Data Wrangling Project

This project presents a foundational data wrangling pipeline designed to **transform raw, disparate operational data into a unified, clean, and analyzable dataset**. The central thesis is that by systematically collecting, organizing, and cleaning data from four separate sources—invoices, vials, dispense logs, and claims—it's possible to conduct a preliminary but critical analysis that immediately highlights key areas of financial and operational concern for an ophthalmology practice.

Execution

1. **Data Collection and Integration:** The project begins by loading four distinct CSV files. A crucial first step involves standardizing the 'Purchase Price' column by removing currency symbols and converting the data type to a float. These separate tables are then systematically merged using common keys (Invoice Number, Vial Number, Dispense ID) into a single, comprehensive DataFrame (full_df). This integration creates a holistic view, linking the entire lifecycle from purchase to reimbursement for each drug vial.
2. **Data Organization and Initial Analysis:** With the data unified, the project conducts an initial analysis to derive immediate business insights. This phase focuses on creating actionable reports by segmenting the data to identify anomalies:
    - **Unscanned Vials Report:** Identifies vials that were dispensed but not scanned, pointing directly to process gaps and potential revenue loss.
    - **Denied Claims Report:** Filters for all claims denied by insurance, highlighting another source of lost revenue.
    - **Profitability Analysis:** Calculates total revenue, cost, and profit, and identifies the top five most and least profitable vials, providing a quick overview of financial performance.
3. **Data Cleaning:** The final stage ensures the dataset's integrity. The code explicitly checks for and confirms the absence of duplicate rows. It then identifies columns with missing values—primarily in the dispense and claims sections, which is logical for vials not yet dispensed or claimed. These missing values are handled using a forward-fill (ffill) method, a straightforward approach to ensure data completeness for subsequent analysis.

Conclusion

This project successfully demonstrates a classic data wrangling workflow. It proves that by methodically combining and cleaning fragmented data, an organization can move from raw information to actionable intelligence. The resulting clean dataset and preliminary reports effectively expose operational inefficiencies and financial leakage, providing a solid foundation for more advanced analytics and data-driven decision-making.