IDS

Post-graduation in Data Science for Finance

Insurance Data Science

Group Project

title

June 2025

Group:

Francisco Perestrello - 20241560
Gonçalo Gomes - 20211007
Nuno Vieira - 20241111
Petr Terletskiy - 20211580

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA IMS Information Management School

INDEX

ABSTRACT

This project presents a comprehensive analysis of automobile insurance claims data for the IMS Insurance Company, focusing on Third Party Liability coverage.

The first part conducts an Exploratory Data Analysis (EDA) of claim counts and severity, identifying key trends, distributions, and variable interactions through statistical and graphical techniques. Probability distributions are fitted to both claim counts and severity, with insights into excluded claims' implications for premiums.

The second one proposes a pricing structure using Generalized Linear Models (GLMs) to estimate claim frequency and severity for common claims, supplemented by a Machine Learning approach for large claims. The analysis identifies risk profiles, estimates premiums, and discusses fairness and adequacy.

The findings provide actionable recommendations for optimizing IMS Insurance Company's pricing strategy.

KEYWORDS

Automobile Insurance, Third Party Liability, Exploratory Data Analysis, Claim Frequency, Claim Severity, Generalized Linear Models, Machine Learning, Pricing Structure, Distribution Fitting, Risk Profiles.

INTRODUCTION

This report addresses the evaluation of the automobile insurance portfolio's claim data, with a focus on Third Party Liability coverage. The analysis leverages two datasets, where the first one contains the policy details, and the other one details the claim records.

The primary objective is to provide actionable insights into claim behavior, enabling informed pricing strategies through a structured examination of claim counts, severity, variable interactions, distribution fitting, and predictive modeling.

The project and the report are divided into two main parts. **Part I** begins with an Exploratory Data Analysis (EDA) of claim counts, assessing key variables such as driver age, vehicle age, and geographic zone, followed by a descriptive analysis of claim severity. Graphical techniques, including scatterplots, boxplots, heatmaps, and interaction plots, explore variable interactions, while distribution fitting addresses claim counts and severity. **Part II** proposes a pricing structure, employing Generalized Linear Models (GLM) to model claim frequency and severity for common claims, and a logistic regression approach for large claims, culminating in a comprehensive premium framework.

Utilizing R for statistical analysis and visualization, this study integrates robust methodologies to identify risk profiles, estimate claim metrics, and recommend pricing adjustments.

NOVA IMS Information Management School

PART 1

DATA ANALYSIS AND INSIGHTS

1. **EXPLORATORY ANALYSIS OF CLAIM COUNTS**
2. **EXPLORATORY ANALYSIS OF CLAIM SEVERITY**
3. **GRAPHICAL ANALYSIS OF VARIABLE INTERACTIONS**
4. **DISTRIBUTION FITTING FOR CLAIMS DATA**

1. DATA ANALYSIS AND INSIGHTS

Through Exploratory Data Analysis (EDA), descriptive statistics, and graphical techniques, we will investigate Number of Claims and Claim Severity to identify key patterns, trends, and variable interactions.

We will also fit probability distributions to both datasets, after removing the highest claim count outlier and applying an upper threshold for claim severity, while analyzing excluded claims to inform premium considerations, laying a foundation for the pricing structure in Part II.

1.1 EXPLORATORY ANALYSIS OF CLAIM COUNTS

NUMERICAL FEATURES

By conducting a brief analysis on the number of the contract variable, it can be confirmed that each contract is uniquely identified, ensuring that our claim count analysis accurately reflects individual contract behavior without any aggregation or duplication issues

The number of claims averages 0.05394 with a standard deviation of 0.2534, showing high skewness since 95.02% of policyholders report zero claims, with an outlier at 16 claims. The plots of nclaims (Figure 1) reveal a sharp peak at zero followed by a steep drop and a single occurrence at 16, suggesting the presence of an extreme outlier. This outlier will be removed for further distribution fitting, as it does not represent typical claim behavior.

The exposition variable, representing exposure time, averages 0.53 years with a standard deviation of 0.36. Its plots (Figure 2) show a concentration around 1.0 year, indicating most policyholders have nearly full-year exposure. The portfolio claim frequency, calculated as 0.1 claims per policy-year, aligns with the observed low claim incidence.

The agedriver variable (Figure 3), with a mean of 45.5 years and a standard deviation of 14.1, shows a distribution peaking around 35-50 years, indicating a predominance of middle-aged drivers. The agevehicle variable (Figure 4), with a mean of 7.1 years and a standard deviation of 5.7, peaks at younger vehicle ages (0-10 years) and declines sharply afterwards, reflecting a fleet skewed toward newer vehicles.

CATEGORICAL FEATURES

Zone frequency (Figure 5) speaks in Zone C with over 14,000 policies, followed by Zones D and E, while Zone F is the least frequent at around 1,000. Claim frequencies range from 0.0782 in Zone A to a high of 0.1351 in Zone E, suggesting that there are geographic risk factors.

Brand distribution (Figure 6) shows dominance by Brands 1, 2, and 12 (around 12,000 policies each), with Brand 14 at just 349, and claim frequencies vary from 0.0754 (Brand 14) to 0.1315 (Brand 10), suggesting brand-specific risk profiles.

Fuel type (Figure 7) is nearly evenly split between diesel (D) and gasoline (E), though diesel exhibits a slightly higher claim frequency (0.1097 vs. 0.0952).

Power categories (Figure 8) peak in the 5-7 range, declining sharply beyond 9, with claim frequencies ranging from 0.0838 (power 11) to 0.1417 (power 14).

1.2 EXPLORATORY ANALYSIS OF CLAIM SEVERITY

NUMERICAL FEATURES

Figure 9 shows a highly right-skewed claim cost distribution, with most values near zero, including negative or low amounts likely due to adjustments like reimbursements or cancellations, alongside high-value outliers reflecting large payouts. To enhance analysis reliability, claims below €25 (790 rows) are excluded. Post-exclusion (Figure 10), the distribution remains right-skewed, with a mean claim cost of €1,730.71 and a median of €1,172, underscoring the impact of high-cost claims, and a maximum of €75,000 highlighting significant outliers.

The dataset shows an average driver age of 45.2 years with a standard deviation of 14.6 years, reflecting a wide age distribution. The scatter plot (Figure 11) reveals a non-linear relationship between driver age and claim severity: younger drivers (18-30) exhibit higher severities, possibly due to riskier behavior or inexperience, while those aged 30-50 show a moderate decline, suggesting more stable driving. The lowest severities occur between 50-65, marking the least risky group, but severity rises again after 65, likely due to slower reflexes or health issues increasing personal injury claims.

The vehicle age variable averages 7.5 years (median also 7.5 years) with a standard deviation of 5.38, showing moderate right-skewness due to a long tail and outliers up to 69 years, though 90% of vehicles are under 14 years. The scatter plot (Figure 12) indicates a negative relationship between vehicle age and claim severity: newer vehicles (0-10 years) show higher severities (€1,500-€2,000), declining as age increases, likely due to higher repair costs for advanced components in newer models versus cheaper repairs or write-offs for older ones.

CATEGORICAL FEATURES

The overall differences in average claim severity across zones are modest, indicating geographic zone has a moderate effect and is not a dominant risk factor alone. Zone E has the highest average, closely followed by Zone D, with a noticeable but small gap to the lowest in Zone F, while Zones A, B, and C exhibit similar severities.

Average claim severity across power categories shows little variation, except for power levels 11 and 14 (and possibly 13), which exhibit notably lower costs, suggesting distinct usage patterns. Other power groups display consistent severities with minor fluctuations.

The chart highlights distinct claim severity variations among vehicle brands, with most showing similar average costs, though Brand 13 stands out with the highest at over €2,500, well above the norm, while Brands 5 and 6 exhibit lower severities, suggesting less costly damage or repairs. Significant peaks at the upper end suggest brand influences claim severity. Fuel D presents a higher average cost, however this difference is not significant.

1.3 GRAPHICAL ANALYSIS OF VARIABLE INTERACTIONS

The graphical analysis of variable interactions for claim frequency and claim severity in IMS Insurance Company's Automobile Insurance portfolio reveals significant insights into the relationships between claim counts and claim severity and categorical or numerical variables.

CLAIM FREQUENCY

The correlation heatmap (Figure 13) reveals only weak linear relationships among all numerical features. The strongest correlation found, between exposition and vehicle age, is less than 0.2, which isn't a strong correlation.

The vehicle age by fuel type boxplot (Figure 14) reveals diesel (D) and gasoline (E) vehicles share similar median ages (5-10 years) with outliers up to 100 years, suggesting a wide age distribution that may affect claim patterns, notably with diesel's higher frequency (0.1097 vs. 0.0952 for gasoline). The boxplot of vehicle age by driver age (Figure 15) shows younger drivers (18-24) have a slightly higher median age (5-10 years), while older drivers tend toward newer vehicles.

The brand vs. power heatmap (Figure 16) shows most combinations have low claim frequencies, except for Brand 6, where frequencies spike at power levels 12 and 13, indicating elevated claim risk for these vehicles. The fuel type vs. vehicle age heatmap (Figure 17) reveals Diesel cars generally have higher claim frequencies than Gasoline, peaking at [5-10] years, while Gasoline peaks at [10-15] years.

The zone vs. driver age heatmap (Figure 18) highlights younger drivers (18-24) with the highest frequencies across all zones, decreasing with age, with Zone A showing the lowest frequencies overall, especially for 64-101 age group. The driver age vs. vehicle age heatmap (Figure 19) indicates a significant interaction, with the highest claim frequency for 18-24-year-olds with vehicles aged [25-50] years, while other age groups show lower frequencies.

The analysis of insurance claim frequency across multiple variables reveals critical insights into risk factors and their interactions.

Younger drivers, particularly those aged 18-24, exhibit the highest claim frequency (Figure 20) at approximately, with a clear decline as age increases. However, estimates for younger drivers are less precise due to lower exposure (around 1,000 units) compared to the peak exposure of 10,000 units for the 44-64 group. Geographically, Zones D, E, and F show (Figure 21) elevated claim frequencies, while Zone A and B show lower claim frequencies.

Vehicle power indicate higher claim frequencies (Figure 22) at higher levels (from 12 to 15), though estimates are less reliable due to limited exposure at higher levels (e.g., 200 units at 14 and 15). Vehicle age shows (Figure 23) stable frequencies with only a few peaks. These peaks, however, are based on limited data and are therefore less reliable indicators of the overall distribution. Diesel cars slightly exceeds the Gasoline ones (Figure 24) in claim frequency, a small but potentially significant distinction.

CLAIM SEVERITY

To explore the relationship between driver age and claim severity, various binning strategies were tested, like fixed-width intervals and quantile-based groupings. After evaluating these groupings shown in (Figure 26), custom bins ([18–24], [25–34], [35–44], [45–64], 65+) were selected, aligning with life stages (young adulthood, early career, midlife, pre-retirement, retirement) and industry underwriting practices that reflect age-related behavior and risk shifts.

Visualizations confirm a U-shaped severity pattern, with higher costs for youngest and oldest drivers, declining in middle age, especially pre-retirement. Notably, the 35–44 age group shows a slight increase in average claim severity compared to the 25–34 and 45–64 groups, deviating from the expected middle-age decline but fitting the broader U-shaped pattern. This elevation may stem from transitional factors like increased vehicle use for family or work, ownership of costlier cars, or younger family members (e.g., sons) driving, temporarily raising claim costs.

A similar method was applied to analyze the relationship between vehicle age and claim severity, using fixed-width intervals and quantile-based deciles. All visualizations (Figure 27) reveal a clear negative trend, with average claim severity decreasing as vehicle age increases. This supports fixed age-based segmentation, which enhances interpretability and precision, offering practical insights into how depreciation and usage patterns affect claim costs, making it ideal for insurance modeling.

The interactions observed in Categorical Features (Figure 28) reinforce the trends identified in our prior analysis of categorical variables, enhancing the overall data interpretation.

1.4 DISTRIBUTION FITTING FOR CLAIMS DATA

To model the Number of Claims and Claim Severity in the car insurance dataset, suitable probability distributions were fitted to reflect underlying patterns, aiding accurate risk assessment for pricing. The analysis involved removing the highest outlier for Number of Claims, setting an upper threshold for Claim Severity to fit an Exponential Family distribution, and assessing excluded claims. Subsequent sections outline the methodology, results, and implications for insurance pricing.

NUMBER OF CLAIMS

The Number of Claims, represented by the variable nclaims in the dataset, quantifies the frequency of claims per policyholder. The dataset initially contained 50,000 observations, with the number of claims ranging from 0 to a maximum of 16. To ensure a robust fit, the highest outlier—corresponding to one policyholder with 16 claims—was removed, resulting in a cleaned dataset of 49,999 observations, with a maximum of 4 claims. This removal was justified because a single extreme value could disproportionately influence the distribution fit, especially for a dataset where most policyholders have few or no claims, as is typical in insurance data.

A histogram of the cleaned Number of Claims (Figure 30) revealed a right-skewed distribution, with the majority of policyholders reporting zero claims, steadily decreasing until the maximum of four claims (2 observations). This pattern suggested a count distribution, such as the Poisson or Negative Binomial, which are commonly used for claim frequency in insurance. The Negative Binomial distribution was selected due to its ability to model overdispersion, where the variance (0.059) exceeds the mean (0.054), a common feature in claim count data due to heterogeneity in policyholder risk.

To test the suitability of the Negative Binomial distribution, the goodfit function was employed, using maximum likelihood estimation (MLE) (Figure 31). The fitted model (Figure 32 & Table 1) closely matched the observed frequencies, with the Pearson residuals ranging from -0.24 to 0.78, indicating a good fit. A likelihood ratio test (Table 2) a p-value of 0.622, failing to reject the hypothesis that the data follows a Negative Binomial distribution at the 5% significance level, as the p-value exceeds 0.05. Parameter estimates were obtained using the fitdist function. The Negative Binomial distribution was parameterized by size (dispersion parameter of 0.525) and probability of success (0.054) (Table 3). These results indicate that the Negative Binomial distribution effectively captures the frequency of claims, providing a reliable model for estimating expected claim counts in pricing models.

CLAIM SEVERITY

Claim Severity, represented by the cost variable in the dataset, measures the monetary amount of individual claims. The goal was to fit a distribution from the Exponential Family, such as the Gamma distribution, which is widely used for modeling positive, right-skewed claim amounts in insurance. Initial analysis using the gamma_test function rejected the Gamma distribution for the full dataset, with a p-value less than 2.2e-16, indicating significant deviation from a Gamma distribution at the 5% significance level. This rejection was likely due to extreme high-cost claims, which are common in insurance data and can distort distribution fits.

To address this, an upper threshold was defined to exclude large claims, enabling a Gamma distribution fit. Thresholds were tested at quantiles from 90% to 99% in 0.5% increments, and the Gamma test was applied to the data below each threshold (Figure 33). The 96% percentile yielded the highest p-value of 0.5937, indicating the best fit for a Gamma distribution. This threshold, corresponding to a claim cost of 5,915.54, was selected, resulting in the exclusion of 77 claims (approximately 4% of the dataset). These excluded claims were deemed large claims, likely representing severe accidents or high-value damages, which require separate modeling for pricing, as will be discussed later.

The filtered dataset's Gamma fit was retested, yielding a p-value of 0.5937, supporting the hypothesis that the data now follows a Gamma distribution at the 5% significance level. Descriptive statistics showed a minimum cost of 25.59 and a maximum of 5,870.78, a mean of 1,235.36, a standard deviation of 982.32, a skewness of 2.03, and a kurtosis of 8.27, confirming the right-skewed, heavy-tailed nature suitable for a Gamma distribution. A boxplot of the filtered costs (Figure 35) visualized the distribution's spread and central tendency.

The Gamma distribution was fitted using fitdist with MLE, yielding a shape parameter of 1.569 and a rate parameter of 0.00127 (Table 4). These parameters indicate a moderately skewed distribution, consistent with typical claim severity profiles. Visualizations included a histogram, density plot, Q-Q plot, and P-P plot from fit.gamma (Figure 36), confirming the Gamma distribution's adequacy. The fitted density closely followed the observed data, reinforcing the model's appropriateness for non-large claims. Finally, the Cullen and Frey graph (Figure 34) showed the dataset's distribution compared to typical distributions, further confirming the similarity to the Gamma Distribution.

The 77 claims exceeding the 96% percentile threshold were analyzed to understand their characteristics and inform pricing strategies. These large claims had a mean cost of 13,503.23 and a standard deviation of 11,441.70, indicating high variability and significantly larger amounts than the filtered dataset. A histogram and boxplot showed the distribution of these large claims, revealing a right-skewed pattern with a concentration of extremely high values, suggesting that a small subset of claims drives the high mean and variability. These large claims likely represent catastrophic events, such as total vehicle losses or severe accidents, which are rare but financially impactful. Their high standard deviation reflects the unpredictability of such events, complicating risk assessment.

For pricing, insurers might account for these claims by modeling them separately, potentially using a different distribution or incorporating them into a layered pricing structure. For instance, a base premium could cover typical claims, while a surcharge or reinsurance layer could address large claims, ensuring financial stability. Alternatively, predictive models, such as those developed in other project sections, could estimate the probability of large claims to adjust premiums dynamically.

NOVA IMS Information Management School

PART 2

PRICING STRUCTURE PROPOSAL

1. **MODELING CLAIM FREQUENCY WITH GLM**
2. **MODELING CLAIM SEVERITY FOR COMMON CLAIMS WITH GLM**
3. **PROPOSING A PRICING STRUCTURE FOR COMMON CLAIMS**
4. **MODELING AND PRICING FOR LARGE CLAIMS**

2. PRICING STRUCTURE PROPOSAL

2.1 MODELLING CLAIM FREQUENCY WITH GLM

To maintain consistency with the exploratory analysis, we applied the same binning strategy for both driver age and vehicle age in the modeling phase. These variables were categorized into risk-relevant intervals to improve interpretability and capture potential non-linear effects.

We began by building a baseline model that includes all available predictors: driver age ("age_bin"), vehicle age ("veh_age_bin"), geographical zone ("zone"), vehicle brand ("brand"), power ("power"), and fuel type ("fuel"), using the number of claims ("nclaims") as the target variable. Prior to inclusion, we verified that driver age and vehicle age were not strongly correlated (correlation ≈ -0.05), supporting the decision to retain both without risk of multicollinearity.

Given the observed distribution of the claim counts, we selected a Negative Binomial model. Exposure was included as an offset on the logarithmic scale to normalize claim counts by the time each policy was at risk.

MODEL IMPROVEMENT

After estimating the full baseline model, we proceeded to evaluate the contribution of each predictor using statistical testing. The objective was to simplify the model where possible, by grouping similar factor levels or dropping non-informative variables, without compromising its predictive quality.
We analyzed each input variable sequentially:

- **Driver Age**: All age bins were statistically significant at the 95% confidence level in the full model, indicating that each segment contributes meaningfully to explaining claim frequency. As a result, we retained the original bin structure without modification.
- **Vehicle Age**: In contrast, none of the vehicle age bins showed statistical significance. Given the lack of explanatory power, we excluded this variable from the model to reduce complexity and avoid potential noise.
- **Zone**: To refine this categorical variable, we applied pairwise comparison tests using the General Linear Hypothesis Testing (GLHT) framework. The results suggested that certain zone levels, specifically A with B, and E with F, had similar effects on the response variable and could be grouped without loss of information. These groupings were implemented, and the updated model was tested against the original using a likelihood ratio test (ANOVA). With a p-value of **0.07**, we failed to reject the null hypothesis that the simplified model performs equivalently to the full model, supporting the validity of the level grouping.

At this stage, we re-tested the grouped zone variable to confirm no further combinations were statistically justified.

We extended the stepwise model simplification process to the remaining categorical variables—brand, power, and fuel:

- **Brand**: We iteratively applied pairwise tests to identify levels with similar effects. The initial results indicated that several brand groups could be merged. Groupings were formed based on statistical similarity (brands 1–4–6–14 and 3–5–10–11–13). After adjusting the model accordingly, an ANOVA test yielded a high p-value (0.94), indicating that the grouped model did not significantly differ from the full one. Further grouping (adding brand 2 to the first merged group) was also tested and accepted, with a p-value of 0.85. No additional level reductions were supported by the simultaneous tests.
- **Power**: Although pairwise comparisons showed little evidence of substantial differences across most power levels, aside from level 4, we grouped all others accordingly and retained level 4 separately. The updated model yielded a p-value of 0.64 in the ANOVA test, indicating that the simplified version performed equivalently to the full model. As a result, the grouped structure was accepted.
- **Fuel**: This variable is binary and statistically significant in the baseline model, so it was retained without modification.

Through this targeted evaluation, we ensured that each variable was retained in its most efficient form—either fully preserved, partially grouped, or excluded—based on both statistical rigor and model performance criteria.

FINAL MODEL EVALUATION

After refining the categorical predictors through grouping and selection, we constructed the final claim frequency model using the simplified set of variables: age bin, zone (grouped), brand (grouped), power (grouped), and fuel. The model retained the Negative Binomial structure and incorporated exposure as a log offset to account for varying policy durations.

To assess the quality of this final model, we compared it to the original full model using two standard goodness-of-fit metrics:

- **Deviance**: A measure of model error analogous to the residual sum of squares in linear models. The final model achieved a slightly lower deviance (12922.62) than the full model (12923.76), indicating a marginal improvement in fit.
- **AIC**: A penalized likelihood criterion that balances model fit with complexity. The final model produced a lower AIC (20397.42) compared to the full model (20420.05), confirming that it achieves a better trade-off between accuracy and complexity.

These results support the conclusion that the simplified model not only maintains the explanatory power of the full model but also benefits from reduced complexity, making it more interpretable and generalizable.

STANDARD INSURER

To define the Standard Insured, used as the model's reference category, we selected the combination of risk characteristics associated with the most statistical information and the narrowest confidence intervals (Figure 38).
Based on the claim frequency plots, the following profile was identified as the Standard Insured:

- **Zone**: A–B
- **Brand**: Group 1
- **Power**: Group 5
- **Age**: [44, 64[
- **Fuel**: E

These levels were defined as the model's reference categories. The expected claim frequency for this baseline profile corresponds to the model's intercept. Based on the final model, the estimated annual claim frequency for the Standard Insured is **0.068**. This means that, on average, a policyholder with one year of exposure and all baseline characteristics is expected to file approximately 0.068 claims per year.

HIGHEST AND LOWEST RISK PROFILES

To better understand the range of risk in the portfolio, we identified the insured profiles with the highest and lowest predicted claim frequencies.

We generated all possible combinations of the categorical variables used in the final model, assuming one year of exposure. Using the trained model, we then predicted the annual claim frequency for each profile:

- The highest risk profile corresponds to:
  - **Age**: [18, 24[
  - **Zone**: E
  - **Brand**: Group 3
  - **Power**: Group 5
  - **Fuel**: D
  - **Predicted claim frequency**: 0.3903
- The lowest risk profile corresponds to:
  - **Age**: [64, 100]
  - **Zone**: A
  - **Brand**: Group 12
  - **Power**: Group 4
  - **Fuel**: E
  - **Predicted claim frequency**: 0.0462

These results illustrate the significant variation in expected claim frequency across risk profiles, driven primarily by the joint effects of age, zone, and brand. Younger drivers, in particular, show a markedly higher risk, while older drivers in lower-risk zones and conservative brands exhibit the lowest expected claim frequencies.

2.2 MODELLING CLAIM SEVERITY WITH GLM

To model the severity of claims, we focused on a subset of the data consisting of common claims, which are those not classified as extreme or outliers. Given the continuous and right-skewed nature of claim costs, a Gamma distribution with a log link was selected.

We then used the same baseline levels as in the claim frequency model to ensure consistency in interpretation across both models. The initial model included all predictors: **zone**, **power**, **vehicle age** (binned), **driver age** (binned), **brand**, and **fuel**.

MODEL IMPROVEMENT

After fitting the model, we examined the significance of each variable. The vehicle age variable showed no statistical significance at the 95% confidence level and was therefore excluded. This decision also aligns the structure of the severity model with that of the frequency model, aiding coherence in the downstream pricing model.

To confirm that this simplification did not significantly degrade model performance, we conducted an ANOVA test between the full and reduced models. The resulting p-value of **0.51** indicates no significant difference, justifying the exclusion of the vehicle age variable.

FINAL MODEL EVALUATION

To assess the quality of the final claim severity model, we relied on the same two complementary metrics:

- **Deviance**: The deviance for the final model (1267.83) was slightly higher than that of the more complex full model (1262.75), indicating a marginal loss in fit. However, the difference was small and considered acceptable given the gains in simplicity and interpretability.
- **AIC**: Although the final model had a slightly worse deviance, its AIC was lower (29525.87) than that of the full model (29535.71). This suggests that the reduction in complexity outweighed the slight decrease in fit, making the final model preferable from a model selection perspective.
- **Residual Analysis**: We visually inspected the distribution of deviance residuals for both models using histograms and diagnostic plots. No concerning patterns or extreme outliers were observed, confirming that the final model provides a statistically sound and stable fit to the data.

In conclusion, the final severity model achieves a balance between interpretability and predictive performance.

STANDARD INSURER

To ensure consistency across models, the Standard Insured profile defined for claim frequency was also used as the reference baseline in the claim severity model. This profile corresponds to an individual with the following characteristics:

- **Zone**: A–B
- **Brand**: Group 1
- **Power**: Group 5
- **Age**: [44, 64[
- **Fuel**: E
-

As with frequency modeling, the expected claim severity for the Standard Insured is derived from the intercept of the fitted Gamma model. Based on the model output, the estimated average claim cost for a common claim filed by this baseline profile is **€1191.51**.

This serves as a benchmark for comparing expected severity across other insured profiles with different characteristics, as we will do next.

HIGHEST AND LOWEST RISK PROFILES

To assess the range of expected claim costs across the portfolio, we identified the insured profiles associated with the highest and lowest predicted claim severity. This was done by computing predictions for all possible combinations of risk factor levels included in the final Gamma model.

- The highest severity risk profile corresponds to:
  - **Age**: [18, 24[
  - **Zone**: D
  - **Brand**: Group 12
  - **Power**: Group 4
  - **Fuel**: D
  - **Predicted severity**: €1,966.79
- The lowest severity risk profile corresponds to:
  - **Age**: [24, 34[
  - **Zone**: C
  - **Brand**: Group 3
  - **Power**: Group 5
  - **Fuel**: E
  - **Predicted severity**: €1,041.43

These results illustrate that expected claim costs can vary substantially depending on the policyholder's characteristics. In this case, younger drivers with less favorable combinations of vehicle and fuel attributes are associated with significantly higher claim costs, while moderate-age drivers with more standard configurations tend to incur lower severity claims.

2.3 PROPOSING A PRICING STRUCTURE FOR COMMON CLAIMS

Based on the previously developed risk models, the proposed pricing structure is tailored to common claim scenarios by incorporating key risk factors as per below:

Risk Factors	Beta_N	E(N)	beta_Y	**E(Y)**	**Pure Premium**	**Tariff**
Standard Insured	-2,6885	0,0680	7,0830	1191,5139	81,00 €	81,00 €
Age 18-24	0,8297	0,1559		1191,5139	185,70 €	2,2926
Age 24-34	0,1963	0,0827		1191,5139	98,57 €	1,2169
Age 34-44	0,0420	0,0709		1191,5139	84,48 €	1,0429
Age 64-100	-0,0887	0,0622		1191,5139	74,13 €	0,9151
Zone C	0,1640	0,0801		1191,5139	95,44 €	1,1782
Zone D	0,4038	0,1018		1191,5139	121,30 €	1,4975
Zone E	0,5738	0,1207		1191,5139	143,77 €	1,7750
Brand 3	0,1764	0,0811		1191,5139	96,63 €	1,1929
Brand 12	-0,1498	0,0585	0,1606	1399,1623	81,88 €	1,0109
Brand 4	-0,1480	0,0586		1191,5139	69,86 €	0,8625
Fuel D	0,1677	0,0804		1191,5139	95,79 €	1,1826

Highest risk profile

- Total premium: €465.02
  - Age 18-24
  - Zone E-F
  - Vehicle brand group 3
  - Standard Insurer Power
  - Fuel Type D

Lowest risk profile

- Total premium: €63.92
  - Age 64-100
  - Standard Zone A-B
  - Standard vehicle brand group 1
  - Power Group 4
  - Standard Fuel Type E

The proposed premium structure adheres to the principle of actuarial fairness, meaning that each policyholder contributes a premium proportionate to their individual risk. High risk profiles are charged higher premiums, reflecting their elevated claim frequency and severity. Conversevely, low-risk profiles benefit from reduced premiums. This ensures a transparent and justifiable pricing model.

Premium adequacy is achieved by grounding the structure in expected claim cossts (pure premium), ensuring that the collected premiums are sufficient to cover antecipated losses. This strengthens the insurer's financial sustainability and helps maintain solvency. The use of both frequency and severity components in premium calculation improves the model's responsiveness to underlying risk.

While statistically sound, the premium structure may rise practical concerns, particularly regarding affordability for high-risk groups - such as young drivers who face premiums over €465.

Another important limitation to consider is that the current model was developed excluding large claims, meaning that extreme losses are not reflected in the premium structure. As a result, if this model were implemented in practice, the insurer could face significant financial vulnerability when confronted with high-severity claims.
Incorporating these lements would likely lead to a notable increase in the insurance standard, but its essential to maintain solvency and financial requirements.

2.4 MODELING AND PRICING FOR LARGE CLAIMS

To incorporate large claims into the car insurance pricing structure, a modeling approach was developed to estimate the probability of a large claim being reported, leveraging both traditional statistical methods and machine learning techniques. The dataset, containing 1,907 claim records, included 77 large claims (approximately 4% prevalence), as established in a previous section. This section justifies the choice of models, discusses the variables used, explains the rationale for separating large claims from common claims, and proposes a method to integrate these results into a comprehensive premium structure.

MODEL CHOICE AND JUSTIFICATION

Two types of models were developed to predict the probability of a large claim: a logistic regression model as a benchmark and a gradient boosting model using XGBoost as the machine learning approach. Logistic regression was selected as the benchmark due to its interpretability and widespread use in insurance for modeling binary outcomes, such as large claim occurrence. Its linear structure allows for straightforward interpretation of risk factors through odds ratios, facilitating communication with stakeholders. However, logistic regression assumes linear relationships between predictors and the log-odds of the outcome, which may limit its ability to capture complex patterns in the data.
To address this limitation, XGBoost was chosen as the machine learning approach. XGBoost is well-suited for imbalanced datasets, as it can model non-linear relationships and interactions among variables, potentially improving predictive performance for rare events like large claims. Its ability to handle categorical variables through one-hot encoding and incorporate exposure adjustments aligns with insurance modeling requirements. Three XGBoost variants were tested: Option 1 used all available features without preprocessing, Option 2 used only the top two features, and Option 3 used the same preprocessed features as the logistic regression. Stratified 5-fold cross-validation was employed to ensure robust performance evaluation, maintaining class balance across folds due to the low prevalence of large claims.

VARIABLES USED AND PREPROCESSING

The logistic regression model utilized the same five predictors as the previous GLMs: zone_grouped, power_grouped, age_bin, brand_grouped, and fuel, with log(exposition) as an offset to adjust for exposure, as well as veh_age_bin, which had been discarded for its statistical insignificance in previous models. Here, this variable could prove to be important in predicting the probability of large claims occurring.
For XGBoost Option 1, the raw features were used without preprocessing, leveraging one-hot encoding to handle categorical variables. Option 2 focused on agevehicle and agedriver as predictors, identified as the most important features via feature importance analysis of Option 1, to simplify the model while retaining predictive power. Option 3 mirrored the logistic regression's preprocessed variables for consistency. The exposure offset (log(exposure)) was incorporated in all models, ensuring predictions reflect risk per exposure unit, a standard practice in insurance pricing.

RATIONALE FOR SEPARATING LARGE CLAIMS

Separating large claims from common claims was motivated by their distinct characteristics and impact on pricing. Large claims had a mean cost of 13,503.23 and a standard deviation of 11,441.70, compared to a mean of 1,235.36 for common claims. Their high cost and variability, as shown in a previous section's analysis, indicate rare but severe events, which significantly affect insurer profitability. Modeling these claims separately allows for targeted risk assessment, as their predictors may differ in influence from those of common claims. This separation aligns with insurance practice, where large claims are often handled through reinsurance or special reserves, requiring precise probability estimates to allocate appropriate premiums or coverage layers.

MODEL PERFORMANCE AND EVALUATION

The logistic regression model, fitted on the full dataset, provided interpretable odds ratios (Table 5), such as a 2.12-fold increase in large claim odds for brand 12, indicating higher risk of large claims for Japanese and Korean vehicles. However, its performance was poor, with zero true positives at a 0.5 threshold, reflecting the challenge of imbalanced data. Even after threshold optimization using Youden's J statistic, the model failed to predict any large claims, likely due to its linear assumptions and lack of class weighting.

The XGBoost models were evaluated using stratified 5-fold cross-validation, with AUC as the primary metric due to its robustness for imbalanced data. Additionally, to address this imbalance, a weighting scheme was applied to penalize errors on the minority class more highly. All three options yielded mean AUCs between 0.508 and 0.550, suggesting limited discriminative ability, possibly due to insufficient feature informativeness or the extreme imbalance. Option 1 (all features) achieved a balanced accuracy of 0.5049 and an F1-score of 0.0506 at a 0.5 threshold, improving to 0.5443 and 0.0936 with an optimized threshold. Option 2 (driver age, vehicle age) performed best, with a balanced accuracy of 0.5645 and an F1-score of 0.1163 at the optimized threshold, reflecting a better balance between sensitivity and specificity. Option 3 (preprocessed features) had both a lower balanced accuracy (0.5527) and F1-score (0.0976), indicating higher false positives.

Feature importance analysis for XGBoost Option 1 (Figure 41) highlighted driver age and vehicle age as the most influential predictors, justifying the selection of Option 2 for its simplicity and comparable performance. The modest AUC and sensitivity underscore the challenge of predicting rare events, but the optimized threshold improved the model's ability to identify large claims, critical for pricing applications. These metrics reflect improved performance on the minority class compared to other options, critical for identifying large claims in an imbalanced dataset.

INTEGRATION INTO COMPREHENSIVE PRICING

To produce a comprehensive premium, the XGBoost model's predicted probabilities of large claims can be combined with the models for common claims and claim frequency. The expected claim cost per policyholder is calculated as:

E[Cost] = P(Common Claim) * E[Cost | Common Claim] +
P(Large Claim) * E[Cost | Large Claim]

- Common Claim Models: The previous section fitted a Gamma distribution to common claim severity and a Negative Binomial distribution to claim frequency. These provide **P(Common Claim)** and **E[Cost | Common Claim]**.
- Large Claim Probability: The XGBoost model provides **P(Large Claim)**, adjusted for exposure via the offset. For a policyholder, the predicted probability is obtained by applying the model on their feature values (driver age, vehicle age).
- Expected Large Claim Cost: The severity of large claims was modeled using a lognormal distribution, with parameters meanlog = 9.3089 and sdlog = 0.5705, yielding an expected value of 12,986.05 (**E[Cost | Large Claim]**). This distribution was selected based on goodness-of-fit statistics, outperforming Gamma, Weibull, and Pareto distributions (Table 14 & Figure 46). To visually compare these fits, CDF and QQ plots were also created (Figures 47 & 48)
- Implementation and Premium Calculation: A final XGBoost model, trained on the full dataset with driver age and vehicle age, can be deployed to predict large claim probabilities for new policyholders. These probabilities are then multiplied by the expected large claim cost (12,986.05) and combined with common claim costs, adjusted for exposure, to set the total premium.

This approach ensures that the high cost of large claims is explicitly accounted for, preventing underpricing of high-risk policyholders. The model's simplicity (using only two features) facilitates integration into pricing systems, while its exposure adjustment ensures fairness across policy durations. By integrating the model's probability estimates with the lognormal expected cost of large claims and models for common claims, a comprehensive premium can be calculated, balancing risk and affordability for insurers. Future improvements could involve hyperparameter tuning or additional features to enhance AUC and sensitivity, ensuring even more accurate pricing for rare but costly large claims.

ANNEX

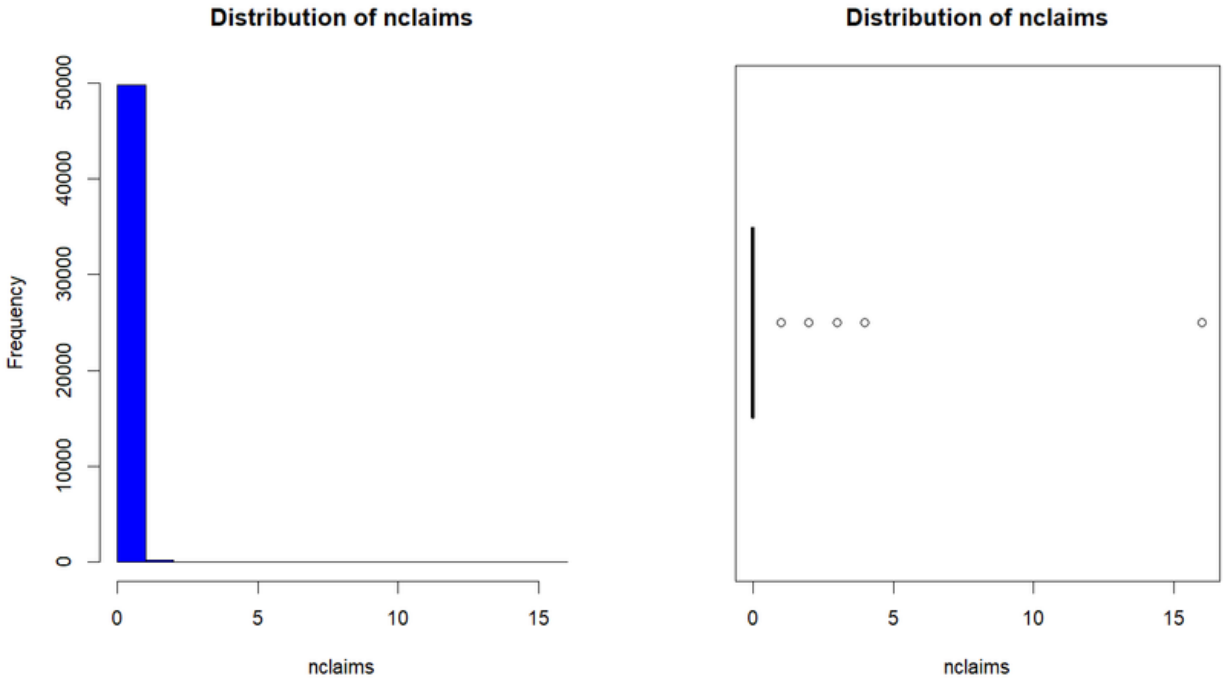


Figure 1 - Histogram and Boxplot of nclaims

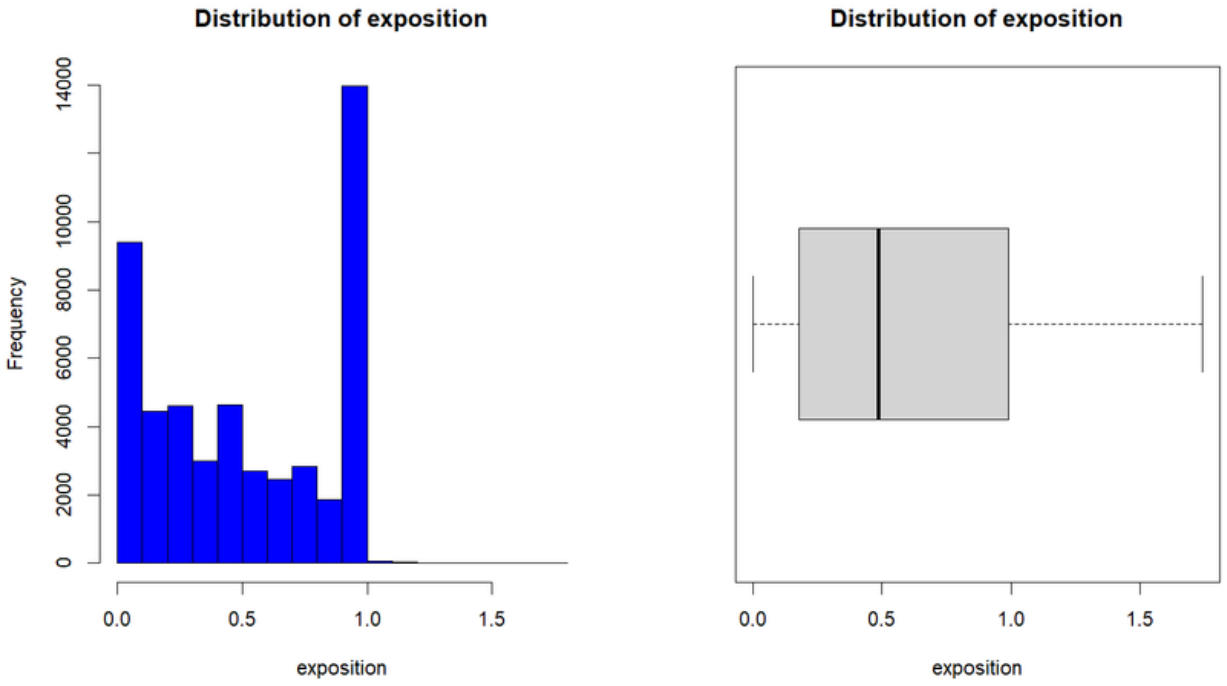


Figure 2 - Histogram and Boxplot of exposition

ANNEX

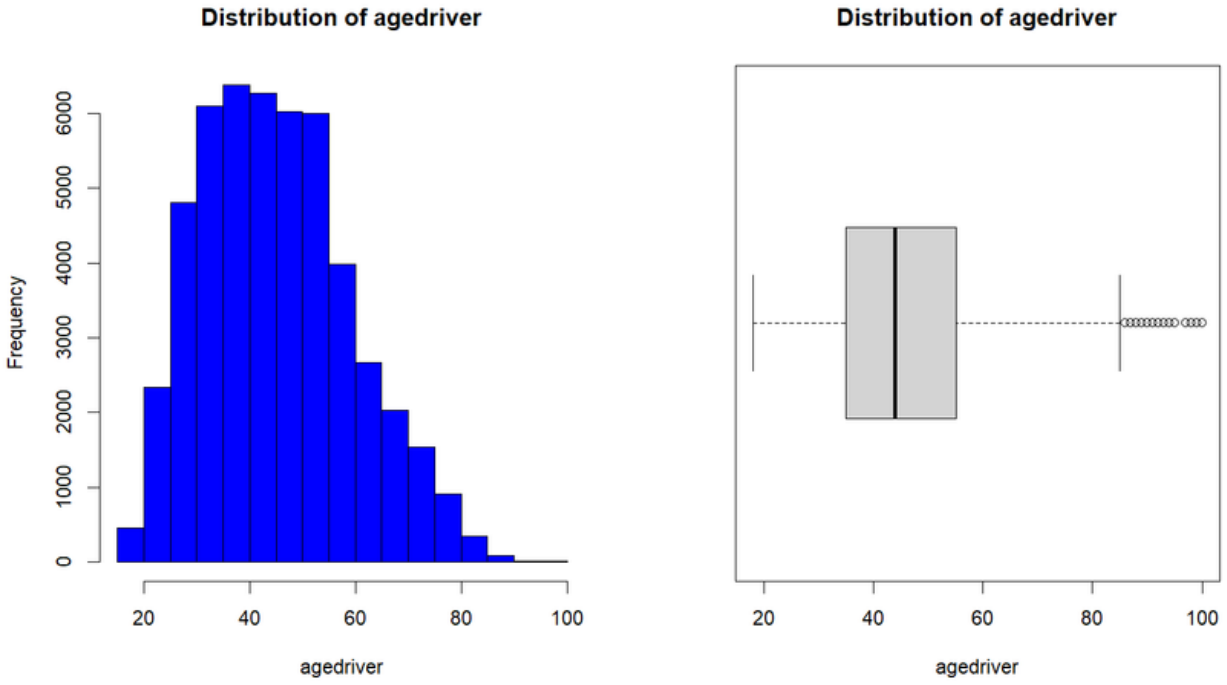


Figure 3 - Histogram and Boxplot of agedriver

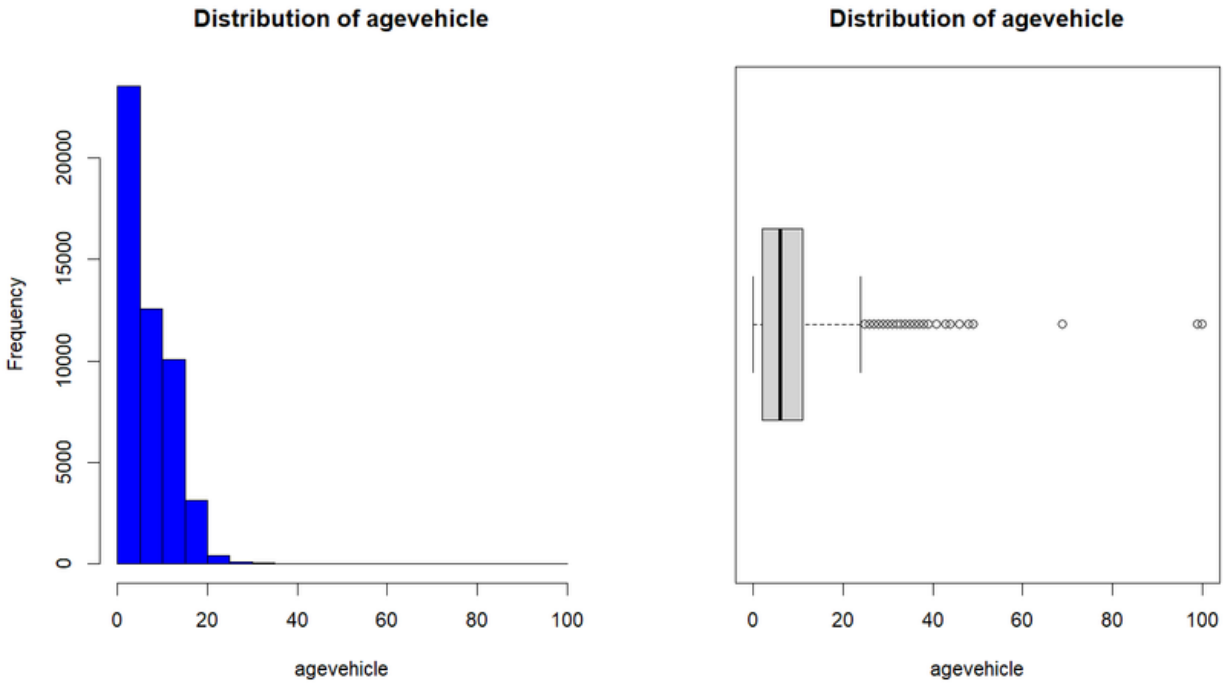


Figure 4 - Histogram and Boxplot of exposition

ANNEX

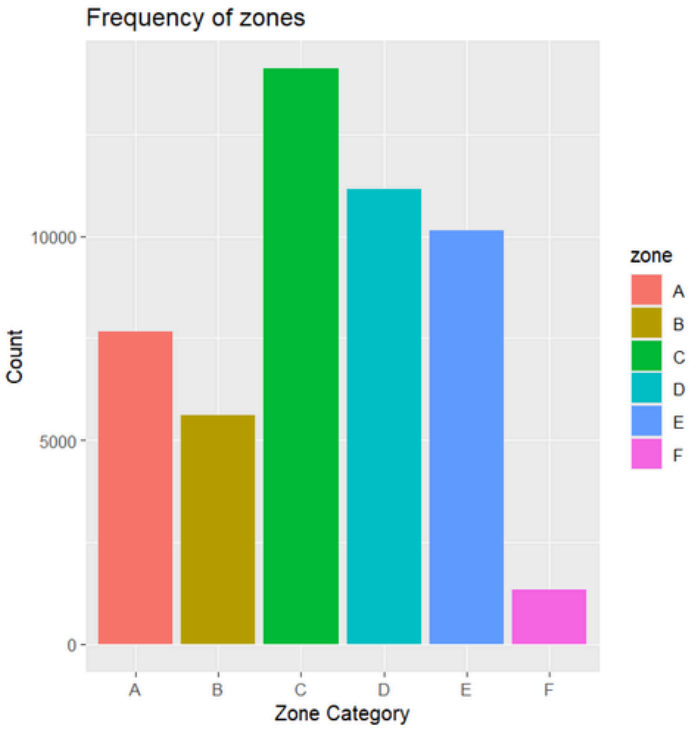


Figure 5 - Barchart of Zones Frequency

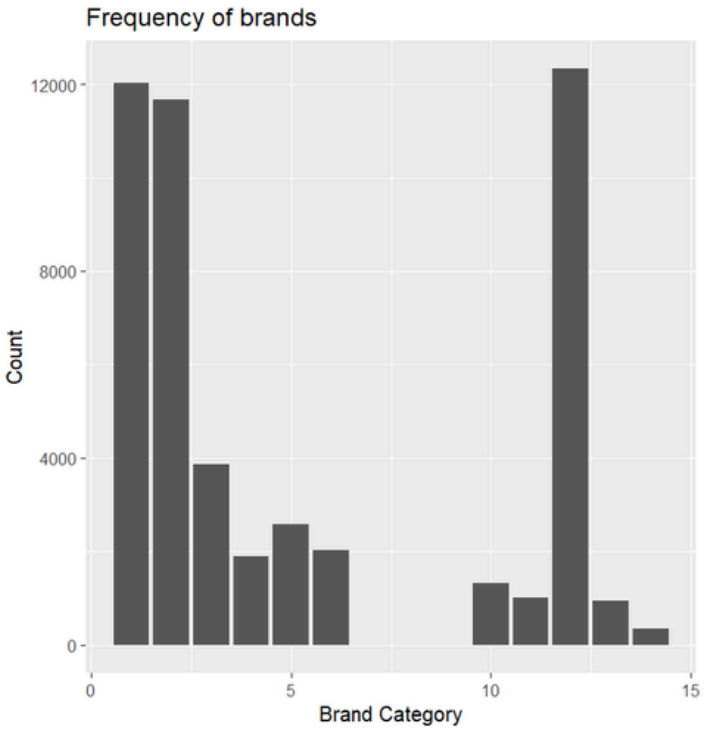


Figure 6 - Barchart of Brands Frequency

ANNEX

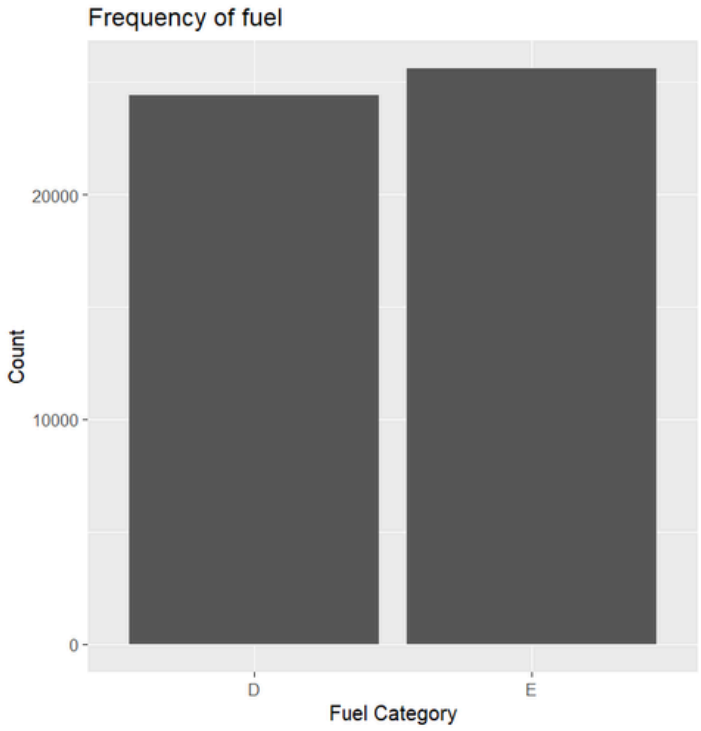


Figure 7 - Barchart of Fuel Frequency

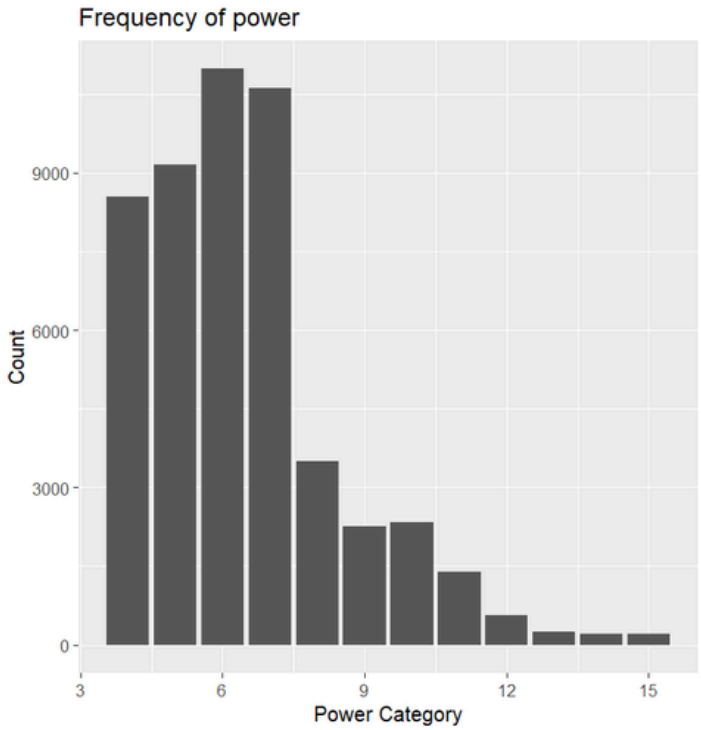


Figure 8 - Barchart of Power Frequency

ANNEX

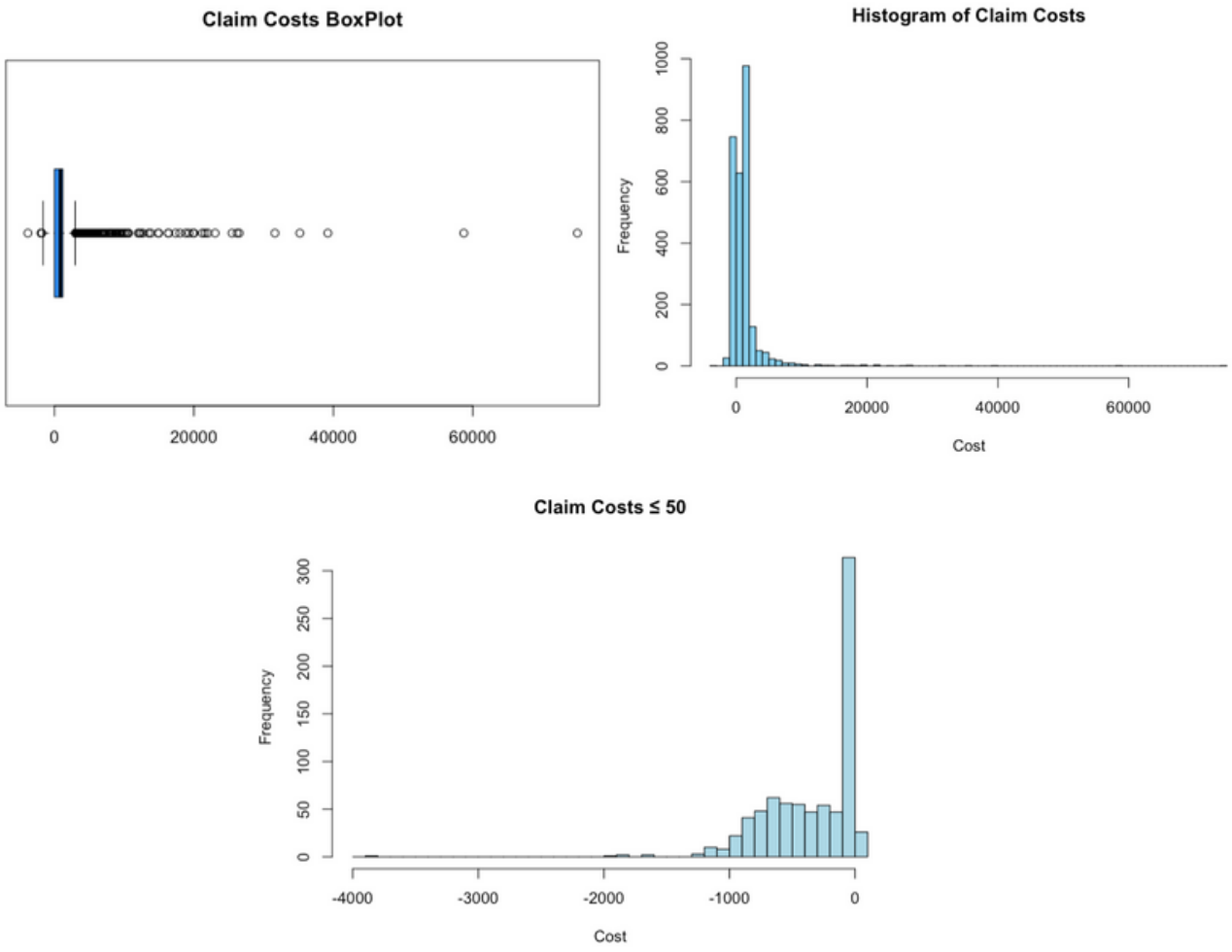


Figure 9 - Histogram and Boxplot of claim costs (with <25 included)

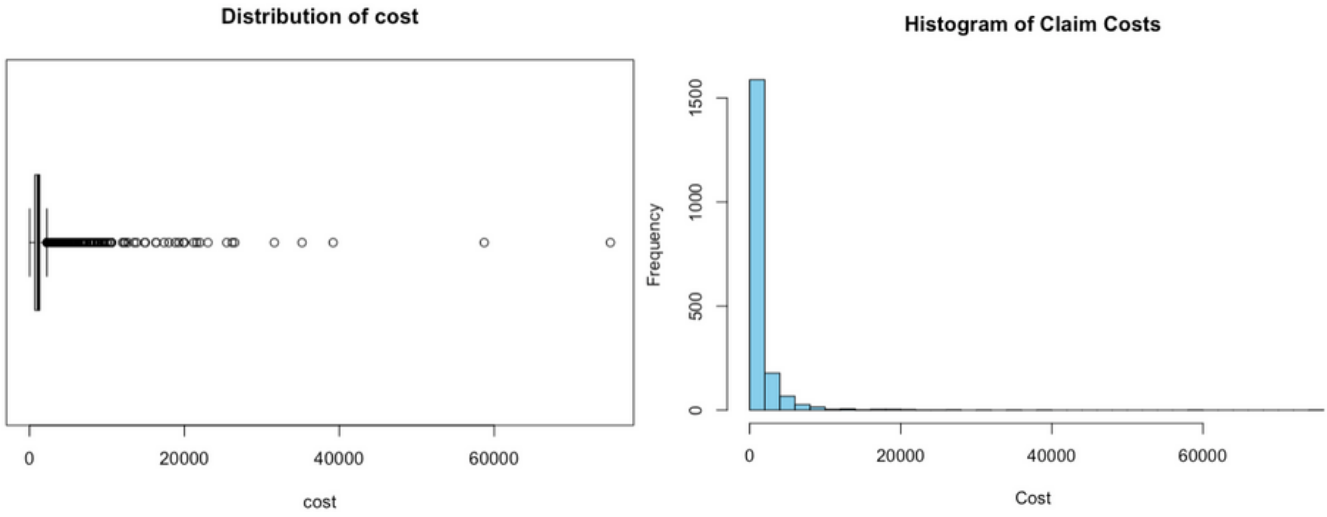


Figure 10 - Histogram and Boxplot of claim cost (with <25 removed)

ANNEX

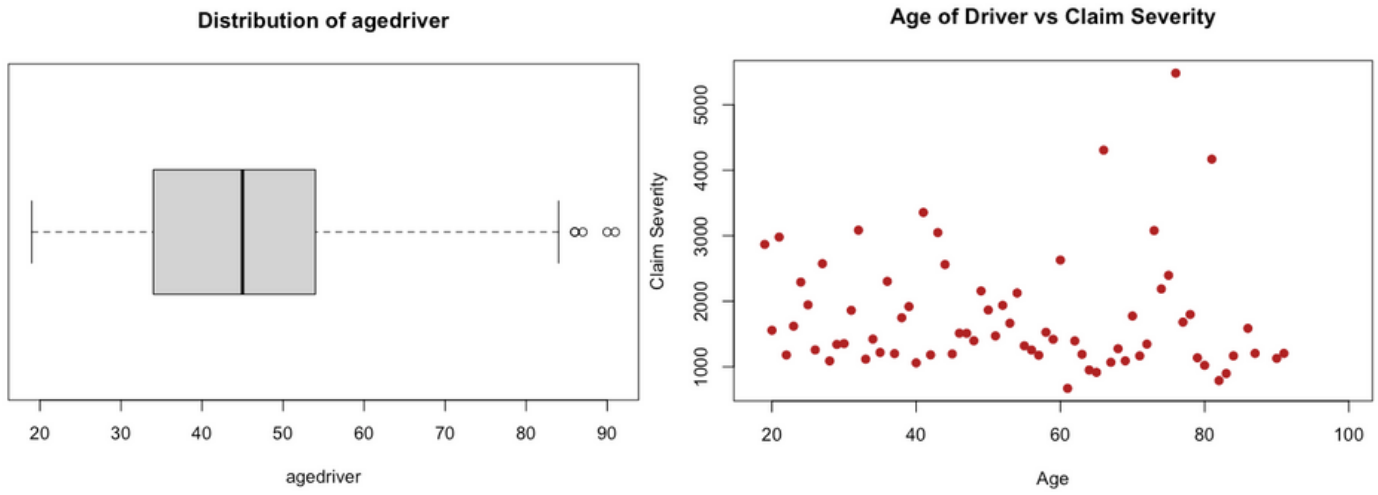


Figure 11 - Boxplot and Scatter plot of Age drive and Age Driver with Claim Severity

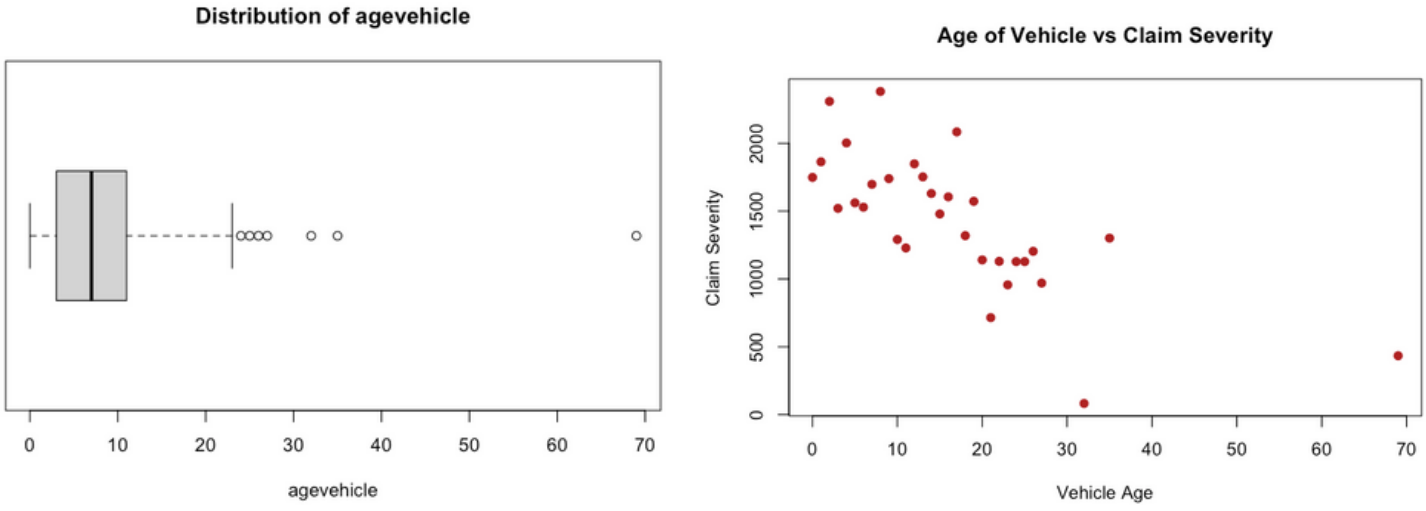


Figure 12 - Boxplot and Scatter plot of Vehicle age and Vehicle age with Claim Severity

ANNEX

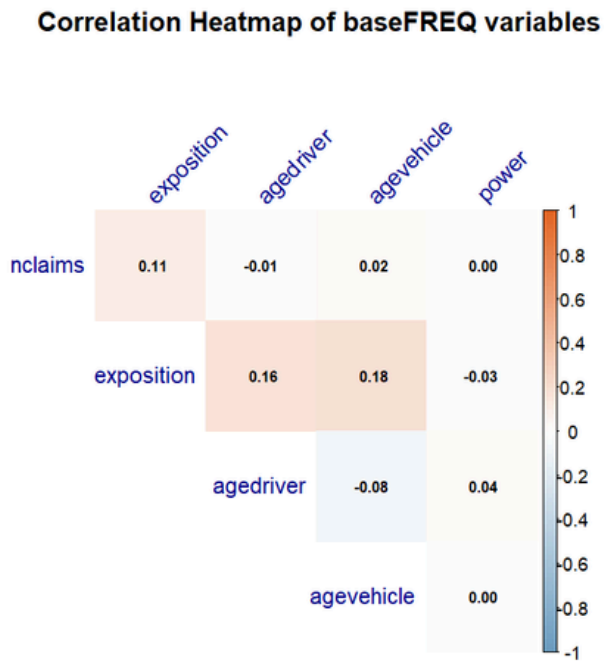


Figure 13 - Numerical Variables Correlation Heatmap

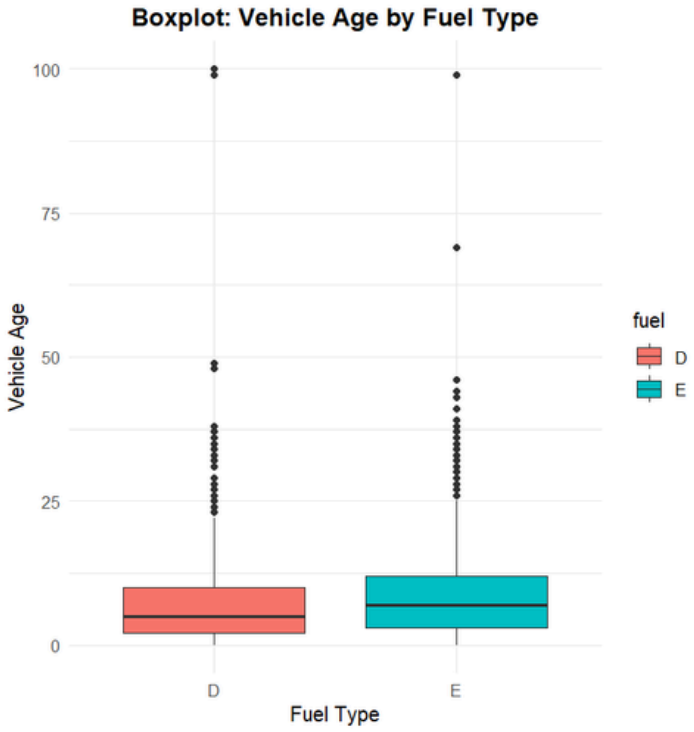


Figure 14 - Vehicle Age by Fuel Type Boxplot

ANNEX

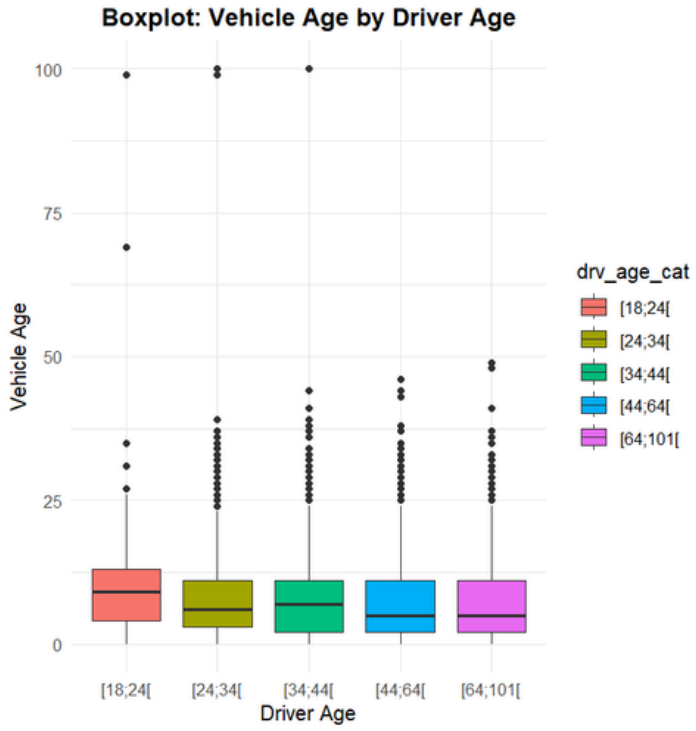


Figure 15 - Vehicle Age by Driver Age Boxplot

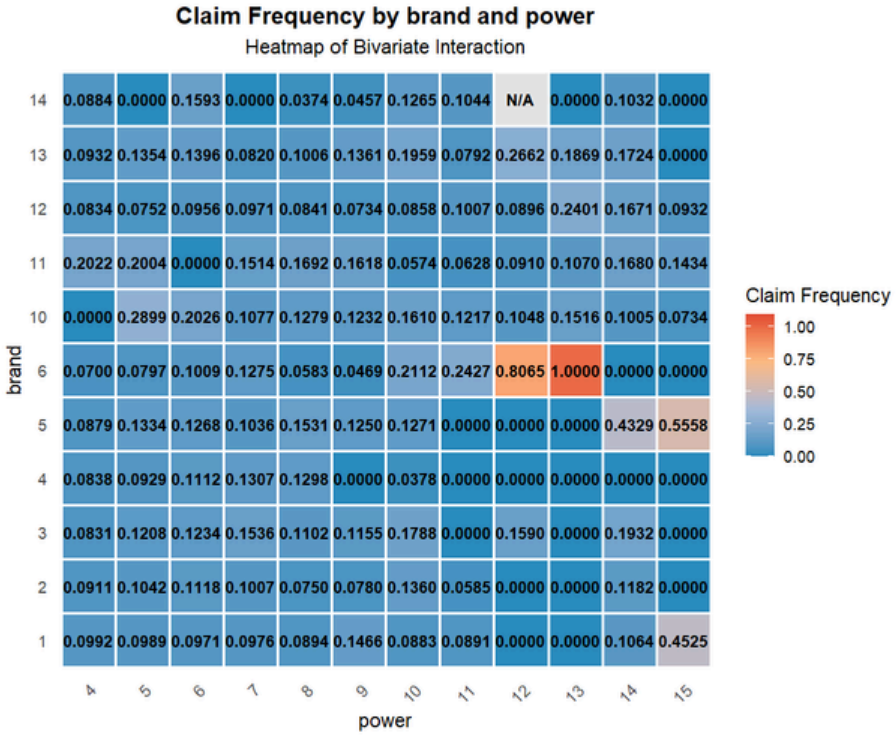


Figure 16 - Claim Frequency by Power and Brand Heatmap

ANNEX

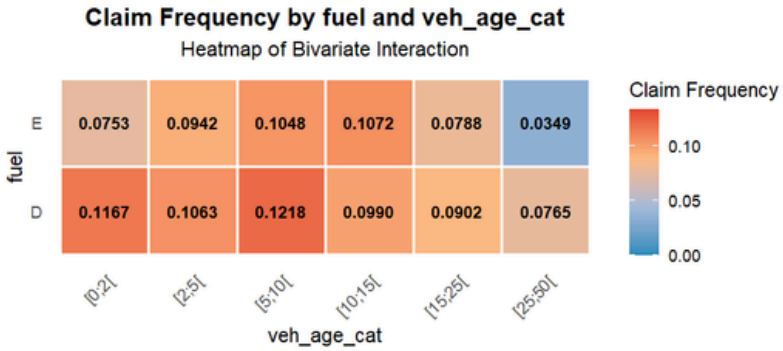


Figure 17 - Claim Frequency by Fuel and Vehicle Age Bins Heatmap

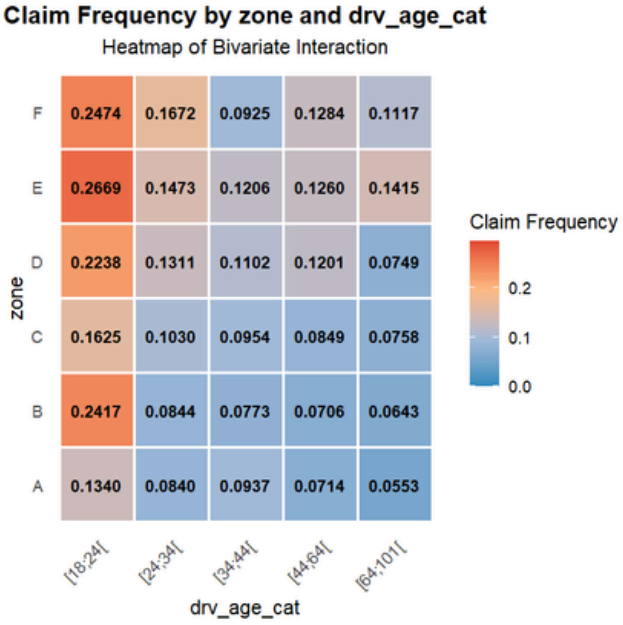


Figure 18 - Claim Frequency by Zone and Driver Age Heatmap

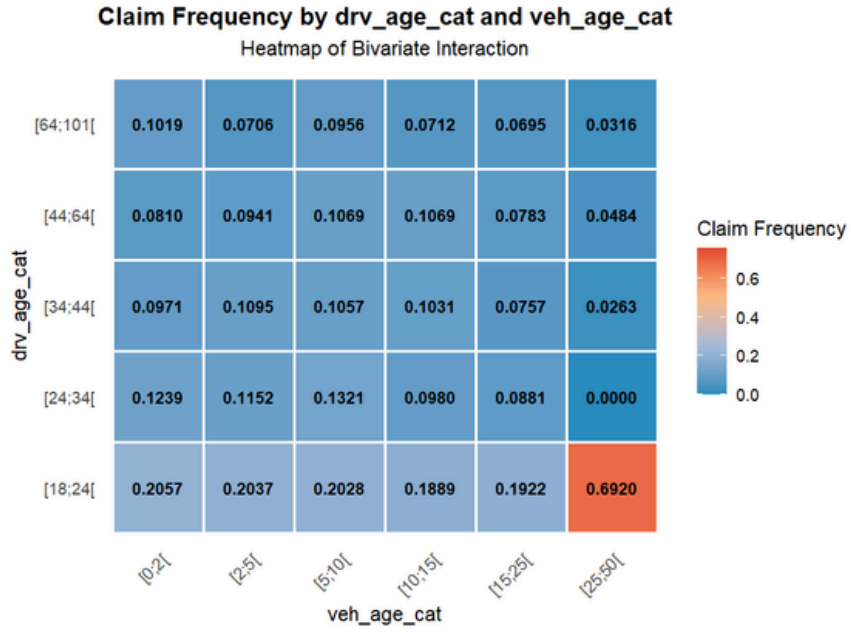


Figure 19 - Claim Frequency by Driver Age vs Vehicle Age Heatmap

ANNEX

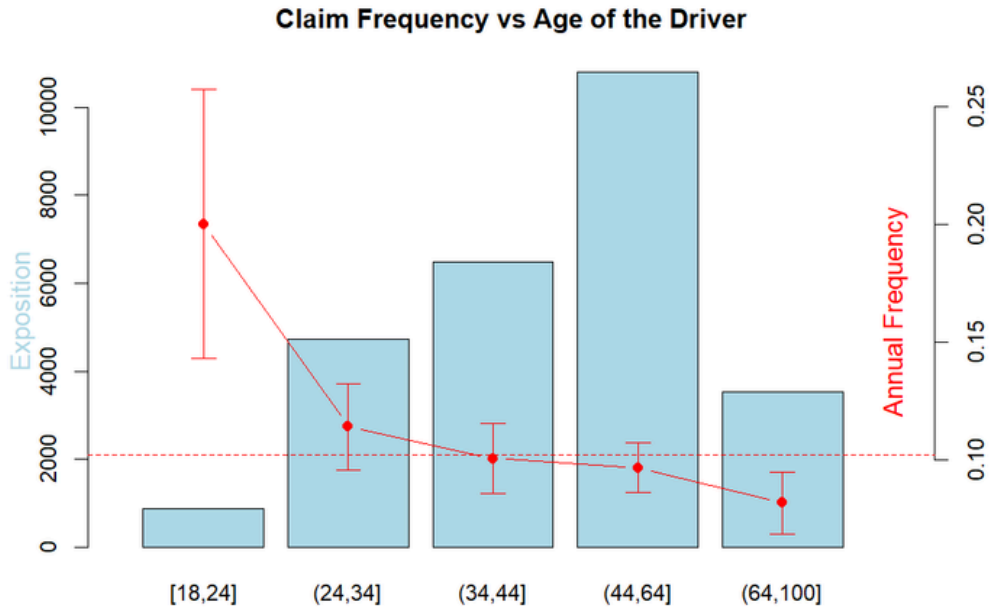


Figure 20 - Claim Frequency vs Age of Driver

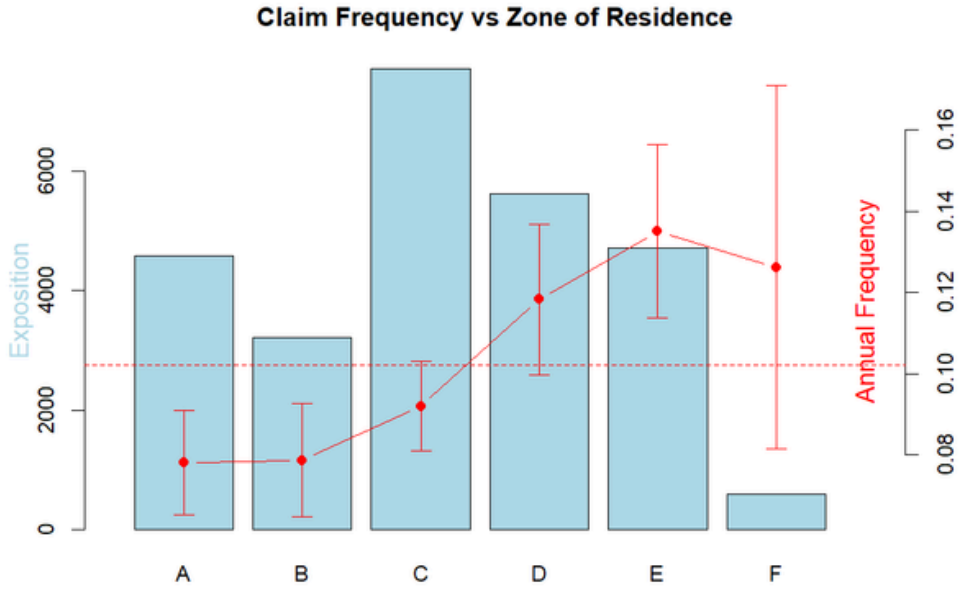


Figure 21 - Claim Frequency vs Zone of Residence

ANNEX

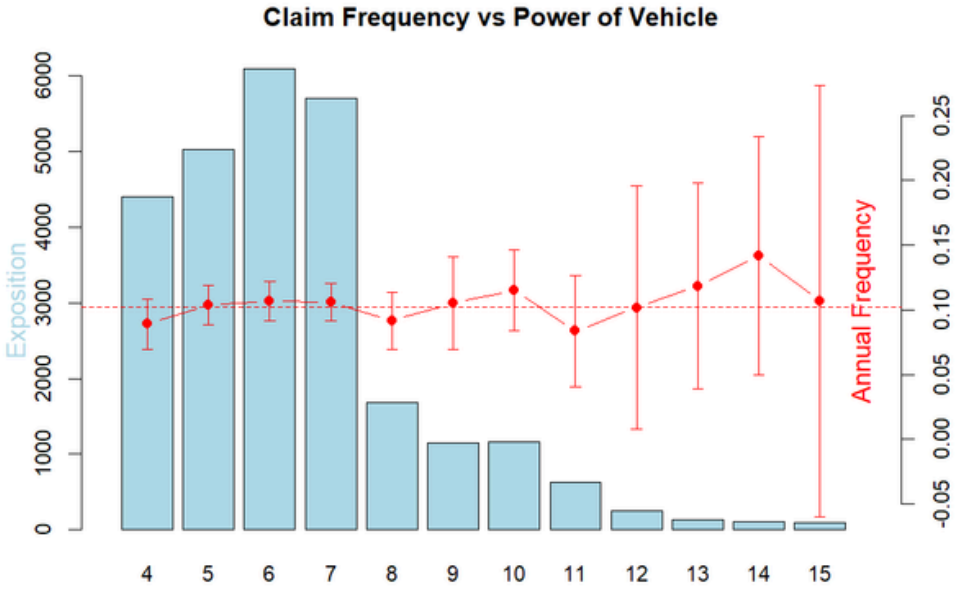


Figure 22 - Claim Frequency vs Power

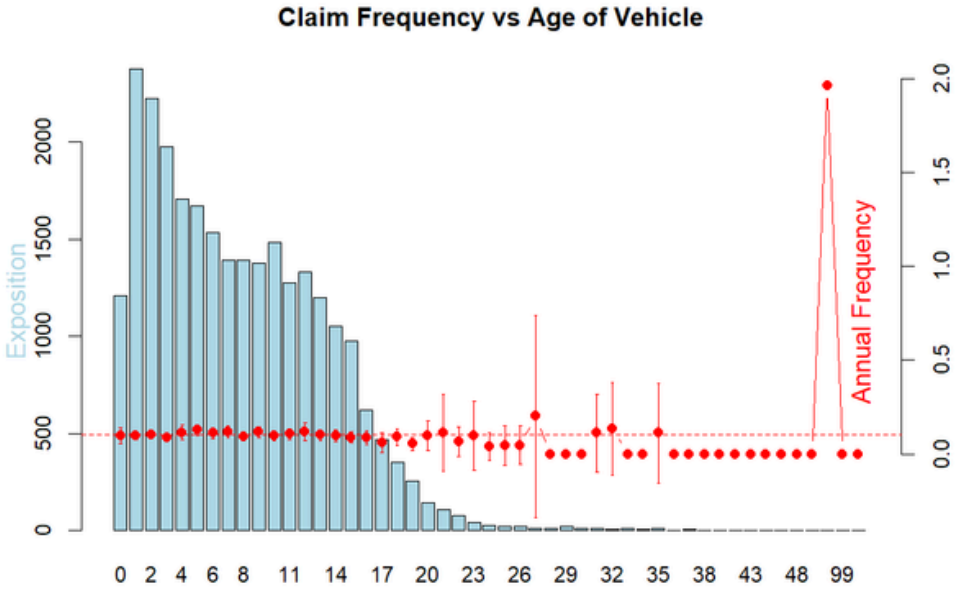


Figure 23 - Claim Frequency vs Age of Vehicle

ANNEX

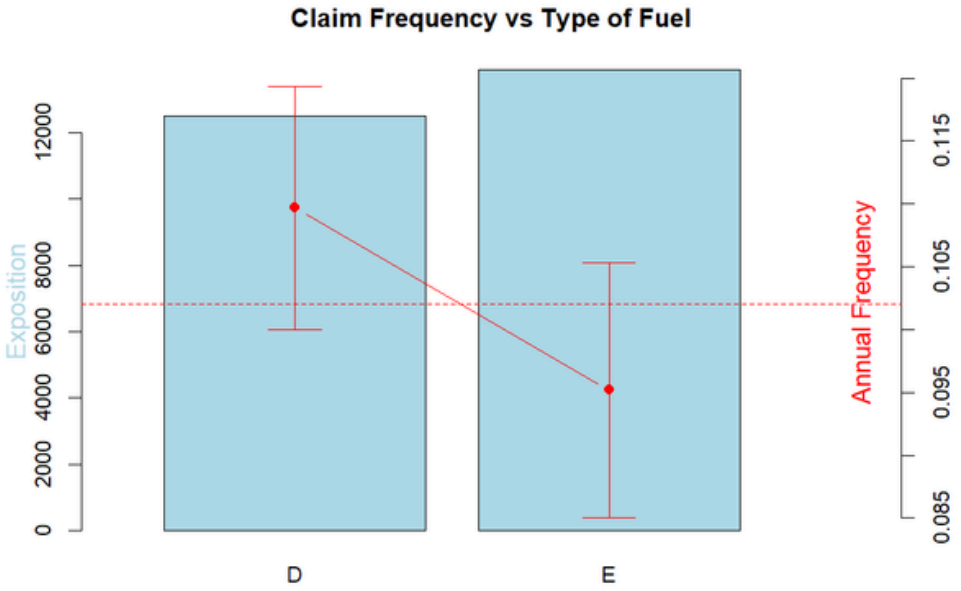


Figure 24 - Claim Frequency vs Age of Driver

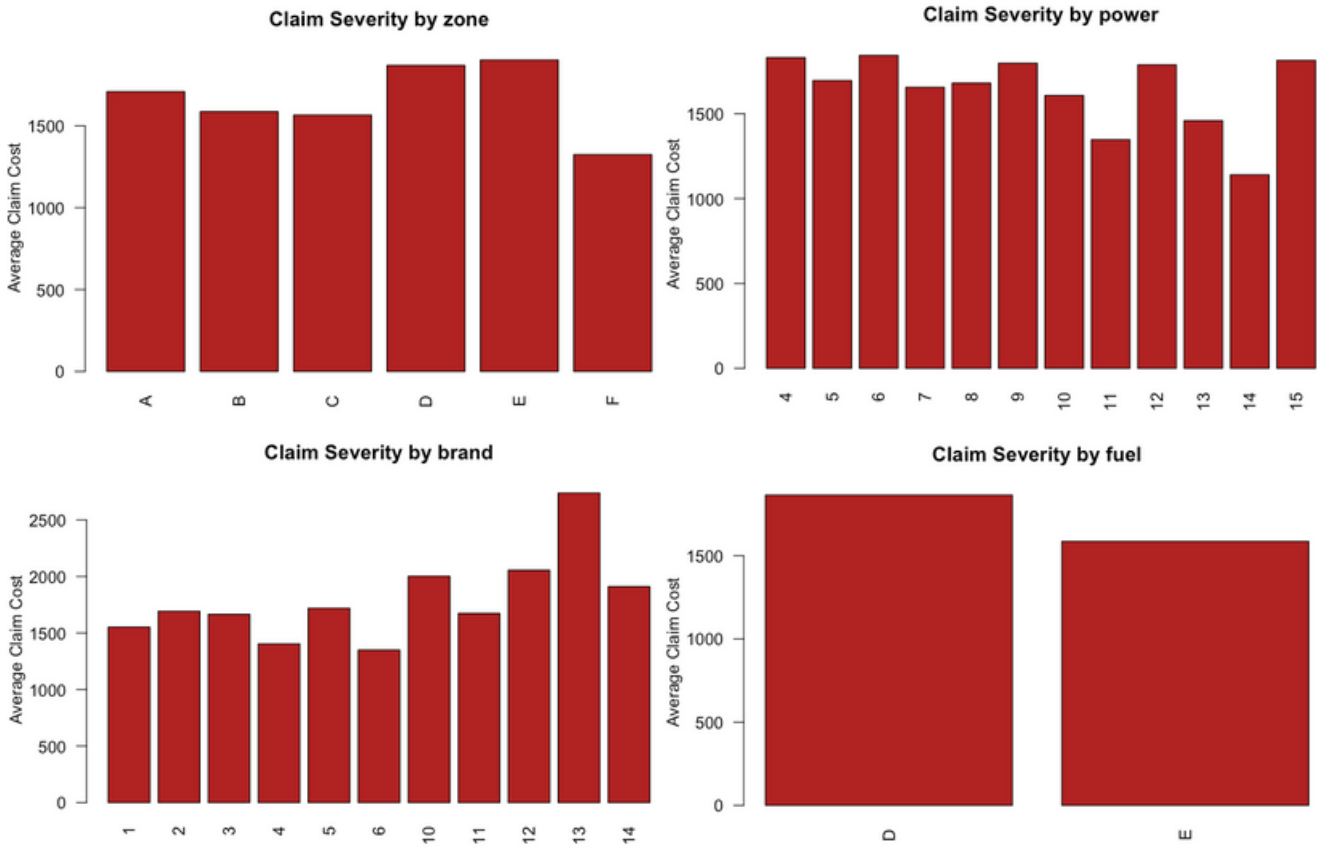


Figure 25 - Bar chart of Categorical Features with Claim Severity

ANNEX

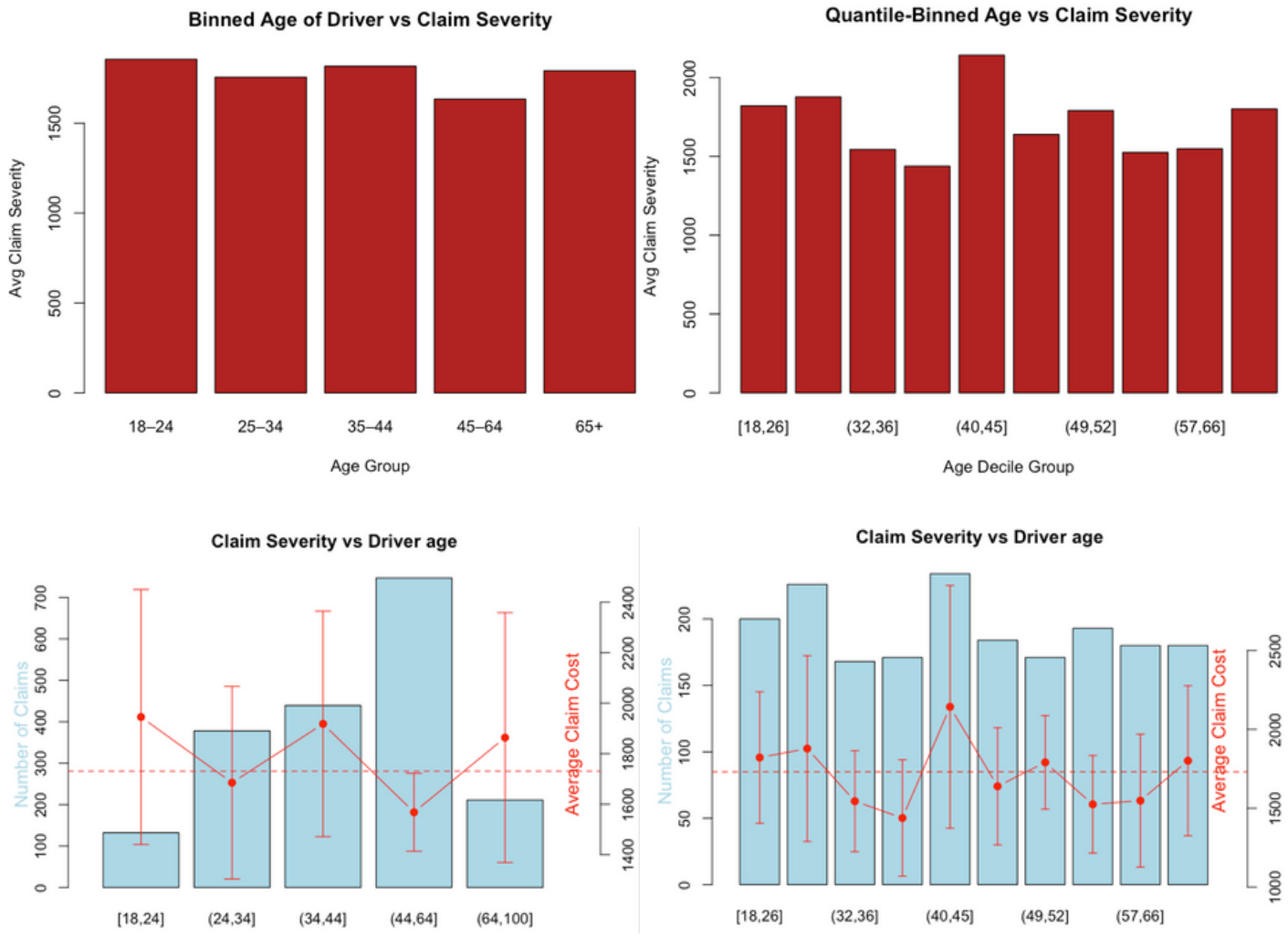


Figure 26 - Graphics Driver age vs Claim Severity

ANNEX

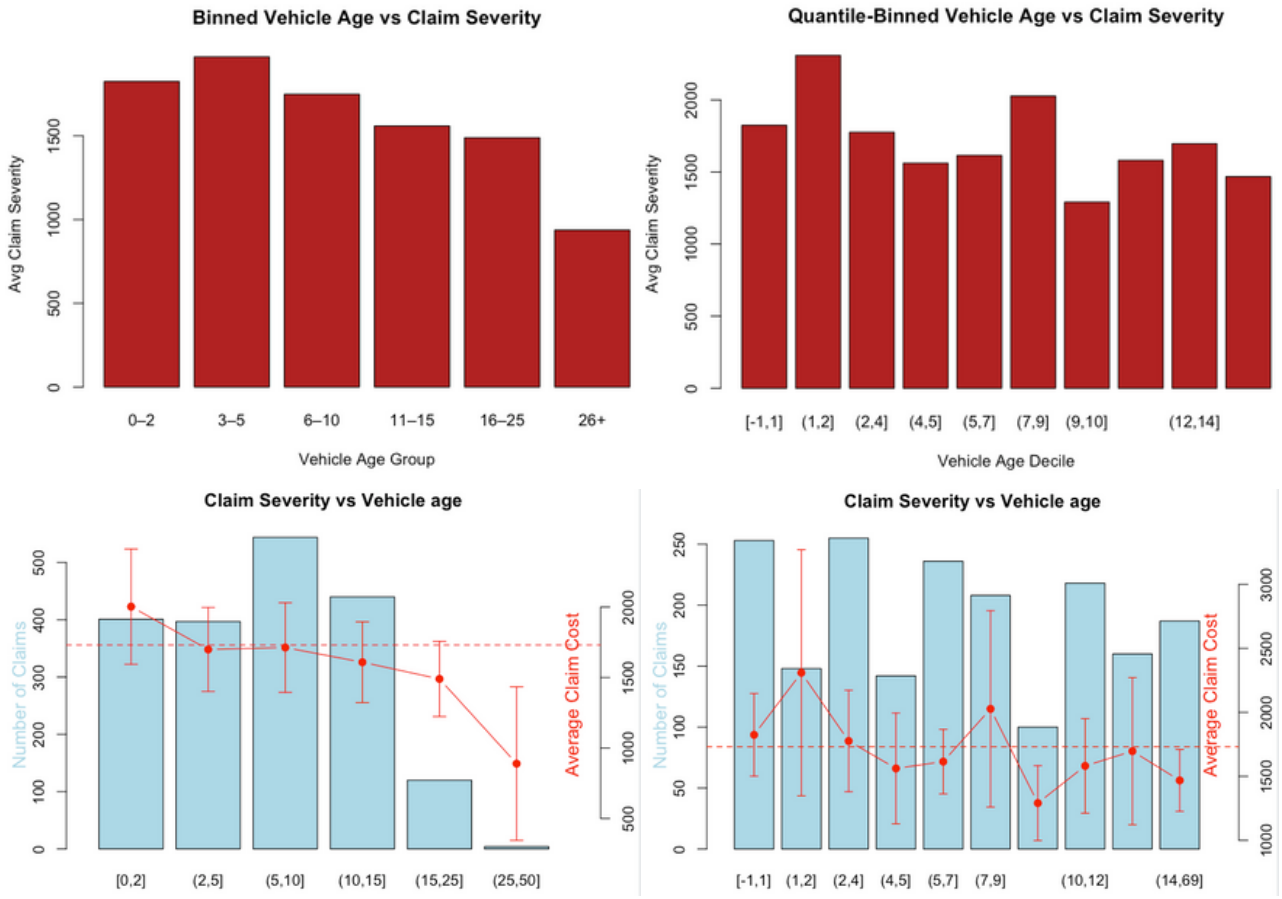


Figure 27 - Graphics Vehicle age vs Claim Severity

ANNEX

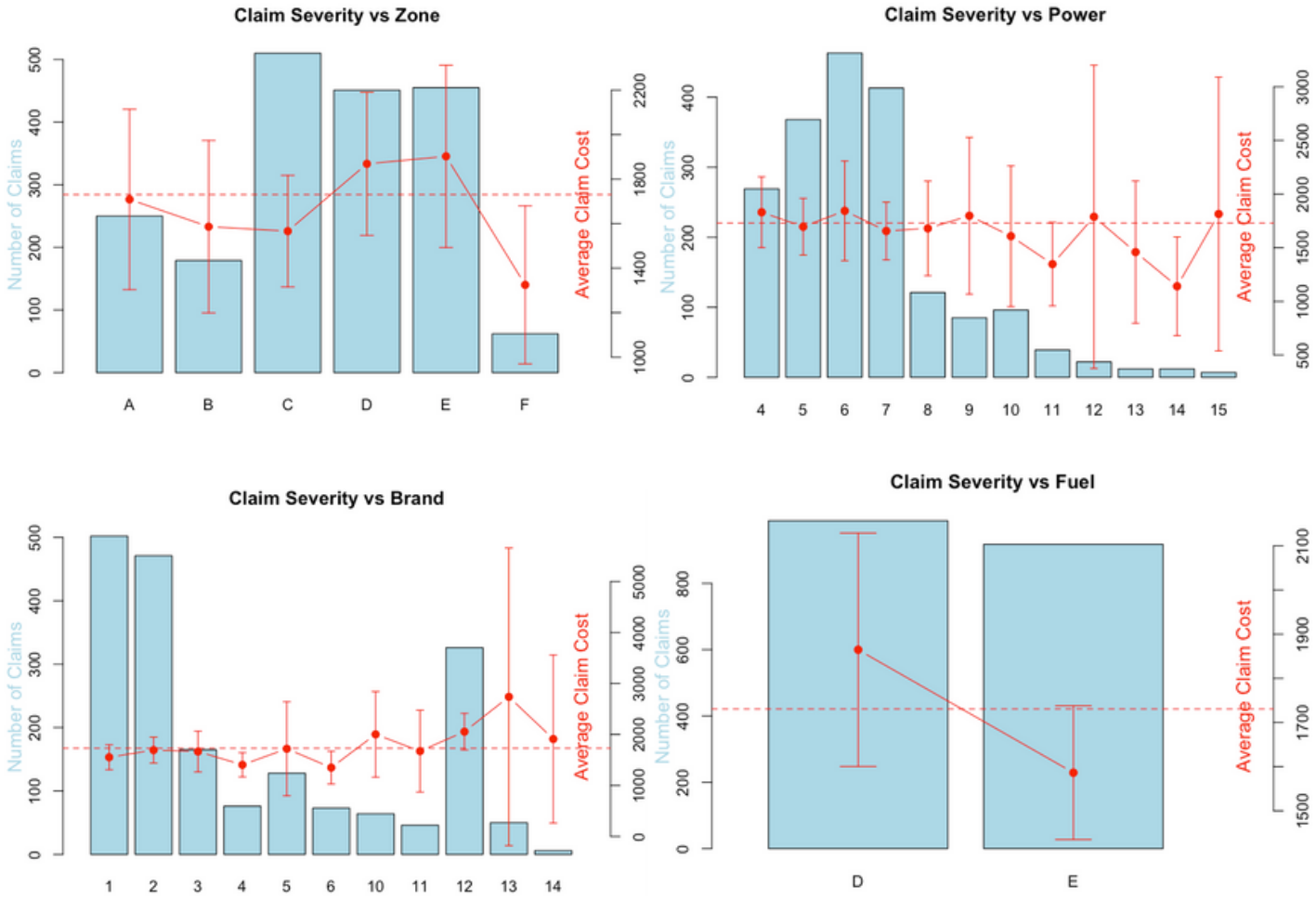
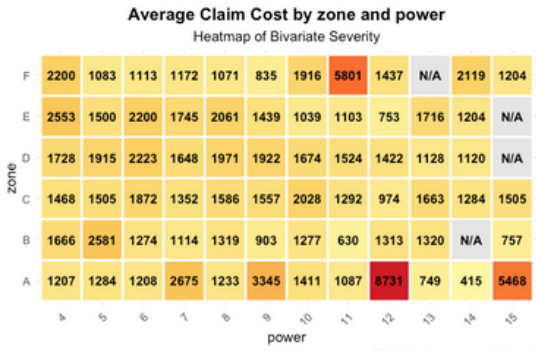
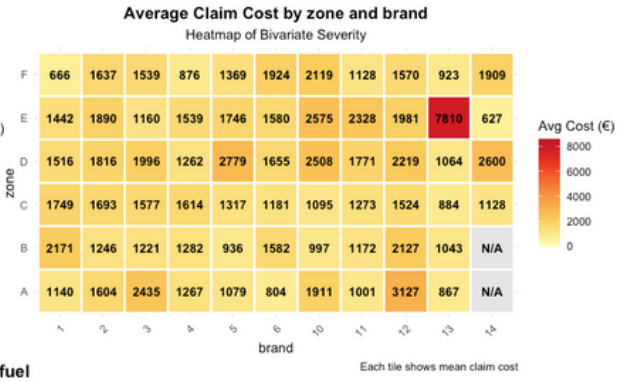


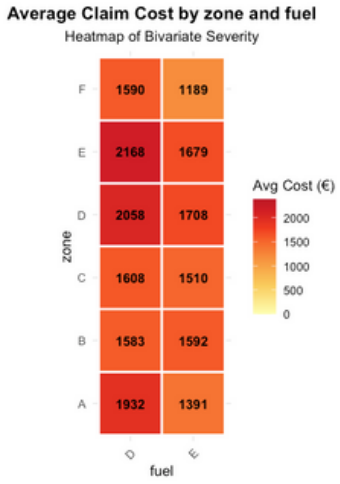
Figure 28 - Graphics Categorical Features vs Claim Severity

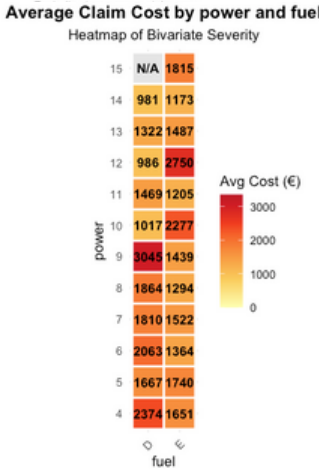
NOVA IMS Information Management School

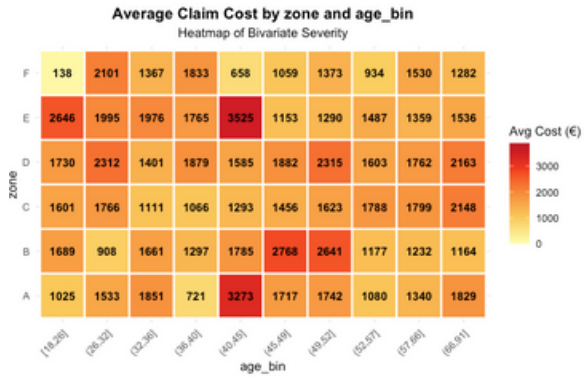
ANNEX

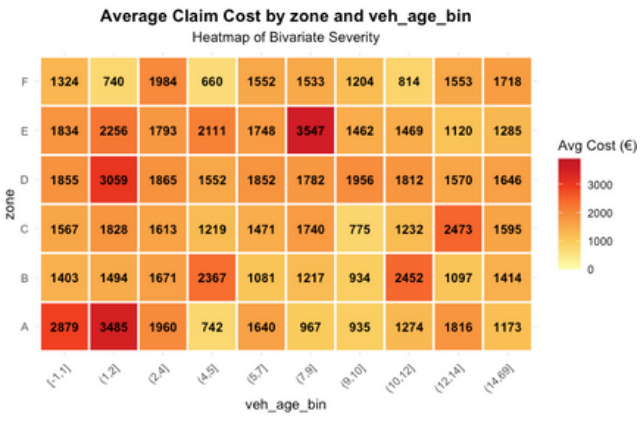


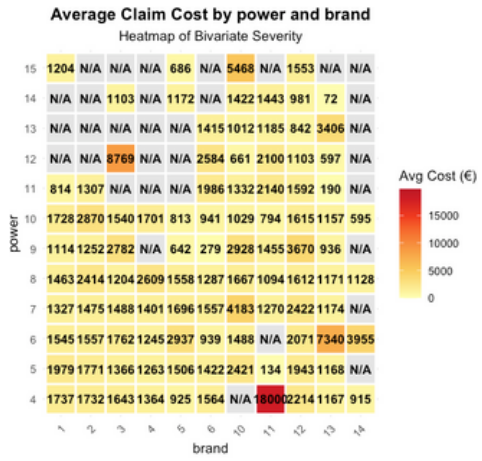




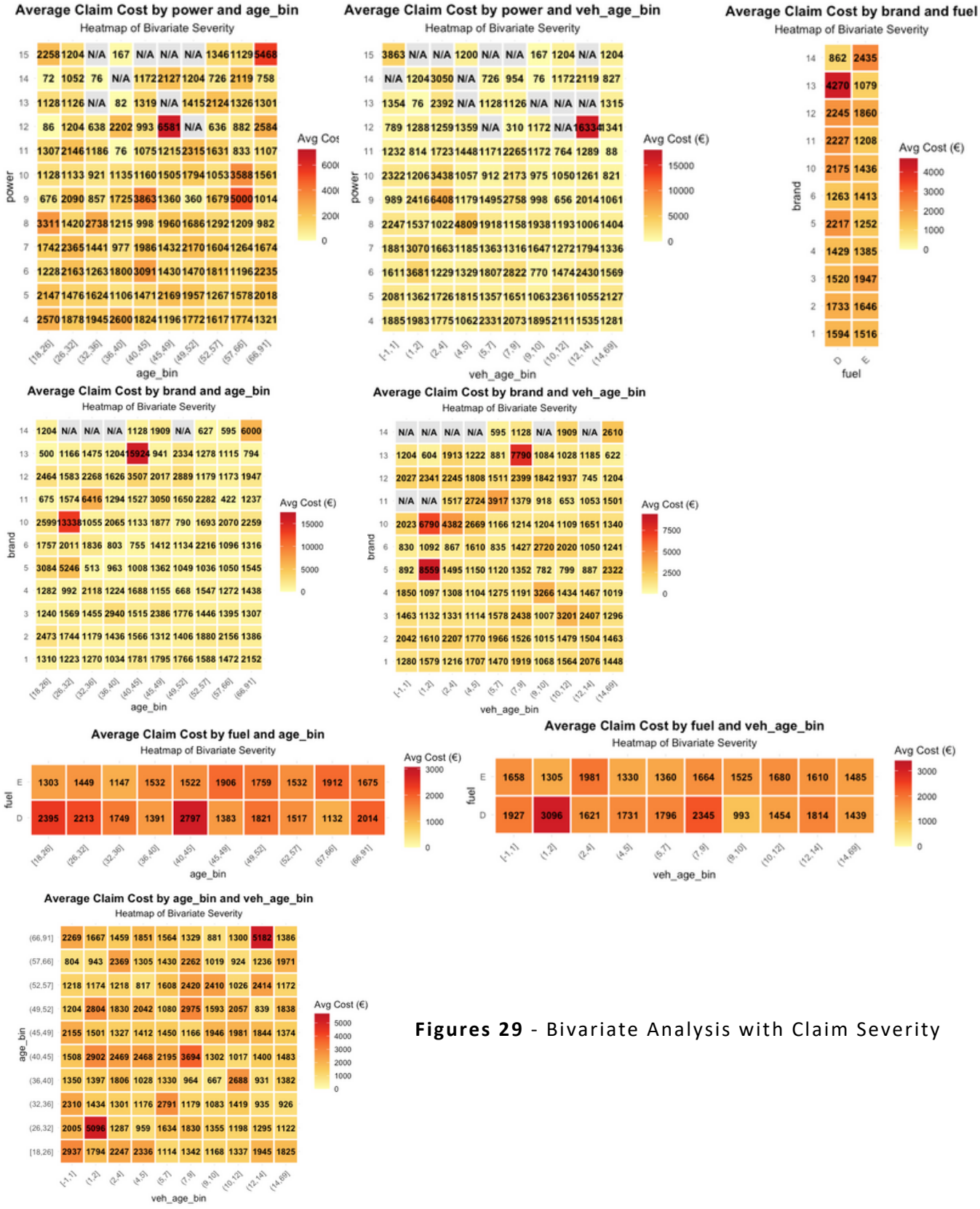








ANNEX



Figures 29 - Bivariate Analysis with Claim Severity

ANNEX

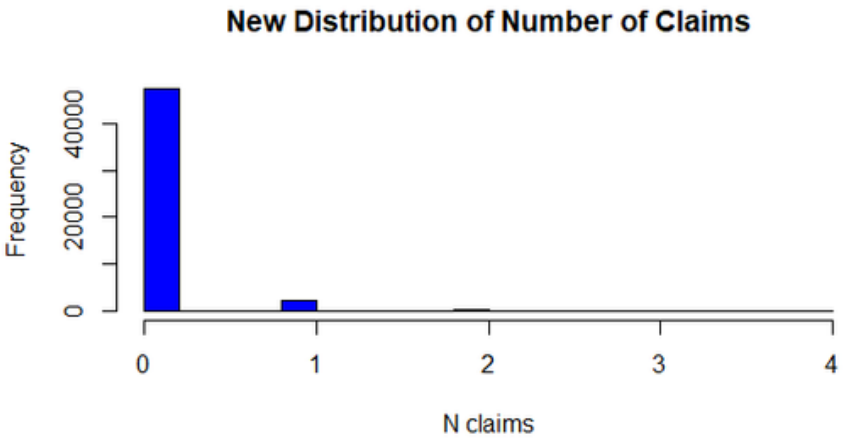


Figure 30 - Histogram Number of Claims after Outlier Removal

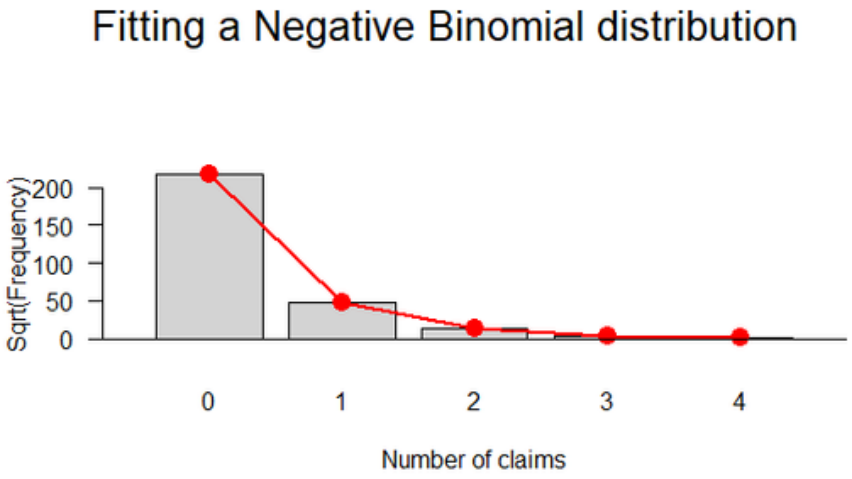


Figure 31 - Number of Claims Negative Binomial Fit

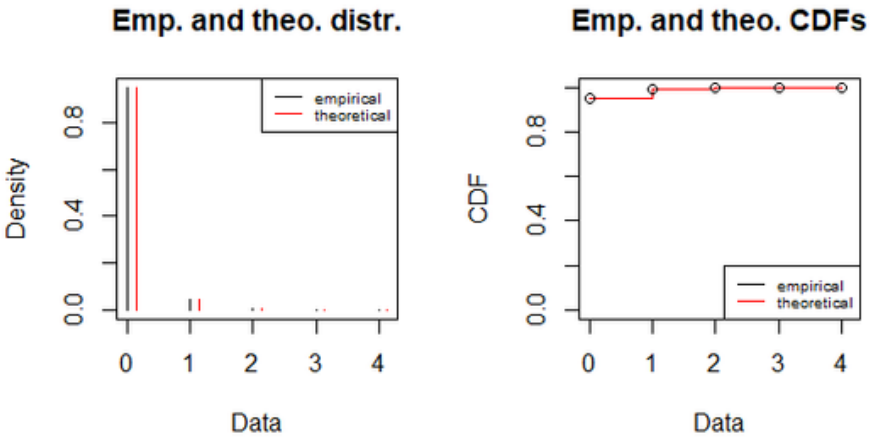


Figure 32 - Empirical versus Neg. Binom. Distribution and CDF

ANNEX

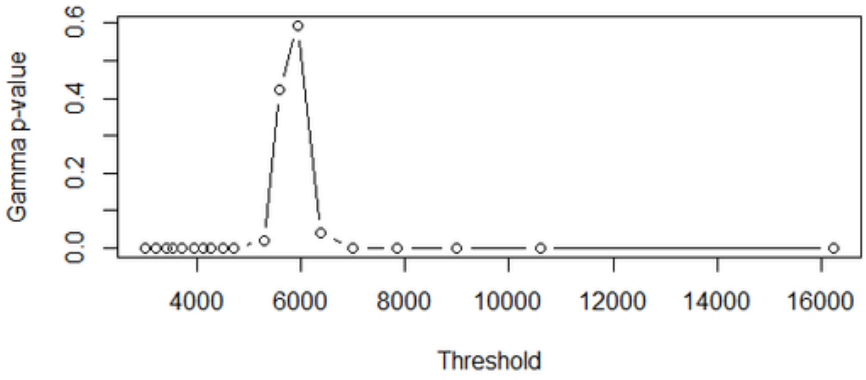


Figure 33 - Thresholds P-values

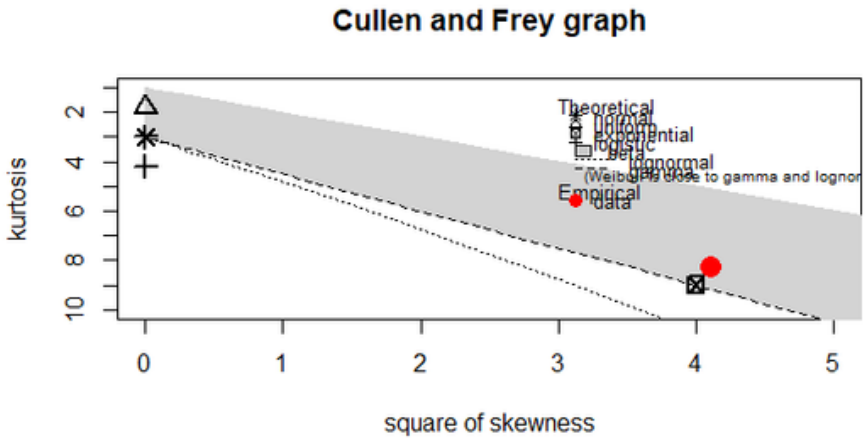


Figure 34 - Cullen and Frey Graph

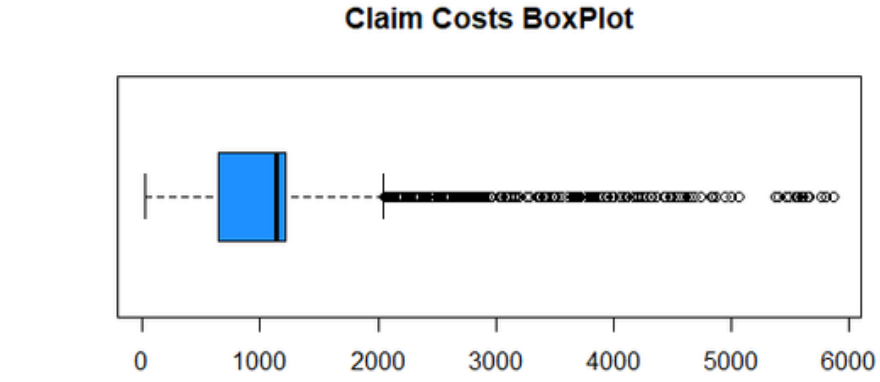


Figure 35 - Claim Cost Boxplot

ANNEX

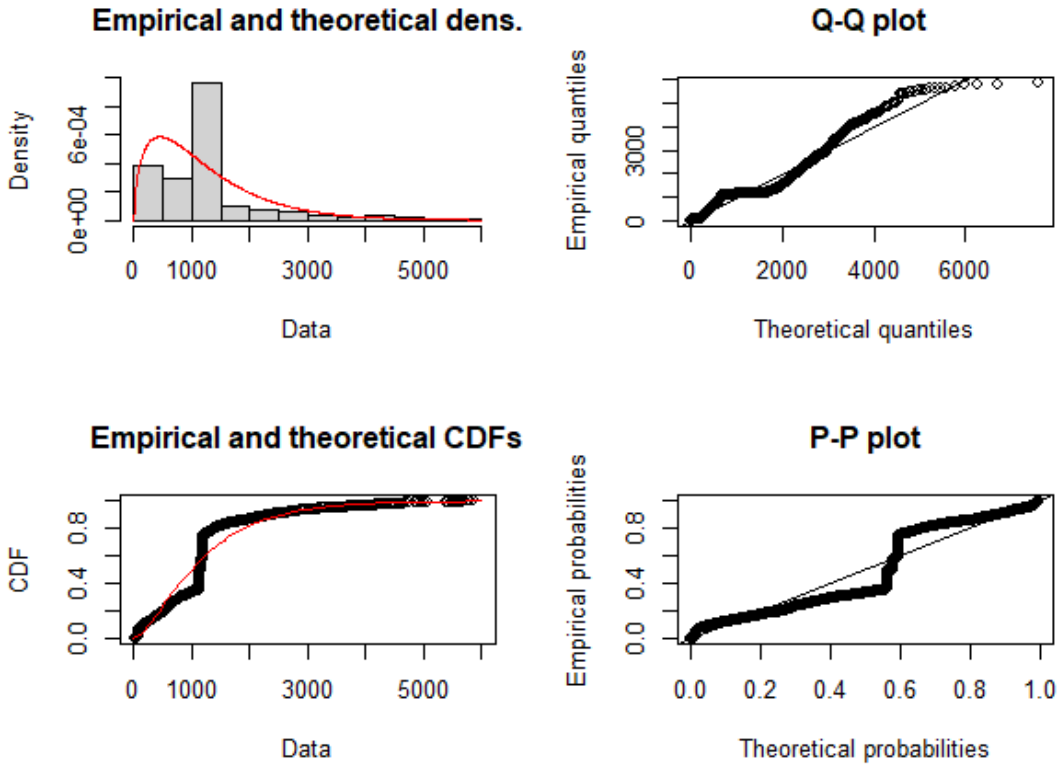


Figure 36 - Claim Cost Gamma Distribution Fit

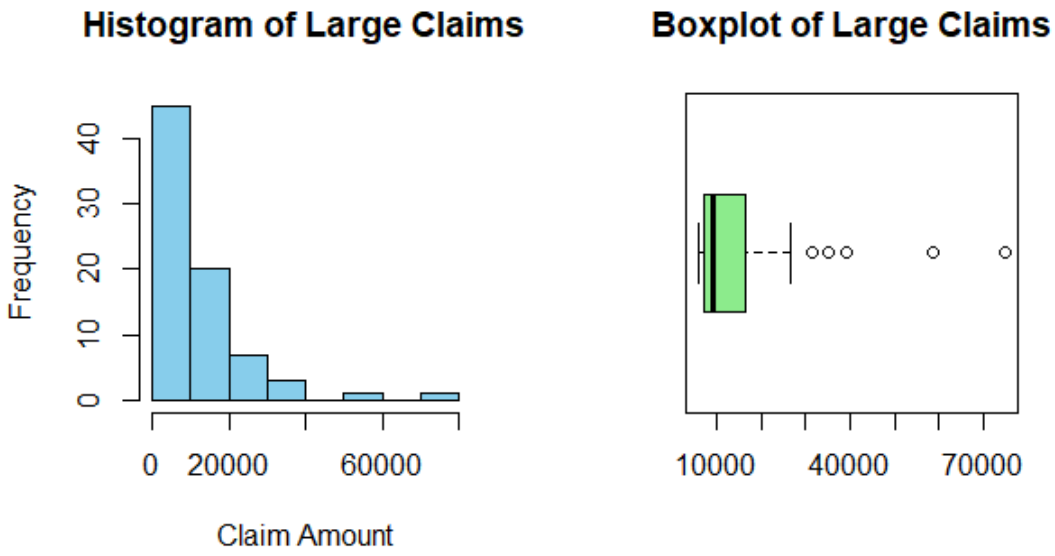


Figure 37 - Histogram and Boxplot of Large Claims

ANNEX

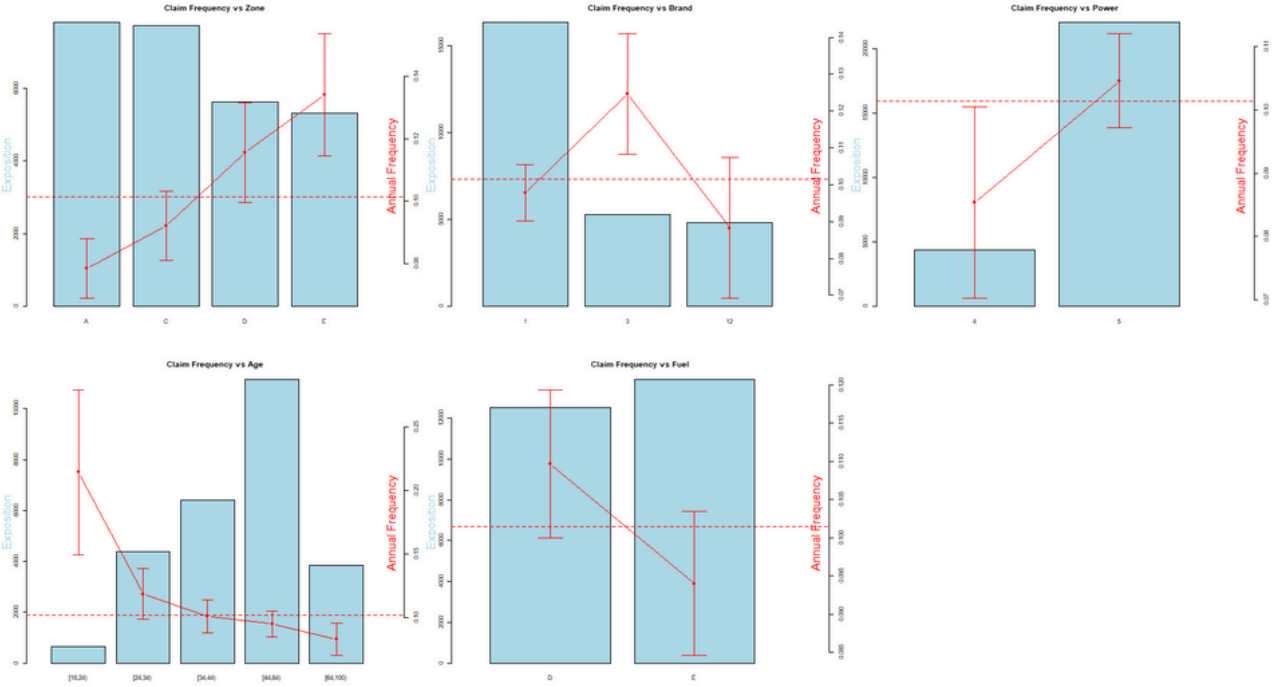


Figure 38 - Claim Frequency and Exposition vs Variable Plots

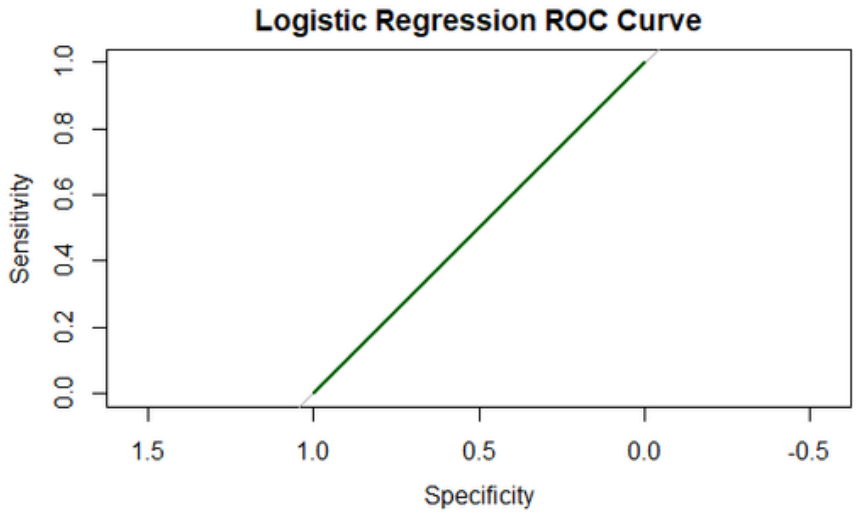


Figure 39 - Large Claims Logistic Regression ROC Curve

ANNEX

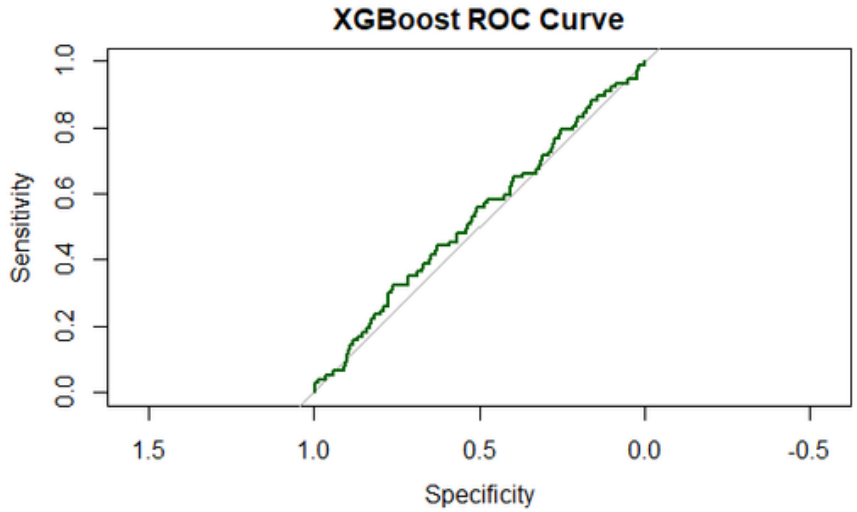


Figure 40 - Large Claims XGBoost ROC Curve (Option 1)

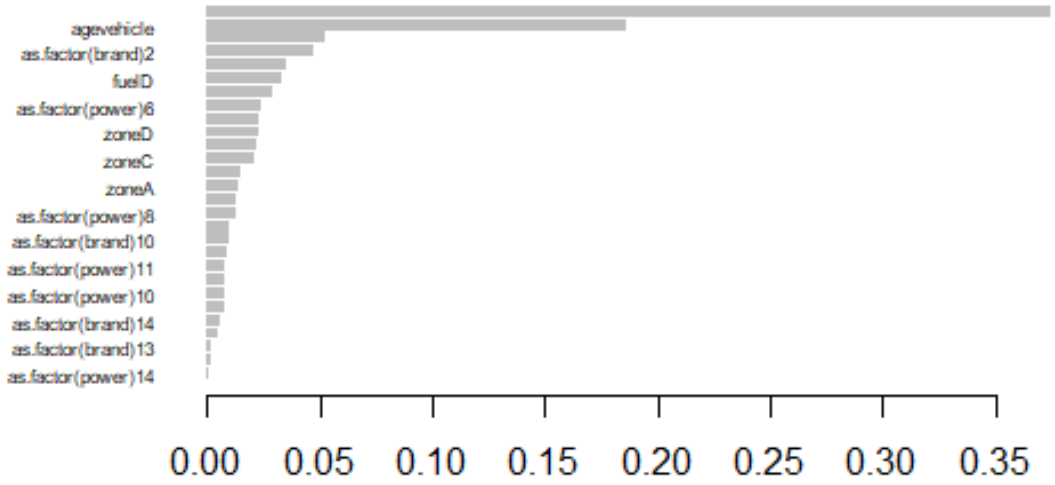


Figure 41 - Large Claims XGBoost Feature Importance (Option 1)

ANNEX

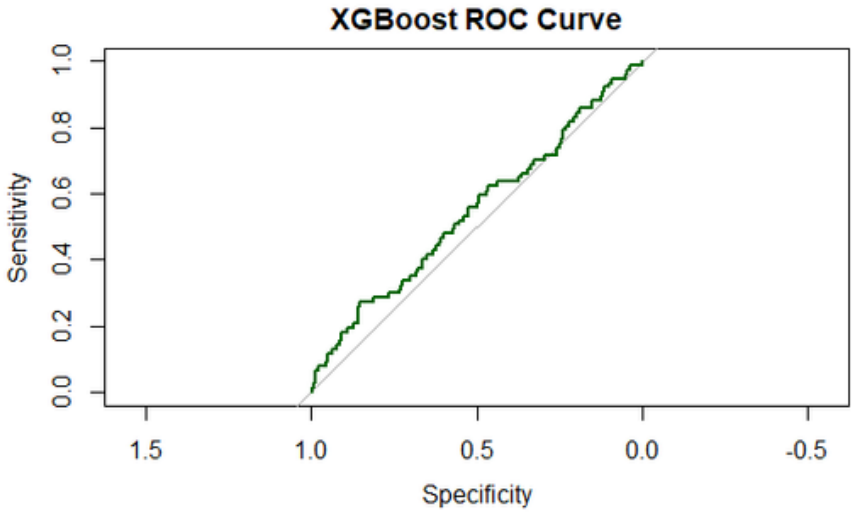


Figure 42 - Large Claims XGBoost ROC Curve (Option 2)

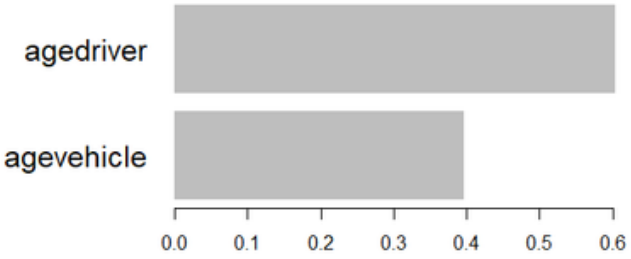


Figure 43 - Large Claims XGBoost Feature Importance (Option 2)

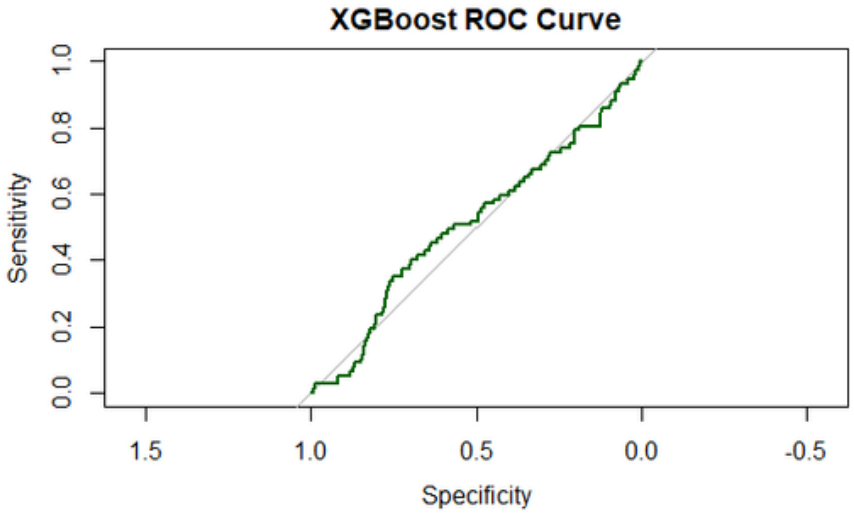


Figure 44 - Large Claims XGBoost ROC Curve (Option 3)

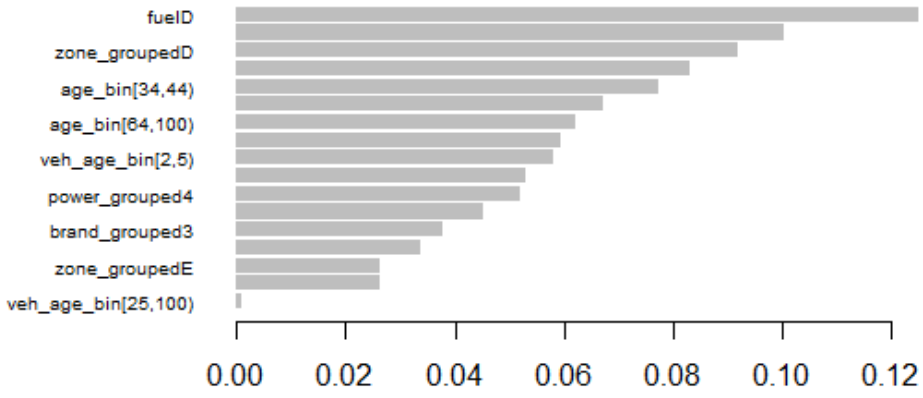


Figure 45 - Large Claims XGBoost Feature Importance (Option 3)

ANNEX

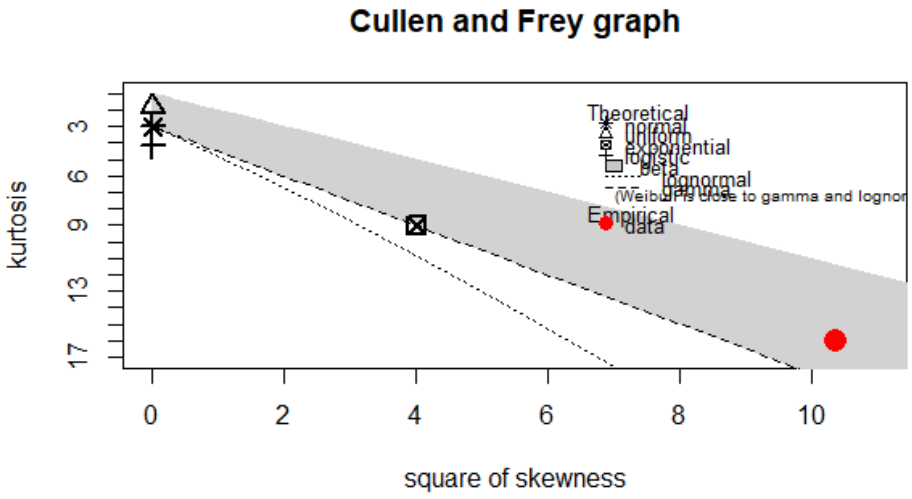


Figure 46 - Cullen and Frey Graph

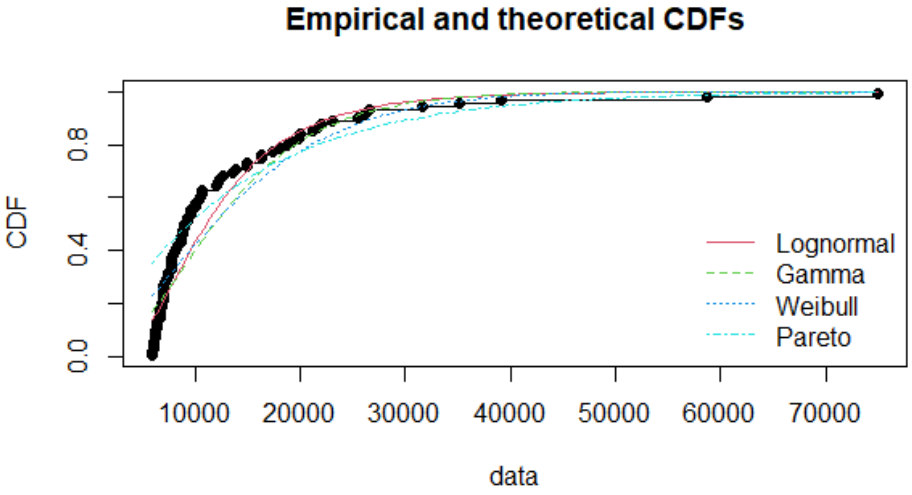


Figure 47 - Empirical and Theoretical CDFs for Large Claims Distribution versus Standard Distributions

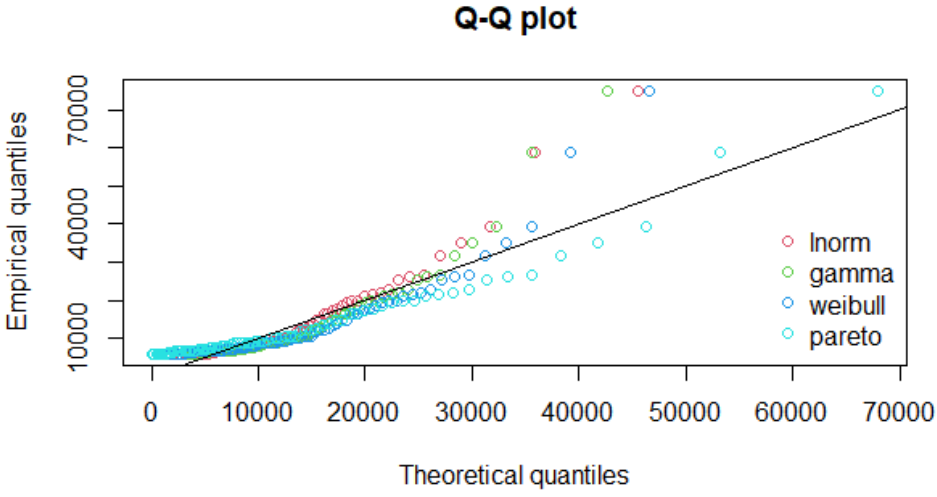


Figure 48 - Empirical and Theoretical CDFs for Large Claims Distribution versus Standard Distributions

TABLES

Observed and fitted values for nbinomial distribution with parameters estimated by `ML'

count	observed	fitted	pearson residual
0	47510	47511.328836	-0.006096392
1	2313	2309.608036	0.070580079
2	162	164.031572	-0.158623821
3	12	12.875047	-0.243869252
4	2	1.058666	0.784342472

Table 1 - Negative Binomial Fit to Number of Claims

Goodness-of-fit test for nbinomial distribution

	X^2	df	P(> X^2)
Likelihood Ratio	0.9487168	2	0.6222842

Table 2 - Goodness-of-fit test

Fitting of the distribution ' nbinom ' by maximum likelihood

Parameters:

	estimate	Std. Error
size	0.52463091	0.059365240
mu	0.05362151	0.001086816

Table 3 - Negative Binomial Fit to Number of Claims

Fitting of the distribution ' gamma ' by maximum likelihood

Parameters:

	estimate	Std. Error
shape	1.569159234	2.490969e-02
rate	0.001270114	4.315860e-08

Table 4 - Gamma Fit to Common Claims

(Intercept)	zone_groupedC	zone_groupedD
4.505436e-02	8.858949e-01	1.834442e+00
zone_groupedE	power_grouped4	veh_age_bin[2,5)
1.456037e+00	1.200860e+00	6.191821e-01
veh_age_bin[5,10)	veh_age_bin[10,15)	veh_age_bin[15,25)
6.513504e-01	5.998081e-01	6.388424e-01
veh_age_bin[25,100)	age_bin[18,24)	age_bin[24,34)
3.156601e-06	1.849817e+00	1.283507e+00
age_bin[34,44)	age_bin[64,100)	brand_grouped3
1.180411e+00	8.281506e-01	1.089247e+00
brand_grouped12	fuelD	
2.115081e+00	1.386685e+00	

Table 5 - Logistic Regression Odds Ratios

Confusion Matrix and Statistics

Prediction	Reference 0	1
0	1830	77
1	0	0

Table 6 - Logistic Regression Confusion Matrix

ANNEX

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1830   77
         1    0    0
```

Table 7 - Logistic Regression Confusion Matrix - Threshold optimization

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1753   73
         1   77    4
```

Table 8 - XGBoost (Option 1) Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1398   52
         1  432   25
```

Table 9 - XGBoost (Option 1) Confusion Matrix - Threshold optimization

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1629   62
         1  201   15
```

Table 10 - XGBoost (Option 2) Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1567   56
         1  263   21
```

Table 11 - XGBoost (Option 2) Confusion Matrix - Threshold optimization

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1529   66
         1  301   11
```

Table 12 - XGBoost (Option 3) Confusion Matrix

ANNEX

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1381   50
         1  449   27
```

Table 13 - XGBoost (Option 3) Confusion Matrix - Threshold optimization

```
Goodness-of-fit statistics
                              1-mle-lnorm 2-mle-gamma 3-mle-weibull
Kolmogorov-Smirnov statistic   0.1642354   0.2024326     0.2304568
Cramer-von Mises statistic     0.5144643   0.7773614     0.8528031
Anderson-Darling statistic     2.9414348   4.3100405     4.9693038
                              4-mle-pareto
Kolmogorov-Smirnov statistic    0.3555398
Cramer-von Mises statistic      1.5520828
Anderson-Darling statistic      8.3221818

Goodness-of-fit criteria
                                1-mle-lnorm 2-mle-gamma 3-mle-weibull
Akaike's Information Criterion    1569.646    1591.207      1605.107
Bayesian Information Criterion    1574.334    1595.895      1609.795
                                4-mle-pareto
Akaike's Information Criterion     1622.645
Bayesian Information Criterion     1627.333
```

Table 14 - Goodness-of-fit Statistics for Large Claims Cost

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa