |
| MYSQL: Administracion de bases de datos | 2020 | 6,269.00 |
| MYSQL: Administracion de bases de datos | 2022 | 6,649.00 |
| MYSQL: Administracion de bases de datos | 2021 | 36,421.00 |
| MYSQL: Analisis de datos avanzado | 2020 | 25,627.00 |
| MYSQL: Analisis de datos avanzado | 2022 | 68,155.00 |
| MYSQL: Analisis de datos avanzado | 2021 | 146,706.00 |
| Mysql: Introduccion a las bases de datos | 2021 | 2,463.00 |
| Power BI: Como crear dashboards inteligentes | 2022 | 11,955.00 |
| Power BI: Como crear dashboards inteligentes | 2020 | 49,144.00 |
| Power BI: Como crear dashboards inteligentes | 2021 | 71,616.00 |
| Recursos gratis de un analista de datos | 2021 | 1,798.00 |
| Recursos gratis de un analista de datos | 2022 | 4,176.00 |

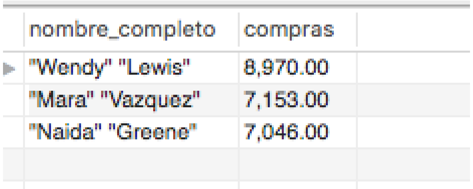
7. **What are the sales by months of the year 2021? Orders the sales from highest to lowest.**



| mes | ventas |
|---|---|
| 1 | 24,289.00 |
| 2 | 21,450.00 |
| 3 | 24,496.00 |
| 4 | 28,425.00 |
| 5 | 15,418.00 |
| 6 | 45,711.00 |
| 7 | 19,037.00 |
| 8 | 21,990.00 |
| 9 | 60,209.00 |
| 10 | 31,902.00 |
| 11 | 82,367.00 |
| 12 | 16,541.00 |

**8. What are the top 3 customers who buy in monetary terms?**

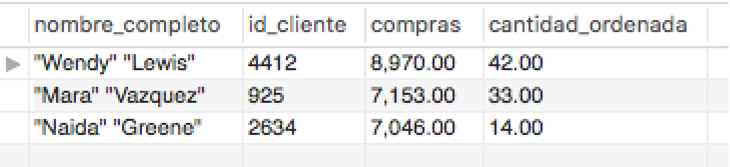
**-- We need to bring the full name with last name in a single field.**



| nombre_completo | compras |
|---|---|
| "Wendy" "Lewis" | 8,970.00 |
| "Mara" "Vazquez" | 7,153.00 |
| "Naida" "Greene" | 7,046.00 |

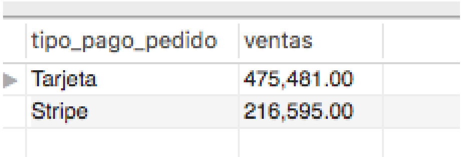
**9. What are the top 3 customers by purchase?**

**-- We need to bring the full name with last name**



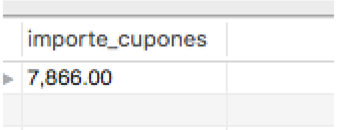
| nombre_completo | id_cliente | compras | cantidad_ordenada |
|---|---|---|---|
| "Wendy" "Lewis" | 4412 | 8,970.00 | 42.00 |
| "Mara" "Vazquez" | 925 | 7,153.00 | 33.00 |
| "Naida" "Greene" | 2634 | 7,046.00 | 14.00 |

**10. What is the most payment method used by  customers (monetary terms?**



| tipo_pago_pedido | ventas |
|---|---|
| Tarjeta | 475,481.00 |
| Stripe | 216,595.00 |

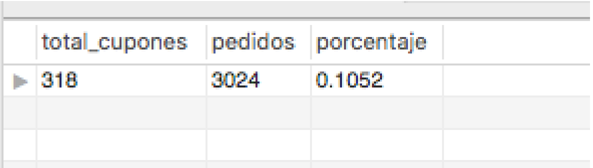
**11. How much is the total spending on coupons?**



| importe_cupones |
|---|
| 7,866.00 |

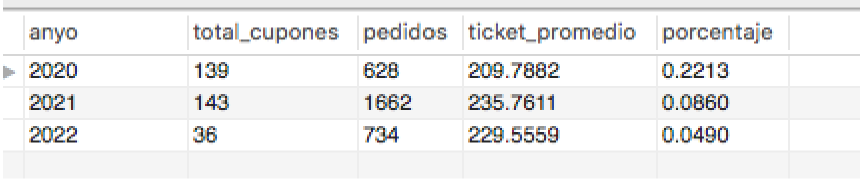
**12. What is the total number of coupons used in sales in quantitative terms?**

**-- Compare it with all sales and calculate the percentage in quantitative terms.**



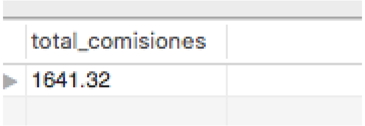
| total_cupones | pedidos | porcentaje |
|---|---|---|
| 318 | 3024 | 0.1052 |

**13.Make the same calculation but broken down by year and calculate the average ticket.**



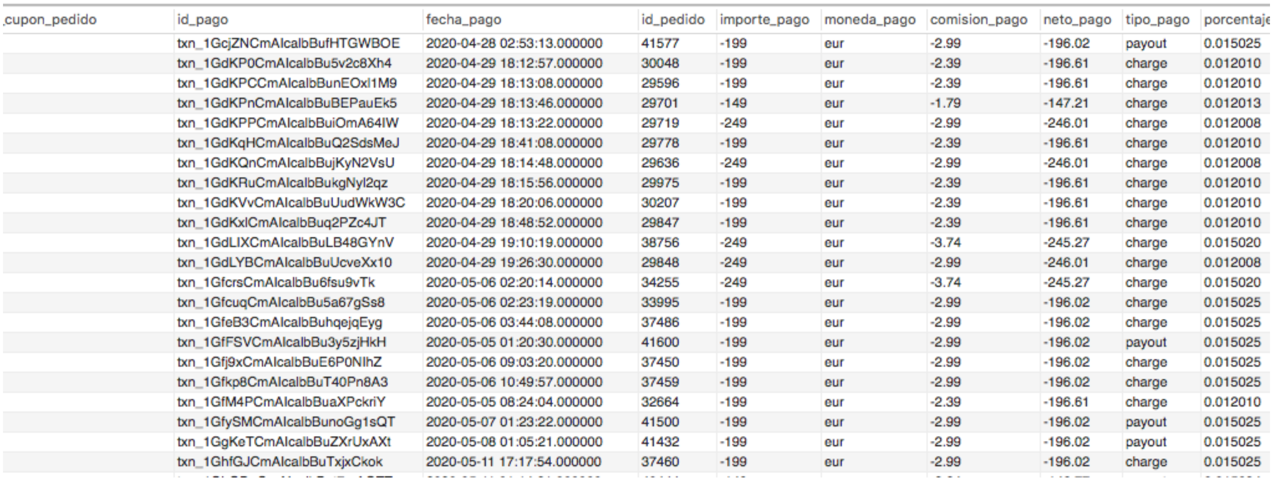
| anyo | total_cupones | pedidos | ticket_promedio | porcentaje |
|---|---|---|---|---|
| 2020 | 139 | 628 | 209.7882 | 0.2213 |
| 2021 | 143 | 1662 | 235.7611 | 0.0860 |
| 2022 | 36 | 734 | 229.5559 | 0.0490 |

**14. What is the total commission paid to stripe?**



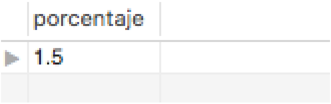
| total_comisiones |
|---|
| 1641.32 |

## 15. What is the commission rate for each order placed by Stripe?



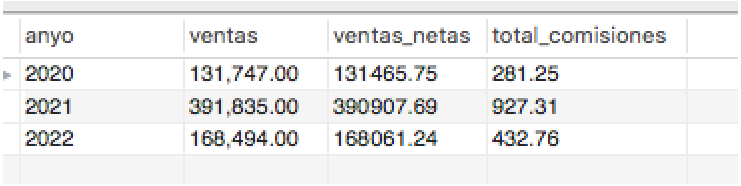
| cupon_pedido | id_pago | fecha_pago | id_pedido | importe_pago | moneda_pago | comision_pago | neto_pago | tipo_pago | porcentaje |
|---|---|---|---|---|---|---|---|---|---|
| | txn_1GcjZNCmAlcalbBufHTGWBOE | 2020-04-28 02:53:13.000000 | 41577 | -199 | eur | -2.99 | -196.02 | payout | 0.015025 |
| | txn_1GdKP0CmAlcalbBu5v2c8Xh4 | 2020-04-29 18:12:57.000000 | 30048 | -199 | eur | -2.39 | -196.61 | charge | 0.012010 |
| | txn_1GdKPCCmAlcalbBunEOxl1M9 | 2020-04-29 18:13:08.000000 | 29596 | -199 | eur | -2.39 | -196.61 | charge | 0.012010 |
| | txn_1GdKPnCmAlcalbBuBEPauEk5 | 2020-04-29 18:13:46.000000 | 29701 | -149 | eur | -1.79 | -147.21 | charge | 0.012013 |
| | txn_1GdKPPCmAlcalbBuiOmA64IW | 2020-04-29 18:13:22.000000 | 29719 | -249 | eur | -2.99 | -246.01 | charge | 0.012008 |
| | txn_1GdKqHCmAlcalbBuQ2SdsMeJ | 2020-04-29 18:41:08.000000 | 29778 | -199 | eur | -2.39 | -196.61 | charge | 0.012010 |
| | txn_1GdKQnCmAlcalbBujKyN2VsU | 2020-04-29 18:14:48.000000 | 29636 | -249 | eur | -2.99 | -246.01 | charge | 0.012008 |
| | txn_1GdKRuCmAlcalbBukgNyl2qz | 2020-04-29 18:15:56.000000 | 29975 | -199 | eur | -2.39 | -196.61 | charge | 0.012010 |
| | txn_1GdKVvCmAlcalbBuUudWkW3C | 2020-04-29 18:20:06.000000 | 30207 | -199 | eur | -2.39 | -196.61 | charge | 0.012010 |
| | txn_1GdKxlCmAlcalbBuq2PZc4JT | 2020-04-29 18:48:52.000000 | 29847 | -199 | eur | -2.39 | -196.61 | charge | 0.012010 |
| | txn_1GdLIXCmAlcalbBuLB48GYnV | 2020-04-29 19:10:19.000000 | 38756 | -249 | eur | -3.74 | -245.27 | charge | 0.015020 |
| | txn_1GdLYBCmAlcalbBuUcveXx10 | 2020-04-29 19:26:30.000000 | 29848 | -249 | eur | -2.99 | -246.01 | charge | 0.012008 |
| | txn_1GfcrsCmAlcalbBu6fsu9vTk | 2020-05-06 02:20:14.000000 | 34255 | -249 | eur | -3.74 | -245.27 | charge | 0.015020 |
| | txn_1GfcuqCmAlcalbBu5a67gSs8 | 2020-05-06 02:23:19.000000 | 33995 | -199 | eur | -2.99 | -196.02 | charge | 0.015025 |
| | txn_1GfeB3CmAlcalbBuhqejqEyg | 2020-05-06 03:44:08.000000 | 37486 | -199 | eur | -2.99 | -196.02 | charge | 0.015025 |
| | txn_1GfFSVCmAlcalbBu3y5zjHkH | 2020-05-05 01:20:30.000000 | 41600 | -199 | eur | -2.99 | -196.02 | payout | 0.015025 |
| | txn_1Gfj9xCmAlcalbBuE6P0NIhZ | 2020-05-06 09:03:20.000000 | 37450 | -199 | eur | -2.99 | -196.02 | charge | 0.015025 |
| | txn_1Gfkp8CmAlcalbBuT40Pn8A3 | 2020-05-06 10:49:57.000000 | 37459 | -199 | eur | -2.99 | -196.02 | charge | 0.015025 |
| | txn_1GfM4PCmAlcalbBuaXPckriY | 2020-05-05 08:24:04.000000 | 32664 | -199 | eur | -2.39 | -196.61 | charge | 0.012010 |
| | txn_1GfySMCmAlcalbBunoGg1sQT | 2020-05-07 01:23:22.000000 | 41500 | -199 | eur | -2.99 | -196.02 | payout | 0.015025 |
| | txn_1GgKeTCmAlcalbBuZXrUxAXt | 2020-05-08 01:05:21.000000 | 41432 | -199 | eur | -2.99 | -196.02 | payout | 0.015025 |
| | txn_1GhfGJCmAlcalbBuTxjxCkok | 2020-05-11 17:17:54.000000 | 37460 | -199 | eur | -2.99 | -196.02 | charge | 0.015025 |

## 16. From the previous year. What is the average of the total percentage rounded to two decimal digits?



| porcentaje |
|---|
| 1.5 |

## 17. Calculate total sales, sales without STRIPE commission and STRIPE commissions per year



| anyo | ventas | ventas_netas | total_comisiones |
|---|---|---|---|
| 2020 | 131,747.00 | 131465.75 | 281.25 |
| 2021 | 391,835.00 | 390907.69 | 927.31 |
| 2022 | 168,494.00 | 168061.24 | 432.76 |

# Reflecting on the process

<u>We have identified the following outliers:</u>

1. Customer with Id_cliente = 3855 made a purchase with order id 38753, with a total amount of €8,236 for a single product. The selling price of the product is €149. Payment was made with Stripe, but the payment is not recorded in the Stripe payment table.
2. Customer with id_cliente = 2666 made a purchase with order id 40794 for a total amount of €5,640. The selling price of the product is €149. Payment was made with Stripe, but the payment is not recorded in the Stripe payment table.
3. Customer with id_cliente = 108 made a purchase with order id 41358 for a total amount of €3,988. The selling price of the product is €219. Payment was made with Stripe, but the payment is not recorded in the Stripe payment table.
4. Customer with id_cliente = 445 made a purchase with order id 41355 for a total amount of €4,460. The selling price of the product is €199. Payment was made with Stripe, but the payment is not recorded in the Stripe payment table.
5. Customer with id_cliente = 917 made a purchase with order id 38798 for a total amount of €4,696. The selling price of the product is €199. Payment was made with Stripe, but the payment is not recorded in the Stripe payment table.
6. Customer with id_cliente = 1800 made a purchase with order id 44333 for a total amount of €4,696. The selling price of the product is €199. Payment was made with Stripe, but the payment is not recorded in the Stripe payment table.
7. Customer with id_cliente = 1834 made a purchase with order id 42004 for a total amount of €4,696. The selling price of the product is €199. Payment was made with Stripe, but the payment is not recorded in the Stripe payment table.
8. Customer with id_cliente = 2646 made a purchase with order id 44182 for a total amount of €4,696. The selling price of the product is €199. Payment was made with Stripe, but the payment is not recorded in the Stripe payment table.

We are uncertain if the company had an issue with the Stripe payment screen, but it is a possibility. We will proceed to remove these records from the order table as they may impact the sales results. Additionally, since these orders were paid with Stripe and are not registered in the Stripe payment table, it is possible that they were not processed. It would be advisable to consult with the customer and, in turn, with Stripe to confirm whether these orders were indeed processed or not.

When it comes to data cleaning, we encountered the following challenges:

1. Products with string errors, meaning they were written in different ways.
2. Confidential information in customer card numbers. We had different payment methods and normalized this data to payment with a card or Stripe.
3. Different date formats and time zones.
4. In the Stripe payment table with raw data, we had numerical values stored as text. This is because decimals are separated by a comma instead of a period. We replaced the commas with periods and converted the text value to a number.
5. In the Stripe payment table containing raw data, there is a column called 'description' that makes data analysis challenging because it contains both string and numeric values in a single row. In this column, there are order numbers and course types. I recommend separating this data into two columns: 1. Course type and 2. Order ID.
6. Outliers and null values: We have different customers who placed orders with unusually high amounts, especially considering the product's selling price. Since payments were made with Stripe, we checked the Stripe payment table to see if these order numbers were processed, and they are not registered. I proceeded to remove these order numbers from the order table as they affect total net sales.
7. In 2022, customer orders decreased by 56%. It is not possible to determine the exact reason since we lack information such as customer experience, customer retention, and website interactions to understand the actions taken before making a purchase.

# Conclusions

1. The total sales of the company over its 3 years of existence amount to €692,076.00.
2. The product with the highest net sales is "MYSQL: Advanced Data Analysis" with a total amount of €240,488.00 and a total of 935 orders. In contrast, the product with the lowest sales is 'Mysql: Introduction to Databases,' with a net amount of €2,463.00 and only 17 orders.
3. The year 2021 stands out as the year with the highest net sales, reaching an amount of €391,835.00. November was the month with the highest net sales at €82,367. This growth is attributed to the introduction of the course 'MYSQL: Advanced Data Analysis,' which experienced a sales surge in 2021, with a year-over-year (YOY) increase of 472%.
4. On the other hand, all products experienced an increase in sales from 2020 to 2021. However, most products showed a decrease from 2021 to 2022. The only products that saw an increase were 'Become a Data Analyst from zero to 100' and 'Free Resources from a Data Analyst,' with a 17% and 132% increase, respectively.
5. The year 2021 represents the year with the highest sales discounts, with a total of 143 coupons created. This had a direct impact on the average ticket, with an average spending of €235.76 per customer, representing a total of 1662 orders. Additionally, the average spending per customer increased by 12% year-over-year from 2020 to 2021. In contrast, in 2022, the number of offered coupons decreased by 75%, the average ticket is €229.56, and there is a total of 734 orders. In other words, orders decreased by 56% in the last year.
6. Credit/debit cards are the most commonly used payment method by customers, while Stripe is the least used. The company pays a total of €1,641.32 in commissions to Stripe. The average commission percentage per order is 1.5%.

# Recommendations

Analyse the reasons for the decrease in sales in 2022, consider the following questions:

1. Why did the product "MYSQL: Advanced Data Analysis" experience a decline of 54%? What is the customer feedback?

2. What is the abandonment rate for the course? What percentage of students successfully completes the program? Are they satisfied with the provided content?

3. The course "MySQL: Introduction to Databases" generated no sales in 2022. What is the reason for this? How much does it cost to keep this content available on the website? Should we retain or remove it from the course catalog?

4. The product "MySQL: Introduction to Databases" has five different prices. Determine the final price, as there is up to a €100 difference in order costs.

# Future Research

- Attempt to maintain a record of payments made by credit card, similar to the way payments through Stripe are tracked.

- Define metrics to measure leads and thus calculate the conversion rate of website visitors. In other words, determine the percentage of customers who made a purchase within a specific period. This allows for the measurement of the effectiveness of sales strategies, such as assessing the impact of discounts through coupons on sales.

- Record costs related to the customer acquisition cost, which may include marketing expenses, discounts, sales costs, among others. Calculating the customer acquisition cost is crucial since acquiring a new customer is typically more expensive than retaining an existing one.