Monthly Revenue as a Stochastic System

Modelling,Simulation, and Uncertainty Quantification

Dataset: Monthly panel aggregated from Online Retail II transactions (2010–2011; ∼24 monthly observations)

Observed variables: Revenue R_t, Orders O_t, AOV A_t with $R_t = O_t \cdot A_t$

Purpose

Recast a real-world revenue process into an explicit, computable stochastic model and build a reproducible simulation pipeline: (i) construct a single source of truth via SQL ETL, (ii) formalize revenue as $R_t = O_t \cdot A_t$, (iii) quantify predictive and distributional uncertainty through Monte Carlo UQ, and (iv) stress-test structural assumptions by varying dependence via a copula-based counterfactual experiment.

Methodology

- **Data engineering:** SQL ETL + validation to obtain a clean monthly panel (single source of truth).
- **System characterization:** trend/seasonality diagnostics; driver decomposition into volume (O_t) and ticket-size (A_t).
- **Predictive modelling (time-consistent):** linear baseline as interpretable structural model; Random Forest as a residual corrector (holdout, no shuffle).
- **Uncertainty quantification:** distribution fitting for (O_t, A_t) and Monte Carlo propagation to obtain a revenue distribution and tail metrics (VaR/CVaR).
- **Counterfactual simulation:** keep marginals fixed and vary only dependence via a Gaussian copula (ρ sweep).
- **Robustness:** replicated simulations across seeds with 95% confidence intervals for tail-risk metrics.

Author: Zhong Yuyang
Email: zhongyuyangm4@gmail.com
Background: BSc Applied Physics, Dalian Maritime University
Code: https://github.com/temper-z-debug/monthly_revenue_analytics
Date: 2026/01/18

Key Findings

1. **Stable seasonal signature enables structured modelling (Fig. 01–02)**
   Smoothed revenue and month-aligned YoY curves show repeatable late-Q4 peaks, consistent with a systematic seasonal component.
2. **Parsimonious decomposition: $R_t = O_t \cdot A_t$ with volume-dominant variability (Fig. 03, 04, and 06)**
   Revenue co-moves with order volume while AOV is comparatively stable, supporting a volume-led stochastic driver at monthly granularity.
3. **Seasonal regime concentrates anomalies and tail exposure (Fig. 05 & 07)**
   Growth dynamics and outliers cluster around Q4, suggesting regime-dependent variance amplification.
4. **Time-consistent holdout reveals nonlinear residual structure (Fig. 08)**
   A linear baseline captures the dominant mapping; Random Forest reduces systematic residual patterns as a residual corrector.
5. **Monte Carlo enables distributional reasoning and tail quantification (Fig. 09)**
   Propagating fitted marginals through $R_t = O_t A_t$ yields an empirical revenue distribution with VaR/CVaR summaries.
6. **Dependence is a first-order tail-risk amplifier; robustness via replication (Fig. 10–11)**
   With fixed marginals, increasing copula dependence (ρ) worsens the downside tail; replicated seeds preserve VaR/CVaR trends with 95% CIs.

Next Steps

1. **Seasonal dynamics:** Explicit seasonal dynamics: Introduce STL decomposition with SARIMA or state-space models for interpretable seasonal components.
2. **Tail dependence:** t-copula / vine copulas to model asymmetric and tail dependence.
3. **Parameter uncertainty:** bootstrap/Bayesian propagation for credible intervals of VaR/CVaR.
4. **Scaling:** variance reduction + profiling-driven optimization for higher-frequency panels.

Reproducibility: All figures (Fig. 01–Fig. 11) are generated from the script pipeline; a one-click runner executes the workflow in a fixed order and writes artifacts to disk.

Computational Pipeline and Reproducible Experiment Design

Design principle

Each stage consumes the same validated monthly panel, writes explicit artifacts, and can be re-run end-to-end with controlled randomness (fixed seeds).

Reproducibility guarantees

- **Single source of truth:** all analyses use the same monthly panel.
- **Time-consistent evaluation:** chronological split; no shuffle.
- **Controlled randomness:** fixed seeds for modelling and simulation.
- **Named outputs:** figures, metrics, and parameters written to disk.
- **Full rerunability:** a one-click script regenerates all artifacts.

Workflow (compact ASCII)

```
Raw Data -> SQL ETL -> monthly_panel.csv
          |-> Diagnostics & Decomposition     (Fig01-Fig07)
          |-> LR + RF Residual Modelling      (Fig08)
          |-> Monte Carlo UQ                  (Fig09)
          |-> Copula Dependence Sweep         (Fig10-Fig11)
```

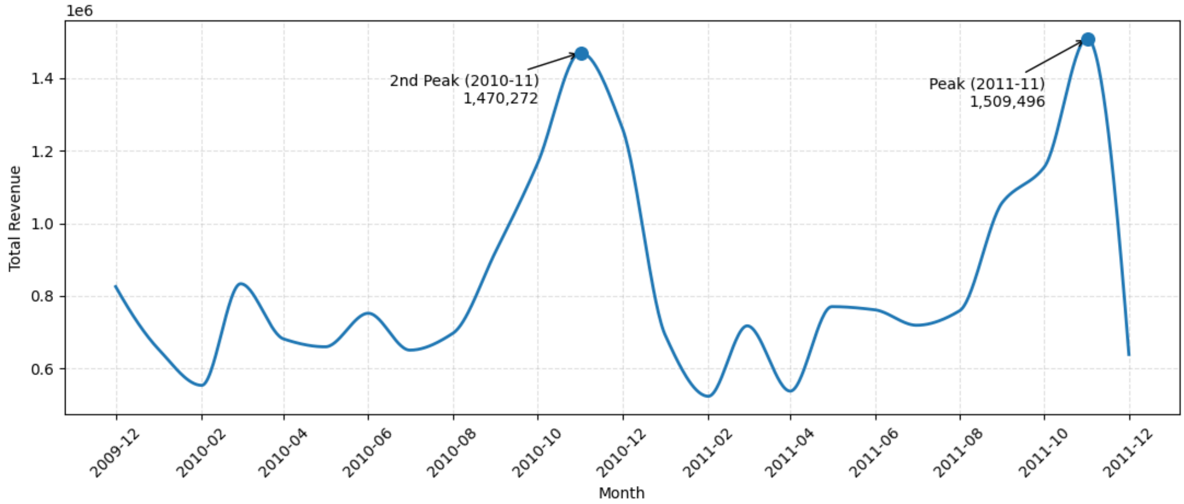
Short pseudo-code

```
panel = sql_etl_and_validate(raw_tables)
make_fig01_to_fig07(panel)

Xtr, ytr, Xte, yte = time_holdout_split(panel)
lr = fit_linear_baseline(Xtr, ytr)
rf = fit_residual_corrector(Xtr, ytr, base=lr)
evaluate_and_plot(lr, rf, Xte, yte)            # Fig08

params = fit_marginals(panel)
mc_uq(params, N=50000)                         # Fig09
copula_sweep(params, RHO_GRID)                 # Fig10-Fig11
```

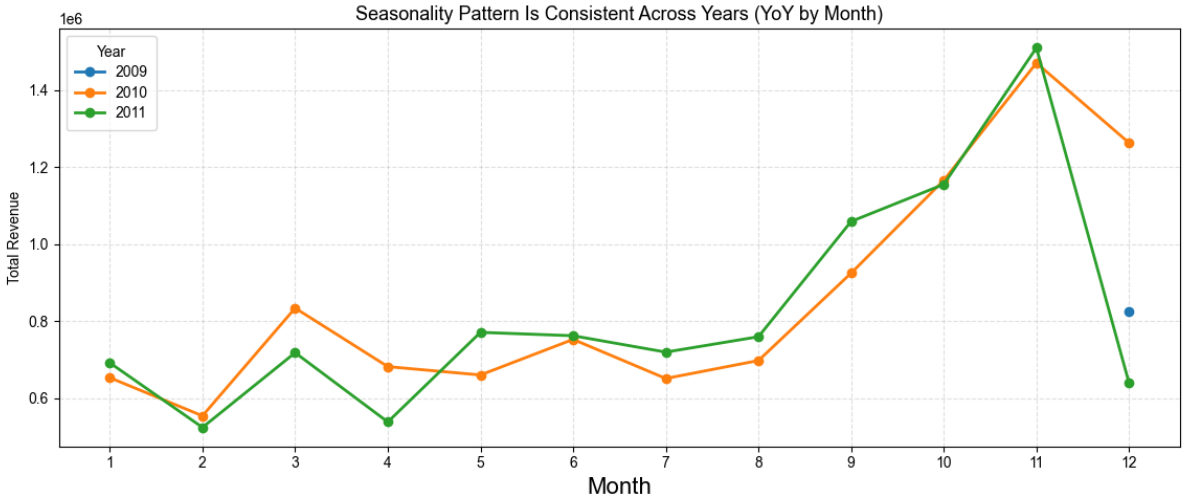
Fig. 01 — Revenue Exhibits Strong Seasonality with Concentrated Year-End Peaks



Note: Smoothed monthly revenue series with peak months highlighted to summarize macro seasonality.

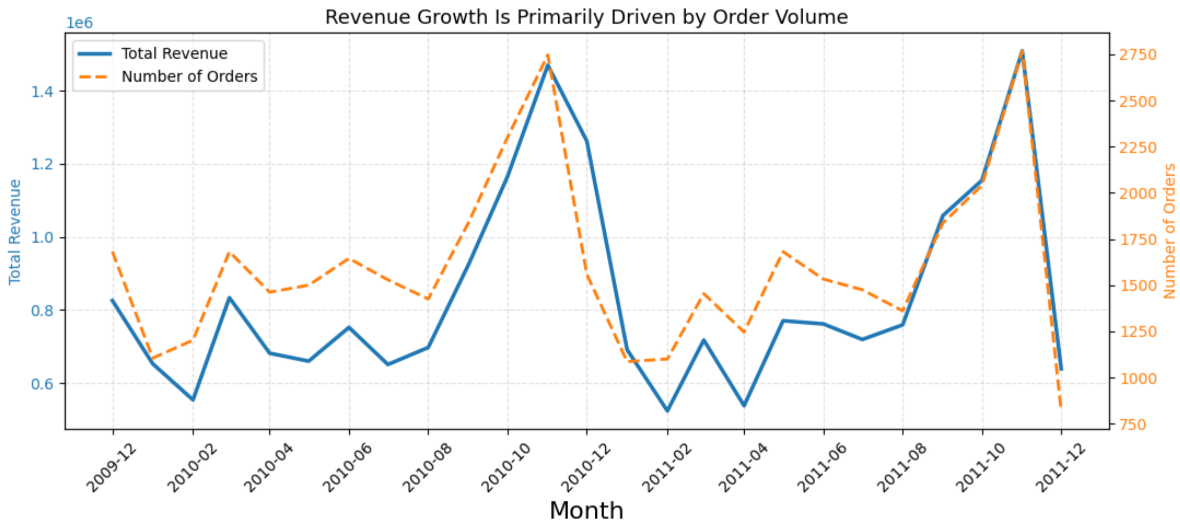
Takeaway: Peak revenue clusters in late Q4, suggesting a stable seasonal regime rather than steadily compounding growth.

Fig. 02 — Seasonality Pattern Is Consistent Across Years (Month-Aligned YoY)



Note: Month-aligned revenue comparison across years. If a partial-year is present, it should be flagged explicitly.

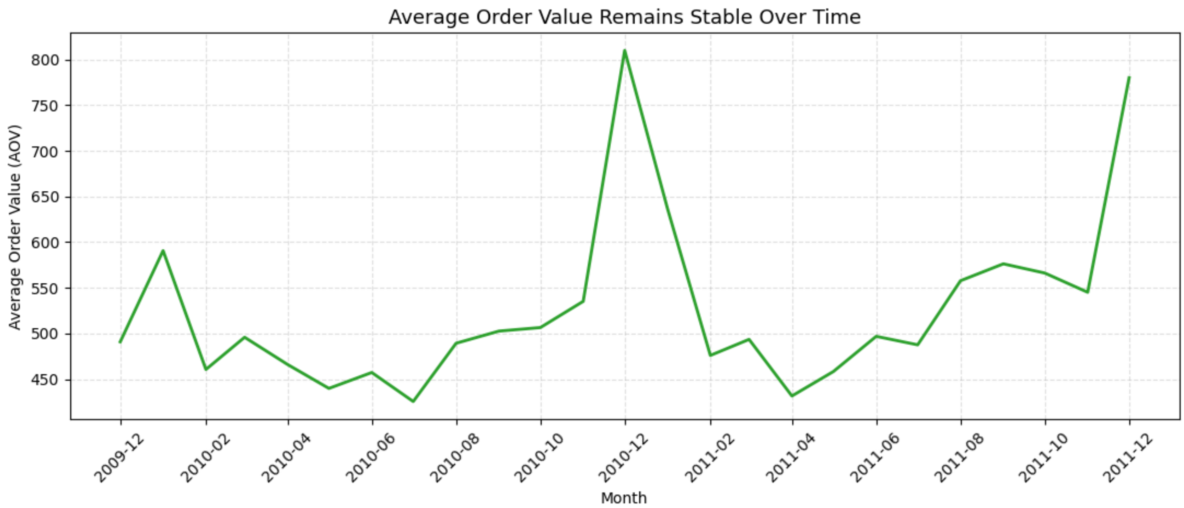
Fig. 03 — Revenue Closely Tracks Order Volume, Indicating Volume-Led Dynamics



Note: Dual-axis time series comparing revenue vs. orders on the same monthly timeline.

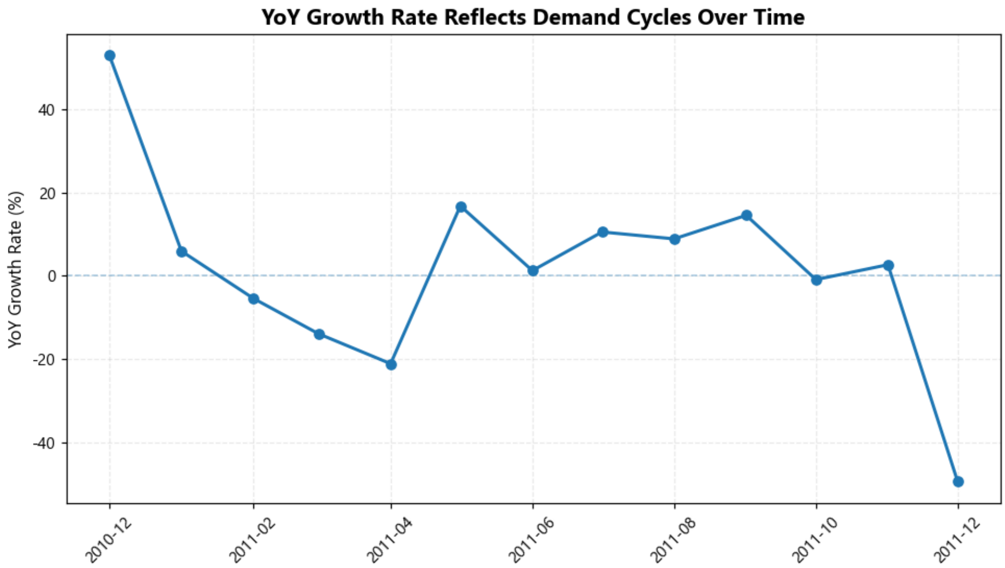
Takeaway: Revenue and orders co-move over time, implying demand volume is the primary driver of revenue fluctuations.

Fig. 04 — Average Order Value (AOV) Is Relatively Stable



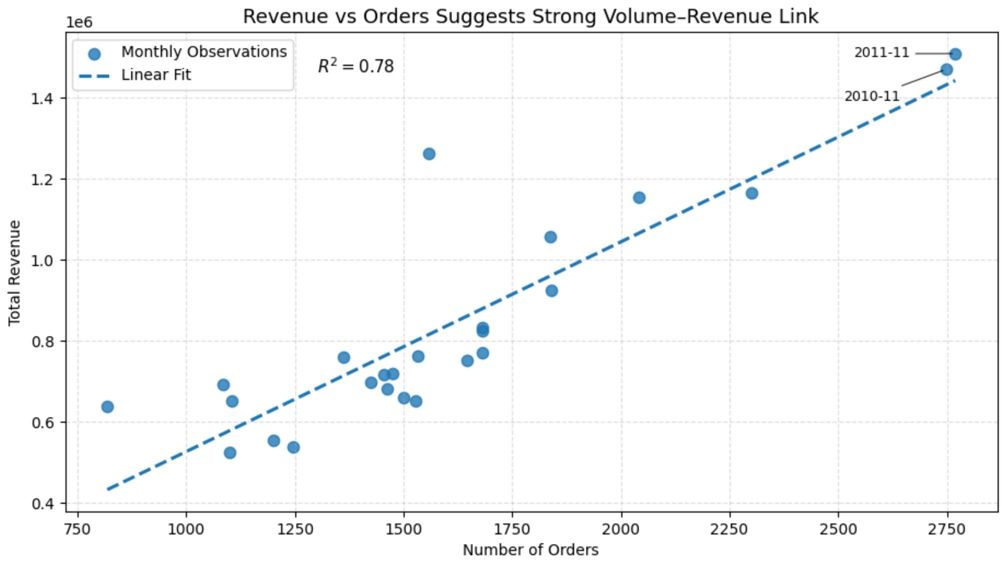
Note: AOV is computed as revenue per order to separate ticket-size effects from volume effects.

Fig. 05 — YoY Growth Highlights Acceleration and Cool-Down Phases



Note: YoY(t) = Revenue(t) / Revenue(t-12) - 1. Months without prior-year comparisons are excluded by design.

Fig. 06 — Orders Explain Most Revenue Variance (Linear Fit with R^2)

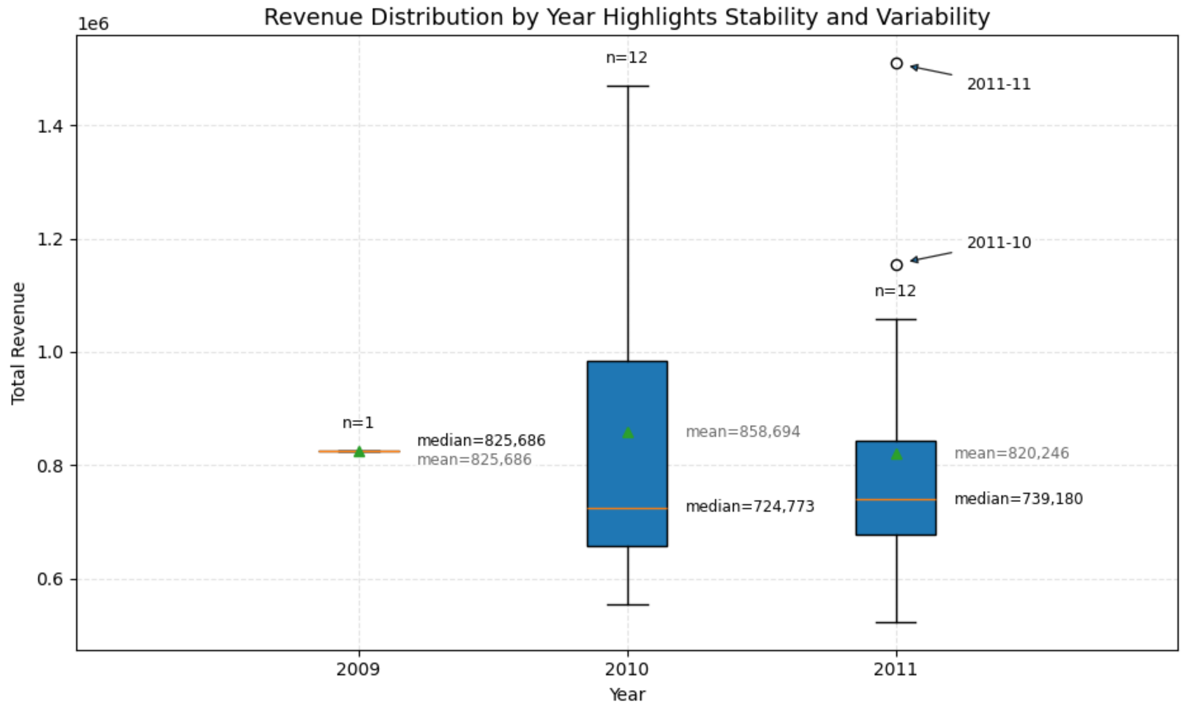


Takeaway: Revenue exhibits a strong linear relationship with order volume, supported by a high R^2, confirming volume-led growth dynamics.

Note: Scatter with linear fit to quantify the strength of the volume–revenue relationship.

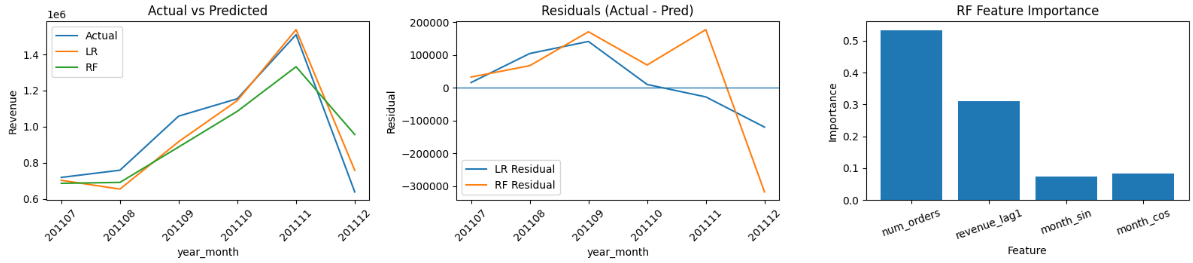
Takeaway: A linear model explains a substantial share of revenue variance, consistent with volume-led revenue dynamics.

Fig. 07 — Annual Revenue Distribution Shows Stability with Q4-Driven Outliers



Note: Yearly boxplots with sample size (n), mean/median labels, and annotated outlier months.

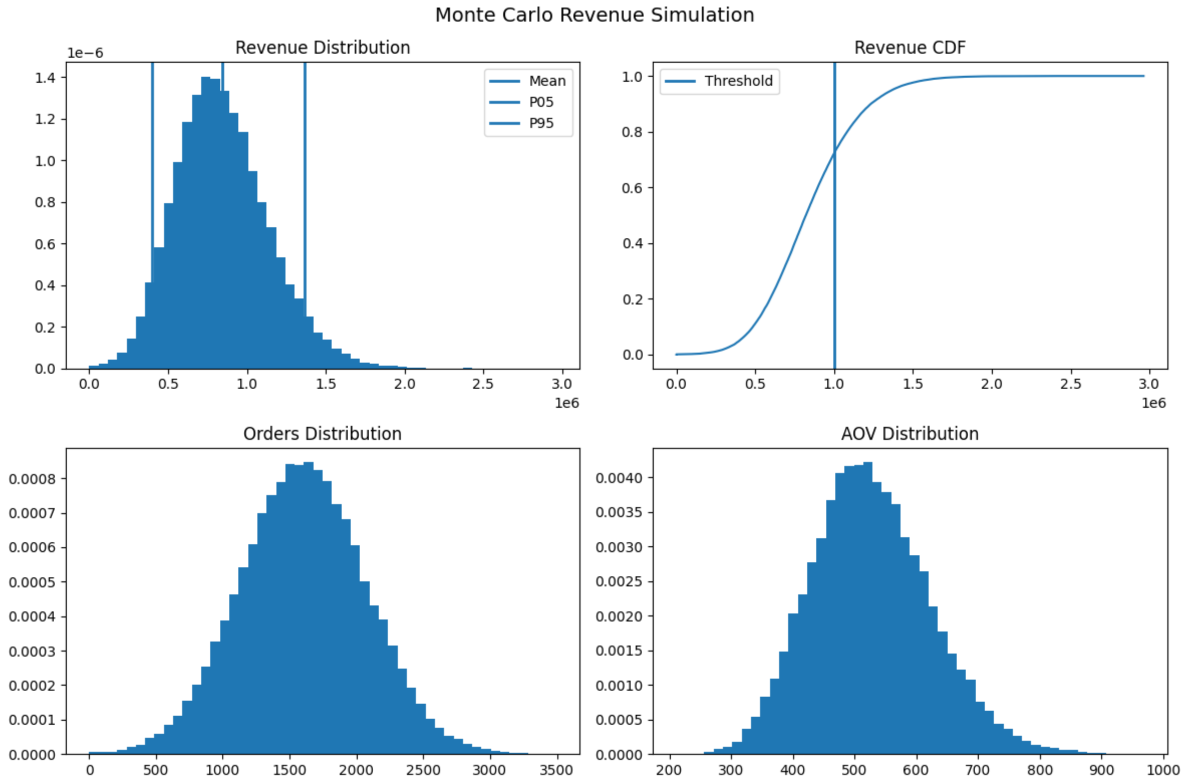
Fig. 08 — Revenue Prediction: Linear Regression vs. Random Forest (Holdout Test)



Note: Holdout evaluation using lagged revenue, order volume, and cyclical month features; panels show prediction fit, residual diagnostics, and feature importance.
Model interpretation: Linear regression provides a transparent baseline; Random Forest functions as a non-parametric residual corrector, approximating unmodelled curvature and interactions revealed by residual structure.

Takeaway: The baseline captures the dominant structure while the RF reduces systematic residual patterns, indicating learnable nonlinear corrections rather than a purely linear mapping.

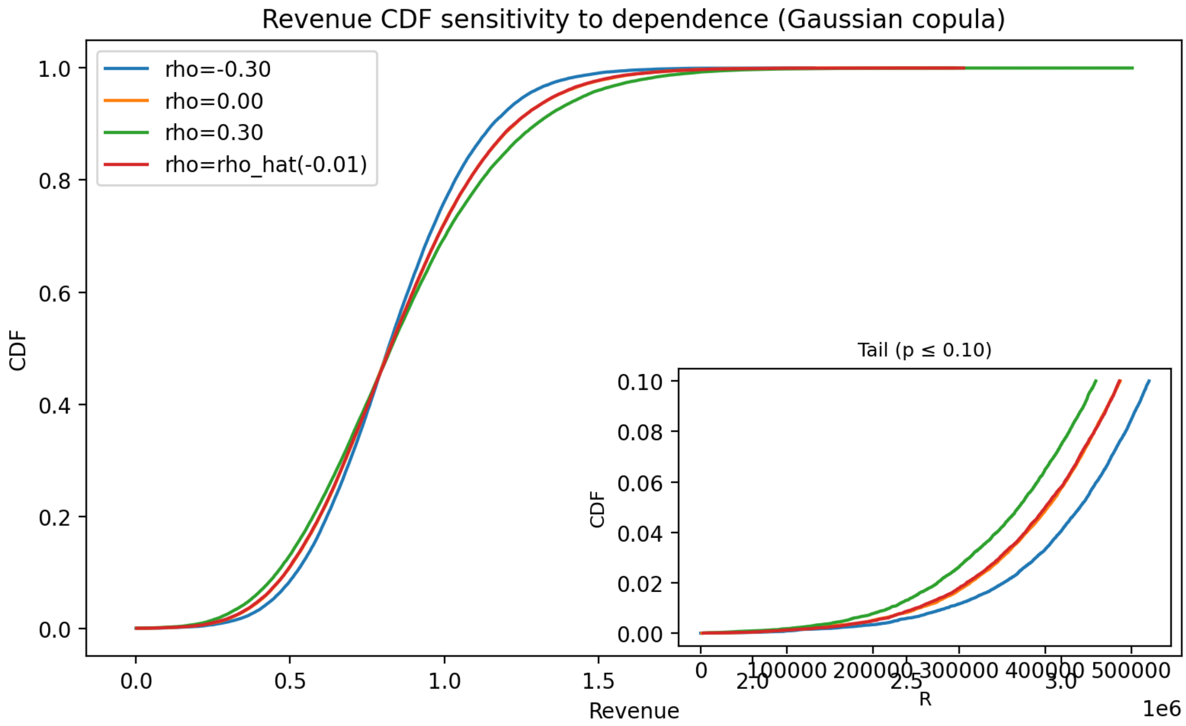
Fig. 09 — Monte Carlo Revenue Simulation (Uncertainty Propagation)



Note: Revenue simulated as Orders × AOV with Orders modeled as Normal (clipped at 0) and AOV modeled as Gamma; panels show distribution, CDF, and component distributions.
Sampling design: $N = 50{,}000$ draws provide a stable empirical approximation at moderate cost; forward simulation is appropriate because the product distribution is not convenient to derive analytically.

Takeaway: Monte Carlo propagates uncertainty through $R_t = O_t A_t$, enabling distributional reasoning and quantile-based tail risk.

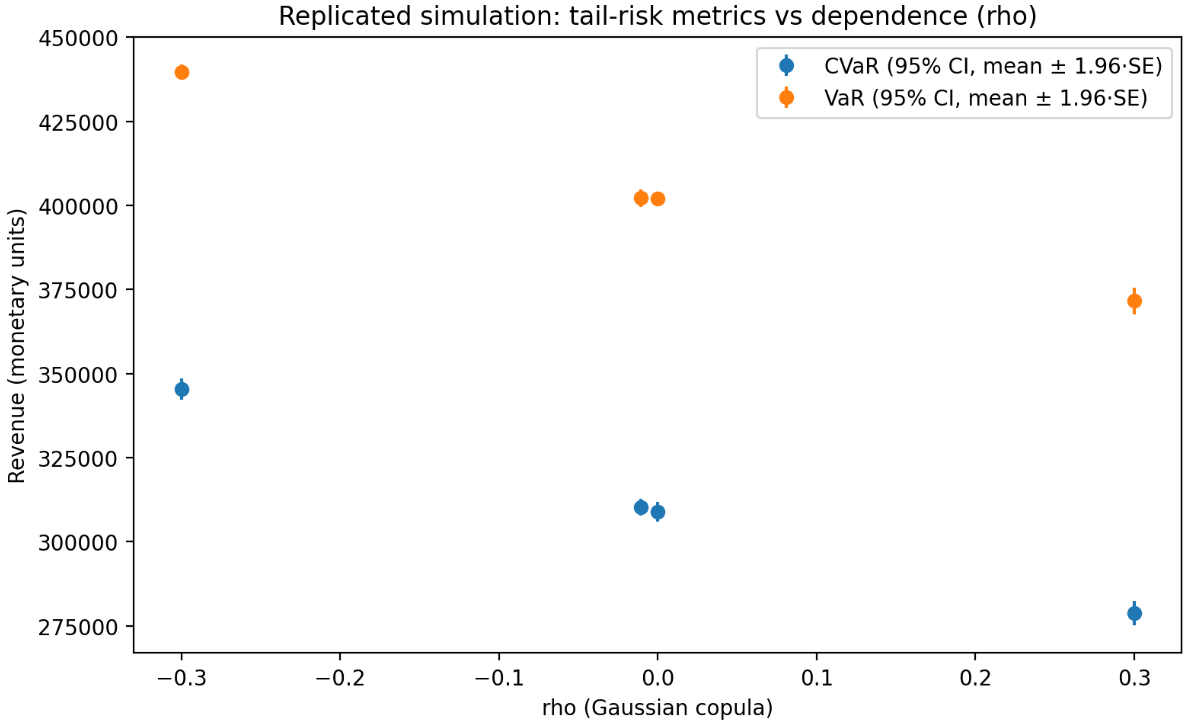
Fig. 10 — Copula Dependence Stress Test: Revenue CDF with Tail Inset (rho sweep)



Note: Controlled counterfactual: fitted marginals of Orders and AOV are held fixed while only dependence varies via a Gaussian copula parameterized by ρ. The CDF includes a left-tail inset ($p \leq 0.10$) to visualize downside shifts.
Design: $N = 50{,}000$ draws per setting.

Takeaway: Increasing ρ shifts the left tail downward, indicating heavier downside risk under stronger positive dependence even when marginals remain unchanged.

Fig. 11 — Robustness via Replicated Simulation: VaR/CVaR vs. rho with 95% Confidence Intervals



Note: Replicated simulation under the copula stress test. Points denote the mean across seeds ($n_{\mathrm{rep}} = 3$); error bars are 95% confidence intervals $mean \pm 1.96 \cdot \mathrm{SE}$.

Takeaway: $\mathrm{VaR}_{5\%}$ and $\mathrm{CVaR}_{5\%}$ decrease monotonically with ρ; replication preserves the trend, supporting robustness of the dependence-driven tail-risk effect.

System Formulation, Uncertainty Quantification and Assumptions

System model

We formalize monthly revenue as a multiplicative stochastic system:

$$R_t = O_t \cdot A_t,$$

where O_t is monthly order volume and A_t is average order value (AOV). The modelling goal is not only point prediction, but distributional reasoning and downside-tail quantification.

Marginal distributions

Based on empirical shape and support, we model

$$O_t \sim \mathcal{N}(\mu_O, \sigma_O^2)\ truncatedat0, \qquad A_t \sim \text{Gamma}(k, \theta).$$

Parameters are estimated via method-of-moments; μ_O, σ_O from sample mean/SD of orders, and

$$k = \left(\frac{E[A]}{\text{SD}(A)}\right)^2, \qquad \theta = \frac{\text{Var}(A)}{E[A]}.$$

Uncertainty propagation (Monte Carlo UQ)

We draw $\{(O_t^{(i)}, A_t^{(i)})\}_{i=1}^N$ from the fitted component model and propagate through

$$R_t^{(i)} = O_t^{(i)} A_t^{(i)}.$$

The resulting empirical distribution supports quantile-based risk metrics and tail inspection.

Risk metrics

For revenue distribution R, VaR_α is the α-quantile and $\text{CVaR}_\alpha = E[R \mid R \leq \text{VaR}_\alpha]$ (here $\alpha = 0.05$).

Counterfactual dependence experiment

To test a falsifiable structural assumption, we keep the fitted marginals of (O_t, A_t) fixed and vary *only* the dependence structure via a Gaussian copula parameterized by ρ (Fig. 10–11). This isolates the causal effect of dependence on downside tail risk under controlled conditions.

Limitations

Finite-sample and identifiability

The panel spans only $\sim$24 months, limiting the stability of distribution fitting and restricting the ability to validate multi-year seasonal persistence. Risk metrics should therefore be interpreted as *model-based* summaries under a short-sample regime.

Missing temporal dependence

The current UQ treats draws as i.i.d. and does not model autocorrelation or regime-switching dynamics (e.g., seasonal states). A state-space or ARIMA-type process for O_t would better represent temporal structure and enable multi-step scenario simulation.

Dependence class restriction

The stress test uses a Gaussian copula, capturing rank dependence but not tail dependence or asymmetric dependence. Alternative copulas (e.g., t-copula, vine) may further amplify or reshape downside tail behaviour.

Computational scope

Simulation budgets are sufficient for stable empirical tails at the current scale ($N = 50{,}000$), but scaling to higher-frequency data would require variance reduction and profiling-driven optimization to maintain accuracy under compute constraints.

Executive Summary

This project constructs an end-to-end modelling and simulation workflow for a real-world revenue process. Starting from invoice-level Online Retail II transactions, a transparent SQL ETL pipeline produces a validated monthly panel (single source of truth). The process is then formalized as a multiplicative stochastic system $R_t = O_t \cdot A_t$, separating demand volume and ticket-size mechanisms.

Beyond descriptive seasonality diagnostics and time-consistent predictive evaluation (linear baseline plus a non-parametric residual corrector), the core contribution is uncertainty quantification: fitted component marginals for (O_t, A_t) are propagated via Monte Carlo to obtain distributional revenue forecasts and explicit tail-risk summaries (VaR/CVaR). To interrogate a key modelling assumption, a controlled counterfactual experiment holds marginals fixed while varying dependence via a Gaussian copula (ρ sweep); replicated simulations provide 95% confidence intervals and support a falsifiable conclusion that dependence can materially amplify downside tail risk.

Overall, the workflow demonstrates the MSMS-aligned capability to translate an applied system into a computable model, design simulation experiments to test structural assumptions, and implement a reproducible computational pipeline for data-driven modelling and risk-aware reasoning.