Advanced Business Analytics

Group X6

Ky Anh Le 140044

Giang Nguyen 140043

Mengzhen Shang 140028

Liyuan Cao 140029

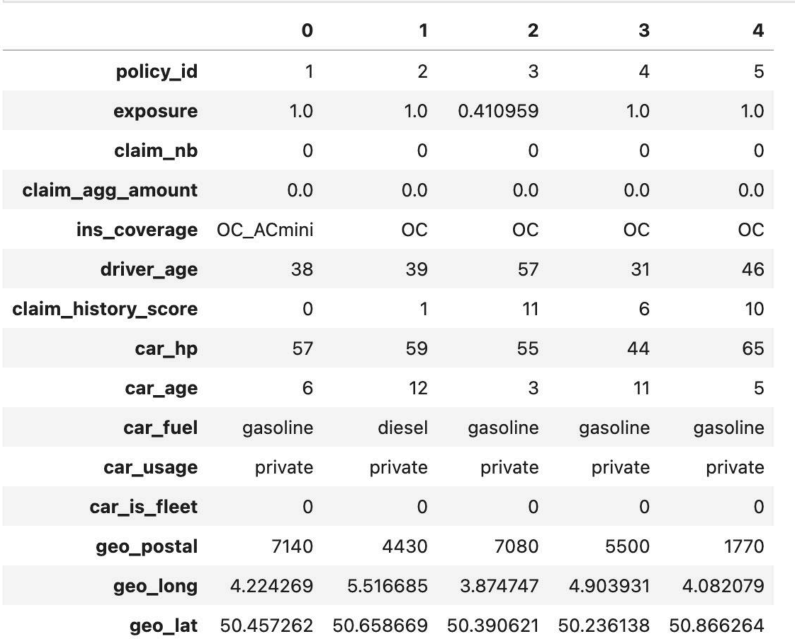
Advanced Business Analytics

Professor Adam Korczyński

Professor Łukasz Głąb

Part 1: Exploratory Data Analysis and Feature Engineering

1.1 Data overview



	0	1	2	3	4
policy_id	1	2	3	4	5
exposure	1.0	1.0	0.410959	1.0	1.0
claim_nb	0	0	0	0	0
claim_agg_amount	0.0	0.0	0.0	0.0	0.0
ins_coverage	OC_ACmini	OC	OC	OC	OC
driver_age	38	39	57	31	46
claim_history_score	0	1	11	6	10
car_hp	57	59	55	44	65
car_age	6	12	3	11	5
car_fuel	gasoline	diesel	gasoline	gasoline	gasoline
car_usage	private	private	private	private	private
car_is_fleet	0	0	0	0	0
geo_postal	7140	4430	7080	5500	1770
geo_long	4.224269	5.516685	3.874747	4.903931	4.082079
geo_lat	50.457262	50.658669	50.390621	50.236138	50.866264

Figure 1: Dataset overview

Data Structure Overview: The dataset has been successfully loaded using the semicolon (`;`) separator. From the initial inspection of the first 5 rows, we can observe:

- Granularity: Each row represents a single insurance policy (`policy_id`).
- Target Variables: `claim_nb` (frequency) and `claim_agg_amount` (severity) are present and appear to be zero-inflated (mostly 0s), which is typical for insurance data.
- Feature Types: The dataset contains a mix of numerical features (e.g., `driver_age`, `car_hp`, `exposure`) and categorical features (e.g., `ins_coverage`, `car_fuel`, `car_usage`) that will require preprocessing.
- Spatial Data: Geographic information is provided via postal codes (`geo_postal`) and coordinates (`geo_long`, `geo_lat`).

Dataset Dimensions: The dataset consists of **163,212 records** and **15 variables**.

- Sample Size: With over 160k observations, the dataset is sufficiently large to ensure statistical significance for our proposed segmentation (K-Means) and predictive modeling tasks.
- Dimensionality: The 15 initial features provide a manageable starting point, which we will expand through feature engineering to capture more complex risk patterns.

Data Integrity & Type Check: The `info()` output validates the technical quality of the dataset:

1. Completeness: There are 163,212 non-null entries for all 15 columns. This confirms that the dataset contains no missing values (NaNs), and thus, no imputation techniques (e.g., mean filling, KNN) are required at this stage.
2. Data Types:

- Numerical: 12 columns (float/int) are correctly identified, including the targets (`claim_nb`, `claim_agg_amount`) and key features (`driver_age`, `car_hp`).
- Categorical: 3 columns (`ins_coverage`, `car_fuel`, `car_usage`) are of object type and will require encoding (One-Hot or Label Encoding) for the K-Means model.

Descriptive Statistics Summary:
The summary statistics provide initial insights into the portfolio's risk profile and feature distributions:

- **Risk Exposure (`exposure`):** The average policy duration is **0.89 years**, with a maximum of 1 year and a minimum of 0.0027 years (~1 day). This indicates that most policies are standard annual contracts, although a small portion were started late or terminated early during the reporting period.
- **Claim Frequency (`claim_nb`):** The mean number of claims per policy is **0.124**, while the median and 75th percentile are **0**, indicating that most policies did not have any claims. This confirms the dataset is **highly zero-inflated**, typical for insurance portfolios with rare-event claims.
- **Claim Severity (`claim_agg_amount`):** The data is extremely right-skewed, with a mean of **162.15**, a median of **0**, and a maximum of **140,032**. The high standard deviation (~1375) and presence of extreme outliers highlight the potential influence of large losses on standard regression models.
- **Driver Demographics:** The average driver is **47 years old** (median 46), ranging from 18 to 95, representing a mature driving population. The `claim_history_score` has a mean of 3.27 and ranges from 0 to 22, showing variability in prior claim behavior.
- **Vehicle Characteristics:**
    - **Car Age (`car_age`):** The average vehicle age is **7.37 years**, with a maximum of 48 years, suggesting the presence of some vintage or classic cars.
    - **Horsepower (`car_hp`):** Ranging from 10 to 243, with a mean of 56, implying most vehicles are standard family cars while high-performance vehicles are rare.
    - **Fleet Status (`car_is_fleet`):** Only **3.17%** of vehicles are part of a fleet.
- **Geographic Information (`geo_postal`, `geo_long`, `geo_lat`):** Postal codes range from 1000 to 9990, with longitude and latitude consistent with the target region (mean longitude ≈ 4.41, mean latitude ≈ 50.76).
- **Overall Observations:** The portfolio is dominated by policies with **low exposure and rare claims**, typical for property/casualty insurance. Claim amounts are highly skewed, emphasizing the need for modeling techniques that account for **zero-inflation and heavy tails**, such as frequency-severity models or generalized linear models with appropriate distributions.

1.2 Data Quality Check (Missing Values)

Although the initial `df.info()` summary indicated that the dataset contains no null entries, we perform an **explicit missing value check** here to rigorously confirm data completeness. This step serves to formally address the project requirement regarding "Data Imputation Techniques" and verifies that no specific handling or filling of missing data is needed before proceeding.

We observed upon inspection, the dataset contains no missing values. Both explicit missing values (NaN) and implicit missing values (e.g., placeholders like "?" or empty strings) were checked, and none were found. Therefore, no data imputation techniques are required for this dataset.

1.3 Exploratory Data Analysis (EDA)

1.3.1 Univariate Analysis

We analyze the distribution of key variables to understand the portfolio composition.

- Target Variables: `claim_nb` (Frequency) and `claim_agg_amount` (Severity).
- Features: Driver age, Car age, Car HP, etc.

1.3.1.1 Categorical Variable Analysis

The frequency distribution of categorical variables highlights several important structural characteristics of the insurance portfolio. With respect to insurance coverage, the portfolio is dominated by policies offering only the mandatory MTPL coverage (OC), which account for approximately 58% of all contracts. Mini Casco coverage (OC_ACmini) represents around 28% of the portfolio, while Full Casco (OC_AC) constitutes the smallest segment at approximately 13.5%.

This relatively low penetration of comprehensive Casco insurance may reflect a customer base dominated by owners of older vehicles, for whom extended coverage is perceived as economically inefficient, or a generally price-sensitive portfolio. Regarding vehicle fuel type, the portfolio is largely composed of gasoline-powered vehicles (about 69%), with diesel vehicles accounting for the remaining 31%, a distribution that is consistent with typical passenger car fleets in the region. Finally, the analysis of vehicle usage and fleet status reveals a strong class imbalance: more than 95% of vehicles are used privately, and over 96% are non-fleet vehicles.

Although business-use and fleet vehicles represent only a small fraction of the portfolio, these categories are often associated with substantially different risk profiles, such as higher mileage and exposure, which may influence segmentation results. In particular, such minority groups may either form small, distinct clusters or be absorbed into larger clusters if not properly accounted for during model preprocessing and scaling.

1.3.1.2 Target Variable Analysis: Claim Frequency (claim_nb)

The histogram of claim counts highlights the fundamental nature of insurance risk within the portfolio.

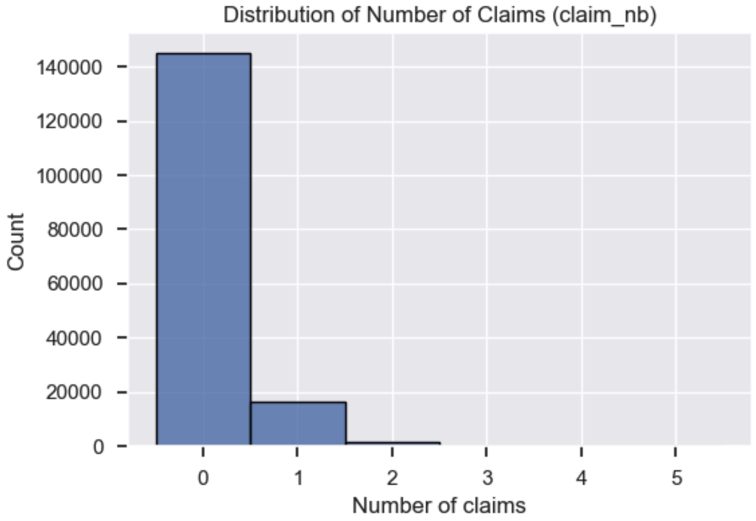


Figure 2: Distribution of Number of Claims (claim_nb)

Key Observations:

1. **Zero Inflation:** The distribution is highly skewed, with the vast majority of policies reporting zero claims. This confirms that claim occurrence is a rare-event phenomenon.
2. **Discrete Structure:** Claim counts take only non-negative integer values, with frequencies declining rapidly as the number of claims increases.

Implications for Modeling (Task 4):

- **Limitations of Linear Models:** The strong concentration of zeros violates the assumptions of standard linear regression models, potentially leading to biased estimates and invalid (negative) predictions.
- **Appropriate Modeling Framework:** The observed distribution motivates the use of count-data models, such as Poisson regression as a baseline, and more flexible alternatives (e.g., Negative Binomial or Zero-Inflated Poisson) to address over-dispersion and excess zeros.

1.3.1.3 Target Variable Analysis: Claim Severity (claim_agg_amount)

The distribution of total claim amounts complements the frequency analysis and illustrates the financial risk structure of the portfolio.

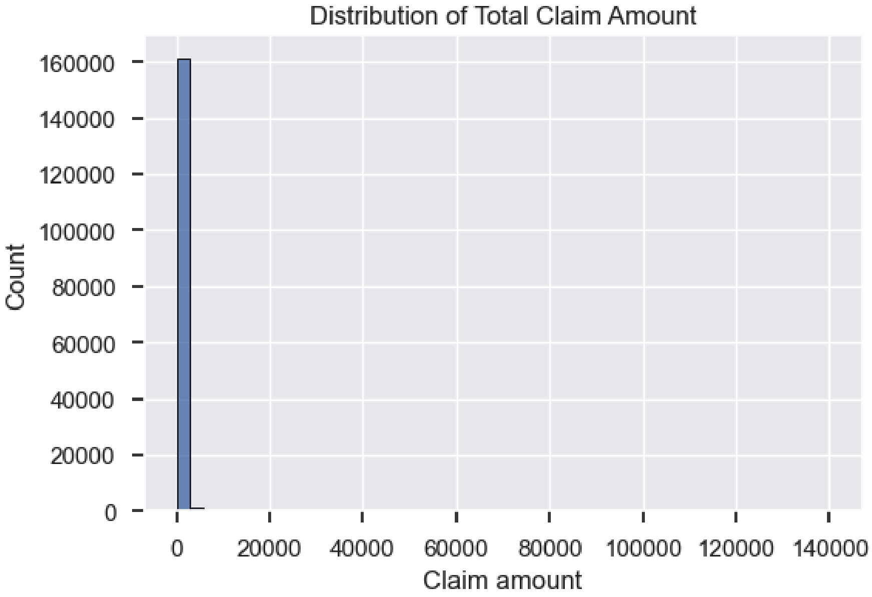


Figure 2: Distribution of Total Claim Amount

1. **Extreme Right Skewness:** Similar to claim frequency, this variable is dominated by zero values corresponding to policies with no reported claims.
2. **Heavy-Tailed Behavior:** While the mean claim amount is relatively low (~162), the maximum observed value reaches approximately 140,000. This indicates the presence of rare but extremely costly claims.
3. **Visualization Challenge:** Due to the overwhelming mass at zero, the tail of the distribution is barely visible on a standard scale. This pattern is characteristic of compound distributions such as Tweedie or Compound Poisson–Gamma models.

Business and Modeling Implications:

- **Tail-Driven Risk:** Financial risk in insurance portfolios is driven by extreme losses rather than average outcomes. Metrics based solely on mean claim amounts therefore provide a misleading representation of risk.
- **Modeling Considerations (Task 4):** Directly modeling `claim_agg_amount` using standard linear regression is ill-suited due to non-normality and sensitivity to outliers.
- **Recommended Strategy:** In subsequent modeling stages, it is preferable to either model claim frequency and claim severity separately or adopt generalized linear models with a Tweedie distribution, which naturally accommodates zero mass and heavy-tailed positive losses.

1.3.1.4 Outlier Analysis: Claim Severity Boxplot

The boxplot of `claim_agg_amount` provides a different perspective on the distribution's extreme skewness, highlighting the limitation of standard statistical visualization for this type of data.

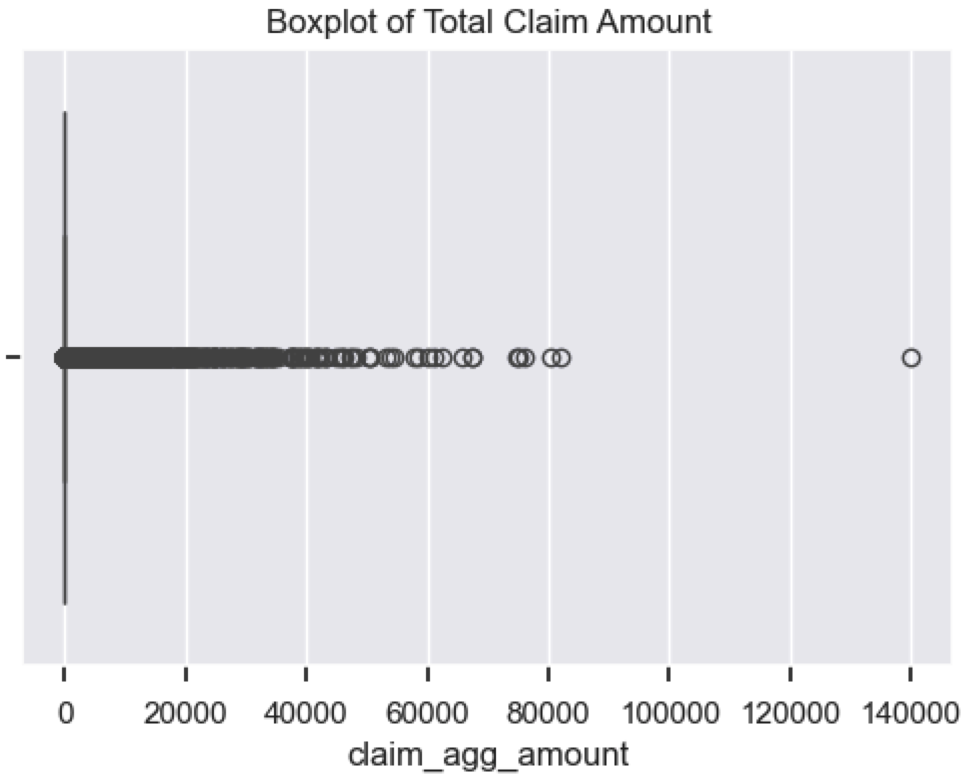


Figure 3: Boxplot of Total Claim Amount

Visual Interpretation:

- **Collapsed Box:** The "box" (representing the Interquartile Range, IQR) is essentially collapsed to a single line at 0. This is because the 75th percentile of the data is still **0** (i.e., fewer than 25% of policies have a claim).
- **"Wall" of Outliers:** The plot is dominated by a dense stream of data points extending to the right. In a standard normal distribution, outliers are rare dots beyond the whiskers. Here, **every single claim** is technically classified as an "outlier" because the baseline risk for the majority is zero.

Business Insight: This visualization confirms that "Average" behavior is meaningless for this portfolio. The portfolio dynamics are entirely driven by these "outliers."

- **Segmentation Strategy:** We cannot simply remove these outliers (e.g., using the 1.5*IQR rule) to clean the data. In insurance, these outliers **are** the business. The segmentation model must group customers to isolate the *probability* of becoming such an outlier, rather than treating them as noise.

1.3.1.5 Distribution of Key Numerical Risk Factors

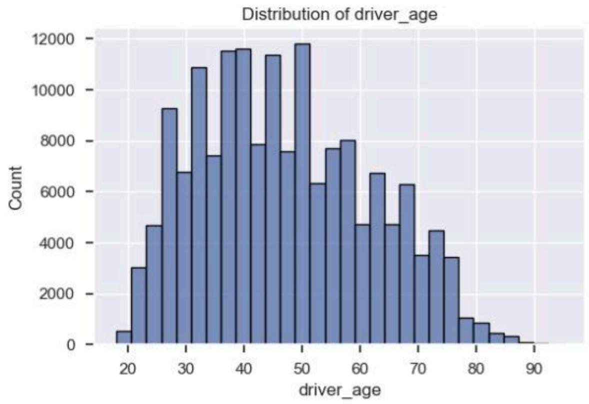


Figure 4: Distribution of driver_age

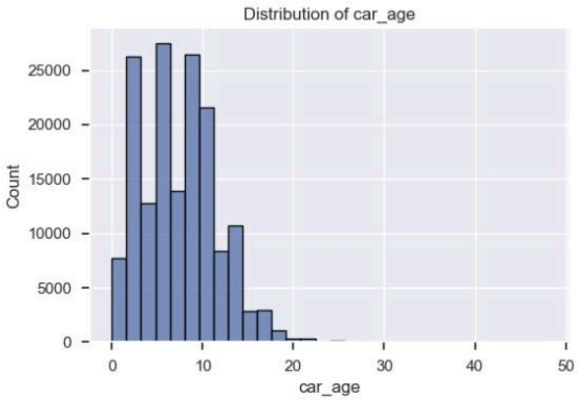


Figure 5: Distribution of car_age

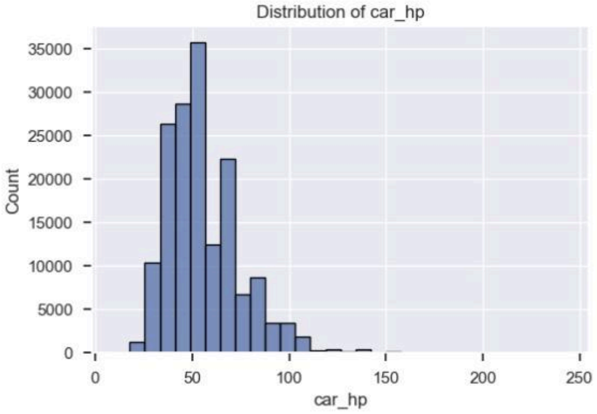


Figure 6: Distribution of car_hp

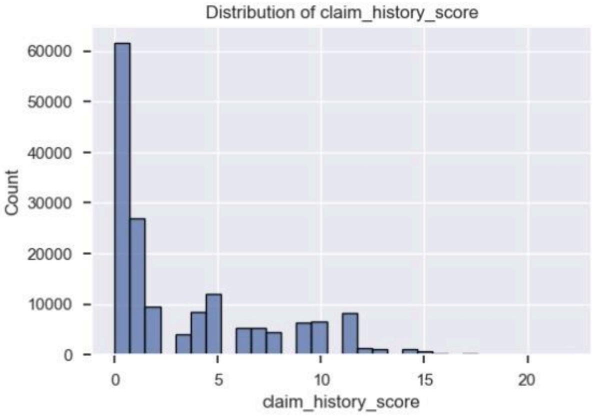


Figure 7: Distribution of claim_history_score

The distributions of selected numerical variables provide additional insight into the heterogeneity of driver and vehicle risk characteristics:

Driver Age (`driver_age`):
The age distribution is unimodal and concentrated between approximately 30 and 60 years. Very young and very old drivers represent a relatively small share of the portfolio. This suggests a mature and relatively stable customer base, with age remaining a relevant but not extreme risk differentiator.

Vehicle Age (`car_age`):
Car age exhibits a right-skewed distribution, with most vehicles being relatively new (under 10 years old), but a non-negligible tail of older vehicles. Older vehicles may be associated with higher mechanical failure risk and potentially higher claim frequency or severity.

Engine Power (`car_hp`):
The distribution of engine horsepower is approximately bell-shaped with a mild right tail. Most vehicles

fall within a moderate power range, while a small subset of high-powered cars may indicate more aggressive driving behavior and elevated risk exposure.

Claim History Score (`claim_history_score`):
This variable is highly right-skewed, with the majority of policyholders having low or zero scores, and a small group exhibiting very high values. This pattern highlights the strong risk segmentation potential of historical claim behavior, making this variable particularly informative for customer clustering and risk assessment.

1.3.2 Bivariate Analysis & Relationships

We examine relationships between key features and risk metrics (`claim_nb`, `claim_agg_amount`) to identify potential risk drivers.

1.3.2.1 Insurance Coverage vs Claim Frequency

The boxplot analysis compares claim counts across different insurance coverage types, revealing subtle but important differences in risk profiles.

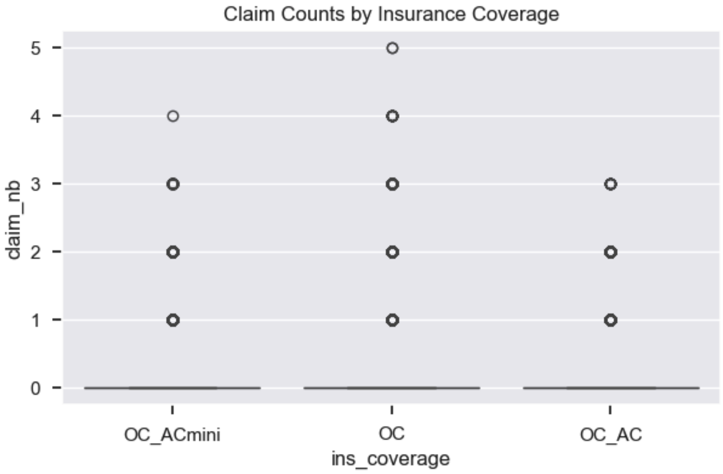


Figure 8: Claim Counts by Insurance Coverage

- **Median at Zero:** Across all coverage levels (`OC`, `OC_ACmini`, `OC_AC`), the median number of claims remains zero. This reconfirms the dominant zero-inflation pattern across the entire portfolio.
- **Tail Behavior:** However, policies with broader coverage (specifically `OC_AC`) exhibit a wider range of outliers in the upper quantiles, indicating a higher probability of multiple claims compared to basic coverage.

This pattern is likely driven by a combination of factors:

1. **Scope of Coverage:** `OC_AC` covers a wider array of perils (e.g., theft, own damage) than basic `OC`, naturally leading to more reportable events.
2. **Self-Selection (Adverse Selection):** Customers who perceive themselves as higher risk (or own more expensive vehicles) are more likely to opt for comprehensive coverage.
3. **Moral Hazard:** Policyholders with full coverage may be more inclined to file claims for minor damages that those with basic coverage would pay out-of-pocket or ignore.

Consequently, `ins_coverage` is not just a contractual label but a powerful proxy for customer risk behavior, making it a valid candidate for segmentation.

1.3.2.2 Fleet Status vs Claim Severity

The boxplot highlights a notable contrast in claim severity dynamics between fleet and non-fleet vehicles.

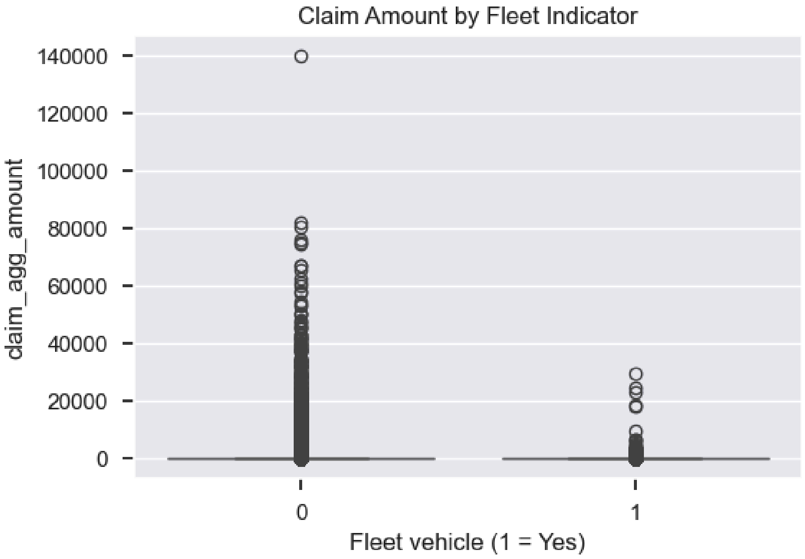


Figure 9: Claim Amount by Fleet Indicator

- **Outlier Dominance in Non-Fleets:** Non-fleet vehicles (`car_is_fleet = 0`) exhibit a much wider range of outliers, including the largest observed claim amounts in the dataset.
- **Controlled Severity in Fleets:** Fleet vehicles (`car_is_fleet = 1`) show a comparatively narrower distribution of positive claim amounts.

Business & Statistical Interpretation:

1. **Sample Size Effect:** It is important to note that non-fleet vehicles constitute ~97% of the portfolio. Statistically, extreme tail events (outliers) are naturally more likely to appear in the larger sub-population.
2. **Operational Factors:** Even accounting for sample size, the "narrower" severity distribution for fleets likely reflects professional fleet management, standardized repair contracts, or bulk insurance policies that cap per-claim payouts.
3. **Risk Profile:** Although fleet vehicles often have higher usage frequency (mileage), their per-accident cost appears more predictable.

Modeling Implication: Fleet status is a meaningful categorical variable. Its distinct distribution suggests it may interact differently with claim severity compared to private cars, confirming the need to include `car_is_fleet` as a predictor in the final GLM models.

1.3.2.3 Driver Age vs Claim History Score

The scatter plot reveals no strong linear relationship between driver age and claim history score, challenging the assumption that age alone is a direct predictor of history.

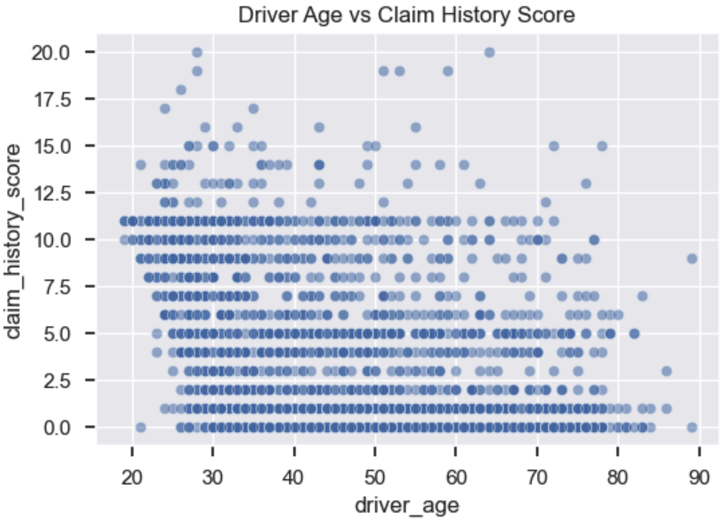


Figure 10: Driver Age vs Claim History Score

- **Dispersion Variance:** Younger drivers display greater dispersion in their scores. This indicates a high degree of volatility in this segment—while some are cautious, others rapidly accumulate poor history.
- **Convergence with Age:** As driver age increases, the variability in claim history generally decreases. Older drivers tend to settle into more stable, predictable risk patterns.
- **Experience Factor:** This pattern suggests that `driver_age` acts as a proxy for **driving experience** and **maturity**. The initial years of driving are a "discovery phase" characterized by high variance.
- **Non-Linearity:** The relationship is clearly non-monotonic. Consequently, treating Age as a simple linear variable in the predictive model would be suboptimal. We recommend using **Age Bands (Binning)** or **Polynomial features** (e.g., Age^2) to capture this non-linear risk heterogeneity effectively.

1.3.2.4 Geographical Distribution of Policies

The spatial scatter plot visualizes the geographic footprint of the portfolio based on latitude and longitude coordinates.

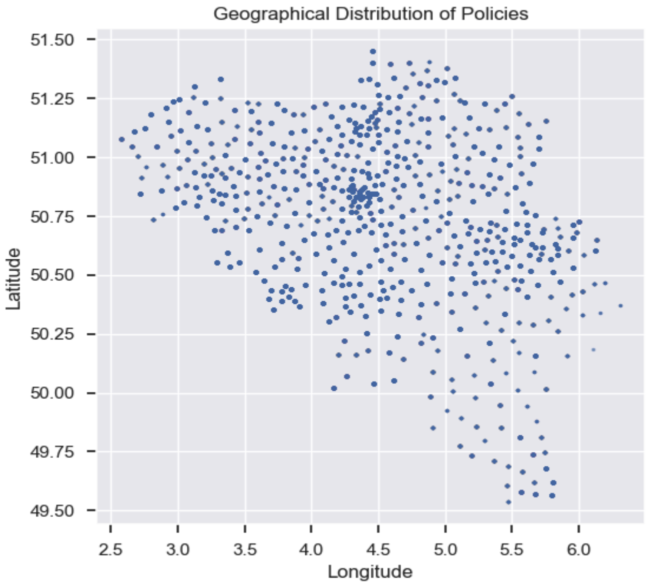


Figure 11: Geographical Distribution of Policies

- Geographic Identity: The coordinate range (Lat ~50, Long ~4) and the distinct shape strongly suggest the data covers Belgium.
- Density Clustering: The distribution is not uniform. We observe a high-density cluster in the center (likely Brussels and surrounding metropolitan areas) and higher density in the north (Flanders) compared to the sparser southern region (Wallonia).
- Urban vs. Rural Dynamics: This clustering highlights the need to differentiate between urban and rural risks. Urban areas (high density) typically correlate with higher claim frequency due to traffic congestion, while rural areas may be associated with higher severity (due to higher speeds).
- Proxy for External Factors: Location acts as a powerful composite variable capturing traffic density, road quality, and local weather patterns.
- Modeling Strategy: Since raw coordinates can be noisy, we recommend exploring spatial binning, using the postal code (`geo_postal`), or creating a "distance to city center" feature to capture these underlying risk gradients in the model.

1.4 Feature Engineering

Based on the insights from our EDA, we implemented the following transformations to better capture risk patterns:

1. **`geo_region` (Dimensionality Reduction):**
- *Reasoning:* The raw `geo_postal` variable has high cardinality. By truncating it to the **first 2 digits** (as implemented in the code), we aggregate policyholders into broader administrative regions (e.g., Provinces), capturing regional risk without overfitting.
2. **`car_age_bin` & `car_hp_bin` (Non-Linearity Handling):**
- *Reasoning:* Risk does not scale linearly. We discretized these variables based on business logic:
■ **Car Age:** Grouped into lifecycle stages: **'New' (0-3 yrs)**, **'Medium' (3-6 yrs)**, **'Old' (6-10 yrs)**, and **'Very Old' (10+ yrs)**.
■ **Horsepower:** Segmented into **'Low' (<60 hp)**, **'Medium' (60-90 hp)**, **'High' (90-120 hp)**, and **'Very High' (>120 hp)**.
3. **`usage_category` (Interaction Feature):**
- *Reasoning:* As noted in EDA, `car_is_fleet` and `car_usage` are highly imbalanced. We created a combined text feature (e.g., **`Private_private`**, **`Fleet_work`**) to explicitly isolate distinct risk profiles, such as commercial fleet usage versus personal private usage, which might otherwise be lost if treated independently.

Overall, these transformations aim to reduce dimensionality, capture non-linear risk effects, and make latent interactions explicit, thereby improving both model interpretability and segmentation stability in downstream tasks.

1.5 Variable Selection for Segmentation

Based on the EDA and feature engineering, we select the subset of variables to be used for the **K-Means Segmentation Model** in Part 2.

1. **Relevance:** Variables must describe the *customer* or the *asset* (policyholder profile), not the outcome (claims).
2. **Redundancy:** We analyze the correlation matrix to avoid multicollinearity.

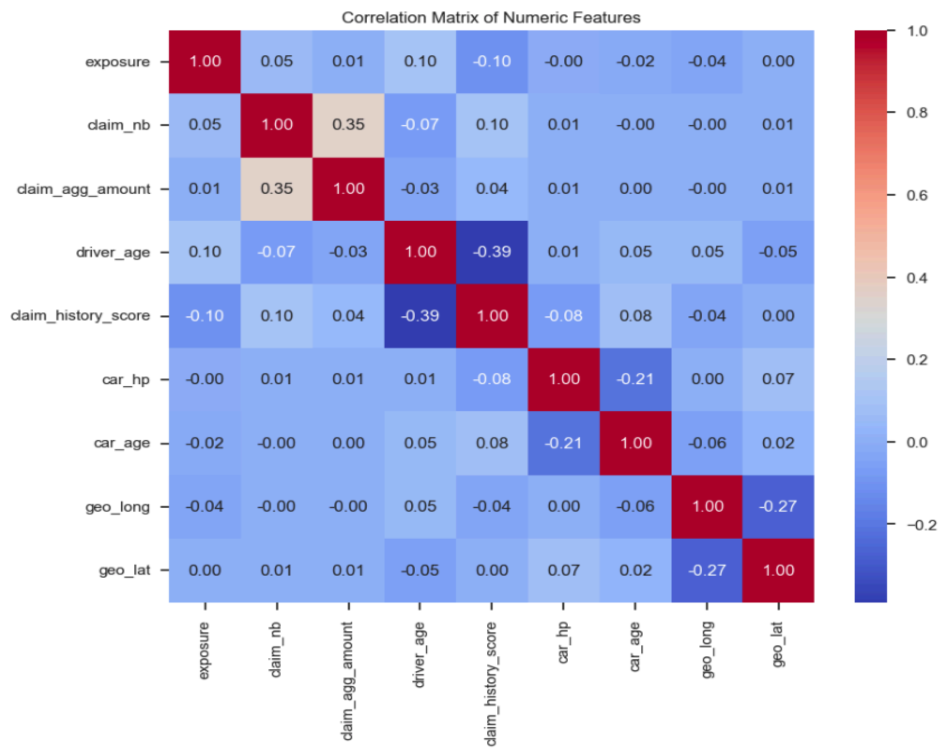


Figure 12: Correlation Matrix of Numeric Features

Based on the EDA, correlation analysis, and feature engineering performed above, we have finalized the set of variables to be used for the K-Means clustering model in Task 2.

We selected the following **6 features** to build the customer profiles:

- **`driver_age`** (Driver Profile)*:* Age is a primary determinant of risk appetite and driving experience. Young drivers and senior drivers often have distinct risk profiles.
- **`claim_history_score`** (Driver History): This score serves as a direct proxy for past driving behavior and riskiness. It differentiates "safe" drivers from "high-risk" ones regardless of their age.
- **`car_hp`** (Vehicle Characteristic): Engine power correlates with the type of vehicle (sports car vs. city car) and potential accident severity.
- **`car_age`** (Vehicle Characteristic): Distinguishes between owners of new, high-value vehicles (often needing Casco) and older, depreciated vehicles.
- **`geo_long`** & **`geo_lat`** (Location): Insurance risk is highly location-dependent (urban vs. rural, traffic density). Using coordinates allows the model to spatially cluster customers without needing high-cardinality postal codes.

Excluded Variables:

- **`claim_nb`** **&** **`claim_agg_amount`:** Excluded because these are *outcomes* (post-event), not *attributes* (pre-event). We want to segment customers based on who they are, not just how many claims they had this specific year.
- **`ins_coverage`:** Excluded from the core clustering to see if the derived risk groups naturally align with specific product choices.
- **Highly Correlated Variables:** We avoided using both `car_age` and `car_age_bin` simultaneously to prevent multicollinearity issues in K-Means (Euclidean distance distortion).

This subset of variables provides a balanced view of the **Driver**, the **Vehicle**, and the **Location**, creating a solid foundation for multi-dimensional segmentation.

Part 2：Building the Segmentation Model (K-means)

2.1 Objective and Overview

In this part, we build a customer segmentation model using the K-means clustering algorithm. Based on the selected risk-related variables from Part 1, we apply appropriate preprocessing steps to ensure K-means performs reliably under Euclidean distance. We then determine the optimal number of clusters using two different criteria (Elbow/Inertia and Silhouette score), and justify the choice of cluster initialization strategy.

2.2 Feature Selection for Clustering

We use a set of continuous risk drivers identified in Part 1 (e.g., driver age, vehicle characteristics, and claim history score). Outcome variables such as claim_nb and claim_agg_amount are not included in the clustering inputs to avoid creating clusters driven purely by historical losses. Geographic variables (geo_long, geo_lat) are evaluated separately to check whether segmentation becomes overly location-driven.

2.3 Preprocessing Strategy (Two Schemes)

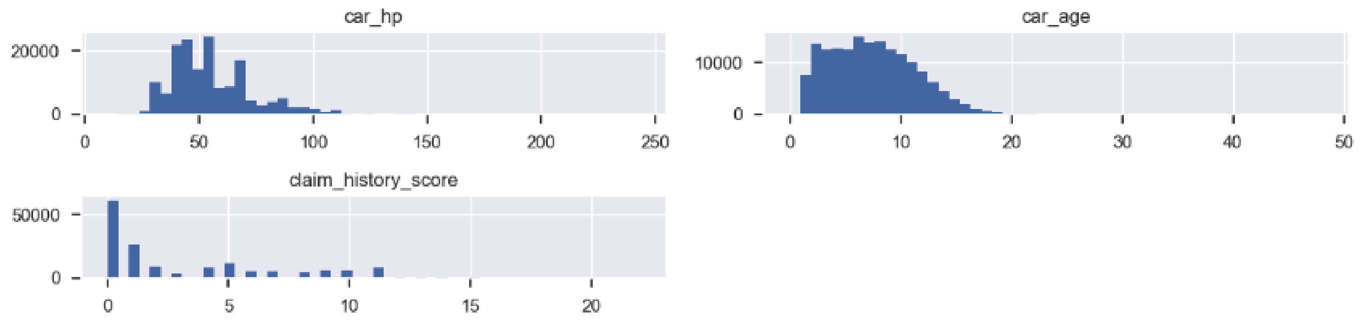


Figure 13: Marginal distributions of selected numerical features

K-means relies on Euclidean distance, which means two practical issues can strongly distort the clustering results: outliers and scale differences. If a small number of observations contain extremely large values (e.g., unusually high car_hp or very old car_age), these points can dominate distance calculations and pull cluster centroids toward them. All numeric variables are therefore standardized (using StandardScaler), because K-means is distance-based; otherwise variables with larger scales would dominate the clustering. To mitigate this, we apply winsorization by clipping selected numeric variables to the 1st and 99th percentiles, reducing the influence of extreme tails while preserving the majority of the distribution.

In addition, claim_history_score shows a strong right-skewed distribution (long right tail), where a large share of policies have very low scores (often near 0–2), but a smaller group has much higher values extending to the upper end. This creates an uneven distance structure: differences among high-score observations can become disproportionately large compared to differences in the main mass of the data. To address this, we consider a log1p transformation on claim_history_score (after winsorization). The log1p transform compresses the upper tail while keeping zero values valid (since log1p(0)=0), leading to a more balanced contribution of this feature in Euclidean space.

We therefore evaluate two preprocessing schemes:

- **S1 (Baseline):** Winsorization (1%–99%) + Standardization (StandardScaler) This setup reduces outlier impact and ensures all numeric variables contribute comparably to distance computations.
- **S2 (Enhanced):** Winsorization (1%–99%) + log1p(claim_history_score) + Standardization

This setup further addresses the pronounced right-skewness of claim_history_score by reducing tail dominance and making distances less sensitive to extreme claim history values.

Both schemes are compared using clustering quality metrics (Elbow/Inertia and Silhouette score) to determine whether the additional skewness correction in S2 improves separation and stability of the resulting segments.

Implementation note：The two schemes are implemented via a data preparation function and a scikit-learn pipeline. The function prepare_data_for_kmeans() applies winsorization and optionally log1p to claim_history_score (S2). The pipeline build_kmeans_pipeline() standardizes numeric variables (and optionally one-hot encodes categorical variables) before fitting K-means. The same pipeline is reused across all experiments to ensure a fair comparison.

2.4 Choosing the Optimal Number of Clusters (k)

2.4.1 Experimental Design

To determine the optimal number of clusters, we evaluate **k = 2 to 8** using two complementary criteria:

- **Elbow method (Inertia / SSE):** identifies the point where adding more clusters yields diminishing returns in reducing the within-cluster sum of squares.
- **Silhouette score:** measures both within-cluster cohesion and between-cluster separation (higher is better).

We compare four configurations formed by:

- **Feature sets:**
  A = Risk-only (driver_age, claim_history_score, car_hp, car_age)
  B = Risk + Geo (A + geo_long, geo_lat)
- **Preprocessing schemes:**
  S1 = Winsorization (1%–99%) + StandardScaler
  S2 = Winsorization (1%–99%) + log1p(claim_history_score) + StandardScaler

2.4.2 Results:Visual inspection (Elbow + Silhouette)

Below we report the Elbow (inertia/SSE) and Silhouette curves for each configuration. The Elbow plots provide a sanity check for diminishing returns as k increases, while the Silhouette score provides a direct measure of clustering quality in terms of separation and compactness.

2.4.2.1 A-S1 (Risk-only, winsorize + scale)

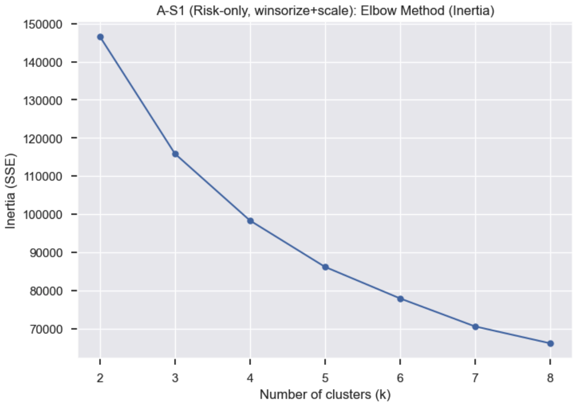


Figure 14: Elbow plot for A-S1 (Risk-only, winsorize + scale)

Elbow plot: inertia decreases sharply at small k and then flattens gradually, indicating diminishing returns after a small number of clusters.

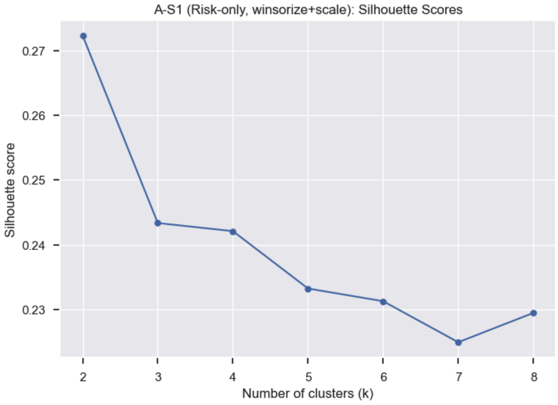


Figure 15: Silhouette plot for A-S1 (Risk-only, winsorize + scale)

Silhouette plot: the highest silhouette is at k = 2, followed by a clear drop for k ≥ 3, suggesting the best separation occurs with two clusters.

2.4.2.2 A-S2 (Risk-only, winsorize + log1p(history) + scale)

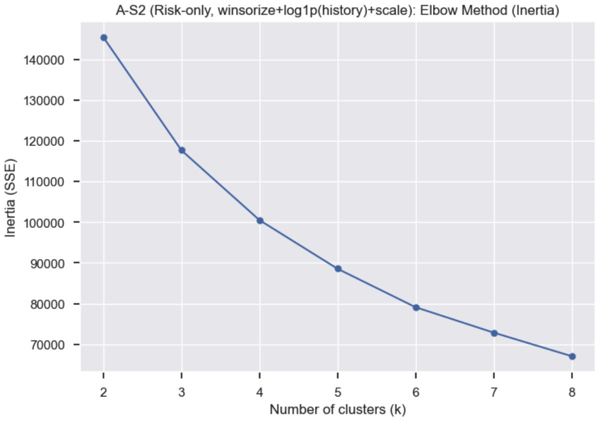


Figure 16: Elbow plot for A-S2 (Risk-only, winsorize + log1p(history) + scale)

Elbow plot: similar shape to A-S1 (monotonic decrease with diminishing returns).

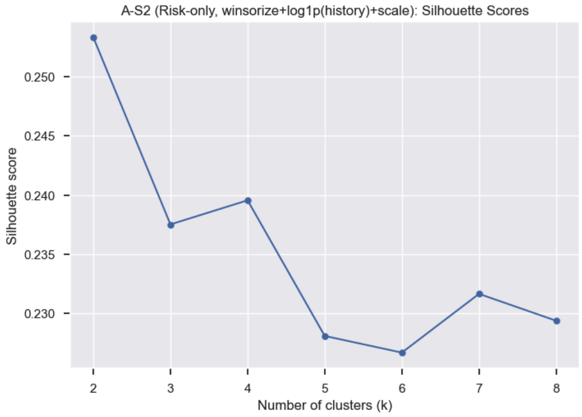


Figure 17: Silhouette plot for A-S2 (Risk-only, winsorize + log1p(history) + scale)

Silhouette plot: again peaks at k = 2, with lower values for larger k.

2.4.2.3 B-S1 (Risk + Geo, winsorize + scale)

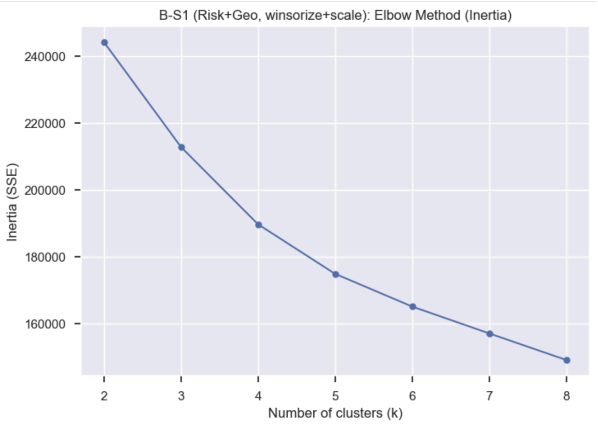


Figure 18: Elbow plot for B-S1 (Risk + Geo, winsorize + scale)

Elbow plot: inertia decreases as expected with k, but without a strong improvement beyond small k.

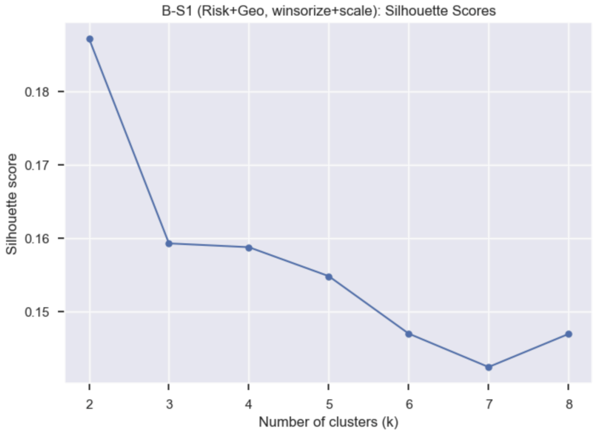


Figure 19: Silhouette plot for B-S1 (Risk + Geo, winsorize + scale)

Silhouette plot: the peak is at k = 2, but overall scores are noticeably lower than the risk-only case, indicating weaker separation when geographic coordinates are included.

2.4.2.4 B-S2 (Risk + Geo, winsorize + log1p(history) + scale)

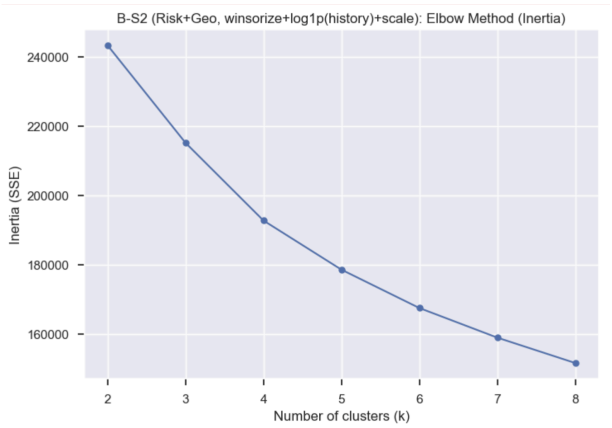


Figure 20: Elbow plot for B-S2 (Risk + Geo, winsorize + log1p(history) + scale)

Elbow plot: similar diminishing-return behavior.

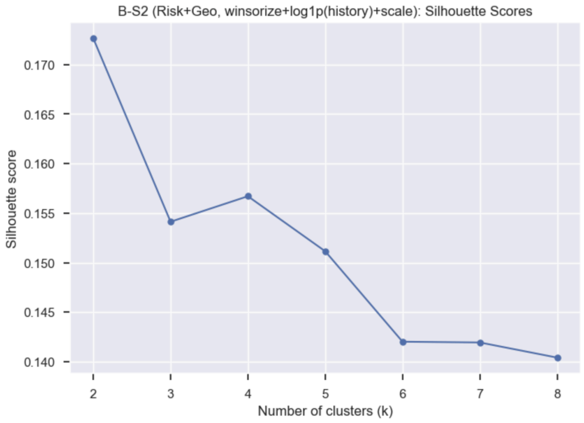


Figure 21: Silhouette plot for B-S2 (Risk + Geo, winsorize + log1p(history) + scale)

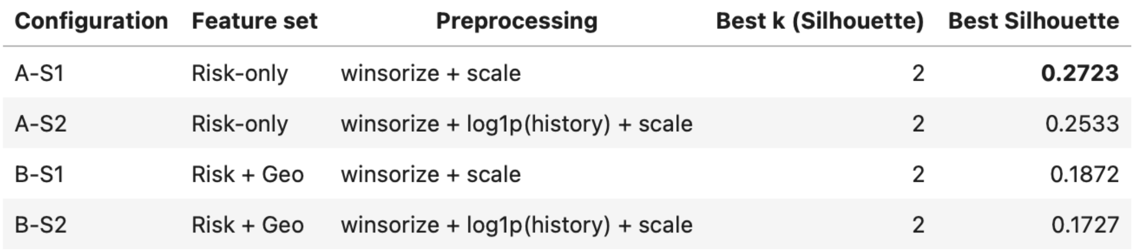
Silhouette plot: highest at k = 2, but again with the lowest overall separation among the four configurations.

Overall, both methods consistently point to a small number of clusters, and the Silhouette criterion clearly selects k = 2 in all configurations.

2.4.3 Quantitative comparison across configurations

The best k values (by silhouette) and the corresponding silhouette scores are:

Figure 22: Model selection summary for K-means segmentation (best k by silhouette) Table



Configuration	Feature set	Preprocessing	Best k (Silhouette)	Best Silhouette
A-S1	Risk-only	winsorize + scale	2	**0.2723**
A-S2	Risk-only	winsorize + log1p(history) + scale	2	0.2533
B-S1	Risk + Geo	winsorize + scale	2	0.1872
B-S2	Risk + Geo	winsorize + log1p(history) + scale	2	0.1727

- k = 2 is the optimal choice in all four configurations, indicating a stable preference for two segments.

- Risk-only models (A) outperform models that include geography (B): adding geo_long/geo_lat substantially lowers silhouette scores (≈0.17–0.19 vs. ≈0.25–0.27), meaning clusters become less well-separated when geography is included.
- The log1p transformation (S2) does not improve clustering quality: within both A and B, S2 yields lower silhouettes than S1, suggesting that winsorization + scaling already provides sufficient robustness, and further compression of claim_history_score reduces useful variation for segmentation.

2.4.4 Final decision

Elbow is used as a supporting check (diminishing returns), while the final k is chosen by maximizing the Silhouette score. Based on the consistent Silhouette peak at k = 2 and diminishing returns in inertia reduction (Elbow), we select:

- Final number of clusters: k = 2
- Final configuration: A-S1 (Risk-only, winsorization + StandardScaler)

This configuration provides the **highest silhouette score (0.2723)**, indicating the best balance of compact clusters and clear separation, while remaining interpretable as a **risk-based segmentation** rather than geographic grouping.

2.4.5 Practical considerations

Evaluating K-means across multiple values of k and four configurations is computationally expensive on the full dataset (163,212 observations), particularly for the silhouette score. Therefore, we perform k-selection on a 30% random subsample to reduce runtime. We additionally verified on the full dataset that the conclusions are essentially unchanged (the optimal k remains 2 and A-S1 remains the best-performing configuration). Based on this stability check, we keep subsampling for model selection and refit the final model on the full prepared dataset.

2.5 Initialization Strategy

K-means can converge to different solutions depending on the initial centroid positions, because the optimization problem is non-convex and may get trapped in a local minimum. To improve stability and reduce sensitivity to initialization, we apply two strategies:

- k-means++ initialization: spreads initial centroids farther apart than random initialization, which typically improves convergence quality and stability.
- Multiple random restarts (n_init): runs K-means multiple times with different initializations and selects the best solution based on the lowest inertia (SSE).

In our workflow, we use a smaller n_init during k-selection (for computational efficiency), and then refit the final model on the full dataset with a larger n_init to obtain a stable final segmentation.

2.6 FInal Model Fitting (K-means)

Based on Section Choosing the Optimal Number of Clusters, we choose:

- Final configuration: A-S1 (Risk-only, winsorize + scale)
- Final number of clusters: k = 2

In this section, we refit the final K-means model on the full prepared dataset (not the sampling subset used for k-selection), using a larger n_init to ensure stability. The resulting cluster labels are then merged back into the original dataset for the next step

2.7 Conclusion

In this part, we built a customer segmentation model using K-means based on the variables selected in Part 1. Because K-means relies on Euclidean distance and is sensitive to scale and extreme values, we applied robust preprocessing and compared alternative modeling choices before selecting the final specification.

2.7.1 Preprocessing and feature sets

We evaluated two preprocessing schemes:

- **S1:** winsorization (1%–99%) + standardization (StandardScaler)
- **S2:** winsorization (1%–99%) + log1p(claim_history_score) + standardization

We also compared two feature sets:

- **A (Risk-only):** driver_age, claim_history_score, car_hp, car_age
- **B (Risk + Geo):** A + geo_long, geo_lat

The best-performing configuration was **A-S1**, which achieved the highest separation quality (Silhouette = **0.2723**) and outperformed both the log-transformed variant (A-S2: 0.2533) and the geo-augmented variants (B-S1: 0.1872; B-S2: 0.1727). Therefore, we proceeded with **risk-only features** and the **baseline preprocessing**.

2.7.2 Choice of the optimal number of clusters (two methods)

We selected the optimal number of clusters by combining:

1. **Silhouette score:** maximized at **k = 2** across all tested configurations, indicating the best balance of within-cluster cohesion and between-cluster separation.
2. **Elbow method (Inertia/SSE):** inertia decreases strongly at small k and then shows diminishing returns as k increases, supporting a small number of clusters. Together with the silhouette peak, this supports choosing **k = 2**.

2.7.3 Initialization strategy

To reduce sensitivity to initial centroids and avoid poor local minima, we used:

- k-means++ initialization (better-spread starting centroids), and
- multiple random restarts (`n_init`), selecting the best solution based on the lowest inertia (SSE).

2.7.4 Final segmentation model

The final model is a **K-means (k = 2)** segmentation fitted on the full prepared dataset using **A-S1** (winsorization + standardization) with **k-means++** and a larger `n_init` for stability. The resulting cluster labels are saved back to the dataset for subsequent profiling and interpretation in Task IV.

Part 3: Business Analysis

3.1 Segment size (Portfolio composition)

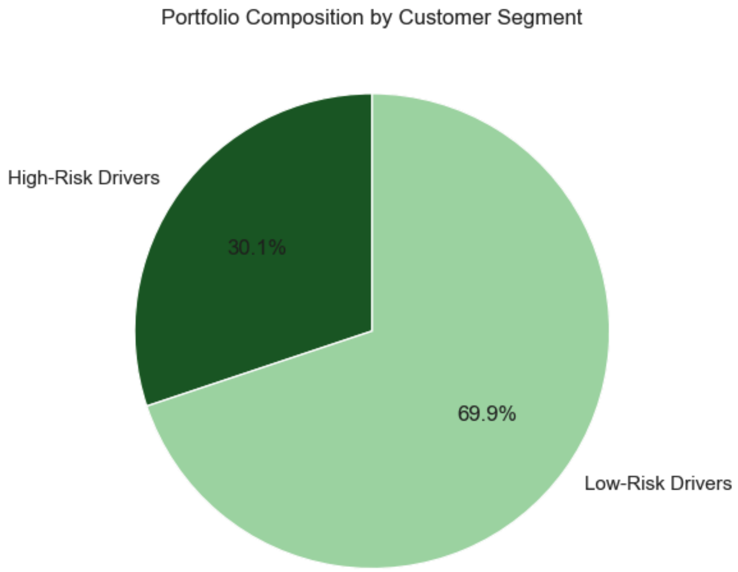
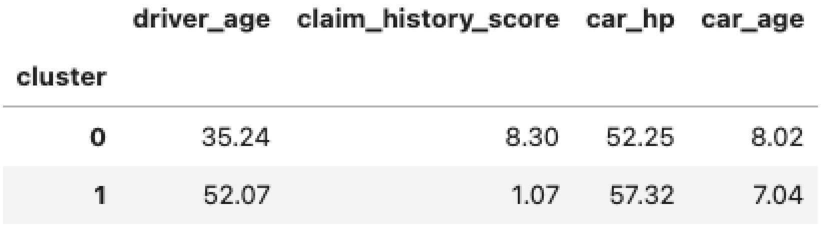


Figure 23: Portfolio Composition by Customer Segment

Cluster 0 represents 30% of the portfolio, while Cluster 1 accounts for 70%.

3.2 Core Segment Profiling



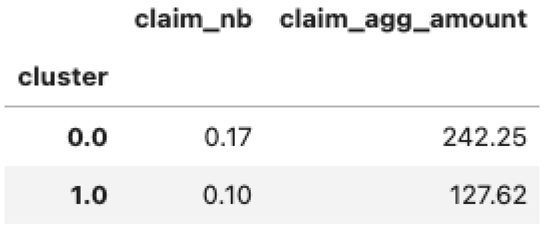
	driver_age	claim_history_score	car_hp	car_age
cluster				
0	35.24	8.30	52.25	8.02
1	52.07	1.07	57.32	7.04

Figure 24: Variables profiling Table

Cluster 0 is characterized by relatively younger drivers with a significantly higher claims history score, indicating elevated behavioral risk. Despite owning vehicles of similar age and horsepower to the other group, this segment exhibits a substantially riskier driving profile driven primarily by past claims behavior rather than vehicle characteristics.

In contrast, Cluster 1 consists of older, more experienced drivers with a very low claims history score, representing a low-risk segment. The segmentation is therefore mainly driven by driver-related risk factors, particularly age and historical claims behavior, rather than vehicle attributes.

- Cluster 0: High-Risk Younger Drivers
- Cluster 1: Low-Risk Experienced Drivers



	claim_nb	claim_agg_amount
cluster		
0.0	0.17	242.25
1.0	0.10	127.62

Figure 25: Claims Outcomes by Cluster Table

Cluster 0 has a claim frequency approximately 70% higher than Cluster 1. Cluster 0 causes approximately 90% more damage.

Although claim variables were not used in the clustering process, the resulting segments differ substantially in realized risk. Cluster 0 exhibits both higher claim frequency (0.17 vs. 0.10) and higher claim severity (242.25 vs. 127.62), indicating a significantly riskier segment. In contrast, Cluster 1 represents a low-risk group with fewer and less costly claims. This validates the risk-based nature of the proposed segmentation.

The segmentation is not only statistically meaningful but also economically relevant.

3.3 Visualization

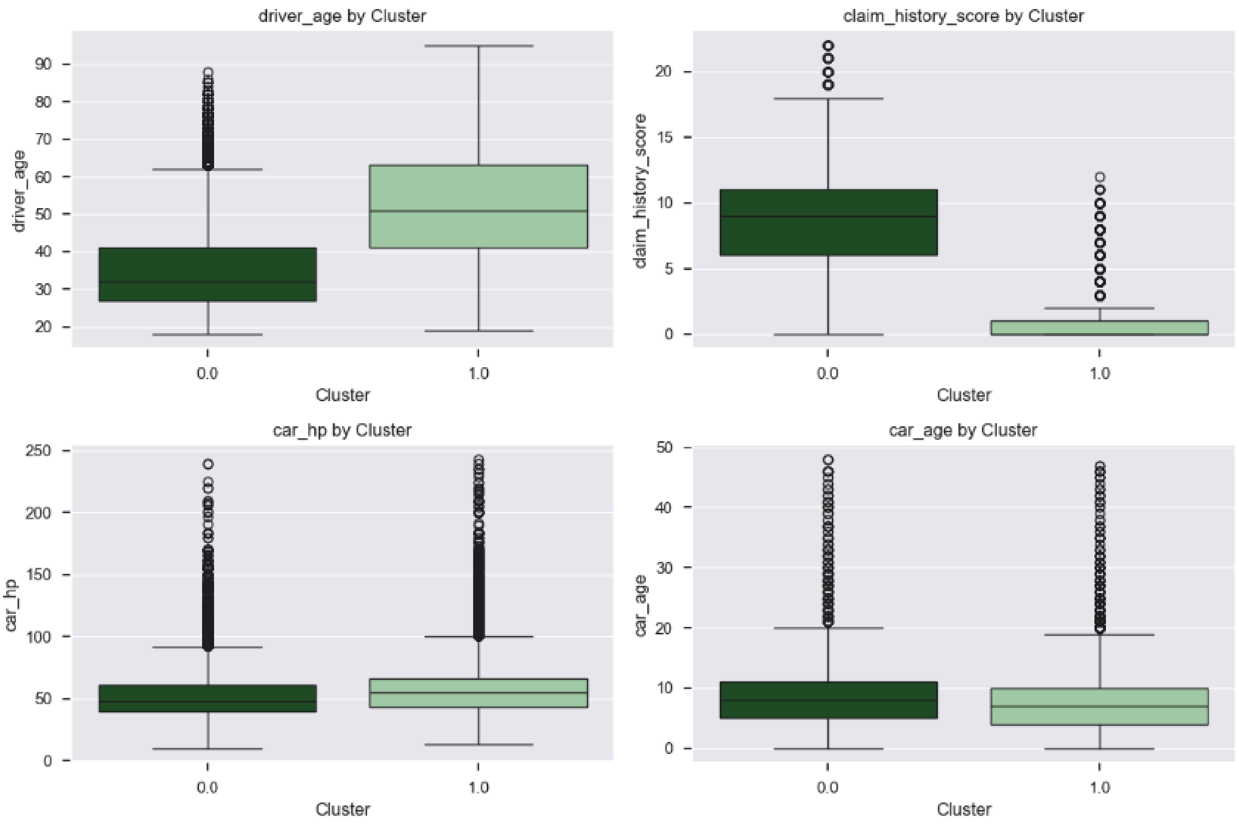


Figure 26: Boxplots of Key Risk Drivers by Cluster

1. Driver Age

The boxplot shows a clear separation between the two clusters. Cluster 0 consists mainly of younger drivers, with a median age of around 30–32 years, while Cluster 1 is dominated by older drivers, with a median age close to 50 years. The limited overlap in medians indicates that driver age is a strong discriminating variable in the segmentation. It means Cluster 0 represents younger drivers, whereas Cluster 1 consists of older and more experienced drivers.

2. Claim History Score

The difference between clusters is particularly pronounced for the claim history score. Cluster 0 exhibits a high median claim history score (around 8–9), with a right-skewed distribution, indicating frequent past claims. In contrast, Cluster 1 has a median close to zero, meaning that most drivers in this group have little or no prior claims history. This variable is the strongest driver of the segmentation. Cluster 0 is characterized by high behavioral risk, while Cluster 1 represents drivers with a clean claims history.

3. Car Horsepower

The distributions of car horsepower largely overlap across clusters. Although Cluster 1 shows a slightly higher median horsepower, the difference is modest. Vehicle power plays only a secondary role in the segmentation and is not a primary driver of risk differentiation.

4. Car Age

The distributions of vehicle age are very similar across clusters, with only minor differences in medians. Car age does not significantly contribute to the separation of clusters, suggesting that vehicle age is not a key determinant of the identified risk profiles. The segmentation is primarily driven by driver-related characteristics-especially driver age and claim history-while vehicle characteristics play a secondary role.

More evidence confirming for 2 Cluster groups; **Cluster 0**: High-Risk Younger Drivers and **Cluster 1**: Low-Risk Experienced Drivers.

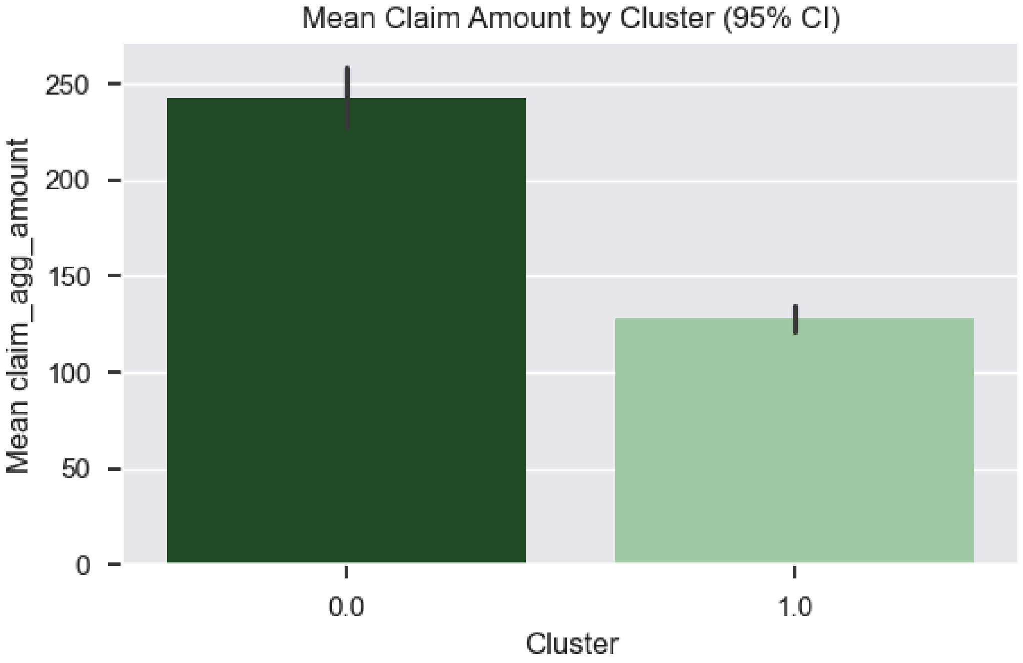


Figure 27: Mean Claim Amount by Cluster (95% Confidence Interval)

The bar chart shows a substantial difference in average claim severity between the two clusters. Cluster 0 has a mean claim aggregate amount of approximately 242, whereas Cluster 1's mean is around 128. The confidence intervals show limited overlap, indicating a meaningful difference in realized losses.

Although claim amounts were not used in the clustering process, the resulting segments exhibit significantly different loss levels. This provides strong economic validation of the segmentation.

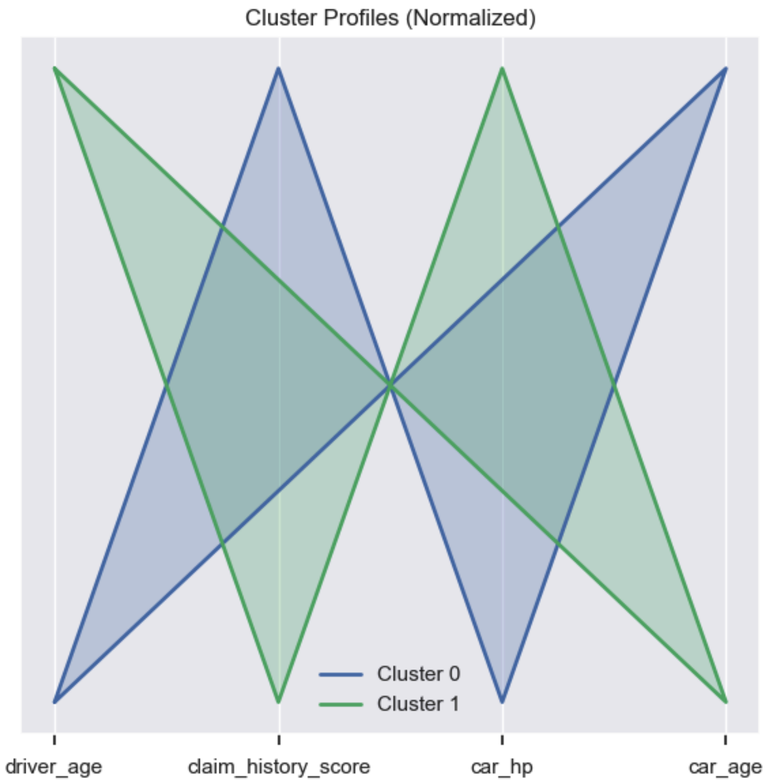


Figure 28: Radar Chart – Normalized Cluster Profiles

The radar chart summarizes the normalized profiles of both clusters across the four segmentation variables. Cluster 0 shows low values for driver age and very high values for claim history score, with moderate values for vehicle characteristics. Cluster 1 displays the opposite pattern, with high driver age, very low claim history score, and similar vehicle characteristics to Cluster 0.The radar chart clearly illustrates two contrasting risk profiles. The clusters are primarily separated along the dimensions of driver experience and historical claims behavior.

3.4 Business Insights and Interpretation

Cluster 0 – Younger High-Risk Drivers

This segment is characterized by younger drivers with a significantly worse claims history. Despite owning vehicles that are broadly similar to those in the other cluster, these policyholders generate both higher claim frequency and higher claim severity. This indicates that risk in this segment is mainly driven by behavioral factors rather than vehicle attributes.

This group may require higher premiums, stricter underwriting criteria, and additional risk mitigation measures such as higher deductibles or coverage limitations.

Cluster 1 – Experienced Low-Risk Drivers

This segment consists of older, more experienced drivers with a clean claims history. Even though their vehicles are comparable in age and power to those of Cluster 0, these drivers exhibit substantially lower claim frequency and severity.

This group represents a low-risk and potentially highly profitable segment, suitable for retention strategies, loyalty programs, and more competitive pricing.

Part 4: Claim frequency modeling

4.1 Models selection

Both the baseline Poisson model and the alternative Negative Binomial model were estimated within the Generalized Linear Model (GLM) framework. This choice was motivated primarily by the need for interpretability in insurance risk modeling.

The GLM framework provides a clear separation between systematic risk effects, captured by the regression coefficients, and random variability in claim counts. In particular, the regression coefficients can be interpreted as multiplicative effects on the expected claim frequency, which is essential for understanding risk drivers and for actuarial applications such as pricing and underwriting.

Table 1: Mean and variance of number of claim

Mean of target	Variance of target
0.123	0.135

The empirical distribution of the target variable number of claims exhibits a mean of 0.123 and a variance of 0.135. By explicitly allowing the variance to exceed the mean through an additional dispersion parameter, the Negative Binomial model can be better suited to capture this characteristic of the data, thereby providing a more flexible and robust framework for modeling claim frequency.

4.2 Feature engineering

In order to improve the explanatory power of the claim frequency models and to better capture the underlying risk structure of the insurance portfolio, several engineered features were created based on actuarial and behavioral considerations as the table below:

Table 2: New features and descriptions

New features	Description
drive age square	captures the well-known U-shaped relationship between driver age and accident risk, where very young and elderly drivers tend to have higher claim frequencies than middle-aged drivers
car age square	allows the model to capture non-linear effects of vehicle aging. Claim frequency typically increases as vehicles age due to wear and reduced safety features, but very old vehicles are often driven less, leading to a potential decline in claims
history exposure interaction	= claim history score × exposure; drivers with a high claim history score are already high-risk, so the marginal impact of additional exposure on claim frequency may differ from that of low-risk drivers
driver young	identifies young drivers (below 25 years), who are typically associated with higher accident risk due to limited driving experience

driver old	flags elderly drivers (above 65 years), for whom declining reaction time and health may increase accident risk
old car	indicates vehicles older than 10 years, which are more likely to lack modern safety features and may be more prone to mechanical failures

4.3 Feature selection

Starting from a full set of original and engineered variables, the backward selection procedure removed two variables: ins_coverage_OC_AC, which represents extended full casco coverage relative to the baseline MTPL product, and driver_young_1, an indicator for young drivers. Both variables were removed because their exclusion led to a reduction in AIC, meaning that they did not contribute additional explanatory power to the model once other variables were included.

The set of selected features were subsequently used consistently in both the Poisson and Negative Binomial models to ensure comparability and interpretability of the results.

4.4 Models comparison

Table 3: Poisson and Negative Binomial GLM models metrics

Model	Train AIC	Train LogLik	Test LogLik	Test RMSE	CV mean LogLik
Poisson	87144.905241	-43558.452621	-18725.379156	0.362862	-8714.973558185638
Negative Binomial	87039.353285	-43505.676643	-18727.196081	0.362864	-8703.911478704034

Both the Poisson and Negative Binomial claim frequency models were estimated on the training dataset and subsequently evaluated on a hold-out test set as well as through k-fold cross-validation. On the training data, the Negative Binomial model achieves a lower AIC and a higher log-likelihood than the Poisson model, indicating a superior in-sample fit. When evaluated on the test set, the two models exhibit nearly identical predictive performance, with only negligible differences in log-likelihood and RMSE. However, cross-validation results show that the Negative Binomial model consistently attains a higher average log-likelihood across folds, suggesting more robust generalization. Taken together, these results indicate that while both models perform similarly on a single test split, accounting for overdispersion through the Negative Binomial specification leads to a more stable and reliable model across different data partitions.

4.5 Negative Binomial model interpretation

Table 4: Negative Binomial model's coefficient

Variable	Coefficient	exp(coef)	p-value	Stat. significance

const	-1.7586	0.172	0.000	***
driver_age	-0.0193	0.981	0.000	***
claim_history_score	0.1580	1.171	0.000	***
car_hp	0.0039	1.004	0.000	***
car_age	0.0237	1.024	0.001	***
driver_age_sq	0.0001	1.000	0.020	**
car_age_sq	-0.0018	0.998	0.000	***
hist_exposure_inter	-0.1052	0.900	0.000	***
ins_coverage_OC_ACmini	-0.0377	0.963	0.070	n.s.
car_fuel_gasoline	-0.1826	0.833	0.000	***
car_usage_work	-0.0630	0.939	0.135	n.s.
car_is_fleet_1	-0.0984	0.906	0.071	n.s.
driver_old_1	-0.0812	0.922	0.139	n.s.
old_car_1	0.0931	1.098	0.008	***

Significance codes:

- *** $p < 0.01$

- ** p < 0.05
- n.s. = not statistically significant

The results indicate that several key risk drivers are statistically significant at conventional levels. Claim history score is the most influential variable, with a highly significant positive effect (p < 0.01), implying that an increase of one unit raises expected claim frequency by approximately 17%, holding other factors constant. Driver age and vehicle age exhibit statistically significant non-linear effects, as confirmed by the significance of both linear and squared terms, supporting the presence of U-shaped risk patterns.

Vehicle-related characteristics such as engine power and the old car indicator are also statistically significant (p < 0.01), indicating higher claim frequency for more powerful and older vehicles. The interaction between claim history and exposure is strongly significant and negative, suggesting diminishing marginal risk from additional exposure among drivers with poor claim histories.

Some variables, including mini casco coverage and fleet status, are only weakly significant at the 10% level, while work usage and the elderly driver indicator are not statistically significant once other risk factors are controlled for. Overall, the statistically significant coefficients are economically intuitive and confirm that the model captures the main structural drivers of claim frequency, while weaker effects may reflect secondary or context-dependent risk factors.

4.6 Extension: Discrete Negative Binomial model

In addition to the GLM-based models, a fully parametric discrete Negative Binomial specification was estimated as an extension to assess potential gains in predictive performance. By jointly estimating the regression coefficients and the dispersion parameter, this model provides greater flexibility and achieves improved likelihood-based fit measures.

Table 5: Discrete Negative Binomial model metrics

Model	Train AIC	Train LogLik	Test LogLik	Test RMSE	CV mean LogLik
Poisson	87144.905241	-43558.452621	-18725.379156	0.362862	-8714.973558185638
Negative Binomial	87039.353285	-43505.676643	-18727.196081	0.362864	-8703.911478704034
Discrete Negative Binomial	86928.62	-43449.31	-18693.51	-	-9820.21

This fully parametric model achieves a substantially better in-sample fit, with a higher training log-likelihood and a notably lower AIC. These results confirm that jointly estimating the regression coefficients and the dispersion parameter provides additional flexibility in fitting the observed claim counts. The improvement in fit also carries over to out-of-sample evaluation. On the test set, the discrete Negative Binomial model attains a higher log-likelihood (−18,694) compared to both the Poisson and the GLM Negative Binomial models, indicating superior predictive performance on unseen data.

However, its cross-validation performance is substantially worse than that of the GLM-based models. This indicates that the fully parametric specification, although highly flexible, suffers from reduced stability across different data partitions. The numerical warnings observed during the optimization process further corroborate the instability of the fully parametric approach compared to the robust GLM framework. In contrast, the GLM-based Negative Binomial model provides the highest average cross-validated log-likelihood, suggesting superior robustness and generalization. The estimated dispersion parameter ($\alpha \approx 0.51$) provides clear evidence of substantial overdispersion in the claim count data, reflecting pronounced unobserved heterogeneity across policyholders.

The three claim frequency models exhibit clear trade-offs between fit, predictive performance, and stability. The Poisson GLM serves as a robust baseline, showing competitive performance on the test set but weaker in-sample fit. The GLM-based Negative Binomial model improves upon the Poisson specification by accounting for overdispersion, achieving a lower AIC, higher training log-likelihood, and the best cross-validated performance, indicating superior robustness across different data partitions. The discrete Negative Binomial model delivers the strongest in-sample fit and the highest test-set log-likelihood, but its substantially poorer cross-validation performance suggests reduced stability and a higher risk of overfitting. Taken together, these results support the GLM-based Negative Binomial model as the preferred specification, offering the best balance between predictive robustness and interpretability, while the discrete Negative Binomial model is retained as a robustness check illustrating the trade-off between model flexibility and stability.