Swedish Motor Insurance: SAS Case Study

Objective

The purpose of this case study is to explore and analyze patterns in motor insurance claims using the Swedish Motor Insurance dataset. The focus is on understanding risk indicators such as claim frequency and severity, identifying variations across risk zones, and evaluating the potential to predict claim costs using policy features.

This type of analysis closely mirrors the responsibilities found in product and pricing analyst roles within the insurance industry, particularly those involving SAS and predictive modeling.

Data Overview

The dataset contains 2,182 observations and 7 variables: - **Zone**: Risk classification of the geographical zone (1–7) - **Kilometers**: Annual driving distance band - **Bonus**: No-claim bonus level (1 = low bonus, 7 = high bonus) - **Make**: Car make (coded 1–9) - **Insured**: Number of policyholders - **Claims**: Number of claims - **Payment**: Total claim payment in Swedish Kronor

Supplemental Variables Created:

- **Frequency** = Claims / Insured
- **Severity** = Payment / Claims
- **AvgCostPerPolicy** = Payment / Insured

Analysis Summary

1. Data Cleaning & Variable Creation

- Verified dataset shape and structure
- Created core insurance KPIs (Frequency, Severity, AvgCostPerPolicy)
- Removed records with missing or zero `Claims`, `Payment`, or `Insured`

Reference Visuals:

- insurance_dataset_structure.png
- insurance_data_sample_preview.png
- insurance_frequency_severity_sample.png

2. Descriptive Analysis

Key Findings by Zone:

Zone	Avg Cost per Policy	Claim Frequency	Severity
1	512.10	0.10	4644.52

Zone	Avg Cost per Policy	Claim Frequency	Severity
7	199.74	0.05	1842.45

Observations:

- **Zone 1** has the highest average cost per policy and claim frequency — suggesting it is the riskiest zone.
- **Zone 7** has the lowest frequency and average cost per policy — potentially an area of lower risk and high profitability.

Recommendations:

- **Zone 1** should be considered for premium adjustment or targeted underwriting interventions.
- **Zone 7** could be a candidate for promotional pricing or customer acquisition efforts.

Reference Visuals:

- avg_cost_per_policy_by_zone.png
- claim_frequency_by_zone.png
- risk_metrics_by_zone_summary.png

3. Regression Modeling

A linear regression model was built to predict `Payment` using all independent variables. Key results: - **R-Square** = 0.9952 (very high explanatory power) - All predictors (Claims, Insured, Bonus, kilometers, Zone, Make) were statistically significant ($p < 0.05$)

Key Takeaways:

- **Claims** and **Insured** have the largest positive coefficients, aligning with expectations.
- **Zone** also significantly influences claim payouts, validating the earlier descriptive insights.

Reference Visuals:

- regression_model_table.png
- regression_model_diagnostics.png

Conclusions & Recommendations

1. **Zone-Based Pricing Strategy**:
    - Zone 1 poses the highest risk and should be subject to stricter underwriting and pricing scrutiny.
    - Zone 7 offers potential as a low-risk, cost-efficient market segment.
2. **Bonus Level Implications**:
    - Lower bonus levels correlate with higher average payments, reaffirming the logic of rewarding claim-free driving.
3. **Claims and Insured Volumes**:

    - These are the strongest predictors of total payout, which underscores their utility in pricing and forecasting.
4. **Kilometers Driven**:
    - More mileage may increase exposure, though its influence is smaller than claimed volume.

Limitations & Future Work

Although the model is statistically strong, the conclusions should be validated with additional data and business context.

Additional Data That Would Improve This Analysis:

- **Policy duration**: Helps normalize frequency/severity
- **Driver demographics**: Age, experience, past claims
- **Vehicle information**: Year, model, safety rating
- **Historical rate data**: Premium charged per policy
- **Claims type classification**: Collision, theft, glass, etc.
- **Temporal trends**: Over time analysis to spot cyclical risk

Next Steps:

- Incorporate time-series data if available to study trends
- Create a risk-adjusted profitability score by combining premium (if known) and loss ratio
- Visualize interaction effects (e.g., zone × kilometers) using interaction terms or grouped visualizations

---

This case study demonstrates a foundational insurance pricing and claims analysis, aligning directly with responsibilities in analyst roles that require SAS and a product/pricing mindset.