Amazon India Sales Analysis Summary

A concise overview of the Amazon India Sales data analysis project — written for the recruiter who wants the full picture in the shortest time.

Tool used across all projects: MySQL Workbench · Datasets sourced from Kaggle

Project 3

Amazon India Sales Analysis — E-Commerce Performance Across 120,000+ Orders

What This Project Was About

This was a commercial dataset — real sales data from Amazon's Indian marketplace, covering over 120,000 orders across product categories, geographies, customer segments, shipping methods, and time periods. I designed 16 SQL queries organised around five business questions: how is the business performing overall, which products and categories drive revenue, where geographically is demand strongest, who are the customers, and how efficient are the operations? The goal was to produce analysis that a business decision-maker could actually use — not just a description of the data, but actionable insight.

What I Discovered

Revenue at scale is real here — ₹78.6 million in total, with over 109,000 orders successfully shipped. But the number that kept drawing my attention was the cancellation rate: 14.21%. That is not a rounding error. At the volume this business operates, a 14% cancellation rate represents a significant and measurable revenue leak, and it is concentrated in the most valuable products and the busiest states. Maharashtra and Karnataka — the top two revenue-generating states — are also among the states with the most cancellations. That tells me the problem is not marginal; it sits right at the heart of operations.

On the product side, the Set category alone accounts for nearly 50% of all revenue, and the top three categories together account for 91%. The business is not diversified — it is built on a small number of high-performing product lines. That is not necessarily a problem, but it means any disruption to stock availability for key SKUs like Style J0230 or JNE3797 would have an outsized impact on overall performance. Protecting those products is not optional.

The B2B vs B2C comparison was one of the more interesting findings. B2C dominates by volume — 119,584 orders vs just 794 for B2B. But B2B customers spend roughly 15% more per order and cancel far less frequently. That is the profile of a high-value, underserved segment. The business has the infrastructure to serve them; it just is not doing so at scale yet.

What I Think Could Be Improved

The promotional analysis (Query 12) is the one area I would revisit. Every single order in the dataset was associated with a promotion, which makes it impossible to draw any conclusions about whether promotions actually drive incremental sales — there is no control group. The finding that 'promotions are effective' cannot really be substantiated with this data. A proper A/B test comparing promotional vs non-promotional periods would be needed to answer that question honestly.

I would also push the geographic analysis further. Knowing that Kerala has a 17.84% cancellation rate is useful, but it raises a follow-up question the data cannot currently answer: is that a logistics

problem, a product availability problem, or a customer behaviour pattern? Layering in fulfilment data by region would sharpen the recommendations considerably.

My Overall Take

What I appreciate most about this project is that it covers the full picture of a business — not just revenue, but operations, geography, customer segments, products, and time. Each query builds on the others. By the time you reach Query 16 (high-value order analysis), you already have enough context from the earlier queries to know exactly why those premium customers behave the way they do. That layered approach is what separates a business intelligence report from a collection of statistics. The SQL itself is technically solid — window functions, CTEs, conditional aggregations — but what I am most proud of is that every query was written to answer a specific business question, not just to demonstrate a technique.