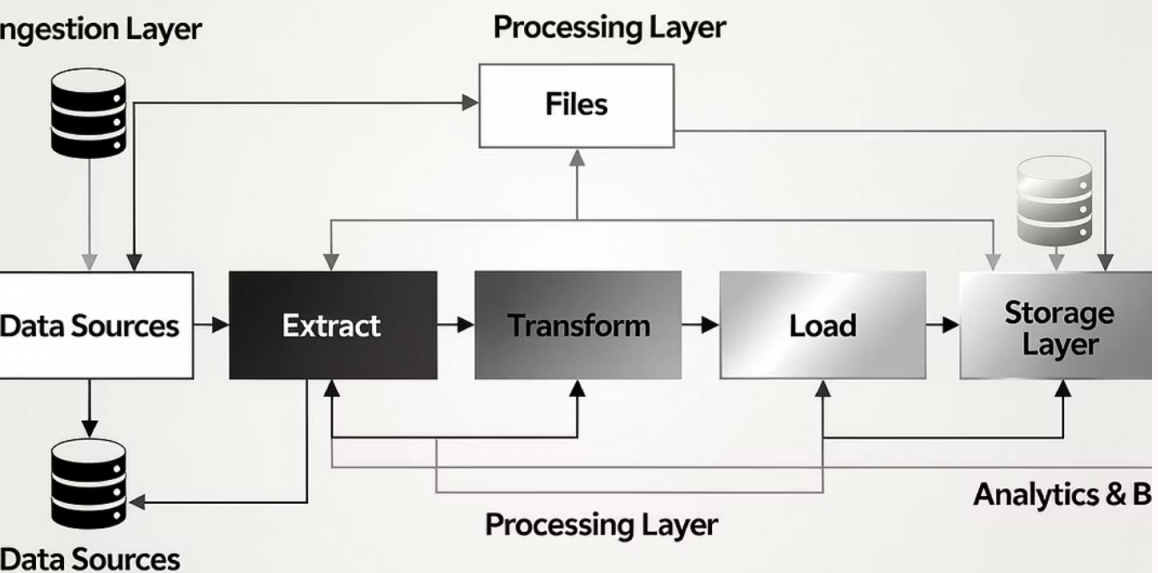# End-to-End Supply Chain Data Architecture and Analytical Framework

A comprehensive technical deep-dive into enterprise-grade ETL pipeline development and advanced SQL analytics for supply chain optimization

Executive Overview: Two-Phase Architecture

Phase I: Data Engineering

Python-based ETL pipeline transforming raw legacy CSV data into a normalized PostgreSQL database. Implements rigorous data quality protocols, feature engineering, and privacy-compliant data masking to establish a single source of truth for analytical workflows.

Phase II: Strategic Analytics

Comprehensive SQL analytical suite delivering multi-dimensional insights across logistics efficiency, financial performance, customer segmentation, and temporal trends. Advanced queries uncover actionable intelligence for data-driven decision making.

Phase I: Python ETL Pipeline Architecture

The foundational data engineering layer processes the DC Supply Chain dataset through a sophisticated transformation workflow. Built using Pandas and SQLAlchemy, this pipeline ensures data integrity, consistency, and analytical readiness.

01

Data Ingestion

Raw CSV ingestion using Pandas with latin-1 encoding to resolve character conflicts. Initial schema inspection validates structural integrity and identifies null distributions.

02

Standardization

Nomenclature normalization to snake_case prevents SQL syntax errors. Semantic renaming clarifies ambiguous fields. High-cardinality and PII fields removed for optimization and compliance.

03

Quality Validation

Datetime parsing enables temporal arithmetic. Logic checks verify shipping record consistency. Feature engineering creates delay_variance metric for efficiency quantification.

04

Database Migration

SQLAlchemy connector establishes PostgreSQL interface. Replace strategy ensures analytical database reflects current pipeline state. Ready for downstream SQL analysis.

Data Input   Cleaning   Transformation   Output
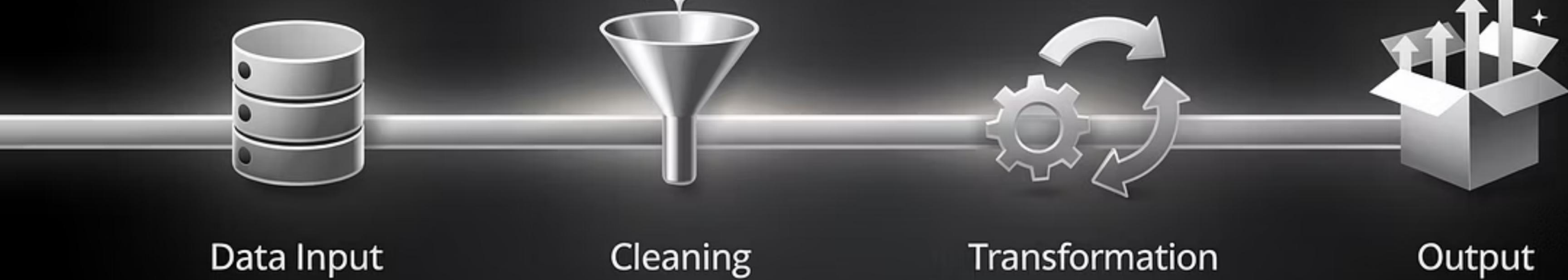
Data Cleaning and Standardization Protocol

1	2	3
Column Nomenclature All headers converted to snake_case (lowercase with underscores) to prevent SQL syntax errors from spaces or mixed-case identifiers.	**Semantic Mapping** - days_for_shipping_(real) → actual_shipping_days - days_for_shipment_(scheduled) → scheduled_shipping_days - shipping_date_(dateorders) → shipping_date	**Privacy Optimization** Removed product_description, customer_email, customer_password, and customer_zipcode fields to optimize query performance and ensure GDPR compliance.

Data Integrity and Derived Metrics

Temporal Feature Engineering

Date fields (order_date, shipping_date) parsed to datetime objects enabling sophisticated temporal arithmetic and time-series analysis capabilities.

Logic Validation Framework

Implemented data quality checks comparing calculated shipping duration (shipping_date - order_date) against recorded actual_shipping_days to identify and flag inconsistencies in source data.

Performance Metric Generation

Created delay_variance derived metric by calculating the difference between actual_shipping_days and scheduled_shipping_days. This quantifies logistics efficiency and enables bottleneck identification across geographic and operational dimensions.

Phase II: Strategic Data Analysis Framework

The analytical suite leverages advanced SQL techniques to extract actionable insights across four critical business dimensions. Each query is designed to answer specific strategic questions and inform operational decisions.

Logistics & Operations

Geographic bottleneck identification and shipment mode reliability analysis

Financial Health

Market profitability ranking and department revenue performance assessment

Customer Intelligence

Segment value analysis, CLV calculation, and fraud pattern detection

Temporal Dynamics

Time-series trend analysis with MoM/YoY growth and seasonality patterns

# Critical Finding: Geographic Shipping Bottlenecks

Analysis of delay variance across global distribution nodes reveals significant unpredictability in specific locations. Thiais, France emerges as the most problematic logistics hub with extreme shipping variability.
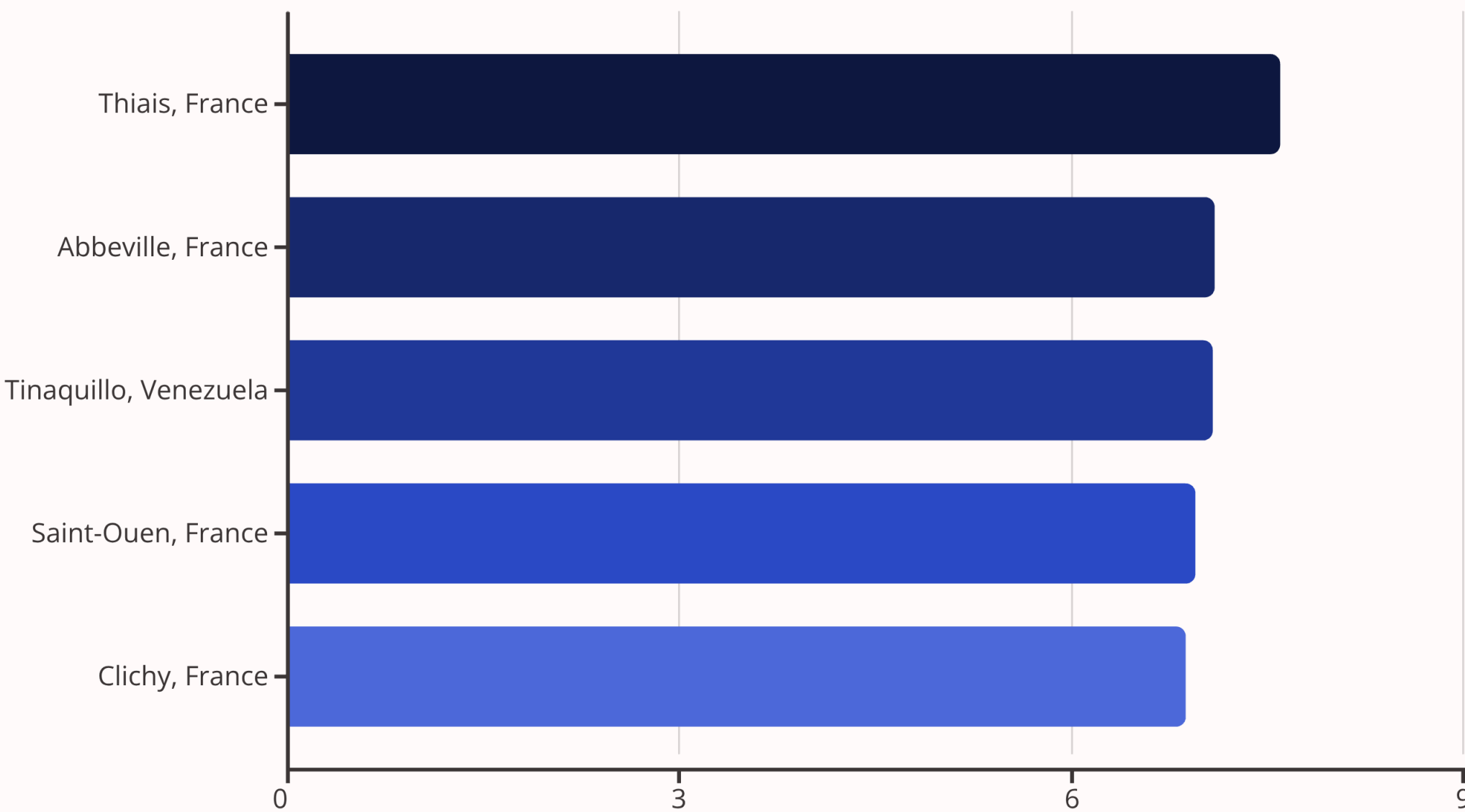
France dominates the top 5 most unpredictable shipping locations, suggesting systemic logistics infrastructure challenges requiring immediate operational review and potential carrier relationship reassessment.
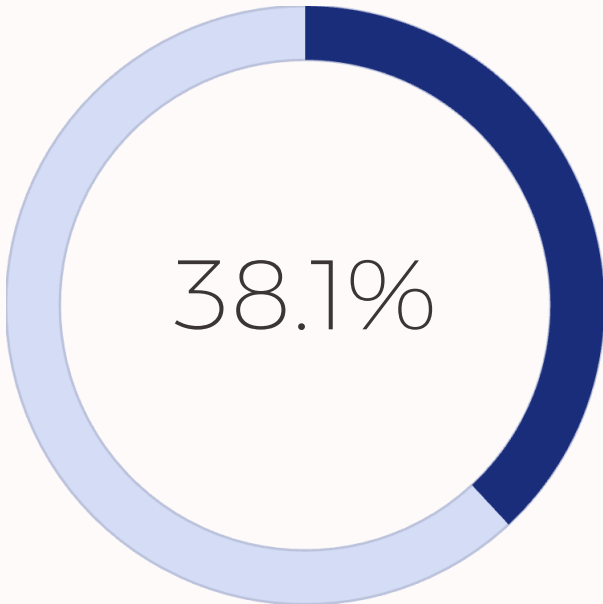
Shipment Mode Reliability Analysis

Comparative analysis of late delivery rates versus on-time performance reveals critical reliability gaps across shipment modes. First Class shipping presents the most significant operational risk.
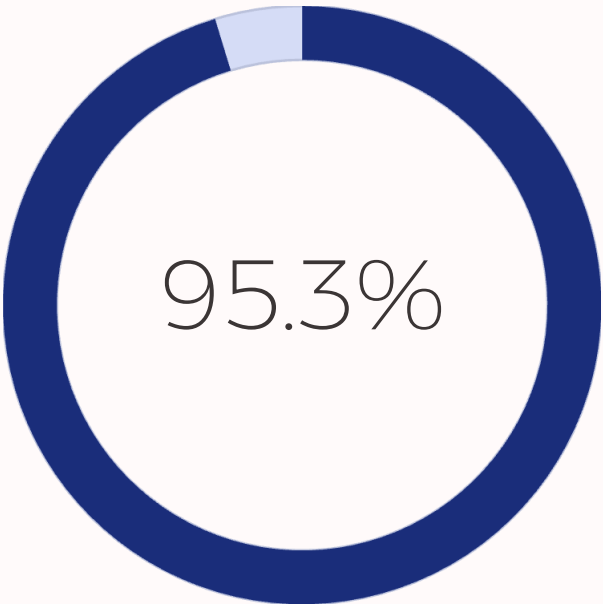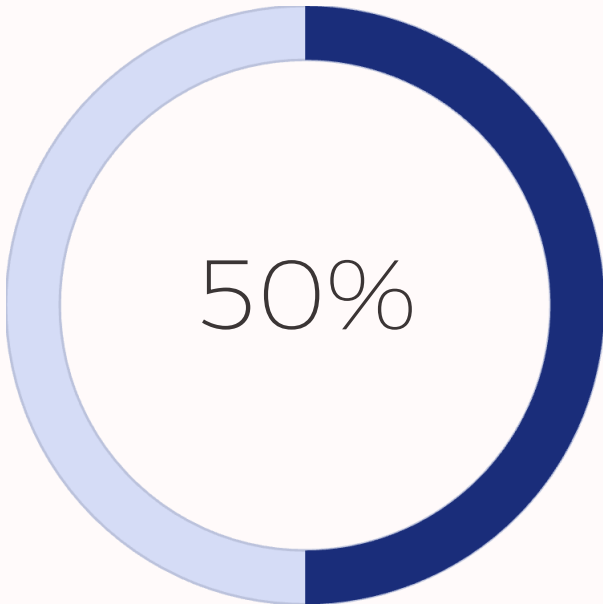

95.3%	38.1%	50%
First Class Late Rate	**Standard Class Late Rate**	**Same Day On-Time**
Critically unreliable with near-total late delivery performance	Moderate reliability but still misses deadlines in over one-third of shipments	Most reliable option, though only achieving baseline 50% punctuality

These findings indicate urgent need for carrier performance reviews and potential contract renegotiations, particularly for First Class service level agreements.

Market Profitability and Revenue Performance

Regional Profit Margins

Geographic market analysis reveals USCA (US & Canada) leading profitability at 11.14%, with Africa (10.99%) and Latin America (10.93%) showing competitive margins. All major markets maintain healthy double-digit profitability.

11.14%

USCA Market

10.99%

Africa Market

10.93%

LatAm Market

Department Revenue Leaders

Fan Shop dominates revenue generation at $17.1M, significantly outperforming other categories. Apparel follows at $7.9M, while Book Shop and Pet Shop underperform dramatically.

- **Fan Shop:** $17.1M (Top performer)
- **Apparel:** $7.9M (Strong second)
- **Outdoor:** $5.2M
- **Book Shop:** $12.6K (Underperformer)
- **Pet Shop:** $41.5K (Critical review needed)

Loss Leader Identification: The Margin Challenge

High-volume categories with disproportionately low profit margins represent strategic risks. The Strength Training category demonstrates this phenomenon most severely, generating substantial sales but minimal profit contribution.

Strength Training

Margin: 0.60% - Major loss leader with critically thin profitability despite high sales volume. Immediate pricing strategy review recommended.

As Seen on TV

Margin: 3.47% - Significantly underperforming relative to operational costs. Consider category rationalization or supplier renegotiation.

Strategic Implications

These loss leaders require urgent attention: either optimize pricing/costs or evaluate strategic value of maintaining these product lines for customer acquisition purposes.

Customer Segmentation and Lifetime Value

Segment Revenue Contribution

The Consumer segment drives the majority of revenue at $19.1M, representing the largest addressable market. Interestingly, Average Order Value (AOV) remains remarkably consistent across all segments at approximately $204, suggesting price sensitivity is uniform regardless of customer type.

$19.1M
Consumer Revenue

$204
Consistent AOV

Across all segments

Top 1% Customer Value

207 customers comprise the elite tier with spending exceeding $6,471. The highest-value customer (ID 791) has contributed $10,524 lifetime value, representing critical retention targets for personalized engagement programs.

Fraud Risk Geography: High-Alert Locations

Geographic analysis of suspected fraud patterns (minimum 50 orders threshold) identifies specific cities with disproportionately high fraud rates, requiring enhanced verification protocols and payment security measures.
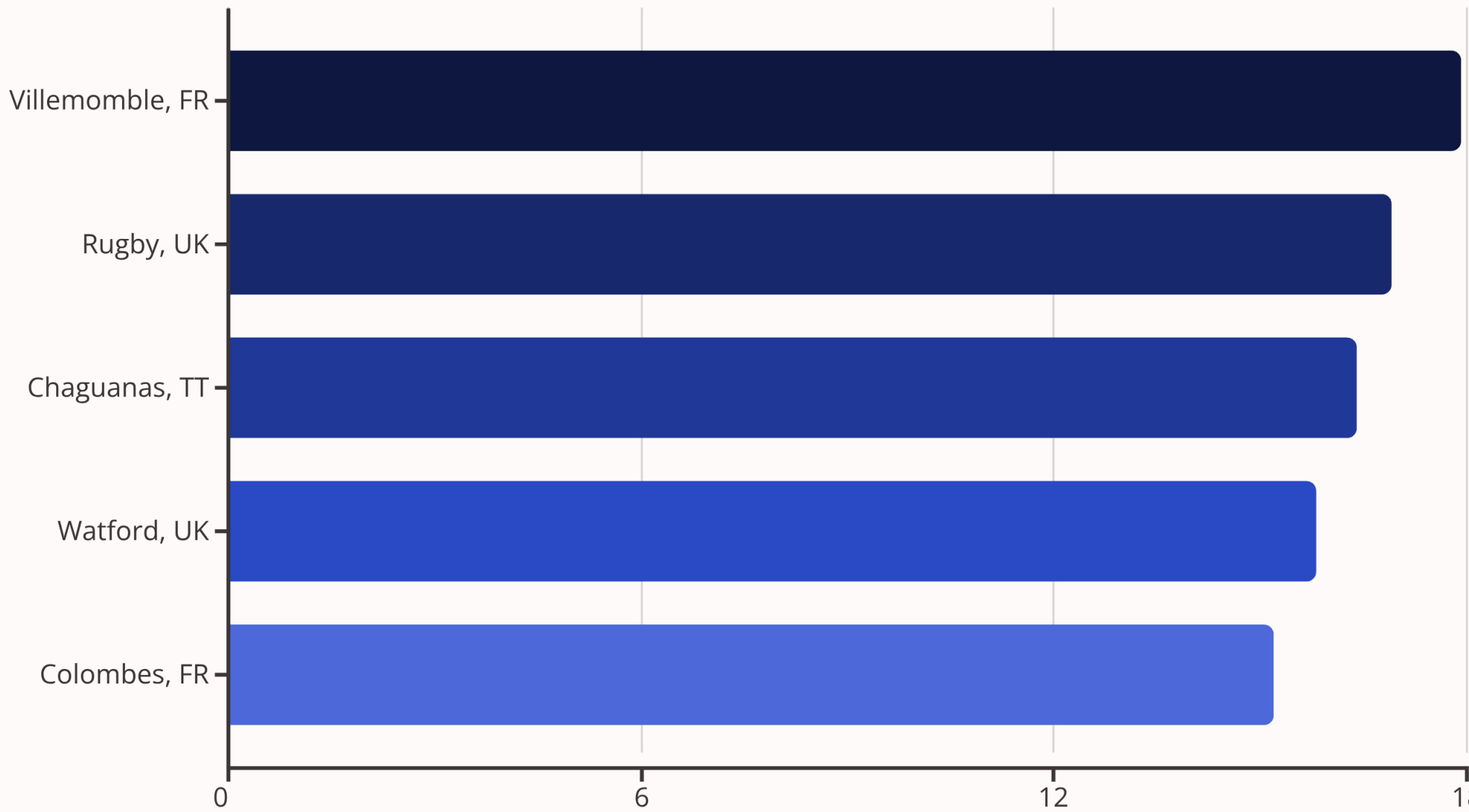
France and UK locations dominate fraud risk zones. Implement enhanced identity verification, address validation, and payment authentication for orders originating from these high-risk geographies.

Critical Alert: Revenue Decline in Late 2017

Time-series analysis reveals a severe downturn in business performance during Q4 2017 and continuing into early 2018. This represents a critical inflection point requiring immediate investigation and remediation.

November 2017

-41.6% MoM Sales Growth

Precipitous month-over-month sales collapse indicating systemic operational or market disruption.

January 2018

-34.2% MoM | -70.5% YoY Profit

Continued deterioration with devastating year-over-year profit decline suggesting structural business challenges.

Root cause analysis required to identify whether decline stems from operational failures, market conditions, competitive pressure, or internal strategic missteps. Recovery plan development is urgent priority.

Seasonality Patterns: Pre-Holiday Surge

October Peak Performance

Monthly sales analysis identifies October as the strongest revenue period with an average order value of $244.79, representing a significant 22% premium over the annual baseline of approximately $200.

$244.79
October AOV

Peak month performance

+22%
Above Baseline

Premium vs. annual average

Strategic Implications

This pre-holiday surge pattern suggests strong opportunity for targeted October marketing campaigns, inventory optimization, and promotional strategies to maximize this naturally high-performing period. Capacity planning should account for this predictable seasonal lift.

Key Takeaways and Strategic Recommendations

The implemented data architecture demonstrates professional-grade workflow capabilities, successfully transforming raw legacy data into actionable intelligence across multiple business dimensions.

Pipeline Excellence

Python ETL workflow ensures clean, consistent, privacy-compliant data entry with sophisticated feature engineering creating analytical-ready datasets for downstream SQL analysis.

Critical Bottlenecks Identified

France logistics infrastructure presents severe operational risks with extreme delay variance. First Class shipping reliability at 95.3% late rate requires immediate carrier intervention and SLA renegotiation.

Financial Risk Areas

Strength Training category operates as loss leader with 0.60% margin. Book Shop and Pet Shop departments critically underperform. Strategic review needed for category rationalization or optimization.

Revenue Crisis Alert

Q4 2017 downturn with -70.5% YoY profit decline in January 2018 represents existential threat requiring urgent root cause analysis and comprehensive recovery strategy development.

Report Prepared By

Anshu Raj

ar.anshu2611@gmail.com