**Swiss Insurance - R Project**

getwd()

View(swiss)

# Handling missing data

mean(swin$Insured, na.rm = TRUE)

[1] 1092.195

range(swin$Insured, na.rm = TRUE)

[1] 0.01 127687.27

sd(swin$Insured, na.rm = TRUE)

[1] 5661.156

median(swin$Insured, na.rm = TRUE)

[1] 81.525

tapply(swin$Claims,swin$Payment, FUN=mean, na.rm = TRUE)

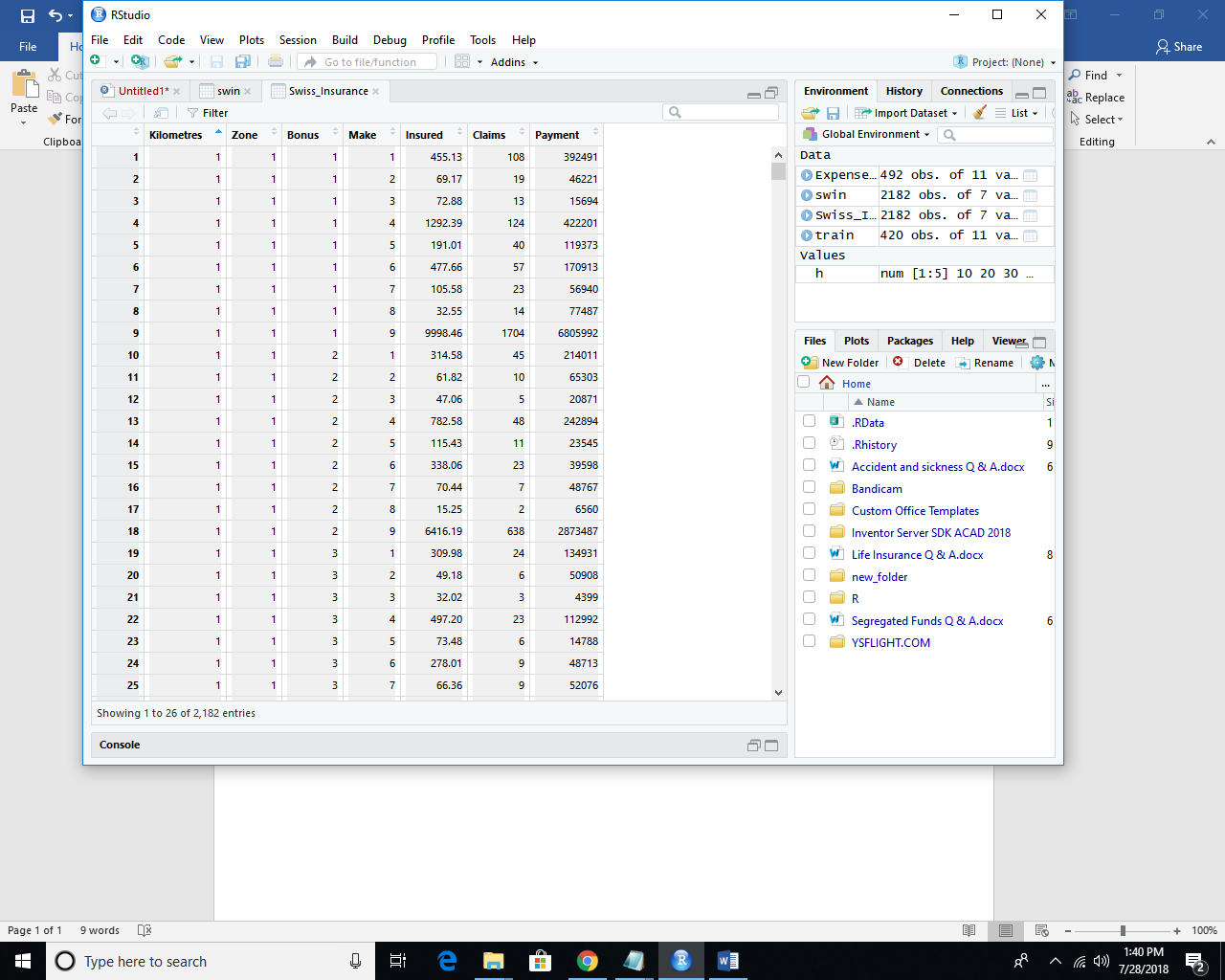
tapply(swin$Claims,swin$Payment, FUN=range, na.rm = TRUE)

tapply(swin$Claims,swin$Payment, FUN=sd, na.rm = TRUE)

tapply(swin$Claims,swin$Payment, FUN=median, na.rm = TRUE)

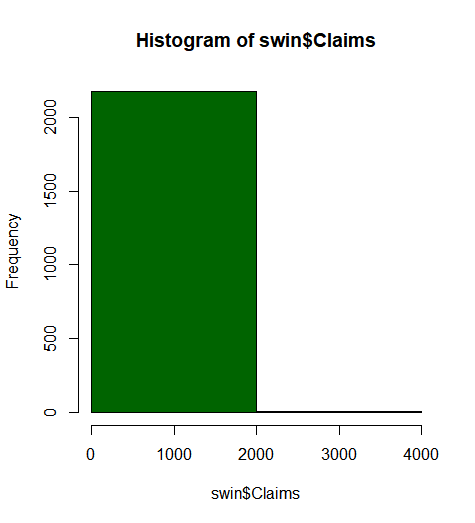
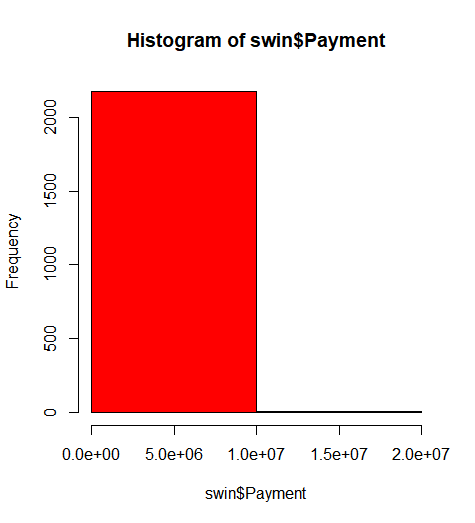
str(swin)

|  |
| --- |
| Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 2182 obs. of 7 variables:  $ Kilometres: num 1 1 1 1 1 1 1 1 1 1 ...  $ Zone : num 1 1 1 1 1 1 1 1 1 1 ...  $ Bonus : num 1 1 1 1 1 1 1 1 1 2 ...  $ Make : num 1 2 3 4 5 6 7 8 9 1 ...  $ Insured : num 455.1 69.2 72.9 1292.4 191 ...  $ Claims : num 108 19 13 124 40 ...  $ Payment : num 392491 46221 15694 422201 119373 ... |
| head(swin)  # A tibble: 6 x 7  Kilometres Zone Bonus Make Insured Claims Payment  *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*  1 1 1 1 1 455. 108 392491  2 1 1 1 2 69.2 19 46221  3 1 1 1 3 72.9 13 15694  4 1 1 1 4 1292. 124 422201  5 1 1 1 5 191. 40 119373  6 1 1 1 6 478. 57 170913  nrow(swin)  [1] 2182  ncol(swin)  [1] 7  swin$Zone <- as.factor(swin$Zone)  swin$Make <- as.factor(swin$make)  swin$Bonus <- as.factor(swin$Bonus)  swin$Make <- as.factor(swin$Make)  swin$Kilometres <- as.factor(swin$Kilometres) |



hist(swin$Claims, breaks = 2, col ="darkgreen")

hist(swin$Payment, breaks = 2, col = "red")

|  |
| --- |
|  |
| |  | | --- | |  | |
|  |
|  |
|  |

# The committee has decided to find whether payment is related to number of claims and the number of insured policy years. They want to visualize the results.

plot(swin$Payment,swin$Claims)

cor(swin$Payment,swin$Claims)

plot(swin$Payment,swin$Insured)

cor(swin$Payment,swin$Insured)

cor(df2)

mat <-cor(df2)

corrplot(mat)

swin.Payment swin.Claims swin.Insured

swin.Payment 1.0000000 0.9954003 0.9332170

swin.Claims 0.9954003 1.0000000 0.9103478

swin.Insured 0.9332170 0.9103478 1.0000000

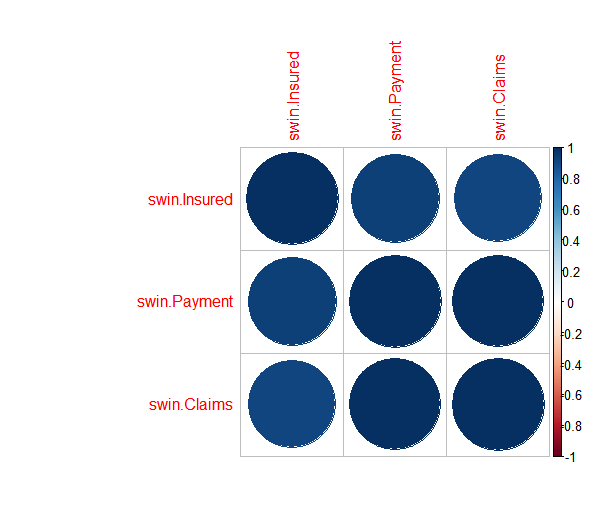
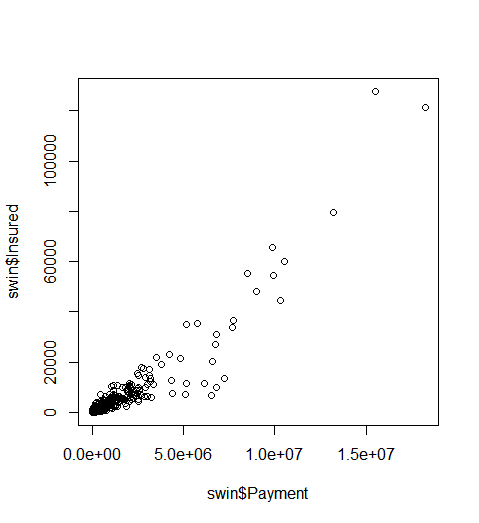
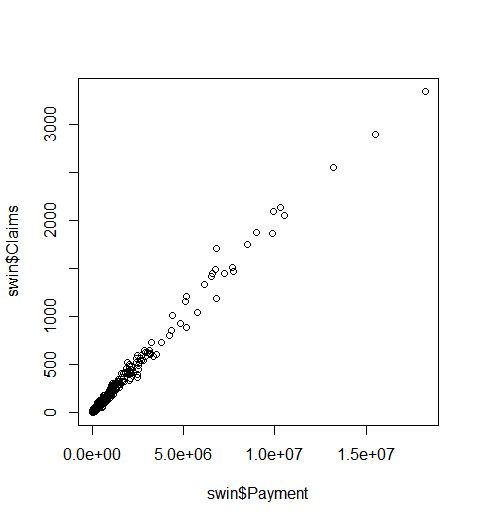
)

> cor(swin$Payment,swin$Claims)

[1] 0.9954003

> cor(swin$Payment,swin$Insured)

[1] 0.933217



#####3 question Whether distance , location bonus,make , insured amount or claim is effecting the payment

pairs(swin)

library(caTools)

set.sead(1234)

split<- sample.split(swin,SplitRatio = 0.75)

trn <-subset(swin,split = TRUE)

test <-subset(swin,split = FALSE)

head(trn)

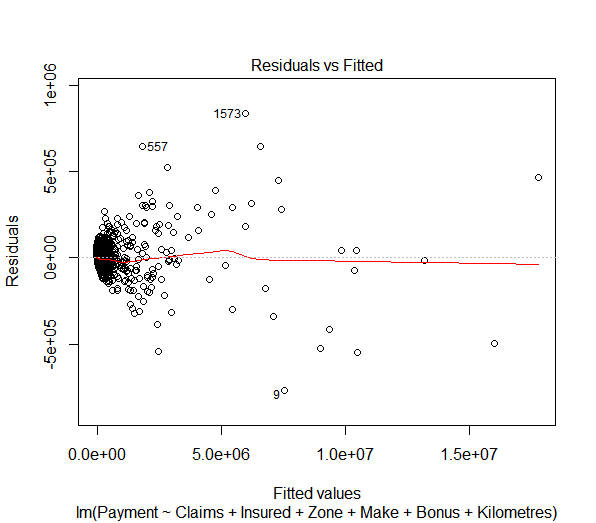
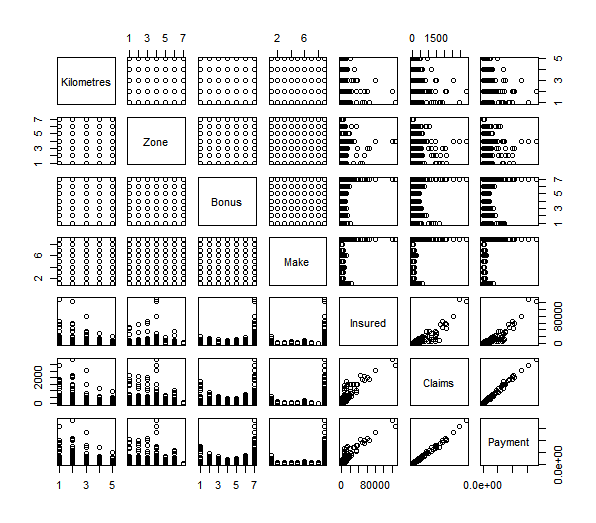
head(test)

### Fit a regression model

attach(trn)

ml <-lm(Payment ~ Claims + Insured + Zone + Make + Bonus + Kilometres, data = swin)

summary(ml)



#the committee is interested to know at what level their bonus level, location,kilometer should be increased for the Insured amount, claims and payment to be increased

df1 <- data.frame(swin$Kilometres,swin$Zone,swin$Bonus,swin$Payment)

df1

str(df1)

combined\_groups <- data.frame(cbind.data.frame(df1))

summary(combined\_groups)

|  |
| --- |
| swin.Kilometres swin.Zone swin.Bonus swin.Payment  Min. :1.000 Min. :1.00 Min. :1.000 Min. : 0  1st Qu.:2.000 1st Qu.:2.00 1st Qu.:2.000 1st Qu.: 2989  Median :3.000 Median :4.00 Median :4.000 Median : 27404  Mean :2.986 Mean :3.97 Mean :4.015 Mean : 257008  3rd Qu.:4.000 3rd Qu.:6.00 3rd Qu.:6.000 3rd Qu.: 111954  Max. :5.000 Max. :7.00 Max. :7.000 Max. :18245026 |
| Here the kilometers is max at 5 ie >25000 the zone is Gotland Bonus is 7 to get the max Payment |
| |  | | --- | | > | |

#The committee wants to decide the right premium, hence they need to find the insured amount,zone,kilometer, bonus, or make affects the rates to what extent.

#Decision Tree Predictive model

# Install package rpart

# Install package rpart.plot

# Install package caret

library(rpart)

library(rpart.plot)

library(caret)

set.seed(1234)

trn <-subset(swin,split = TRUE)

test <-subset(swin,split = FALSE)

head(trn)

head(test)

class(test)

library(rpart)

library(rpart.plot)

m1 <-rpart(Payment ~., data=trn, method="class")

m1

rpart.plot(m1,type = 3, digits = 3,extra = 101 )

print(ml)

test$predict.payment <- predict(m1,test,type = "class")

t <-table(test$predict.payment,test$Payment)

c<-cofusionmatrix(text$predict.payment,test$Payment)

n= 2182

#Interpretation

node), split, n, deviance, yval

\* denotes terminal node

1) root 2182 2.257038e+15 257007.60

2) Claims< 763.5 2156 3.991262e+14 164248.10

4) Claims< 175 2054 4.297311e+13 80932.12

8) Claims< 48.5 1844 5.727989e+12 39731.63 \*

9) Claims>=48.5 210 6.629248e+12 442711.60 \*

5) Claims>=175 102 5.477928e+13 1842003.00

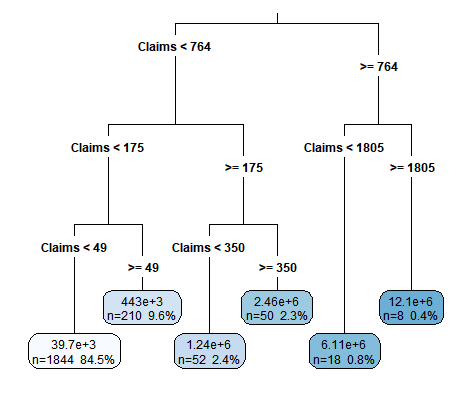
10) Claims< 350 52 3.000499e+12 1243570.00 \*

11) Claims>=350 50 1.378922e+13 2464373.00 \*

3) Claims>=763.5 26 3.010592e+14 7948916.00

6) Claims< 1804.5 18 2.814420e+13 6113205.00 \*

7) Claims>=1804.5 8 7.577968e+13 12079270.00 \*



The Predicted Payment rate is 33442