ML1

2024-11-27

# Introduction

Once 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 enormely 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

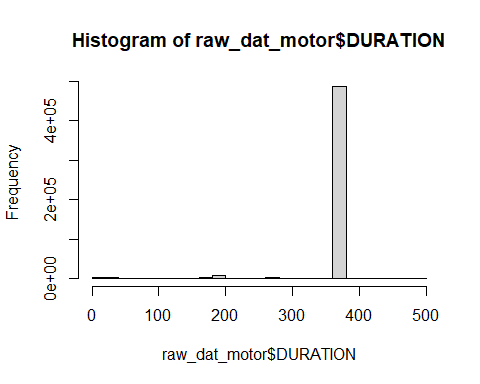
In order to apply such algorithms, the data had to be pre-processed. This process can be found below.

## [1] 508499 16

## spc\_tbl\_ [508,499 × 16] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ SEX : num [1:508499] 0 0 0 0 0 0 0 0 1 1 ...  
## $ INSR\_BEGIN : chr [1:508499] "08-AUG-17" "08-AUG-16" "08-AUG-15" "08-AUG-14" ...  
## $ INSR\_END : chr [1:508499] "07-AUG-18" "07-AUG-17" "07-AUG-16" "07-AUG-15" ...  
## $ EFFECTIVE\_YR : chr [1:508499] "08" "08" "08" "08" ...  
## $ INSR\_TYPE : num [1:508499] 1202 1202 1202 1202 1202 ...  
## $ INSURED\_VALUE : num [1:508499] 519755 519755 519755 519755 1400000 ...  
## $ PREMIUM : num [1:508499] 5098 6557 6557 5103 13305 ...  
## $ OBJECT\_ID : num [1:508499] 5e+09 5e+09 5e+09 5e+09 5e+09 ...  
## $ PROD\_YEAR : num [1:508499] 2007 2007 2007 2007 2010 ...  
## $ SEATS\_NUM : num [1:508499] 4 4 4 4 4 4 4 4 0 0 ...  
## $ CARRYING\_CAPACITY: chr [1:508499] "6" "6" "6" "6" ...  
## $ TYPE\_VEHICLE : chr [1:508499] "Pick-up" "Pick-up" "Pick-up" "Pick-up" ...  
## $ CCM\_TON : num [1:508499] 3153 3153 3153 3153 2494 ...  
## $ MAKE : chr [1:508499] "NISSAN" "NISSAN" "NISSAN" "NISSAN" ...  
## $ USAGE : chr [1:508499] "Own Goods" "Own Goods" "Own Goods" "Own Goods" ...  
## $ CLAIM\_PAID : num [1:508499] NA NA NA NA NA ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. SEX = col\_double(),  
## .. INSR\_BEGIN = col\_character(),  
## .. INSR\_END = col\_character(),  
## .. EFFECTIVE\_YR = col\_character(),  
## .. INSR\_TYPE = col\_double(),  
## .. INSURED\_VALUE = col\_double(),  
## .. PREMIUM = col\_double(),  
## .. OBJECT\_ID = col\_double(),  
## .. PROD\_YEAR = col\_double(),  
## .. SEATS\_NUM = col\_double(),  
## .. CARRYING\_CAPACITY = col\_character(),  
## .. TYPE\_VEHICLE = col\_character(),  
## .. CCM\_TON = col\_double(),  
## .. MAKE = col\_character(),  
## .. USAGE = col\_character(),  
## .. CLAIM\_PAID = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

## Anzahl der entfernten Duplikate: 113

##   
## Legal entity Male Female   
## 247026 217734 43626



## Fehlende Werte in INSURED\_VALUE: 0

## Zusammenfassung der statistischen Kennzahlen von INSURED\_VALUE:

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 62500 544064 800000 67824388

## Anzahl der Einträge mit dem Wert 0 in INSURED\_VALUE: 181149

## Anzahl der Datensätze mit INSURED\_VALUE = 0: 181149

## Verteilung der Versicherungstypen (INSR\_TYPE) bei INSURED\_VALUE = 0:

##   
## Private Commercial Motor trade road risk   
## 92929 284178 141

##   
## Private Commercial Motor trade road risk   
## 35649 145462 38

##   
## Verteilung der Fahrzeugtypen (TYPE\_VEHICLE) bei INSURED\_VALUE = 0:

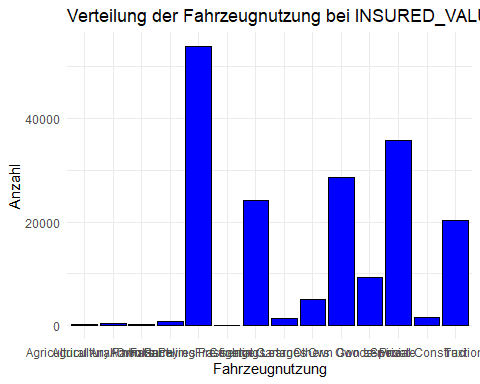
##   
## Automobile Bus Motor-cycle   
## 25016 23749 73368   
## Pick-up Special construction Station Wagones   
## 19747 2397 5183   
## Tanker Tractor Trailers and semitrailers   
## 1377 420 2016   
## Truck   
## 27876

##   
## Verteilung der Fahrzeugnutzung (USAGE) bei INSURED\_VALUE = 0:

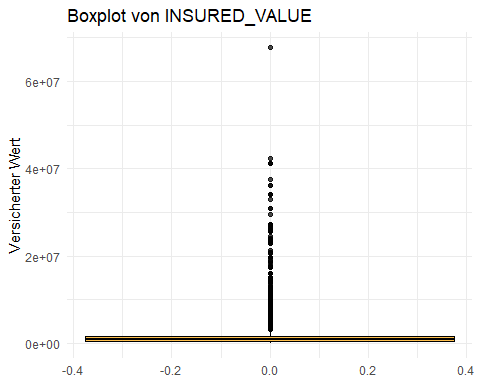
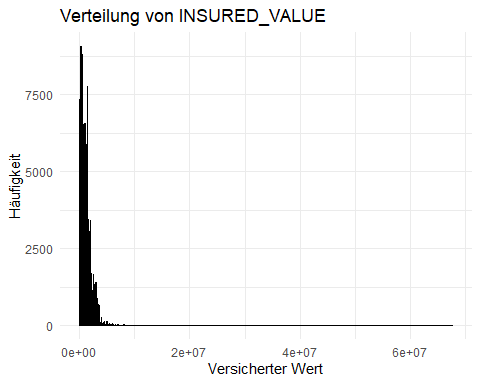
##   
## Agricultural Any Farm Agricultural Own Farm Ambulance   
## 174 341 186   
## Car Hires Fare Paying Passengers Fire fighting   
## 793 53905 6   
## General Cartage Learnes Others   
## 24168 1306 4945   
## Own Goods Own service Private   
## 28565 9290 35736   
## Special Construction Taxi   
## 1493 20241

##   
## Zusammenfassung der Prämien (PREMIUM) bei INSURED\_VALUE = 0:

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0 347.7 647.4 1370.7 1830.5 33645.3 4

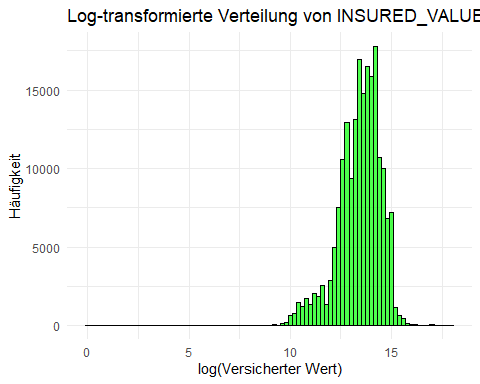


## Anzahl der verbleibenden Datensätze: 196099



## Zusammenfassung der statistischen Kennzahlen ohne 0-Werte:

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 397751 795000 1046650 1480050 67824388



## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 397751 795000 1046650 1480050 67824388

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2665 397751 795000 1046698 1480050 67824388

## PREMIUM\_0\_Percent PREMIUM\_NA\_Percent PREMIUM\_MORE\_Percent  
## 1 0.0031 0.002 99.9969

## Anzahl der entfernten Duplikate: 0

## Die OBJECT\_IDs sind NICHT einmalig.  
## Anzahl der Duplikate: 100652   
## Durchschnittliche Häufigkeit der OBJECT\_ID: 2.055   
## Maximale Häufigkeit der OBJECT\_ID: 8   
## Durchschnittliche Häufigkeit der Kombination (OBJECT\_ID, INSR\_BEGIN, INSR\_END, INSURED\_VALUE, PREMIUM): 1   
## Maximale Häufigkeit der Kombination (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

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 1298 2779 3764 4200 19798

## CCM\_TON\_0\_Percent CCM\_TON\_MORE\_Percent  
## 1 13.76464 86.23536

##   
## Automobile Bus Motor-cycle   
## 704 2559 678   
## Pick-up Special construction Station Wagones   
## 1374 1287 598   
## Tanker Tractor Trailers and semitrailers   
## 425 3455 13000   
## Truck   
## 2902

##   
## Automobile Bus Motor-cycle   
## 29028 20034 10267   
## Pick-up Special construction Station Wagones   
## 45063 2601 21028   
## Tanker Tractor Trailers and semitrailers   
## 2816 1602 816   
## Truck   
## 35787

## Anzahl der entfernten Duplikate: 0

##   
## ABAY   
## 595   
## ACHIVER   
## 1   
## ADDIS GEELY   
## 34   
## ADDIS GELLY   
## 3   
## ADGE   
## 749   
## AEOLUS   
## 55   
## AFRO   
## 951   
## AIRCARGO MOBILE TRUCK   
## 6   
## ALAMI   
## 6   
## ALFAROMEO   
## 3   
## AMBULANCE   
## 24   
## AMI   
## 2   
## APACHE   
## 3   
## ARBE EMERET   
## 3   
## AREB EMERATE   
## 2   
## ARTICULATED DUMP TRUCK   
## 2   
## ASNAKE ENGINERING   
## 2   
## ASNAKE ENGNERING   
## 2   
## ASTRA   
## 669   
## ATOZ   
## 148   
## AU   
## 2   
## AUDI   
## 33   
## AUTO   
## 15   
## AUTOMOBIL   
## 1   
## AUTOMOBILE   
## 5   
## AWASH   
## 47   
## AXION   
## 7   
## BAIC   
## 20   
## BAIC AUTOMOBIL   
## 22   
## BAJAJ   
## 321   
## BAJAJI   
## 1597   
## BARTOLETI   
## 94   
## BAYBEN HIGHBAD   
## 2   
## BAYBEN HIGHBAD TRAILER   
## 1   
## BAYBEN TRUCK HIGHBED   
## 3   
## BEBEN   
## 406   
## BEBEN HIGHBAD   
## 2   
## BEBEN SEMI TRAILER   
## 4   
## BEBIEN TANKER   
## 22   
## BEL TRACTOR   
## 3   
## BELARUS   
## 667   
## BELARUS TRACTOR   
## 166   
## BELL   
## 28   
## BELL TRACTOR   
## 8   
## BEYBEN TRUCK   
## 1   
## BISHEFTU   
## 3   
## BISHOFTU   
## 3736   
## BISHOFTU P/UP D/CAB   
## 6   
## BISHOFTU/FAW   
## 88   
## BISHOFTUKAMA   
## 271   
## BJC   
## 2   
## BMB   
## 1   
## BMP SONIC   
## 3   
## BMW   
## 393   
## BMW AUTO   
## 2   
## BOX   
## 41   
## BOXER   
## 332   
## BRIDGE   
## 3   
## BUS   
## 21   
## BYD   
## 42   
## CACCIAMALLI   
## 11   
## CADILLAC   
## 17   
## CALABRASE   
## 612   
## CALABRESE   
## 1422   
## CANEHAULAGE   
## 5   
## CARENZI   
## 22   
## CARGO   
## 2   
## CASE   
## 1   
## CAT   
## 29   
## CAT DOZER   
## 46   
## CATERPILLAR   
## 3   
## CATERPILLAR TRACTOR   
## 4   
## CATO   
## 3   
## CHANA   
## 64   
## CHANGHE   
## 3   
## CHARRY   
## 22   
## CHENGLONG MOTOR OF CHINA   
## 3   
## CHERRY   
## 34   
## CHEVROLET   
## 287   
## CHINA   
## 7   
## CHINA - BBN   
## 5   
## CHINA SELONG   
## 1   
## CHINA SPECIAL POWER TRUCK   
## 24   
## CHINA SPECIAL SEMI TRAILER   
## 82   
## CHINA ZENGIZO   
## 1   
## CITROEN   
## 6   
## CLASS   
## 126   
## CLASS COMBINE   
## 5   
## CO   
## 6   
## CO BUS   
## 6   
## COASTER   
## 2   
## COASTER BUS   
## 2   
## COMBI   
## 5   
## COMPACT YARIS   
## 1   
## CORDES   
## 44   
## CORE DRILLING RING   
## 4   
## CRANE   
## 18   
## CRANE ZUMLIN   
## 1   
## CRANE ZUMLIN 70 TON   
## 2   
## DACIA   
## 77   
## DAEWOO   
## 1179   
## DAF   
## 294   
## DAHATSUN   
## 34   
## DAIHATSU   
## 6   
## DAIHATSU TERIOS   
## 69   
## DAMAS   
## 39   
## DANDO GEATECH 7.5 HYDOLIC TOP ROTATYING   
## 3   
## DATSUN   
## 20   
## DAW BUS   
## 6   
## DAWOO   
## 266   
## DAWWO   
## 1   
## DAYUN   
## 14   
## DEAWOO   
## 1   
## DEAWOO USE   
## 1   
## DEUTZ FAHR   
## 102   
## DFAC   
## 1   
## DFM   
## 3   
## DGOIX   
## 2   
## DHATSU   
## 3   
## DIAHATSU   
## 169   
## DIATSU   
## 1   
## DISCOVERY   
## 10   
## DOCC   
## 4   
## DONFING   
## 288   
## DONG FENG   
## 38   
## DONG FENGSHEN   
## 1   
## DONGFANG   
## 5   
## DONGFENG   
## 4   
## DORSEY   
## 3   
## DOZER   
## 58   
## DSIT   
## 17   
## DUBI   
## 4   
## DUNGFINF   
## 53   
## DUNGFING   
## 24   
## EICHER   
## 50   
## EMGRAND   
## 7   
## EMGTAND   
## 3   
## ENGLAND   
## 1   
## ENGLAND TRACTOR   
## 13   
## ETHIOPIA   
## 9   
## EURO TRUCKER   
## 3   
## EXCAVATOR   
## 22   
## FARID   
## 382   
## FAW   
## 625   
## FAWBELLA   
## 4   
## FENGXING   
## 4   
## FIAT   
## 1571   
## FOED   
## 3   
## FORCE   
## 256   
## FORD   
## 2217   
## FORD CARGO   
## 1   
## FORKLIFT   
## 9   
## FORLAND   
## 98   
## FORSCHE   
## 3   
## FOTON   
## 163   
## FOTTON   
## 23   
## FPRD   
## 4   
## FRANKON   
## 2   
## FRANKUN   
## 1   
## FRANKUN ET   
## 3   
## FRANKUN IVECO   
## 3   
## G9   
## 3   
## GEELY   
## 946   
## GEEP   
## 2   
## GELION   
## 470   
## GELYION   
## 3   
## GENLION   
## 1   
## GENLYON   
## 30   
## GENLYONIVECO   
## 3   
## GETZ   
## 1   
## GLEEY   
## 50   
## GMC   
## 75   
## GMS   
## 3   
## GOLZ-PLUS   
## 2   
## GORICA   
## 79   
## GRADER   
## 33   
## GREAT WALL   
## 36   
## H.H   
## 2   
## HAFEI RULY   
## 4   
## HAFREI   
## 3   
## HANVE   
## 4   
## HERO   
## 154   
## HH   
## 10   
## HIGBAN HIGHBAD   
## 2   
## HIGER   
## 53   
## HIGER BUS   
## 13   
## HIGH BED   
## 3   
## HIGH BED TRAILER   
## 19   
## HIGHBED   
## 7   
## HIGHBENCARGOTRAUCK   
## 1   
## HIGHBIN HIGHBAD   
## 2   
## HIGHER   
## 37   
## HILUX   
## 2   
## HINO   
## 315   
## HOLAND CAR   
## 3   
## HONDA   
## 331   
## HONGYAN   
## 4   
## HOVER   
## 15   
## HOWO   
## 82   
## HOYGYAN   
## 6   
## HUANGHA   
## 3   
## HUMMER   
## 3   
## HUYBED   
## 2   
## HYDROLIC   
## 1   
## HYUNDAI   
## 2405   
## HYUNDI GETZ   
## 2   
## ILSBO   
## 186   
## INDOFARMO   
## 1   
## INFINITY   
## 1   
## INTERNATIONAL   
## 22   
## INTERNATIONAL USE   
## 1704   
## ISUSU   
## 6   
## ISUZU   
## 13768   
## ISUZU FVR   
## 3   
## ITALY   
## 1   
## IVECO   
## 7296   
## IVECO/CHINA   
## 16   
## JAC   
## 5   
## JAK   
## 2   
## JCB WORK MAX   
## 3   
## JEEP   
## 16   
## JERMEN   
## 1   
## JIEFANG   
## 134   
## JILI SABA   
## 3   
## JIN BEI   
## 150   
## JMC   
## 18   
## JOHN DEER   
## 504   
## JOHNDEER   
## 16   
## KAINUO   
## 8   
## KAMA   
## 1   
## KAMA MINI TRUCK   
## 8   
## KAMA NINI TRUCK   
## 2   
## KAMAMI   
## 1   
## KAMAZ   
## 8   
## KAMZ   
## 6   
## KAT   
## 21   
## KAT TRACTOR   
## 2   
## KAT TRAILER   
## 2   
## KG   
## 2   
## KIA   
## 277   
## KING LONG   
## 14   
## KM.UAG   
## 2   
## KOMATSU   
## 1   
## KOREA   
## 18   
## KORIA   
## 3   
## KORYA   
## 1   
## KYRON   
## 2   
## LADA   
## 10   
## LAND CRUISER   
## 3   
## LAND ROVER   
## 166   
## LANDINI   
## 209   
## LANDINI DT125   
## 1   
## LANDROVER   
## 282   
## LANJIAN   
## 21   
## LEXUS   
## 9   
## LIBERR MOBILE CRANE   
## 1   
## LIBERR MOBILECRANE   
## 1   
## LIEBERR MOBILE CRANE   
## 1   
## LIFAN   
## 2382   
## LIFAN 520   
## 3   
## LIFAN AUTOMOBILE   
## 2   
## LISBO   
## 16   
## LITON   
## 12   
## LOADER   
## 27   
## LOADER POERR PLUS   
## 1   
## LOADER POWER PLUS   
## 1   
## LOBADE TRUCK   
## 1   
## LOBED   
## 20   
## LOGAN   
## 3   
## LONG BASE TRAILER   
## 1   
## LONG JIANG   
## 3   
## LONGJIANG   
## 22   
## LOW BED   
## 14   
## LOWBED   
## 218   
## MACK   
## 291   
## MAHANDRA   
## 163   
## MAHINDRA   
## 115   
## MAMMUT   
## 13   
## MAN   
## 130   
## MARU   
## 504   
## MASIL FERGUSAN   
## 77   
## MASSY FUREGUSON   
## 386   
## MATIZ   
## 5   
## MAZ   
## 20   
## MAZDA   
## 1456   
## ME   
## 3   
## MERCEDES   
## 855   
## MERCEDICE   
## 2   
## MERCEEDES   
## 833   
## MERCEEDICE   
## 252   
## MERCHEDES   
## 3   
## MESFIN   
## 5935   
## MESIFIN   
## 3   
## MF5340   
## 3   
## MINI BUS   
## 2   
## MISTIBUSH   
## 1   
## MITSUBISHI   
## 6167   
## MITSUBISHI\*   
## 3   
## MIXER   
## 2   
## MOBILE GUARAGE   
## 1   
## MOTOR CYCLE   
## 12   
## MOTOR CYCLE (TWOCYCLE)   
## 10   
## MOTORCYCLE   
## 10   
## MTE   
## 60   
## MUSSO   
## 6   
## NAM   
## 1   
## NAMI   
## 313   
## NATFA   
## 17   
## NEW HOLAND   
## 2   
## NEW HOLLAND   
## 273   
## NIO   
## 4   
## NISAN   
## 7   
## NISSAN   
## 11604   
## NISSAN SUNNY   
## 2   
## NISSAN UD   
## 174   
## NISSAN X-TRIAL   
## 1   
## NISSAN\*   
## 42   
## NIVA   
## 11   
## NKG ENG   
## 2   
## OD   
## 3   
## OHNDEERE   
## 1   
## OPEL   
## 41   
## ORAL   
## 134   
## OTOYOL   
## 36   
## P/UP   
## 10   
## PAGOT   
## 2   
## PEJOT   
## 1   
## PEUGEOT   
## 371   
## PEUGEOT AUTOMOBILE   
## 4   
## PEUGEOUT   
## 2   
## PLATENA   
## 110   
## PORCHE   
## 2   
## PORSCHE   
## 2   
## POWER PLUS   
## 5   
## POWER PLUS DAM   
## 2   
## POWER PLUS DAMP   
## 2   
## POWER PLUS DOSER   
## 1   
## POWER PLUS TRUCK   
## 5   
## POWRPLUS TRUCK   
## 2   
## PREGIO   
## 6   
## R425DOHC   
## 1   
## RANDON   
## 62   
## RANGE ROVER   
## 3   
## RANGEROVER   
## 18   
## RAV4   
## 1   
## RAVA   
## 3   
## RED FOX   
## 89   
## RENALT   
## 19   
## RENAULT   
## 1086   
## RENAULT\*   
## 3   
## RENGE ROVER   
## 7   
## RENUALT   
## 2   
## REXTON   
## 8   
## RIG   
## 18   
## RIO LS   
## 7   
## RIO JAMES   
## 3   
## RIO JAMES TRUCK PALLET   
## 1   
## ROLD   
## 2   
## ROLF   
## 2   
## ROLFO   
## 521   
## ROLLER   
## 39   
## ROZA   
## 31   
## S/W   
## 21   
## SAMI   
## 70   
## SANIA   
## 2   
## SANY   
## 5   
## SCANIA   
## 643   
## SCHMITZ   
## 109   
## SCRAPER   
## 1   
## SEDEN   
## 1   
## SEECOME   
## 2   
## SHACMAN   
## 75   
## SHNAY   
## 56   
## SINALIKE   
## 27   
## SINO   
## 1035   
## SINO HOWO   
## 6970   
## SINO TRUCK   
## 7   
## SINOTRUK   
## 63   
## SINOTRUK HOWO   
## 1   
## SKODA   
## 40   
## SKY BUS   
## 13   
## SMART   
## 1   
## SOCOOL   
## 2   
## SONALIKA   
## 67   
## SPAIN   
## 6   
## SPORTAGE   
## 3   
## STAYER   
## 5   
## STEYER   
## 84   
## SUGERCANE TRAILER   
## 72   
## SUNLONG   
## 225   
## SUNLONGBUS   
## 20   
## SUV   
## 13   
## SUZIKE   
## 3   
## SUZUKI   
## 2860   
## SUZUKI GRAND VITARA   
## 3   
## T0Y0TA   
## 70   
## TAIWAN   
## 2   
## TALER   
## 3   
## TALIAN   
## 2   
## TATA   
## 971   
## TEKEZE   
## 3   
## TERIOS   
## 7   
## TICO   
## 6   
## TOMSON   
## 4   
## TOYATA   
## 1   
## TOYOTA   
## 73371   
## TOYOTA AUTOMOBILE   
## 2   
## TOYOTA 4 RUNNER   
## 1   
## TOYOTA AUTOMOBILE   
## 2   
## TOYOTA COROLLA   
## 3   
## TOYOTA HIACE   
## 2   
## TOYOTA HILUX   
## 2   
## TOYOTA L/C PRADO   
## 2   
## TOYOTA L/CRUISER   
## 3   
## TOYOTA MERCHEDIS   
## 1   
## TOYOTA MINIBUS   
## 1   
## TOYOTA P/UP   
## 2   
## TOYOTA PICK-UP   
## 2   
## TOYOTA PLATZ   
## 1   
## TOYOTA RAV4   
## 9   
## TOYOTA RAVA4   
## 1   
## TOYOTA VANZE   
## 1   
## TOYOTA VITZ   
## 11   
## TOYOTA YARIS   
## 6   
## TOYOTA\*   
## 31   
## TOYOTAA   
## 4   
## TOYTA   
## 3   
## TRACTOR   
## 289   
## TRACTOR 4WD   
## 23   
## TRACTOR BELARUS   
## 9   
## TRACTOR TRAILER   
## 34   
## TRACTOR4WD   
## 3   
## TRAILED TANKER WITH FIRE EXTINGUISHER   
## 4   
## TRAILED TANKKER WITH FIRE FIRE EXTINGUIS   
## 2   
## TRAILER   
## 1124   
## TRAKER   
## 39   
## TRAKKER   
## 344   
## TRUCK   
## 13   
## TURBO   
## 3   
## TURBO BUS   
## 105   
## TVS   
## 869   
## TVS125   
## 3   
## UAE   
## 4   
## UAI   
## 3   
## URSUS   
## 194   
## URSUS TRACTOR   
## 15   
## URSUS TRACTOR URSUS TRACTOR   
## 2   
## USA   
## 3   
## VALTRA TRACTOR   
## 2   
## VAN TRUCK   
## 6   
## VERCYA   
## 53   
## VERSATILE   
## 101   
## VERSATILE TRACTOR   
## 5   
## VERYCA   
## 1   
## VIBERTI   
## 178   
## VITZ   
## 1185   
## VITZ AUTOMOBILE   
## 1   
## VOLKS WAGON   
## 53   
## VOLKSWAGEN   
## 16   
## VOLKSWAGON   
## 104   
## VOLVO   
## 925   
## WAFA   
## 2   
## WAZ   
## 7   
## WETER TRUCK STAYER   
## 2   
## WHEEL LOADER   
## 126   
## WINEGEL   
## 12   
## WUCING   
## 3   
## X60   
## 2   
## XERION TARCTOR   
## 1   
## XERION TRACTOR   
## 1   
## YAMAHA   
## 3534   
## YAMHA   
## 4   
## YARIS   
## 5   
## YOUNGMAN   
## 3   
## YOUTOGNMIDBUS   
## 1   
## YOUTONG   
## 5   
## YOUTONG BUS   
## 2   
## YTO   
## 25   
## YTO TRACTOR   
## 5   
## YUTONG   
## 6   
## YVS   
## 12   
## ZAMAJ   
## 28   
## ZEPPLIN   
## 36   
## ZILE SHOPAN   
## 4   
## ZNA   
## 13   
## ZOBLE   
## 16   
## ZONGSHEN   
## 68   
## ZONGUSHEN   
## 71   
## ZOOM LION CRANE   
## 20   
## ZORZI   
## 43   
## ZOTYE   
## 30   
## ZOTYE, NOMAD II   
## 3   
## ZOYTE   
## 18   
## ZOYTE, NOMAD II   
## 3   
## ZTLTRUCK   
## 1   
## ZUMLIN CRANE   
## 1   
## ZUNGSHUN   
## 21   
## ZX-TOP   
## 1   
## ZX TOP   
## 1   
## ZZ   
## 28

## 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

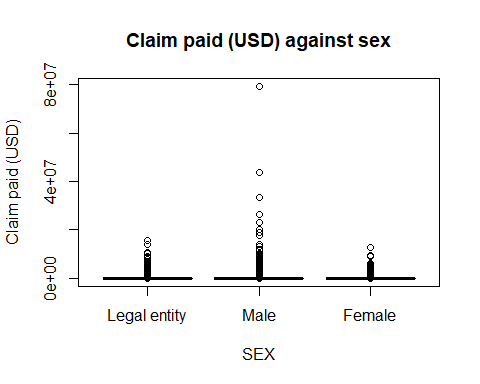
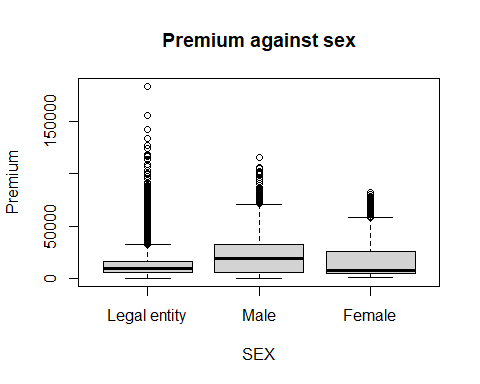
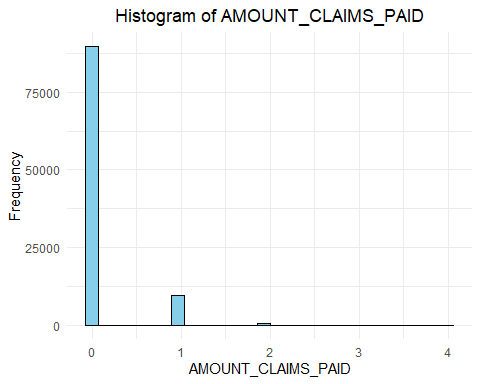
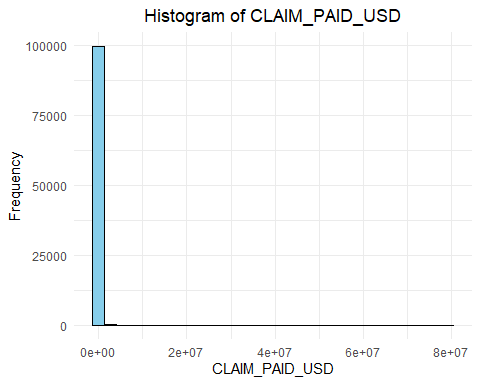
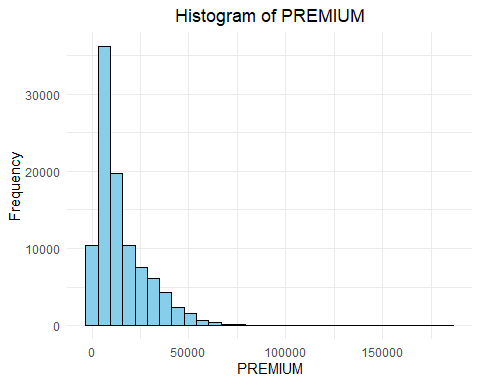
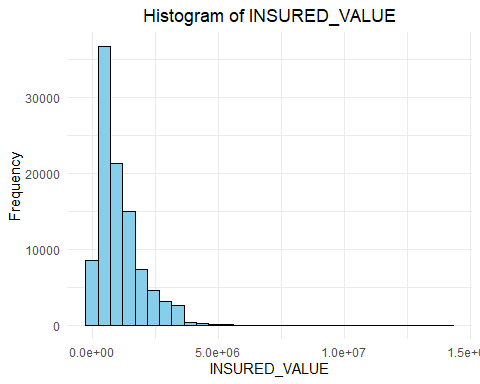
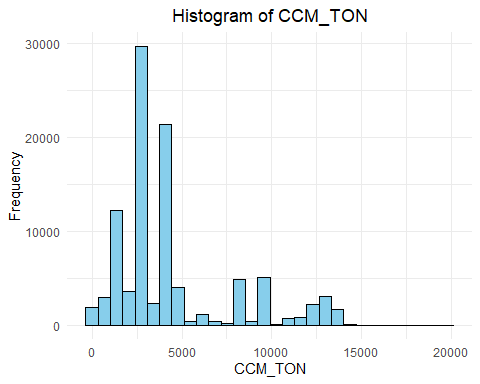
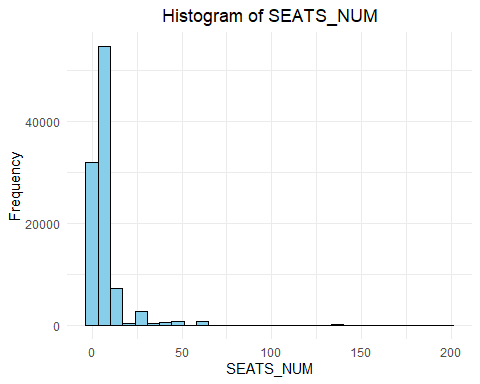
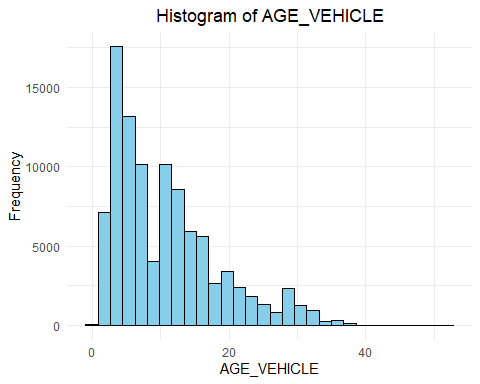
## CLAIM\_PAID\_0\_Percent CLAIM\_PAID\_MORE\_THAN\_0\_Percent  
## 1 86.15231 13.84769

## 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 is analysed to determine whether any transformations are necessary.



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 can be inserted in the later regression models instead of the original variables.

# Models

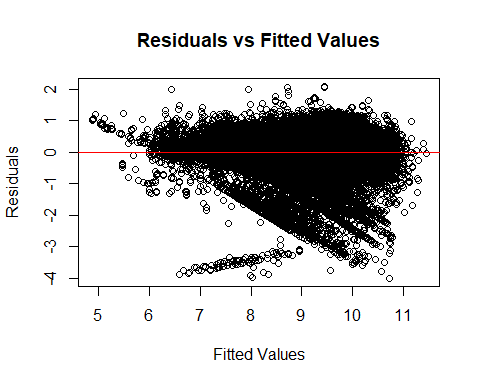
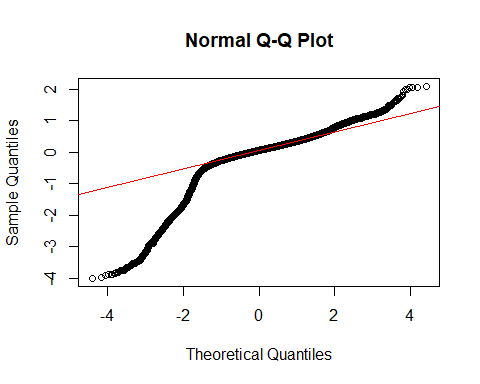
## Linear Model

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.

##   
## Call:  
## lm(formula = PREMIUM\_log ~ SEX + INSR\_TYPE + USAGE + TYPE\_VEHICLE +   
## MAKE + AGE\_VEHICLE + SEATS\_NUM + CCM\_TON\_log + INSURED\_VALUE\_log +   
## AMOUNT\_CLAIMS\_PAID, data = clean\_dat\_motor)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.0103 -0.1433 0.0632 0.2506 2.0985   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.3149332 0.0561212 -23.430 < 2e-16  
## SEXMale -0.0853692 0.0053558 -15.940 < 2e-16  
## SEXFemale -0.0796192 0.0076065 -10.467 < 2e-16  
## INSR\_TYPEPrivate 0.0886106 0.0200423 4.421 9.83e-06  
## USAGECar Hires -0.2156283 0.0415019 -5.196 2.04e-07  
## USAGEFare Paying Passengers 0.5417051 0.0291570 18.579 < 2e-16  
## USAGEGeneral Cartage 0.1532115 0.0496041 3.089 0.00201  
## USAGEOwn Goods -0.5047742 0.0491266 -10.275 < 2e-16  
## USAGEOwn service -0.0673302 0.0275711 -2.442 0.01461  
## USAGEPrivate -0.4625354 0.0281823 -16.412 < 2e-16  
## TYPE\_VEHICLEBus 0.1307061 0.0222932 5.863 4.56e-09  
## TYPE\_VEHICLEMotor-cycle -0.0076512 0.0235611 -0.325 0.74538  
## TYPE\_VEHICLEPick-up 0.0832227 0.0455400 1.827 0.06763  
## TYPE\_VEHICLESpecial construction 0.0108042 0.0417602 0.259 0.79585  
## TYPE\_VEHICLEStation Wagones 0.1688239 0.0075031 22.501 < 2e-16  
## TYPE\_VEHICLETanker 0.3463080 0.0479742 7.219 5.29e-13  
## TYPE\_VEHICLETrailers and semitrailers 0.0638634 0.0642133 0.995 0.31996  
## TYPE\_VEHICLETruck 0.3246406 0.0464155 6.994 2.68e-12  
## MAKEDAEWOO 0.1845153 0.0193497 9.536 < 2e-16  
## MAKEFIAT 0.1413248 0.0201080 7.028 2.10e-12  
## MAKEFORD 0.0950234 0.0171017 5.556 2.76e-08  
## MAKEGEELY 0.0994460 0.0234607 4.239 2.25e-05  
## MAKEGENLYON -0.1549654 0.0299558 -5.173 2.31e-07  
## MAKEHYUNDAI 0.0521274 0.0167280 3.116 0.00183  
## MAKEISUZU 0.3090353 0.0130996 23.591 < 2e-16  
## MAKEIVECO -0.0380251 0.0147388 -2.580 0.00988  
## MAKELIFAN 0.1311950 0.0175353 7.482 7.39e-14  
## MAKEMAZDA 0.0517023 0.0191657 2.698 0.00698  
## MAKEMERCEDES 0.1962833 0.0177475 11.060 < 2e-16  
## MAKEMITSUBISHI 0.1273932 0.0135116 9.428 < 2e-16  
## MAKENISSAN 0.1472757 0.0125553 11.730 < 2e-16  
## MAKERENAULT -0.0981452 0.0219967 -4.462 8.14e-06  
## MAKESINO -0.0114553 0.0230703 -0.497 0.61952  
## MAKESINO HOWO 0.0243778 0.0147721 1.650 0.09889  
## MAKESUZUKI 0.1146855 0.0209568 5.472 4.45e-08  
## MAKETATA 0.1495387 0.0221466 6.752 1.46e-11  
## MAKETOYOTA 0.1612054 0.0115209 13.992 < 2e-16  
## AGE\_VEHICLE 0.0029448 0.0003430 8.585 < 2e-16  
## SEATS\_NUM -0.0017519 0.0002502 -7.003 2.52e-12  
## CCM\_TON\_log 0.0109448 0.0043773 2.500 0.01241  
## INSURED\_VALUE\_log 0.7682069 0.0035157 218.510 < 2e-16  
## AMOUNT\_CLAIMS\_PAID 0.1362875 0.0047779 28.525 < 2e-16  
##   
## (Intercept) \*\*\*  
## SEXMale \*\*\*  
## SEXFemale \*\*\*  
## INSR\_TYPEPrivate \*\*\*  
## USAGECar Hires \*\*\*  
## USAGEFare Paying Passengers \*\*\*  
## USAGEGeneral Cartage \*\*   
## USAGEOwn Goods \*\*\*  
## USAGEOwn service \*   
## USAGEPrivate \*\*\*  
## TYPE\_VEHICLEBus \*\*\*  
## TYPE\_VEHICLEMotor-cycle   
## TYPE\_VEHICLEPick-up .   
## TYPE\_VEHICLESpecial construction   
## TYPE\_VEHICLEStation Wagones \*\*\*  
## TYPE\_VEHICLETanker \*\*\*  
## TYPE\_VEHICLETrailers and semitrailers   
## TYPE\_VEHICLETruck \*\*\*  
## MAKEDAEWOO \*\*\*  
## MAKEFIAT \*\*\*  
## MAKEFORD \*\*\*  
## MAKEGEELY \*\*\*  
## MAKEGENLYON \*\*\*  
## MAKEHYUNDAI \*\*   
## MAKEISUZU \*\*\*  
## MAKEIVECO \*\*   
## MAKELIFAN \*\*\*  
## MAKEMAZDA \*\*   
## MAKEMERCEDES \*\*\*  
## MAKEMITSUBISHI \*\*\*  
## MAKENISSAN \*\*\*  
## MAKERENAULT \*\*\*  
## MAKESINO   
## MAKESINO HOWO .   
## MAKESUZUKI \*\*\*  
## MAKETATA \*\*\*  
## MAKETOYOTA \*\*\*  
## AGE\_VEHICLE \*\*\*  
## SEATS\_NUM \*\*\*  
## CCM\_TON\_log \*   
## INSURED\_VALUE\_log \*\*\*  
## AMOUNT\_CLAIMS\_PAID \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5036 on 99958 degrees of freedom  
## Multiple R-squared: 0.7308, Adjusted R-squared: 0.7307   
## F-statistic: 6619 on 41 and 99958 DF, p-value: < 2.2e-16

## Single term deletions  
##   
## Model:  
## PREMIUM\_log ~ SEX + INSR\_TYPE + USAGE + TYPE\_VEHICLE + MAKE +   
## AGE\_VEHICLE + SEATS\_NUM + CCM\_TON\_log + INSURED\_VALUE\_log +   
## AMOUNT\_CLAIMS\_PAID  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 25348 -137161   
## SEX 2 68.8 25417 -136894 135.5764 < 2.2e-16 \*\*\*  
## INSR\_TYPE 1 5.0 25353 -137144 19.5468 9.827e-06 \*\*\*  
## USAGE 6 2157.4 27506 -129005 1417.8731 < 2.2e-16 \*\*\*  
## TYPE\_VEHICLE 8 299.6 25648 -136002 147.6921 < 2.2e-16 \*\*\*  
## MAKE 19 593.8 25942 -134884 123.2348 < 2.2e-16 \*\*\*  
## AGE\_VEHICLE 1 18.7 25367 -137089 73.6994 < 2.2e-16 \*\*\*  
## SEATS\_NUM 1 12.4 25361 -137114 49.0394 2.524e-12 \*\*\*  
## CCM\_TON\_log 1 1.6 25350 -137157 6.2517 0.01241 \*   
## INSURED\_VALUE\_log 1 12108.2 37457 -98117 47746.8041 < 2.2e-16 \*\*\*  
## AMOUNT\_CLAIMS\_PAID 1 206.3 25555 -136352 813.6607 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## (Intercept) SEXMale   
## -1.314933235 -0.085369195   
## SEXFemale INSR\_TYPEPrivate   
## -0.079619233 0.088610579   
## USAGECar Hires USAGEFare Paying Passengers   
## -0.215628313 0.541705075   
## USAGEGeneral Cartage USAGEOwn Goods   
## 0.153211453 -0.504774189   
## USAGEOwn service USAGEPrivate   
## -0.067330170 -0.462535383   
## TYPE\_VEHICLEBus TYPE\_VEHICLEMotor-cycle   
## 0.130706079 -0.007651153   
## TYPE\_VEHICLEPick-up TYPE\_VEHICLESpecial construction   
## 0.083222728 0.010804232   
## TYPE\_VEHICLEStation Wagones TYPE\_VEHICLETanker   
## 0.168823870 0.346308019   
## TYPE\_VEHICLETrailers and semitrailers TYPE\_VEHICLETruck   
## 0.063863449 0.324640609   
## MAKEDAEWOO MAKEFIAT   
## 0.184515271 0.141324825   
## MAKEFORD MAKEGEELY   
## 0.095023368 0.099446048   
## MAKEGENLYON MAKEHYUNDAI   
## -0.154965367 0.052127411   
## MAKEISUZU MAKEIVECO   
## 0.309035283 -0.038025106   
## MAKELIFAN MAKEMAZDA   
## 0.131194965 0.051702264   
## MAKEMERCEDES MAKEMITSUBISHI   
## 0.196283343 0.127393220   
## MAKENISSAN MAKERENAULT   
## 0.147275719 -0.098145229   
## MAKESINO MAKESINO HOWO   
## -0.011455266 0.024377755   
## MAKESUZUKI MAKETATA   
## 0.114685452 0.149538684   
## MAKETOYOTA AGE\_VEHICLE   
## 0.161205353 0.002944844   
## SEATS\_NUM CCM\_TON\_log   
## -0.001751879 0.010944818   
## INSURED\_VALUE\_log AMOUNT\_CLAIMS\_PAID   
## 0.768206885 0.136287464



## 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-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%.

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.

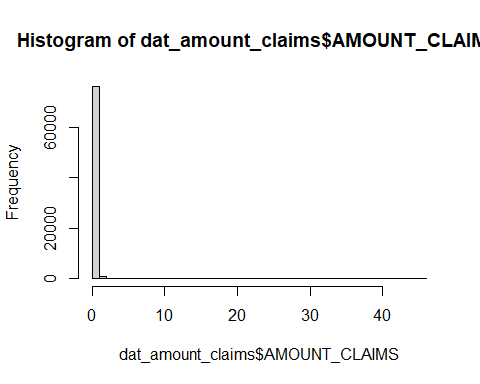
To improve the model, various measures could be considered. One approach would be to transform the target variable, for example, using a Box-Cox transformation, to reduce heteroskedasticity and achieve a more stable residual variance. Additionally, incorporating non-linear relationships by including 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

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.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.1791 0.0000 46.0000



## [1] 0.1790586

## [1] 0.3650433

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.

##   
## Call:  
## glm(formula = AMOUNT\_CLAIMS ~ SEX + INSR\_TYPE + TYPE\_VEHICLE +   
## MAKE + AGE\_VEHICLE + SEATS\_NUM + CCM\_TON + INSURED\_VALUE +   
## PREMIUM, family = poisson(link = "log"), data = dat\_amount\_claims)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.200e+00 9.032e-02 -13.281 < 2e-16  
## SEXMale -4.059e-01 2.468e-02 -16.449 < 2e-16  
## SEXFemale -4.728e-01 3.972e-02 -11.903 < 2e-16  
## INSR\_TYPEPrivate 2.501e-01 6.834e-02 3.659 0.000253  
## TYPE\_VEHICLEBus -1.671e-01 7.769e-02 -2.151 0.031503  
## TYPE\_VEHICLEMotor-cycle -3.240e+00 4.601e-01 -7.042 1.89e-12  
## TYPE\_VEHICLEPick-up 1.139e-01 7.331e-02 1.554 0.120126  
## TYPE\_VEHICLESpecial construction 8.007e-02 1.939e-01 0.413 0.679604  
## TYPE\_VEHICLEStation Wagones -3.233e-01 3.645e-02 -8.869 < 2e-16  
## TYPE\_VEHICLETanker -4.640e-01 1.250e-01 -3.712 0.000206  
## TYPE\_VEHICLETrailers and semitrailers -5.598e-01 3.283e-01 -1.705 0.088183  
## TYPE\_VEHICLETruck 4.702e-02 9.021e-02 0.521 0.602217  
## MAKEDAEWOO -7.534e-01 1.033e-01 -7.294 3.00e-13  
## MAKEFIAT -7.671e-01 1.269e-01 -6.045 1.49e-09  
## MAKEFORD -2.683e-01 8.605e-02 -3.118 0.001823  
## MAKEGEELY 3.922e-02 1.052e-01 0.373 0.709333  
## MAKEGENLYON -7.784e-01 1.571e-01 -4.955 7.25e-07  
## MAKEHYUNDAI -3.562e-01 8.916e-02 -3.995 6.46e-05  
## MAKEISUZU -5.545e-01 6.277e-02 -8.833 < 2e-16  
## MAKEIVECO -7.208e-01 7.584e-02 -9.504 < 2e-16  
## MAKELIFAN -2.035e-01 8.220e-02 -2.475 0.013308  
## MAKEMAZDA -4.090e-02 8.634e-02 -0.474 0.635671  
## MAKEMERCEDES -9.091e-01 1.053e-01 -8.631 < 2e-16  
## MAKEMITSUBISHI -4.308e-01 6.630e-02 -6.497 8.20e-11  
## MAKENISSAN -2.715e-01 5.849e-02 -4.643 3.44e-06  
## MAKERENAULT -3.912e-01 1.498e-01 -2.611 0.009027  
## MAKESINO -5.413e-01 1.328e-01 -4.077 4.56e-05  
## MAKESINO HOWO -9.059e-01 7.570e-02 -11.967 < 2e-16  
## MAKESUZUKI -6.848e-01 1.443e-01 -4.745 2.09e-06  
## MAKETATA -1.079e+00 1.247e-01 -8.650 < 2e-16  
## MAKETOYOTA -1.417e-01 5.122e-02 -2.766 0.005673  
## AGE\_VEHICLE -4.480e-02 1.804e-03 -24.843 < 2e-16  
## SEATS\_NUM 8.748e-03 1.144e-03 7.644 2.11e-14  
## CCM\_TON 3.125e-05 6.073e-06 5.145 2.67e-07  
## INSURED\_VALUE -1.384e-07 1.747e-08 -7.925 2.28e-15  
## PREMIUM 1.895e-05 9.430e-07 20.093 < 2e-16  
##   
## (Intercept) \*\*\*  
## SEXMale \*\*\*  
## SEXFemale \*\*\*  
## INSR\_TYPEPrivate \*\*\*  
## TYPE\_VEHICLEBus \*   
## TYPE\_VEHICLEMotor-cycle \*\*\*  
## TYPE\_VEHICLEPick-up   
## TYPE\_VEHICLESpecial construction   
## TYPE\_VEHICLEStation Wagones \*\*\*  
## TYPE\_VEHICLETanker \*\*\*  
## TYPE\_VEHICLETrailers and semitrailers .   
## TYPE\_VEHICLETruck   
## MAKEDAEWOO \*\*\*  
## MAKEFIAT \*\*\*  
## MAKEFORD \*\*   
## MAKEGEELY   
## MAKEGENLYON \*\*\*  
## MAKEHYUNDAI \*\*\*  
## MAKEISUZU \*\*\*  
## MAKEIVECO \*\*\*  
## MAKELIFAN \*   
## MAKEMAZDA   
## MAKEMERCEDES \*\*\*  
## MAKEMITSUBISHI \*\*\*  
## MAKENISSAN \*\*\*  
## MAKERENAULT \*\*   
## MAKESINO \*\*\*  
## MAKESINO HOWO \*\*\*  
## MAKESUZUKI \*\*\*  
## MAKETATA \*\*\*  
## MAKETOYOTA \*\*   
## AGE\_VEHICLE \*\*\*  
## SEATS\_NUM \*\*\*  
## CCM\_TON \*\*\*  
## INSURED\_VALUE \*\*\*  
## PREMIUM \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 56192 on 77186 degrees of freedom  
## Residual deviance: 53113 on 77151 degrees of freedom  
## AIC: 77497  
##   
## Number of Fisher Scoring iterations: 7

## (Intercept) SEXMale   
## -1.199531e+00 -4.059098e-01   
## SEXFemale INSR\_TYPEPrivate   
## -4.727981e-01 2.500586e-01   
## TYPE\_VEHICLEBus TYPE\_VEHICLEMotor-cycle   
## -1.670950e-01 -3.240007e+00   
## TYPE\_VEHICLEPick-up TYPE\_VEHICLESpecial construction   
## 1.139349e-01 8.006750e-02   
## TYPE\_VEHICLEStation Wagones TYPE\_VEHICLETanker   
## -3.232992e-01 -4.640321e-01   
## TYPE\_VEHICLETrailers and semitrailers TYPE\_VEHICLETruck   
## -5.597884e-01 4.701695e-02   
## MAKEDAEWOO MAKEFIAT   
## -7.534160e-01 -7.671059e-01   
## MAKEFORD MAKEGEELY   
## -2.682717e-01 3.921656e-02   
## MAKEGENLYON MAKEHYUNDAI   
## -7.784331e-01 -3.562094e-01   
## MAKEISUZU MAKEIVECO   
## -5.544635e-01 -7.208167e-01   
## MAKELIFAN MAKEMAZDA   
## -2.034788e-01 -4.090397e-02   
## MAKEMERCEDES MAKEMITSUBISHI   
## -9.090543e-01 -4.307733e-01   
## MAKENISSAN MAKERENAULT   
## -2.715283e-01 -3.911515e-01   
## MAKESINO MAKESINO HOWO   
## -5.412876e-01 -9.059321e-01   
## MAKESUZUKI MAKETATA   
## -6.847925e-01 -1.078515e+00   
## MAKETOYOTA AGE\_VEHICLE   
## -1.416733e-01 -4.480383e-02   
## SEATS\_NUM CCM\_TON   
## 8.747609e-03 3.124821e-05   
## INSURED\_VALUE PREMIUM   
## -1.384306e-07 1.894822e-05

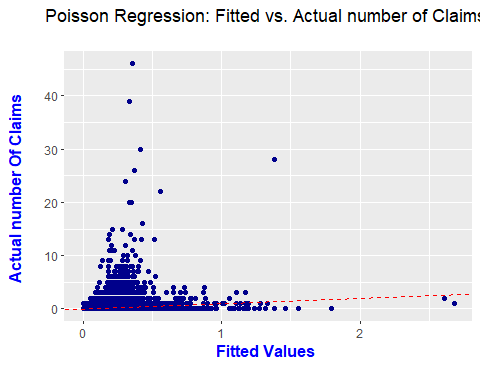
## (Intercept) SEXMale   
## 0.30133547 0.66637028   
## SEXFemale INSR\_TYPEPrivate   
## 0.62325590 1.28410071   
## TYPE\_VEHICLEBus TYPE\_VEHICLEMotor-cycle   
## 0.84611923 0.03916362   
## TYPE\_VEHICLEPick-up TYPE\_VEHICLESpecial construction   
## 1.12067918 1.08336020   
## TYPE\_VEHICLEStation Wagones TYPE\_VEHICLETanker   
## 0.72375725 0.62874336   
## TYPE\_VEHICLETrailers and semitrailers TYPE\_VEHICLETruck   
## 0.57132996 1.04813977   
## MAKEDAEWOO MAKEFIAT   
## 0.47075570 0.46435500   
## MAKEFORD MAKEGEELY   
## 0.76470002 1.03999568   
## MAKEGENLYON MAKEHYUNDAI   
## 0.45912487 0.70032593   
## MAKEISUZU MAKEIVECO   
## 0.57438032 0.48635487   
## MAKELIFAN MAKEMAZDA   
## 0.81588754 0.95992131   
## MAKEMERCEDES MAKEMITSUBISHI   
## 0.40290508 0.65000623   
## MAKENISSAN MAKERENAULT   
## 0.76221367 0.67627769   
## MAKESINO MAKESINO HOWO   
## 0.58199841 0.40416499   
## MAKESUZUKI MAKETATA   
## 0.50419482 0.34010015   
## MAKETOYOTA AGE\_VEHICLE   
## 0.86790475 0.95618504   
## SEATS\_NUM CCM\_TON   
## 1.00878598 1.00003125   
## INSURED\_VALUE PREMIUM   
## 0.99999986 1.00001895

## Overdispersion ratio (Deviance / DF): 0.6884257

## Goodness-of-Fit p-value: 1

## Variance Inflation Factor (VIF) values for all predictor variables:

## GVIF Df GVIF^(1/(2\*Df))  
## SEX 1.624635 2 1.128987  
## INSR\_TYPE 14.051227 1 3.748497  
## TYPE\_VEHICLE 276.509321 8 1.421042  
## MAKE 16.507471 19 1.076575  
## AGE\_VEHICLE 1.888760 1 1.374322  
## SEATS\_NUM 2.739027 1 1.655001  
## CCM\_TON 5.312166 1 2.304814  
## INSURED\_VALUE 3.556136 1 1.885772  
## PREMIUM 2.909685 1 1.705780

 The analysis of the Poisson model for predicting claim frequency indicated no overdispersion. The calculated overdispersion value, represented by the ratio of deviance to degrees of freedom, is 0.688, which is significantly below 1. This suggests that the model does not overestimate variance in the data, and overdispersion is not an issue. The Goodness-of-Fit test further confirms the adequacy of the model, as the p-value of 1 supports the null hypothesis that the model sufficiently describes the data.

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 analysis of the Poisson model reveals significant multicollinearity, reflected in extremely high VIF values for some variables. Notably, the variables TYPE\_VEHICLEMotor-cycle (VIF of 200.88), TYPE\_VEHICLETruck (107.23), TYPE\_VEHICLEPick-up (82.92), INSR\_TYPEPrivate (80.14), and MAKETOYOTA (50.55) stand out. These high values indicate that these variables are highly correlated with other predictors, especially among the vehicle type variables, suggesting redundancy within the model. Other variables, such as CCM\_TON (33.28), MAKEISUZU (28.86), MAKEIVECO (23.94), and MAKETATA (30.58), also display moderate multicollinearity, while some variables, like MAKEGEELY (3.99), show lower VIF values and are less strongly correlated with other predictors.

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.

### Massnahme 1): Zero-Inflated Poisson (ZIP) model (eventuell weglassen, machts naemlich ned besser)

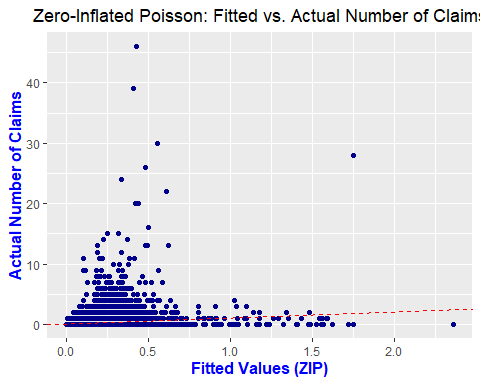
##   
## Call:  
## zeroinfl(formula = AMOUNT\_CLAIMS ~ SEX + INSR\_TYPE + TYPE\_VEHICLE + MAKE +   
## AGE\_VEHICLE + SEATS\_NUM + CCM\_TON + INSURED\_VALUE + PREMIUM, data = dat\_amount\_claims,   
## dist = "poisson")  
##   
## Pearson residuals:  
## Min 1Q Median 3Q Max   
## -1.2570 -0.4249 -0.3606 -0.2780 62.1189   
##   
## Count model coefficients (poisson with log link):  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -7.191e-01 NaN NaN NaN  
## SEXMale -4.053e-01 NaN NaN NaN  
## SEXFemale -4.184e-01 NaN NaN NaN  
## INSR\_TYPEPrivate 6.189e-01 NaN NaN NaN  
## TYPE\_VEHICLEBus 1.909e-03 NaN NaN NaN  
## TYPE\_VEHICLEMotor-cycle -1.402e+00 NaN NaN NaN  
## TYPE\_VEHICLEPick-up 6.602e-01 NaN NaN NaN  
## TYPE\_VEHICLESpecial construction 1.170e-01 NaN NaN NaN  
## TYPE\_VEHICLEStation Wagones -4.263e-01 NaN NaN NaN  
## TYPE\_VEHICLETanker 4.055e-01 NaN NaN NaN  
## TYPE\_VEHICLETrailers and semitrailers 2.259e+00 NaN NaN NaN  
## TYPE\_VEHICLETruck 3.814e-01 NaN NaN NaN  
## MAKEDAEWOO -1.268e+00 NaN NaN NaN  
## MAKEFIAT -1.415e+00 NaN NaN NaN  
## MAKEFORD -1.975e-01 NaN NaN NaN  
## MAKEGEELY -2.777e-01 NaN NaN NaN  
## MAKEGENLYON -1.688e+00 NaN NaN NaN  
## MAKEHYUNDAI -5.009e-01 NaN NaN NaN  
## MAKEISUZU -1.033e+00 NaN NaN NaN  
## MAKEIVECO -1.635e+00 NaN NaN NaN  
## MAKELIFAN -4.342e-01 NaN NaN NaN  
## MAKEMAZDA -6.912e-01 NaN NaN NaN  
## MAKEMERCEDES -1.686e+00 NaN NaN NaN  
## MAKEMITSUBISHI -8.388e-01 NaN NaN NaN  
## MAKENISSAN -6.548e-01 NaN NaN NaN  
## MAKERENAULT 1.907e-01 NaN NaN NaN  
## MAKESINO -1.318e+00 NaN NaN NaN  
## MAKESINO HOWO -1.629e+00 NaN NaN NaN  
## MAKESUZUKI -1.470e+00 NaN NaN NaN  
## MAKETATA -1.354e+00 NaN NaN NaN  
## MAKETOYOTA -3.732e-01 NaN NaN NaN  
## AGE\_VEHICLE -5.247e-02 NaN NaN NaN  
## SEATS\_NUM 1.113e-02 NaN NaN NaN  
## CCM\_TON 8.380e-05 NaN NaN NaN  
## INSURED\_VALUE -1.374e-07 NaN NaN NaN  
## PREMIUM 4.805e-06 NaN NaN NaN  
##   
## Zero-inflation model coefficients (binomial with logit link):  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.394e+00 NaN NaN NaN  
## SEXMale -2.390e-02 NaN NaN NaN  
## SEXFemale 2.226e-01 NaN NaN NaN  
## INSR\_TYPEPrivate 1.712e+00 NaN NaN NaN  
## TYPE\_VEHICLEBus 1.145e+00 NaN NaN NaN  
## TYPE\_VEHICLEMotor-cycle 1.431e+01 NaN NaN NaN  
## TYPE\_VEHICLEPick-up 2.022e+00 NaN NaN NaN  
## TYPE\_VEHICLESpecial construction 2.980e-01 NaN NaN NaN  
## TYPE\_VEHICLEStation Wagones -3.216e-01 NaN NaN NaN  
## TYPE\_VEHICLETanker 2.980e+00 NaN NaN NaN  
## TYPE\_VEHICLETrailers and semitrailers 6.227e+00 NaN NaN NaN  
## TYPE\_VEHICLETruck 1.365e+00 NaN NaN NaN  
## MAKEDAEWOO -1.438e+00 NaN NaN NaN  
## MAKEFIAT -1.796e+00 NaN NaN NaN  
## MAKEFORD 1.158e-01 NaN NaN NaN  
## MAKEGEELY -6.284e-01 NaN NaN NaN  
## MAKEGENLYON -3.602e+00 NaN NaN NaN  
## MAKEHYUNDAI -2.584e-01 NaN NaN NaN  
## MAKEISUZU -8.665e-01 NaN NaN NaN  
## MAKEIVECO -4.163e+00 NaN NaN NaN  
## MAKELIFAN -4.345e-01 NaN NaN NaN  
## MAKEMAZDA -1.507e+00 NaN NaN NaN  
## MAKEMERCEDES -1.779e+01 NaN NaN NaN  
## MAKEMITSUBISHI -8.424e-01 NaN NaN NaN  
## MAKENISSAN -8.768e-01 NaN NaN NaN  
## MAKERENAULT 1.410e+00 NaN NaN NaN  
## MAKESINO -2.008e+00 NaN NaN NaN  
## MAKESINO HOWO -1.975e+00 NaN NaN NaN  
## MAKESUZUKI -1.288e+01 NaN NaN NaN  
## MAKETATA -3.657e-01 NaN NaN NaN  
## MAKETOYOTA -3.687e-01 NaN NaN NaN  
## AGE\_VEHICLE -2.360e-02 NaN NaN NaN  
## SEATS\_NUM 1.021e-02 NaN NaN NaN  
## CCM\_TON 2.110e-04 NaN NaN NaN  
## INSURED\_VALUE 4.069e-07 NaN NaN NaN  
## PREMIUM -1.012e-04 NaN NaN NaN  
##   
## Number of iterations in BFGS optimization: 146   
## Log-likelihood: -3.815e+04 on 72 Df

## count\_(Intercept)   
## -7.190522e-01   
## count\_SEXMale   
## -4.052865e-01   
## count\_SEXFemale   
## -4.183935e-01   
## count\_INSR\_TYPEPrivate   
## 6.189175e-01   
## count\_TYPE\_VEHICLEBus   
## 1.908625e-03   
## count\_TYPE\_VEHICLEMotor-cycle   
## -1.401767e+00   
## count\_TYPE\_VEHICLEPick-up   
## 6.602076e-01   
## count\_TYPE\_VEHICLESpecial construction   
## 1.170154e-01   
## count\_TYPE\_VEHICLEStation Wagones   
## -4.263271e-01   
## count\_TYPE\_VEHICLETanker   
## 4.055105e-01   
## count\_TYPE\_VEHICLETrailers and semitrailers   
## 2.258682e+00   
## count\_TYPE\_VEHICLETruck   
## 3.814461e-01   
## count\_MAKEDAEWOO   
## -1.267855e+00   
## count\_MAKEFIAT   
## -1.415463e+00   
## count\_MAKEFORD   
## -1.975283e-01   
## count\_MAKEGEELY   
## -2.776996e-01   
## count\_MAKEGENLYON   
## -1.688237e+00   
## count\_MAKEHYUNDAI   
## -5.008562e-01   
## count\_MAKEISUZU   
## -1.032635e+00   
## count\_MAKEIVECO   
## -1.634795e+00   
## count\_MAKELIFAN   
## -4.342070e-01   
## count\_MAKEMAZDA   
## -6.912153e-01   
## count\_MAKEMERCEDES   
## -1.685992e+00   
## count\_MAKEMITSUBISHI   
## -8.387773e-01   
## count\_MAKENISSAN   
## -6.548468e-01   
## count\_MAKERENAULT   
## 1.906575e-01   
## count\_MAKESINO   
## -1.318192e+00   
## count\_MAKESINO HOWO   
## -1.628814e+00   
## count\_MAKESUZUKI   
## -1.469926e+00   
## count\_MAKETATA   
## -1.354117e+00   
## count\_MAKETOYOTA   
## -3.732259e-01   
## count\_AGE\_VEHICLE   
## -5.247376e-02   
## count\_SEATS\_NUM   
## 1.113326e-02   
## count\_CCM\_TON   
## 8.379862e-05   
## count\_INSURED\_VALUE   
## -1.374022e-07   
## count\_PREMIUM   
## 4.805172e-06   
## zero\_(Intercept)   
## -1.394381e+00   
## zero\_SEXMale   
## -2.390421e-02   
## zero\_SEXFemale   
## 2.226171e-01   
## zero\_INSR\_TYPEPrivate   
## 1.712252e+00   
## zero\_TYPE\_VEHICLEBus   
## 1.145035e+00   
## zero\_TYPE\_VEHICLEMotor-cycle   
## 1.431269e+01   
## zero\_TYPE\_VEHICLEPick-up   
## 2.022223e+00   
## zero\_TYPE\_VEHICLESpecial construction   
## 2.980475e-01   
## zero\_TYPE\_VEHICLEStation Wagones   
## -3.216156e-01   
## zero\_TYPE\_VEHICLETanker   
## 2.979559e+00   
## zero\_TYPE\_VEHICLETrailers and semitrailers   
## 6.226884e+00   
## zero\_TYPE\_VEHICLETruck   
## 1.364634e+00   
## zero\_MAKEDAEWOO   
## -1.437761e+00   
## zero\_MAKEFIAT   
## -1.795543e+00   
## zero\_MAKEFORD   
## 1.158134e-01   
## zero\_MAKEGEELY   
## -6.283996e-01   
## zero\_MAKEGENLYON   
## -3.602155e+00   
## zero\_MAKEHYUNDAI   
## -2.584406e-01   
## zero\_MAKEISUZU   
## -8.664861e-01   
## zero\_MAKEIVECO   
## -4.163215e+00   
## zero\_MAKELIFAN   
## -4.345035e-01   
## zero\_MAKEMAZDA   
## -1.507104e+00   
## zero\_MAKEMERCEDES   
## -1.778728e+01   
## zero\_MAKEMITSUBISHI   
## -8.423770e-01   
## zero\_MAKENISSAN   
## -8.767542e-01   
## zero\_MAKERENAULT   
## 1.410242e+00   
## zero\_MAKESINO   
## -2.008013e+00   
## zero\_MAKESINO HOWO   
## -1.974716e+00   
## zero\_MAKESUZUKI   
## -1.288104e+01   
## zero\_MAKETATA   
## -3.656970e-01   
## zero\_MAKETOYOTA   
## -3.686973e-01   
## zero\_AGE\_VEHICLE   
## -2.360178e-02   
## zero\_SEATS\_NUM   
## 1.021244e-02   
## zero\_CCM\_TON   
## 2.109916e-04   
## zero\_INSURED\_VALUE   
## 4.069011e-07   
## zero\_PREMIUM   
## -1.011932e-04

## count\_(Intercept)   
## 4.872138e-01   
## count\_SEXMale   
## 6.667858e-01   
## count\_SEXFemale   
## 6.581032e-01   
## count\_INSR\_TYPEPrivate   
## 1.856917e+00   
## count\_TYPE\_VEHICLEBus   
## 1.001910e+00   
## count\_TYPE\_VEHICLEMotor-cycle   
## 2.461616e-01   
## count\_TYPE\_VEHICLEPick-up   
## 1.935194e+00   
## count\_TYPE\_VEHICLESpecial construction   
## 1.124137e+00   
## count\_TYPE\_VEHICLEStation Wagones   
## 6.529028e-01   
## count\_TYPE\_VEHICLETanker   
## 1.500068e+00   
## count\_TYPE\_VEHICLETrailers and semitrailers   
## 9.570464e+00   
## count\_TYPE\_VEHICLETruck   
## 1.464401e+00   
## count\_MAKEDAEWOO   
## 2.814347e-01   
## count\_MAKEFIAT   
## 2.428132e-01   
## count\_MAKEFORD   
## 8.207569e-01   
## count\_MAKEGEELY   
## 7.575243e-01   
## count\_MAKEGENLYON   
## 1.848452e-01   
## count\_MAKEHYUNDAI   
## 6.060116e-01   
## count\_MAKEISUZU   
## 3.560673e-01   
## count\_MAKEIVECO   
## 1.949924e-01   
## count\_MAKELIFAN   
## 6.477782e-01   
## count\_MAKEMAZDA   
## 5.009669e-01   
## count\_MAKEMERCEDES   
## 1.852606e-01   
## count\_MAKEMITSUBISHI   
## 4.322387e-01   
## count\_MAKENISSAN   
## 5.195216e-01   
## count\_MAKERENAULT   
## 1.210045e+00   
## count\_MAKESINO   
## 2.676187e-01   
## count\_MAKESINO HOWO   
## 1.961622e-01   
## count\_MAKESUZUKI   
## 2.299424e-01   
## count\_MAKETATA   
## 2.581751e-01   
## count\_MAKETOYOTA   
## 6.885097e-01   
## count\_AGE\_VEHICLE   
## 9.488792e-01   
## count\_SEATS\_NUM   
## 1.011195e+00   
## count\_CCM\_TON   
## 1.000084e+00   
## count\_INSURED\_VALUE   
## 9.999999e-01   
## count\_PREMIUM   
## 1.000005e+00   
## zero\_(Intercept)   
## 2.479864e-01   
## zero\_SEXMale   
## 9.763792e-01   
## zero\_SEXFemale   
## 1.249342e+00   
## zero\_INSR\_TYPEPrivate   
## 5.541429e+00   
## zero\_TYPE\_VEHICLEBus   
## 3.142553e+00   
## zero\_TYPE\_VEHICLEMotor-cycle   
## 1.644085e+06   
## zero\_TYPE\_VEHICLEPick-up   
## 7.555101e+00   
## zero\_TYPE\_VEHICLESpecial construction   
## 1.347226e+00   
## zero\_TYPE\_VEHICLEStation Wagones   
## 7.249768e-01   
## zero\_TYPE\_VEHICLETanker   
## 1.967913e+01   
## zero\_TYPE\_VEHICLETrailers and semitrailers   
## 5.061759e+02   
## zero\_TYPE\_VEHICLETruck   
## 3.914289e+00   
## zero\_MAKEDAEWOO   
## 2.374587e-01   
## zero\_MAKEFIAT   
## 1.660373e-01   
## zero\_MAKEFORD   
## 1.122786e+00   
## zero\_MAKEGEELY   
## 5.334448e-01   
## zero\_MAKEGENLYON   
## 2.726490e-02   
## zero\_MAKEHYUNDAI   
## 7.722549e-01   
## zero\_MAKEISUZU   
## 4.204263e-01   
## zero\_MAKEIVECO   
## 1.555746e-02   
## zero\_MAKELIFAN   
## 6.475861e-01   
## zero\_MAKEMAZDA   
## 2.215506e-01   
## zero\_MAKEMERCEDES   
## 1.884007e-08   
## zero\_MAKEMITSUBISHI   
## 4.306856e-01   
## zero\_MAKENISSAN   
## 4.161314e-01   
## zero\_MAKERENAULT   
## 4.096946e+00   
## zero\_MAKESINO   
## 1.342552e-01   
## zero\_MAKESINO HOWO   
## 1.388008e-01   
## zero\_MAKESUZUKI   
## 2.545869e-06   
## zero\_MAKETATA   
## 6.937130e-01   
## zero\_MAKETOYOTA   
## 6.916347e-01   
## zero\_AGE\_VEHICLE   
## 9.766746e-01   
## zero\_SEATS\_NUM   
## 1.010265e+00   
## zero\_CCM\_TON   
## 1.000211e+00   
## zero\_INSURED\_VALUE   
## 1.000000e+00   
## zero\_PREMIUM   
## 9.998988e-01

## Variance Inflation Factor (VIF) values for all predictor variables:

## GVIF Df GVIF^(1/(2\*Df))  
## SEX NaN 2 NaN  
## INSR\_TYPE NaN 1 NaN  
## TYPE\_VEHICLE NaN 8 NaN  
## MAKE NaN 19 NaN  
## AGE\_VEHICLE NaN 1 NaN  
## SEATS\_NUM NaN 1 NaN  
## CCM\_TON NaN 1 NaN  
## INSURED\_VALUE NaN 1 NaN  
## PREMIUM NaN 1 NaN



## Binomial

# TODO

##   
## Call:  
## glm(formula = CLAIM\_PAID ~ SEX + INSR\_TYPE + USAGE + TYPE\_VEHICLE +   
## MAKE + AGE\_VEHICLE + SEATS\_NUM + CCM\_TON\_log + INSURED\_VALUE\_log +   
## PREMIUM\_log + AMOUNT\_CLAIMS\_PAID, family = binomial(link = "logit"),   
## data = clean\_dat\_motor)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.564095 0.329873 -4.742 2.12e-06 \*\*\*  
## SEXMale -0.229080 0.032602 -7.026 2.12e-12 \*\*\*  
## SEXFemale -0.254521 0.046106 -5.520 3.38e-08 \*\*\*  
## INSR\_TYPEPrivate 0.009493 0.137611 0.069 0.945003   
## USAGECar Hires -0.509402 0.277510 -1.836 0.066414 .   
## USAGEFare Paying Passengers -0.129315 0.169739 -0.762 0.446152   
## USAGEGeneral Cartage 0.210532 0.358874 0.587 0.557441   
## USAGEOwn Goods -0.162464 0.356428 -0.456 0.648526   
## USAGEOwn service -0.179487 0.160614 -1.118 0.263778   
## USAGEPrivate 0.041574 0.168232 0.247 0.804814   
## TYPE\_VEHICLEBus -0.095838 0.146551 -0.654 0.513140   
## TYPE\_VEHICLEMotor-cycle -3.413157 0.466174 -7.322 2.45e-13 \*\*\*  
## TYPE\_VEHICLEPick-up 0.029870 0.344943 0.087 0.930995   
## TYPE\_VEHICLESpecial construction 0.202535 0.225865 0.897 0.369875   
## TYPE\_VEHICLEStation Wagones -0.219052 0.043670 -5.016 5.27e-07 \*\*\*  
## TYPE\_VEHICLETanker -0.405600 0.360252 -1.126 0.260217   
## TYPE\_VEHICLETrailers and semitrailers -1.731847 0.513830 -3.370 0.000750 \*\*\*  
## TYPE\_VEHICLETruck -0.067082 0.349488 -0.192 0.847786   
## MAKEDAEWOO -0.489734 0.112875 -4.339 1.43e-05 \*\*\*  
## MAKEFIAT -0.620431 0.135434 -4.581 4.63e-06 \*\*\*  
## MAKEFORD -0.182604 0.096159 -1.899 0.057568 .   
## MAKEGEELY -0.038824 0.120341 -0.323 0.746985   
## MAKEGENLYON -0.656277 0.172400 -3.807 0.000141 \*\*\*  
## MAKEHYUNDAI -0.503550 0.098210 -5.127 2.94e-07 \*\*\*  
## MAKEISUZU -0.572474 0.072726 -7.872 3.50e-15 \*\*\*  
## MAKEIVECO -0.497240 0.083226 -5.975 2.31e-09 \*\*\*  
## MAKELIFAN -0.096894 0.094482 -1.026 0.305113   
## MAKEMAZDA 0.071264 0.098610 0.723 0.469873   
## MAKEMERCEDES -0.720153 0.113784 -6.329 2.47e-10 \*\*\*  
## MAKEMITSUBISHI -0.296876 0.074916 -3.963 7.41e-05 \*\*\*  
## MAKENISSAN -0.265349 0.067550 -3.928 8.56e-05 \*\*\*  
## MAKERENAULT -1.487104 0.162638 -9.144 < 2e-16 \*\*\*  
## MAKESINO -0.920258 0.145515 -6.324 2.55e-10 \*\*\*  
## MAKESINO HOWO -0.938988 0.085814 -10.942 < 2e-16 \*\*\*  
## MAKESUZUKI -0.619487 0.153757 -4.029 5.60e-05 \*\*\*  
## MAKETATA -0.713283 0.133991 -5.323 1.02e-07 \*\*\*  
## MAKETOYOTA -0.155056 0.060318 -2.571 0.010151 \*   
## AGE\_VEHICLE -0.028568 0.002219 -12.875 < 2e-16 \*\*\*  
## SEATS\_NUM 0.002766 0.001296 2.134 0.032841 \*   
## CCM\_TON\_log -0.086870 0.025573 -3.397 0.000682 \*\*\*  
## INSURED\_VALUE\_log -0.286045 0.028108 -10.177 < 2e-16 \*\*\*  
## PREMIUM\_log 0.541475 0.023015 23.527 < 2e-16 \*\*\*  
## AMOUNT\_CLAIMS\_PAID 0.228941 0.024596 9.308 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 80340 on 99999 degrees of freedom  
## Residual deviance: 77630 on 99957 degrees of freedom  
## AIC: 77716  
##   
## Number of Fisher Scoring iterations: 7

## Single term deletions  
##   
## Model:  
## CLAIM\_PAID ~ SEX + INSR\_TYPE + USAGE + TYPE\_VEHICLE + MAKE +   
## AGE\_VEHICLE + SEATS\_NUM + CCM\_TON\_log + INSURED\_VALUE\_log +   
## PREMIUM\_log + AMOUNT\_CLAIMS\_PAID  
## Df Deviance AIC LRT Pr(>Chi)   
## <none> 77630 77716   
## SEX 2 77688 77770 57.63 3.056e-13 \*\*\*  
## INSR\_TYPE 1 77630 77714 0.00 0.9450168   
## USAGE 6 77688 77762 57.66 1.341e-10 \*\*\*  
## TYPE\_VEHICLE 8 77855 77925 224.89 < 2.2e-16 \*\*\*  
## MAKE 19 77914 77962 284.04 < 2.2e-16 \*\*\*  
## AGE\_VEHICLE 1 77800 77884 170.25 < 2.2e-16 \*\*\*  
## SEATS\_NUM 1 77635 77719 4.47 0.0344357 \*   
## CCM\_TON\_log 1 77641 77725 11.29 0.0007782 \*\*\*  
## INSURED\_VALUE\_log 1 77735 77819 104.45 < 2.2e-16 \*\*\*  
## PREMIUM\_log 1 78270 78354 639.38 < 2.2e-16 \*\*\*  
## AMOUNT\_CLAIMS\_PAID 1 77713 77797 83.06 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## AIC des logit-Modells: 77716.2

## Pseudo.R.squared  
## McFadden 0.0337310  
## Cox and Snell (ML) 0.0267356  
## Nagelkerke (Cragg and Uhler) 0.0484169

## Globale Teststatistik (TD): 2709.954

## p-Wert des globalen Tests: 0

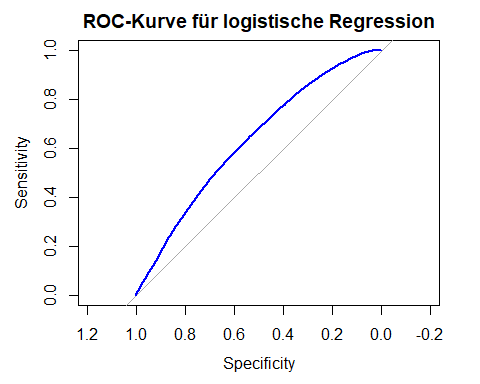
## Zusammenfassung der Deviance-Residuals:

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.1184 -0.5935 -0.5083 -0.1861 -0.4025 3.5157

## Overdispersion Ratio (Deviance / DF): 0.7766359

## Likelihood-Ratio-Test-Statistik gegen Nullmodell: 2709.954

## AUC-Wert des Modells: 0.6292681



Modellgüte Die Berechnung von Pseudo-R²-Werten ergab niedrige Werte (McFadden: 3,4% und Nagelkerke: 4,8%). Diese niedrigen Werte zeigen, dass das Modell nur einen kleinen Teil der Varianz in der Zielvariablen CLAIM\_PAID erklären kann. Es handelt sich also um ein Modell mit begrenzter Erklärungskraft.

Die Modellgüte wurde mittels Pseudo-R²-Werten bewertet. Der McFadden-Wert (0.0337) und der Nagelkerke-Wert (0.0484) sind relativ niedrig und deuten darauf hin, dass das Modell nur einen kleinen Anteil der Varianz in der Zielvariablen erklärt. Diese Werte deuten darauf hin, dass das Modell nur begrenzte Vorhersagekraft besitzt und möglicherweise noch wichtige Prädiktoren fehlen oder die Variablenkategorien weitere Anpassungen benötigen.

Globale Modellsignifikanz: Der globale F-Test (Wald-Test) zeigt, dass die Gesamtmodellstatistik (TD) 2709.95 beträgt, und der p-Wert des Tests ist nahe null. Dies deutet darauf hin, dass das Modell als Ganzes statistisch signifikant ist. Das bedeutet, dass zumindest eine der erklärenden Variablen in signifikantem Zusammenhang mit der Zielvariablen steht und somit das Modell besser ist als ein Nullmodell (Modell ohne erklärende Variablen).

Ergebnisse des Likelihood-Ratio-Tests Gemäss dem Likelihood-Ratio-Test, scheint (INSR\_TYPE) keinen signifikanten Einfluss zu haben.

Der AUC-Wert (Fläche unter der Kurve) von 0,629 deutet auf eine relativ moderate Diskriminierungsfähigkeit hin, was bedeutet, dass das Modell etwas besser als zufälliges Raten ist, aber noch Raum für Verbesserungen bietet.

Modellanpassung und Signifikanz: Der AIC (Akaike Informationskriterium) des Modells beträgt 77716,2. Ein niedrigerer AIC-Wert zeigt in der Regel eine bessere Anpassung an. Das Overdispersion-Verhältnis beträgt 0,7766, was unter 1 liegt und auf keine signifikante Overdispersion hinweist. Mehrere Prädiktoren sind statistisch signifikant (z.B. SEX, TYPE\_VEHICLE, MAKE, PREMIUM\_log, usw.).

Likelihood-Ratio-Test:

Der Likelihood-Ratio-Test gegen das Nullmodell ergibt einen Wert von 2709,954, was darauf hinweist, dass das vollständige Modell signifikant besser ist als das Nullmodell.

ROC und AUC: Die ROC-Kurve zeigt den Kompromiss zwischen Sensitivität und Spezifität. Der AUC-Wert beträgt 0,629, was auf eine gewisse Vorhersagekraft des Modells hindeutet, jedoch möglicherweise weitere Optimierungen erfordert.

Pseudo R-Quadrat: Der McFadden-Pseudo-R-Quadrat beträgt 0,0337, was auf eine moderate Anpassung hinweist, während der Nagelkerke-Pseudo-R-Quadrat 0,0484 beträgt, was ebenfalls eine begrenzte Erklärungskraft des Modells andeutet.

### Massnahme 1): Entfernen von Variablen

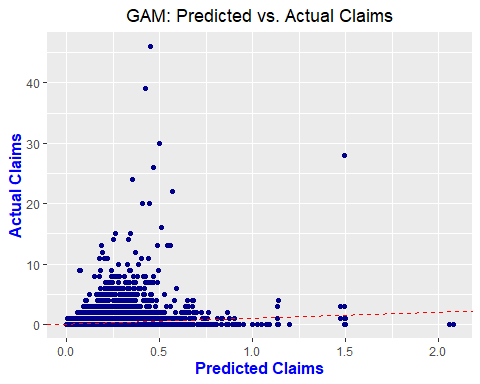
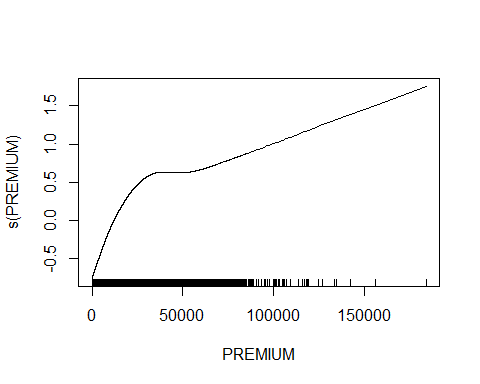
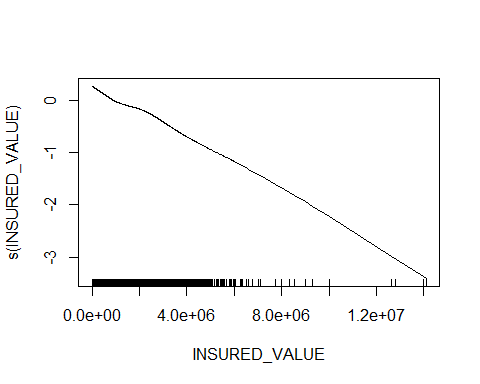
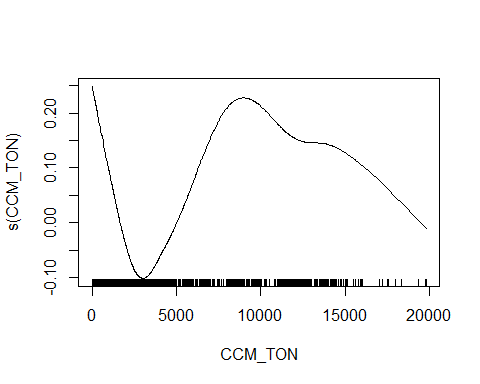
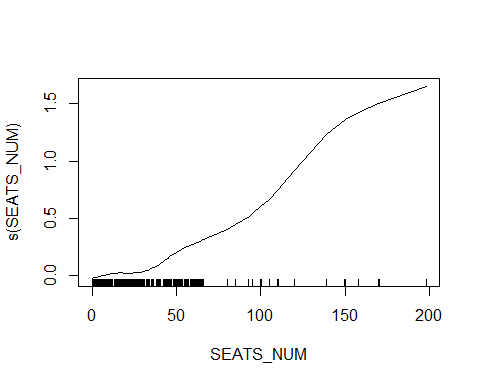
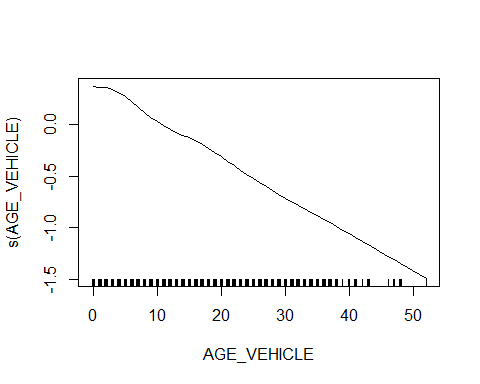
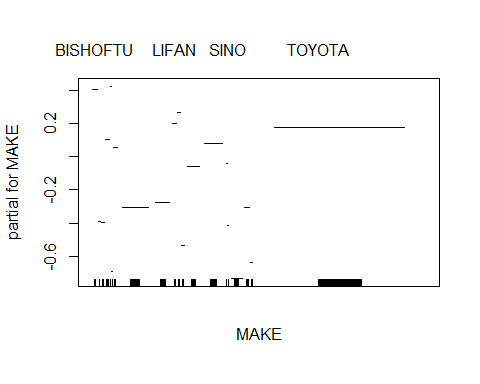
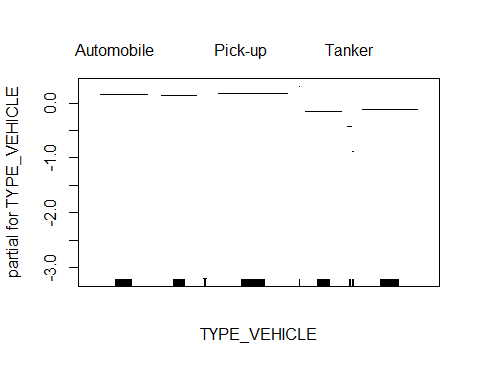
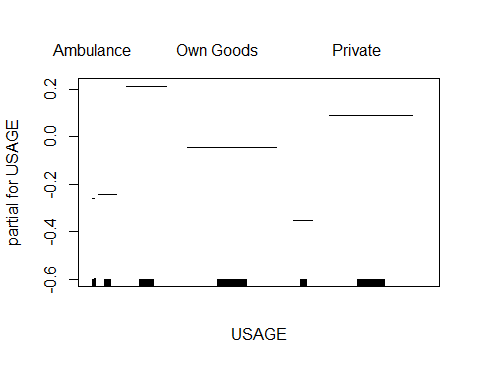
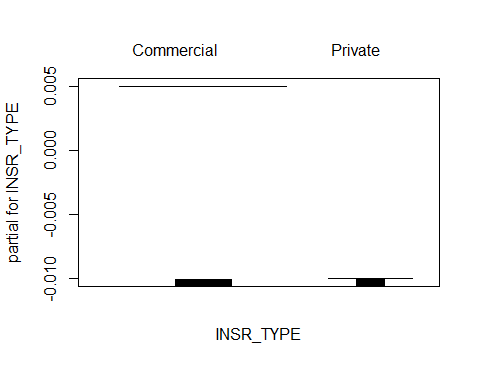
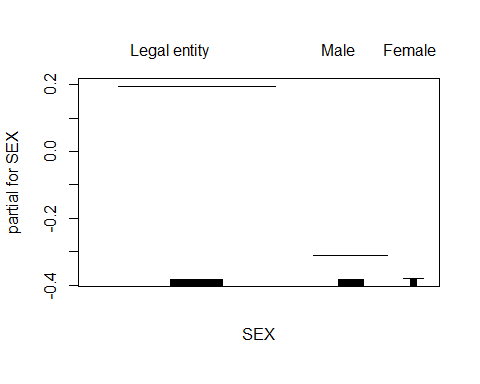
AIC-Wert: Der AIC des neuen logistischen Modells beträgt 77714.2, was einen minimalen Rückgang gegenüber dem vorherigen Modell darstellt, aber die Anpassung nicht wesentlich beeinflusst.

Zusammengefasst hat das Entfernen der Variable INSR\_TYPE nur geringe Auswirkungen auf die Modellgüte und die Erklärungskraft. Falls gewünscht, können weitere Anpassungen oder zusätzliche Prädiktoren in Betracht gezogen werden, um die Modellleistung zu verbessern.

## Generalised Additive Model (GAM)

# TODO description

##   
## Call: gam(formula = AMOUNT\_CLAIMS ~ SEX + INSR\_TYPE + USAGE + TYPE\_VEHICLE +   
## MAKE + s(AGE\_VEHICLE) + s(SEATS\_NUM) + s(CCM\_TON) + s(INSURED\_VALUE) +   
## s(PREMIUM), family = poisson(link = "log"), data = dat\_amount\_claims)  
## Deviance Residuals:  
## Min 1Q Median 3Q Max   
## -2.0389 -0.6421 -0.5314 -0.3996 18.3018   
##   
## (Dispersion Parameter for poisson family taken to be 1)  
##   
## Null Deviance: 56192.36 on 77186 degrees of freedom  
## Residual Deviance: 52472.32 on 77130 degrees of freedom  
## AIC: 76898.29   
##   
## Number of Local Scoring Iterations: NA   
##   
## Anova for Parametric Effects  
## Df Sum Sq Mean Sq F value Pr(>F)   
## SEX 2 451 225.39 151.3069 < 2.2e-16 \*\*\*  
## INSR\_TYPE 1 7 7.27 4.8811 0.02716 \*   
## USAGE 6 136 22.73 15.2556 < 2.2e-16 \*\*\*  
## TYPE\_VEHICLE 8 295 36.82 24.7160 < 2.2e-16 \*\*\*  
## MAKE 19 497 26.15 17.5547 < 2.2e-16 \*\*\*  
## s(AGE\_VEHICLE) 1 862 862.48 578.9823 < 2.2e-16 \*\*\*  
## s(SEATS\_NUM) 1 44 43.60 29.2663 6.327e-08 \*\*\*  
## s(CCM\_TON) 1 8 7.86 5.2742 0.02165 \*   
## s(INSURED\_VALUE) 1 7 6.90 4.6338 0.03135 \*   
## s(PREMIUM) 1 561 560.90 376.5322 < 2.2e-16 \*\*\*  
## Residuals 77130 114897 1.49   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Anova for Nonparametric Effects  
## Npar Df Npar Chisq P(Chi)   
## (Intercept)   
## SEX   
## INSR\_TYPE   
## USAGE   
## TYPE\_VEHICLE   
## MAKE   
## s(AGE\_VEHICLE) 3 37.18 4.205e-08 \*\*\*  
## s(SEATS\_NUM) 3 18.73 0.0003113 \*\*\*  
## s(CCM\_TON) 3 101.34 < 2.2e-16 \*\*\*  
## s(INSURED\_VALUE) 3 21.37 8.834e-05 \*\*\*  
## s(PREMIUM) 3 380.89 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



## Deviance of the GAM model: 52472.32

## AIC of the GAM model: 76898.29

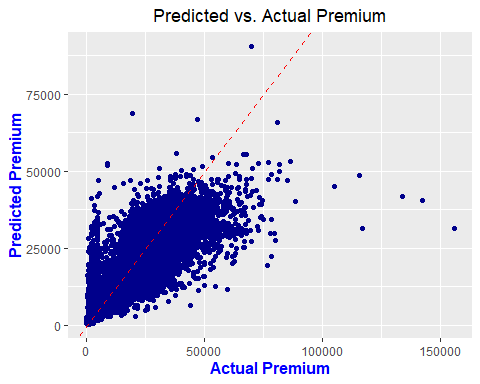
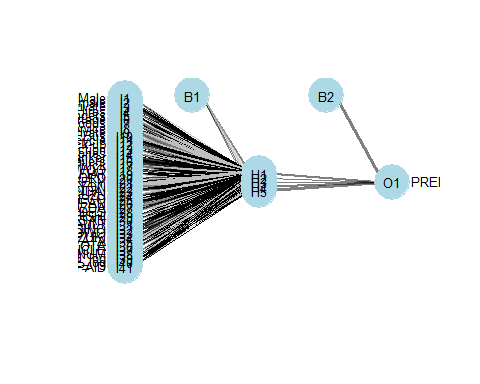
## Pseudo.R.squared  
## McFadden NA  
## Cox and Snell (ML) NA  
## Nagelkerke (Cragg and Uhler) NA

## Overdispersion ratio (Deviance / DF) for the GAM model: 0.6803101

## p-value for the GAM model: 1

## Neural Network

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.

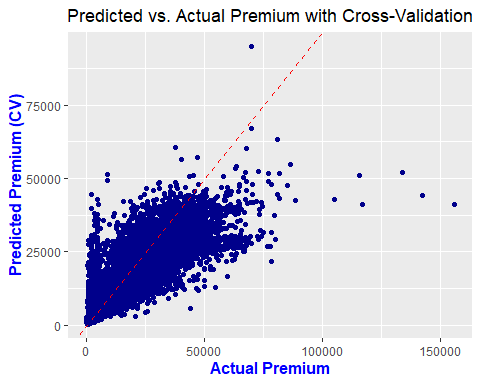
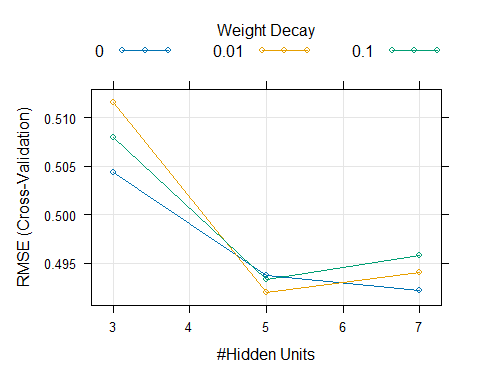


The neural network model was trained to predict the log-transformed premium amount based on the characteristics of the insured vehicles. The model was fitted using the nnet package, with the training data split into 80% training and 20% testing sets. The neural network model was trained with a hidden 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 % |

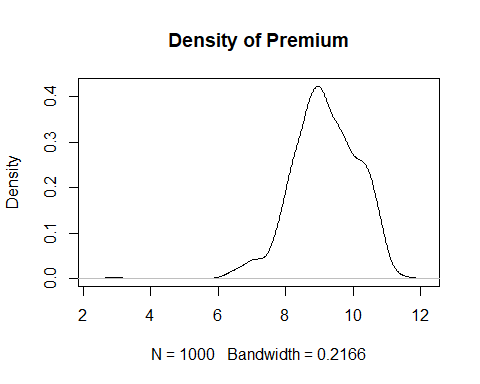
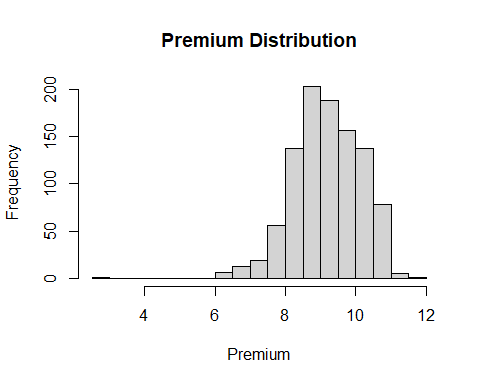
The weights decay plot shows how the RMSE (Root Mean Square Error) from cross-validation varies with the number of hidden units in a neural network for different weight decay values (0, 0.01, and 0.1). As the number of hidden units increases from 3 to 5, RMSE decreases across all weight decay values, 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)

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

# TODO

# TODO remove of do whatever you want

### Massnahme 1): Quasi-Poisson-Regression

Switching to a quasi-Poisson model to account for overdispersion led to improvements compared to the original Poisson model. The residuals vs. fitted plot shows a reduced dispersion of the residuals at higher estimated values, which indicates a better fit of the variance, although heteroscedasticity still exists. The QQ plot of the residuals shows an improved fit to the theoretical normal distribution, especially in the middle range, while deviations at the edges remain, indicating extreme values or modelling errors. By adjusting the dispersion parameter (29.765) in the quasi-Poisson model, the increased variance compared to the Poisson model is adequately taken into account. The F-test confirms the significance of the variables ‘SEX’, ‘TYPE\_VEHICLE’, ‘MAKE’, ‘AGE\_VEHICLE’ and ‘SEATS\_NUM’. Despite these improvements, there are still slight anomalies in the residuals.