

Exploratory Analysis of Credit loan dataset

Reading the csv file

```
set.seed(2345)

loan_data <- read.csv("Residential Mortgages.csv")

summary(loan_data)

##          i..ContractRef      LoanBalance      InterestType
## 00000CONTR00000001:     1  55,602 : 517  Fixed   : 45349
## 00000CONTR00000002:     1  1,18,155: 502  Tracker :118744
## 00000CONTR00000003:     1  60,437 : 500  Variable: 73296
## 00000CONTR00000004:     1  1,20,723: 497
## 00000CONTR00000005:     1  1,25,860: 497
## 00000CONTR00000006:     1  1,15,838: 490
## (Other)           :237383 (Other) :234386
## InterestRate NIM                  MortgageType    NewLoan
## 0%:16106     0%: 25639  Buy to Let     : 52085  No :225779
## 1%:32718     1%: 48679  Owner Occupied:185304 Yes: 11610
## 2%:32427     2%: 48813
## 3%:32616     3%:104217
## 4%:38358     4%: 10041
## 5%:82199
## 6%: 2965
## ProbationaryLoans DefaultedLoans CreditRating InterestIncome
## No :228852      No :226595    Min.   :1.000   0       : 3003
## Yes: 8537      Yes: 10794   1st Qu.:1.000   592     : 168
##                               Median :2.000   623     : 165
##                               Mean   :2.108   302     : 160
##                               3rd Qu.:3.000   324     : 160
##                               Max.   :5.000   453     : 157
##                               (Other):233576
##          LTV          LTVCategory      MortgageYears   PropertyValue
## 92%   : 10925    70 to 80%:67253  Min.   :28.00  1,02,717: 408
## 94%   :  9922    90 to 100%:48503  1st Qu.:30.00  1,09,137: 404
## 93%   :  9560    60 to 70%:48483  Median :32.00  1,05,927: 383
## 91%   :  9417    30 to 40%:17628  Mean   :31.51  98,437 : 382
## 73%   :  7913    10 to 20%:16112  3rd Qu.:34.00  1,04,857: 376
## 72%   :  7570    20 to 30%: 9994  Max.   :35.00  1,03,787: 365
## (Other):182082 (Other)   :29416   (Other) :235071
##          AnnualPYMT      MaturityDate      BookingDate
## 3,210  :  268  26-06-2041: 71  15-01-2007: 96
## 6,420  :  188  04-11-2041: 69  05-08-2008: 95
## 2,140  :  167  11-12-2040: 67  19-07-2006: 94
## 4,280  :  151  31-05-2039: 67  27-02-2008: 94
## 2,996  :  144  31-08-2040: 67  04-08-2008: 92
## 3,668  :  144  06-09-2041: 66  17-01-2008: 92
## (Other):236327 (Other)   :236982 (Other)   :236826
##          LastValuationDate ValuationAgeYears CreditRatingMovement
```

```

## 24-12-2016: 263 Min. :0.010      Min. :-4.0000
## 09-11-2016: 261 1st Qu.:1.110    1st Qu.: 0.0000
## 20-12-2016: 260 Median :2.410    Median : 0.0000
## 18-06-2016: 258 Mean  :2.524     Mean  :-0.1806
## 10-10-2016: 255 3rd Qu.:3.880    3rd Qu.: 0.0000
## 11-10-2016: 255 Max.  :5.510     Max.  : 4.0000
## (Other)   :235837

##          County           Branch
## Dublin   :77892  Branch 234: 1610
## Cork     :26133  Branch 77 : 1591
## Galway   :12870  Branch 21 : 1582
## Kildare  :10806  Branch 230: 1581
## Limerick : 8698  Branch 27 : 1577
## Meath    : 8282  Branch 62 : 1577
## (Other)  :92708  (Other)  :227871

##                                         Address
## 21 Fforster Dr, Ballyowen, Lucan, Co. Dublin, Ireland      : 2
## 21 Forge Ave, Racecourse Common, Lusk, Co. Dublin, Ireland : 2
## 21 The Alders, Poulacurra South, Co. Cork, Ireland        : 2
## 22 Christendom Mews, Christendom, Waterford, Co. Kilkenny, Ireland: 2
## 23 Fforster Dr, Ballyowen, Lucan, Co. Dublin, Ireland      : 2
## 23 Ossory Rd, North Dock, Dublin, Ireland                 : 2
## (Other)          :237377

##          Town       AddressLatitude AddressLongitude InArrears
## Dublin    : 5022  Min.   :51.43   Min.   :-10.464  No :212002
## Galway    : 3380  1st Qu.:52.67   1st Qu.: -8.431  Yes: 25387
## Lucan     : 2524  Median :53.29   Median : -6.933
## Swords    : 2280  Mean    :53.12   Mean    : -7.328
## Drogheda  : 2267  3rd Qu.:53.42   3rd Qu.: -6.288
## Blackrock : 2179  Max.    :55.38   Max.    : -6.015
## (Other)   :219737

##          ValueInArrears DaysInArrears ArrearsCategory
## Min.   : 0.0   Min.   : 0.00 < 1 year   : 62
## 1st Qu.: 0.0   1st Qu.: 0.00 < 6 months : 5974
## Median : 0.0   Median : 0.00 < 90 days  :19263
## Mean   :113.6   Mean   : 8.38 >2 years   : 17
## 3rd Qu.: 0.0   3rd Qu.: 0.00 1 to 2 years:  71
## Max.   :22163.6  Max.   :798.00 N/A       :212002
## 

##          X_HousePriceMovement UpdatedPropertyValue LTVUpdated
## 0%      :121395          0       : 694    92%      : 6287
## 13%     : 4509          1,02,717: 201    73%      : 6242
## 15%     : 4263          1,09,137: 191    72%      : 5953
## 3%      : 3945          1,05,927: 182    77%      : 5864
## 19%     : 3332          98,437  : 179    75%      : 5781
## 8%      : 3316          1,04,857: 177    94%      : 5647
## (Other): 96629         (Other) :235765  (Other):201615

##          LTVCategoryUpdated
## 70 to 80% :53030
## 60 to 70% :42201
## 90 to 100%:30373
## 50 to 60% :18111
## 30 to 40% :18108
## 10 to 20% :16807

```

```
##  (Other) :58759
```

Check for missing and unknown values

```
library(caret)

## Warning: package 'caret' was built under R version 3.3.3
## Loading required package: lattice
## Loading required package: ggplot2
## Warning: package 'ggplot2' was built under R version 3.3.3
apply(loan_data, 2, function(x) sum(is.na(x)==TRUE))
```

```
##      i..ContractRef          LoanBalance        InterestType
##                  0                  0                      0
##      InterestRate                 NIM        MortgageType
##                  0                  0                      0
##      NewLoan     ProbationaryLoans    DefaultedLoans
##                  0                  0                      0
##      CreditRating       InterestIncome          LTV
##                  0                  0                      0
##      LTVCategory      MortgageYears      PropertyValue
##                  0                  0                      0
##      AnnualPYMT        MaturityDate      BookingDate
##                  0                  0                      0
##      LastValuationDate   ValuationAgeYears CreditRatingMovement
##                  0                  0                      0
##      County             Branch            Address
##                  0                  0                      0
##      Town           AddressLatitude AddressLongitude
##                  0                  0                      0
##      InArrears        ValueInArrears    DaysInArrears
##                  0                  0                      0
##      ArrearsCategory X._HousePriceMovement UpdatedPropertyValue
##                  0                  0                      0
##      LTVUpdated      LTVCategoryUpdated
##                  0                  0
```

CrossTable data

```
library(gmodels)

## Warning: package 'gmodels' was built under R version 3.3.3
print("-----")
## [1] "-----"
print("          Cross Table of Defaulted Loans")
## [1] "          Cross Table of Defaulted Loans"
```

```

print("-----")

## [1] "-----"
CrossTable(loan_data$DefaultedLoans)

##
##
##      Cell Contents
## |-----|-----|
## |           N |
## |   N / Table Total |-----|
## |-----|-----|
## 
## 
## Total Observations in Table:  237389
## 
## 
## 
##          |      No |      Yes |
## |-----|-----|
## |    226595 |    10794 |
## |    0.955  |    0.045 |
## |-----|-----|
## 
## 
## 
## 
print("-----")

## [1] "-----"
print("      Cross Table of Defaulted Loans and Credit Rating")
## [1] "      Cross Table of Defaulted Loans and Credit Rating"
print("-----")

## [1] "-----"
CrossTable(loan_data$CreditRating, loan_data$DefaultedLoans, prop.r = TRUE, prop.c = FALSE, prop.t = FALSE)

##
##
##      Cell Contents
## |-----|-----|
## |           N |
## |   N / Row Total |-----|
## |-----|-----|
## 
## 
## Total Observations in Table:  237389
## 
## 
##          | loan_data$DefaultedLoans
## loan_data$CreditRating |      No |      Yes | Row Total |
## -----|-----|-----|-----|

```

```

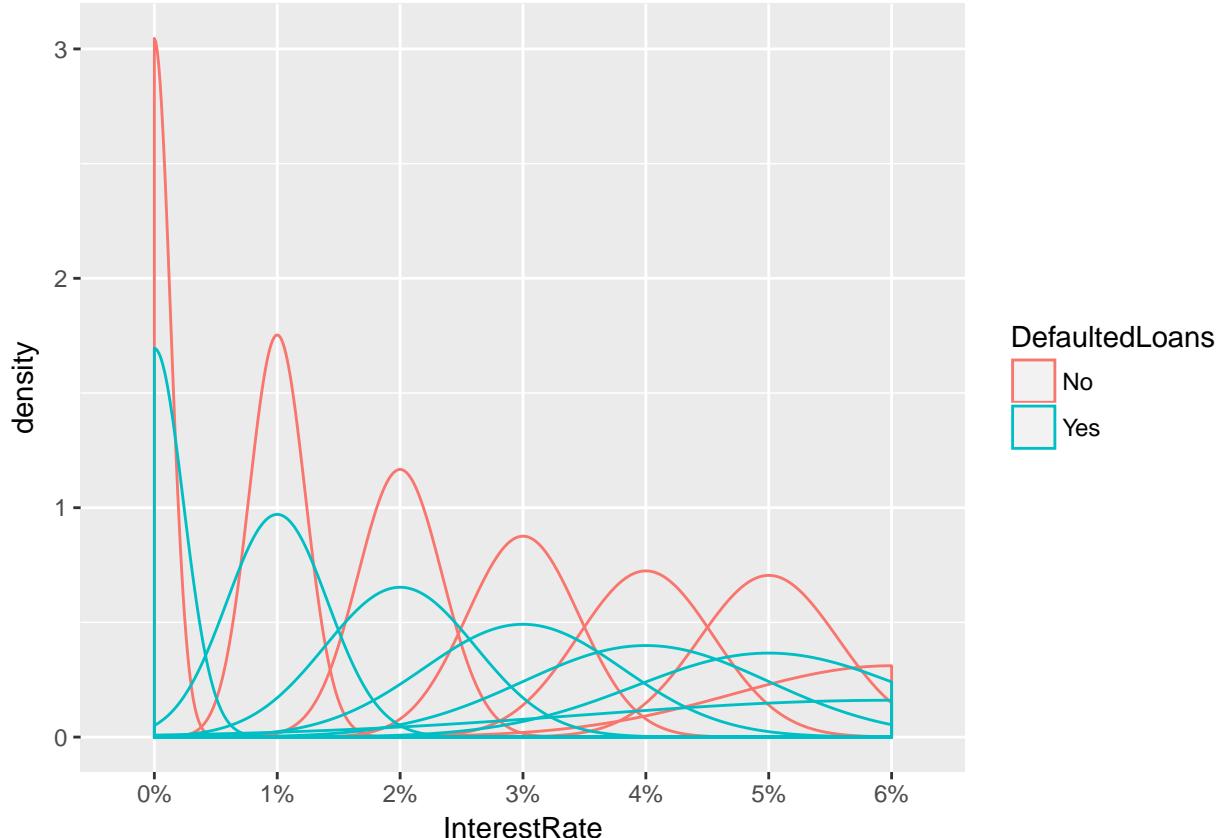
##          1 | 106525 |      0 | 106525 |
##          | 1.000 | 0.000 | 0.449 |
## -----
##          2 | 50294 |    14 | 50308 |
##          | 1.000 | 0.000 | 0.212 |
## -----
##          3 | 41748 |    12 | 41760 |
##          | 1.000 | 0.000 | 0.176 |
## -----
##          4 | 25917 |      0 | 25917 |
##          | 1.000 | 0.000 | 0.109 |
## -----
##          5 | 2111 | 10768 | 12879 |
##          | 0.164 | 0.836 | 0.054 |
## -----
##          Column Total | 226595 | 10794 | 237389 |
## -----
##
```

Plots to explore data

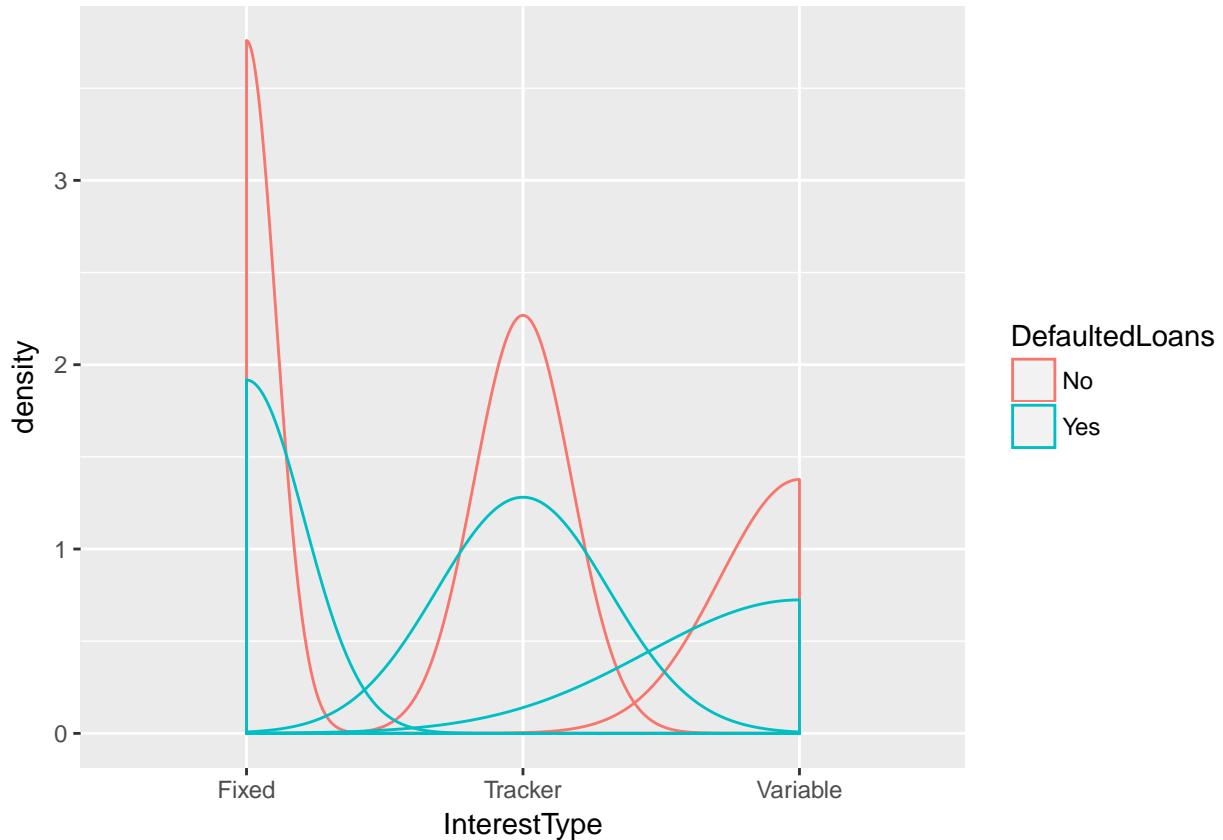
```

library(ggplot2)
qplot(InterestRate, colour=DefaultedLoans, data = loan_data, geom = "density")

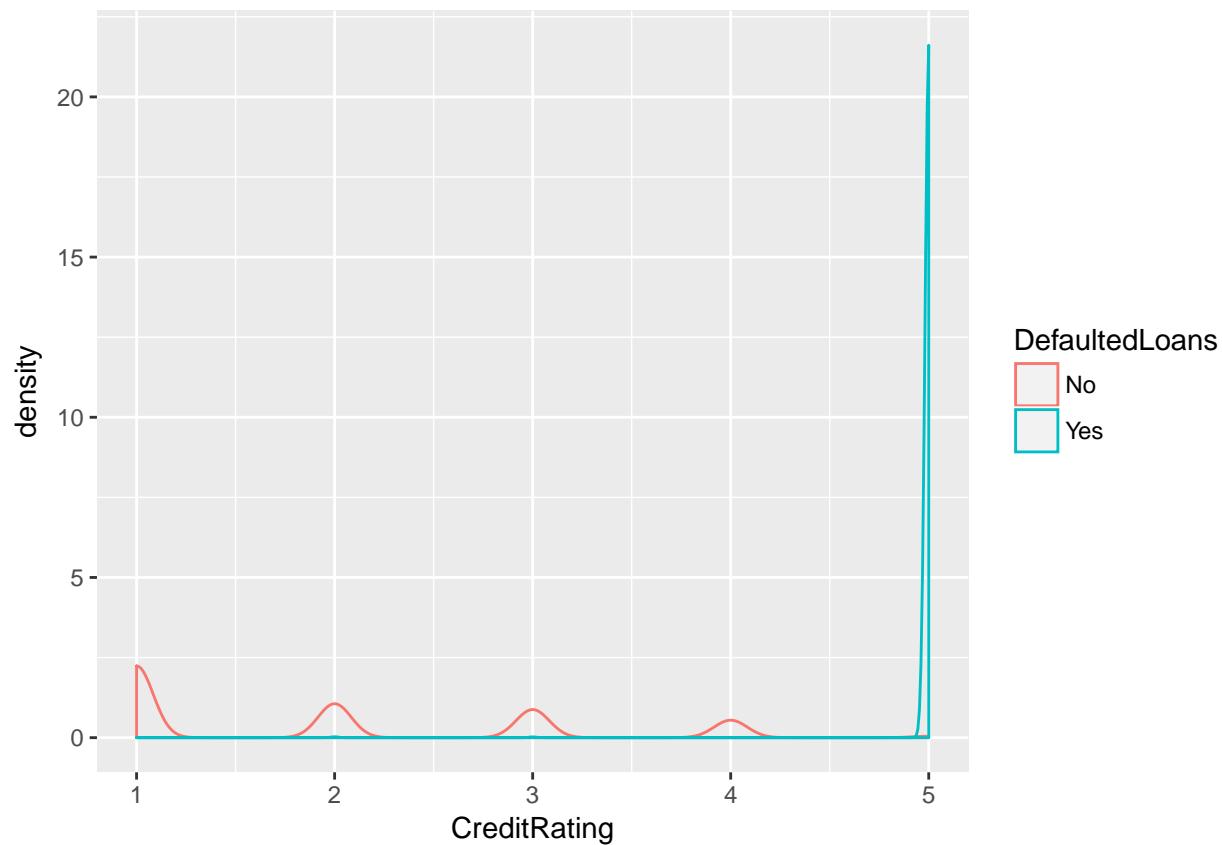
```



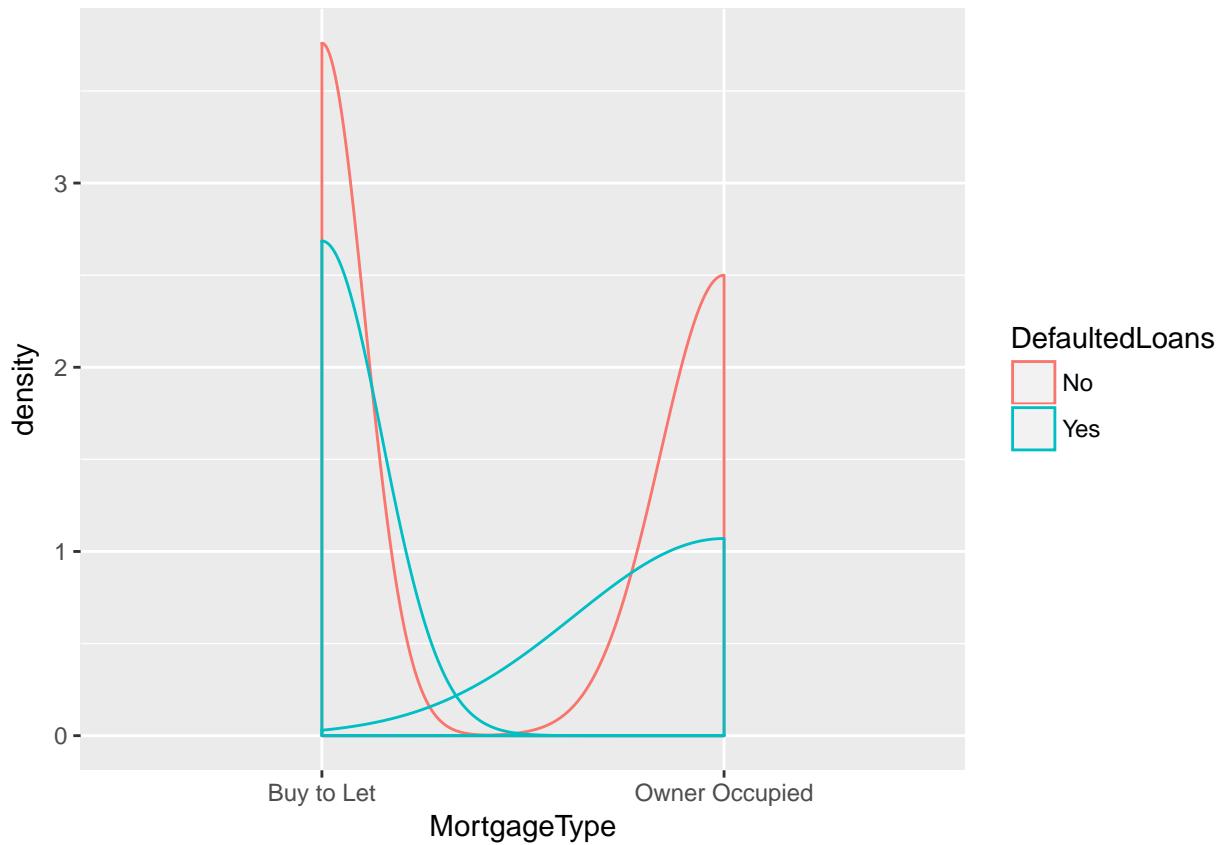
```
qplot(InterestType, colour=DefaultedLoans, data = loan_data, geom = "density")
```

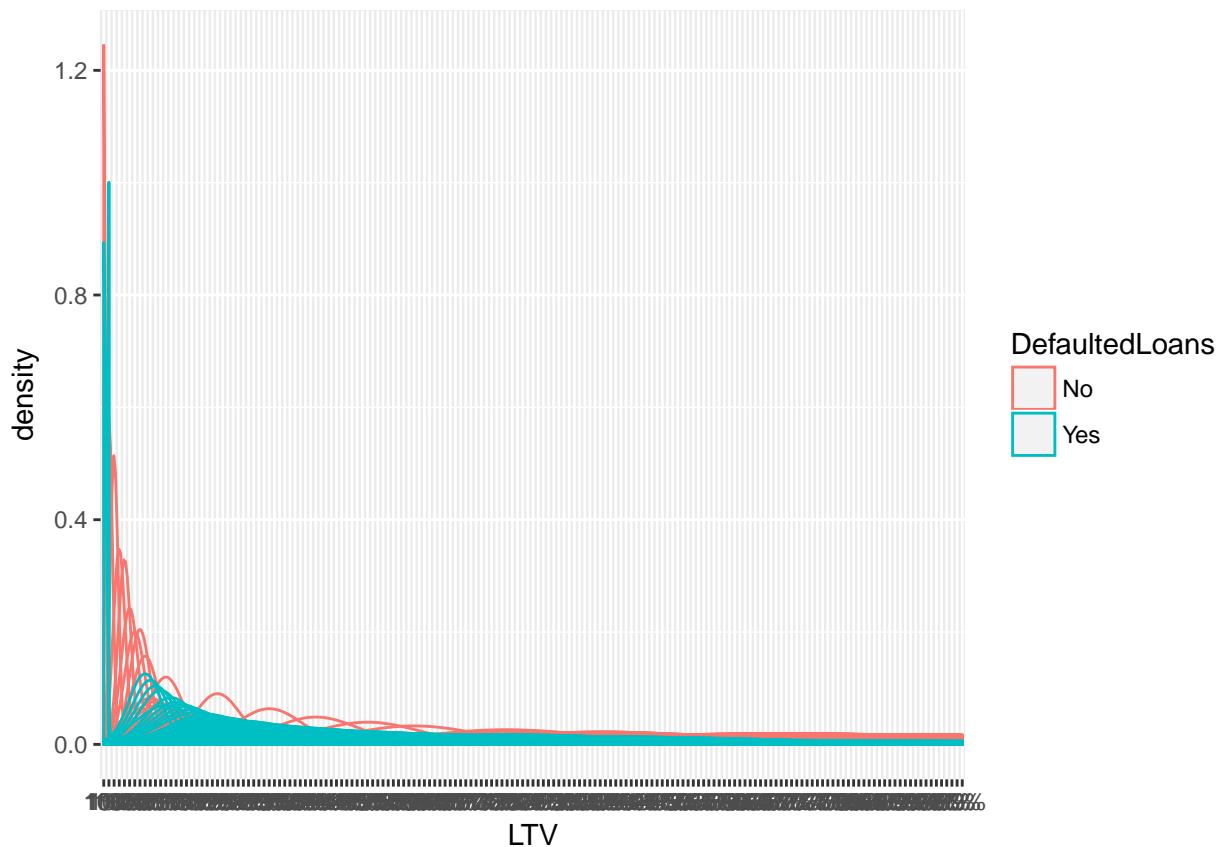


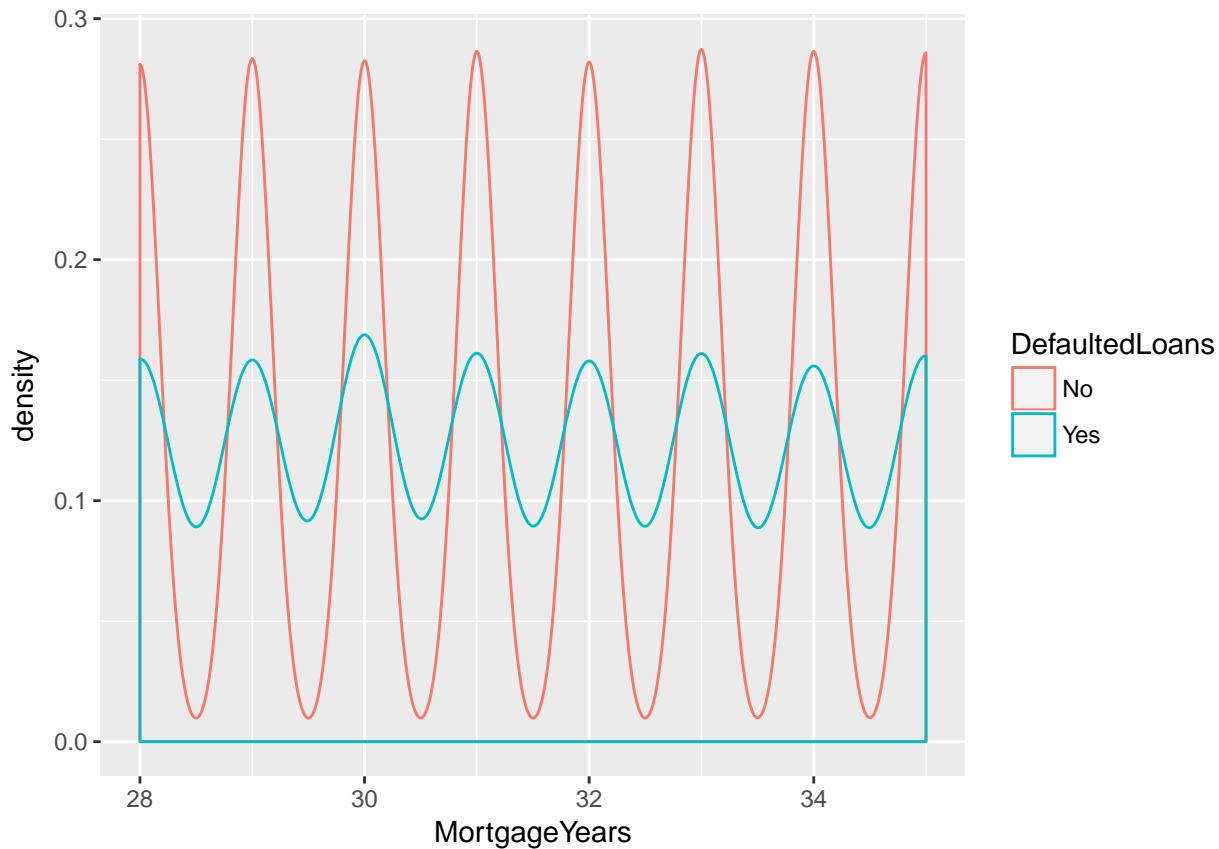
```
qplot(CreditRating, colour=DefaultedLoans, data = loan_data, geom = "density")
```



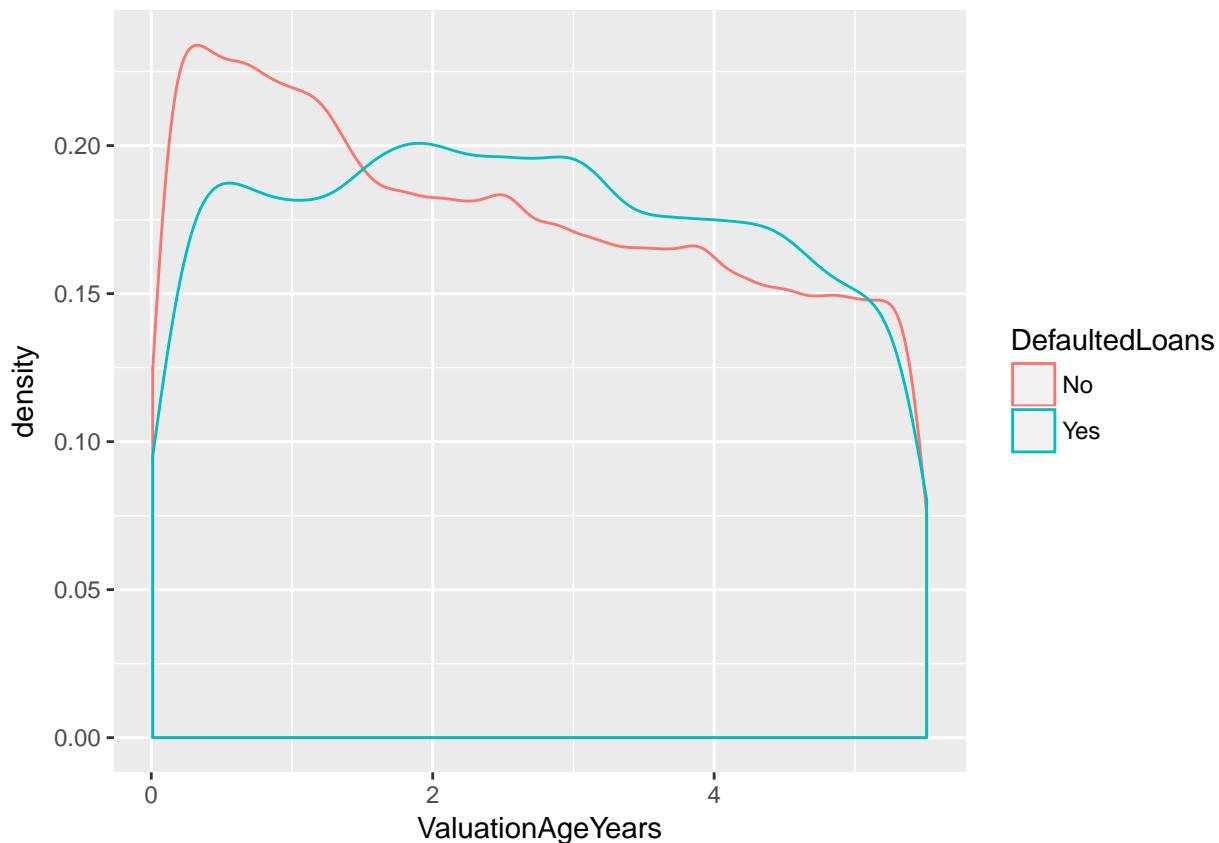
```
qplot(MortgageType, colour=DefaultedLoans, data = loan_data, geom = "density")
```







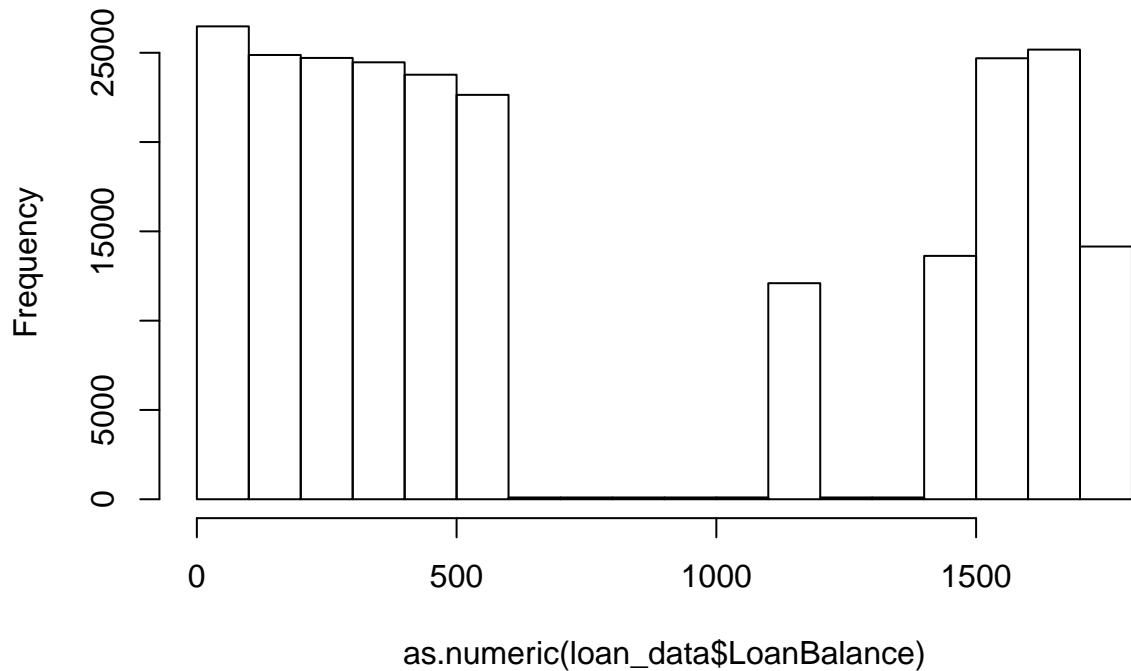
```
qplot(ValuationAgeYears, colour=DefaultedLoans, data = loan_data, geom = "density")
```



Spotting outliers

```
hist_1 <- hist(as.numeric(loan_data$LoanBalance))
```

Histogram of as.numeric(loan_data\$LoanBalance)

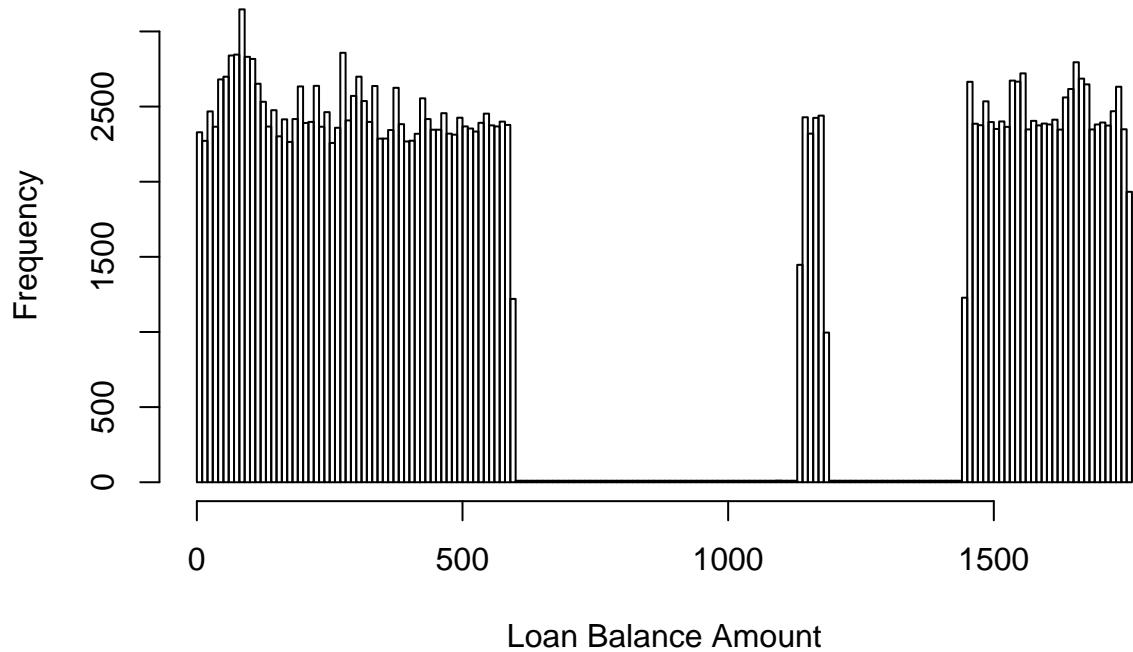


```
hist_1$breaks
```

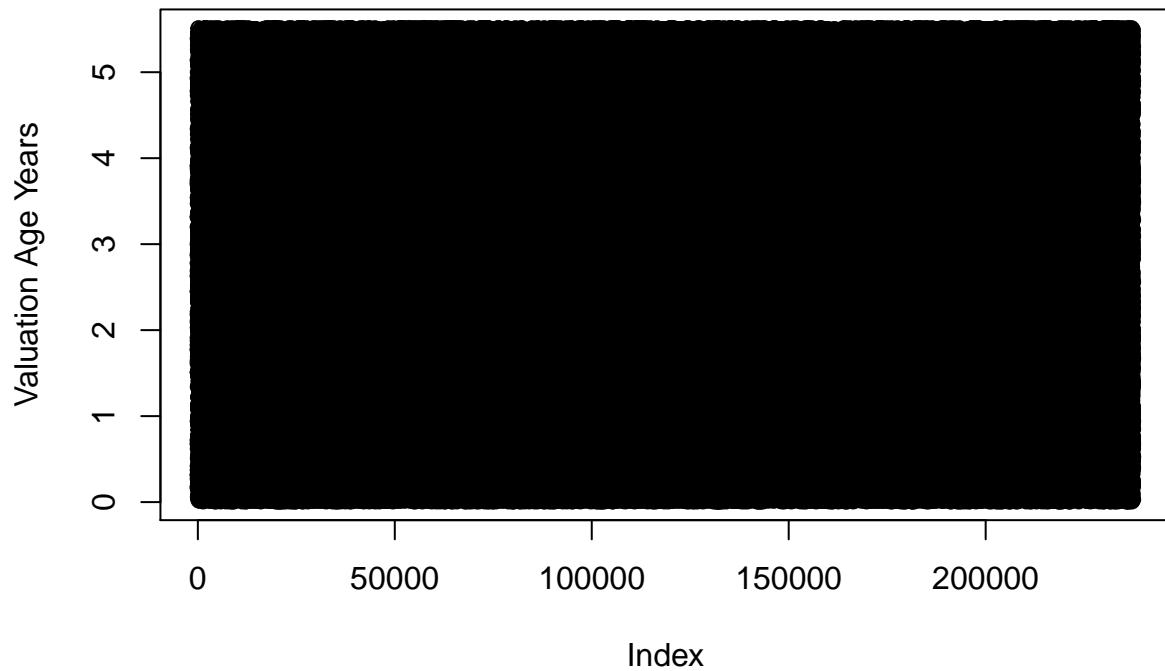
```
## [1] 0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300  
## [15] 1400 1500 1600 1700 1800
```

```
hist_2 <- hist(as.numeric(loan_data$LoanBalance), breaks = 200, xlab = "Loan Balance Amount",  
               main = "Histogram of the loan balance amount")
```

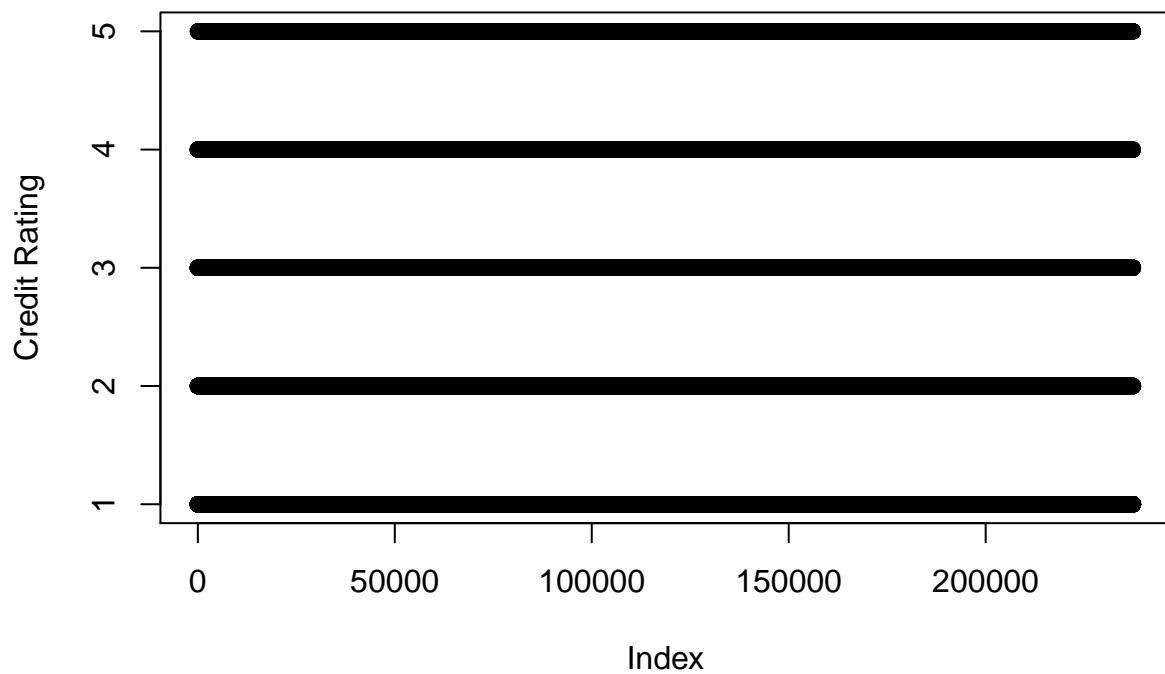
Histogram of the loan balance amount



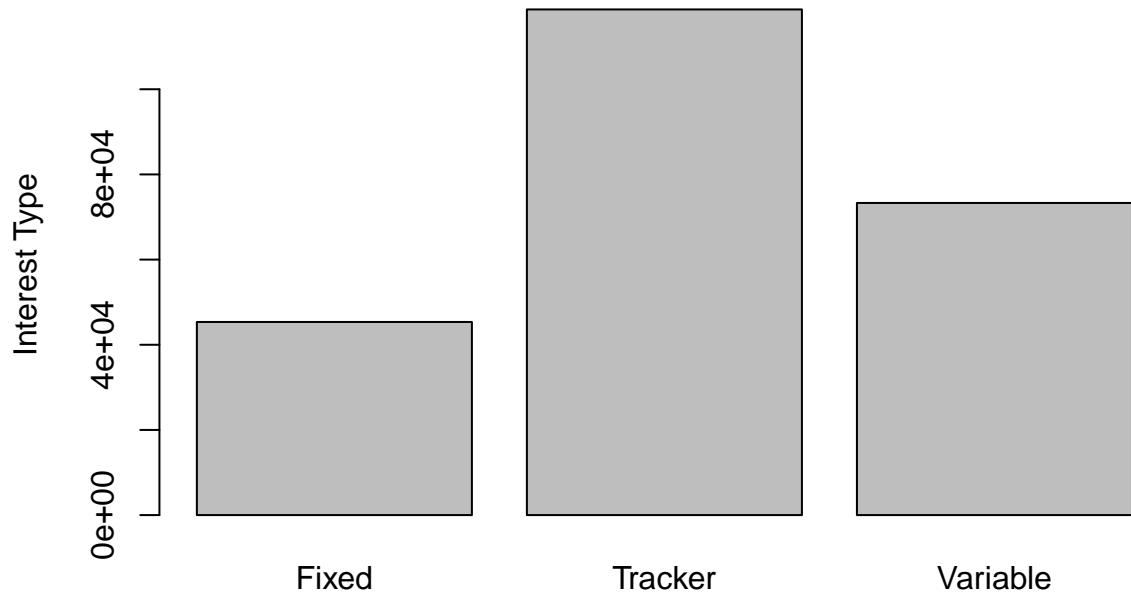
```
#Plot outliers  
plot(loan_data$ValuationAgeYears, ylab = "Valuation Age Years")
```



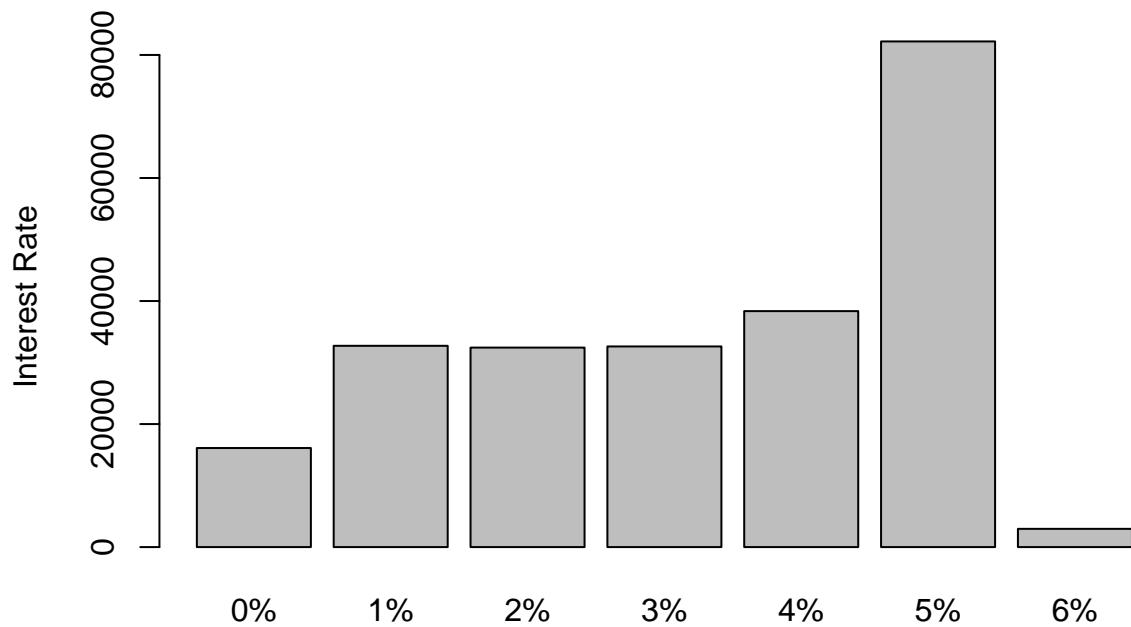
```
plot(loan_data$CreditRating, ylab = "Credit Rating")
```



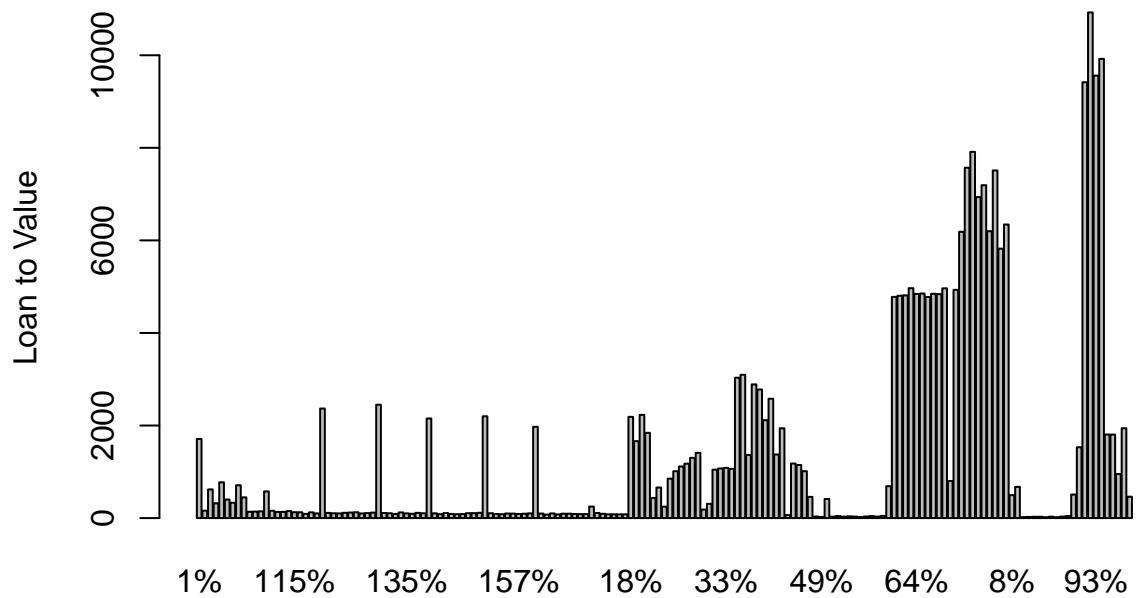
```
plot(loan_data$InterestType, ylab = "Interest Type")
```



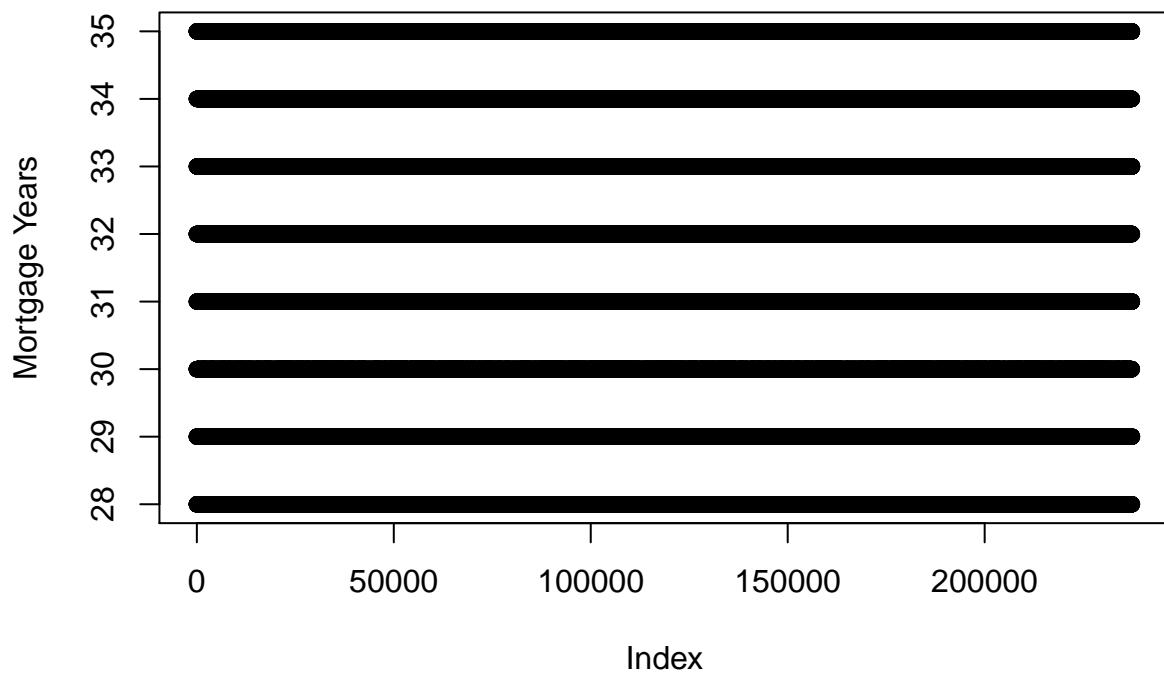
```
plot(loan_data$InterestRate, ylab = "Interest Rate")
```



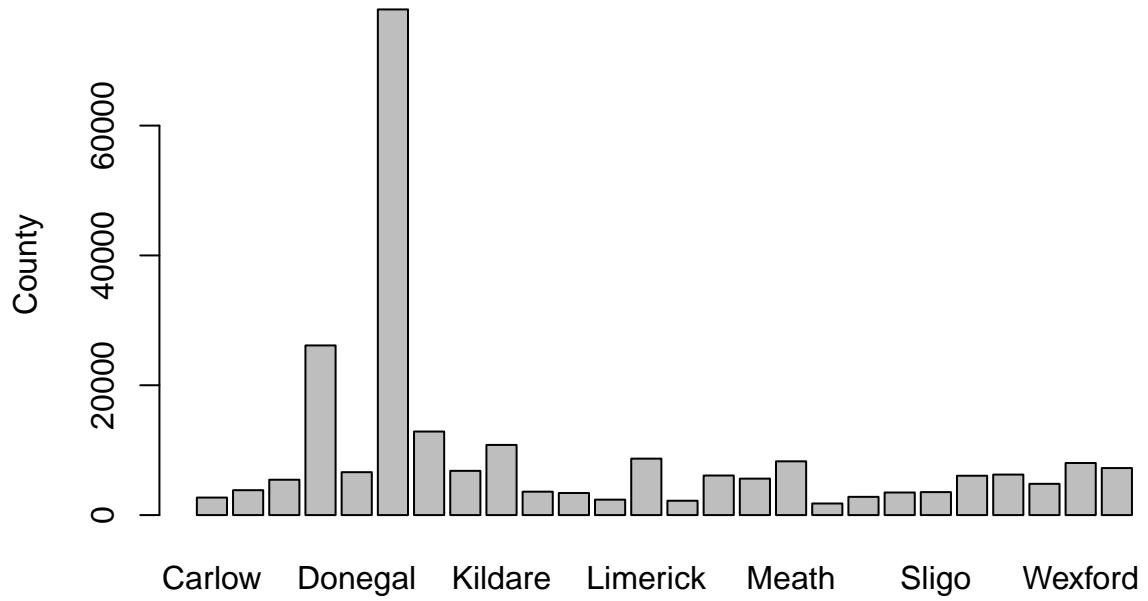
```
plot(loan_data$LTV, ylab = "Loan to Value")
```



```
plot(loan_data$MortgageYears, ylab = "Mortgage Years")
```



```
plot(loan_data$County, ylab = "County")
```



```
plot(loan_data$InArrears, ylab = "In Arrears")
```

