From Spreadsheets to Strategy

A Predictive Retail Journey: AI / ML / ROI

A four-part challenge series demonstrating the evolution of data-driven decision-making in grocery retail. The project traces the journey from untangling raw data to modeling customer behavior and quantifying business impact.

The Data Maturity Roadmap: A Four-Stage Evolution

The project moves logically from foundational data structure to high-level business impact. Each challenge builds directly on the last, creating a reproducible and robust analytics capability.

Data Foundation • Data Integrity • Temporal Insight • Behavioral ROI

Stage 1: From Spaghetti to Schema

Building a Single Source of Truth from Excel Chaos

STRUCTURE → INTEGRITY → SIGNAL → STRATEGY

The Business Challenge

A growing grocery chain is operationally constrained by running on inconsistent, disconnected spreadsheets. Issues like duplicated headers, mixed data types, and naming drift block analytics.

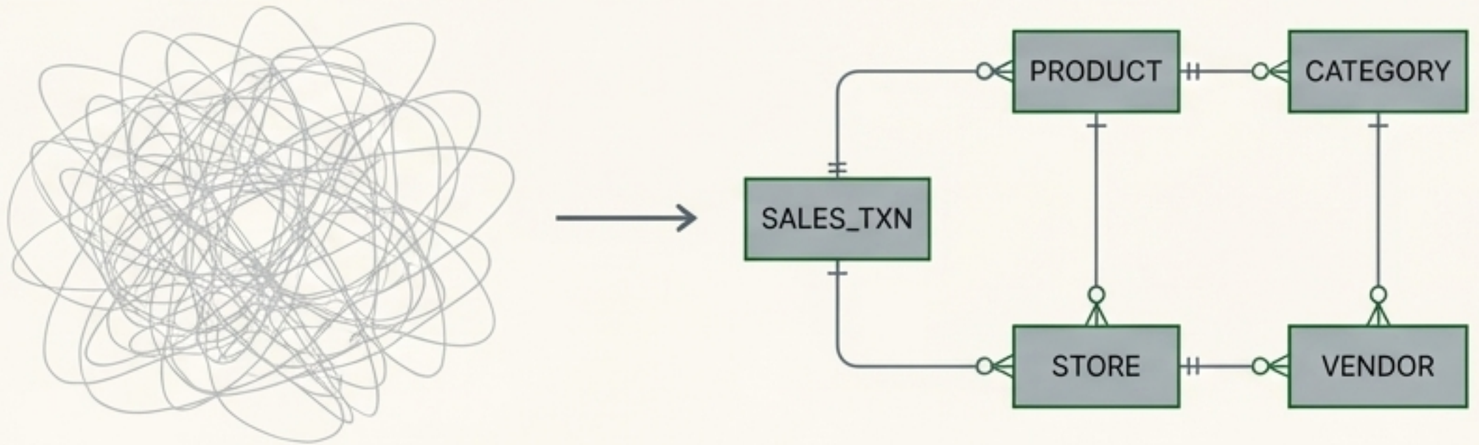
The Technical Action

Reverse-engineered a conceptual model from the spreadsheets, designed a normalized 3rd Normal Form (3NF) relational database, and implemented a reliable, idempotent ETL pipeline in Python and Postgres.

The Strategic Outcome

A structured, clean, and relational data foundation. This “single source of truth” is the essential first step for any trusted analytics, reporting, or predictive modeling.

Before → **After**



Technical Snapshot

Key Deliverables

- ERD (Mermaid)
- SQL DDL
- Python ETL scripts (`load_data.py`)
- Baseline model (`03_model_baseline.ipynb`)

Core Technologies

- Postgres
- Python
- pandas
- sqlalchemy

Acceptance Criteria

- Schema builds cleanly; ETL is idempotent
- ≥95% of rows load with referential integrity

AI Collaboration 💡

- **Tutor Mode:** An LLM proposed ERD variants to compare against the human-designed schema.
- **Coder Mode:** An AI-generated data dictionary draft was human-edited and finalized.

Stage 2: From Errors to Integrity

STRUCTURE
Spaghetti → Schema

INTEGRITY
Errors → Integrity

SIGNAL
Dalasot → Detection

STRATEGY
Saibers → Bewategy

Auditing, Repairing, and Validating Data for Trustworthy Analytics

The Business Challenge

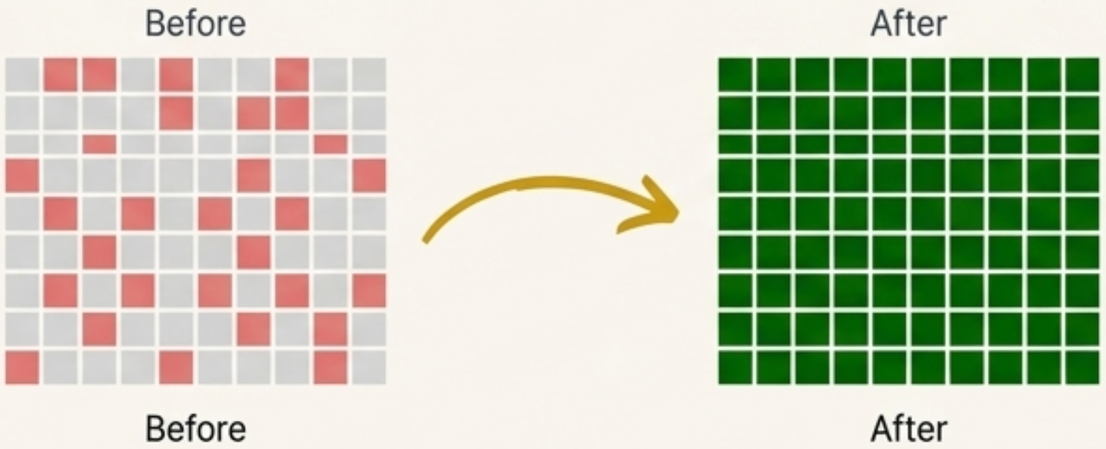
The new database contains messy, real-world data issues: missing values (MNAR/MAR), duplicates, category drift, and outlier spikes. Without correction, any downstream analysis or model would be unreliable.

The Technical Action

Executed a systematic data quality audit to detect all injected issues. Applied principled imputation methods (mean/median, kNN, regression, time-aware interpolation) with clear justification for each choice.

The Strategic Outcome

A reliable, auditable dataset ready for modeling. This step builds organizational trust in the data and ensures that business decisions are based on a valid and reproducible foundation.



Technical Snapshot

Key Deliverables

- Data quality audit notebook (`02_data_quality.ipynb`)
- Auto-generated quality report (`data_quality.md`)
- Validation tests (`evaluation.py`)

Core Techniques

- Missingness matrix
- Distributional shift analysis
- Deduplication, canonicalization
- Multiple imputation

Acceptance Criteria

- All introduced issues detected and fixed/justified
- Imputation rationale is documented
- Baseline model re-trains without suspicious metric infiation

AI Collaboration 💡

- **Tutor Mode:** An LLM generated a comprehensive Data Quality Checklist to guide the audit.
- **Coder Mode:** AI suggested imputation candidates; the data scientist selected and validated the final methods with statistical tests.

Stage 3: From Data to Signal

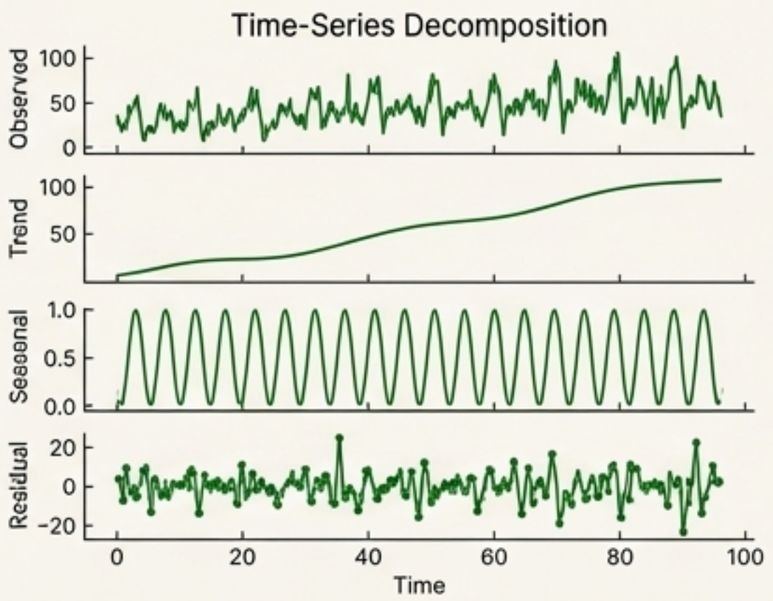
Discovering Trends and Forecasting Future Demand

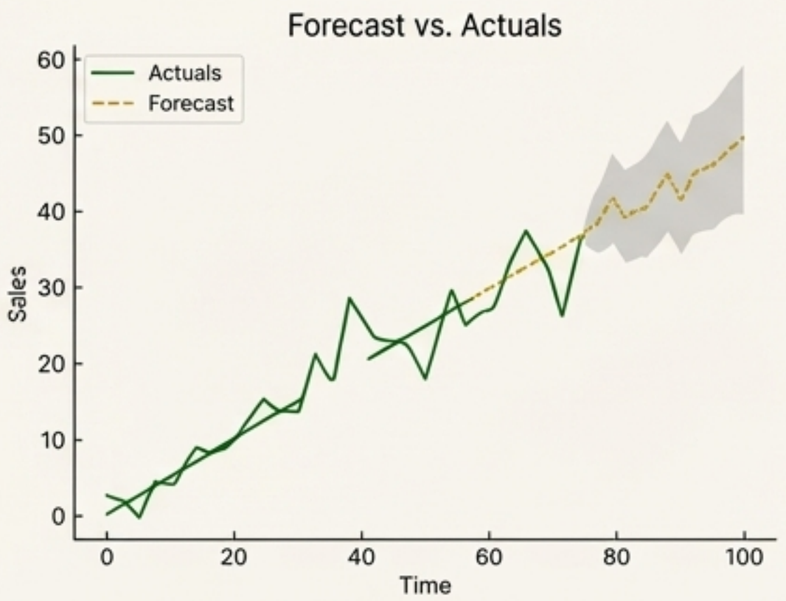
STRUCTURE → INTEGRITY → **SIGNAL** Data to Signal → STRATEGY

The Business Challenge: With clean historical data, leadership needs to understand demand drivers. Key questions: How do holidays impact sales? What are the seasonal patterns? How can we accurately forecast for better staffing and promotion planning?

The Technical Action: Engineered features from calendar and external weather data. Used time-series decomposition (STL/LOESS) to isolate trends. Compared forecasting models (SARIMA/Prophet vs. ML like XGBoost) using robust backtesting.

The Strategic Outcome: Quantified, actionable insights. The business can now see the specific uplift from holidays and understand weather elasticity. Accurate forecasts provide a data-driven basis for operational planning.





Technical Snapshot

Key Deliverables

- Trend analysis notebook (`03_trends.ipynb`)
- Forecasting model notebook (`04_forecasting.ipynb`)
- Visual dashboards

Core Techniques

- Time-series decomposition
- k-means clustering (on demand patterns)
- SARIMA, Prophet, XGBoost
- Rolling-window backtests

Acceptance Criteria

- Clear evidence of holiday, seasonality, and weather trends
- At least two forecasting approaches compared
- Effect sizes are quantified with business interpretation (e.g., "a hot day drives a 15% increase in beverage sales")

AI Collaboration 💡

- **Tutor Mode:** An LLM was used for hypothesis generation and feature ideation (e.g., "what if we create a feature for `lead/lag` days around a holiday?").
- **Coder Mode:** AI-assisted automated chart captions were generated and then edited by a human for clarity and insight.

Mini-Roadmap

STRUCTURE → INTEGRITY → SIGNAL → STRATEGY

Stage 4: From Shoppers to Strategy

Modeling Membership Effects and Simulating Promotion ROI

The Business Challenge

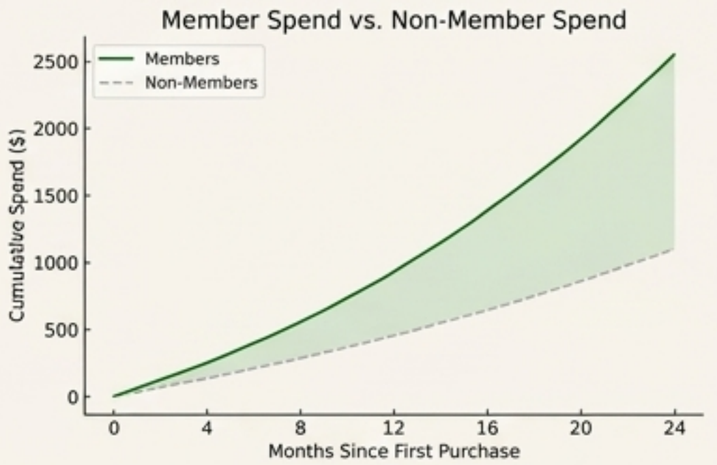
40% of customers are loyalty members. Leadership needs to know if the program is working: How does membership change shopping behavior? What is the actual profit uplift from targeted promotions?

The Technical Action

Performed cohort analysis to identify differences in frequency, basket size (AOV), and category mix between members and non-members. Built causal uplift models (T-learner) to isolate the effect of promotions from self-selection bias.

The Strategic Outcome

A strategic decision-making tool. The analysis delivers a defensible measure of the membership program's value and a simple ROI simulator to help managers design more profitable promotions, moving from guesswork to data-driven strategy.



Technical Snapshot

Key Deliverables

- Uplift models (`modeling.py`)
- ROI simulation notebook (`06_uplift_simulation.ipynb`)
- 'Promotion Playbook' executive summary

Core Techniques

- Cohort analysis
- Uplift modeling (T-learner, causal forests)
- Scenario simulation, sensitivity analysis

Acceptance Criteria

- Statistically clear member vs. non-member differences
- Validated uplift model
- ROI simulator reports incremental revenue, cost, and margin

AI Collaboration 💡

- **Tutor Mode:** An LLM was used to draft 'promotion scenario cards' (inputs and expected outputs) to test the simulator's logic.
- **Agent Mode (Prototype):** Auto-generated cohort summary statistics were human-reviewed for final reporting.

Human–AI Collaboration: A Practical Research Framework

This project doubles as a living lab to study how data science workflows evolve with AI.
We observed and applied three distinct modes of collaboration across the analytics lifecycle.

1. AI as Tutor: Guidance & Ideation

Using AI to propose variants, generate checklists, and brainstorm hypotheses. The human remains the primary creator, using AI as an expert guide.

Example from Project: LLM proposed ERD variants in C1 and generated a data quality checklist in C2.

2. AI as Coder: Co-Creation & Acceleration

AI generates candidate code, diagnostics, and report drafts. The human validates the logic, ensures correctness, and refined the final output.

Example from Project: AI suggested imputation methods in C2 and drafted chart captions in C3.

3. AI as Agent: Delegation & Oversight

AI executes well-defined tasks (e.g., running models, summarizing data) based on human intent. The human's role shifts to oversight, governance, and auditing the results.

Example from Project: Auto-generated cohort summaries in C4 were human-reviewed before inclusion in the final report.

The goal is not automation for its own sake, but shared understanding. We documented prompts and tracked where human validation overruled AI suggestions in `docs/ai_workflow.md` to ensure traceability.

An Open Invitation to a Living Laboratory

This project is an evolving research ecosystem. It serves as a learning laboratory and a collaboration framework for academics, data practitioners, and industry partners.

We welcome collaboration.

- **Contribute:** Collaborate on model design or advanced feature engineering.
- **Teach:** Integrate modules into university courses or professional training programs.
- **Partner:** Provide real or synthetic datasets to test the framework on new challenges.
- **Sponsor:** Support outreach or mentorship for the ongoing Challenge 4 (Membership Modeling & ROI).

Interested collaborators can reach out via the StarShipTutor repository or by opening a discussion on the project wiki.

`github.com/StarShipTutor/Patrick-Horgan-Portfolio`

The long-term goal is not automation for its own sake, but collaboration. We seek to build a shared understanding of how humans and AI co-create insight responsibly, reproducibly, and at scale.