## **What Was Done**

* Aggregated claims and patient demographics at the provider level, creating 30+ fraud-relevant features.
* Merged inpatient and outpatient claims with a “ClaimType” flag.
* Calculated core metrics: claim counts, reimbursement sums/means/medians/std, deductible totals, max claim value.
* Added patient demographic metrics: unique beneficiary count, average age, % deceased, gender ratio, race diversity, chronic condition rates.
* Built claim pattern features: claims per beneficiary, % with multiple claims, diagnosis/procedure diversity, average days between claims.
* Created high-risk billing indicators: % claims > $10,000, % weekend claims, % with all diagnosis/procedure slots filled.
* All features merged with fraud labels for modeling.
* Nulls handled sensibly for providers with missing data.

## **Quality Checks**

* No missing Provider IDs.
* No duplicate providers.
* All engineered features checked for outliers and reasonable ranges.
* Providers with no claims have 0s for claim-based features.
* Feature distributions and correlations ready for further review.

## **Output**

* Main feature table: /data/processed/feature\_table.parquet
* Train/test split feature files as needed
* This documentation for audit and reproducibility