prosperLoanData

Ubirajara Theodoro Schier

17 de julho de 2017

1. REQUISITOS DO PROJETO

1.1. Uma análise passo a passo da análise e exploração dos dados.

a. Cabeçalhos e texto devem organizar seus pensamentos e refletir sua análise conforme você explora os dados.

b. Gráficos na sua análise não precisam ser rebuscados com rótulos, unidades e títulos; estes gráficos são exploratório (rápidos e simples). Eles devem, contudo, ser dos tipos apropriados e efetivamente representar a informação que você deseja obter a partir deles.

c. Você pode iterar sobre um gráfico no mesmo código R, mas você não precisa mostrar todas as iterações na sua análise.

1.2. Uma seção no final chamada “Gráficos Finais e Sumário”

Você irá escolher três gráficos da sua análise para polir e compartilhar nesta seção. Os três gráficos deverão demonstrar diferentes tendências e devem ser detalhados com os rótulos apropriados, unidades e títulos.

1.3. Uma seção final chamada “Reflexão”

Esta deve conter algumas frases sobre suas dificuldades, sucessos, e idéias para explorações futures neste mesmo conjunto de dados.

2. PREPARAÇÃO DO AMBIENTE E DO CONJUNTO DE DADOS

2.1. Carga, Limpeza e Transformações do conjunto de dados

```
## [1] "> Ações realizadas nesta etapa:"
```

```
## [1] "> carga do conjunto de dados prosperLoanData.csv"
```

```
## [1] "> remoção de registros duplicados"
```

```
## # A tibble: 827 x 2
##                    LoanKey     n
##                     <fctr> <int>
##  1 001D3702020659484457 65E     2
##  2 009C3695912584260075 7B0     2
##  3 0166370410175471 5CF7C56     2
##  4 0176370480684030 15E470D     2
##  5 01C83707282210922790778     2
##  6 0216370080923136 5A56A1C     2
##  7 0224369791751850396C2F3     2
##  8 0275370118102352450205B     2
##  9 0298370462888699448 2743     2
## 10 02D137079845719197ACFE8     2
## # ... with 817 more rows
```

```
## [1] "> reordenação do campo ProsperRating..Alpha."
```

```
## [1] "> reordenação do campo IncomeRange"
```

```
## [1] "> reordenação do campo LoanOriginationQuarter"
```

```
## [1] "> conversão do campo ListingCreationDate para o formato DateTime"
```

```
## [1] "> conversão do campo ListingCreationDate para o formato DateTime"
```

```
## [1] "> conversão do campo LstCDate para o formato DateMonth"
```

```
## [1] "> conversão do campo FirstRecordedCreditLine para o formato DateTime"
```

```
## [1] "> Conversão do campo Term de numérico para um novo campo vetorial"
```

```
## [1] "> criação do campo DaysFromFirstRecCredLine que acumula o tempo em dias \n     desde
 o primeiro crédito"
```

```
## [1] "> criação do campo CreditScoreRangeMean para simplificação das\n     análises"
```

```
## [1] "> criação do campo FLoanOriginalAmount (versão fatorial do campo \n     LoanOriginalA
mount)"
```

```
## [1] "> criação do campo FInvestors (versão fatorial do campo Investors"
```

```
## [1] "> criação do campo FBorrowerAPR (versão fatorial do campo BorrowerAPR"
```

```
## [1] "> criação do campo ProsperGain (ganho da financeira)"
```

```
## [1] "> filtrando os registros com LoanStatus=Completed"
```

```
## [1] "> filtrando os registros com DebtToIncomeRatio<10"
```

```
## [1] "> filtrando os registros posteriores à 2005 (22 movimentos) e anteriores à 2014 (66 m
ovimentos)"
```

```
#table(pld$IncomeRange)
```

2.2. Seleção das variáveis de interesse

> O objetivo geral nesta etapa é selecionar as variáveis de interesse (LVI) que possam caracterizar alguma relação entre os principais indicadores financeiros relativos ao mutuário e à taxa de juros proposta no fechamento do empréstimo.

Caracterização do Conjunto de Dados

```
## [1] "> Número de Variáveis do Conjunto de Dados: 80"
```

```
## [1] "> Número de registros do conjuntos de dados:  35116"
```

```
## [1] "> Relação das variáveis disponíveis no conjunto de dados:"
```

```
## 'data.frame':    35116 obs. of  91 variables:
##  $ ListingKey                        : Factor w/ 113066 levels "00003546482094282EF90E
5",..: 7180 6647 6720 7161 6825 6827 6840 6880 7420 6919 ...
##  $ ListingNumber                     : int  193129 81716 213551 241498 713131 463083 5552
13 478891 24135 378497 ...
##  $ ListingCreationDate               : Factor w/ 113064 levels "2005-11-09 20:44:28.84700
0000",..: 14184 6429 15374 16841 69285 33837 49717 35437 1930 26897 ...
##  $ CreditGrade                       : Factor w/ 9 levels "","A","AA","B",..: 5 8 5 6 1 1
1 1 4 5 ...
##  $ Term                              : int  36 36 36 36 60 36 36 36 36 36 ...
##  $ LoanStatus                        : Factor w/ 12 levels "Cancelled","Chargedoff",..: 3
3 3 3 3 3 3 3 3 ...
##  $ ClosedDate                        : Factor w/ 2803 levels "","2005-11-25 00:00:00",..:
1138 1263 538 1623 2664 2553 2357 2359 1207 1306 ...
##  $ BorrowerAPR                       : num  0.165 0.283 0.15 0.215 0.307 ...
##  $ BorrowerRate                      : num  0.158 0.275 0.133 0.207 0.281 ...
##  $ LenderYield                       : num  0.138 0.24 0.122 0.198 0.271 ...
##  $ EstimatedEffectiveYield           : num  NA NA NA NA 0.247 ...
##  $ EstimatedLoss                     : num  NA NA NA NA 0.122 ...
##  $ EstimatedReturn                   : num  NA NA NA NA 0.125 ...
##  $ ProsperRating..numeric.           : int  NA NA NA NA 2 6 6 7 NA NA ...
##  $ ProsperRating..Alpha.             : Ord.factor w/ 7 levels "AA"<"A"<"B"<"C"<..: NA NA
 NA NA NA 2 2 1 NA NA ...
##  $ ListingCategory..numeric.         : int  0 0 0 0 1 1 20 7 0 1 ...
##  $ BorrowerState                     : Factor w/ 52 levels "","AK","AL","AR",..: 7 12 1 2
4 36 7 6 16 1 16 ...
##  $ Occupation                        : Factor w/ 68 levels "","Accountant/CPA",..: 37 37
 68 43 49 21 37 37 1 2 ...
##  $ EmploymentStatus                  : Factor w/ 9 levels "","Employed",..: 9 4 3 3 2 3 2
2 1 3 ...
##  $ EmploymentStatusDuration          : int  2 NA 19 36 25 10 44 7 NA 32 ...
##  $ IsBorrowerHomeowner               : Factor w/ 2 levels "False","True": 2 1 1 1 1 2 1 1
1 1 ...
##  $ CurrentlyInGroup                  : Factor w/ 2 levels "False","True": 2 2 1 2 1 1 1 1
1 2 ...
##  $ GroupKey                          : Factor w/ 707 levels "","0034337690131242316873
1",..: 1 335 1 307 1 1 1 1 287 467 ...
##  $ DateCreditPulled                  : Factor w/ 112992 levels "2005-11-09 00:30:04.48700
0000",..: 14347 6446 15488 16463 69245 33566 49592 35389 2011 26980 ...
##  $ CreditScoreRangeLower             : int  640 480 640 620 660 700 740 760 680 660 ...
##  $ CreditScoreRangeUpper             : int  659 499 659 639 679 719 759 779 699 679 ...
##  $ FirstRecordedCreditLine           : Factor w/ 11586 levels "","1947-08-24 00:00:0
0",..: 8639 8927 8310 9255 776 8094 3883 9418 6117 7767 ...
##  $ CurrentCreditLines                : int  5 NA 2 4 7 16 4 6 NA 10 ...
##  $ OpenCreditLines                   : int  4 NA 2 4 6 16 4 3 NA 9 ...
##  $ TotalCreditLinespast7years        : int  12 3 10 13 12 49 19 17 6 28 ...
##  $ OpenRevolvingAccounts             : int  1 0 1 3 6 11 4 3 4 6 ...
##  $ OpenRevolvingMonthlyPayment       : num  24 0 40 15 95 294 105 58 278 151 ...
##  $ InquiriesLast6Months              : int  3 0 3 0 2 2 0 0 1 1 ...
##  $ TotalInquiries                    : num  3 1 5 8 3 8 0 3 3 7 ...
##  $ CurrentDelinquencies              : int  2 1 3 1 0 0 1 0 0 0 ...
##  $ AmountDelinquent                  : num  472 NA 2224 5200 0 ...
##  $ DelinquenciesLast7Years           : int  4 0 1 5 0 0 10 0 0 29 ...
##  $ PublicRecordsLast10Years          : int  0 0 0 1 0 0 1 1 0 0 ...
##  $ PublicRecordsLast12Months         : int  0 NA 0 0 0 0 0 0 NA 0 ...
##  $ RevolvingCreditBalance            : num  0 NA 1220 134 2033 ...
##  $ BankcardUtilization               : num  0 NA 0.32 0.08 0.3 0.09 0.13 0.01 NA 0.68 ...
```

```
##  $ AvailableBankcardCredit            : num  1500 NA 2580 1366 3619 ...
##  $ TotalTrades                        : num  11 NA 7 6 10 40 18 14 NA 27 ...
##  $ TradesNeverDelinquent..percentage. : num  0.81 NA 0.4 0.37 1 1 0.72 0.92 NA 0.81 ...
##  $ TradesOpenedLast6Months            : num  0 NA 0 0 1 2 1 0 NA 0 ...
##  $ DebtToIncomeRatio                  : num  0.17 0.06 0.27 0.09 0.11 0.26 0.11 0.05 0.06
 0.27 ...
##  $ IncomeRange                        : Ord.factor w/ 6 levels "Not employed"<..: 3 NA 2 3
3 5 3 6 NA 3 ...
##  $ IncomeVerifiable                   : Factor w/ 2 levels "False","True": 2 2 2 2 2 2 2 2
2 2 ...
##  $ StatedMonthlyIncome                : num  3083 2083 1667 3750 3886 ...
##  $ LoanKey                            : Factor w/ 113066 levels "00003683605746079487FF
7",..: 100337 46303 107474 75508 25952 55706 111176 64426 72310 9196 ...
##  $ TotalProsperLoans                  : int  NA NA NA NA NA NA NA 1 NA 1 ...
##  $ TotalProsperPaymentsBilled         : int  NA NA NA NA NA NA NA 8 NA 9 ...
##  $ OnTimeProsperPayments              : int  NA NA NA NA NA NA NA 8 NA 9 ...
##  $ ProsperPaymentsLessThanOneMonthLate: int  NA NA NA NA NA NA NA 0 NA 0 ...
##  $ ProsperPaymentsOneMonthPlusLate    : int  NA NA NA NA NA NA NA 0 NA 0 ...
##  $ ProsperPrincipalBorrowed           : num  NA NA NA NA NA NA NA 4000 NA 6000 ...
##  $ ProsperPrincipalOutstanding        : num  NA NA NA NA NA ...
##  $ ScorexChangeAtTimeOfListing        : int  NA NA NA NA NA NA NA 161 NA 0 ...
##  $ LoanCurrentDaysDelinquent          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LoanFirstDefaultedCycleNumber      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanMonthsSinceOrigination         : int  78 86 77 75 13 45 25 41 92 67 ...
##  $ LoanNumber                         : int  19141 6466 20907 23565 85538 43110 60054 4498
7 1901 35508 ...
##  $ LoanOriginalAmount                 : int  9425 3001 1000 3000 4000 4000 10000 16000 100
00 4000 ...
##  $ LoanOriginationDate                : Factor w/ 1873 levels "2005-11-15 00:00:00",..: 42
6 260 451 488 1609 941 1348 1026 134 659 ...
##  $ LoanOriginationQuarter             : Ord.factor w/ 34 levels "Q4 2005"<"Q1 2006"<..: 8
 6 9 9 30 19 26 21 4 12 ...
##  $ MemberKey                          : Factor w/ 90831 levels "00003397697413387CAF96
6",..: 11071 33781 31599 14155 74985 63754 18434 6778 12488 54178 ...
##  $ MonthlyLoanPayment                 : num  330.4 123.3 33.8 112.6 124.8 ...
##  $ LP_CustomerPayments                : num  11396 4187 1012 4061 4725 ...
##  $ LP_CustomerPrincipalPayments       : num  9425 3001 1000 3000 4000 ...
##  $ LP_InterestandFees                 : num  1971.1 1185.6 11.6 1061 725.4 ...
##  $ LP_ServiceFees                     : num  -133.18 -24.2 -0.88 -51.12 -25.81 ...
##  $ LP_CollectionFees                  : num  0 0 0 0 0 ...
##  $ LP_GrossPrincipalLoss              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NetPrincipalLoss                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NonPrincipalRecoverypayments    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PercentFunded                      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Recommendations                    : int  0 0 0 0 0 0 0 0 0 2 ...
##  $ InvestmentFromFriendsCount         : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ InvestmentFromFriendsAmount        : num  0 0 0 0 0 ...
##  $ Investors                          : int  258 41 53 53 37 121 30 326 44 103 ...
##  $ FLoanOriginationYear               : Ord.factor w/ 8 levels "2006"<"2007"<..: 2 2 2 2 8
5 7 5 1 3 ...
##  $ LstCDate                           : POSIXct, format: "2007-08-26" "2007-01-05" ...
##  $ LstCMonth                          : Date, format: "2007-08-01" "2007-01-01" ...
##  $ FirstRecCredLine                   : Date, format: "2001-10-11" "2002-07-27" ...
##  $ FTerm                              : Ord.factor w/ 3 levels "12"<"36"<"60": 2 2 2 2 3 2
2 2 2 2 ...
##  $ DaysFromFirstRecCredLine           : num  5806 5517 6139 5189 15288 ...
##  $ CreditScoreRangeMean               : num  650 490 650 630 670 ...
##  $ FLoanOriginalAmount                : Ord.factor w/ 9 levels "(0-4.8K]"<"(4.8K-8.6K]"
```

```
<..: 3 1 1 1 1 1 3 4 3 1 ...
##  $ FInvestors                        : Ord.factor w/ 5 levels "(0,100]"<"(100,200]"<..: 3
1 1 1 1 2 1 4 1 2 ...
##  $ FBorrowerAPR                      : Ord.factor w/ 6 levels "0"<"0.1"<"0.2"<..: 3 4 3 3
4 2 2 2 3 3 ...
##  $ ProsperGain                       : num  188 105 10 30 40 ...
```

Pré-seleção e criação da lsita de variáveis de interesse (LVI)

Nesta etapa serão armazenadas em uma lista apenas as variáveis do dataset que atendam à dois critérios: 1- percentual de registros inválidos 2- número de categorias de classificação (válido apenas para variáveis fatoriais) Como uma relação completa das variáveis disponíveis no conjunto de dados já foi apresentado, relacionaremos abaixo apenas as variáveis que não atenderam aos requisitos acima e que portanto não foram inseridas na LVI.

```
## [1] "> Pré-Seleção de variáveis:"
```

```
## [1] "Critério1: Percentual de registros inválidos < 10 %"
```

```
## [1] "Critério2: Variáveis fatoriais com até  24 classes"
```

```
## [1] "> ListingKey : variável descartada   - Motivo: núm. valores únicos   =  35116"
## [1] "> ListingCreationDate : variável descartada   - Motivo: núm. valores únicos   =  3511
6"
## [1] "> CreditGrade : variável descartada   - Motivo: % registros inválidos =  50.48"
## [1] "> LoanStatus : variável descartada   - Motivo: núm. valores únicos   =  1"
## [1] "> ClosedDate : variável descartada   - Motivo: núm. valores únicos   =  2513"
## [1] "> EstimatedEffectiveYield : variável descartada   - Motivo: % registros inválidos =
 49.84"
## [1] "> EstimatedLoss : variável descartada   - Motivo: % registros inválidos =  49.84"
## [1] "> EstimatedReturn : variável descartada   - Motivo: % registros inválidos =  49.84"
## [1] "> ProsperRating..numeric. : variável descartada   - Motivo: % registros inválidos =
 49.84"
## [1] "> ProsperRating..Alpha. : variável descartada   - Motivo: % registros inválidos =  5
5.84"
## [1] "> BorrowerState : variável descartada   - Motivo: núm. valores únicos   =  52"
## [1] "> Occupation : variável descartada   - Motivo: núm. valores únicos   =  68"
## [1] "> EmploymentStatusDuration : variável descartada   - Motivo: % registros inválidos =
12.58"
## [1] "> GroupKey : variável descartada   - Motivo: núm. valores únicos   =  590"
## [1] "> DateCreditPulled : variável descartada   - Motivo: núm. valores únicos   =  35114"
## [1] "> FirstRecordedCreditLine : variável descartada   - Motivo: núm. valores únicos   =
 8010"
## [1] "> CurrentCreditLines : variável descartada   - Motivo: % registros inválidos =  12.5
7"
## [1] "> OpenCreditLines : variável descartada   - Motivo: % registros inválidos =  12.57"
## [1] "> AmountDelinquent : variável descartada   - Motivo: % registros inválidos =  12.56"
## [1] "> PublicRecordsLast12Months : variável descartada   - Motivo: % registros inválidos =
  12.57"
## [1] "> RevolvingCreditBalance : variável descartada   - Motivo: % registros inválidos =  1
2.57"
## [1] "> BankcardUtilization : variável descartada   - Motivo: % registros inválidos =  12.5
7"
## [1] "> AvailableBankcardCredit : variável descartada   - Motivo: % registros inválidos =
 12.45"
## [1] "> TotalTrades : variável descartada   - Motivo: % registros inválidos =  12.45"
## [1] "> TradesNeverDelinquent..percentage. : variável descartada   - Motivo: % registros in
válidos =  12.45"
## [1] "> TradesOpenedLast6Months : variável descartada   - Motivo: % registros inválidos =
 12.45"
## [1] "> IncomeRange : variável descartada   - Motivo: % registros inválidos =  12.8"
## [1] "> LoanKey : variável descartada   - Motivo: núm. valores únicos   =  35116"
## [1] "> TotalProsperLoans : variável descartada   - Motivo: % registros inválidos =  79.07"
## [1] "> TotalProsperPaymentsBilled : variável descartada   - Motivo: % registros inválidos
 =  79.07"
## [1] "> OnTimeProsperPayments : variável descartada   - Motivo: % registros inválidos =  7
9.07"
## [1] "> ProsperPaymentsLessThanOneMonthLate : variável descartada   - Motivo: % registros i
nválidos =  79.07"
## [1] "> ProsperPaymentsOneMonthPlusLate : variável descartada   - Motivo: % registros inval
idos =  79.07"
## [1] "> ProsperPrincipalBorrowed : variável descartada   - Motivo: % registros inválidos =
79.07"
## [1] "> ProsperPrincipalOutstanding : variável descartada   - Motivo: % registros inválidos
=  79.07"
## [1] "> ScorexChangeAtTimeOfListing : variável descartada   - Motivo: % registros inválidos
=  79.29"
## [1] "> LoanCurrentDaysDelinquent : variável descartada   - Motivo: núm. valores únicos   =
```

```
  1"
## [1] "> LoanFirstDefaultedCycleNumber : variável descartada   - Motivo: % registros inválid
os =  99.88"
## [1] "> LoanOriginationDate : variável descartada   - Motivo: núm. valores únicos   =  180
5"
## [1] "> LoanOriginationQuarter : variável descartada   - Motivo: núm. valores únicos   =  3
1"
## [1] "> MemberKey : variável descartada   - Motivo: núm. valores únicos   =  27742"
## [1] "> LP_GrossPrincipalLoss : variável descartada   - Motivo: núm. valores únicos   =  1"
## [1] "> LP_NetPrincipalLoss : variável descartada   - Motivo: núm. valores únicos   =  1"
## [1] "> LP_NonPrincipalRecoverypayments : variável descartada   - Motivo: núm. valores únic
os   =  1"
```

```
## [1] ">  IncomeRange : variável adicionada - Motivo:  Reinserida, por que pois após remover
os registros iguais à Not-displayed o percentual de missing values ficou maior que 10% (12,
8%)."
```

Remoção de variáveis secundárias

> Nesta etapa, em função do alto número de variáveis do conjunto de dados, serão removidas da LVI as variáveis que foram consideradas sem uma relevância significativa em um primeiro momento. Entretanto, no decorrer das análises, estas variáveis estarão sendo sempre reavaliadas, permanecendo na lista se necessárias.

```
## [1] "> Relação de variáveis secundárias descartadas"
```

```
## [1] "- Motivo:  variável sem relevância significativa ao tema investigado."
```

```
## [1] ">  CurrentlyInGroup : variável descartada "
```

```
## [1] ">  ListingNumber : variável descartada "
```

```
## [1] ">  Term : variável descartada "
```

```
## [1] ">  OpenRevolvingAccounts : variável descartada "
```

```
## [1] ">  OpenRevolvingMonthlyPayment : variável descartada "
```

```
## [1] ">  InquiriesLast6Months : variável descartada "
```

```
## [1] ">  CurrentDelinquencies : variável descartada "
```

```
## [1] ">  PublicRecordsLast10Years : variável descartada "
```

```
## [1] ">  StatedMonthlyIncome : variável descartada "
```

```
## [1] ">  LoanMonthsSinceOrigination : variável descartada "
```

```
## [1] ">  LoanNumber : variável descartada "
```

```
## [1] ">  LP_CustomerPayments : variável descartada "
```

```
## [1] ">  LP_CustomerPrincipalPayments : variável descartada "
```

```
## [1] ">  LP_InterestandFees : variável descartada "
```

```
## [1] ">  LP_ServiceFees : variável descartada "
```

```
## [1] ">  LP_CollectionFees : variável descartada "
```

```
## [1] ">  PercentFunded : variável descartada "
```

```
## [1] ">  Recommendations : variável descartada "
```

```
## [1] ">  InvestmentFromFriendsCount : variável descartada "
```

```
## [1] ">  InvestmentFromFriendsAmount : variável descartada "
```

```
## [1] ">  CreditScoreRangeLower : variável descartada "
```

```
## [1] ">  CreditScoreRangeUpper : variável descartada "
```

```
## [1] ">  BorrowerRate : variável descartada "
```

```
## [1] ">  LenderYield : variável descartada "
```

```
## [1] ">  MonthlyLoanPayment : variável descartada "
```

```
## [1] ">  LstCDate : variável descartada "
```

```
## [1] ">  LstCMonth : variável descartada "
```

```
## [1] ">  FirstRecCredLine : variável descartada "
```

```
## [1] ">  ListingCategory..numeric. : variável descartada "
```

Relação das variáveis de interesse selecionadas para análise

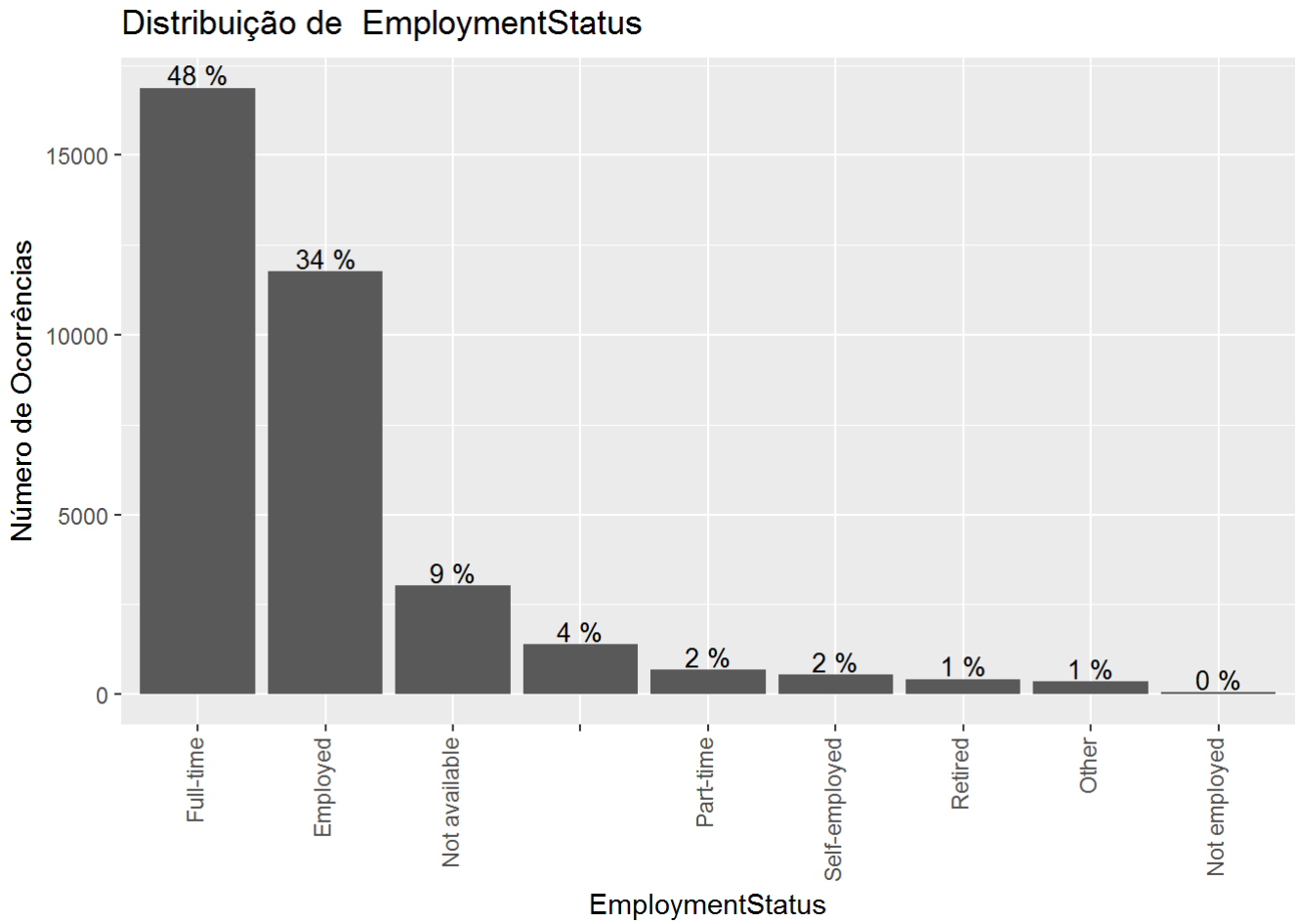
> Nesta etapa apresentamos a relação final da LVI com o nome de cada variável e seu tipo. Serão estas as variáveis que serão utilizadas nas análises seguintes.

```
##  [1] "[ n ] BorrowerAPR"                "[ f ] EmploymentStatus"
##  [3] "[ f ] IsBorrowerHomeowner"        "[ n ] TotalCreditLinespast7years"
##  [5] "[ n ] TotalInquiries"             "[ n ] DelinquenciesLast7Years"
##  [7] "[ n ] DebtToIncomeRatio"          "[ f ] IncomeVerifiable"
##  [9] "[ n ] LoanOriginalAmount"         "[ n ] Investors"
## [11] "[ f ] FLoanOriginationYear"       "[ f ] FTerm"
## [13] "[ n ] DaysFromFirstRecCredLine"   "[ n ] CreditScoreRangeMean"
## [15] "[ f ] FLoanOriginalAmount"        "[ f ] FInvestors"
## [17] "[ f ] FBorrowerAPR"               "[ n ] ProsperGain"
## [19] "[ f ] IncomeRange"
```

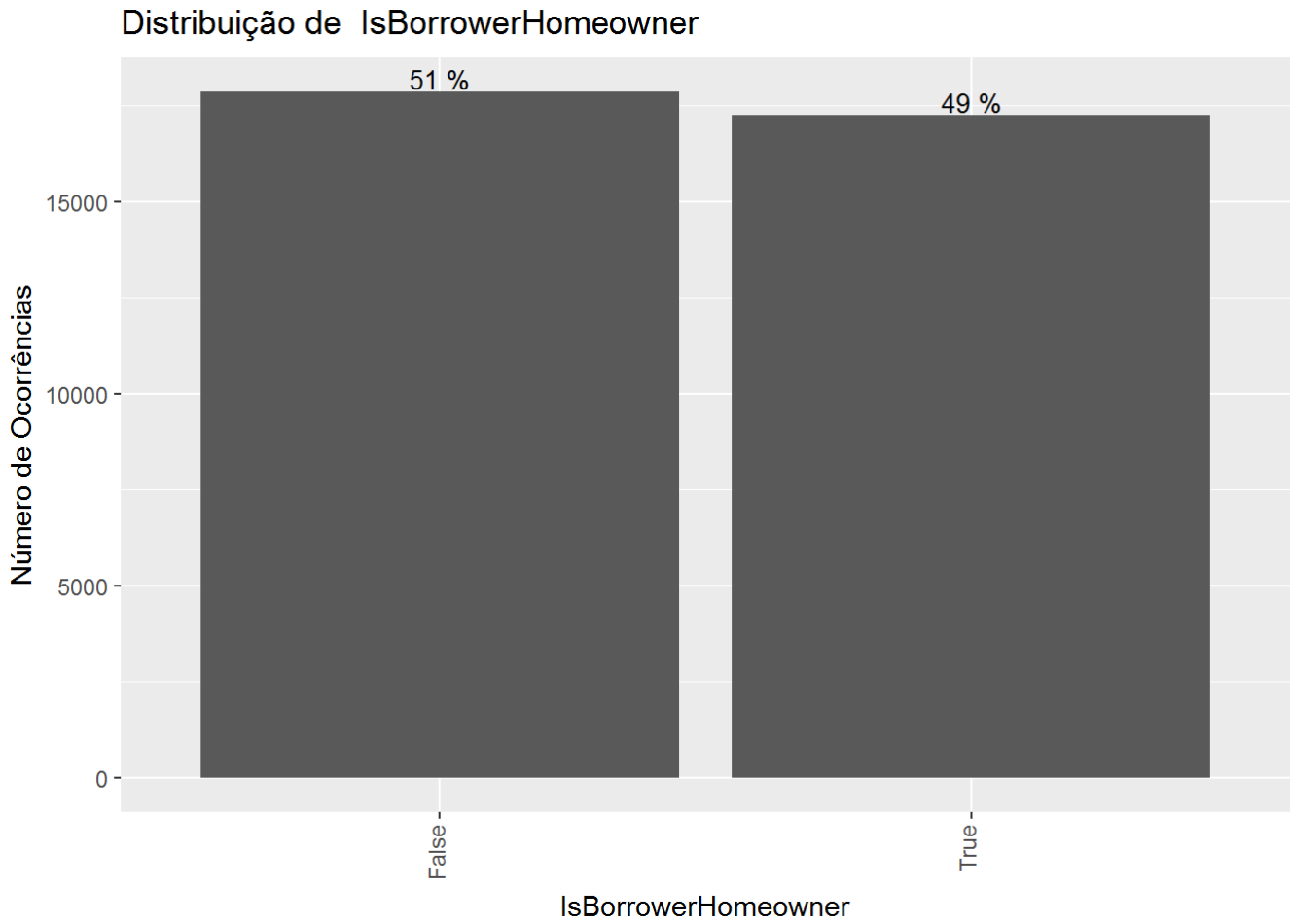
3. ANÁLISE E SESSÃO DE GRÁFICOS UNIVARIADOS

3.1. Distribuição das variáveis de interesse fatoriais

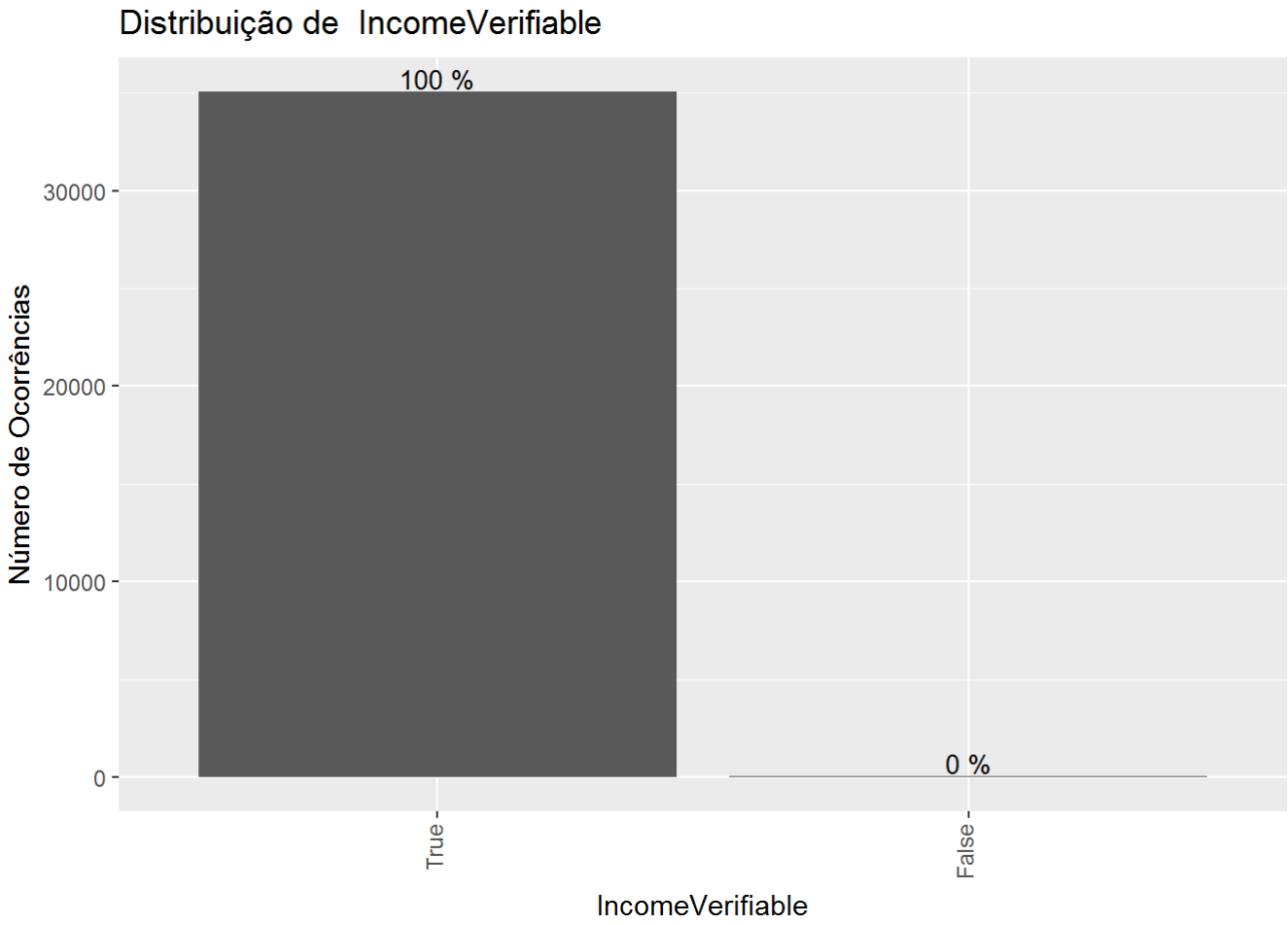
> Nesta etapa analisaremos a distribuição das ocorrências das variáveis de interesse fatoriais (categóricas) que foram selecionadas a fim de analisarmos o conteúdo e frequência das informações categóricas disponíveis.



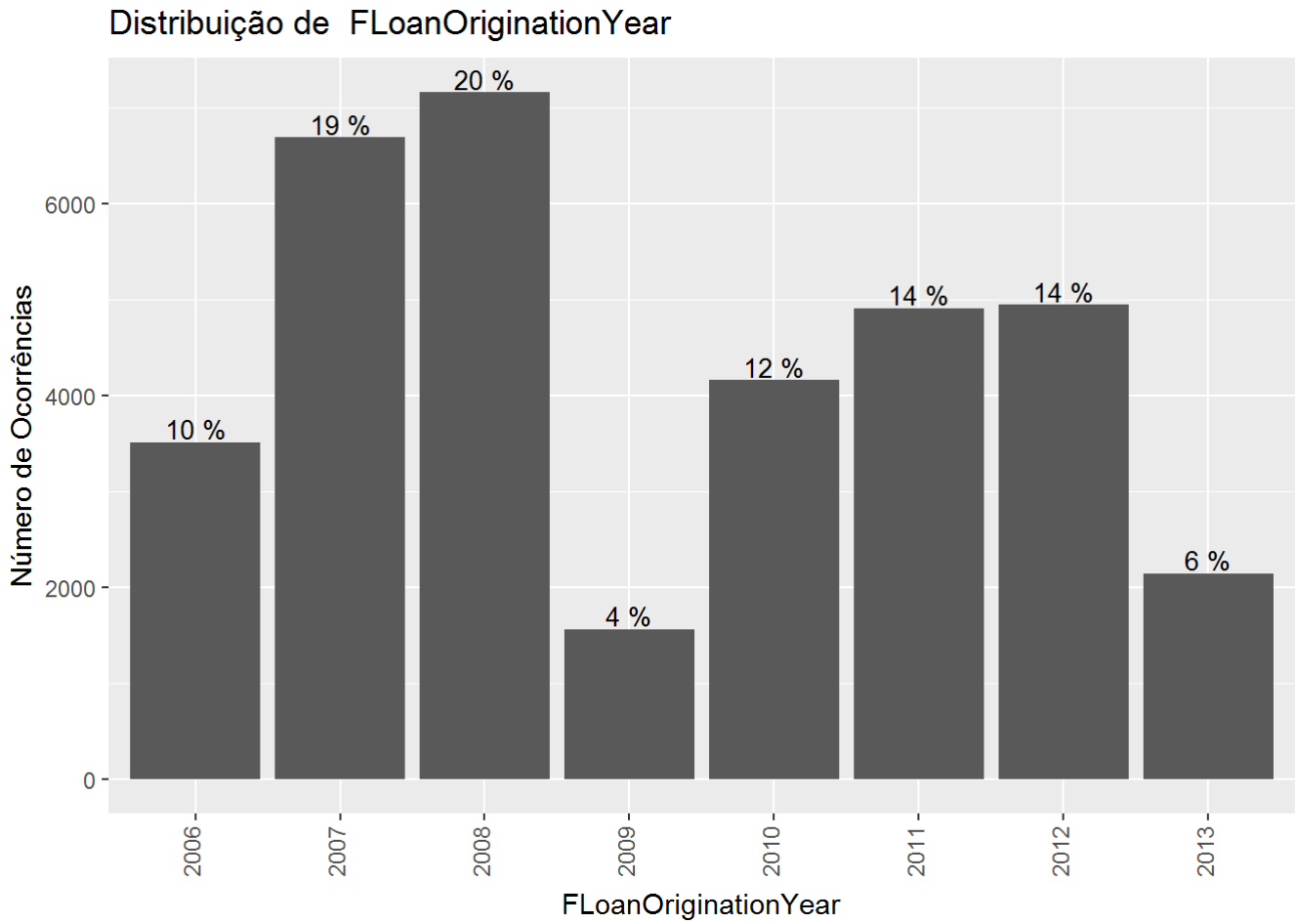

```
##
## [1] "> EmploymentStatus - Observações:  Cerca de 82% dos empréstimos foram destinados à pe
ssoas empregadas ou com trabalho full-time."
```



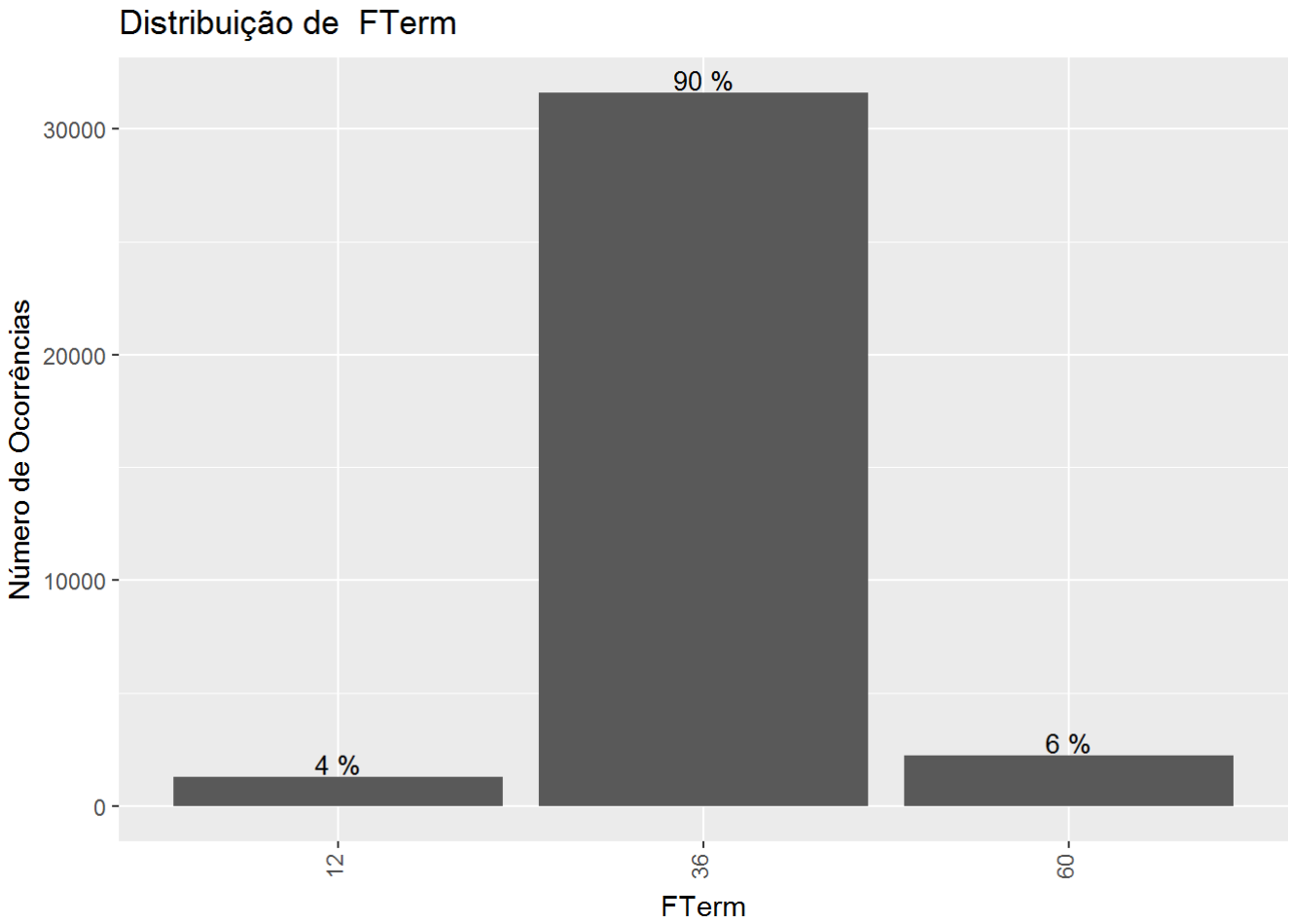

```
##
## [1] "> IsBorrowerHomeowner - Observações:  O percentual de empréstimos concedidos para pes
soas proprietárias de sua residência foi aproximadamente igual ao de pessoas sem residência p
ópria."
```



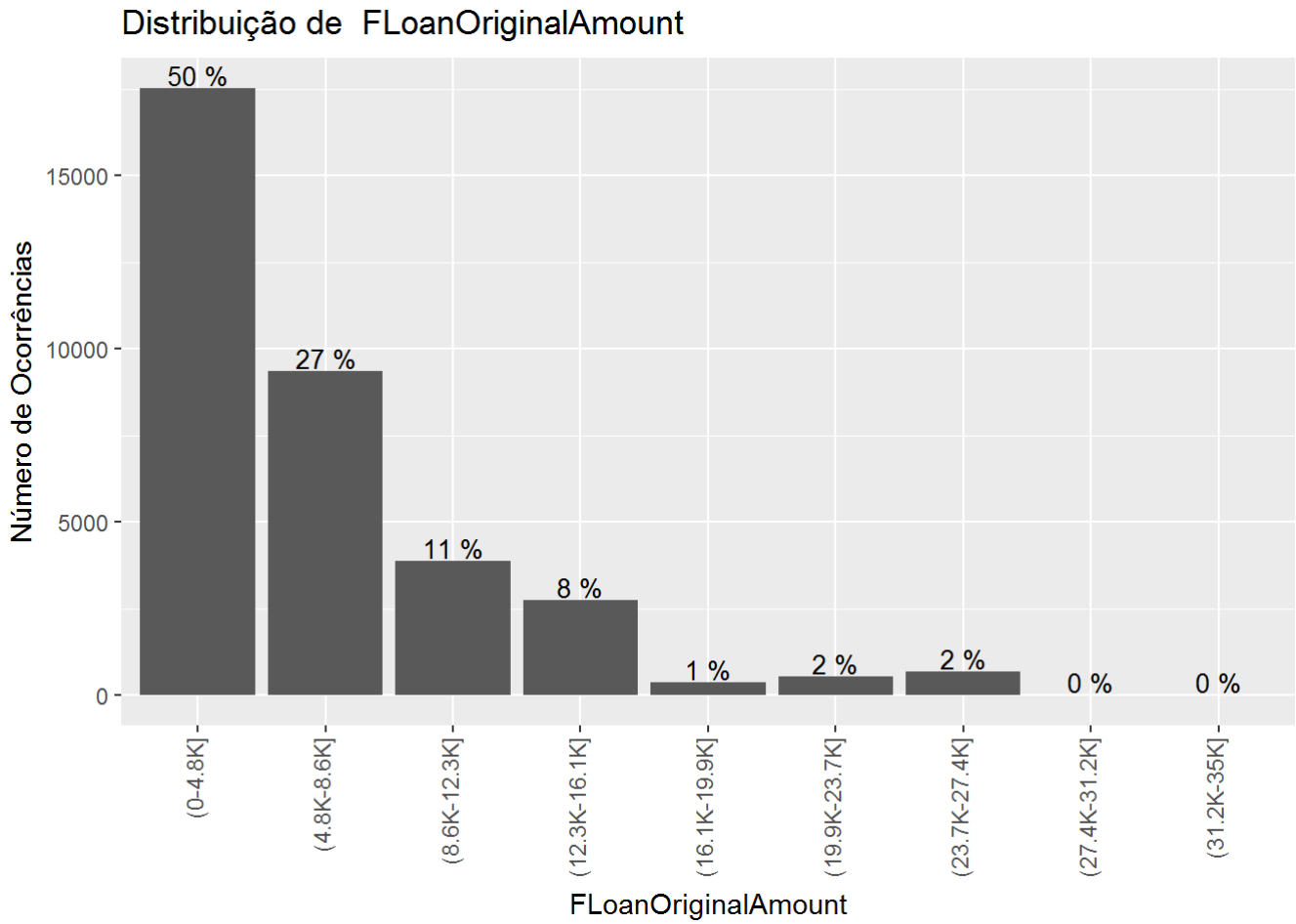

```
##
## [1] "> IncomeVerifiable - Observações:  Variável sem variância (constante) para o sub-conj
unto de dados analisado (LoanStatus=Completed)."
```



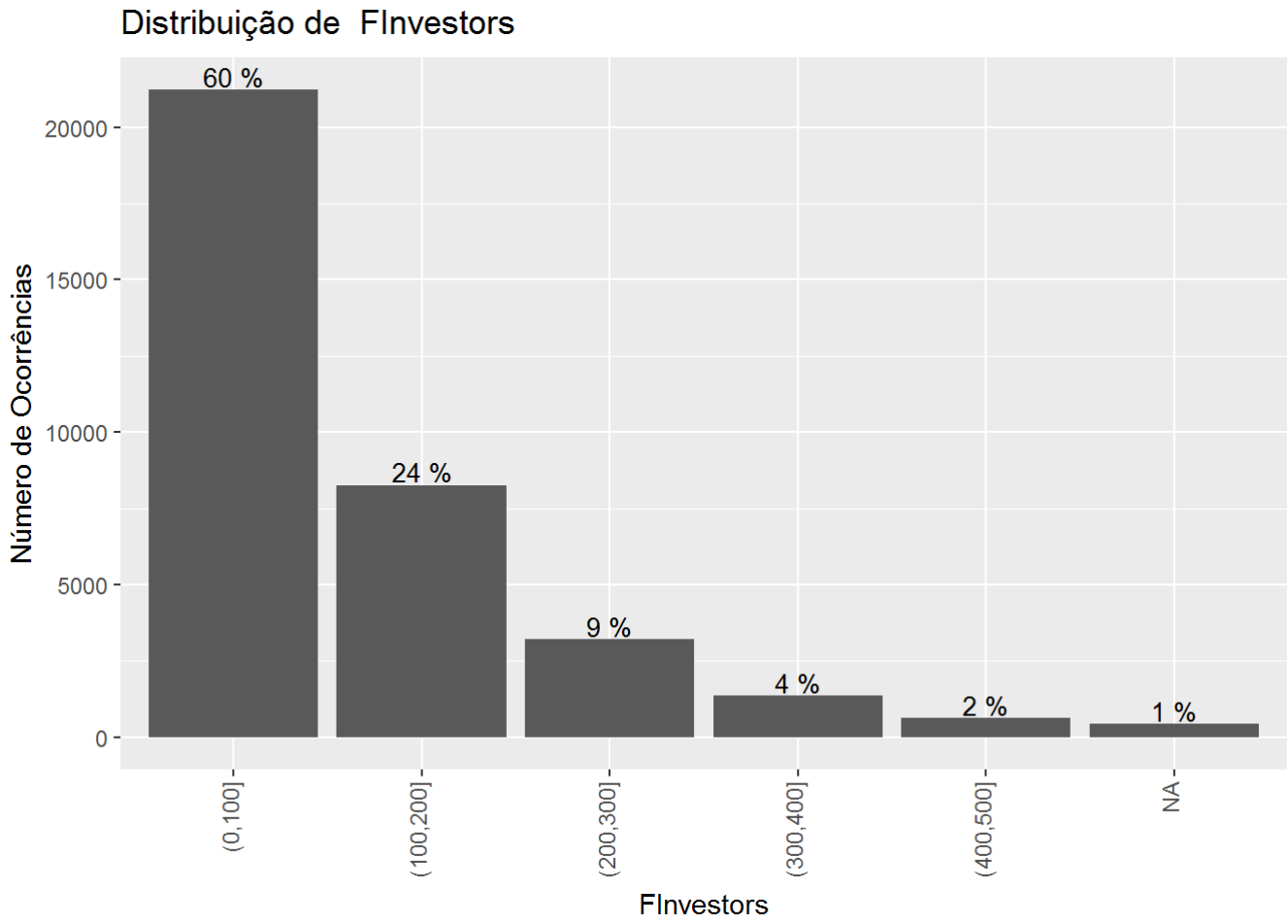

```
##
## [1] "> FLoanOriginationYear - Observações:  Percebe-se quedas significativas no volume de
 empréstimo concedidos nos anos de 2009 e 2013."
```



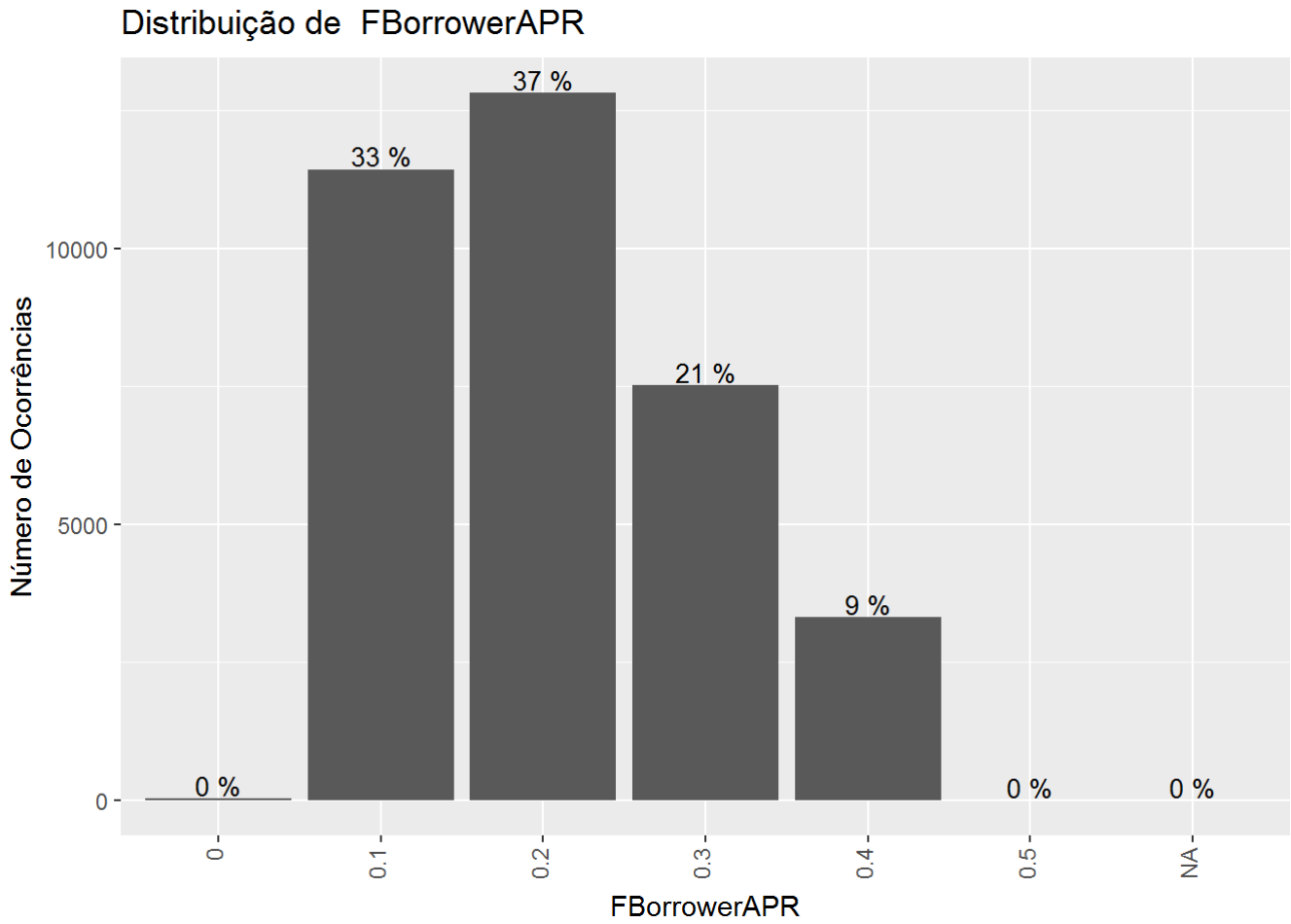

```
##
## [1] "> FTerm - Observações:  A grande maioria dos empréstimos concedidos (90%) tiveram pra
zo de pagamento de 3 anos."
```



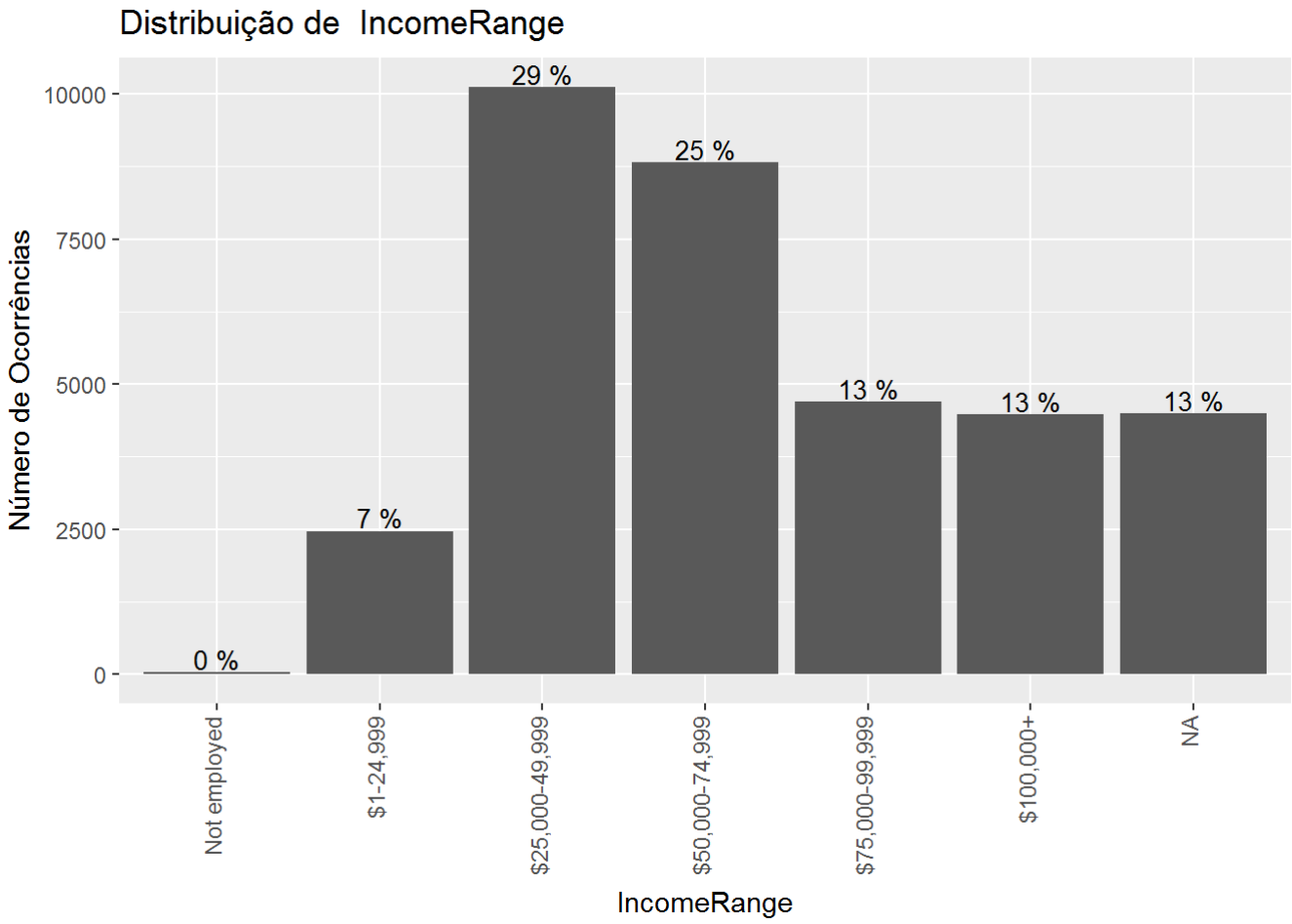

```
##
## [1] "> FLoanOriginalAmount - Observações:  O volume de empréstimos concedidos aumenta quan
to menor o valor do empréstimo solicitado, diferentemente do esperado, que seria de se concen
trar em torno do valor médio."
```




```
##
## [1] "> FInvestors - Observações:  A maioria dos empréstimos são fechado com no máximo até
 100 investidores (63%)."
```




```
##
## [1] "> FBorrowerAPR - Observações:  Percebe-se uma grande concentração dos emrpréstimos co
m taxa anuais de juros aprovadas entre 0,1 e 0,299 (69%)."
```




```
##
## [1] "> IncomeRange - Observações:  Cerca de 54% dos empréstimos foram concedidos à pessoas com rendimento na faixa entre 25 e 75 mil dólares."
```

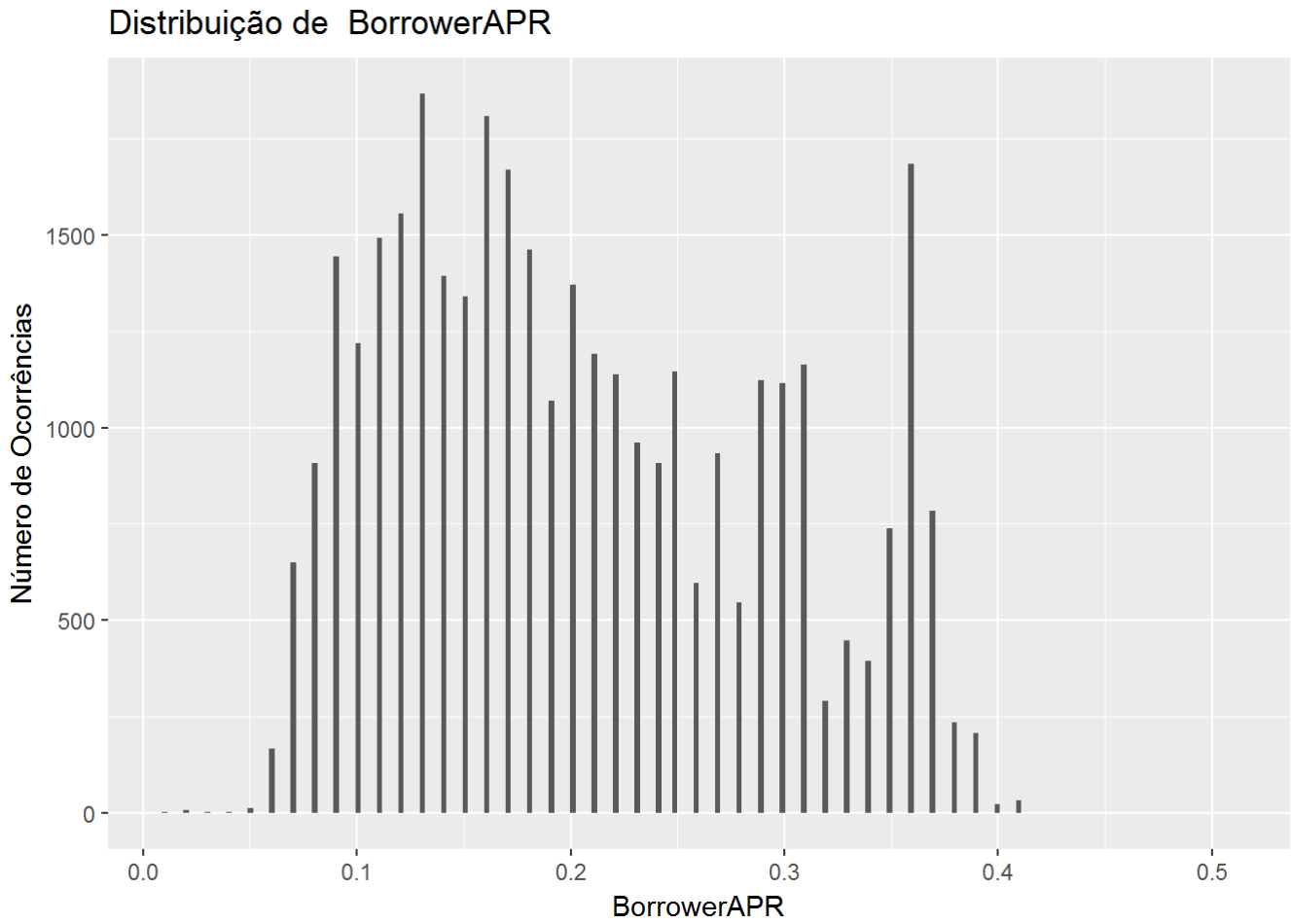
Atualização da lista de variáveis de interesse

> Após analisadas as distribuições do item anterior, será feito nessa etapa uma atualização da LVI, removendo e/ou inserindo variáveis conforme necessidade.

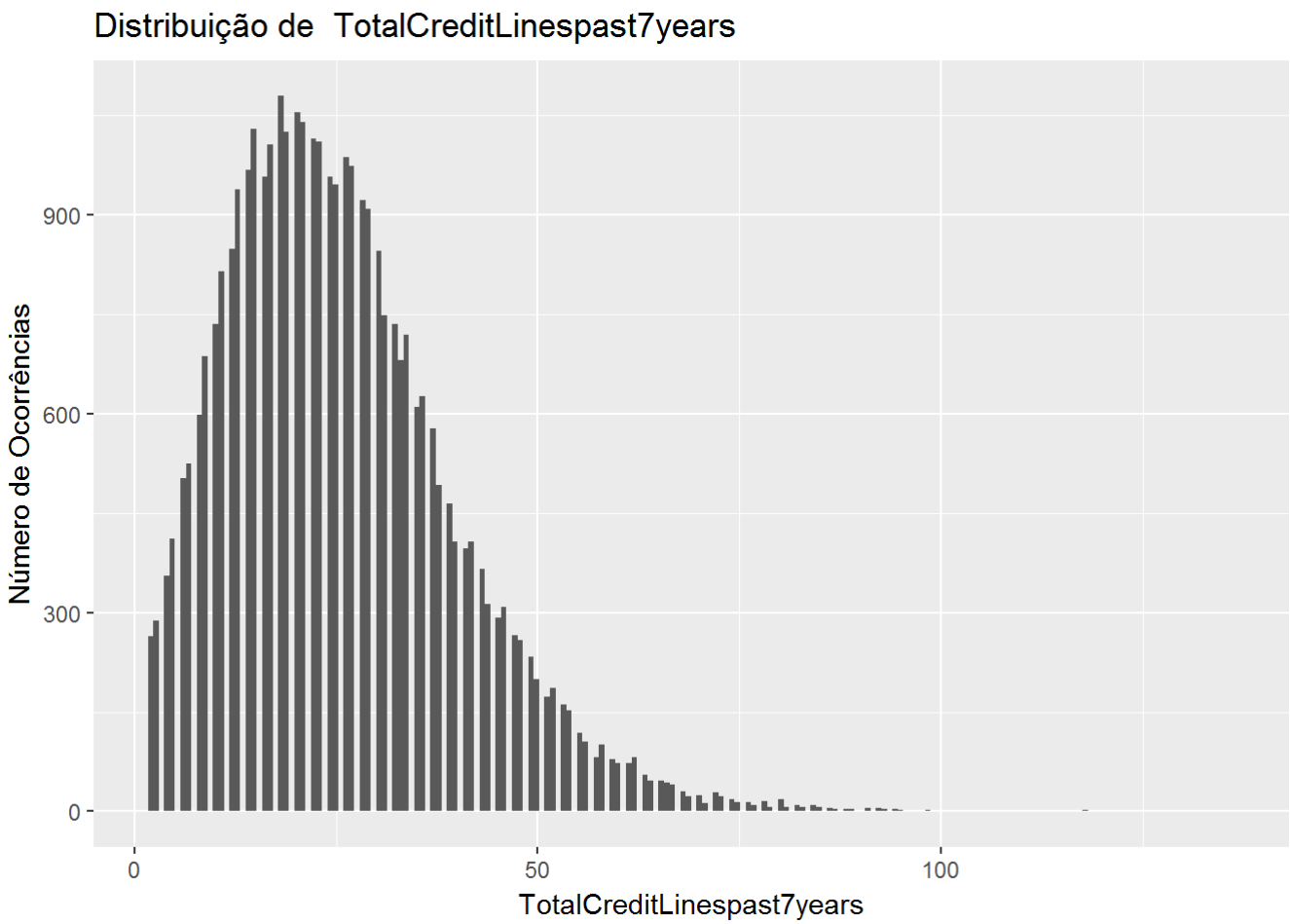
```
## [1] ">  IncomeVerifiable : variável descartada - Motivo:  Variável sem variância (constante) para o sub-conjunto de dados analisado (LoanStatus=Completed)."
```

3.2. Distribuição das variáveis de interesse numéricas

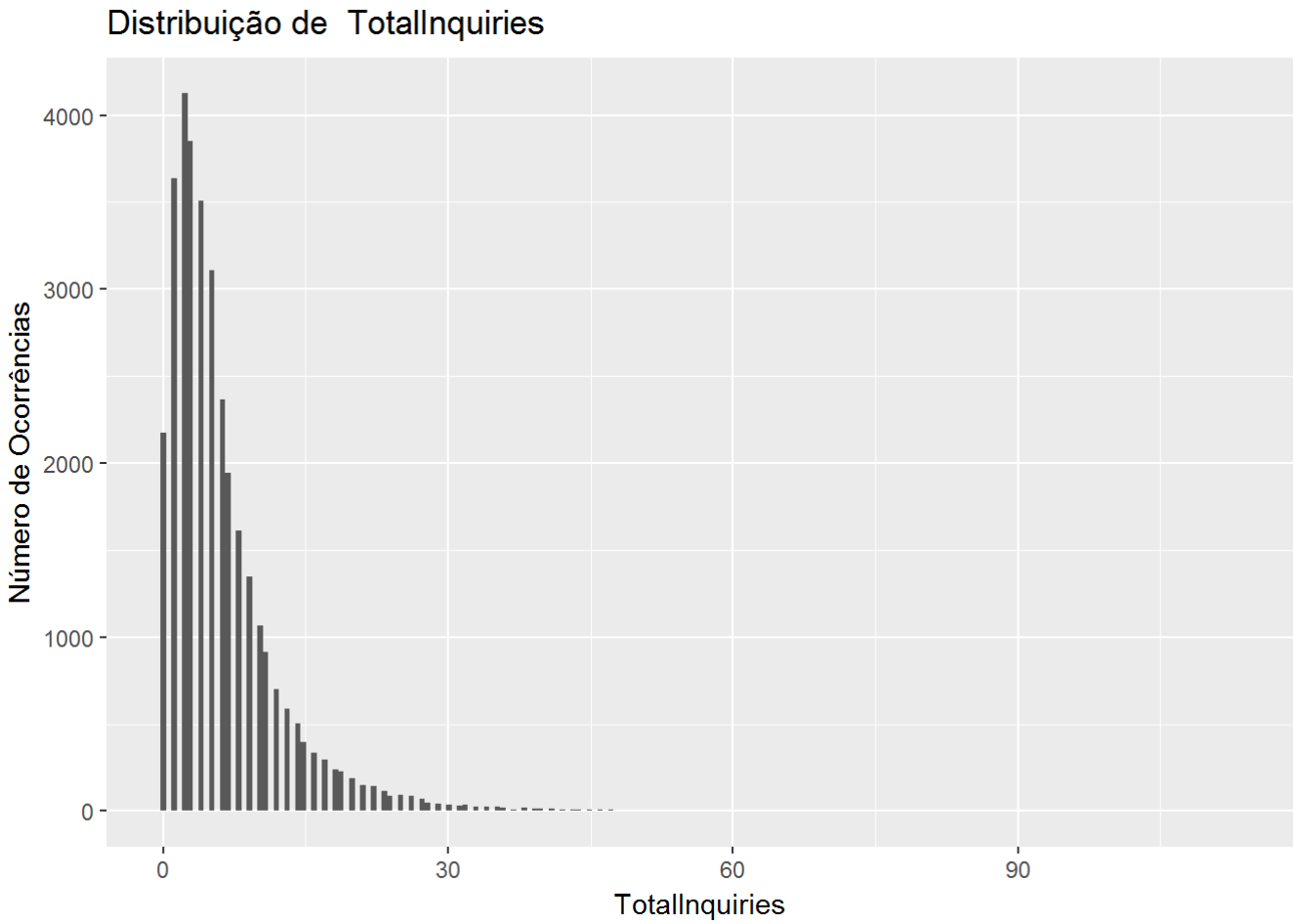
> Nesta etapa analisaremos a distribuição das ocorrências das variáveis de interesse numéricas que foram selecionadas a fim de analisarmos a forma de distribuição da frequência das informações disponíveis.



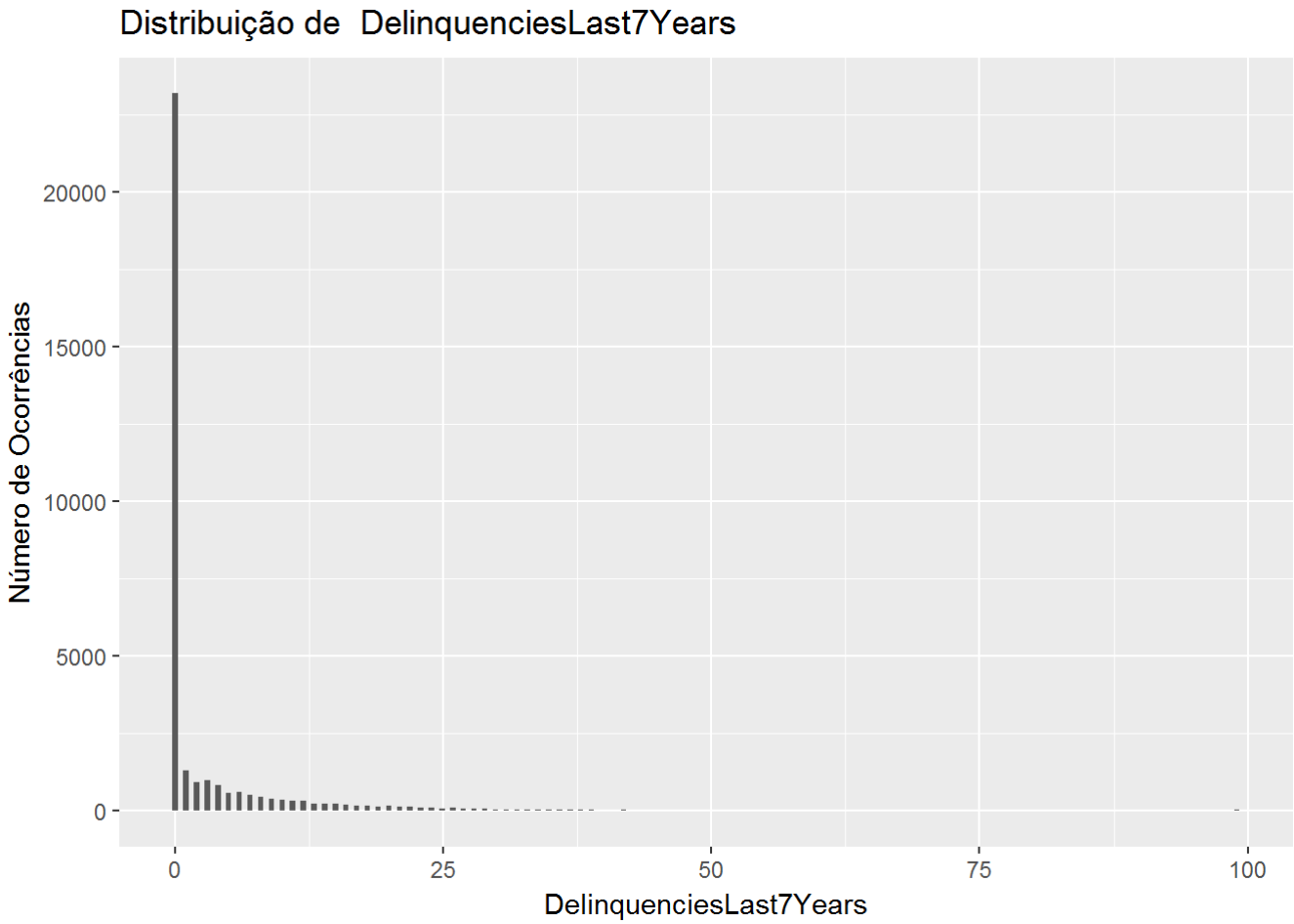

```
##
## [1] "> BorrowerAPR - Observações:  A distribuição das taxas de juros anuais em termos de n
úmero de ocorrências tende a se concentrar em torno da média, embora apresente picos signific
ativos nas taxas de 0,3 e 0,4%."
##
## [1] "> BorrowerAPR - Estatísticas: "
## [1] "- Valor Mín./Máx.:  0.007 / 0.512"
## [1] "- Média          :  0.205"
## [1] "- Mediana        :  0.19"
## [1] "- Desvio Padrão  :  0.088"
```



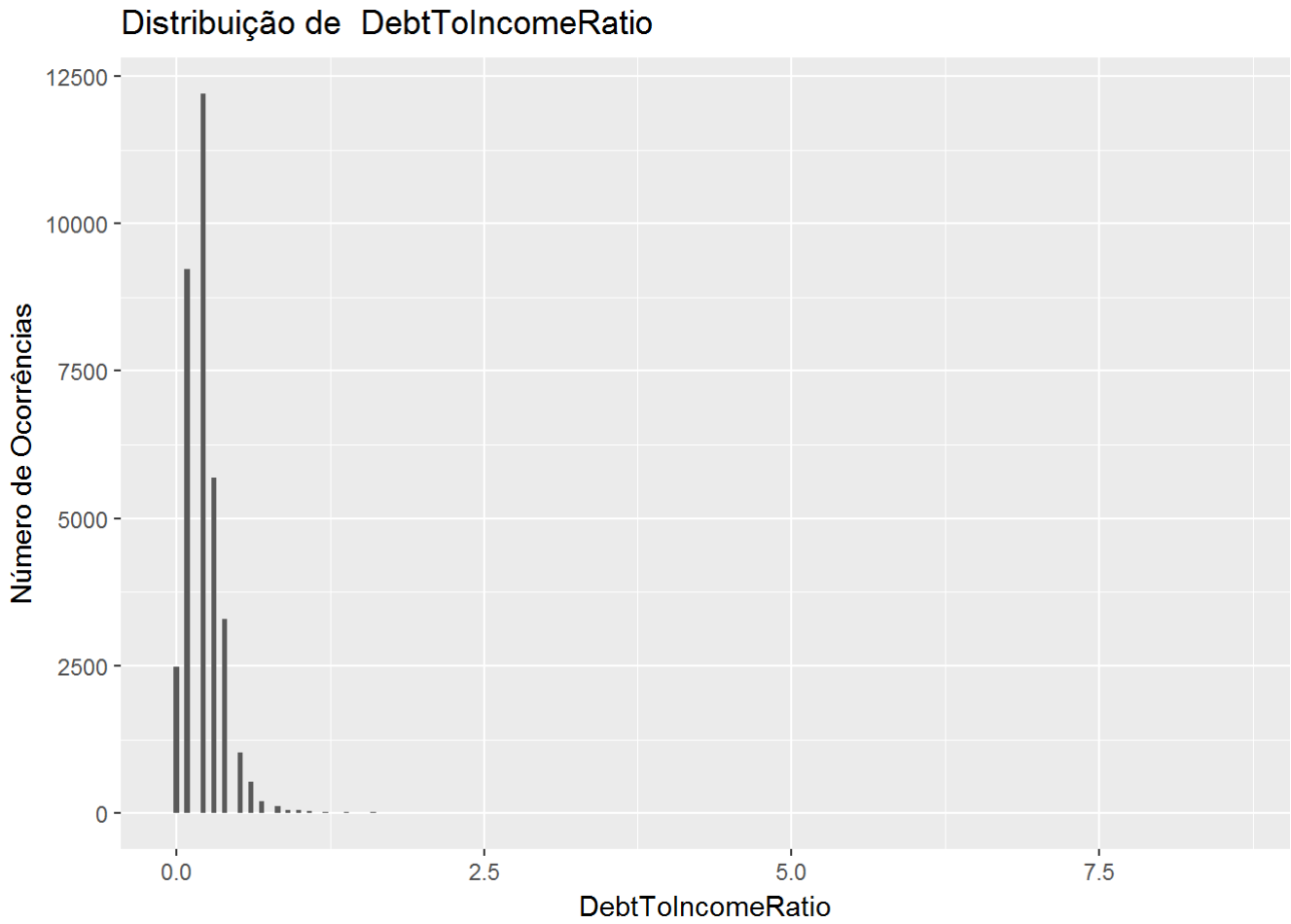

```
##
## [1] "> TotalCreditLinespast7years - Observações:  Este indicador apresenta uma distribuiçã
o normal, deslocada à esquerda, pois o mesmo não poderia apresentar valores negativos."
##
## [1] "> TotalCreditLinespast7years - Estatísticas: "
## [1] "- Valor Mín./Máx.:  2 / 136"
## [1] "- Média           :  25.667"
## [1] "- Mediana         :  24"
## [1] "- Desvio Padrão   :  14.115"
```



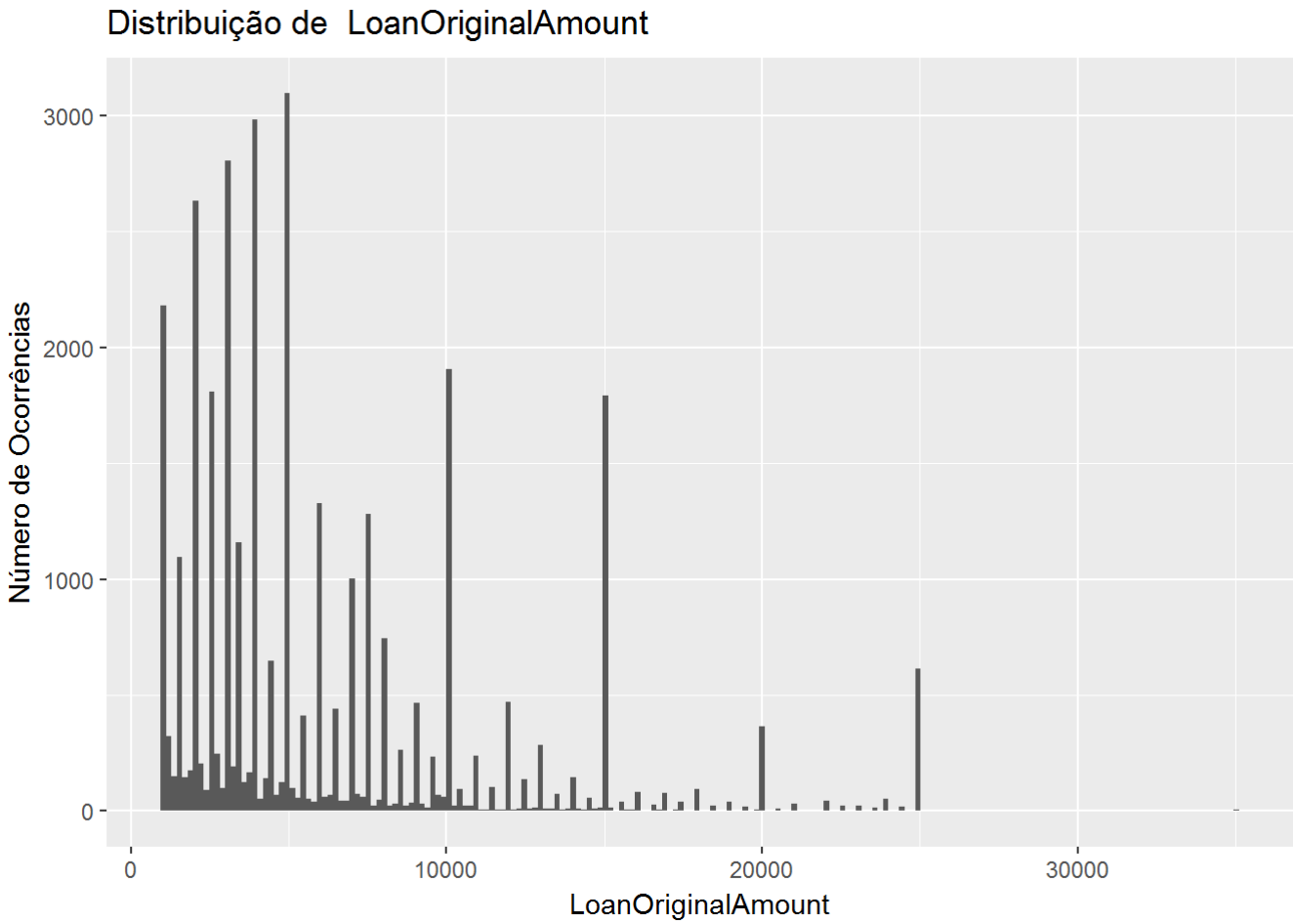

```
##
## [1] "> TotalInquiries - Observações:  A distribuição desta variável não está muito clara,
 pois confunde o comportamento de uma distribuição normal com uma exponencial negativa. No it
em seguinte o gráfico será transformado para melhor visualização."
##
## [1] "> TotalInquiries - Estatísticas: "
## [1] "- Valor Mín./Máx.:  0 / 113"
## [1] "- Média          :  6.243"
## [1] "- Mediana        :  4"
## [1] "- Desvio Padrão  :  6.559"
```



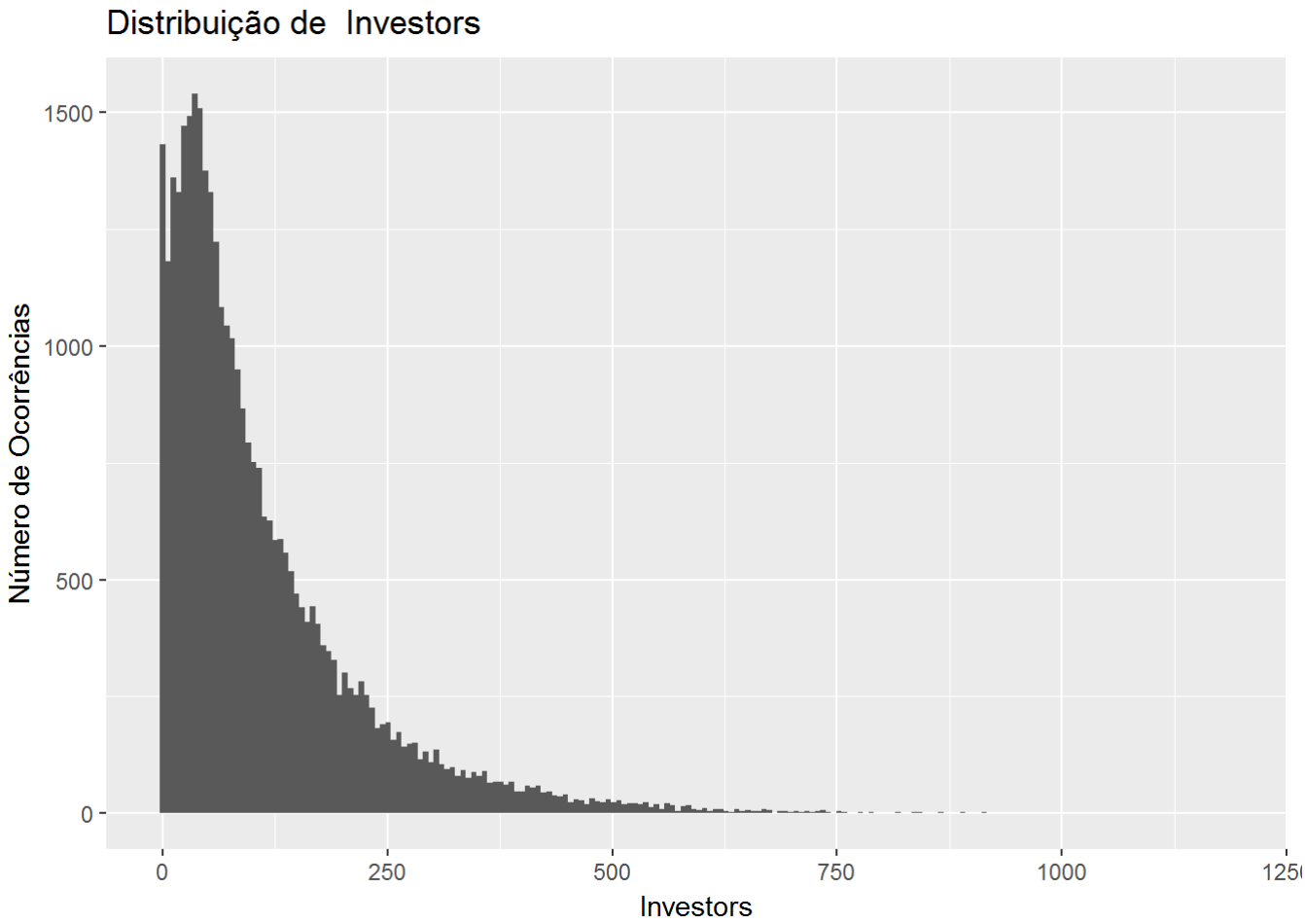

```
##
## [1] "> DelinquenciesLast7Years - Observações:  Esta variável possui uma grande concentraçã
o de ocorrências em torno do valor 0. Para uma melhor visualização será preciso transformar a
variável a fim de verificar a distribuição nos valores acima de 0 no eixo X."
##
## [1] "> DelinquenciesLast7Years - Estatísticas: "
## [1] "- Valor Mín./Máx.:  0 / 99"
## [1] "- Média          :  4.04"
## [1] "- Mediana        :  0"
## [1] "- Desvio Padrão  :  9.974"
```



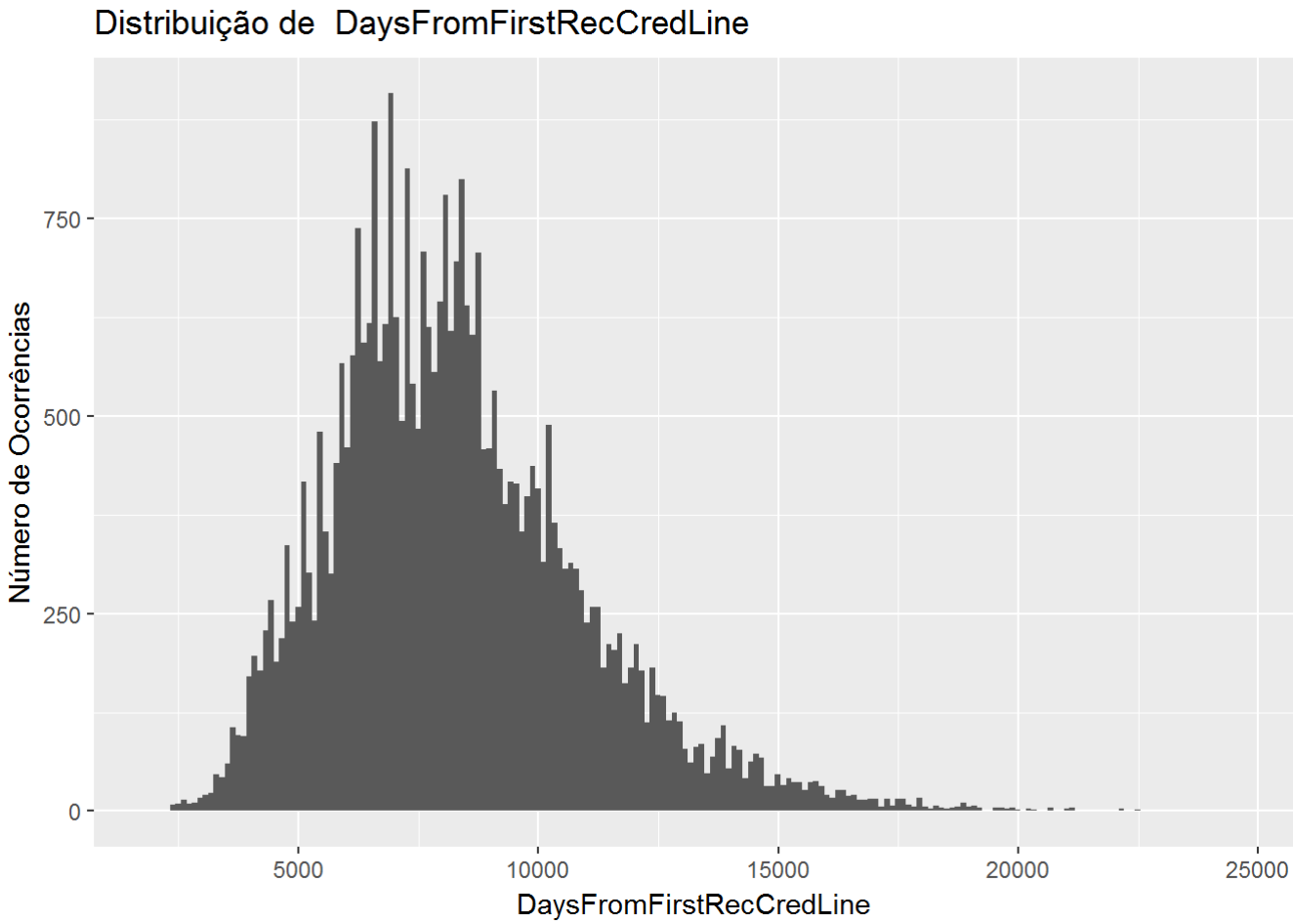

```
##
## [1] "> DebtToIncomeRatio - Observações:  A distribuição desta variável não está muito clar
a, pois confunde o comportamento de uma distribuição normal com uma exponencial negativa. No
 item seguinte o gráfico será transformado para melhor visualização."
##
## [1] "> DebtToIncomeRatio - Estatísticas: "
## [1] "- Valor Mín./Máx.:  0 / 8.63"
## [1] "- Média          :  0.23"
## [1] "- Mediana        :  0.19"
## [1] "- Desvio Padrão  :  0.248"
```



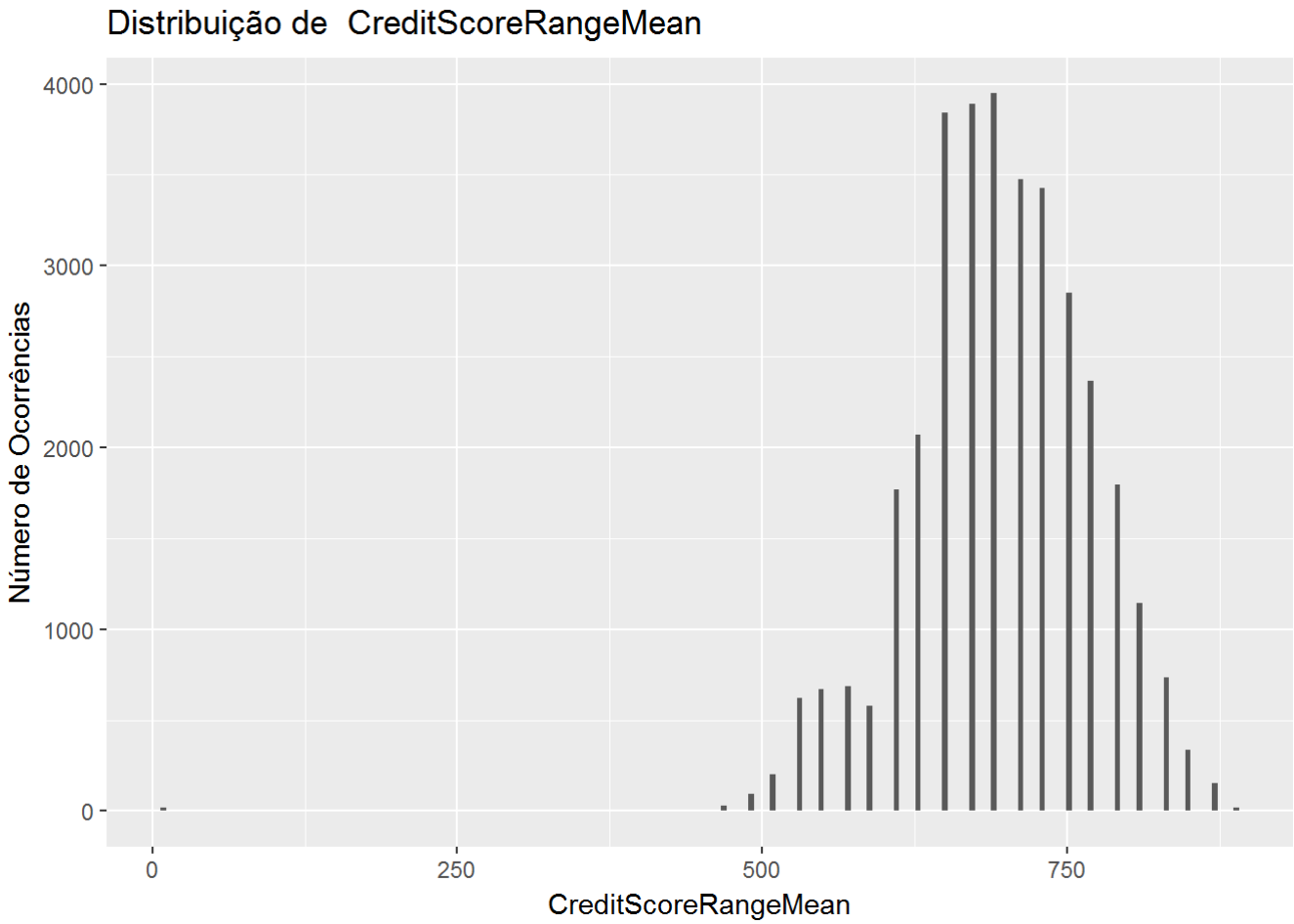

```
##
## [1] "> LoanOriginalAmount - Observações:  os valores de empréstimos apresentam picos de co
ncentração em determinados pontos do gráfico. Possivelmente valores cheios de empréstimos ten
dem a ser mais requisitados."
##
## [1] "> LoanOriginalAmount - Estatísticas: "
## [1] "- Valor Mín./Máx.:  1000 / 35000"
## [1] "- Média          :  6257.787"
## [1] "- Mediana        :  4800"
## [1] "- Desvio Padrão  :  5118.819"
```



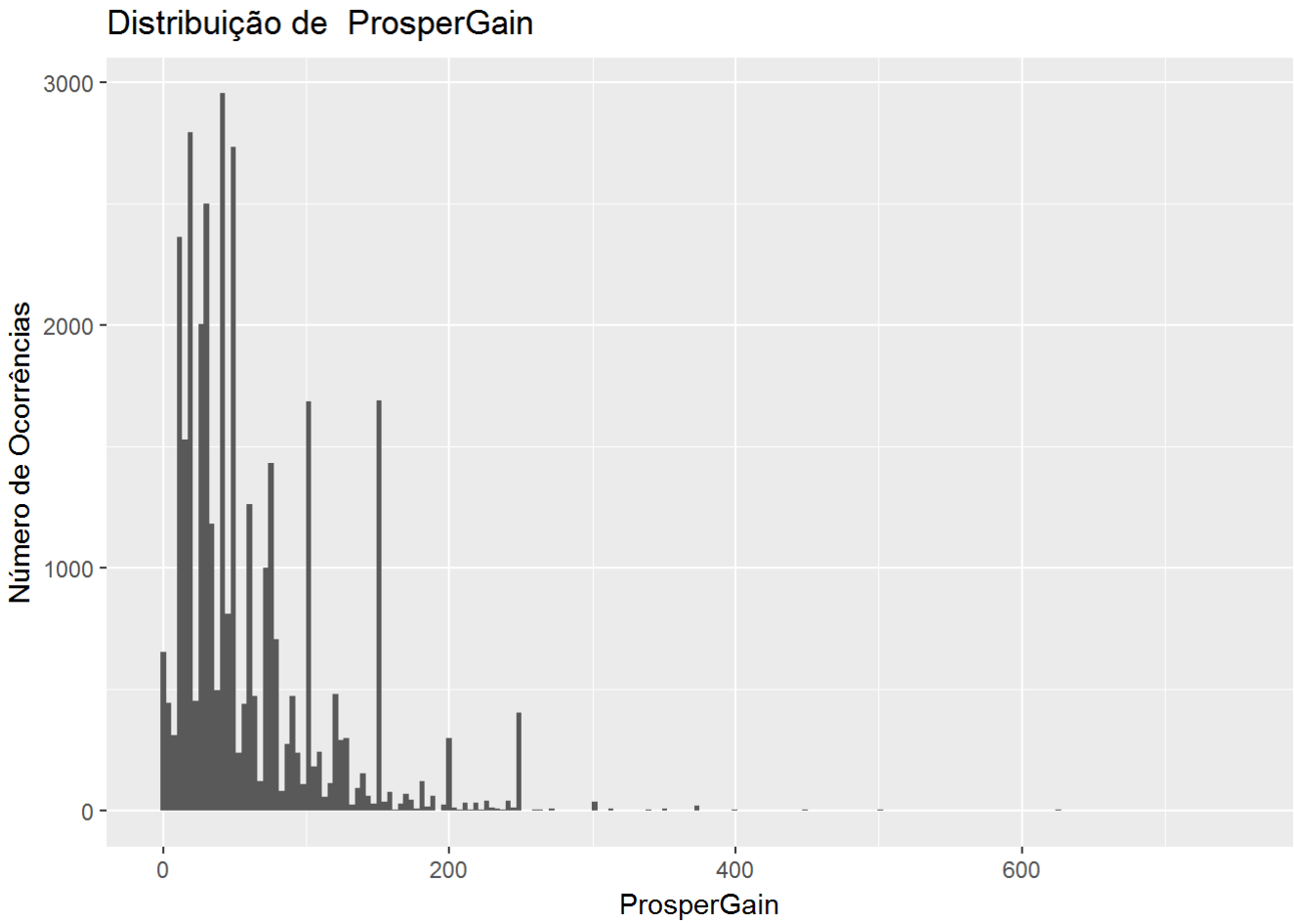

```
##
## [1] "> Investors - Observações:  Na distribuição do número de investidores, pode-se perceb
er uma distribuição do tipo exponencial negativa onde diminuem as ocorrências de empréstimos
 para maiores número de investidores."
##
## [1] "> Investors - Estatísticas: "
## [1] "- Valor Mín./Máx.:  1 / 1189"
## [1] "- Média          :  110.847"
## [1] "- Mediana        :  76"
## [1] "- Desvio Padrão  :  111.662"
```




```
##
## [1] "> DaysFromFirstRecCredLine - Observações:  Nota-se que a distribuição desta variável
 segue uma distribuição normal, apesar da presença de determinados picos de concentração, pos
sivelmente decorrentes de prazos específicos para encaminhamento dos empréstimos."
##
## [1] "> DaysFromFirstRecCredLine - Estatísticas: "
## [1] "- Valor Mín./Máx.:  1902 / 24505"
## [1] "- Média          :  8311.338"
## [1] "- Mediana        :  8005"
## [1] "- Desvio Padrão  :  2633.144"
```



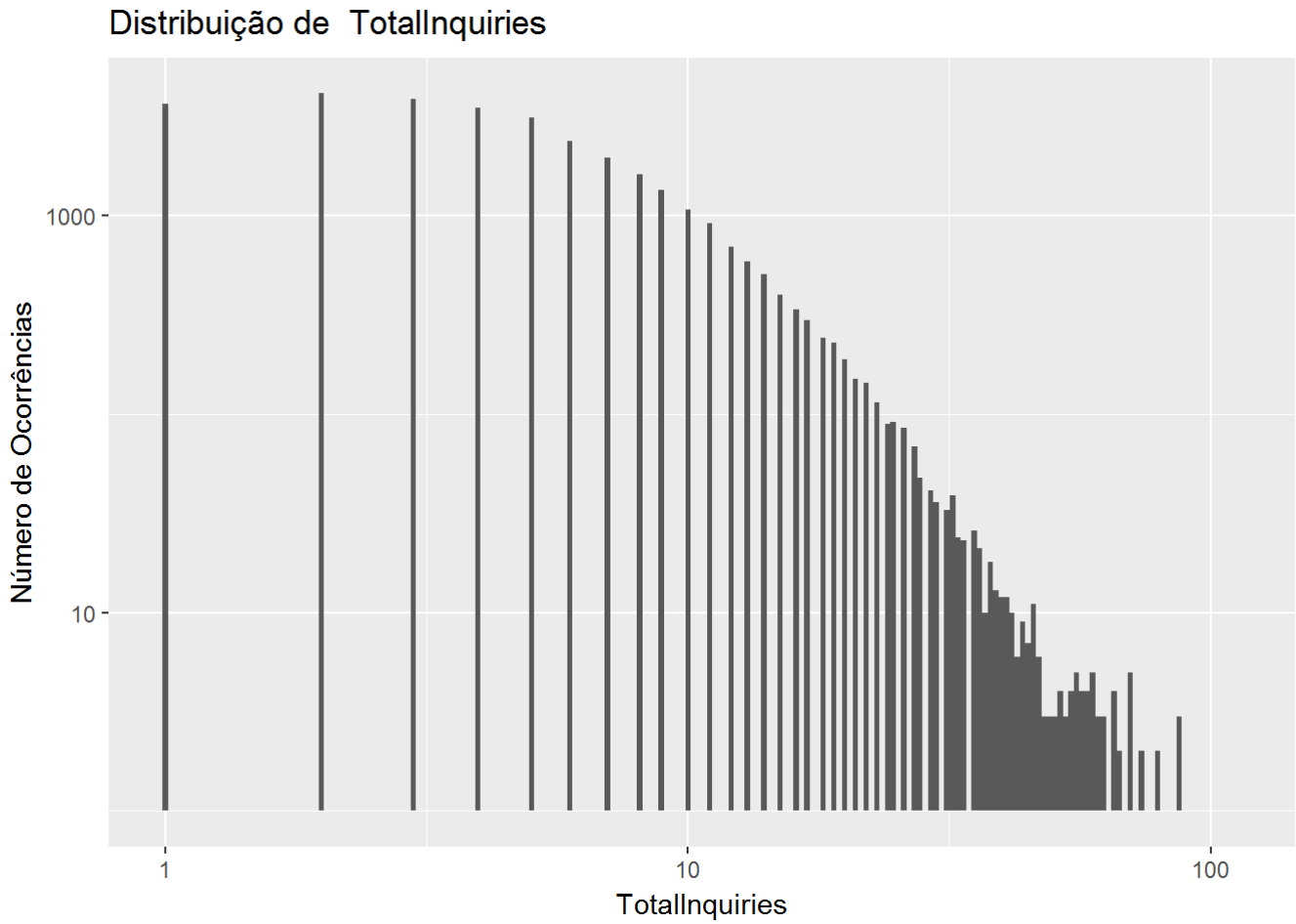

```
##
## [1] "> CreditScoreRangeMean - Observações:  Repare-se nesta variável uma distribuição que
 se assemelha à uma distribuição normal, concentrando os valores em torno da média."
##
## [1] "> CreditScoreRangeMean - Estatísticas: "
## [1] "- Valor Mín./Máx.:  9.5 / 889.5"
## [1] "- Média          :  694.256"
## [1] "- Mediana        :  689.5"
## [1] "- Desvio Padrão  :  73.419"
```



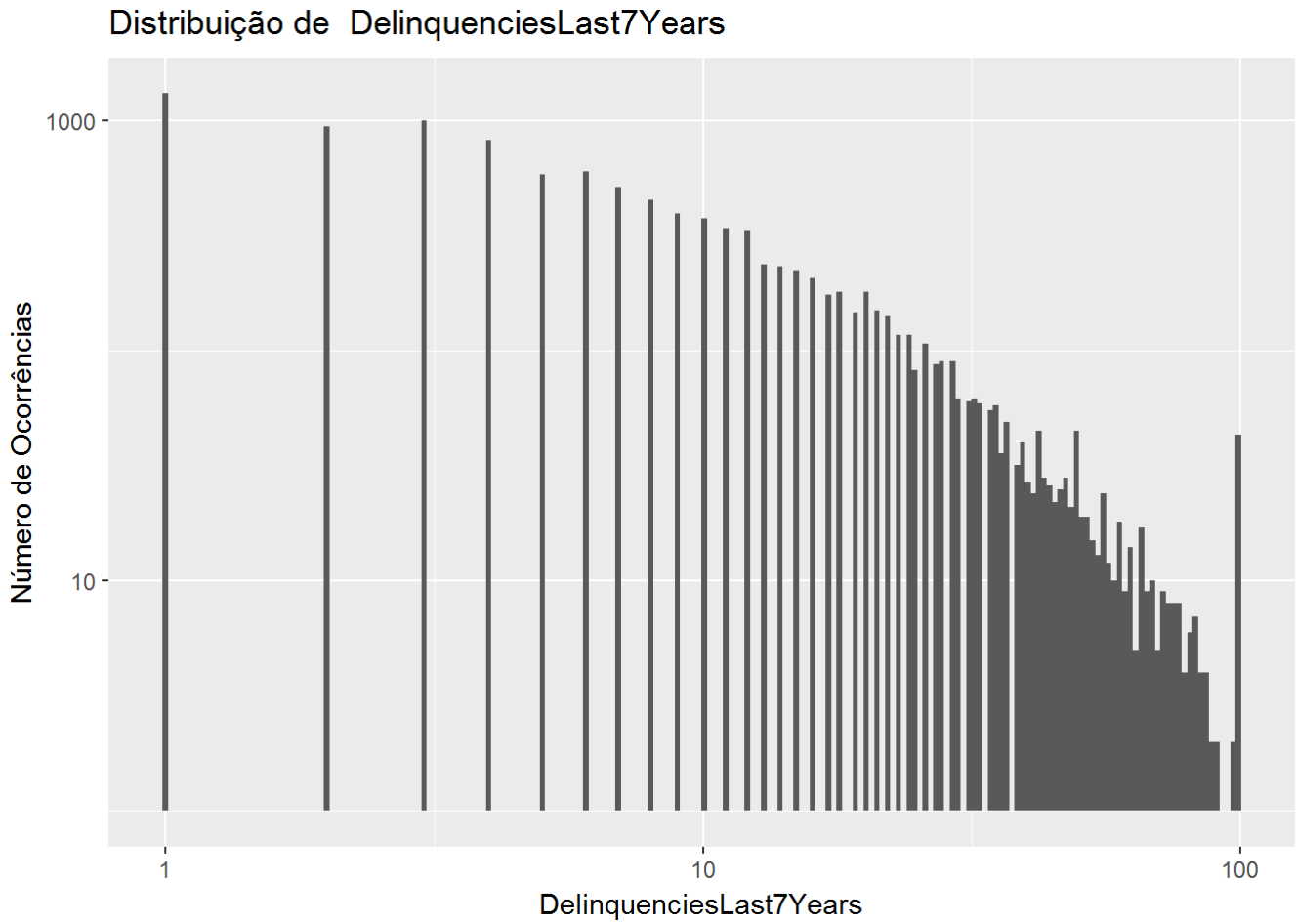

```
##
## [1] "> ProsperGain - Observações:  Nesta variável perce-se a dificuldade de se identificar
uma distribuição clara, pois o ganho obtido em cada empréstimo varia unicamente em função do
 valor do empréstimo em si."
##
## [1] "> ProsperGain - Estatísticas: "
## [1] "- Valor Mín./Máx.:  0 / 750"
## [1] "- Média          :  60.36"
## [1] "- Mediana        :  42"
## [1] "- Desvio Padrão  :  53.687"
```

Distribuições transformadas

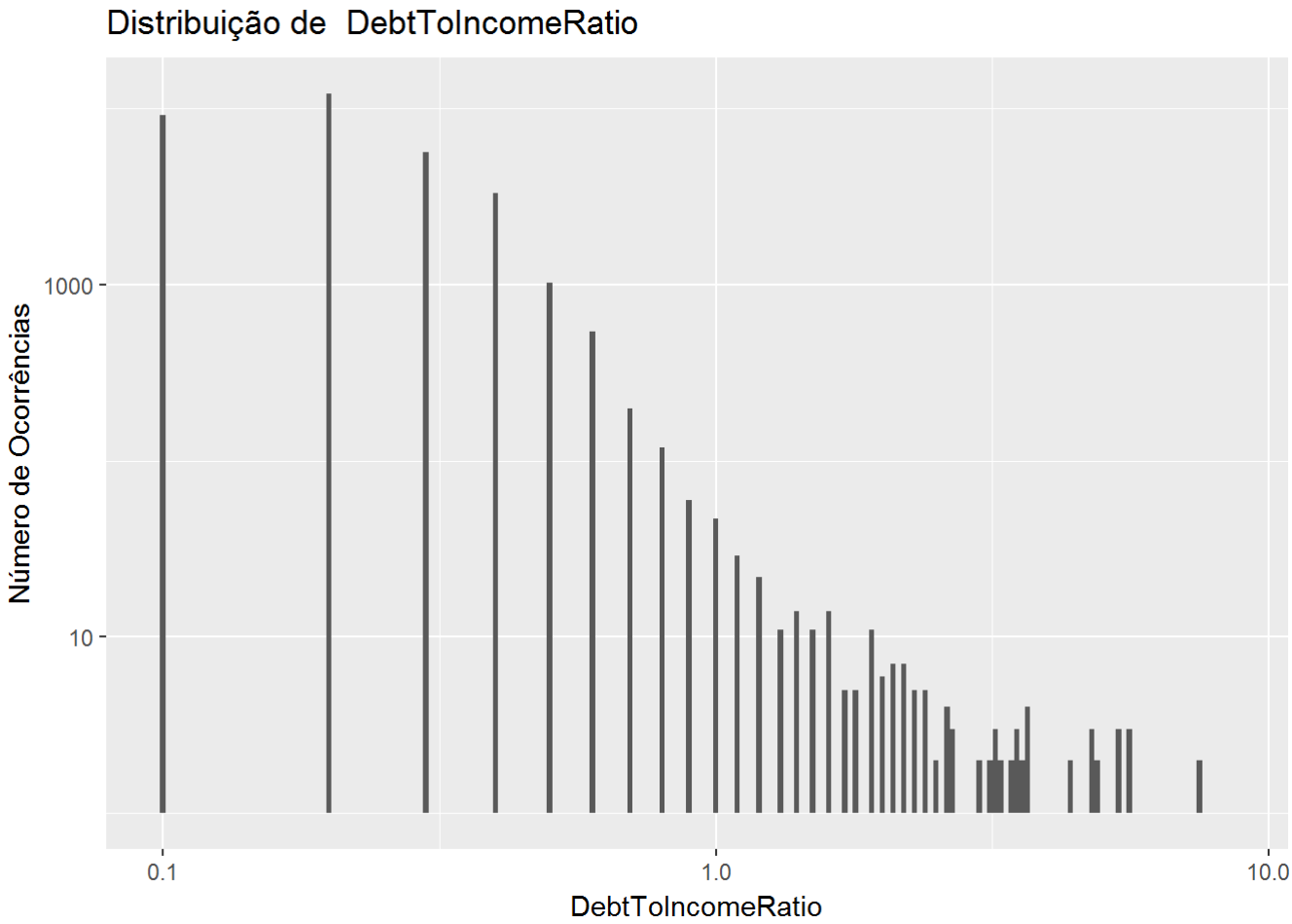
Nesta etapa serão replotados os gráficos do item anterior em outra escala para possibilitar uma melhor visualização da distribuição.




```
##
## [1] "> TotalInquiries - Observações:  Nesta variável podemos observar mais claramente após
transformar o eixo X na escala log10, uma distribuição que se assemelha à uma meia normal."
##
## [1] "> TotalInquiries - Estatísticas: "
## [1] "- Valor Mín./Máx.:  0 / 113"
## [1] "- Média          :  6.243"
## [1] "- Mediana        :  4"
## [1] "- Desvio Padrão  :  6.559"
```




```
##
## [1] "> DelinquenciesLast7Years - Observações:  Após transformarmos a escala do eixo Y para Log10, é possível agora observar o comportamento de uma distribuição exponencial negativa, em bora ocorra um pico de concentração de mutuários com apenas 1 ocorrência."
##
## [1] "> DelinquenciesLast7Years - Estatísticas: "
## [1] "- Valor Mín./Máx.:  0 / 99"
## [1] "- Média          :  4.04"
## [1] "- Mediana        :  0"
## [1] "- Desvio Padrão  :  9.974"
```




```
##
## [1] "> DebtToIncomeRatio - Observações:  Ao transformarmos a escala de ambox o eixos para
 log10, é possível observar uma distribuição que se assemelha à uma meia parábola"
##
## [1] "> DebtToIncomeRatio - Estatísticas: "
## [1] "- Valor Mín./Máx.:  0 / 8.63"
## [1] "- Média          :  0.23"
## [1] "- Mediana        :  0.19"
## [1] "- Desvio Padrão  :  0.248"
```

3.3. Conclusões

Características do Conjunto de Dados

- O conjunto de dados é constituído por apenas uma tabela extraída de um arquivo texto (csv) contendo ao todo NN variáveis (colunas).
- Em função da complexidade do conjunto de dados, foi criada uma lista de variáveis de interesse, que foi sendo atualizada durante todas as etapas do projeto, de acordo com os resultados obtidos nas análises.
- Verificou-se que o objetivo principal do conjunto de dados é trazer informações relativas aos empréstimos de uma determinada instituição financeira. Trata-se pois de uma tabela contendo em sua essência dados analíticos que registram a intermediação entre a pessoa que toma emprestado (Borrower) e as pessoas que entram nesse empréstimo como investidores (Investors).

- Além de variáveis que discriminam as condições de empréstimo e investimentos, existem também variáveis que são atualizadas de acordo com a evolução dos pagamentos mensais destes empréstimos, como por exemplo: total já pago, total em atraso, etc. Estas variáveis contém em sua essência as atualizações dos empréstimos enquanto abertos.
- Na seleção das variáveis de interesse, procuramos reunir as informações que de acordo com o assunto especificado no item 2.4., à fim de que possamos fazer essas distinções.

Atributos de interesse mantidos do conjunto de dados:

Na etapa 2.4. selecionamos as variáveis de interesse que podem ser classificadas nos seguintes grupos:

- informações relativas ao empréstimo (li=loan info)
- condições do Empréstimo (lc=loan conditions)
- informações do Mutuário (bi=borrower info)
- análise de Crédito (ci=credit info)
- situação do Empréstimo (lr=loan results)

A relação final das variáveis de interesse selecionadas encontram-se listadas no item 2.4.5.

Atributos descartados do conjunto de dados

Durante as etapas 2.4.2. e 2.4.4. foram descartadas as variáveis que:

- Variáveis que apresentaram um percentual de registros inválidos superior à 10%.
- Variáveis secundárias, onde não se reconheceu de que forma estas poderiam agregar valor às análises.
- Variáveis substituídas por uma variável similar criada para facilitar o processamento das análises.
- Variáveis que apresentaram pouca ou quase nenhuma distribuição observada nos gráficos.

Atributos auxiliares sugeridos

Verificou-se após uma avaliação das variáveis de interesse, que os atributos abaixo sugeridos poderiam complementar positivamente o conjunto de dados:

- idade do mutuário
- renda total familiar
- número de dependentes
- se mora em imóvel próprio ou alugado
- valor do aluguel (quando não proprietário de imóvel)

Atributos novos criadas à partir do conjunto de dados

Durante a etapa 2.3., para facilitar as análises, foram criados novas variáveis de acordo conforme segue:

- variáveis numéricas convertidas para fatorial
- variáveis contendo o valor médio entre duas variáveis
- variáveis do tipo data contendo o formato datetime adequado

As novas variáveis criadas foram adicicionadas também à lista de variáveis de interesse conforme já listado previamente no item 2.4.5

Atributos inconsistentes e ações corretivas adotadas

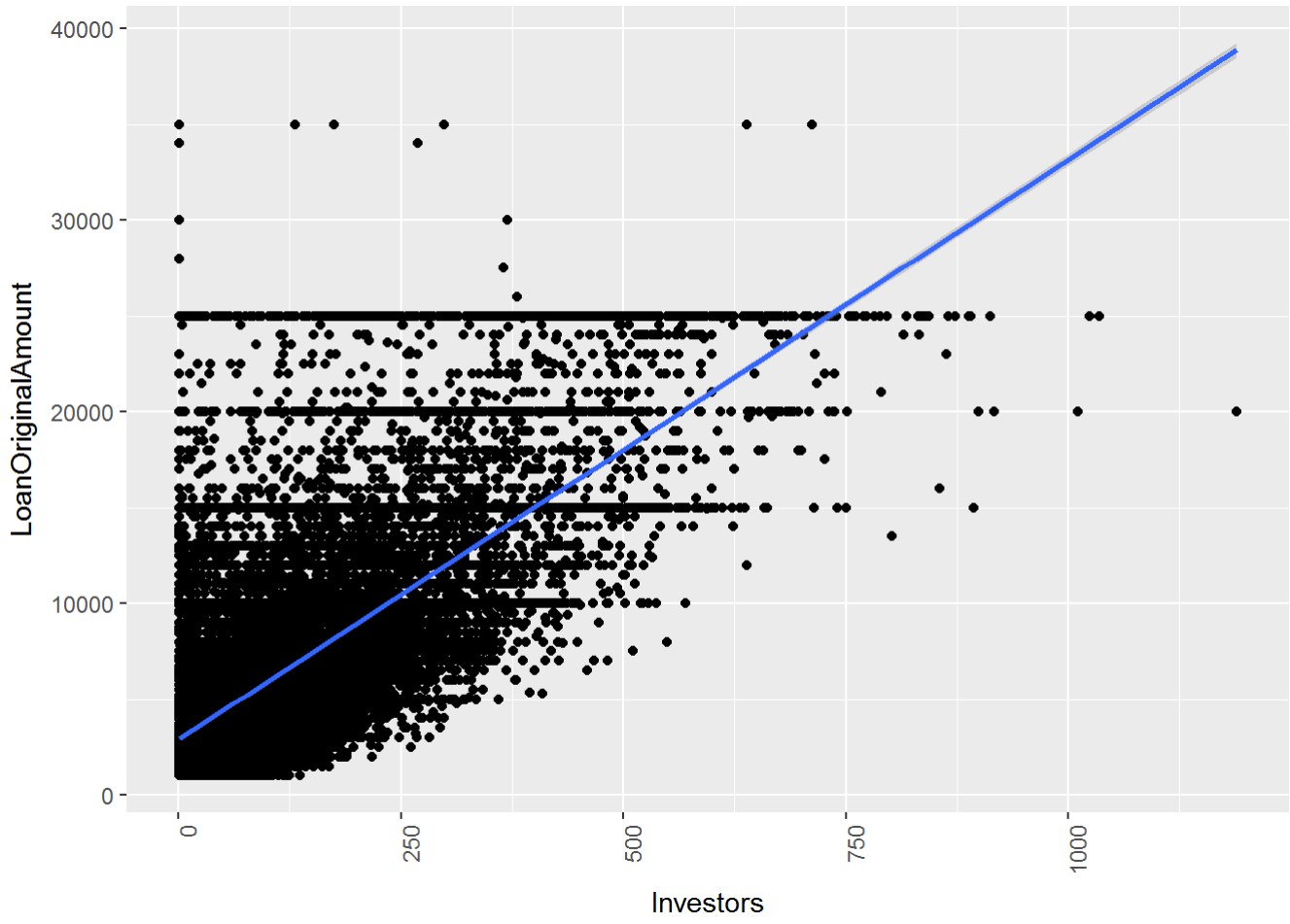
O tratamento de dados foi realizado onde foram tratados as seguintes situações:

- remoção de registros duplicados (871 registros)
- reordenação de campos

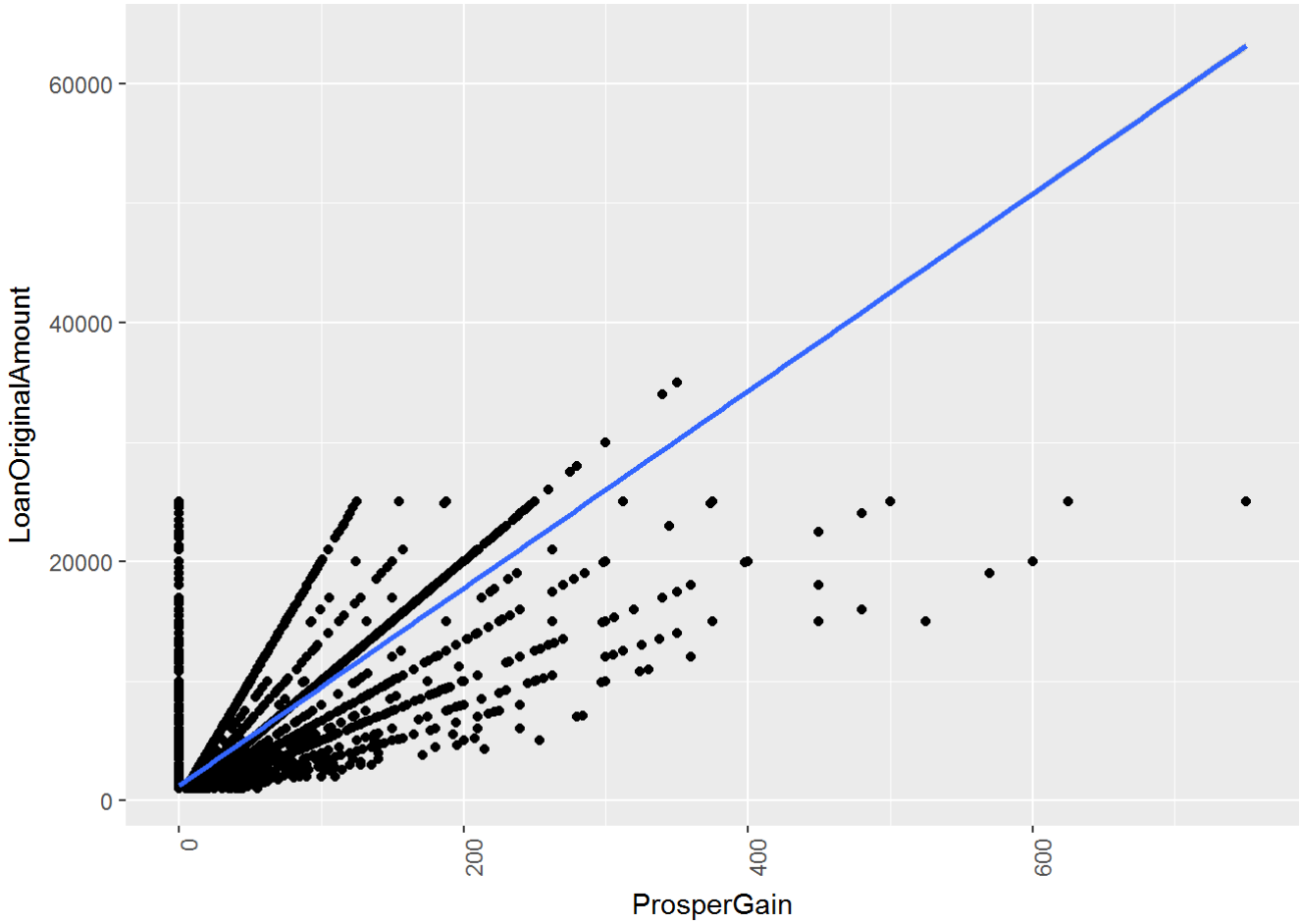
4: ANÁLISE E SESSÃO DE GRÁFICOS BIVARIADOS

4.1. Gráficos de Dispersão entre as variáveis de Resultado numéricas investigadas e demais variáveis (também numéricas)

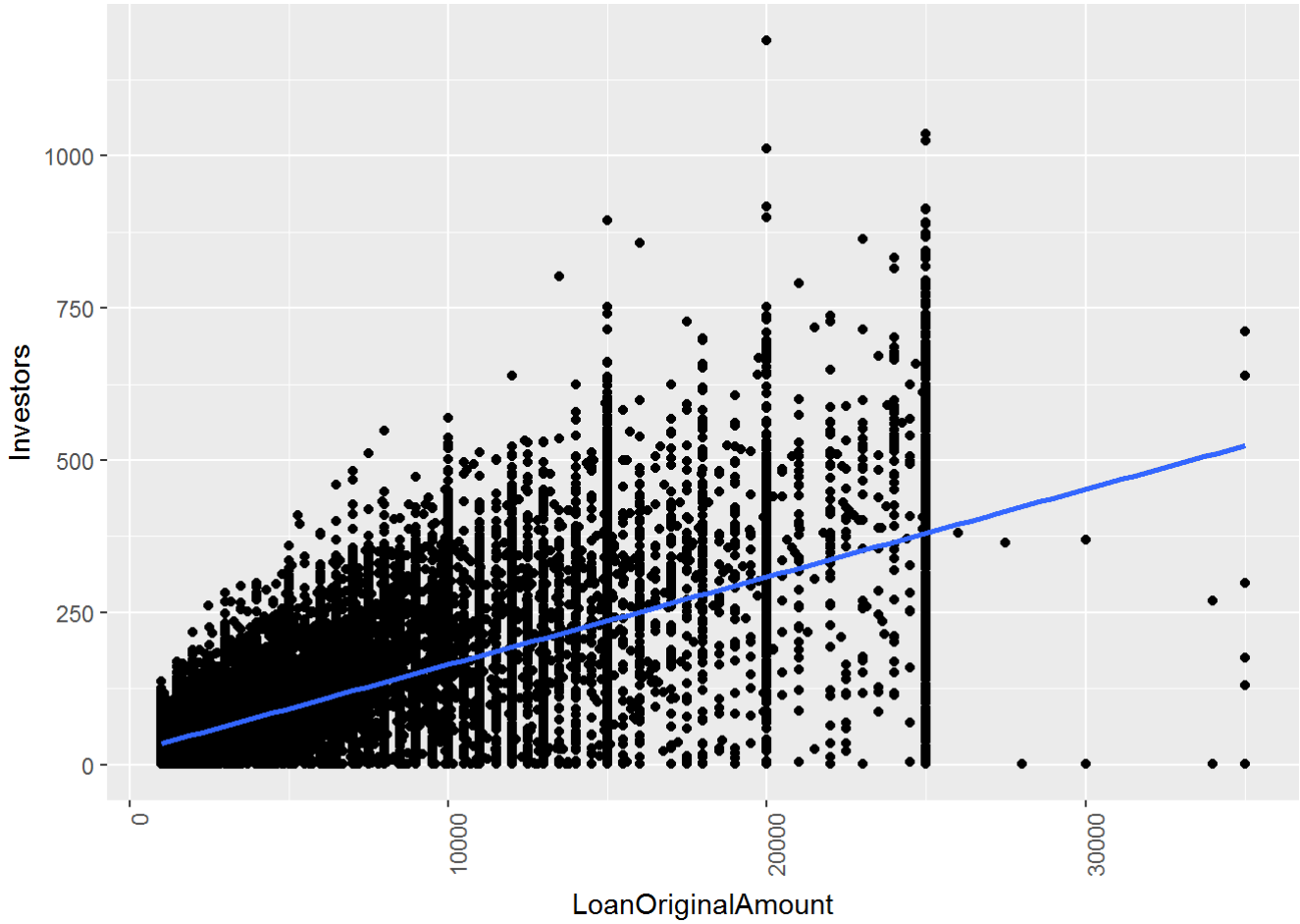
Nesta etapa foram selecionadas entre as variáveis numéricas do conjunto de dados, aquelas que possuem maior relevância com o tema investigado. Para cada variável selecionada foram calculadas as correlações com todas as demais variáveis numéricas e estas expostas por meio de gráfico de dispersão quando as correlações apresentaram um coeficiente maior ou igual a 0,5.



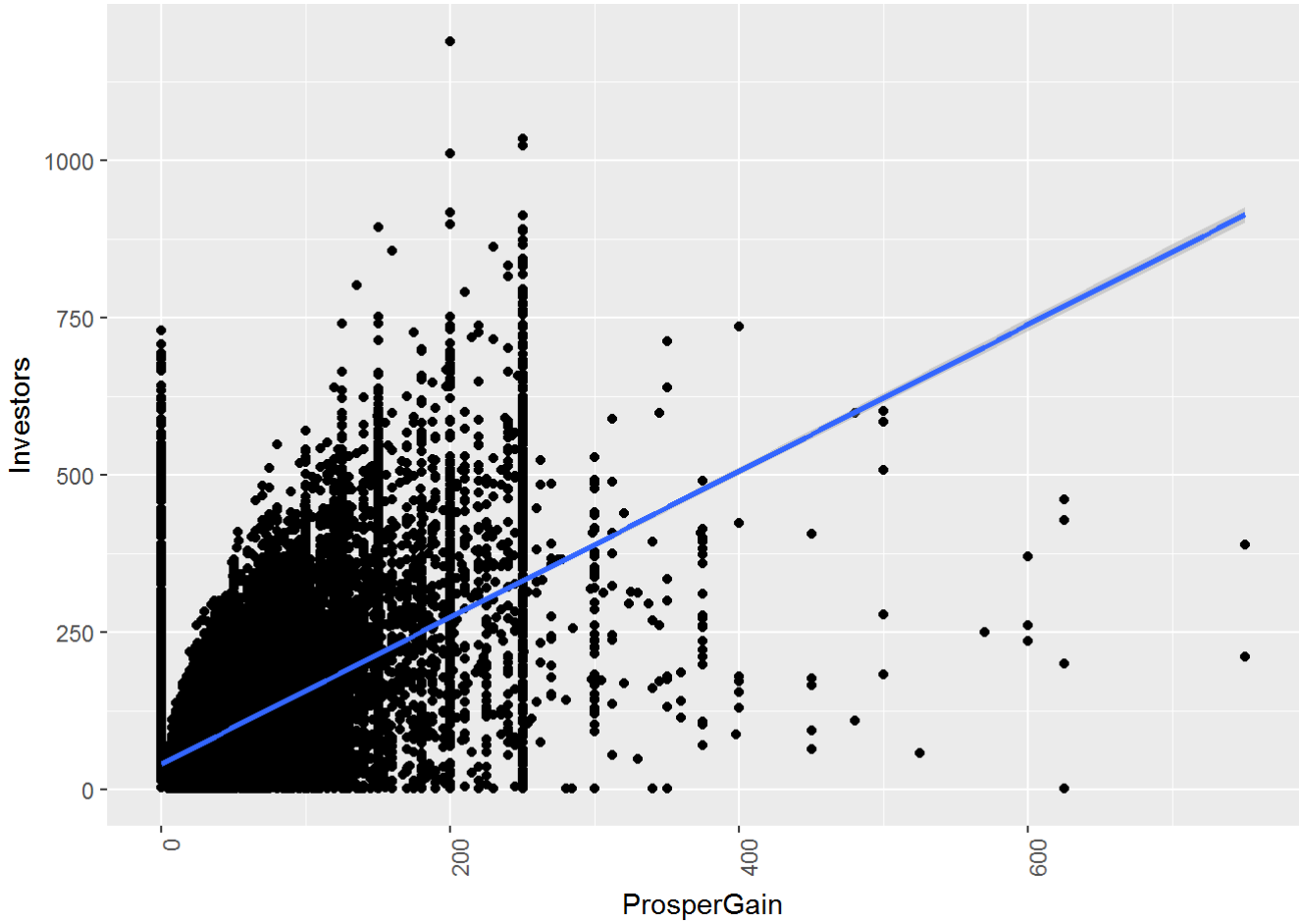
```
## [1] "> Estatísticas de Correlação entre:  LoanOriginalAmount  x  Investors"
## [1] "- Método           :  Pearson's product-moment correlation"
## [1] "- p.value          :  0.66"
## [1] "- Interv.Confiança :  0.654  /  0.666"
##
## [1] "- Observações:  Verifica-se por meio do coeficiente uma correlação moderada, onde é p
ossível observar no gráfico de dispersão a tendência a ser necessário mais investidores para
 completar o valor total do empréstimo requisitado à medida que aumenta o valor do empréstim
o."
```




```
## [1] "> Estatísticas de Correlação entre:  LoanOriginalAmount  x  ProsperGain"
## [1] "- Método           :  Pearson's product-moment correlation"
## [1] "- p.value          :  0.865"
## [1] "- Interv.Confiança :  0.863  /  0.868"
##
## [1] "- Observações:  Na correlação com o ganho da financeira, observa-se uma correlação fo
rte, sendo possível observar várias retas, caracterizando a forte correspondência linear."
```



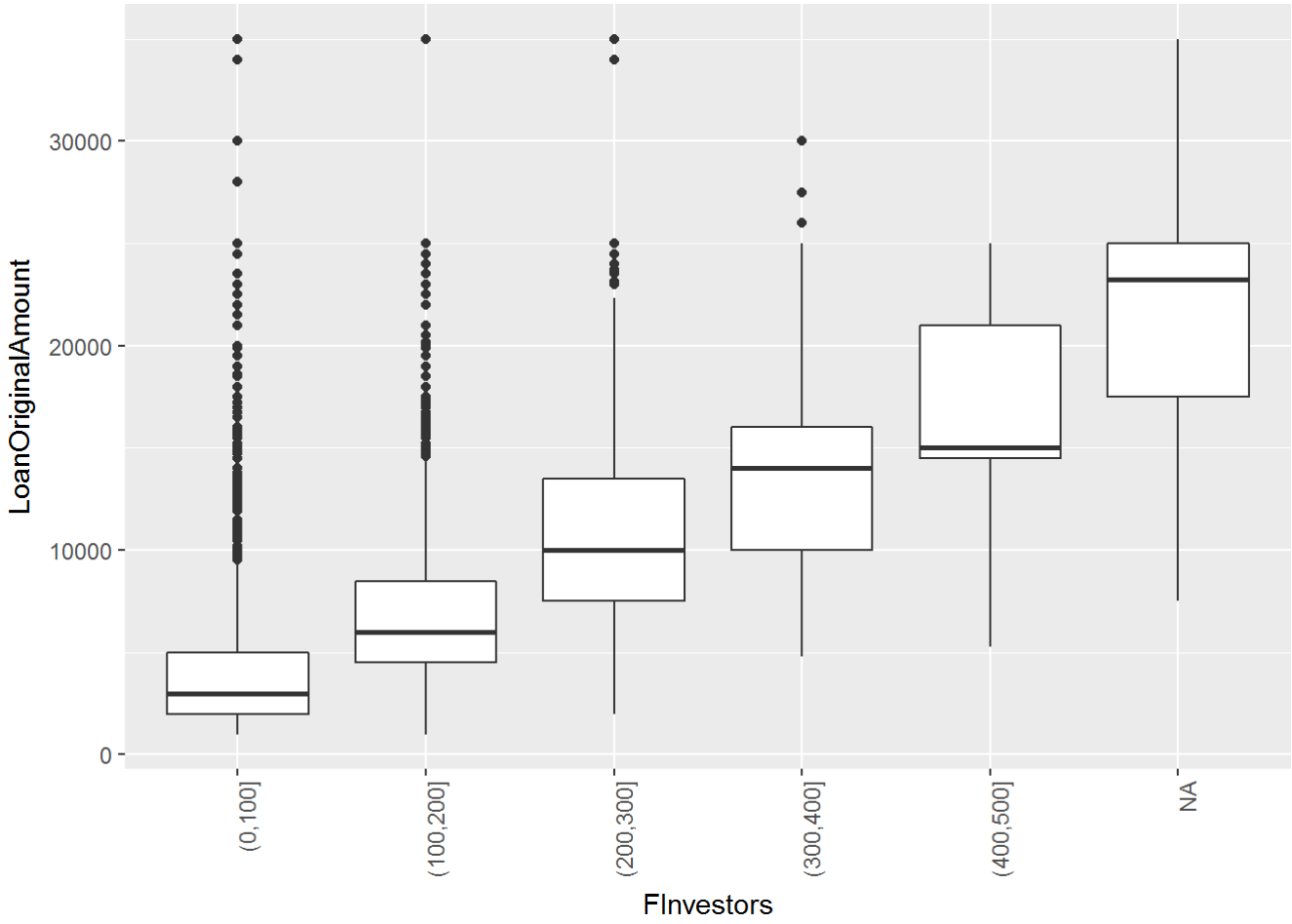

```
## [1] "> Estatísticas de Correlação entre:  Investors  x  LoanOriginalAmount"
## [1] "- Método           :  Pearson's product-moment correlation"
## [1] "- p.value          :  0.66"
## [1] "- Interv.Confiança :  0.654  /  0.666"
##
## [1] "- Observações:  Neste invertem-se os eixos do gráfico apresentado anteriormente. Apes
ar disso, é possível observar novamente a tendência natural de precisar de mais investidores
 para valores mais altos de empréstimos."
```



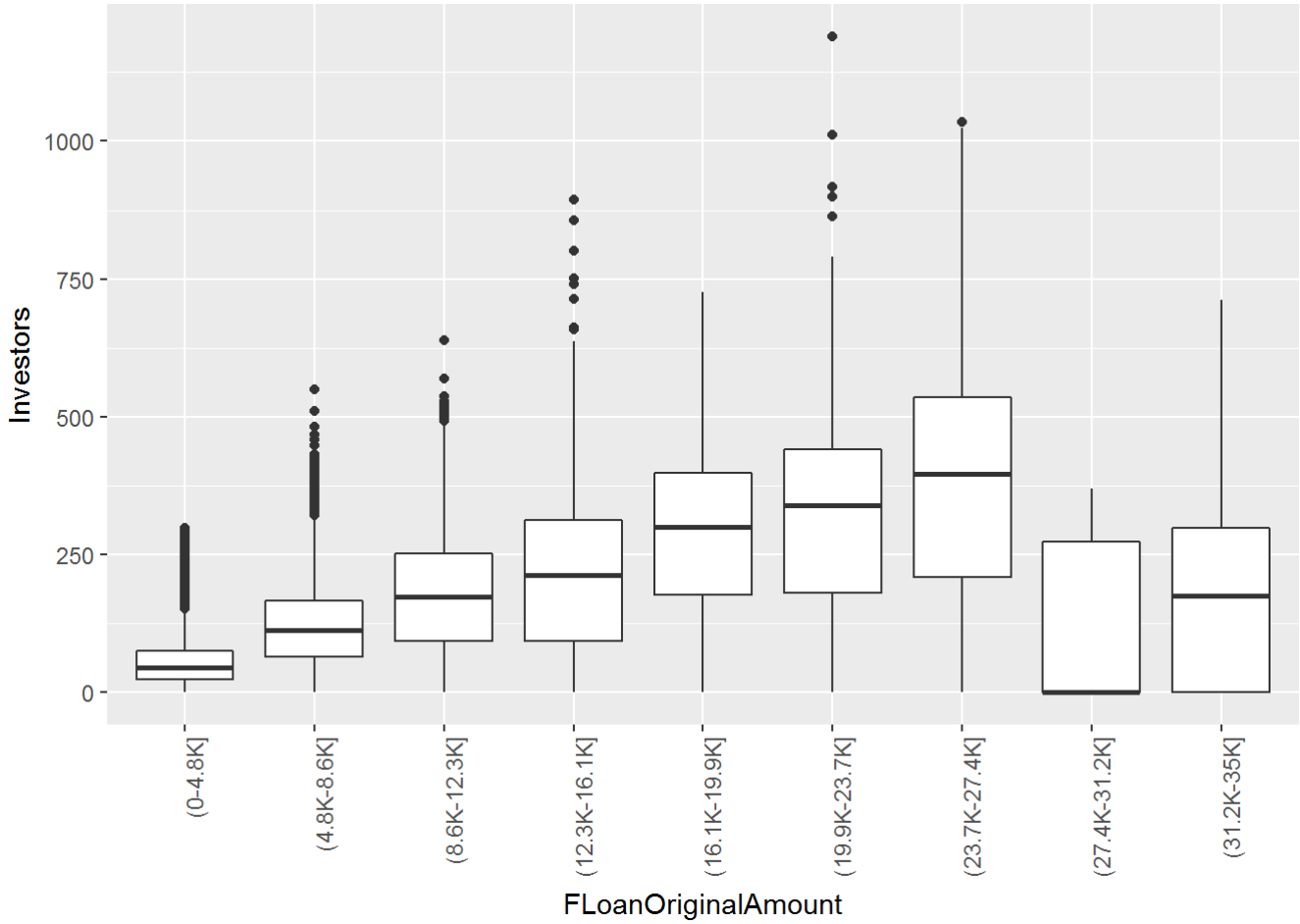

```
## [1] "> Estatísticas de Correlação entre:  Investors  x  ProsperGain"
## [1] "- Método           :  Pearson's product-moment correlation"
## [1] "- p.value          :  0.56"
## [1] "- Interv.Confiança :  0.552  /  0.567"
##
## [1] "- Observações:  A correlação entre as variáveis pode ser considerada fraca, não sendo
possível identificar qual a tendência de relação entre elas."
```

4.2. Gráficos BoxPlot entre as variáveis de Resultado numéricas investigadas e demais variáveis (fatoriais)

Nesta etapa foram selecionadas entre as variáveis fatoriais do conjunto de dados, aquelas que possuem maior relevância com o tema investigado. Para cada variável selecionada foram calculadas as correlações com todas as demais variáveis numéricas e estas expostas por meio de gráfico de dispersão quando as correlações apresentaram um coeficiente maior ou igual a 0,4.




```
## [1] "> Estatísticas de Correlação entre:  FInvestors  x  LoanOriginalAmount"
## [1] "- Método           : eta squared"
## [1] "- p.value          :  0.614"
##   FInvestors    x.mean      x.sd     x.min     x.max
## 1    (0,100]  4195.464  3522.026  1000.000 35000.000
## 2  (100,200]  7029.646  3944.118  1000.000 35000.000
## 3  (200,300] 10535.494  4539.260  2000.000 35000.000
## 4  (300,400] 13932.391  4916.040  4800.000 30000.000
## 5  (400,500] 17225.189  5027.298  5300.000 25000.000
##
## [1] "- Observações:  Confirmando os resultados das correlações entre as versões numéricas
 das variáveis de interesse selecionadas, observa-se novamente a tendência de ocorrer um núme
ro maior de investidores para valores de empréstimos mais elevados."
```




```
## [1] "> Estatísticas de Correlação entre:  FLoanOriginalAmount  x  Investors"
## [1] "- Método           : eta squared"
## [1] "- p.value          :  0.655"
##   FLoanOriginalAmount     x.mean       x.sd      x.min      x.max
## 1            (0-4.8K]   54.28661   41.36403    1.00000  299.00000
## 2         (4.8K-8.6K]  120.40812   73.72687    1.00000  549.00000
## 3        (8.6K-12.3K]  177.47500  110.67847    1.00000  639.00000
## 4       (12.3K-16.1K]  213.41093  147.16760    1.00000  894.00000
## 5       (16.1K-19.9K]  292.33149  163.55544    1.00000  727.00000
## 6       (19.9K-23.7K]  322.74906  193.26996    1.00000 1189.00000
## 7       (23.7K-27.4K]  376.82267  216.56673    1.00000 1035.00000
## 8       (27.4K-31.2K]  123.16667  189.26639    1.00000  370.00000
## 9         (31.2K-35K]  247.55556  267.68970    1.00000  712.00000
##
## [1] "- Observações:  Confirmando os resultados das correlações entre as versões numéricas
 das variáveis de interesse selecionadas, observa-se novamente a tendência de ocorrer um núme
ro maior de investidores para valores de empréstimos mais elevados."
```

4.3. Conclusões

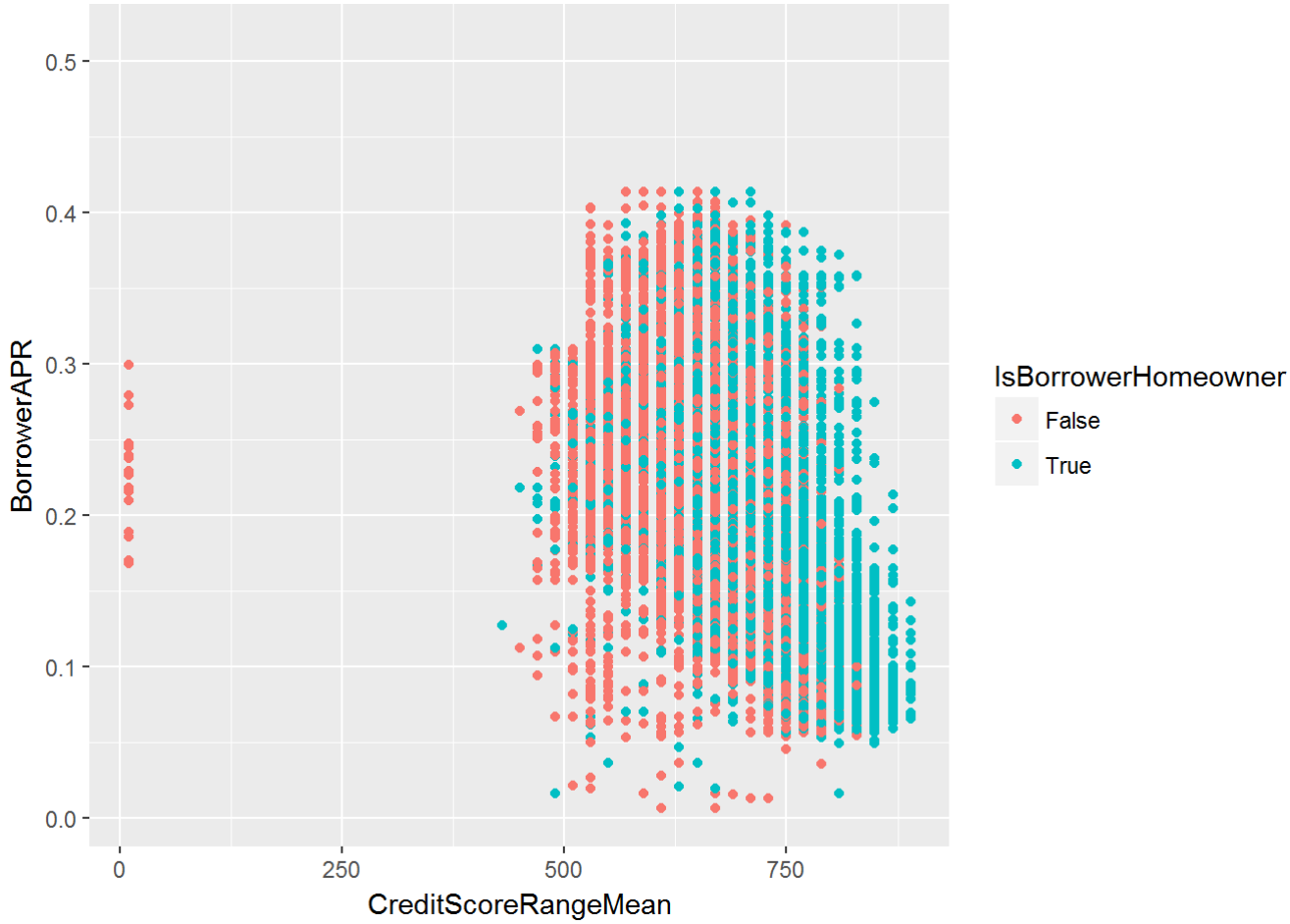
Verificou-se nesta etapa que as únicas correlações fortes apresentadas ocorrem entre as variáveis LoanOriginalAmount (valor do empréstimo) e Investors (número de investidores). Tanto nos gráficos de dispersão quanto nos boxplots, verificou-se a tendência de valores de empréstimos mais elevados precisam de mais investidores para completar o valor total.

5: ANÁLISE E SESSÃO DE GRÁFICOS MULTIVARIADOS:

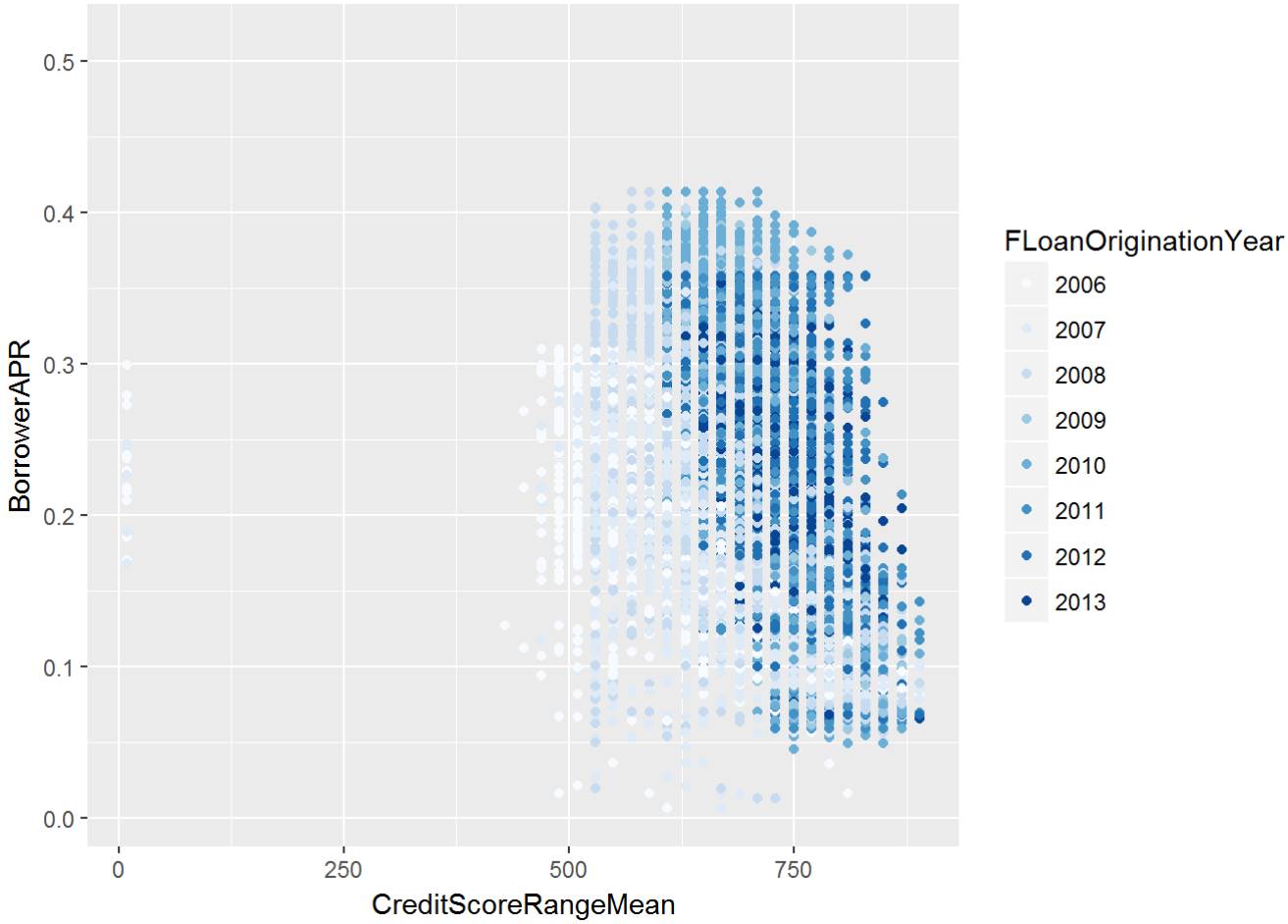
5.1. Sessão de Gráficos Multivariados com as variáveis BorrowerAPR e CreditScoreRangeMean:

Nesta etapa foi selecionada a variável CreditScoreRangeMean enquanto indicador da capacidade de crédito do mutuário. O objetivo é verificar o comportamento deste indicador em relação às taxas de juros anuais aprovadas (BorrowerAPR) agrupando nas categorias de cada variável fatorial disponível no conjunto de dados.

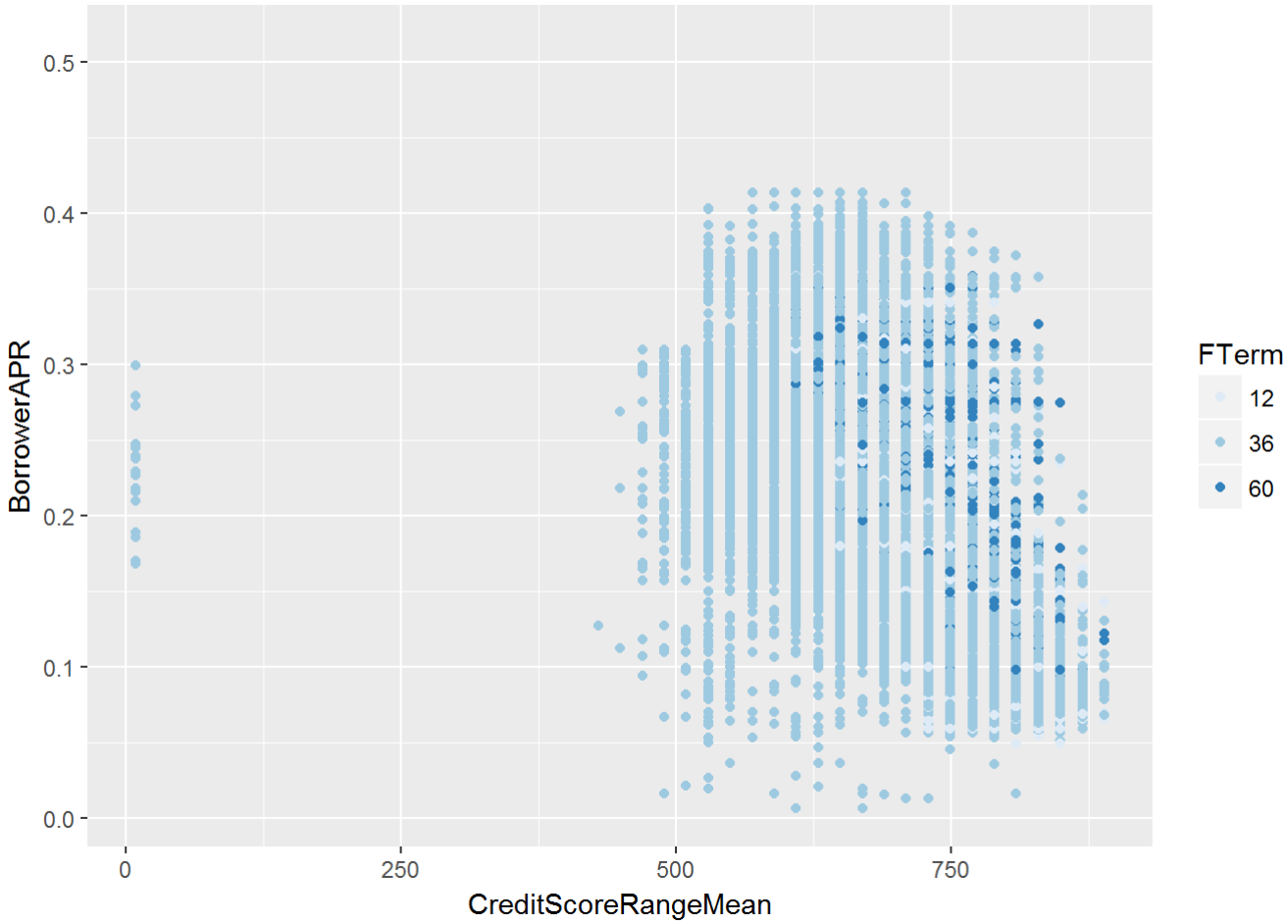

```
##
## [1] "- Observações:  Observa-se que a variável EmploymentStatus tem certa influência sobre
a variável CreditScoreRangeMean, pois mutuários com o status de Not Emplyed tendem a receber
 scores baixos. A taxa de juros aprovada (BorrowerAPR) não apresentou variância em relação ao
crédito do mutuário e à variável EmploymentStatus."
```



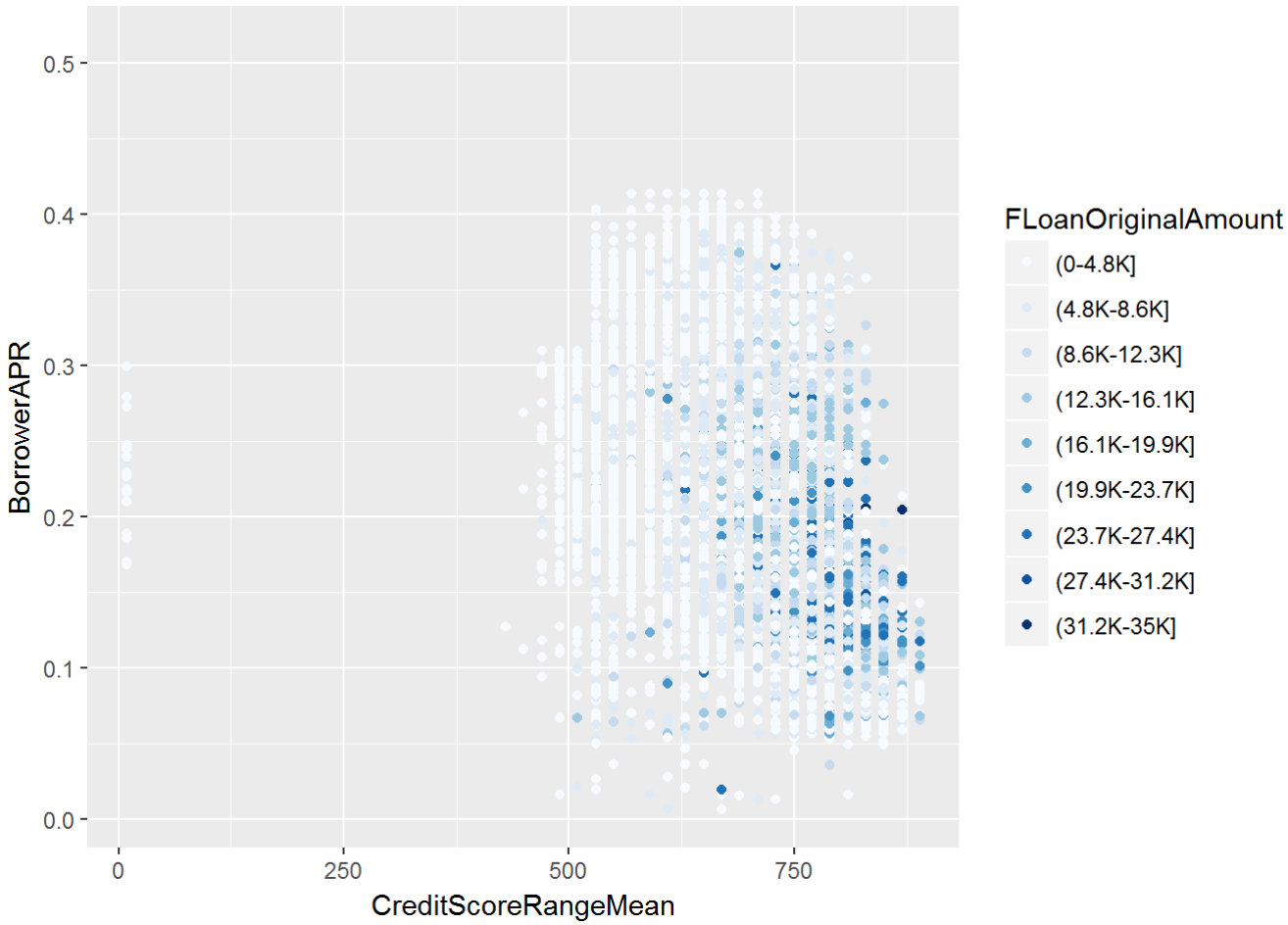

```
##
## [1] "- Observações:  Observa-se que a variável IsBorrowerHomeowner tem certa influência so
bre a variável CreditScoreRangeMean, pois mutuários que não moram em casa própria tendem a re
ceber scores baixos. A taxa de juros aprovada (BorrowerAPR) não apresentou variância em relaç
ão ao crédito do mutuário e à variável IsBorrowerHomeowner."
```



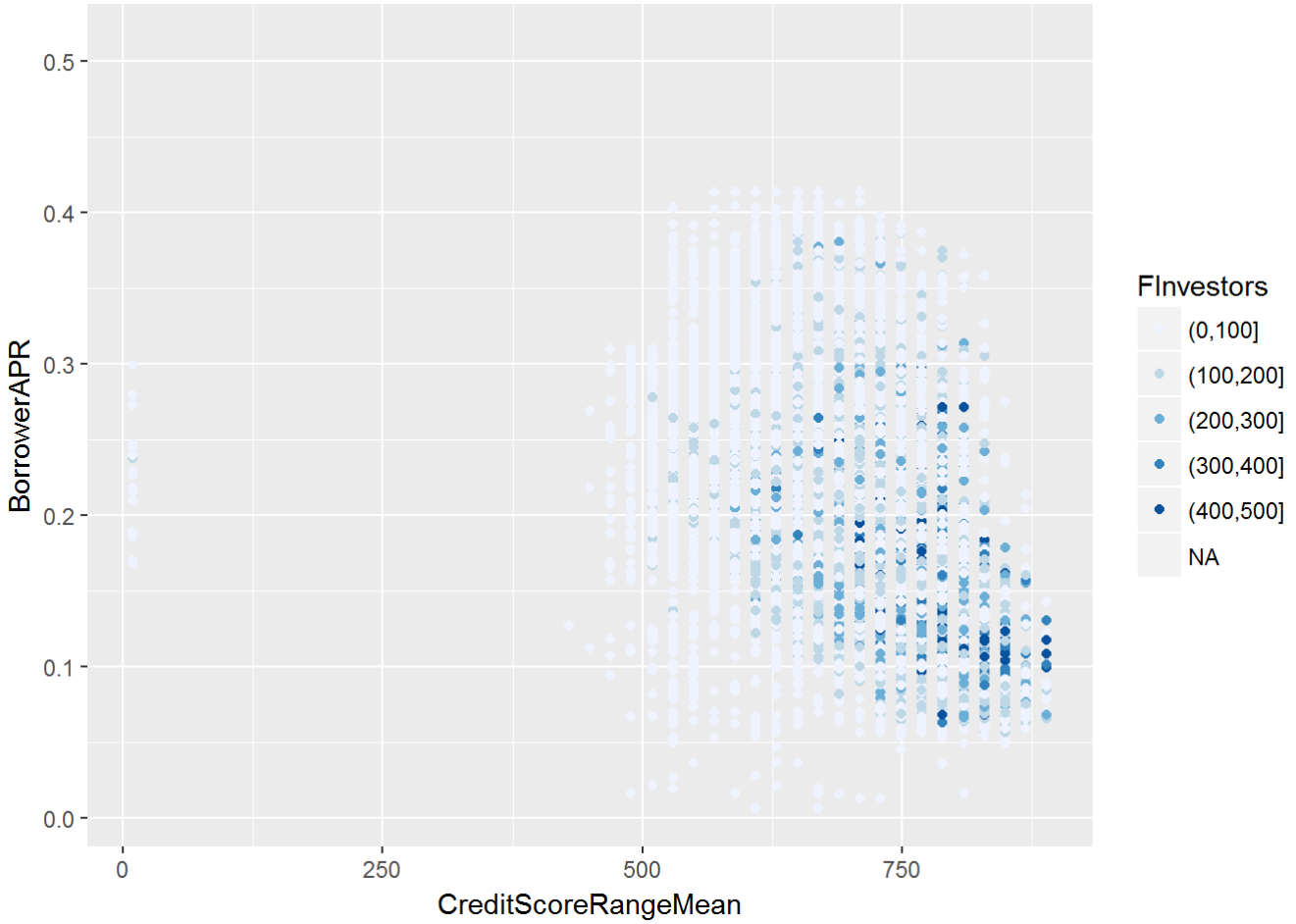

```
##
## [1] "- Observações:  Pecebe-se um aumento gradual no score de crédito ao longo dos anos."
```



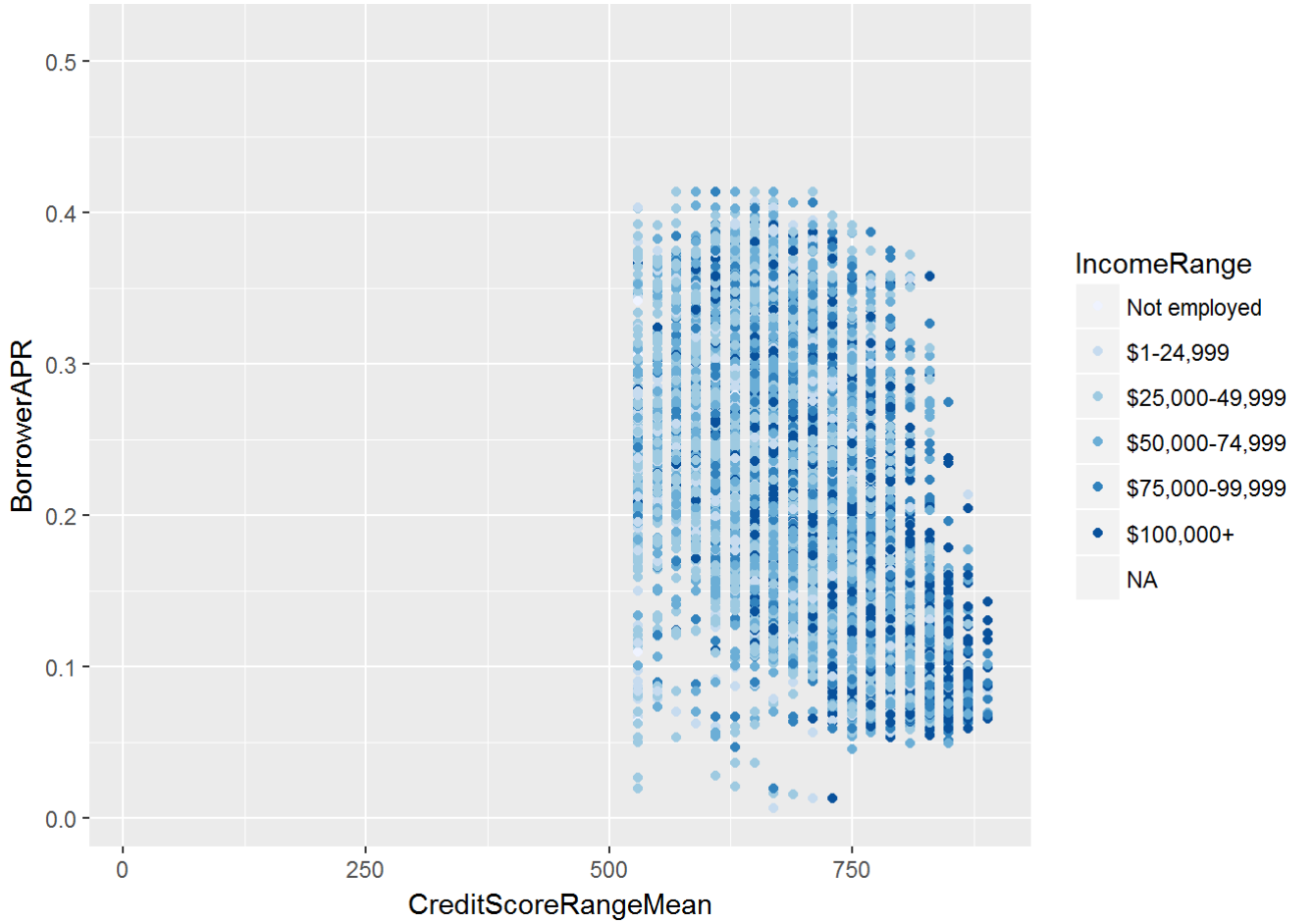
```
##
## [1] "- Observações:  Em termos de prazo de pagamento do empréstimo, destaca-se que a grand
e maioria deles ocorrem no prazo de 36 meses. A taxa de juros aprovada não (BorrowerAPR) apre
sentou variância em relação ao crédito do mutuário e à variável FTerm"
```




```
##
## [1] "- Observações:  Em relação à faixa do valor dos empréstimos, observa-se uma certa ten
dência onde empréstimos com valores altos tendem a ser concedidos à mutuários com score de cr
édito mais altos. A taxa de juros aprovada (BorrowerAPR) não apresentou variância em relação
 ao crédito do mutuário e à variável FLoanOriginalAmount"
```

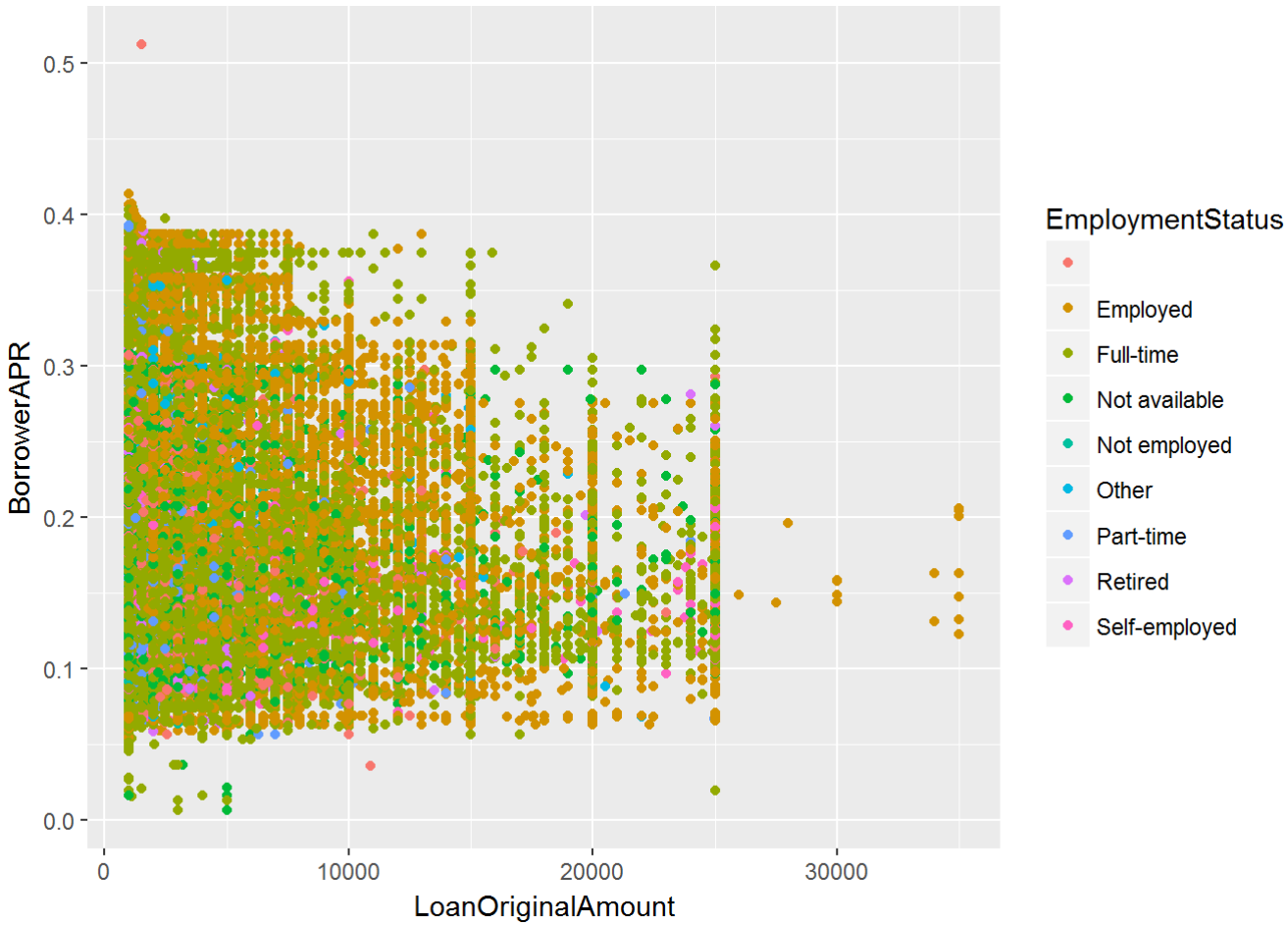



```
##
## [1] "- Observações:  Não percebe-se na variável FInvestors alguma tendência em relação à i
nfluência da média de salário sobre o score de crédito do mutuário. A taxa de juros aprovada
 (BorrowerAPR) não apresentou variância em relação ao crédito do mutuário e à variável FInves
tors"
```

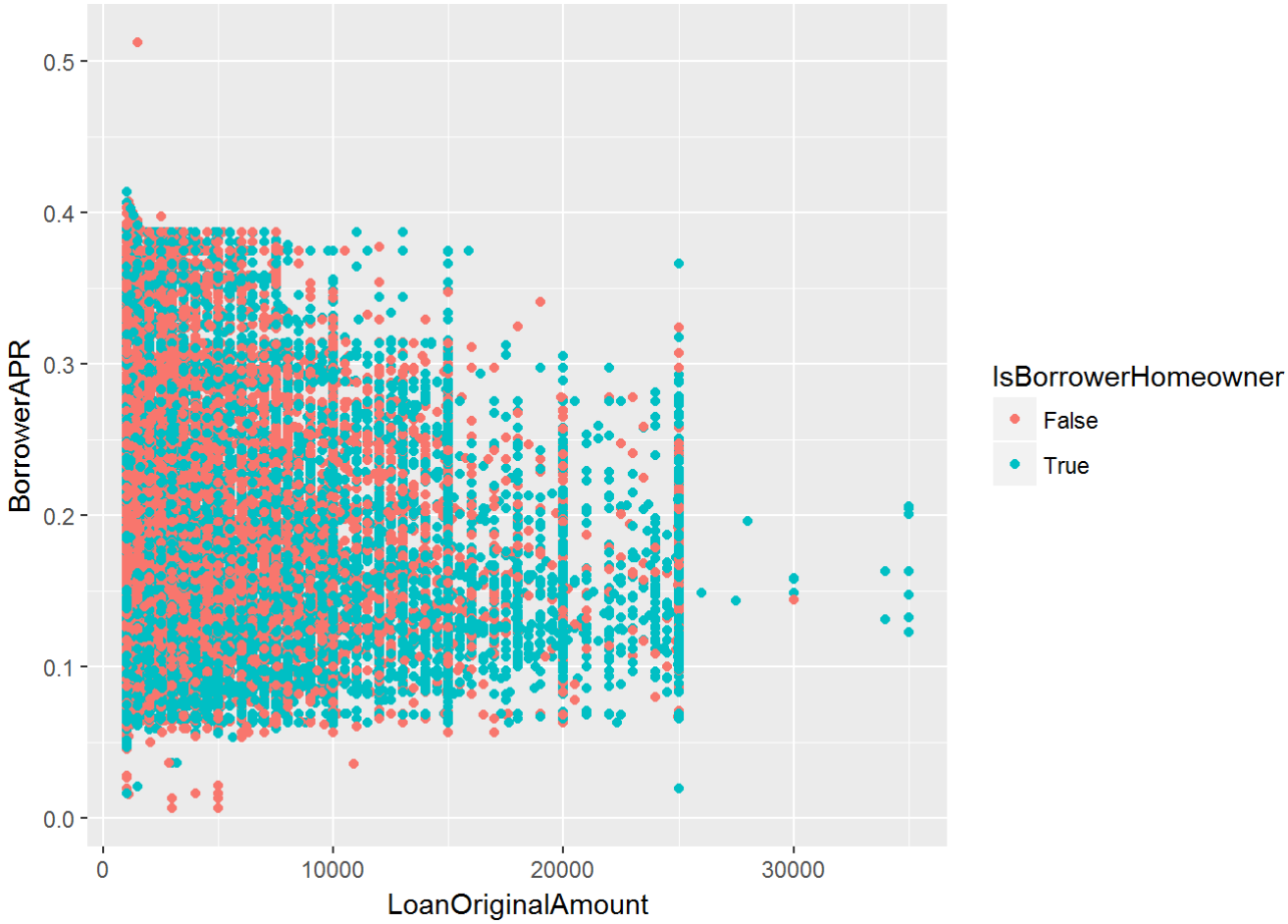



```
##
## [1] "- Observações:  Percebe-se que o score de crédito do mutuário tende a ser maior quant
o maior sua faixa de salário. A taxa de juros aprovada (BorrowerAPR) não apresentou variância
em relação ao crédito do mutuário e à variável IncomeRange"
```

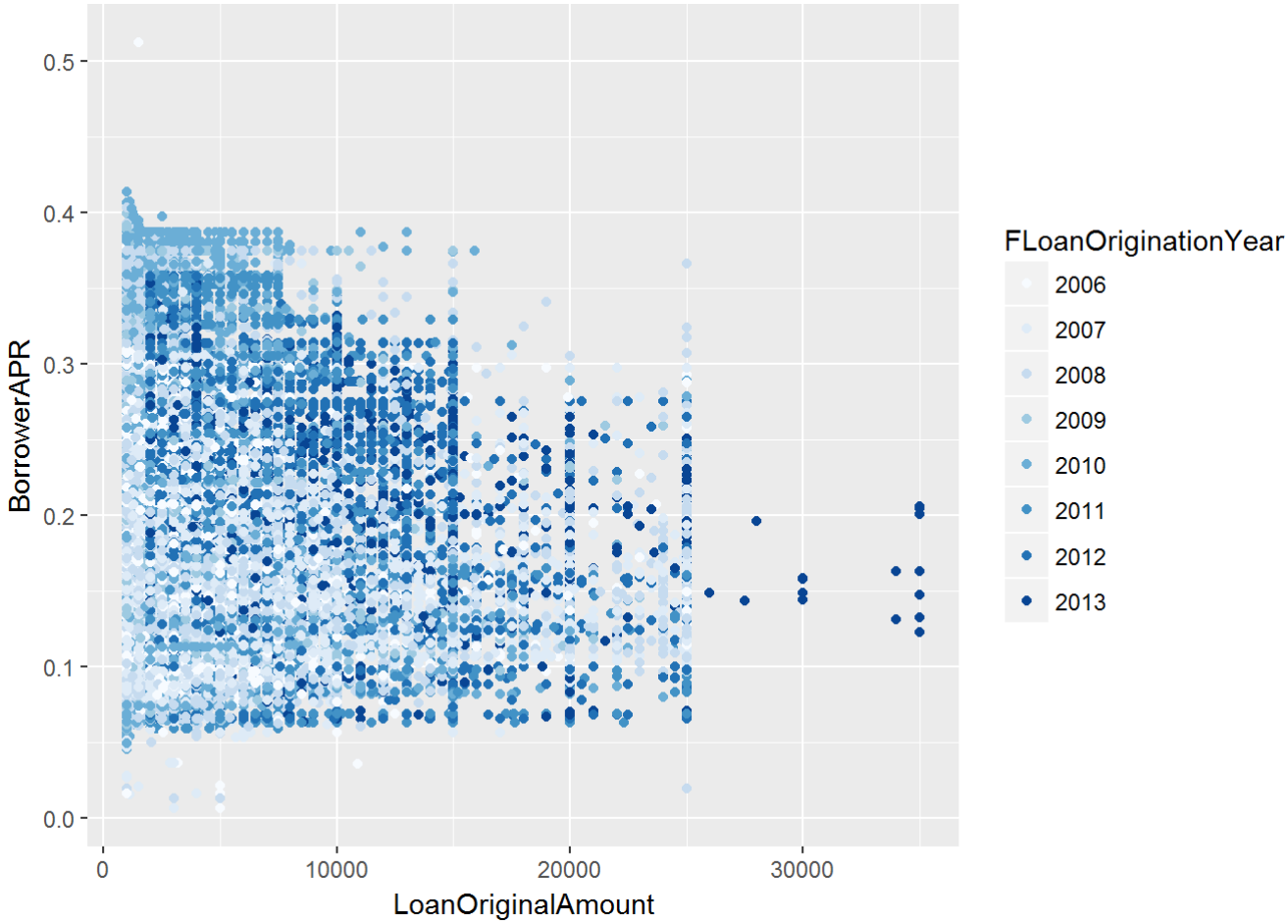
5.2. Sessão de Gráficos Multivariados com as variáveis BorrowerAPR e LoanOriginalAmount:



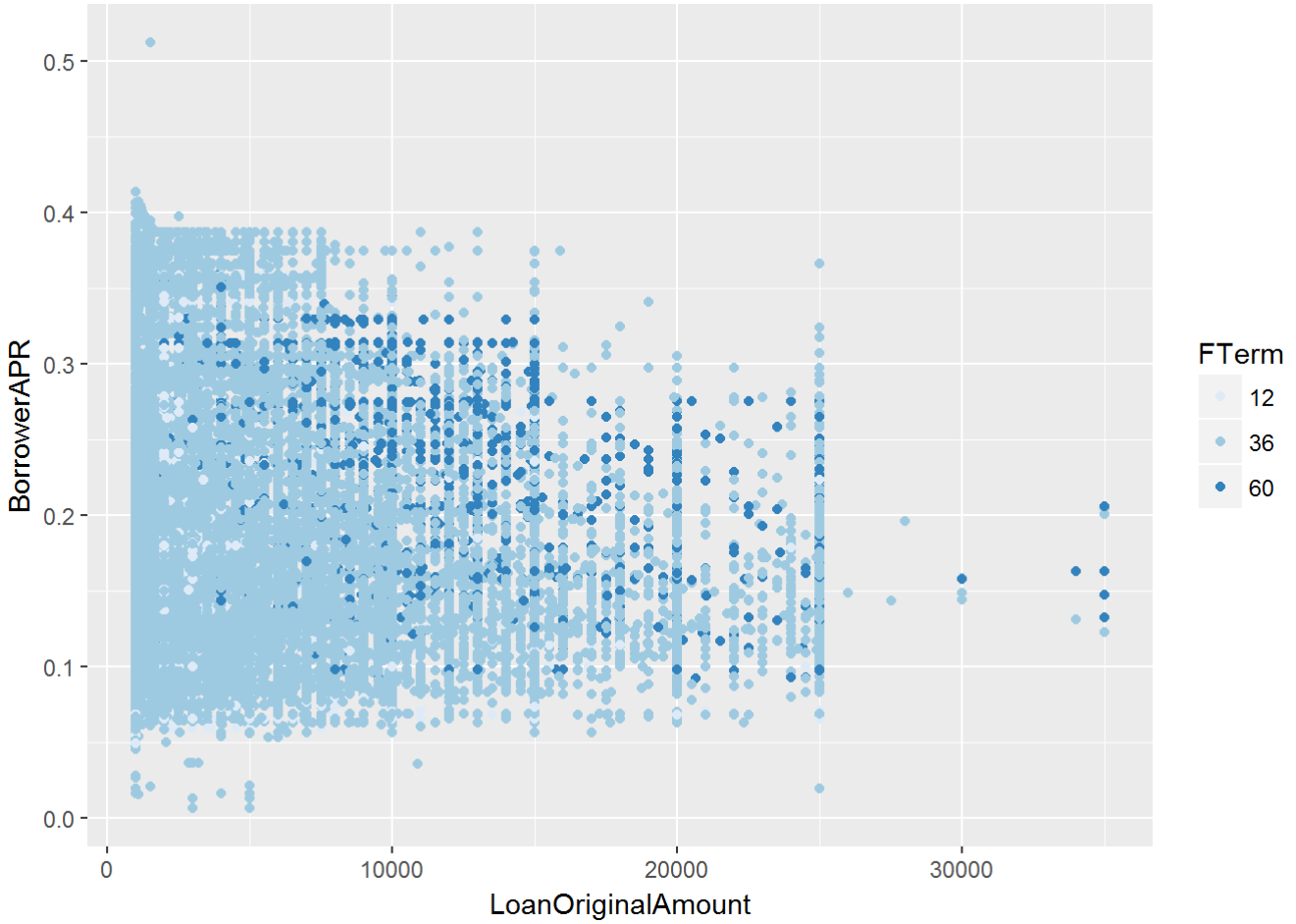

```
##
## [1] "- Observações:  Não observa-se no comportamento estre as variáveis qualquer indício d
e que exista alguma relação entre elas."
```



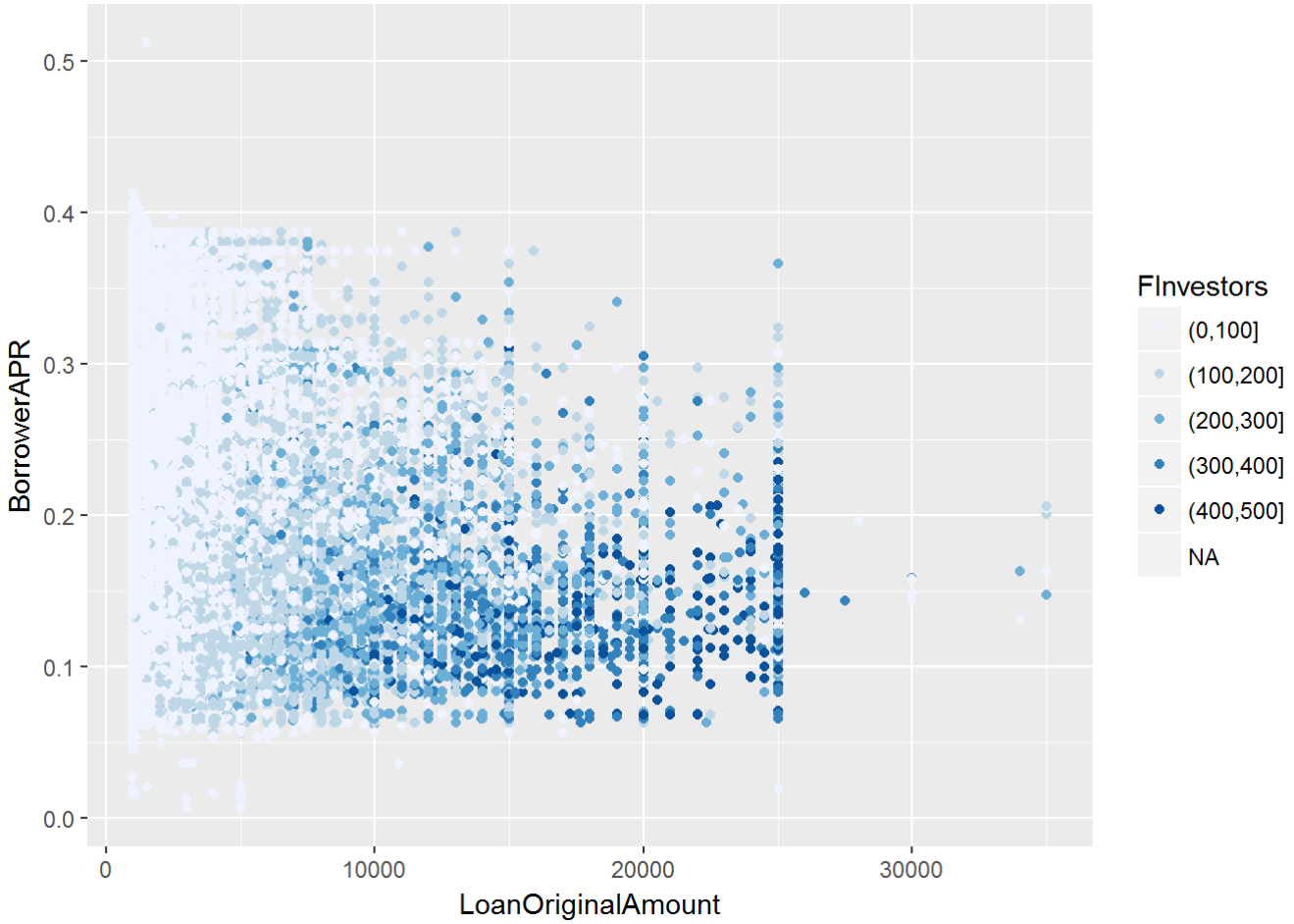

```
##
## [1] "- Observações:  Não observa-se no comportamento estre as variáveis qualquer indício d
e que exista alguma relação entre elas."
```



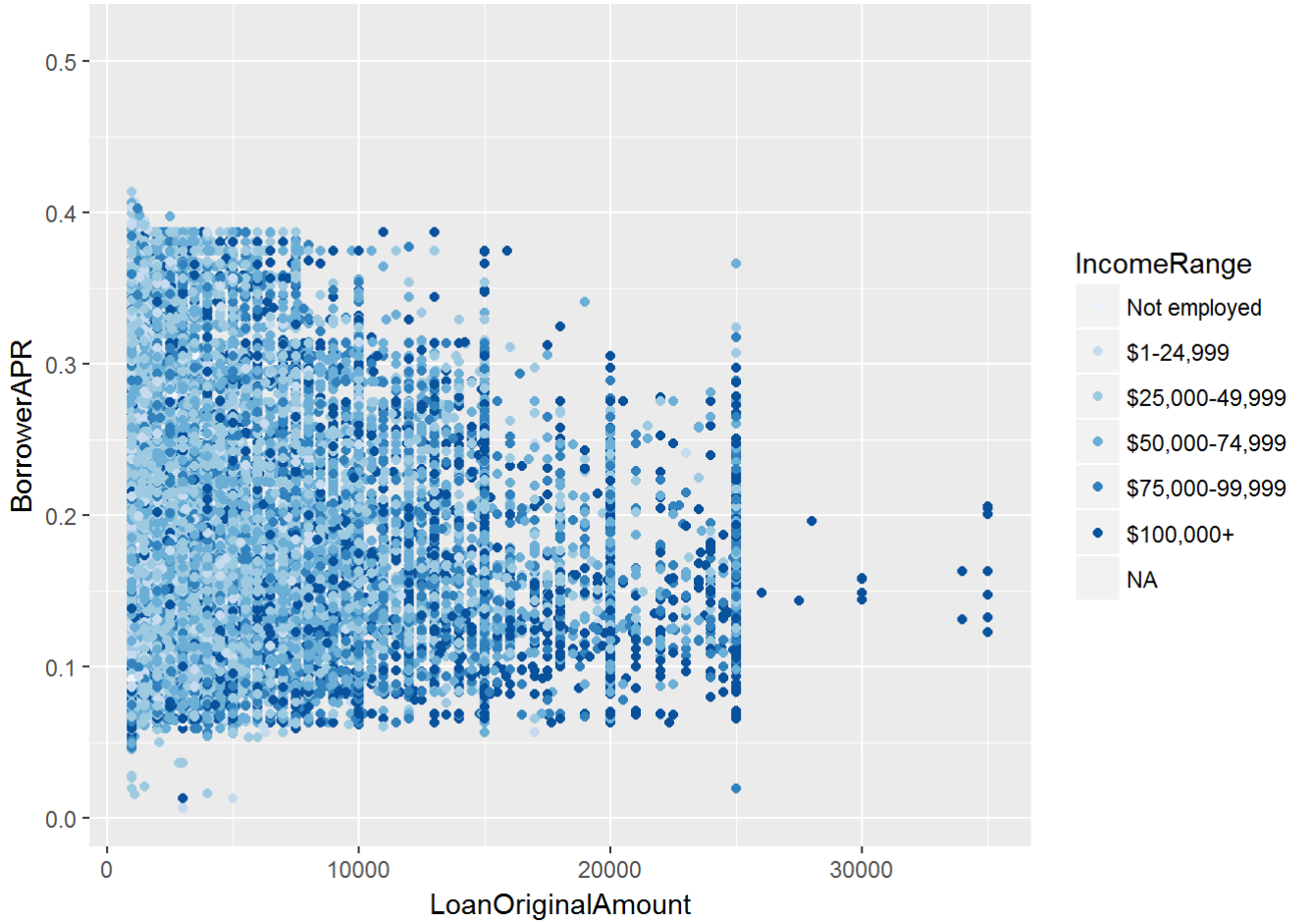

```
##
## [1] "- Observações:  Percebe-se uma tendência à um aumento no valor dos empréstimos conced
idos e também na taxa de juros anuais aplicada."
```




```
##
## [1] "- Observações:  Nesta distribuição por pontos, fica difícil perceber a ocorrência dos
empréstimos com prazos diferentes de 36 meses, pois estes representam a grande maioria."
```




```
##
## [1] "- Observações:  Como foi observado nas correlações bivariadas, novamente percebe-se u
ma certa tendência em empréstimos de valores maiores apresentarem um maior número de investid
ores."
```

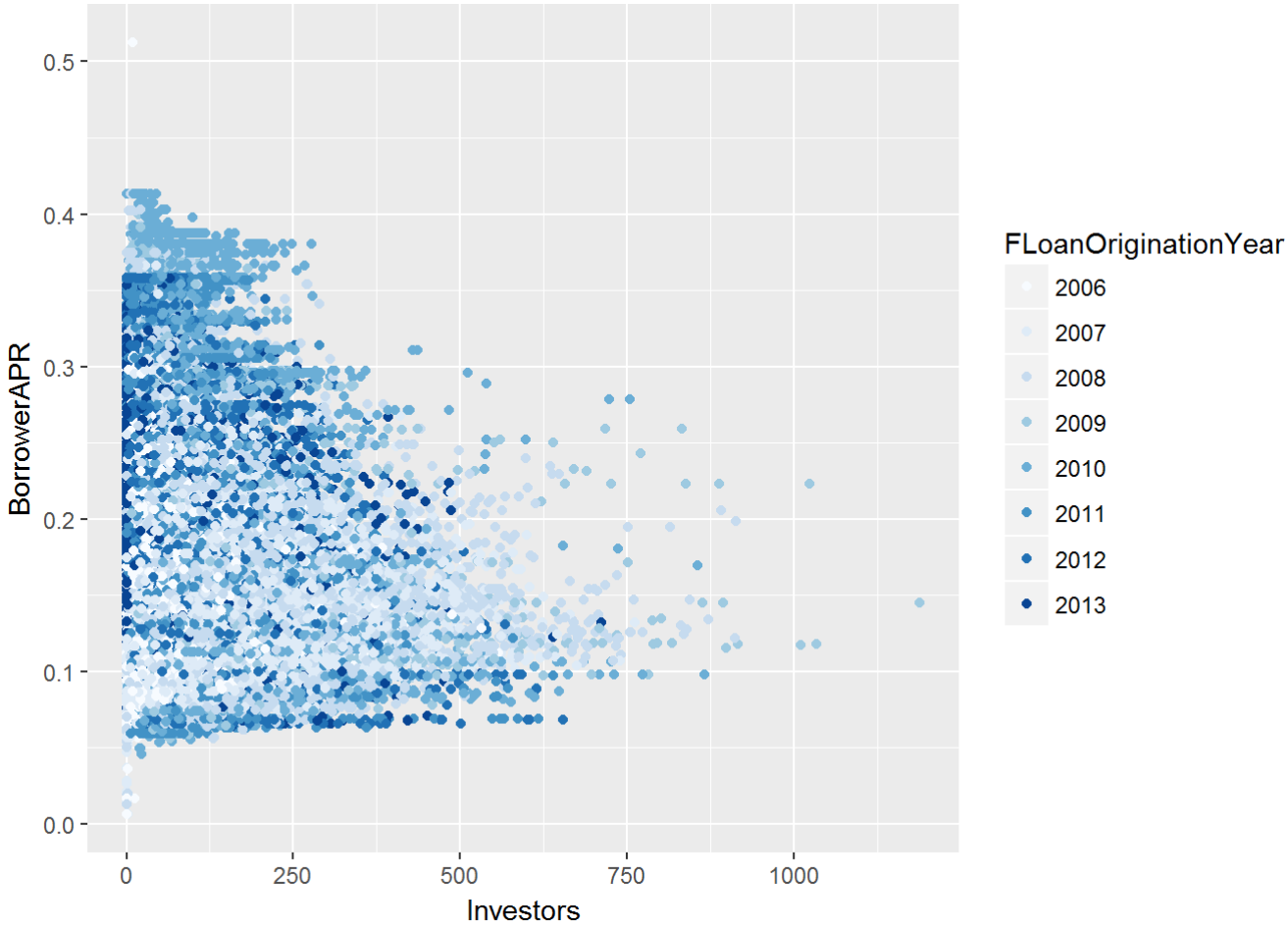



```
##
## [1] "- Observações:  Não observa-se no comportamento estre as variáveis qualquer indício d
e que exista alguma relação entre elas."
```

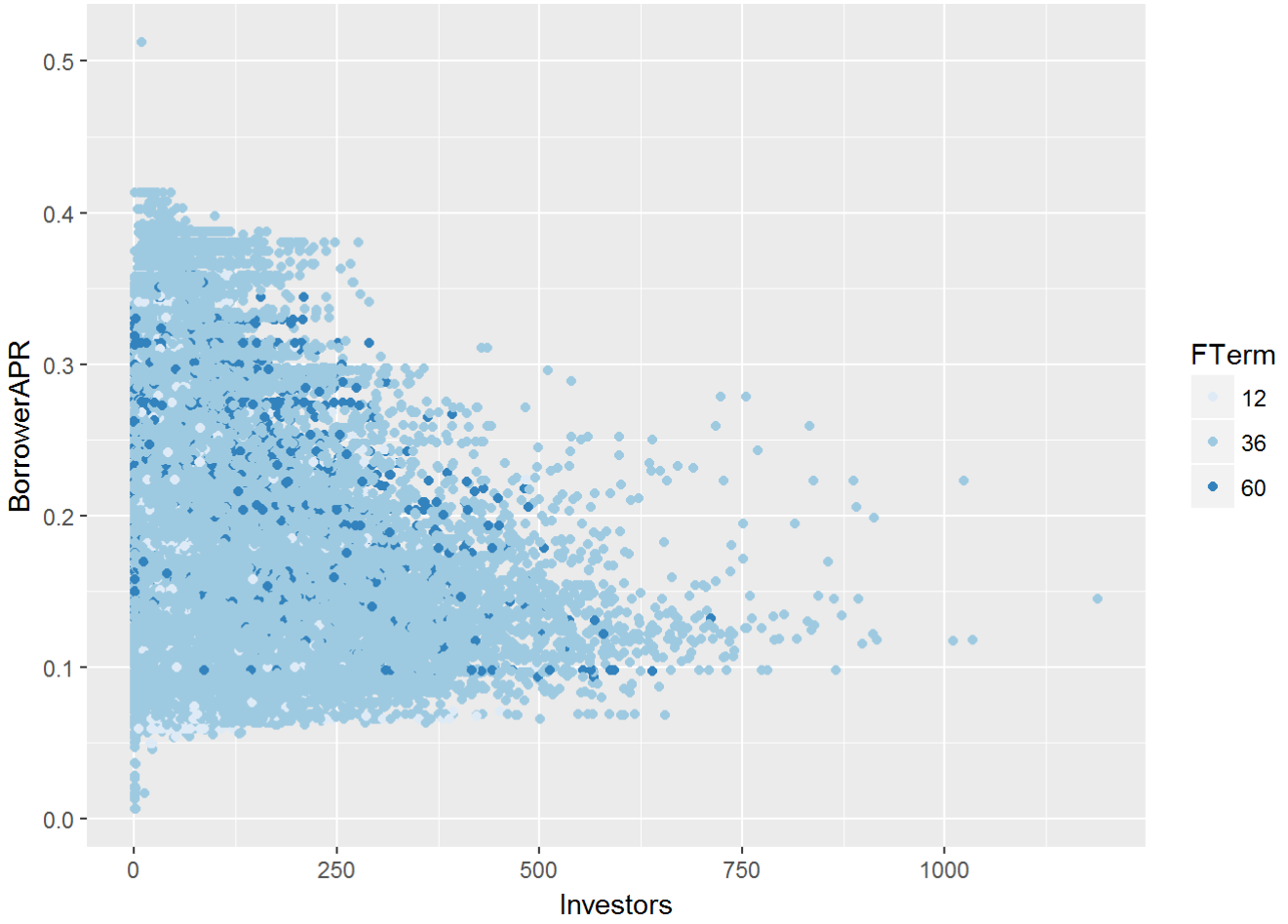
5.3. Sessão de Gráficos Multivariados com as variáveis BorrowerAPR e Investors:


```
##
## [1] "- Observações:  Não observa-se no comportamento estre as variáveis qualquer indício d
e que exista alguma relação entre elas."
```

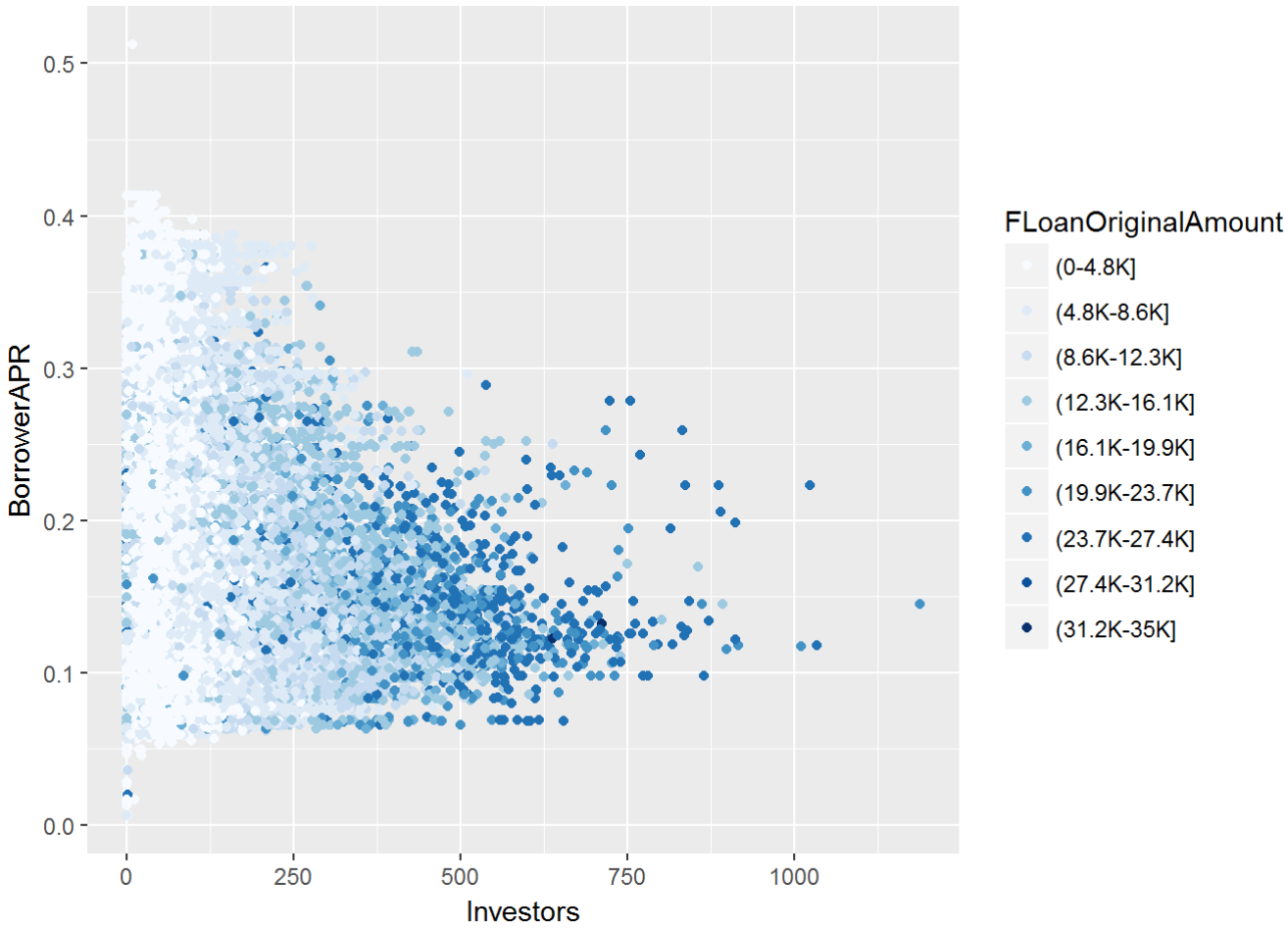

```
##
## [1] "- Observações:  Não observa-se no comportamento estre as variáveis qualquer indício d
e que exista alguma relação entre elas."
```



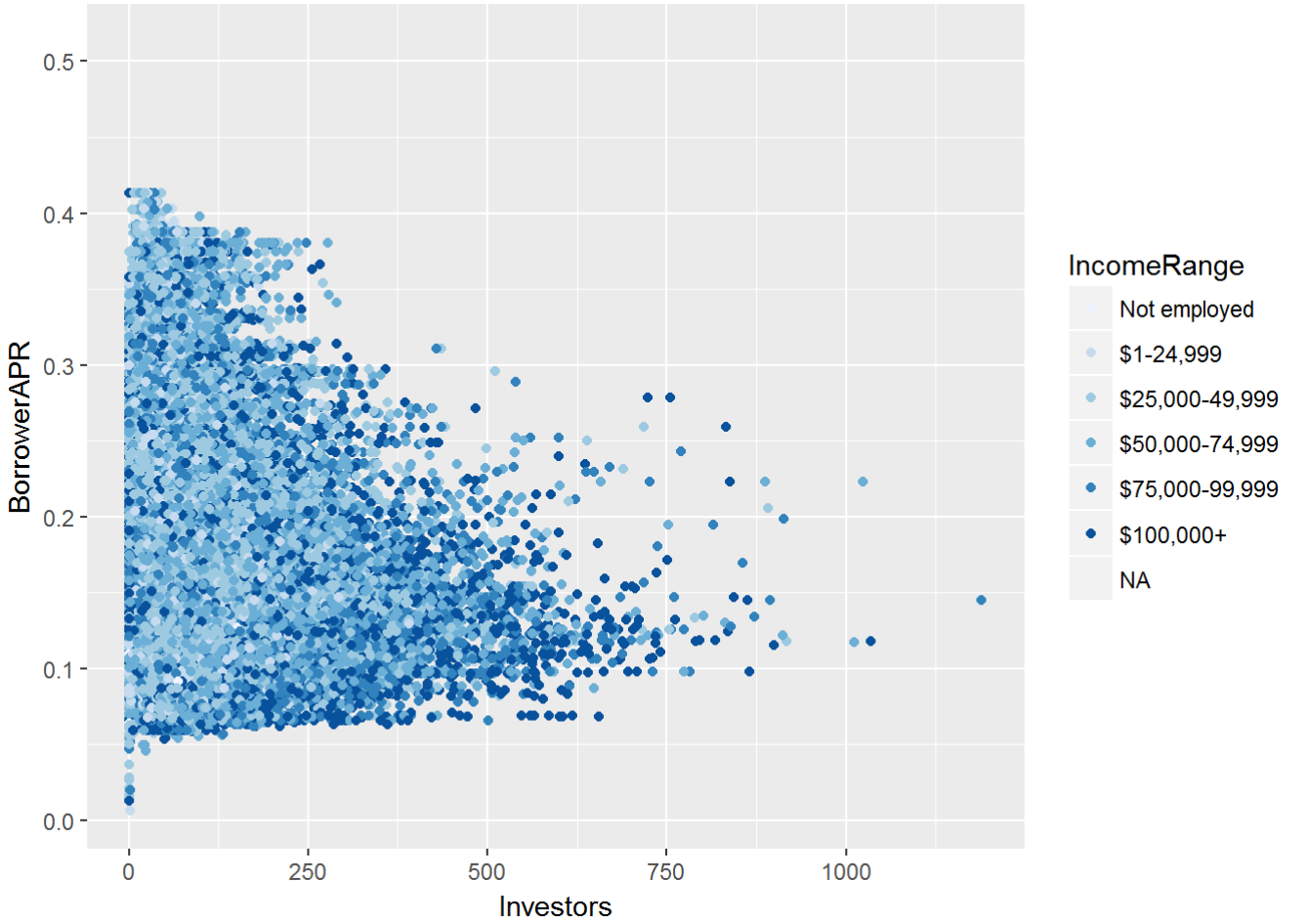

```
##
## [1] "- Observações:  Percebe-se na distribuição um aumento gradual da taxa de juros aprova
da e uma diminuição do número de investidores por empréstimo ao longo dos anos."
```




```
##
## [1] "- Observações:  Nesta distribuição por pontos, fica difícil perceber a ocorrência dos
empréstimos com prazos diferentes de 36 meses, pois estes representam a grande maioria."
```




```
##
## [1] "- Observações:  Observa-se que o número de investidores aumenta conforme aumenta o va
lor do empréstimo, entretanto, a taxa de juros anuais aprovada tende à diminuir."
```




```
##
## [1] "- Observações:  Não observa-se no comportamento estre as variáveis qualquer indício d
e que exista alguma relação entre elas."
```

5.4. Conclusões das Análises MULTIVARIADAS

Observando as semelhanças e diferenças entre os conjuntos de gráficos, destaca-se:

- Constata-se que da mesma forma que Investors e LoanOriginalAmount são influenciadas da mesma maneira também em relação às demais variáveis do conjunto de dados. Neste caso, os gráficos de dispersão da variável LoanOriginalAmount com as demais variáveis possuem o mesmo formato dos gráficos de disersão da variável Investors com as demais variáveis.

Entretanto, observando a variável Investors contra a variável LoanOriginalAmount, não obtêm-se uma distribuição clara, apesar desta distribuição tender a explicar que quanto maior o valor de LoanOriginalAmount, maior o número de investidores.

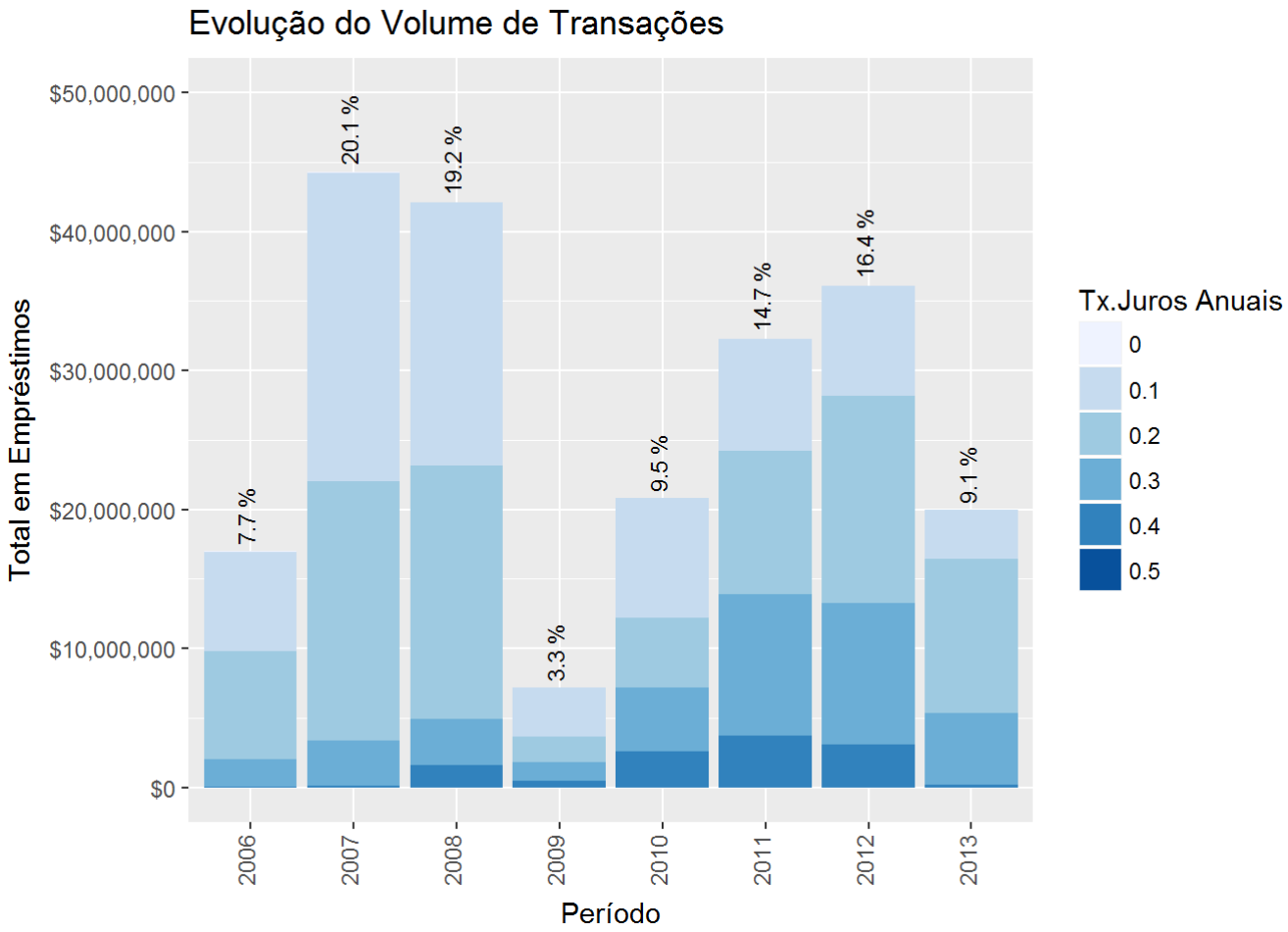
6. SESSÃO DE GRÁFICOS FINAIS

6.1. Evolução do Volume de Transações

Inicialmente faz-se necessária a apresentação de um gráfico com o propósito de caracterizar o contexto referente ao comportamento do conjunto de dados em relação ao volume de concessão dos empréstimos ao longo do tempo e as respectivas taxas de juros anuais aplicadas aos mutuários nestes empréstimos.

O gráfico apresentou uma clara proporcionalidade na distribuição dos empréstimos dentro das faixas de taxas de juros anuais (BorrowerAPR). Nota-se que essa proporcionalidade se mantêm ao longo do tempo, com a diferença de que a partir de 2008, empréstimos com juros anuais superiores à 0,4% começaram a ser concedidos.

Observou-se também que cerca de 76% do volume de empréstimo foi concedido com taxas de juros menores que 0,3%.




```
## [1] "> Estatísticas de Total em Empréstimos: "
```

```
##   FBorrowerAPR   x.mean x.min x.max    x.sum  x.perc
## 1            0 3647.917  1000 25000    87550  0.04 %
## 2          0.1 6995.378  1000 35000 79957175 36.39 %
## 3          0.2 6849.457  1000 35000 87803183 39.96 %
## 4          0.3 5320.492  1000 25000 40015419 18.21 %
## 5          0.4  3576.85  1000 25000 11867989   5.4 %
## 6          0.5     1500  1500  1500     1500     0 %
```

```
## [1] "- Total ->  $219,748,436"
```

6.2. DebtToIncomeRatio do mutuário x Juros aplicados

Um dos fatores importantes para caracterizar de que forma as taxas de juros são definidas para os empréstimos, é verificar quais as condições financeiras dos mutuários. Entre os indicadores disponíveis no conjunto de dados, foi selecionado o indicador “DebtToIncomeRatio”, pois este parece ser resultante da

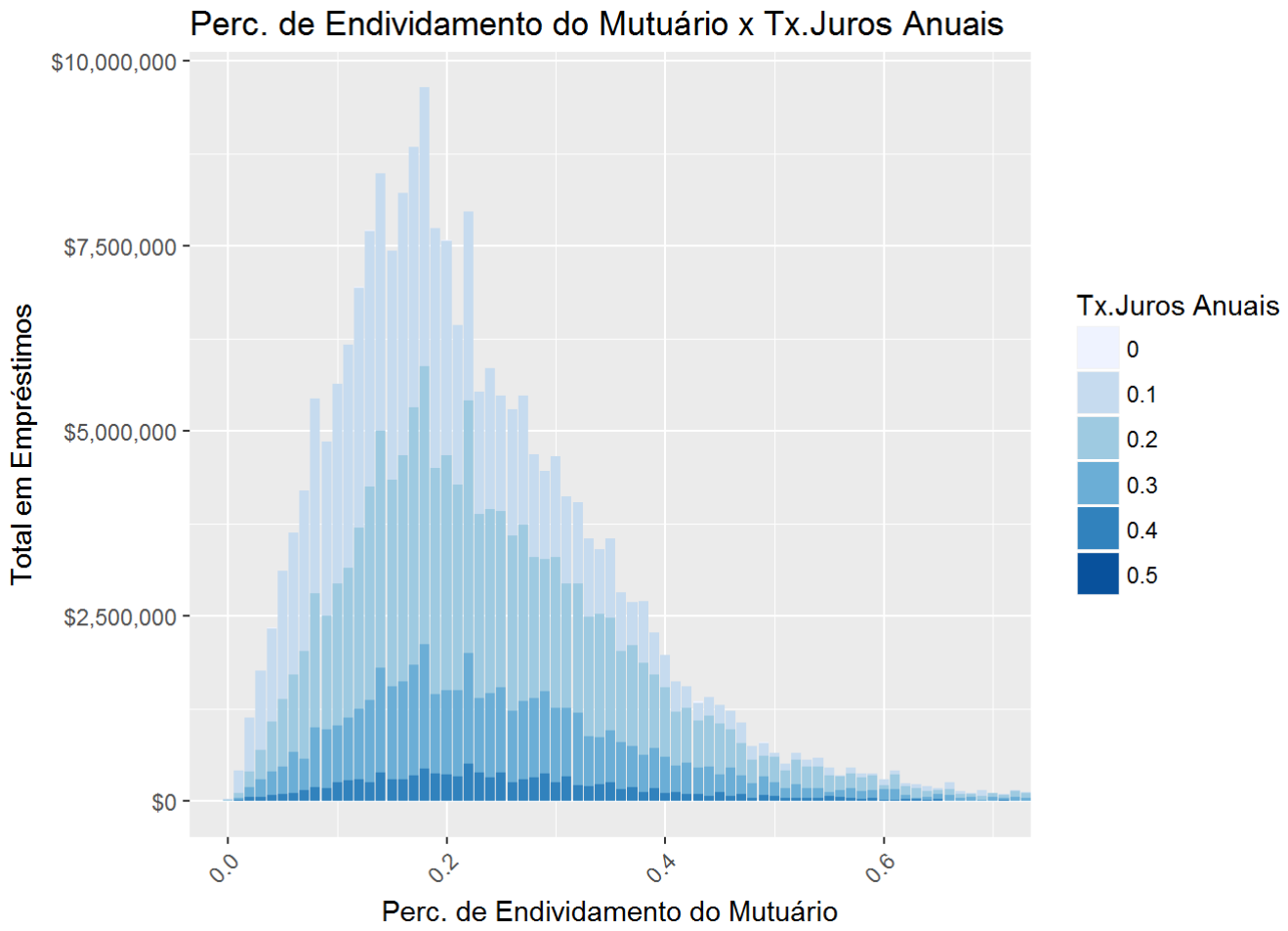
aplicação direta de uma fórmula considerando diversos outros indicadores financeiros (diferentemente dos indicadores “CreditScoreRangeLower” e “CreditScoreRangeUpper”, que tendem a ser menos precisos e talvez até depender da avaliação subjetiva de um avaliador de crédito).

Avaliando este indicador em relação às taxas de juros que foram praticadas nos empréstimos, não é possível perceber nenhuma influência do mesmo em relação à taxa de juros contratada na concessão dos empréstimos. Verifica-se novamente uma clara proporcionalidade na distribuição dos empréstimos dentro das faixas de taxas de juros anuais (BorrowerAPR), concentrando o maior volume de empréstimos em torno do valor médio de DebtToIncomeRatio (cerca de)

Como por exemplo, para empréstimos negociados com taxas de juros anuais (BorrowerAPR) entre 0,2 e 0,299:

% Volume Faixa DebtToIncomeRatio 8.35 % 0.1 14.94 % 0.2 7.83 % 0.3 4.90 % 0.4 1.39 % 0.5 0.73 % 0.6

Assim, pode-se concluir no exemplo acima, que empréstimos dentro de uma mesma faixa de juros foram concedidos para mutuários independente do resultado do seu indicador DebtToIncomeRatio.




```
## [1] 0.2053796
```

```
## [1] "> Estatísticas de Total em Empréstimos: "
```

```
##    FBorrowerAPR FDebtToIncomeRatio    x.sum  x.perc
## 7             0                0.1    12500  0.01 %
## 12            0                0.2    16850  0.01 %
## 17            0                0.3     5100     0 %
## 22            0                0.4    25000  0.01 %
## 27            0                0.5     1000     0 %
## 8           0.1                0.1 25072673 11.41 %
## 13          0.1                0.2 29846478 13.58 %
## 18          0.1                0.3 11594181  5.28 %
## 23          0.1                0.4  5704454   2.6 %
## 28          0.1                0.5  1339262  0.61 %
## 32          0.1                0.6   600062  0.27 %
## 9           0.2                0.1 18338744  8.35 %
## 14          0.2                0.2 32838418 14.94 %
## 19          0.2                0.3 17197891  7.83 %
## 24          0.2                0.4 10771711   4.9 %
## 29          0.2                0.5  3059927  1.39 %
## 33          0.2                0.6  1611199  0.73 %
## 10          0.3                0.1  7676790  3.49 %
## 15          0.3                0.2 13892532  6.32 %
## 20          0.3                0.3  8435564  3.84 %
## 25          0.3                0.4  5194174  2.36 %
## 30          0.3                0.5  1836720  0.84 %
## 34          0.3                0.6   938851  0.43 %
## 11          0.4                0.1  2101173  0.96 %
## 16          0.4                0.2  4053899  1.84 %
## 21          0.4                0.3  2486912  1.13 %
## 26          0.4                0.4  1537956   0.7 %
## 31          0.4                0.5   554131  0.25 %
## 35          0.4                0.6   410772  0.19 %
```

```
## [1] "- Total ->  $219,748,436"
```

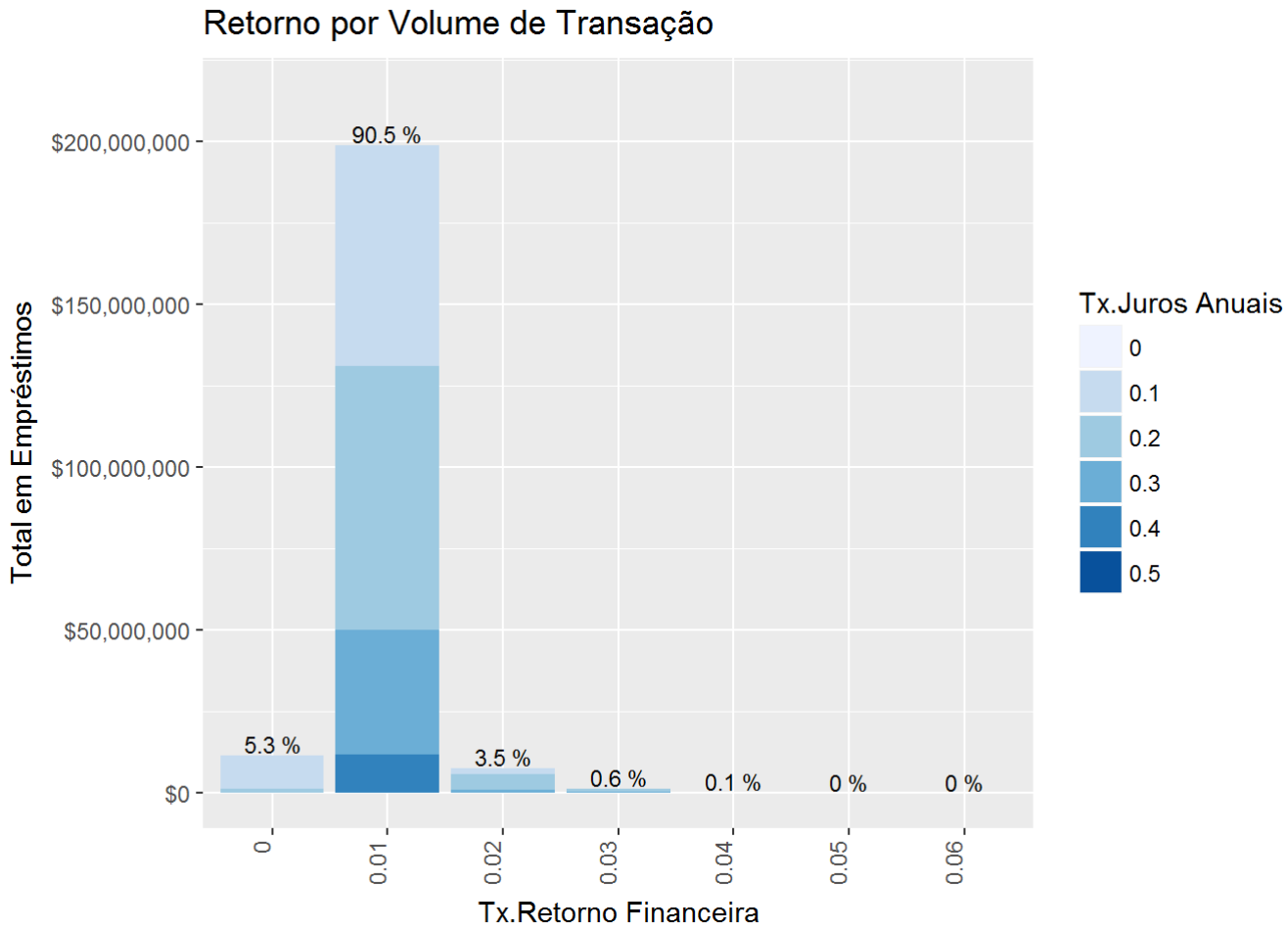
```
## [1] "- Obs. Considerados apenas empréstimos com FDebtToIncomeRatio até 0.6"
```

6.3. Retorno por Volume de Transação

Além da taxa de juros anuais aprovada, BorrowerAPR, temos também 2 outras informações importantes: que é o juros efetivamente aplicado nos empréstimos aos mutuários, BorrowerRate; e os juros de retorno aos investidores, LenderYield. A diferença entre essas taxas é o que caracterizamos como ganho da financeira, ProsperGain.

Assim, é importante verificarmos se a distribuição dos ganhos da financeira dependem ou não dos valores dos empréstimos e/ou da taxa de juros anuais aprovada. Ou seja, o volume de cada transação e à taxa de juros aprovada refletem no ganho de intermediação do serviço?

Conclui-se por meio de análises gráficas e estatísticas que os ganhos de intermediação não diferem em função da taxa de juros (BorrowerAPR) e do valor do empréstimo (LoanOriginalAmount), sendo somente proporcional ao volume dos empréstimos concedidos. Nestas condições, 90,5% dos ganhos da financeira, de um total de $2,119,610, foram obtidos à uma taxa de 0,01% (diferença entre BorrowerRate e LenderYield).




```
## [1] "> Estatísticas de Retorno Financeira: "
```

```
##    FLoanOriginalAmount FBorrowerAPR     x.sum  x.perc
## 1             (0-4.8K]            0     285.5  0.01 %
## 5             (0-4.8K]          0.1 105787.39  4.99 %
## 14            (0-4.8K]          0.2 154912.26  7.31 %
## 23            (0-4.8K]          0.3 127174.93     6 %
## 30            (0-4.8K]          0.4  78118.18  3.69 %
## 35            (0-4.8K]          0.5       7.5     0 %
## 2           (4.8K-8.6K]            0       125  0.01 %
## 6           (4.8K-8.6K]          0.1 180731.08  8.53 %
## 15          (4.8K-8.6K]          0.2 229725.53 10.84 %
## 24          (4.8K-8.6K]          0.3 127782.14  6.03 %
## 31          (4.8K-8.6K]          0.4  37613.19  1.77 %
## 3          (8.6K-12.3K]            0      54.5     0 %
## 7          (8.6K-12.3K]          0.1 143291.51  6.76 %
## 16         (8.6K-12.3K]          0.2 163866.29  7.73 %
## 25         (8.6K-12.3K]          0.3  70554.08  3.33 %
## 32         (8.6K-12.3K]          0.4   1657.51  0.08 %
## 8         (12.3K-16.1K]          0.1 106099.67  5.01 %
## 17        (12.3K-16.1K]          0.2 211359.15  9.97 %
## 26        (12.3K-16.1K]          0.3   72087.8   3.4 %
## 33        (12.3K-16.1K]          0.4      1469  0.07 %
## 9         (16.1K-19.9K]          0.1  29778.09   1.4 %
## 18        (16.1K-19.9K]          0.2  26590.91  1.25 %
## 27        (16.1K-19.9K]          0.3      3179  0.15 %
## 10        (19.9K-23.7K]          0.1  46096.63  2.17 %
## 19        (19.9K-23.7K]          0.2  43132.68  2.03 %
## 28        (19.9K-23.7K]          0.3      7473  0.35 %
## 4         (23.7K-27.4K]            0       250  0.01 %
## 11        (23.7K-27.4K]          0.1  62666.24  2.96 %
## 20        (23.7K-27.4K]          0.2  74304.97  3.51 %
## 29        (23.7K-27.4K]          0.3      8210  0.39 %
## 34        (23.7K-27.4K]          0.4       250  0.01 %
## 12        (27.4K-31.2K]          0.1       875  0.04 %
## 21        (27.4K-31.2K]          0.2       880  0.04 %
## 13          (31.2K-35K]          0.1      1390  0.07 %
## 22          (31.2K-35K]          0.2      1740  0.08 %
```

```
## [1] "- Total ->  $2,119,610"
```

7. REFLEXÃO

Abaixo relacionaremos alguns pontos de reflexão e aprendizado obtidos durante a análise do conjunto de dados “prosperLoanData”.

1. Contexto:

- O grau de dificuldade em realizar uma análise de dados pode ser maior ou menor dependendo do conhecimento e experiência do analista de dados na área relativa aos dados representados como também ao acesso deste à informações contextuais do conjunto de dados. Neste case, pesquisamos e encontramos o site www.prosper.com, que, ao que tudo indica, parece ser o site da instituição financeira provedora do conjunto de dados. Entretanto, apesar do nome do conjunto de dados sugerir ser esta a instituição, como também as similaridades entre as informações obtidas no site e o conjunto de dados, não podemos correr o risco de inferir que as regras dispostas no site referem-se ao conjunto de dados fornecido para análise. Creio que seria interessante se, junto com o conjunto de dados,

pudessem ser fornecidas mais informações contextuais, como por exemplo, o site que apresenta informações do negócio relativo à empresa provedora dos dados.

2. Exploração dos Dados:

- Quando o conjunto de dados possui muitas variáveis é praticamente inevitável desenvolver funções que busquem automatizar os procedimentos de análises à serem aplicados às mesmas. Entretanto, é preciso tomar cuidado com a automatização de código, pois pode nos levar à conclusões equivocadas em relação às variáveis cujo comportamento saia do padrão mas sem ocasionar erro durante a execução do procedimento.

3. Conclusões:

- As conclusões referente cada etapa devem ser feitas somente quando exaurem-se as perguntas e descobertas relativas ao modelo de dados, pois trata-se de um processo de conhecimento do conjunto de dados, onde a revelação de determinado comportamento de uma variável pode levar à necessidade de revisão de todas as etapas de análise anteriores.

4. Conjunto de Dados: Em relação ao conjunto de dados, podemos concluir que:

- A taxa de juros a ser aplicada ao mutuário é o indicador que irá determinar também a taxa de retorno que terá os investidores que escolheram esse empréstimo.
- Os investidores escolhem os empréstimos e investem neles a quantia desejada até que a soma total dos valores investidos cubra o valor total do empréstimo solicitado pelo mutuário.
- Ainda que o indicadores de endividamento e de crédito, como por exemplo os indicadores “DebtToIncomeRatio”" e “CreditScoreRange” sejam utilizados para determinar a taxa de juros do empréstimo (e consequentemente a tax de retorno dos investidores), parecem existir variáveis reguladoras, baseados na oferta e na procura, que são efetivas na determinação da taxa no momento da contratação. Tais variáveis parecem visar garantir um equilíbrio entre a concessão de empréstimos à juros baixos e de investimentos à juros atrativos aos investidores.

5. Sugestões para trabalhos futuros:

- Um dos pontos interessante seria estudar uma forma de determinar a taxa de juros em função do risco da transação. Ou seja, dependendo do valor do empréstimo e do perfil do mutuário, a taxa de juros só seria determinada quando no fechamento total do valor pelos investidores. Nessas condições, quanto menor o número de investidores, maior o valor do empréstimo e mais arriscado o perfil do mutuário, maior é o risco para cada investidor (proporcional ao valor investido). Nesse cenário, seria interessante que o retorno aos investidores, e também à financeira, fosse proporcional ao risco envolvido.