Question 1:

The committee is interested to know each field of the data collected through descriptive analysis to gain basic insights into the data set and to prepare for further analysis.

Solution: Here the committee want to see the summary of the data set. For this we need to clean the data, sort it and represent it using the Summary option to get the descriptive stat of all the variables.

Code:

getwd()

DF <- read.csv('SwedishMotorInsurance.csv', header = T, sep = ',')

str(DF)

summary(DF)

DF$Zone <- as.factor(DF$Zone)

DF$Make = as.factor(DF$Make)

DF$Bonus = as.factor(DF$Bonus)

DF$Kilometres = as.factor(DF$Kilometres)

str(DF)

head(DF)

summary(DF)

colSums(is.na(DF))

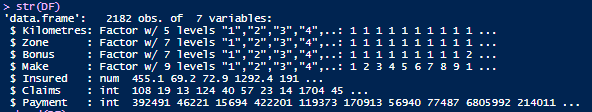
View(DF)

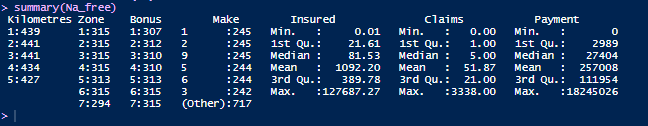
is.na(DF)

Na\_free <- na.omit(DF)

summary(Na\_free)

Output:



The two tables show the structure of the data set and also the summary statistic for each variable in the dataset. This dataset has now been sorted and manipulated and can be used for further analysis.

The results provide the minimum and maximum values. It also provides the mean and median values of all variables. From this you can understand the spread of data. We can see that claims and payment also have null or zero values, however the insured column does not have a zero value. This specifies that there are few entries where the car has been insured for a given period of time. However, no claim or payment has been made for that combination of car make, zone, and kilometers.

Question2:

The total value of payment by an insurance company is an important factor to be monitored. So, the committee has decided to find whether this payment is related to number of claims and the number of insured policy years. They also want to visualize the results for better understanding.

Solution:

Here we should run a correlation model to check if the total value of payment is related to number of claims and number of insured policies. Also, we will plot the graphs of these relationship to get a visual summary of relationships.

Code:

plot(DF$Claims,DF$Payment,ylab = 'Payment',xlab = 'Number of Claims')

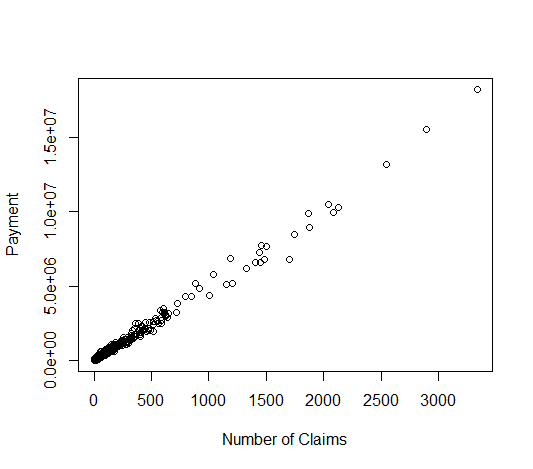
plot(DF$Insured,DF$Payment,ylab = 'Payment',xlab = 'Numnber of insured policy years')

cor(DF$Claims,DF$Payment)

cor(DF$Insured,DF$Payment)

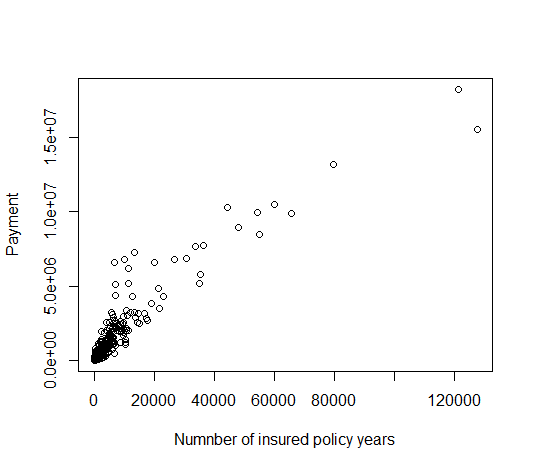
Outputs:

plot(DF$Claims,DF$Payment,ylab = 'Payment',xlab = 'Number of Claims')



**Figure1: Scatterplot Showing relationship between Payment and Number of claims**

plot(DF$Insured,DF$Payment,ylab = 'Payment',xlab = 'Numnber of insured policy years')



**Figure2: Scatterplot depicting relationship between payment and the number of insured policy years.**



**Figure3: Correlation results**

These Data show that there exists a strong positive correlation between Claims and Payment (r=0.9954) and also a strong positive correlation between Number of insured policy years and payment(r=0.9332). This means that as either Claims and Insured increases so does payment, and vice versa. This payment is related to number of claims and the number of insured policy years. the two scatterplots also show the same results for both.

Question3

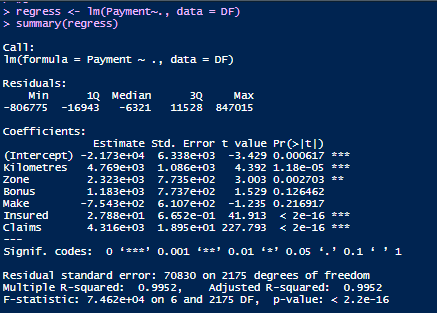
Find whether distance, location, bonus, make, and insured amount or claims are affecting the payment or all or some of these are affecting it.

Code:

regress <- lm(Payment~., data = DF)

summary(regress)

Output:



**Figure4: Summary statistics of the anova model ran to assess the impact of other variables on Payment.**

Solution:

Here we performed a regression model using payment as a dependent variable to determine the relation that exists between it and other variables.

Ho= Payment is not dependent on any other variable in the data set

Ha= payment is dependent on some variables in the data set

Figure4 shows that only Kilometers, Zone, Insured, Claims are the only variables that have an effect on payment. They all generated p values that were less than alpha (0.05). the variable, Bonus and Make, have no effect on payment. So, the null hypothesis was rejected, payment is dependent on some variables, namely the Kilometers, Zone, Insured and claims.

Question 4:

The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilometer, and bonus level their insured amount, claims, and payment get increased.

Question5:

The committee wants to understand what affects their claim rates so as to decide the right premiums for a certain set of situations. Hence, they need to find whether the insured amount, zone, kilometer, bonus, or make affects the claim rates and to what extent.

Code:

regress2 <- lm(Claims~Insured+Make+Bonus+Zone+Kilometres,data = DF)

summary(regress2)

Output:

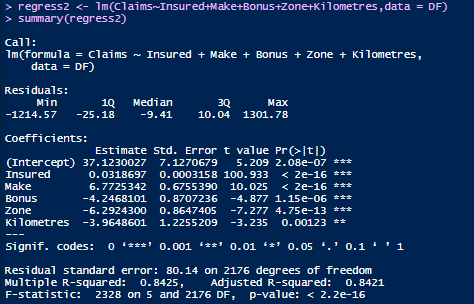


Figure5: Results of the regression model ran to determine which factors affect claims.

Figure5, show that all these factors (Insured, Make, Bonus, Zone and Kilometers) have a significant effect on Claims. So, the null hypothesis of no effect of these variables is rejected. In deciding on the right premium, the company should know that all these factors have a significant effect on Claims.