web scraping

krishna

18 March 2018

To find the number of missing values

```
colSums(is.na(ins))
```

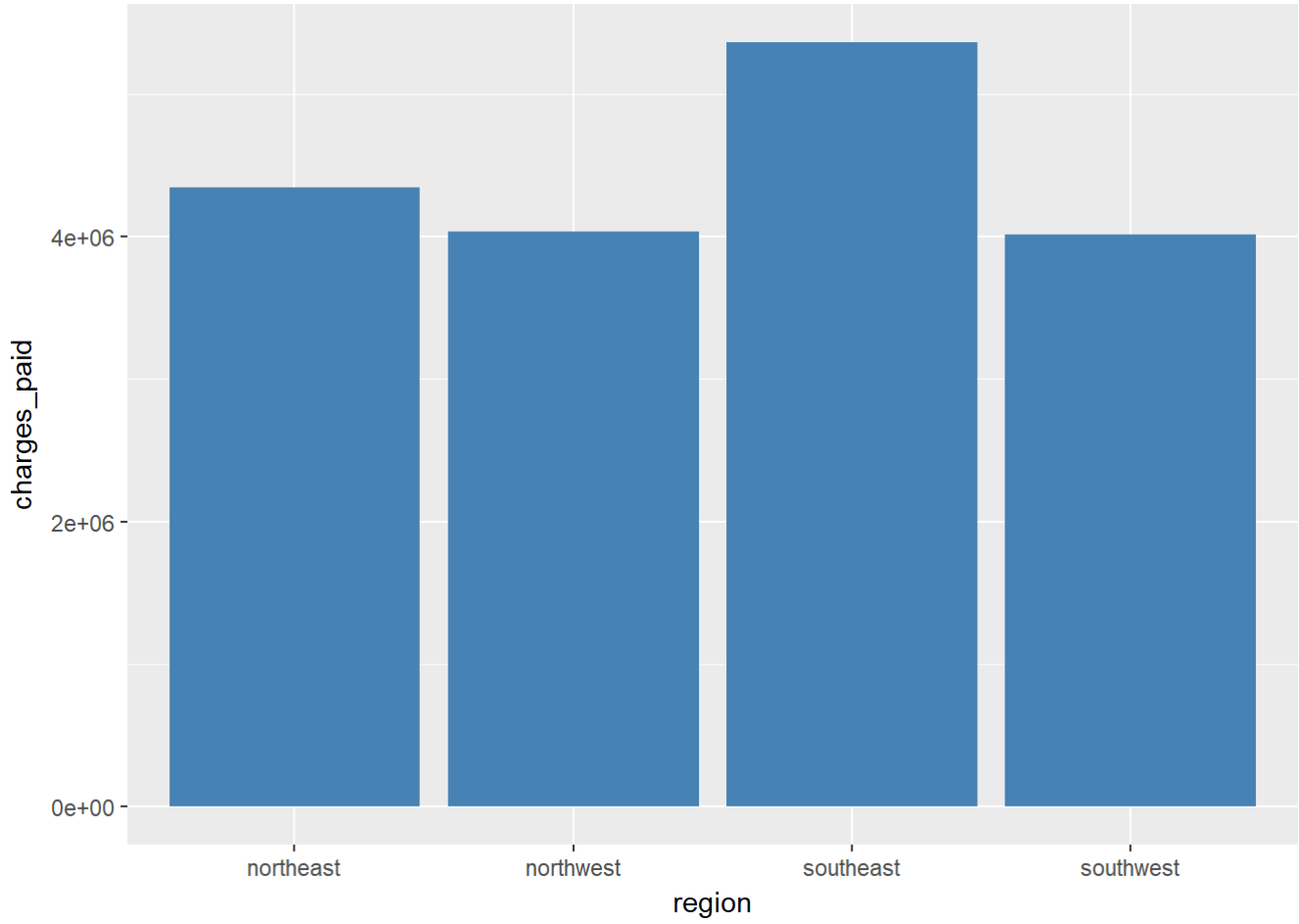
```
##      age      sex      bmi children   smoker   region  charges
##        0        0        0        0        0        0        0
```

```
str(ins)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

insurance claims is maximum in which region

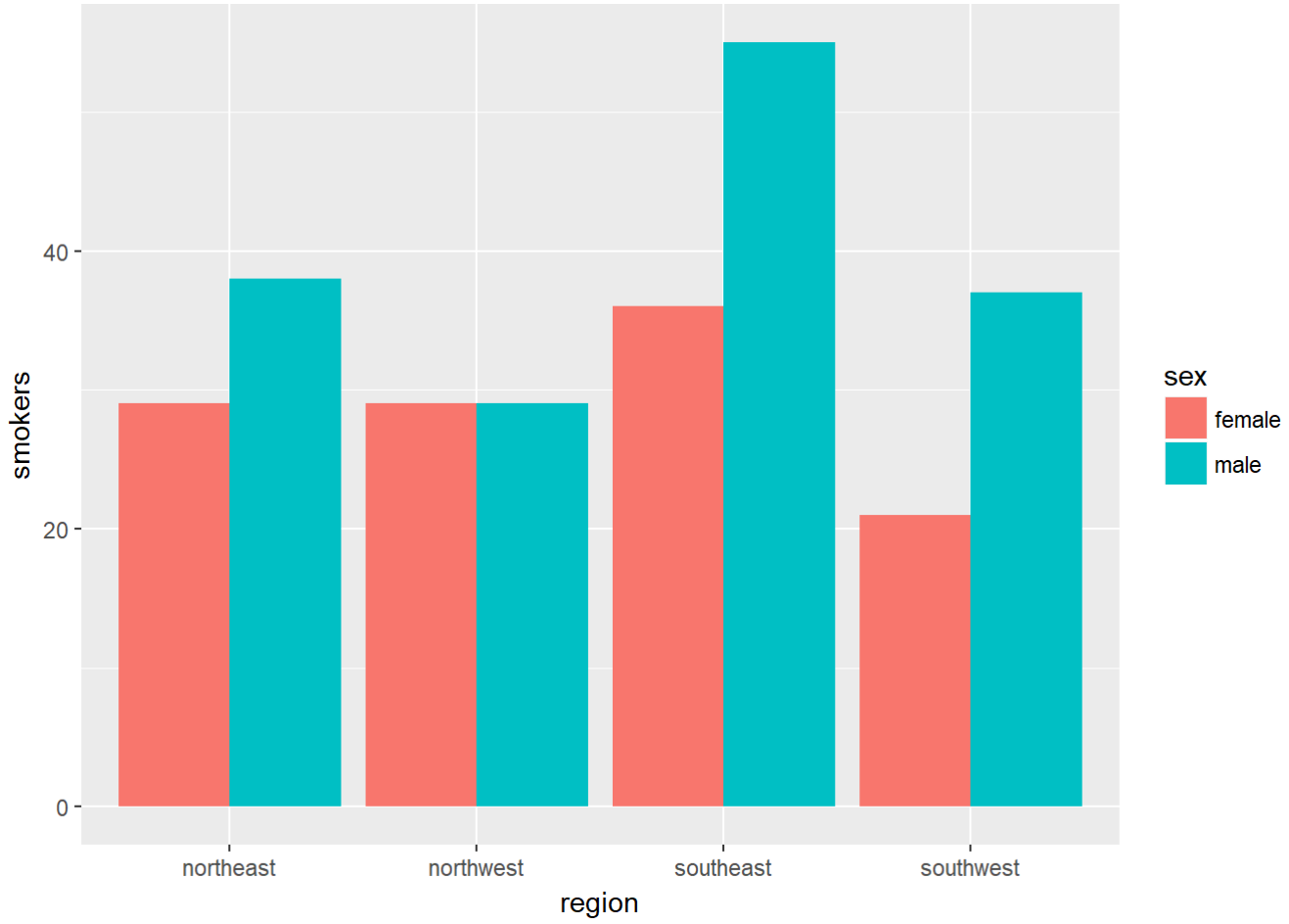
```
ins %>% group_by(region) %>% summarise(charges_paid=sum(charges)) %>% ggplot(aes(x=region,y=c
harges_paid))+geom_bar(stat = "identity",fill="steel blue")
```



smokers

```
ins %>% group_by(sex,region) %>% summarise(smokers=sum(smoker=="yes")) %>% ggplot(aes(x=regio
n,y=smokers,fill=sex))+geom_bar(stat = "identity",position = "dodge")
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.3
```

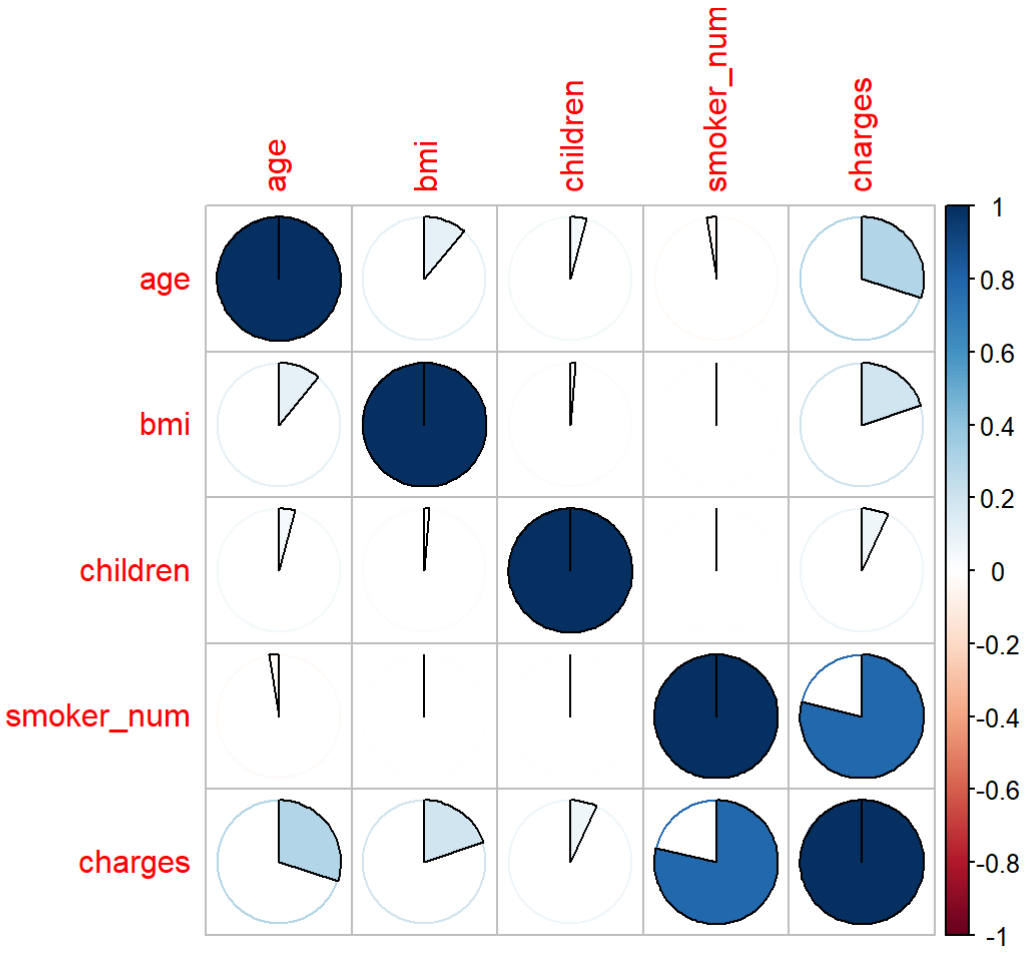


converting yes and no into numeric

```
ins$smoker_num=ifelse(ins$smoker=="yes",1,0)
ins$smoker_num=as.factor(ins$smoker_num)
```

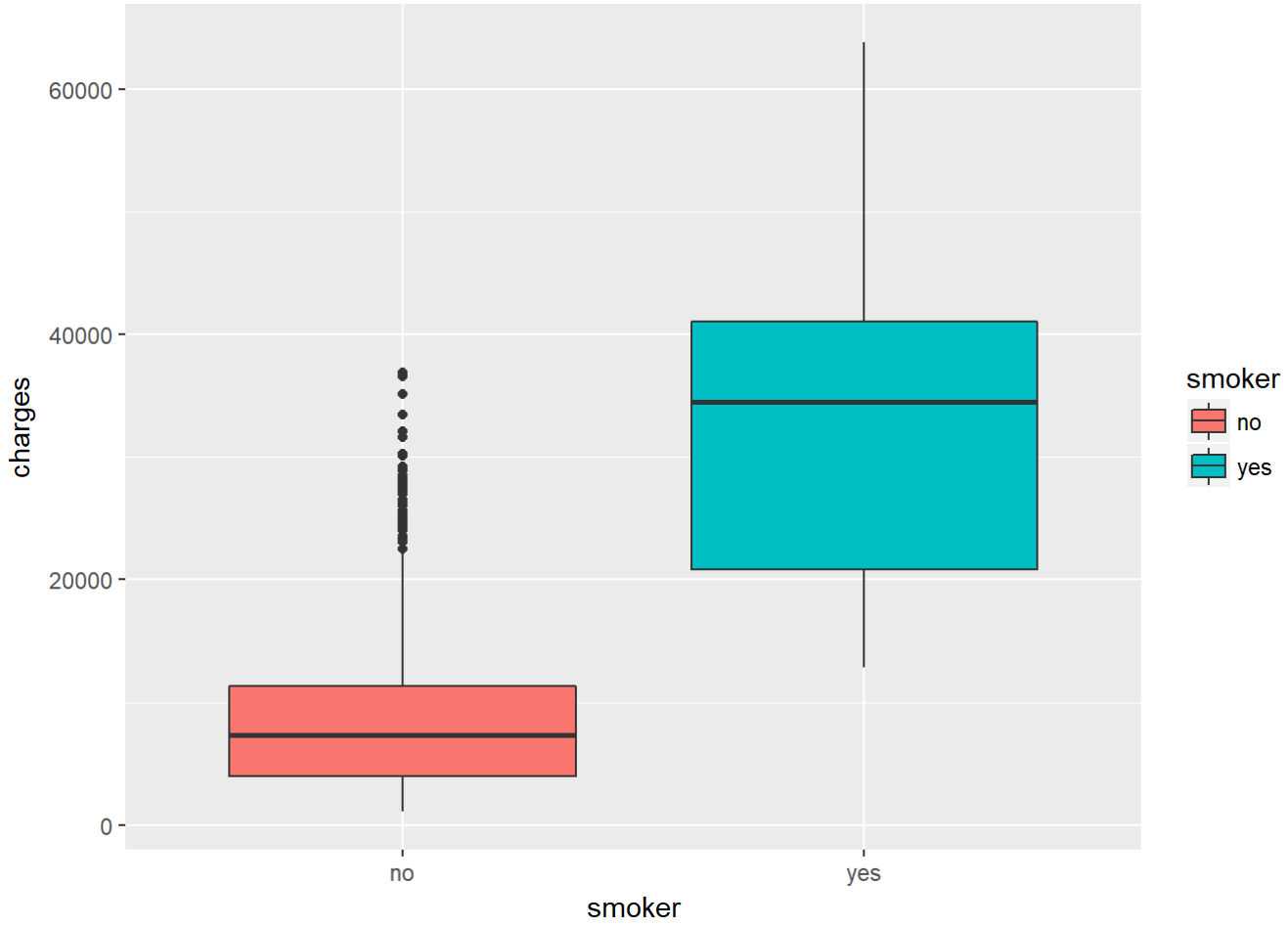
correlation between different categories

```
ins1=ins %>% select(-sex,-region,-smoker)
ins1 = ins %>% select(age, bmi, children, smoker_num, charges)
ins1$smoker_num=as.numeric(ins1$smoker_num)
corrplot(cor(ins1), method= "pie")
```



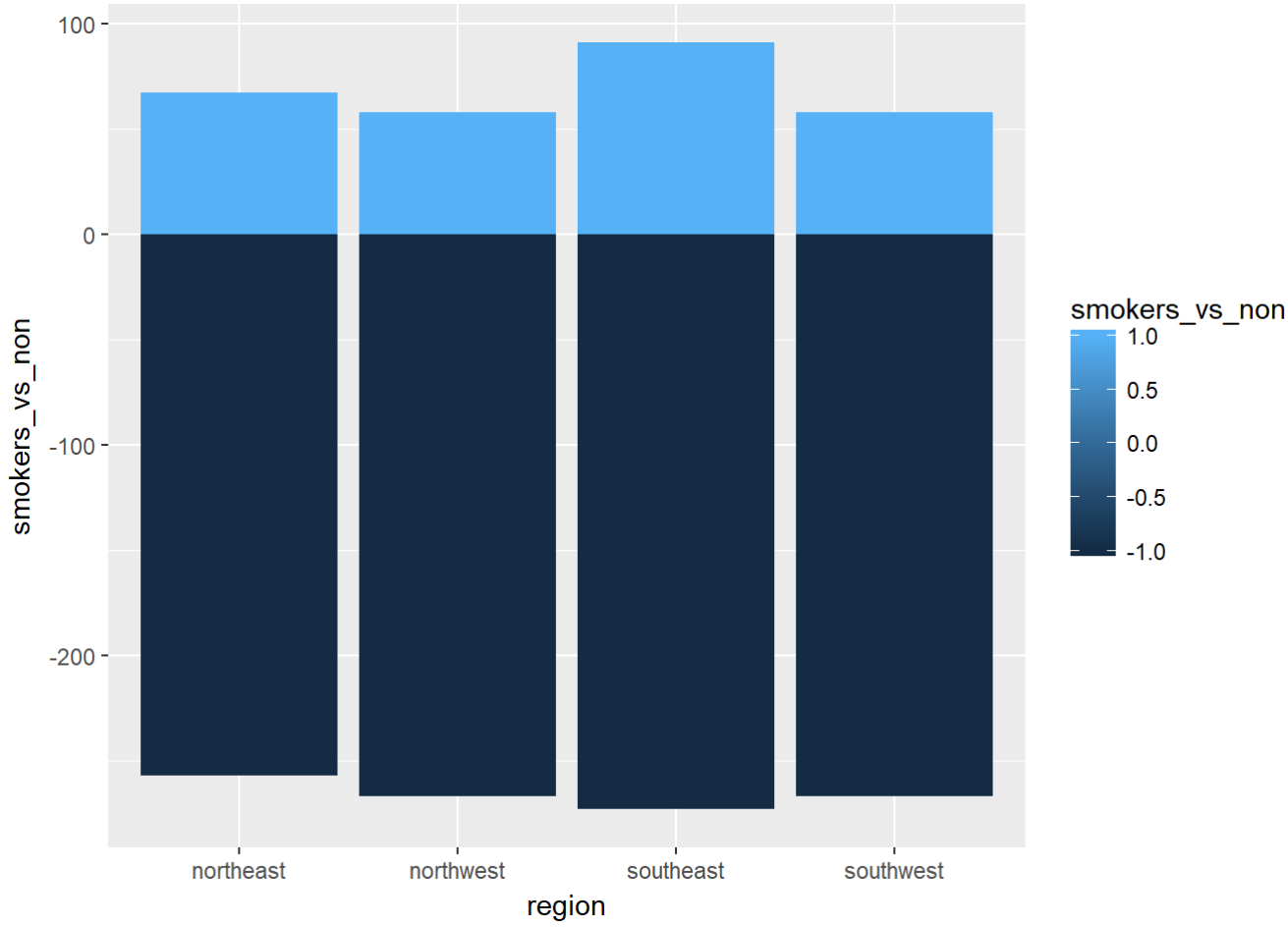
smoker vs non smoker

```
ggplot(data=ins,aes(x=smoker,y=charges,fill=smoker)) + geom_boxplot()
```



concentration of smokers vs non smokers by region

```
ins$smokers_vs_non=ifelse(ins$smoker_num==0,-1,1)
ggplot(data=ins,aes(x=region,y=smokers_vs_non))+geom_bar(stat='identity', aes(fill=smokers_vs_non))
```

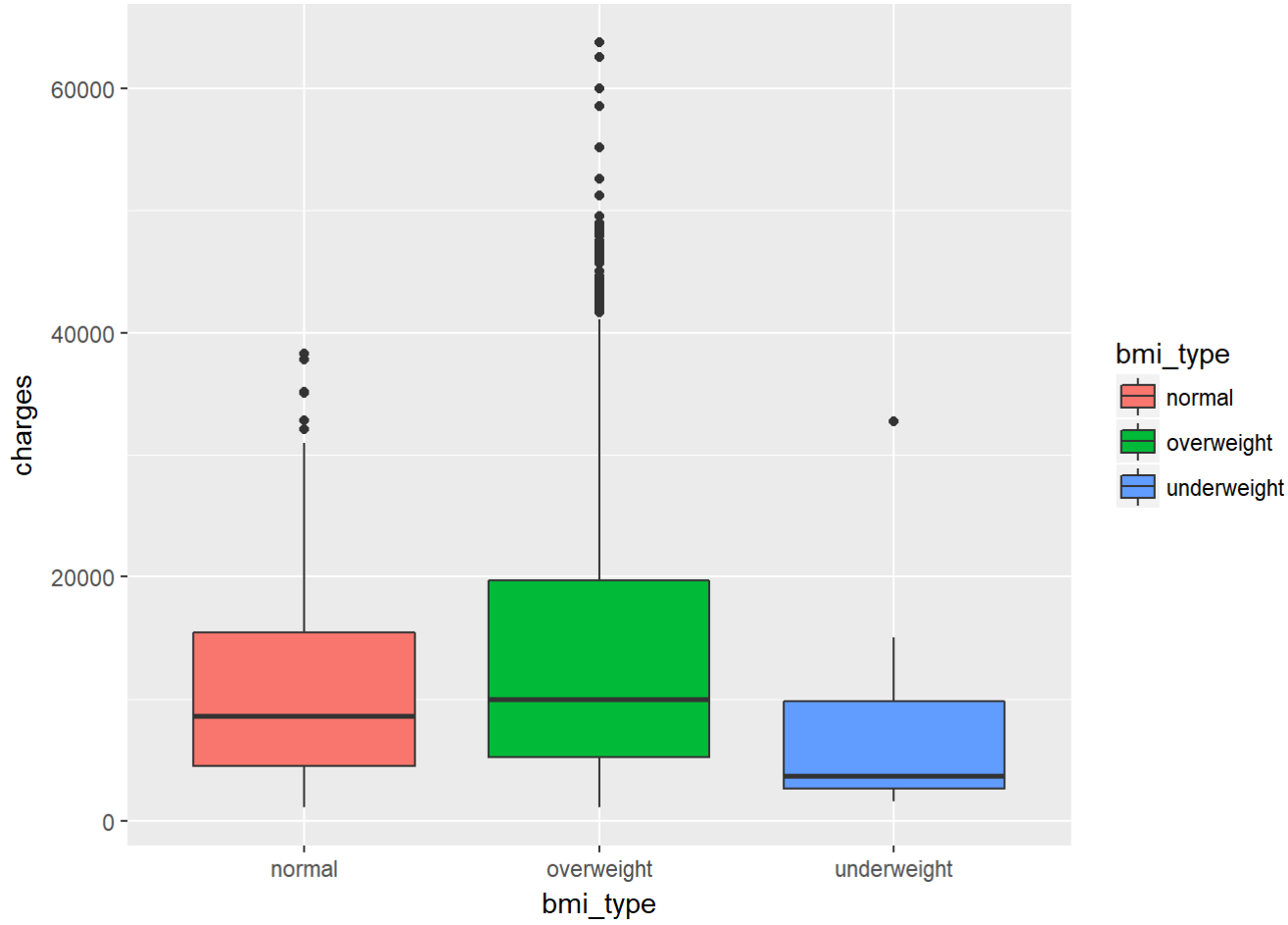


charges by different age group

```
ins %>% group_by(sex,age) %>% summarise(total=sum(charges)) %>% ggplot(aes(x=age,y=total,col=
sex))+geom_jitter(aes(size=total))
```

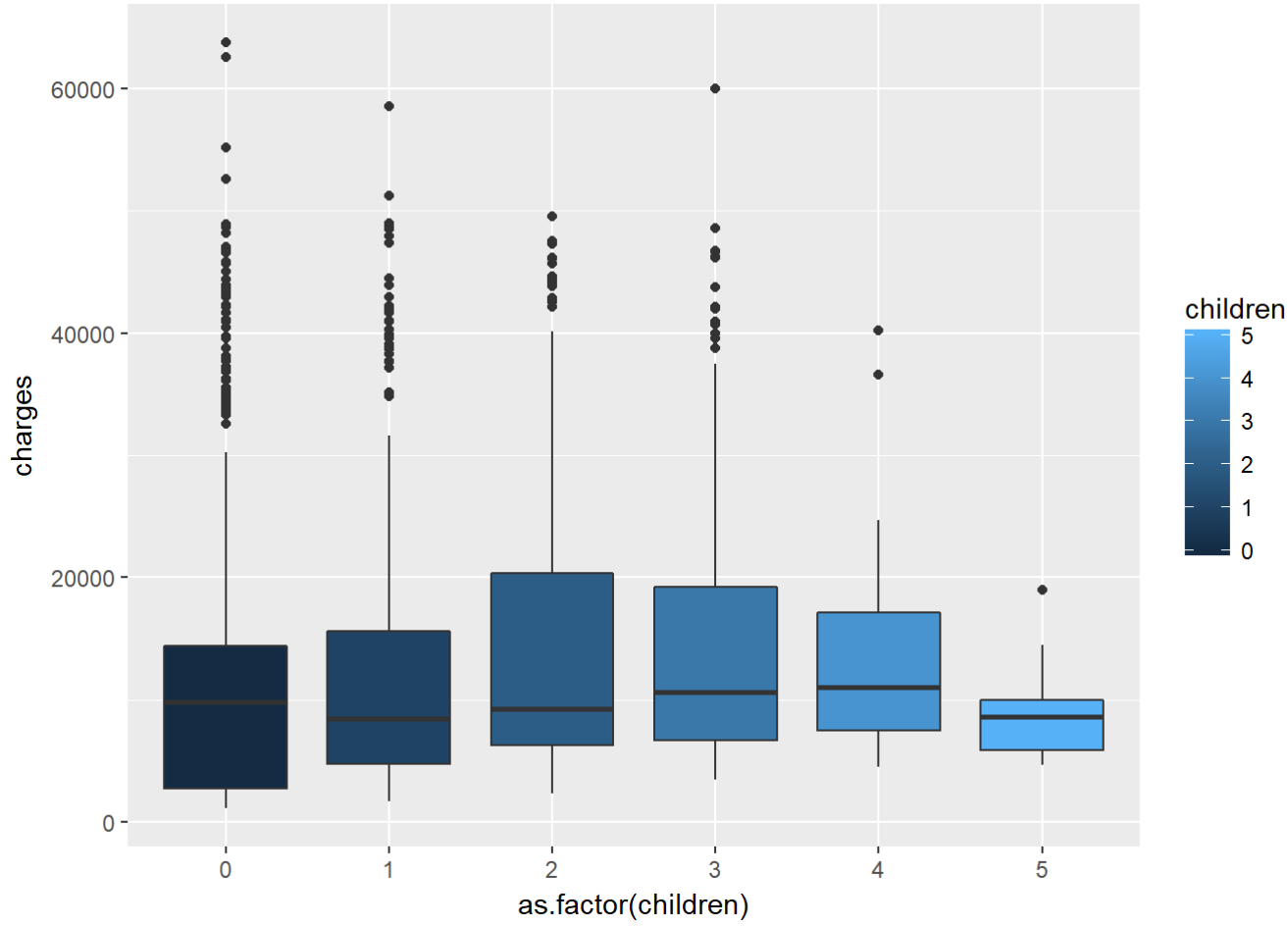
charges based on bmi

```
for (i in 1:nrow(ins)) {
  if(ins$bmi[i]<18){
  ins$bmi_type[i] = "underweight"
}else if(ins$bmi[i]>30){
      ins$bmi_type[i] = "overweight"
}else{
       ins$bmi_type[i] = "normal"
}
}

ggplot(ins,aes(x=bmi_type,y=charges))+geom_boxplot(aes(fill=bmi_type))
```



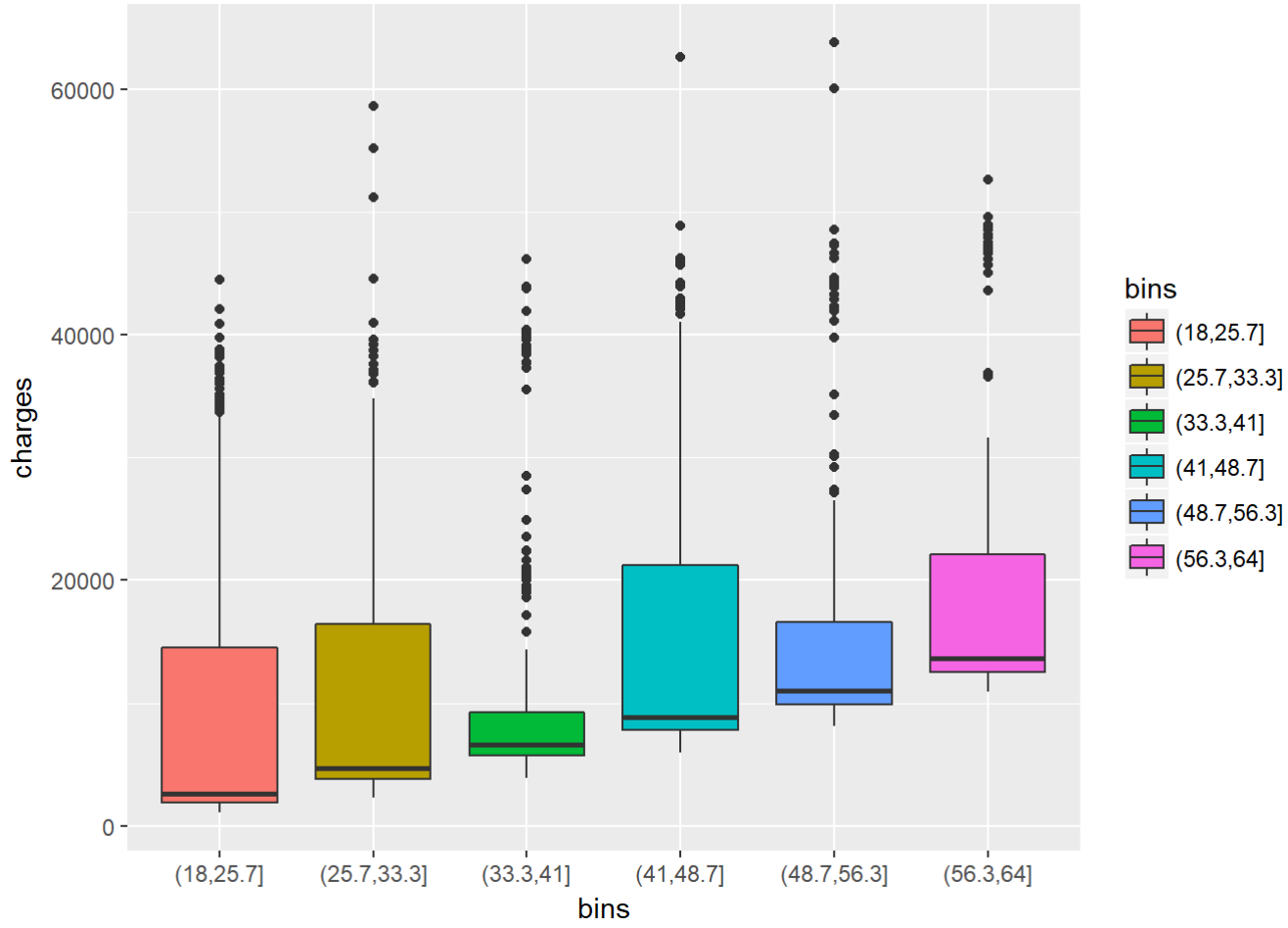
charges based on no of children

```
ggplot(data = ins,aes(x=as.factor(children),y=charges))+geom_boxplot(aes(fill=children))
```

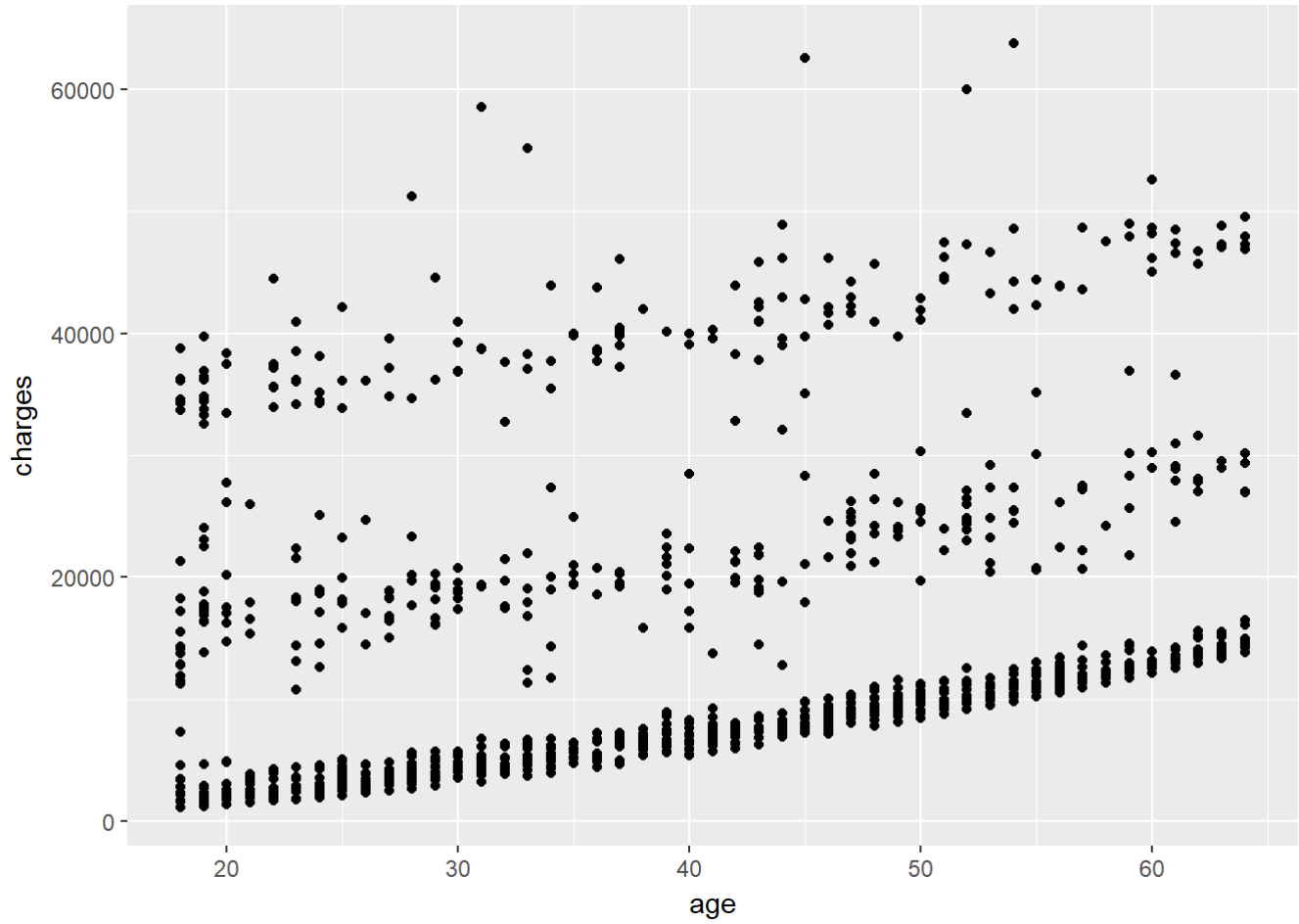


###charges based on age

```
bins=cut(ins$age,breaks = 6)
ggplot(ins,aes(x=bins,y=charges))+geom_boxplot(aes(fill=bins))
```

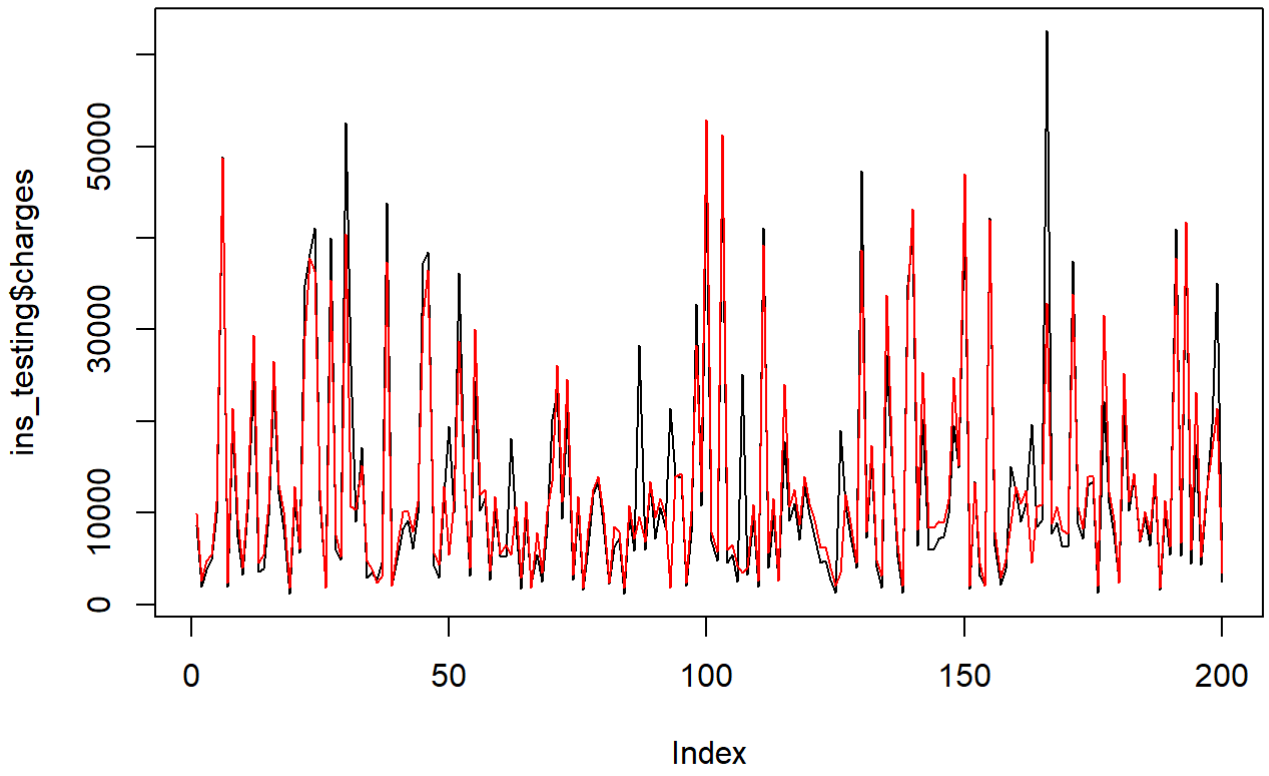



```
ins %>% select(age,charges) %>% ggplot(aes(x=age,y=charges))+geom_point()
```



Linear regression

```
ins$smoker_num=ifelse(ins$smoker=="yes",1,0)
ins1=ins %>% select(-sex,-region,-smoker)
ins1 = ins %>% select(age, bmi, children, smoker_num, charges)

ins1$child_cat = as.factor(ins1$children)

ins2 = ins1
ins_training=ins2[sample(1:1338,0.85*nrow(ins2)),]
ins_testing=ins2[sample(1:1338,0.15*nrow(ins2)),]


m1 = lm(charges~., data=ins_training %>% filter(smoker_num==0 ) %>% select(-smoker_num))
m2 = lm(charges~., data=ins_training %>% filter(smoker_num==1 ) %>% select(-smoker_num))


ins_testing$pred_hybrid = ifelse(ins_testing$smoker_num==0 ,
                          m1$coefficients[1]+
                          m1$coefficients[2]*ins_testing$age+
                          m1$coefficients[3]*ins_testing$bmi +
                          m1$coefficients[4]*ins_testing$children ,

                          m2$coefficients[1]+
                          m2$coefficients[2]*ins_testing$age+
                          m2$coefficients[3]*ins_testing$bmi +
                          m2$coefficients[4]*ins_testing$children)

{{plot(ins_testing$charges, type='l')
  lines(ins_testing$pred_hybrid, col='red')}}
```

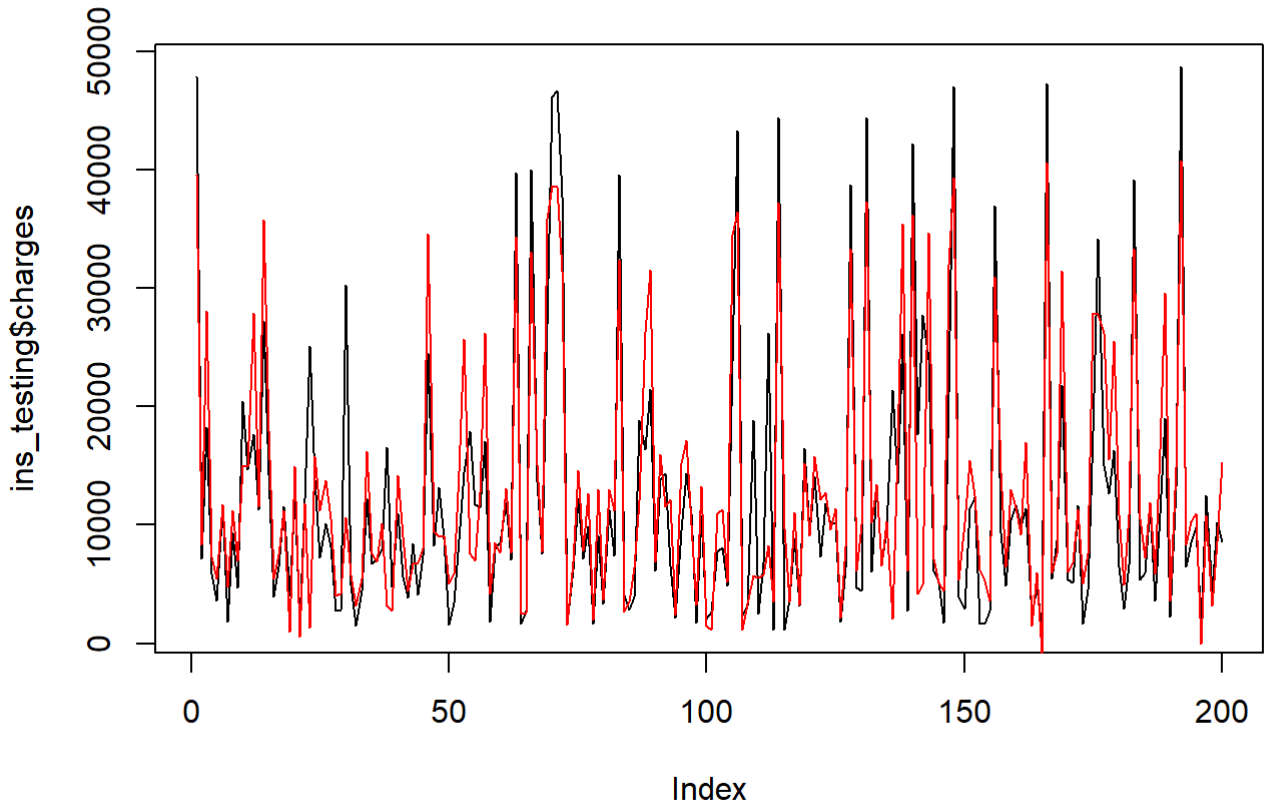



```
RMSE(ins_testing$charges, ins_testing$pred_hybrid)
```

```
## [1] 4785.783
```

Linear regression with scaled data

```
ins2 = ins1

s_dev=sd(ins2$charges)
mean_val=mean(ins2$charges)

ins2$bmi = scale(ins2$bmi)
ins2$charges=scale(ins2$charges)
ins_training=ins2[sample(1:1338,0.85*nrow(ins2)),]
ins_testing=ins2[sample(1:1338,0.15*nrow(ins2)),]

linear_model=lm(charges~.,data=ins_training)

ins_testing$pred_ins=predict(linear_model,ins_testing)

ins_testing$charges=(ins_testing$charges*s_dev)+mean_val
ins_testing$pred_ins=(ins_testing$pred_ins*s_dev)+mean_val

{{plot(ins_testing$charges,type = "l")
  lines(ins_testing$pred_ins,type = "l",col ="red")
}}
```

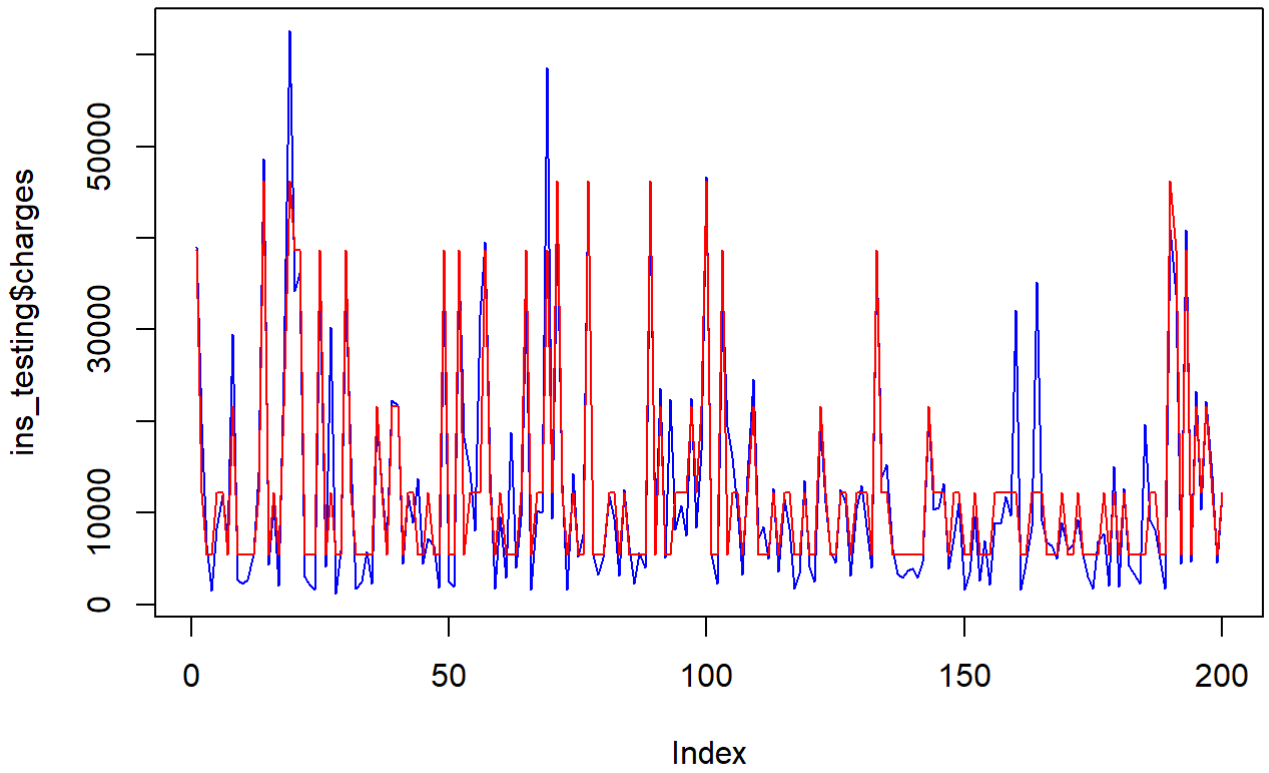



```
RMSE(ins_testing$charges, ins_testing$pred_ins)
```

```
## [1] 5728.071
```

Decision tree to identify the factor which is affecting the most

```
ins2 = ins1
ins_training=ins2[sample(1:1338,0.85*nrow(ins2)),]
ins_testing=ins2[sample(1:1338,0.15*nrow(ins2)),]

mod=tree(charges~.,data=ins_training)

ins_testing$pred=predict(mod,ins_testing)

{{plot(ins_testing$charges,type = "l",col = "blue")
lines(ins_testing$pred, type = "l",col ="red")}}
```

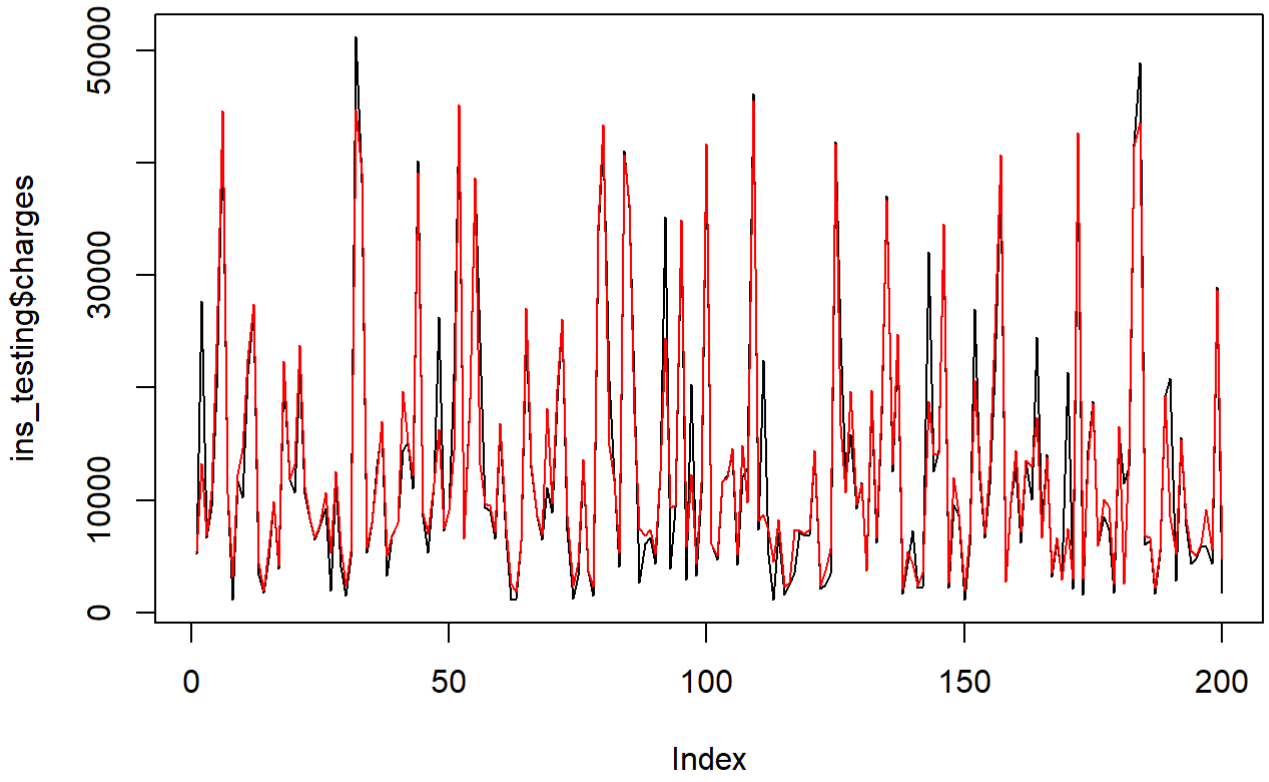



```
RMSE(ins_testing$pred,ins_testing$charges)
```

```
## [1] 4858.43
```

Random forest

```
ins2 = ins1
ins_training=ins2[sample(1:1338,0.85*nrow(ins2)),]
ins_testing=ins2[sample(1:1338,0.15*nrow(ins2)),]

mod=randomForest(charges~.,data=ins_training,ntree=500,mtry =3)

ins_testing$pred=predict(mod,ins_testing)


{{plot(ins_testing$charges,type = "l")
  lines(ins_testing$pred,type = "l",col ="red")
}}
```




```
RMSE(ins_testing$pred,ins_testing$charges)
```

```
## [1] 3263.488
```