1. Project Overview

This project aims to analyze the claims performance of an auto insurance company by examining key metrics such as **claim frequency, severity, payout accuracy, processing time, and fraud detection effectiveness**. By leveraging historical claims data, the project will identify **trends, outliers, and inefficiencies**. The objective is to gain actionable insights that improve **operational efficiency,** enhance the **accuracy of reimbursements**, and support strategic decision-making. Ultimately, the analysis will guide the company toward more data-driven claims management practices, optimizing both customer experience and financial performance.

2. Executive Summary

A. Key Findings:

- **Loss ratio volatility is severity-driven**, not frequency-driven. Claim counts are stable to declining, while average claim severity rises materially mid-year.
- A **small proportion of severe claims** contributes disproportionately to incurred losses, settlement duration, and reserve uncertainty.
- Loss emergence follows a **predictable lag**: claim intake peaks before severity, closures, and loss ratio deterioration.
- Current **pricing, reserving, and estimation frameworks underperform in the tail**, increasing earnings and capital volatility.
- Operational performance is working well. The **primary risk is inflation-driven severity exceeding model and pricing assumptions**.
- Without intervention, declining frequency may mask **structural profitability deterioration**.

B. Conclusion:

The portfolio faces **severity inflation and tail-risk concentration**, requiring targeted actuarial, underwriting, reserving, and reinsurance responses.

3. Business Context & Objective

Auto insurance claims are a major cost driver and directly impact insurer profitability and customer satisfaction. Efficient claims handling is essential to controlling loss ratios, reducing operational costs, and maintaining policyholder trust. However, claims operations often face challenges such as long processing times, inconsistent assessments, and uneven performance across agencies and assignees.

The objective of this project is to evaluate the end-to-end claims handling performance using historical claims data. Specifically, the analysis aims to:

- Measure claim processing efficiency across different workflow stages
- Identify factors driving claim severity and payout variability
- Compare performance across agencies, claim adjusters, coverage types, and claim severities
- Detect operational bottlenecks and potential areas of cost leakage

The insights from this analysis will support data-driven decision-making to improve claims operations, enhance payout accuracy, and optimize overall claims management performance.

4. Scope & Key Questions

A. Scope of Analysis

This analysis focuses on auto insurance claims processed within the selected reporting period (January to November 2022). Each claim is represented at its latest available status after data consolidation and cleansing.

The scope of the analysis includes:

- Claims across all coverage types handled by internal and external agencies
- Both approved and non-approved claims (paid out, denied, and cancelled)
- Operational timelines from claim submission through final resolution
- Financial metrics related to estimated and final claim compensation

B. Key Business Questions

The analysis seeks to answer the following key questions:

Claims Efficiency & Operational Performance

- What is the average claim processing time, and how does it vary by claim severity, coverage type, vehicle segments and customers?

- Which workflow stages contribute most to processing delays?

- How does claim processing efficiency differ across agencies and claim handlers?

Claims Volume & Throughput Trends

- How does claim volume change over time?

- Are there monthly or seasonal fluctuations in claim activity?

- Are there monthly or seasonal fluctuations in claim processing performance?

- What percentage of claims are approved, denied, or cancelled?

Financial Performance

- What is the average paid claim amount, and how does it vary by severity, coverages and vehicles?

- How closely do initial estimates align with final compensation amounts?

- Are there patterns indicating potential overpayment or cost leakage?

Risk & Control Indicators

- Are there claims with unusually long processing times or large estimate-to-payout gaps?

- Do these patterns suggest potential operational inefficiencies or fraud risks?

- Are claims disproportionately filed shortly after policy inception or near policy expiration?
  Do these timing patterns correlate with higher payouts or abnormal behaviors?

Strategic & Predictive Insights

- Are there seasonal or monthly spikes in claim volume or severity?

- How can staffing and resources be optimized during peak periods?

Predictive Risk Factors

- Which factors predict:
  - High-cost claims
  - Long processing times
  - High likelihood of rework or denial?

Portfolio Performance

- What is the claims ratio (claims paid ÷ premium earned)?
- How does it vary by product, coverage, or customer segment?

5. Data Overview & Preparation

A. Data Sources

The analysis is based on historical auto insurance claims data sourced from two primary datasets:

- **Claims Dataset (DF1)** – the main claims-level dataset containing claim status, workflow dates, coverage details, and transaction history.
- **Claims Staging Dataset (DF2)** – a supplementary dataset providing compensation amounts, estimated payouts, and updated financial information.

Both datasets required extensive preprocessing to ensure data consistency, accuracy, and analytical readiness.

B. Initial Data Quality Assessment

A preliminary data quality review identified several issues that required remediation before analysis:

- A high volume of **missing (NULL) values**

- **Duplicate columns** and ambiguously named fields
- **Inconsistent data types** across date, numeric, and categorical variables
- **Multiple records per claim** caused by coverage-level granularity and workflow updates

To address these issues, a multi-stage data cleaning and transformation pipeline was implemented.

C. Data Cleaning & Transformation Process

Stage 1 – Data Standardization (DF1_Staging1)

- Corrected and standardized data types (dates, numeric fields, categorical values).
- Translated and normalized column values to ensure consistency across the dataset.
- Reviewed vague or ambiguous column names and assessed potential duplicates.

Stage 2 – Deduplication (DF1_Staging2)

- Removed exact duplicate rows to eliminate redundant records.
- Ensured each record represented a unique claim event or update.

Stage 3 – Workflow Normalization & Data Enrichment (DF1_Staging3)

Claim Status Standardization

Claim statuses were mapped and normalized into a unified workflow to reflect the end-to-end claims lifecycle:

> Submitted → Assessment Pending → Estimate Pending → Documents Pending → In Process → Approved / Denied / Cancelled → Paid Out

This standardized workflow enables consistent measurement of processing times and operational performance.

Workflow Date Engineering

- Created new date columns corresponding to each major workflow milestone.
- Enabled calculation of key cycle-time metrics such as FNOL-to-Open, Open-to-Decision, and Approved-to-Paid durations.

Missing Value Resolution

- Cross-referenced DF1 with DF2 to identify and fill missing compensation and estimated payout values.
- For claims present in both datasets, financial fields were populated from DF2 where missing in DF1.

Record Consolidation Logic

When multiple records existed for the same claim:

- Retained the **most recently updated record** based on updated_date.
- Removed obsolete historical rows to ensure one consolidated record per claim.

D. Data Dictionary

This section defines the key variables used in the analysis, including their business meaning, data type, and analytical purpose. All fields listed below are part of the final cleaned and consolidated dataset, with one record per claim.

Claim & Policy Identification

Field Name	Data Type	Description
claimnumber	Text	Unique identifier for each insurance claim. Primary key across workflow stages.
policycode	Text	Identifier of the insurance policy associated with the claim.

coverage	Text	Type of insurance coverage applicable to the claim.
customer_name	Text	Name of the policyholder or insured customer.
policy_start_date	Date	Policy coverage start date.
policy_end_date	Date	Policy coverage end date.

Claim Lifecycle Dates

Field Name	Data Type	Description
claim_accident_date	Date	Date the insured event occurred.
claim_submitted_date	Date	Date the claim was reported (FNOL).
claim_opened_date	Date	Date the claim was opened for processing.
process_start_date	Date	Date the claim entered active processing.

assessment_pending_date	Date	Date the claim entered assessment pending.
estimate_pending_date	Date	Date the claim entered estimate pending.
documents_pending_date	Date	Date the claim entered documents pending.
claim_approved_date	Date	Date the claim was approved.
updateddate	Date	Date of the most recent update to the claim record.
payment_date	Date	Date the policy premium payment was made by the customer (not claim payout).

Claim Characteristics

Field Name	Data Type	Description
claim_status	Text	Current standardized status of the claim.
assignee_fullname	Text	Name of the assigned claim handler or adjuster.

agencycompensation	Text	Descriptor related to agency compensation.
processdays	Integer	Total number of days taken to process the claim.
damage_type	Text	Classification of damage sustained.
vehicle_type	Text	Type of insured vehicle.
vehicle_make	Text	Vehicle manufacturer or brand.
customer_type	Text	Customer classification.

Financial & Compensation

Field Name	Data Type	Description
claim_estimate_first	Decimal	Initial estimated claim amount.
total_claim_estimate	Decimal	Final estimated claim amount after reassessment.
compensation_total	Decimal	Final claim compensation amount.

cost_copay_deductible	Decimal	Deductible or co-payment amount borne by the policyholder.

6. Methodology & Analytical Approach

A. Descriptive Analytics

Establish a factual baseline of portfolio performance and claim behavior.

Key techniques:

- Time-series analysis of claim frequency, severity, and loss ratio (incurred basis)
- Distribution analysis of claims by severity band, vehicle type, and coverage type
- Settlement duration
- Actual vs Estimated Compensation comparison (% Estimation Error)

B. Diagnostic Analytics

Apply to identify root causes and structural drivers behind observed trends.

Key techniques:

- Loss ratio decomposition (frequency vs severity effects)
- Lag analysis linking claim intake, closure, severity, and loss ratio emergence
- Outlier and dispersion analysis to identify tail risk
- Backlog and capacity analysis

C. Predictive Analytics (Light / Directional)

While no machine learning models were built, forward-looking risk behavior was inferred using observed lag patterns and sensitivity analysis.

Key techniques:

- Severity-based stress testing (+5%, +10%, +15%)

- Lag-based early warning indicators (intake → severity → loss ratio)
- Scenario analysis on settlement duration and backlog effects

7. Key Findings & Insights

A. Profitability pressure is severity-driven, not volume-driven

- Claim frequency is declining and total claim counts remain stable.
- Average reimbursement and loss severity rise steadily, peaking in August–October.
- Loss ratio spikes (>60%) align tightly with increases in high-severity and severe claims, not claim volume.
- Two-wheelers dominate policy count, but private four-wheelers and severe claims dominate financial impact.

★ **Insight:**
Profitability risk is driven by claim size inflation, not operational failure or claim count growth. The reduction of frequency creates a false sense of improvement while overall profitability weakens.

B. Claim costs follow intake with a predictable lag

- New claim submissions peak **before** severity and loss ratio peaks.
- Closed claims volume increases later and coincides with severity spikes.
- Mid-year (July–September) closure capacity lags submissions, forming a backlog ahead of the September peak.

★ **Insight:**
Loss ratio deterioration follows a predictable lag, giving an early warning window that is not currently used. Costs are addressed only after they have escalated.

C. Severe claims disproportionately drive volatility and operational strain

- Severe claims are low frequency but drive:
  - A disproportionate share of total loss
  - Longer settlement durations

    - Higher reserve uncertainty

- Severity is strongly correlated with settlement time and estimation error.

- Collision and partial/total loss claims dominate loss exposure; personal injury is minor by comparison.

★ Insight:
This is a tail-risk concentration problem. Managing average claims will not stabilize results, severe claims require distinct controls.

D. Estimation accuracy breaks down for large losses

- Actual vs. estimated compensation aligns well at low–mid severities.
- Dispersion and extreme outliers increase sharply at higher claim values.
- % error rises when:
    - Severity increases
    - Closed-claim volume drops
- Errors cluster by assignee, agency, and claim type.

★ **Insight:**
The estimation framework works for routine claims but breaks down for severe losses. Errors are systematic, not random, indicating control and capability gaps, not noise.

E. Early policy-period risk signals potential anti-selection

- Claims concentrate heavily in the first month after policy inception, stabilizing after 10-25 days.
- Claim frequency declines steadily over the policy term.

★ **Insight:**
The concentration of early claims signals potential anti-selection risk, requiring strengthened underwriting and claims controls at policy inception.

8. Recommendations & Action Plan

- Recalibrate financial assumptions for severity inflation

    - Update pricing, reserving, and reinsurance assumptions to reflect **severity growth**, not just frequency trends.
    - Stress-test profitability under +5%, +10%, +15% severity scenarios.
- Introduce severity-based claim controls
    - Implement **early large-loss flags at FNOL** with escalation thresholds.
    - Apply distinct reserving logic for top 5–10% of claims.
    - Route severe claims to specialized handlers earlier.
- Apply targeted pricing, deductibles, and underwriting controls to private four-wheelers, collision and total loss coverage
- Strengthen operational resilience mid-year
    - Add surge capacity (staffing or vendors) during July–September.
    - Prevent backlog accumulation that amplifies severity and reserve uncertainty.
- Institutionalize leading risk indicators. Track monthly indicators that move before loss ratio deterioration:
    - Severity mix (% Severe + High)
    - % estimation error
    - Backlog size
    - Settlement duration for severe claims

9. Executive-Level Questions

1. How much of severity growth is inflation vs coverage or mix?
2. Are severe claims developing upward after initial reserve?
3. Are collision losses exceeding expected severity curves?
4. Why does closure capacity drop mid-year — people, process, or tools?

5. Are deductibles and limits still fit for current repair inflation?
6. What happens to profitability if severe claim frequency rises by 0.1%?

10. Data Limitations, Assumptions & Constraints

A. Data Limitations

The analysis is subject to the following constraints:

- Limited cost decomposition
  - Claim-level cost drivers (e.g., labor, parts, medical, repair vendor pricing) are not available, restricting root-cause attribution of severity inflation.
- Lack of granular process timestamps
  - Detailed dates for claim lifecycle stages (assessment, examination, documentation, approval) are not fully available, limiting precise bottleneck and cycle-time diagnostics.
- No vendor or repair network data
  - Prevents evaluation of cost concentration by workshop, hospital, or service provider.
- Limited exposure and usage data
  - Mileage, usage intensity, or behavioral risk indicators are not captured.
- No external inflation benchmarks
  - Inflation effects are inferred from claims outcomes rather than linked to CPI, repair indices, or medical inflation data.

B. Key Assumptions

- Claims data is complete and consistently reported across periods.
- Severity classification is stable over time.
- No material changes to claims handling or underwriting rules during the observation period.
- Incurred losses reflect best-estimate valuations at each valuation date.

C. Implications of Limitations

Due to these constraints:

- Results should be interpreted as risk-directional rather than actuarial point estimates.
- Findings are most suitable for strategy, pricing direction, reserving governance, and capital risk management, not for individual claim adjudication.

11. Conclusion:

Profitability pressure is driven by rising claim severity, not claim volume or operational failure. While frequency is stabilizing, higher average claim costs and severe losses are eroding results.

Loss ratio deterioration follows a predictable lag from claim intake to severity to incurred losses, meaning actions are taken after costs have already escalated. The estimation framework works for routine claims but underperforms on large and complex losses, where errors are systematic rather than random.

Severe claims, though few, drive disproportionate cost and volatility, and early-policy claim spikes indicate potential opportunistic behavior, requiring stronger early controls.

Overall, sustainable profitability requires severity-focused pricing, reserving, and early large-loss intervention.