ACME Legacy Reimbursement - Project Summary

Species: **default species**

Project Overview

ACME's legacy reimbursement engine has operated as a black box for decades. Our mandate is to reproduce its outputs and explain its embedded logic so stakeholders can compare legacy and modern systems with confidence. The effort balances statistical fidelity with business clarity, preserving quirks such as tiered adjustments and rounding artifacts that materially affect payouts.

Team Roles & Responsibilities

Team Member	Role	Focus
Ayushi	Technical Lead / ML Engineer	Feature engineering, modeling approach, reproducibility, pipeline design
Mike	Business Analyst	PRD/interview synthesis, business logic hypotheses, narrative alignment
Colyn	Documentation & Communication	Report structure, clarity of written deliverables, presentation flow
Matt	QA / Testing / Data Wrangler	Data integrity checks, spot-testing model outputs, edge-case validation

Executive Overview

We are reverse-engineering ACME's legacy reimbursement engine to match and explain its outputs. This report summarizes progress across three phases: initial discovery and behavioral hypotheses, feature engineering with baseline models, and correlation-driven insights to guide next modeling steps. The aim is to balance technical rigor with business clarity while preserving legacy artifacts (tiering, rounding quirks) that stakeholders care about.

Dataset Reviewed

The dataset includes 6,000 records (1,000 public with reimbursement labels and 5,000 unlabeled private cases). We examined the historical reimbursement dataset, which includes:

- **Trip Duration (days)**
- **Total Mileage**
- **Total Receipts**
- **Reimbursement Amount** (system output)

A data cleaning notebook was created to: - Remove formatting inconsistencies - Validate numeric ranges - Prepare features for further modeling

Result: A clean, analysis-ready dataset.

Key Observed Patterns

Factor	Observation	Interpretation
Trip Length	Reimbursement is **more generous** around **4–6 days**, declines for >7 days trips	Suggests a **sweet spot and long-trip penalty**
Mileage	Value-per-mile decreases after ~100 miles/day	Indicates a **non-linear mileage adjustment curve**
Receipts	Higher receipts do **not** consistently produce higher reimbursement	Suggests **diminishing returns and upper/lower spend penalties**
Non-linear Behavior	Adjustments change together, not independently	Legacy system likely uses **multiple interacting rules**
Rounding Artifacts	Some reimbursements show small irregularities	Suggests **bugs/features that must be preserved**

Phase 1 - Discovery, Data Quality, and Business Logic Hypotheses

Goal

Understand the structure and quality of the public/private reimbursement data and form testable hypotheses about the legacy system's business rules.

What we did

- Flattened public and private JSON cases into tabular data.
- Checked for missing values, duplicates, and non-positive entries across duration, miles, receipts, and reimbursement; none required removal.
- Ran IQR-based outlier scans and confirmed all points remained within reasonable business bounds.
- Produced distributions and correlation views for duration, miles, receipts, and reimbursement.
- Compared public vs private datasets to check for domain drift.

Key evidence

- **Trip duration:** mostly 1-5 days with a small tail beyond a week; moderate correlation with reimbursement (r ~ 0.45), so duration matters but is not the main driver.
- **Miles traveled:** right-skewed; most cases under ~500 miles; strong correlation with reimbursement (r ~ 0.80).
- **Receipts:** strongest correlation with reimbursement (r ~ 0.85); high spend drives payouts but hints at diminishing returns at the top end.
- **Public vs private:** similar ranges and shapes; only minor variance differences, so combined training is reasonable.

See Figure 1 for the correlation heatmap and distribution views.

Takeaway (business view)

The legacy engine is not purely linear. It appears to reward balanced, mid-length trips with reasonable mileage and disciplined spend, using tiered mileage adjustments and diminishing-return curves rather than straight-line rules.

Phase 1 visuals

The charts below summarize the main patterns:

- Correlation: receipts and miles dominate; duration contributes modestly.
- Receipts skew: long right tail underscores diminishing returns and the need for non-linear handling of high spend.
- Public vs private: overlapping distributions suggest shared logic across datasets.
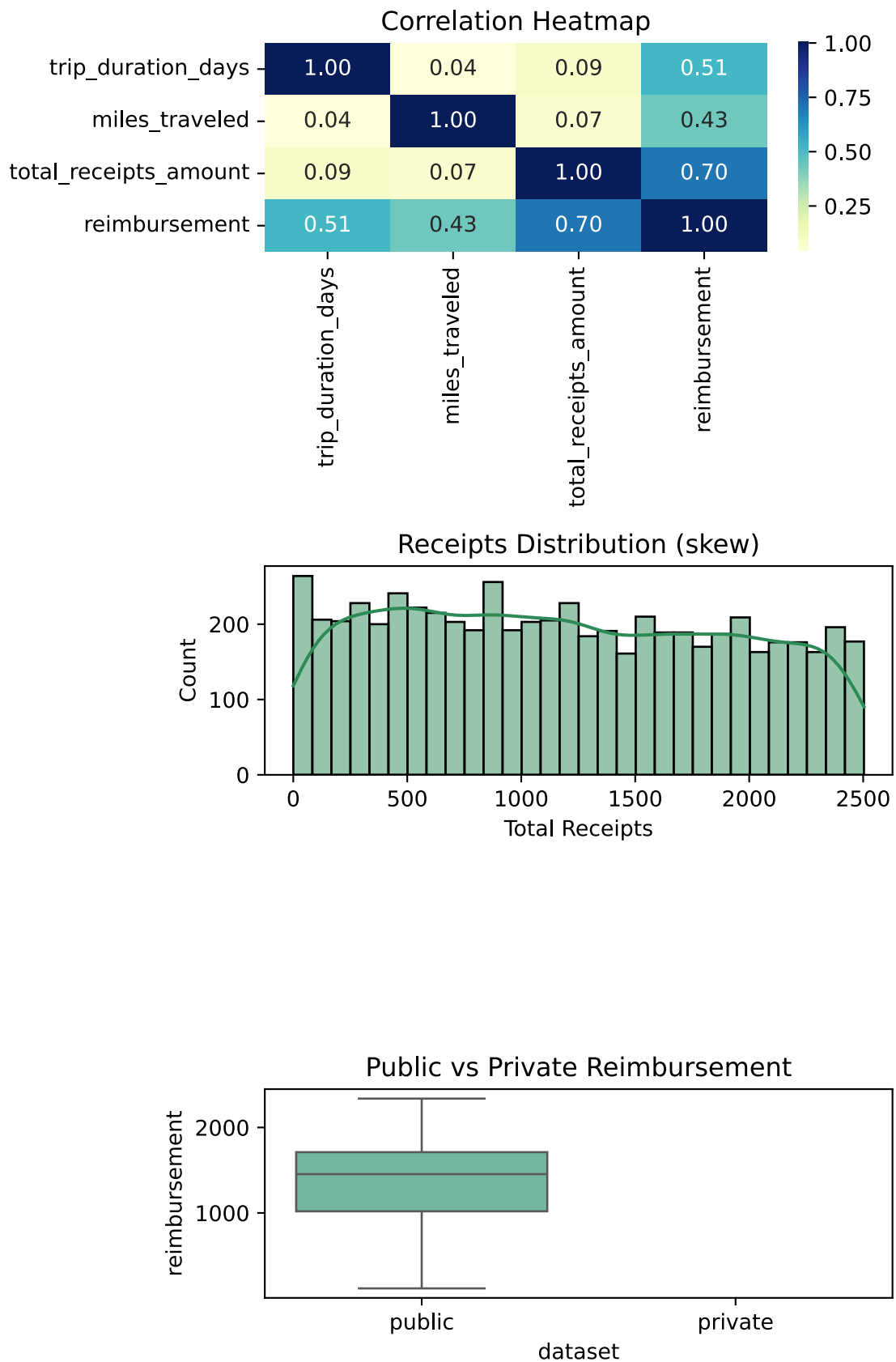

Figure 1: Phase 1 visuals: correlation structure, receipts skew, and public vs private alignment.

Phase 2 - Feature Engineering and Baseline Modeling

Goal

Translate Phase 1 behavioral findings into engineered efficiency features and measure how well simple baselines approximate the legacy engine.

What we did

- Engineered rate and balance features: cost_per_day, cost_per_mile, miles_per_day, cost_ratio (cost_per_day / cost_per_mile).
- Guarded divides by zero, replaced inf/NaN, confirmed IQR checks required no row drops; used a 75/25 train/test split.
- Trained baseline models: Linear, Ridge, Lasso, and Polynomial (degree 2) on the seven engineered features.

Key evidence

- Correlations with reimbursement: receipts (~0.70) strongest; trip duration (~0.51) and miles traveled (~0.43) moderate; rate features are weak alone but help via interactions.
- Model fit: Linear/Ridge/Lasso R^2 ~0.78 with RMSE ~200; Polynomial R^2 ~0.89 with RMSE ~142, showing non-linear terms are needed.
- Residuals for the polynomial model center near zero with a few long-tail trips driving the extremes, consistent with the skew observed in Phase 1.

Takeaway (business view)

Linear structure covers much of the variance, but the legacy engine's tiering and diminishing-return behavior needs non-linear interactions. The engineered efficiency features add signal when combined non-linearly; regularization keeps coefficients stable.

Phase 2 visuals

The charts below mirror the Phase 1 layout: - Correlation heatmap across engineered features and reimbursement. - Baseline model fit (R^2) comparison. - Actual vs predicted for the best baseline model (polynomial degree 2).
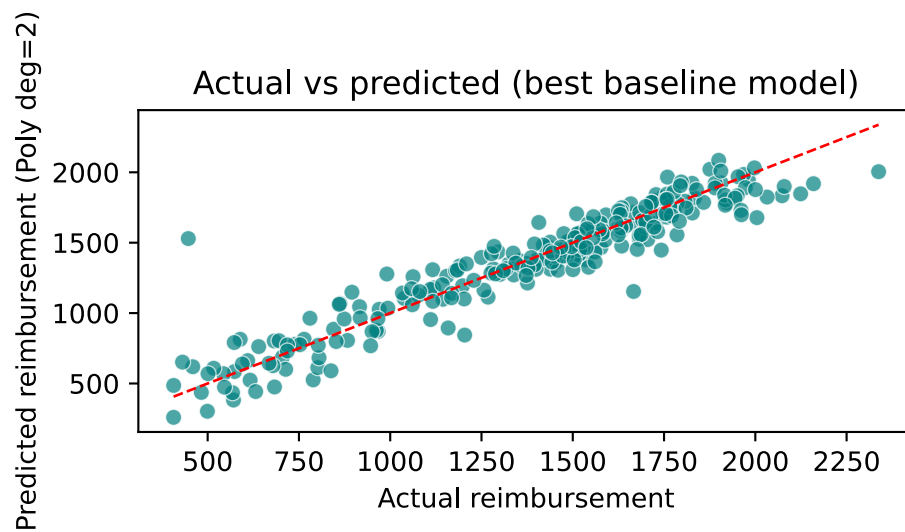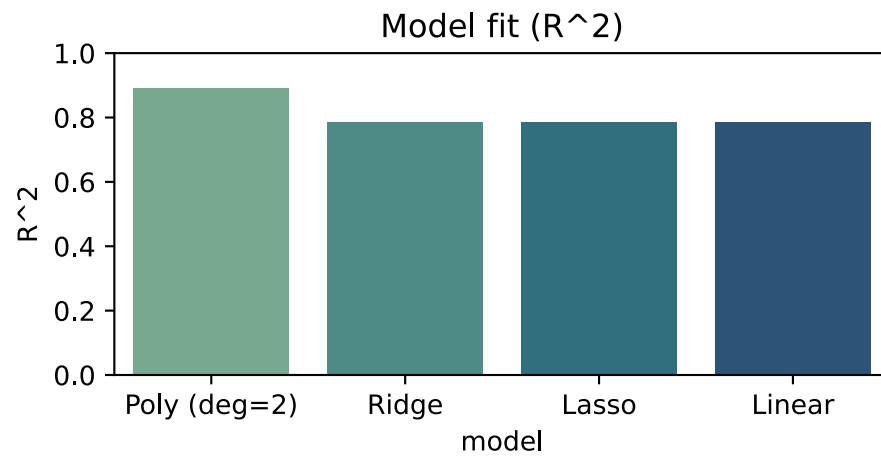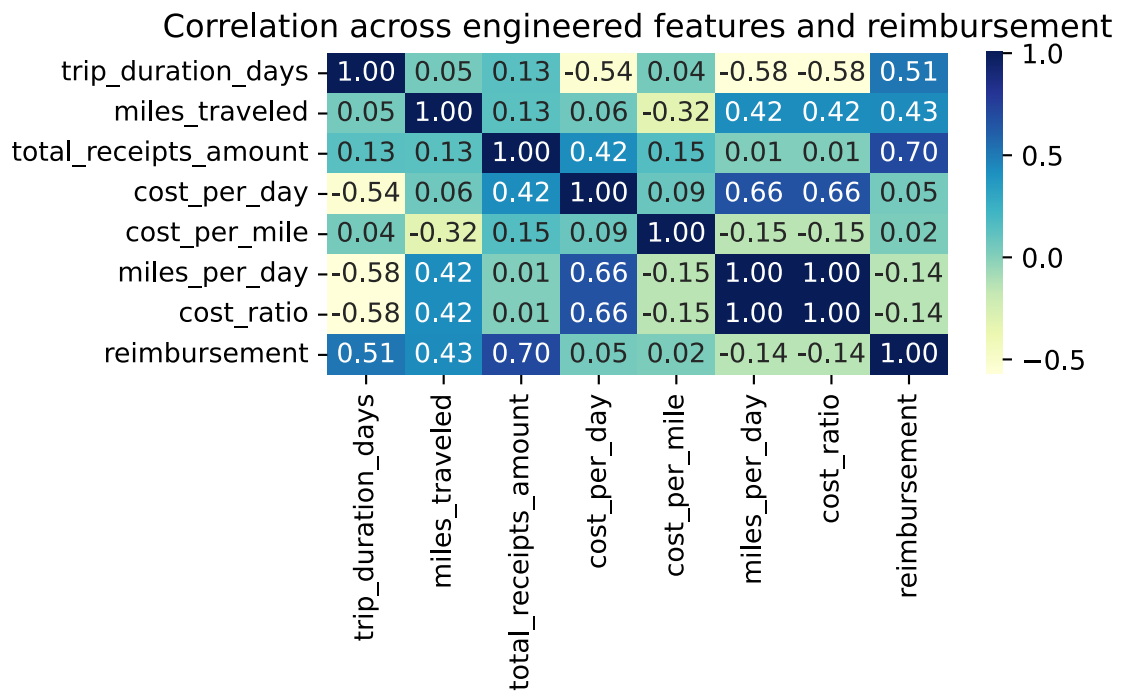
Figure 2: Phase 2 visuals: engineered feature correlations, model fit comparison, and polynomial predictions.

Phase 3 - Modeling Outlook & Integration Plan

Objectives - Move beyond polynomial baselines to tree/ensemble methods that can capture tiering, interaction, and diminishing-return effects without manual specification. - Harden the pipeline toward production requirements (performance, determinism, minimal dependencies).

Planned actions - Add random forest and gradient boosting models; tune hyperparameters with cross-validation. - Evaluate against business thresholds (within $0.01 / within $1.00) in addition to MAE/RMSE. - Apply interpretability tools (SHAP, partial dependence) to translate model behavior into stakeholder-ready rules. - Implement explicit business constraints where needed (e.g., long-trip penalties, capped daily spend, monotonic mileage effects).

Next Steps and Recommendations

- Broaden models: add tree/ensemble approaches (random forest, gradient boosting) alongside polynomial baselines.
- Strengthen evaluation: introduce cross-validation and report MAE/RMSE plus business thresholds (percent within +/-$0.01 and +/-$1.00).
- Encode business effects: implement explicit long-trip penalties and spend-discipline rules (piecewise mileage rates, capped daily spend, monotonic constraints).
- Interpretability: apply SHAP/partial dependence to translate model behavior into stakeholder-friendly rules; document rounding/post-processing to preserve legacy artifacts.
- Production readiness: package the pipeline into the required 3-argument script, ensure <5s runtime, remove external dependencies, and fix seeds/versioning for deterministic outputs.