# Swedish Motor Insurance Case Study  
  
This case study explores insights from a Kaggle dataset on Swedish Motor Insurance, using SAS to perform exploratory data analysis and predictive modeling. It is designed to demonstrate practical SAS proficiency relevant to insurance-focused analyst roles.  
  
## 📁 Dataset Source  
  
Kaggle: [Swedish Motor Insurance Dataset](https://www.kaggle.com/datasets/floser/swedish-motor-insurance/data)  
  
The dataset contains 2,182 observations with 7 variables:  
- Zone: Geographical region (1–7)  
- Bonus: No-claims bonus index (1–7)  
- Make: Car model category (1–9)  
- Kilometres: Annual distance driven (1–5)  
- Insured: Policy exposure in thousands of Swedish Kronor  
- Claims: Number of claims  
- Payment: Total payment for all claims  
  
## 📌 Objectives  
  
- Identify risk patterns across geographical zones  
- Analyze average policy cost and claim frequency  
- Predict claim payment amounts using regression modeling  
  
## 📊 Methodology Overview  
  
### 1. Data Cleaning & Preview  
- File: `insurance\_data\_sample\_preview.png`  
- File: `insurance\_dataset\_structure.png`  
- Dropped 3 missing rows  
- Verified structure and completeness  
  
### 2. Derived Metrics  
- Severity = Payment / Claims  
- Frequency = Claims / Insured  
- AvgCostPerPolicy = Payment / Insured  
- File: `insurance\_frequency\_severity\_sample.png`  
  
### 3. Descriptive Statistics & Trends  
- Zone-level frequency and severity summaries  
- File: `risk\_metrics\_by\_zone\_summary.png`  
- File: `avg\_cost\_per\_policy\_by\_zone.png`  
- File: `claim\_frequency\_by\_zone.png`  
  
### 4. Predictive Modeling  
- Multiple linear regression using all variables to predict Payment  
- Diagnostic & residual analysis  
- File: `regression\_model\_diagnostics.png`  
- File: `regression\_model\_table.png`  
  
## ✅ Key Insights  
  
1. \*\*Zone 1\*\* presents the highest claim frequency and cost per policy, suggesting a higher risk region.  
2. \*\*Zone 7\*\* has the lowest cost per policy and claim activity, representing a potentially safer customer base.  
3. \*\*Bonus level\*\*, \*\*Kilometers driven\*\*, and \*\*Vehicle Make\*\* show strong correlation to claim amounts.  
4. The regression model explains 99.5% of variance (R² = 0.9952), but extreme outliers exist.  
  
## 💡 Recommendations  
  
- Reassess premium structures for high-risk zones (e.g., Zone 1).  
- Investigate drivers with high kilometer categories and low bonus levels.  
- Create targeted incentives to encourage safer driving and lower claim frequency.  
  
## 🔍 Further Data Needed  
  
To expand and validate this analysis, the following data would be valuable:  
- Driver age and gender  
- Policy duration and lapse behavior  
- Weather or seasonal incident context  
- Location-specific road safety ratings  
- Incident causes (e.g., theft, collision, weather-related)  
  
## 🧠 Skills Demonstrated  
  
- Data wrangling, aggregation, and transformation in SAS  
- Use of `PROC MEANS`, `PROC FREQ`, and `PROC REG`  
- Creating custom metrics for insurance KPIs  
- Diagnostic regression analysis  
- Visual insight communication  
  
---