diabetesfinalproject.R

Pranshu

Sat Jun 11 22:31:01 2016

```
setwd("~/Research/Diabetes analysis2/NewDiabetes")
pima.indians.diabetes <- read.csv("C:/Users/IBM_ADMIN/Desktop/pima-indians-diab
etes.data", header=FALSE)
names(pima.indians.diabetes)<-c('TimesPregnant','GlucoseLV','DiastolicBP','Tric
epsThickness','SerumInsulin','BMI','Heridarymarkup','Age','Classification')
View(pima.indians.diabetes)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.5
```

```
##
## 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)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 3.2.5
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.2.5
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 3.2.5
```

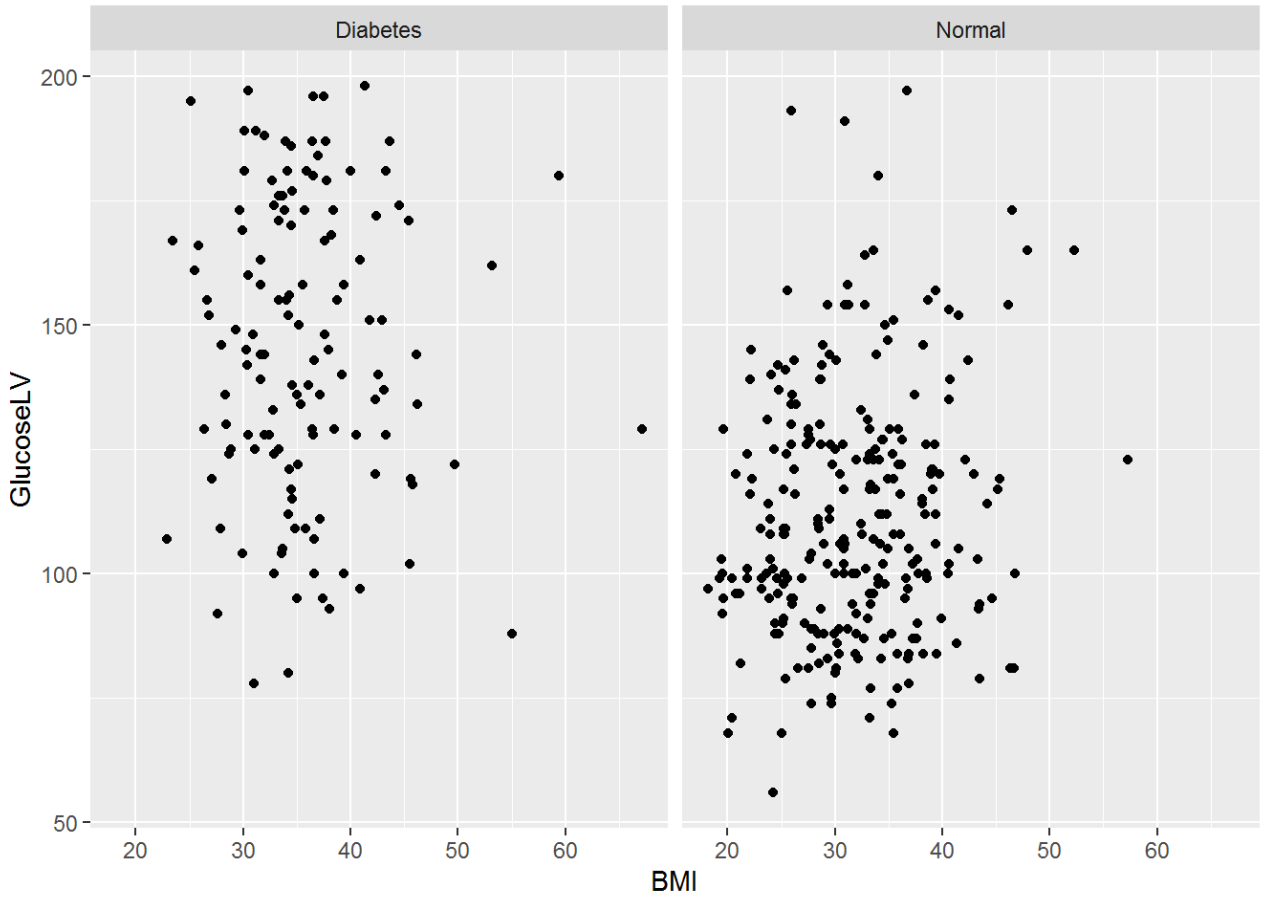
```
pima.indians.diabetes$DiastolicBP[pima.indians.diabetes$DiastolicBP==0]<-NA
pima.indians.diabetes$GlucoseLV[pima.indians.diabetes$GlucoseLV==0]<-NA
pima.indians.diabetes$TricepsThickness[pima.indians.diabetes$TricepsThickness==
0]<-NA
pima.indians.diabetes$SerumInsulin[pima.indians.diabetes$SerumInsulin==0]<-NA
pima.indians.diabetes$BMI[pima.indians.diabetes$BMI==0]<-NA
pima.indian.diabetes2<-na.omit(pima.indians.diabetes)
summary(pima.indian.diabetes2)
```

```
##  TimesPregnant      GlucoseLV      DiastolicBP     TricepsThickness
##  Min.   : 0.000   Min.   : 56.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:21.00
##  Median : 2.000   Median :119.0   Median : 70.00   Median :29.00
##  Mean   : 3.301   Mean   :122.6   Mean   : 70.66   Mean   :29.15
##  3rd Qu.: 5.000   3rd Qu.:143.0   3rd Qu.: 78.00   3rd Qu.:37.00
##  Max.   :17.000   Max.   :198.0   Max.   :110.00   Max.   :63.00
##   SerumInsulin         BMI        Heridarymarkup        Age
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0850   Min.   :21.00
##  1st Qu.: 76.75   1st Qu.:28.40   1st Qu.:0.2697   1st Qu.:23.00
##  Median :125.50   Median :33.20   Median :0.4495   Median :27.00
##  Mean   :156.06   Mean   :33.09   Mean   :0.5230   Mean   :30.86
##  3rd Qu.:190.00   3rd Qu.:37.10   3rd Qu.:0.6870   3rd Qu.:36.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
##  Classification
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3316
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

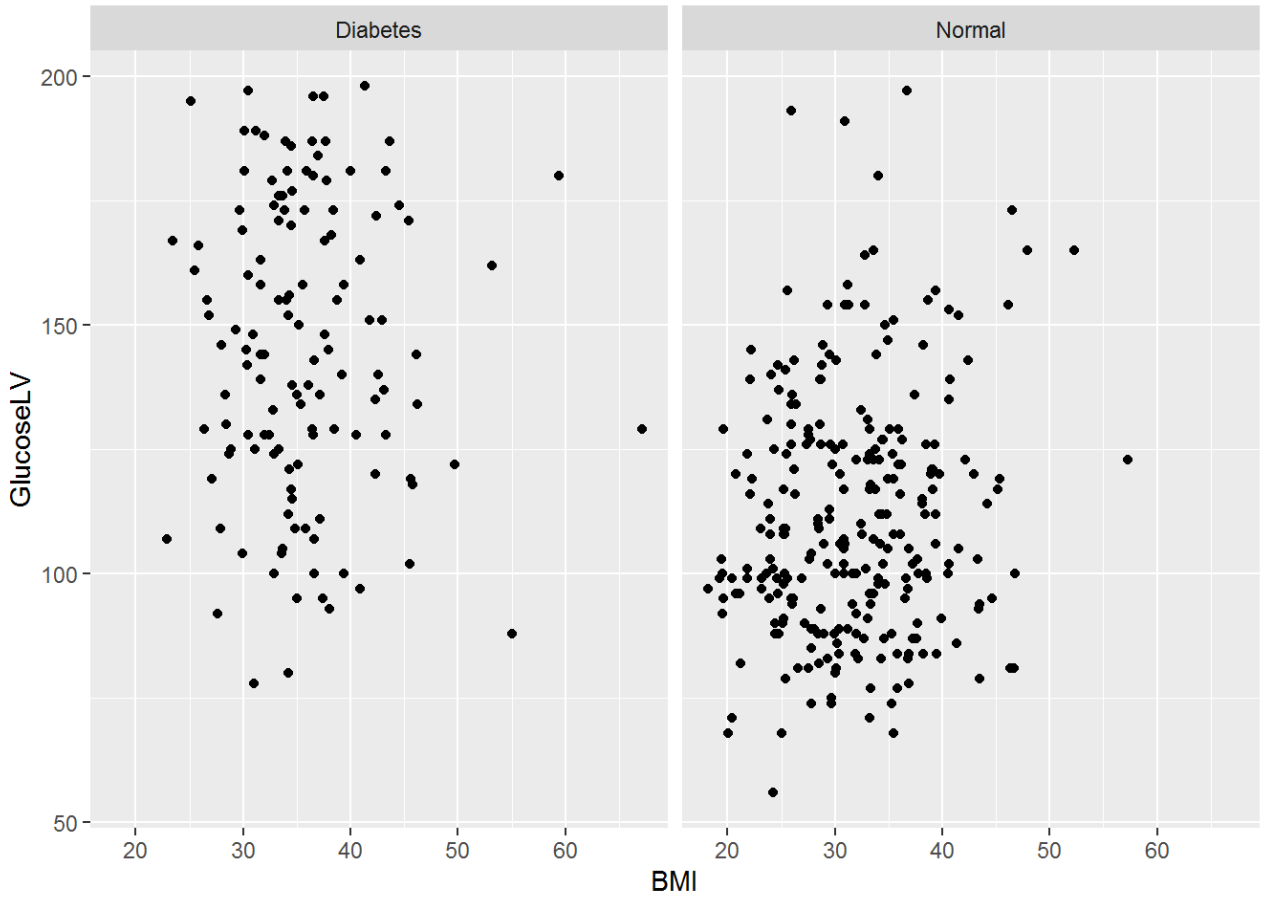
```
View(pima.indian.diabetes2)
pima.indian.diabetes4<-pima.indian.diabetes2%>% mutate(Group=ifelse(pima.india
n.diabetes2$Classification==1,"Diabetes","Normal"))
head(pima.indian.diabetes4)
```

```
##   TimesPregnant GlucoseLV DiastolicBP TricepsThickness SerumInsulin  BMI
## 1             1        89          66               23           94 28.1
## 2             0       137          40               35          168 43.1
## 3             3        78          50               32           88 31.0
## 4             2       197          70               45          543 30.5
## 5             1       189          60               23          846 30.1
## 6             5       166          72               19          175 25.8
##   Heridarymarkup Age Classification    Group
## 1          0.167  21              0   Normal
## 2          2.288  33              1 Diabetes
## 3          0.248  26              1 Diabetes
## 4          0.158  53              1 Diabetes
## 5          0.398  59              1 Diabetes
## 6          0.587  51              1 Diabetes
```

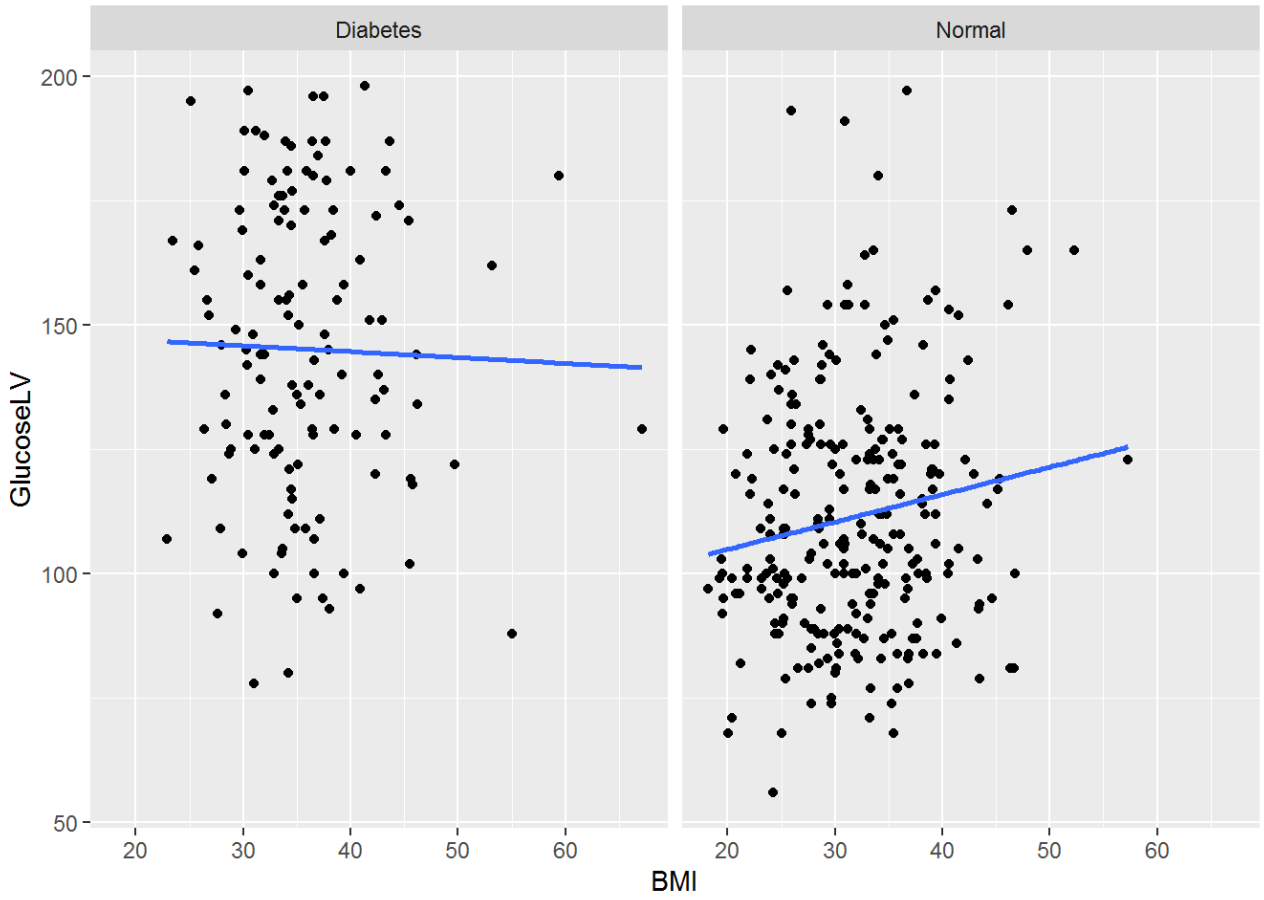
```
ggplot(pima.indian.diabetes4,aes(x=BMI,y=GlucoseLV))+geom_point()+facet_grid(.
~ Group)
```




```
ggplot(pima.indian.diabetes4,aes(x=BMI,y=GlucoseLV))+geom_point()+facet_grid(.
~ Group)
```




```
ggplot(pima.indian.diabetes4,aes(x=BMI,y=GlucoseLV))+geom_point()+geom_smooth(m
ethod= "lm",se=FALSE)+facet_grid(. ~ Group)
```



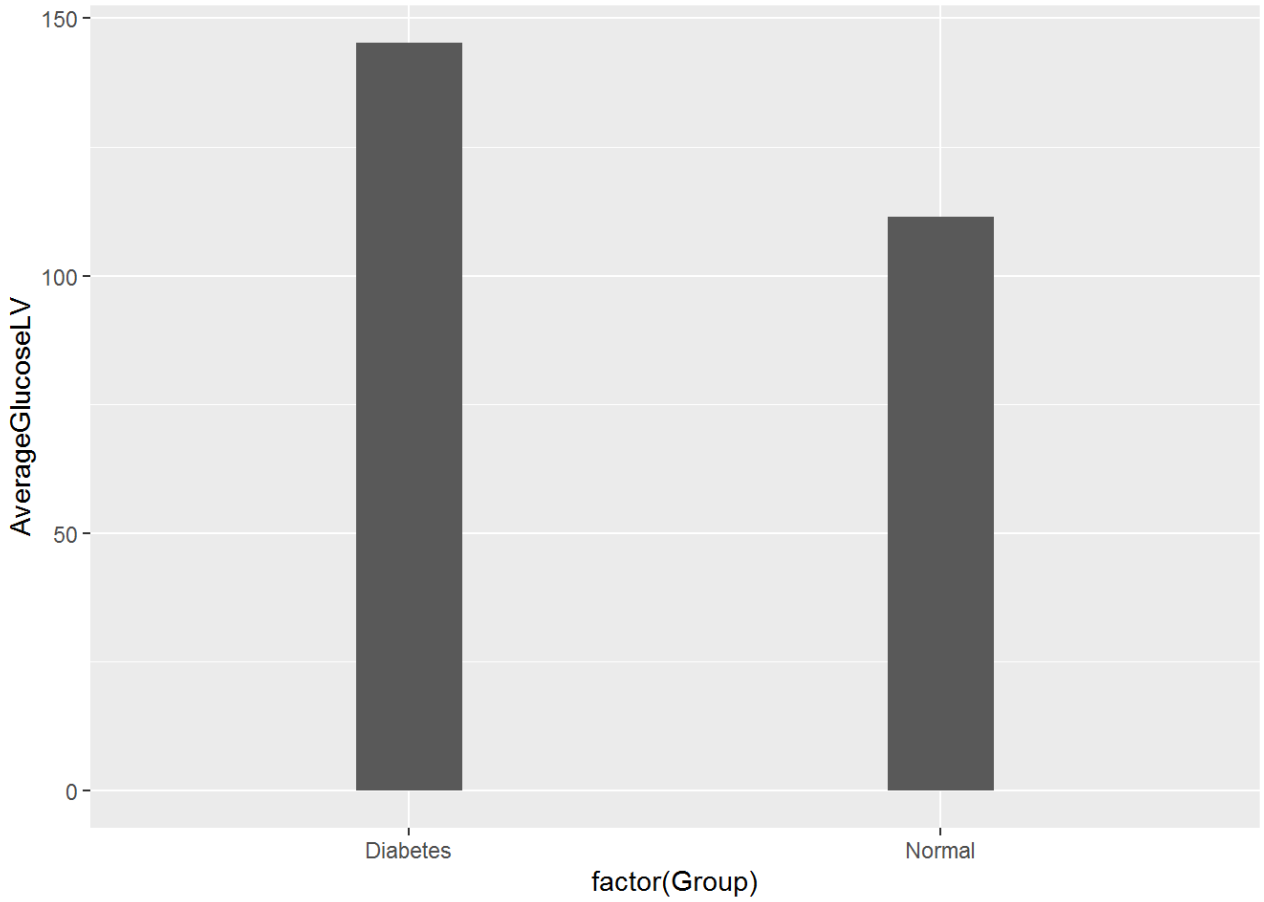

```
pima.indian.diabetes3<-pima.indian.diabetes4 %>% group_by(Group)%>% summarise_e
ach(funs(mean),GlucoseLV,DiastolicBP,SerumInsulin,BMI,TricepsThickness,Age)
View(pima.indian.diabetes3)
names(pima.indian.diabetes3)
```

```
## [1] "Group"            "GlucoseLV"        "DiastolicBP"
## [4] "SerumInsulin"     "BMI"              "TricepsThickness"
## [7] "Age"
```

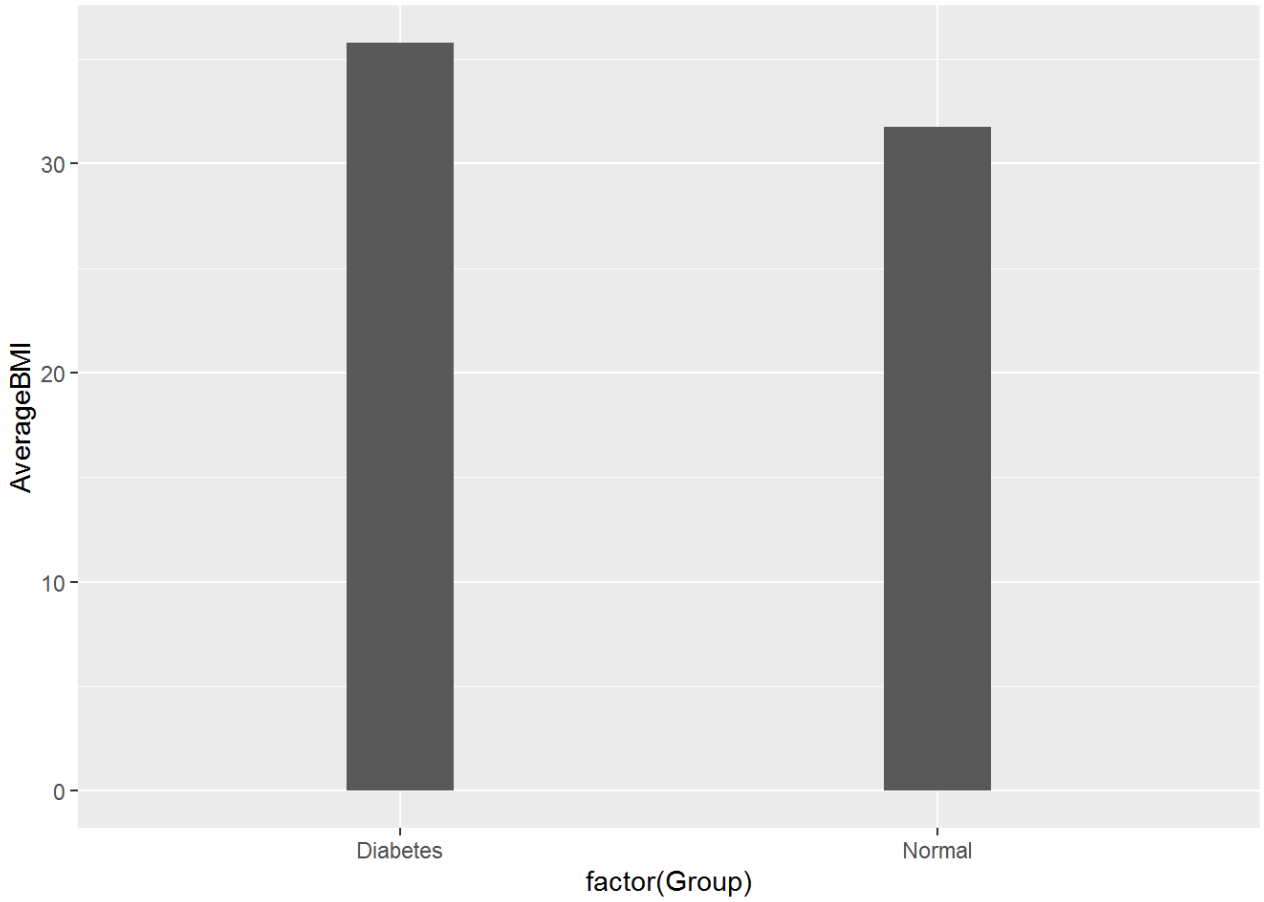
```
names(pima.indian.diabetes3)[2]<-'AverageGlucoseLV'
names(pima.indian.diabetes3)[3]<-'AverageDiastolicBP'
names(pima.indian.diabetes3)[4]<-'AverageSerumInsulin'
names(pima.indian.diabetes3)[5]<-'AverageBMI'
names(pima.indian.diabetes3)[6]<-'AverageTricepsThickness'
names(pima.indian.diabetes3)[7]<-'AverageAge'
names(pima.indian.diabetes3)
```

```
## [1] "Group"                  "AverageGlucoseLV"
## [3] "AverageDiastolicBP"     "AverageSerumInsulin"
## [5] "AverageBMI"             "AverageTricepsThickness"
## [7] "AverageAge"
```

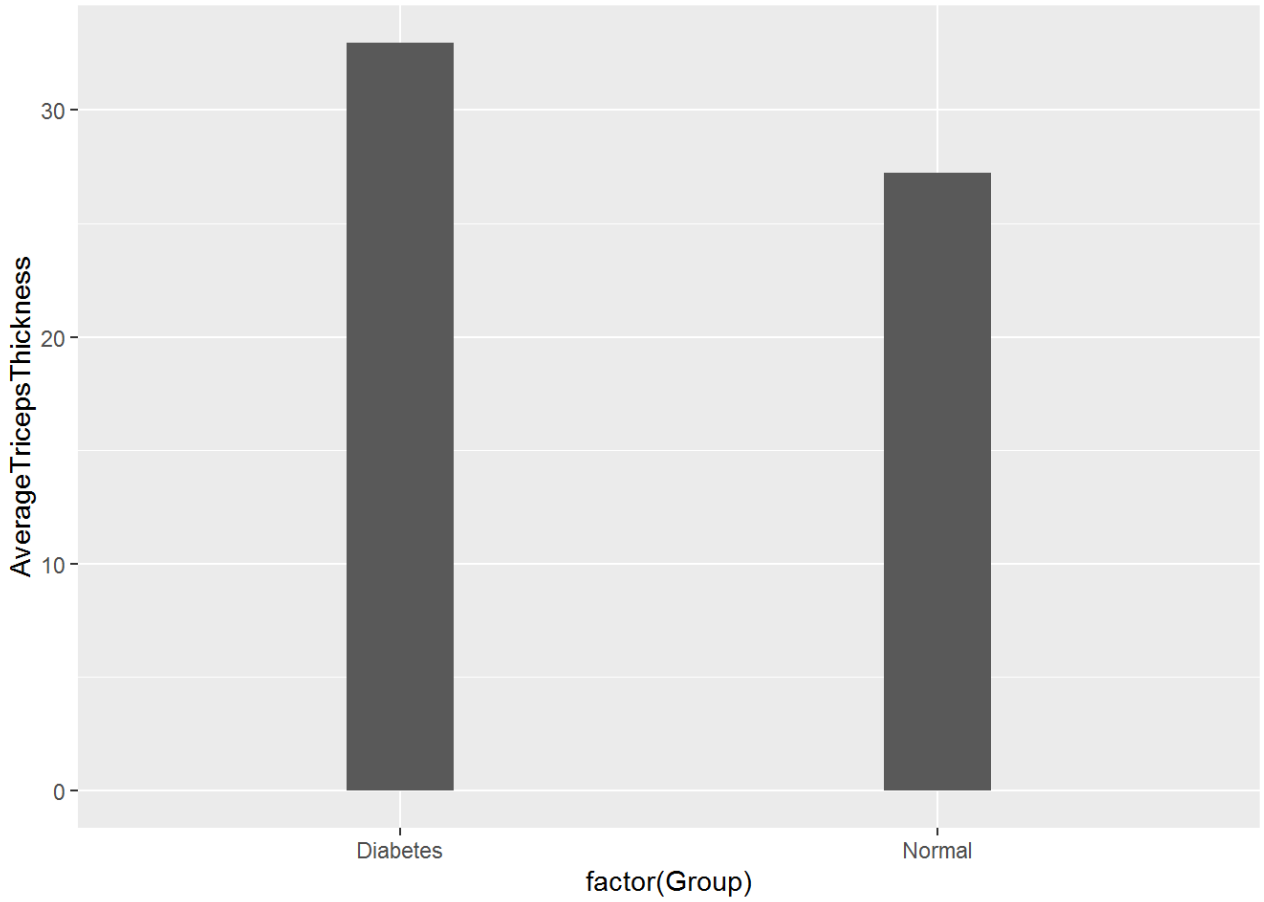
```
ggplot(pima.indian.diabetes3,aes(x=factor(Group),y=AverageGlucoseLV))+geom_bar
(stat="identity",width = 0.2)
```



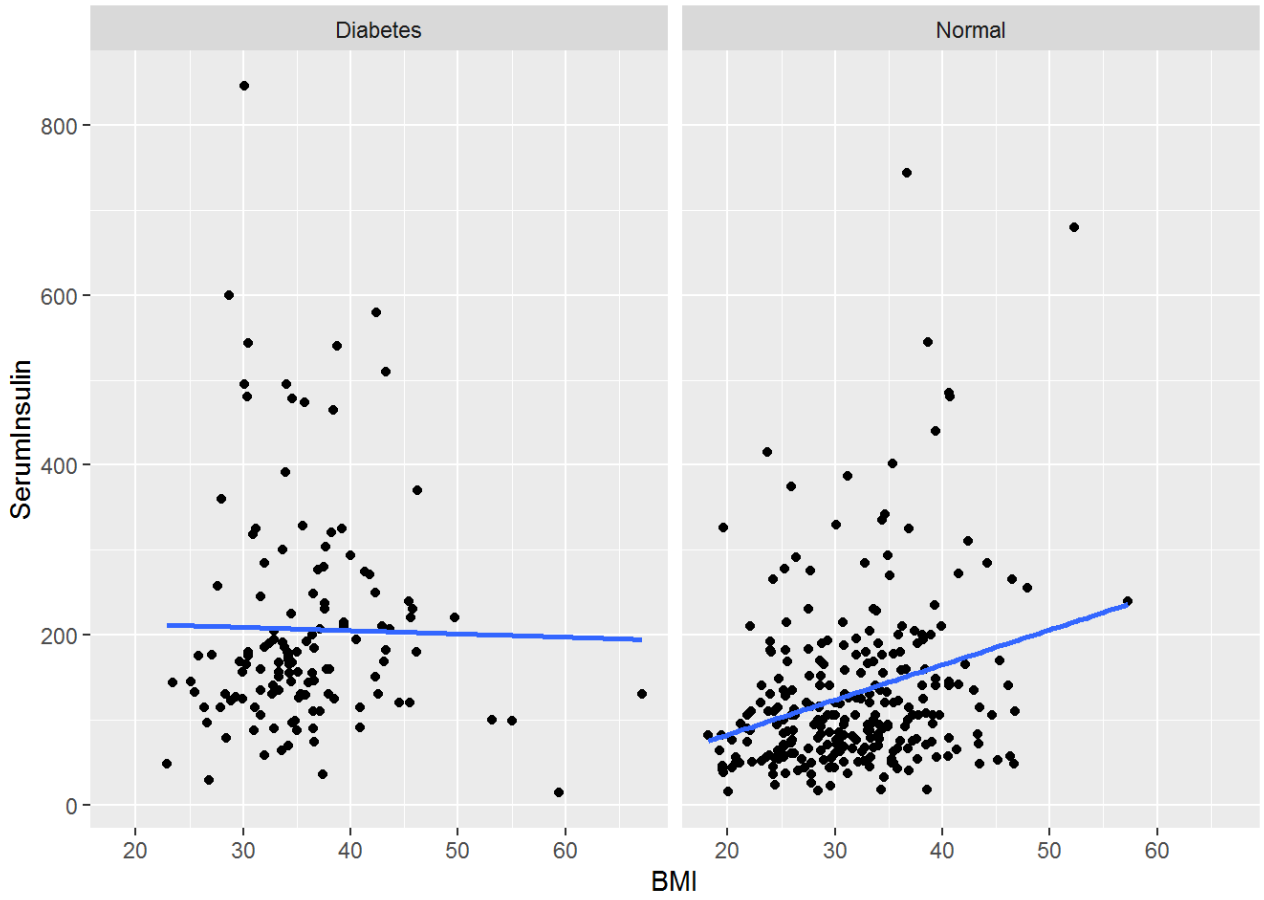

```
ggplot(pima.indian.diabetes3,aes(x=factor(Group),y=AverageBMI))+geom_bar(stat
="identity",width = 0.2)
```



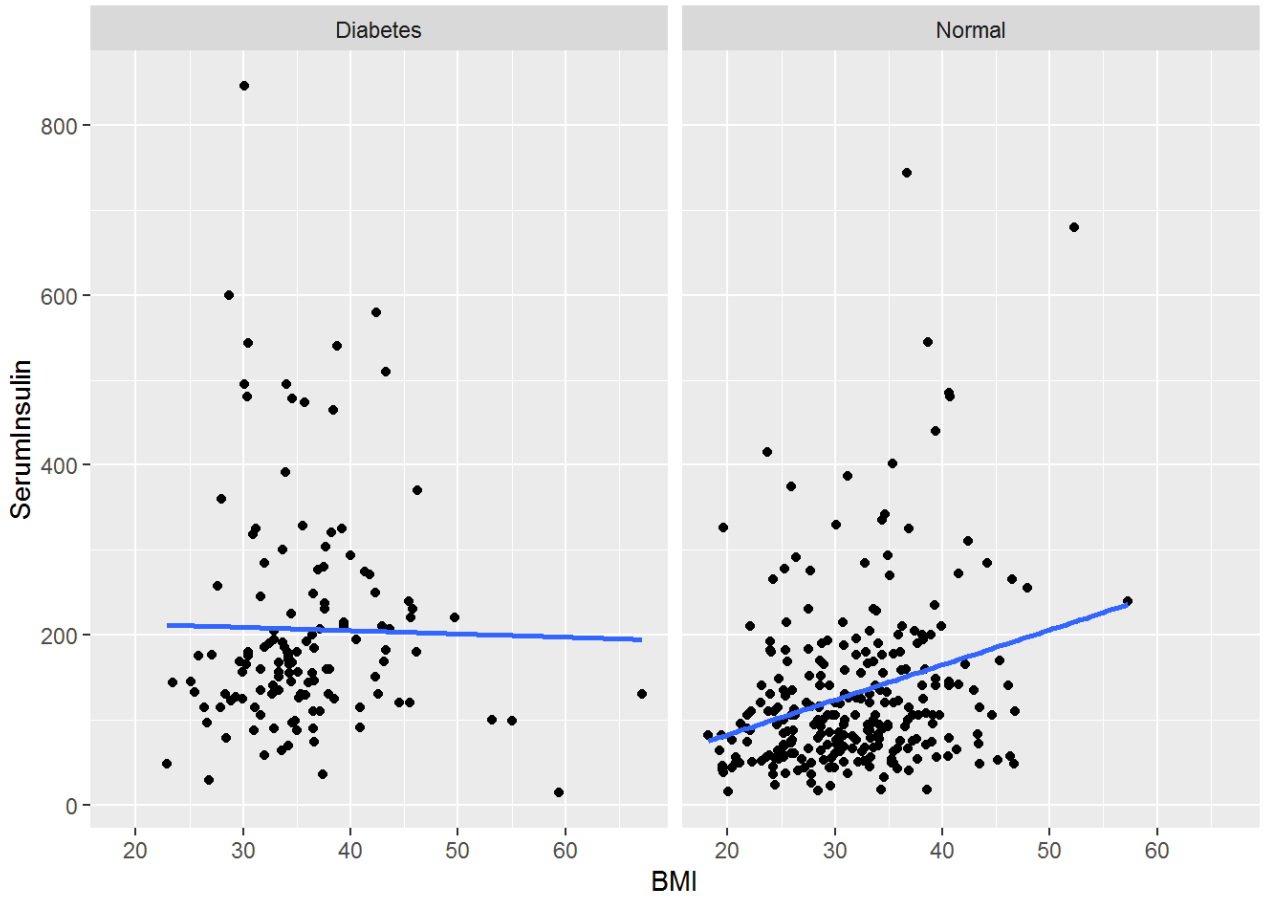

```
ggplot(pima.indian.diabetes3,aes(x=factor(Group),y=AverageTricepsThickness))+ge
om_bar(stat="identity",width = 0.2)
```



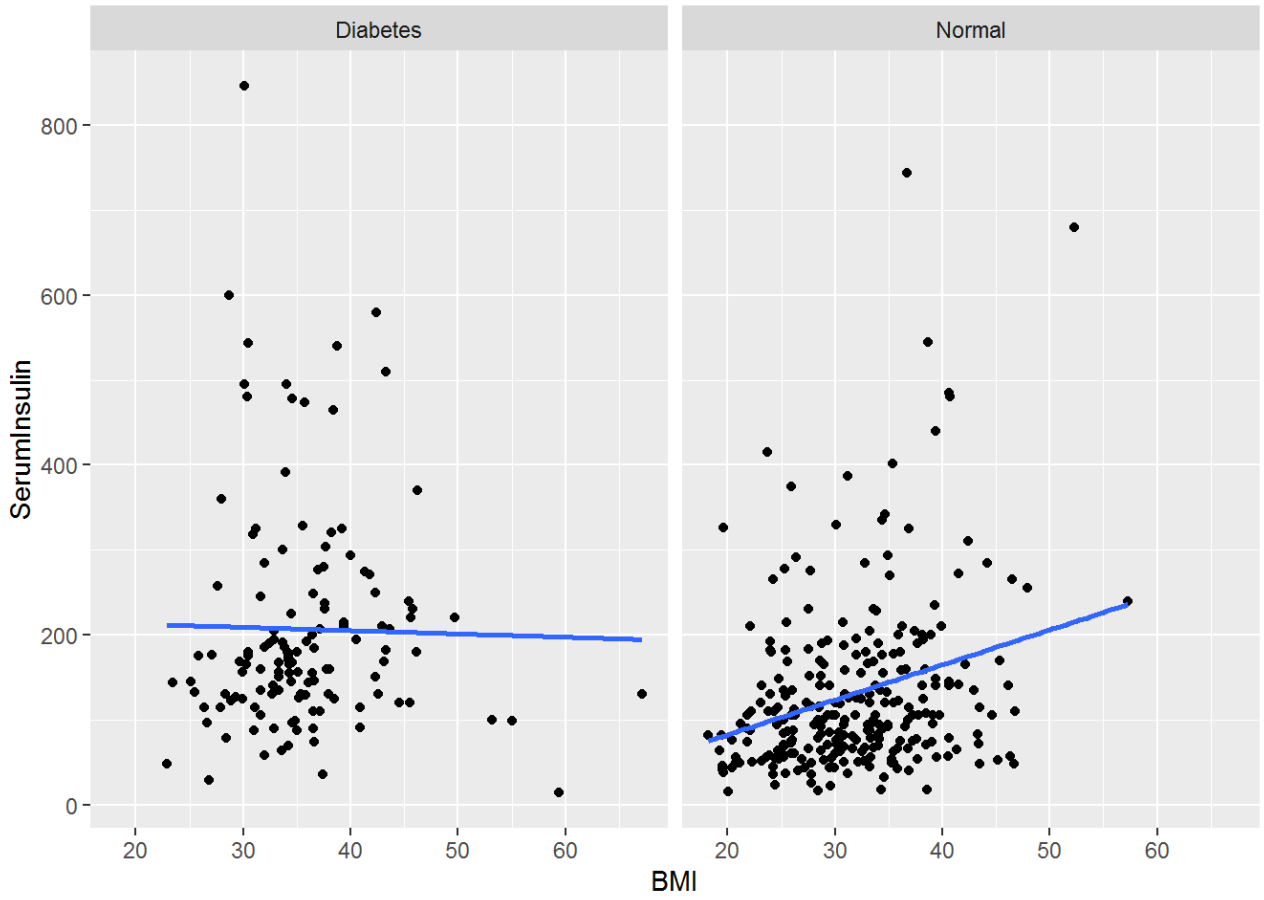

```
ggplot(pima.indian.diabetes4,aes(x=BMI,y=SerumInsulin))+geom_point()+geom_smoot
h(method="lm",se=FALSE)+facet_grid(.~ Group)
```



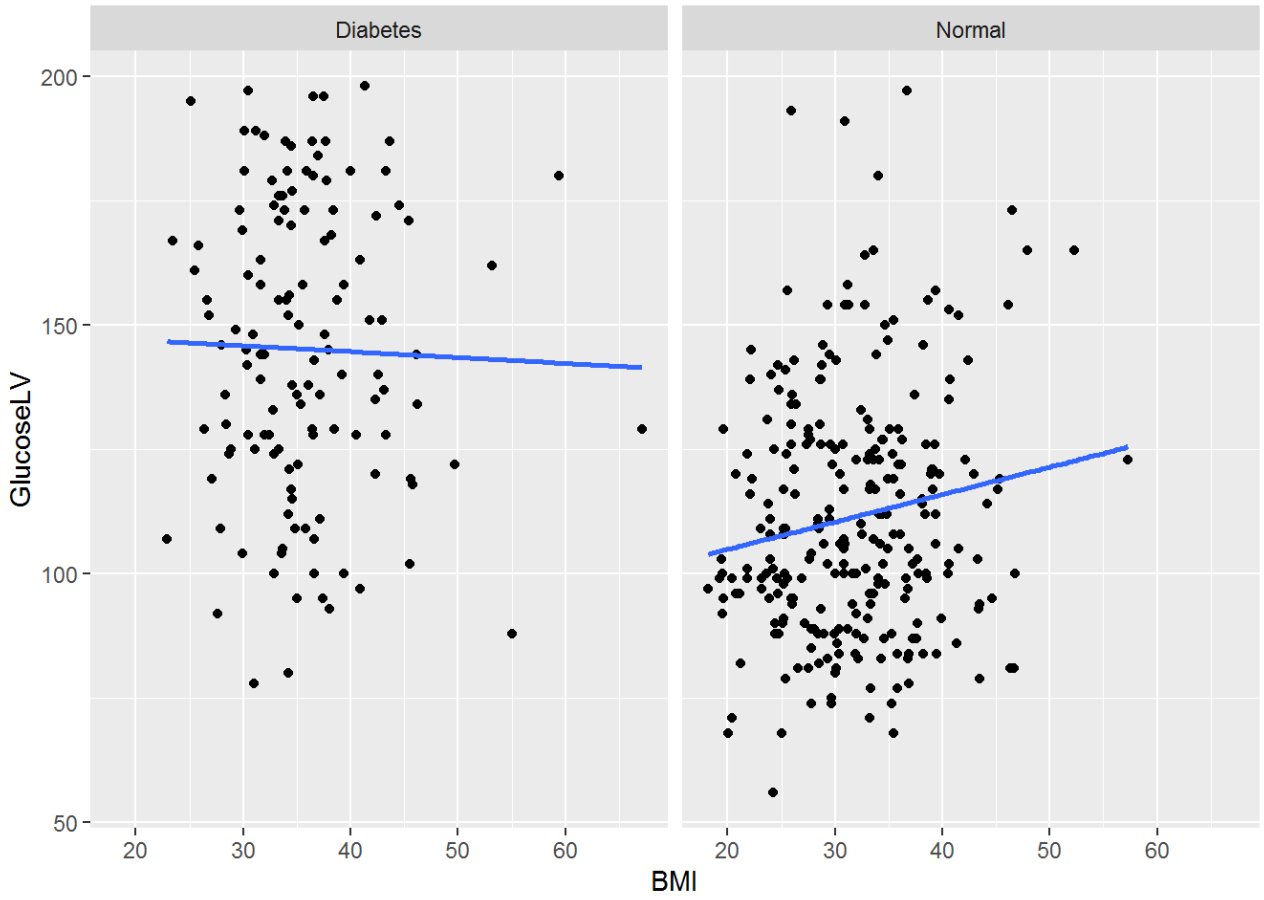

```
ggplot(pima.indian.diabetes4,aes(x=BMI,y=SerumInsulin))+geom_point()+geom_smoot
h(method="lm",se=FALSE)+facet_grid(.~ Group)
```



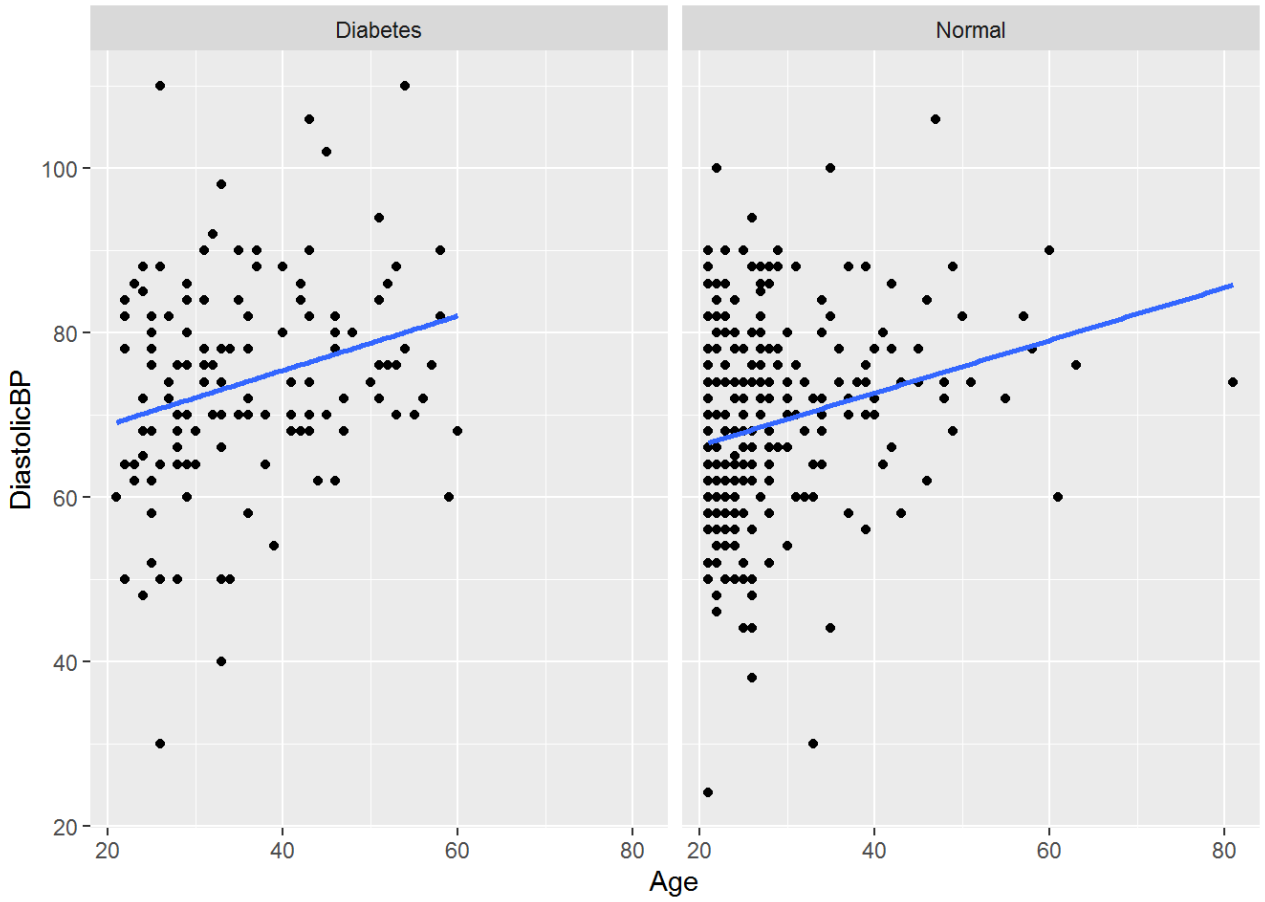

```
ggplot(pima.indian.diabetes4,aes(x=BMI,y=SerumInsulin))+geom_point()+geom_smoot
h(method="lm",se=FALSE)+facet_grid(.~ Group)
```



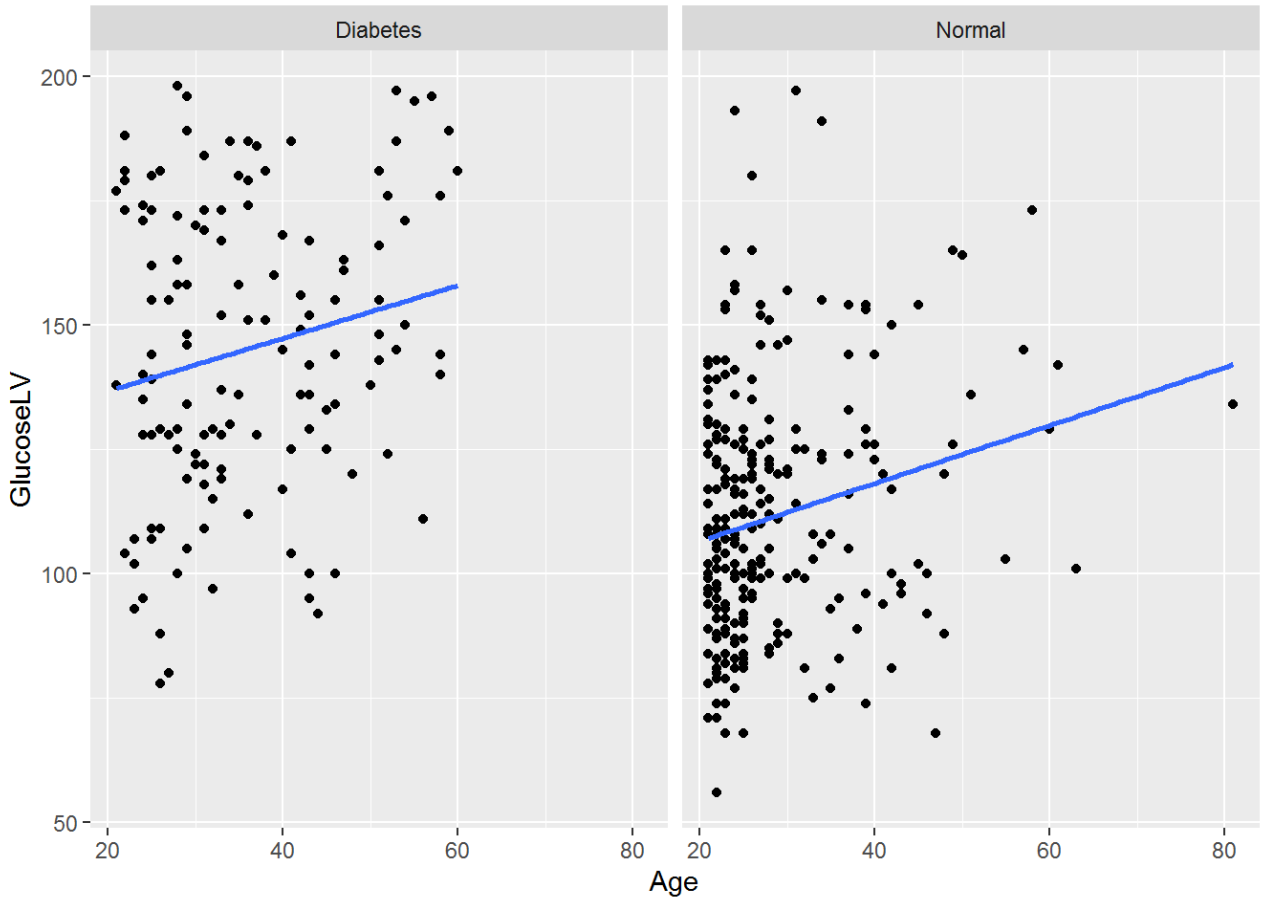

```
ggplot(pima.indian.diabetes4,aes(x=BMI,y=GlucoseLV))+geom_point()+geom_smooth(m
ethod= "lm",se=FALSE)+facet_grid(. ~ Group)
```



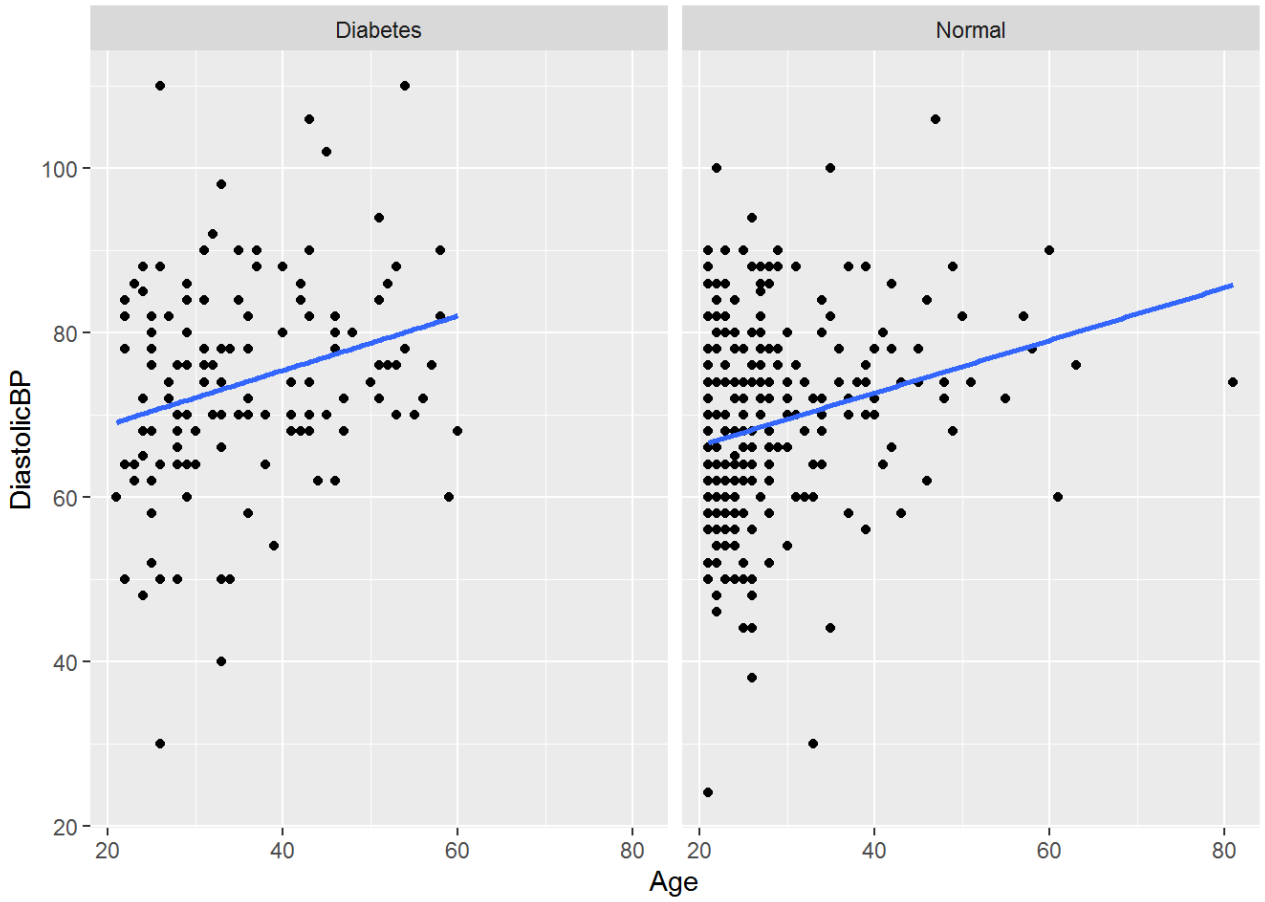

```
ggplot(pima.indian.diabetes4,aes(x=Age,y=DiastolicBP))+geom_point()+geom_smooth
(method = "lm",se=FALSE)+facet_grid(. ~ Group)
```



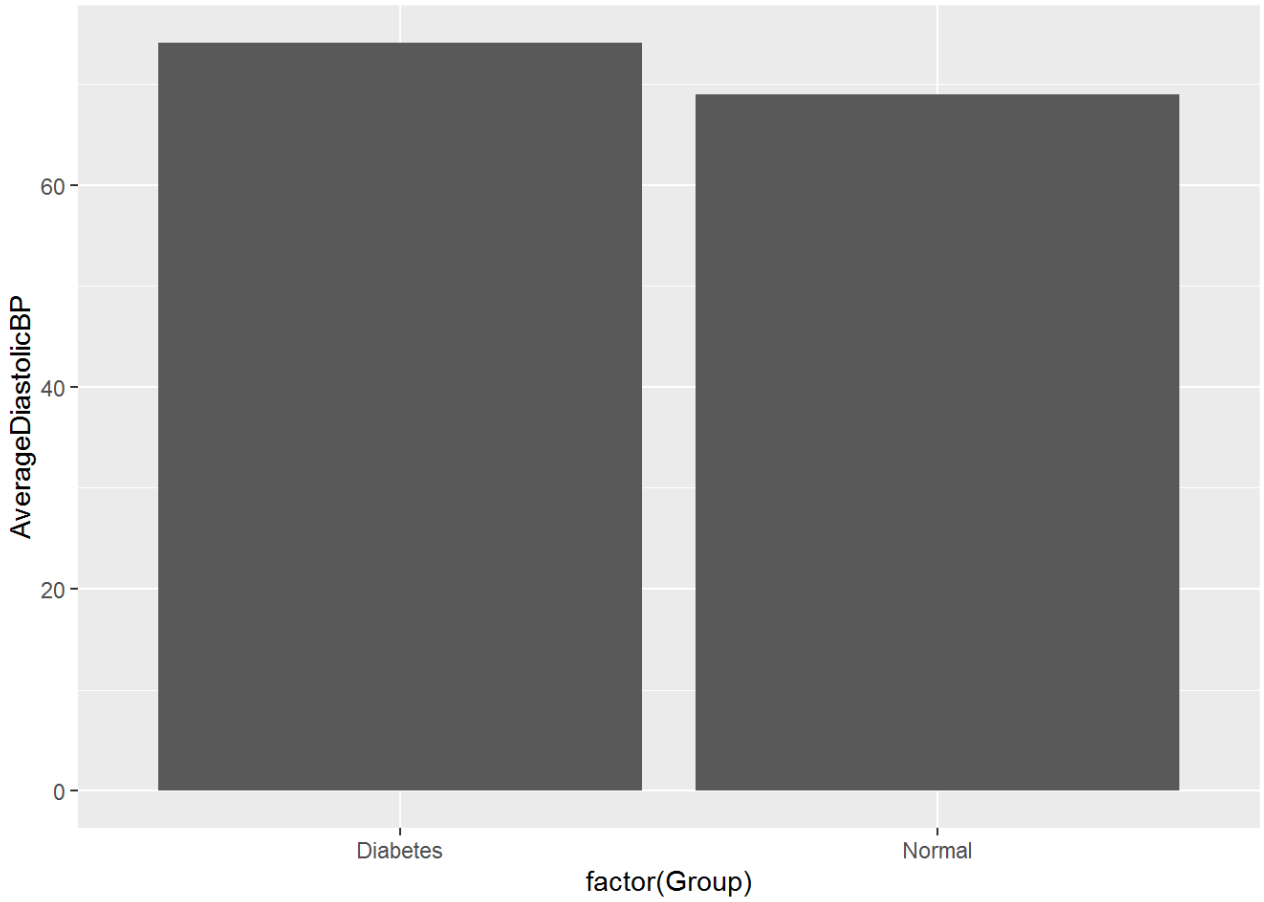

```
ggplot(pima.indian.diabetes4,aes(x=Age,y=GlucoseLV))+geom_point()+geom_smooth(m
ethod= "lm",se=FALSE)+facet_grid(. ~ Group)
```




```
ggplot(pima.indian.diabetes4,aes(x=Age,y=DiastolicBP))+geom_point()+geom_smooth
(method = "lm",se=FALSE)+facet_grid(. ~ Group)
```




```
ggplot(pima.indian.diabetes3,aes(x=factor(Group),y=AverageDiastolicBP))+geom_bar(stat="identity")
```




```
names(pima.indian.diabetes2)[9]<-'Diabetics'
set.seed(88)
split<-sample.split(pima.indian.diabetes2$Diabetics,SplitRatio = 0.75)
split
```

```
##   [1]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [12] FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE
##  [23]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  [34] FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE
##  [45]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
##  [56]  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE
##  [67]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
##  [78] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
##  [89]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## [100]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## [111]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [122]  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
## [133]  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
## [144]  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## [155]  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE
## [166]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE
## [177]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE
## [188] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
## [199] FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
## [210] FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE
## [221]  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
## [232]  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [243]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [254]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## [265]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [276]  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## [287] FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
## [298]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE
## [309]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [320]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
## [331]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
## [342] FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [353]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
## [364]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## [375]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE
## [386]  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE
```

```
pima.indian.diabetes2train<-subset(pima.indian.diabetes2,split==TRUE)
pima.indian.diabetes2test<-subset(pima.indian.diabetes2,split==FALSE)
nrow(pima.indian.diabetes2train)
```

```
## [1] 294
```

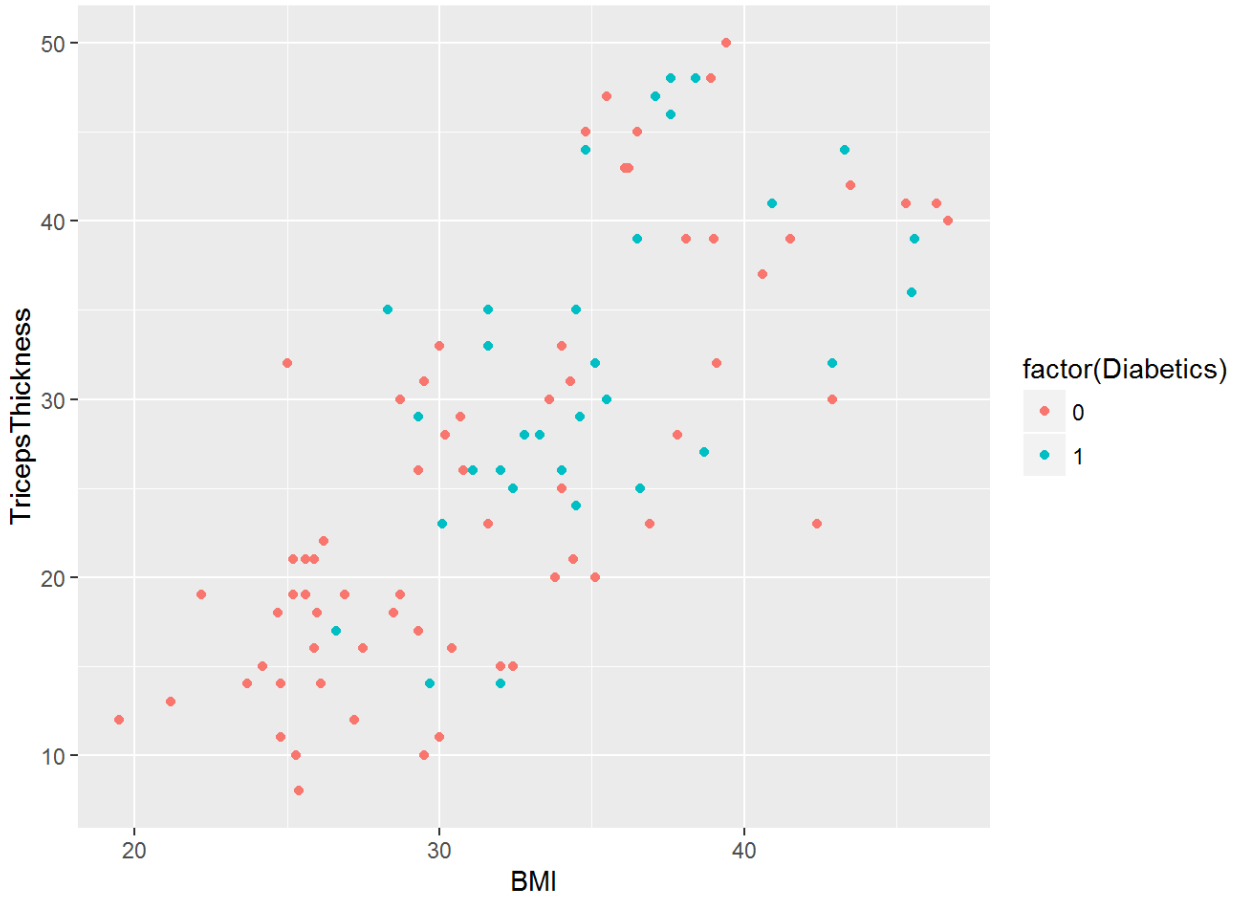
```
nrow(pima.indian.diabetes2test)
```

```
## [1] 98
```

```
Trainlog<-glm(Diabetics~TricepsThickness+BMI+DiastolicBP,data=pima.indian.diabe
tes2train,family=binomial)
summary(Trainlog)
```

```
##
## Call:
## glm(formula = Diabetics ~ TricepsThickness + BMI + DiastolicBP,
##     family = binomial, data = pima.indian.diabetes2train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6724  -0.8828  -0.6676   1.2292   2.0112
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -4.38257    0.91402  -4.795 1.63e-06 ***
## TricepsThickness  0.02709    0.01656   1.636   0.1019
## BMI               0.05312    0.02449   2.169   0.0301 *
## DiastolicBP       0.01506    0.01107   1.361   0.1737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 374.27  on 293  degrees of freedom
## Residual deviance: 347.13  on 290  degrees of freedom
## AIC: 355.13
##
## Number of Fisher Scoring iterations: 4
```

```
ggplot(pima.indian.diabetes2test,aes(x=BMI,y=TricepsThickness,col=factor(Diabet
ics)))+geom_point()
```




```
predictTrain=predict(Trainlog,type ="response")
summary(predictTrain)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1095  0.2261  0.3174  0.3333  0.4224  0.8893
```

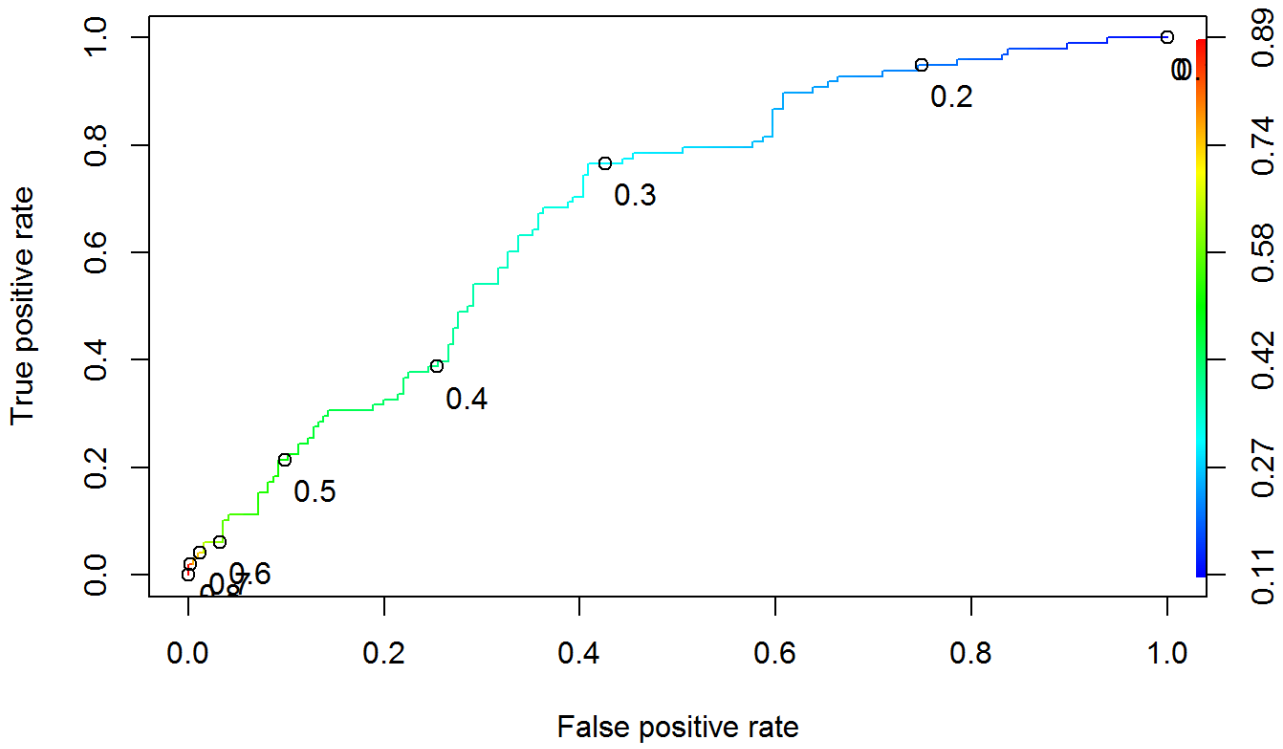
```
tapply(predictTrain,pima.indian.diabetes2train$Diabetics,mean)
```

```
##         0         1
## 0.304273 0.391454
```

```
table(pima.indian.diabetes2train$Diabetics,predictTrain>0.35)
```

```
##
##     FALSE TRUE
##   0   132   64
##   1    41   57
```

```
ROCRpred=prediction(predictTrain,pima.indian.diabetes2train$Diabetics)
ROCRperf=performance(ROCRpred,"tpr","fpr")
plot(ROCRperf,colorize=TRUE,print.cutoffs.at=seq(0,1,0.1),text.adj=c(-0.2,1.8))
```




```
predicttest=predict(Trainlog,type = "response",newdata=pima.indian.diabetes2tes
t)
summary(predicttest)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1256  0.2115  0.2936  0.3224  0.4348  0.6129
```

```
tapply(predicttest,pima.indian.diabetes2test$Diabetics,mean)
```

```
##         0         1
## 0.2919504 0.3851939
```

```
table(pima.indian.diabetes2test$Diabetics,predicttest>0.25)
```

```
##
##      FALSE TRUE
##    0    32   34
##    1     3   29
```

```
ROCRpred=prediction(predicttest,pima.indian.diabetes2test$Diabetics)
ROCRperf=performance(ROCRpred,"tpr","fpr")
plot(ROCRperf,colorize=TRUE,print.cutoffs.at=seq(0,1,0.1),text.adj=c(-0.2,1.8))
```