THC - Jr Data Scientist Assessment - Alejandro Castano V

9/17/2025

Key Definitions

Shortage: A shortage occurs when Amazon believes that the supplier did not send the full inventory ordered by Amazon. As a result, Amazon will only pay part of the invoice

Shortage Categories:

Current Shortages: Shortages that occurred within the last 90 calendar days past the due date.

Aged Shortages: Shortages that are older than 90 calendar days past the due date.

Summary

The dataset provided for this case contains multiple rows categorized by the following columns:

- Marketplace
- Invoice Date
- Payment Due Date
- Invoice Status
- Actual Paid Amount
- Paid Amount Currency
- Payee
- Invoice Creation Date
- Randomized Invoice
- Invoice Amount
- Invoice Currency
- Any Deductions
- Quantity Variance Amount
- Price Variance Amount
- Quick Pay Discount Amount
- Randomized Latest Child Invoice
- Randomized PO

However, in order to develop the analysis was required to create the additional columns:

- Invoice year
- Due year
- PaidAmountCorrect
- InvoiceAmountCorrect
- ShortageAmount
- ShortageTime
- ShortageCategory
- Any DeductionsText

Where; Invoice year, Due year, PaidAmountCorrect, InvoiceAmountCorrect and Any DeductionsText don’t include any additional information, they were just made to fix the formats in how the data is presented in order to be used for further calculations.

Where; ShortageAmount, ShortageTime, ShortageCategory adds information or is the result of an operation between exists information.

- **ShortageAmount =** InvoiceAmountCorrect - PaidAmountCorrect
- **ShortageTime =** Payment Due Date - Invoice Creation Date
- **ShortageCategory:** classify the types of shortages, where:
  - **Aged and Current Shortaged** are defined at the beginning of the document, in the key definitions section
  - **No shortage:** means there are no discrepancies in between the paid and invoice amount
  - **Overpaid:** a rare case (seven cases found in 2024) where the difference between invoice and paid amount is negative, this means that the seller/company received more money than expected.

Key questions

1. Total Shortage Amount (in dollars) from the dataset.
2. Annual Breakdown of Shortages to understand trends over time. (The breakdown is done annually, based on the payment due date as the year indicator.)
3. Aged Shortages Amount per year.

Based on the analysis performed, the following key insights were identified:

1. **Total Shortage Amount**
   There is an interesting classification that was found in the dataset, where the column invoice status column presents four classifications:
   - PAID
   - PAID, PRICE_DISCREPANCY
   - PROCESSING, PENDING_AMAZON_ACTION
   - QUEUED_FOR PAYMENT

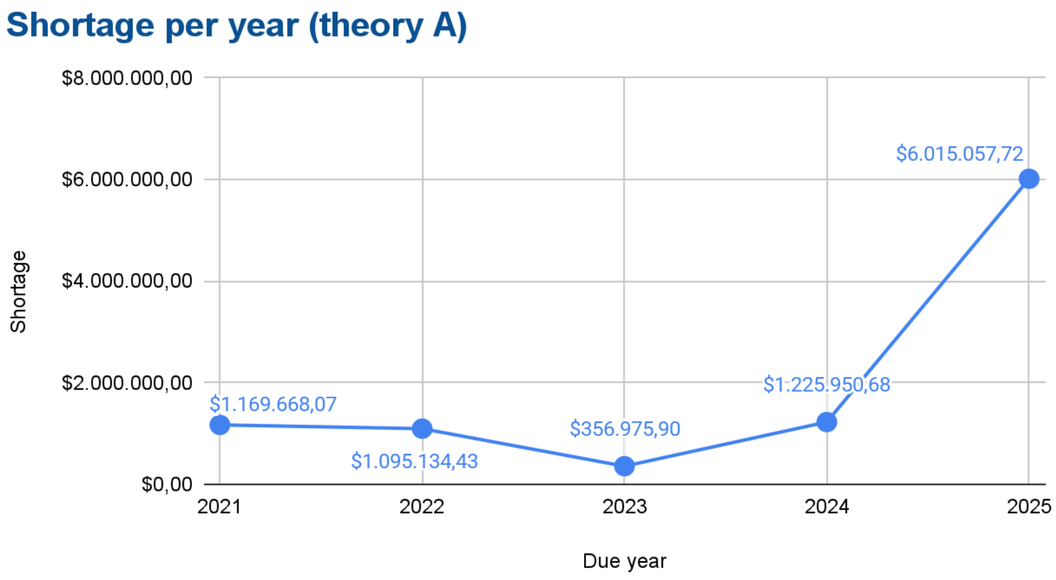
   This will divide this report into **two theories**. The first (A) is where these labels are irrelevant and financial calculations between invoice and paid amount should be prioritized. This theory assumes that the seller/company shipped 100% of the products. The other theory (B) suggests that both financial calculations and invoice status should be prioritized and the seller/company does not deliver 100% of the merchandise or products, however, they consider this in their dataset. Therefore, only orders with a discrepancy between the invoice and paid amount, and with an invoice status PAID_DISCREPANCY, are taken into account when calculating the shortage.

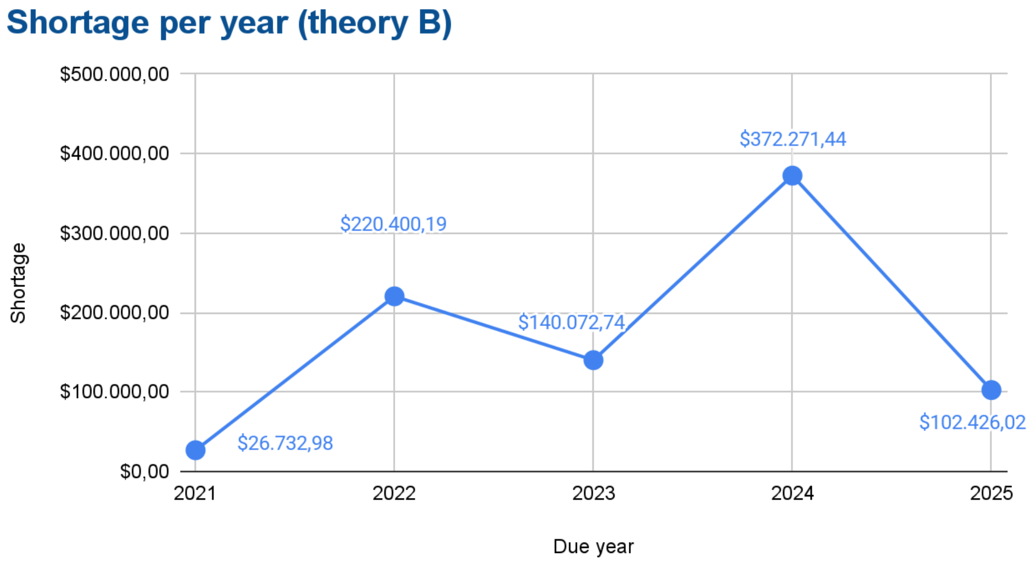
   In summary:
   Theory A: Shortages are calculated only from the financial difference between invoiced and paid amounts, assuming the seller shipped 100% of the products.

   Theory B: Shortages are calculated considering both financial discrepancies and the invoice status, reflecting cases where the seller may not have shipped the full order.

   With this said, the overall shortage identified in the dataset amounts to **$9.862.786,80 in theory A**, while **$861.903,37 in theory B** representing the gap between invoiced amounts and payments received.

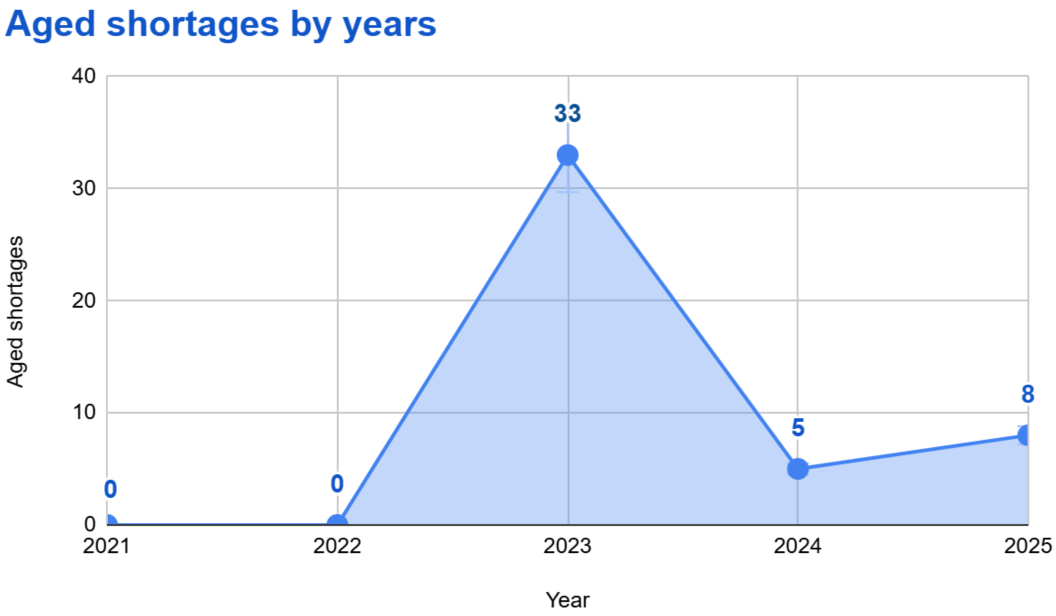
2. **Annual Breakdown of Shortages**
When shortages are analyzed by payment due year, the results show the following distribution:





This annual view highlights the evolution of shortages over time and allows us to better understand potential recurring issues.

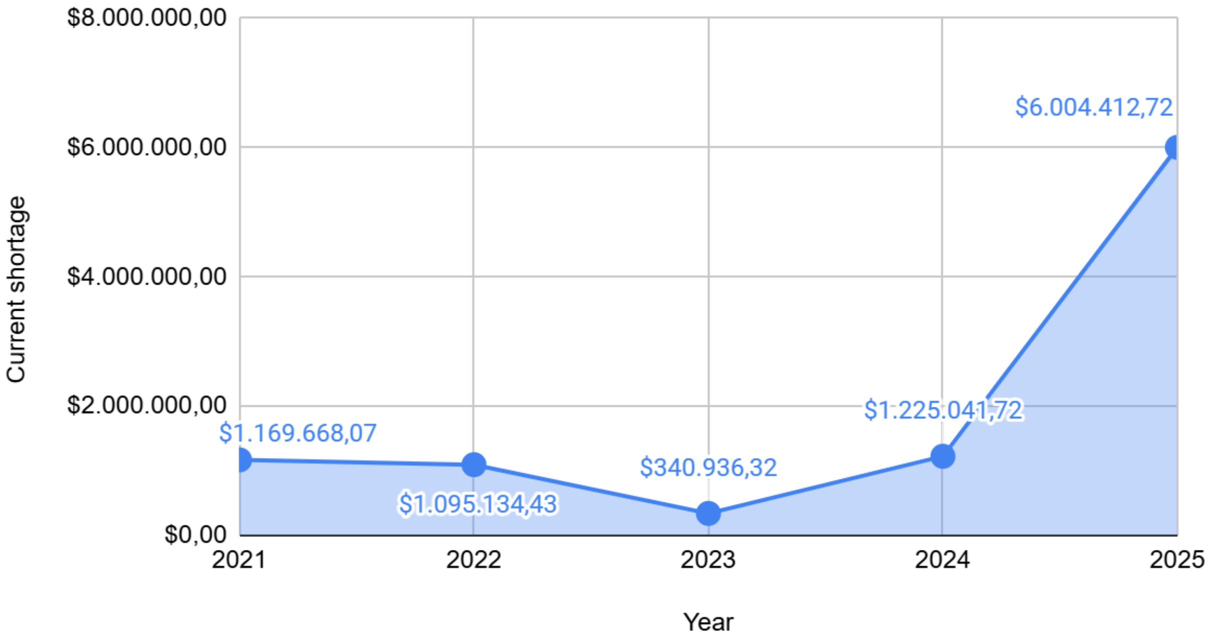
3. **Aged Shortages per Year**
   Focusing on aged shortages (those older than 90 days past the payment due date), the yearly breakdown is as follows:



These figures indicate that a significant portion of shortages remain unresolved beyond the 90-day threshold, emphasizing the importance of addressing aged cases to secure full recovery. Even thought the amounts of cases are low, those cases represents a loss **total loss of $27.850,98:**

- **2023: $16.039,58**
- **2024: $1.166,40**
- **2025: $10.645,00**

Additionally, the graph is presented showing the economic embezzlement generated by the current shortages, which represents a current loss of **$9.835.193,26.**


Current shortages by years

Overall, the analysis confirms that shortages represent a material financial impact, with aged shortages contributing a meaningful share of the total. Addressing these aged cases should be a key priority in order to maximize payment recovery and improve cash flow consistency.

Although both aged and current shortages represent a relevant financial impact, the data clearly shows that current shortages account for the vast majority of the total loss (99,72%, or $9.835.193,26), compared to only 0.28% from aged shortages ($27.850,98). Therefore, recovery efforts should be primarily focused on addressing current shortages, as this is where the greatest opportunity lies to secure payment recovery and significantly improve overall financial results.

[Analysis ends here]

Link where Google sheet was cleaned, developed and analyzed:

Junior Data Scientist Technical test