Data Journal: Inventory Analysis - Capstone Project

Date: Apr 29, 2025	**Course/topic:** Capstone Project - Phase 2: Data Preparing
Prompt:	What do I hope to take away from this capstone project? What is one important skill you think you'll learn? Which skills do you most look forward to demonstrating? What are some issues you might encounter?
Journal Entry:	Through this capstone project, I hope to bring together all the skills I've developed into a real-world data analytics case study that I can confidently add to my portfolio. One important skill I believe I'll strengthen is storytelling with data. Using clear visuals and structured reasoning to support insights. I'm especially excited to demonstrate my skills in cleaning messy data, building dashboards, and forming actionable recommendations. I might face challenges with time management or deciding how deep to go in my analysis. or my capstone, I want to focus not just on what the data says, but on how I guide my audience through a journey, from problem to solution.
Other thoughts or questions:	How can I make sure my project stands out to employers?

Date: May 3, 2025	**Course/topic:** Capstone Project - Phase 2: Data Preparing
Prompt:	What challenges do you expect when organizing the data? What tools or strategies did you choose to overcome them?
Journal Entry:	Today I started Phase 2 by importing raw datasets into RStudio. The main challenge I faced was the size of the sales data file, which had over 12 million rows and couldn't be opened in R. To solve this, I switched to Python and used a Jupyter Notebook in VS Code to split the sales data into four files by quarter. This allowed me to import the data in chunks later, avoiding memory issues in R. This situation highlighted the importance of choosing the right tool for each task. While I prefer doing most of the Capstone in R, being flexible and using Python for heavy data operations was crucial.
Other thoughts or questions:	Is it acceptable in a professional context to use multiple tools in a single project, or is that considered bad practice? I wonder whether Big Query should have been my first choice instead of splitting the CSV. Would that have saved me time?

Date: May 5, 2025	**Course/topic:** Capstone Project - Phase 2: Data Preparing
Prompt:	What cleaning and organization techniques did you apply? How did you decide what data to keep, transform, or remove?
Journal Entry:	Today I focused on preparing the purchase, stock, and sales datasets. I grouped them into three logical lists and renamed key columns for consistency. I standardized the formats of text fields by removing white spaces and converting all strings to uppercase. I also converted fields like StoreId, BrandId, and Vendor to character types to prevent mismatches during joints. Using functions like map_dfr (), tabyl (), and skim () helped me explore data quality and spot inconsistencies. I documented missing values and deleted irrelevant columns like Approval or constant date fields.
Other thoughts or questions:	I found that VendorName and VendorNumber don't match perfectly. Should I merge them into a separate dimension table? Some Size values contain packaging details like "3 PK". Should this be split into two variables? What's the best way to document this cleaning logic in a corporate environment?

Date: May 10, 2025	**Course/topic:** Capstone Project - Phase 2: Data Preparing
Prompt:	How did you ensure the data was ready for the next phase? What assumptions did you make, and how did you validate them?
Journal Entry:	Before closing Phase 2, I ran a final verification on keys (InventoryId, PONumber) and ensured that all columns were clean, renamed, and had proper data types. I created summary tables to understand missing data distribution and calculated the percentage of nulls for each dataset. I also created a clean dim inventory table and removed records with missing vendor information and no associated stock. I made the choice to document and store all cleaned datasets as CSV files in a dedicated "/Data/processed" folder. Now that everything is clean and documented, I feel ready to move on to Phase 3: Process.
Other thoughts or questions:	Should I impute the missing volume or package information now or during the modeling phase? Would it be better to use a data catalog or metadata file to document field transformations?

~~**Date: May 25, 2025**~~	~~**Course/topic:** Capstone Project – Phase 3: Data Processing~~
~~**Prompt:**~~	~~What do I hope to take away from this phase of the capstone project?~~ ~~What is one important skill I think I'll learn?~~ ~~Which skills do I most look forward to demonstrating?~~ ~~What are some issues I might encounter?~~
~~**Journal Entry:**~~	~~Today I kicked off Phase 3 by ingesting all my Phase 2 "prepared" CSVs (sales_Q1–Q4, stock_jan/dec, po_prices, purchase_orders, po_invoices) and standardizing every key field to character. I confirmed that InventoryId == paste (Store, City, Brand, sep ="_") in stock_jan with zero mismatches. I committed the typecasting, and derivation checks as reusable R functions.~~ - ~~Takeaway: Mastering an end-to-end ETL pipeline: from CSV —› clean R data frames —› ready for modelling.~~ - ~~Skill: Writing robust, vectorized R code (across (), if else (), anti-joins) that can be re-run safely.~~ - ~~Demonstrating: Referential-integrity testing and modular function design.~~ - ~~Challenges: Handling very large joins purely in memory; edge cases in key construction.~~
~~**Other thoughts or questions:**~~	~~How can I optimize multi-million-row dplyr joins (e.g. chunking, indexing, or using Arrow tables) without relying on an external database?~~

~~**Date: May 25,2025**~~	~~**Course/topic:** Capstone Project – Phase 3: Data Processing~~
~~**Prompt:**~~	~~What do I hope to take away from this phase of the capstone project?~~ ~~What is one important skill I think I'll learn?~~ ~~Which skills do I most look forward to demonstrating?~~ ~~What are some issues I might encounter?~~
~~**Journal Entry:**~~	~~I spent today resolving data-quality snags: harmonizing VendorNumber ↔ VendorName, correcting "TARMSWORTH" → "TAMWORTH", realigning Store 81 →80, and imputing 1 284 missing City entries to "TYWARDREATH." Next, I purged zero-price rows across all sales and purchase tables. Each change was wrapped in clear R scripts with before/after counts logged.~~ - ~~Takeaway: The importance of precise, reproducible data fixes and thorough logging.~~ - ~~Skill: Judicious decision-making on drop vs. impute and applying the same fix across multiple tables.~~ - ~~Demonstrating: Cohesive QA workflows—counts, filters, and vectorized updates.~~ - ~~Challenges: Ensuring no unintended overwrites when applying global replacements.~~
~~**Other thoughts or questions:**~~	~~Could I integrate simple unit tests so that any future data drift or typo re-introduction triggers an R-based alert?~~

~~**Date: May 28, 2025**~~	~~**Course/topic:** Capstone Project – Phase 3: Data Processing~~
~~**Prompt:**~~	~~What do I hope to take away from this phase of the capstone project?~~ ~~What is one important skill I think I'll learn?~~ ~~Which skills do I most look forward to demonstrating?~~ ~~What are some issues I might encounter?~~
~~**Journal Entry:**~~	~~On my final day I built three dimensions—Brand (12 256 SKUs), Inventory (275 882 IDs with on-hand snapshots), and Purchase Order (5 543 headers + freight)—then pruned my fact tables to keys + measures using a single clean_fact() helper. I ran full QA (anti-joins → zero orphans, distinct-price checks → one price per brand) and skimmed each table for sanity.~~ ~~Takeaway: Crafting a clean star schema and automating repetitive tasks with custom functions.~~ ~~Skill: Dimension modelling and automated integrity validation.~~ ~~Challenges: Keeping memory usage in check and managing long run-times on large data frames.~~
~~**Other thoughts or questions:**~~	~~What's the best way to schedule this R script nightly (cron job or RStudio Add-ins) so that every time new raw data lands, Phase 3 runs automatically and notifies me of any integrity failures?~~

~~**Date: May 30, 2025**~~	~~**Course/topic:** Capstone Project – Phase 4: Analyze~~
~~**Prompt:**~~	~~What analytical methods did you apply to the cleaned data?~~ ~~What did you discover about inventory behaviour and purchasing patterns?~~
~~**Journal Entry:**~~	~~Today I finalized the bulk of my Phase 4 analysis. I started by calculating the Inventory Turnover Ratio using adjusted average inventory levels (including purchases). This showed that while many SKUs have healthy turnover rates around 1.5 to 2 per year, there's a significant tail of items with near-zero turnover—clear candidates for delisting or closer review.~~ ~~I also ran an ABC Analysis at both InventoryId and Brand levels. This revealed that roughly 15–20% of SKUs account for ~80% of sales value. That insight will help prioritize attention on the A-class items for stockout prevention and purchasing strategy.~~ ~~Additionally, I mapped stock variation between January and December to identify cities with net stock accumulation vs consumption. Visualizing this geographically highlighted potential overstock issues in specific locations, suggesting opportunities for balancing inventory between stores.~~
~~**Other thoughts or questions:**~~	~~How should I approach recommendations for SKUs with *extreme* coverage months (e.g., >24)?~~ ~~Should they be delisted entirely or managed via discounting strategies?~~

~~**Date: June 15, 2025**~~	~~**Course/topic:** Capstone Project – Phase 4: Analyze~~
~~**Prompt:**~~	~~What challenges did you encounter when exploring the data?~~ ~~How did you address these challenges?~~
~~**Journal Entry:**~~	~~Today I spent most of my time tackling the challenges around data outliers and vendor concentration. Freight cost per unit showed heavy skew—some POs had extremely high per-unit freight. I filtered those out (above 0.5) to produce a cleaner distribution for interpretation, realizing the need to flag those anomalies for procurement review.~~ ~~I also calculated the Herfindahl–Hirschman Index (HHI) for vendor purchases and found a value around 0.06, indicating moderate concentration but with several dominant suppliers. I made sure to visualize both value and quantity shares for the top 20 vendors.~~
~~**Other thoughts or questions:**~~	~~Should I consider building predictive models for demand at the SKU level in Phase 5, or is descriptive analysis sufficient for stakeholder needs?~~

Date: June 29, 2025	**Course/topic:** Capstone Project – Phase 3: Data Processing
Prompt:	What challenges did you encounter that led you to restart Phase 3? How did you recognize that the initial approach wasn't working?
Journal Entry:	Key problems I identified: • Duplicated information: Brand pricing appeared in 3-4 different tables with conflicting values • Complex joins: Needed 5-6 LEFT JOINs just to get basic inventory information • Artificial NAs: InventoryIds that existed in sales but not in certain dimension tables created calculation errors • Quarterly fragmentation: Sales data split across Q1-Q4 made temporal analysis unnecessarily complex The wake-up call came when I spent 2 hours debugging why turnover calculations were giving different results depending on which table I pulled stock data from. That's when I realized I had violated a core data warehouse principle: single source of truth. Takeaway: Sometimes the best decision is to acknowledge that your current approach isn't working and start fresh with a better design. The time spent restarting is far less than the time wasted debugging a flawed architecture.
Other thoughts or questions:	How do I prevent making similar architectural mistakes in future projects? Should I have sketched out the entity relationships before building the tables?

Date: June 30, 2025	**Course/topic:** Capstone Project – Phase 3: Data Processing
Prompt:	What was your new approach and how did it differ from the original design?
Journal Entry:	Key changes made: 1. Consolidated sales data: Combined all quarterly files into one master sales_orders table (12.8M records) 2. Unified inventory dimension: Single dimtable_inventory with 274K records containing ALL inventory attributes 3. Simplified purchases: Streamlined purchase_orders with proper foreign key relationships Immediate improvements: • Faster queries: Single joins instead of complex multi-table operations • Consistent calculations: No more conflicting data sources • Cleaner code: sales_orders %>% left_join(dimtable_inventory, by = "InventoryId") vs. the previous 6-table join nightmare Validation success: The composite key hypothesis (InventoryId = Store_City_Brand) validated with 100% accuracy, which gave me confidence the data structure was sound.
Other thoughts or questions:	This feels much more like a real data warehouse now. Why didn't I start with this simpler approach? Was I overcomplicating things because I thought "more tables = more professional"?

Date: July 1, 2025	**Course/topic:** Capstone Project – Phase 3: Data Processing
Prompt:	How did you handle data quality issues in the revised approach?
Journal Entry:	Today I focused on systematic data quality resolution and master table construction. The simplified architecture made quality issues much easier to identify and fix. Major data quality fixes: 1. Vendor standardization 2. Geographic cleaning 3. Missing value imputation 4. Zero-value removal Quality validation: • Zero missing values in critical fields • 100% referential integrity between fact and dimension tables • Consistent data types across all joins
Other thoughts or questions:	The quality issues were actually easier to spot and fix with the simplified structure. In the old approach, I was spending so much time managing table relationships that I wasn't focusing on the actual data quality. Should I have done this quality assessment earlier in the process? It seems like Phase 2 vs Phase 3 boundaries got blurred when I had to restart

Date: July 2, 2025	**Course/topic:** Capstone Project – Phase 3: Data Processing
Prompt:	What performance improvements did you observe with the new architecture?
Journal Entry:	The performance improvements with the new architecture are dramatic. What used to take several minutes now runs in seconds. • Automated integrity checks show zero orphaned records • Price consistency verified across all brand records • Temporal calculations now consistent across all quarters • All tables ready for skim() summary statistics • Export to Data/cleaned/ directory completed successfully • Foundation set for Phase 4 analytical processing ***** **CRITICAL LESSON LEARNED: MORE TABLES != BETTER ANALYSIS. SOMETIMES THE SIMPLEST SOLUTION THAT MAINTAINS DATA INTEGRITY IS THE BEST SOLUTION** ***********
Other thoughts or questions:	How do I avoid similar architectural mistakes in the future? Should I always start with the minimal viable structure and only add complexity when necessary?

Date: July 3, 2025	**Course/topic:** Capstone Project – Phase 4: Analyze
Prompt:	What were your initial findings when you started analyzing the cleaned data? How did the scope of the inventory crisis become apparent?
Journal Entry:	Today marked the moment when the abstract became devastatingly concrete. After importing my beautifully cleaned datasets from Phase 3, I began calculating basic inventory metrics, expecting to find some areas for improvement. Instead, I discovered a full-scale inventory management crisis This was my first real "holy shit" moment in data analysis. When I created the coverage categories (Stockout, Low Stock, Optimal, High Stock, Overstock) and saw the distribution, I realized this wasn't just an academic exercise anymore. Real business implications stared back at me from the screen. Initially struggled with the coverage calculation due to division by zero issues when products had zero sales. I had to implement proper case_when () logic to handle edge cases without breaking the analysis. When I generated the first visualization showing the risk distribution, the bars were almost equal across all categories except optimal. It looked like a completely random distribution
Other thoughts or questions:	How do companies operate with this level of inventory dysfunction? What's the human cost of these stockouts in terms of lost sales and frustrated customers?

Date: July 4, 2025	**Course/topic:** Capstone Project – Phase 4: Analyse
Prompt:	What were your initial findings when you started analyzing the cleaned data? How did the scope of the inventory crisis become apparent?
Journal Entry:	If Day 1 was about discovering the problem, Day 2 was about understanding what it costs the business. When I calculated the financial impact and saw $28.8M tied up in excess inventory, I sat back in my chair and just stared at the screen. The fact that stockout categories show $0M in my financial view highlighted the hidden cost - lost sales that don't show up in inventory valuation but represent real revenue impact.
Other thoughts or questions:	At what point does working capital efficiency become more important than service levels? How do you communicate financial impact to different stakeholder levels (CFO vs. warehouse manager)? |

Date: July 5, 2025	**Course/topic:** Capstone Project – Phase 4: Analyse
Prompt:	How did you develop the turnover analysis and ABC classification? What insights emerged about the root causes of the inventory crisis?
Journal Entry:	Today was about moving from "what's wrong" to "why it's wrong" and "how to fix it." The turnover analysis revealed the underlying dysfunction, while ABC classification provided the strategic framework for solutions. I also ran an ABC Analysis at both InventoryId and Brand levels. This revealed that roughly 15–20% of SKUs account for ~80% of sales value. That insight will help prioritize attention on the A-class items for stockout prevention and purchasing strategy. Additionally, I mapped stock variation between January and December to identify cities with net stock accumulation vs consumption. Visualizing this geographically highlighted potential overstock issues in specific locations, suggesting opportunities for balancing inventory between stores.
Other thoughts or questions:	How do you balance the theoretical elegance of ABC classification with the practical realities of vendor minimum orders and storage constraints? What's the right balance between data-driven decisions and human judgment in inventory management?

Date: July 12, 2025	**Course/topic:** Capstone Project -- Phase 5: Share
Prompt:	How did you approach designing dashboards that would communicate complex inventory insights to different stakeholder levels?
Journal Entry:	Moving from R analysis to Tableau visualization felt like learning a completely different language. My initial attempts were terrible - I was trying to cram every insight into single dashboards that looked like Excel spreadsheets. The breakthrough came when I realized I needed to design for the audience journey: executives need high-level KPIs and trends, operations managers need actionable risk alerts, and procurement teams need detailed vendor performance metrics. Dashboard 1 became the executive summary with the big financial numbers. Dashboard 2 focused on operational risks that need immediate action. Dashboard 3 quantified the savings opportunities. Dashboard 4 provided strategic vendor insights.
Other thoughts or questions:	How do you know when a dashboard is useful vs just pretty? Should I have tested these with real business users before finalizing?

Date: July 15, 2025	**Course/topic:** Capstone Project -- Phase 5: Share
Prompt:	What challenges did you encounter in Tableau and how did your visualization approach evolve?
Journal Entry:	Tableau's learning curve was steeper than expected. Simple things like calculated fields that were easy in R became complex in Tableau syntax. I spent hours figuring out why my ABC classification wasn't sorting correctly (turns out I needed to convert the field to a dimension and manually sort the values). The color-coding strategy evolved through trial and error. Initially, I used random colours that didn't convey meaning. Eventually settled on red for risks/problems, green for opportunities/good performance, and purple for strategic/high-level metrics. Interactive filters were game changers. Being able to filter by classification, city, or vendor across all dashboards made the analysis much more powerful.
Other thoughts or questions:	Why didn't I start with pen and paper sketches before jumping into Tableau? Would that have saved time on iterations? How do professional dashboard designers approach the design process?

Date: August 2, 2025	**Course/topic:** Capstone Project -- Phase 6: Act
Prompt:	How did you develop the implementation roadmap and what was your approach to prioritizing actions?
Journal Entry:	Phase 6 felt different from the previous phases - instead of analysing what IS, I was designing what SHOULD BE. The implementation roadmap had to balance urgency (preventing lost sales) with strategic value (long-term capability building). The immediate actions (0-30 days) focus on revenue protection - addressing those 27,997 Class A SKUs at stockout risk. The financial impact here is clear and measurable. Medium-term initiatives (3-12 months) focus on process standardization and technology deployment. Long-term strategy (12-24 months) emphasizes predictive capabilities and organizational excellence. The ROI calculation (1,390% in Year 1) gives me confidence that this isn't just academic theory - these recommendations could drive real business value.
Other thoughts or questions:	How do you communicate urgency without creating panic? What's the right balance between quick wins and sustainable change? How would I handle resistance to data-driven decisions in a traditional organization?

Date: August 10, 2025	**Course/topic:** Capstone Project -- Project Reflection
Prompt:	Looking back on the entire 5-month journey, what are your key takeaways about data analytics as a discipline and about your own capabilities?
Journal Entry:	This project taught me that data analytics isn't just about technical skills - it's about solving real business problems. The most sophisticated analysis means nothing if you can't communicate it effectively or translate it into actionable recommendations. My biggest personal growth was learning to restart when something isn't working. The Phase 3 architectural failure was devastating at the time, but it taught me that persistence means knowing when to change direction, not just pushing harder in the wrong direction. I also learned that storytelling with data is a skill unto itself. Moving from R analysis to Tableau dashboards to written recommendations required different types of thinking and communication.
Other thoughts or questions:	Am I ready for a professional data analyst role? What would I do differently in my next project? How do I continue building this foundation?