PRODIGY\_DS\_02.R

#EXPLORATORY DATA ANALYSIS OF CREDIT SCORE DATA  
#Question:Perform data cleaning and EDA on a dataset of choice. Explore the relationship between variables and identify patterns and trends.

#DATA SOURCE

#<https://www.kaggle.com/datasets/yashkmd/credit-profile-two-wheeler-loan-dataset?resource=download>  
#DATA DESCRIPTION  
#The data is 'Credit Profile (Two-wheeler loan) Dataset'. This dataset provides a comprehensive overview of potential loan applicants' profiles in India, specifically tailored for the Indian demographic. It encapsulates a range of features, from basic demographics to financial details, that can be instrumental in assessing the creditworthiness of an individual.  
#DATASET  
library(readxl)  
credit\_data <- read\_excel("D:/OneDrive/Documents/edu/prodigy int/2/credit\_data.xlsx")

df<-data.frame(credit\_data)  
head(df,2)

## Age Gender Income Credit.Score Credit.History.Length Number.of.Existing.Loans  
## 1 31 Male 36000 604 487 5  
## 2 25 Male 50000 447 386 2  
## Loan.Amount Loan.Tenure Existing.Customer State City LTV.Ratio  
## 1 109373 221 No Karnataka Mysuru 90.94343  
## 2 150000 89 No Karnataka Bengaluru 91.13525  
## Employment.Profile Profile.Score Occupation  
## 1 Salaried 77 Doctor  
## 2 Salaried 43 Software Engineer

dim(df)

## [1] 279856 15

#This shows our data has 279856 rows and 15 columns.  
  
#DATA CLEANING  
# Checking for missing values  
missing\_values <- colSums(is.na(df))  
print(missing\_values)

## Age Gender Income   
## 0 0 0   
## Credit.Score Credit.History.Length Number.of.Existing.Loans   
## 0 0 0   
## Loan.Amount Loan.Tenure Existing.Customer   
## 0 0 0   
## State City LTV.Ratio   
## 0 0 0   
## Employment.Profile Profile.Score Occupation   
## 0 0 0

#We observe there are no missing values in our data.  
# Removing duplicate rows  
df <- unique(df)  
# Converting data types   
str(df)

## 'data.frame': 179042 obs. of 15 variables:  
## $ Age : num 31 25 62 69 52 64 29 30 52 39 ...  
## $ Gender : chr "Male" "Male" "Other" "Female" ...  
## $ Income : num 36000 50000 178000 46000 132000 127000 15000 82000 119000 101000 ...  
## $ Credit.Score : num 604 447 850 668 601 850 378 424 753 575 ...  
## $ Credit.History.Length : num 487 386 503 349 553 158 89 610 271 424 ...  
## $ Number.of.Existing.Loans: num 5 2 10 6 5 10 1 2 8 5 ...  
## $ Loan.Amount : num 109373 150000 69099 150000 150000 ...  
## $ Loan.Tenure : num 221 89 110 148 157 111 108 92 251 12 ...  
## $ Existing.Customer : chr "No" "No" "Yes" "Yes" ...  
## $ State : chr "Karnataka" "Karnataka" "Uttar Pradesh" "Karnataka" ...  
## $ City : chr "Mysuru" "Bengaluru" "Kanpur" "Bengaluru" ...  
## $ LTV.Ratio : num 90.9 91.1 40 87.4 66.2 ...  
## $ Employment.Profile : chr "Salaried" "Salaried" "Salaried" "Self-Employed" ...  
## $ Profile.Score : num 77 43 90 86 90 92 25 58 100 87 ...  
## $ Occupation : chr "Doctor" "Software Engineer" "Banker" "Contractor" ...

# since all the variables have the correct data type, we don't need to convert any.  
  
#EXPLORATORY DATA ANALYSIS (EDA):  
summary(df)

## Age Gender Income Credit.Score   
## Min. :18.00 Length:179042 Min. : 9000 Min. :300.0   
## 1st Qu.:31.00 Class :character 1st Qu.: 42000 1st Qu.:447.0   
## Median :44.00 Mode :character Median : 68000 Median :584.0   
## Mean :43.99 Mean : 76437 Mean :583.1   
## 3rd Qu.:57.00 3rd Qu.:104000 3rd Qu.:722.0   
## Max. :70.00 Max. :209000 Max. :850.0   
## Credit.History.Length Number.of.Existing.Loans Loan.Amount   
## Min. : 6 Min. : 0.000 Min. : 5294   
## 1st Qu.:156 1st Qu.: 2.000 1st Qu.: 72220   
## Median :307 Median : 5.000 Median :111284   
## Mean :308 Mean : 4.705 Mean :105793   
## 3rd Qu.:460 3rd Qu.: 7.000 3rd Qu.:150000   
## Max. :611 Max. :10.000 Max. :150000   
## Loan.Tenure Existing.Customer State City   
## Min. : 12.0 Length:179042 Length:179042 Length:179042   
## 1st Qu.: 62.0 Class :character Class :character Class :character   
## Median :100.0 Mode :character Mode :character Mode :character   
## Mean :133.3   
## 3rd Qu.:201.0   
## Max. :359.0   
## LTV.Ratio Employment.Profile Profile.Score Occupation   
## Min. :40.00 Length:179042 Min. : 0.00 Length:179042   
## 1st Qu.:58.08 Class :character 1st Qu.: 61.00 Class :character   
## Median :72.15 Mode :character Median : 89.00 Mode :character   
## Mean :71.63 Mean : 77.38   
## 3rd Qu.:86.21 3rd Qu.: 98.00   
## Max. :95.00 Max. :100.00

table(df$Gender)

##   
## Female Male Other   
## 82955 82746 13341

table(df$Existing.Customer)

##   
## No Yes   
## 111244 67798

table(df$State)

##   
## Delhi Gujarat Karnataka Kerala Maharashtra   
## 17950 17954 18003 17903 17929   
## Rajasthan Tamil Nadu Telangana Uttar Pradesh West Bengal   
## 17654 17813 17959 17846 18031

table(df$City)

##   
## Ahmedabad Bengaluru Bishanpura Channarayapatna   
## 7569 7510 5378 5316   
## Chennai Coimbatore Dhulagori Hyderabad   
## 7550 7533 5444 15152   
## Jaipur Kanpur Kochi Kolkata   
## 7334 7770 7504 15352   
## Lucknow Manjari Mumbai Mysuru   
## 7501 5469 5185 7794   
## Nagpur Nellikuppam New Delhi Pune   
## 5065 5287 15308 5005   
## Surat Thiruvananthapuram Udaipur   
## 7695 7641 7680

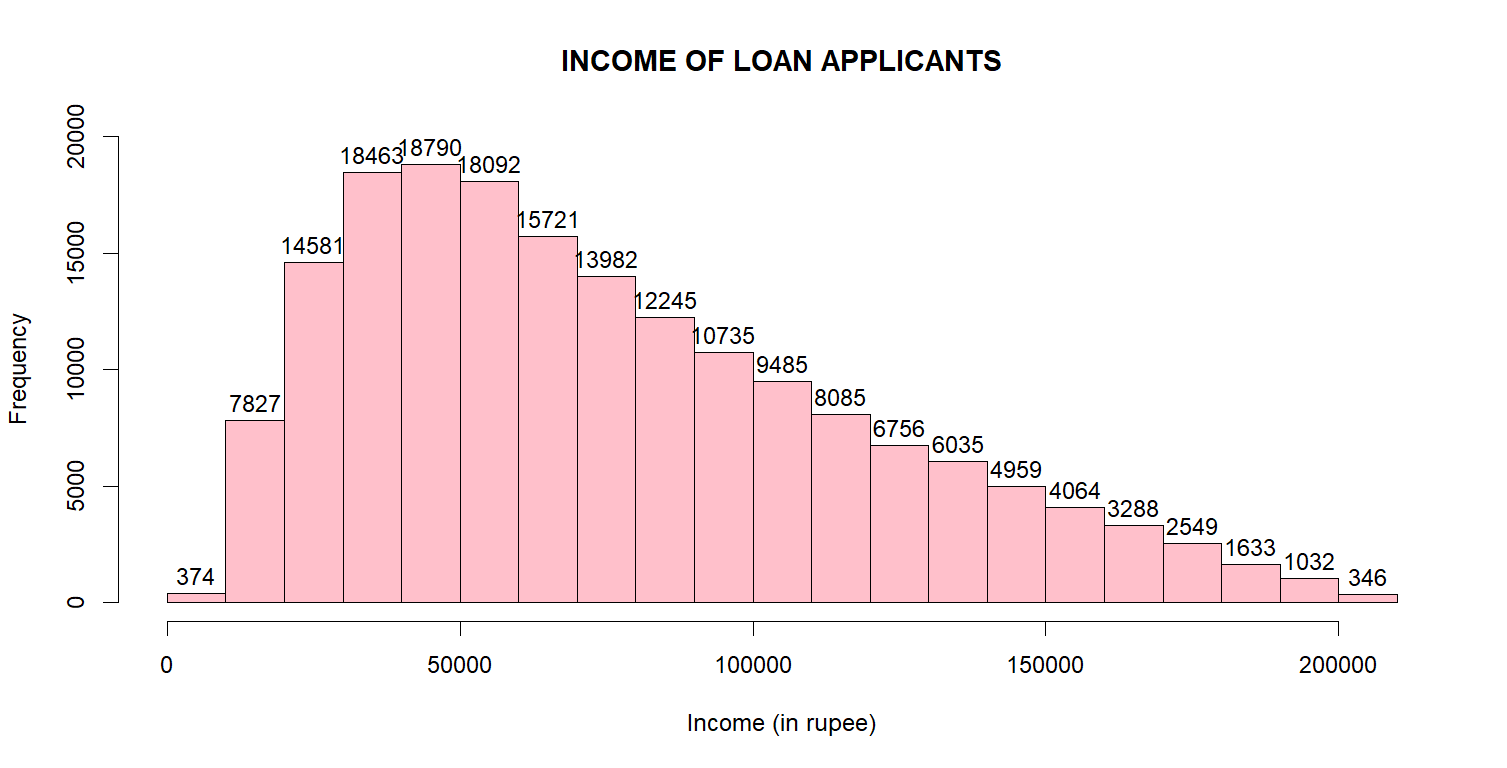
table(df$Employment.Profile)

##   
## Freelancer Salaried Self-Employed Student Unemployed   
## 14553 87138 53822 11839 11690

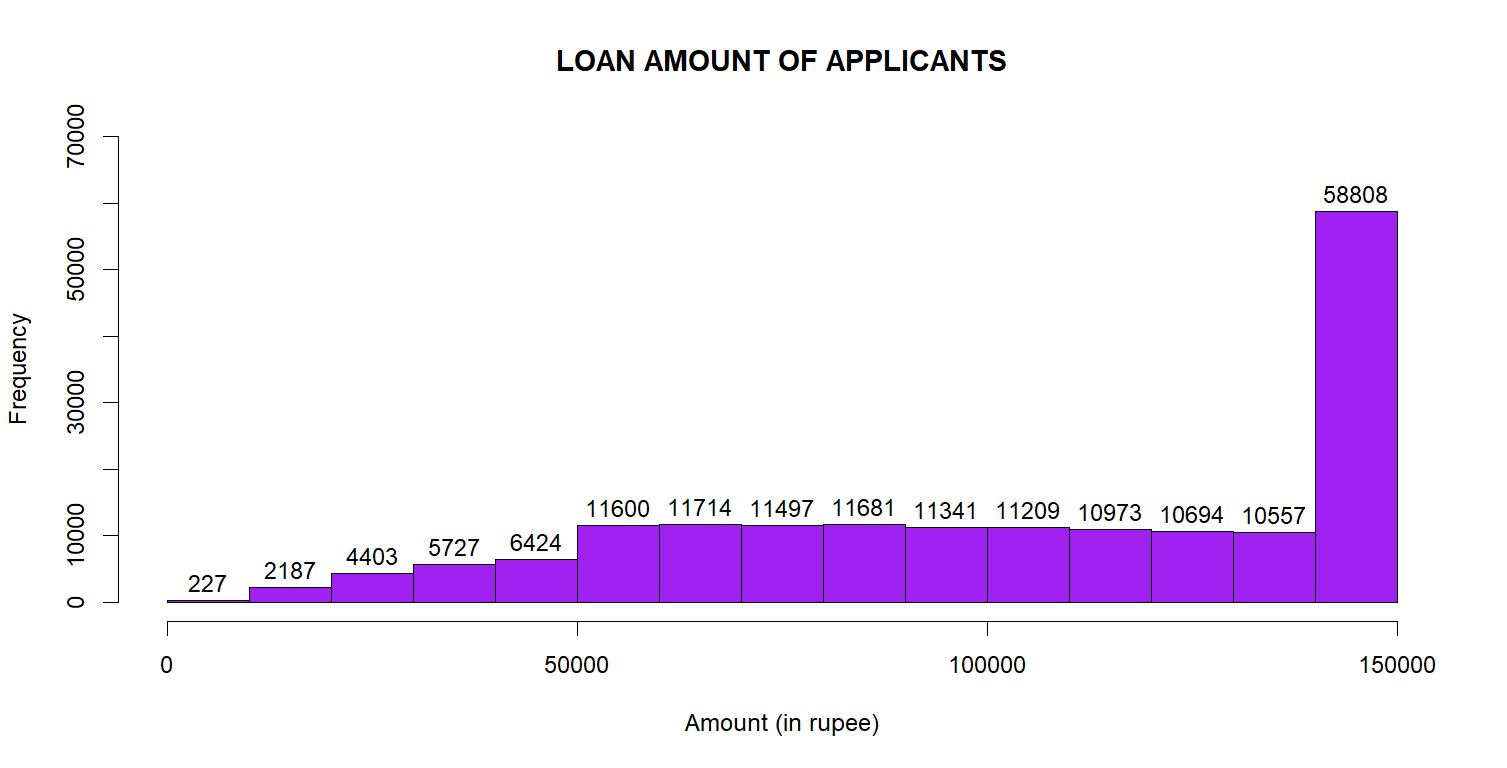
table(df$Occupation)

##   
## Banker Business Owner Civil Servant   
## 17788 13299 17486   
## Contractor Doctor Farmer   
## 13523 16997 13351   
## Graphic Designer Independent Consultant NA   
## 3694 3623 11690   
## Photographer Shopkeeper Software Engineer   
## 3687 13649 17312   
## Student Teacher Writer   
## 11839 17555 3549

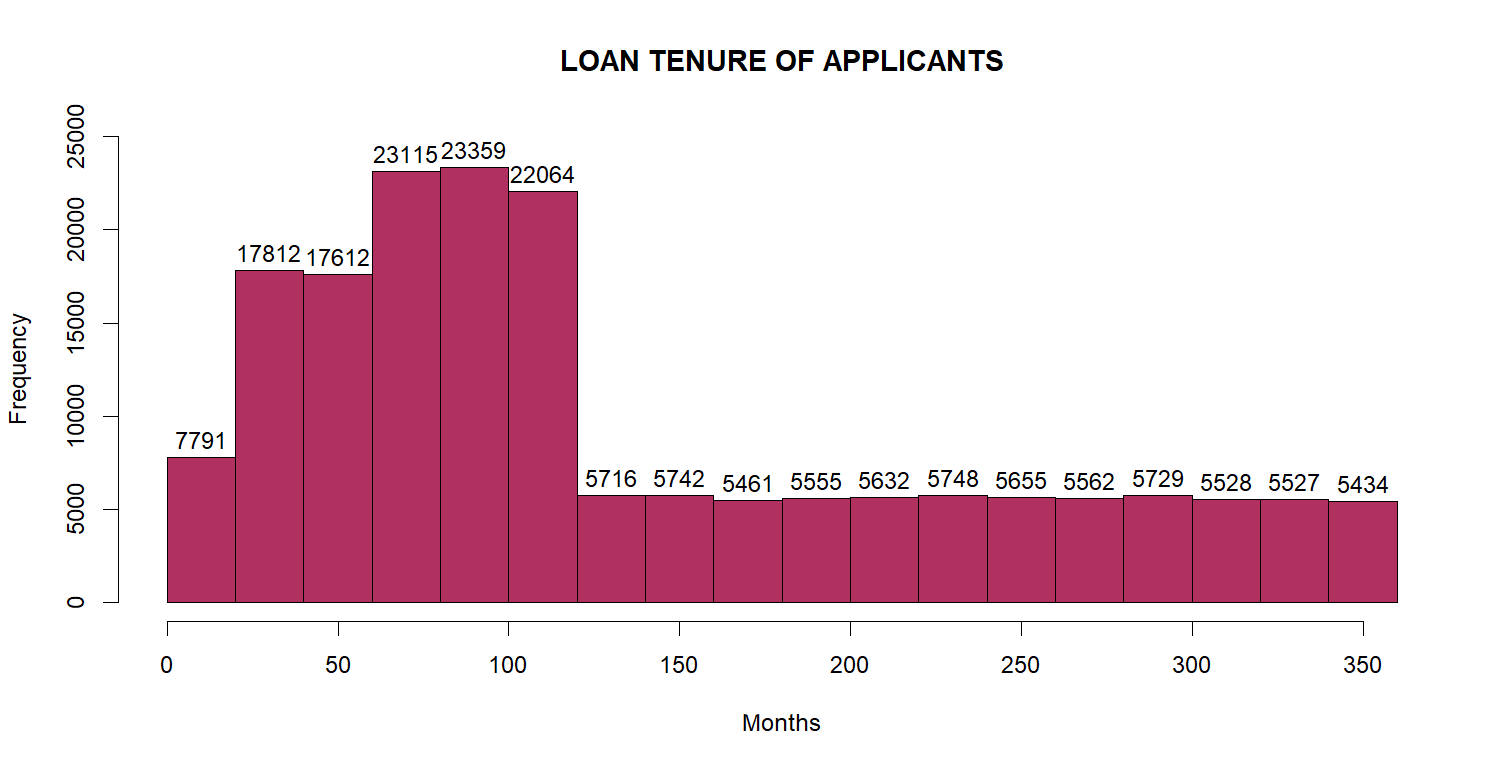
#GRAPHICAL ANALYSIS  
#HISTOGRAMS  
h1<-hist(df$Income,main="INCOME OF LOAN APPLICANTS", xlab="Income (in rupee)", col="pink", ylim=c(0,20000))  
text(h1$mids,h1$counts,labels=h1$counts,adj=c(0.5,-0.5))



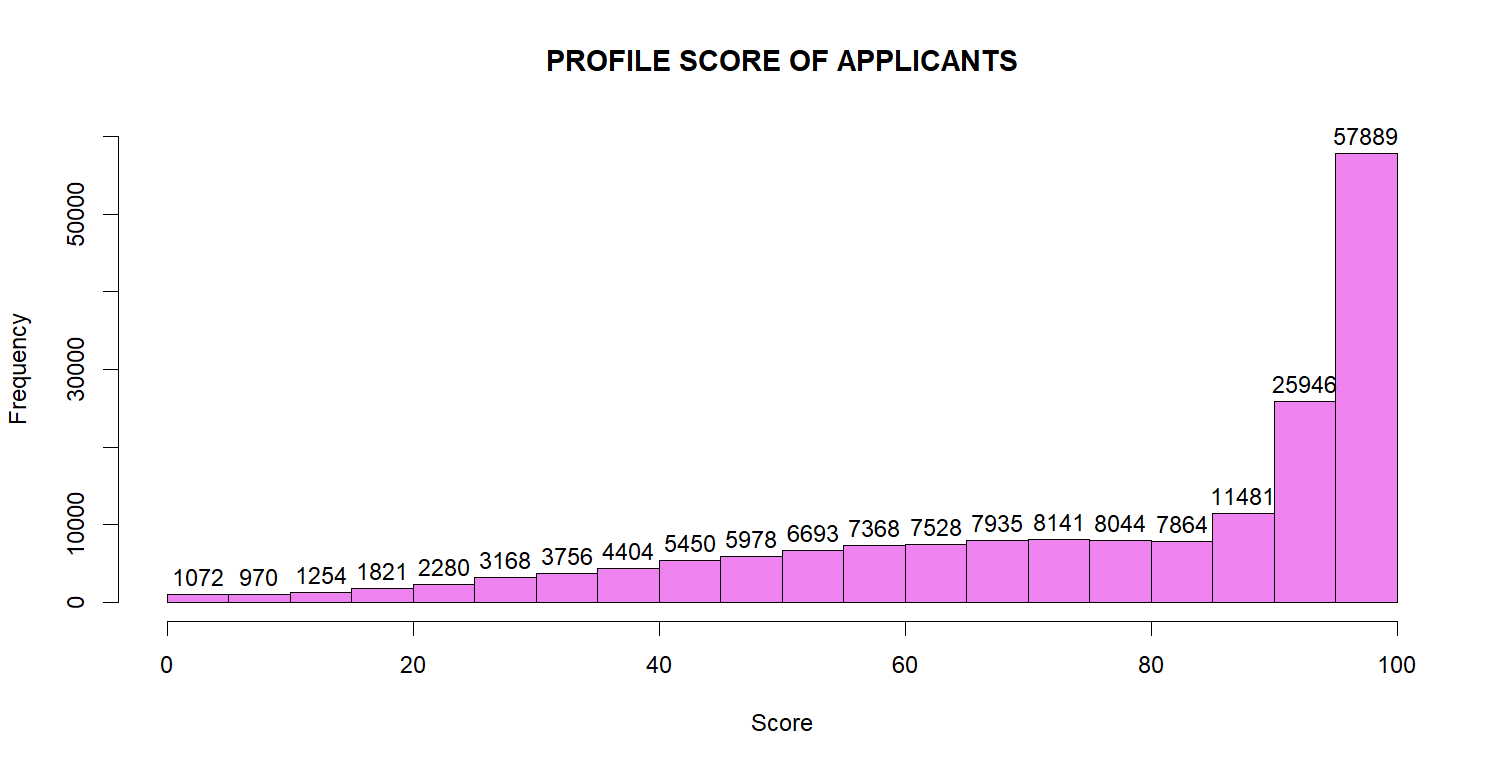
#we observe the distribution of income among loan applicants is right-skewed. Majority of people have income less than 1L  
h2<-hist(df$Loan.Amount,main="LOAN AMOUNT OF APPLICANTS", xlab="Amount (in rupee)", col="purple", ylim=c(0,70000))  
text(h2$mids,h2$counts,labels=h2$counts,adj=c(0.5,-0.5))



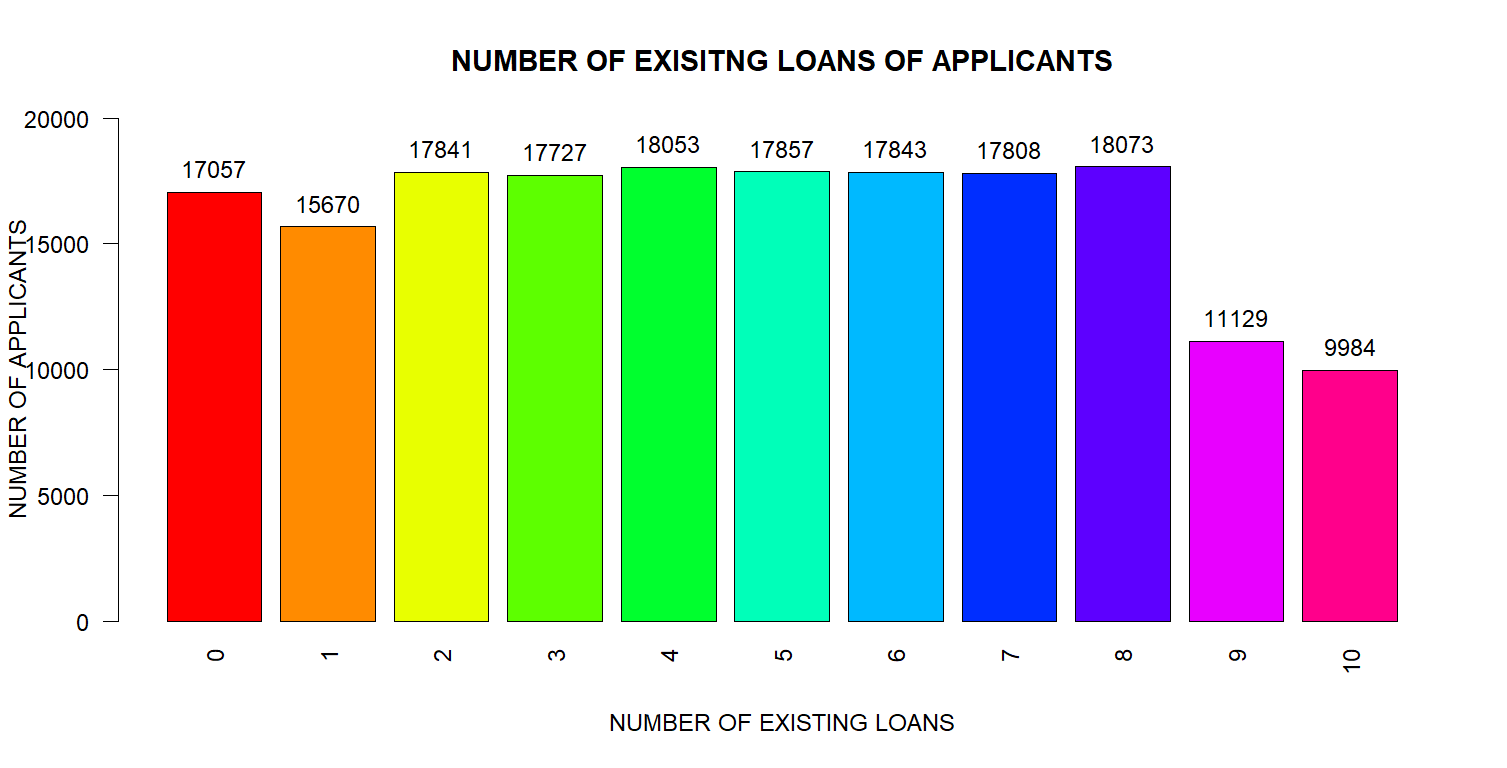
#we observe majority of applicants apply for loan of about 1.5L  
h3<-hist(df$Loan.Tenure,main="LOAN TENURE OF APPLICANTS", xlab="Months", col="pink", ylim=c(0,25000))  
text(h3$mids,h3$counts,labels=h3$counts,adj=c(0.5,-0.5))



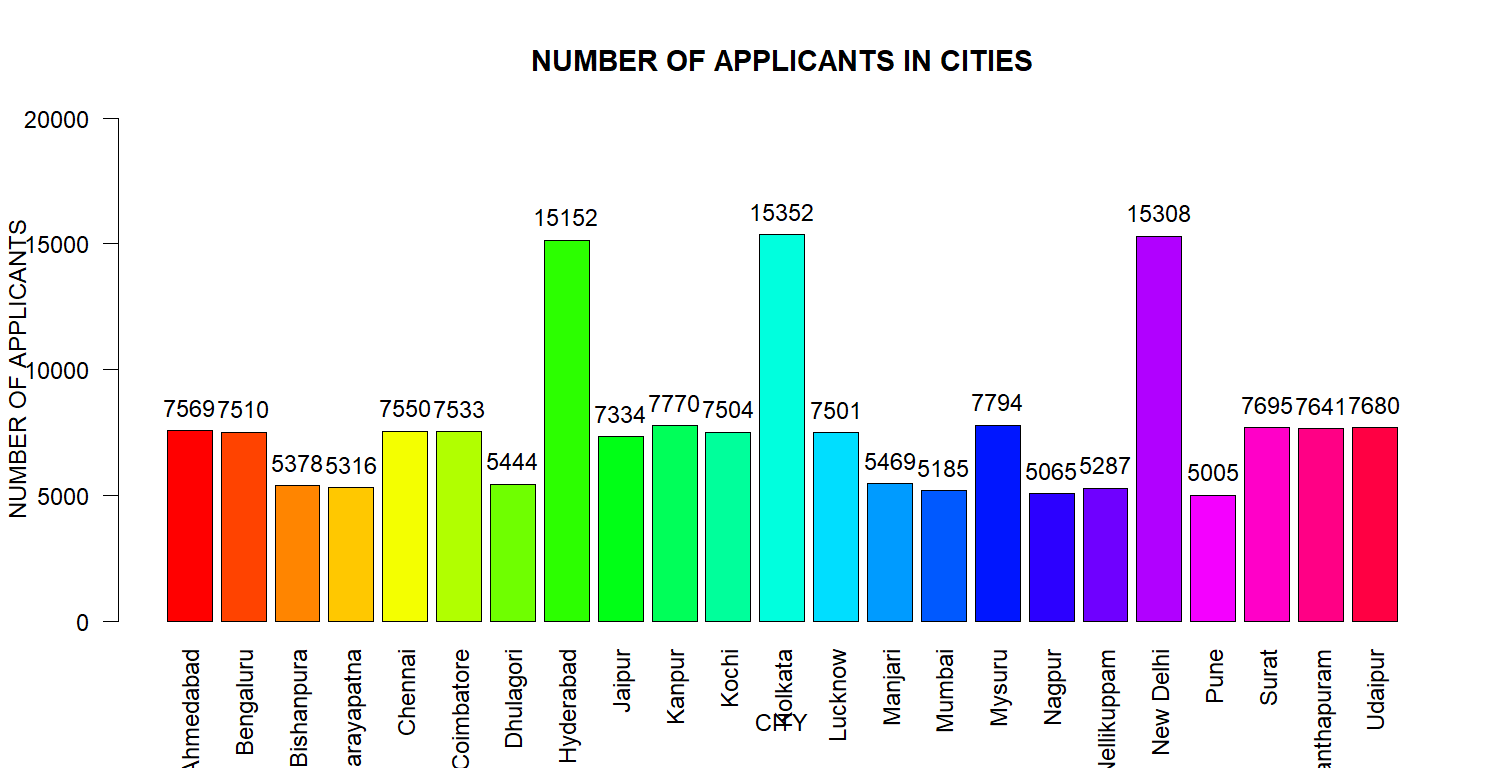
#We observe the distribution of loan tenure is right- skewed. Majority of applicants want a tenure shorter than 100 momths (~= 8yrs)  
h4<-hist(df$Profile.Score,main="PROFILE SCORE OF APPLICANTS", xlab="Score", col="pink", ylim=c(0,60000))  
text(h4$mids,h4$counts,labels=h4$counts,adj=c(0.5,-0.5))



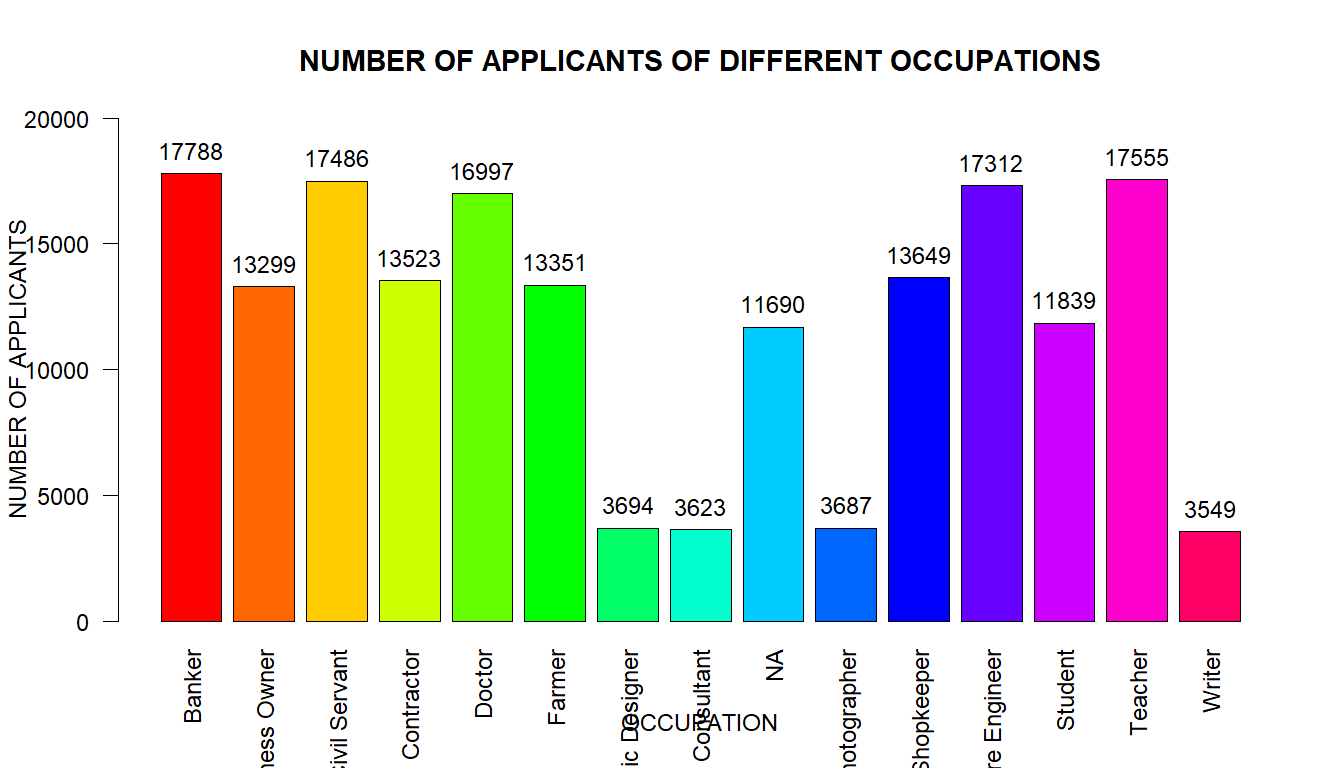
#We observe that the distribution of profile scores is left-skewed, with majority of applicants having a score of 100  
  
# Bar chart   
b1<-barplot(table(df$Number.of.Existing.Loans),xlab="NUMBER OF EXISTING LOANS", ylab="NUMBER OF APPLICANTS",main="NUMBER OF EXISITNG LOANS OF APPLICANTS", col=rainbow(length(table(df$Number.of.Existing.Loans))),las=2,ylim=c(0,20000))  
text(b1, table(df$Number.of.Existing.Loans), round(table(df$Number.of.Existing.Loans), 1),cex=1,pos=3)



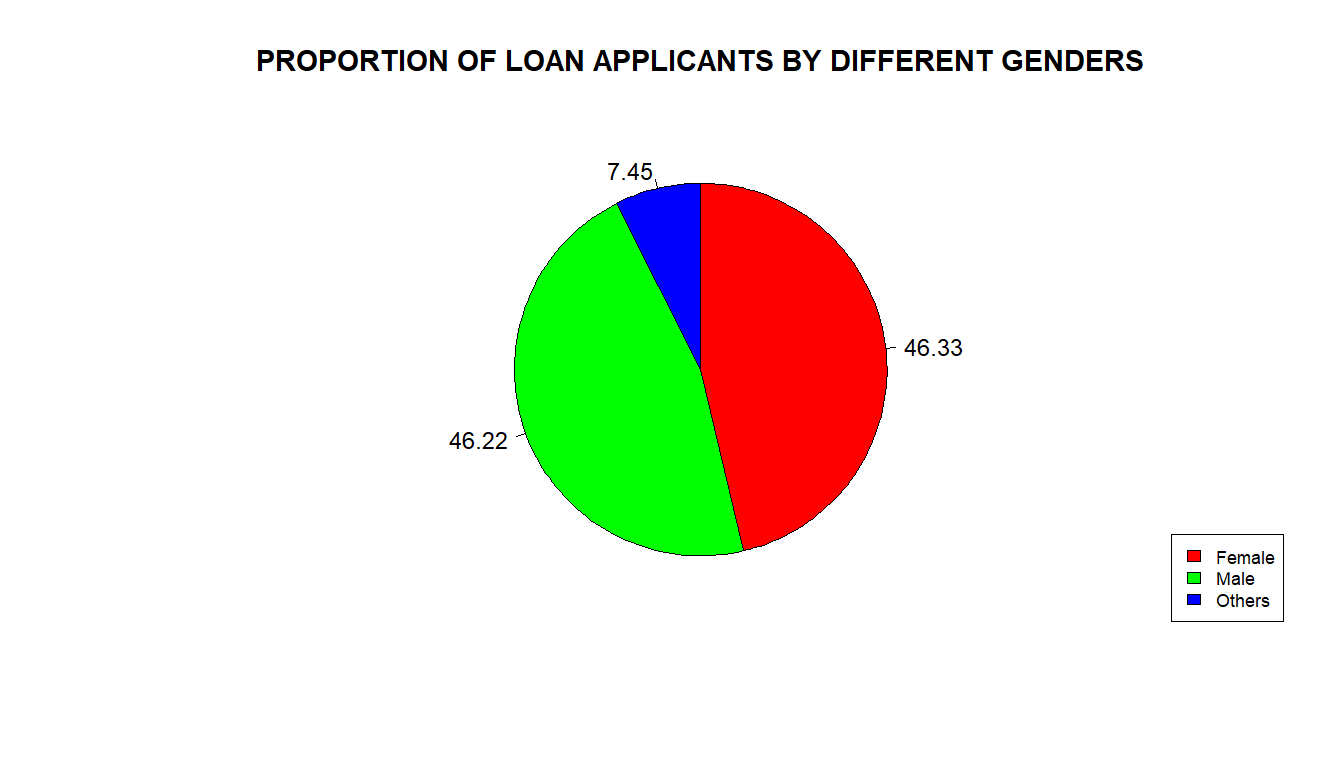
#least number of applicants had 10 existing loans.  
b2<-barplot(table(df$City),xlab="CITY", ylab="NUMBER OF APPLICANTS",main="NUMBER OF APPLICANTS IN CITIES", col=rainbow(length(table(df$City))),las=2,ylim=c(0,20000))  
text(b2, table(df$City), round(table(df$City), 1),cex=1,pos=3)



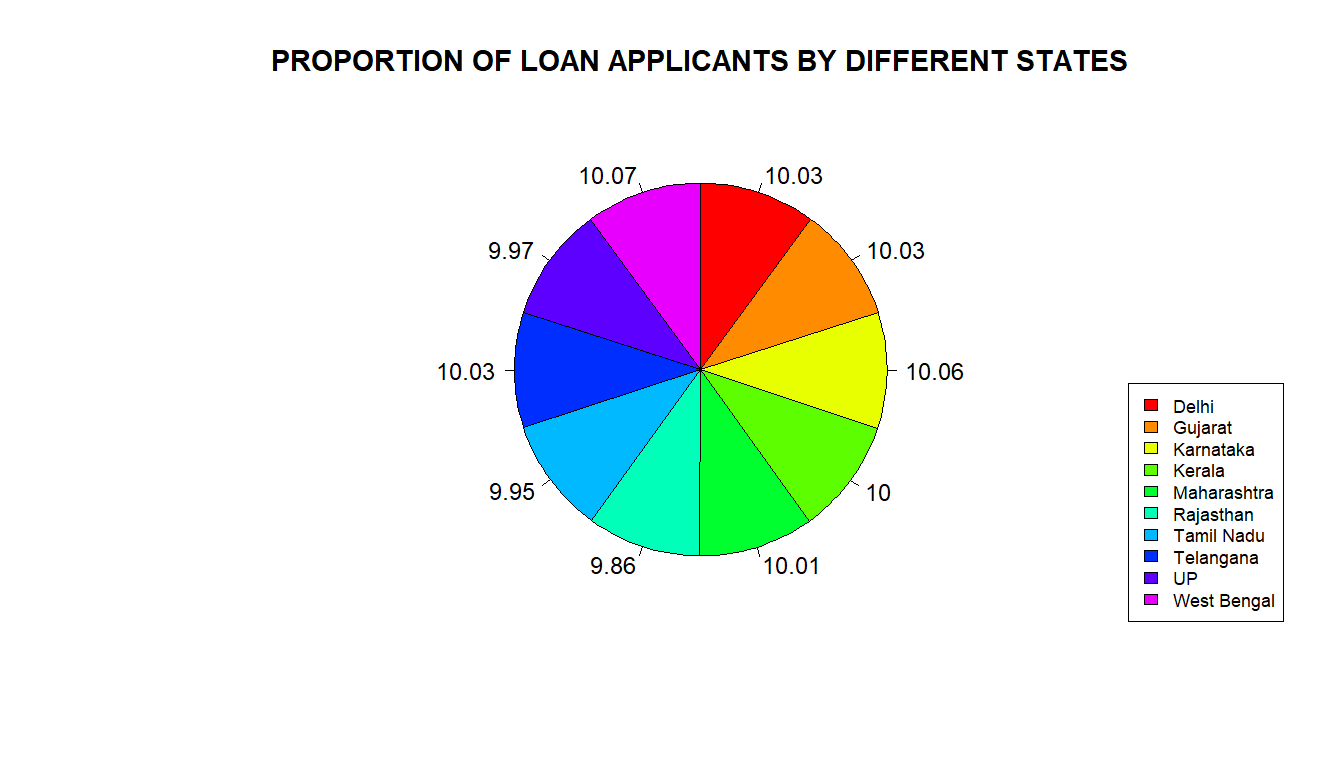
#New Delhi, Hyderabad and Kolkata had the maximum number of loan applicants.  
b3<-barplot(table(df$Occupation),xlab="OCCUPATION", ylab="NUMBER OF APPLICANTS",main="NUMBER OF APPLICANTS OF DIFFERENT OCCUPATIONS", col=rainbow(length(table(df$Occupation))),las=2,ylim=c(0,20000))  
text(b3, table(df$Occupation), round(table(df$Occupation), 1),cex=1,pos=3)



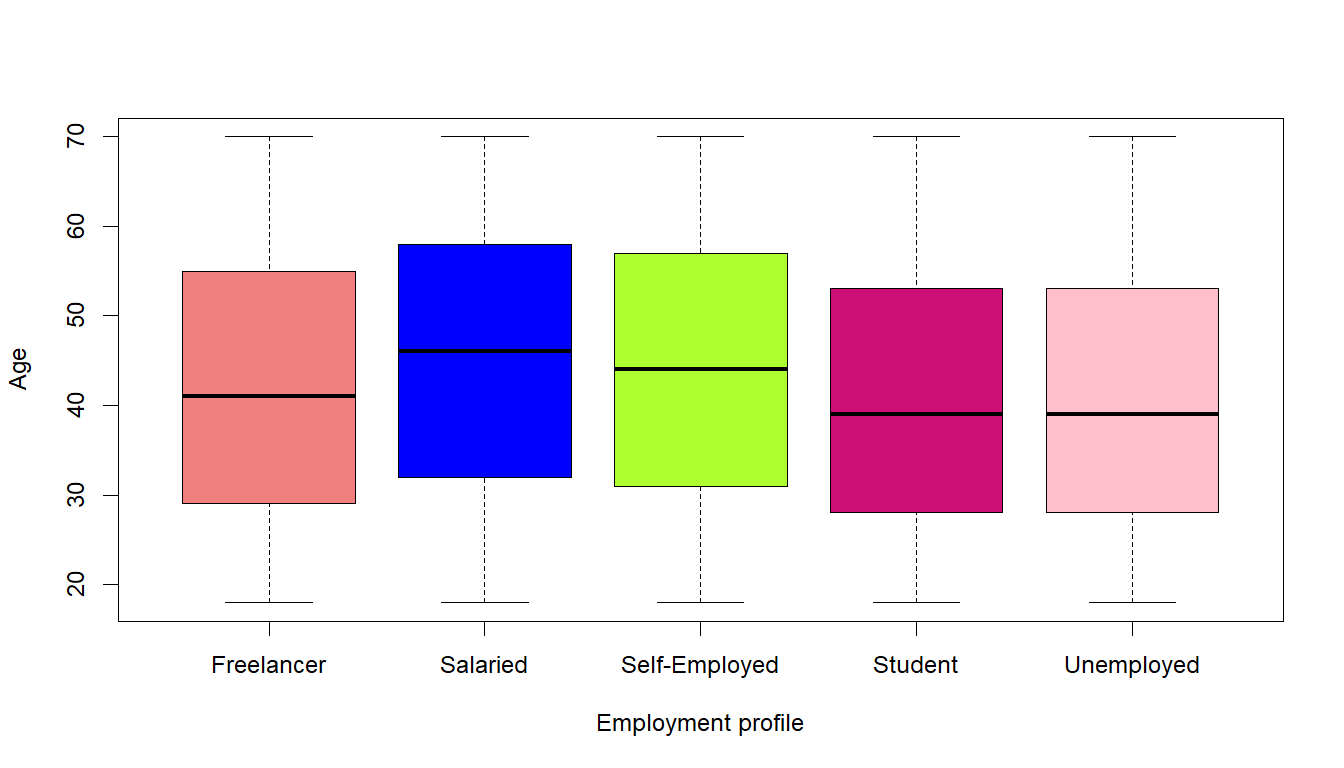
#Bankers, Civil servants,Software engineers and Teachers have the most number of loan applicants.  
  
#Piechart  
table1<-table(df$Gender)   
piepercent1<- round(100 \* table1 / sum(table1), 2)   
pie1<-pie(table1,piepercent1, main="PROPORTION OF LOAN APPLICANTS BY DIFFERENT GENDERS", col=rainbow(length(table1)), clockwise=TRUE)   
legend("bottomright", c("Female","Male","Others"), cex = 0.75, fill = rainbow(length(table1)))



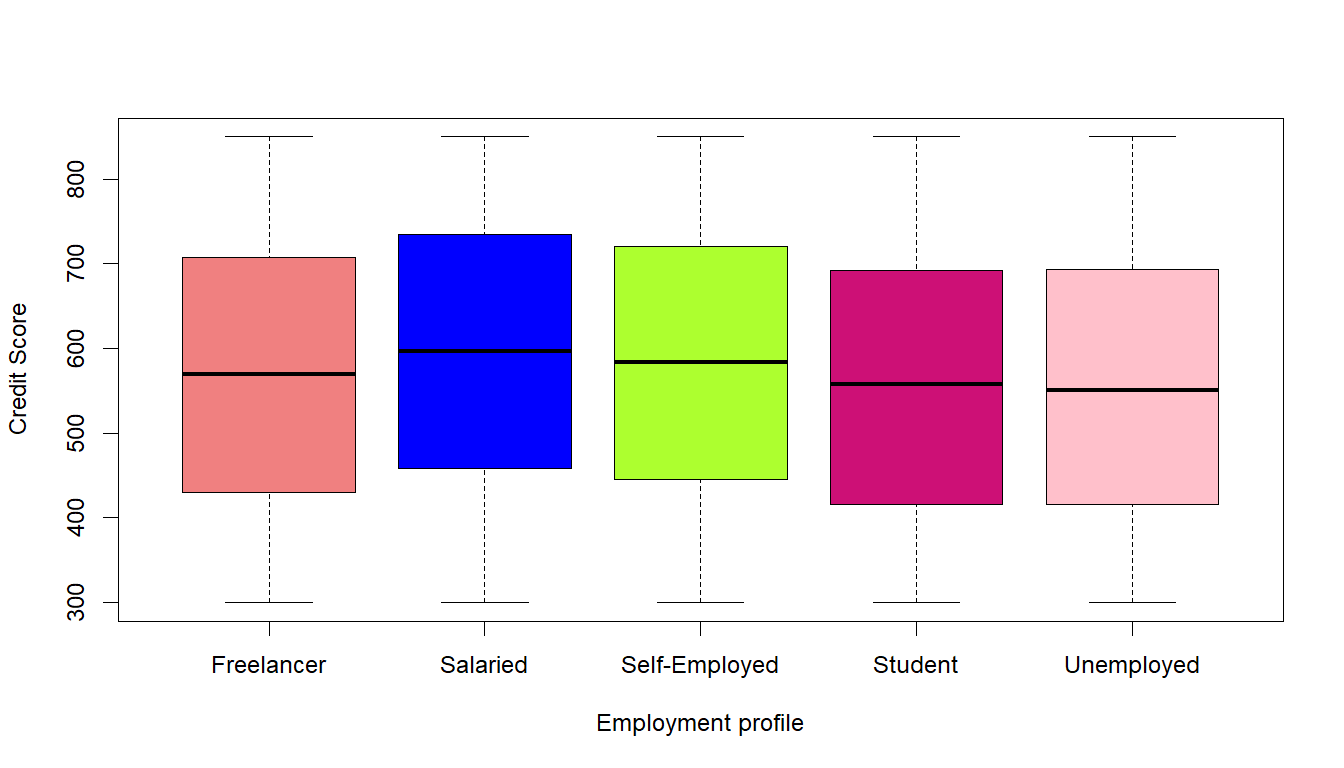
#Proportion of female and male applicants is almost equal (~=46%)  
table2<-table(df$State)   
piepercent2<- round(100 \* table2 / sum(table2), 2)   
pie2<-pie(table2,piepercent2, main="PROPORTION OF LOAN APPLICANTS BY DIFFERENT STATES", col=rainbow(11), clockwise=TRUE)   
legend("bottomright", c("Delhi","Gujarat","Karnataka","Kerala","Maharashtra","Rajasthan","Tamil Nadu","Telangana","UP","West Bengal"), cex = 0.75, fill = rainbow(11))



#Proportion of applicants is almost same for all states  
  
#Boxplot  
p1<-boxplot(df$Age~df$Employment.Profile, data=df, ylab="Age",xlab="Employment profile", col=c("lightcoral","blue","greenyellow","deeppink3","pink"))



#Median age for loan applicants is highest for salaried applicants (46 yrs).   
p2<-boxplot(df$Credit.Score~df$Employment.Profile, data=df, ylab="Credit Score",xlab="Employment profile", col=c("lightcoral","blue","greenyellow","deeppink3","pink"))



#Salaried people have a slightly higher median credit score compared to other employment profiles. median credit scores for freelancers, students and unemployed people is almost the same  
  
#Making a df of only numerical variables to check for correlation  
df1 <- subset(df, select = -c(2, 9,10,11,13,15))  
# Correlation matrix  
library(Hmisc)

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

rcorr(as.matrix(df1))

## Age Income Credit.Score Credit.History.Length  
## Age 1.00 0.62 0.13 0  
## Income 0.62 1.00 0.22 0  
## Credit.Score 0.13 0.22 1.00 0  
## Credit.History.Length 0.00 0.00 0.00 1  
## Number.of.Existing.Loans 0.13 0.22 0.99 0  
## Loan.Amount 0.27 0.39 0.08 0  
## Loan.Tenure 0.06 0.11 0.65 0  
## LTV.Ratio -0.04 -0.07 -0.38 0  
## Profile.Score 0.10 0.18 0.78 0  
## Number.of.Existing.Loans Loan.Amount Loan.Tenure  
## Age 0.13 0.27 0.06  
## Income 0.22 0.39 0.11  
## Credit.Score 0.99 0.08 0.65  
## Credit.History.Length 0.00 0.00 0.00  
## Number.of.Existing.Loans 1.00 0.08 0.64  
## Loan.Amount 0.08 1.00 0.05  
## Loan.Tenure 0.64 0.05 1.00  
## LTV.Ratio -0.38 -0.03 -0.24  
## Profile.Score 0.76 0.07 0.49  
## LTV.Ratio Profile.Score  
## Age -0.04 0.10  
## Income -0.07 0.18  
## Credit.Score -0.38 0.78  
## Credit.History.Length 0.00 0.00  
## Number.of.Existing.Loans -0.38 0.76  
## Loan.Amount -0.03 0.07  
## Loan.Tenure -0.24 0.49  
## LTV.Ratio 1.00 -0.54  
## Profile.Score -0.54 1.00  
##   
## n= 179042   
##   
##   
## P  
## Age Income Credit.Score Credit.History.Length  
## Age 0.0000 0.0000 0.2760   
## Income 0.0000 0.0000 0.3344   
## Credit.Score 0.0000 0.0000 0.1232   
## Credit.History.Length 0.2760 0.3344 0.1232   
## Number.of.Existing.Loans 0.0000 0.0000 0.0000 0.1180   
## Loan.Amount 0.0000 0.0000 0.0000 0.5149   
## Loan.Tenure 0.0000 0.0000 0.0000 0.1862   
## LTV.Ratio 0.0000 0.0000 0.0000 0.6784   
## Profile.Score 0.0000 0.0000 0.0000 0.5475   
## Number.of.Existing.Loans Loan.Amount Loan.Tenure  
## Age 0.0000 0.0000 0.0000   
## Income 0.0000 0.0000 0.0000   
## Credit.Score 0.0000 0.0000 0.0000   
## Credit.History.Length 0.1180 0.5149 0.1862   
## Number.of.Existing.Loans 0.0000 0.0000   
## Loan.Amount 0.0000 0.0000   
## Loan.Tenure 0.0000 0.0000   
## LTV.Ratio 0.0000 0.0000 0.0000   
## Profile.Score 0.0000 0.0000 0.0000   
## LTV.Ratio Profile.Score  
## Age 0.0000 0.0000   
## Income 0.0000 0.0000   
## Credit.Score 0.0000 0.0000   
## Credit.History.Length 0.6784 0.5475   
## Number.of.Existing.Loans 0.0000 0.0000   
## Loan.Amount 0.0000 0.0000   
## Loan.Tenure 0.0000 0.0000   
## LTV.Ratio 0.0000   
## Profile.Score 0.0000

#Age and Income have a strong positive correlation (0.62)  
#Credit.Score and Income also have a positive correlation (0.22).  
#LTV.Ratio and Profile.Score have a moderate negative correlation (-0.54).  
#Credit.Score and Number.of.Existing.Loans have a very strong positive correlation (0.99)  
#Credit.History.Length and all other variables have zero correlation, suggesting no linear relationship.