Data visualization

b327zhan

7/12/2021

(This pdf is gonna show some demo)

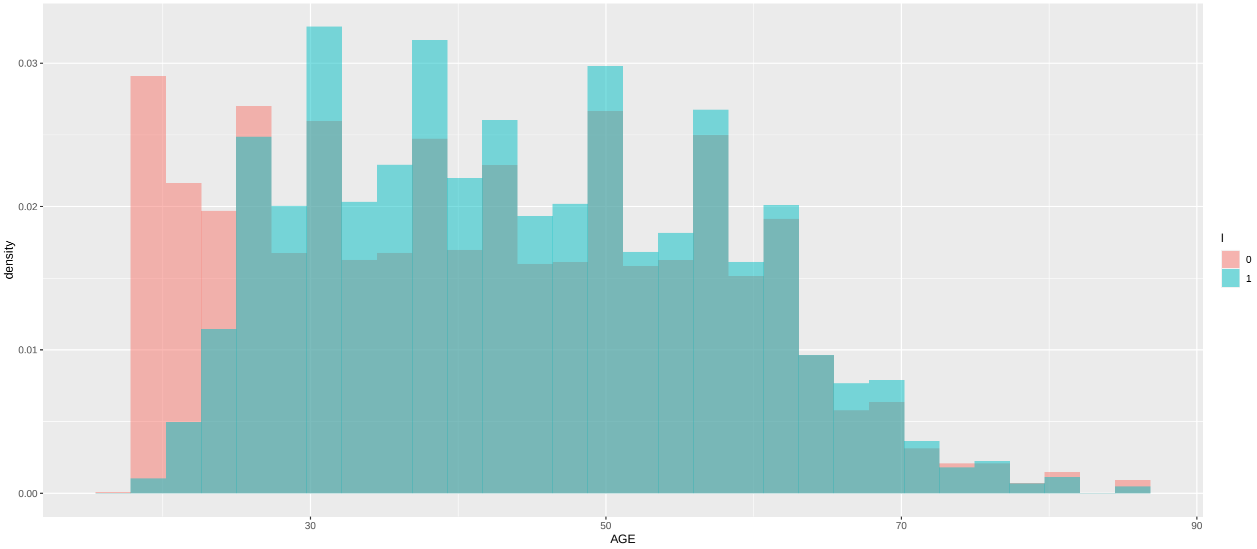
Data visualization

```
library(ggplot2)
X = read.csv('train_xx.csv')
Y = read.csv('train_y.csv')
l = sapply(Y$label, as.character)
data = data.frame(X,Y$label,l)
```

Generate a graph with x-axis being the age period, and condition on only showing those with positive WAGP

Red one with label 0; Blue one with label 1.

```
library(cowplot)
data_copy = data
# Condition
data_copy1 = data_copy[data$WAGP > 0
                      ,]
ggplot(data_copy1, aes(AGE, fill = l)) +
  geom_histogram(alpha = 0.5, aes(y = ..density..), position = 'identity',bins=30)
```



Generate a graph with x-axis being the age period, and condition on only showing those with positive WAGP

Red one with label 0; Blue one with label 1.

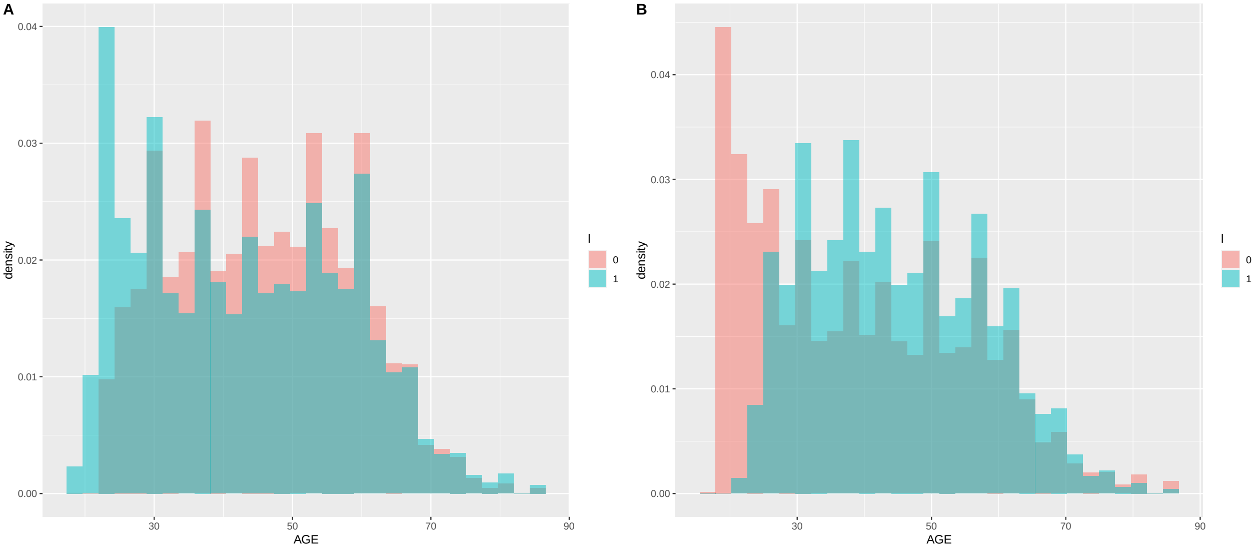
Left one is incorrectly fitted

Right one is correctly fitted

```
library(cowplot)
par(mfrow = c(1,2))
data_copy = data
data_copy1 = data_copy[data$correct==0
                       & data$WAGP > 0
                      ,]
iris1 = ggplot(data_copy1, aes(AGE, fill = l)) +
  geom_histogram(alpha = 0.5, aes(y = ..density..), position = 'identity',bins=30)

data_copy = data
data_copy2 = data_copy[data$correct==1
                       & data$WAGP > 0
                      ,]
iris2 = ggplot(data_copy2, aes(AGE, fill = l)) +
  geom_histogram(alpha = 0.5, aes(y = ..density..), position = 'identity',bins=30)

plot_grid(iris1, iris2, labels = "AUTO")
```



This one detects correlation.

Data correlation

```
# Should be dropped
library(caret)
```

```
## Loading required package: lattice
```

```
Cor = function(dataset) {
  print(names(dataset)[findCorrelation(cor(dataset))])
}
Cor(X)
```

```
##  [1] "NET"    "AGE_N"  "N_I"    "INCOME" "S_I"    "N_N"    "S_N"    "AGE_I"
##  [9] "NS_I"   "ASSIST" "CITSHP"
```