DATA 606 Final Project

Forhad Akbar

12/04/2019

Libraries

```
library(tidyverse)
library(caTools)
library(ROCR)
library(rpart)
library(rmdformats)
library(randomForest)
```

Introduction

Research question
About Company: Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan.
Problem: Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers.

Data

This data source was given as part of a data science challenge or practice problem. I downloaded the data and loaded to my git-hub account. I will read the data into R from my git-hub account using raw link of the csv file using read.csv command.

Source: https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/

```
# load data
my_loan_data<- read.csv("https://raw.githubusercontent.com/forhadakbar/data606fall2019stat/master/Final
```

```
head(my_loan_data)
```

```
##    Loan_ID Gender Married Dependents    Education Self_Employed
## 1 LP001002   Male      No          0     Graduate            No
## 2 LP001003   Male     Yes          1     Graduate            No
## 3 LP001005   Male     Yes          0     Graduate           Yes
## 4 LP001006   Male     Yes          0 Not Graduate            No
## 5 LP001008   Male      No          0     Graduate            No
## 6 LP001011   Male     Yes          2     Graduate           Yes
##   ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term
## 1            5849                 0         NA              360
```

```
## 2            4583              1508        128               360
## 3            3000                 0         66               360
## 4            2583              2358        120               360
## 5            6000                 0        141               360
## 6            5417              4196        267               360
##   Credit_History Property_Area Loan_Status
## 1              1         Urban           Y
## 2              1         Rural           N
## 3              1         Urban           Y
## 4              1         Urban           Y
## 5              1         Urban           Y
## 6              1         Urban           Y
```

```r
dim(my_loan_data)
```

```
## [1] 614  13
```

There are 614 cases and 13 columns. Each case or observation represent a loan application.

Exploratory Data Analysis & Inference

Dependent Variable

Loan_Status is the response variable. It is a categorical variable which gives us yes and no for loan approval status.

Independent Variable

I have few independent variables that i will consider for now. I will choose the most appropiate variables after doing exploratory analysis.

Applicants took a loan before. Credit history is the variable which answers that.
Applicants with higher incomes. So, we might look at the applicant income variable.
Applicants with higher education.
Gender of the applicant.
Number of Dependens an applicant has.
Property area contains location information of the loan property applied for.

Relevant summary statistics

```r
str(my_loan_data)
```

```
## 'data.frame':    614 obs. of  13 variables:
##  $ Loan_ID          : Factor w/ 614 levels "LP001002","LP001003",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender           : Factor w/ 3 levels "","Female","Male": 3 3 3 3 3 3 3 3 3 3 ...
##  $ Married          : Factor w/ 3 levels "","No","Yes": 2 3 3 3 2 3 3 3 3 3 ...
##  $ Dependents       : Factor w/ 5 levels "","0","1","2",..: 2 3 2 2 2 4 2 5 4 3 ...
##  $ Education        : Factor w/ 2 levels "Graduate","Not Graduate": 1 1 1 2 1 1 2 1 1 1 ...
##  $ Self_Employed    : Factor w/ 3 levels "","No","Yes": 2 2 3 2 2 3 2 2 2 2 ...
##  $ ApplicantIncome  : int  5849 4583 3000 2583 6000 5417 2333 3036 4006 12841 ...
```

```
##  $ CoapplicantIncome: num  0 1508 0 2358 0 ...
##  $ LoanAmount       : int  NA 128 66 120 141 267 95 158 168 349 ...
##  $ Loan_Amount_Term : int  360 360 360 360 360 360 360 360 360 360 ...
##  $ Credit_History   : int  1 1 1 1 1 1 1 0 1 1 ...
##  $ Property_Area    : Factor w/ 3 levels "Rural","Semiurban",..: 3 1 3 3 3 3 3 2 3 2 ...
##  $ Loan_Status      : Factor w/ 2 levels "N","Y": 2 1 2 2 2 2 2 1 2 1 ...
```

```r
summary(my_loan_data)
```

```
##      Loan_ID       Gender    Married   Dependents        Education  
##  LP001002:  1         : 13      :  3     : 15     Graduate    :480  
##  LP001003:  1   Female:112   No :213   0 :345     Not Graduate:134  
##  LP001005:  1   Male  :489   Yes:398   1 :102                       
##  LP001006:  1                          2 :101                       
##  LP001008:  1                          3+: 51                       
##  LP001011:  1                                                       
##  (Other) :608                                                       
##  Self_Employed ApplicantIncome CoapplicantIncome   LoanAmount   
##     : 32       Min.   :  150   Min.   :    0     Min.   :  9.0  
##  No :500       1st Qu.: 2878   1st Qu.:    0     1st Qu.:100.0  
##  Yes: 82       Median : 3812   Median : 1188     Median :128.0  
##                Mean   : 5403   Mean   : 1621     Mean   :146.4  
##                3rd Qu.: 5795   3rd Qu.: 2297     3rd Qu.:168.0  
##                Max.   :81000   Max.   :41667     Max.   :700.0  
##                                                  NA's   :22     
##  Loan_Amount_Term Credit_History     Property_Area Loan_Status
##  Min.   : 12      Min.   :0.0000   Rural    :179   N:192      
##  1st Qu.:360      1st Qu.:1.0000   Semiurban:233   Y:422      
##  Median :360      Median :1.0000   Urban    :202              
##  Mean   :342      Mean   :0.8422                              
##  3rd Qu.:360      3rd Qu.:1.0000                              
##  Max.   :480      Max.   :1.0000                              
##  NA's   :14       NA's   :50
```

Data Cleaning

LoanAmount variable has 22 Null Value -Loan_Amount_Term has 14 null values -Credit_History has 50 Null values Data set observation.

```r
#Store backup before removing missing values
my_loan_data_backup <- my_loan_data

#Retrun all rows with missing values
my_loan_data[!complete.cases(my_loan_data),]
```

```
##      Loan_ID Gender Married Dependents    Education Self_Employed
## 1   LP001002   Male      No          0     Graduate            No
## 17  LP001034   Male      No          1 Not Graduate            No
## 20  LP001041   Male     Yes          0     Graduate              
## 25  LP001052   Male     Yes          1     Graduate              
## 31  LP001091   Male     Yes          1     Graduate              
## 36  LP001106   Male     Yes          0     Graduate            No
```

```
## 37  LP001109   Male     Yes          0     Graduate            No
## 43  LP001123   Male     Yes          0     Graduate            No
## 45  LP001136   Male     Yes          0 Not Graduate           Yes
## 46  LP001137 Female      No          0     Graduate            No
## 64  LP001213   Male     Yes          1     Graduate            No
## 74  LP001250   Male     Yes         3+ Not Graduate            No
## 80  LP001264   Male     Yes         3+ Not Graduate           Yes
## 82  LP001266   Male     Yes          1     Graduate           Yes
## 84  LP001273   Male     Yes          0     Graduate            No
## 87  LP001280   Male     Yes          2 Not Graduate            No
## 96  LP001326   Male      No          0     Graduate
## 103 LP001350   Male     Yes                Graduate            No
## 104 LP001356   Male     Yes          0     Graduate            No
## 113 LP001391   Male     Yes          0 Not Graduate            No
## 114 LP001392 Female      No          1     Graduate           Yes
## 118 LP001405   Male     Yes          1     Graduate            No
## 126 LP001443 Female      No          0     Graduate            No
## 128 LP001449   Male      No          0     Graduate            No
## 130 LP001465   Male     Yes          0     Graduate            No
## 131 LP001469   Male      No          0     Graduate           Yes
## 157 LP001541   Male     Yes          1     Graduate            No
## 166 LP001574   Male     Yes          0     Graduate            No
## 182 LP001634   Male      No          0     Graduate            No
## 188 LP001643   Male     Yes          0     Graduate            No
## 198 LP001669 Female      No          0 Not Graduate            No
## 199 LP001671 Female     Yes          0     Graduate            No
## 203 LP001682   Male     Yes         3+ Not Graduate            No
## 220 LP001734 Female     Yes          2     Graduate            No
## 224 LP001749   Male     Yes          0     Graduate            No
## 233 LP001770   Male      No          0 Not Graduate            No
## 237 LP001786   Male     Yes          0     Graduate
## 238 LP001788 Female      No          0     Graduate           Yes
## 260 LP001864   Male     Yes         3+ Not Graduate            No
## 261 LP001865   Male     Yes          1     Graduate            No
## 280 LP001908 Female     Yes          0 Not Graduate            No
## 285 LP001922   Male     Yes          0     Graduate            No
## 306 LP001990   Male      No          0 Not Graduate            No
## 310 LP001998   Male     Yes          2 Not Graduate            No
## 314 LP002008   Male     Yes          2     Graduate           Yes
## 318 LP002036   Male     Yes          0     Graduate            No
## 319 LP002043 Female      No          1     Graduate            No
## 323 LP002054   Male     Yes          2 Not Graduate            No
## 324 LP002055 Female      No          0     Graduate            No
## 336 LP002106   Male     Yes                Graduate           Yes
## 339 LP002113 Female      No         3+ Not Graduate            No
## 349 LP002137   Male     Yes          0     Graduate            No
## 364 LP002178   Male     Yes          0     Graduate            No
## 368 LP002188   Male      No          0     Graduate            No
## 378 LP002223   Male     Yes          0     Graduate            No
## 388 LP002243   Male     Yes          0 Not Graduate            No
## 393 LP002263   Male     Yes          0     Graduate            No
## 396 LP002272   Male     Yes          2     Graduate            No
## 412 LP002319   Male     Yes          0     Graduate
## 422 LP002357 Female      No          0 Not Graduate            No
```

```
## 424 LP002362   Male     Yes          1     Graduate            No
## 436 LP002393 Female                         Graduate            No
## 438 LP002401   Male     Yes          0     Graduate            No
## 445 LP002424   Male     Yes          0     Graduate            No
## 450 LP002444   Male      No          1 Not Graduate           Yes
## 452 LP002447   Male     Yes          2 Not Graduate            No
## 461 LP002478            Yes          0     Graduate           Yes
## 474 LP002522 Female      No          0     Graduate           Yes
## 480 LP002533   Male     Yes          2     Graduate            No
## 491 LP002560   Male      No          0 Not Graduate            No
## 492 LP002562   Male     Yes          1 Not Graduate            No
## 498 LP002588   Male     Yes          0     Graduate            No
## 504 LP002618   Male     Yes          1 Not Graduate            No
## 507 LP002624   Male     Yes          0     Graduate            No
## 525 LP002697   Male      No          0     Graduate            No
## 531 LP002717   Male     Yes          0     Graduate            No
## 534 LP002729   Male      No          1     Graduate            No
## 545 LP002757 Female     Yes          0 Not Graduate            No
## 551 LP002778   Male     Yes          2     Graduate           Yes
## 552 LP002784   Male     Yes          1 Not Graduate            No
## 557 LP002794 Female      No          0     Graduate            No
## 566 LP002833   Male     Yes          0 Not Graduate            No
## 584 LP002898   Male     Yes          1     Graduate            No
## 601 LP002949 Female      No         3+     Graduate
## 606 LP002960   Male     Yes          0 Not Graduate            No
##     ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term
## 1              5849                 0         NA              360
## 17             3596                 0        100              240
## 20             2600              3500        115               NA
## 25             3717              2925        151              360
## 31             4166              3369        201              360
## 36             2275              2067         NA              360
## 37             1828              1330        100               NA
## 43             2400                 0         75              360
## 45             4695                 0         96               NA
## 46             3410                 0         88               NA
## 64             4945                 0         NA              360
## 74             4755                 0         95               NA
## 80             3333              2166        130              360
## 82             2395                 0         NA              360
## 84             6000              2250        265              360
## 87             3333              2000         99              360
## 96             6782                 0         NA              360
## 103           13650                 0         NA              360
## 104            4652              3583         NA              360
## 113            3572              4114        152               NA
## 114            7451                 0         NA              360
## 118            2214              1398         85              360
## 126            3692                 0         93              360
## 128            3865              1640         NA              360
## 130            6080              2569        182              360
## 131           20166                 0        650              480
## 157            6000                 0        160              360
## 166            3707              3166        182               NA
```

```
## 182            1916            5063        67            360
## 188            2383            2138        58            360
## 198            1907            2365       120             NA
## 199            3416            2816       113            360
## 203            3992               0        NA            180
## 220            4283            2383       127            360
## 224            7578            1010       175             NA
## 233            3189            2598       120             NA
## 237            5746               0       255            360
## 238            3463               0       122            360
## 260            4931               0       128            360
## 261            6083            4250       330            360
## 280            4100               0       124            360
## 285           20667               0        NA            360
## 306            2000               0        NA            360
## 310            7667               0       185            360
## 314            5746               0       144             84
## 318            2058            2134        88            360
## 319            3541               0       112            360
## 323            3601            1590        NA            360
## 324            3166            2985       132            360
## 336            5503            4490        70             NA
## 339            1830               0        NA            360
## 349            6333            4583       259            360
## 364            3013            3033        95            300
## 368            5124               0       124             NA
## 378            4310               0       130            360
## 388            3010            3136        NA            360
## 393            2583            2115       120            360
## 396            3276             484       135            360
## 412            6256               0       160            360
## 422            2720               0        80             NA
## 424            7250            1667       110             NA
## 436           10047               0        NA            240
## 438            2213            1125        NA            360
## 445            7333            8333       175            300
## 450            2769            1542       190            360
## 452            1958            1456        60            300
## 461            2083            4083       160            360
## 474            2500               0        93            360
## 480            2947            1603        NA            360
## 491            2699            2785        96            360
## 492            5333            1131       186            360
## 498            4625            2857       111             12
## 504            4050            5302       138            360
## 507           20833            6667       480            360
## 525            4680            2087        NA            360
## 531            1025            5500       216            360
## 534           11250               0       196            360
## 545            3017             663       102            360
## 551            6633               0        NA            360
## 552            2492            2375        NA            360
## 557            2667            1625        84            360
## 566            4467               0       120            360
```

```
## 584            1880                 0        61              360
## 601             416             41667       350              180
## 606            2400              3800        NA              180
##     Credit_History Property_Area Loan_Status
## 1                1         Urban           Y
## 17              NA         Urban           Y
## 20               1         Urban           Y
## 25              NA     Semiurban           N
## 31              NA         Urban           N
## 36               1         Urban           Y
## 37               0         Urban           N
## 43              NA         Urban           Y
## 45               1         Urban           Y
## 46               1         Urban           Y
## 64               0         Rural           N
## 74               0     Semiurban           N
## 80              NA     Semiurban           Y
## 82               1     Semiurban           Y
## 84              NA     Semiurban           N
## 87              NA     Semiurban           Y
## 96              NA         Urban           N
## 103              1         Urban           Y
## 104              1     Semiurban           Y
## 113              0         Rural           N
## 114              1     Semiurban           Y
## 118             NA         Urban           Y
## 126             NA         Rural           Y
## 128              1         Rural           Y
## 130             NA         Rural           N
## 131             NA         Urban           Y
## 157             NA         Rural           Y
## 166              1         Rural           Y
## 182             NA         Rural           N
## 188             NA         Rural           Y
## 198              1         Urban           Y
## 199             NA     Semiurban           Y
## 203              1         Urban           N
## 220             NA     Semiurban           Y
## 224              1     Semiurban           Y
## 233              1         Rural           Y
## 237             NA         Urban           N
## 238             NA         Urban           Y
## 260             NA     Semiurban           N
## 261             NA         Urban           Y
## 280             NA         Rural           Y
## 285              1         Rural           N
## 306              1         Urban           N
## 310             NA         Rural           Y
## 314             NA         Rural           Y
## 318             NA         Urban           Y
## 319             NA     Semiurban           Y
## 323              1         Rural           Y
## 324             NA         Rural           Y
## 336              1     Semiurban           Y
```

```
## 339             0        Urban           N
## 349            NA    Semiurban           Y
## 364            NA        Urban           Y
## 368             0        Rural           N
## 378            NA    Semiurban           Y
## 388             0        Urban           N
## 393            NA        Urban           Y
## 396            NA    Semiurban           Y
## 412            NA        Urban           Y
## 422             0        Urban           N
## 424             0        Urban           N
## 436             1    Semiurban           Y
## 438             1        Urban           Y
## 445            NA        Rural           Y
## 450            NA    Semiurban           N
## 452            NA        Urban           Y
## 461            NA    Semiurban           Y
## 474            NA        Urban           Y
## 480             1        Urban           N
## 491            NA    Semiurban           Y
## 492            NA        Urban           Y
## 498            NA        Urban           Y
## 504            NA        Rural           N
## 507            NA        Urban           Y
## 525             1    Semiurban           N
## 531            NA        Rural           Y
## 534            NA    Semiurban           N
## 545            NA    Semiurban           Y
## 551             0        Rural           N
## 552             1        Rural           Y
## 557            NA        Urban           Y
## 566            NA        Rural           Y
## 584            NA        Rural           N
## 601            NA        Urban           N
## 606             1        Urban           N
```

```r
#store only data without missing values (removed 85 rows)
my_loan_data<- my_loan_data[complete.cases(my_loan_data),]
```

Visual Analysis

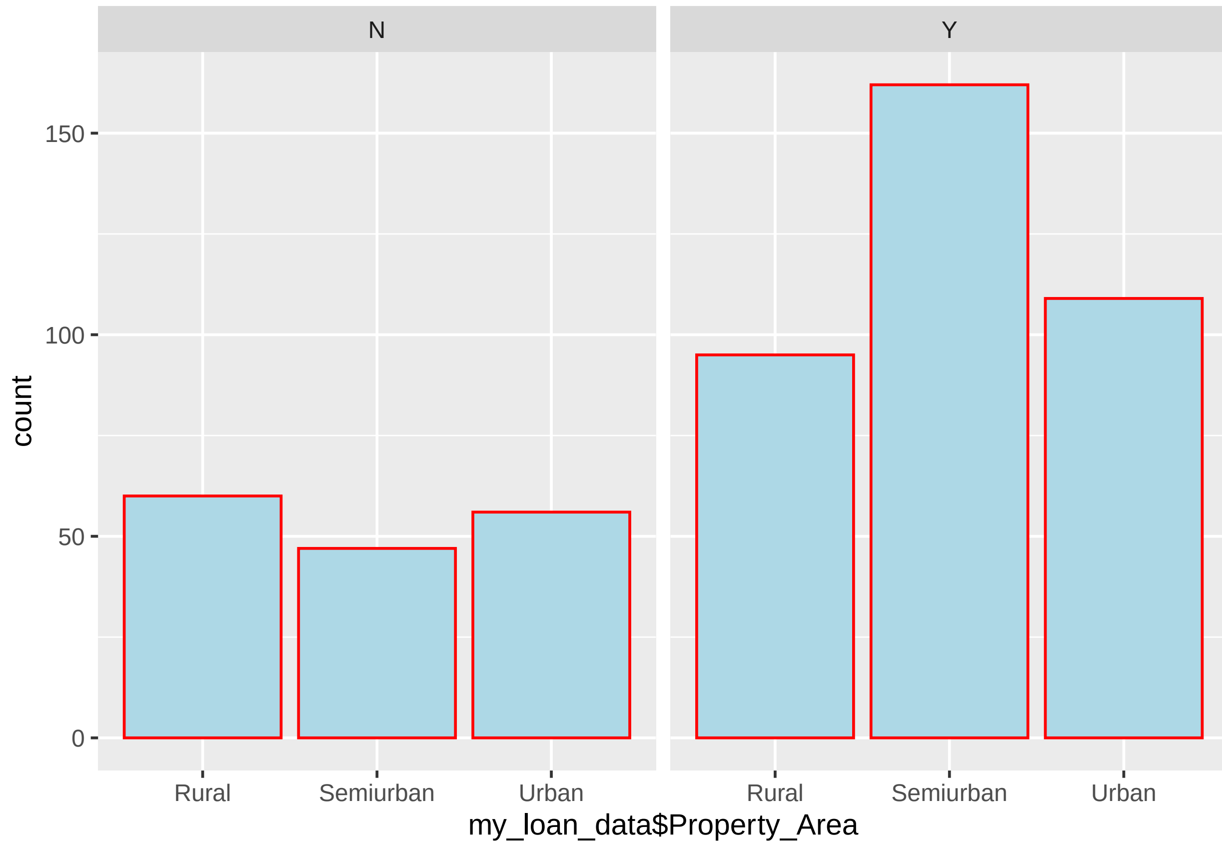
Property Area:

```r
summary(my_loan_data$Property_Area)
```

```
##     Rural Semiurban     Urban
##       155       209       165
```

```r
ggplot(data=my_loan_data, aes(my_loan_data$Property_Area)) +
  geom_histogram(col="red",fill="lightblue",stat="count" ) +
  facet_grid(~my_loan_data$Loan_Status)+
  scale_x_discrete()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



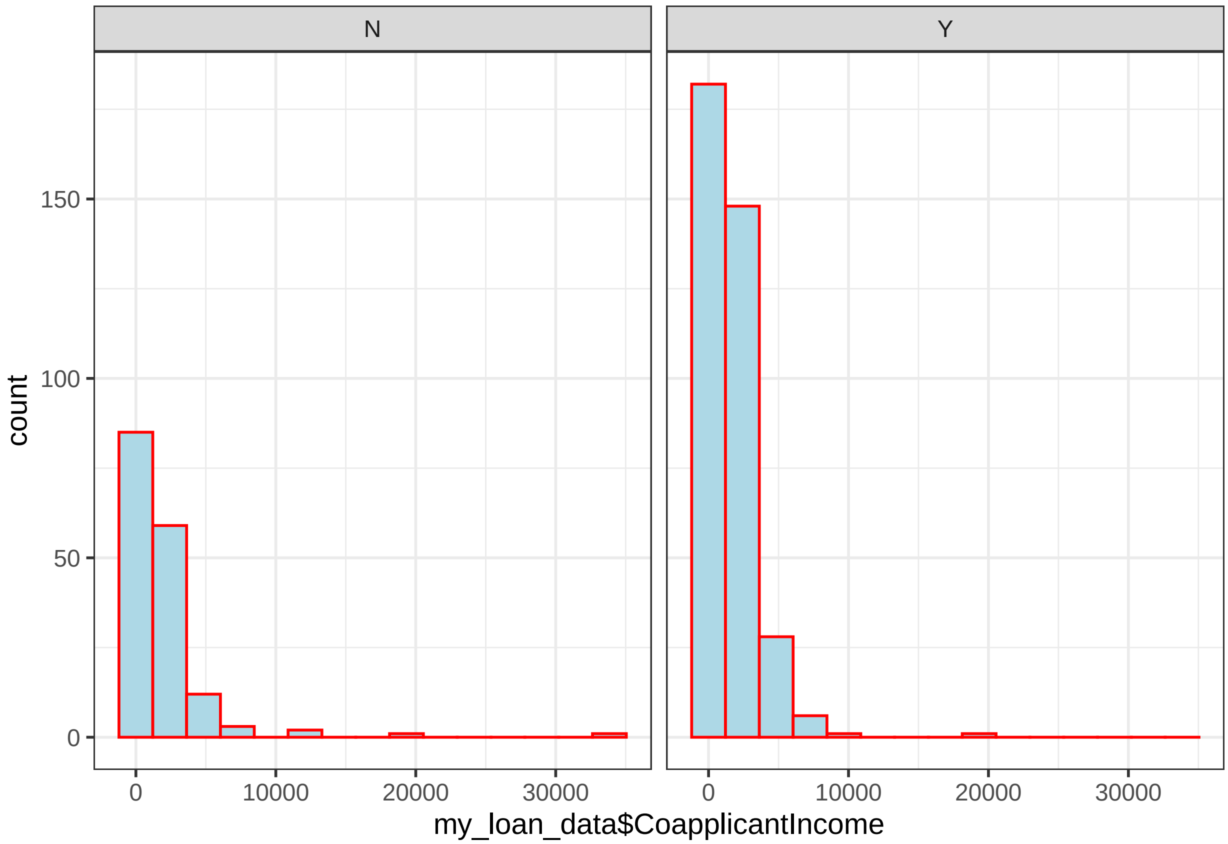
Histogram of Property Area shows that Loan approval is more into Semiurban area than Rural and Urban. Urban area has lowest loan approval. Loan rejection is lowest in Rural area. Semiurban & Urban has same loan rejection

Coapplicant Income:

```
summary(my_loan_data$CoapplicantIncome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0    1086    1542    2232   33837
```

```
ggplot(data=my_loan_data, aes(x= my_loan_data$CoapplicantIncome)) +
  geom_histogram(col="red",fill="lightblue", bins = 15) +
  facet_grid(~my_loan_data$Loan_Status)+
  theme_bw()
```



Histogram shows that low income peoples are mainly applying for loans and number of loan rejection is more in the lowest income segment

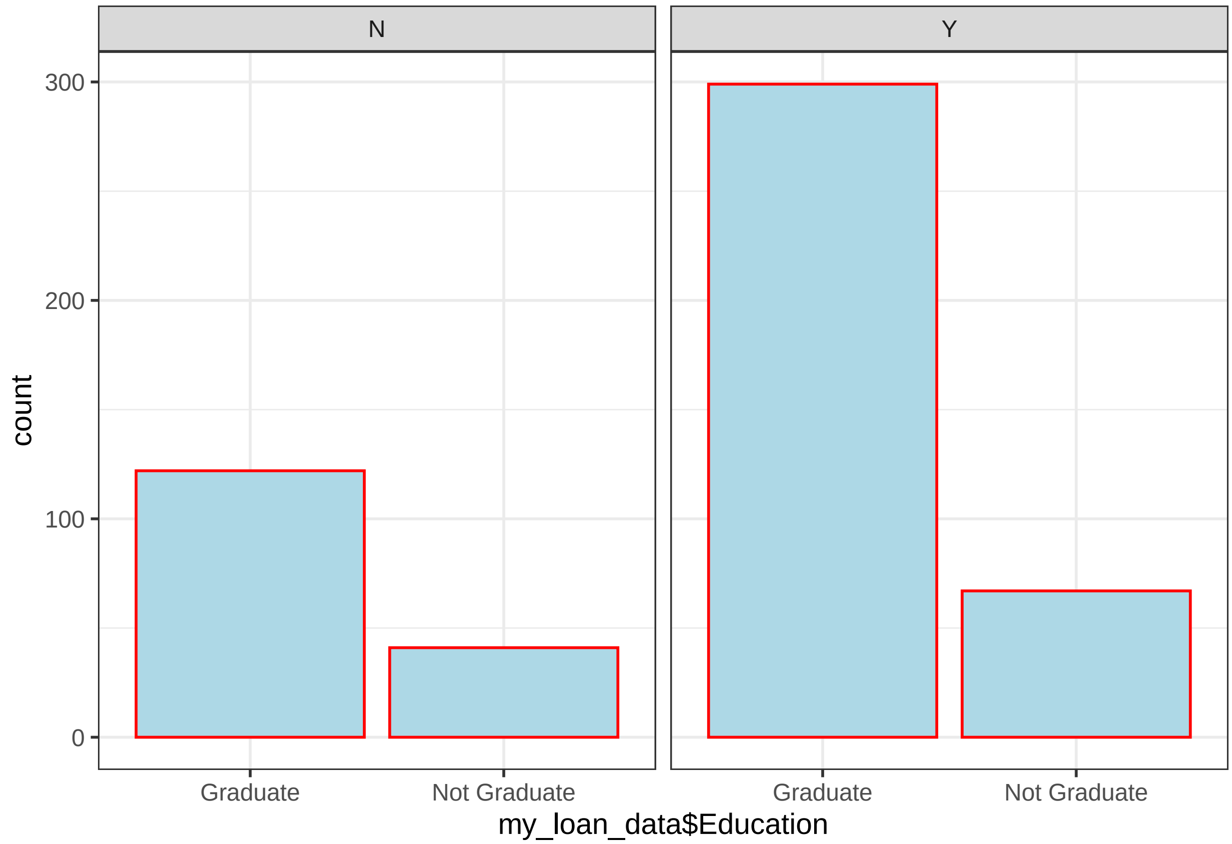
Education:

```
summary(my_loan_data$Education)
```

```
##     Graduate Not Graduate
##          421          108
```

```
ggplot(data=my_loan_data, aes(my_loan_data$Education)) +
  geom_histogram(col="red",fill="lightblue",stat="count" ) +
  facet_grid(~my_loan_data$Loan_Status)+
  scale_x_discrete()+
  theme_bw()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Based on loan approval flag shows that - loan approval rate for graduate is more than non graduate

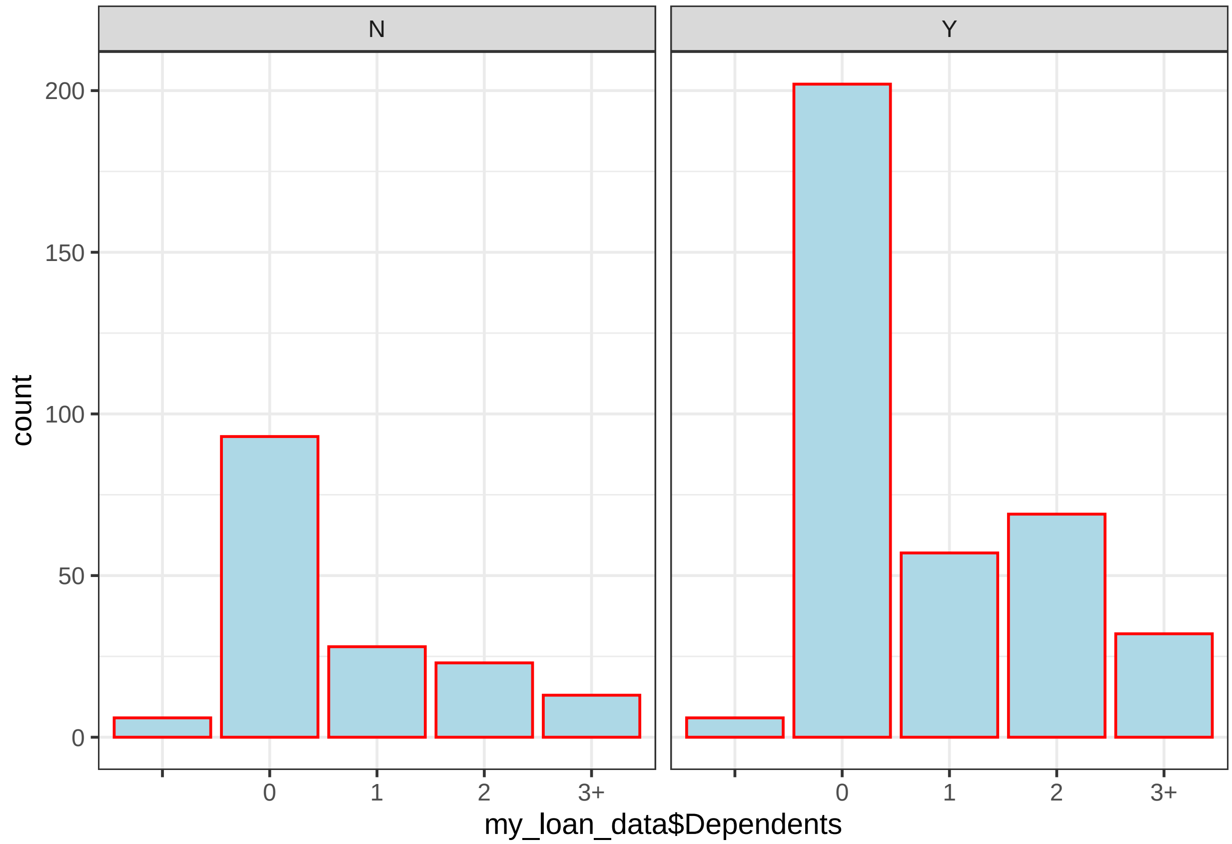
Number of Dependents:

```
summary(my_loan_data$Dependents)
```

```
##         0    1    2   3+
##   12 295   85   92   45
```

```
ggplot(data=my_loan_data, aes(my_loan_data$Dependents)) +
  geom_histogram(col="red",fill="lightblue",stat="count" ) +
  facet_grid(~my_loan_data$Loan_Status)+
  scale_x_discrete()+
  theme_bw()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Loan approval shows that -People having no dependents have maximum loan approval and rejection count

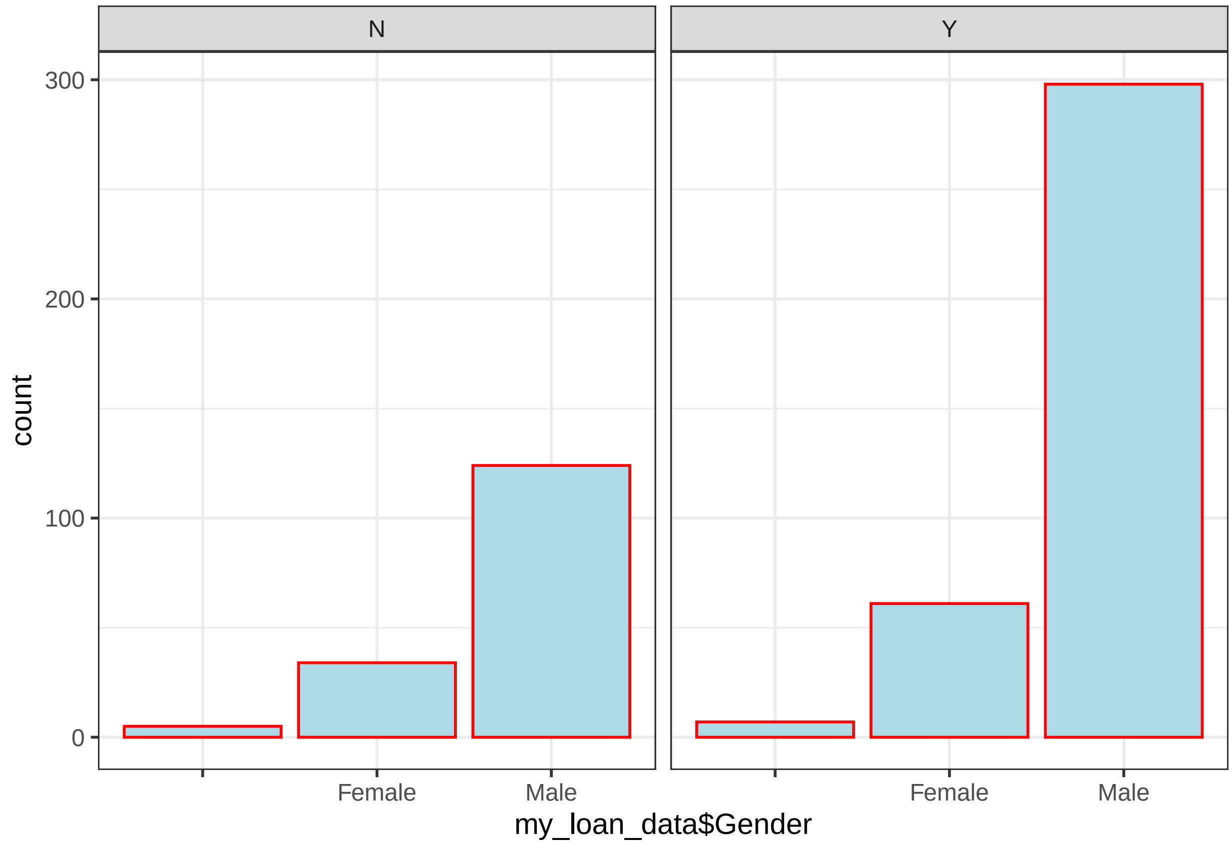
Gender:

```
summary(my_loan_data$Gender)
```

```
##        Female   Male
##     12     95    422
```

```
ggplot(data=my_loan_data, aes(my_loan_data$Gender)) +
  geom_histogram(col="red",fill="lightblue",stat="count") +
  facet_grid(~my_loan_data$Loan_Status)+
  scale_x_discrete()+
  theme_bw()
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Male applicant has higher loan approval and rejection count than female applicant. So this looks to be an influencing factor

Logestic Regression

Logistic Regression, in simple terms, predicts the probability of occurrence of an event by fitting data to a logit function. Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. This type of models is part of a larger class of algorithms known as Generalized Linear Model or GLM.

Preparing Data for The Model:

```
my_loan_data_1 <- my_loan_data[,2:13]
ind <- sample.split (Y=my_loan_data_1$Loan_Status, SplitRatio=0.8)
traindf<- my_loan_data_1 [ind,]
testdf<- my_loan_data_1 [!ind,]
```

Logistic Regression Model

```
#Logistic regression
LRmodel<-glm(Loan_Status~.,traindf,family = "binomial")
summary(LRmodel)
```

```
##
## Call:
## glm(formula = Loan_Status ~ ., family = "binomial", data = traindf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5496  -0.3734   0.5130   0.6757   2.4697
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.205e+01  8.827e+02   0.014  0.98911
## GenderFemale              1.369e-01  8.671e-01   0.158  0.87455
## GenderMale                5.509e-01  8.160e-01   0.675  0.49958
## MarriedNo                -1.374e+01  8.827e+02  -0.016  0.98758
## MarriedYes               -1.337e+01  8.827e+02  -0.015  0.98792
## Dependents0               1.674e-01  9.500e-01   0.176  0.86016
## Dependents1              -1.602e-01  9.831e-01  -0.163  0.87056
## Dependents2               3.984e-01  9.922e-01   0.402  0.68805
## Dependents3+              5.108e-01  1.057e+00   0.483  0.62897
## EducationNot Graduate    -4.764e-01  3.318e-01  -1.436  0.15103
## Self_EmployedNo          -7.479e-01  6.996e-01  -1.069  0.28504
## Self_EmployedYes         -7.380e-01  7.621e-01  -0.968  0.33282
## ApplicantIncome          -1.176e-06  3.116e-05  -0.038  0.96989
## CoapplicantIncome         6.125e-05  8.482e-05   0.722  0.47020
## LoanAmount               -2.609e-03  2.015e-03  -1.295  0.19527
## Loan_Amount_Term         -2.604e-03  2.295e-03  -1.135  0.25651
## Credit_History            3.873e+00  4.774e-01   8.114  4.9e-16 ***
## Property_AreaSemiurban  9.281e-01  3.292e-01   2.819  0.00482 **
## Property_AreaUrban        1.521e-01  3.398e-01   0.447  0.65455
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 521.94  on 422  degrees of freedom
## Residual deviance: 372.47  on 404  degrees of freedom
## AIC: 410.47
##
## Number of Fisher Scoring iterations: 13
```

Most significant variables are

- Credit_History
- Property_AreaSemiurban

```r
res<-predict(LRmodel,testdf,type="response")
res
```

```
##          4         13         21         24         28         29
## 0.74928663 0.87489537 0.05279797 0.03706114 0.78061219 0.68989070
##         30         39         41         42         44         53
## 0.92552715 0.79011769 0.76061170 0.78894225 0.90951430 0.80748935
##         57         60         77         78         92         93
```

```
## 0.91399090 0.78464801 0.76786873 0.78953476 0.95551288 0.80119926
##        112        119        127        134        136        141
## 0.93273851 0.77883309 0.59627649 0.95624582 0.92928084 0.81165299
##        146        147        151        156        167        168
## 0.89244961 0.79572185 0.04872897 0.10389029 0.70665776 0.81084973
##        170        173        176        192        214        215
## 0.90191078 0.84698727 0.80287525 0.84592233 0.75433850 0.82732056
##        216        221        223        229        239        247
## 0.89850834 0.09200073 0.87399276 0.99999969 0.54454405 0.84161059
##        251        255        269        270        278        279
## 0.10728029 0.04677858 0.76755053 0.61887466 0.83343268 0.79642448
##        284        286        289        291        296        298
## 0.71382120 0.76278599 0.80628902 0.79778867 0.93205018 0.65556623
##        308        316        317        320        325        346
## 0.04401245 0.77109606 0.91926647 0.70202283 0.75655656 0.90456808
##        355        357        366        369        375        383
## 0.85738906 0.82856661 0.59537135 0.84906482 0.81786750 0.63201914
##        386        391        392        395        399        402
## 0.87641137 0.74401126 0.76951754 0.85335620 0.63497248 0.67347820
##        408        409        414        416        418        423
## 0.53786085 0.12653164 0.72097884 0.68888665 0.92978081 0.82993128
##        431        441        442        444        449        468
## 0.73336359 0.87685180 0.77279510 0.79842513 0.08770957 0.84041897
##        472        490        503        510        520        522
## 0.05605869 0.76322684 0.92730018 0.62421406 0.51865809 0.88064584
##        527        529        535        538        546        556
## 0.90268765 0.76357705 0.53517848 0.89017770 0.78627187 0.89846631
##        558        564        570        572        579        582
## 0.89861247 0.79038264 0.09288200 0.06627842 0.75375113 0.96100870
##        586        591        604        614
## 0.84708372 0.95466245 0.73638710 0.07693284
```

```r
table(Actualvalue=testdf$Loan_Status,Predictedvalue=res>0.5)
```

```
##            Predictedvalue
## Actualvalue FALSE TRUE
##           N    13   20
##           Y     1   72
```

```r
(16+71)/ (16+17+2+71)
```

```
## [1] 0.8207547
```

Accuracy: 82.07%

Decision Tree

Decision trees create a set of binary splits on the predictor variables in order to create a tree that can be used to classify new observations into one of two groups. Here, we will be using classical trees. The algorithm of this model is the following:

Choose the predictor variable that best splits the data into two groups;

Separate the data into these two groups;

Repeat these steps until a subgroup contains fewer than a minimum number of observations;

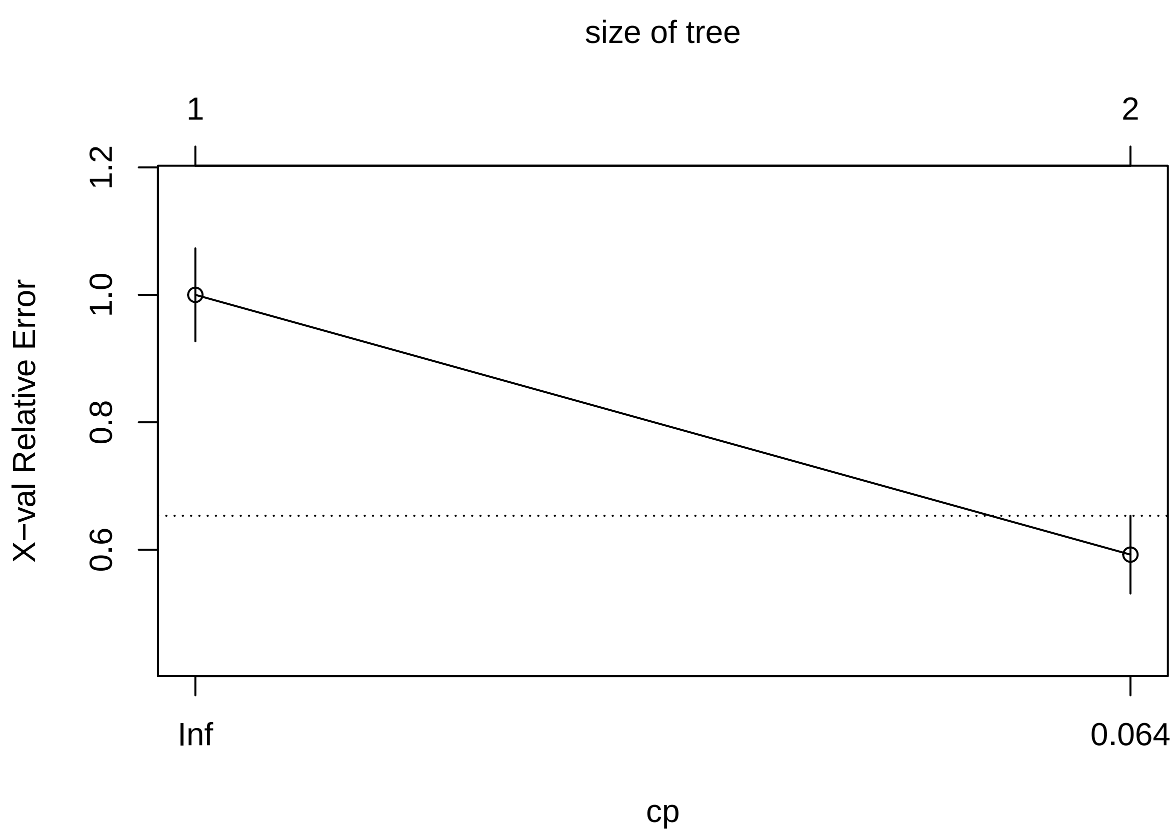
To classify a case, run it down the tree to a terminal node, and assign it the model outcome value assigned in the previous step.

```
set.seed(42)
sample <- sample.int(n = nrow(my_loan_data_1), size = floor(.70*nrow(my_loan_data_1)), replace = F)
trainnew <- my_loan_data_1[sample, ]
testnew  <- my_loan_data_1[-sample, ]
```

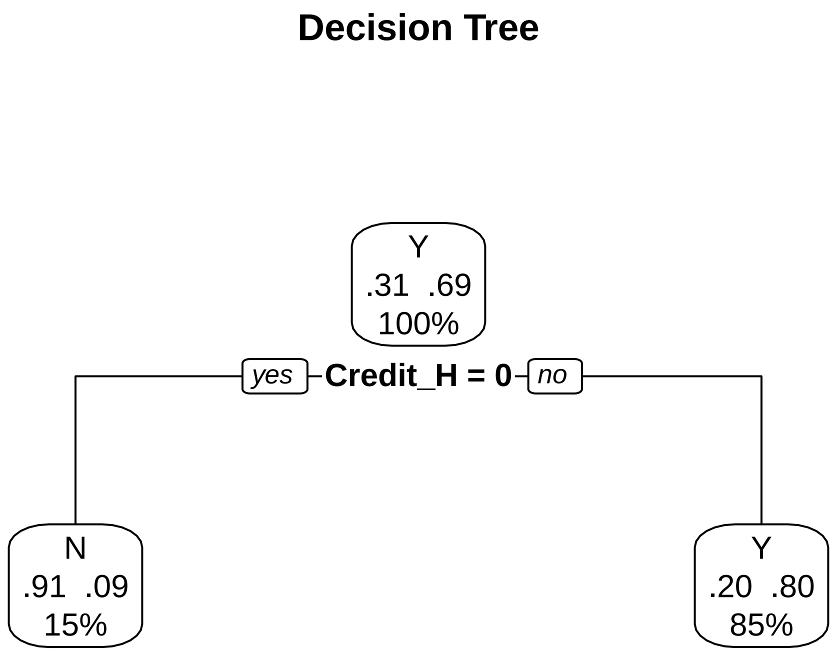
```
dtree <- rpart(Loan_Status ~ Credit_History + Education + Self_Employed + Property_Area + LoanAmount +
                 ApplicantIncome, method="class", data=traindf,parms=list(split="information"))
dtree$cptable
```

```
##          CP nsplit rel error    xerror       xstd
## 1 0.4076923      0 1.0000000 1.0000000 0.07299480
## 2 0.0100000      1 0.5923077 0.5923077 0.06104778
```

```
plotcp(dtree)
```




```
dtree.pruned <- prune(dtree, cp=.02290076)
library(rpart.plot)
prp(dtree.pruned, type = 2, extra = 104,
    fallen.leaves = TRUE, main="Decision Tree")
```



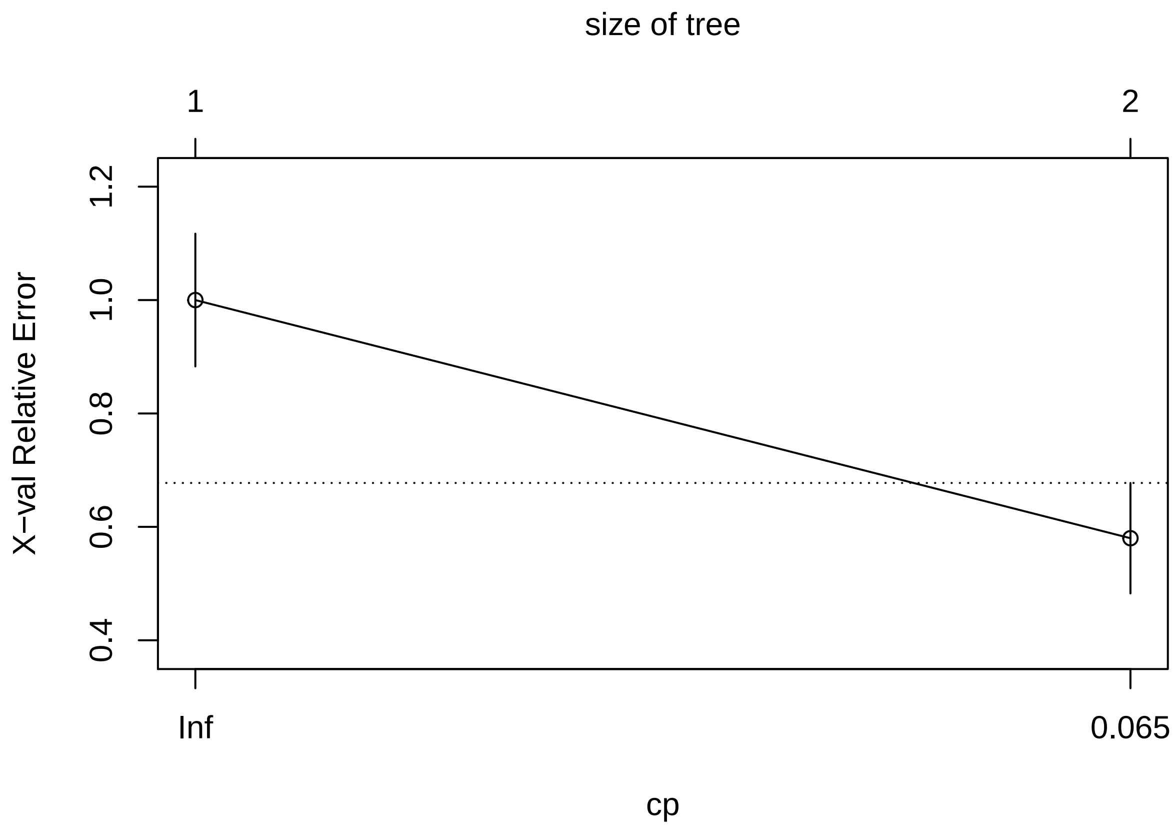

```
dtree.pred <- predict(dtree.pruned, trainnew, type="class")
dtree.perf <- table(trainnew$Loan_Status, dtree.pred,
                    dnn=c("Actual", "Predicted"))
dtree.perf
```

```
##        Predicted
## Actual   N   Y
##      N  49  64
##      Y   5 252
```

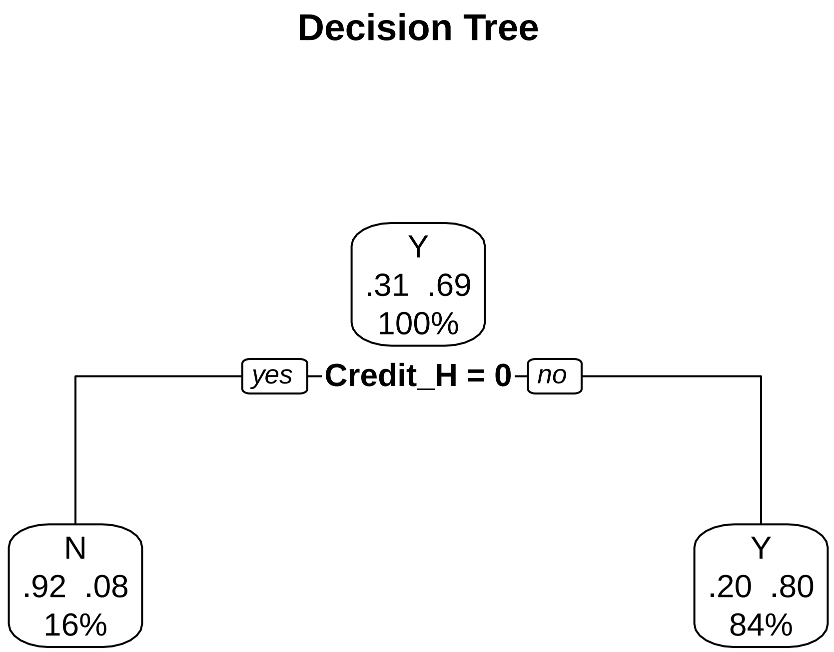
```
dtree_test <- rpart(Loan_Status ~ Credit_History+Education+Self_Employed+Property_Area+LoanAmount+
                 ApplicantIncome,method="class", data=testnew,parms=list(split="information"))
dtree_test$cptable
```

```
##     CP nsplit rel error xerror       xstd
## 1 0.42      0      1.00   1.00 0.11709266
## 2 0.01      1      0.58   0.58 0.09738725
```

```
plotcp(dtree_test)
```




```r
dtree_test.pruned <- prune(dtree_test, cp=.01639344)
prp(dtree_test.pruned, type = 2, extra = 104,
    fallen.leaves = TRUE, main="Decision Tree")
```



Accuracy: 84% Results show better performance than the logistic model.

Random Forest

```
set.seed(42)
fit.forest <- randomForest(Loan_Status ~ Credit_History+Education+Self_Employed+Property_Area+LoanAmoun
                             ApplicantIncome, data=trainnew,
                           na.action=na.roughfix,
                           importance=TRUE)
fit.forest
```

```
##
## Call:
##  randomForest(formula = Loan_Status ~ Credit_History + Education +      Self_Employed + Property_Are
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 18.65%
## Confusion matrix:
##    N   Y class.error
## N 53  60  0.53097345
## Y  9 248  0.03501946
```

```r
importance(fit.forest, type=2)
```

```
##                  MeanDecreaseGini
## Credit_History          41.890029
## Education                3.328873
## Self_Employed            4.921181
## Property_Area            8.718707
## LoanAmount              29.449887
## ApplicantIncome         29.320255
```

```r
forest.pred <- predict(fit.forest, testnew)
forest.perf <- table(testnew$Loan_Status, forest.pred,
                     dnn=c("Actual", "Predicted"))
forest.perf
```

```
##       Predicted
## Actual   N   Y
##      N  24  26
##      Y   5 104
```

Here is the accuracy of the model: 80.50%

Conclusion

After analyzing the data from the loan prediction dataset, the data shows that Credit History and Property_AreaSemiurban are most significant variables to predict whether a loan application will approved or not. We can predict the loan approval using different models. Here, we got 82.07% accuracy for logistic regresission, 84% accuracy for Decesion tree and 80.50% accuracy for random forest.

The dataset is relatively small. A larger dataset will help to improve the model accuracy.

We can conclude that the company should target customers with Credit history and customer who lives in Semiurban area.

Reference

https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/