

Insurance Portfolio Analysis Report

Prepared for: AlphaCare Insurance Solutions (ACIS)

Prepared By: Mariam Gustavo

Date: December 9, 2025

Subject: Analysis of Insurance Portfolio Data for Risk, Profitability & Pricing Insights

1. Executive Summary

This project aims to optimize insurance premium pricing through exploratory data analysis (EDA), hypothesis testing, and predictive modelling. EDA revealed key trends, such as Gauteng province's high loss ratio and the influence of vehicle make and model on claims, providing descriptive insights into risk and profitability. Hypothesis testing validated significant risk differences across provinces but found no meaningful impact from postal codes or gender, offering diagnostic insights to refine pricing strategies. Predictive modelling focused on estimating OptimalPremium using regression models, with Linear Regression performing marginally better ($R^2 \approx 0.00007$, $MSE \approx 4.57M$) but lacking strong predictive power. SHAP analysis highlighted RegistrationYear, kilowatts, and SubCrestaZone as key drivers of premium adjustments, suggesting discounts for newer, low-power vehicles and surcharges for older, high-power ones, while emphasizing detailed geographic zones for pricing. Despite limitations in predictive accuracy, the project provides actionable recommendations for pricing and marketing strategies, with future work focusing on richer features, improved encoding, and advanced modelling techniques.

2. EDA Methodology (Condensed)

This report summarizes an exploratory data analysis (EDA) of an insurance portfolio containing policy, vehicle, geographic, premium, and claim information. The goals were to:

- Understand the distribution and quality of key variables (`TotalPremium`, `TotalClaims`, `CustomValueEstimate`, `SumInsured`).
- Quantify risk and profitability via **Loss Ratio** (`TotalClaims / TotalPremium`).
- Identify high-risk segments by province, vehicle type, make, gender, and postal code.

- Establish a reproducible data pipeline using **DVC (Data Version Control)**.

2.1 Step 1 — Data Summary

Objectives

- Inspect structure (`head`, `info`, `dtypes`)
- Compute descriptive statistics for `TotalPremium`, `TotalClaims`, `CustomValueEstimate`, `SumInsured`.

Findings

- Numeric columns are generally typed correctly; dates required conversion.
- All four key monetary variables show **high variance, strong right skew, and extreme outliers**.
-

Figure- Descriptive Statistics Table

Title: Descriptive Statistics for Key Numeric Columns

X-axis: Key variables

Y-axis: Statistical values (table: mean, std, quartiles, max)

	TotalPremium	TotalClaims	CustomValueEstimate	SumInsured
count	1.000098e+06	1.000098e+06	2.204560e+05	1.000098e+06
mean	6.190550e+01	6.486119e+01	2.255311e+05	6.041727e+05
std	2.302845e+02	2.384075e+03	5.645157e+05	1.508332e+06
min	-7.825768e+02	-1.200241e+04	2.000000e+04	1.000000e-02
25%	0.000000e+00	0.000000e+00	1.350000e+05	5.000000e+03
50%	2.178333e+00	0.000000e+00	2.200000e+05	7.500000e+03
75%	2.192982e+01	0.000000e+00	2.800000e+05	2.500000e+05
max	6.528260e+04	3.930921e+05	2.655000e+07	1.263620e+07

Emphasize skew, long right tails, and concentration of volume in a small subset of policies.

2.2 Step 2 — Data Quality

Objectives

- Measure missingness and duplicates.
- Visualize outliers for `TotalPremium`, `TotalClaims`, `CustomValueEstimate`.

Findings

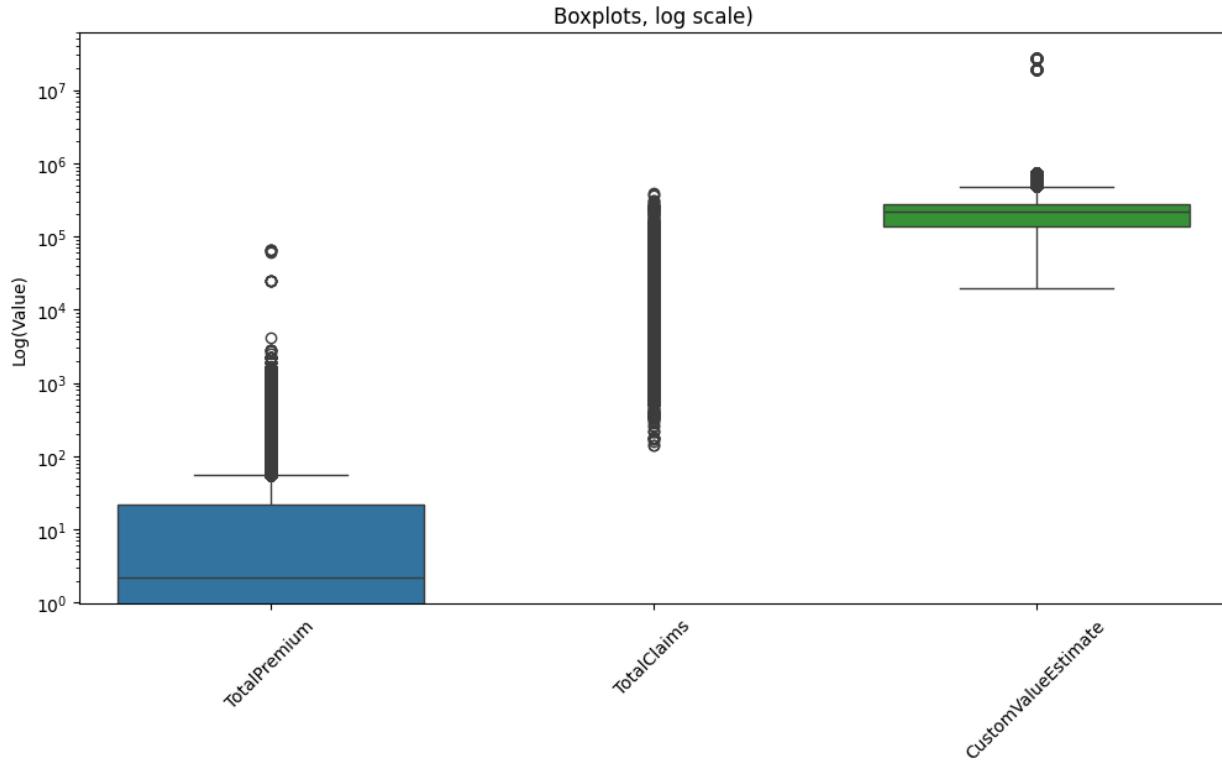
- `CustomValueEstimate` has ~77% missing values; several other features (e.g., `WrittenOff`, `Rebuilt`, `Converted`, `CrossBorder`, `NumberOfSupplements`) are also sparse.
- `NumberOfVehiclesInFleet` is effectively unusable (near 100% missing) and can be dropped.
- Raw boxplots are dominated by zeros; a **log scale** helps reveal distribution and outliers.

Figure-Boxplots (Log Scale)

Title: Boxplots of `TotalPremium`, `TotalClaims`, `CustomValueEstimate` (Log Scale)

X-axis: Variable

Y-axis: Log(Value)



This plot shows that most observations are low-value, with clear high-value outliers.

2.3 Step 3 — Univariate Analysis

Objectives

- Profile distributions of `TotalPremium` and `TotalClaims`.
- Understand portfolio composition by `Province`, `make`, `Gender`, `VehicleType`.

Findings

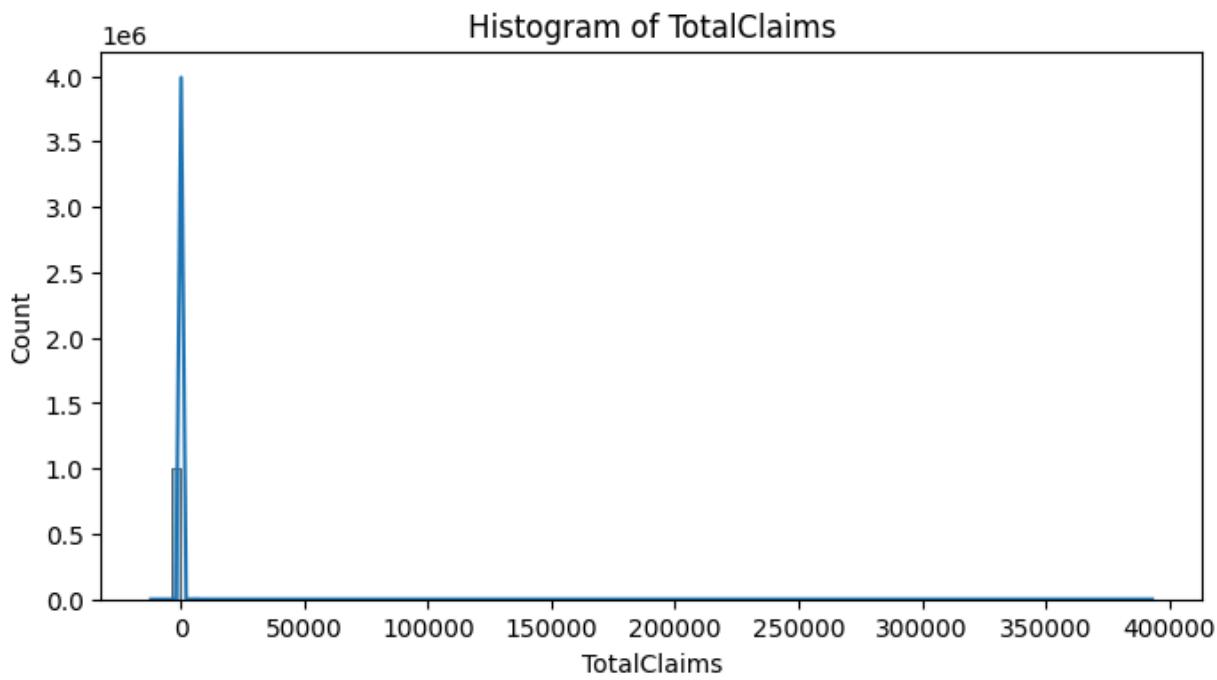
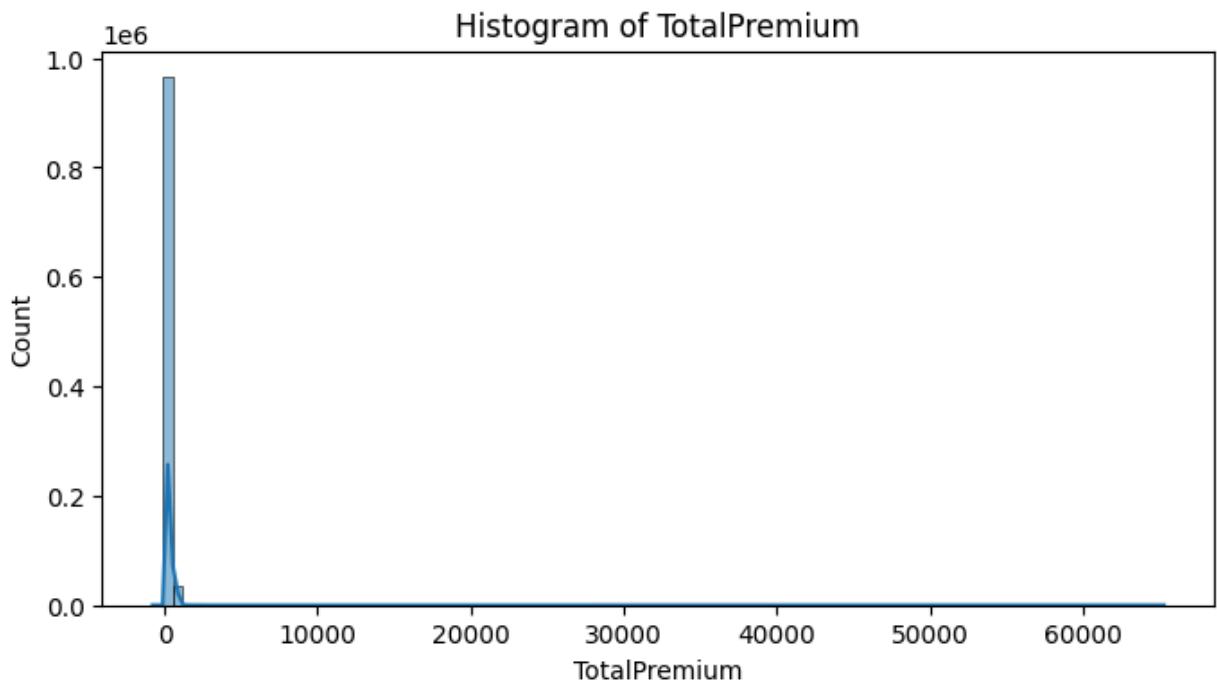
- Premiums and claims are **heavily right-skewed** with many zeros.
- Gauteng dominates policy counts, followed by Western Cape and KwaZulu-Natal.
- Toyota is the most common make; passenger vehicles are the dominant vehicle type.
- Gender is often unspecified; among specified records, male policies are more common.

Figure- Histograms

Title: Distribution of TotalPremium and TotalClaims

X-axis: TotalPremium Amount, TotalClaims

Y-axis: Policy count



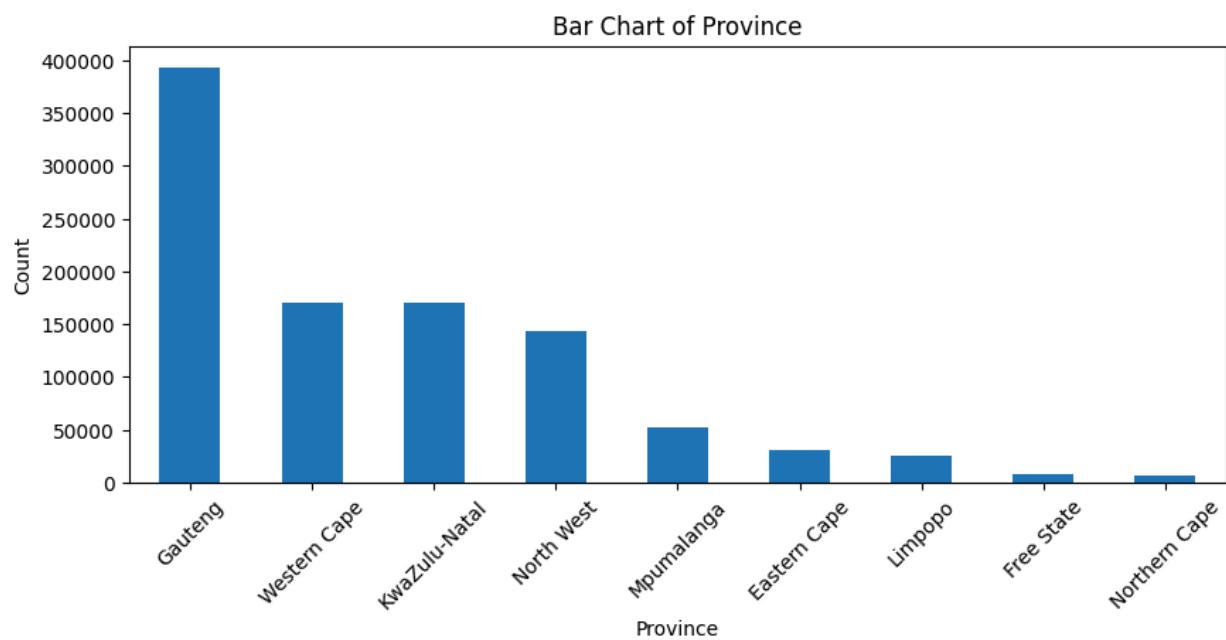
This plot shows zero-inflation and implications for modeling (e.g., two-part or zero-inflated models).

Figure- Categorical Bar Charts

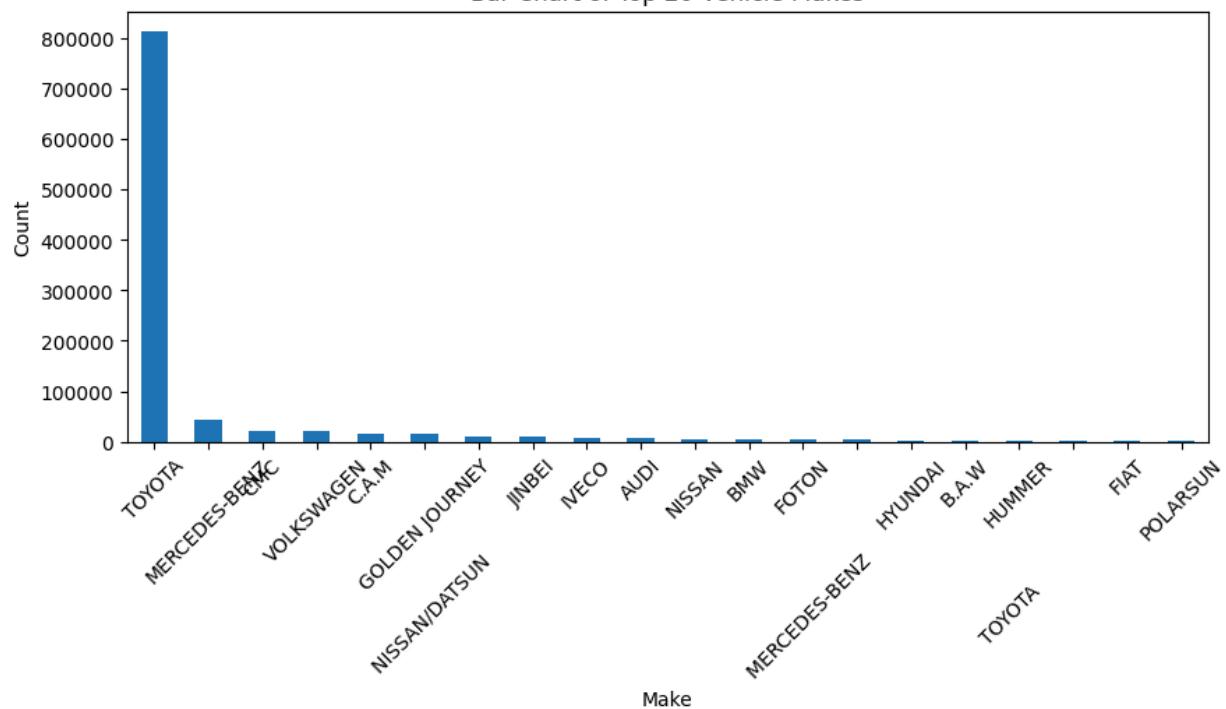
Title: Policy Counts by Province, Make, Gender, Vehicle Type

X-axis: Category

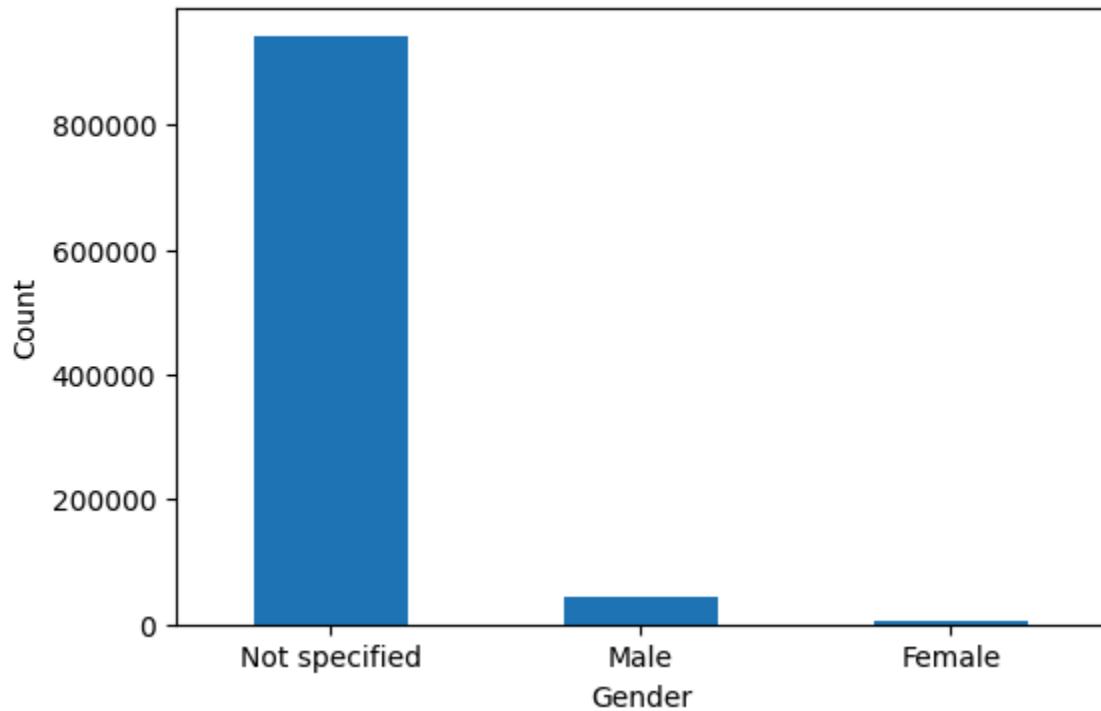
Y-axis: Count



Bar Chart of Top 20 Vehicle Makes



Bar Chart of Gender



2.4 Step 4 — Bivariate & Multivariate Analysis

Loss Ratio by Segment

- Defined `LossRatio = TotalClaims / TotalPremium` (with zero-premium cases handled as missing).
- Computed mean loss ratio by **Province**, **VehicleType**, **make**, and **Gender**.

Key Patterns

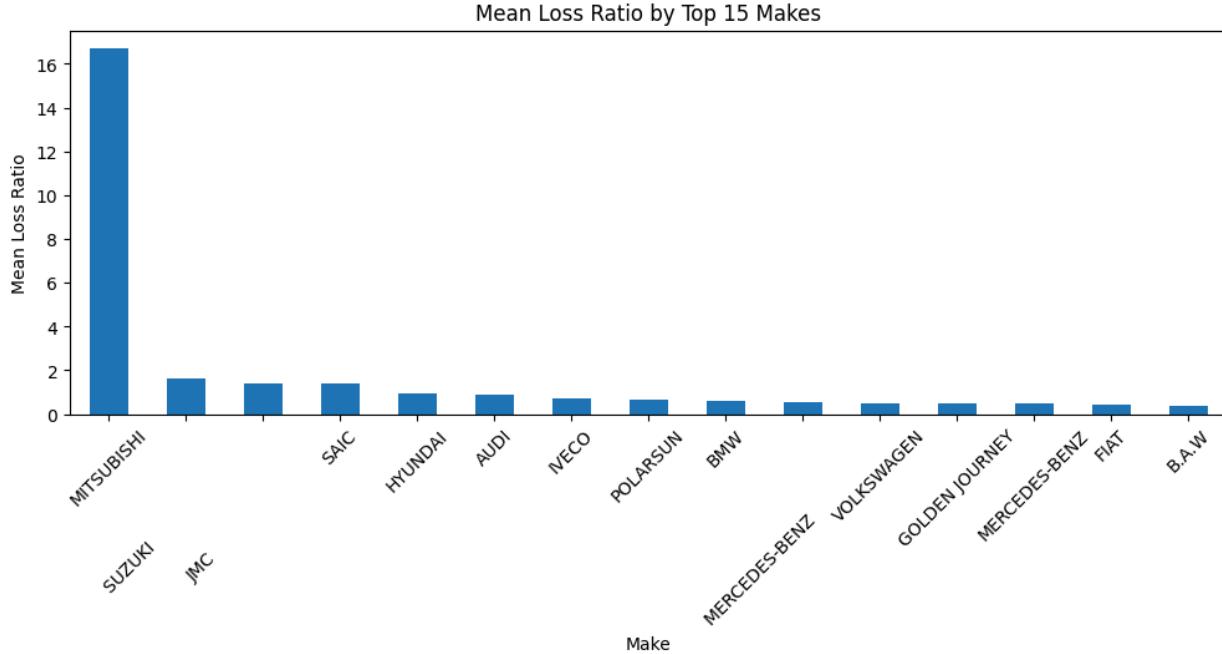
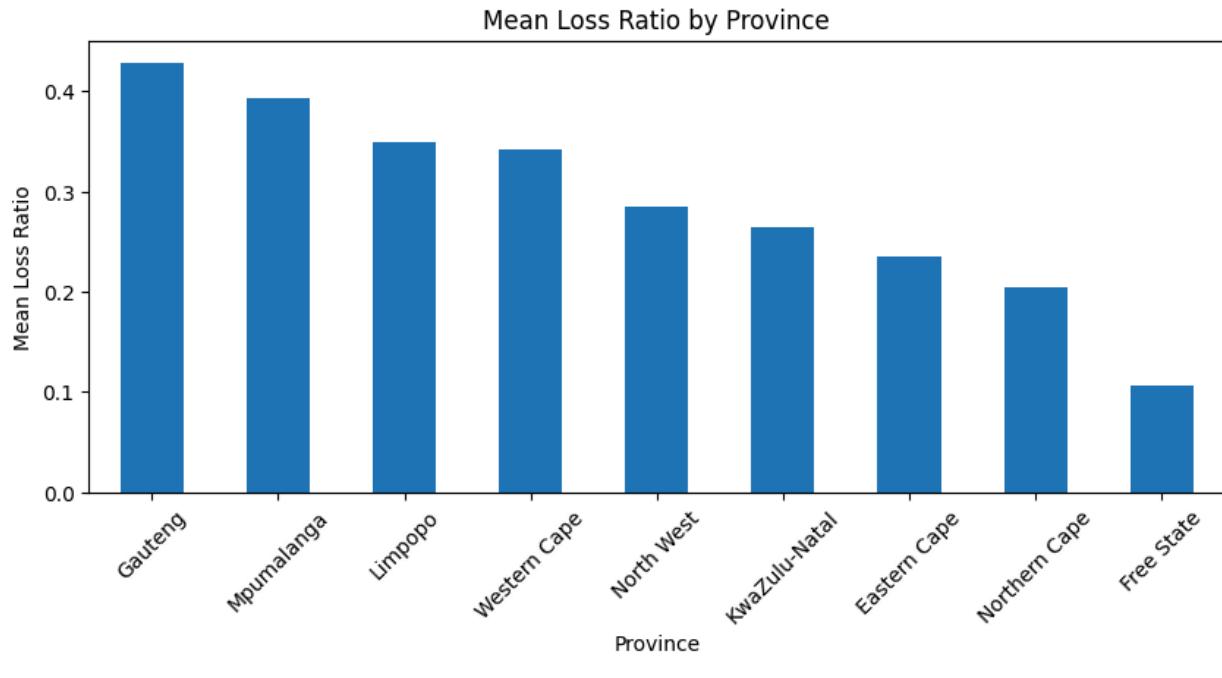
- Provinces: Gauteng, Mpumalanga, and Limpopo have **higher average loss ratios**; Free State shows lower ratios.
- Vehicle types: **Heavy commercial** and **light commercial** vehicles have higher loss ratios than passenger vehicles.
- Makes: Mitsubishi, Suzuki, and similar brands stand out with higher loss ratios; some makes like B.A.W are lower.
- Gender: Among specified records, females show a slightly higher mean loss ratio than males.

Figure- Ratio by Segment

Title: Mean Loss Ratio by Province, Vehicle Type, Make, Gender

X-axis: Segment

Y-axis: Mean Loss Ratio



Correlation Analysis

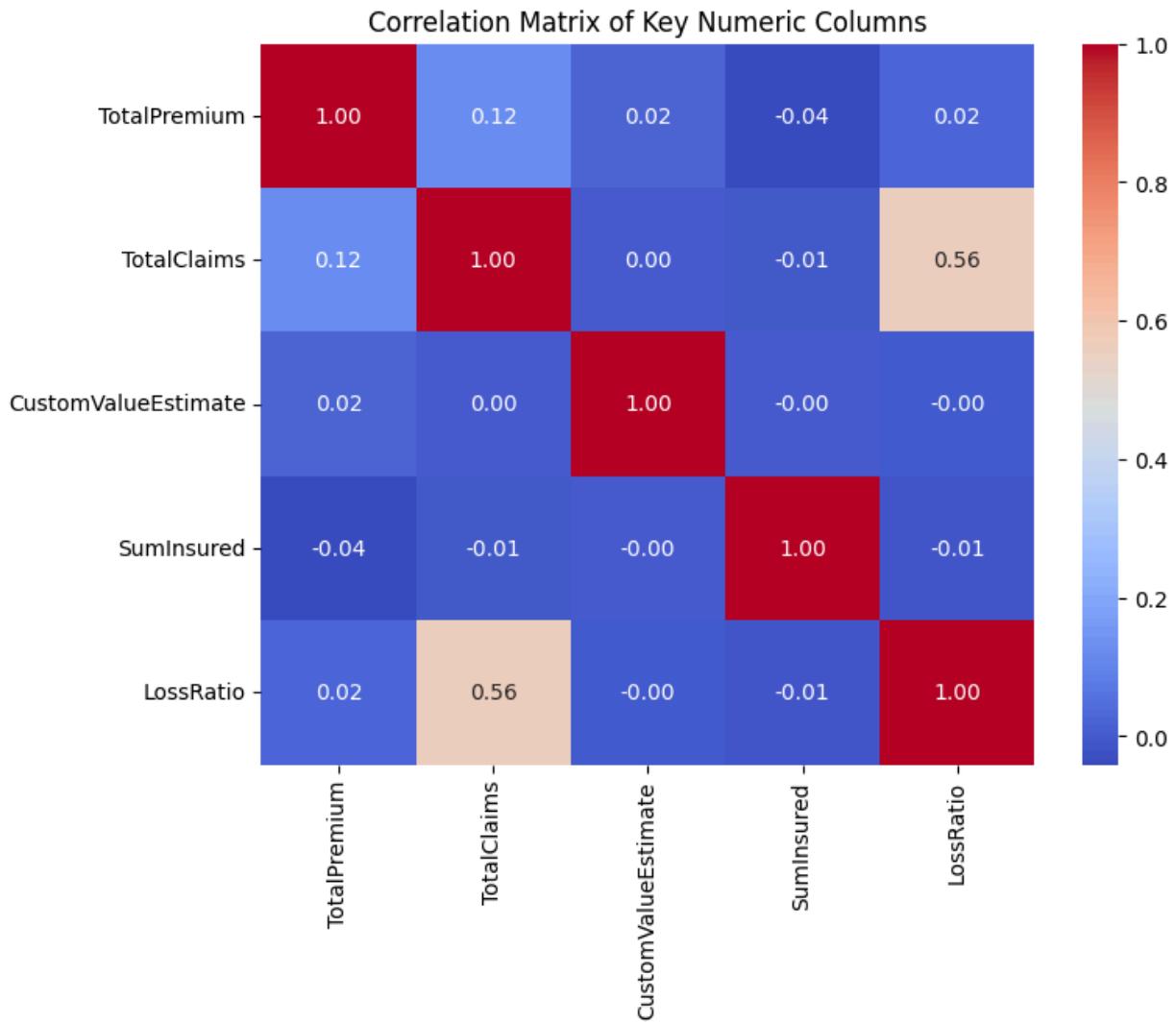
- Examined correlations among `TotalPremium`, `TotalClaims`, `CustomValueEstimate`, `SumInsured`, and `LossRatio`.
- Premiums and claims are positively correlated, with expected links between loss ratio and both components.

- Not very strong correlations

Figure- Correlation Heatmap

Title: Correlation Matrix of Key Numeric Variables

X-/Y-axis: Variables



Multivariate: Monthly x Postal Code

- Aggregated monthly **TotalPremium** and **TotalClaims** by **PostalCode**.
- Focused on top 10 postal codes by premiums and claims (e.g., 2000, 122, 8000).

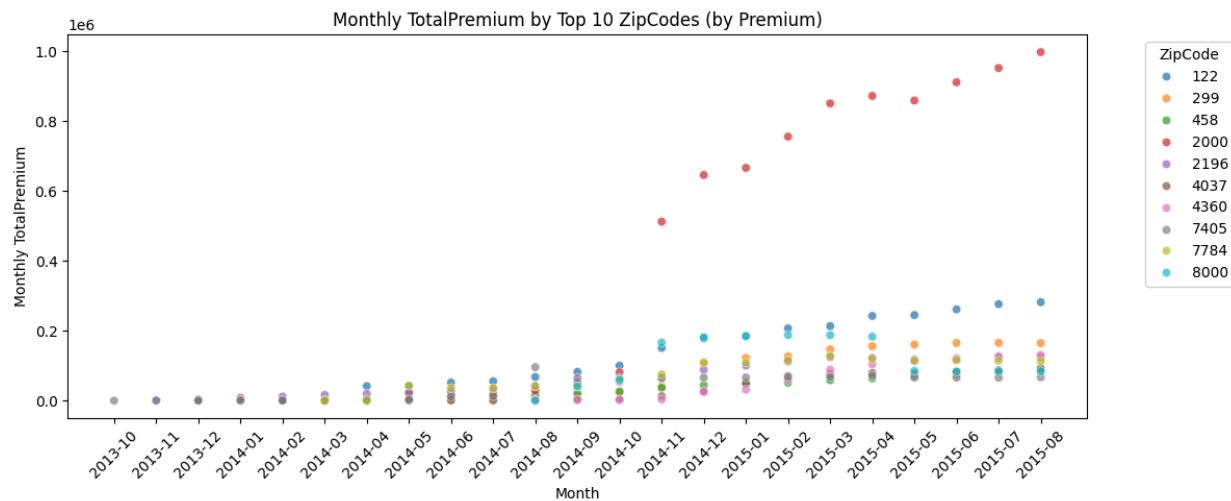
Figure- Monthly Premiums by Postal Code

Title: Monthly **TotalPremium** by Top Postal Codes

X-axis: Month

Y-axis: Monthly total premiums

Color: Postal code



This plot emphasizes postal code 2000 as a high-volume, high-exposure region.

2.5 Step 5 — Trend Analysis

Objectives

- Track monthly **TotalPremium**, **TotalClaims**, and **LossRatio**.

Findings

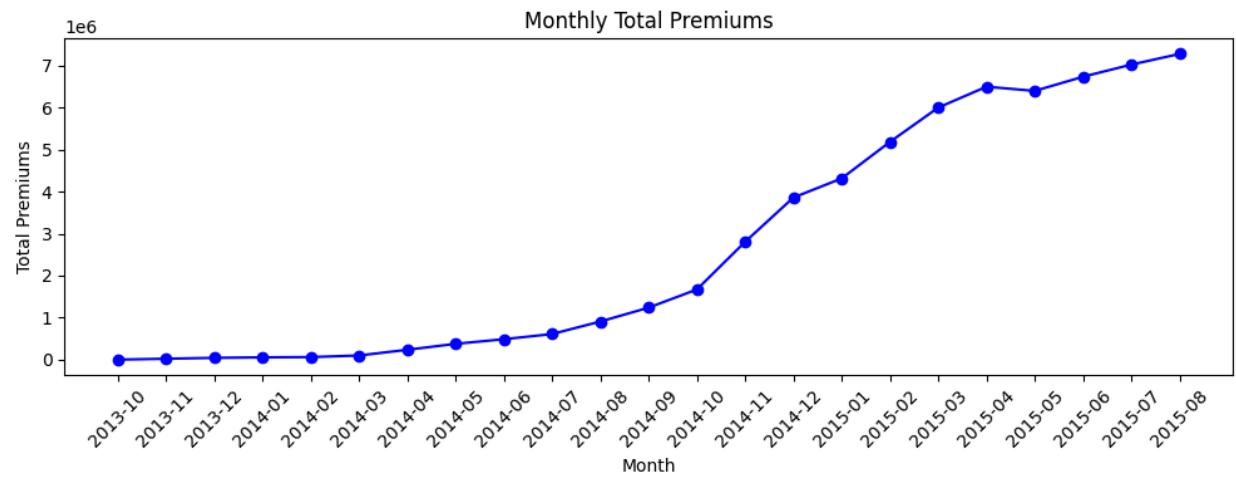
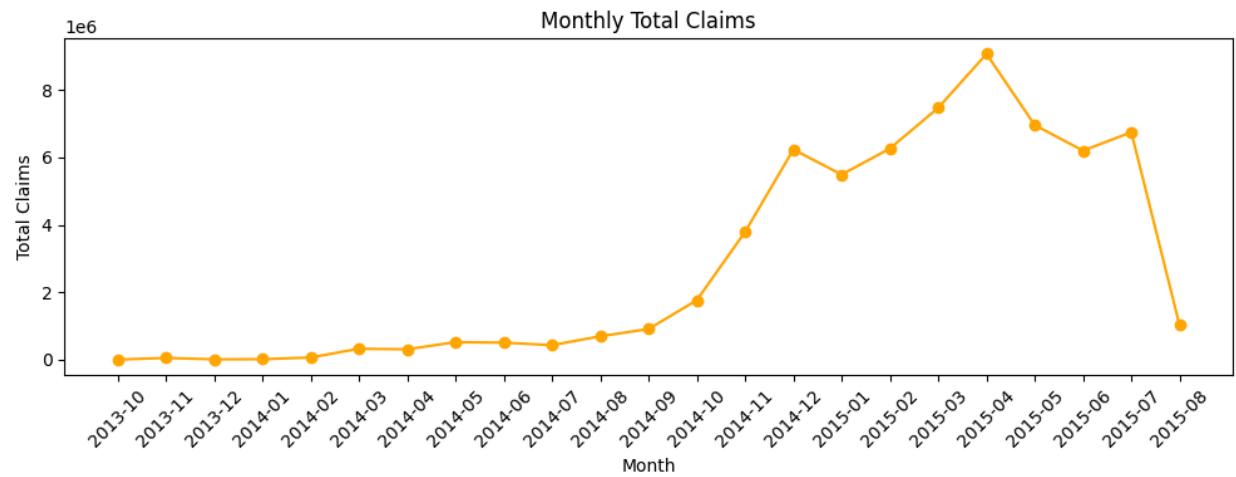
- Claims rise over time, peaking around mid-2015 before easing.
- Premiums grow steadily and strongly, indicating portfolio expansion.
- Loss ratio is volatile early (low volume + large claims) but stabilizes and **gradually improves** as the book grows.

Figure- Monthly Trends

Title: Monthly **TotalClaims**, **TotalPremium**, and **LossRatio**

X-axis: Month

Y-axis: Aggregate value





This plot shows an improving loss ratio to potential gains in profitability and scale.

2.6 Step 6 — Geography & Car Insights

Objectives

- Identify high-claim makes/models and high-premium/high-claim postal codes.

Findings

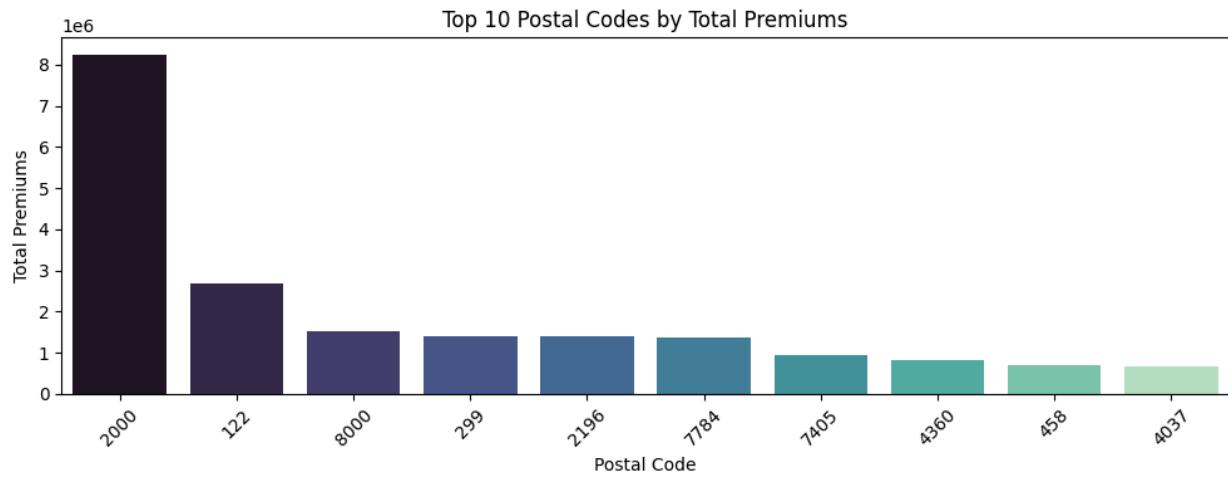
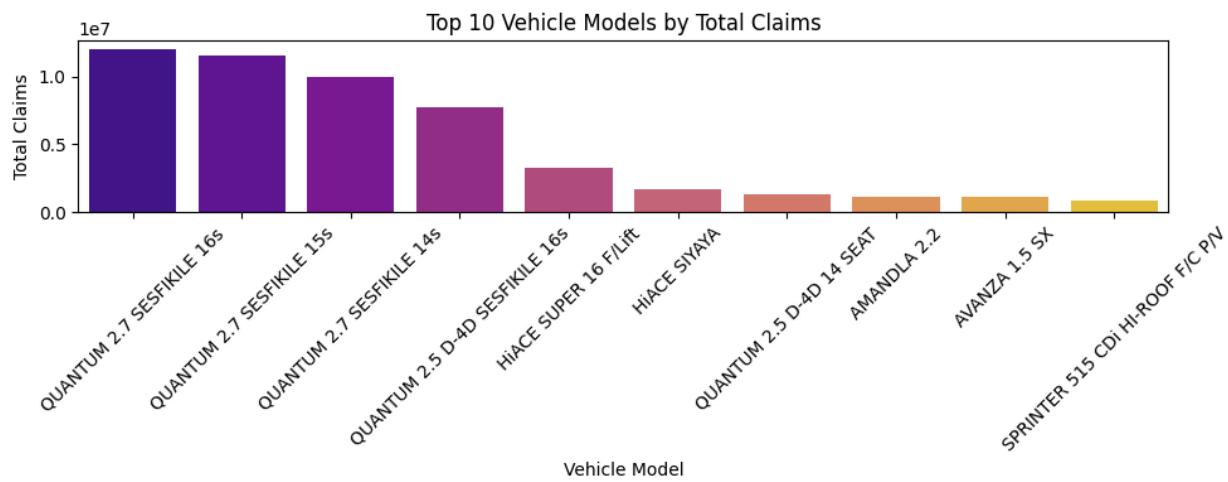
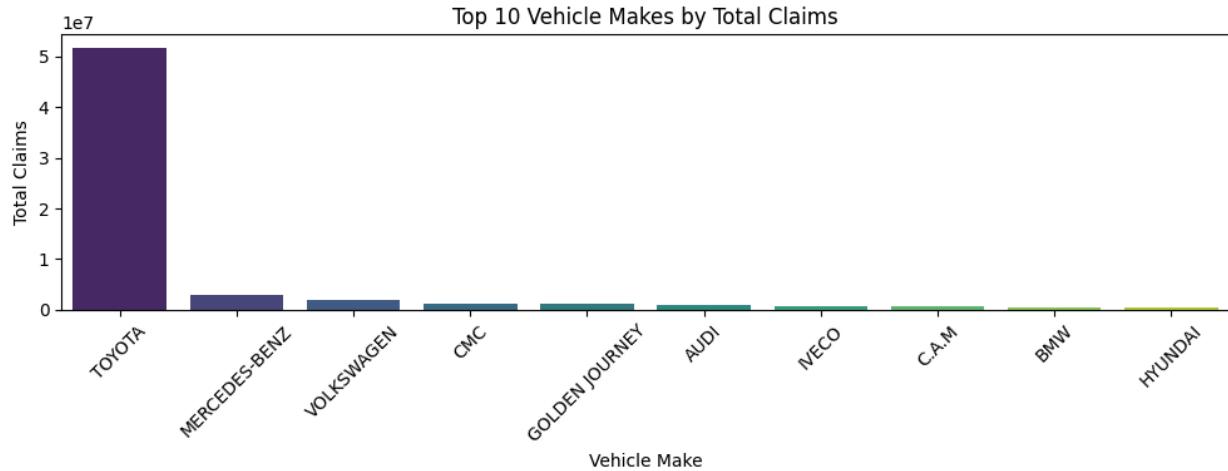
- Toyota and particularly **Toyota Quantum** models feature prominently in total claims.
- Postal codes 2000, 122, 8000 and others are major contributors to premiums and claims.

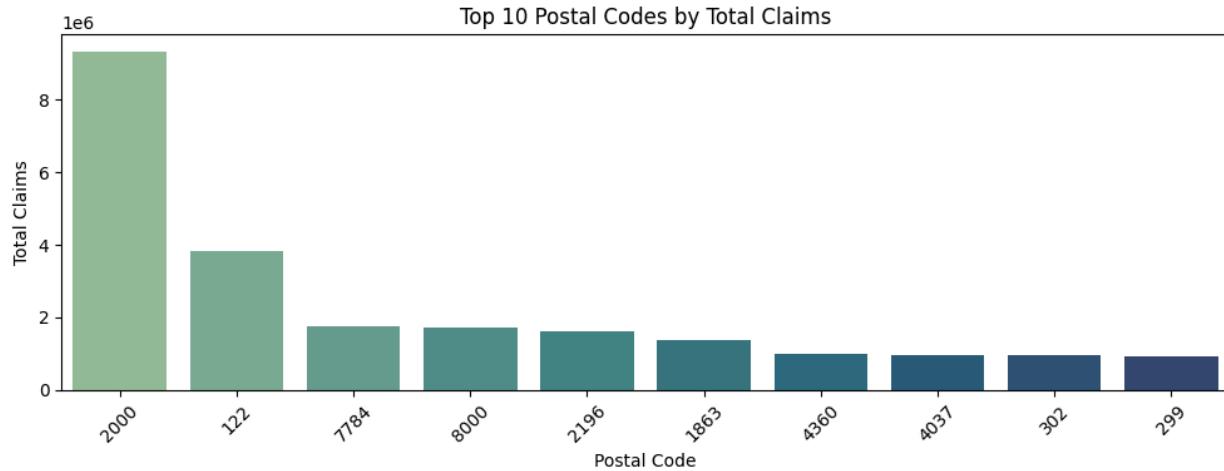
Figure- Vehicle & Postal Code Insights

Title: Top 10 Vehicle Makes/Models and Postal Codes by Claims/Premiums

X-axis: Category (make, model, postal code)

Y-axis: TotalClaims or TotalPremium





3. DVC Setup Methodology (High Level)

To keep analysis reproducible:

- A DVC stage `initial_process_csv` is defined in `dvc.yaml`:
 1. `cmd: python scripts/initial_preprocess.py`
 2. `deps: data/raw/MachineLearningRating_v3.txt, scripts/initial_preprocess.py`
 3. `outs: data/processed/MachineLearningRating_v3.csv`
- Raw and processed data are tracked with DVC; only small `.dvc` files and pipeline metadata live in Git.
- A local DVC remote (e.g., `C:/Users/yeget/LocalStorage`) is configured for storing versioned data.
- Any collaborator can:
 1. Clone the repo and run `dvc pull` to fetch data.
 2. Run `dvc repro` to regenerate the processed CSV from raw inputs.

This ensures that all EDA and modeling work can be rerun consistently, with clear auditability of data and code versions.

4. Hypothesis Testing

Objective:

The hypothesis testing aims to evaluate whether specific factors (e.g., province, postal code, gender) significantly influence risk (TotalClaims) or profitability (Profit). This helps identify key drivers of risk and profitability for the insurance business.

Steps:

Define Null and Alternate Hypotheses:

Null Hypothesis: Assumes no significant difference across groups (e.g., provinces, postal codes, genders).

Alternate Hypothesis: Assumes significant differences exist.

Select KPIs:

Risk is measured using TotalClaims.

Profitability is measured using Profit (calculated as TotalPremium - TotalClaims).

Run Statistical Tests:

ANOVA Test: Used for comparing means across multiple groups (e.g., provinces, postal codes).

A/B Test: Used for comparing means between two groups (e.g., male vs. female).

Interpret Results:

Accept or reject the null hypothesis based on the p-value.

Use insights to validate or challenge observations from exploratory data analysis (EDA).

5. Statistical Modelling

Methodology for Modelling

Objective:

The modelling process aims to predict OptimalPremium, the ideal premium price for each customer, ensuring profitability while maintaining competitive pricing.

Steps:

Feature Engineering:

Create OptimalPremium as the target variable based on TotalPremium and TotalClaims.

Model Training:

Train four regression models: Linear Regression, Random Forest, Gradient Boosting, and Decision Tree.

Evaluate models using R^2 (explained variance) and MSE (mean squared error).

Comparison of Models:

Linear Regression:

MSE $\approx 4.57M$, $R^2 \approx 0.00007$

Captures almost no meaningful structure in the data but performs marginally better than other models.

Gradient Boosting Regressor:

MSE $\approx 4.58M$, $R^2 \approx -0.00004$

Performs worse than Linear Regression, indicating no additional predictive power from non-linear trees.

Random Forest Regressor:

MSE $\approx 5.02M$, $R^2 \approx -0.098$

Overfits noise in the training data and fails to generalize.

Decision Tree Regressor:

MSE $\approx 5.08M$, $R^2 \approx -0.110$

The weakest model, with the largest error and most negative R^2 .

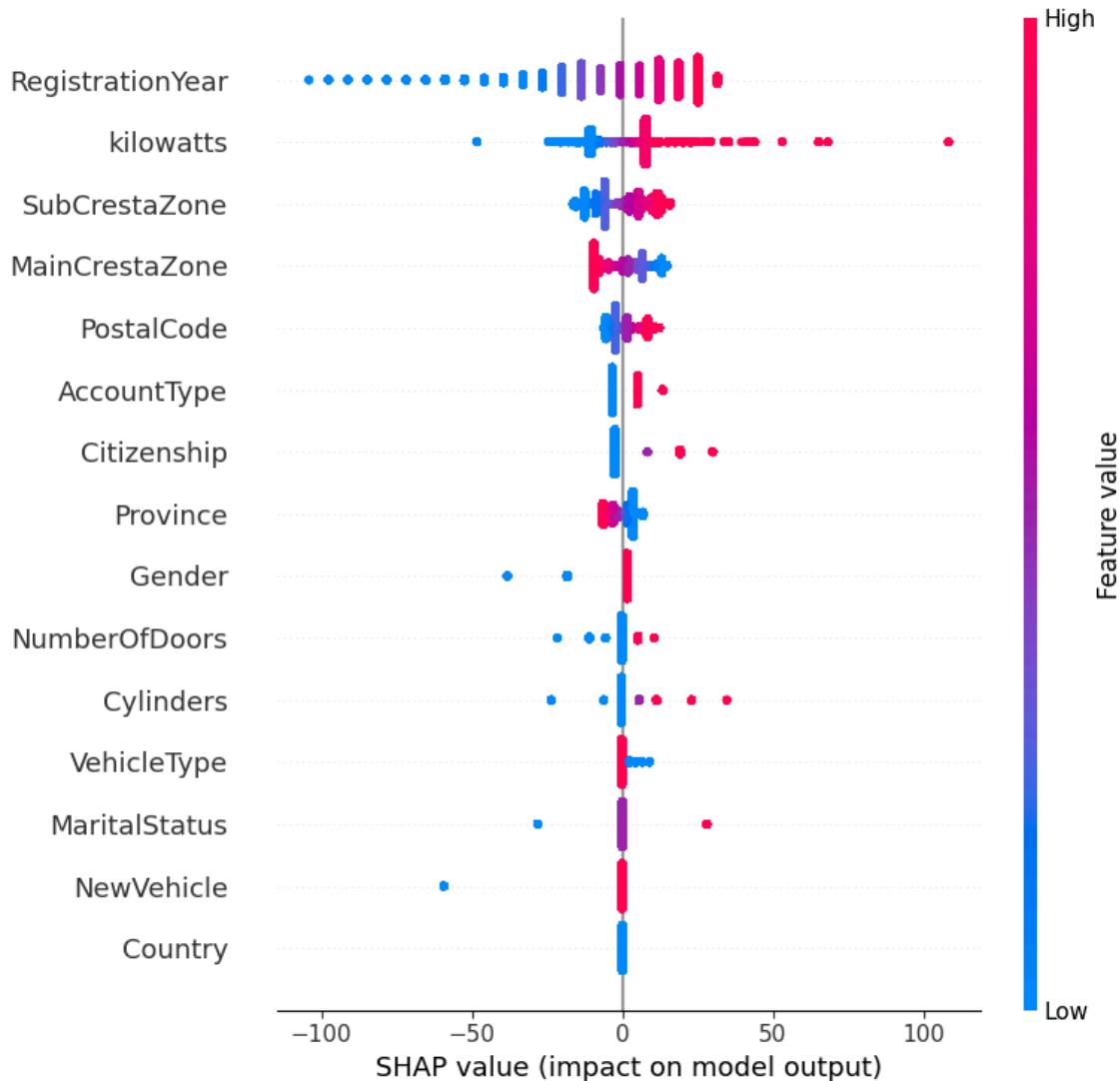
Conclusion:

Linear Regression is selected as the best model due to its lowest MSE and least negative R^2 . However, its predictions should be treated as approximate and used mainly for directional insights rather than production pricing.

SHAP Analysis:

Use SHAP to interpret feature importance and understand the drivers of OptimalPremium.

SHAP plot showing the importance of features



Plot Interpretation:

- Y-Axis (Features): Lists the features used in the model, such as RegistrationYear, kilowatts, MainCrestaZone, SubCrestaZone, Province, AccountType, PostalCode, Citizenship, Gender, NumberOfDoors, VehicleType, MaritalStatus, NewVehicle, Cylinders, and Country.
- X-Axis (SHAP Value): Represents the impact of each feature on the model's output. The values range from approximately -100 to 100.
 - * Positive SHAP Values (bars extending to the right): Indicate that higher values of the feature increase the model's prediction.
 - * Negative SHAP Values (bars extending to the left): Indicate that higher values of the feature decrease the model's prediction.
 - * Color Gradient: The color of the bars ranges from blue (low feature value) to red (high feature value), showing the magnitude of the feature's value.

Insights and Recommendations:

Translate SHAP results into actionable business strategies for pricing and marketing. Insights and recommendations are explicitly discussed in separate section below.

6. Business Impact - Insights and Marketing and Pricing Strategy recommendations

The analysis supports several business decisions:

1) EDA

Key insights:

- Monetary variables are **highly skewed and zero-inflated**, with many low/zero values and a few very large ones.
- **Gauteng**, heavy commercial vehicles, and certain makes (e.g., Mitsubishi, Suzuki) show **elevated loss ratios** compared with other segments.
- Postal code **2000** exhibits especially high premium and claim volumes, with several other postal codes (e.g., 122, 8000) also material.
- Over time, **premiums grow faster than claims**, leading to a gradually improving and more stable loss ratio.

Business Impact of EDA insights:

- **Risk-Based Pricing:**
Elevated loss ratios in specific provinces, vehicle types, makes, and postal codes provide a basis for refined rating factors and surcharges/discounts.
- **Portfolio Management:**
High-loss segments (e.g., heavy commercial vehicles, certain makes or postal codes) can be targeted with stricter underwriting rules, reinsurance strategies, or risk-management programs.

- **Profitable Growth:**
Improving loss ratios as the portfolio grows suggest that expansion within well-understood segments may be profitable, provided risk appetite is controlled.

2)DVC

- **Governance & Compliance:**
DVC-backed data lineage improves transparency, enabling regulators and stakeholders to trace how datasets used for pricing and capital modeling were produced.

3)Hypothesis Testing

Province:

Result: Risk differences across provinces are significant.

Impact: Provinces like Gauteng (high loss ratio) pose higher risks, confirming the need for geographic-based pricing adjustments.

Postal Code:

Result: No significant risk or profit differences across postal codes.

Impact: Postal codes do not independently influence risk or profitability, suggesting broader geographic factors (e.g., Cresta zones) may be more relevant.

Gender:

Result: Gender does not significantly influence risk.

Impact: Gender-based pricing adjustments are not supported by the data.

4) Modelling

SHAP Insights:

RegistrationYear:

Insight: Newer vehicles reduce optimal premiums, while older vehicles increase them.

Impact: Newer cars are less risky, likely due to better safety features and fewer mechanical failures.

kilowatts:

Insight: Higher kilowatt values increase optimal premiums, while lower kilowatt values reduce them.

Impact: High-power vehicles are riskier, likely due to higher accident severity or repair costs.

SubCrestaZone:

Insight: SubCrestaZone has moderate SHAP impact, suggesting local territory differences in risk.

Impact: Detailed geographic zones are more predictive than broader categories like province or postal code.

Recommendations:**Pricing Strategy:**

Adjust premiums based on vehicle age and engine power:

Discounts for newer, low-power vehicles.

Surcharges for older, high-power vehicles.

Use SubCrestaZone as the primary geographic rating factor.

Marketing Strategy:

Target customers with newer, low-risk vehicles for loyalty programs or discounts.

Position “safe driver” products for low-power vehicles.

7. Model Limitations and Next Steps

Limitations:

Low Predictive Power:

The best model explains almost none of the variation in OptimalPremium.

Limited Features:

Key insurance-related features like vehicle value, cover types, deductibles, prior claims, and policy tenure were not included.

Encoding Issues:

Label encoding for categorical variables may oversimplify relationships.

Future Work:

Expand Predictor Columns:

Include richer features such as CustomValueEstimate, SumInsured, cover types, deductibles, prior claims, policy tenure, vehicle usage, make/model, and driver age.

Improve Feature Encoding:

Use one-hot encoding or target encoding for categorical variables.

Alternative Targets:

Model pure premium (expected claim cost) or loss ratio instead of OptimalPremium.

Deeper SHAP Analysis:

Aggregate SHAP values by category (e.g., SubCrestaZone, Make, VehicleType) to create actionable pricing tables.

Model Refinement:

Experiment with non-linear models (e.g., XGBoost, LightGBM) and hyperparameter tuning.