HealthCareCost-Linear\_Regression-.R

user

2021-09-27

# Health Care Cost w/ Linear Regression  
  
# Understanding the data  
# Age:age of primary beneficiary  
  
# Sex: gender, [female, male]  
  
# BMI: Body mass index, providing an understanding of body, weights that are relatively  
# high or low relative to height, objective index of body weight (kg / m ^ 2) using the   
# ratio of height to weight, ideally 18.5 to 24.9  
  
# Children: number of children   
  
# Smoker: smoking, [yes, no]  
  
# Region: the beneficiary's residential area in the US, [northeast, southeast, southwest, northwest]  
  
# Charges: Individual medical costs billed by health insurance, $ #predicted value  
   
   
# load required libraries  
library(ggplot2)  
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(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.1.1

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(psych)

## Warning: package 'psych' was built under R version 4.1.1

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.1.1

## corrplot 0.90 loaded

# Load the dataset  
d<-read.csv("D:/Rprograms/insurance.csv")  
#print(d)  
  
  
# Print head   
print(head(d))

## age sex bmi children smoker region charges  
## 1 19 female 27.900 0 yes southwest 16884.924  
## 2 18 male 33.770 1 no southeast 1725.552  
## 3 28 male 33.000 3 no southeast 4449.462  
## 4 33 male 22.705 0 no northwest 21984.471  
## 5 32 male 28.880 0 no northwest 3866.855  
## 6 31 female 25.740 0 no southeast 3756.622

# Print tail  
print(tail(d))

## age sex bmi children smoker region charges  
## 1333 52 female 44.70 3 no southwest 11411.685  
## 1334 50 male 30.97 3 no northwest 10600.548  
## 1335 18 female 31.92 0 no northeast 2205.981  
## 1336 18 female 36.85 0 no southeast 1629.833  
## 1337 21 female 25.80 0 no southwest 2007.945  
## 1338 61 female 29.07 0 yes northwest 29141.360

# To View the contents in the dataset  
View(d)  
  
  
# To print column names  
print(colnames(d))

## [1] "age" "sex" "bmi" "children" "smoker" "region" "charges"

# Dimention of data  
print(dim(d))

## [1] 1338 7

# Print Statistical summary  
describe(d)

## vars n mean sd median trimmed mad min max  
## age 1 1338 39.21 14.05 39.00 39.01 17.79 18.00 64.00  
## sex\* 2 1338 1.51 0.50 2.00 1.51 0.00 1.00 2.00  
## bmi 3 1338 30.66 6.10 30.40 30.50 6.20 15.96 53.13  
## children 4 1338 1.09 1.21 1.00 0.94 1.48 0.00 5.00  
## smoker\* 5 1338 1.20 0.40 1.00 1.13 0.00 1.00 2.00  
## region\* 6 1338 2.52 1.10 3.00 2.52 1.48 1.00 4.00  
## charges 7 1338 13270.42 12110.01 9382.03 11076.02 7440.81 1121.87 63770.43  
## range skew kurtosis se  
## age 46.00 0.06 -1.25 0.38  
## sex\* 1.00 -0.02 -2.00 0.01  
## bmi 37.17 0.28 -0.06 0.17  
## children 5.00 0.94 0.19 0.03  
## smoker\* 1.00 1.46 0.14 0.01  
## region\* 3.00 -0.04 -1.33 0.03  
## charges 62648.55 1.51 1.59 331.07

# Summary of the dataset  
print(summary(d))

## age sex bmi children   
## Min. :18.00 Length:1338 Min. :15.96 Min. :0.000   
## 1st Qu.:27.00 Class :character 1st Qu.:26.30 1st Qu.:0.000   
## Median :39.00 Mode :character Median :30.40 Median :1.000   
## Mean :39.21 Mean :30.66 Mean :1.095   
## 3rd Qu.:51.00 3rd Qu.:34.69 3rd Qu.:2.000   
## Max. :64.00 Max. :53.13 Max. :5.000   
## smoker region charges   
## Length:1338 Length:1338 Min. : 1122   
## Class :character Class :character 1st Qu.: 4740   
## Mode :character Mode :character Median : 9382   
## Mean :13270   
## 3rd Qu.:16640   
## Max. :63770

# Internal structure of R object  
print(str(d))

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

# Display columns and display some portions of the data  
#print(glimpse(d))  
  
  
# To print unique columns  
#print(unique(d$age))  
#print(unique(d$bmi))  
#print(unique(d$charges))  
  
  
# statistical values  
#print(is.na(d))  
print(ncol(d))

## [1] 7

print(nrow(d))

## [1] 1338

print(max(d$charges))

## [1] 63770.43

print(min(d$charges))

## [1] 1121.874

#print(sort(d$charges))  
print(which.max(d$charges))# Return the index of the first maximum value

## [1] 544

print(which.min(d$charges))# Return the index of the first minimum value

## [1] 941

print(mean(d$charges))

## [1] 13270.42

print(mean(d$charges,trim=0.10))

## [1] 11076.02

print(var(d$charges))

## [1] 146652372

print(median(d$charges))

## [1] 9382.033

print(mad(d$charges))# mean absolute division

## [1] 7440.809

print(sd(d$charges))

## [1] 12110.01

print(range(d$charges))

## [1] 1121.874 63770.428

print(quantile(d$charges))

## 0% 25% 50% 75% 100%   
## 1121.874 4740.287 9382.033 16639.913 63770.428

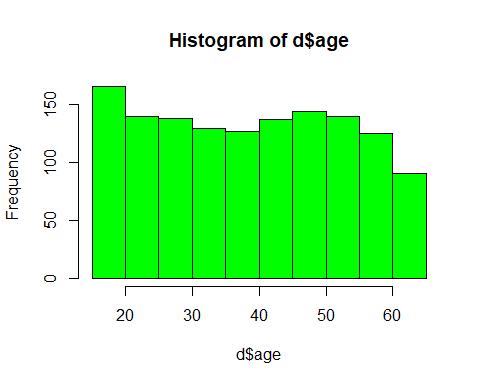
print(IQR(d$charges))

## [1] 11899.63

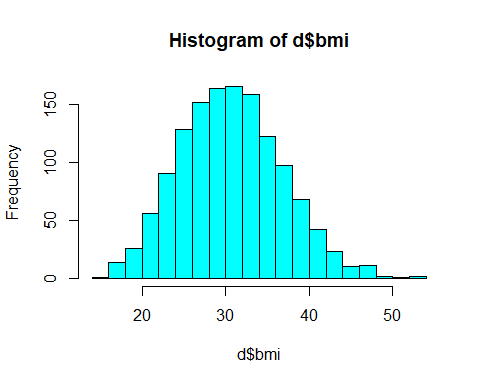
print(t.test(d$charges))

##   
## One Sample t-test  
##   
## data: d$charges  
## t = 40.084, df = 1337, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 12620.95 13919.89  
## sample estimates:  
## mean of x   
## 13270.42

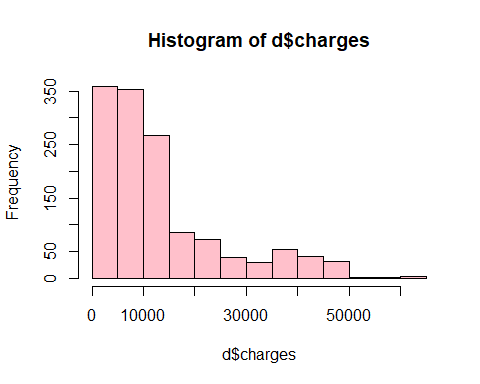
# Data visualisation  
  
# Histogram of Numerical data  
hist(d$age,breaks=15,col="green")



hist(d$bmi,breaks=15,col="cyan")



# BMI values are normally disributed.  
  
  
hist(d$charges,breaks=15,col="pink")



# As we expected, the figure shows right skewed distribution  
  
  
# To see the distribution of data  
table(d$region)

##   
## northeast northwest southeast southwest   
## 324 325 364 325

table(d$age)

##   
## 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43   
## 69 68 29 28 28 28 28 28 28 28 28 27 27 27 26 26 26 25 25 25 25 25 27 27 27 27   
## 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64   
## 27 29 29 29 29 28 29 29 29 28 28 26 26 26 25 25 23 23 23 23 22

table(d$sex)

##   
## female male   
## 662 676

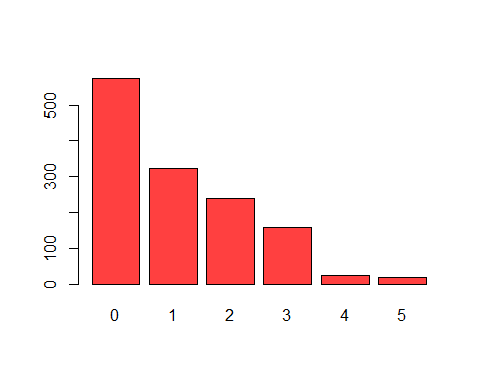
table(d$smoker)

##   
## no yes   
## 1064 274

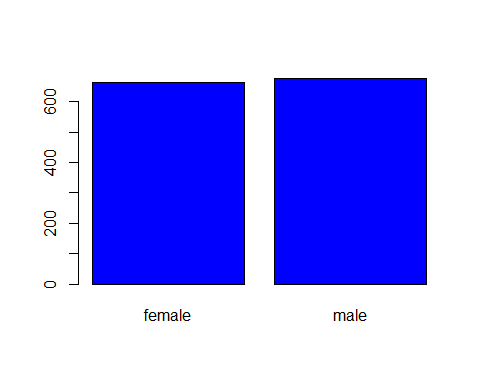
table(d$children)

##   
## 0 1 2 3 4 5   
## 574 324 240 157 25 18

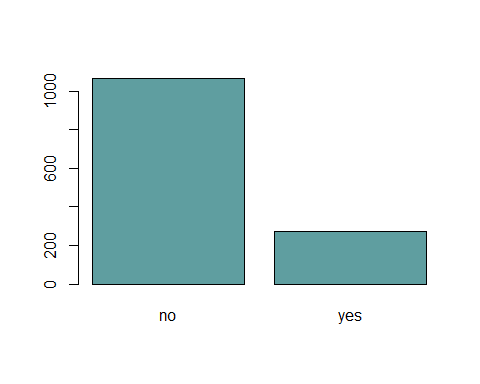
#table(d$bmi)  
#table(d$charges)  
  
  
# Barplot of Categorical data  
  
barplot(table(d$children),col="brown1")



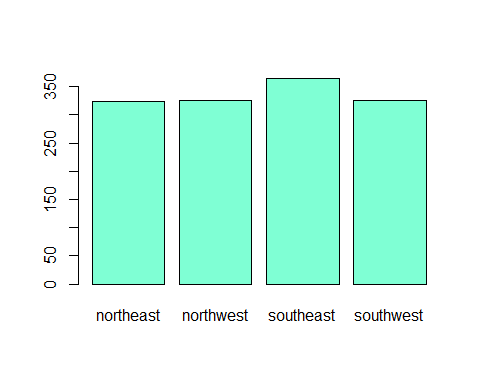
# majority of them having no children.  
  
barplot(table(d$sex),col="blue1")



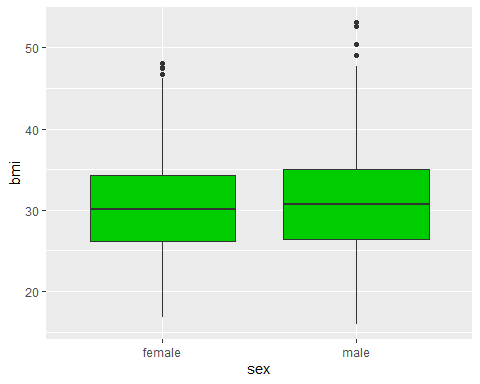
# Here the graph shows,number of males are more than females.  
  
  
barplot(table(d$smoker),col="cadetblue")



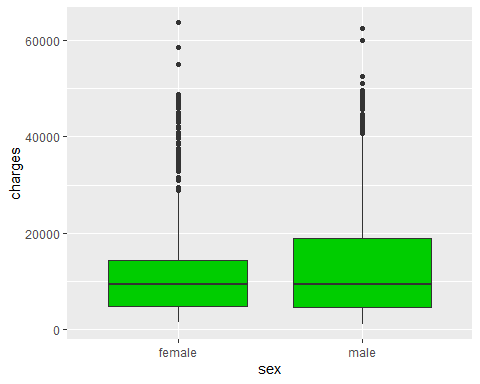
# The number of persons without smoke are more than others.  
  
  
barplot(table(d$region),col="aquamarine")



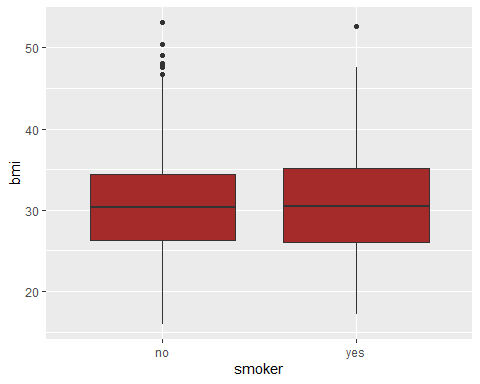
# Shows,more number of persons are from southeast.  
  
  
# Boxplot male and female with BMI values  
sex\_bmi<-ggplot(d,aes(x=sex,y=bmi))+geom\_boxplot(fill="green3")  
print(sex\_bmi)



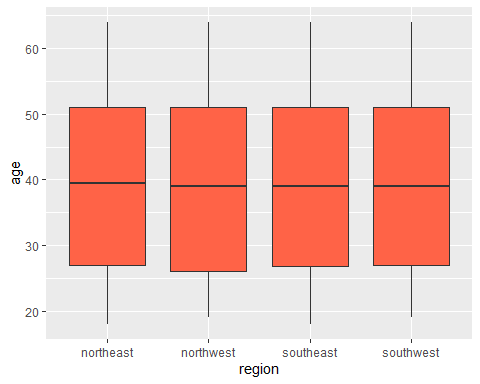
# BMI value is more for male than female  
  
  
## Boxplot of male and female with charges  
sex\_chr<-ggplot(d,aes(x=sex,y=charges))+geom\_boxplot(fill="green3")  
print(sex\_chr)



# More charges are paid by male  
  
  
# Boxplot of smoker and nonsmoker with BMI values  
smok\_bmi<-ggplot(d,aes(x=smoker,y=bmi))+geom\_boxplot(fill="brown")  
print(smok\_bmi)



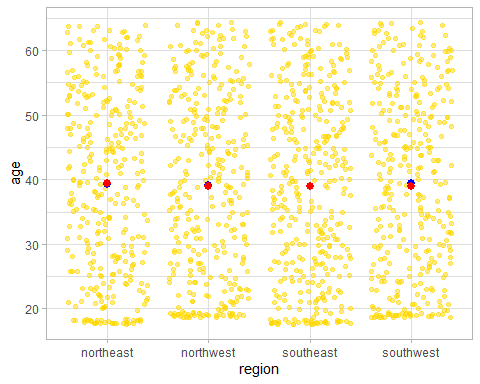
# BMI value of smokers are more than without smokers  
  
  
# Boxplot of age with region  
age\_reg<-ggplot(d,aes(x=region,y=age))+geom\_boxplot(fill="tomato")  
print(age\_reg)



# Here Maximum age from all regions are almost same  
  
  
  
# geom\_jitter with region and age  
g1 <- ggplot(d, aes(region, age)) +  
 geom\_jitter(color = "gold", alpha = 0.5) +  
 theme\_light()+  
 stat\_summary(aes(x=region,y=age),fun=mean,color="blue")+  
 stat\_summary(aes(x=region,y=age),fun=median,color="red")  
print(g1)

## Warning: Removed 4 rows containing missing values (geom\_segment).

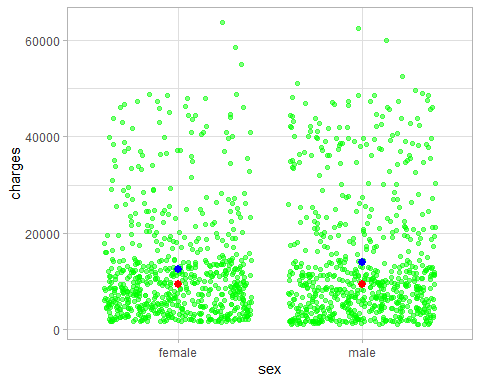
## Warning: Removed 4 rows containing missing values (geom\_segment).



# Here all the region shows almost same mean and median value for age  
  
  
# geom\_jitter with sex and charges  
g2 <- ggplot(d, aes(sex, charges)) +  
 geom\_jitter(color = "green", alpha = 0.5) +  
 theme\_light()+  
 stat\_summary(aes(x=sex,y=charges),fun=mean,color="blue")+  
 stat\_summary(aes(x=sex,y=charges),fun=median,color="red")  
print(g2)

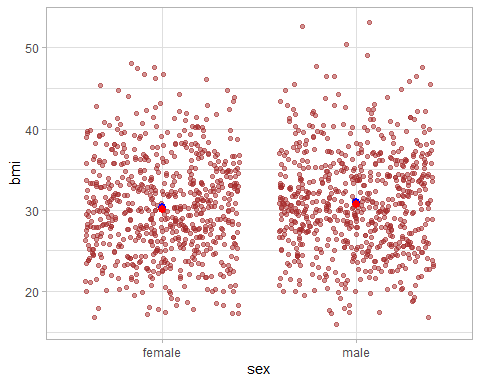
## Warning: Removed 2 rows containing missing values (geom\_segment).

## Warning: Removed 2 rows containing missing values (geom\_segment).



# Here, there is a small differnce in mean value of male and female w.r.t  
# charges  
  
  
# geom\_jitter of sex and bmi  
g3 <- ggplot(d, aes(sex, bmi)) +  
 geom\_jitter(color = "brown", alpha = 0.5) +  
 theme\_light()+  
 stat\_summary(aes(x=sex,y=bmi),fun=mean,color="blue")+  
 stat\_summary(aes(x=sex,y=bmi),fun=median,color="red")  
print(g3)

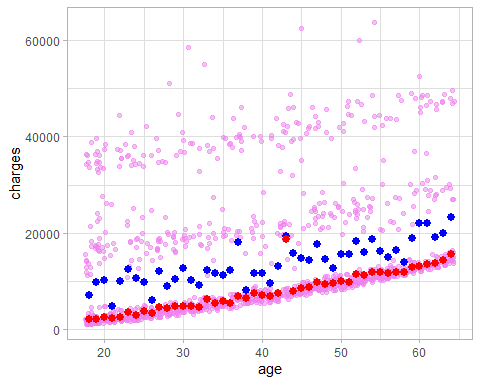
## Warning: Removed 2 rows containing missing values (geom\_segment).  
  
## Warning: Removed 2 rows containing missing values (geom\_segment).



# There is a small difference in mean and median value of male and female   
# w.r.t bmi values  
  
# geom\_jitter of age and charges  
g4<-ggplot(d, aes(age, charges)) +  
 geom\_jitter(color = "violet", alpha = 0.5) +  
 theme\_light()+  
 stat\_summary(aes(x=age,y=charges),fun=mean,color="blue")+  
 stat\_summary(aes(x=age,y=charges),fun=median,color="red")  
print(g4)

## Warning: Removed 47 rows containing missing values (geom\_segment).

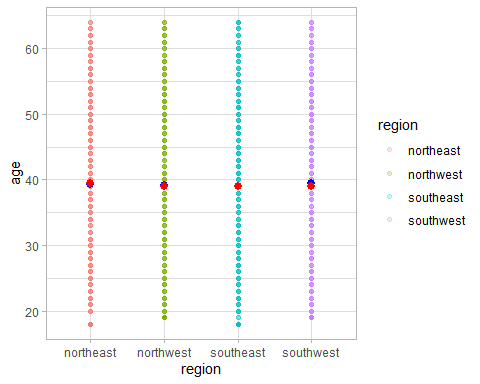
## Warning: Removed 47 rows containing missing values (geom\_segment).



# Here,mean and median values are different w.r.t age and charges  
  
  
# geom\_point with region and age  
p1<-ggplot(data=d)+geom\_point(aes(x=region,y=age,color=region),alpha=.2)+  
 theme\_light()+  
 stat\_summary(aes(x=region,y=age),fun=mean,color="blue")+  
 stat\_summary(aes(x=region,y=age),fun=median,color="red")  
print(p1)

## Warning: Removed 4 rows containing missing values (geom\_segment).

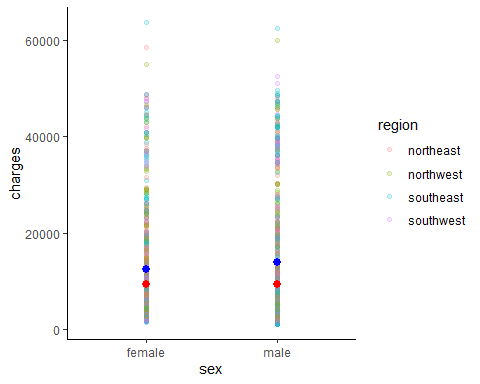
## Warning: Removed 4 rows containing missing values (geom\_segment).



# geom\_point with sex and charges  
p2<-ggplot(data=d)+geom\_point(aes(x=sex,y=charges,color=region),alpha=.2)+  
 theme\_classic()+  
 stat\_summary(aes(x=sex,y=charges),fun=mean,color="blue")+  
 stat\_summary(aes(x=sex,y=charges),fun=median,color="red")  
print(p2)

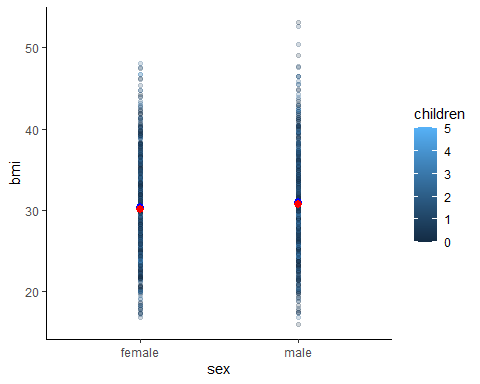
## Warning: Removed 2 rows containing missing values (geom\_segment).

## Warning: Removed 2 rows containing missing values (geom\_segment).



# geom\_point with sex and bmi  
p3<-ggplot(data=d)+geom\_point(aes(x=sex,y=bmi,color=children),alpha=.2)+  
 theme\_classic()+  
 stat\_summary(aes(x=sex,y=bmi),fun=mean,color="blue")+  
 stat\_summary(aes(x=sex,y=bmi),fun=median,color="red")  
print(p3)

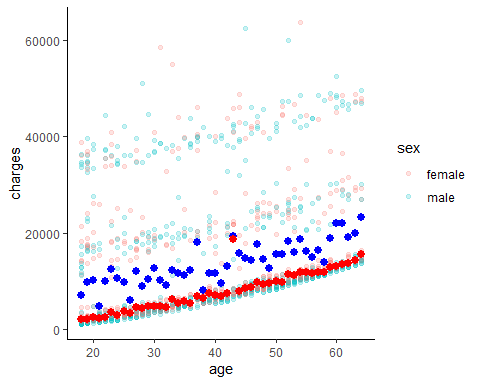
## Warning: Removed 2 rows containing missing values (geom\_segment).  
  
## Warning: Removed 2 rows containing missing values (geom\_segment).



# geom\_point with age and charges  
p4<-ggplot(data=d)+geom\_point(aes(x=age,y=charges,color=sex),alpha=.2)+  
 theme\_classic()+  
 stat\_summary(aes(x=age,y=charges),fun=mean,color="blue")+  
 stat\_summary(aes(x=age,y=charges),fun=median,color="red")  
print(p4)

## Warning: Removed 47 rows containing missing values (geom\_segment).

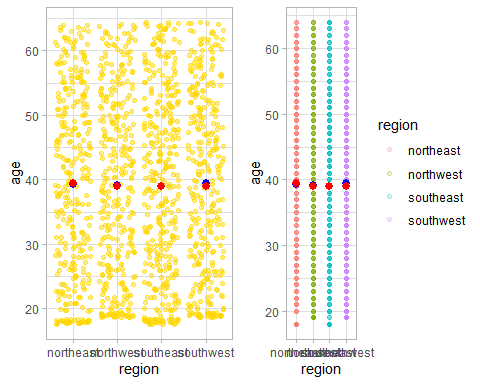
## Warning: Removed 47 rows containing missing values (geom\_segment).



# Combination of geom\_jitter and geom\_point  
print(grid.arrange(g1,p1,nrow=1))

## Warning: Removed 4 rows containing missing values (geom\_segment).

## Warning: Removed 4 rows containing missing values (geom\_segment).  
  
## Warning: Removed 4 rows containing missing values (geom\_segment).  
  
## Warning: Removed 4 rows containing missing values (geom\_segment).

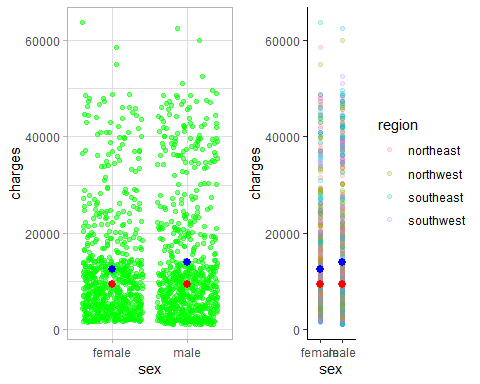


## TableGrob (1 x 2) "arrange": 2 grobs  
## z cells name grob  
## 1 1 (1-1,1-1) arrange gtable[layout]  
## 2 2 (1-1,2-2) arrange gtable[layout]

print(grid.arrange(g2,p2,nrow=1))

## Warning: Removed 2 rows containing missing values (geom\_segment).

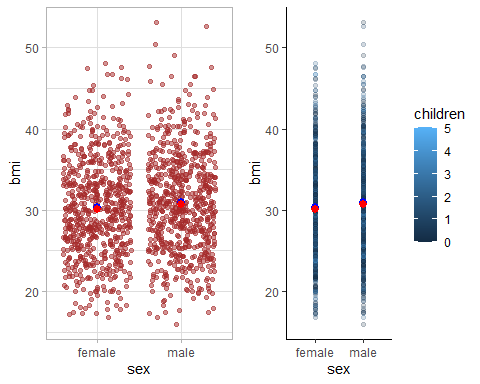
## Warning: Removed 2 rows containing missing values (geom\_segment).  
  
## Warning: Removed 2 rows containing missing values (geom\_segment).  
  
## Warning: Removed 2 rows containing missing values (geom\_segment).



## TableGrob (1 x 2) "arrange": 2 grobs  
## z cells name grob  
## 1 1 (1-1,1-1) arrange gtable[layout]  
## 2 2 (1-1,2-2) arrange gtable[layout]

print(grid.arrange(g3,p3,nrow=1))

## Warning: Removed 2 rows containing missing values (geom\_segment).  
  
## Warning: Removed 2 rows containing missing values (geom\_segment).  
  
## Warning: Removed 2 rows containing missing values (geom\_segment).  
  
## Warning: Removed 2 rows containing missing values (geom\_segment).

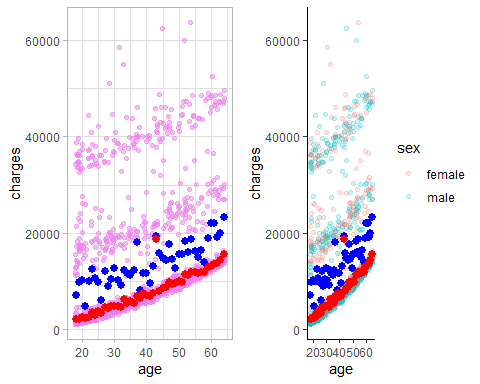


## TableGrob (1 x 2) "arrange": 2 grobs  
## z cells name grob  
## 1 1 (1-1,1-1) arrange gtable[layout]  
## 2 2 (1-1,2-2) arrange gtable[layout]

print(grid.arrange(g4,p4,nrow=1))

## Warning: Removed 47 rows containing missing values (geom\_segment).

## Warning: Removed 47 rows containing missing values (geom\_segment).  
  
## Warning: Removed 47 rows containing missing values (geom\_segment).  
  
## Warning: Removed 47 rows containing missing values (geom\_segment).

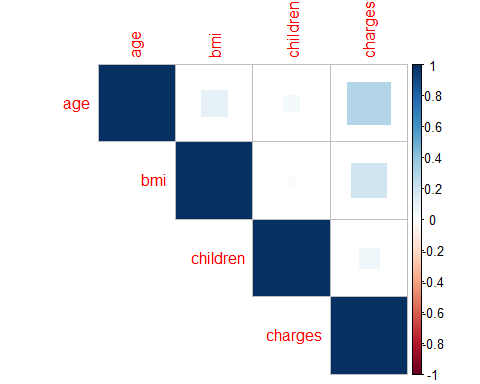


## TableGrob (1 x 2) "arrange": 2 grobs  
## z cells name grob  
## 1 1 (1-1,1-1) arrange gtable[layout]  
## 2 2 (1-1,2-2) arrange gtable[layout]

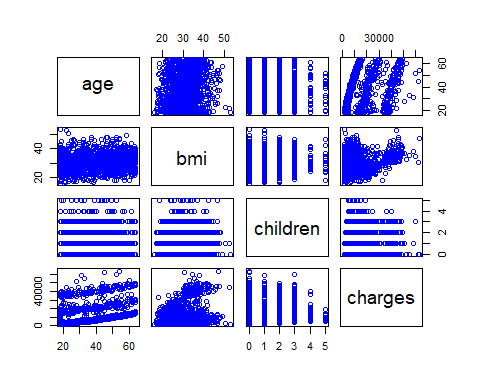
# To check the summary of charges  
summary(d$charges)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1122 4740 9382 13270 16640 63770

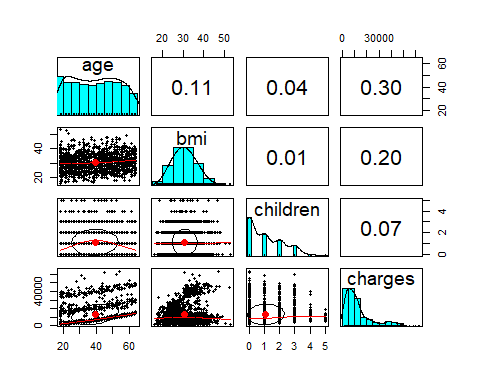
# To find the relation among variables.So we will use correlation matrix  
corr<-cor(d[c("age","bmi","children","charges")])  
corrplot(corr,method="square",type="upper")



# Scatterplot matrix  
pairs(d[c("age","bmi","children","charges")],col="blue")



# To add more information to scatterplot.  
# To enhance the plot,already load the package "psych"  
pairs.panels(d[c("age","bmi","children","charges")])



# To train a model on to the data  
# To fit the linear regression model to the data with R,we will use  
# the function lm()  
model<-lm(charges~age+children+bmi+sex+region,data=d)  
model<-lm(charges~.,data=d)  
   
  
# To build the model  
model

##   
## Call:  
## lm(formula = charges ~ ., data = d)  
##   
## Coefficients:  
## (Intercept) age sexmale bmi   
## -11938.5 256.9 -131.3 339.2   
## children smokeryes regionnorthwest regionsoutheast   
## 475.5 23848.5 -353.0 -1035.0   
## regionsouthwest   
## -960.1

# To view more information about the model  
summary(model)

##   
## Call:  
## lm(formula = charges ~ ., data = d)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11304.9 -2848.1 -982.1 1393.9 29992.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11938.5 987.8 -12.086 < 2e-16 \*\*\*  
## age 256.9 11.9 21.587 < 2e-16 \*\*\*  
## sexmale -131.3 332.9 -0.394 0.693348   
## bmi 339.2 28.6 11.860 < 2e-16 \*\*\*  
## children 475.5 137.8 3.451 0.000577 \*\*\*  
## smokeryes 23848.5 413.1 57.723 < 2e-16 \*\*\*  
## regionnorthwest -353.0 476.3 -0.741 0.458769   
## regionsoutheast -1035.0 478.7 -2.162 0.030782 \*   
## regionsouthwest -960.0 477.9 -2.009 0.044765 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6062 on 1329 degrees of freedom  
## Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494   
## F-statistic: 500.8 on 8 and 1329 DF, p-value: < 2.2e-16

# In this analysis,applied linear regression.  
# As we can see, summary of a model showed us the significance of variable.