Link to dataset: <https://www.kaggle.com/mirichoi0218/insurance> Predicting cost of medical bill provided age, sex, bmi, number of children, smoking habit, and region.

Loading the CSV file

insuranceDataset = read.csv("insurance.csv", header = TRUE)

Exploring data

head(insuranceDataset)

## 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

tail(insuranceDataset)

## 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

summary(insuranceDataset)

## age sex bmi children smoker   
## Min. :18.00 female:662 Min. :15.96 Min. :0.000 no :1064   
## 1st Qu.:27.00 male :676 1st Qu.:26.30 1st Qu.:0.000 yes: 274   
## Median :39.00 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   
## region charges   
## northeast:324 Min. : 1122   
## northwest:325 1st Qu.: 4740   
## southeast:364 Median : 9382   
## southwest:325 Mean :13270   
## 3rd Qu.:16640   
## Max. :63770

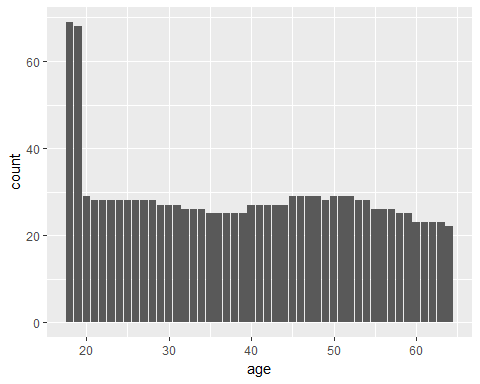
Visualizing data using ggplot2

install.packages('ggplot2', repos = "http://cran.us.r-project.org")

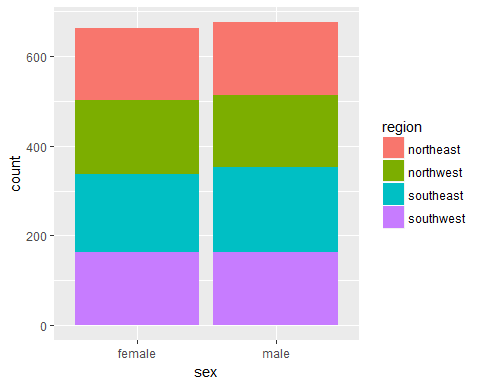
## Installing package into 'C:/Users/Faraz Khalid/Documents/R/win-library/3.4'  
## (as 'lib' is unspecified)

## package 'ggplot2' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\Faraz Khalid\AppData\Local\Temp\RtmpC4rs3u\downloaded\_packages

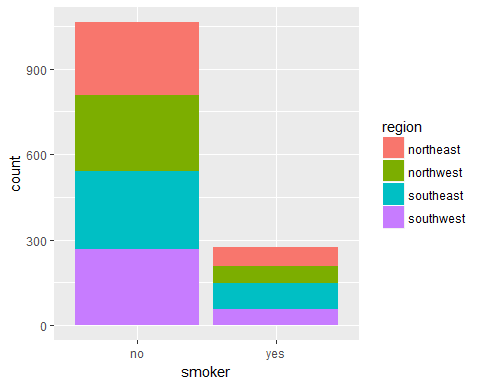
library(ggplot2)  
  
#Plotting the age of participants and their count with respective age.  
ggplot(insuranceDataset, aes(age, fill = age ) ) +  
 geom\_bar()



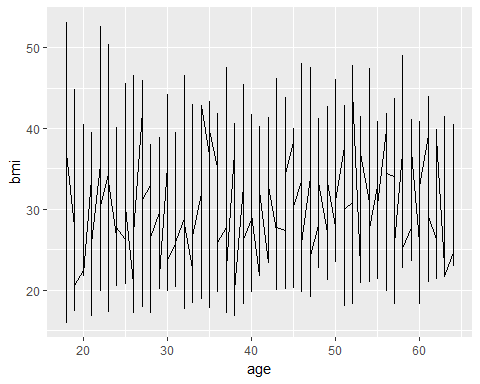
#Plotting number of males and number of females in the dataset with respect to their region.  
ggplot(insuranceDataset, aes(sex, fill = region) ) +  
 geom\_bar(position = "stack")



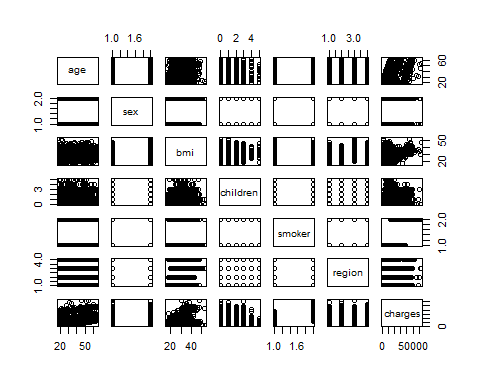
#Checking to see if their is a pattern between region and person smoking.  
ggplot(insuranceDataset, aes(smoker, fill = region) ) +  
 geom\_bar(position = "stack")



#Plotting age and BMI to see if their is a noticible pattern.  
ggplot(insuranceDataset, aes(age, bmi)) +  
 geom\_line()



#Corelation Matrix  
pairs(insuranceDataset)



Cleaning and transforming data to so that it better data can be used to train and test model.

#Males are represented by 1's while females are given a numeric of 0 for identification purposes.  
insuranceDataset$sex<-ifelse(insuranceDataset$sex =='male', 1,0)  
  
#If a person is a smoker, it is indicated as 1 and if they are not, they are indicated with a 0.  
insuranceDataset$smoker<-ifelse(insuranceDataset$smoker =='yes', 1,0)

Dividing data into training and test set.

set.seed(1234)  
i <- sample(1:nrow(insuranceDataset), nrow(insuranceDataset)\*0.75, replace=FALSE)  
train <- insuranceDataset[i,]  
test <- insuranceDataset[-i,]

Linear Regression Model to predict charge

lm1 <- lm(charges ~ ., data=train)  
lm1$coefficients

## (Intercept) age sex bmi   
## -12605.5783 259.0506 -349.6416 342.3090   
## children smoker regionnorthwest regionsoutheast   
## 554.2580 23671.1676 253.6374 -638.4677   
## regionsouthwest   
## -376.6793

summary(lm1)

##   
## Call:  
## lm(formula = charges ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10969.1 -2943.5 -936.3 1484.2 25880.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12605.58 1126.64 -11.189 < 2e-16 \*\*\*  
## age 259.05 13.42 19.305 < 2e-16 \*\*\*  
## sex -349.64 376.87 -0.928 0.353762   
## bmi 342.31 32.42 10.558 < 2e-16 \*\*\*  
## children 554.26 158.81 3.490 0.000504 \*\*\*  
## smoker 23671.17 464.58 50.952 < 2e-16 \*\*\*  
## regionnorthwest 253.64 545.29 0.465 0.641930   
## regionsoutheast -638.47 543.77 -1.174 0.240617   
## regionsouthwest -376.68 542.29 -0.695 0.487460   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5949 on 994 degrees of freedom  
## Multiple R-squared: 0.7575, Adjusted R-squared: 0.7556   
## F-statistic: 388.2 on 8 and 994 DF, p-value: < 2.2e-16

#Evaluating the model  
pred <- predict(lm1, newdata=test)  
cor(pred, test$charges)

## [1] 0.8554837

Creating another linear regression modell but only with predictors that have \*\*\*

lm2 <- lm(charges ~ age+bmi+children+smoker, data=train)  
summary(lm2)

##   
## Call:  
## lm(formula = charges ~ age + bmi + children + smoker, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11446.1 -2964.1 -941.2 1430.9 26446.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12487.27 1076.83 -11.596 < 2e-16 \*\*\*  
## age 259.69 13.40 19.375 < 2e-16 \*\*\*  
## bmi 325.72 31.05 10.492 < 2e-16 \*\*\*  
## children 553.93 158.63 3.492 0.000501 \*\*\*  
## smoker 23599.20 462.75 50.998 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5949 on 998 degrees of freedom  
## Multiple R-squared: 0.7566, Adjusted R-squared: 0.7556   
## F-statistic: 775.4 on 4 and 998 DF, p-value: < 2.2e-16

pred2 <- predict(lm2, newdata=test)  
cor(pred2, test$charges)

## [1] 0.8554768

Comparision of two Linear Regression Models

anova(lm1, lm2)

## Analysis of Variance Table  
##   
## Model 1: charges ~ age + sex + bmi + children + smoker + region  
## Model 2: charges ~ age + bmi + children + smoker  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 994 3.5179e+10   
## 2 998 3.5320e+10 -4 -141693054 1.0009 0.4061