# Amazon International Apparel Sales Data Cleaning Challenge

Hi, guys. My name is Joseph (Joebass) Edet.

I got this data to teach my students Data Cleaning, turned out it was extremely dirty and not suitable for teaching beginners - it was a chore cleaning it. It was fun though and I liked my approach. So, I thought, why not share it and let other data professionals and enthusiasts have fun.

Decided to make it a challenge. Although there would be no "winners," I'll announce my favourite submissions. It's also a great project to add to your portfolio.

Use any tool, it doesn't matter. But make sure to document your process

Access the data here

Just cleaning the data alone isn't fun, make sure to derive some insights. There's a flexible list of business questions you can use your cleaned data to answer in this document. Pick a few most important and run with it

When you're done cleaning and deriving insights, share a link to your solution here

I'm particularly interested in how you deal with the null values in SKU column 🌚

Here is a list of insightful business questions you can derive from your cleaned sales data, broken down by business area:

## 📊 Sales & Financial Performance

These questions focus on where and when money is being made.

1. **Revenue Drivers:** Which specific **Styles** or **SKUs** account for the top 80% of total gross revenue (Pareto Principle/80-20 Rule)?
2. **Profitability Index:** Which SKUs or Styles generate the highest average **RATE** (price per unit) and **GROSS AMT** per transaction?
3. **Seasonal Trends:** How does monthly **GROSS AMT** change over the year, and which product categories (Styles) are most popular in specific months (Months column)?
4. **Sales Velocity:** What is the average daily or monthly sales volume (PCS) for the top 10 best-selling SKUs?

## 🛍️ Customer & Market Analysis

These questions help you understand who is buying your products and how to engage them.

1. **Top Tier Customers:** Who are the top 10 **CUSTOMERS** based on total revenue generated, and what is the characteristic (Style/Size) of the products they purchase?
2. **Customer Concentration:** What percentage of total revenue comes from the top 5% of your customers?
3. **Customer Basket Analysis:** Do certain Styles or Sizes frequently appear together in the same customer's transaction? (This requires more detailed line-item analysis, but the question is valid).

### 👕 Product & Inventory Management

These questions address the efficiency of your product portfolio and potential stock issues.

1. **Dead Stock/Underperforming Products:** Which Styles have sold less than 50 PCS over the last 12 months in the data, indicating potential products to discontinue?
2. **Popular Size Distribution:** What is the breakdown of sales (`PCS`) by **Size** (S, M, L, XL, etc.) across all products?
3. **SKU Ambiguity (Data Quality):** In the SKU column, are there any **Styles** that map to more than one distinct "Middle Code" (e.g., uses both `-KR-` and `-SKD-`) that should be standardized?
4. **Pricing Consistency:** Is there a significant variance in the **RATE** charged for the same **SKU** to different **CUSTOMERS** or across different months, and if so, why?

# Deadline: Christmas, 2025