ML1

2024-12-16

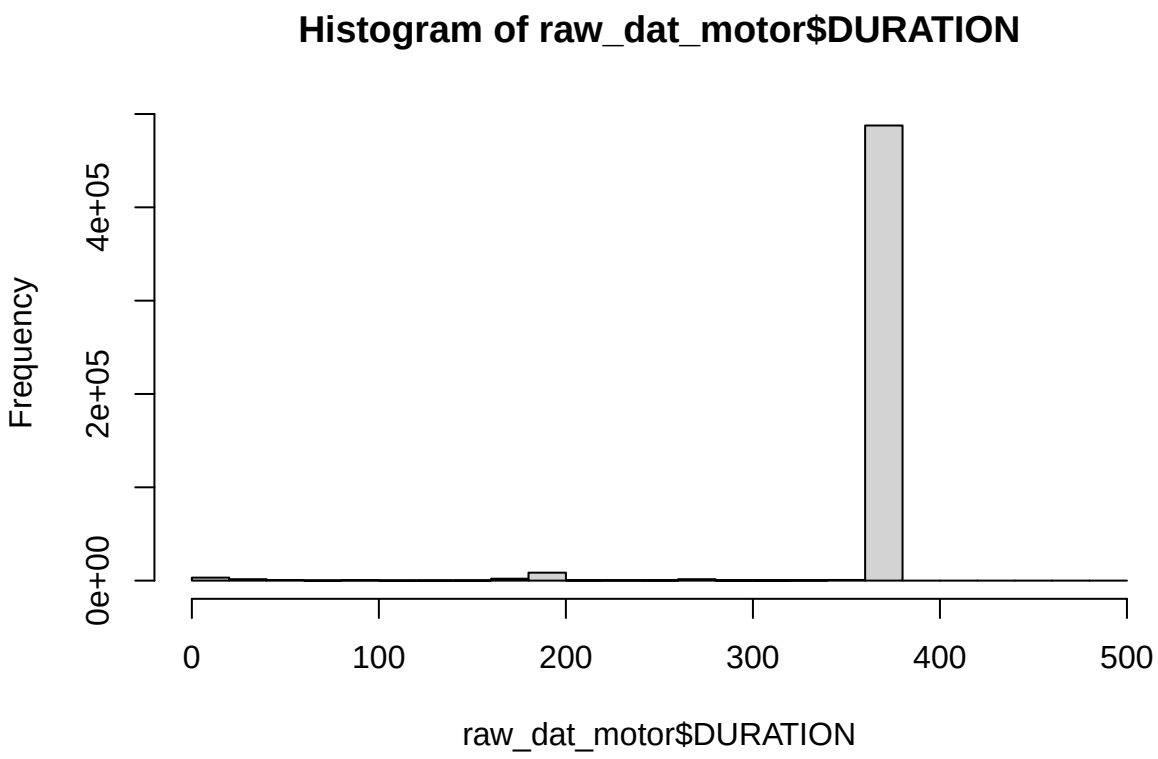
Introduction

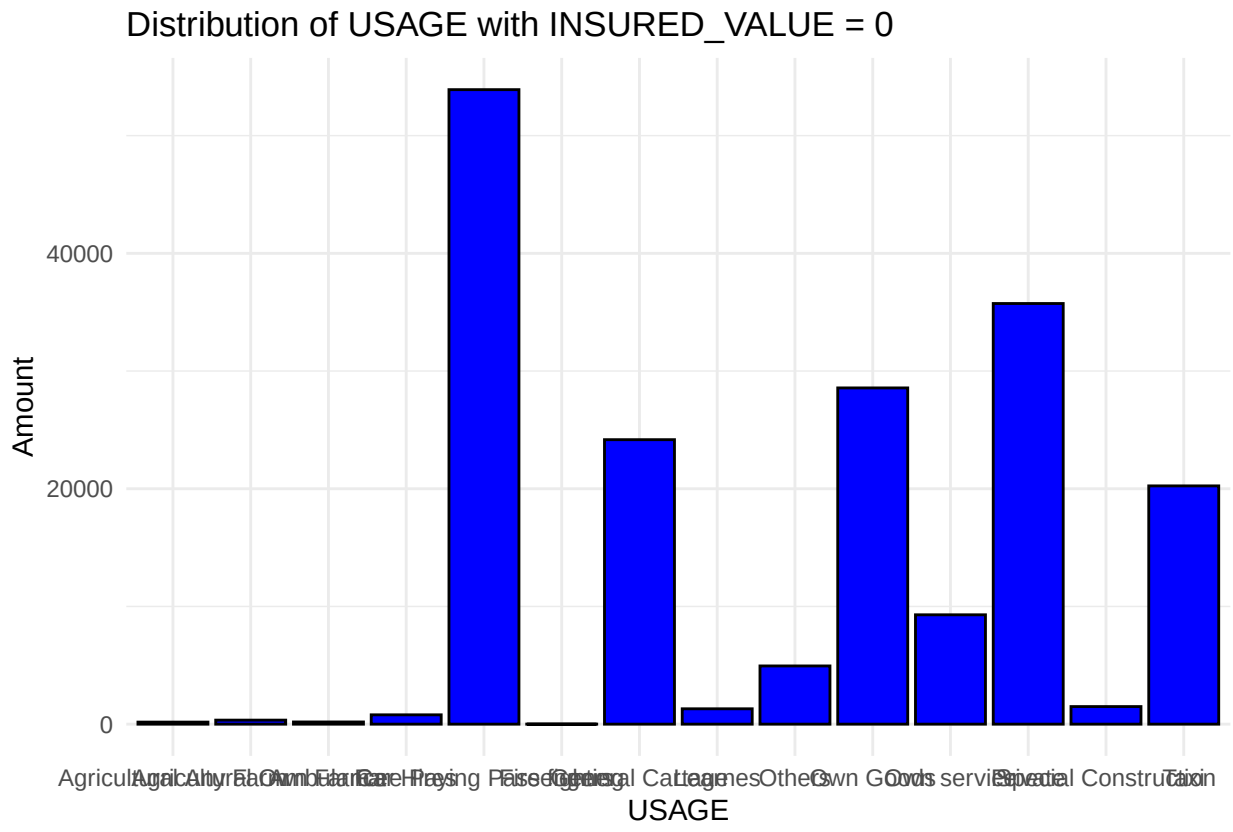
One of the major challenges for insurances is to estimate the appropriate premiums to charge each customer while not risking to lose any money. Therefore, this project aims at supporting an Ethiopian Insurance company to understand how their customers can benefit from having the most accurate and fair premium as they need and have to pay. Machine Learning helps in this case enormously to understand, what factors have a larger impact on the premium and how customers can be classified accordingly.

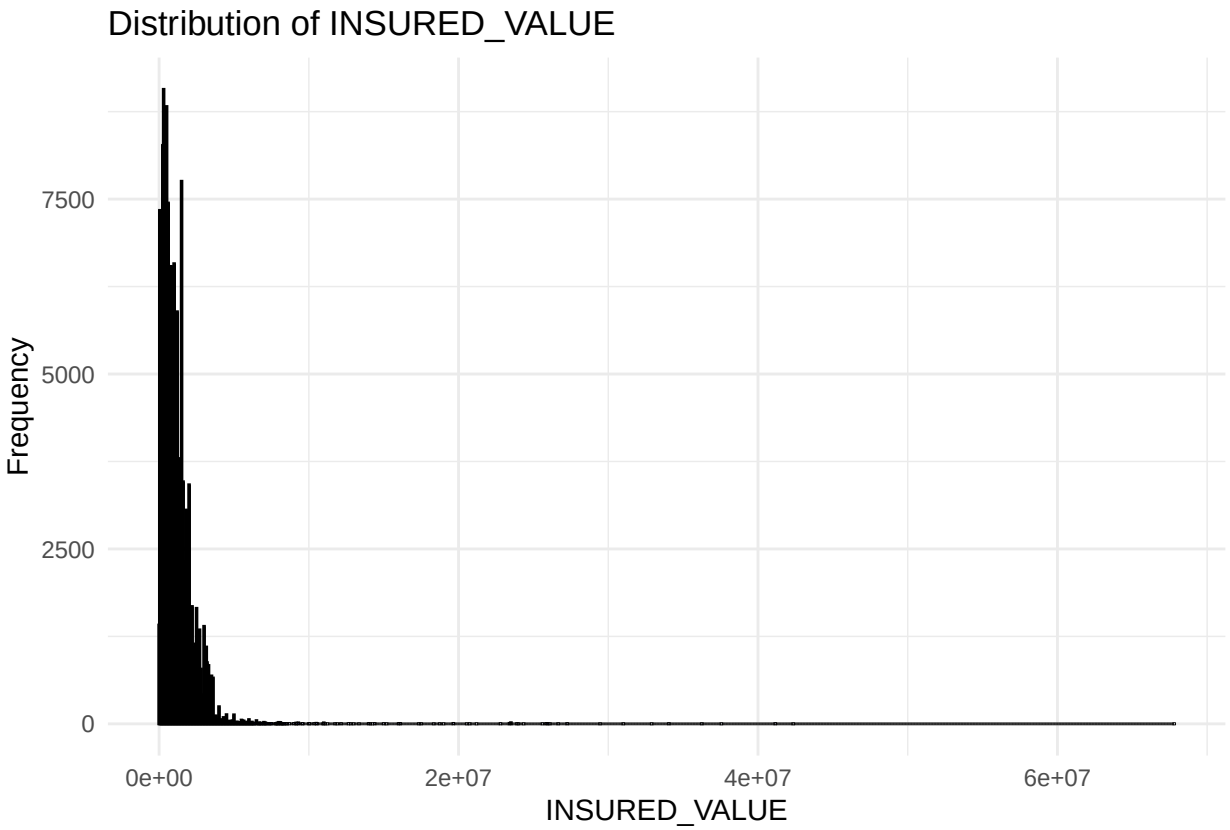
In this document, the reader may find different algorithms to solve various aspects of the premium-calculations.

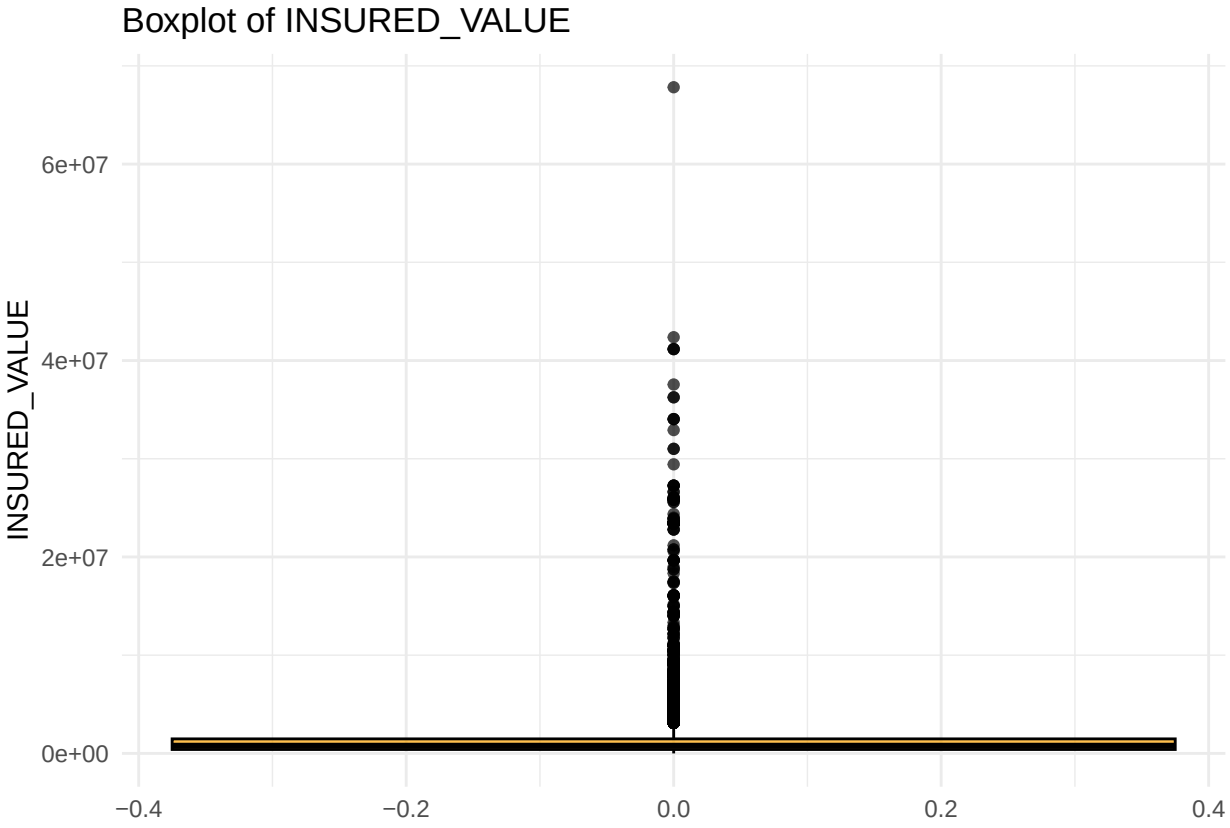
Data Preprocessing

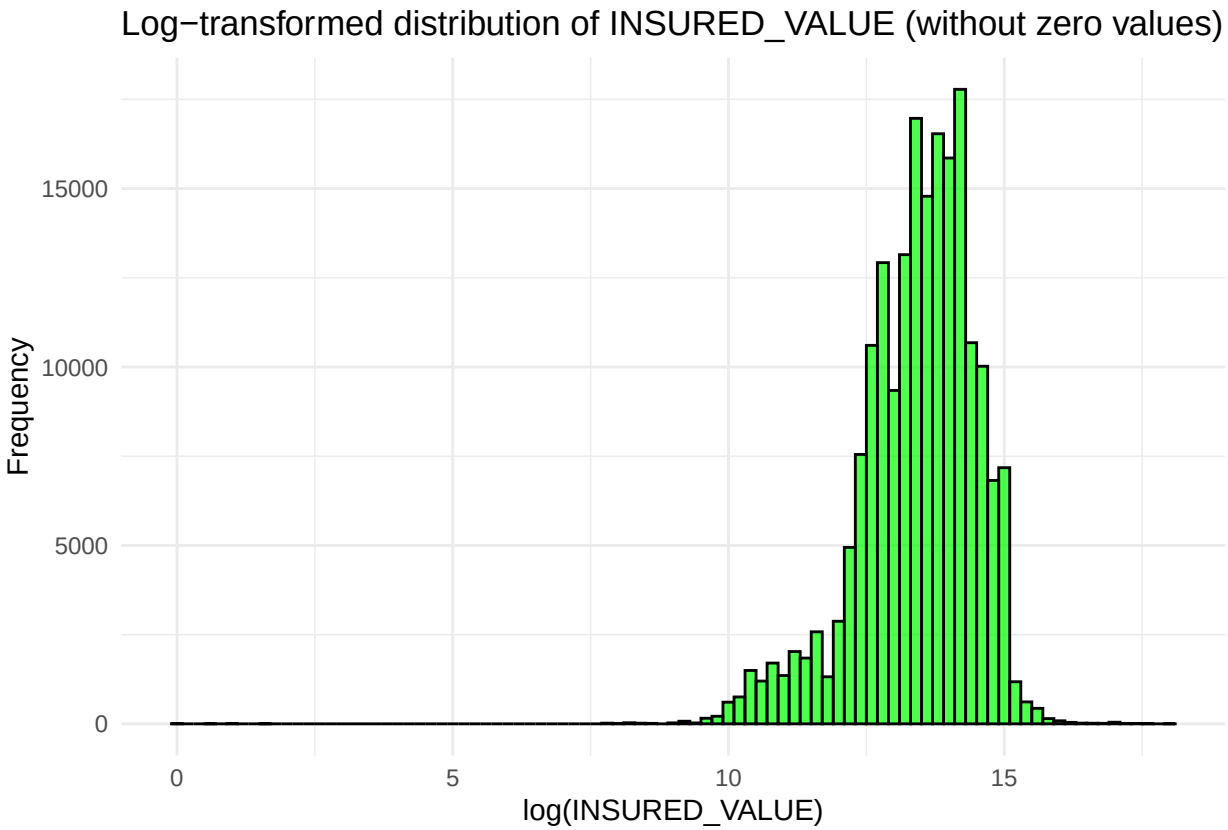
In order to apply such algorithms, the data has to be pre-processed. In a very brief summary, the script removes unnecessary columns and duplicates, and handles missing and zero values, particularly for columns like INSURED_VALUE and SEATS_NUM. It converts certain columns to more meaningful categories, such as transforming SEX into factors representing legal entities and genders. The script also filters data to exclude irrelevant vehicle types and usage, ensuring the final dataset contains only pertinent records. Finally, it summarizes and adjusts the dataset further by converting appropriate columns into factors and removing variables not required for analysis.


Histogram of raw_dat_motor$DURATION
Frequency
0e+00
2e+05
4e+05
0
100
200
300
400
500
raw_dat_motor$DURATION


Distribution of USAGE with INSURED_VALUE = 0
Amount
0
20000
40000
USAGE


Distribution of INSURED_VALUE
Frequency
7500
5000
2500
0
0e+00
2e+07
4e+07
6e+07
INSURED_VALUE

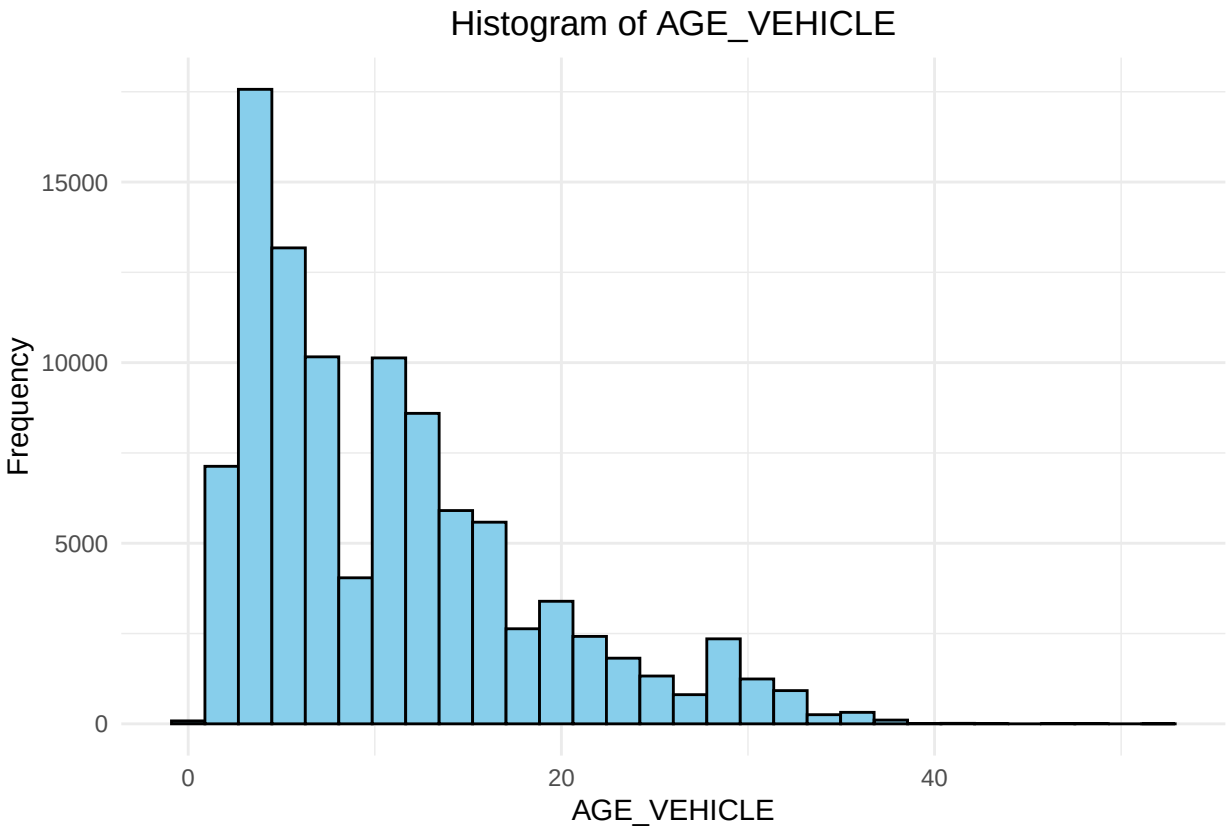

Boxplot of INSURED_VALUE
INSURED_VALUE
6e+07
4e+07
2e+07
0e+00
−0.4
−0.2
0.0
0.2
0.4

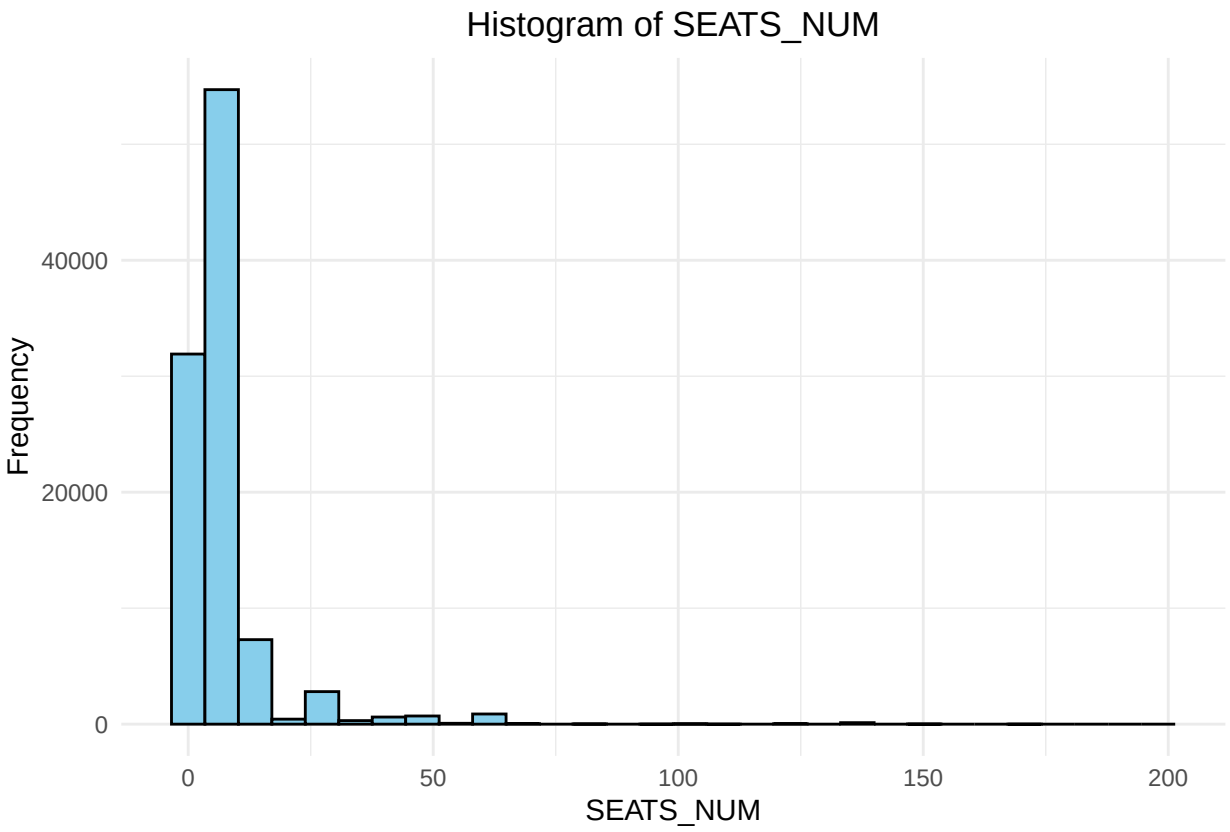


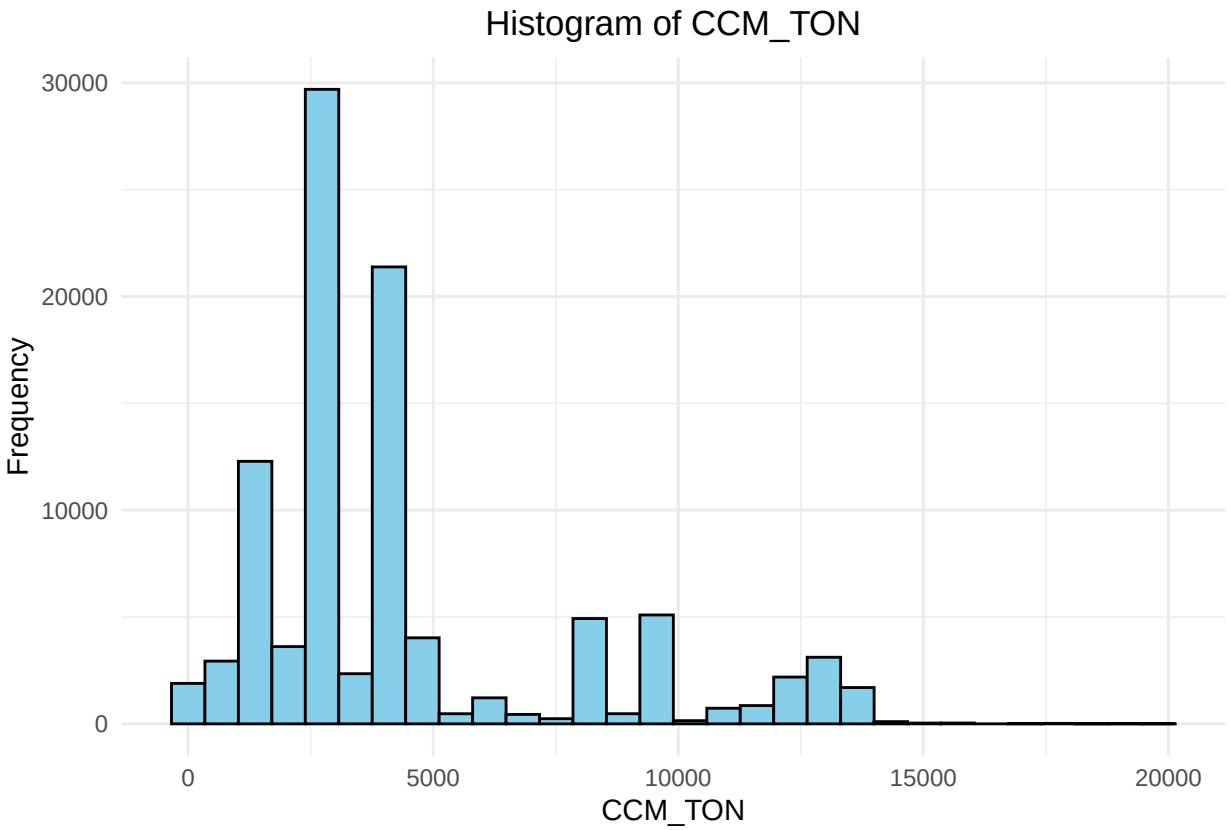
PREMIUM_0_Percent PREMIUM_NA_Percent PREMIUM_MORE_Percent 1 0.0031 0.002 99.9969 Number of removed duplicates: 0 The OBJECT_IDs are NOT unique. Number of duplicates: 100652 Average frequency of the OBJECT_ID: 2.055 Maximum frequency of the OBJECT_ID: 8 Average frequency of the combination (OBJECT_ID, INSR_BEGIN, INSR_END, INSURED_VALUE, PREMIUM): 1 Maximum frequency of the combination (OBJECT_ID, INSR_BEGIN, INSR_END, INSURED_VALUE, PREMIUM): 2 Min. 1st Qu. Median Mean 3rd Qu. Max. 1960 2003 2010 2007 2014 2018 SEATS_NUM_0 SEATS_NUM_NA SEATS_NUM_OTHER 1 19940 10 176087 SEATS_NUM_0_or_NA_Percent SEATS_NUM_OTHER_Percent 1 10.17665 89.82335 Min. 1st Qu. Median Mean 3rd Qu. Max. 0.000 1.000 4.000 6.086 4.000 198.000 CCM_TON_0_Percent CCM_TON_MORE_Percent 1 13.76464 86.23536 CLAIM_PAID_0 CLAIM_PAID_MORE_THAN_0 1 125007 20093 CLAIM_PAID_0_Percent CLAIM_PAID_MORE_THAN_0_Percent 1 86.15231 13.84769 CLAIM_PAID_0 CLAIM_PAID_MORE_THAN_0 1 162658 22489 CLAIM_PAID_0_Percent CLAIM_PAID_MORE_THAN_0_Percent 1 87.85344 12.14656 CLAIM_PAID_0_Percent CLAIM_PAID_MORE_THAN_0_Percent 1 86.15231 13.84769 SEX INSR_BEGIN INSR_END INSR_TYPE INSURED_VALUE 0 0 0 0 0 PREMIUM OBJECT_ID PROD_YEAR SEATS_NUM TYPE_VEHICLE 0 0 0 0 0 CCM_TON MAKE USAGE CLAIM_PAID CLAIM_PAID_USD 0 0 0 0 0 DURATION START_INS_YR 0 0 SEX INSR_BEGIN INSR_END INSR_TYPE 0 0 0 0 INSURED_VALUE PREMIUM OBJECT_ID SEATS_NUM 0 0 0 0 TYPE_VEHICLE CCM_TON MAKE USAGE 0 0 0 0 CLAIM_PAID CLAIM_PAID_USD DURATION START_INS_YR 0 0 0 0 AGE_VEHICLE AMOUNT_CLAIMS_PAID 0 0

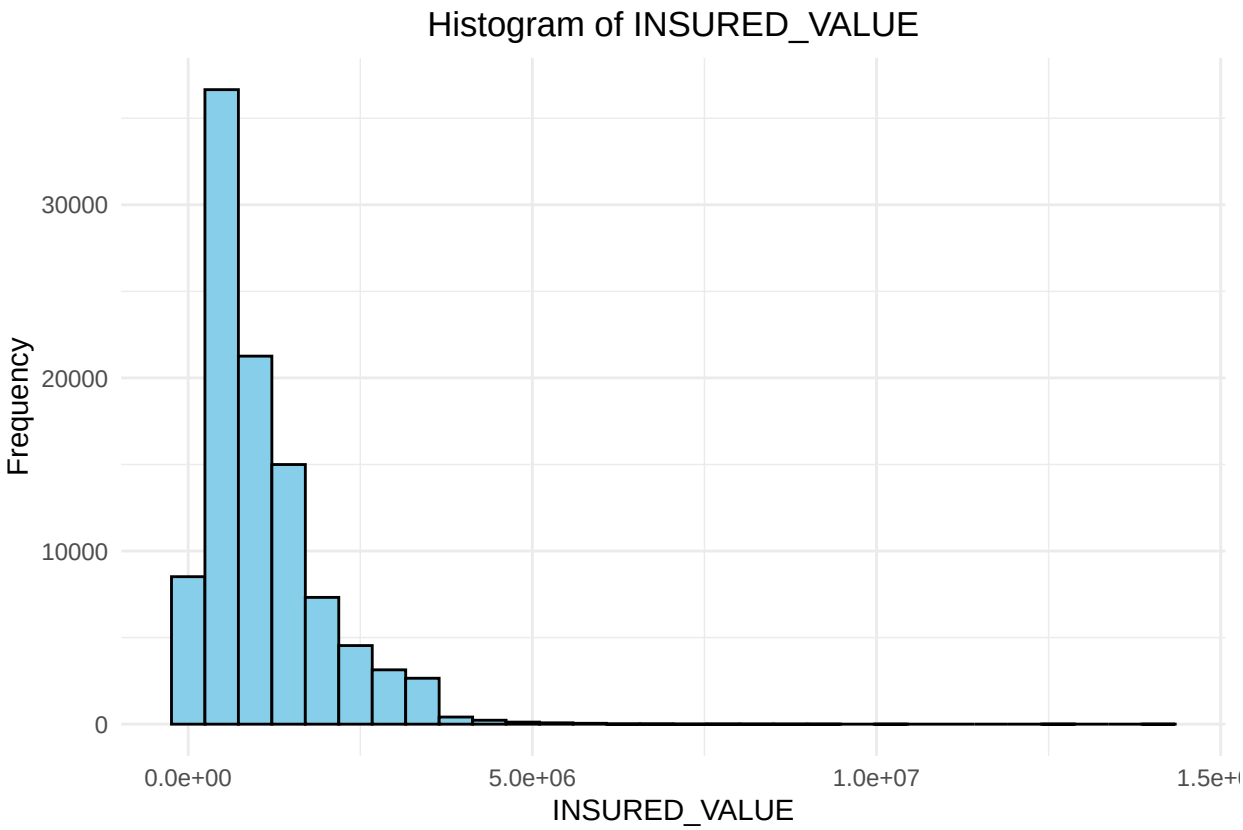
Graphical Data Analysis

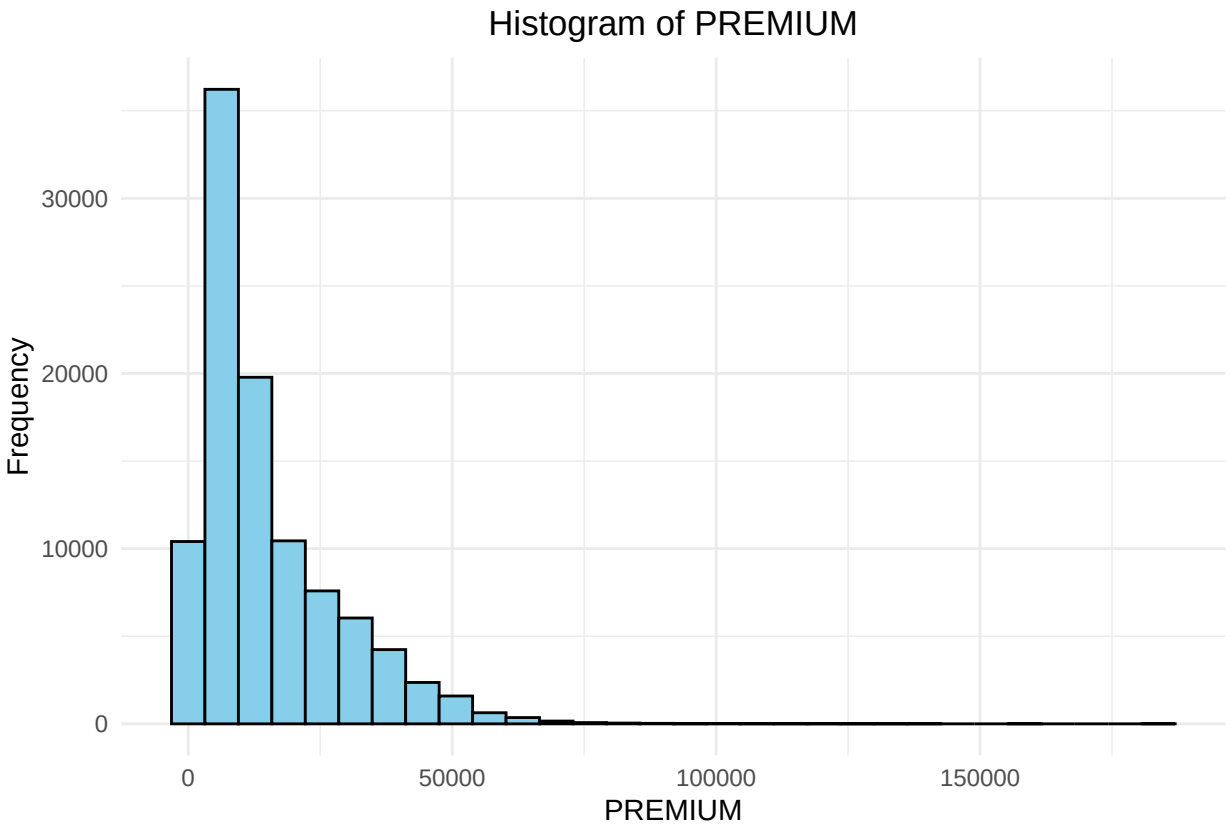
First, the distribution of the individual numerical variables was analysed to determine whether any transformations were necessary.

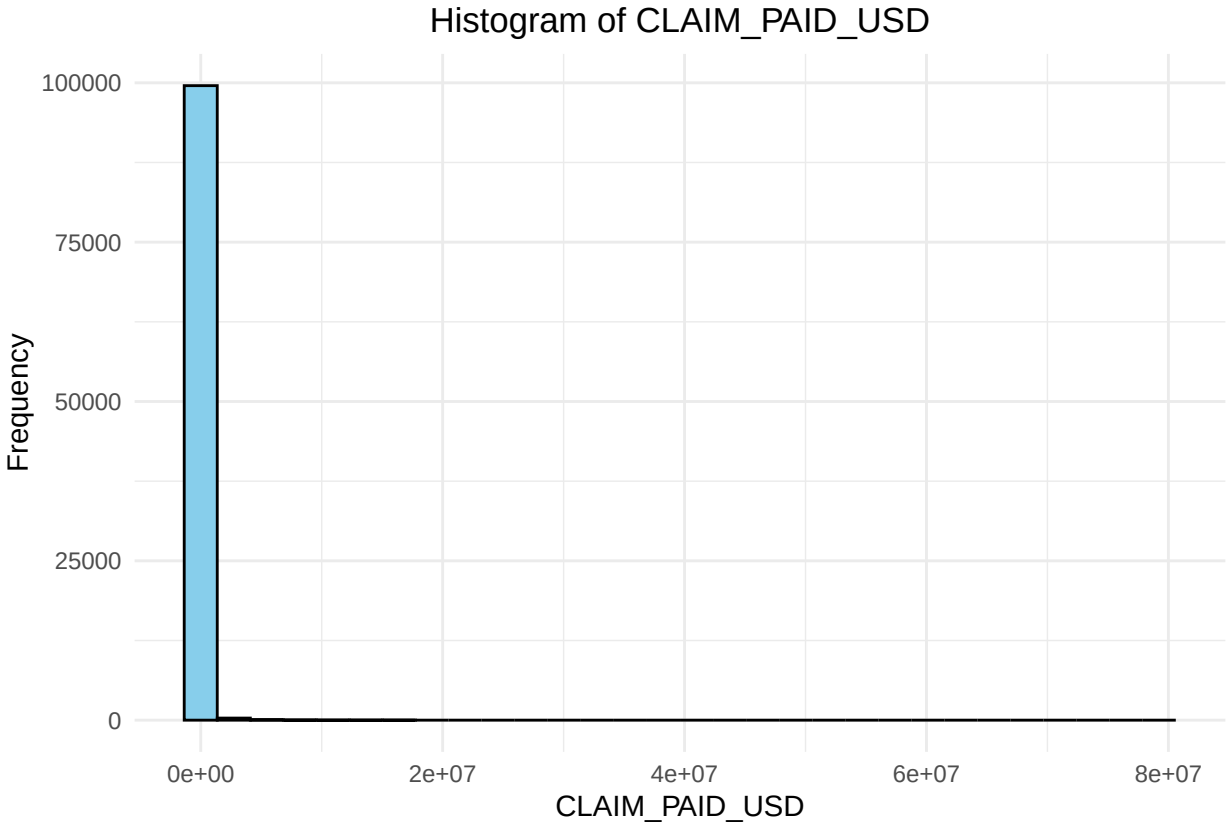

Histogram of AGE_VEHICLE
Frequency
15000
10000
5000
0
0
20
40
AGE_VEHICLE

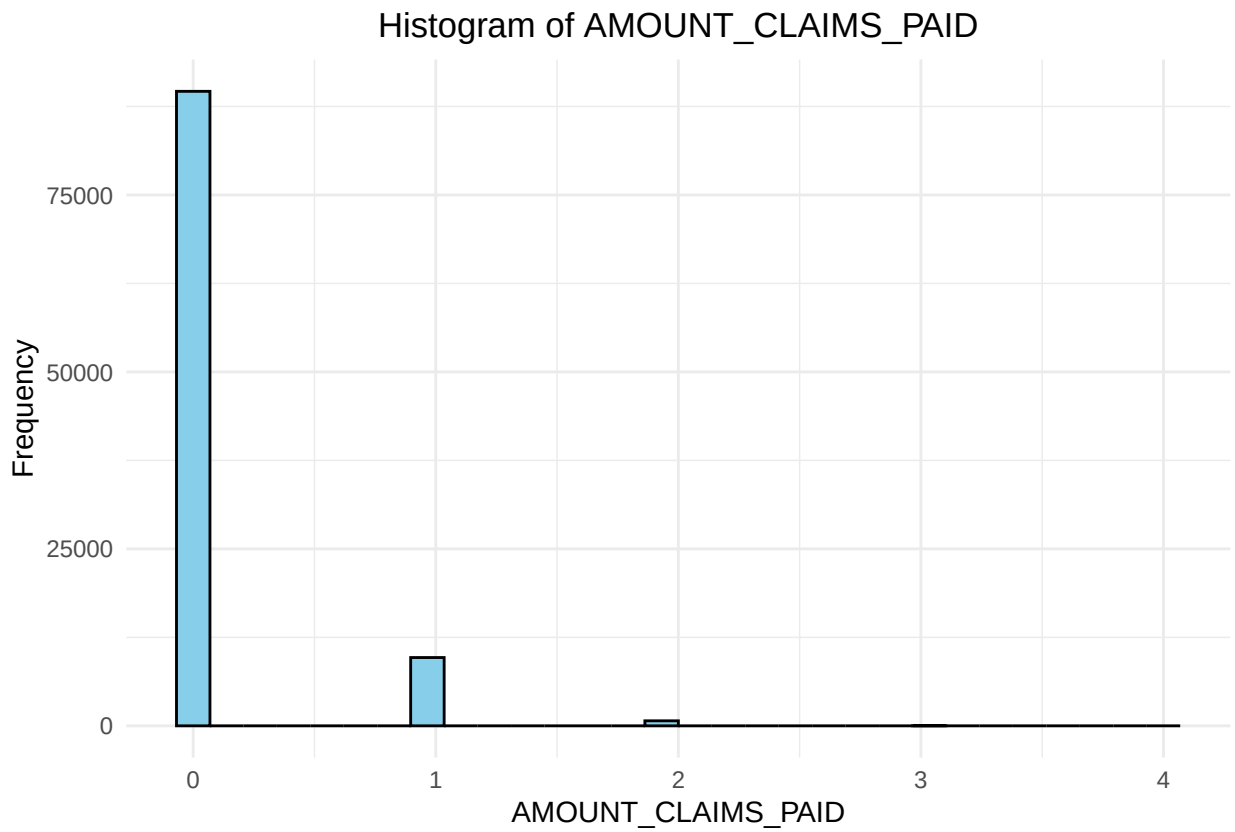

Histogram of SEATS_NUM
Frequency
40000
20000
0
0
50
100
150
200
SEATS_NUM

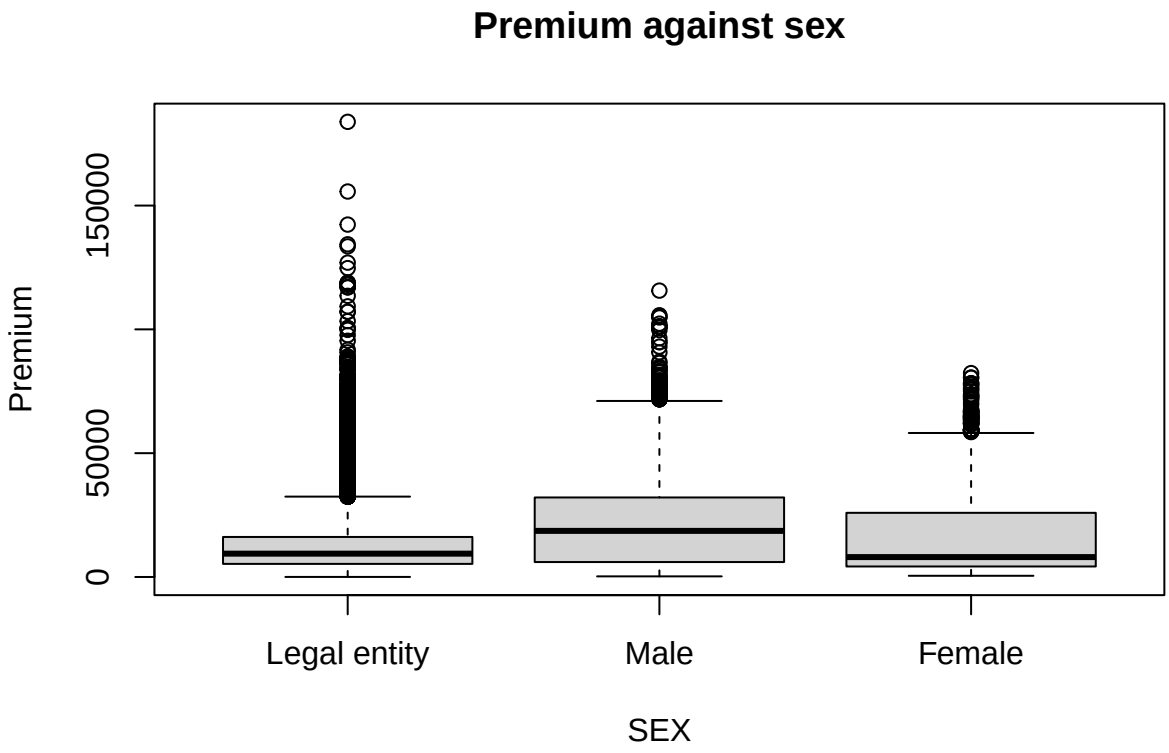

Histogram of CCM_TON
Frequency
30000
20000
10000
0
0
5000
10000
15000
20000
CCM_TON

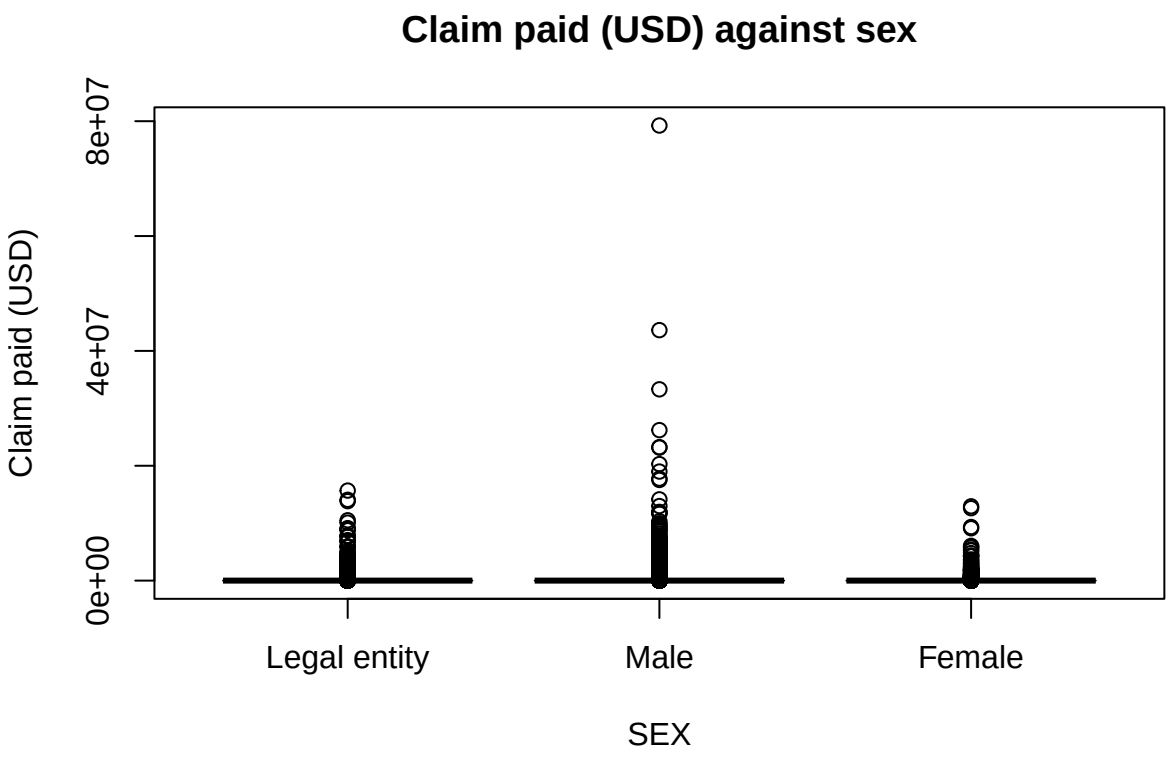

Histogram of INSURED_VALUE
Frequency
30000
20000
10000
0
0.0e+00
5.0e+06
1.0e+07
INSURED_VALUE


Histogram of PREMIUM
Frequency
30000
20000
10000
0
0
50000
100000
150000
PREMIUM


Histogram of CLAIM_PAID_USD
Frequency
100000
75000
50000
25000
0
0e+00
2e+07
4e+07
6e+07
8e+07
CLAIM_PAID_USD


Histogram of AMOUNT_CLAIMS_PAID
Frequency
75000
50000
25000
0
0
1
2
3
4
AMOUNT_CLAIMS_PAID


Premium against sex
Premium
150000
50000
0
Legal entity
Male
Female
SEX



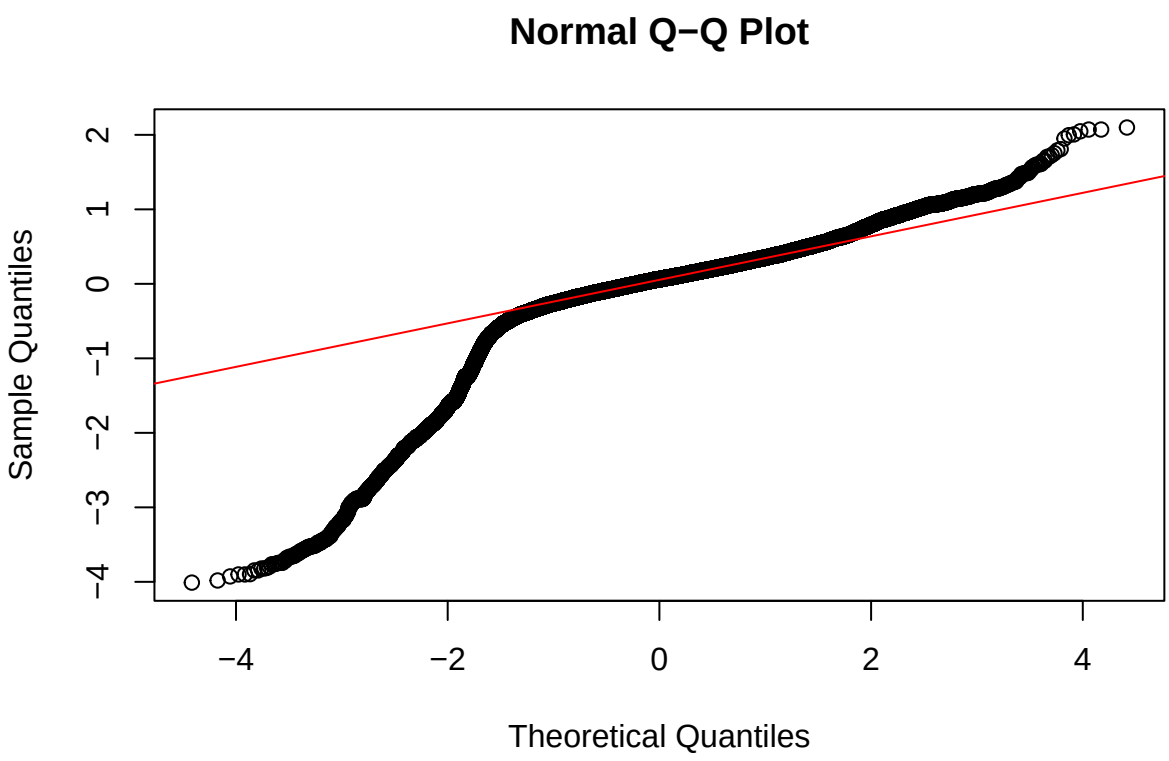
The histograms show that the variables INSURED_VALUE, PREMIUM, CLAIM_PAID_USD and CCM_TON are right-skewed and require a log transformation. The transformed variables will be inserted in the later regression models instead of the original variables.

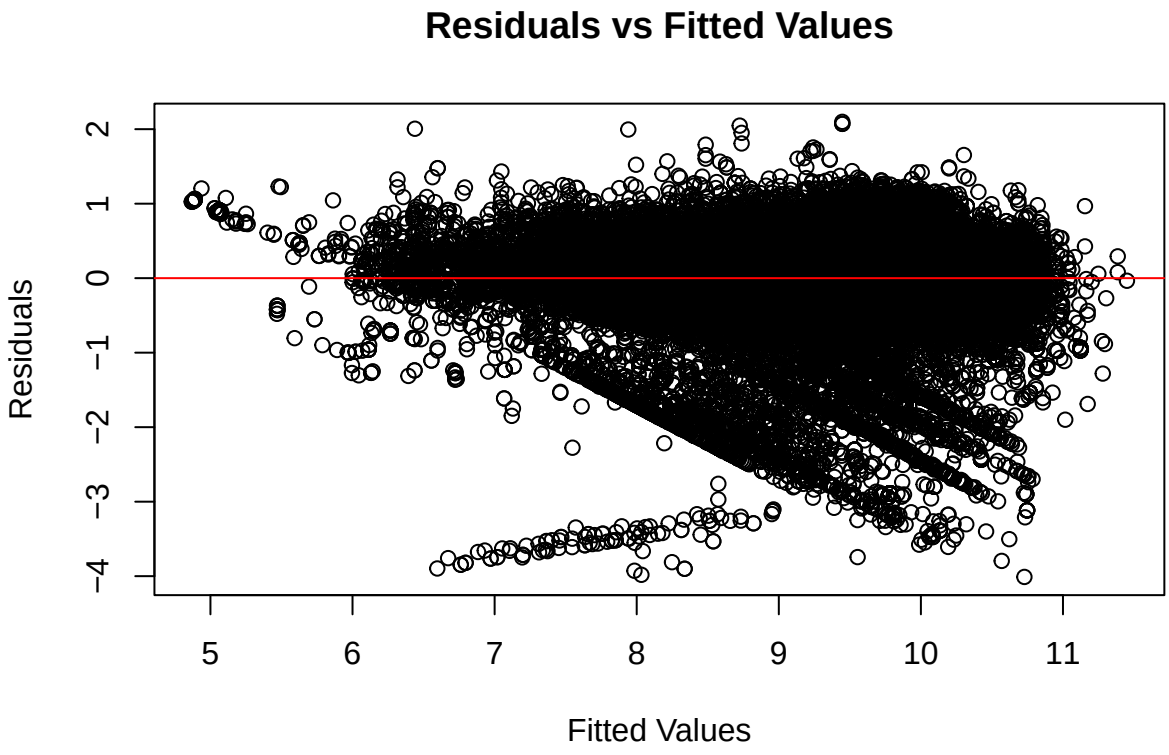
Models

Once the pre-processing was completed and a good overview about the given data and domain knowledge what acquired, the team focused on defining models using several different methods shown in followed subsections.

Linear model

A linear model is adapted, whereby CLAIM_PAID_USD_log was not included, as the premium is incurred at the start of the contract and this would therefore not make technical sense. Instead, a bonus-malus system is taken into account by adding AMOUNT_CLAIMS_PAID.





Mean Squared Error (MSE): 0.2534847 R-squared: 0.7308074

The model summary shows that the Multiple R-squared value is 0.7308, indicating that the model can explain approximately 73.08% of the variance in premiums. This suggests that the model provides a good fit to the data. The F-test for the overall model is significant ($p < 2.2e\text{-}16$), indicating that the predictors as a group have a substantial effect on the premium.

All predictors have a significant impact on the target variable PREMIUM_log. For instance, the categories SEX and USAGE (usage) have a significant effect on PREMIUM_log. Men pay slightly less compared to women, while certain usages, such as "Fare Paying Passengers," lead to higher premiums. In contrast, usages like "Own Goods" and "Private" are associated with lower premiums.

The coefficient of INSURED_VALUE_log (0.7682) in the model shows that the insured value of the vehicle has a strong influence on the premium level. Since both the insured value and the premium are logarithmically transformed, this means that a 1% increase in the insured value results in approximately a 0.7682% increase in the premium. This illustrates the direct and positive relationship between vehicle value and premium: higher-insured vehicles attract proportionally higher premiums, as they represent a greater financial risk for the insurer. Overall, this coefficient confirms that vehicle value is one of the most significant factors in premium calculation.

The coefficient of AMOUNT_CLAIMS_PAID, with a value of 0.1363, indicates that an increase in the number of claims leads to an increase in the log-transformed premium by approximately 0.1363. This means that each additional claim results in a proportional increase in the premium by about 13.63%. This coefficient highlights that an insured's claim history has a significant impact on the premium level.

The coefficient of AGE_VEHICLE is 0.0029, indicating that with each additional year of vehicle age, the log-transformed premium increases by about 0.0029. Since the target variable is logarithmic, this implies that an additional year in vehicle age leads to a minimal increase in the premium of approximately 0.29%.

The coefficient of SEATS_NUM is -0.00175, which means that with each additional seat, the log-transformed premium decreases by approximately 0.00175. Given the logarithmic nature of the target variable, this can be interpreted as each additional seat leading to a slight reduction in the premium by around 0.175%.

The variable CCM_TON_log has a positive coefficient of 0.0109, indicating a small but statistically significant ($p = 0.01241$) relationship with the log-transformed premium (PREMIUM_log). This suggests that an increase in the log-transformed vehicle capacity (CCM_TON) is associated with a slight increase in the premium. However, its relatively small effect size indicates it plays a minor role compared to stronger predictors like AMOUNT_CLAIMS_PAID or INSURED_VALUE_log.

VIF: An analysis of multicollinearity revealed that the Variance Inflation Factor (VIF) for the variable INSR_TYPE is 5.85, which suggests possible multicollinearity. This could affect the model's stability and interpretability and should be considered in further model optimization.

Residuals Analysis Residuals vs. Fitted Plot: The Residuals vs. Fitted Plot displays a funnel-shaped pattern, indicating heteroskedasticity. The variance of the residuals increases with higher predicted values, meaning that the model is less accurate for larger premium values. This violates the assumption of constant variance, suggesting that homoskedasticity is not fully met.

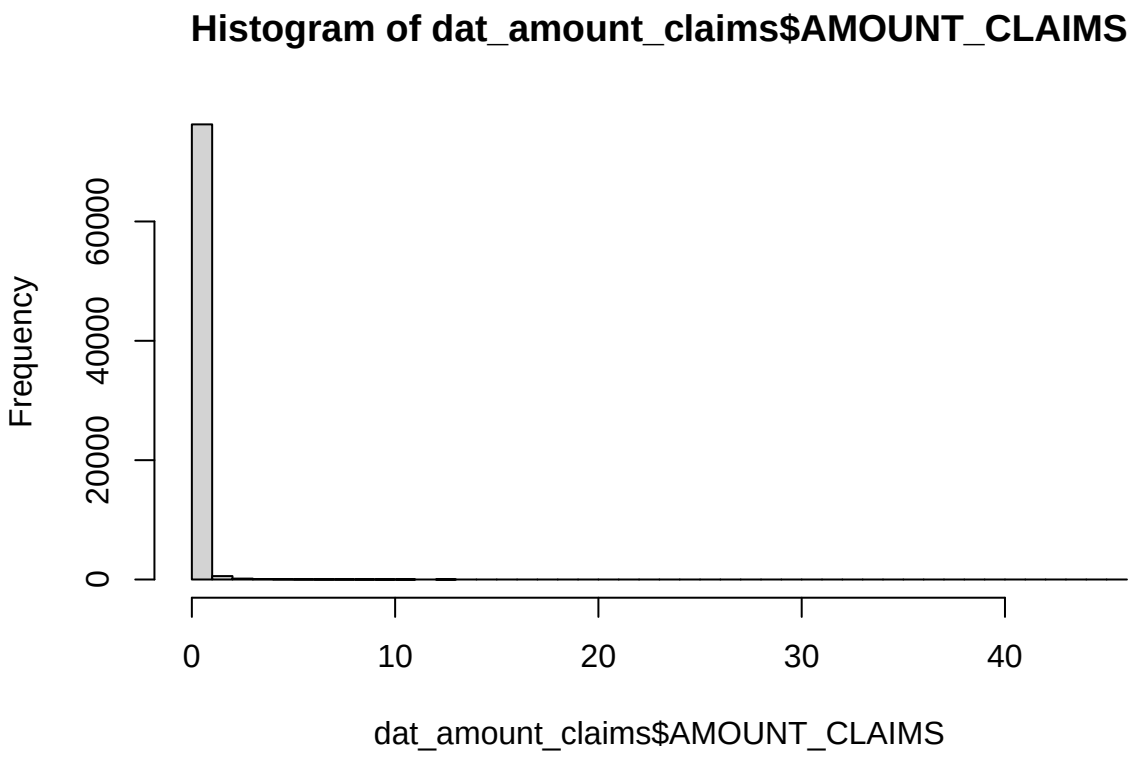
Normal Q-Q Plot: The Normal Q-Q Plot shows that the residuals do not lie perfectly along the line, indicating significant deviations from the theoretical normal distribution, particularly at the tails. These "heavy tails" suggest a non-normal distribution of residuals, potentially due to outliers or unmodeled non-linear relationships.

To improve the model, various measures could be considered. One approach would be to transform the target variable, for example, using a Box-Cox transformation, to reduce heteroskedasticity and achieve a more stable residual variance. Additionally, incorporating non-linear relationships by including polynomial terms or using a generalized linear model (GLM) could be beneficial. This would allow the model to better capture complex relationships between variables, thereby enhancing predictive accuracy.

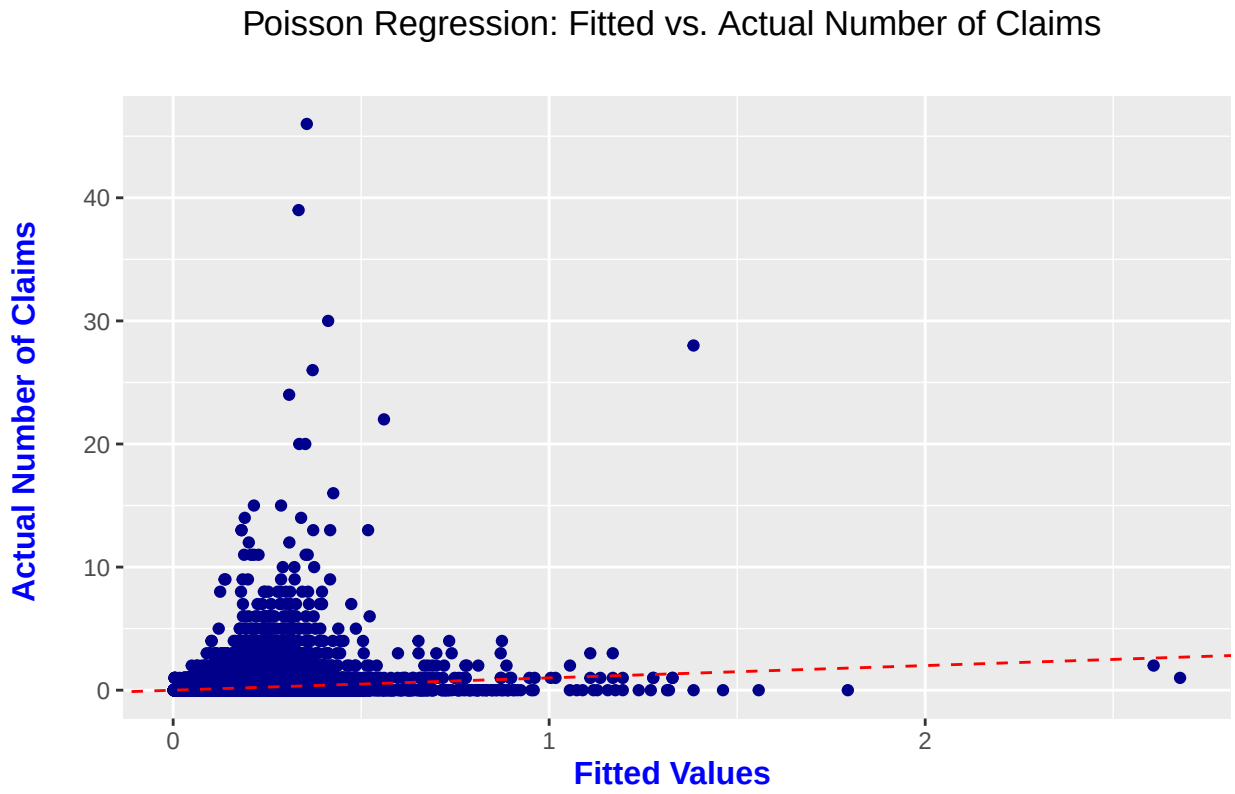
Poisson model

A Poisson model is fitted to predict the number of claims over a 5-year period based on the characteristics SEX, INSR_TYPE, USAGE, TYPE_VEHICLE, MAKE, AGE_VEHICLE, SEATS_NUM, CCM_TON, INSURED_VALUE, and PREMIUM.

First, the data is grouped accordingly, and the results are analyzed to gather insights.



The analysis of the distribution of the target variable AMOUNT_CLAIMS reveals that a large portion of the values are zero. This concentration of zero values is confirmed by the median, as well as the 1st and 3rd quartiles, which are also at zero. Additionally, the distribution shows some high outliers with a maximum value of 46, indicating an uneven distribution with a few high values. The low mean of 0.1791 further supports this observation, suggesting a significant number of zero values. Given these distribution characteristics, the use of a Zero-Inflated Poisson (ZIP) model could be appropriate, as such a model can account for both random and structural zeros. Initially, however, a Poisson model will be fitted.



The Poisson model analysis shows no signs of overdispersion, as the overdispersion value of 0.688 is below 1. This indicates a possible underdispersion, but this is confirmed by the goodness-of-fit test (p-value = 1), which confirms a well-fitting model.

The Poisson regression model for predicting the number of claims reveals that several variables show statistically significant relationships with claim frequency. The model indicates statistically significant differences in claim frequency across categories (p-value < 0.001). The group of legal entities, which serves as the reference category, exhibits the highest claim rate. Compared to legal entities, males have a rate ratio of 0.666, reflecting a 33.4% lower claim rate, while females have the lowest claim frequency, with a rate ratio of 0.623, or 37.7% below that of legal entities.

For the insurance type (INSR_TYPE), it was found that INSR_TYPEPrivate has a rate ratio of 1.284, indicating that private insurers have a 28.4% higher claim probability compared to the reference category INSR_TYPECommercial. The variables TYPE_VEHICLE and MAKE also show significant differences in claim rates. Among vehicle types, Pick-up has the highest claim rate, with a rate ratio of 1.121, representing a 12.1% increase in claim probability compared to the reference category Automobile; however, this effect is not statistically significant (p-value = 0.120). Conversely, Motor-cycle has the lowest claim rate, with a rate ratio of 0.039, indicating an approximately 96% reduced claim probability and a highly significant result (p-value < 0.001).

Among vehicle brands, GEELY shows the highest claim rate with a rate ratio of 1.040, which, however, represents no meaningful change compared to the reference brand BISHOFTU and is statistically insignificant (p-value = 0.709). Conversely, MERCEDES has the lowest claim rate, with a rate ratio of 0.403, indicating a 59.7% lower claim probability compared to BISHOFTU and is highly significant (p-value < 0.001).

These results suggest that Pick-up and GEELY exhibit the highest, though statistically insignificant, claim rates, while Motor-cycle and MERCEDES show the lowest and statistically significant claim rates relative to their respective reference categories.

Further analysis indicates that vehicle age (AGE_VEHICLE) has a rate ratio of 0.956, meaning that the

claim rate decreases by approximately 4.4% with each additional year (p-value < 0.001). The number of seats (SEATS_NUM) shows a rate ratio of 1.009, indicating that each additional seat slightly increases the claim probability, though significantly. Engine capacity (CCM_TON) shows no practical change in claim rate with a rate ratio of 1.000031, though it is statistically significant (p-value < 0.001). Insured value (INSURED_VALUE) has a rate ratio of 0.99999986, effectively showing no influence on claim frequency, although the effect is statistically significant. Premium amount (PREMIUM) exhibits a rate ratio of 1.000019, suggesting a minimal increase in claim probability with rising premiums; again, the effect is significant but very small.

The VIF values of all predictors are below the critical limit of 5, which indicates that there are no multicollinearity problems. The variables INSR_TYPE (3.75) and TYPE_VEHICLE (1.42) have the highest values, which indicates a moderate correlation with other predictors, but does not cause any stability problems in the model. Overall, the low VIF values support the robustness of the model estimates.

The plot illustrates the relationship between the predicted values (Fitted Values) and the actual number of claims (Actual Number of Claims). Most actual values are concentrated in the lower range (0 to 10), while the predicted values are almost entirely clustered near zero. The model struggles to predict higher claim counts (>10), as evident from the significant deviations for extreme values. The red dashed line, representing the ideal fit between predicted and actual values, shows that many points fall below the line, indicating a systematic underestimation of actual claims by the model. High variability or extreme values in the data can cause the model to perform poorly in capturing such cases.

The plots of estimated vs. actual values show that the Poisson model has difficulties in accurately modelling the distribution of claims, especially for higher claims values. Most of the predicted values are close to zero and systematically underestimate the actual loss frequencies as they increase. This systematic underestimation and the high number of zero claims indicate that the simple distribution of the Poisson model may not be sufficient to fully represent the structure of the data.

Given the high number of zero values in the data, a Zero-Inflated Poisson (ZIP) model could represent a useful alternative. Such a model can distinguish between structural zeros (cases where no claims occur) and random zeros (cases where claims could occur but did not), potentially improving predictive accuracy for higher claim counts without violating model assumptions about variance. As a further alternative, simplifying the model, for example by removing fewer significant variables, could be a sensible measure to improve the model.

Binomial model

`Pseudo.R.squared`

McFadden 0.0337310 Cox and Snell (ML) 0.0267356 Nagelkerke (Cragg and Uhler) 0.0484169

The analysis shows that several variables have significant associations with the likelihood of CLAIMS_PAID = Yes. According to the likelihood ratio test, the variable INSR_TYPE does not appear to have a significant impact.

Gender emerges as a key predictor: men have a 20.5% lower likelihood of receiving a claim payout compared to the reference group (legal entities), while women have a 22.4% reduced likelihood. These effects are highly significant ($p < 0.001$).

The vehicle type also shows significant differences. Motorcycles have approximately a 96% lower likelihood of claim payouts, making them the group with the lowest payout frequency. Similarly, station wagons and trailers exhibit significantly lower probabilities, with reductions of around 20% and 82%, respectively. In contrast, pick-ups do not show significant differences compared to the reference category (automobiles).

The vehicle brand further influences payout frequency. Vehicles from the brand DAEWOO have a 38% lower likelihood of receiving a claim payout compared to the reference brand, FIAT shows a reduction of 46%, and

MERCEDES has a 51% lower likelihood. These effects are all statistically significant. Other brands, such as GEELY or MAZDA, do not show significant differences compared to the reference category.

The premium amount proves to be one of the strongest predictors. An increase in the logarithmic premium amount leads to a significant 71% increase in the likelihood of a claim payout. Conversely, a higher logarithmic insured sum reduces the payout likelihood by 25%. The claim amount also positively correlates with payout frequency: higher claim amounts significantly increase the likelihood of payouts.

The age of the vehicle has a negative effect on payout frequency: with each additional year, the likelihood of a payout decreases by approximately 2.9%. Other predictors, such as the insurance type or vehicle usage, do not have significant effects on the likelihood of a claim payout and could be excluded in a simplified model.

The pseudo-R^2 values (McFadden: 3.4%; Nagelkerke: 4.8%) indicate that the model explains only a small portion of the variance in the target variable, suggesting limited model performance. Nevertheless, the global test statistic (TD = 2709.954, $p < 0.001$) is highly significant, confirming that the model performs significantly better than a null model.

The overdispersion ratio (deviance / degrees of freedom) is 0.7766, which is less than 1, indicating that no overdispersion is present in the model. This suggests that the assumptions of the binomial model regarding data dispersion are met.

The AIC value of the logistic regression model is 77716.2.

Method 1): Removal of (non-significant) explanatory variables

The analysis shows that some variables, such as SEX, TYPE_VEHICLE, MAKE, AGE_VEHICLE, INSURED_VALUE_log, PREMIUM_log, and AMOUNT_CLAIMS_PAID, are highly significant ($p < 0.001$). These variables make a substantial contribution to explaining the variance. On the other hand, certain categories, particularly within the variables USAGE and MAKE, show low significance, indicating a limited explanatory effect of these predictors.

The model has an AIC value of 77714.2, showing only minimal improvement. However, the comparison of the null deviance (80340) with the residual deviance (77630) indicates a significant improvement over the null model. This is confirmed by the Likelihood-Ratio Test (LRT). The test comparing the model with and without the variable INSR_TYPE shows no significant difference ($p = 0.945$), suggesting that INSR_TYPE has no significant impact and can therefore be excluded.

The pseudo-R^2 values (e.g., Nagelkerke: 0.0484) indicate limited explanatory power for the model. This is supported by the AUC (Area Under the Curve) of 0.629, reflecting a low to moderate discriminatory ability. The model is only slightly better than random at distinguishing between positive and negative cases.

The analysis of deviance residuals shows a distribution close to zero, but with a maximum of 3.5154, which could indicate potential outliers. The overdispersion ratio (deviance/DF) is 0.777, suggesting no significant overdispersion in the model.

The ROC curve confirms the moderate discriminatory ability of the model. With an AUC value of 0.629, the curve demonstrates that the model predicts the target variable better than random probabilities, though it lacks strong predictive power.

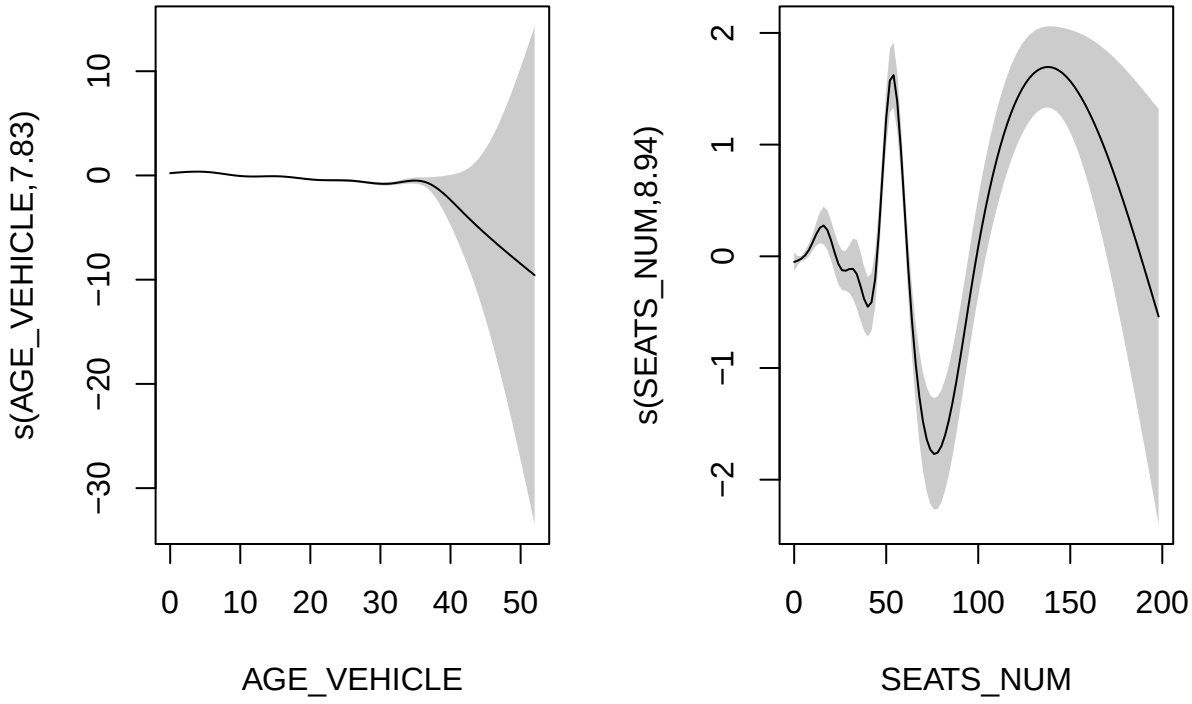
In summary, the model shows some degree of significance and stability but offers only moderate explanatory power and discriminatory ability.

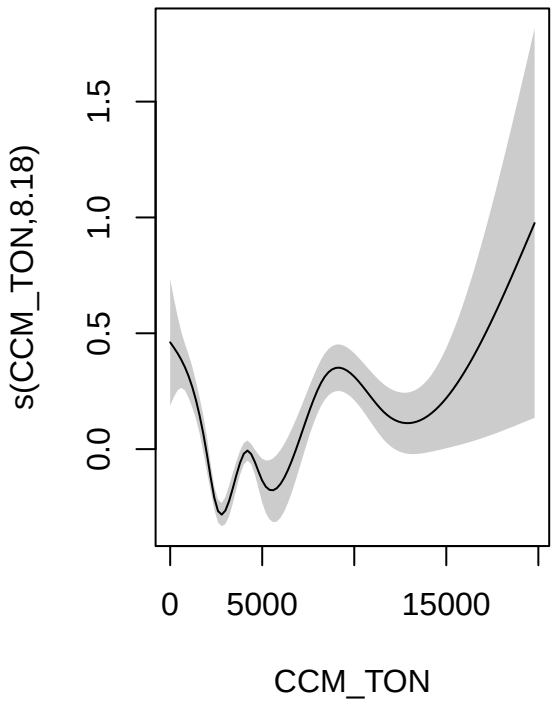
The imbalanced distribution of classes (CLAIMS_PAID: “YES”/“NO”) in the dataset significantly impacts the model’s discriminatory ability. The moderate AUC of 0.629 indicates that the model struggles to reliably recognize the minority class (“YES”). To improve discriminatory performance, balancing strategies such as oversampling, class weighting, or adjusting the decision threshold could be implemented. Additionally, exploring interaction effects might further enhance the model’s performance.

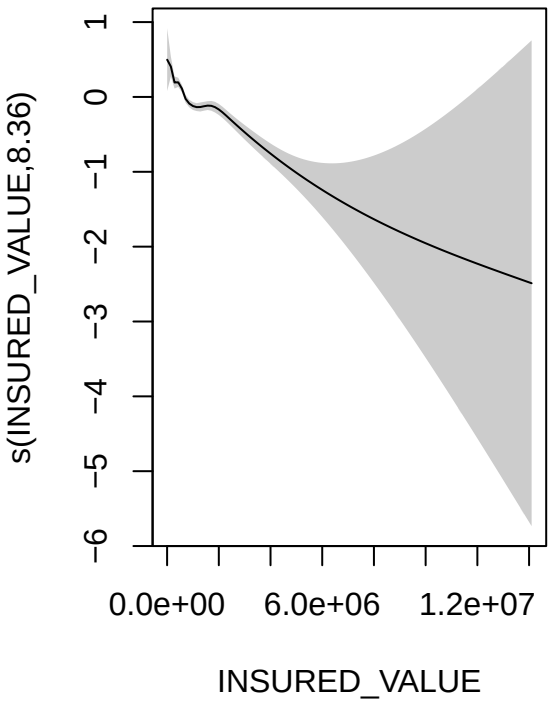
Generalised Additive Model (GAM)

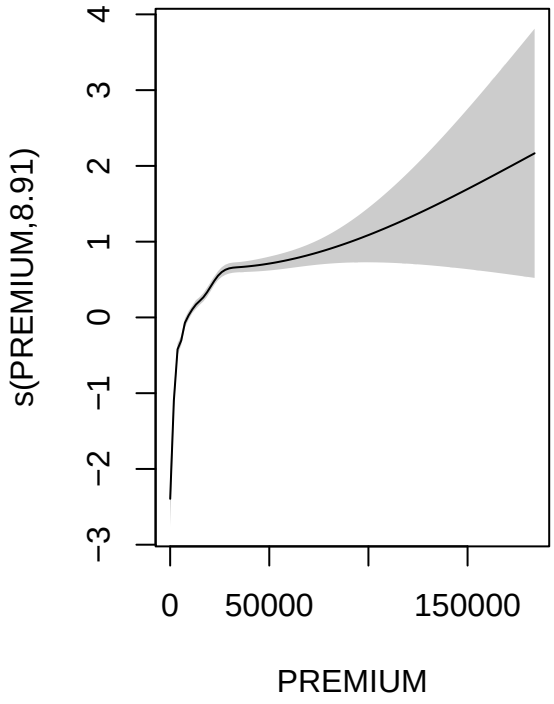
A General Additive Model (GAM) is fitted to predict the number of claims based on the characteristics SEX, INSR_TYPE, USAGE, TYPE_VEHICLE, MAKE, AGE_VEHICLE, SEATS_NUM, CCM_TON, INSURED_VALUE, and PREMIUM. Similar to the Poisson model, the GAM model aims to capture the relationship between the predictors and the number of claims, allowing for non-linear relationships and interactions between variables. AUC value of the GAM model: 0.6175357 RMSE of the GAM model: 0.5938353

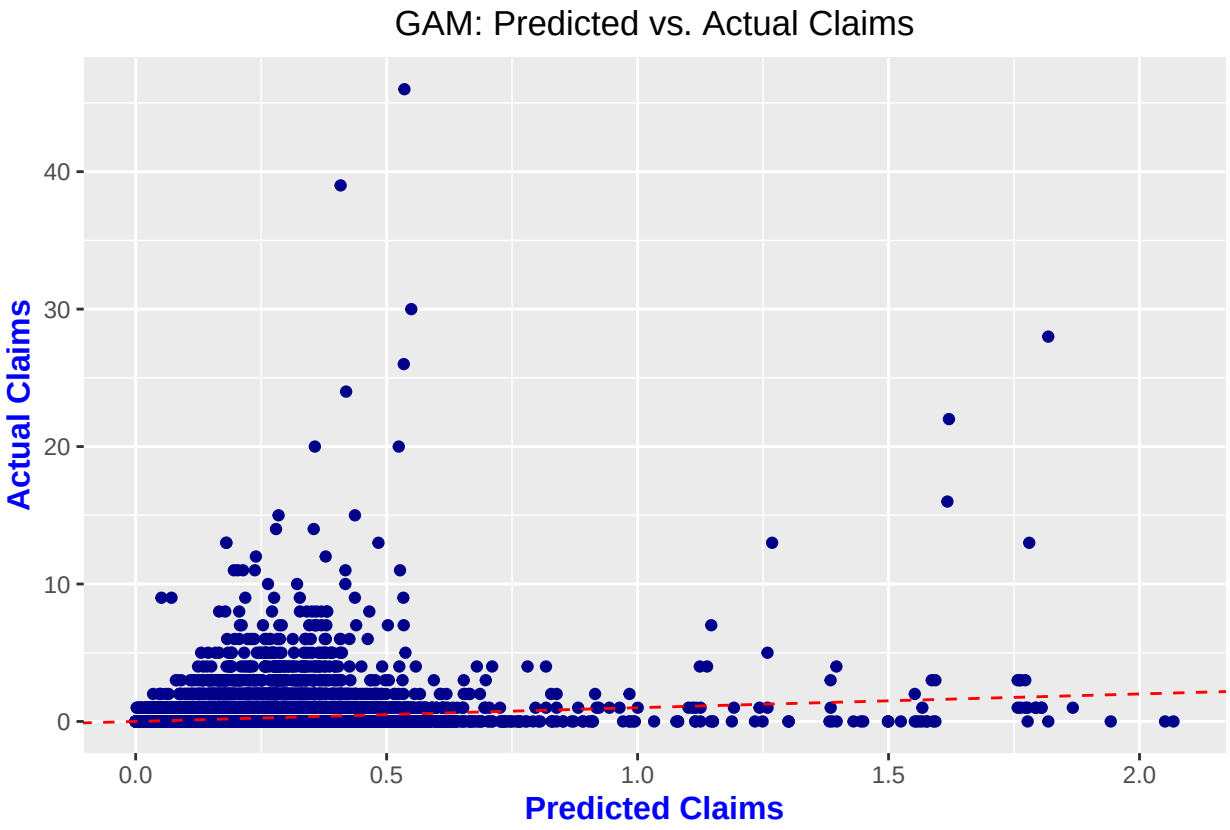
Different variations were used, with and without smoothing predictos and using B-splines with cubic regression. Cubic regression was chosen as it provided the best fit for the data with the smallest RMSE value, all values where ~0.02 of difference. The AUC value of the GAM model is 0.61, very similar to the results in the Poisson model. This indicates that the GAM model has a moderate ability to discriminate between the number of claims and the predictors. The RMSE value of the GAM model is 0.59, which is relatively low and indicates that the model's predictions are close to the actual values.

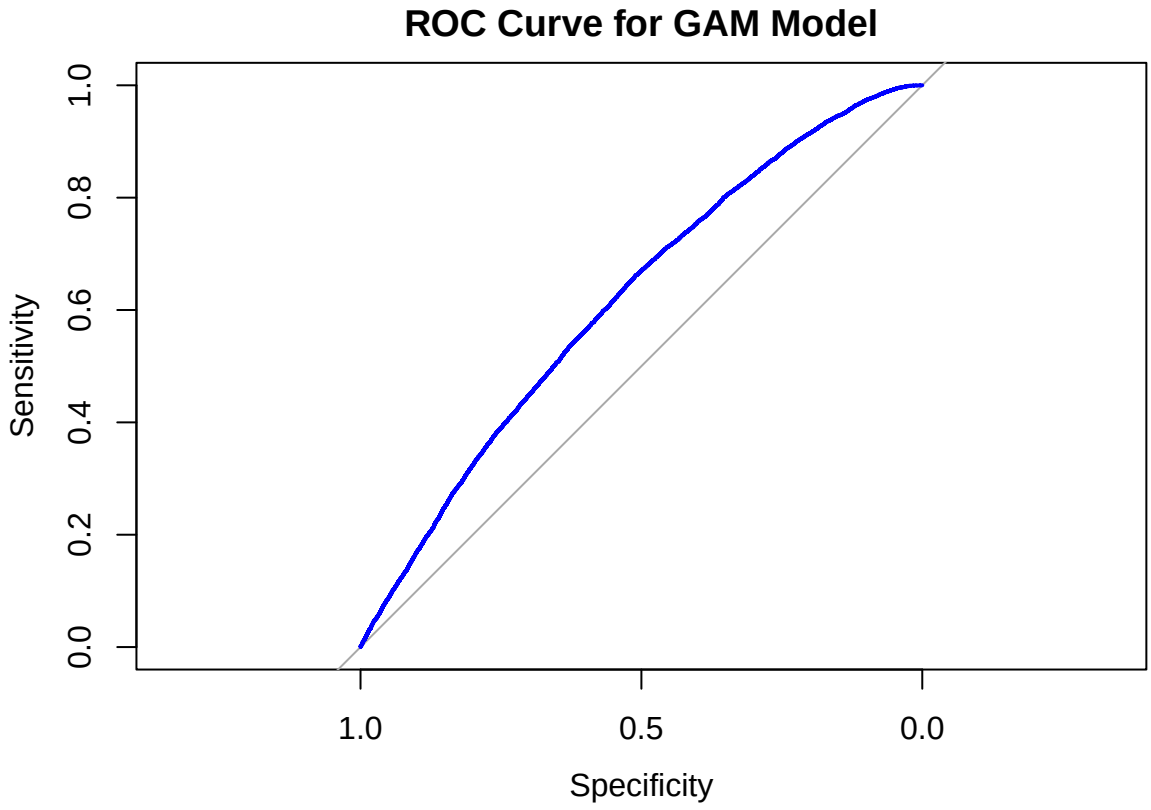



s(CCM_TON,8.18)
0.0
0.5
1.0
1.5
0
5000
15000
CCM_TON


s(INSURED_VALUE,8.36)
−6
−5
−4
−3
−2
−1
0
1
0.0e+00
6.0e+06
1.2e+07
INSURED_VALUE


s(PREMIUM,8.91)
−3
−2
−1
0
1
2
3
4
0
50000
150000
PREMIUM


GAM: Predicted vs. Actual Claims
Actual Claims
Predicted Claims
0
10
20
30
40
0.0
0.5
1.0
1.5
2.0


ROC Curve for GAM Model
Sensitivity
Specificity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0

Some interesting plots are shown, to illustrate how the predictors are related to the number of claims. The plots show the smooth functions of the predictors AGE_VEHICLE, SEATS_NUM, CCM_TON, INSURED_VALUE, and PREMIUM. The plots illustrate the non-linear relationships between these predictors and the number of claims, capturing the complex interactions and patterns in the data.

For AGE_VEHICLE, is steady until 40 years where id declines rapidly, indicating that older vehicles have fewer claims.

For SEATS_NUM, the number of claims increases with the number of seats related to "consumer vehicles", up to ~15 seats, where it starts to decline. There is a spike about 50 seats, probably related to commercial vehicles. The next valley and peak are related to very high number of seats, which are outliers and related to commercial or custom vehicles.

For CCM_TON, the number of claims decreases with the engine capacity untill 2500cc, which is where most of the vehicles are, from motorcicles and utility cars. After that, the number of claims increases, probably related to commercial and sport vehicles. Around 5000cc is a common engine size for commercial vehicles such as busses and trucks, after that the number of claims increases rapidly.

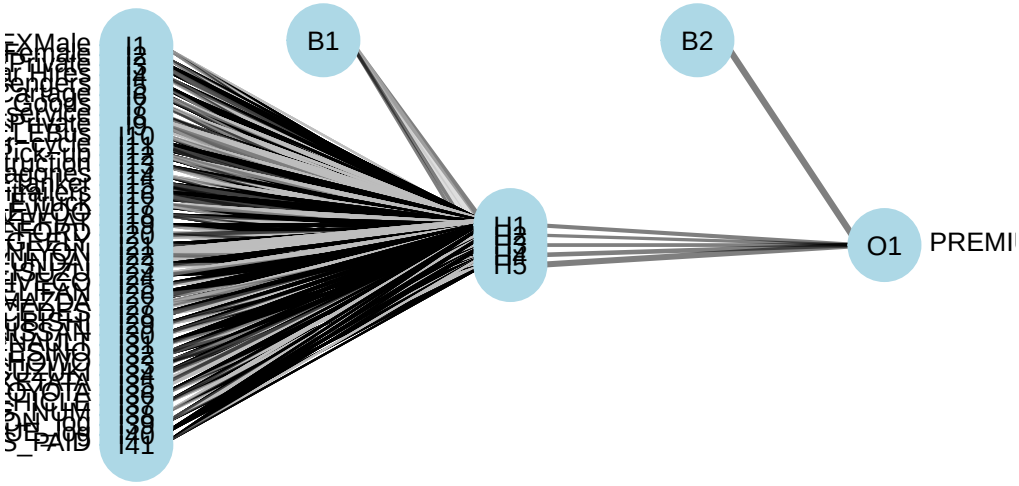
For INSURED_VALUE, the number of claims decreases with the insured value, indicating that more expensive vehicles have fewer claims. This is expected as more expensive vehicles are usually driven more carefully and less often.

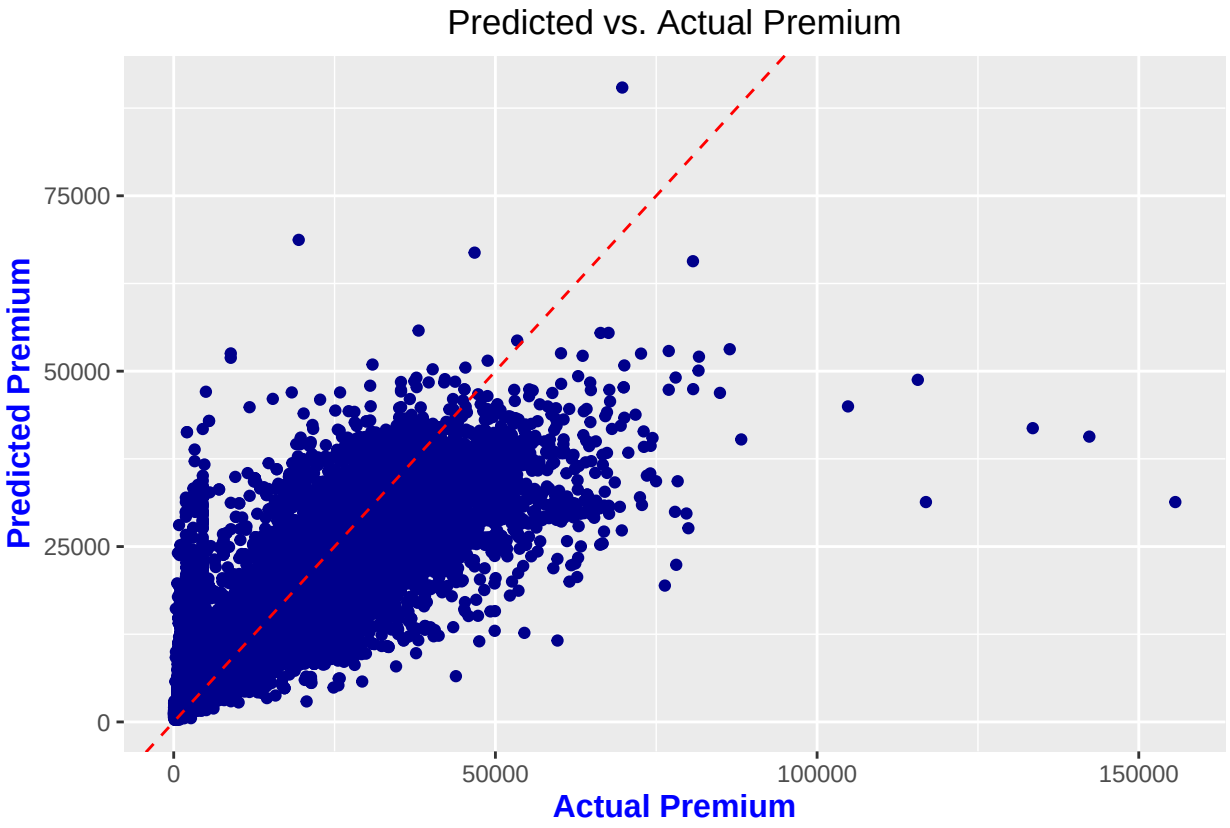
For PREMIUM, the number of claims increases with the premium amount, starting in negative for the cheapest premium values, probably related to the insurance type and the insured value. After that, the number of claims increases with the premium amount, indicating that higher premiums are associated with more claims. This could be due to higher premiums for higher-risk drivers or vehicles or simply due to the increased value of the insured vehicles and the related need to keep it in good condition, making smaller defects a claim, which would not be claimed in cheaper vehicles.

Neural Network

Lead: Alvaro Cervan

A neural network model is fitted to predict the premium amount based on the characteristics of the insured vehicles and the driver. The model is trained using the cleaned and transformed data, and the results are analyzed to evaluate the model's performance.


B1
B2
H1
H2
H3
O1
PREMI

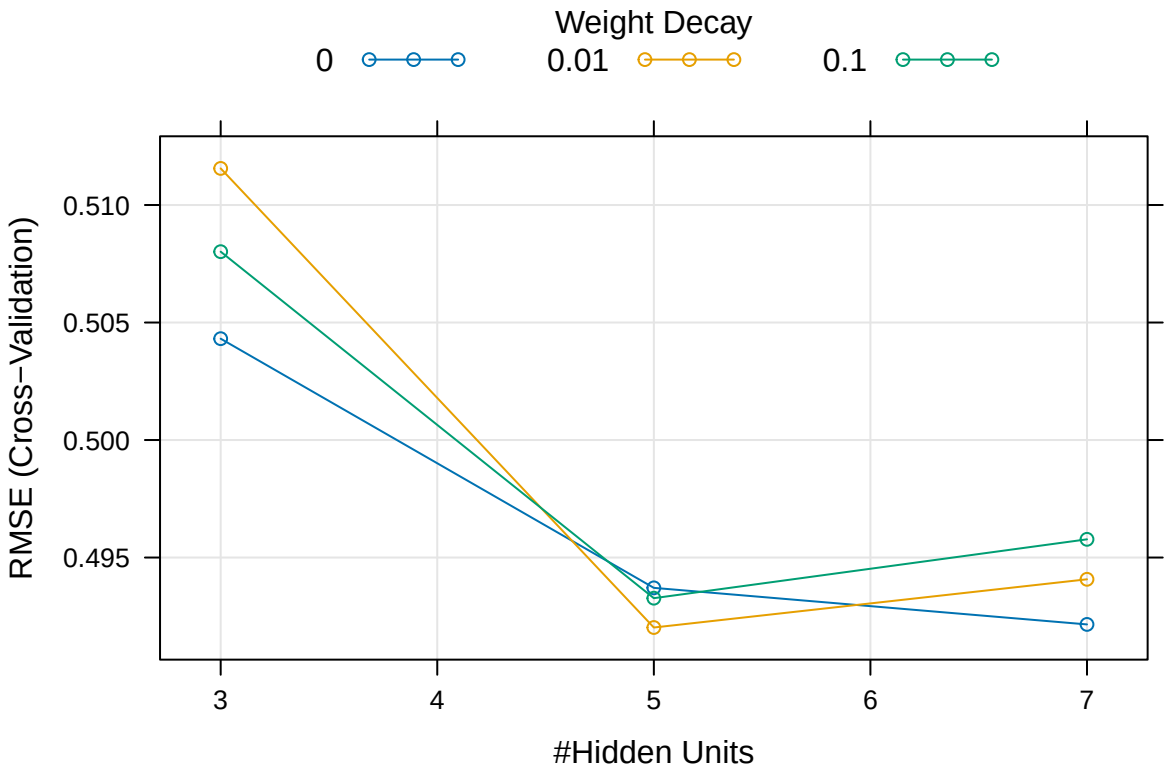

Predicted vs. Actual Premium
Predicted Premium
Actual Premium
0
25000
50000
75000
0
50000
100000
150000

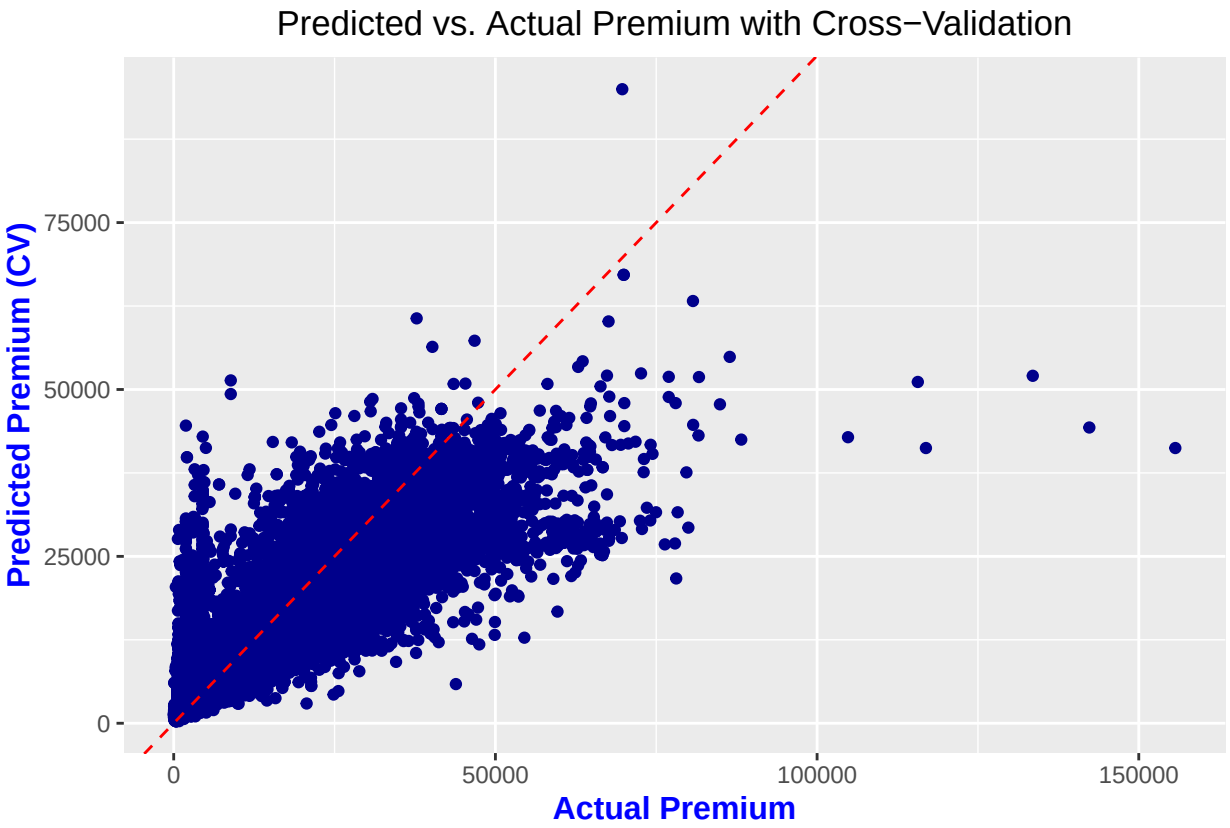
The neural network model was trained to predict the log-transformed premium amount based on the characteristics of the insured vehicles. The model was fitted using the nnet package, with the training data split into 80% training and 20% testing sets. The neural network model was trained with a hidden layer size of 5 neurons, linear output for regression tasks, and a maximum of 100 iterations for training.

The plots of predicted vs. actual premium amounts show that the neural network model generally performs well in predicting the premium amounts. The points are clustered around the diagonal line, indicating a good alignment between the actual and predicted values. The model captures the general trend of the premiums, with some deviations for higher premium values. The model's performance can be further evaluated by considering additional metrics such as the R-squared value, RMSE, MAE, and MAPE, which provide insights into the model's accuracy and predictive power.

Neural Network Cross Validation

Nevertheless, we cannot be sure that those values for the model above are truly correct or it was luck that the model performs well at a first instance. To solve this question, the NN will be run again using **k-fold Cross Validation** with hyperparameter tuning. This approach will help ensure that the model's performance is robust and not due to overfitting or random chance. The k-fold Cross Validation will produce a more reliable estimate of the model's performance by splitting the data into k = 10 subsets, training the model on k-1 subsets, and validating it on the remaining subset. This process is repeated k times, and the results are averaged to provide a comprehensive evaluation of the model.





Results

Model	Mean Squared Error (MSE)	R-squared	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	Mean Absolute Percentage Error (MAPE)
Neural Network	0.2345529	0.7510002	0.4843067	0.3009332	3.488952 %
Neural Network Cross Validation	0.2343864	0.751177	0.4841347	0.2981682	3.462526 %

The weights decay plot shows how the RMSE (Root Mean Square Error) from cross-validation varies with the number of hidden units in a neural network for different weight decay values (0, 0.01, and 0.1). As the number of hidden units increases from 3 to 5, RMSE decreases across all weight decay values, suggesting improved model accuracy with additional capacity. However, beyond 5 hidden units, RMSE levels off or slightly increases, especially when weight decay is low or absent, indicating potential overfitting. Weight decay, a regularization technique to prevent overfitting, has a noticeable effect as the number of hidden units increases; while it slightly raises RMSE at lower hidden units, it helps to control error at higher hidden units. The optimal configuration, with the lowest RMSE, occurs at 5 hidden units regardless of weight decay, though weight decay of 0.1 becomes more beneficial as the model complexity increases, particularly at 6 and 7 hidden units.

The results from both the neural network and the neural network with cross-validation are very similar, with only minor differences in the evaluation metrics. This consistency suggests that the model is robust

and performs well regardless of the validation method used. The cross-validation approach confirms the reliability of the neural network model, indicating that it is not overfitting and generalizes well to unseen data.

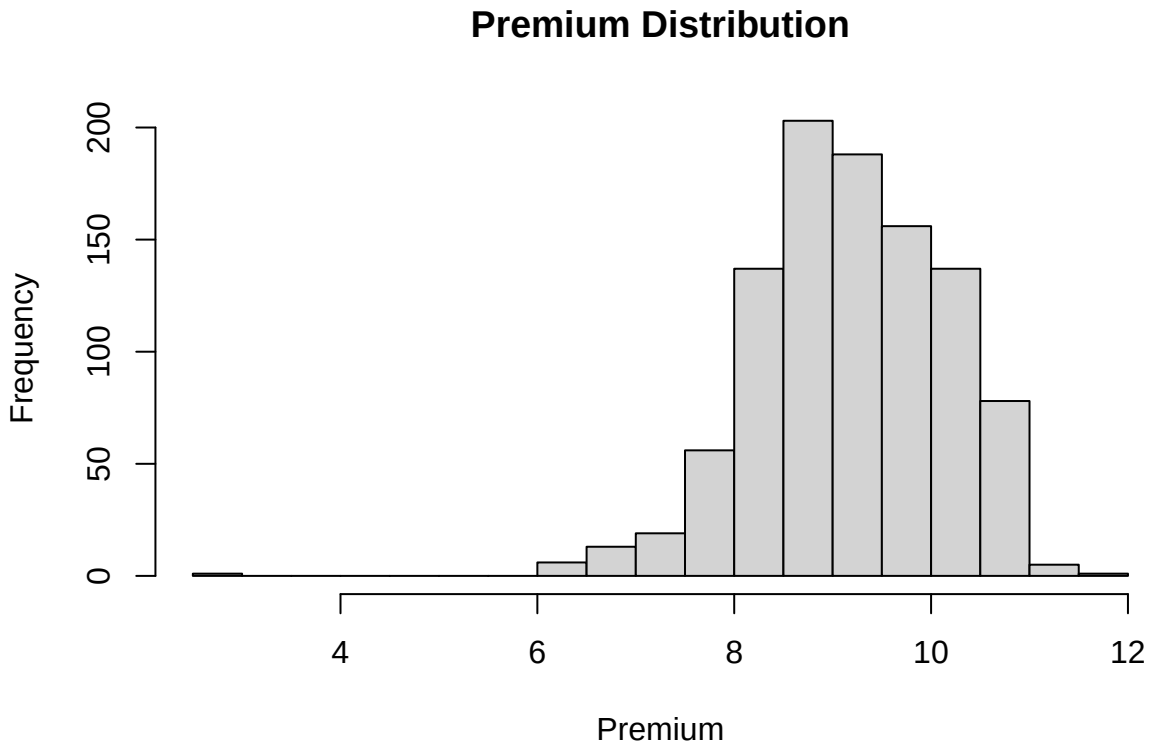
The evaluation of the neural network model revealed a Mean Squared Error (MSE) of 0.23, indicating the average squared difference between the actual and predicted log-transformed premium amounts. The R-squared value of 0.75 suggests that the model can explain approximately 75.10% of the variance in the log-transformed premiums, indicating a good fit to the data. The Root Mean Squared Error (RMSE) of 0.48 represents the square root of the MSE, providing a measure of the model's prediction accuracy. The Mean Absolute Error (MAE) of 0.30 indicates the average absolute difference between the actual and predicted log-transformed premiums. The Mean Absolute Percentage Error (MAPE) of 3.46% represents the average percentage difference between the actual and predicted premiums, providing a measure of the model's relative accuracy.

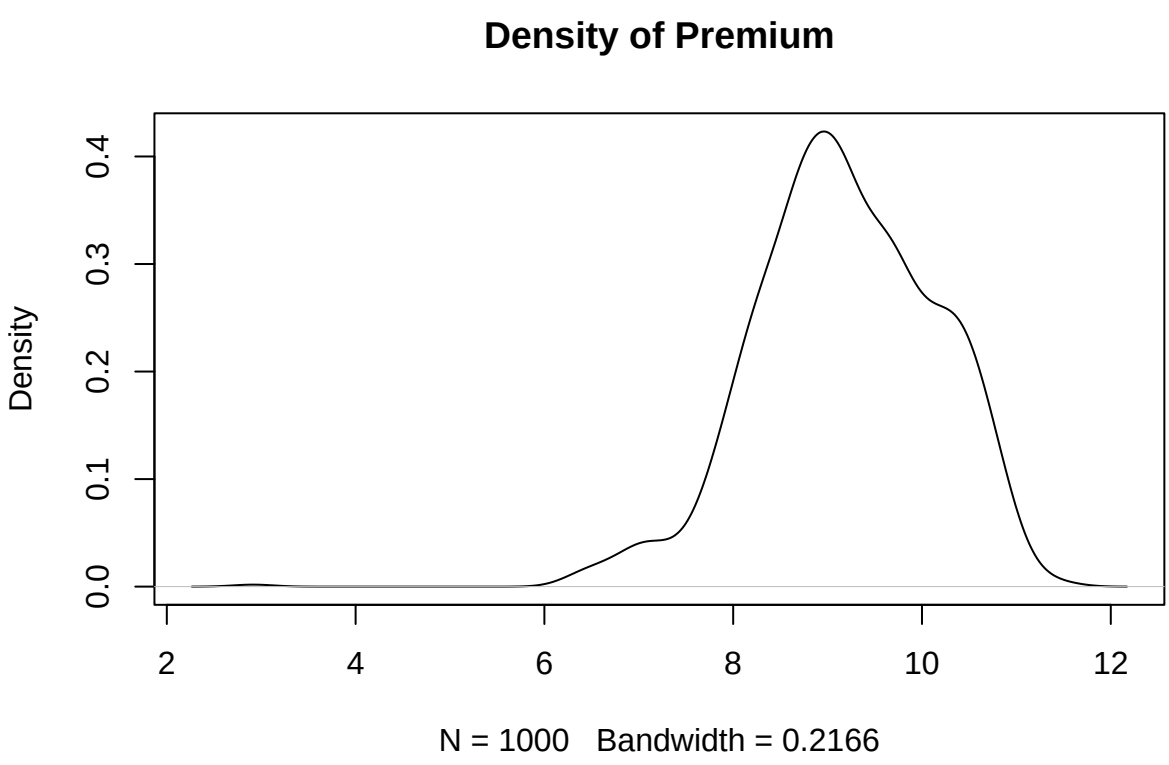
Overall, the neural network model demonstrates good performance in predicting the premium amounts based on the characteristics of the insured vehicles and drivers. The model captures the underlying patterns in the data and provides accurate predictions of the premium amounts. The evaluation metrics indicate that the model has a high level of accuracy and predictive power, which could make it a valuable tool for premium prediction in the insurance industry.

Support Vector Machine (SVM)

Lead: Luca Renz

For the scenario of SVM models, it has been decided to do multiple-classifications for the premiums and divide it into 4 levels from low to very high.





low medium high very_high 250 250 250 250 sigma C 7 0.01 10 Support Vector Machine object of class "ksvm"

SV type: C-svc (classification) parameter : cost C = 10

Gaussian Radial Basis kernel function. Hyperparameter : sigma = 0.01

Number of Support Vectors : 511

Objective Function Value : -1621.026 -807.8499 -396.511 -1517.487 -437.1526 -1202.273 Training error : 0.222857 Support Vector Machines with Radial Basis Function Kernel

700 samples 7 predictor 4 classes: 'low', 'medium', 'high', 'very_high'

No pre-processing Resampling: Cross-Validated (10 fold, repeated 3 times) Summary of sample sizes: 631, 631, 632, 629, 629, 631, ... Resampling results across tuning parameters:

C sigma Accuracy Kappa
0.1 0.01 0.5231537 0.3656084 0.1 0.10 0.5401605 0.3869747 0.1 0.50 0.5430138 0.3910158 1.0 0.01 0.6389542 0.5185779 1.0 0.10 0.6249998 0.4999076 1.0 0.50 0.6073559 0.4764562 10.0 0.01 0.6874802 0.5832099 10.0 0.10 0.6585251 0.5446075 10.0 0.50 0.6339281 0.5119300

Accuracy was used to select the optimal model using the largest value. The final values used for the model were sigma = 0.01 and C = 10. [1] "Confusion metrics for TEST_DATA" Confusion Matrix and Statistics

```
          Reference
```

Prediction low medium high very_high low 133 15 0 1 medium 28 132 29 1 high 9 25 128 22 very_high 5 3 18 151

Overall Statistics

```
           Accuracy : 0.7771
             95% CI : (0.7445, 0.8075)
No Information Rate : 0.25
P-Value [Acc > NIR] : < 2.2e-16

              Kappa : 0.7029
```

Mcnemar's Test P-Value : 0.008264

Statistics by Class:

```
                     Class: low Class: medium Class: high Class: very_high
```

Sensitivity 0.7600 0.7543 0.7314 0.8629 Specificity 0.9695 0.8895 0.8933 0.9505 Pos Pred Value 0.8926 0.6947 0.6957 0.8531 Neg Pred Value 0.9238 0.9157 0.9089 0.9541 Prevalence 0.2500 0.2500 0.2500 0.2500 Detection Rate 0.1900 0.1886 0.1829 0.2157 Detection Prevalence 0.2129 0.2714 0.2629 0.2529 Balanced Accuracy 0.8648 0.8219 0.8124 0.9067 [1] "MCC for Train Data: 0" [1] "MCC Train manually calculated: 0.7238" [1] "Confusion metrics for TEST_DATA" Confusion Matrix and Statistics

```
          Reference
```

Prediction low medium high very_high low 54 11 0 2 medium 17 48 12 1 high 2 14 49 17 very_high 2 2 14 55

Overall Statistics

```
           Accuracy : 0.6867
             95% CI : (0.6309, 0.7387)
No Information Rate : 0.25
P-Value [Acc > NIR] : <2e-16

              Kappa : 0.5822
```

Mcnemar's Test P-Value : 0.6681

Statistics by Class:

```
                     Class: low Class: medium Class: high Class: very_high
```

Sensitivity 0.7200 0.6400 0.6533 0.7333 Specificity 0.9422 0.8667 0.8533 0.9200 Pos Pred Value 0.8060 0.6154 0.5976 0.7534 Neg Pred Value 0.9099 0.8784 0.8807 0.9119 Prevalence 0.2500 0.2500 0.2500 0.2500 Detection Rate 0.1800 0.1600 0.1633 0.1833 Detection Prevalence 0.2233 0.2600 0.2733 0.2433 Balanced Accuracy 0.8311 0.7533 0.7533 0.8267 [1] "MCC for Test Data: 0" [1] "MCC manually calculated: 0.5717"

This analysis explores the performance of a multiclass classification task using an SVM model with a radial kernel. The goal was to classify PREMIUM_log into categories: "low," "medium," "high," and "very_high," employing hyperparameter tuning and parallel computation for efficiency. Data preparation included sampling 1000 entries from the cleaned dataset and visualizing the distribution of PREMIUM_log, followed by a 70/30 split for training and testing. It is important to note that the dataset is imbalanced, which may pose challenges in model training and evaluation, potentially impacting the reliability of certain performance metrics.

Model training was carried out with 10-fold cross-validation repeated three times, leveraging parallel processing to speed up the evaluation. A grid search was performed to find optimal values for the hyperparameters c (cost) and sigma, ensuring the model's robustness and generalizability.

The evaluation showed that the "very_high" category had the highest sensitivity at 0.8629, while the "low" category excelled in specificity at 0.9695 and positive predictive value at 0.8926. The MCC for each class revealed strong performance overall: ~0.986 for "low," ~0.802 for "medium," ~0.781 for "high," and ~1.038 for "very_high," though the latter may indicate overestimation and warrants further review.

The code process incorporated parallelized cross-validation and grid search, facilitating comprehensive hyperparameter tuning. The findings highlighted an overall accuracy of 0.7771 and a Kappa statistic of 0.7029, with McNemar's test yielding a significant P-value of 0.008264 and mcc-value of roughly 0.57, suggesting a noteworthy difference from random classification.

To improve the model, checking the training set's class distribution and considering resampling techniques like random oversampling or SMOTE could be performed to further improve the model.

In conclusion, while the model showed strong results for the "low" and "very_high" categories, further optimization is needed for "medium" and "high" to enhance overall performance.

Nevertheless, the model demonstrates reliable performance in classifying insurance premiums into the four categories with MCC of about 0.57 indicating a moderate to strong correlation between prediction and actual category.Therefore, the robust model can be used to classify premiums.Further refinement of tailored features may improve overall performance, especially for medium and high categories.

Conclusion

This project successfully evaluated multiple machine learning models to improve insurance premium calculations for an Ethiopian insurance company. The analysis provided critical insights into predictive accuracy and practical applications, enabling data-driven recommendations.

Among the models evaluated, the neural network emerged as the most robust for predicting premium amounts, achieving an R^2 of 75.1% with minimal overfitting, as confirmed by cross-validation. Its ability to handle complex interactions and deliver high accuracy makes it the most reliable choice for implementation. The linear regression model, while simpler, also performed well, explaining 73.08% of the variance in premiums. However, challenges with heteroskedasticity and residual normality limit its utility, especially for larger premium values.

For claims prediction, the Poisson model provided valuable insights into claims frequency and highlighted the influence of factors such as vehicle age and insured value. However, systematic underestimation for higher claims suggests that alternative approaches, such as Zero-Inflated Poisson models, could address these shortcomings. Similarly, the binomial logistic regression model was effective in identifying predictors of claim payouts, but its limited explanatory power ($R^2 \approx 4.8\%$) indicates a need for further refinement.

The support vector machine (SVM) demonstrated strong potential for classifying premiums into discrete categories, achieving an accuracy of 77.7%. Its strength lies in distinguishing "low" and "very_high" premium categories effectively. However, its performance for "medium" and "high" premiums could be improved through better feature engineering and data balancing techniques. The generalized additive model (GAM) captured non-linear relationships in the data, but it did not outperform simpler models like linear regression or neural networks in terms of accuracy, limiting its practical application.

Based on these findings, we recommend adopting the neural network model for premium prediction due to its high accuracy and flexibility. For claims prediction, incorporating a Zero-Inflated Poisson model would provide a more nuanced approach to handling data with a high frequency of zero claims and outliers. Additionally, the SVM model is well-suited for premium categorization and can aid in customer segmentation, especially if combined with strategies to enhance accuracy for the "medium" and "high" categories.

In conclusion, integrating insights from these models into pricing policies will enable the client to implement fair and accurate premiums, optimize risk management, and improve overall profitability. Factors such as insured value, vehicle brand, and age should be strategically incorporated into pricing and marketing strategies. With these data-driven enhancements, the company is well-positioned to leverage machine learning for sustained competitive advantage in the insurance market.

Usage of Generative AI

In the group project, generative AI was employed to facilitate coding tasks, generate text, and clarify complex concepts. This technology proved beneficial for automating repetitive tasks and assisting in the assembly of report sections, especially in presenting complex ideas in a clear manner.

However, challenges were encountered in the precise formulation of prompts; imprecise prompts occasionally led to AI-generated solutions that did not meet specific project needs. Consequently, all AI-generated outputs required thorough verification to ensure their relevance and accuracy. In some instances, modifications were necessary to align the AI-produced code with project specifications or to optimize performance. The text generated by the AI also needed careful examination to confirm its alignment with project objectives and adherence to academic standards.

Generative AI struggled with tasks requiring deep contextual understanding or specialized knowledge unique to the project. While it significantly enhanced productivity and facilitated the drafting process, active human oversight was crucial to not apply irrelevant or incorrect changes.

Verification of AI suggestions against trusted sources and empirical data was consistently performed, particularly in the context of complex statistical analyses and interpretations.Using AI offered considerable advantages but required a focused and hands-on approach to fully leverage its capabilities in the academic context. Summing up, it has definitely supported the team in terms of explaining difficult concepts while also increasing efficiency.