First, we have to import Insurance\_factor\_identification.csv file to the R Lab

* insu <- read.csv('Insurance\_factor\_identification.csv')

Then we have to check dimension and structure of data

* dim(insu)
* str(insu)

After that we have to convert Kilometres, Zone & Make into factors variables (because they are categorical variables but showing as integers in the data structure)

* insu$Kilometres <- as.factor(insu$Kilometres)
* insu$Zone <- as.factor(insu$Zone)
* insu$Make <- as.factor(insu$Make)

so now we are ready to do analysis for the given problem

**Analysis Tasks:**After understanding the data, you need to help the committee with the following by the use of the R tool:

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.

Answer-1

summary(insu)

Kilometres Zone Bonus Make Insured

1:439 1:315 Min. :1.000 1 :245 Min. : 0.01

2:441 2:315 1st Qu.:2.000 2 :245 1st Qu.: 21.61

3:441 3:315 Median :4.000 9 :245 Median : 81.53

4:434 4:315 Mean :4.015 5 :244 Mean : 1092.20

5:427 5:313 3rd Qu.:6.000 6 :244 3rd Qu.: 389.78

6:315 Max. :7.000 3 :242 Max. :127687.27

7:294 (Other):717

Claims Payment

Min. : 0.00 Min. : 0

1st Qu.: 1.00 1st Qu.: 2989

Median : 5.00 Median : 27404

Mean : 51.87 Mean : 257008

3rd Qu.: 21.00 3rd Qu.: 111954

Max. :3338.00 Max. :18245026

1. 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 the number of claims and the number of insured policy years. They also want to visualize the results for better understanding.

Answer-2

So to found whether this payment is related to the number of claims and the number of insured policy years.

So for finding find correlation between payment and claims and to plot it we have to use following codes

* cor(insu$Payment,insu$Claims)
* plot(insu$Payment,insu$Claims)
* ggplot(data=insu,mapping=aes(x=Claims,y=Payment))+geom\_point()

So for finding find correlation between payment and insured and to plot it we have to use following codes

* cor(insu$Payment,insu$Insured)
* plot(insu$Payment,insu$Insured)
* ggplot(data=insu,aes(x=Insured,y=Payment))+geom\_point()

1. The committee wants to figure out the reasons for insurance payment increase and decrease. So, they have decided to find whether distance, location, bonus, make, and insured amount or claims are affecting the payment or all or some of these are affecting it.

Answer-3

So first we want to split our data into test and training data, so follow the following codes for splitting our data

* sel <- sample.split(insu$Zone,SplitRatio = 0.7)
* ins\_train <- subset(insu,sel==TRUE)
* ins\_test <- subset(insu,!sel)

after splitting our data we have to find out model on which we get the variables that are significant and affect the payment increase and decrease

* payment\_model <- lm(Payment~.,data=ins\_train)
* summary(payment\_model)

Since F statistics value is very less so we reject the null hypothesis that the entire variable coefficient are zero.

so only Kilometres, Insured and Claims are significant so we only include that in our model

and that are the parameters that affect increase and decrease in payment

* payment\_model <- lm(Payment ~ Kilometres+Insured+Claims,data=ins\_train)
* summary(payment\_model)

1. The insurance company is planning to establish a new branch office, so they are interested to find at what location, kilometre, and bonus level their insured amount, claims, and payment gets increased. (Hint: Aggregate Dataset)

Answer-4

So, to find out at what location, kilometre, and bonus level their insured amount, claims, and payment gets increased

We use aggregate function to find the Insured, Claims and payment by Zone, Kilometres and bonus

* insu %>% group\_by(Zone) %>% summarise(avg\_ins=mean(Insured),avg\_claims=mean(Claims),avg\_payment=mean(Payment))

so, they have to select zone 4 for the payment to get increased.

* insu %>% group\_by(Kilometres) %>% summarise(avg\_ins=mean(Insured),avg\_claims=mean(Claims),avg\_payment=mean(Payment))

so, they have to select distance 1000-15000 cat 2 for the new branch to get increase in payment

* insu %>% group\_by(Bonus) %>% summarise(avg\_ins=mean(Insured),avg\_claims=mean(Claims),avg\_payment=mean(Payment))

so they have to select 7 years as bonus year to get increase in payment.

1. 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, kilometre, bonus, or make affects the claim rates and to what extent.

Answer-5

* claim\_model <- lm(Claims~.,data=insu)
* summary(claim\_model)

so zones are insignificant for the model so we to exclude that and in problem part it is not given to take payment so we have to exclude payment also.

* claim\_model <- lm(Claims ~ Kilometres+Bonus+Insured,data=insu)
* summary(claim\_model)

after excluding payment and zone, Kilometeres become insignificant so now we have to exclude kilometeres also.

* claim\_model <- lm(Claims ~ Bonus+Insured,data=insu)
* summary(claim\_model)

now there is no insignificant variable left and we got our solution as

claim = 35.0731814 - 4.7159993 x Bonus + 0.0327120 X Insured