Project One

Dana Saeed

2/26/2021

Project Objective

The purpose of this project is to build a risk analytics model to understand the renewal potential and claim propensity of Existing Customers under Personal Auto Insurance Lines. This data contains information about 127 variables which also includes the personal information of the driver, their age, profession, marital status, and other demographics. The data also includes the details related to the car like to model and make of the car, not only that it also contains details regarding the different type of coverages and their code. Some of the data is relevant to the model that we are building but a lot of it is of no use , because either it has weak or no correlation with the target variable or it does not add anything significant to the predictions like the name of the driver and their distance to work and their gender. We would use Logistic regression, Random Forest Classification and KNN to predict the target variable. We would also use the Ensemble techniques to fine tune our model and then select the best fit if the model without overfitting.

#Assumptions There are a few assumptions considered: * The Sample size is adequate to perform techniques like logistic regression and Random Forest Classification. * All the necessary packages are installed in R * Working Directly is set to appropriate folder and file is in CSV format

Imprting required libraries

```
#first we check that if the required libraries are downloaded or not
if(!require("ggplot2"))
{
  install.packages("ggplot2",repos = "http://cran.us.r-project.org")

}
```

```
## Loading required package: ggplot2
```

```
if(!require("caTools"))
{
    install.packages("caTools",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: caTools

## Warning: package 'caTools' was built under R version 4.0.5
```

```r
if(!require("tidyverse"))
{
    install.packages("tidyverse",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: tidyverse

## Warning: package 'tidyverse' was built under R version 4.0.5

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.4

## Warning: package 'dplyr' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
if(!require("Hmisc"))
{
    install.packages("Hmisc",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: Hmisc

## Warning: package 'Hmisc' was built under R version 4.0.4

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
if(!require("gensvm"))
{
    install.packages("gensvm",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: gensvm
```

```
if(!require("randomForest"))
{
    install.packages("randomForest",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: randomForest

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
if(!require("glmnet"))
{
    install.packages("glmnet",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: glmnet

## Warning: package 'glmnet' was built under R version 4.0.4

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-1
```

```
if(!require("caret"))
{
    install.packages("caret",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: caret
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
if(!require("pROC"))
{
    install.packages("pROC",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: pROC
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
if(!require("corrplot"))
{
    install.packages("corrplot",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: corrplot
```

```
## Warning: package 'corrplot' was built under R version 4.0.5
```

```
## corrplot 0.88 loaded
```

```
if(!require("ROCR"))
{
    install.packages("ROCR",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: ROCR
```

```
if(!require("gbm"))
{
    install.packages("gbm",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: gbm
```

```
## Warning: package 'gbm' was built under R version 4.0.4
```

```
## Loaded gbm 2.1.8
```

```
if(!require("readxl"))
{
    install.packages("readxl",repos = "http://cran.us.r-project.org")
}
```

```
## Loading required package: readxl
```

```
## Warning: package 'readxl' was built under R version 4.0.4
```

```
#load the libraries
options(warn=-1)
library(ggplot2)
library(caTools)
library(tidyverse)
library(Hmisc)
library(gensvm)
library(randomForest)
library(glmnet)
library(caret)
library(pROC)
library(corrplot)
library(ROCR)
library(gbm)
#download the dataset
#download.file("https://drive.google.com/u/0/uc?id=1mnjmZmXp_ej1G4k7rj-cKA7tGKqiB5cc&export=download","
df <- readxl::read_excel('dataset.xlsx')
df = df %>% distinct()
df$ClaimStatus = factor(df$ClaimStatus, levels = c(0, 1)) #convert the target variable to the encoded v

total_cells <- prod(dim(df)) #check total number of cells
missing_vals <- sum(is.na(df)) #check the misisng values
percent_of_missing_data <- (missing_vals/total_cells)*100 #check percenatge of missing vals
colSums(is.na(df))
```

```
##                  Sr No            ClaimStatus        ClaimFrequency
##                     0                      0                     0
##                Premium          Billing_Term               Renewed
##                     0                      0                     0
##                   DOB1                   DOB2                  DOB3
##                     0                   8625                 13268
```

```
##                  DOB4                  DOB5       Number_of_Driver
##                 13992                 14135                     0
##         AgeUSdriving_1        AgeUSdriving_2        AgeUSdriving_3
##                     0                     0                     0
##         AgeUSdriving_4        AgeUSdriving_5             Amendment
##                     0                     0                     0
##      CoverageLiability            CoverageMP          CoveragePD_1
##                     0                    57                    52
##        CoveragePIP_CDW          CoverageUMBI          CoverageUMPD
##                    50                     6                     6
##       DistanceToWork_1      DistanceToWork_2      DistanceToWork_3
##                     0                     0                     0
##       DistanceToWork_4      DistanceToWork_5      DriverAssigned_1
##                     0                     0                     0
##              Engine_1 ExcludedDriverName_01 ExcludedDriverName_02
##                   941                  2817                  6747
##  ExcludedDriverName_03 ExcludedDriverName_04 ExcludedDriverName_05
##                  8950                 10686                 11850
##  ExcludedDriverName_06 ExcludedDriverName_07 ExcludedDriverName_08
##                 12639                 13194                 13542
##  ExcludedDriverName_09 ExcludedDriverName_10 ExcludedDriverName_11
##                 13789                 13927                 14123
##  ExcludedDriverName_12 ExcludedDriverName_13 ExcludedDriverName_14
##                 14128                 14131                 14138
##  ExcludedDriverName_15 ExcludedDriverName_16 ExcludedDriverName_17
##                 14143                 14145                 14146
##  ExcludedDriverName_18 ExcludedDriverName_19 ExcludedDriverName_20
##                 14149                 14151                 14152
##          GaragedZIP_1       MaritalStatus_1       MaritalStatus_2
##                     0                     0                  9097
##       MaritalStatus_3       MaritalStatus_4       MaritalStatus_5
##                 13423                 14042                 14159
##          Occupation_1          Occupation_2          Occupation_3
##                  5851                 11572                 13882
##          Occupation_4          Occupation_5            Relation_1
##                 14136                 14171                     0
##            Relation_2            Relation_3            Relation_4
##                  9137                 13430                 14043
##            Relation_5              Rental_1                 Sex_1
##                 14159                     0                     0
##                 Sex_2                 Sex_3                 Sex_4
##                  9097                 13423                 14042
##                 Sex_5      Surcharge1Unit_1      Surcharge2Unit_1
##                 14159                   897                   899
##       Surcharge3Unit_1             Towing_1                 Units
##                   897                     0                     0
##     VehicleInspected_1   ViolPoints1Driver_1   ViolPoints1Driver_2
##                     0                     0                     0
##    ViolPoints1Driver_3   ViolPoints1Driver_4   ViolPoints1Driver_5
##                     0                     0                     0
##    ViolPoints2Driver_1   ViolPoints2Driver_2   ViolPoints2Driver_3
##                     0                     0                     0
##    ViolPoints2Driver_4   ViolPoints2Driver_5   ViolPoints3Driver_1
##                     0                     0                     0
```

```
##    ViolPoints3Driver_2    ViolPoints3Driver_3    ViolPoints3Driver_4
##                      0                      0                      0
##    ViolPoints3Driver_5    ViolPoints4Driver_1    ViolPoints4Driver_2
##                      0                      0                      0
##    ViolPoints4Driver_3    ViolPoints4Driver_4    ViolPoints4Driver_5
##                      0                      0                      0
##    ViolPoints5Driver_1    ViolPoints5Driver_2    ViolPoints5Driver_3
##                      0                      0                      0
##    ViolPoints5Driver_4    ViolPoints5Driver_5    ViolPoints6Driver_1
##                      0                      0                      0
##    ViolPoints6Driver_2    ViolPoints6Driver_3    ViolPoints6Driver_4
##                      0                      0                      0
##    ViolPoints6Driver_5    ViolPoints7Driver_1    ViolPoints7Driver_2
##                      0                      0                      0
##    ViolPoints7Driver_3    ViolPoints7Driver_4    ViolPoints7Driver_5
##                      0                      0                      0
##    ViolPoints8Driver_1    ViolPoints8Driver_2    ViolPoints8Driver_3
##                      0                      0                      0
##    ViolPoints8Driver_4    ViolPoints8Driver_5                Year_1
##                      0                      0                      0
##                 Make_1                Model_1                    Zip
##                     76                    138                      0
## Total_Distance_To_Work           NoLossSigned                   Type
##                      0                      0                      0
##       CancellationType
##                  13857
```

```
describe(df) #checking basic stats of the data
```

```
## df
##
##  127  Variables      14177  Observations
## ---------------------------------------------------------------------------
## Sr No
##        n  missing distinct
##    14177        0    14177
##
## lowest : P1     P10    P100   P1000  P10000, highest: P9995  P9996  P9997  P9998  P9999
## ---------------------------------------------------------------------------
## ClaimStatus
##        n  missing distinct
##    14177        0        2
##
## Value          0     1
## Frequency  13399   778
## Proportion 0.945 0.055
## ---------------------------------------------------------------------------
## ClaimFrequency
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        6    0.156  0.07406   0.1417
##
## lowest : 0 1 2 3 4, highest: 1 2 3 4 5
##
## Value          0     1     2     3     4     5
```

```
## Frequency  13399   580   144    36    16     2
## Proportion 0.945 0.041 0.010 0.003 0.001 0.000
## ---------------------------------------------------------------------------
## Premium
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0     1689        1    191.2    224.7       32       34
##      .25      .50      .75      .90      .95
##       40       71      239      556      718
##
## lowest :    0.00    8.00    8.87    9.06    9.87
## highest: 2064.00 2085.00 2095.00 2357.00 2869.00
## ---------------------------------------------------------------------------
## Billing_Term
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        3    0.774    3.527    2.497
##
## Value          1     3     6
## Frequency   6724   482  6971
## Proportion 0.474 0.034 0.492
## ---------------------------------------------------------------------------
## Renewed
##        n  missing distinct     Info      Sum     Mean      Gmd
##    14177        0        2     0.75     6970   0.4916   0.4999
##
## ---------------------------------------------------------------------------
## DOB1
##          n    missing   distinct       Info       Mean        Gmd        .05
##      14177          0       8360          1 1969-09-14   4.33e+08 1947-08-15
##        .10        .25        .50        .75        .90        .95
## 1952-10-14 1961-08-08 1971-02-24 1979-05-21 1984-06-10 1986-09-20
##
## lowest : 1931-03-09 1931-03-10 1931-03-19 1931-04-20 1931-06-30
## highest: 1992-05-04 1992-05-21 1992-10-13 1992-12-03 1993-05-28
## ---------------------------------------------------------------------------
## DOB2
##          n    missing   distinct       Info       Mean        Gmd        .05
##       5552       8625       4142          1 1970-06-02  418557875 1948-11-02
##        .10        .25        .50        .75        .90        .95
## 1953-06-21 1962-08-05 1971-11-30 1979-08-27 1984-08-01 1986-11-16
##
## lowest : 1931-01-30 1931-03-03 1931-03-10 1931-03-19 1931-11-25
## highest: 1992-07-12 1992-08-19 1992-10-28 1993-01-25 1993-03-07
## ---------------------------------------------------------------------------
## DOB3
##          n    missing   distinct       Info       Mean        Gmd        .05
##        909      13268        836          1 1971-12-29  493680788 1946-09-14
##        .10        .25        .50        .75        .90        .95
## 1952-02-09 1961-09-24 1974-10-06 1983-10-09 1987-10-22 1989-05-14
##
## lowest : 1931-12-09 1932-12-06 1932-12-24 1934-01-03 1934-02-12
## highest: 1993-02-15 1993-05-29 1993-10-01 1993-10-26 1994-04-06
## ---------------------------------------------------------------------------
## DOB4
##        n  missing distinct
```

```
##      185    13992      165
##
## lowest : 1/0/1900 13225    16677    16975    17112
## highest: 33099    33212    33367    33931    34046
## ---------------------------------------------------------------------------
## DOB5
##        n  missing distinct
##       42    14135       34
##
## lowest : 1/0/1900 16875    16880    17040    17210
## highest: 29967    30173    31705    32050    32224
## ---------------------------------------------------------------------------
## Number_of_Driver
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        5     0.74    1.472    0.628
##
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
##
## Value          1     2     3     4     5
## Frequency   8624  4636   733   150    34
## Proportion 0.608 0.327 0.052 0.011 0.002
## ---------------------------------------------------------------------------
## AgeUSdriving_1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       63    0.999    37.97    13.58       21       23
##      .25      .50      .75      .90      .95
##       28       36       46       55       60
##
## lowest : 17 18 19 20 21, highest: 75 76 77 78 79
## ---------------------------------------------------------------------------
## AgeUSdriving_2
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       65    0.737    13.48    18.96        0        0
##      .25      .50      .75      .90      .95
##        0        0       30       44       51
##
## lowest :  0 16 17 18 19, highest: 75 76 78 82 83
## ---------------------------------------------------------------------------
## AgeUSdriving_3
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       63    0.152    1.961    3.757        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        0       20
##
## lowest :  0 16 17 18 19, highest: 75 76 80 83 89
## ---------------------------------------------------------------------------
## AgeUSdriving_4
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       47    0.029   0.3384   0.6717        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        0        0
##
## lowest :  0 17 18 19 20, highest: 60 62 63 64 89
## ---------------------------------------------------------------------------
```

```
## AgeUSdriving_5
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       17    0.004  0.05149   0.1029        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        0        0
##
## lowest :  0 21 22 24 25, highest: 53 57 58 60 62
##
## Value          0    21    22    24    25    26    30    33    39    42    48
## Frequency  14159     1     2     1     1     1     1     1     1     1     1
## Proportion 0.999 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
##
## Value         51    53    57    58    60    62
## Frequency      1     1     2     1     1     1
## Proportion 0.000 0.000 0.000 0.000 0.000 0.000
## ---------------------------------------------------------------------------
## Amendment
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       10    0.113  0.06391   0.1243        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        0        0
##
## lowest :  0  1  2  3  4, highest:  5  6  7  8 10
##
## Value          0     1     2     3     4     5     6     7     8    10
## Frequency  13622   346   135    42    14    11     3     1     1     2
## Proportion 0.961 0.024 0.010 0.003 0.001 0.001 0.000 0.000 0.000 0.000
## ---------------------------------------------------------------------------
## CoverageLiability
##        n  missing distinct
##    14177        0        4
##
## Value      20/40/15 25/50/25 30/60/25     None
## Frequency      7061     7077       38        1
## Proportion    0.498    0.499    0.003    0.000
## ---------------------------------------------------------------------------
## CoverageMP
##        n  missing distinct
##    14120       57        2
##
## Value        535  None
## Frequency      1 14119
## Proportion     0     1
## ---------------------------------------------------------------------------
## CoveragePD_1
##        n  missing distinct
##    14125       52        3
##
## Value      1000/1000   500/500      None
## Frequency          1      3094     11030
## Proportion     0.000     0.219     0.781
## ---------------------------------------------------------------------------
## CoveragePIP_CDW
##        n  missing distinct
```

```
##    14127       50        3
##
## Value       2535  2569  None
## Frequency      1    90 14036
## Proportion 0.000 0.006 0.994
## ---------------------------------------------------------------------------
## CoverageUMBI
##        n  missing distinct
##    14171        6        2
##
## Value      Accepted     None
## Frequency       316    13855
## Proportion    0.022    0.978
## ---------------------------------------------------------------------------
## CoverageUMPD
##        n  missing distinct
##    14171        6        2
##
## Value      Accepted     None
## Frequency       316    13855
## Proportion    0.022    0.978
## ---------------------------------------------------------------------------
## DistanceToWork_1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       64    0.852    5.442    4.514        0        1
##      .25      .50      .75      .90      .95
##        3        5        5       10       15
##
## lowest :   0   1   2   3   4, highest: 180 200 260 480 500
## ---------------------------------------------------------------------------
## DistanceToWork_2
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       46    0.632    1.636    2.674        0        0
##      .25      .50      .75      .90      .95
##        0        0        1        5        5
##
## lowest :   0   1   2   3   4, highest:  65  70 100 260 425
## ---------------------------------------------------------------------------
## DistanceToWork_3
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       24    0.116   0.2367    0.461        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        0        0
##
## lowest :   0   1   2   3   4, highest:  45  50  60  70 176
## ---------------------------------------------------------------------------
## DistanceToWork_4
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       13     0.02  0.03358  0.06681        0        0
##      .25      .50      .75      .90      .95
##        0        0        0        0        0
##
## lowest :  0  1  2  3  4, highest:  8 10 12 16 20
##
```

```
## Value          0     1     2     3     4     5     6     7     8    10    12
## Frequency  14083     8     7     1     1    66     1     1     2     4     1
## Proportion 0.993 0.001 0.000 0.000 0.000 0.005 0.000 0.000 0.000 0.000 0.000
##
## Value         16    20
## Frequency      1     1
## Proportion 0.000 0.000
## ---------------------------------------------------------------------------
## DistanceToWork_5
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        3    0.003 0.004303 0.008599
##
## Value          0     1     5
## Frequency  14164     1    12
## Proportion 0.999 0.000 0.001
## ---------------------------------------------------------------------------
## DriverAssigned_1
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        5    0.314    1.141   0.2535
##
## lowest : 1 2 3 4 5, highest: 1 2 3 4 5
##
## Value          1     2     3     4     5
## Frequency  12495  1400   248    30     4
## Proportion 0.881 0.099 0.017 0.002 0.000
## ---------------------------------------------------------------------------
## Engine_1
##        n  missing distinct
##    13236      941      221
##
## lowest : /    1    1.3  1.5  1.6 , highest: 91/4 97   97/4 98   98/4
## ---------------------------------------------------------------------------
## ExcludedDriverName_01
##        n  missing distinct
##    11360     2817    10718
##
## lowest : .                  A,FIDELRUIZ        AaronAlexanderHolg AaronBallina       AARONMELISSAPAT
## highest: ZULEMAVILLALOBOS   ZulemeGonzalez     ZUNIGA,IDEANA      ZUNIGA,JUAN        ZUNIGA,LISANDRA
## ---------------------------------------------------------------------------
## ExcludedDriverName_02
##        n  missing distinct
##     7430     6747     7123
##
## lowest : AARONDAVIDLOZANO    AARONLARA           AaronMedinaMurillo  ABBYSEGURA          ABELARDOCAS
## highest: ZUNIGA,JOSEALEJANDR ZUNIGA,MARIALUISA   ZUNIGA,RENE         ZUNIGA,SUSANA       ZUREYAALFAR
## ---------------------------------------------------------------------------
## ExcludedDriverName_03
##        n  missing distinct
##     5227     8950     5054
##
## lowest : AARONJAMESHOPKINS   AaronJrMurillo      AARONLARA           AaronMartinez       ABELARDOFLO
## highest: ZEPEDA,LORENZOAMBRO ZoilaMartinez-Grand ZulemaCastilloTrev  ZulemaZapataAguero  ZUNIGA,GUAD
## ---------------------------------------------------------------------------
## ExcludedDriverName_04
```

```
##        n  missing distinct
##     3491    10686     3387
##
## lowest : 0CUELLAR,LORENZOMAN AbelUribe           ABIGAILFLORES       AbigailParedes      ABRAHAMDIAZ
## highest: ZochieFernandez     ZoraidaSanchez      ZorinaMercado       ZULEMAJUAREZ        ZUNIGA,HECT
## ------------------------------------------------------------------------------
## ExcludedDriverName_05
##        n  missing distinct
##     2327    11850     2256
##
## lowest : (PrevOwner)        AbelJoelSalazarJr  AbelLermaJr        ABREGO,MARIA       ACOSTA,EVANGELI
## highest: ZETINA,MARIATERESA ZeusOCourtois      ZORINAMERCADO      ZOROLA,MARIAA      ZulemaLopez
## ------------------------------------------------------------------------------
## ExcludedDriverName_06
##        n  missing distinct
##     1538    12639     1494
##
## lowest : AARONJAMESHOPKINS   ABBYZAMARRON        AbrahmguerraMartin  ACOSTA,JOSEMANUEL   AdolfoNieto
## highest: ZAVALA,MIRNAELIZABE ZAYRAGUERRA         ZEPEDASARA          ZILM,MARIAANA       ZOROLA,ERIC
## ------------------------------------------------------------------------------
## ExcludedDriverName_07
##        n  missing distinct
##      983    13194      956
##
## lowest : 09071973TREVINO,JORG AbelardoGuerraMart   ABREGO,JESUS         ACEVEDO-FRESNILLO,VI AdolfoC
## highest: YvonneMFlores        ZACARIASGONZALEZ     ZAMORA,LETICIAADRIA  ZoraidaGLara         ZUNIGAM
## ------------------------------------------------------------------------------
## ExcludedDriverName_08
##        n  missing distinct
##      635    13542      619
##
## lowest : ACEVEDO,LINDAE       ADANDELAOLA          AdnrewRogleioSalin   ADRIANACAMPOS        ADRIANA
## highest: WOMER,ALICIAMARYAN   YvonneRodriguez      ZAPATA,OLIVIAMENDOZ  ZOROLA-GRACIANO,LUCI ZUNIGAR
## ------------------------------------------------------------------------------
## ExcludedDriverName_09
##        n  missing distinct
##      388    13789      379
##
## lowest : AbrahamMendoza       ACEVEDO-FRESNILLO,VI AGUIRRE,HECTOR       AlbertoEnriqueMadr   Alberto
## highest: YANEZ,VIRGINIAPENA   YOLANDACORTEZ        YolandaRodriguez     ZOROLA-GRACIANO,LUCI ZULEMAC
## ------------------------------------------------------------------------------
## ExcludedDriverName_10
##        n  missing distinct
##      250    13927      237
##
## lowest : **AdditionalExclusi AdanFierrosDenova   AdditionalExclusion ADDITIONALEXCLUSION AGUILAR,JOS
## highest: YAHAIRACOLMENERO    YolandTorresAguila  ZAPATA,GUADALUPEGAR ZAPATA,SANJUANAAVIL ZAVALA,ROSA
## ------------------------------------------------------------------------------
## ExcludedDriverName_11
##        n  missing distinct
##       54    14123       49
##
## lowest : AdameE.Ardner      AdamsStephen       AmaliaTrevinoMckee AngeloLuisMercado  AnselmoBriagas
## highest: VirginiaNavarroMor WensesladaFloresJi XavierPerez        YiWang             YvetteMarie
```

```
## ---------------------------------------------------------------------------
## ExcludedDriverName_12
##        n  missing distinct
##       49    14128       44
##
## lowest : AmandaMendezHenry  AndreaJoyBedford   AntonioEsequielVas BeasleyRoger       BERNALCESAREDUAL
## highest: SeanOGrayP-Owner   SylviaRuiz         VirginiaGarza      WrightKennethRayJ  ZacharyArnoldoGa
## ---------------------------------------------------------------------------
## ExcludedDriverName_13
##        n  missing distinct
##       46    14131       42
##
## lowest : AlbertDelarosa     ArnoldoGarza       BlancaAliciaRodela CantuJose          CarlosChaves
## highest: SeverianoGuevara   SharrenaNicoleCash TarnohTwaylee      TomasGilbertoMoral Vidal,SabrinaSt
## ---------------------------------------------------------------------------
## ExcludedDriverName_14
##        n  missing distinct
##       39    14138       35
##
## lowest : AlfredoHernandez-ca AlmaDeliaEsquivel   AnitaFloresPerez    BeatriceDiazPerez   BlaireAEstra
## highest: Scott,Wayne-PrevO   SylviaTheodoraNara  VernealMarieAdams   WendyLeeBillegas    YracemaSalin
## ---------------------------------------------------------------------------
## ExcludedDriverName_15
##        n  missing distinct
##       34    14143       30
##
## lowest : AmadoPerez         AustinAcevesLimon  BobbieJoyceAdams   CarolynLucilleCald CoryPineock-Prev
## highest: SadieHankinsBullio SamuelRiojas       SergioTDelacruz    TommyLeal          TravisWadeCornel
## ---------------------------------------------------------------------------
## ExcludedDriverName_16
##        n  missing distinct
##       32    14145       29
##
## lowest : AgustinEduardoLimo AsmatAraDurrani    BonnieJeanCorneliu BrookeNicolePerez  CiprianoPenaCar
## highest: RoseldaCastilloJim RoseMarieMartinez  TimmyLeal          VeronicaNunezCast  VirginiaRocha
## ---------------------------------------------------------------------------
## ExcludedDriverName_17
##        n  missing distinct
##       31    14146       29
##
## lowest : AmaliaTrevinoMckee AnabelLucio        AnnetteSplattYance AnthonyMoore       AugustineDMartin
## highest: RosieVidalesEspino RualJuanLimon      SamuelGandara-Son  SoniaGarciaLedezma VeronicaTaffola
## ---------------------------------------------------------------------------
## ExcludedDriverName_18
##        n  missing distinct
##       28    14149       26
##
## lowest : AlbertSalazar     CarolMarieSimmons DanielJacobAlkire DonaldJackJones   EmeteriaCerdaLeal
## highest: PascualLopez      RaulSaucedaJr     REBECCACANTUDOBUK SandraLimon       ZeferinoRodriguez
## ---------------------------------------------------------------------------
## ExcludedDriverName_19
##        n  missing distinct
##       26    14151       24
##
```

```
## lowest : AdamChristopherAhl AndresMartinezNola AntonioMejiaTorres AracelyJuarez      BrendaSRivera
## highest: MelissaHunter      MonicaEvetteSegovi ReynolBernal       RosaAdrianaDeleon  VeronicaTaffolal
## ---------------------------------------------------------------------------
## ExcludedDriverName_20
##        n  missing distinct
##       25    14152       23
##
## lowest : AneySolisChavez    AnnJonesMeek       ArkealiesBDuncan-P BeverlyAnnHeileman BrendaSRivera
## highest: PoncianoSanchezSan RahacelSueVega     SethWayneCrider    SimonaOlguinPadill SylviaGonzalesA
## ---------------------------------------------------------------------------
## GaragedZIP_1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0      167    0.983    78107    278.3    77642    78040
##      .25      .50      .75      .90      .95
##    78041    78046    78119    78503    78801
##
## lowest : 75009 75023 75028 75034 75035, highest: 78934 79601 79603 79605 79713
## ---------------------------------------------------------------------------
## MaritalStatus_1
##        n  missing distinct
##    14177        0        2
##
## Value          M     S
## Frequency  11255  2922
## Proportion 0.794 0.206
## ---------------------------------------------------------------------------
## MaritalStatus_2
##        n  missing distinct
##     5080     9097        2
##
## Value          M     S
## Frequency   4588   492
## Proportion 0.903 0.097
## ---------------------------------------------------------------------------
## MaritalStatus_3
##        n  missing distinct
##      754    13423        2
##
## Value          M     S
## Frequency    583   171
## Proportion 0.773 0.227
## ---------------------------------------------------------------------------
## MaritalStatus_4
##        n  missing distinct
##      135    14042        2
##
## Value          M     S
## Frequency    110    25
## Proportion 0.815 0.185
## ---------------------------------------------------------------------------
## MaritalStatus_5
##        n  missing distinct
##       18    14159        2
##
```

```
## Value          M     S
## Frequency     15     3
## Proportion 0.833 0.167
## ---------------------------------------------------------------------------
## Occupation_1
##        n  missing distinct
##     8326     5851     2466
##
## lowest : A.C.Technitia  A/C            A/CTech        A/CTECH        Abogado
## highest: YardMaintenanc YARDMANAGER    YARDS          YardWork       YolisTacoPlac
## ---------------------------------------------------------------------------
## Occupation_2
##        n  missing distinct
##     2605    11572      881
##
## lowest : A/CTech        AC/TECH        ACCAUNTING     Accountant     ACCOUNTANT
## highest: WindowInstalle WIRELINEOPER   WLEDER         Worker         XRAYTECH
## ---------------------------------------------------------------------------
## Occupation_3
##        n  missing distinct
##      295    13882      148
##
## lowest : ACCT           ACCTSPAYABLE   ACTech         ASSMANAGER     ASST.PRINCIPAL
## highest: WAREHOUSECLERK WELDER         WELLSMACHINE   WELTA          WORKSTUDENT
## ---------------------------------------------------------------------------
## Occupation_4
##        n  missing distinct
##       41    14136       30
##
## lowest : Aircraft     Clerk        CNA          CONSTRUCTION COOK
## highest: TrukDriver   UNEMPLOYED   WAITRESS     Welder       WELLS
## ---------------------------------------------------------------------------
## Occupation_5
##        n  missing distinct
##        6    14171        5
##
## lowest : H.W.         Self         SELFEMPLOYED STUDENT      TEACHER
## highest: H.W.         Self         SELFEMPLOYED STUDENT      TEACHER
##
## Value              H.W.         Self SELFEMPLOYED      STUDENT      TEACHER
## Frequency             2            1            1            1            1
## Proportion        0.333        0.167        0.167        0.167        0.167
## ---------------------------------------------------------------------------
## Relation_1
##        n  missing distinct    value
##    14177        0        1     Self
##
## Value       Self
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## Relation_2
##        n  missing distinct
##     5040     9137      226
```

```
##
## lowest : 2ndinsure  amigo      AMIGO      AQUANTANCE aunt
## highest: uncle      UNCLE      wife       Wife       WIFE
## ---------------------------------------------------------------------------
## Relation_3
##        n  missing distinct
##      747    13430      135
##
## lowest : AMANDA aunt   AUNT   BIL    boss  , highest: UNCEL  uncle  UNCLE  wife   WIFE
## ---------------------------------------------------------------------------
## Relation_4
##        n  missing distinct
##      134    14043       54
##
## lowest : AUNT      boyfriend BR/S      brother   Brother
## highest: SONINLAW  Spouse    SPOUSE    wife      WIFE
## ---------------------------------------------------------------------------
## Relation_5
##        n  missing distinct
##       18    14159       15
##
## lowest : cousin   daghter  DAINLAW  daughter DAUGHTER
## highest: INLAW    MOTHER   SON      spouse   Wife
##
## cousin (1, 0.056), daghter (1, 0.056), DAINLAW (1, 0.056), daughter (1, 0.056),
## DAUGHTER (1, 0.056), DAUGTHER (1, 0.056), FATHER (1, 0.056), FRIEND (1, 0.056),
## Friends (1, 0.056), inlaw (1, 0.056), INLAW (2, 0.111), MOTHER (2, 0.111), SON
## (2, 0.111), spouse (1, 0.056), Wife (1, 0.056)
## ---------------------------------------------------------------------------
## Rental_1
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        5    0.069   0.1083   0.2149
##
## lowest :  0  1 20 25 35, highest:  0  1 20 25 35
##
## Value          0     1    20    25    35
## Frequency  13841   285     7    43     1
## Proportion 0.976 0.020 0.000 0.003 0.000
## ---------------------------------------------------------------------------
## Sex_1
##        n  missing distinct
##    14177        0        2
##
## Value          F     M
## Frequency   5783  8394
## Proportion 0.408 0.592
## ---------------------------------------------------------------------------
## Sex_2
##        n  missing distinct
##     5080     9097        2
##
## Value          F     M
## Frequency   2533  2547
## Proportion 0.499 0.501
```

```
## ---------------------------------------------------------------------------
## Sex_3
##        n  missing distinct
##      754    13423        2
##
## Value         F    M
## Frequency   362  392
## Proportion 0.48 0.52
## ---------------------------------------------------------------------------
## Sex_4
##        n  missing distinct
##      135    14042        2
##
## Value          F     M
## Frequency     70    65
## Proportion 0.519 0.481
## ---------------------------------------------------------------------------
## Sex_5
##        n  missing distinct
##       18    14159        2
##
## Value          F     M
## Frequency     10     8
## Proportion 0.556 0.444
## ---------------------------------------------------------------------------
## Surcharge1Unit_1
##        n  missing distinct
##    13280      897        2
##
## Value          N     Y
## Frequency  13258    22
## Proportion 0.998 0.002
## ---------------------------------------------------------------------------
## Surcharge2Unit_1
##        n  missing distinct
##    13278      899        2
##
## Value          N     Y
## Frequency  13239    39
## Proportion 0.997 0.003
## ---------------------------------------------------------------------------
## Surcharge3Unit_1
##        n  missing distinct
##    13280      897        2
##
## Value          N     Y
## Frequency   9131  4149
## Proportion 0.688 0.312
## ---------------------------------------------------------------------------
## Towing_1
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        4    0.062   0.2401    0.478
##
## Value          0     1    50    70
```

```
## Frequency  13876   254     7    40
## Proportion 0.979 0.018 0.000 0.003
## ---------------------------------------------------------------------------
## Units
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        5     0.47     1.24   0.4199
##
## lowest : 0 1 2 3 4, highest: 0 1 2 3 4
##
## Value          0     1     2     3     4
## Frequency     57 11449  2035   490   146
## Proportion 0.004 0.808 0.144 0.035 0.010
## ---------------------------------------------------------------------------
## VehicleInspected_1
##        n  missing distinct     Info      Sum     Mean      Gmd
##    14177        0        2      0.4     2244   0.1583   0.2665
##
## ---------------------------------------------------------------------------
## ViolPoints1Driver_1
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        5    0.219   0.2047    0.382
##
## lowest : 0 1 2 3 5, highest: 0 1 2 3 5
##
## Value          0     1     2     3     5
## Frequency  13057   249     8   839    24
## Proportion 0.921 0.018 0.001 0.059 0.002
## ---------------------------------------------------------------------------
## ViolPoints1Driver_2
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        4     0.09  0.09085   0.1762
##
## Value          0     1     3     5
## Frequency  13737    17   422     1
## Proportion 0.969 0.001 0.030 0.000
## ---------------------------------------------------------------------------
## ViolPoints1Driver_3
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        2    0.009 0.009522  0.01899
##
## Value          0     3
## Frequency  14132    45
## Proportion 0.997 0.003
## ---------------------------------------------------------------------------
## ViolPoints1Driver_4
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        2    0.003 0.002539 0.005075
##
## Value          0     3
## Frequency  14165    12
## Proportion 0.999 0.001
## ---------------------------------------------------------------------------
## ViolPoints1Driver_5
##         n   missing  distinct      Info      Mean       Gmd
```

```
##     14177         0         2         0 0.0002116 0.0004232
##
## Value          0     3
## Frequency  14176     1
## Proportion     1     0
## ------------------------------------------------------------------------------
## ViolPoints2Driver_1
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        5    0.004 0.002751 0.005497
##
## lowest : 0 1 2 3 5, highest: 0 1 2 3 5
##
## Value          0     1     2     3     5
## Frequency  14159     8     1     8     1
## Proportion 0.999 0.001 0.000 0.001 0.000
## ------------------------------------------------------------------------------
## ViolPoints2Driver_2
##         n   missing  distinct      Info      Mean       Gmd
##     14177         0         3     0.001 0.0007054   0.00141
##
## Value          0     1     3
## Frequency  14171     4     2
## Proportion     1     0     0
## ------------------------------------------------------------------------------
## ViolPoints2Driver_3
##         n   missing  distinct      Info      Mean       Gmd
##     14177         0         2         0 0.0002116 0.0004232
##
## Value          0     3
## Frequency  14176     1
## Proportion     1     0
## ------------------------------------------------------------------------------
## ViolPoints2Driver_4
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ------------------------------------------------------------------------------
## ViolPoints2Driver_5
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ------------------------------------------------------------------------------
## ViolPoints3Driver_1
##         n   missing  distinct      Info      Mean       Gmd
##     14177         0         3         0 0.0002821 0.0005643
##
## Value          0     1     3
## Frequency  14175     1     1
```

```
## Proportion     1     0     0
## ---------------------------------------------------------------------------
## ViolPoints3Driver_2
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints3Driver_3
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints3Driver_4
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints3Driver_5
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints4Driver_1
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints4Driver_2
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints4Driver_3
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
```

```
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints4Driver_4
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints4Driver_5
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints5Driver_1
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints5Driver_2
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints5Driver_3
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints5Driver_4
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints5Driver_5
##        n  missing distinct     Info     Mean      Gmd
```

```
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints6Driver_1
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints6Driver_2
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints6Driver_3
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints6Driver_4
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints6Driver_5
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints7Driver_1
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
```

```
## ViolPoints7Driver_2
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ------------------------------------------------------------------------------
## ViolPoints7Driver_3
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ------------------------------------------------------------------------------
## ViolPoints7Driver_4
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ------------------------------------------------------------------------------
## ViolPoints7Driver_5
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ------------------------------------------------------------------------------
## ViolPoints8Driver_1
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ------------------------------------------------------------------------------
## ViolPoints8Driver_2
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ------------------------------------------------------------------------------
## ViolPoints8Driver_3
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
```

```
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints8Driver_4
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## ViolPoints8Driver_5
##        n  missing distinct     Info     Mean      Gmd
##    14177        0        1        0        0        0
##
## Value          0
## Frequency  14177
## Proportion     1
## ---------------------------------------------------------------------------
## Year_1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       52    0.996     1996    11.99     1989     1992
##      .25      .50      .75      .90      .95
##     1996     2000     2003     2005     2007
##
## lowest :   -3    0  199  988 1957, highest: 2007 2008 2009 2010 2011
##
## Value         -5     0   200   990  1955  1965  1970  1975  1980  1985  1990
## Frequency      1    19     1     1     1     3    21    31   117   333  1192
## Proportion 0.000 0.001 0.000 0.000 0.000 0.000 0.001 0.002 0.008 0.023 0.084
##
## Value       1995  2000  2005  2010
## Frequency   2997  5837  3260   363
## Proportion 0.211 0.412 0.230 0.026
##
## For the frequency table, variable is rounded to the nearest 5
## ---------------------------------------------------------------------------
## Make_1
##        n  missing distinct
##    14101       76      134
##
## lowest : Acura      ACURA      AMG        AUDI       Bmw
## highest: VOLKSWAGEN Volvo      VOLVO      VW         Wrangler
## ---------------------------------------------------------------------------
## Model_1
##        n  missing distinct
##    14039      138     1448
##
## lowest : 150              1500             1g1nd52t3x612814 2.3CL            20
## highest: YUKONXL1500SLE   YUKONXLDENALI    Z3               ZEPHYR           ZX2
## ---------------------------------------------------------------------------
## Zip
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0      139    0.983    78107      278    77642    78040
##      .25      .50      .75      .90      .95
```

```
##    78041    78045    78119    78503    78801
##
## lowest : 75009 75023 75028 75034 75035, highest: 78852 78934 79601 79603 79605
## ---------------------------------------------------------------------------
## Total_Distance_To_Work
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##    14177        0       85    0.913    7.353    6.172        1        2
##      .25      .50      .75      .90      .95
##        5        5       10       12       20
##
## lowest :   1   2   3   4   5, highest: 200 260 435 480 500
## ---------------------------------------------------------------------------
## NoLossSigned
##        n  missing distinct     Info      Sum     Mean      Gmd
##    14177        0        2    0.132      653  0.04606  0.08788
##
## ---------------------------------------------------------------------------
## Type
##        n  missing distinct
##    14177        0        9
##
## lowest : A   AP  DP  FC  P  , highest: P   REN RET VD  XFR
##
## Value          A    AP    DP    FC     P   REN   RET    VD   XFR
## Frequency    221    11 11130     4  1886   675     6     2   242
## Proportion 0.016 0.001 0.785 0.000 0.133 0.048 0.000 0.000 0.017
## ---------------------------------------------------------------------------
## CancellationType
##        n  missing distinct
##      320    13857        2
##
## Value        INS    NP
## Frequency      3   317
## Proportion 0.009 0.991
## ---------------------------------------------------------------------------
```

```
summary(df) #checking basic facts
```

```
##     Sr No           ClaimStatus ClaimFrequency       Premium
##  Length:14177       0:13399     Min.   :0.00000   Min.   :   0.0
##  Class :character   1:  778     1st Qu.:0.00000   1st Qu.:  40.0
##  Mode  :character               Median :0.00000   Median :  71.0
##                                 Mean   :0.07406   Mean   : 191.2
##                                 3rd Qu.:0.00000   3rd Qu.: 239.0
##                                 Max.   :5.00000   Max.   :2869.0
##
##   Billing_Term      Renewed            DOB1
##  Min.   :1.000   Min.   :0.0000   Min.   :1931-03-09 00:00:00
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1961-08-08 00:00:00
##  Median :3.000   Median :0.0000   Median :1971-02-24 00:00:00
##  Mean   :3.527   Mean   :0.4916   Mean   :1969-09-13 16:54:31
##  3rd Qu.:6.000   3rd Qu.:1.0000   3rd Qu.:1979-05-21 00:00:00
##  Max.   :6.000   Max.   :1.0000   Max.   :1993-05-28 00:00:00
##
```

```
##       DOB2                          DOB3                          DOB4
##  Min.   :1931-01-30 00:00:00   Min.   :1931-12-09 00:00:00   Length:14177
##  1st Qu.:1962-08-05 00:00:00   1st Qu.:1961-09-24 00:00:00   Class :character
##  Median :1971-11-30 00:00:00   Median :1974-10-06 00:00:00   Mode  :character
##  Mean   :1970-06-01 21:54:43   Mean   :1971-12-28 16:28:30
##  3rd Qu.:1979-08-26 18:00:00   3rd Qu.:1983-10-09 00:00:00
##  Max.   :1993-03-07 00:00:00   Max.   :1994-04-06 00:00:00
##  NA's   :8625                  NA's   :13268
##      DOB5           Number_of_Driver AgeUSdriving_1  AgeUSdriving_2
##  Length:14177       Min.   :1.000    Min.   :17.00   Min.   : 0.00
##  Class :character   1st Qu.:1.000    1st Qu.:28.00   1st Qu.: 0.00
##  Mode  :character   Median :1.000    Median :36.00   Median : 0.00
##                     Mean   :1.472    Mean   :37.97   Mean   :13.48
##                     3rd Qu.:2.000    3rd Qu.:46.00   3rd Qu.:30.00
##                     Max.   :5.000    Max.   :79.00   Max.   :83.00
##
##  AgeUSdriving_3   AgeUSdriving_4    AgeUSdriving_5       Amendment
##  Min.   : 0.000   Min.   : 0.0000   Min.   : 0.00000   Min.   : 0.00000
##  1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.: 0.00000
##  Median : 0.000   Median : 0.0000   Median : 0.00000   Median : 0.00000
##  Mean   : 1.961   Mean   : 0.3384   Mean   : 0.05149   Mean   : 0.06391
##  3rd Qu.: 0.000   3rd Qu.: 0.0000   3rd Qu.: 0.00000   3rd Qu.: 0.00000
##  Max.   :89.000   Max.   :89.0000   Max.   :62.00000   Max.   :10.00000
##
##  CoverageLiability   CoverageMP        CoveragePD_1       CoveragePIP_CDW
##  Length:14177       Length:14177       Length:14177       Length:14177
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  CoverageUMBI       CoverageUMPD       DistanceToWork_1  DistanceToWork_2
##  Length:14177       Length:14177       Min.   :  0.000   Min.   :  0.000
##  Class :character   Class :character   1st Qu.:  3.000   1st Qu.:  0.000
##  Mode  :character   Mode  :character   Median :  5.000   Median :  0.000
##                                        Mean   :  5.442   Mean   :  1.637
##                                        3rd Qu.:  5.000   3rd Qu.:  1.000
##                                        Max.   :500.000   Max.   :425.000
##
##  DistanceToWork_3   DistanceToWork_4   DistanceToWork_5   DriverAssigned_1
##  Min.   :  0.0000   Min.   : 0.00000   Min.   :0.000000   Min.   :1.000
##  1st Qu.:  0.0000   1st Qu.: 0.00000   1st Qu.:0.000000   1st Qu.:1.000
##  Median :  0.0000   Median : 0.00000   Median :0.000000   Median :1.000
##  Mean   :  0.2366   Mean   : 0.03358   Mean   :0.004303   Mean   :1.141
##  3rd Qu.:  0.0000   3rd Qu.: 0.00000   3rd Qu.:0.000000   3rd Qu.:1.000
##  Max.   :176.0000   Max.   :20.00000   Max.   :5.000000   Max.   :5.000
##
##    Engine_1         ExcludedDriverName_01 ExcludedDriverName_02
##  Length:14177       Length:14177          Length:14177
##  Class :character   Class :character      Class :character
##  Mode  :character   Mode  :character      Mode  :character
##
##
```

```
##
##
##  ExcludedDriverName_03 ExcludedDriverName_04 ExcludedDriverName_05
##  Length:14177          Length:14177          Length:14177
##  Class :character      Class :character      Class :character
##  Mode  :character      Mode  :character      Mode  :character
##
##
##
##
##  ExcludedDriverName_06 ExcludedDriverName_07 ExcludedDriverName_08
##  Length:14177          Length:14177          Length:14177
##  Class :character      Class :character      Class :character
##  Mode  :character      Mode  :character      Mode  :character
##
##
##
##
##  ExcludedDriverName_09 ExcludedDriverName_10 ExcludedDriverName_11
##  Length:14177          Length:14177          Length:14177
##  Class :character      Class :character      Class :character
##  Mode  :character      Mode  :character      Mode  :character
##
##
##
##
##  ExcludedDriverName_12 ExcludedDriverName_13 ExcludedDriverName_14
##  Length:14177          Length:14177          Length:14177
##  Class :character      Class :character      Class :character
##  Mode  :character      Mode  :character      Mode  :character
##
##
##
##
##  ExcludedDriverName_15 ExcludedDriverName_16 ExcludedDriverName_17
##  Length:14177          Length:14177          Length:14177
##  Class :character      Class :character      Class :character
##  Mode  :character      Mode  :character      Mode  :character
##
##
##
##
##  ExcludedDriverName_18 ExcludedDriverName_19 ExcludedDriverName_20
##  Length:14177          Length:14177          Length:14177
##  Class :character      Class :character      Class :character
##  Mode  :character      Mode  :character      Mode  :character
##
##
##
##
##   GaragedZIP_1   MaritalStatus_1    MaritalStatus_2    MaritalStatus_3
##  Min.   :75009   Length:14177       Length:14177       Length:14177
##  1st Qu.:78041   Class :character   Class :character   Class :character
##  Median :78046   Mode  :character   Mode  :character   Mode  :character
```

```
##  Mean   :78107
##  3rd Qu.:78119
##  Max.   :79713
##
##  MaritalStatus_4    MaritalStatus_5    Occupation_1       Occupation_2
##  Length:14177       Length:14177       Length:14177       Length:14177
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Occupation_3       Occupation_4       Occupation_5        Relation_1
##  Length:14177       Length:14177       Length:14177       Length:14177
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Relation_2         Relation_3         Relation_4         Relation_5
##  Length:14177       Length:14177       Length:14177       Length:14177
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Rental_1          Sex_1              Sex_2              Sex_3
##  Min.   : 0.0000   Length:14177       Length:14177       Length:14177
##  1st Qu.: 0.0000   Class :character   Class :character   Class :character
##  Median : 0.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 0.1083
##  3rd Qu.: 0.0000
##  Max.   :35.0000
##
##     Sex_4              Sex_5           Surcharge1Unit_1   Surcharge2Unit_1
##  Length:14177       Length:14177       Length:14177       Length:14177
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Surcharge3Unit_1      Towing_1           Units      VehicleInspected_1
##  Length:14177       Min.   : 0.0000   Min.   :0.00   Min.   :0.0000
##  Class :character   1st Qu.: 0.0000   1st Qu.:1.00   1st Qu.:0.0000
##  Mode  :character   Median : 0.0000   Median :1.00   Median :0.0000
##                     Mean   : 0.2401   Mean   :1.24   Mean   :0.1583
##                     3rd Qu.: 0.0000   3rd Qu.:1.00   3rd Qu.:0.0000
##                     Max.   :70.0000   Max.   :4.00   Max.   :1.0000
##
##  ViolPoints1Driver_1 ViolPoints1Driver_2 ViolPoints1Driver_3
##  Min.   :0.0000      Min.   :0.00000     Min.   :0.000000
```

```
##  1st Qu.:0.0000      1st Qu.:0.00000     1st Qu.:0.000000
##  Median :0.0000      Median :0.00000     Median :0.000000
##  Mean   :0.2047      Mean   :0.09085     Mean   :0.009522
##  3rd Qu.:0.0000      3rd Qu.:0.00000     3rd Qu.:0.000000
##  Max.   :5.0000      Max.   :5.00000     Max.   :3.000000
##
##  ViolPoints1Driver_4 ViolPoints1Driver_5 ViolPoints2Driver_1
##  Min.   :0.000000    Min.   :0.0000000   Min.   :0.000000
##  1st Qu.:0.000000    1st Qu.:0.0000000   1st Qu.:0.000000
##  Median :0.000000    Median :0.0000000   Median :0.000000
##  Mean   :0.002539    Mean   :0.0002116   Mean   :0.002751
##  3rd Qu.:0.000000    3rd Qu.:0.0000000   3rd Qu.:0.000000
##  Max.   :3.000000    Max.   :3.0000000   Max.   :5.000000
##
##  ViolPoints2Driver_2 ViolPoints2Driver_3 ViolPoints2Driver_4
##  Min.   :0.0000000   Min.   :0.0000000   Min.   :0
##  1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0
##  Median :0.0000000   Median :0.0000000   Median :0
##  Mean   :0.0007054   Mean   :0.0002116   Mean   :0
##  3rd Qu.:0.0000000   3rd Qu.:0.0000000   3rd Qu.:0
##  Max.   :3.0000000   Max.   :3.0000000   Max.   :0
##
##  ViolPoints2Driver_5 ViolPoints3Driver_1 ViolPoints3Driver_2
##  Min.   :0           Min.   :0.0000000   Min.   :0
##  1st Qu.:0           1st Qu.:0.0000000   1st Qu.:0
##  Median :0           Median :0.0000000   Median :0
##  Mean   :0           Mean   :0.0002821   Mean   :0
##  3rd Qu.:0           3rd Qu.:0.0000000   3rd Qu.:0
##  Max.   :0           Max.   :3.0000000   Max.   :0
##
##  ViolPoints3Driver_3 ViolPoints3Driver_4 ViolPoints3Driver_5
##  Min.   :0           Min.   :0           Min.   :0
##  1st Qu.:0           1st Qu.:0           1st Qu.:0
##  Median :0           Median :0           Median :0
##  Mean   :0           Mean   :0           Mean   :0
##  3rd Qu.:0           3rd Qu.:0           3rd Qu.:0
##  Max.   :0           Max.   :0           Max.   :0
##
##  ViolPoints4Driver_1 ViolPoints4Driver_2 ViolPoints4Driver_3
##  Min.   :0           Min.   :0           Min.   :0
##  1st Qu.:0           1st Qu.:0           1st Qu.:0
##  Median :0           Median :0           Median :0
##  Mean   :0           Mean   :0           Mean   :0
##  3rd Qu.:0           3rd Qu.:0           3rd Qu.:0
##  Max.   :0           Max.   :0           Max.   :0
##
##  ViolPoints4Driver_4 ViolPoints4Driver_5 ViolPoints5Driver_1
##  Min.   :0           Min.   :0           Min.   :0
##  1st Qu.:0           1st Qu.:0           1st Qu.:0
##  Median :0           Median :0           Median :0
##  Mean   :0           Mean   :0           Mean   :0
##  3rd Qu.:0           3rd Qu.:0           3rd Qu.:0
##  Max.   :0           Max.   :0           Max.   :0
##
```

```
##  ViolPoints5Driver_2 ViolPoints5Driver_3 ViolPoints5Driver_4
##  Min.   :0           Min.   :0           Min.   :0
##  1st Qu.:0           1st Qu.:0           1st Qu.:0
##  Median :0           Median :0           Median :0
##  Mean   :0           Mean   :0           Mean   :0
##  3rd Qu.:0           3rd Qu.:0           3rd Qu.:0
##  Max.   :0           Max.   :0           Max.   :0
##
##  ViolPoints5Driver_5 ViolPoints6Driver_1 ViolPoints6Driver_2
##  Min.   :0           Min.   :0           Min.   :0
##  1st Qu.:0           1st Qu.:0           1st Qu.:0
##  Median :0           Median :0           Median :0
##  Mean   :0           Mean   :0           Mean   :0
##  3rd Qu.:0           3rd Qu.:0           3rd Qu.:0
##  Max.   :0           Max.   :0           Max.   :0
##
##  ViolPoints6Driver_3 ViolPoints6Driver_4 ViolPoints6Driver_5
##  Min.   :0           Min.   :0           Min.   :0
##  1st Qu.:0           1st Qu.:0           1st Qu.:0
##  Median :0           Median :0           Median :0
##  Mean   :0           Mean   :0           Mean   :0
##  3rd Qu.:0           3rd Qu.:0           3rd Qu.:0
##  Max.   :0           Max.   :0           Max.   :0
##
##  ViolPoints7Driver_1 ViolPoints7Driver_2 ViolPoints7Driver_3
##  Min.   :0           Min.   :0           Min.   :0
##  1st Qu.:0           1st Qu.:0           1st Qu.:0
##  Median :0           Median :0           Median :0
##  Mean   :0           Mean   :0           Mean   :0
##  3rd Qu.:0           3rd Qu.:0           3rd Qu.:0
##  Max.   :0           Max.   :0           Max.   :0
##
##  ViolPoints7Driver_4 ViolPoints7Driver_5 ViolPoints8Driver_1
##  Min.   :0           Min.   :0           Min.   :0
##  1st Qu.:0           1st Qu.:0           1st Qu.:0
##  Median :0           Median :0           Median :0
##  Mean   :0           Mean   :0           Mean   :0
##  3rd Qu.:0           3rd Qu.:0           3rd Qu.:0
##  Max.   :0           Max.   :0           Max.   :0
##
##  ViolPoints8Driver_2 ViolPoints8Driver_3 ViolPoints8Driver_4
##  Min.   :0           Min.   :0           Min.   :0
##  1st Qu.:0           1st Qu.:0           1st Qu.:0
##  Median :0           Median :0           Median :0
##  Mean   :0           Mean   :0           Mean   :0
##  3rd Qu.:0           3rd Qu.:0           3rd Qu.:0
##  Max.   :0           Max.   :0           Max.   :0
##
##  ViolPoints8Driver_5     Year_1        Make_1            Model_1
##  Min.   :0           Min.   :  -3   Length:14177       Length:14177
##  1st Qu.:0           1st Qu.:1996   Class :character   Class :character
##  Median :0           Median :2000   Mode  :character   Mode  :character
##  Mean   :0           Mean   :1996
##  3rd Qu.:0           3rd Qu.:2003
```

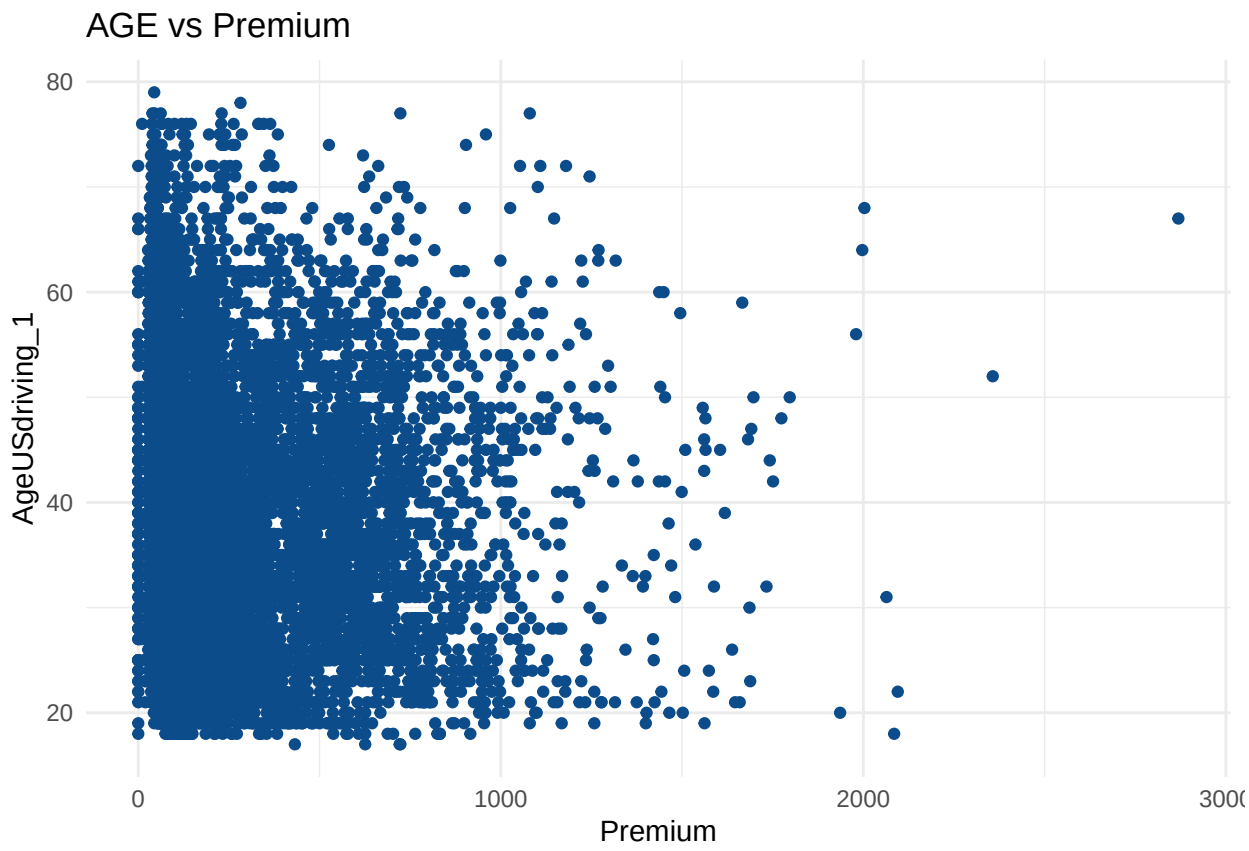
```
##  Max.   :0           Max.   :2011
##
##       Zip        Total_Distance_To_Work  NoLossSigned         Type
##  Min.   :75009   Min.   :  1.000        Min.   :0.00000   Length:14177
##  1st Qu.:78041   1st Qu.:  5.000        1st Qu.:0.00000   Class :character
##  Median :78045   Median :  5.000        Median :0.00000   Mode  :character
##  Mean   :78107   Mean   :  7.353        Mean   :0.04606
##  3rd Qu.:78119   3rd Qu.: 10.000        3rd Qu.:0.00000
##  Max.   :79605   Max.   :500.000        Max.   :1.00000
##
##  CancellationType
##  Length:14177
##  Class :character
##  Mode  :character
##
##
##
##
```

Before moving forward few points need to be considered for the sake of data correction * The missing variables need to be handled. * The missing categorical variables are replaced by the frequency of the occurrence of that respected variables. * The variables who missing percentage is very high are dropped, because the keeping of them would perform our model performance negatively and the thought of imputing those variables were not considered because it would gravely alter the gist of the actual data. * The target variable is in "num" and needs to be converted as a factor so that the classification model can be trained. * The categorical variables need to be encoded so that the models can recognize them correctly. * Outliers exists in the datasets and needs to be treated.

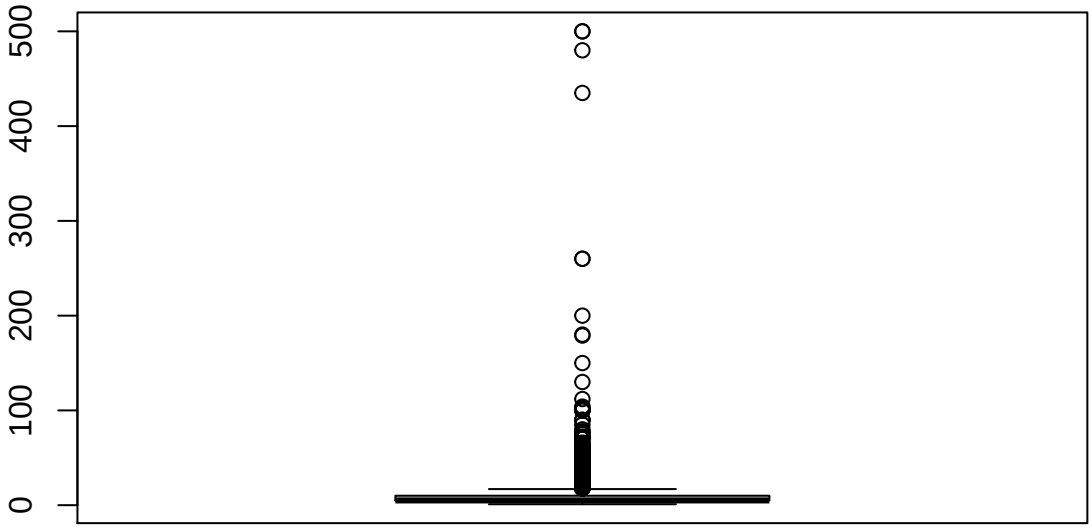
Data Visulization

- The 55% of people that made claim
- The average driving age of a person in USA is 22 years old.
- The 49% of people considered renewing their policy.
- 29 % of the data in missing.
- The age of the driver has nothing to do with the target variable claim status, they have very weak positive correlation.
- Most of the people tend to pay their Premium 6 times a year, after every 2 months.
- Male tend to have more claims than the women
- The violation points do not help much in filing the claim
- After performing Pearson correlation on to the datasets and by using the backward elimination techniques the unsignificant variables were removed and the dimensions of the data set was reduced to (14177,11).
- Violation point have no or negative impact on the ClaimStatus.
- If the distance to works increases, then the ClaimStatus decreases.
- DP is the most common Type of auto insurance

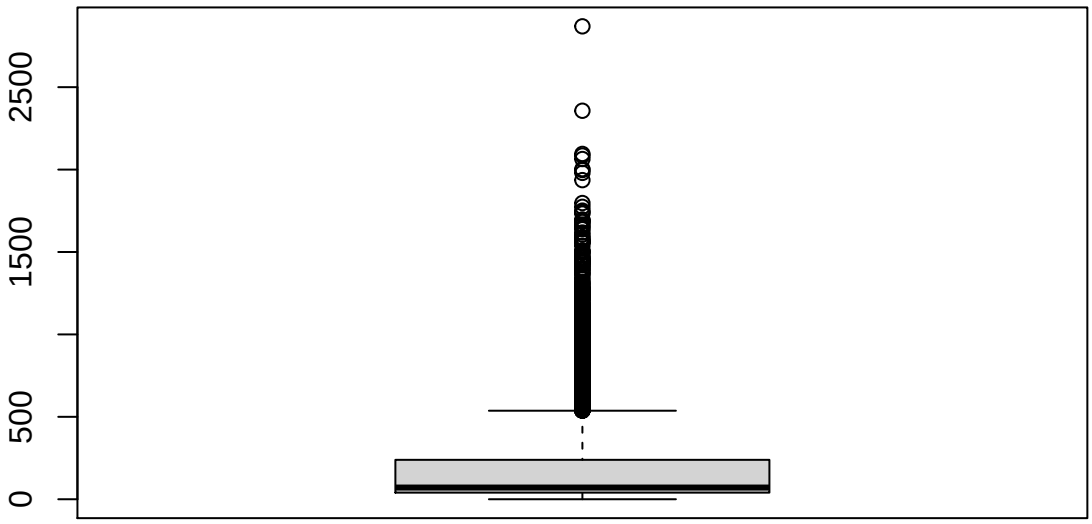
```
#plot the scatterplot to see relationship with Premium and AGEofdriving in US
ggplot(df) +
  aes(x = Premium, y = AgeUSdriving_1) +
  geom_point(colour = "#0c4c8a") +
  theme_minimal()+ggtitle("AGE vs Premium ")
```



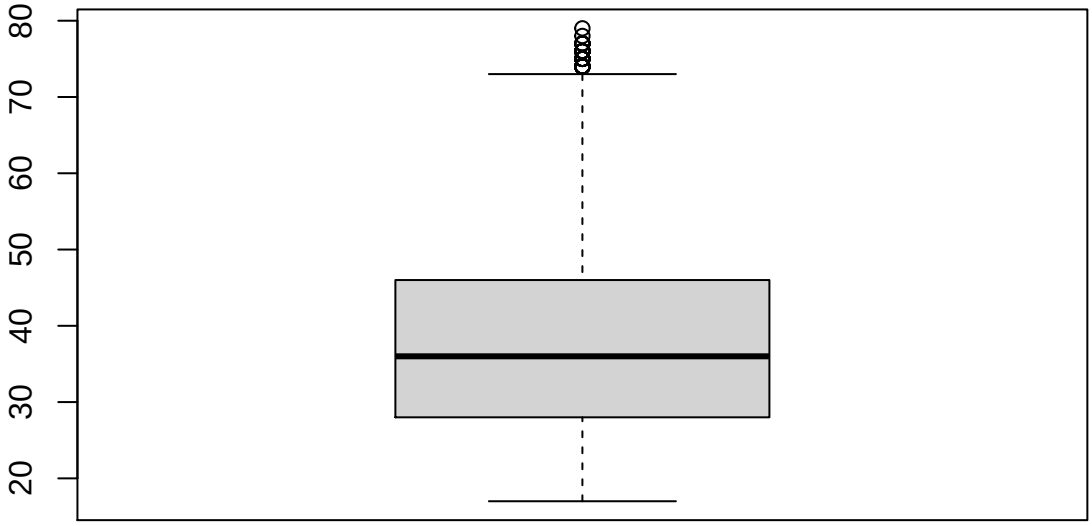

```
#plotting the boxplots to see the outliers and the distributions
boxplot(df$Total_Distance_To_Work)
```



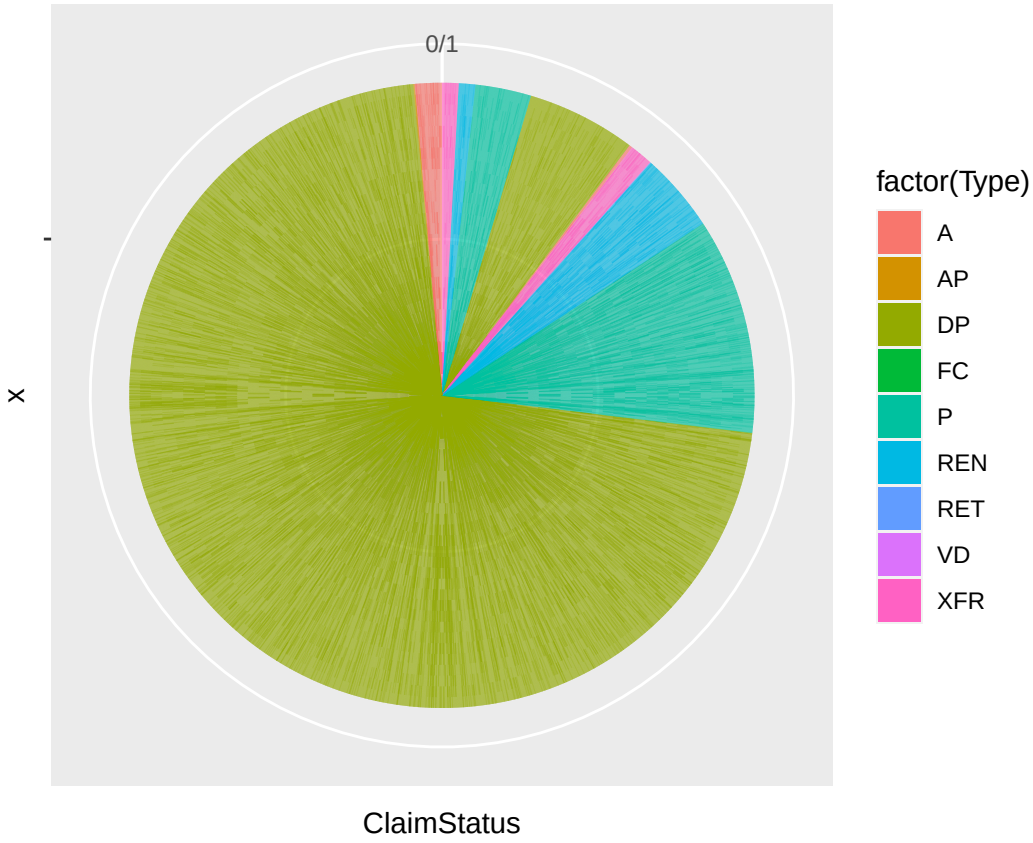

```
boxplot(df$Premium,bins=20)
```



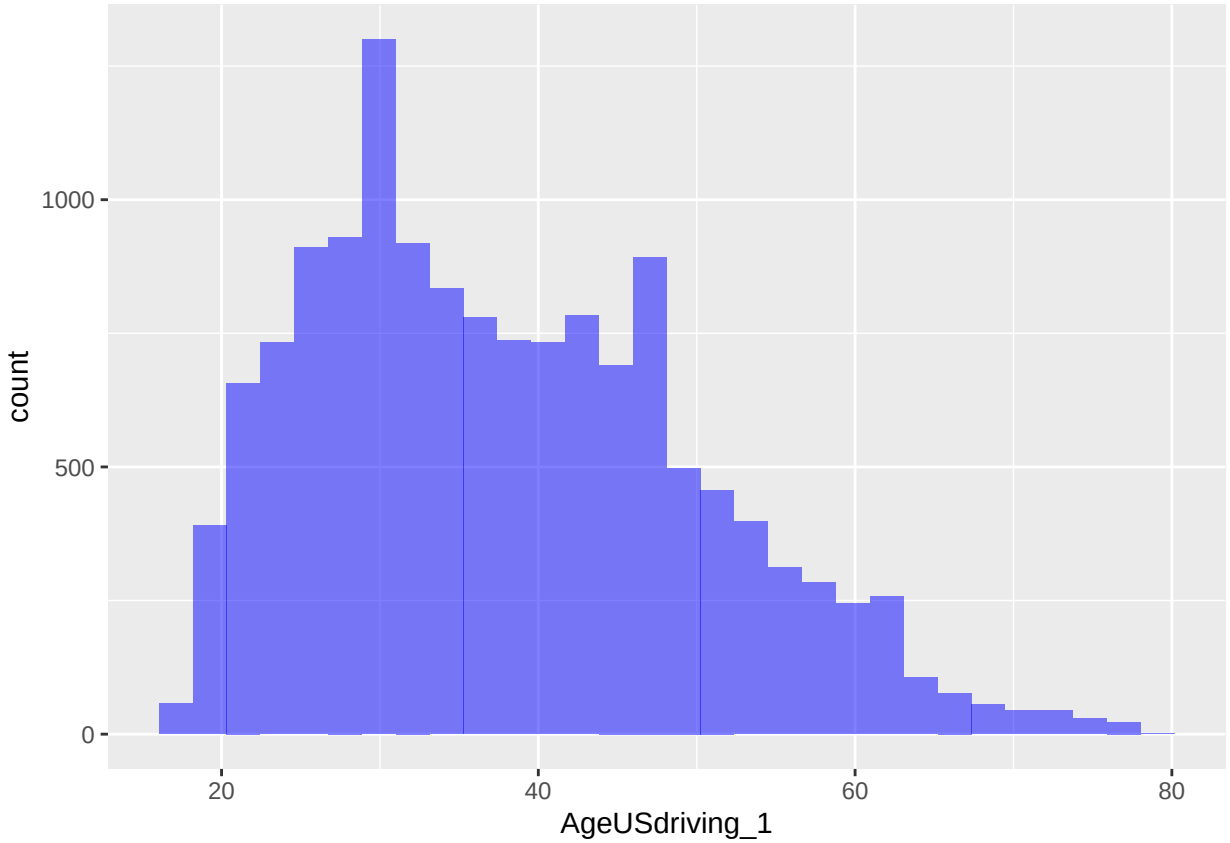

```
boxplot(df$AgeUSdriving_1,bins=20)
```



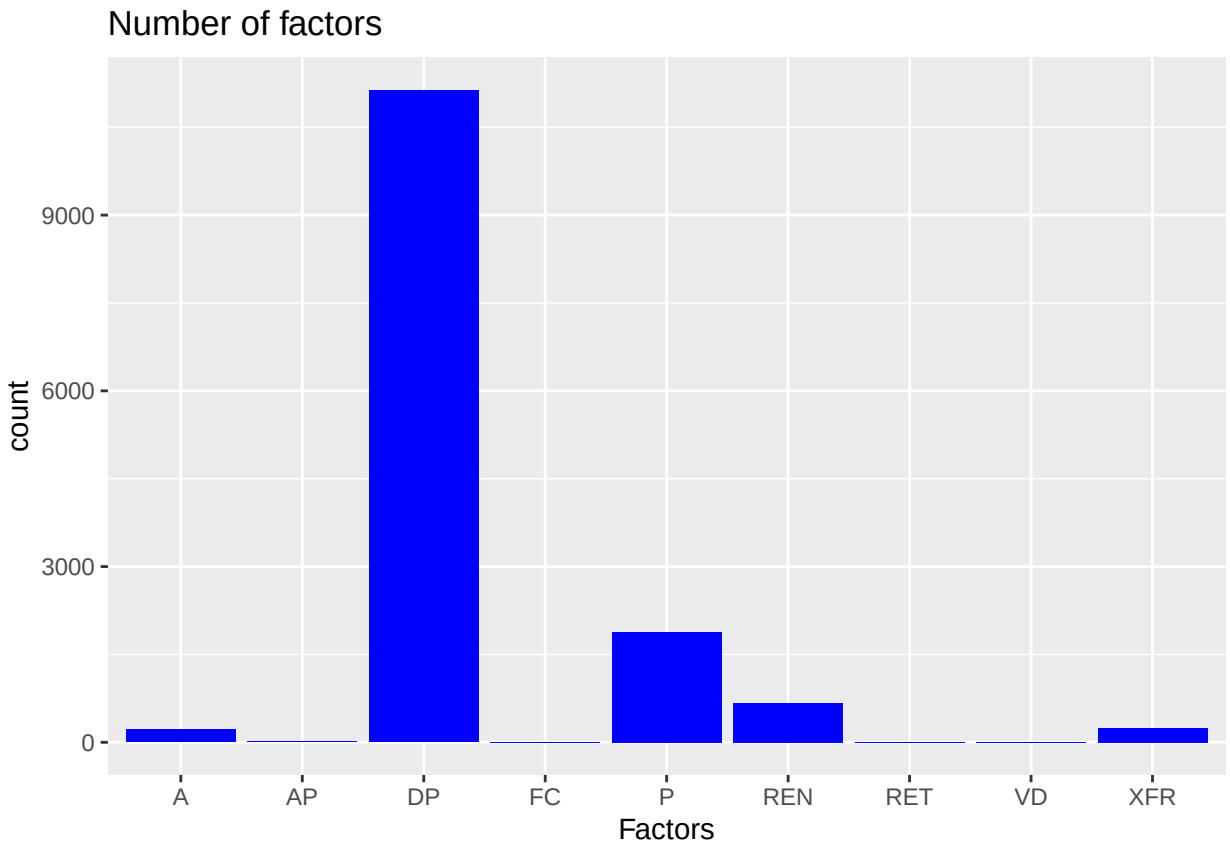

```
#we plot the barplot of the ClaimStatus count and see the types, this would be a pie plot by keeping th
ggplot(df, aes(x="", y=ClaimStatus, fill=factor(Type))) +geom_bar(stat="identity", width=1) +coord_pola
```



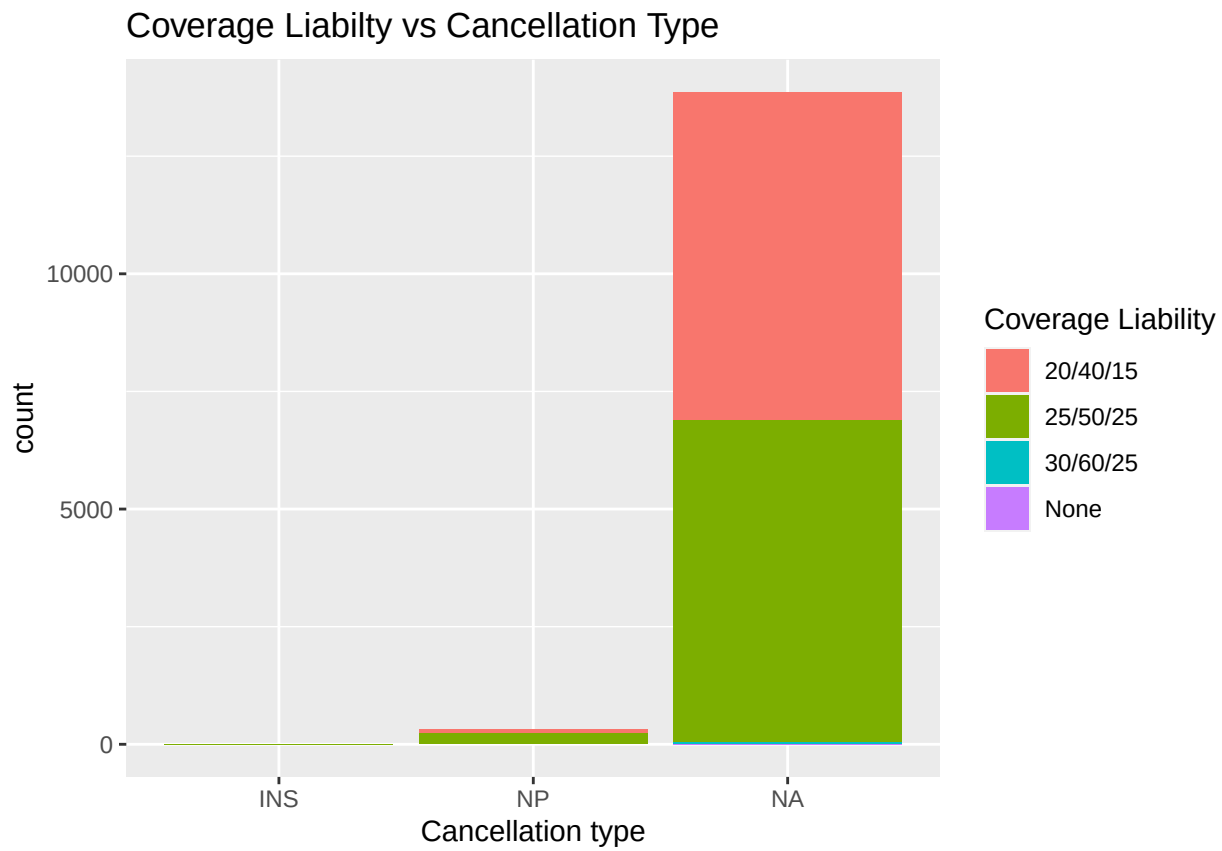

```
#check the histogram the dustribution of the age
ggplot(df,aes(AgeUSdriving_1)) + geom_histogram(fill='blue',bins=30,alpha=0.5)
```



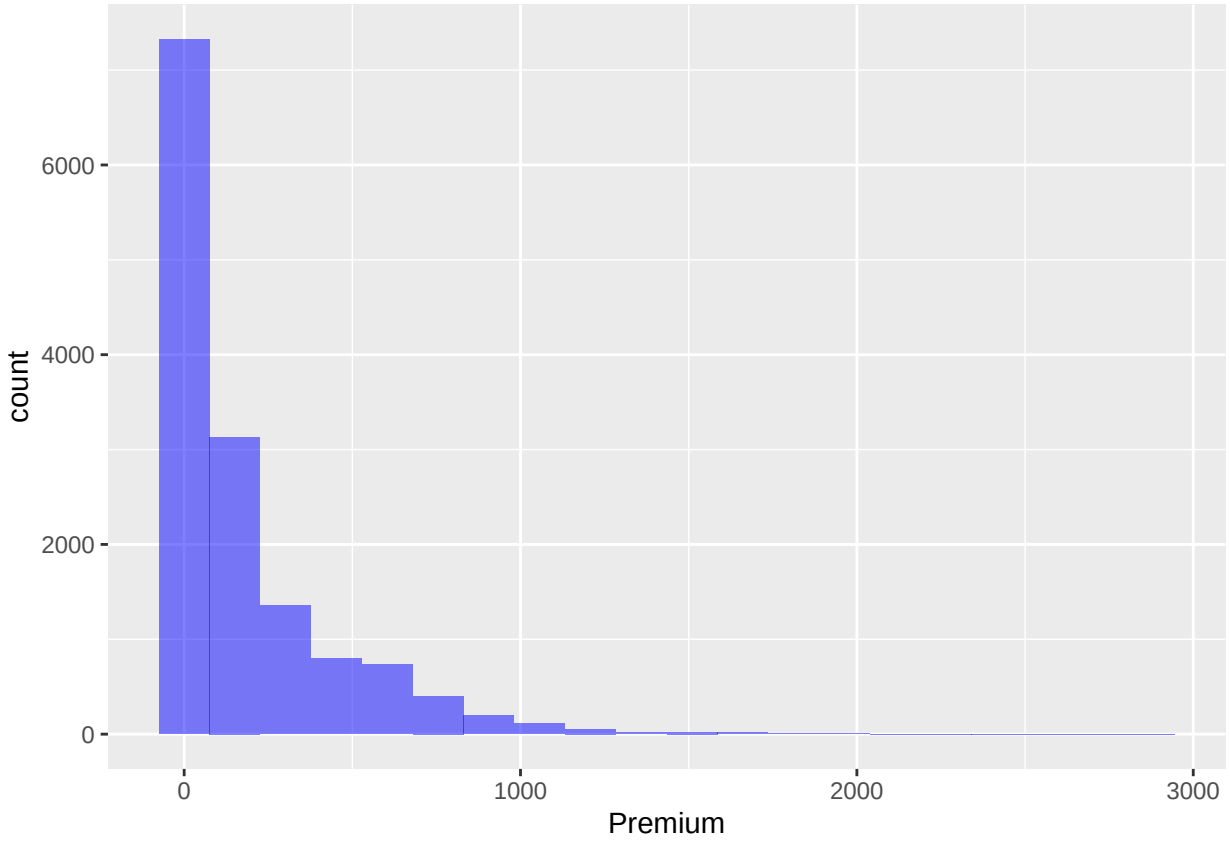

```
#next we plot the geom_bar to see the Type and their count
ggplot(data = df) +geom_bar(mapping = aes(x = factor(Type)),fill="blue")+xlab("Factors")+ggtitle("Numbe
```



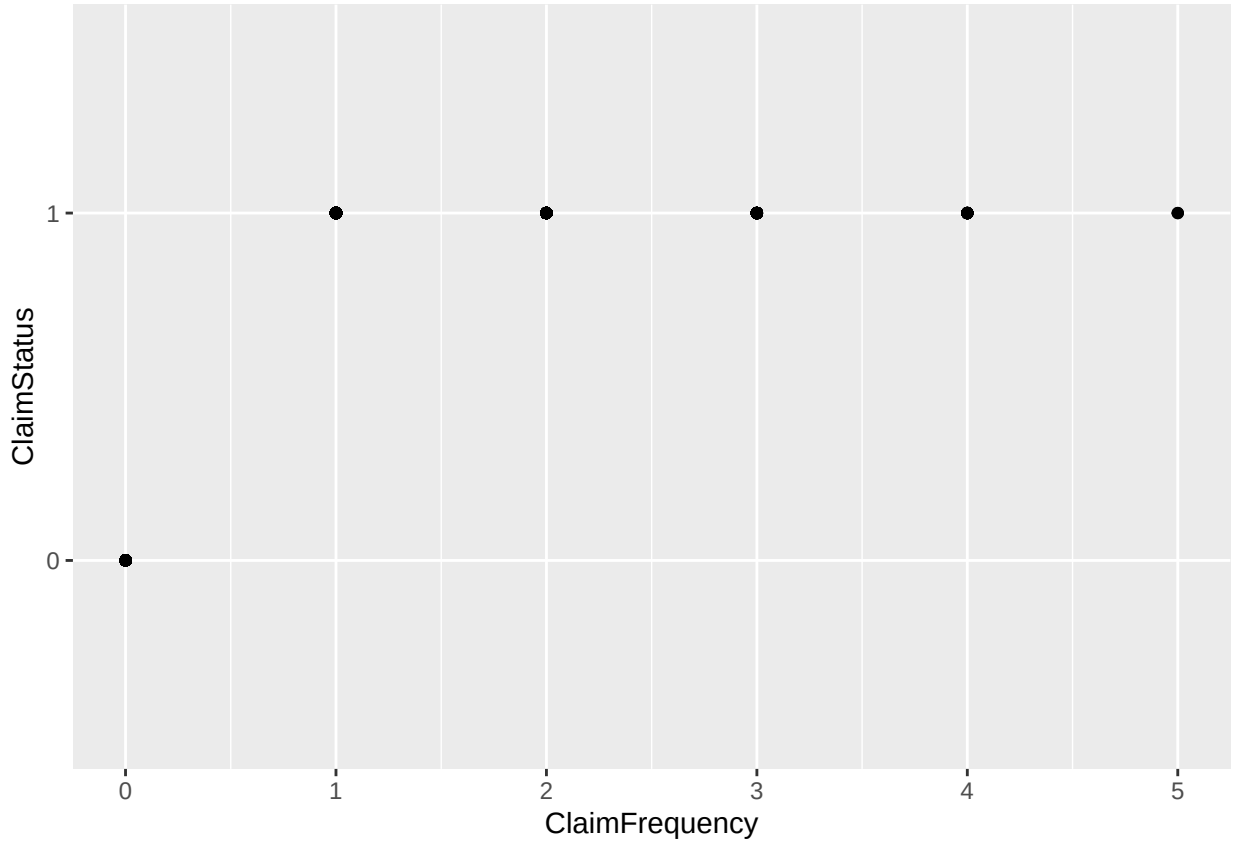

```
#we see the barplots to see the count of CancellationType by the CoverageLiability of a customer
ggplot(data = df) +geom_bar(mapping = aes(x = factor(CancellationType),fill=factor(CoverageLiability)))+
```



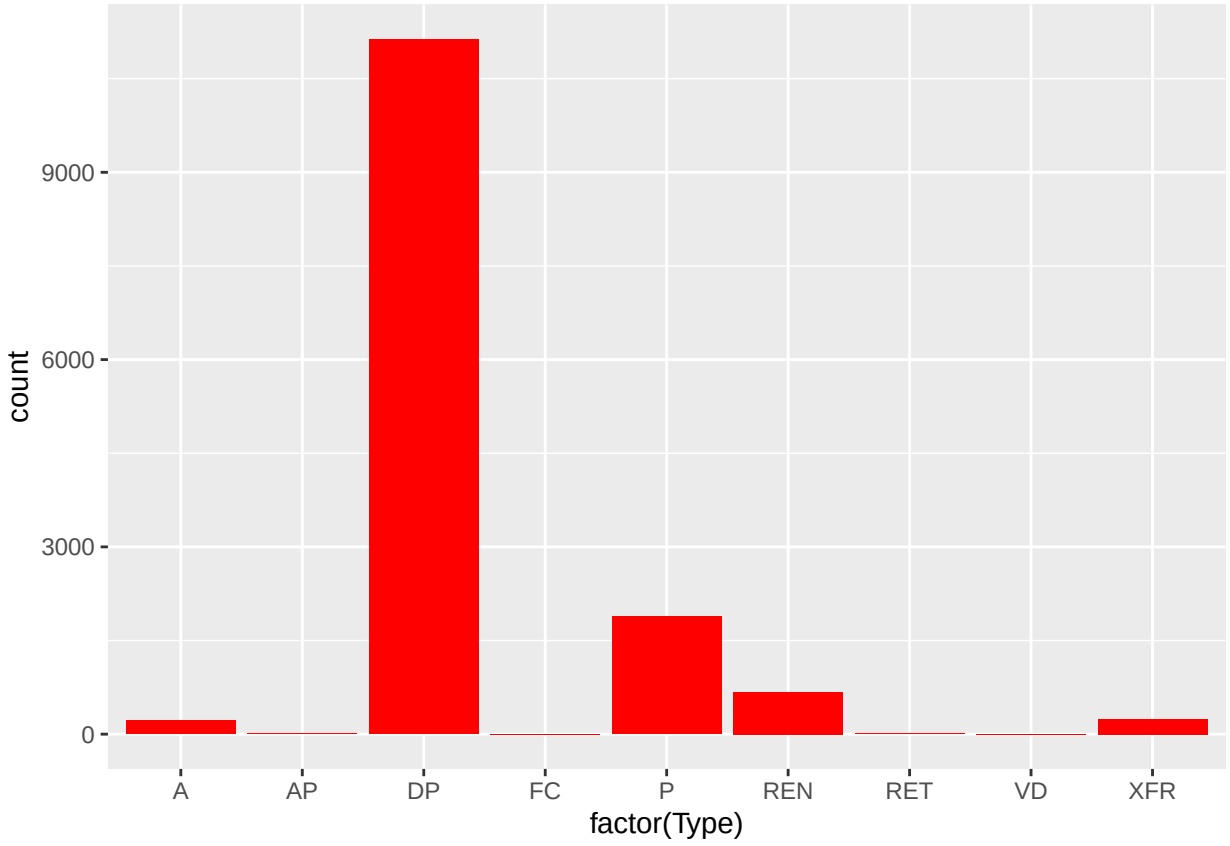

```
ggplot(df,aes(Premium)) + geom_histogram(fill='blue',bins=20,alpha=0.5)# Premium is not noramlized, we
```




```
ggplot(df,aes(x=ClaimFrequency,y=ClaimStatus))+geom_point()
```




```
ggplot(data = df) +geom_bar(mapping = aes(x = factor(Type)),fill="red")
```

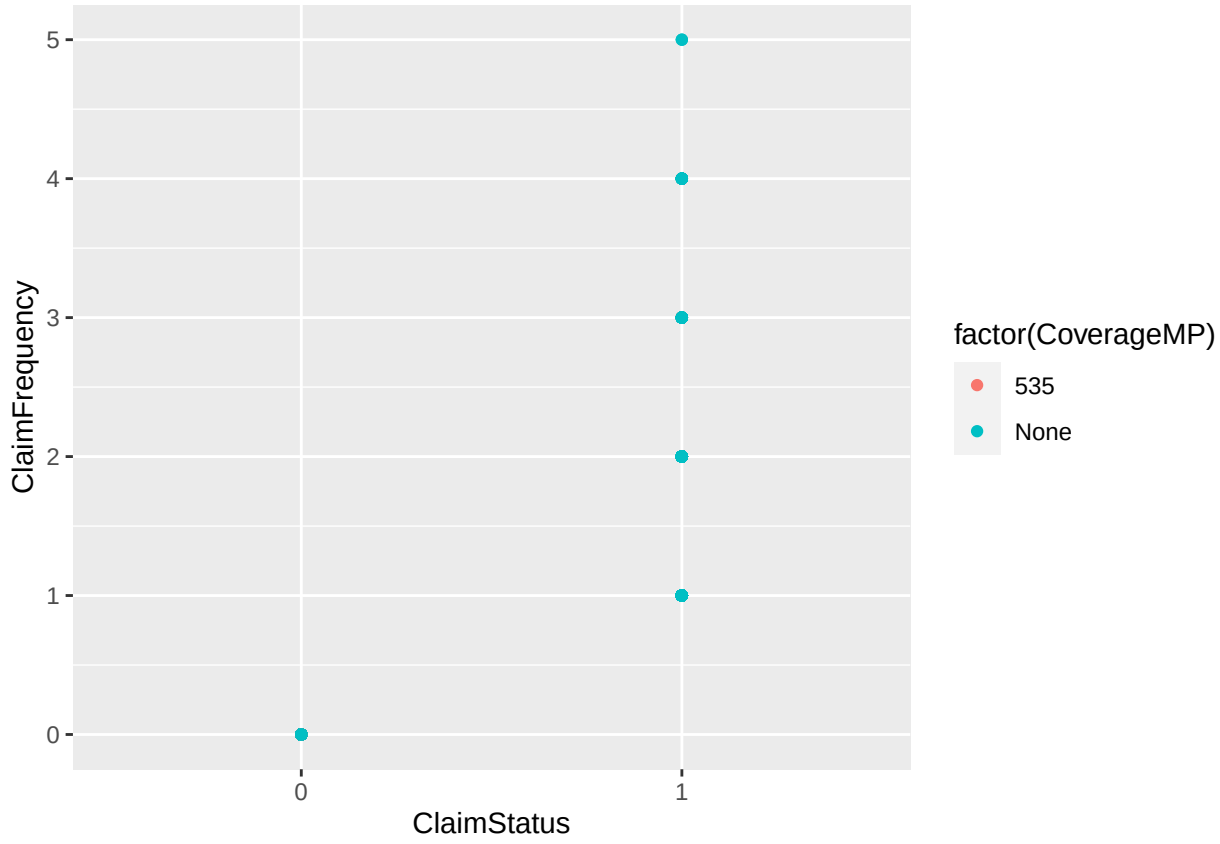


Preprocessing

As stated earlier the data has a lot of missing values which needs to be treated so that our model can be trained on the data and then the conclusion can be drawn from the data. The categorical variables are replaced by the mode of that variable. The numerical variables are replaced by their mean. The algorithms like random forest needs the input data to be in the numerical formats, so the categorical variables are converted to their equivalent numerical aliases using the Label Encoding. The data is also normalized by using Min Max Scaler.

```
#this function is used to find the unique values and then compute the mode (the most occuring values in
getmode <- function(x) {
  uniqv <- unique(x)
  uniqv[which.max(tabulate(match(x, uniqv)))]
}


# Calculate the mode using the user function and then save the values inplace of missing values
result.most.age <- getmode(df$AgeUSdriving_1) #Most common age in 28
result.most.gender <- getmode(df$Sex_1) #Most of the drivers are male

# Filling in the missing values
df$Model_1[is.na(df$Model_1)] <-getmode(df$Model_1)
# df$Make_1[is.na(df$Make_1)] <-getmode(df$Make_1)
```

```
# df$CoverageLiability[is.na(df$CoverageLiability)] <-getmode(df$CoverageLiability)

df$CoverageMP[is.na(df$CoverageMP)] <-getmode(df$CoverageMP)
df$CoveragePD_1[is.na(df$CoveragePD_1)] <-getmode(df$CoveragePD_1)
df$CoveragePIP_CDW[is.na(df$CoveragePIP_CDW)] <-getmode(df$CoveragePIP_CDW)
df$CoverageUMBI[is.na(df$CoverageUMBI)] <-getmode(df$CoverageUMBI)
df$CoverageUMPD[is.na(df$CoverageUMPD)] <-getmode(df$CoverageUMPD)


#only

p <- ggplot(df, aes(ClaimStatus, ClaimFrequency))
p + geom_point(aes(colour=factor(CoverageMP)))
```



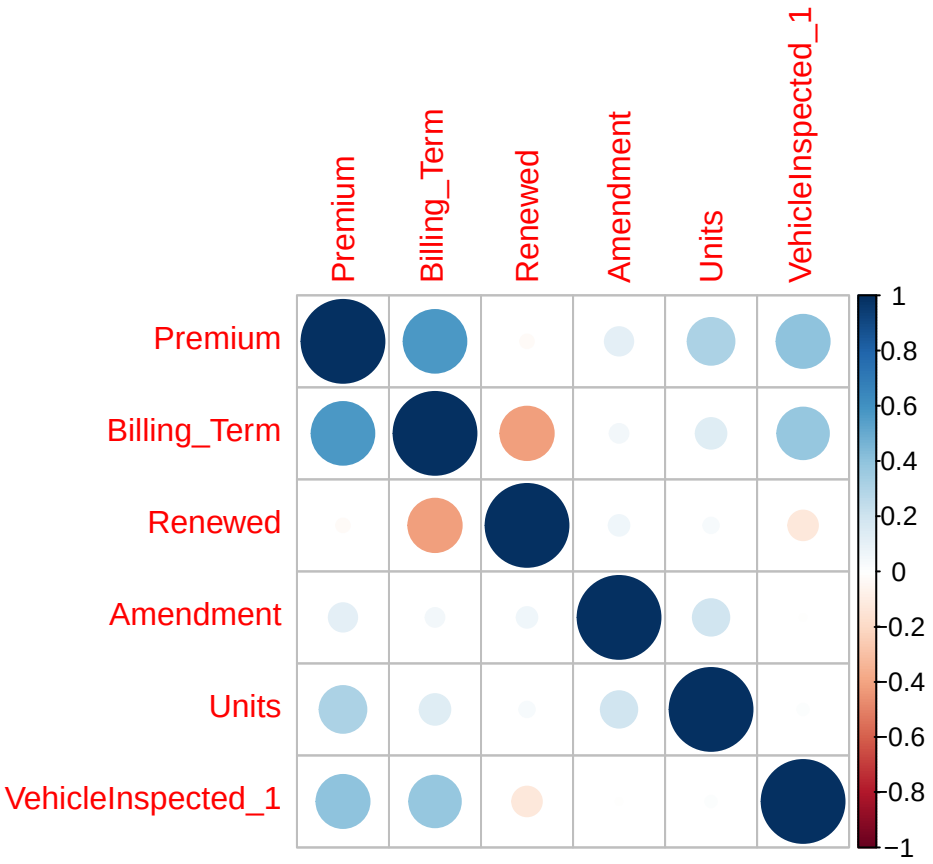

```
#convert the values to factors so that we can have classes
#the logic is to first get the unqiue values and convert to list
#Next is to take the one elemet less than the length because the encoded values
#are starting from zero
#in the factor argument the labels represnts the labels required

list.of.make.1 <- as.list(unique(df$Type))
length.of.make1 <- (length(list.of.make.1)-1)
df$Type<- factor(df$Type,levels = list.of.make.1,labels =c(0:8))

#
```

```
list.of.make.1 <- as.list(unique(df$CoverageLiability))
length.of.make1 <- (length(list.of.make.1)-1)
df$CoverageLiability <- factor(df$CoverageLiability,levels = list.of.make.1,labels =c(0:3))

list.of.make.1 <- as.list(unique(df$Model_1))
length.of.make1 <- (length(list.of.make.1)-1)
df$Model_1 <- factor(df$Model_1,levels = list.of.make.1,labels =c(0:1447)) #ONEHOT
#
list.of.make.1 <- as.list(unique(df$CoverageMP))
length.of.make1 <- (length(list.of.make.1)-1)
df$CoverageMP <- factor(df$CoverageMP,levels = list.of.make.1,labels = c(1:2))

#

#
list.of.make.1 <- as.list(unique(df$CoveragePD_1))
length.of.make1 <- (length(list.of.make.1)-1)
df$CoveragePD_1 <- factor(df$CoveragePD_1 ,levels = list.of.make.1,labels = c(1:3))
#
list.of.make.1 <- as.list(unique(df$CoveragePIP_CDW))
length.of.make1 <- (length(list.of.make.1)-1)
df$CoveragePIP_CDW <- factor(df$CoveragePIP_CDW ,levels = list.of.make.1,labels = c(1:3))
#

#

#

#here we replace the continous values by the MinMaxScaler so that every value is on the same scale

df$Premium <- (df$Premium-min(df$Premium))/(max(df$Premium) - min(df$Premium))
df$ClaimFrequency <- (df$ClaimFrequency-min(df$ClaimFrequency))/(max(df$ClaimFrequency) - min(df$ClaimF
df$VehicleInspected_1 <- (df$VehicleInspected_1-min(df$VehicleInspected_1))/(max(df$VehicleInspected_1)
df$Units <- (df$Units-min(df$Units))/(max(df$Units) - min(df$Units))
df$Billing_Term <- (df$Billing_Term-min(df$Billing_Term))/(max(df$Billing_Term) - min(df$Billing_Term))
df$Renewed <- (df$Renewed-min(df$Renewed))/(max(df$Renewed) - min(df$Renewed))
df$Amendment <- (df$Amendment-min(df$Amendment))/(max(df$Amendment) - min(df$Amendment))

df$VehicleInspected_1 <-(df$VehicleInspected_1-min(df$VehicleInspected_1))/(max(df$VehicleInspected_1) -









#We can see that only few attributes have larger affect on the claimstatus,
#So we have to drop those irreveleant columns
# 88:89,91:119,120,123,1
```

Preparing the data for model training

The dataset is divided into two set one for training the model and the other for predicting the performance of the model. 75% of the data is used for training and the rest of the data is use for testing. The data is shuffled so that each a mix of the data can be the part of the training and predictions.

```
#drop irrelevant features
df <- subset(df,select = -c(1,3,7:11,12,13:17,20,23,24,25:52,53,54:62,63,64:68,69,70:73,74,75,76,77,80:
#see the correlation plot of only of the numerical vals
only_num <- sapply(df, is.numeric)
corrplot(cor(df[,only_num]),method = 'circle')
```




```
M <- as.data.frame(cor(df[,only_num]))
table(df$ClaimStatus)
```

```
##
##     0     1
## 13399   778
```

```
sum(df$ClaimStatus ==1 )/nrow(df)
```

```
## [1] 0.05487762
```

```
#split data into train and test
#TRAIN TEST SPLIT
a=gensvm.train.test.split(x=df, train.size = 0.95,
                          shuffle = T,
                         return.idx = FALSE,random.state = 101)
```

#LOGISTIC REGRESSION Logistic regression is used as the first model to train the classification the model. Binomial family was used in training.

```
#we want to predict the ClaimStatus so we write the formula s given below and then predict on test set
log.model <- glm(formula=ClaimStatus ~ . ,data = a$x.train,family=binomial)
summary(log.model)
```

```
##
## Call:
## glm(formula = ClaimStatus ~ ., family = binomial, data = a$x.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5875  -0.3362  -0.2338  -0.1854   3.1415
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -4.8789     0.1433 -34.040  < 2e-16 ***
## Premium               1.7476     0.5230   3.342 0.000833 ***
## Billing_Term          0.2347     0.1366   1.718 0.085771 .
## Renewed               1.3371     0.1022  13.078  < 2e-16 ***
## Amendment            -2.5976     0.8922  -2.911 0.003600 **
## CoverageLiability1   -0.7452     0.0908  -8.207 2.26e-16 ***
## CoverageLiability2  -12.9569  1455.3976  -0.009 0.992897
## CoverageLiability3  -12.6429   241.5335  -0.052 0.958254
## CoveragePD_12         0.8124     0.1346   6.036 1.58e-09 ***
## CoveragePD_13       -12.9455  1455.3976  -0.009 0.992903
## CoveragePIP_CDW2      0.1006     0.3640   0.276 0.782322
## CoveragePIP_CDW3    -12.7140  1455.3975  -0.009 0.993030
## Units                 2.7040     0.2182  12.394  < 2e-16 ***
## VehicleInspected_1   -0.1759     0.1217  -1.446 0.148230
## Type1                 0.4285     0.1089   3.934 8.34e-05 ***
## Type2                 0.6478     0.1878   3.450 0.000561 ***
## Type3                 0.1097     0.1606   0.683 0.494686
## Type4                -0.1796     0.3110  -0.578 0.563516
## Type5               -11.3386   640.2128  -0.018 0.985870
## Type6               -11.8418   690.8013  -0.017 0.986323
## Type7               -12.6287   428.6933  -0.029 0.976499
## Type8               -14.4238  1022.4139  -0.014 0.988744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5749.7  on 13467  degrees of freedom
## Residual deviance: 4879.8  on 13446  degrees of freedom
## AIC: 4923.8
```

```
##
## Number of Fisher Scoring iterations: 14
```

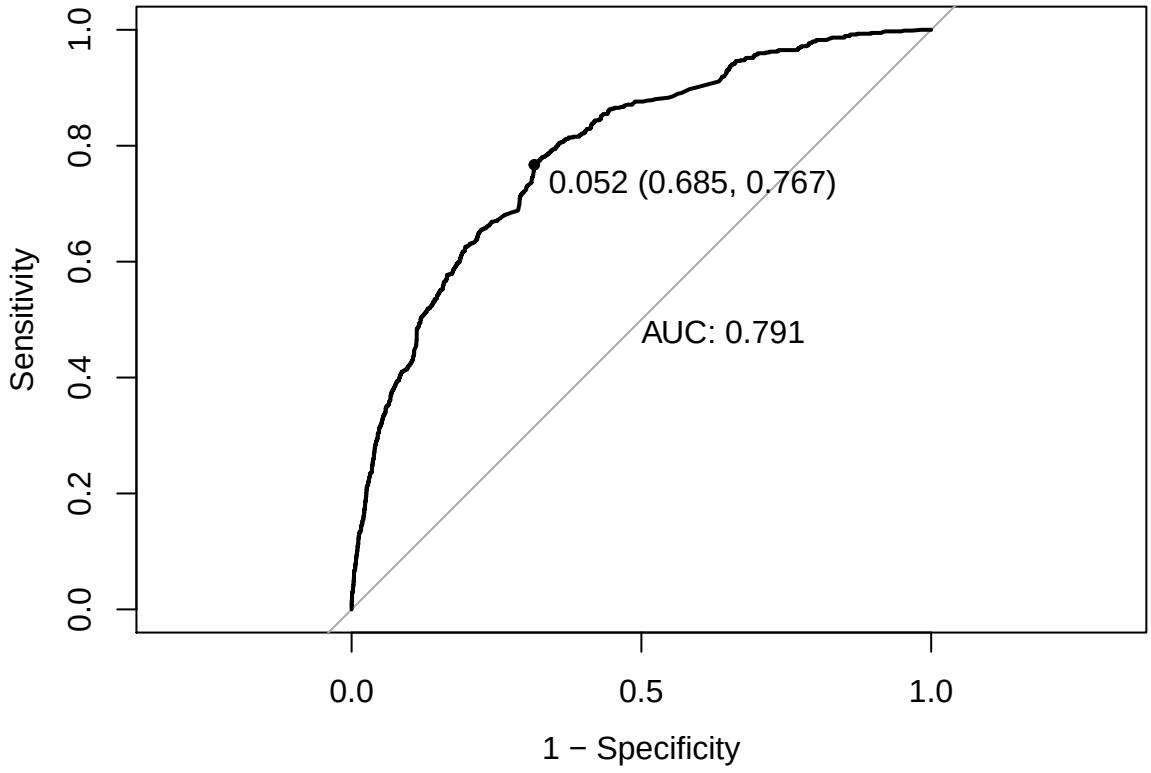
```
prob_pred <- predict(log.model, type = 'response', newdata = a$x.test)

#convert probabilties to classes
fitted.results <- ifelse(prob_pred > 0.5,1,0)
#see the confusion matrix to get the performance of the model
confusionMatrix(as.factor(fitted.results),as.factor(a$x.test$ClaimStatus))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 672  32
##          1   2   3
##
##                Accuracy : 0.952
##                  95% CI : (0.9336, 0.9666)
##     No Information Rate : 0.9506
##     P-Value [Acc > NIR] : 0.4758
##
##                   Kappa : 0.1394
##
##  Mcnemar's Test P-Value : 6.577e-07
##
##             Sensitivity : 0.99703
##             Specificity : 0.08571
##          Pos Pred Value : 0.95455
##          Neg Pred Value : 0.60000
##              Prevalence : 0.95063
##          Detection Rate : 0.94781
##    Detection Prevalence : 0.99295
##       Balanced Accuracy : 0.54137
##
##        'Positive' Class : 0
##
```

```
#see the ROC values and plot the ROC curve

roc(a$x.train$ClaimStatus, log.model$fitted.values, plot=TRUE,legacy.axes=TRUE,print.thres=T,print.auc='
```

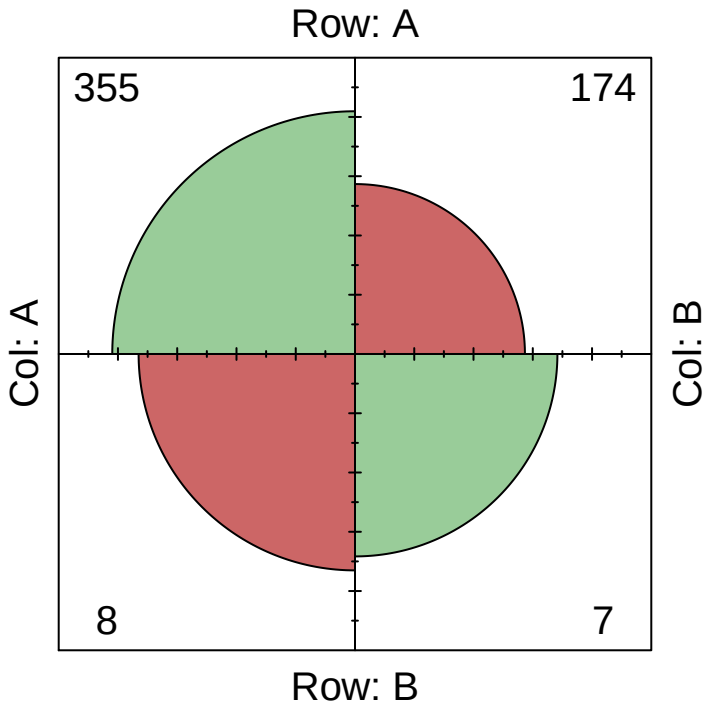
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = a$x.train$ClaimStatus, predictor = log.model$fitted.values,     plot = TRUE, l
##
## Data: log.model$fitted.values in 12725 controls (a$x.train$ClaimStatus 0) < 743 cases (a$x.train$Clai
## Area under the curve: 0.7909
```

```
par(pty = "s")

#this code is used to plot the values the TP,FP,TN,FN values are given as matrirx and the other argumen
ctable <- as.table(matrix(c(355, 174, 8, 7), nrow = 2, byrow = TRUE))
fourfoldplot(ctable, color = c("#CC6666", "#99CC99"),
             conf.level = 0, margin = 1, main = "Logistic Regression Confusion Matrix")
```

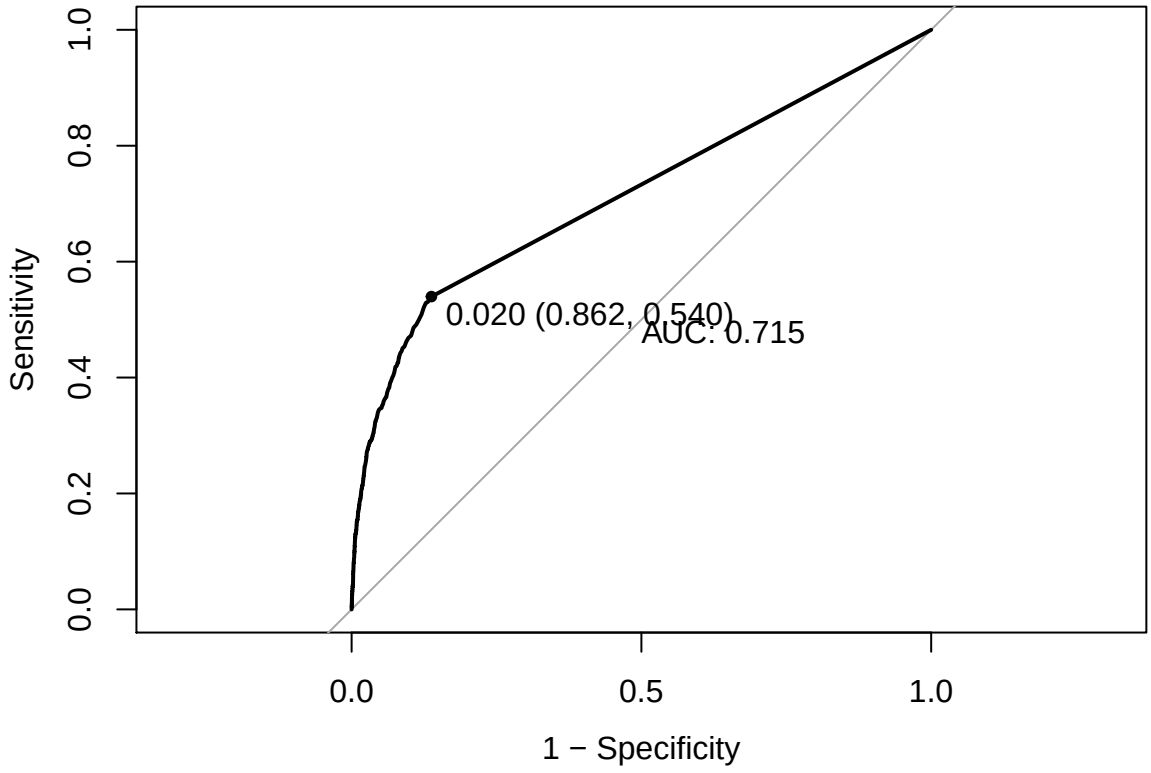


#RANDOMFOREST The second algorithm we used is Random forest, 100 trees were generated to make train the model

```
#get the randomforest fit on the data
forest.model <- randomForest(x = a$x.train[-1],
                             y = a$x.train$ClaimStatus,
                             ntree = 100)
#predict on test set
y_pred.forest = predict(forest.model, newdata = a$x.test)
y_pred.forest.prob = predict(forest.model, newdata = a$x.test,type='prob')

#plot the roc curve
roc(a$x.train$ClaimStatus,forest.model$votes[,2] , plot=TRUE,legacy.axes=TRUE,print.thres=T,print.auc=T
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = a$x.train$ClaimStatus, predictor = forest.model$votes[,     2], plot = TRUE, l
##
## Data: forest.model$votes[, 2] in 12725 controls (a$x.train$ClaimStatus 0) < 743 cases (a$x.train$Clai
## Area under the curve: 0.7153
```

```
par(pty = "s")

confusionMatrix(as.factor(a$x.test$ClaimStatus),as.factor(y_pred.forest))
```

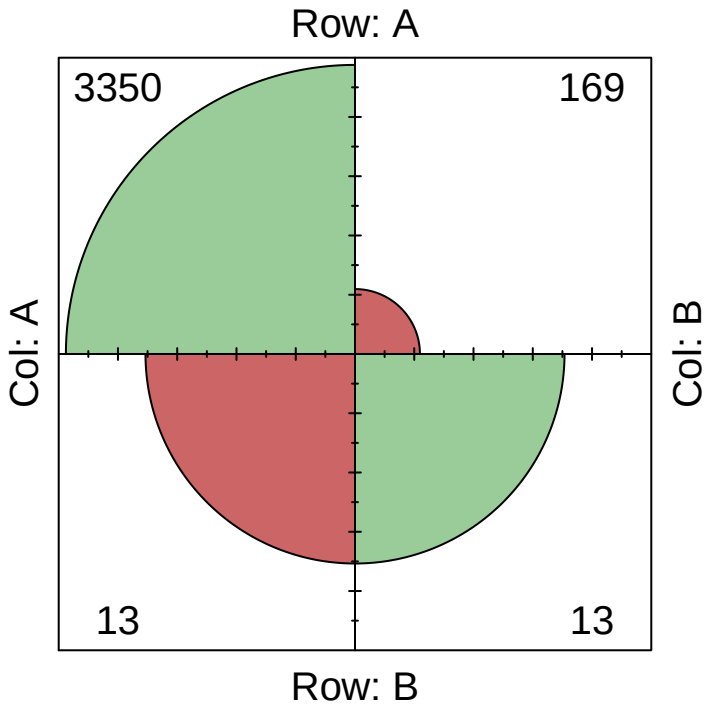
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 673   1
##          1  32   3
##
##                Accuracy : 0.9535
##                  95% CI : (0.9353, 0.9677)
##     No Information Rate : 0.9944
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1452
##
```

```
##  Mcnemar's Test P-Value : 1.767e-07
##
##             Sensitivity : 0.95461
##             Specificity : 0.75000
##          Pos Pred Value : 0.99852
##          Neg Pred Value : 0.08571
##              Prevalence : 0.99436
##          Detection Rate : 0.94922
##    Detection Prevalence : 0.95063
##       Balanced Accuracy : 0.85230
##
##        'Positive' Class : 0
##
```

```
confusionMatrix(as.factor(y_pred.forest),as.factor(a$x.test$ClaimStatus))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 673  32
##          1   1   3
##
##                Accuracy : 0.9535
##                  95% CI : (0.9353, 0.9677)
##     No Information Rate : 0.9506
##     P-Value [Acc > NIR] : 0.4069
##
##                   Kappa : 0.1452
##
##  Mcnemar's Test P-Value : 1.767e-07
##
##             Sensitivity : 0.99852
##             Specificity : 0.08571
##          Pos Pred Value : 0.95461
##          Neg Pred Value : 0.75000
##              Prevalence : 0.95063
##          Detection Rate : 0.94922
##    Detection Prevalence : 0.99436
##       Balanced Accuracy : 0.54212
##
##        'Positive' Class : 0
##
```

```
ctable <- as.table(matrix(c(3350, 169, 13, 13), nrow = 2, byrow = TRUE))
fourfoldplot(ctable, color = c("#CC6666", "#99CC99"),
             conf.level = 0, margin = 1, main = "Random Forest Confusion Matrix")
```



Actionable Insight and recommendations

- It is more likely the people who have a good premium and DP policy are more likely to male the claims so that company must focus on those customers. car related reimbursement in company .
- Model Performance values for Train and test are within the maximum tolerance deviation of +/- 10%. Hence, the all models are not over-fitting.
- Company should introduce more features in data for better analysis and also suggest furthercoverage plans to get better understanding of the customer needs, pooling suggestions to focus group. It would increase the focus group volume.
- Company should keep track of the any amendment in the law because it may cause the working to get changed completely.
- Company should keep the vehicles inspected to reduce the Claims.