Assignment

2023-02-18

**Problem**

Insurance company need to analysis the reason for the increasing claim range and factor affecting the claims.

**Attribute Descriptions**

For analysing the above issue, attributes like age, gender, BMI, Blood pressure, diabetic disorder, smoking habit, region and their claims where collected for the analysis

1. Age-age of the persons
2. Gender-male or female
3. BMI- Normal Body Mass Index (18-26)
4. Blood pressure- Normal blood pressure ranges from (80-120) mm Hg
5. Diabetic-tells the person diabetic or not
6. smoker-smoking habit about the person
7. region- region of the persons which he belongs to
8. claim- amount claimed by the persons

**Assumption**

* To check the blood pressure will affect the claim amount or not
* To analyse which region has more claim occurred
* Diabetic disorder and smoking habit will affect the BMI range
* To check the age range of the people and BMI of the people will affect the claim amount
* To check the BMI is correlated with claim amount

**Subset the data**

* By analysing the attribute of age, the majority of the people comes under the range of 29 and 47 so I have subset the particular range in the name ‘dfage’
* By analysing the attribute of region, the majority of the claims occur under Northwest and Southeast region so I subset the both region in a single subset
* Health condition is one of the important attributes for setting premium. so, by analysing the habits of the people I subset the smoking persons and non-smokers
* With respect to the above point diabetic persons are subseted into single subset

**Exploratory Data Analysis**

library(readr)  
library(lattice)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
df <- read\_csv("Insurance\_data0.csv")

## Rows: 1332 Columns: 9

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): gender, diabetic, smoker, region  
## dbl (5): PatientID, age, bmi, bloodpressure, claim  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

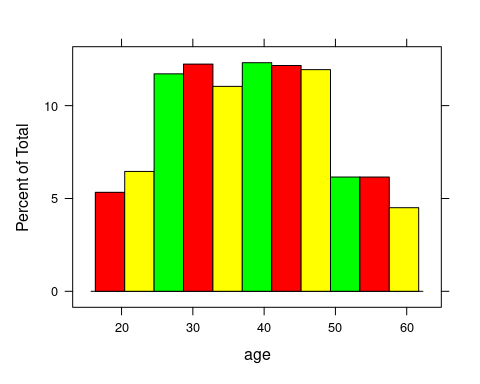
summary(df)

## PatientID age gender bmi   
## Min. : 1.0 Min. :18.00 Length:1332 Min. :16.00   
## 1st Qu.: 341.8 1st Qu.:29.00 Class :character 1st Qu.:26.20   
## Median : 674.5 Median :38.00 Mode :character Median :30.35   
## Mean : 674.5 Mean :38.09 Mean :30.66   
## 3rd Qu.:1007.2 3rd Qu.:47.00 3rd Qu.:34.73   
## Max. :1340.0 Max. :60.00 Max. :53.10   
## bloodpressure diabetic smoker region   
## Min. : 80.00 Length:1332 Length:1332 Length:1332   
## 1st Qu.: 86.00 Class :character Class :character Class :character   
## Median : 92.00 Mode :character Mode :character Mode :character   
## Mean : 94.19   
## 3rd Qu.: 99.00   
## Max. :140.00   
## claim   
## Min. : 1122   
## 1st Qu.: 4760   
## Median : 9413   
## Mean :13325   
## 3rd Qu.:16781   
## Max. :63770

str(df)

## spc\_tbl\_ [1,332 × 9] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ PatientID : num [1:1332] 1 2 8 9 10 11 12 13 17 18 ...  
## $ age : num [1:1332] 39 24 19 20 30 36 37 19 35 41 ...  
## $ gender : chr [1:1332] "male" "male" "male" "male" ...  
## $ bmi : num [1:1332] 23.2 30.1 41.1 43 53.1 19.8 20.3 20.7 34.1 34.4 ...  
## $ bloodpressure: num [1:1332] 91 87 100 86 97 88 90 81 90 84 ...  
## $ diabetic : chr [1:1332] "Yes" "No" "No" "No" ...  
## $ smoker : chr [1:1332] "No" "No" "No" "No" ...  
## $ region : chr [1:1332] "southeast" "southeast" "northwest" "northwest" ...  
## $ claim : num [1:1332] 1122 1132 1147 1149 1163 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. PatientID = col\_double(),  
## .. age = col\_double(),  
## .. gender = col\_character(),  
## .. bmi = col\_double(),  
## .. bloodpressure = col\_double(),  
## .. diabetic = col\_character(),  
## .. smoker = col\_character(),  
## .. region = col\_character(),  
## .. claim = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

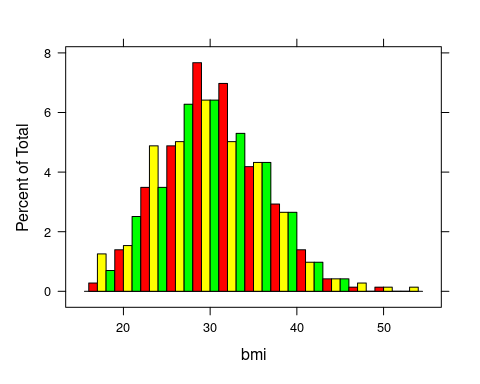
histogram(~age,data = df,col=c('red','yellow','green'))



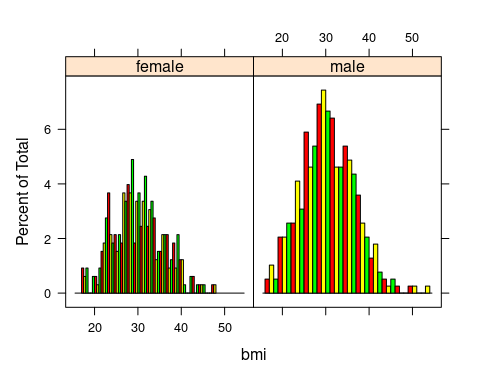
summary(df$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 29.00 38.00 38.09 47.00 60.00

dfage=filter(df,age >=29 & age <=47)  
  
histogram(~bmi,data = dfage,breaks = 50,col=c('red','yellow','green'))



histogram(~bmi|gender,data = dfage,breaks = 50,col=c('red','yellow','green'))



summary(df$bmi)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 26.20 30.35 30.66 34.73 53.10

summary(df$bmi & df$gender=="male")

## Mode FALSE TRUE   
## logical 662 670

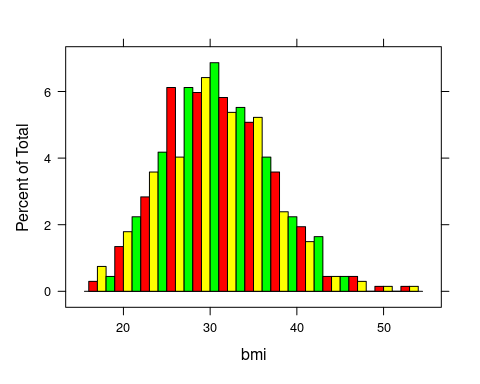
dfm=subset(df,gender=="male")  
dffm=subset(df,gender=="female")  
summary(dfm$bmi)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 26.40 30.70 30.93 35.17 53.10

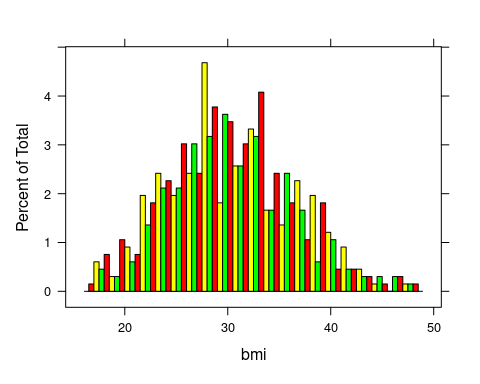
summary(dffm$bmi)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.80 26.10 30.10 30.38 34.30 48.10

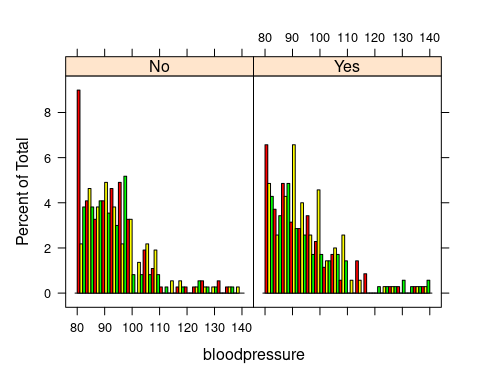
histogram(~bmi,data = dfm,breaks = 50,col=c('red','yellow','green'))



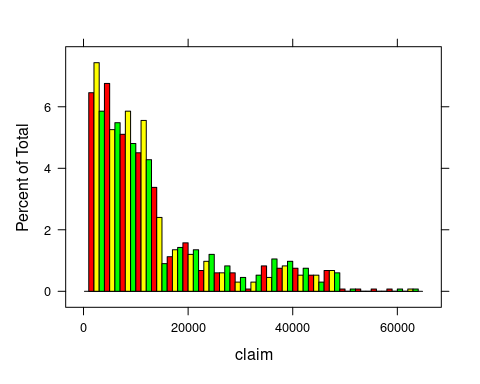
histogram(~bmi,data = dffm,breaks = 50,col=c('red','yellow','green'))



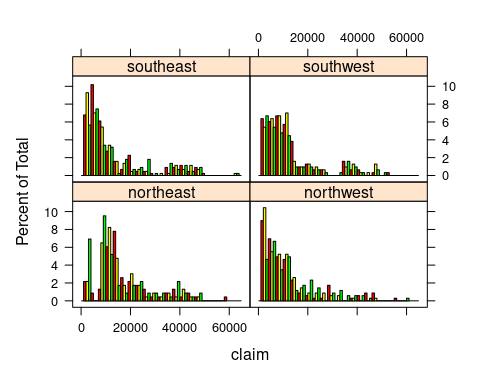
histogram(~bloodpressure|diabetic,data = dfage,breaks = 50,col=c('red','yellow','green'))



histogram(~claim,data = df,breaks = 50,col=c('red','yellow','green'))



histogram(~claim|region,data =df,breaks = 50,col=c('red','yellow','green'))



summary(df$claim & df$region=="northeast")

## Mode FALSE TRUE   
## logical 1101 231

summary(df$claim & df$region=="northwest")

## Mode FALSE TRUE   
## logical 987 345

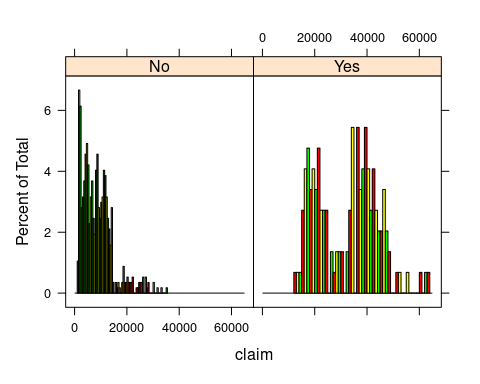
summary(df$claim & df$region=="southeast")

## Mode FALSE TRUE   
## logical 890 442

summary(df$claim & df$region=="southwest")

## Mode FALSE TRUE   
## logical 1018 314

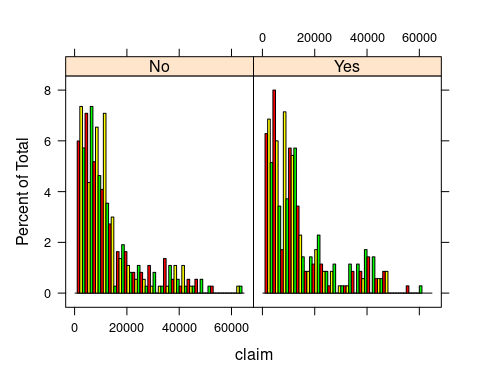
dfreg=subset(df,region =='northwest'|region =='southeast')  
  
histogram(~claim|smoker,data = dfage,breaks = 50,col=c('red','yellow','green'))



summary(dfage$claim & dfage$smoker =="Yes")

## Mode FALSE TRUE   
## logical 570 147

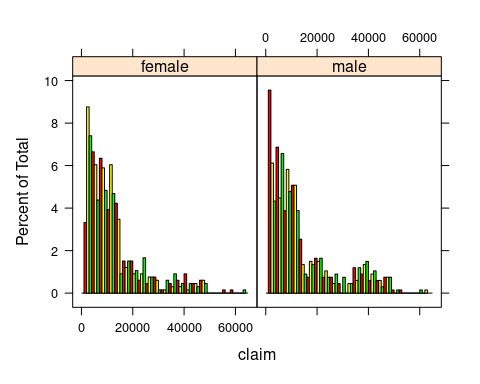
dfsmoker=subset(dfage,smoker=="Yes")  
  
histogram(~claim|diabetic,data = dfage,breaks = 50,col=c('red','yellow','green'))



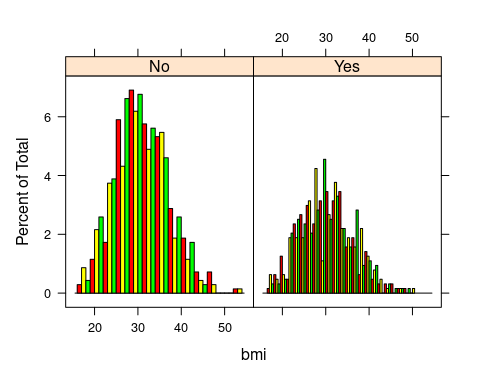
summary(dfage$claim & dfage$diabetic =="Yes")

## Mode FALSE TRUE   
## logical 367 350

dfsugar=subset(dfage,diabetic=="Yes")  
  
histogram(~claim|gender,data = df,breaks = 50,col=c('red','yellow','green'))



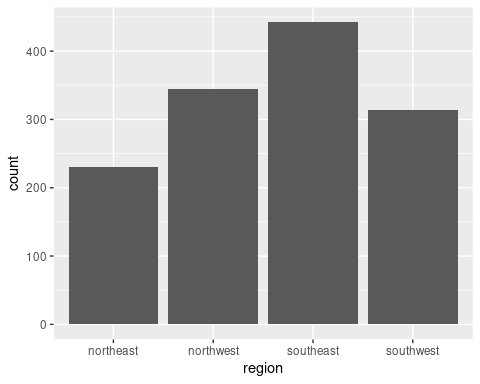
histogram(~bmi|diabetic,data =df,breaks = 50,col=c('red','yellow','green'))



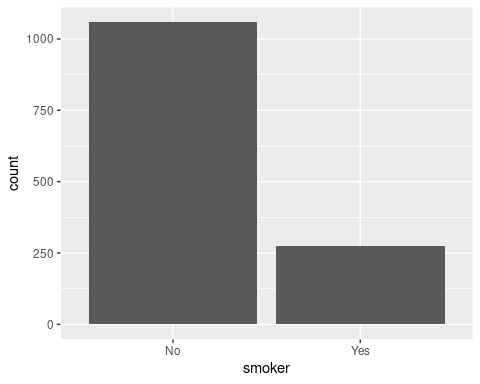
df %>% ggplot(aes(gender))+geom\_bar(stat = "count")



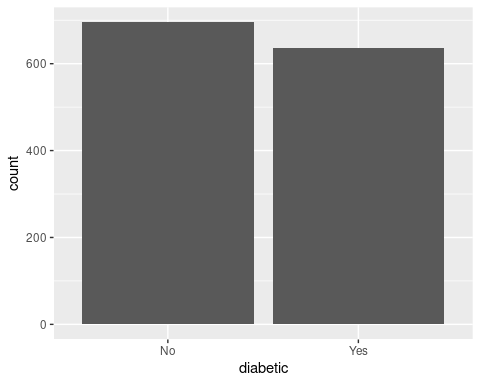
df %>% ggplot(aes(region))+geom\_bar(stat = "count")



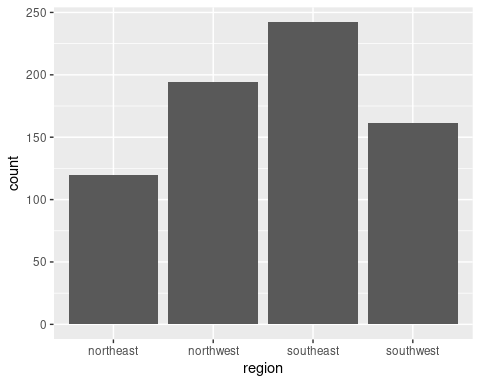
df %>% ggplot(aes(smoker))+geom\_bar(stat = "count")



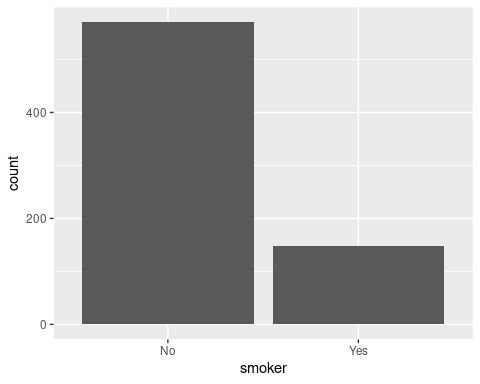
df %>% ggplot(aes(diabetic))+geom\_bar(stat = "count")



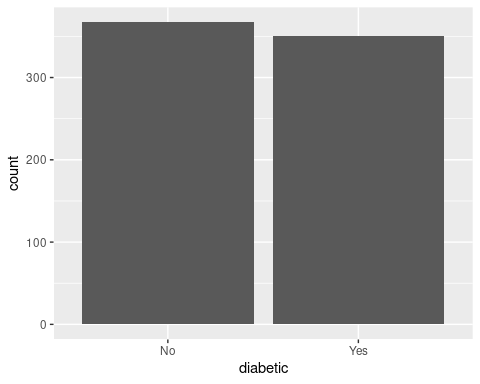
dfage %>% ggplot(aes(region))+geom\_bar(stat = "count")



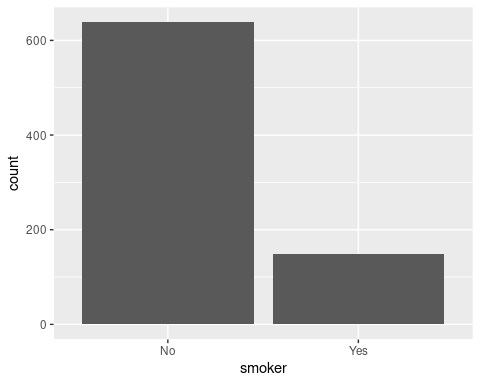
dfage %>% ggplot(aes(smoker))+geom\_bar(stat = "count")



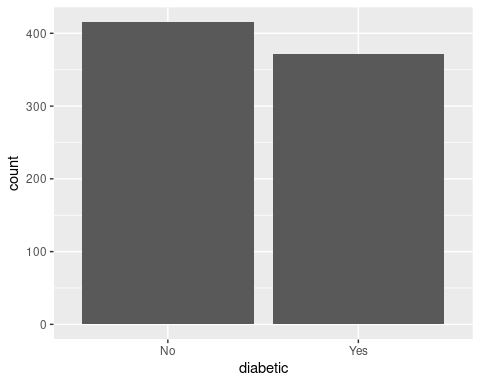
dfage %>% ggplot(aes(diabetic))+geom\_bar(stat = "count")



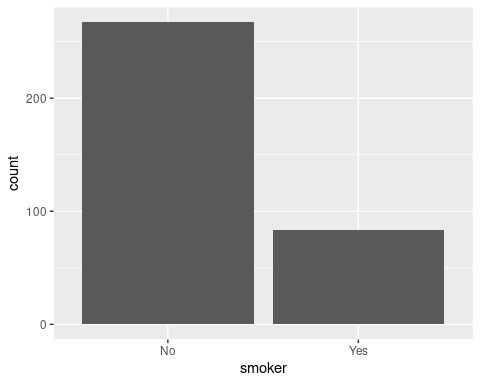
dfreg %>% ggplot(aes(smoker))+geom\_bar(stat = "count")



dfreg %>% ggplot(aes(diabetic))+geom\_bar(stat = "count")



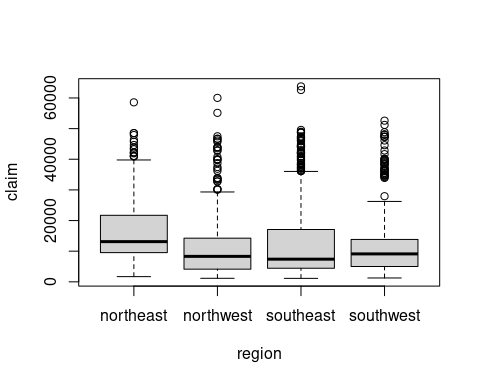
dfsugar %>% ggplot(aes(smoker))+geom\_bar(stat = "count")



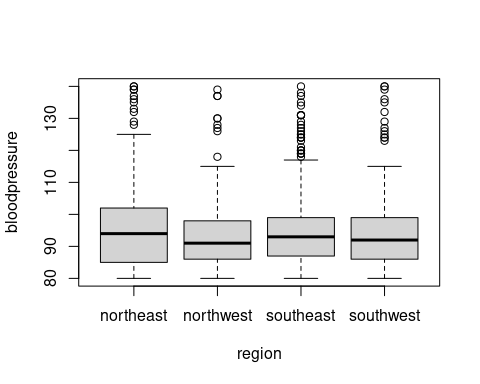
summary(df$claim)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1122 4760 9413 13325 16781 63770

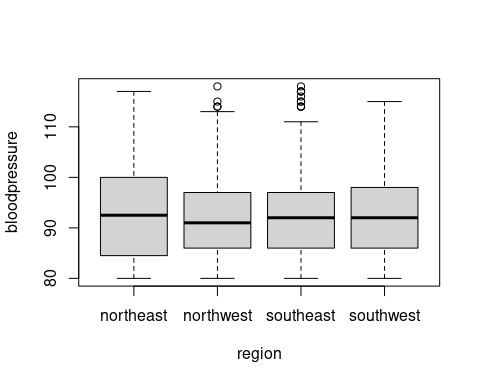
boxplot(claim~region,data = df)



boxplot(bloodpressure~region,data = df)



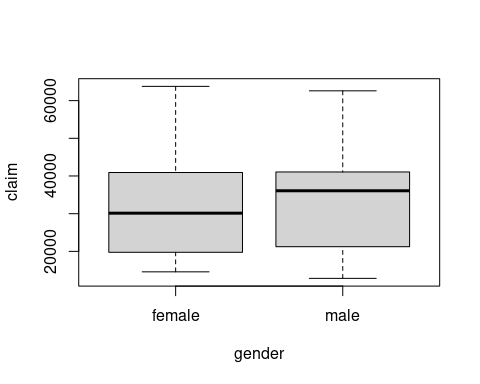
Q1 <- quantile(df$bloodpressure, .25)  
Q3 <- quantile(df$bloodpressure, .75)  
IQR <- IQR(df$bloodpressure)  
df2 <- subset(df, df$bloodpressure > (Q1 - 1.5\*IQR) & df$bloodpressure < (Q3 + 1.5\*IQR))  
boxplot(bloodpressure~region,data = df2)



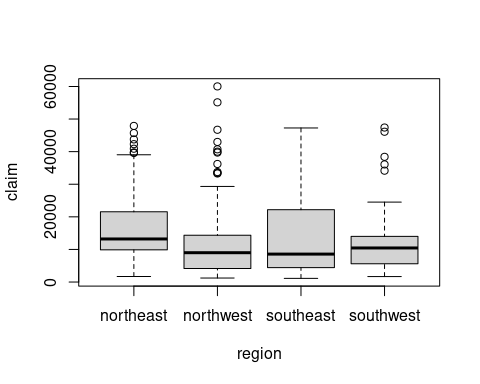
boxplot(claim~gender,data = dfreg)



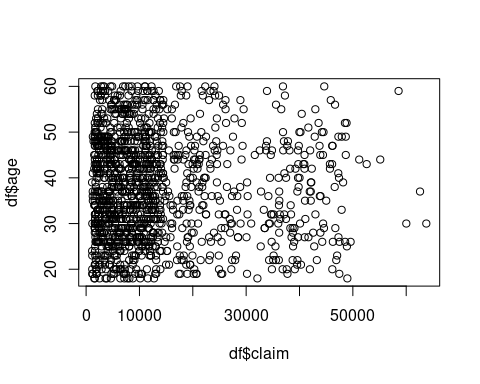
boxplot(claim~gender,data = dfsmoker)



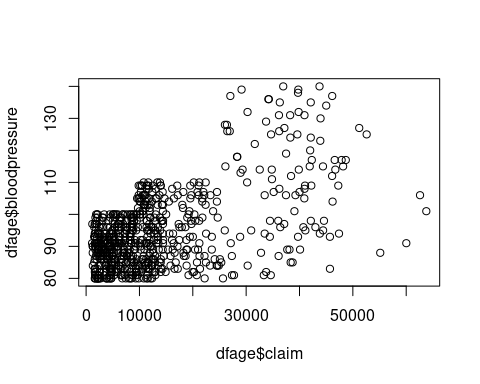
boxplot(claim~region,data = dfsugar)



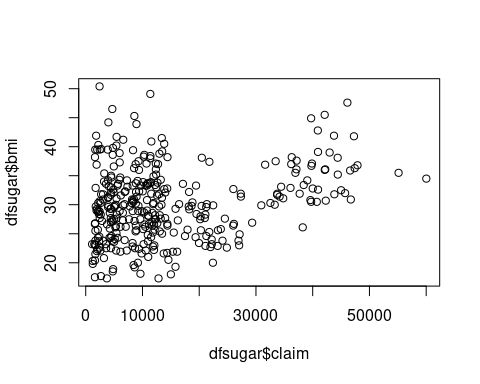
plot(x=df$claim,y=df$age)



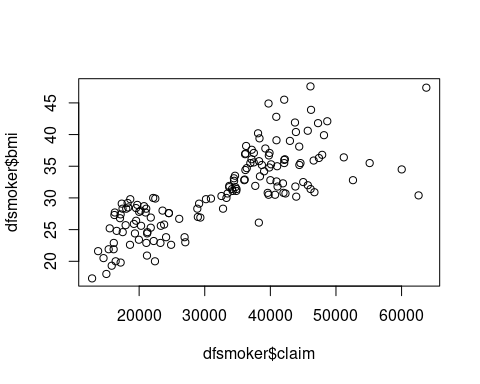
plot(x=dfage$claim,y=dfage$bloodpressure)



plot(x=dfsugar$claim,y=dfsugar$bmi)



plot(x=dfsmoker$claim,y=dfsmoker$bmi)



**Insight**

* The age falls on normal distribution and maximum number of people insured in age between 29 and 47 +
* Here Body mass index (BMI) is also normal distribution where means and median is almost equal. Majority of the people falls under normal BMI range
* The majority of the male falls on normal BMI range when compared to female policy holder BMI range
* The maximum number of claims was occurred in South east and North west and the maximum number of claims occurred between 10,000 range
* When analysing the age range between 29 and 47 the maximum number of smoking people has a average claim between 20000 to 45000, where no smoking people has less than 15000
* And also, diabetic people has also large number of claims in range 40000 when compare to non-diabetic people
* The health condition which includes diabetic disorder is affecting the BMI range of the people and also smoking habits also affecting the BMI range
* The BMI range of the people and the claims occurred for them is moderately correlated.

**Inference**

Here Diabetic disorder and Blood Pressure are affecting BMI range. The affected BMI range affects the claim amount. So, the company should increase the premium for the smokers. And the normal people having a minimal claim amount compare to others.