ML Assignment #2

Tom Scarberry

9/23/2020

##load UniversalBank file and see column names and summary data

```
bank_full_all <- read.csv("UniversalBank.csv")
colnames(bank_full_all)
```

```
##  [1] "ID"                 "Age"                "Experience"
##  [4] "Income"             "ZIP.Code"           "Family"
##  [7] "CCAvg"              "Education"          "Mortgage"
## [10] "Personal.Loan"      "Securities.Account" "CD.Account"
## [13] "Online"             "CreditCard"
```

```
summary(bank_full_all)
```

```
##        ID            Age          Experience       Income          ZIP.Code
##  Min.   :   1   Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   : 9307
##  1st Qu.:1251   1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:91911
##  Median :2500   Median :45.00   Median :20.0   Median : 64.00   Median :93437
##  Mean   :2500   Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :93153
##  3rd Qu.:3750   3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:94608
##  Max.   :5000   Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :96651
##      Family          CCAvg          Education        Mortgage
##  Min.   :1.000   Min.   : 0.000   Min.   :1.000   Min.   :  0.0
##  1st Qu.:1.000   1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0
##  Median :2.000   Median : 1.500   Median :2.000   Median :  0.0
##  Mean   :2.396   Mean   : 1.938   Mean   :1.881   Mean   : 56.5
##  3rd Qu.:3.000   3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0
##  Max.   :4.000   Max.   :10.000   Max.   :3.000   Max.   :635.0
##  Personal.Loan   Securities.Account   CD.Account         Online
##  Min.   :0.000   Min.   :0.0000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000     Median :0.0000   Median :1.0000
##  Mean   :0.096   Mean   :0.1044     Mean   :0.0604   Mean   :0.5968
##  3rd Qu.:0.000   3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000     Max.   :1.0000   Max.   :1.0000
##    CreditCard
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.294
##  3rd Qu.:1.000
##  Max.   :1.000
```

```r
head(bank_full_all)
```

```
##   ID Age Experience Income ZIP.Code Family CCAvg Education Mortgage
## 1  1  25          1     49    91107      4   1.6         1        0
## 2  2  45         19     34    90089      3   1.5         1        0
## 3  3  39         15     11    94720      1   1.0         1        0
## 4  4  35          9    100    94112      1   2.7         2        0
## 5  5  35          8     45    91330      4   1.0         2        0
## 6  6  37         13     29    92121      4   0.4         2      155
##   Personal.Loan Securities.Account CD.Account Online CreditCard
## 1             0                  1          0      0          0
## 2             0                  1          0      0          0
## 3             0                  0          0      0          0
## 4             0                  0          0      0          0
## 5             0                  0          0      0          1
## 6             0                  0          0      1          0
```

```r
class(bank_full_all$Personal.Loan)
```

```
## [1] "integer"
```

```r
str(bank_full_all)
```

```
## 'data.frame':    5000 obs. of  14 variables:
##  $ ID                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age               : int  25 45 39 35 35 37 53 50 35 34 ...
##  $ Experience        : int  1 19 15 9 8 13 27 24 10 9 ...
##  $ Income            : int  49 34 11 100 45 29 72 22 81 180 ...
##  $ ZIP.Code          : int  91107 90089 94720 94112 91330 92121 91711 93943 90089 93023 ...
##  $ Family            : int  4 3 1 1 4 4 2 1 3 1 ...
##  $ CCAvg             : num  1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...
##  $ Education         : int  1 1 1 2 2 2 2 3 2 3 ...
##  $ Mortgage          : int  0 0 0 0 0 155 0 0 104 0 ...
##  $ Personal.Loan     : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Securities.Account: int  1 1 0 0 0 0 0 0 0 0 ...
##  $ CD.Account        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Online            : int  0 0 0 0 0 1 1 0 1 0 ...
##  $ CreditCard        : int  0 0 0 0 1 0 0 1 0 0 ...
```

##load libraries

```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
```

```
library(caret)
```

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

```
## Loading required package: ggplot2
```

```
library(gmodels)
library(ISLR)
library(lattice)
library(ggplot2)
library(FNN)
library(e1071)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

create dummy variables for Education and Family

```
library(fastDummies)
bank_fulla<-dummy_cols(bank_full_all,select_columns ='Education')

bank_full<-dummy_cols(bank_fulla,select_columns ='Family')
summary (bank_full)
```

```
##        ID            Age          Experience       Income          ZIP.Code
##  Min.   :   1   Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   : 9307
##  1st Qu.:1251   1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:91911
##  Median :2500   Median :45.00   Median :20.0   Median : 64.00   Median :93437
##  Mean   :2500   Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :93153
##  3rd Qu.:3750   3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:94608
##  Max.   :5000   Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :96651
##      Family          CCAvg          Education        Mortgage
##  Min.   :1.000   Min.   : 0.000   Min.   :1.000   Min.   :  0.0
##  1st Qu.:1.000   1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0
##  Median :2.000   Median : 1.500   Median :2.000   Median :  0.0
##  Mean   :2.396   Mean   : 1.938   Mean   :1.881   Mean   : 56.5
##  3rd Qu.:3.000   3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0
##  Max.   :4.000   Max.   :10.000   Max.   :3.000   Max.   :635.0
##  Personal.Loan   Securities.Account   CD.Account         Online
##  Min.   :0.000   Min.   :0.0000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000     Median :0.0000   Median :1.0000
##  Mean   :0.096   Mean   :0.1044     Mean   :0.0604   Mean   :0.5968
##  3rd Qu.:0.000   3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000     Max.   :1.0000   Max.   :1.0000
##    CreditCard     Education_1      Education_2      Education_3
```

```
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.294   Mean   :0.4192   Mean   :0.2806   Mean   :0.3002
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     Family_1         Family_2         Family_3        Family_4
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.000   Median :0.0000
##  Mean   :0.2944   Mean   :0.2592   Mean   :0.202   Mean   :0.2444
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
```

create database without ID and Zip Code variables

```
bank<-select(bank_full,Personal.Loan, CD.Account, Education_1, Education_2, Education_3, Family_1, Famil
summary(bank)
```

```
##  Personal.Loan     CD.Account      Education_1      Education_2
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.096   Mean   :0.0604   Mean   :0.4192   Mean   :0.2806
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##   Education_3        Family_1         Family_2         Family_3
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000
##  Mean   :0.3002   Mean   :0.2944   Mean   :0.2592   Mean   :0.202
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
##     Family_4      Securities.Account   CreditCard        Online
##  Min.   :0.0000   Min.   :0.0000     Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000     Median :0.000   Median :1.0000
##  Mean   :0.2444   Mean   :0.1044     Mean   :0.294   Mean   :0.5968
##  3rd Qu.:0.0000   3rd Qu.:0.0000     3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000     Max.   :1.000   Max.   :1.0000
##      Income           CCAvg            Mortgage          Age
##  Min.   :  8.00   Min.   : 0.000   Min.   :  0.0   Min.   :23.00
##  1st Qu.: 39.00   1st Qu.: 0.700   1st Qu.:  0.0   1st Qu.:35.00
##  Median : 64.00   Median : 1.500   Median :  0.0   Median :45.00
##  Mean   : 73.77   Mean   : 1.938   Mean   : 56.5   Mean   :45.34
##  3rd Qu.: 98.00   3rd Qu.: 2.500   3rd Qu.:101.0   3rd Qu.:55.00
##  Max.   :224.00   Max.   :10.000   Max.   :635.0   Max.   :67.00
##    Experience
##  Min.   :-3.0
##  1st Qu.:10.0
##  Median :20.0
##  Mean   :20.1
```

```
##  3rd Qu.:30.0
##  Max.   :43.0
```

##partition data into training and validation sets

```
set.seed(12)
Train_Index=createDataPartition(bank$Personal.Loan,p=0.6, list=FALSE)
Train_Data=bank[Train_Index,]
Test_Data=bank[-Train_Index,]

summary(Train_Data$Personal.Loan)
```

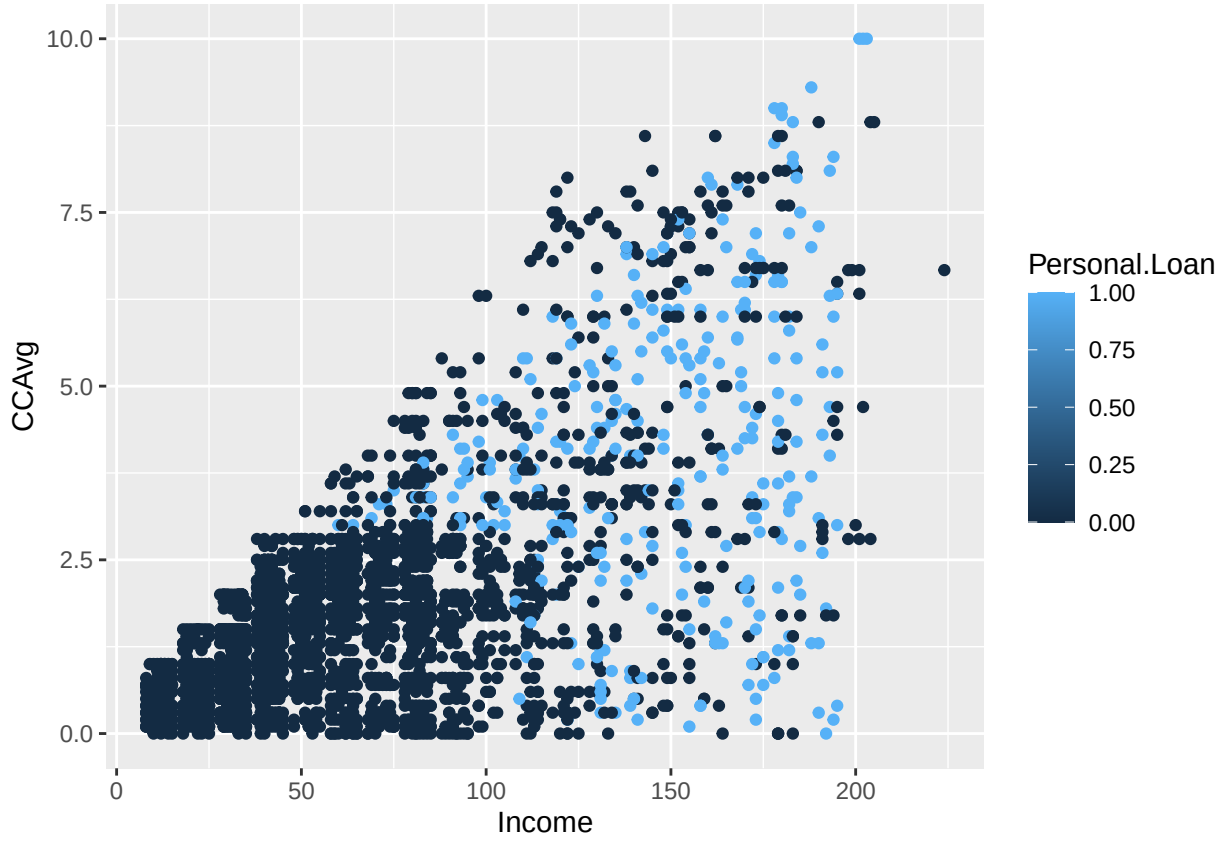
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.09967 0.00000 1.00000
```

```
summary(Test_Data$Personal.Loan)
```

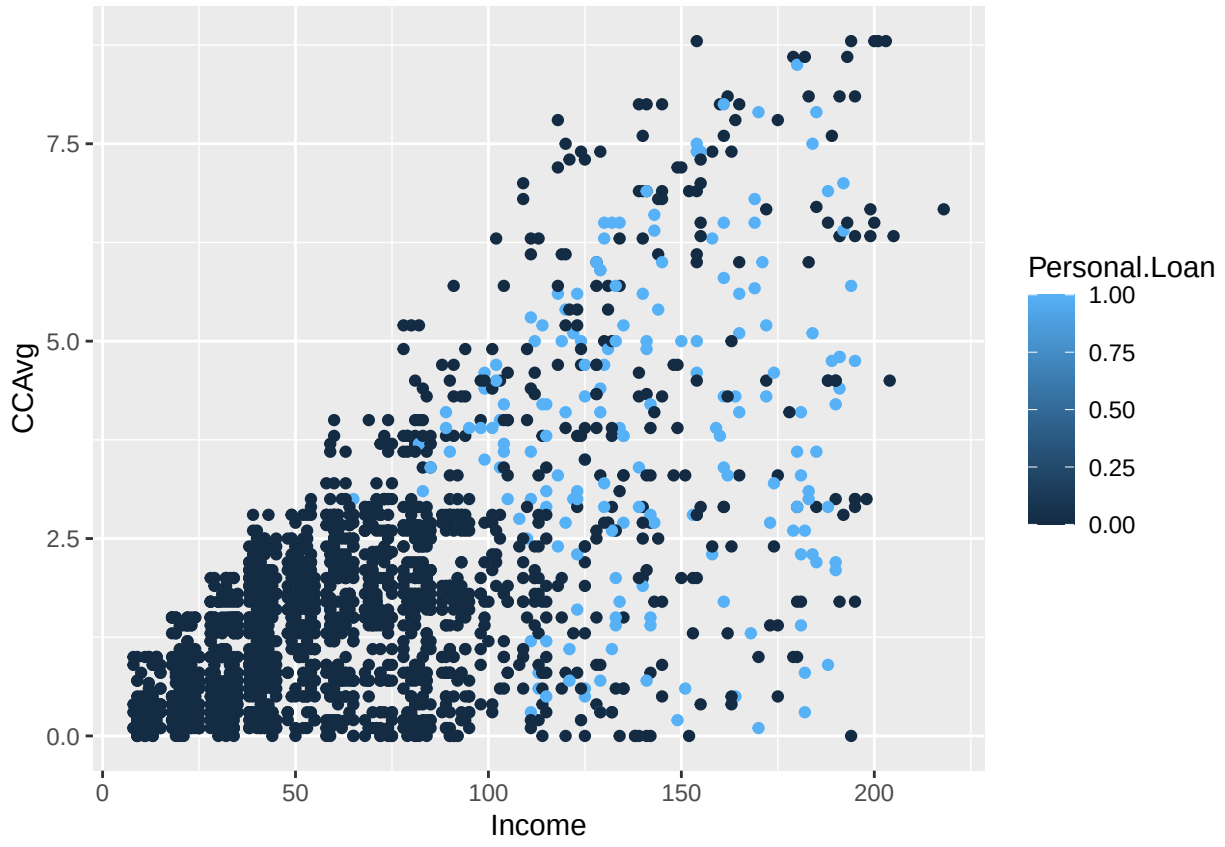
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.0905  0.0000  1.0000
```

###plot the data

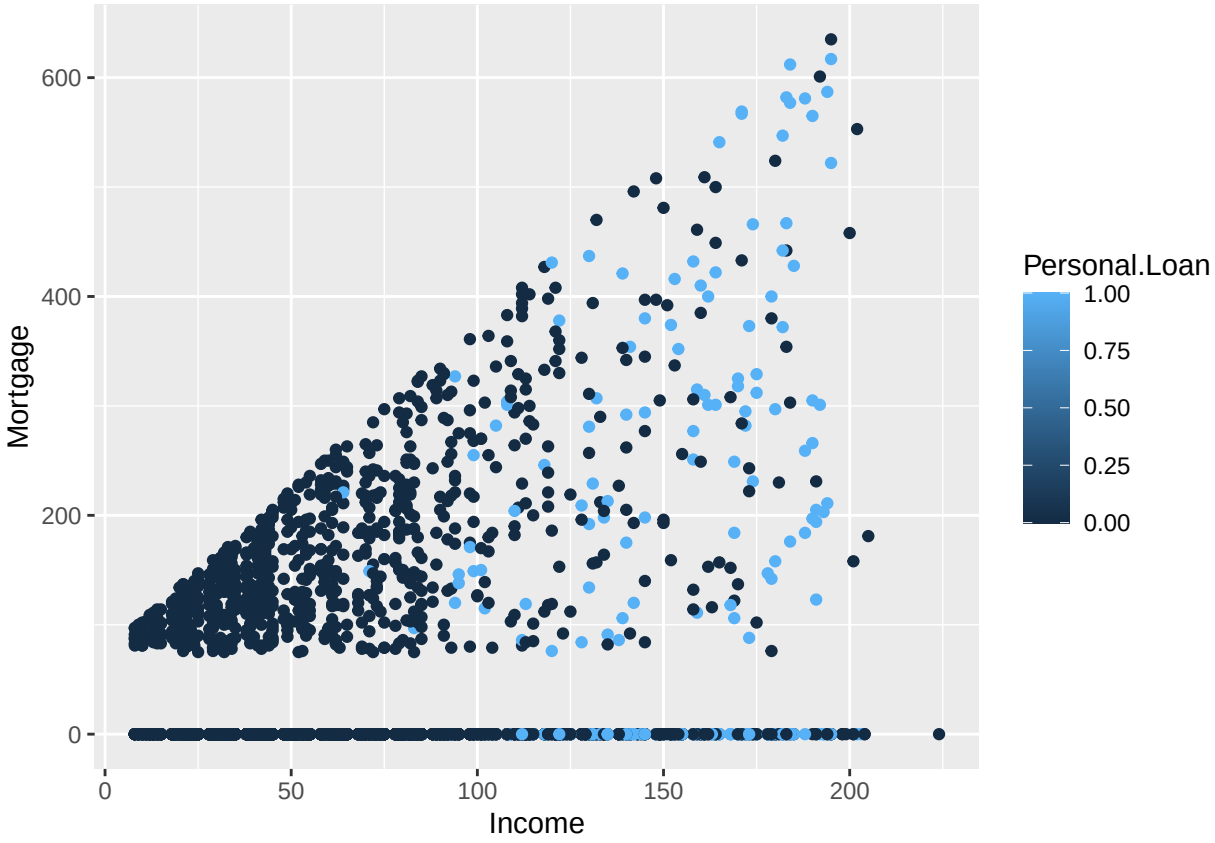
```
ggplot(Train_Data, aes(x=Income,y=CCAvg, color=Personal.Loan)) +geom_point()
```



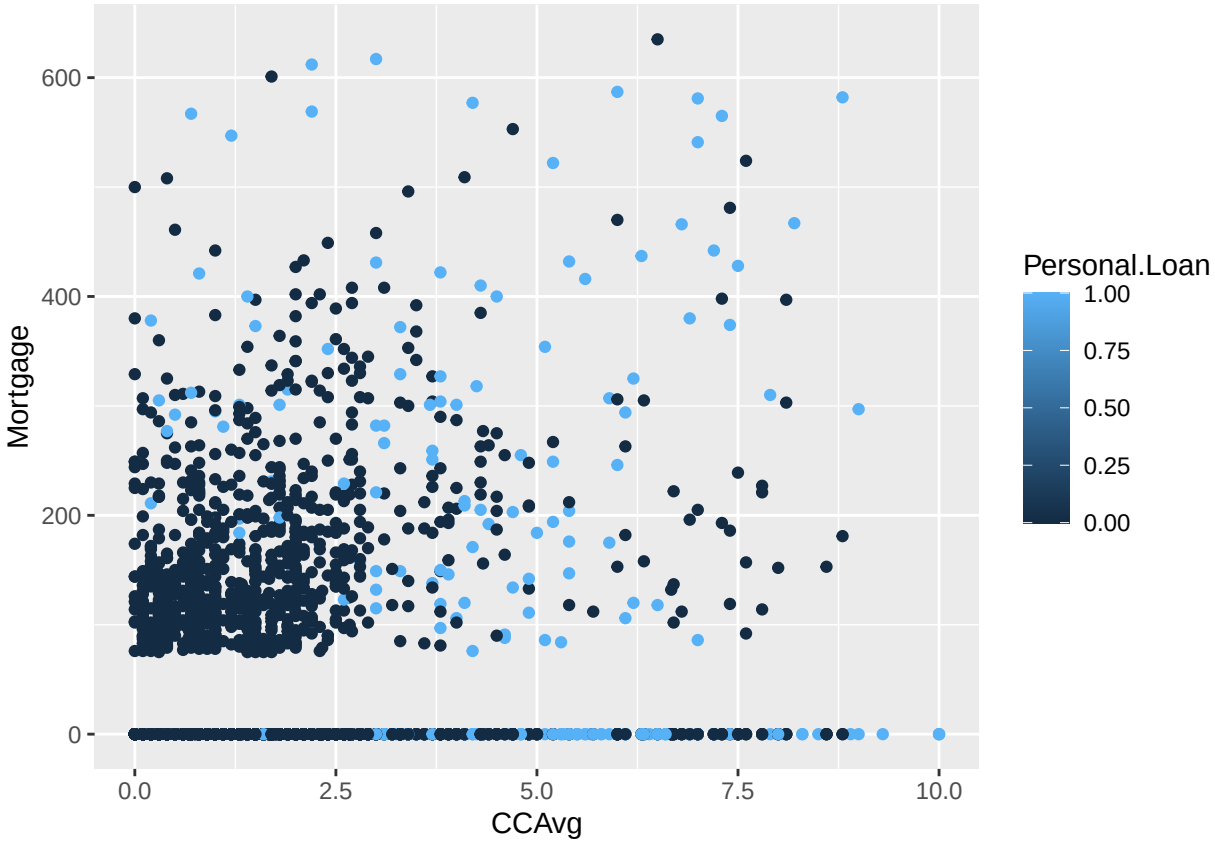
```
ggplot(Test_Data, aes(x=Income,y=CCAvg, color=Personal.Loan)) +geom_point()
```




```
ggplot(Train_Data, aes(x=Income,y=Mortgage, color=Personal.Loan)) +geom_point()
```




```
ggplot(Train_Data, aes(x=CCAvg,y=Mortgage, color=Personal.Loan)) +geom_point()
```



#normalize the data

```
train.norm.df<-Train_Data
test.norm.df<-Test_Data

test_z_norm<-preProcess(Train_Data[ ,13:17],method = c("center","scale"))

train.norm.df[, 13:17]<-predict(test_z_norm, Train_Data[13:17])
test.norm.df[, 13:17]<-predict(test_z_norm, Test_Data[13:17])

summary(train.norm.df)
```

```
##  Personal.Loan       CD.Account       Education_1     Education_2
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.000   Median :0.0000
##  Mean   :0.09967   Mean   :0.05833   Mean   :0.419   Mean   :0.2827
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.000   Max.   :1.0000
##   Education_3        Family_1         Family_2         Family_3
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000
##  Mean   :0.2983   Mean   :0.2867   Mean   :0.2597   Mean   :0.203
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
```

```
##     Family_4      Securities.Account   CreditCard         Online
##  Min.   :0.0000   Min.   :0.0000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000     Median :0.0000   Median :1.0000
##  Mean   :0.2507   Mean   :0.1007     Mean   :0.2983   Mean   :0.5927
##  3rd Qu.:1.0000   3rd Qu.:0.0000     3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000     Max.   :1.0000   Max.   :1.0000
##      Income            CCAvg            Mortgage            Age
##  Min.   :-1.4191   Min.   :-1.1053   Min.   :-0.5524   Min.   :-1.93623
##  1st Qu.:-0.7542   1st Qu.:-0.7101   1st Qu.:-0.5524   1st Qu.:-0.89391
##  Median :-0.2395   Median :-0.2585   Median :-0.5524   Median :-0.02531
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.5379   3rd Qu.: 0.3059   3rd Qu.: 0.4465   3rd Qu.: 0.84330
##  Max.   : 3.2133   Max.   : 4.5396   Max.   : 5.6665   Max.   : 1.88562
##    Experience
##  Min.   :-2.001532
##  1st Qu.:-0.872359
##  Median :-0.003764
##  Mean   : 0.000000
##  3rd Qu.: 0.864831
##  Max.   : 1.994004
```

##k-NN set up

```
Train.Predictors <-train.norm.df[,2:17]
Test.Predictors <- test.norm.df[,2:17]


Train.Labels <-factor(train.norm.df[,1])
Test.Labels <-factor(test.norm.df[,1])
class(Test.Labels)
```

```
## [1] "factor"
```

```
summary(Train.Labels)
```

```
##    0    1
## 2701  299
```

```
summary(Test.Labels)
```

```
##    0    1
## 1819  181
```

train knn model with k=1

```
##k=1
c1<-Train.Labels
Predicted.Test.Labels <- knn (Train.Predictors, Test.Predictors, c1, k=1, prob = TRUE )
CrossTable(x=Test.Labels,y=Predicted.Test.Labels)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  2000
##
##
##              | Predicted.Test.Labels
##  Test.Labels |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |      1804 |        15 |      1819 |
##              |     6.496 |    91.928 |           |
##              |     0.992 |     0.008 |     0.909 |
##              |     0.966 |     0.114 |           |
##              |     0.902 |     0.007 |           |
## -------------|-----------|-----------|-----------|
##            1 |        64 |       117 |       181 |
##              |    65.283 |   923.853 |           |
##              |     0.354 |     0.646 |     0.090 |
##              |     0.034 |     0.886 |           |
##              |     0.032 |     0.058 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1868 |       132 |      2000 |
##              |     0.934 |     0.066 |           |
## -------------|-----------|-----------|-----------|
##
##
```

knn model with k of 1 has fairly high precision scoring 89%, but low recall (sensitivity) with a score of 65%. This model generally errors on the the side of falsely predicting a no for the personal loan acceptance, when the customer would accept the loan. That will lead to lost revenue opportunities for the bank.

```
summary(bank)
```

```
##  Personal.Loan     CD.Account      Education_1      Education_2
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.096   Mean   :0.0604   Mean   :0.4192   Mean   :0.2806
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##   Education_3        Family_1         Family_2         Family_3
```

```
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000
##  Mean   :0.3002   Mean   :0.2944   Mean   :0.2592   Mean   :0.202
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
##     Family_4      Securities.Account   CreditCard        Online
##  Min.   :0.0000   Min.   :0.0000     Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000     Median :0.000   Median :1.0000
##  Mean   :0.2444   Mean   :0.1044     Mean   :0.294   Mean   :0.5968
##  3rd Qu.:0.0000   3rd Qu.:0.0000     3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000     Max.   :1.000   Max.   :1.0000
##      Income           CCAvg            Mortgage          Age
##  Min.   :  8.00   Min.   : 0.000   Min.   :  0.0   Min.   :23.00
##  1st Qu.: 39.00   1st Qu.: 0.700   1st Qu.:  0.0   1st Qu.:35.00
##  Median : 64.00   Median : 1.500   Median :  0.0   Median :45.00
##  Mean   : 73.77   Mean   : 1.938   Mean   : 56.5   Mean   :45.34
##  3rd Qu.: 98.00   3rd Qu.: 2.500   3rd Qu.:101.0   3rd Qu.:55.00
##  Max.   :224.00   Max.   :10.000   Max.   :635.0   Max.   :67.00
##    Experience
##  Min.   :-3.0
##  1st Qu.:10.0
##  Median :20.0
##  Mean   :20.1
##  3rd Qu.:30.0
##  Max.   :43.0
```

combine training and test data and renormalize the data

```
bank_norm_df<-bank
bank_z_norm<-preProcess(bank[13:17], method=c("center", "scale"))

bank_norm_df[,13:17]<-predict(bank_z_norm, bank[13:17])

summary(bank_norm_df)
```

```
##  Personal.Loan     CD.Account       Education_1      Education_2
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.096   Mean   :0.0604   Mean   :0.4192   Mean   :0.2806
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##   Education_3        Family_1         Family_2         Family_3
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000
##  Mean   :0.3002   Mean   :0.2944   Mean   :0.2592   Mean   :0.202
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
##     Family_4      Securities.Account   CreditCard        Online
```

```
##  Min.   :0.0000   Min.   :0.0000     Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000     Median :0.000   Median :1.0000
##  Mean   :0.2444   Mean   :0.1044     Mean   :0.294   Mean   :0.5968
##  3rd Qu.:0.0000   3rd Qu.:0.0000     3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000     Max.   :1.000   Max.   :1.0000
##      Income            CCAvg            Mortgage            Age
##  Min.   :-1.4288   Min.   :-1.1089   Min.   :-0.5555   Min.   :-1.94871
##  1st Qu.:-0.7554   1st Qu.:-0.7083   1st Qu.:-0.5555   1st Qu.:-0.90188
##  Median :-0.2123   Median :-0.2506   Median :-0.5555   Median :-0.02952
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.5263   3rd Qu.: 0.3216   3rd Qu.: 0.4375   3rd Qu.: 0.84284
##  Max.   : 3.2634   Max.   : 4.6131   Max.   : 5.6875   Max.   : 1.88967
##    Experience
##  Min.   :-2.014710
##  1st Qu.:-0.881116
##  Median :-0.009121
##  Mean   : 0.000000
##  3rd Qu.: 0.862874
##  Max.   : 1.996468
```

##k-NN set up

```
Bank_Predictors <-bank_norm_df[,2:17]

Bank_Labels <-factor(bank_norm_df[,1])

summary(Bank_Predictors)
```

```
##    CD.Account      Education_1      Education_2      Education_3
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.0604   Mean   :0.4192   Mean   :0.2806   Mean   :0.3002
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     Family_1         Family_2         Family_3        Family_4
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.000   Median :0.0000
##  Mean   :0.2944   Mean   :0.2592   Mean   :0.202   Mean   :0.2444
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##  Securities.Account   CreditCard        Online           Income
##  Min.   :0.0000     Min.   :0.000   Min.   :0.0000   Min.   :-1.4288
##  1st Qu.:0.0000     1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:-0.7554
##  Median :0.0000     Median :0.000   Median :1.0000   Median :-0.2123
##  Mean   :0.1044     Mean   :0.294   Mean   :0.5968   Mean   : 0.0000
##  3rd Qu.:0.0000     3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.: 0.5263
##  Max.   :1.0000     Max.   :1.000   Max.   :1.0000   Max.   : 3.2634
##      CCAvg            Mortgage            Age             Experience
##  Min.   :-1.1089   Min.   :-0.5555   Min.   :-1.94871   Min.   :-2.014710
##  1st Qu.:-0.7083   1st Qu.:-0.5555   1st Qu.:-0.90188   1st Qu.:-0.881116
```

```
##  Median :-0.2506   Median :-0.5555   Median :-0.02952   Median :-0.009121
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.000000
##  3rd Qu.: 0.3216   3rd Qu.: 0.4375   3rd Qu.: 0.84284   3rd Qu.: 0.862874
##  Max.   : 4.6131   Max.   : 5.6875   Max.   : 1.88967   Max.   : 1.996468
```

```
summary(Bank_Labels)
```

```
##    0    1
## 4520  480
```

New Customer set up

```
Personal.Loan<- c(0)
CD.Account<- c(0)
Education_1<- c(0)
Education_2<- c(1)
Education_3<- c(0)
Family_1<- c(0)
Family_2<- c(1)
Family_3<- c(0)
Family_4<- c(0)
Securities.Account<- c(0)
CreditCard<- c(1)
Online<- c(1)
Income<- c(84)
CCAvg<- c(2)
Mortgage<- c(0)
Age<- c(40)
Experience<- c(10)

New_Customer_1 <- data.frame(Personal.Loan, CD.Account, Education_1, Education_2, Education_3, Family_1
 str(New_Customer_1)
```

```
## 'data.frame':    1 obs. of  17 variables:
##  $ Personal.Loan     : num 0
##  $ CD.Account        : num 0
##  $ Education_1       : num 0
##  $ Education_2       : num 1
##  $ Education_3       : num 0
##  $ Family_1          : num 0
##  $ Family_2          : num 1
##  $ Family_3          : num 0
##  $ Family_4          : num 0
##  $ Securities.Account: num 0
##  $ CreditCard        : num 1
##  $ Online            : num 1
##  $ Income            : num 84
##  $ CCAvg             : num 2
##  $ Mortgage          : num 0
##  $ Age               : num 40
##  $ Experience        : num 10
```

normalize new customer

```
New_Customer_1_norm <- New_Customer_1
New_Customer_1_norm[,13:17]<-predict(bank_z_norm, New_Customer_1[13:17])
summary(New_Customer_1_norm)
```

```
##  Personal.Loan   CD.Account  Education_1  Education_2  Education_3    Family_1
##  Min.   :0     Min.   :0    Min.   :0    Min.   :1    Min.   :0    Min.   :0
##  1st Qu.:0     1st Qu.:0    1st Qu.:0    1st Qu.:1    1st Qu.:0    1st Qu.:0
##  Median :0     Median :0    Median :0    Median :1    Median :0    Median :0
##  Mean   :0     Mean   :0    Mean   :0    Mean   :1    Mean   :0    Mean   :0
##  3rd Qu.:0     3rd Qu.:0    3rd Qu.:0    3rd Qu.:1    3rd Qu.:0    3rd Qu.:0
##  Max.   :0     Max.   :0    Max.   :0    Max.   :1    Max.   :0    Max.   :0
##     Family_2    Family_3    Family_4 Securities.Account   CreditCard
##  Min.   :1   Min.   :0   Min.   :0   Min.   :0          Min.   :1
##  1st Qu.:1   1st Qu.:0   1st Qu.:0   1st Qu.:0          1st Qu.:1
##  Median :1   Median :0   Median :0   Median :0          Median :1
##  Mean   :1   Mean   :0   Mean   :0   Mean   :0          Mean   :1
##  3rd Qu.:1   3rd Qu.:0   3rd Qu.:0   3rd Qu.:0          3rd Qu.:1
##  Max.   :1   Max.   :0   Max.   :0   Max.   :0          Max.   :1
##      Online      Income           CCAvg            Mortgage
##  Min.   :1   Min.   :0.2221   Min.   :0.03551   Min.   :-0.5555
##  1st Qu.:1   1st Qu.:0.2221   1st Qu.:0.03551   1st Qu.:-0.5555
##  Median :1   Median :0.2221   Median :0.03551   Median :-0.5555
##  Mean   :1   Mean   :0.2221   Mean   :0.03551   Mean   :-0.5555
##  3rd Qu.:1   3rd Qu.:0.2221   3rd Qu.:0.03551   3rd Qu.:-0.5555
##  Max.   :1   Max.   :0.2221   Max.   :0.03551   Max.   :-0.5555
##       Age            Experience
##  Min.   :-0.4657   Min.   :-0.8811
##  1st Qu.:-0.4657   1st Qu.:-0.8811
##  Median :-0.4657   Median :-0.8811
##  Mean   :-0.4657   Mean   :-0.8811
##  3rd Qu.:-0.4657   3rd Qu.:-0.8811
##  Max.   :-0.4657   Max.   :-0.8811
```

```
New_Customer_1_norm
```

```
##   Personal.Loan CD.Account Education_1 Education_2 Education_3 Family_1
## 1             0          0           0           1           0        0
##   Family_2 Family_3 Family_4 Securities.Account CreditCard Online    Income
## 1        1        0        0                  0          1      1 0.2221371
##       CCAvg   Mortgage        Age Experience
## 1 0.0355115 -0.5554684 -0.4657003 -0.8811162
```

```
New_customer_loan<-factor(c("0","1"))
str(New_customer_loan)
```

```
##  Factor w/ 2 levels "0","1": 1 2
```

```
summary(New_customer_loan)
```

```
## 0 1
## 1 1
```

##run knn model k=1 for new customer evaluation

```
##k=1
c1<-Bank_Labels
New_customer_loan <- knn (Bank_Predictors, New_Customer_1_norm[2:17], c1, k=1, prob = TRUE )
New_customer_loan
```

```
## [1] 0
## attr(,"prob")
## [1] 1
## attr(,"nn.index")
##      [,1]
## [1,] 4035
## attr(,"nn.dist")
##           [,1]
## [1,] 0.4787594
## Levels: 0
```

```
print(New_customer_loan[1])
```

```
## [1] 0
## Levels: 0
```

The customer should not be a target for the marketing effort as they are unlikely to accept the loan offer

Choosing a k value of 1 allows for significant noise in the knn prediction algorithm because if will find only the one closest value and even if that value is surrounded by other values that lead to a different prediction, the prediction will come back matching the outcome of the closest data point

Choosing a larger value of k will eliminate the noise of one variable as the higher the k value the more searching for close values and the aggregate of those outcomes will be utilized for the predicted outcome

Confustion matrix to identify most accurate k for knn model

```
c1<-Train.Labels

accuracy.df<- data.frame(k=seq(1,20,1), accuracy = rep(0,20))

for (i in 1:20) {
  knn.pred <- knn(Train.Predictors, Test.Predictors,
```

```
              c1, k=i, prob=TRUE)
  accuracy.df[i,2]<- confusionMatrix(knn.pred,Test.Labels)$overall[1]
}
accuracy.df
```

```
##     k accuracy
## 1   1   0.9605
## 2   2   0.9550
## 3   3   0.9620
## 4   4   0.9550
## 5   5   0.9590
## 6   6   0.9500
## 7   7   0.9570
## 8   8   0.9475
## 9   9   0.9505
## 10 10   0.9455
## 11 11   0.9505
## 12 12   0.9445
## 13 13   0.9485
## 14 14   0.9435
## 15 15   0.9460
## 16 16   0.9445
## 17 17   0.9460
## 18 18   0.9425
## 19 19   0.9445
## 20 20   0.9410
```

The best k is 3 with an accuracy of 0.9620 from the confusionMatrix

```
##k=3
c1<-Train.Labels
Predicted.Test.Labels <- knn (Train.Predictors, Test.Predictors, c1, k=3, prob = TRUE )
CrossTable(x=Test.Labels,y=Predicted.Test.Labels)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  2000
##
##
##              | Predicted.Test.Labels
##  Test.Labels |         0 |         1 | Row Total |
```

```
## -------------|-----------|-----------|-----------|
##            0 |      1815 |         4 |      1819 |
##              |     5.685 |    94.929 |           |
##              |     0.998 |     0.002 |     0.909 |
##              |     0.962 |     0.035 |           |
##              |     0.907 |     0.002 |           |
## -------------|-----------|-----------|-----------|
##            1 |        72 |       109 |       181 |
##              |    57.129 |   954.012 |           |
##              |     0.398 |     0.602 |     0.090 |
##              |     0.038 |     0.965 |           |
##              |     0.036 |     0.054 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      1887 |       113 |      2000 |
##              |     0.944 |     0.056 |           |
## -------------|-----------|-----------|-----------|
##
##
```

The most accuracy for k value minimizes the false positive values, which is described as high precision with a score of 96%. However, the model has low recall (sensitivity) because it predicts 72 negative outcomes that are in fact positive outcomes. The recall % is only 60% for the model.

However, for a bank marketing campaign it would be wiser to use k=1 for knn modeling because it captures more true positives with only slightly higher false positives and the cost of sending marketing material vs. profitability of new loans would more than off set sending the additional marketing material to a few customers that ultimately turn down the loan. Comparing the results from k=1 to k=3, the bank would capture 8 more personal loan customers while sending material to 11 individuals that would ultimately not take the loans.

```
##k=1
c1<-Bank_Labels
New_customer_loan_a <- knn (Bank_Predictors, New_Customer_1_norm[2:17], c1, k=3, prob = TRUE)
New_customer_loan_a
```

```
## [1] 0
## attr(,"prob")
## [1] 1
## attr(,"nn.index")
##      [,1] [,2] [,3]
## [1,] 4035 4408 3399
## attr(,"nn.dist")
##           [,1]     [,2]      [,3]
## [1,] 0.4787594 0.496127 0.6343843
## Levels: 0
```

```
print(New_customer_loan_a[1])
```

```
## [1] 0
## Levels: 0
```

Re-running the knn with k=3 (most accurate model) returns the same result of do not market to this banking customer for a personal loan, as the model is more selective in terms of who it markets to than the k=1 knn model.

##partition data into training, validation, and test sets

```
set.seed(15)
Train_Index_1=createDataPartition(bank$Personal.Loan, p=0.5, list=FALSE)
Train_Data_1=bank[Train_Index_1,]
Remaining_Data_a=bank[-Train_Index_1,]

Train_Index=createDataPartition(Remaining_Data_a$Personal.Loan, p=0.6, list=FALSE)
Validation_Data_1=Remaining_Data_a[Train_Index,]
Test_Data_1=Remaining_Data_a[-Train_Index,]



summary(Train_Data_1$Personal.Loan)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1008  0.0000  1.0000
```

```
summary(Validation_Data_1$Personal.Loan)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.09333 0.00000 1.00000
```

```
summary(Test_Data_1$Personal.Loan)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.088   0.000   1.000
```

#normalize the data with Test Data

```
train.norm.1.df<-Train_Data_1
validation.norm.1.df<-Validation_Data_1

test_1_z_norm<-preProcess(Train_Data_1[ ,13:17],method = c("center","scale"))

train.norm.1.df[, 13:17]<-predict(test_1_z_norm, Train_Data_1[13:17])
validation.norm.1.df[, 13:17]<-predict(test_1_z_norm, Validation_Data_1[13:17])

summary(train.norm.1.df)
```

```
##  Personal.Loan      CD.Account      Education_1      Education_2
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1008   Mean   :0.0628   Mean   :0.4236   Mean   :0.2736
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##   Education_3        Family_1         Family_2         Family_3
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.3028   Mean   :0.2988   Mean   :0.2608   Mean   :0.2048
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     Family_4      Securities.Account   CreditCard         Online
##  Min.   :0.0000   Min.   :0.0000     Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000     Median :0.0000   Median :1.000
##  Mean   :0.2356   Mean   :0.1032     Mean   :0.2916   Mean   :0.596
##  3rd Qu.:0.0000   3rd Qu.:0.0000     3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :1.0000     Max.   :1.0000   Max.   :1.000
##      Income            CCAvg            Mortgage            Age
##  Min.   :-1.4425   Min.   :-1.0943   Min.   :-0.5595   Min.   :-1.93365
##  1st Qu.:-0.7705   1st Qu.:-0.7178   1st Qu.:-0.5595   1st Qu.:-0.88896
##  Median :-0.2503   Median :-0.2515   Median :-0.5595   Median :-0.01839
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.5733   3rd Qu.: 0.3666   3rd Qu.: 0.4272   3rd Qu.: 0.85219
##  Max.   : 3.2394   Max.   : 4.5245   Max.   : 5.4682   Max.   : 1.89687
##    Experience
##  Min.   :-2.0015517
##  1st Qu.:-0.8697084
##  Median : 0.0009403
##  Mean   : 0.0000000
##  3rd Qu.: 0.8715890
##  Max.   : 2.0034323
```

```
summary(validation.norm.1.df)
```

```
##  Personal.Loan       CD.Account      Education_1      Education_2
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.09333   Mean   :0.054   Mean   :0.3947   Mean   :0.2887
##  3rd Qu.:0.00000   3rd Qu.:0.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##   Education_3        Family_1         Family_2         Family_3
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000
##  Mean   :0.3167   Mean   :0.2853   Mean   :0.2567   Mean   :0.202
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
##     Family_4     Securities.Account   CreditCard       Online
##  Min.   :0.000   Min.   :0.000      Min.   :0.00   Min.   :0.000
```

```
##  1st Qu.:0.000   1st Qu.:0.000      1st Qu.:0.00   1st Qu.:0.000
##  Median :0.000   Median :0.000      Median :0.00   Median :1.000
##  Mean   :0.256   Mean   :0.108      Mean   :0.28   Mean   :0.588
##  3rd Qu.:1.000   3rd Qu.:0.000      3rd Qu.:1.00   3rd Qu.:1.000
##  Max.   :1.000   Max.   :1.000      Max.   :1.00   Max.   :1.000
##      Income            CCAvg            Mortgage            Age
##  Min.   :-1.44247   Min.   :-1.09429   Min.   :-0.55951   Min.   :-1.93365
##  1st Qu.:-0.79221   1st Qu.:-0.70098   1st Qu.:-0.55951   1st Qu.:-0.80190
##  Median :-0.25032   Median :-0.25148   Median :-0.55951   Median : 0.06867
##  Mean   :-0.04259   Mean   :-0.02123   Mean   :-0.02074   Mean   : 0.04029
##  3rd Qu.: 0.45954   3rd Qu.: 0.31040   3rd Qu.: 0.40766   3rd Qu.: 0.85219
##  Max.   : 2.82757   Max.   : 4.52446   Max.   : 5.64407   Max.   : 1.89687
##    Experience
##  Min.   :-2.00155
##  1st Qu.:-0.78264
##  Median : 0.08801
##  Mean   : 0.03629
##  3rd Qu.: 0.87159
##  Max.   : 1.91637
```

##knn model with Training and Validation data (k=3)

```
##k=3
Train.Labels.1<-Train_Data_1[,1]
Validation.Labels.1<-Validation_Data_1[,1]
Train.Predictors.1<-Train_Data_1[,2:17]
Validation.Predictors.1<-Validation_Data_1[,2:17]
c1<-Train.Labels.1
Predicted.Test.Labels.1 <- knn (Train.Predictors.1, Validation.Predictors.1, c1, k=3, prob = TRUE )
CrossTable(x=Validation.Labels.1,y=Predicted.Test.Labels.1)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1500
##
##
##                      | Predicted.Test.Labels.1
## Validation.Labels.1 |         0 |         1 | Row Total |
## --------------------|-----------|-----------|-----------|
##                   0 |      1285 |        75 |      1360 |
##                     |     1.815 |    18.356 |           |
##                     |     0.945 |     0.055 |     0.907 |
##                     |     0.941 |     0.556 |           |
```

```
##                      |     0.857 |     0.050 |           |
## --------------------|-----------|-----------|-----------|
##                   1 |        80 |        60 |       140 |
##                     |    17.635 |   178.314 |           |
##                     |     0.571 |     0.429 |     0.093 |
##                     |     0.059 |     0.444 |           |
##                     |     0.053 |     0.040 |           |
## --------------------|-----------|-----------|-----------|
##        Column Total |      1365 |       135 |      1500 |
##                     |     0.910 |     0.090 |           |
## --------------------|-----------|-----------|-----------|
##
##
```

##Combine training and Validation Data

```
Train.Validation.Data<-rbind(Train_Data_1, Validation_Data_1)
summary(Train.Validation.Data)
```

```
##  Personal.Loan     CD.Account      Education_1      Education_2
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.098   Mean   :0.0595   Mean   :0.4128   Mean   :0.2792
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##   Education_3       Family_1         Family_2         Family_3
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.308   Mean   :0.2938   Mean   :0.2592   Mean   :0.2037
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     Family_4      Securities.Account   CreditCard         Online
##  Min.   :0.0000   Min.   :0.000      Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.000      1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.000      Median :0.0000   Median :1.000
##  Mean   :0.2432   Mean   :0.105      Mean   :0.2873   Mean   :0.593
##  3rd Qu.:0.0000   3rd Qu.:0.000      3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :1.000      Max.   :1.0000   Max.   :1.000
##      Income           CCAvg            Mortgage           Age
##  Min.   :  8.00   Min.   : 0.000   Min.   :  0.00   Min.   :23.00
##  1st Qu.: 39.00   1st Qu.: 0.700   1st Qu.:  0.00   1st Qu.:35.00
##  Median : 63.00   Median : 1.500   Median :  0.00   Median :45.00
##  Mean   : 73.81   Mean   : 1.933   Mean   : 56.48   Mean   :45.38
##  3rd Qu.: 99.00   3rd Qu.: 2.500   3rd Qu.:101.00   3rd Qu.:55.00
##  Max.   :224.00   Max.   :10.000   Max.   :635.00   Max.   :67.00
##    Experience
##  Min.   :-3.00
##  1st Qu.:10.00
##  Median :20.00
##  Mean   :20.15
##  3rd Qu.:30.00
##  Max.   :43.00
```

#re-normalize the data based on Training and Validation data

```
train.validation.norm.df<-Train.Validation.Data
test.norm.1.df<-Test_Data_1

test_validation_z_norm<-preProcess(Train.Validation.Data[ ,13:17],method = c("center","scale"))

train.validation.norm.df[, 13:17]<-predict(test_validation_z_norm, Train.Validation.Data[13:17])
test.norm.1.df[, 13:17]<-predict(test_z_norm, Test_Data_1[13:17])

summary(train.validation.norm.df)
```

```
##  Personal.Loan     CD.Account      Education_1      Education_2
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.098   Mean   :0.0595   Mean   :0.4128   Mean   :0.2792
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##   Education_3       Family_1         Family_2         Family_3
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.308   Mean   :0.2938   Mean   :0.2592   Mean   :0.2037
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##     Family_4      Securities.Account   CreditCard         Online
##  Min.   :0.0000   Min.   :0.000      Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.000      1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :0.000      Median :0.0000   Median :1.000
##  Mean   :0.2432   Mean   :0.105      Mean   :0.2873   Mean   :0.593
##  3rd Qu.:0.0000   3rd Qu.:0.000      3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :1.000      Max.   :1.0000   Max.   :1.000
##      Income            CCAvg            Mortgage            Age
##  Min.   :-1.4266   Min.   :-1.1041   Min.   :-0.5534   Min.   :-1.95649
##  1st Qu.:-0.7546   1st Qu.:-0.7043   1st Qu.:-0.5534   1st Qu.:-0.90766
##  Median :-0.2344   Median :-0.2475   Median :-0.5534   Median :-0.03363
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.5460   3rd Qu.: 0.3236   3rd Qu.: 0.4363   3rd Qu.: 0.84040
##  Max.   : 3.2555   Max.   : 4.6064   Max.   : 5.6686   Max.   : 1.88923
##    Experience
##  Min.   :-2.02192
##  1st Qu.:-0.88628
##  Median :-0.01271
##  Mean   : 0.00000
##  3rd Qu.: 0.86086
##  Max.   : 1.99650
```

```
summary(test.norm.1.df)
```

```
##  Personal.Loan     CD.Account      Education_1      Education_2
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
```

```
##  Median :0.000   Median :0.000   Median :0.000   Median :0.000
##  Mean   :0.088   Mean   :0.064   Mean   :0.445   Mean   :0.286
##  3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000
##   Education_3       Family_1        Family_2        Family_3
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
##  Median :0.000   Median :0.000   Median :0.000   Median :0.000
##  Mean   :0.269   Mean   :0.297   Mean   :0.259   Mean   :0.195
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.000
##  Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000
##     Family_4     Securities.Account   CreditCard        Online
##  Min.   :0.000   Min.   :0.000      Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000      1st Qu.:0.000   1st Qu.:0.000
##  Median :0.000   Median :0.000      Median :0.000   Median :1.000
##  Mean   :0.249   Mean   :0.102      Mean   :0.321   Mean   :0.612
##  3rd Qu.:0.000   3rd Qu.:0.000      3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :1.000   Max.   :1.000      Max.   :1.000   Max.   :1.000
##      Income             CCAvg              Mortgage             Age
##  Min.   :-1.41908   Min.   :-1.105269   Min.   :-0.552405   Min.   :-1.93623
##  1st Qu.:-0.77569   1st Qu.:-0.710127   1st Qu.:-0.552405   1st Qu.:-0.89391
##  Median :-0.19664   Median :-0.258537   Median :-0.552405   Median :-0.02531
##  Mean   :-0.01171   Mean   :-0.001091   Mean   : 0.001818   Mean   :-0.01202
##  3rd Qu.: 0.42530   3rd Qu.: 0.362399   3rd Qu.: 0.426944   3rd Qu.: 0.84330
##  Max.   : 3.08464   Max.   : 4.539609   Max.   : 5.333483   Max.   : 1.88562
##    Experience
##  Min.   :-2.001532
##  1st Qu.:-0.872359
##  Median :-0.003764
##  Mean   :-0.008889
##  3rd Qu.: 0.864831
##  Max.   : 1.994004
```

##knn model with Training/ Validation and Test data

```
##k=3
Train.Validation.Labels.1<-Train.Validation.Data[,1]
Test.Labels.2<-Test_Data_1[,1]
Train.Validation.Predictors.1<-train.validation.norm.df[,2:17]
Test.Predictors.1<-test.norm.1.df[,2:17]
c1<-Train.Validation.Labels.1
Predicted.Test.Labels.1 <- knn (Train.Validation.Predictors.1, Test.Predictors.1, c1, k=3, prob = TRUE )
CrossTable(x=Test.Labels.2,y=Predicted.Test.Labels.1)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
```

```
## |-------------------------|
##
##
## Total Observations in Table:  1000
##
##
##                | Predicted.Test.Labels.1
## Test.Labels.2 |         0 |         1 | Row Total |
## --------------|-----------|-----------|-----------|
##             0 |       912 |         0 |       912 |
##               |     2.500 |    46.512 |           |
##               |     1.000 |     0.000 |     0.912 |
##               |     0.961 |     0.000 |           |
##               |     0.912 |     0.000 |           |
## --------------|-----------|-----------|-----------|
##             1 |        37 |        51 |        88 |
##               |    25.905 |   482.033 |           |
##               |     0.420 |     0.580 |     0.088 |
##               |     0.039 |     1.000 |           |
##               |     0.037 |     0.051 |           |
## --------------|-----------|-----------|-----------|
##  Column Total |       949 |        51 |      1000 |
##               |     0.949 |     0.051 |           |
## --------------|-----------|-----------|-----------|
##
##
```

Comparing confusion matrix of test set with training and validation set: The Test set is more accurate overall than the training set, but predicts less true positive results than the training set (51 vs. 60). In the business world a small number of false positives (Precision) on a marketing campaign is a far better outcome than a high number of false negatives (Recall or Sensitivity) as the cost of marketing campaigns (often by mail or the internet) are far less costly than the value captured by gaining a new customer transaction, in this case a personal loan. The incremental gain of the 9 new customers (represented by the difference in confirmed positives) will likely outweigh the cost of the 75 customers that receive marketing material but do not accept the personal loans.

##As the k for the knn model increases it becomes more selective with the positive outcomes because the negative outcomes of the overall population is ~90%. Therefore, as more neighbors are utilized to make the determination the likelyhood of a negative outcome becomes more likely and eventually a high enough k value would lead to all negative outcomes as most of the results are negative within the entire data set.