Interim Report

Insurance Risk Analytics & Predictive Modeling

AlphaCare Insurance Solutions (ACIS)

Report Date: December 06, 2025

Covering Task 1 (EDA) and Task 2 (DVC Setup)

1. Executive Summary

This interim report presents progress on the Insurance Risk Analytics project for ACIS, covering comprehensive Exploratory Data Analysis (EDA) and Data Version Control (DVC) infrastructure setup. Analysis of 1,000,098 insurance records (Feb 2014 - Aug 2015) reveals critical risk segmentation opportunities. The overall portfolio loss ratio of 104.77% indicates claims exceed premiums, requiring immediate risk-based pricing optimization. DVC infrastructure ensures reproducible and auditable data pipelines for regulatory compliance.

2. Business Objective

Primary Goal: Identify low-risk customer segments for premium reduction opportunities and develop predictive models for optimal premium pricing. **Key Objectives:** (1) Discover low-risk targets for marketing, (2) Build ML models for premium optimization, (3) Support data-driven marketing strategies, (4) Ensure regulatory compliance through auditable analytics.

3. Task 1: Exploratory Data Analysis - Key Findings

3.1 Data Quality Assessment

Comprehensive data quality analysis was conducted to ensure reliable insights. The dataset contains 1,000,098 records with 52 original features. Data quality assessment revealed:

3.1.1 Missing Value Analysis

Column	Missing Count	Missing %
NumberOfVehiclesInFleet	1,000,098	100.00%
CrossBorder	999,400	99.93%

CustomValueEstimate	779,642	77.96%
Converted	641,901	64.18%
Rebuilt	641,901	64.18%
WrittenOff	641,901	64.18%
LossRatio	381,634	38.16%
NewVehicle	153,295	15.33%
Bank	145,961	14.59%
AccountType	40,232	4.02%

Key Findings: Fleet-related fields show 100% missing values, indicating these features are not applicable to most policies. CustomValueEstimate missing in 77.96% of records requires careful handling in modeling. LossRatio missing in 38.16% of records (where TotalPremium = 0). No duplicate rows detected, indicating clean data ingestion. Missing values in demographic fields (Bank: 14.59%, AccountType: 4.02%) are manageable and can be handled through imputation or exclusion strategies.

3.1.2 Outlier Detection

Outlier detection using Interquartile Range (IQR) method with 1.5x multiplier identified:

- TotalClaims: 2,793 outliers (0.28% of records) - extreme claim values requiring investigation
- CustomValueEstimate: 1,785 outliers (0.81% of non-missing records) - high-value vehicles
- TotalPremium: Negative values detected (min: -782.58) - data quality issue requiring correction
- SumInsured: Wide distribution with high variability (CV: 249.65%)

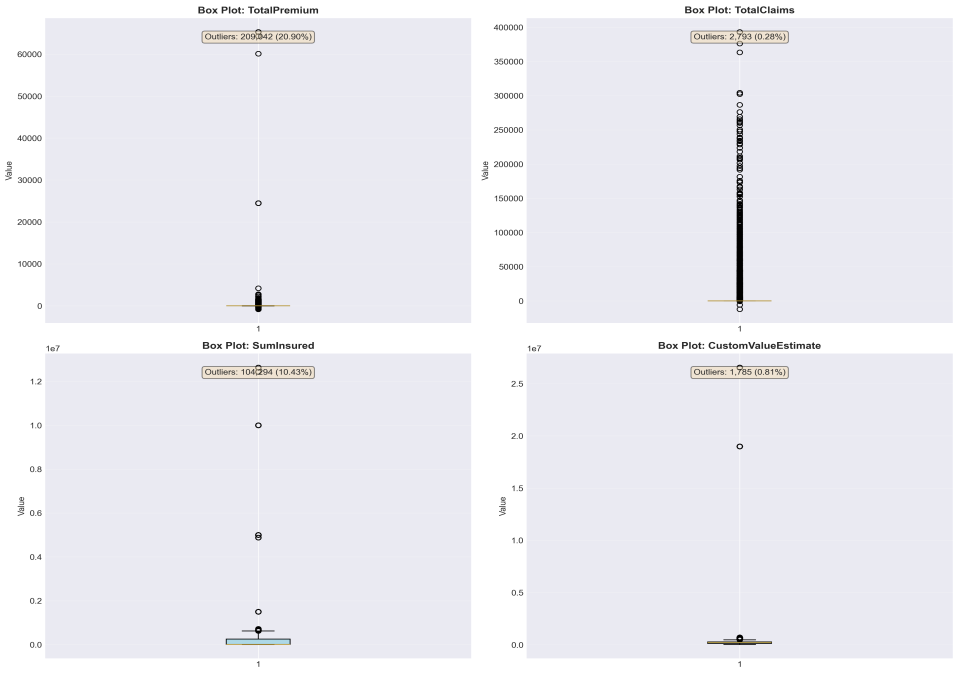

Figure 1: Box plots showing outlier detection for key financial variables

3.2 Overall Portfolio Metrics

Metric	Value

Total Premium	ZAR 61.9M
Total Claims	ZAR 64.9M
Loss Ratio	104.77%
Records Analyzed	1,000,098
Period	Feb 2014 - Aug 2015

3.3 Loss Ratio by Segment

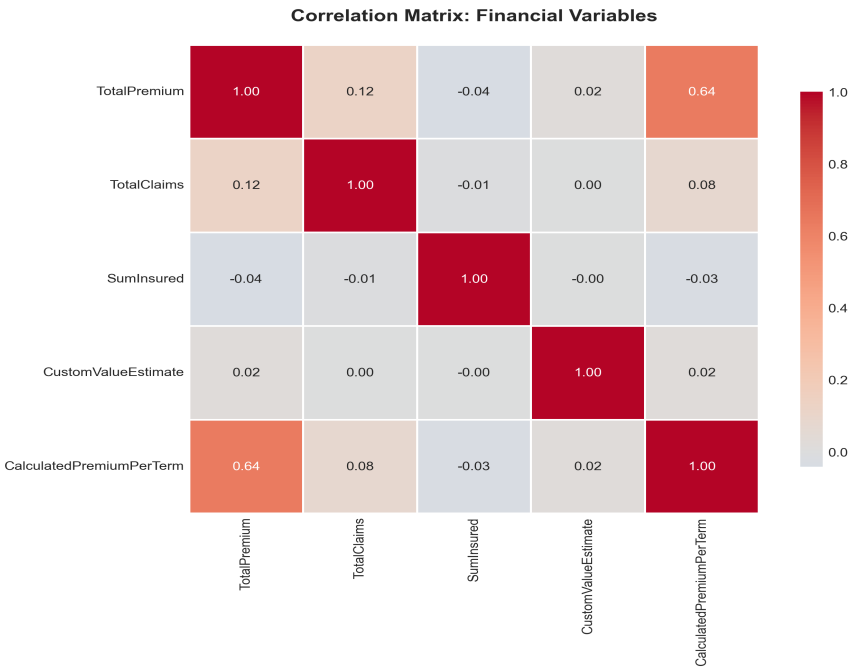

Figure 2: Correlation matrix of key financial variables

Province	Loss Ratio	Risk
Gauteng	122.20%	High
KwaZulu-Natal	108.27%	High
Western Cape	105.95%	High
North West	79.04%	Moderate
Mpumalanga	72.09%	Low
Free State	68.08%	Low
Limpopo	66.12%	Low
Eastern Cape	63.38%	Low
Northern Cape	28.27%	Very Low

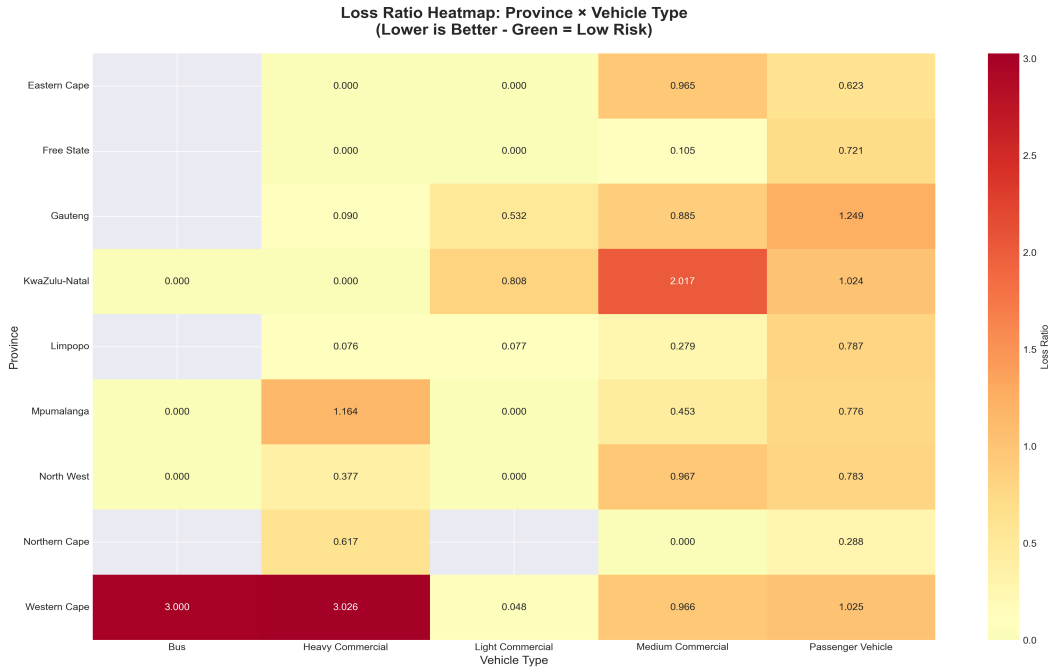

Figure 3: Loss ratio heatmap by Province and Vehicle Type (lower is better)

Key Insights: (1) **Geographic Risk:** 4.3x difference between highest (Gauteng: 122.20%) and lowest (Northern Cape: 28.27%) risk provinces. (2) **Vehicle Type:** Heavy Commercial shows 162.81% loss ratio vs Bus at 13.73% (12x difference). (3) **Gender:** Female drivers (82.19%) lower risk than males (88.39%). (4) **Vehicle Makes:** AUDI (271%) and HYUNDAI (399%) show extreme risk requiring immediate attention.

3.4 Temporal Trends

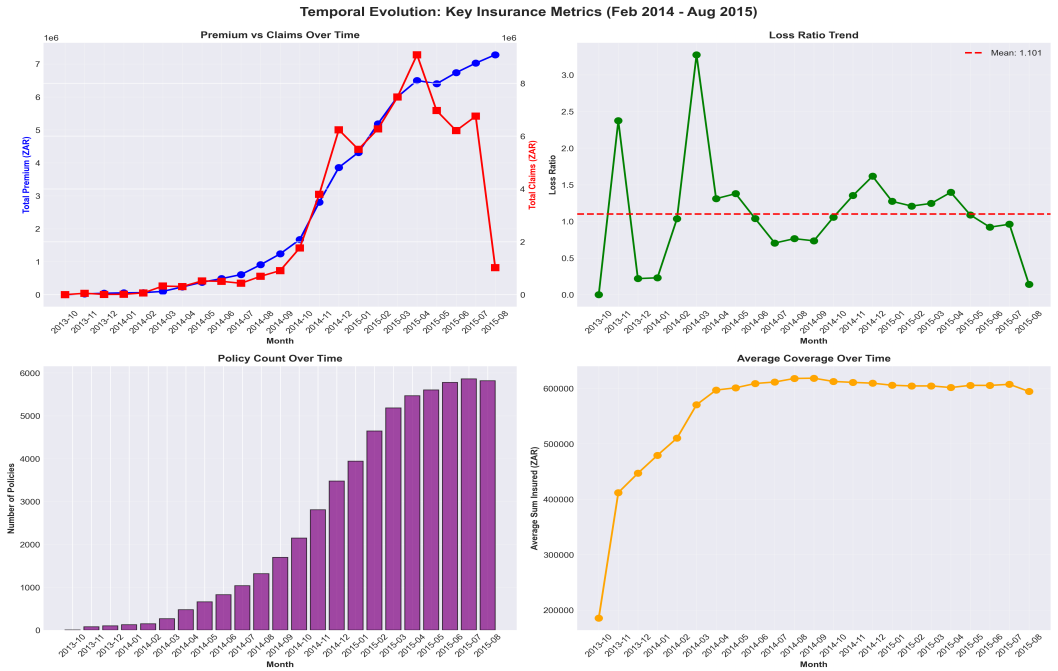

Figure 4: Temporal evolution of key insurance metrics (Feb 2014 - Aug 2015)

Monthly analysis reveals significant volatility: Loss ratio peaked at 139.64% in April 2015, with August 2015 showing unusually low 14.01% (requires investigation). Premium collection remained stable while

claims showed high variability. The temporal dashboard (Figure 4) shows clear patterns in policy count, premium, claims, and loss ratio trends over the 18-month period.

4. Task 2: Data Version Control (DVC) Setup

Implementation: DVC repository initialized with local remote storage at './dvc_storage/'. Primary data file (MachineLearningRating_v3.txt, ~529 MB) added to DVC tracking. Data metadata committed to Git while actual data stored in DVC remote. **Benefits:** (1) Complete reproducibility for regulatory audits, (2) Version control for datasets, (3) Storage efficiency (large files outside Git), (4) Team collaboration with specific data versions.

5. Key Findings and Opportunities

5.1 Low-Risk Opportunities

- Northern Cape, Eastern Cape, Limpopo provinces - premium reduction to attract customers
- Bus and Light Commercial vehicles - expand market share with competitive pricing
- Female driver segment - targeted marketing campaigns
- Specific low-risk vehicle makes - competitive pricing strategies

5.2 High-Risk Segments Requiring Action

- Gauteng, KwaZulu-Natal, Western Cape - premium adjustments needed
- Heavy Commercial vehicles - implement risk-based pricing
- AUDI and HYUNDAI makes - underwriting review required

6. Next Steps and Focus Areas

Immediate Priorities: (1) **A/B Hypothesis Testing:** Validate risk differences across provinces, zipcodes, and gender using statistical tests. (2) **Statistical Modeling:** Develop linear regression per zipcode to predict total claims. (3) **ML Pipeline:** Build predictive model for optimal premium pricing incorporating car features, owner demographics, location, and other risk factors. (4) **Feature Engineering:** Create derived features and assess importance. (5) **Model Evaluation:** Assess performance and ensure interpretability for regulatory compliance.

6.1 Hypothesis Testing Priorities

- H■: No risk differences across provinces → Preliminary evidence suggests rejection
- H■: No risk differences between zipcodes → Requires statistical validation
- H■: No margin differences between zip codes → Critical for pricing strategy
- H■: No risk difference between Women and Men → Evidence suggests difference (82.19% vs 88.39%)

7. Conclusion

The EDA has revealed significant risk segmentation opportunities with 4.3x variation across provinces and 12x variation across vehicle types. Data quality analysis identified manageable missing value patterns and outliers requiring attention. The DVC infrastructure ensures reproducible, auditable analytics meeting regulatory requirements. Low-risk segments (Northern Cape, Bus vehicles, Female drivers) present immediate opportunities for premium optimization and market expansion. The project is on track to deliver actionable recommendations for risk-based pricing and targeted marketing strategies that will improve ACIS profitability through data-driven risk assessment.

Detailed visualizations and analysis outputs available in outputs/figures/ directory. All code and analysis notebooks are version-controlled in the project repository.