Prediction of Loan Defaulters

sammy waiyaki

17/10/2019

importing data

importation of bank data set for analysis

```
bank_data <- read.csv(file.choose(), header = T)
View(bank_data)
attach(bank_data)
```

exploration on the data

```
str(bank_data)
```

```
## 'data.frame':    1118 obs. of  13 variables:
##  $ X            : int  0 2 5 6 7 9 11 13 14 15 ...
##  $ branch       : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ no_customer  : int  3017 3017 3017 3017 3017 3017 3017 3017 3017 3017 ...
##  $ customer     : int  10012 10030 10071 10096 10128 10140 10169 10200 10218 10234 ...
##  $ age          : int  28 40 35 26 25 21 30 18 53 18 ...
##  $ eduction_level: Factor w/ 5 levels "College degree",..: 3 2 2 5 2 5 1 3 2 3 ...
##  $ employ       : int  7 20 2 2 4 0 4 0 9 0 ...
##  $ address      : int  2 12 9 4 2 0 3 0 13 0 ...
##  $ income       : int  44 61 38 38 30 23 39 35 41 15 ...
##  $ debtinc      : num  17.7 4.8 10.9 11.9 14.4 3.9 10.6 3.9 13.3 7.4 ...
##  $ creddebt     : num  2.99 1.04 1.46 0.95 1.05 0.31 2.39 0.17 2.33 0.83 ...
##  $ othdebt      : num  4.8 1.89 2.68 3.57 3.27 0.59 1.74 1.19 3.12 0.28 ...
##  $ default      : Factor w/ 2 levels "No","Yes": 1 1 2 2 1 1 2 1 1 2 ...
```

```
dim(bank_data)
```

```
## [1] 1118   13
```

```
Na_table <- table(is.na(bank_data))
Na_table
```

```
##
## FALSE
## 14534
```

summary statistics on the data

```
summary(bank_data)
```

```
##        X              branch        no_customer      customer
##  Min.   :   0.0   Min.   : 3.00   Min.   :1919   Min.   : 10012
##  1st Qu.: 390.2   1st Qu.:20.00   1st Qu.:2658   1st Qu.: 99390
##  Median : 766.5   Median :64.00   Median :3491   Median :316285
##  Mean   : 765.7   Mean   :53.07   Mean   :3481   Mean   :262067
##  3rd Qu.:1150.8   3rd Qu.:75.00   3rd Qu.:4358   3rd Qu.:371422
##  Max.   :1499.0   Max.   :91.00   Max.   :4809   Max.   :453777
```

```
##       age                              eduction_level     employ
##  Min.   :18.00   College degree               :236    Min.   : 0.000
##  1st Qu.:22.00   Did not complete high school:171    1st Qu.: 0.000
##  Median :28.00   High school degree           :399    Median : 2.000
##  Mean   :29.57   Post-undergraduate degree    : 58    Mean   : 4.045
##  3rd Qu.:36.00   Some college                 :254    3rd Qu.: 6.000
##  Max.   :53.00                                        Max.   :20.000
##     address           income          debtinc          creddebt
##  Min.   : 0.000   Min.   : 12.00   Min.   : 0.000   Min.   :0.000
##  1st Qu.: 1.000   1st Qu.: 25.00   1st Qu.: 4.400   1st Qu.:0.350
##  Median : 3.000   Median : 35.00   Median : 7.800   Median :0.775
##  Mean   : 4.255   Mean   : 41.42   Mean   : 8.408   Mean   :1.121
##  3rd Qu.: 7.000   3rd Qu.: 50.75   3rd Qu.:11.900   3rd Qu.:1.510
##  Max.   :15.000   Max.   :153.00   Max.   :19.900   Max.   :6.360
##     othdebt        default
##  Min.   : 0.000   No :710
##  1st Qu.: 0.890   Yes:408
##  Median : 1.735
##  Mean   : 2.296
##  3rd Qu.: 3.228
##  Max.   :11.770
```

let us drop unwanted columns

we use dplyr which is a data wrangling package in r

```r
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
```

```r
bank_data1 <-select(bank_data, -c(1,2,3,4))
View(bank_data1)
```

Exploratory data analysis using visualizations

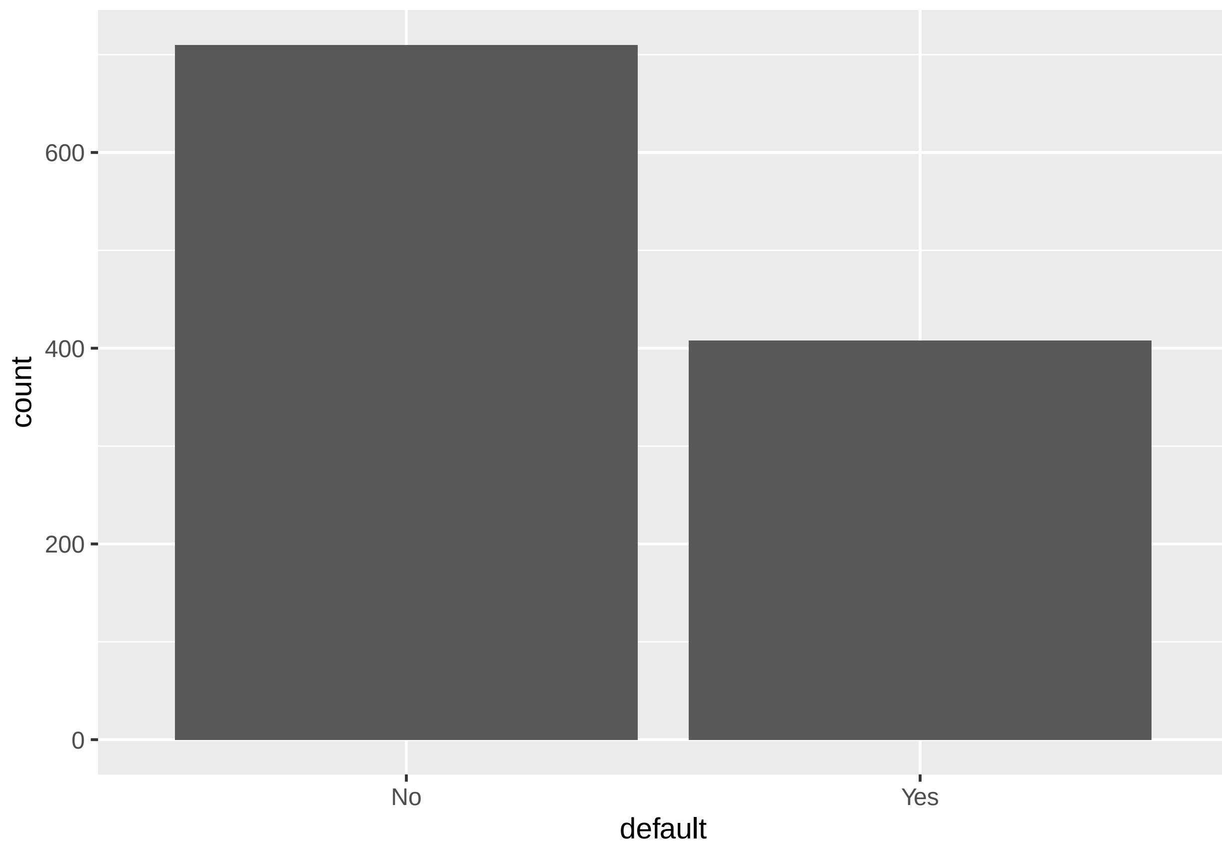
This is important for drawing insights from the data ggplot library is of essence to produce nice visualizations

```r
library(ggplot2)
```

A quick look on the distribution of the defaul which is our our target variable.

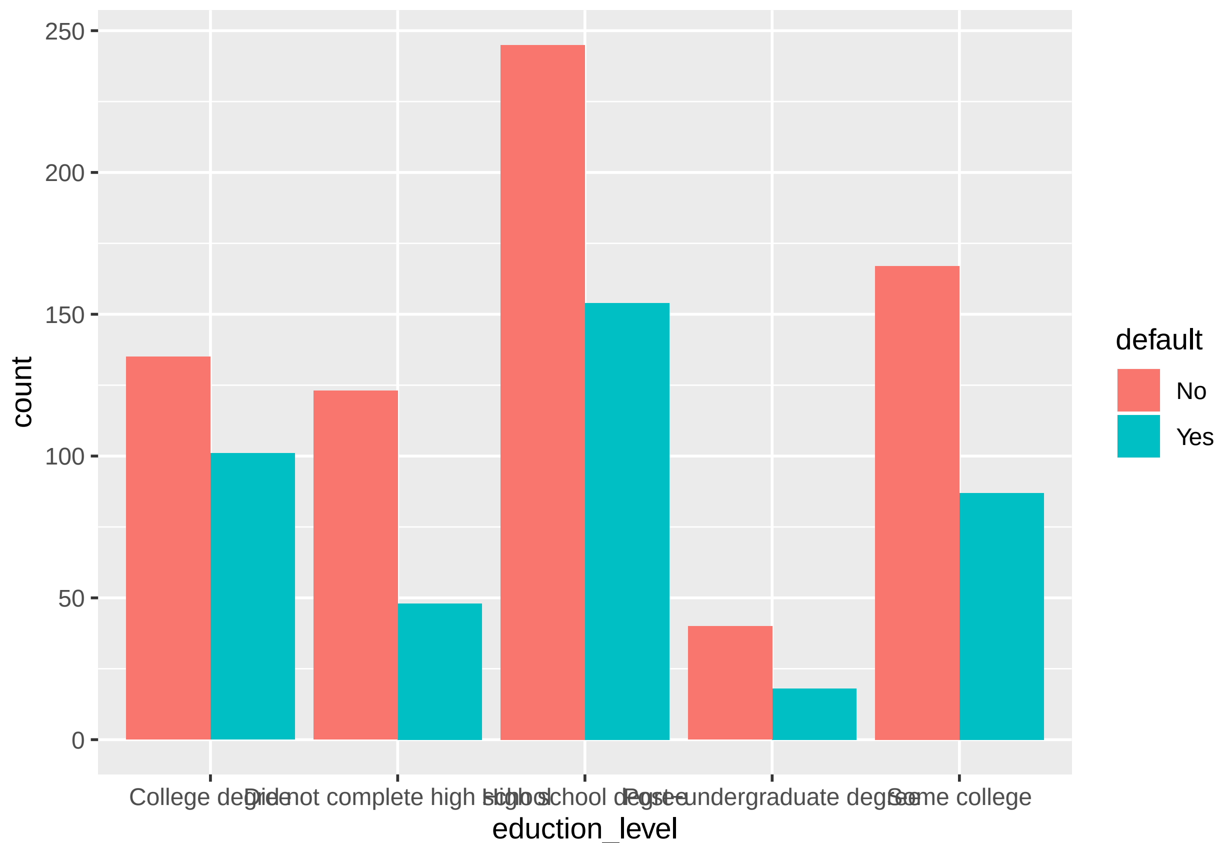
We can see that the number of non defaulters (Yes) is greater that defaulters(No)

```r
ggplot(data = bank_data1) +
  geom_bar(aes(x = default))
```



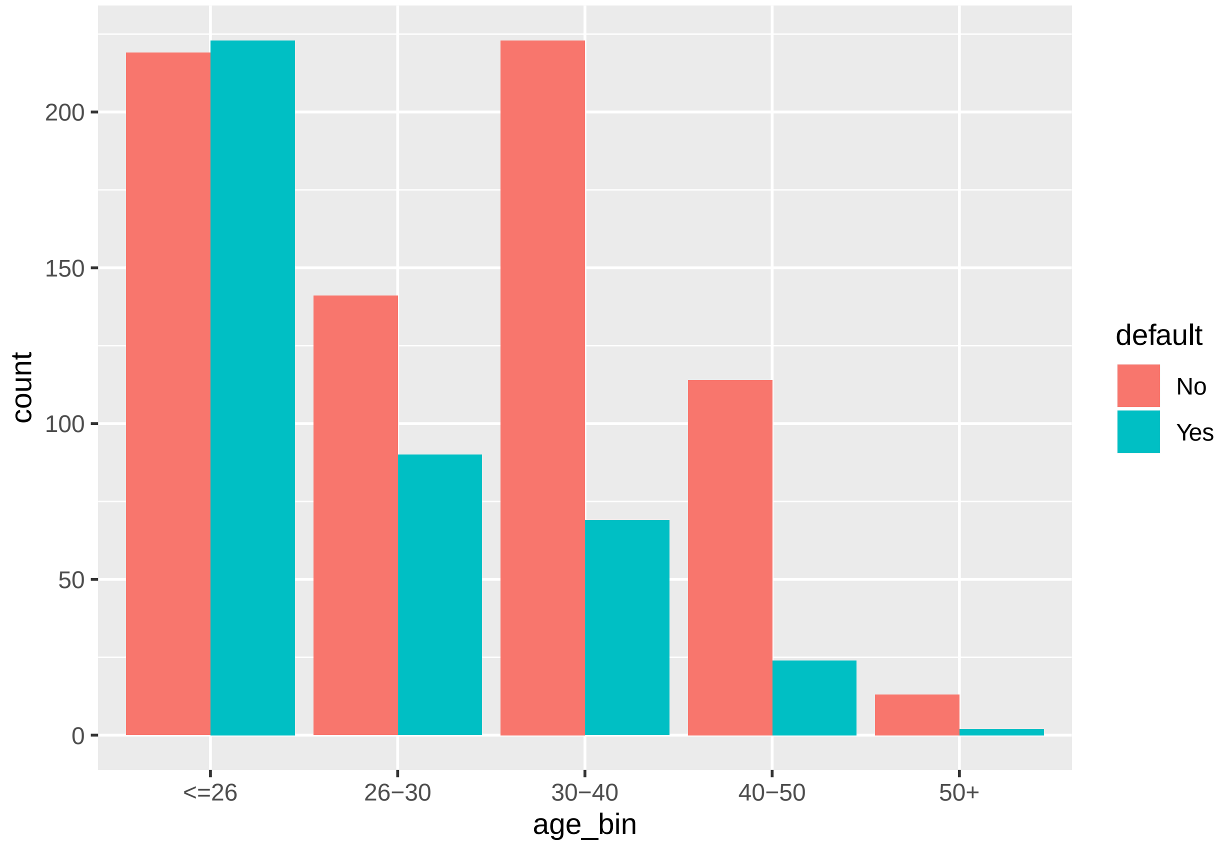
distribution of eduation level with target variable We can see that people with high school degree defaulted than those with post-graduate degree

```
ggplot(bank_data1, aes(eduction_level, ..count..)) + geom_bar(aes(fill = default), position = "dodge")
```



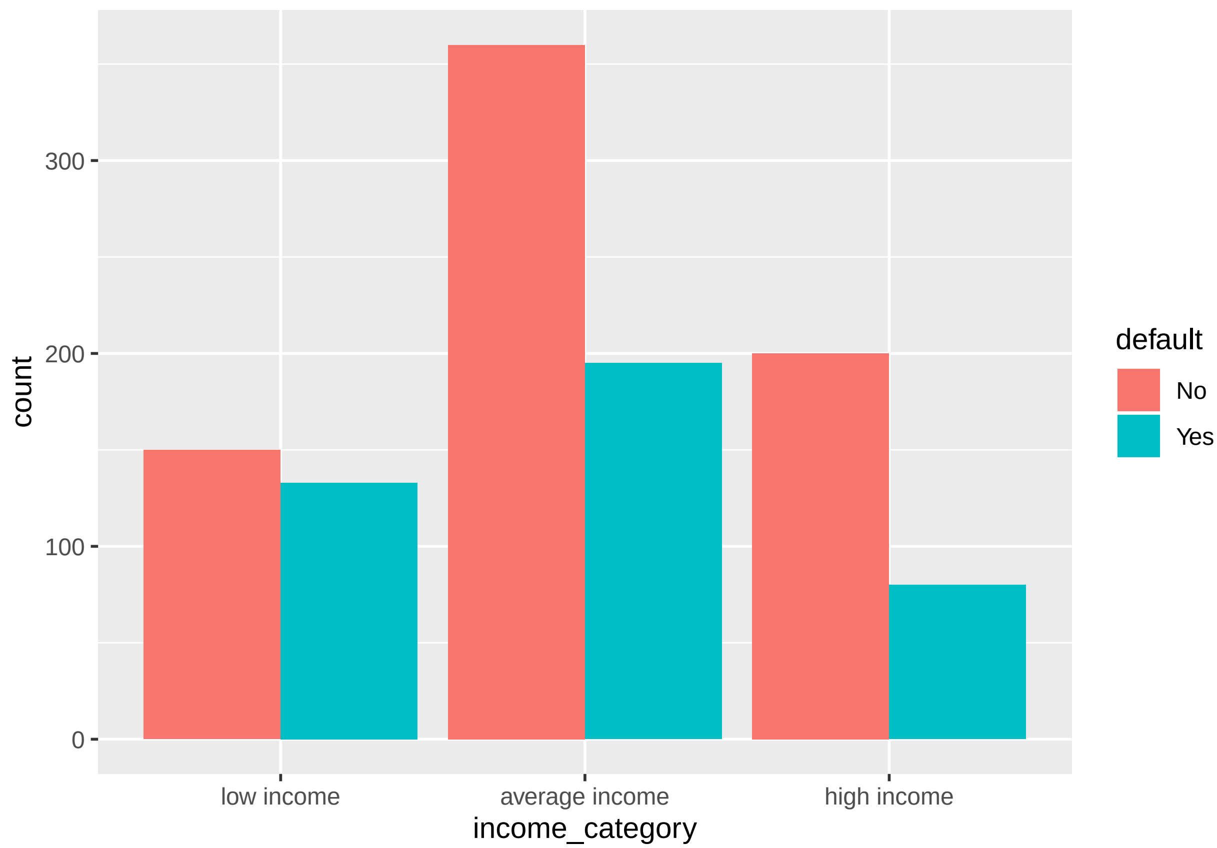
Age groups to visualize their loan repayment behavior we can see that people with less than 26 years defaulted more than those with advance ages

```
bins <- c(0,25,30,40,50,60)
age_cat <- c('<=26','26-30','30-40','40-50','50+')
bank_data1$age_bin <- cut(age, labels = age_cat, breaks = bins)

ggplot(bank_data1, aes(age_bin, ..count..)) + geom_bar(aes(fill = default), position = "dodge")
```



creating income categories to visualize their loan repayment behavior we can see that people under the income category of low income defaulted the most than those in other categories

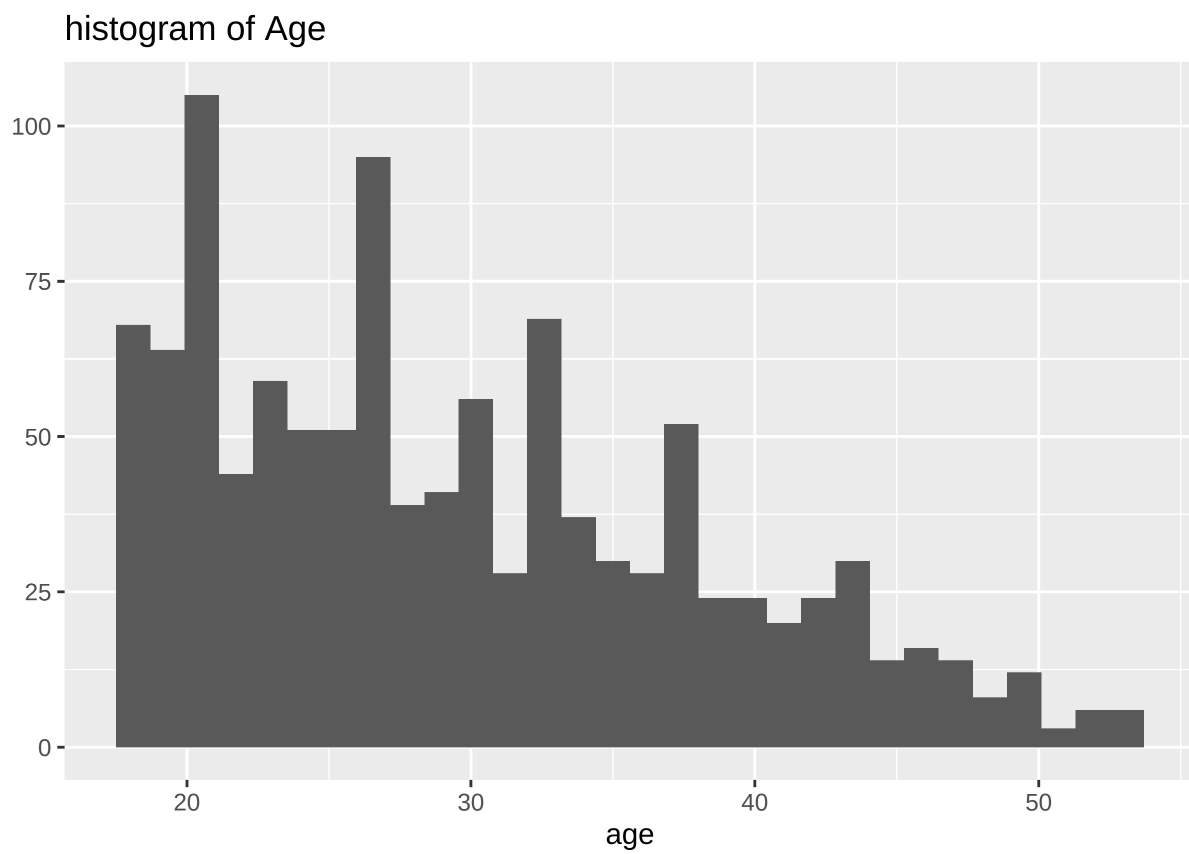
```
income_bins = c(0,25,50,153)
income_cat <- c('low income', 'average income', 'high income')
bank_data1$income_category <-  cut(income, labels = income_cat, breaks = income_bins)

ggplot(bank_data1, aes(income_category, ..count..)) + geom_bar(aes(fill = default), position = "dodge")
```



Distribution of numerical variables ### Histograms to visualize distributions of age and income

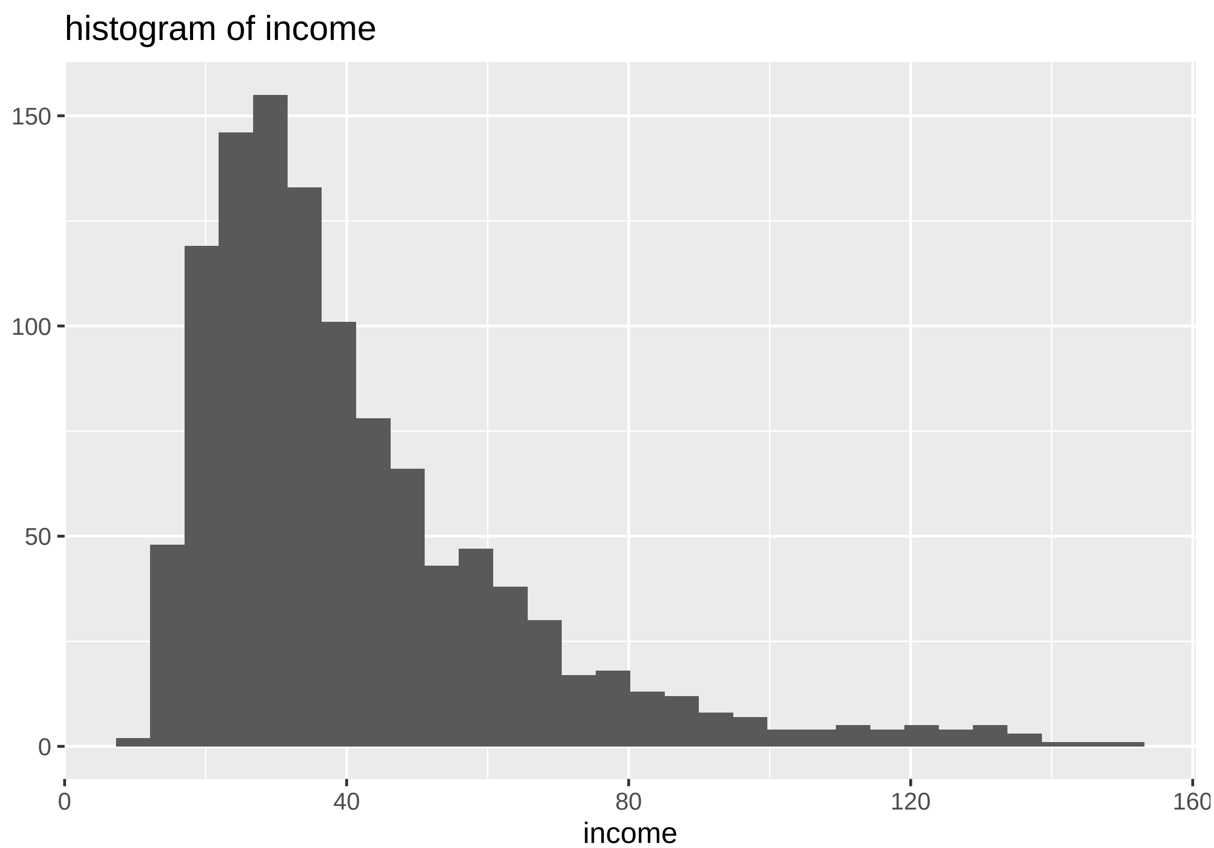
```
qplot(age, geom = "histogram", main = 'histogram of Age')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



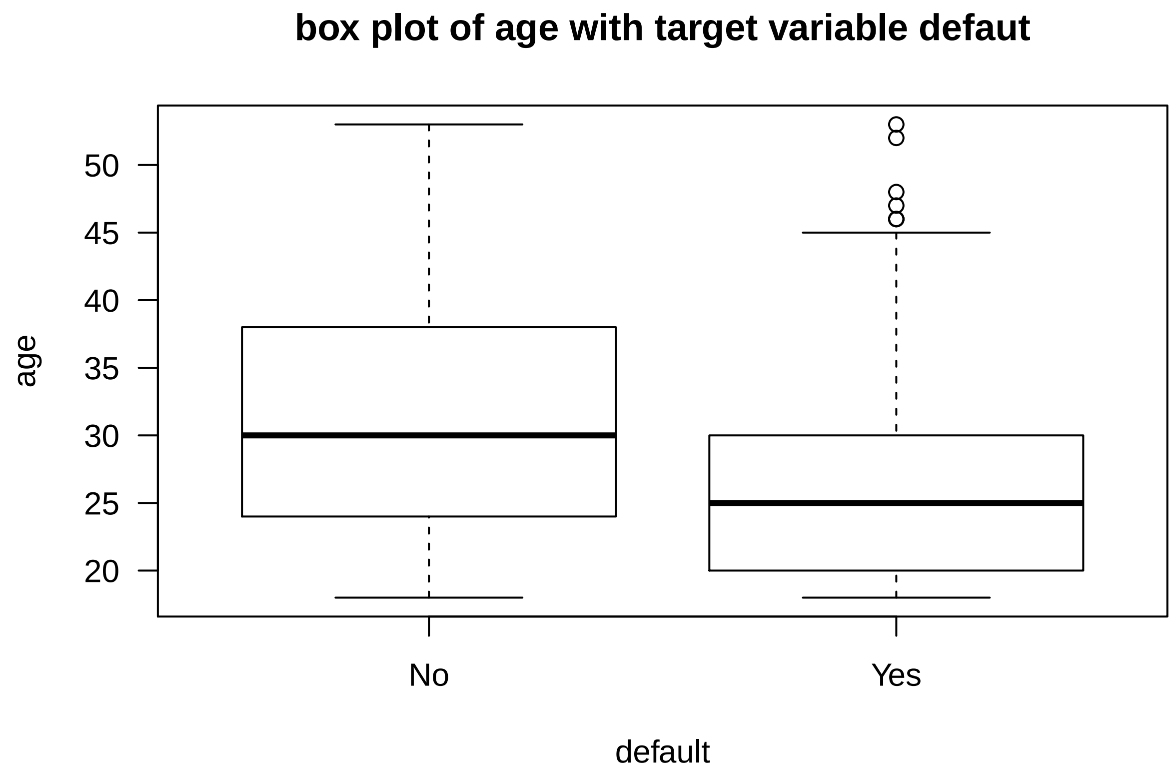

```
qplot(income, geom = 'histogram', main = 'histogram of income')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

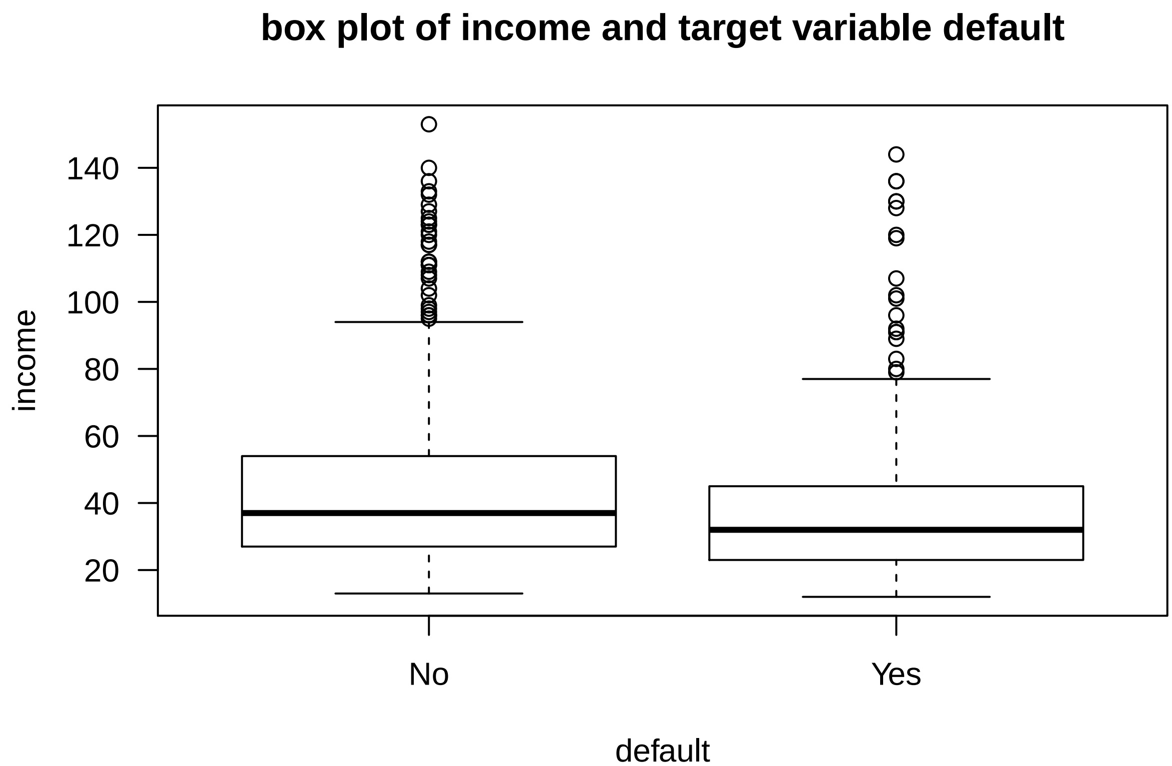


box plots of income and age brouped by default status we can see that people with low mean age defaulted than those with high mean age. The same applies to income.

```
boxplot(age ~ default, main='box plot of age with target variable defaut', las = 1)
```




```
boxplot(income ~ default, main='box plot of income and target variable default', las=1)
```



droping the age and income bins created for visualization purposes

```
bank_data2<- select(bank_data1, -c(10, 11))
View(bank_data2)
```

building logistic regression model to predict defaulters

Logistic regression is a classification algorithm for dichotomous vaiable or binary such as 'Yes' and 'No' or '0' and '1' libraries such as caret are very fundamental in building predictive machine learning model

```
library(caret)
```

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

```
library(klaR)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

creating traning and testing data sets

train set enables the algorithm learn about patterns within data while the testing set is used to evaluate perfomance of the model in classifying the defaulters and non-defaulters

```r
trainIndex <- createDataPartition(bank_data2$default, p=0.80, list=FALSE)
train_set <- bank_data2[ trainIndex,]
test_set <- bank_data2[-trainIndex,]
```

train logistic regression using training set and summary of the model

```r
fit <- glm(default~., data=train_set, family = binomial(link = 'logit'))
summary(fit)
```

```
##
## Call:
## glm(formula = default ~ ., family = binomial(link = "logit"),
##     data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9689  -0.8510  -0.4085   0.9284   2.5545
##
## Coefficients:
##                                           Estimate Std. Error z value
## (Intercept)                              -0.781759   0.669976  -1.167
## age                                      -0.026015   0.025758  -1.010
## eduction_levelDid not complete high school -0.022645   0.305282  -0.074
## eduction_levelHigh school degree         -0.173886   0.226694  -0.767
## eduction_levelPost-undergraduate degree    -0.997749   0.423007  -2.359
## eduction_levelSome college               -0.432526   0.241756  -1.789
## employ                                   -0.277090   0.035264  -7.858
## address                                   0.007798   0.057303   0.136
## income                                    0.012099   0.008833   1.370
## debtinc                                   0.152327   0.040357   3.774
## creddebt                                  0.389107   0.135772   2.866
## othdebt                                  -0.090571   0.103437  -0.876
##                                          Pr(>|z|)
## (Intercept)                               0.24327
## age                                       0.31251
## eduction_levelDid not complete high school  0.94087
## eduction_levelHigh school degree          0.44305
## eduction_levelPost-undergraduate degree     0.01834 *
## eduction_levelSome college                0.07360 .
## employ                                   3.92e-15 ***
## address                                   0.89176
## income                                    0.17076
## debtinc                                   0.00016 ***
## creddebt                                  0.00416 **
## othdebt                                   0.38124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1175.0  on 894  degrees of freedom
## Residual deviance:  924.1  on 883  degrees of freedom
## AIC: 948.1
##
```

```
## Number of Fisher Scoring iterations: 5
```

make predictions based on the test set and visualizing classified classes using confusion matrix

```r
probabilities <- predict(fit, test_set[,1:8,], type = 'response')
predictions <- ifelse(probabilities > 0.5,'Yes','No')
# summarize results
table2 <- table(predictions, test_set$default)
table2
```

```
##
## predictions  No Yes
##         No  122  39
##         Yes  20  42
```

accuracy of the logistic_modelprediction

The model predicted 74.44% default classes accurately.

```r
accuracy <- (127+39)/(127+39+42+15)
accuracy
```

```
## [1] 0.7443946
```