Análise Exploratória - Prosper Marketplace

Vinicius Ferreira Santos

27 de março de 2018

Prosper Marketplace

> O Prosper Marketplace é o primeiro mercado de empréstimos peer-to-peer dos Estados Unidos, com mais de US $ 7 bilhões em empréstimos financiados. Os mutuários solicitam empréstimos pessoais ao Prosper e os investidores (individuais ou institucionais) podem financiar de US $ 2.000 a US $ 35.000 por solicitação de empréstimo. Os investidores podem considerar as pontuações de crédito, as classificações e as histórias dos tomadores e a categoria do empréstimo. A Prosper cuida do serviço do empréstimo e cobra e distribui os pagamentos dos mutuários e os juros aos investidores. Wikipedia (https://en.wikipedia.org/wiki/Prosper_Marketplace)

Visão Geral sobre os dados

Este conjunto de dados possui 113.937 empréstimos com 81 variáveis em cada um, incluindo o valor, taxa de juros, status do pagamento, receita do mutuário, seu emprego atual, histórico do cartão de crédito e informações sobre seu último pagamento. Depois de uma rápida análise sobre os dados, resolvi selecionar as variáveis abaixo:

- ListingCreationDate: A data em que a listagem foi criada.
- Term: A duração do empréstimo expressa em meses.
- LoanStatus: O status atual do empréstimo:
    - Cancelled
    - Chargedoff
    - Completed
    - Current
    - Defaulted
    - FinalPaymentInProgress
    - PastDue (O status PastDue será acompanhado por um intervalo de inadimplência.)
- ClosedDate: A data de encerramento é aplicável para os status de empréstimo Cancelled, Completed, Chargedoff e Defaulted
- BorrowerRate: A taxa de juros do Mutuário para este empréstimo.
- LenderYield: O rendimento do credor no empréstimo. O rendimento do credor é igual à taxa de juros do empréstimo menos a taxa de serviço.
- EstimatedReturn: O retorno estimado atribuído à listagem no momento em que foi criado. O retorno estimado é a diferença entre o rendimento efetivo estimado e a taxa de perda estimada. Aplicável para empréstimos originados após julho de 2009.
- ProsperScore: Uma pontuação de risco personalizada criada usando dados históricos do Prosper. A pontuação varia de 1 a 10, sendo 10 a melhor ou a menor pontuação de risco. Aplicável para

empréstimos originados após julho de 2009.

- ListingCategory: A categoria da listagem que o mutuário selecionou ao postar sua listagem:
  - 0 - Not Available
  - 1 - Debt Consolidation
  - 2 - Home Improvement
  - 3 - Business
  - 4 - Personal Loan
  - 5 - Student Use
  - 6 - Auto
  - 7- Other
  - 8 - Baby&Adoption
  - 9 - Boat
  - 10 - Cosmetic Procedure
  - 11 - Engagement Ring
  - 12 - Green Loans
  - 13 - Household Expenses
  - 14 - Large Purchases
  - 15 - Medical/Dental
  - 16 - Motorcycle
  - 17 - RV
  - 18 - Taxes
  - 19 - Vacation
  - 20 - Wedding Loans
- BorrowerState: A abreviação de duas letras do estado do endereço do mutuário no momento em que a Listagem foi criada.
- Occupation: A Ocupação selecionada pelo Mutuário no momento em que criou a listagem.
- EmploymentStatus: O status de emprego do mutuário no momento em que eles publicaram a listagem.
- EmploymentStatusDuration: A duração em meses do status de emprego no momento em que a listagem foi criada.
- IsBorrowerHomeowner: Um Mutuário será classificado como proprietário se tiver uma hipoteca em seu perfil de crédito ou fornecer documentação confirmando que é um proprietário.
- FirstRecordedCreditLine: A data em que a primeira linha de crédito foi aberta.
- BankcardUtilization: A porcentagem de crédito rotativo disponível que é utilizada no momento em que o perfil de crédito foi retirado.
- AvailableBankcardCredit: O crédito total disponível através de cartão bancário no momento em que o perfil de crédito foi retirado.
- TradesOpenedLast6Months: Número de negociações abertas nos últimos 6 meses no momento em que o perfil de crédito foi retirado.
- StatedMonthlyIncome: A renda mensal que o mutuário declarou no momento em que a listagem foi criada.
- TotalProsperLoans: Número de empréstimos que o mutuário tinha na Prosper no momento em que eles criaram esta listagem. Esse valor será nulo se o mutuário não tiver empréstimos anteriores.
- OnTimeProsperPayments: Número de pagamentos em tempo, que o mutuário efetuou em empréstimos na Prosper no momento em que criaram esta listagem. Esse valor será nulo se o mutuário não tiver empréstimos anteriores.
- LoanCurrentDaysDelinquent: O número de dias em atraso.
- LoanMonthsSinceOrigination: Número de meses desde a origem do empréstimo.
- LoanOriginalAmount: O montante de originação do empréstimo.
- LoanOriginationDate: A data em que o empréstimo foi originado.
- LoanOriginationQuarter: O trimestre em que o empréstimo foi originado.
- MonthlyLoanPayment: O pagamento do empréstimo mensal programado.
- Investors: O número de investidores que financiaram o empréstimo.

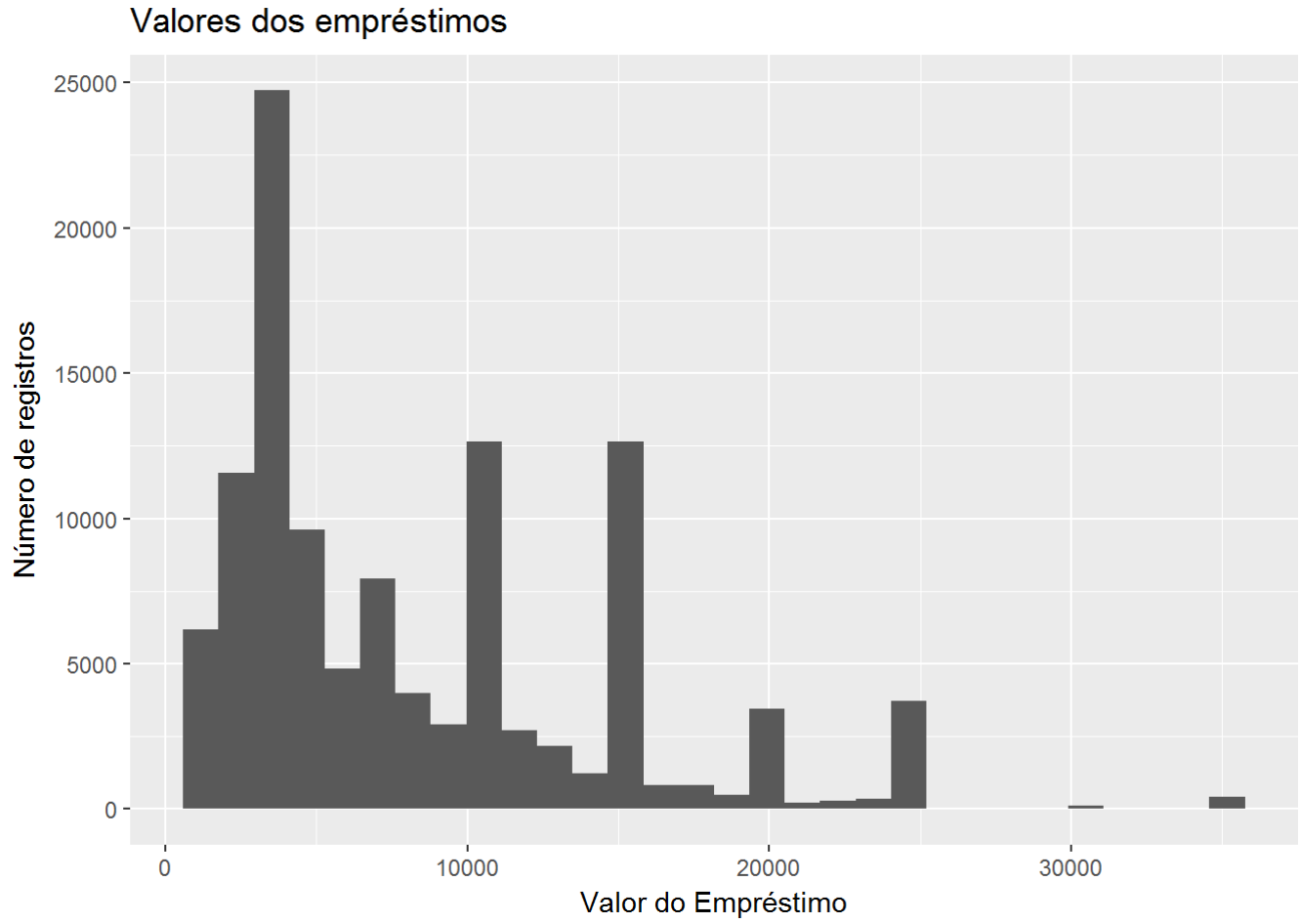
- Recommendations: Número de recomendações que o mutuário tinha no momento em que a listagem foi criada.
- IncomeRange: O intervalo de rendimento do mutuário no momento em que a listagem foi criada.

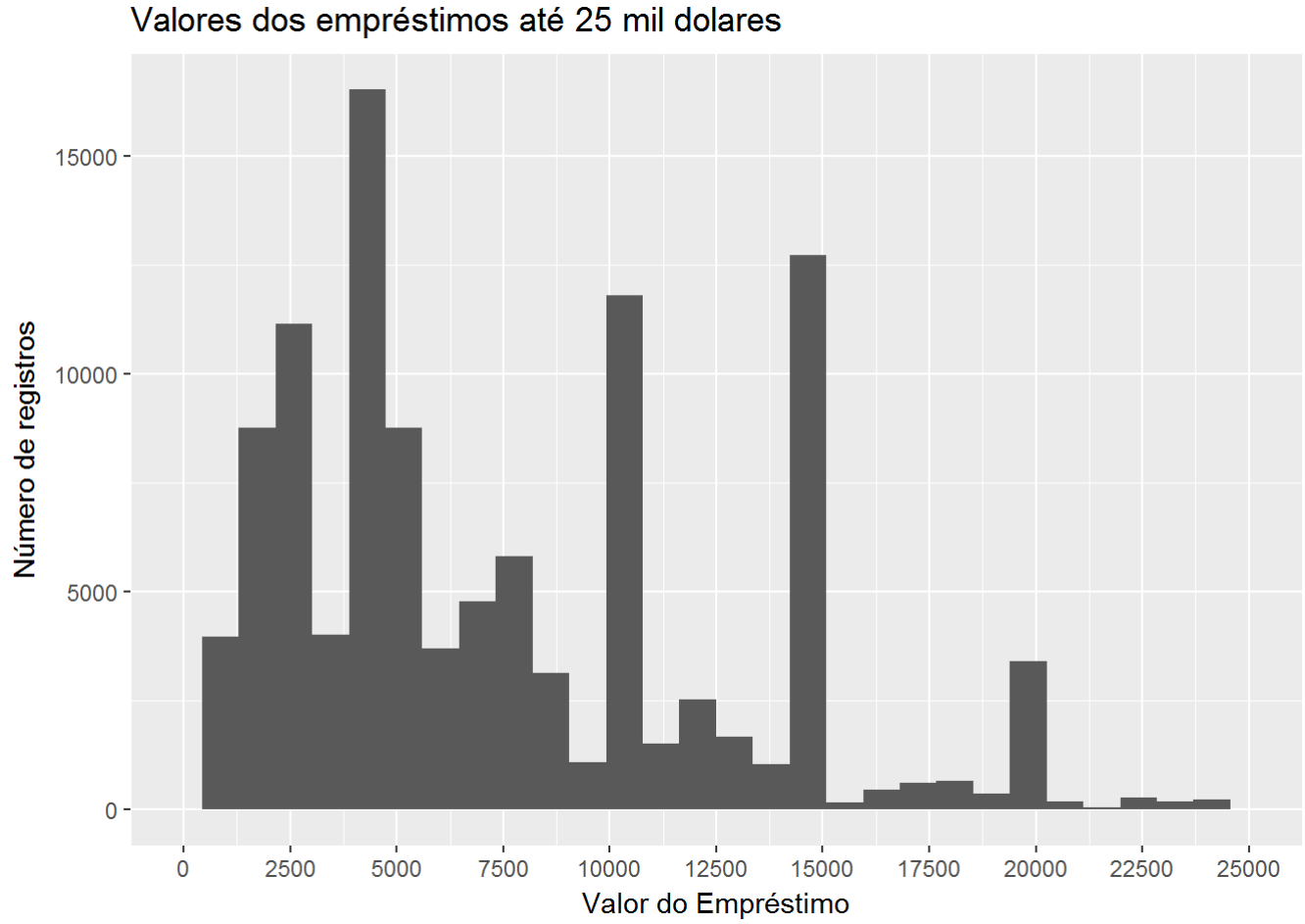
```
## 'data.frame':    113937 obs. of  32 variables:
##  $ ListingCreationDate       : Factor w/ 113064 levels "2005-11-09 20:44:28.847000000",..:
14184 111894 6429 64760 85967 100310 72556 74019 97834 97834 ...
##  $ Term                      : int  36 36 36 36 36 60 36 36 36 36 ...
##  $ LoanStatus                : Factor w/ 12 levels "Cancelled","Chargedoff",..: 3 4 3 4 4
4 4 4 4 4 ...
##  $ ClosedDate                : Factor w/ 2802 levels "2005-11-25 00:00:00",..: 1137 NA 126
2 NA NA NA NA NA NA NA ...
##  $ BorrowerRate              : num  0.158 0.092 0.275 0.0974 0.2085 ...
##  $ LenderYield               : num  0.138 0.082 0.24 0.0874 0.1985 ...
##  $ EstimatedReturn           : num  NA 0.0547 NA 0.06 0.0907 ...
##  $ ProsperScore              : num  NA 7 NA 9 4 10 2 4 9 11 ...
##  $ ListingCategory           : int  0 2 0 16 2 1 1 2 7 7 ...
##  $ BorrowerState             : Factor w/ 51 levels "AK","AL","AR",..: 6 6 11 11 24 33 17 5
15 15 ...
##  $ Occupation                : Factor w/ 67 levels "Accountant/CPA",..: 36 42 36 51 20 42
49 28 23 23 ...
##  $ EmploymentStatus          : Factor w/ 8 levels "Employed","Full-time",..: 8 1 3 1 1 1 1
1 1 1 ...
##  $ EmploymentStatusDuration  : int  2 44 NA 113 44 82 172 103 269 269 ...
##  $ IsBorrowerHomeowner       : Factor w/ 2 levels "False","True": 2 1 1 2 2 2 1 1 2 2 ...
##  $ FirstRecordedCreditLine   : Factor w/ 11585 levels "1947-08-24 00:00:00",..: 8638 6616
8926 2246 9497 496 8264 7684 5542 5542 ...
##  $ BankcardUtilization       : num  0 0.21 NA 0.04 0.81 0.39 0.72 0.13 0.11 0.11 ...
##  $ AvailableBankcardCredit   : num  1500 10266 NA 30754 695 ...
##  $ TradesOpenedLast6Months   : num  0 2 NA 0 2 0 0 0 1 1 ...
##  $ StatedMonthlyIncome       : num  3083 6125 2083 2875 9583 ...
##  $ TotalProsperLoans         : int  NA NA NA NA 1 NA NA NA NA NA ...
##  $ OnTimeProsperPayments     : int  NA NA NA NA 11 NA NA NA NA NA ...
##  $ LoanCurrentDaysDelinquent : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LoanMonthsSinceOrigination: int  78 0 86 16 6 3 11 10 3 3 ...
##  $ LoanOriginalAmount        : int  9425 10000 3001 10000 15000 15000 3000 10000 10000 100
00 ...
##  $ LoanOriginationDate       : Date, format: "2007-09-12" "2014-03-03" ...
##  $ LoanOriginationQuarter    : Factor w/ 33 levels "Q1 2006","Q1 2007",..: 18 8 2 32 24 33
16 16 33 33 ...
##  $ MonthlyLoanPayment        : num  330 319 123 321 564 ...
##  $ Investors                 : int  258 1 41 158 20 1 1 1 1 1 ...
##  $ Recommendations           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ IncomeRange               : Factor w/ 8 levels "$0","$1-24,999",..: 4 5 7 4 3 3 4 4 4 4
...
##  $ ProsperRating             : Factor w/ 7 levels "A","AA","B","C",..: NA 1 NA 1 5 3 6 4 2
2 ...
##  $ YearLoan                  : num  2007 2014 2007 2012 2013 ...
```

Após a criação do novo dataset com as variáveis selecionadas, continuamos com as mesmas 113.937 observações, só que agora distribuídas em 32 variáveis.

Análise Univariada

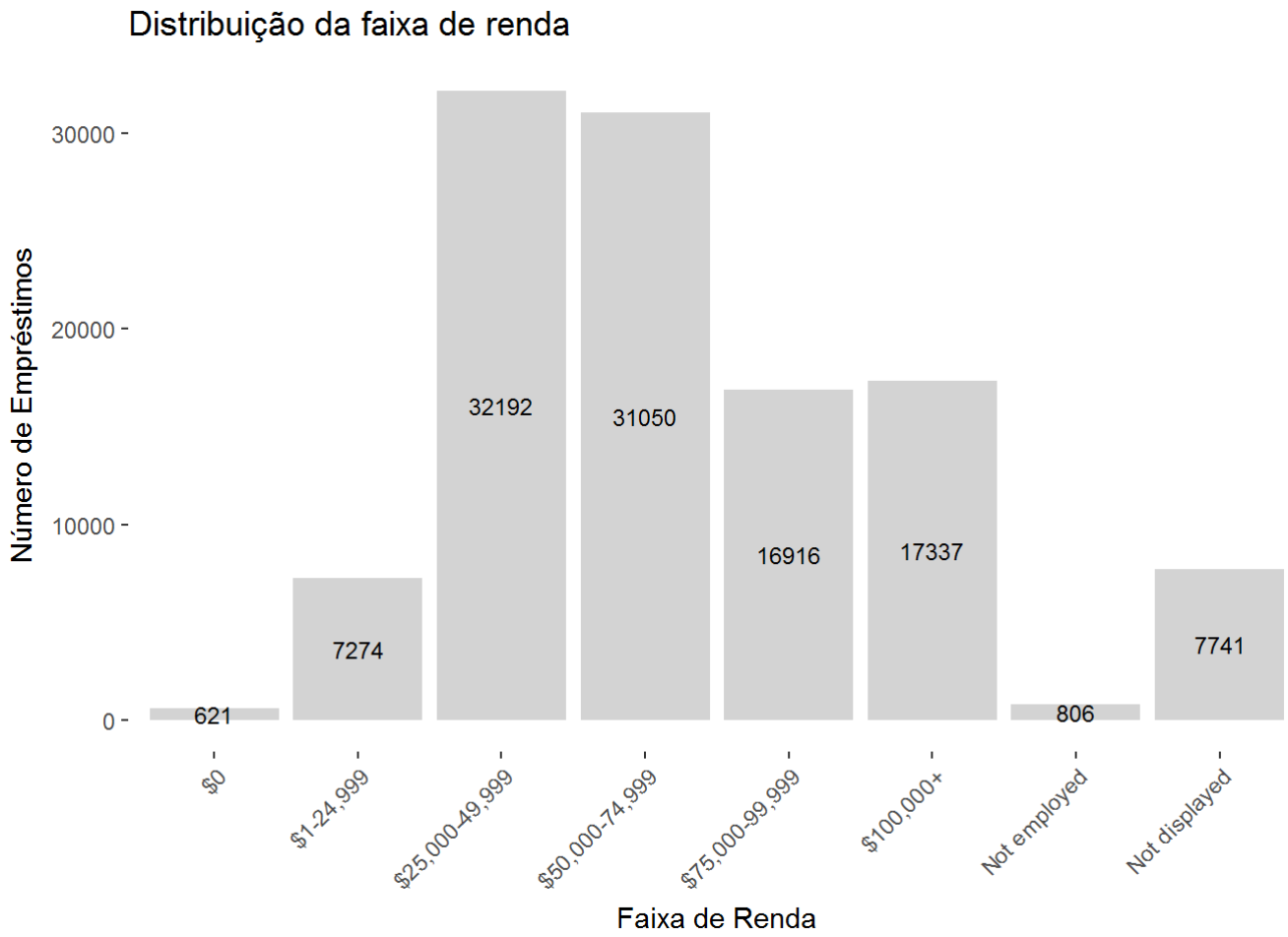
Para iniciar a análise, iremos começar pela distribuição do valor original dos empréstimos





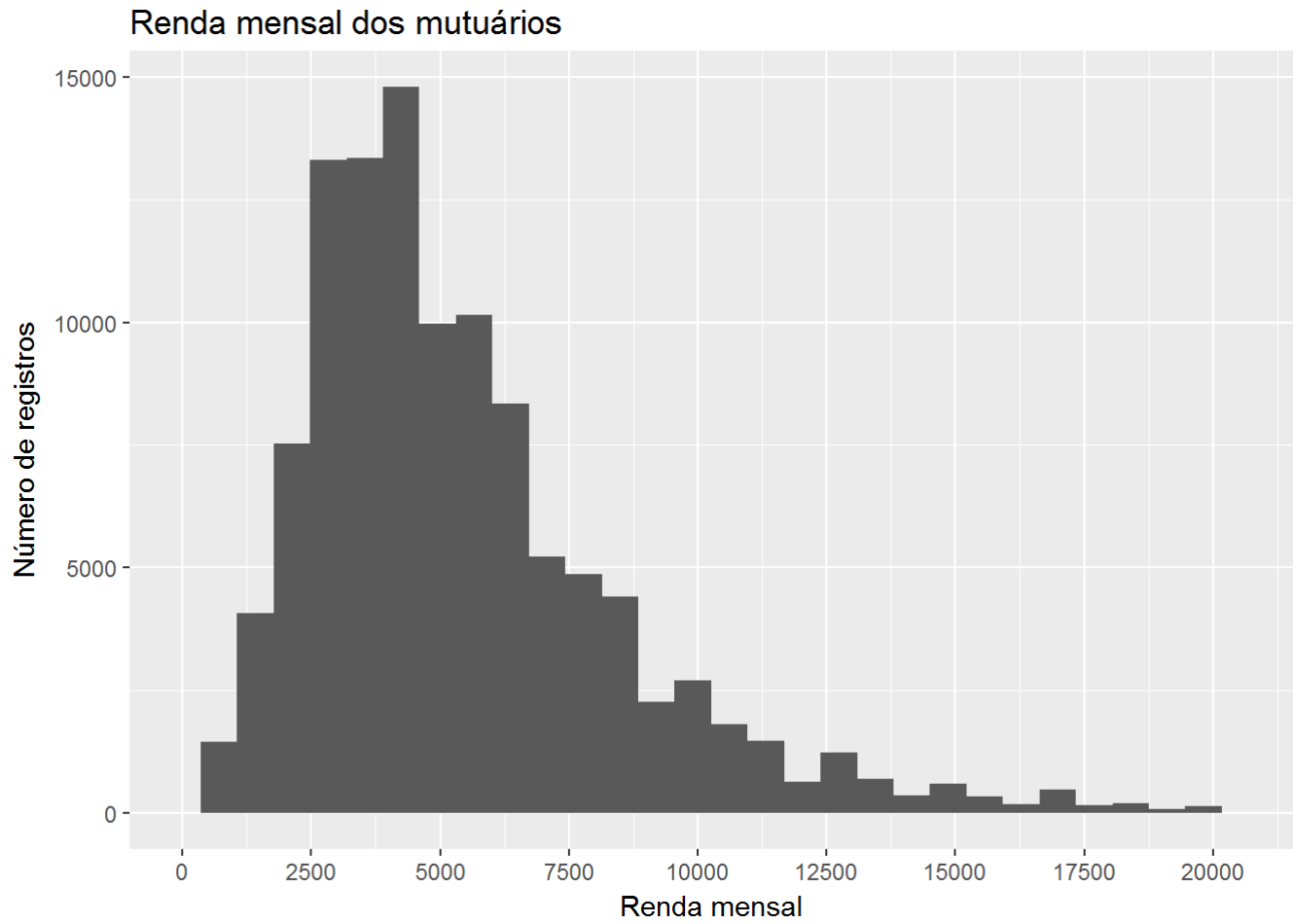
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    6500    8337   12000   35000
```

Como podemos perceber na primeira visualização, a distribuição possui uma assimetria positiva, o que me levou a limitar a segunda visualização até 25000$, a fim de entender melhor a distribuição. Depois dos ajustes, percebemos alguns picos próximos de 5, 10 e 15 mil dolares. Isso é um ponto interessante, porque percebemos que as pessoas costumam pegar valores próximos de múltiplos de 5. Outro ponto é que como temos uma assimetria positiva, nossa média(8.337) tende a ser maior que a mediana(6.500), o que nos levaria a resultados errados se fôssemos utilizar a média salarial como medida de comparação.



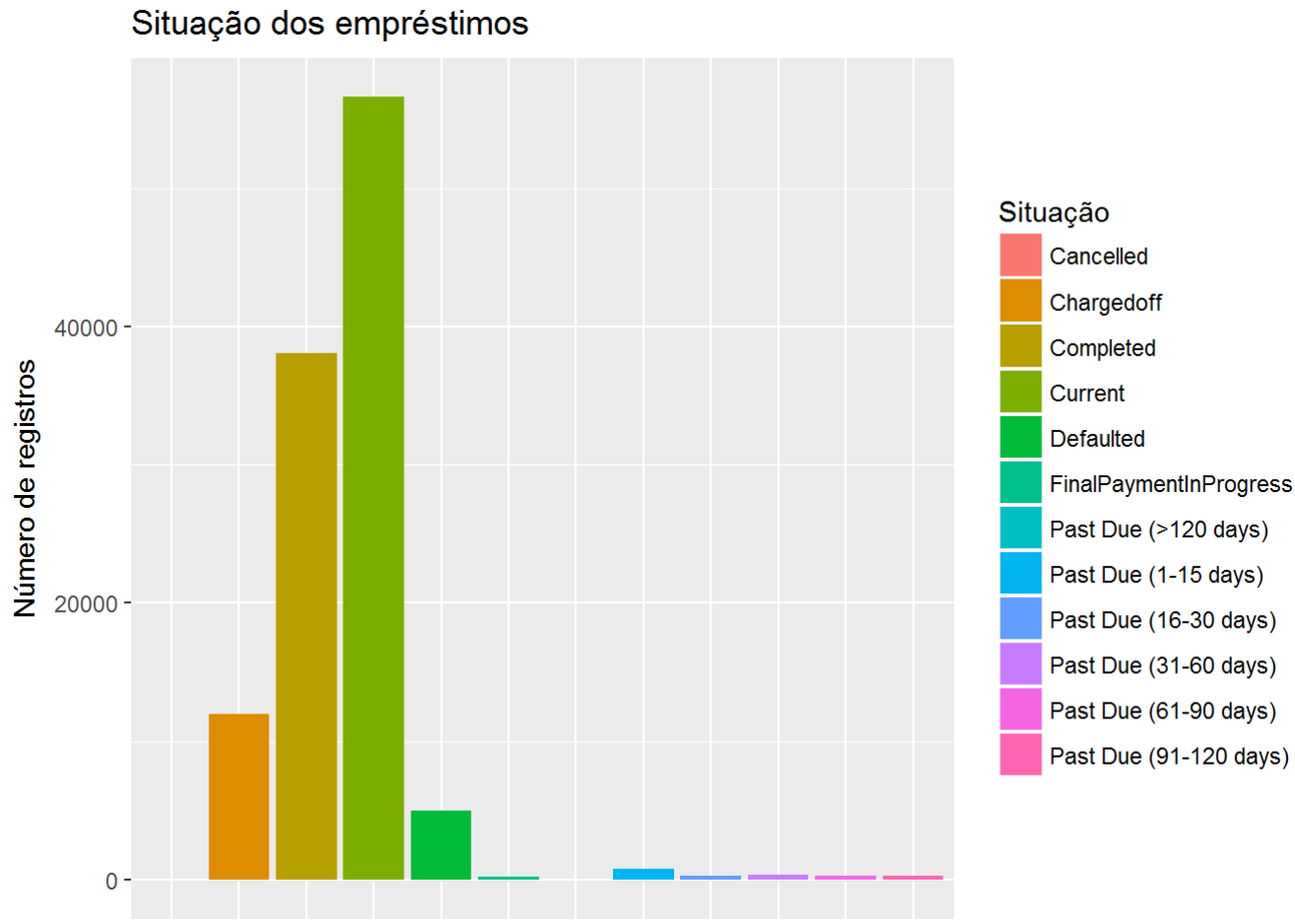

```
##
##             $0      $1-24,999      $100,000+ $25,000-49,999 $50,000-74,999
##            621           7274          17337          32192          31050
## $75,000-99,999  Not displayed   Not employed
##          16916           7741            806
```

A grande concentração de renda dos mutuários está entre $25.000 e $49.999 por ano. A grande descoberta aqui são 1.427 mutuários entre os desempregados e os que não possuem renda estarem conseguindo empréstimos. Como ainda não explorei todas as possíveis informações dos dados, acredito que para os desempregados, a explicação seria informações de renda salarial de empregos anteriores, mas para os com renda $0, acredito que em sua maioria são estudantes ou pessoas com boas recomendações.



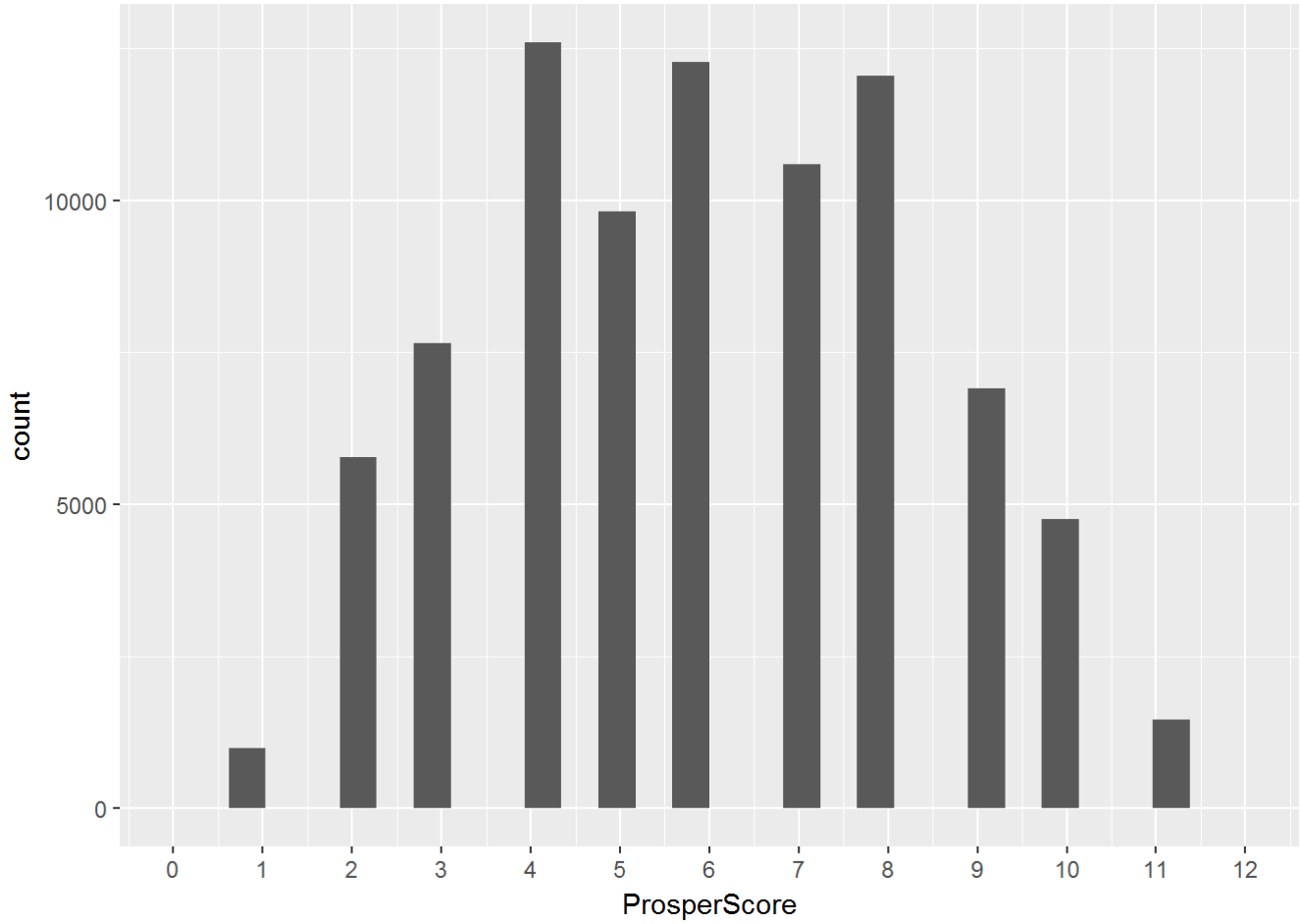

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3200    4667    5608    6825 1750003
```

Como já previa, a distribuição da renda mensal possui uma assimetria positiva, com uma concentração próxima de $ 5.000. Para analisar melhor a distribuição, decidi exclui 1% dos dados superiores, mas, antes disso, descobri que um mutuário com uma renda mensal de quase 2 milhões de dolares, solicitou um empréstimo de 4 mil dolares. Acredito que esse mutuário estava interessado em participar como investidor e fez um empréstimo baixo para verificar a lisura do funcionamento da Prosper.




```
##
##              Cancelled            Chargedoff              Completed
##                      5                 11992                  38074
##                Current             Defaulted FinalPaymentInProgress
##                  56576                  5018                    205
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days)
##                     16                   806                    265
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days)
##                    363                   313                    304
```

Podemos visualizar que a maioria dos dados se encontra com a situação “current”, com surpreendentes 56576 empréstimos, que juntos somam $586.174.602 e outros 38074 “Completed” que somam $235.643.536. Temos também alguns mutuários inadimplentes separados em diferentes categorias que, até então, estão dando um calote de $ 126.356.754.

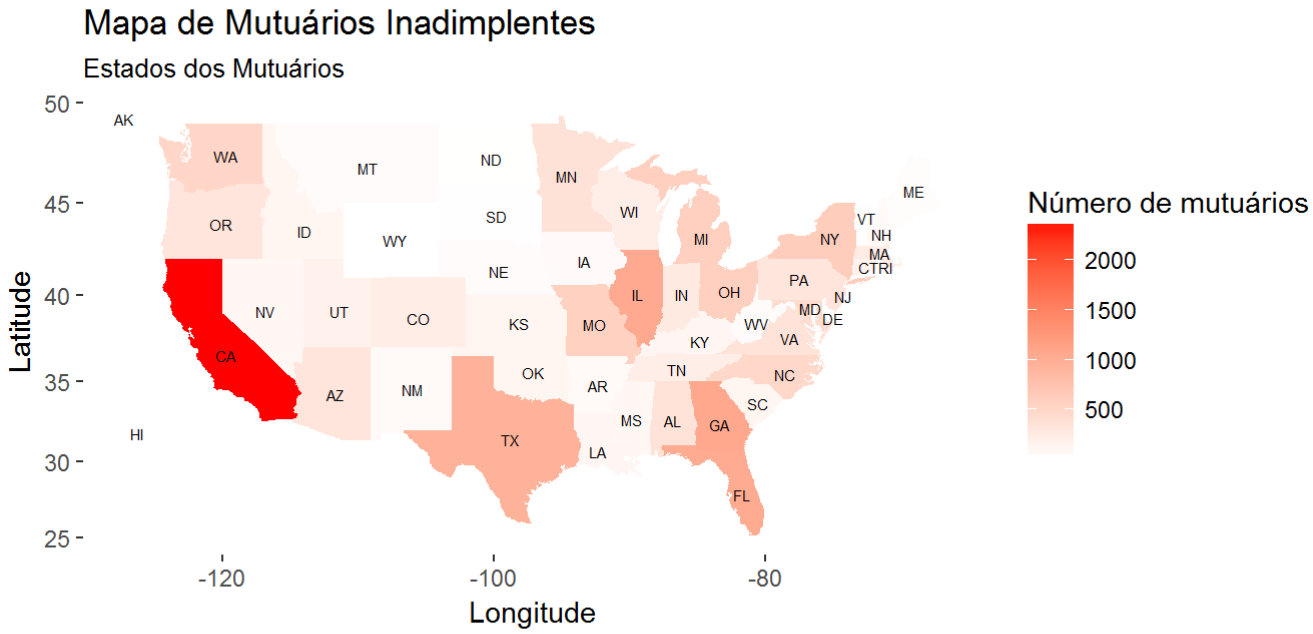


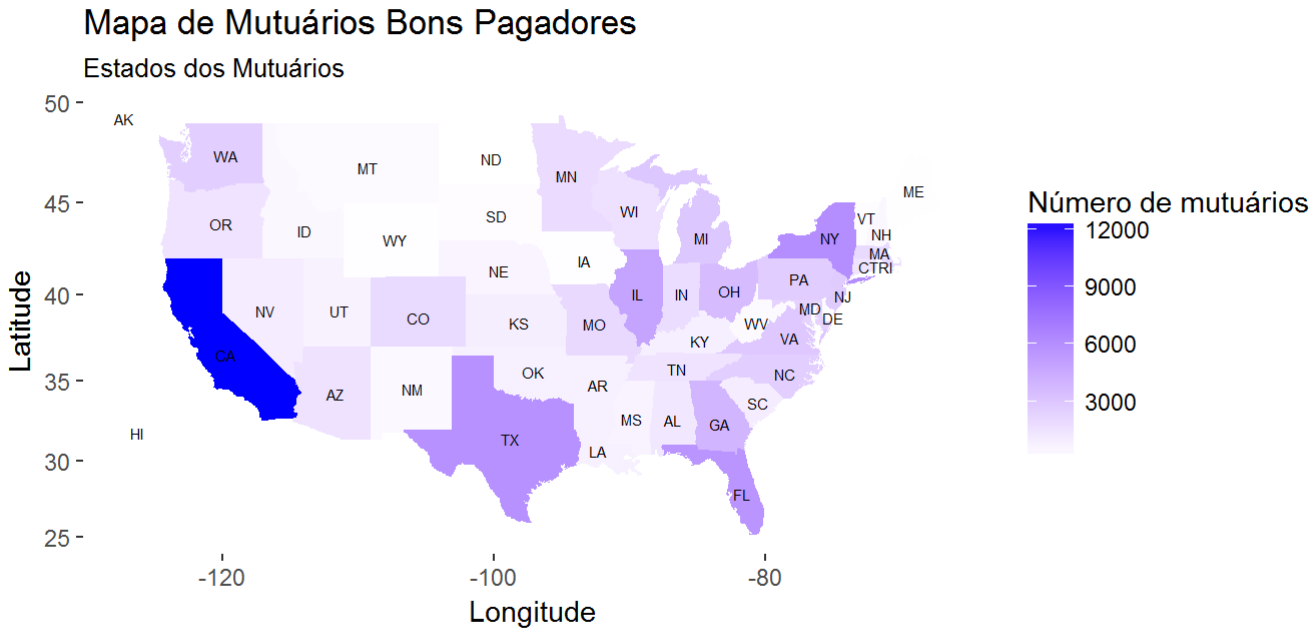

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    4.00    6.00    5.95    8.00   11.00   29084
```

O score dos mutuários demonstram algo parecido com uma distribuição normal, mas como o dicionário de dados indica que o valor 10 pode indicar a melhor ou a menor pontuação, não podemos confiar no score sem uma contextualização melhor de quando o 10 significa algo bom ou ruim.

Adicionando uma nova característica em nossos empréstimos

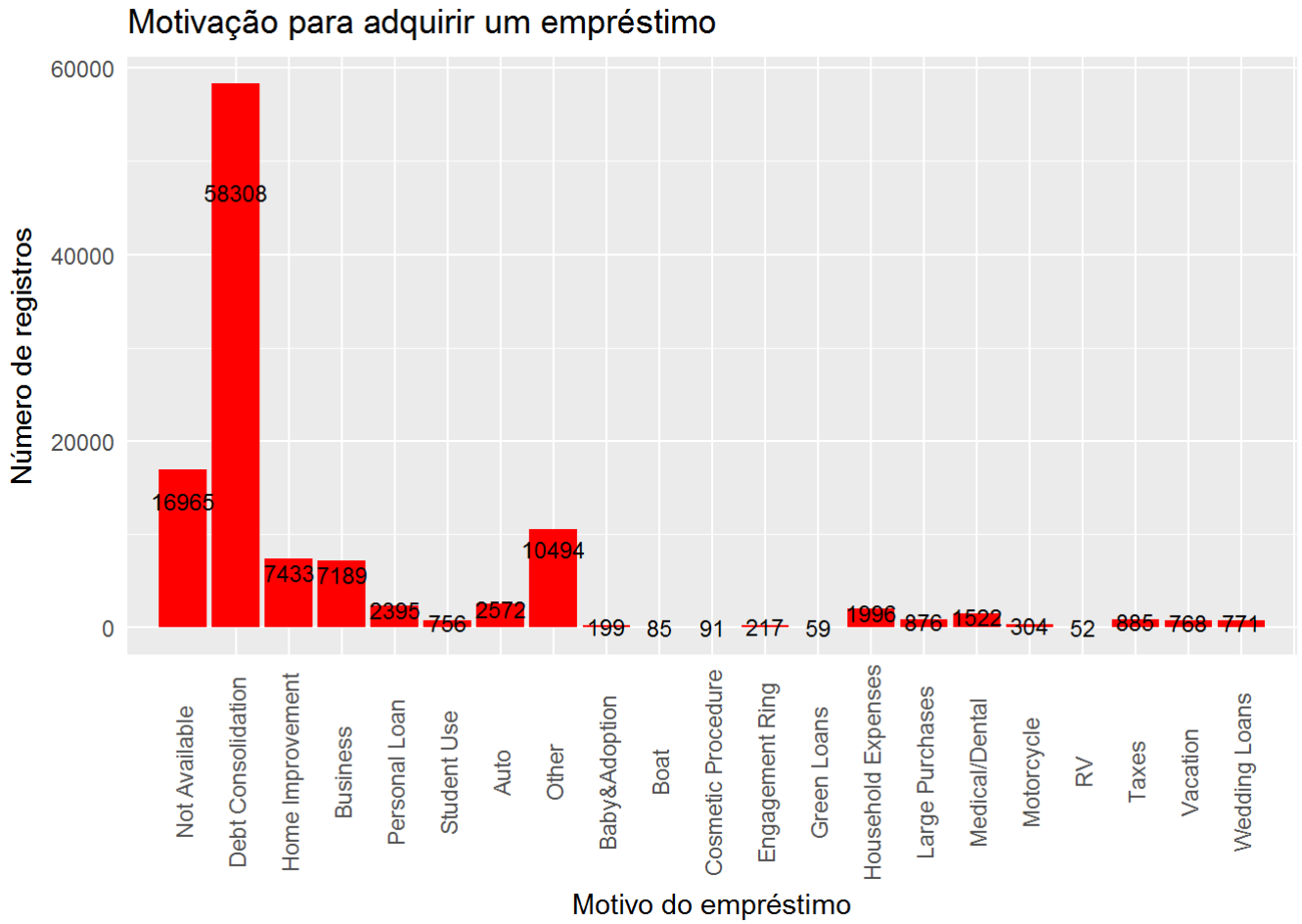
Como identificamos um grande número de inadimplentes separados em diferentes categorias, será criada a variável “inadimplente”, a fim de identificar com mais facilidade um mutuário inadimplente e um bom pagador.





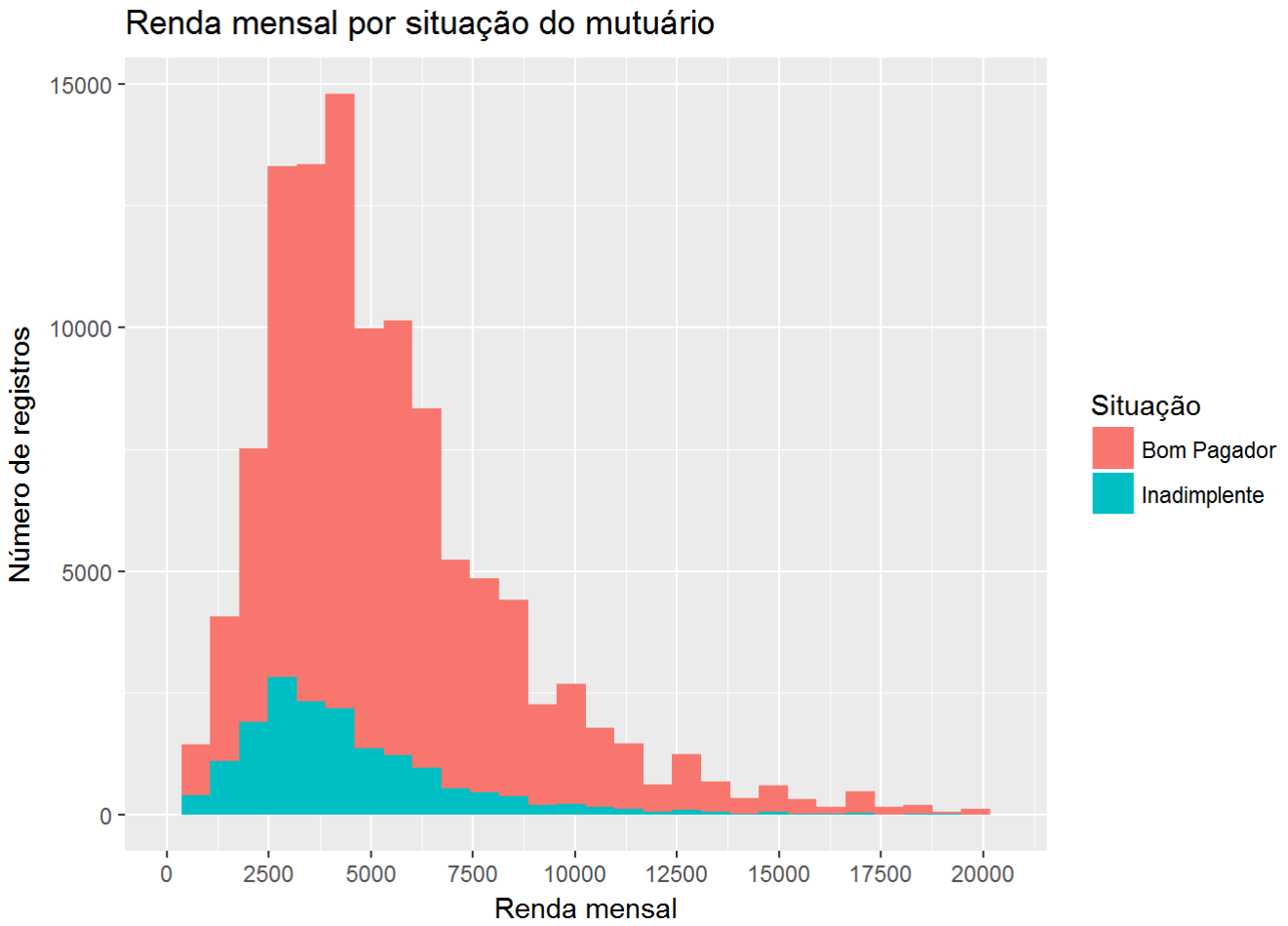
Aqui podemos visualizar a grande concentração de empréstimos na Califórnia com exatos 12.351 mutuários bons pagadores e 2.366 inadimplentes. Acredito que essa grande concentração é devido a sede da empresa está situada na Califórnia, onde a empresa deve ter focado grandes esforços em marketing no seu início,

levando uma grande confiança tanto dos mutuários quanto dos credores. Outra observação importante é que as regiões central e norte do país possuem poucos mutuários, acredito que isso se deve por ser uma região de pouca população, onde grande parte da economia familiar vem da agropecuária.

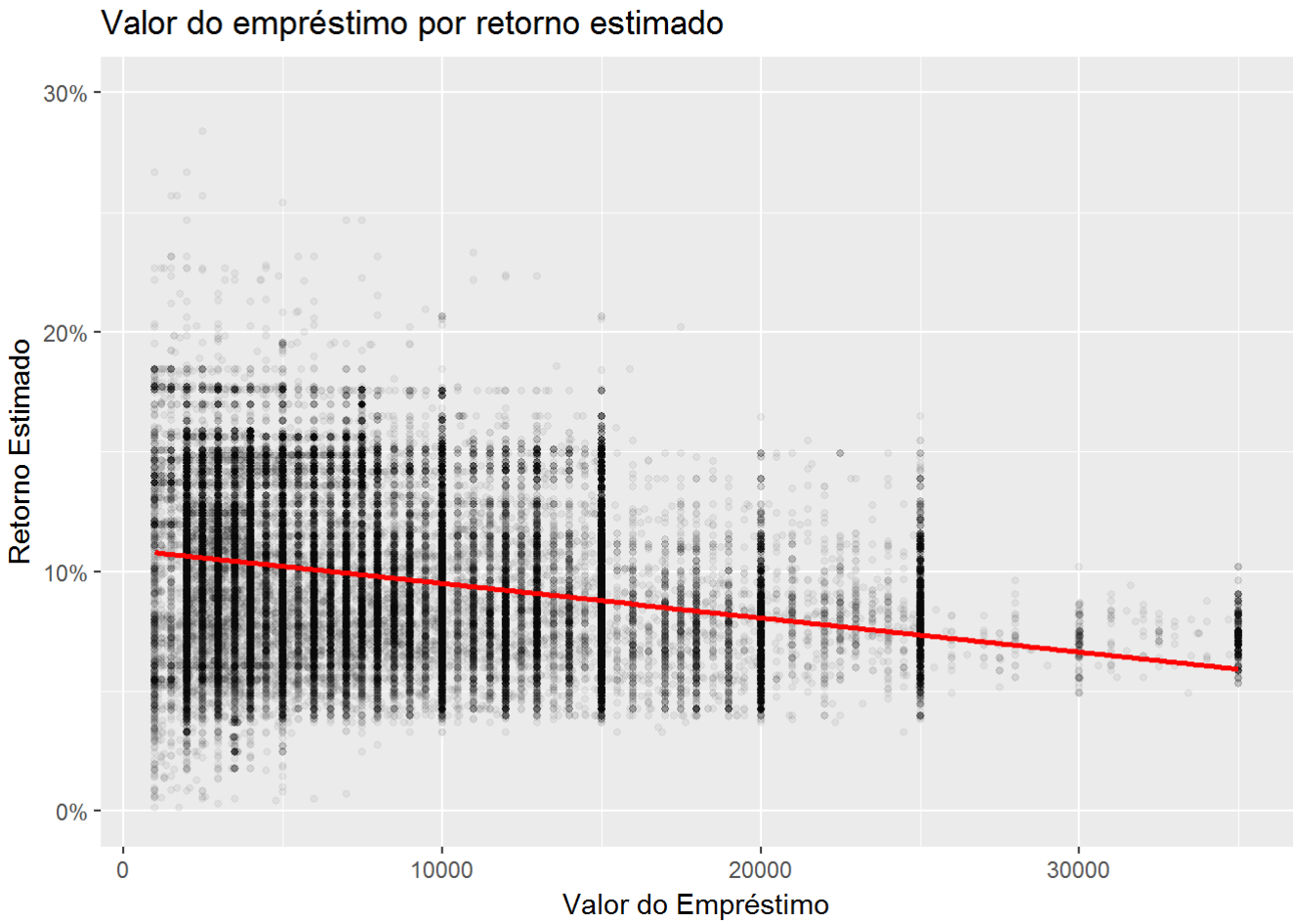


A grande maioria dos empréstimos são para consolidação de dívidas. Isso é algo interessante, porque as pessoas se endividam para pagar outras dívidas, entrando em um círculo vicioso para trocar de credor. Acredito que emprestar dinheiro para alguém com a finalidade de quitar outras dívidas é muito arriscado, se compararmos com “reformas de casa” e “negócios” que se destacam logo em seguida.

Análise Bivariada

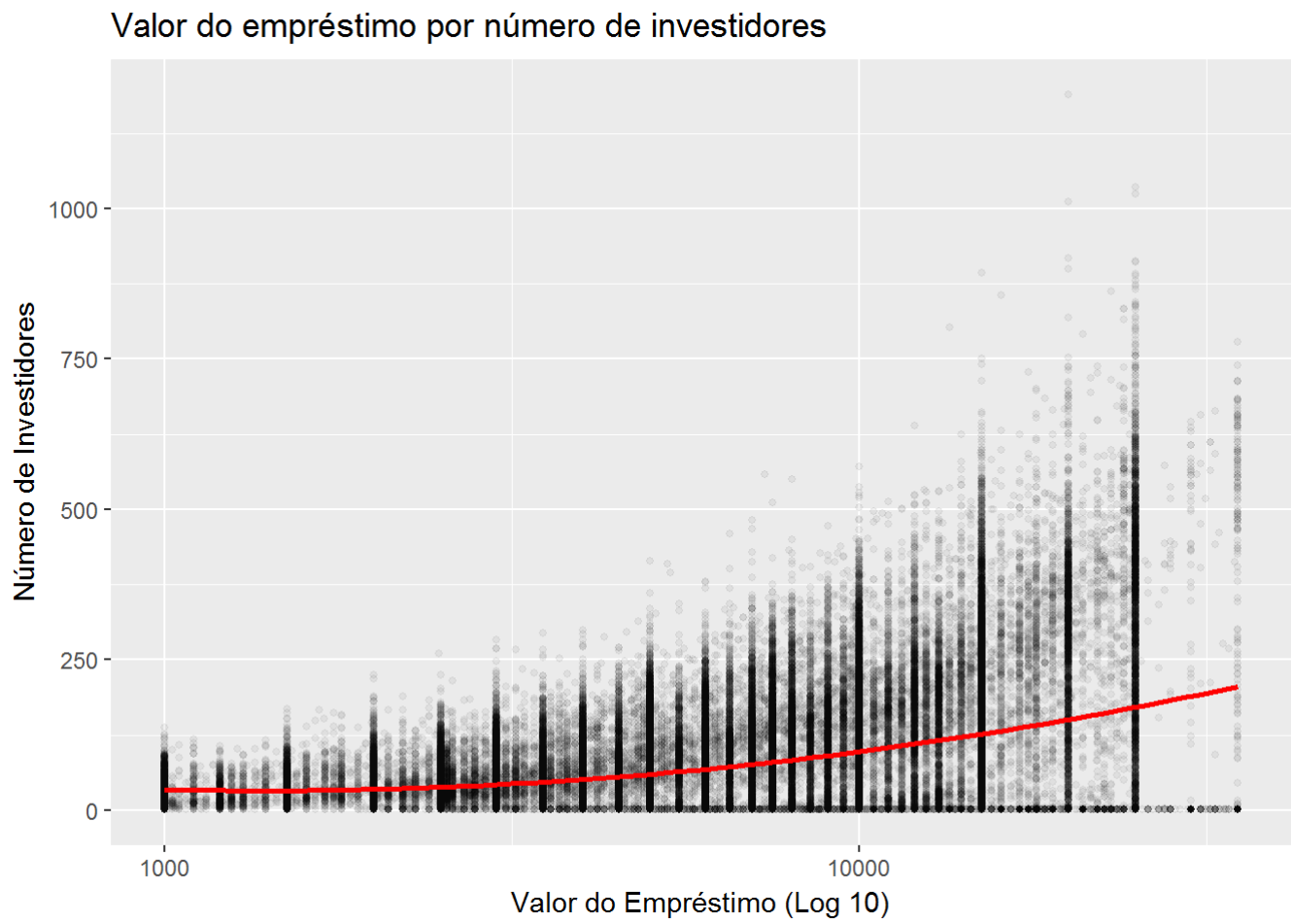


Como agora sabemos quando um mutuário está inadimplente, me surgiu uma dúvida de como ficaria a distribuição de renda mensal por situação. Com a visualização acima, fica claro que somente renda mensal alta não faz um mutuário um bom pagador. Outra observação importante é a confirmação de que a concentração de inadimplentes está próximo da moda da distribuição.



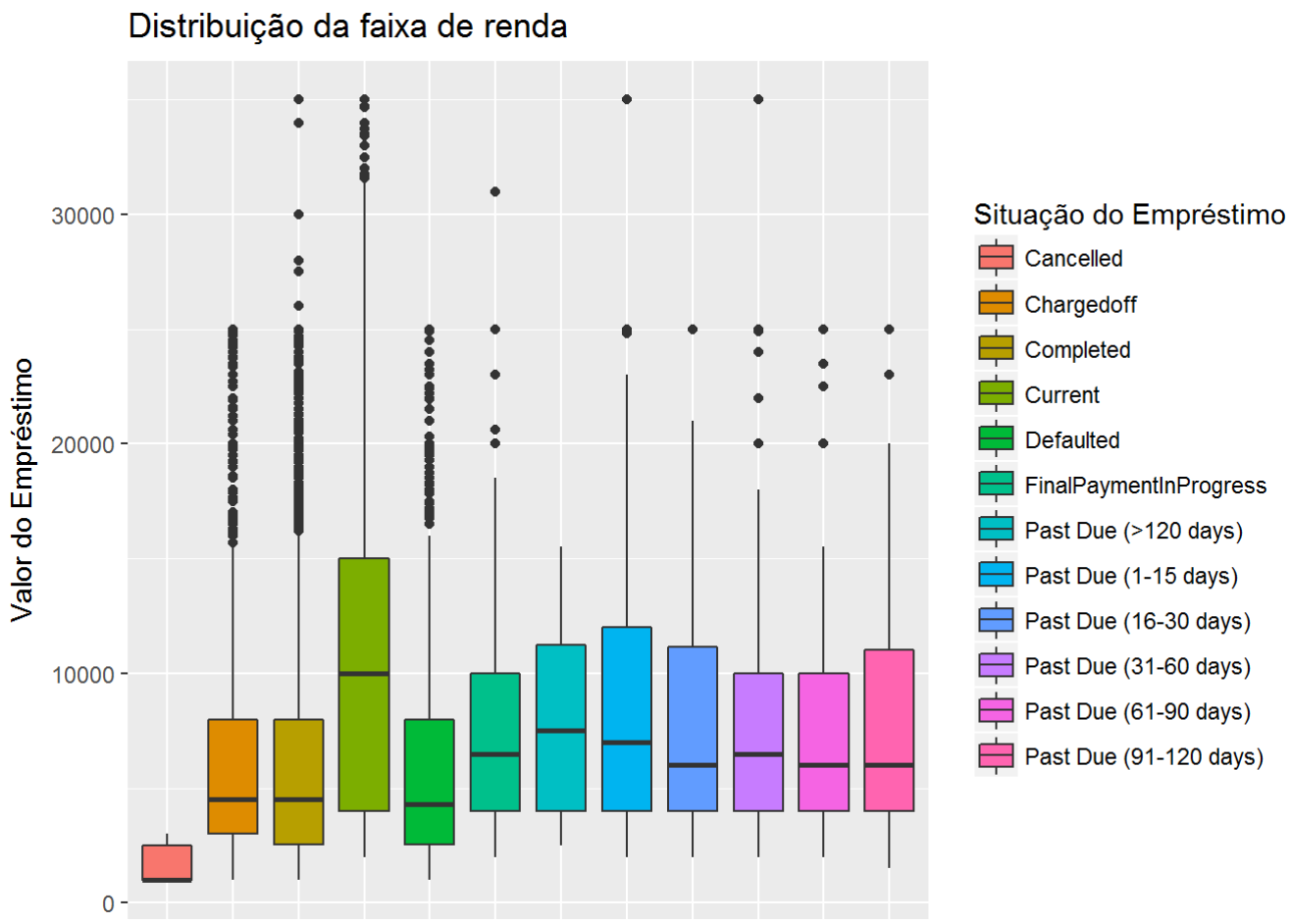

```
##
##  Pearson's product-moment correlation
##
## data:  LoanOriginalAmount and EstimatedReturn
## t = -86.98, df = 84851, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2922833 -0.2799279
## sample estimates:
##        cor
## -0.2861175
```

Aqui foi uma grande surpresa para mim, encontrei uma correlação negativa de -0.286 quando esperava que o retorno estimado fosse maior para os credores nos grandes empréstimos. Depois de uma pesquisa no site da Prosper, descobri que os investidores recebem um uma porcentagem maior para os empréstimos de alto risco, sendo assim, acredito que esses retornos mais altos são de transações de alto risco.



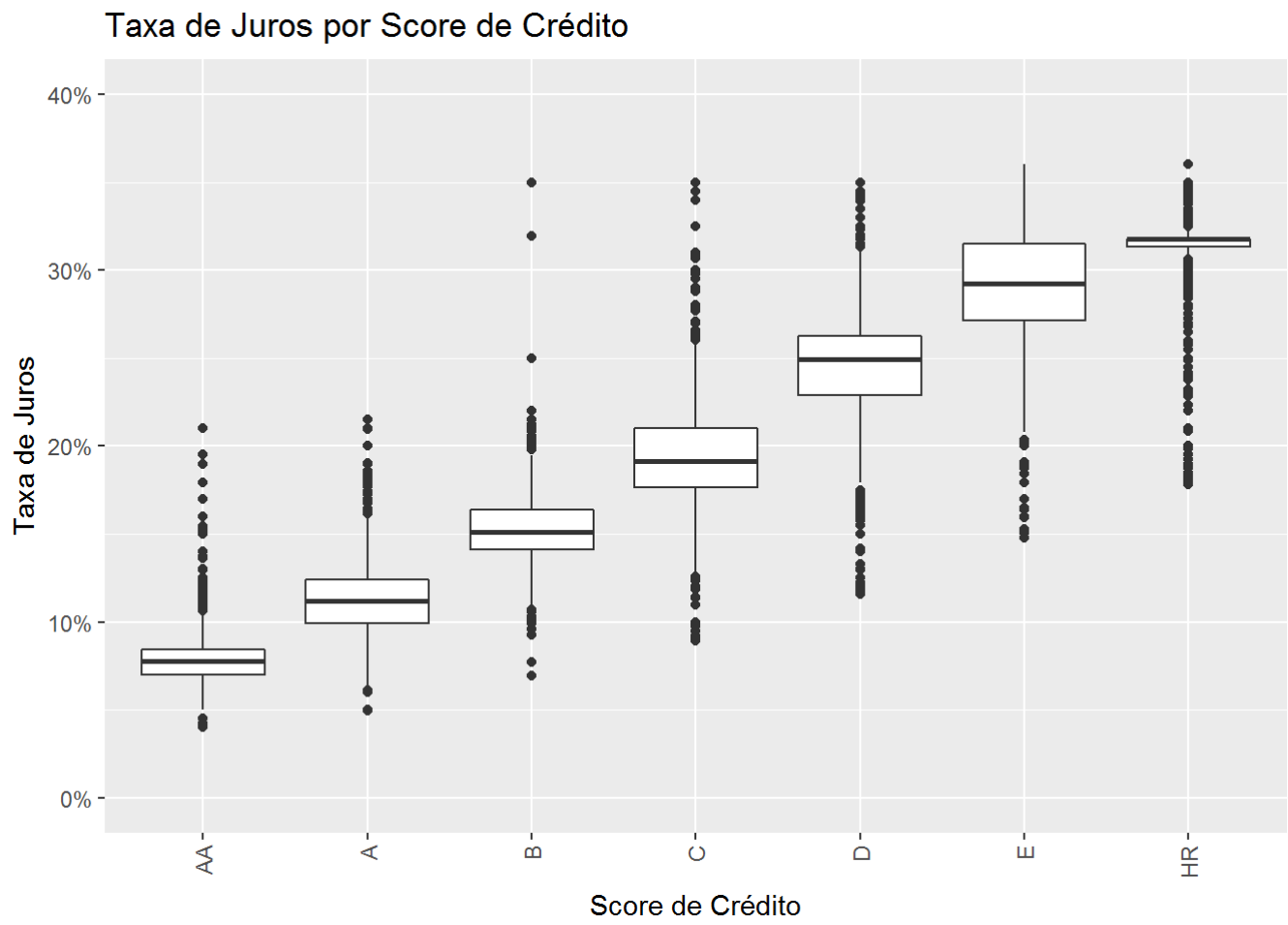

```
##
##  Pearson's product-moment correlation
##
## data:  LoanOriginalAmount and Investors
## t = 138.71, df = 113940, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3751140 0.3850494
## sample estimates:
##       cor
## 0.3800926
```

Aqui foi outra surpresa, esperava um crescimento linear entre investidores e valor do empréstimo, mas encontrei algo parecido com um crescimento polinomial. Percebe-se algumas linhas verticais que são explicadas pela correlação positiva fraca de 0.38.


Distribuição da faixa de renda
Valor do Empréstimo
30000
20000
10000
0
Situação do Empréstimo
Cancelled
Chargedoff
Completed
Current
Defaulted
FinalPaymentInProgress
Past Due (>120 days)
Past Due (1-15 days)
Past Due (16-30 days)
Past Due (31-60 days)
Past Due (61-90 days)
Past Due (91-120 days)

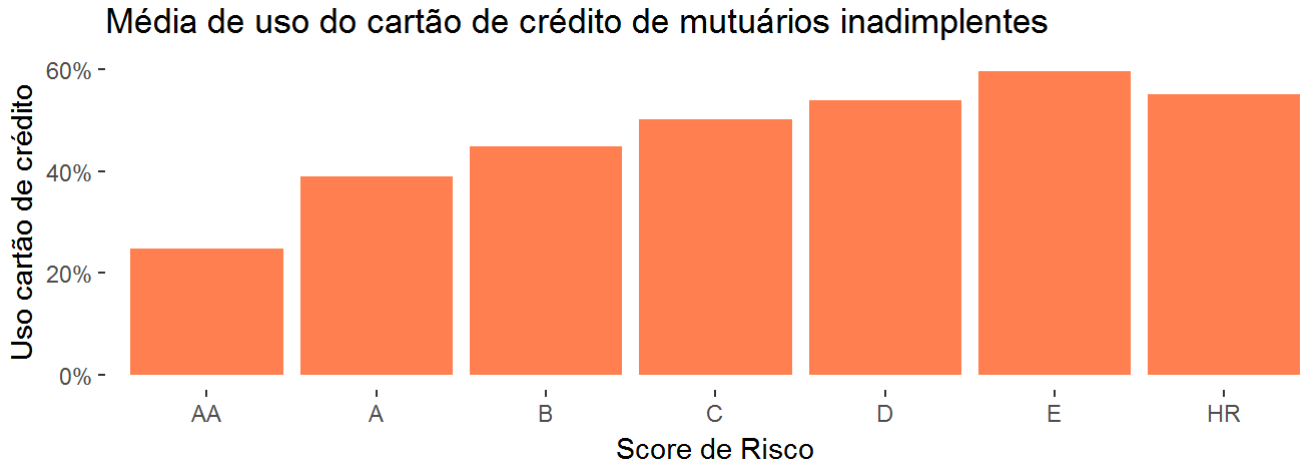
```
## LoanStatus: Cancelled
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    1000    1000    1700    2500    3000
## ------------------------------------------------------------
## LoanStatus: Chargedoff
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    3000    4500    6399    8000   25000
## ------------------------------------------------------------
## LoanStatus: Completed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    2550    4500    6189    8000   35000
## ------------------------------------------------------------
## LoanStatus: Current
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    4000   10000   10361   15000   35000
## ------------------------------------------------------------
## LoanStatus: Defaulted
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    2550    4275    6487    8000   25000
## ------------------------------------------------------------
## LoanStatus: FinalPaymentInProgress
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    4000    6500    8346   10000   31000
## ------------------------------------------------------------
## LoanStatus: Past Due (>120 days)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2500    4000    7500    8281   11250   15500
## ------------------------------------------------------------
## LoanStatus: Past Due (1-15 days)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    4000    7000    8468   12000   35000
## ------------------------------------------------------------
## LoanStatus: Past Due (16-30 days)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    4000    6000    8156   11129   25000
## ------------------------------------------------------------
## LoanStatus: Past Due (31-60 days)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    4000    6500    8534   10000   35000
## ------------------------------------------------------------
## LoanStatus: Past Due (61-90 days)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    4000    6000    7730   10000   25000
## ------------------------------------------------------------
## LoanStatus: Past Due (91-120 days)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1500    4000    6000    8004   11000   25000
```

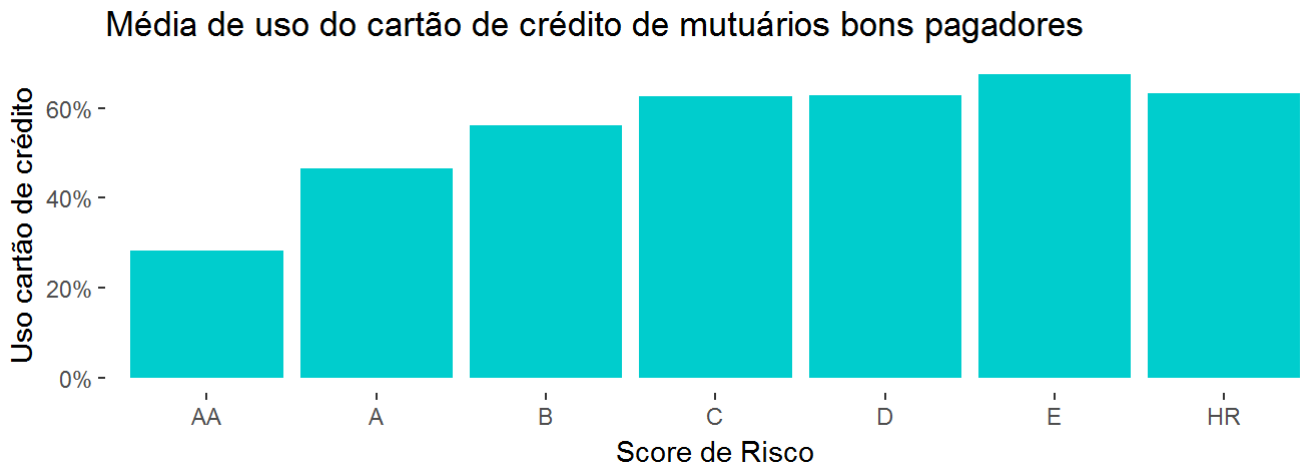
A maioria das categorias possuem o 1º quartil igual, com algumas exceções. Outra observação interessante é que na maioria das categorias de mutuários inadimplentes, o 3 quartil é maior ou igual a 10 mil dolares.




```
## ProsperRating: A
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0498  0.0990  0.1119  0.1129  0.1239  0.2150
## ------------------------------------------------------------
## ProsperRating: AA
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04000 0.06990 0.07790 0.07912 0.08450 0.21000
## ------------------------------------------------------------
## ProsperRating: B
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0693  0.1414  0.1509  0.1545  0.1639  0.3500
## ------------------------------------------------------------
## ProsperRating: C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0895  0.1765  0.1914  0.1944  0.2099  0.3500
## ------------------------------------------------------------
## ProsperRating: D
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1157  0.2287  0.2492  0.2464  0.2625  0.3500
## ------------------------------------------------------------
## ProsperRating: E
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1479  0.2712  0.2925  0.2933  0.3149  0.3600
## ------------------------------------------------------------
## ProsperRating: HR
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1779  0.3134  0.3177  0.3173  0.3177  0.3600
```

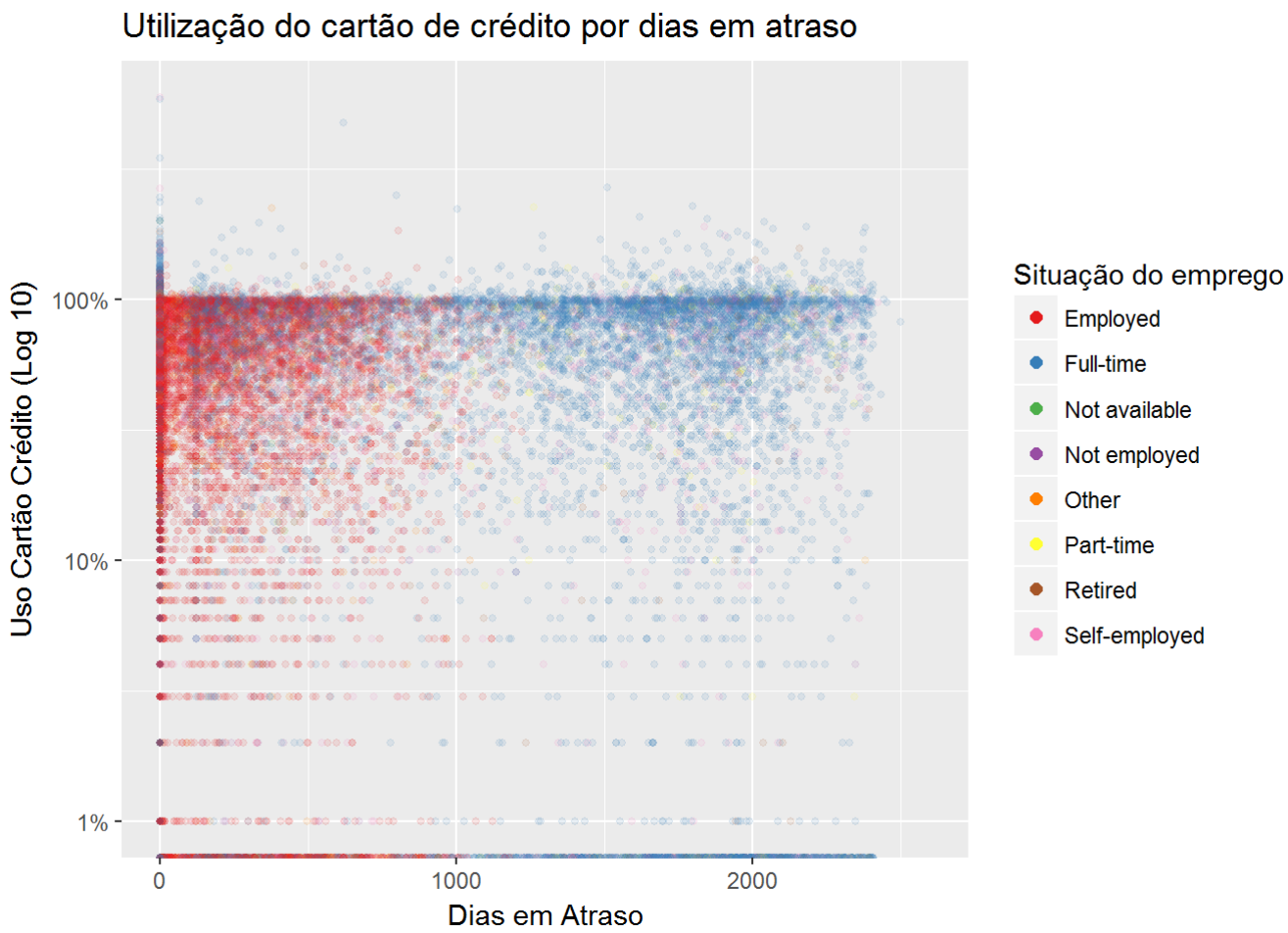
Na visualização acima podemos perceber um crescimento linear, isso já era algo esperado, porque a taxa de juros aumenta a medida que o nível de risco também aumenta. Nós também percebemos um IQR muito baixo no nível de risco mais alto(HR). Outra descoberta nessa visualização é que todos os Scores possuem a média e mediana muito próximas, o que nos leva a inferir uma simetria na distribuição.



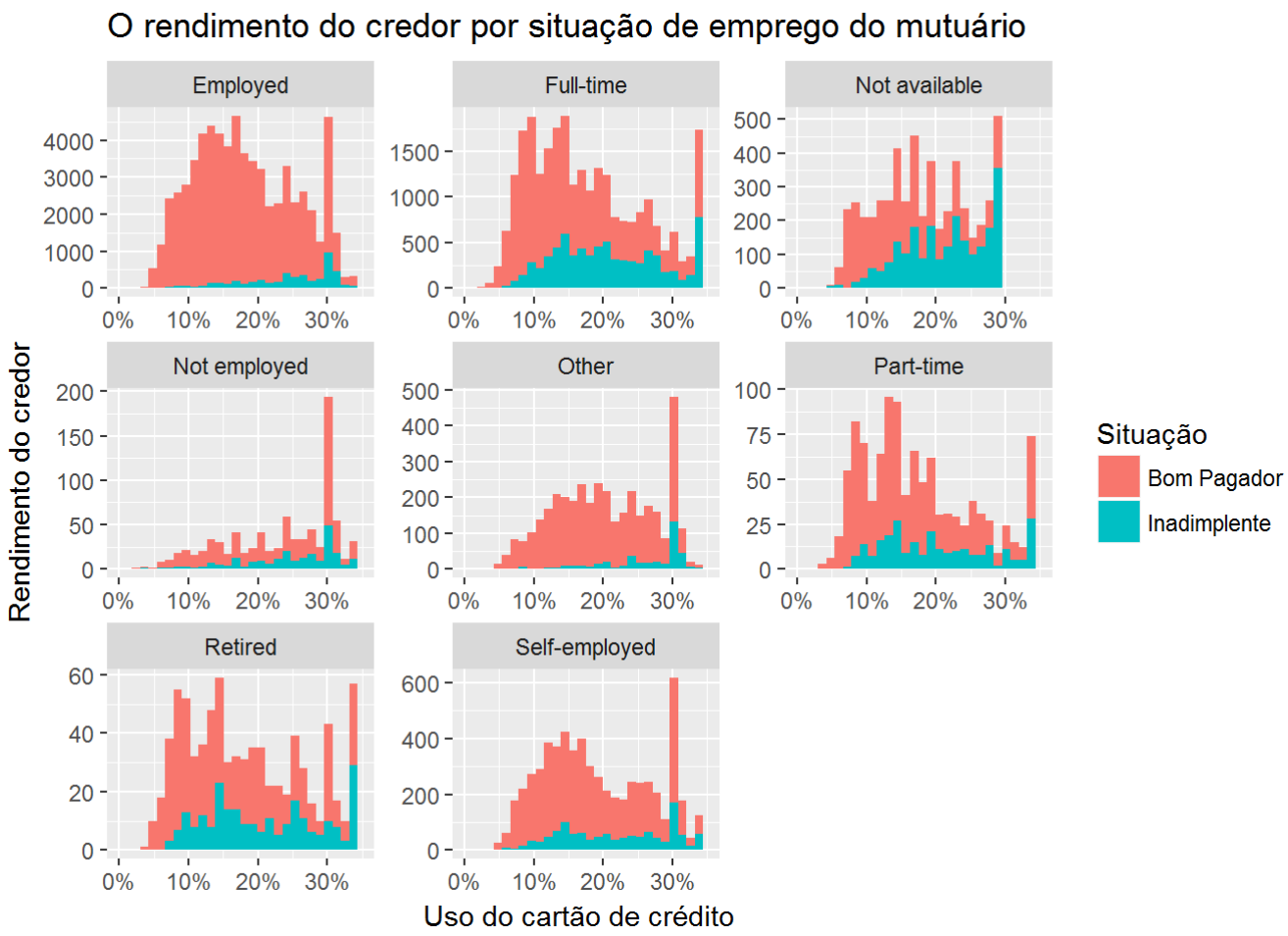


Na visualização acima examinei a média de uso do cartão de crédito por score de risco. Podemos perceber uma tendência entre as duas variáveis, o uso do cartão de crédito aumenta na medida que o score de risco do mutuário também aumenta. Os mutuários de maior risco utilizam em média mais de 50% do crédito rotativo disponível nos cartões, mas identificamos que essa tendência ocorre tanto para os inadimplente quanto para os bons pagadores.

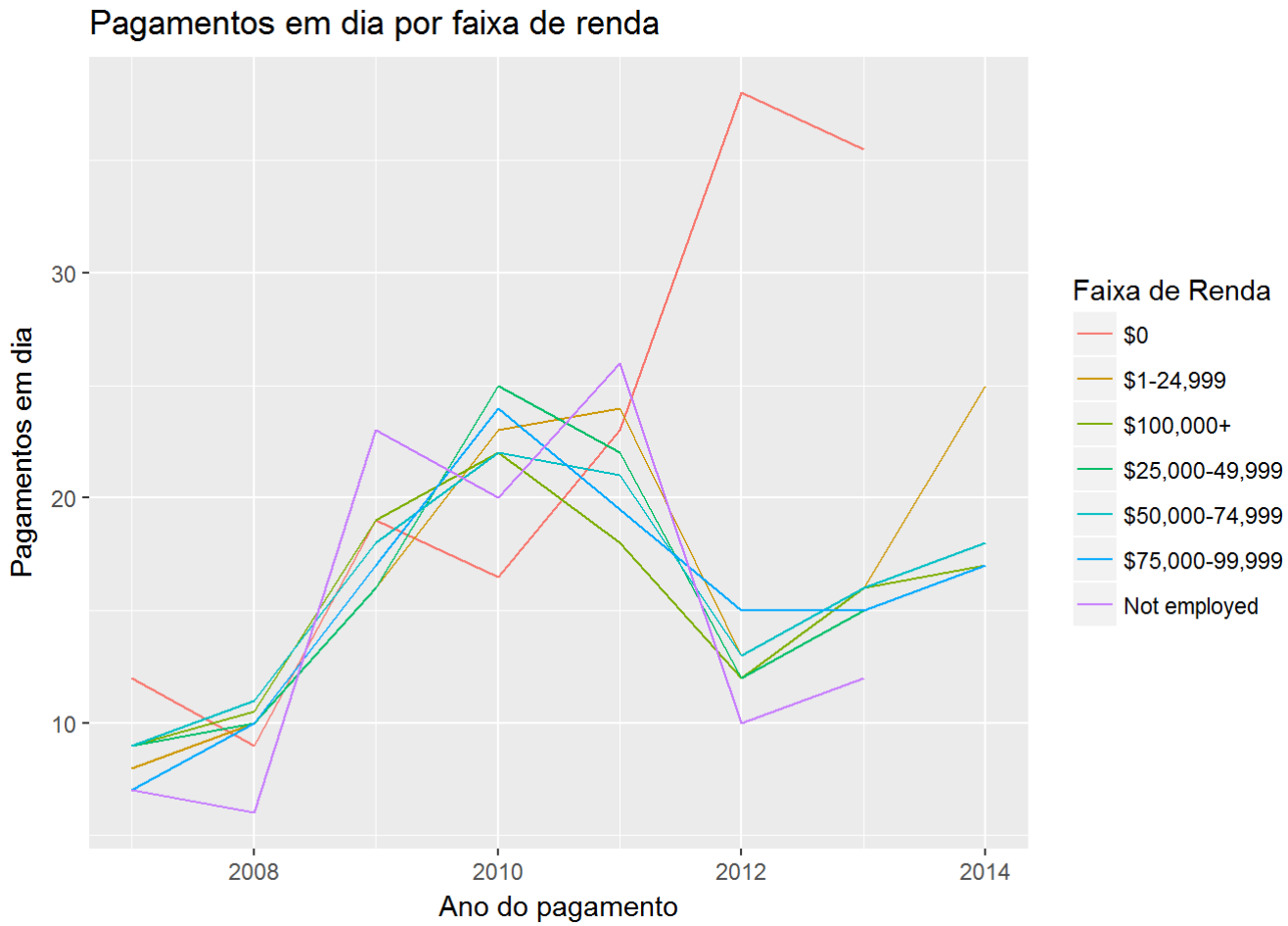
Análise Multivariada



Na visualização acima, eu esperava um crescimento do uso do cartão à medida que os dias em atraso fossem aumentando, mas muitos mutuários quando se tornam inadimplentes, já estão com o limite do cartão estourado. Percebemos que os mutuários que mais utilizam o cartão estão empregados, e isso faz bastante sentido, o estranho aqui é que em certo ponto no tempo, a predominância muda de mutuários com o status de “Employed” para “Full-Time”.



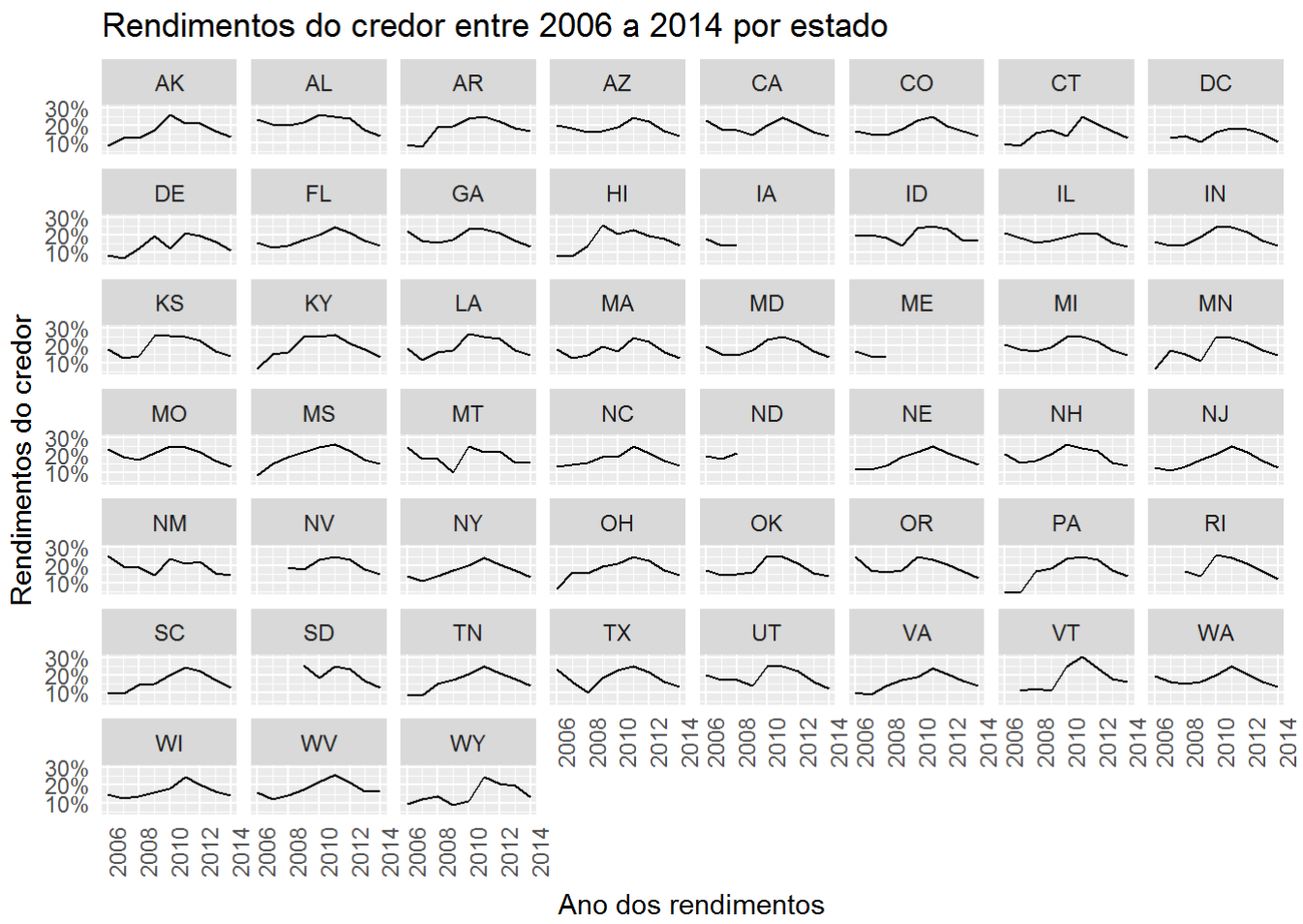
Decidi investigar o retorno do credor para cada situação de emprego do mutuário e encontrei algumas situações interessantes. Aqui identificamos que todas as situações de emprego possuem um pico elevado na casa dos 30% e os mutuários “Retired” são os únicos que possuem 2 picos consideráveis na casa dos 30%. Os mutuários “Full-time” e “Not available” possuem elevadas parcelas de inadimplentes comparado aos bons pagadores.



Podemos visualizar que mesmo no período da crise financeira de 2008, houve um crescimento constante no número de pagamentos em dia, mas logo após o ano de 2010, houve uma queda acentuada nos pagamentos. Outro ponto de interessante é o número de pagamentos crescente entre mutuários desempregados e com renda $0. Aqui poderia ter várias explicações possíveis para isso, mas acredito que estes mutuários devem ter conseguido emprego logo após os empréstimos ou podem ter rendas informais.

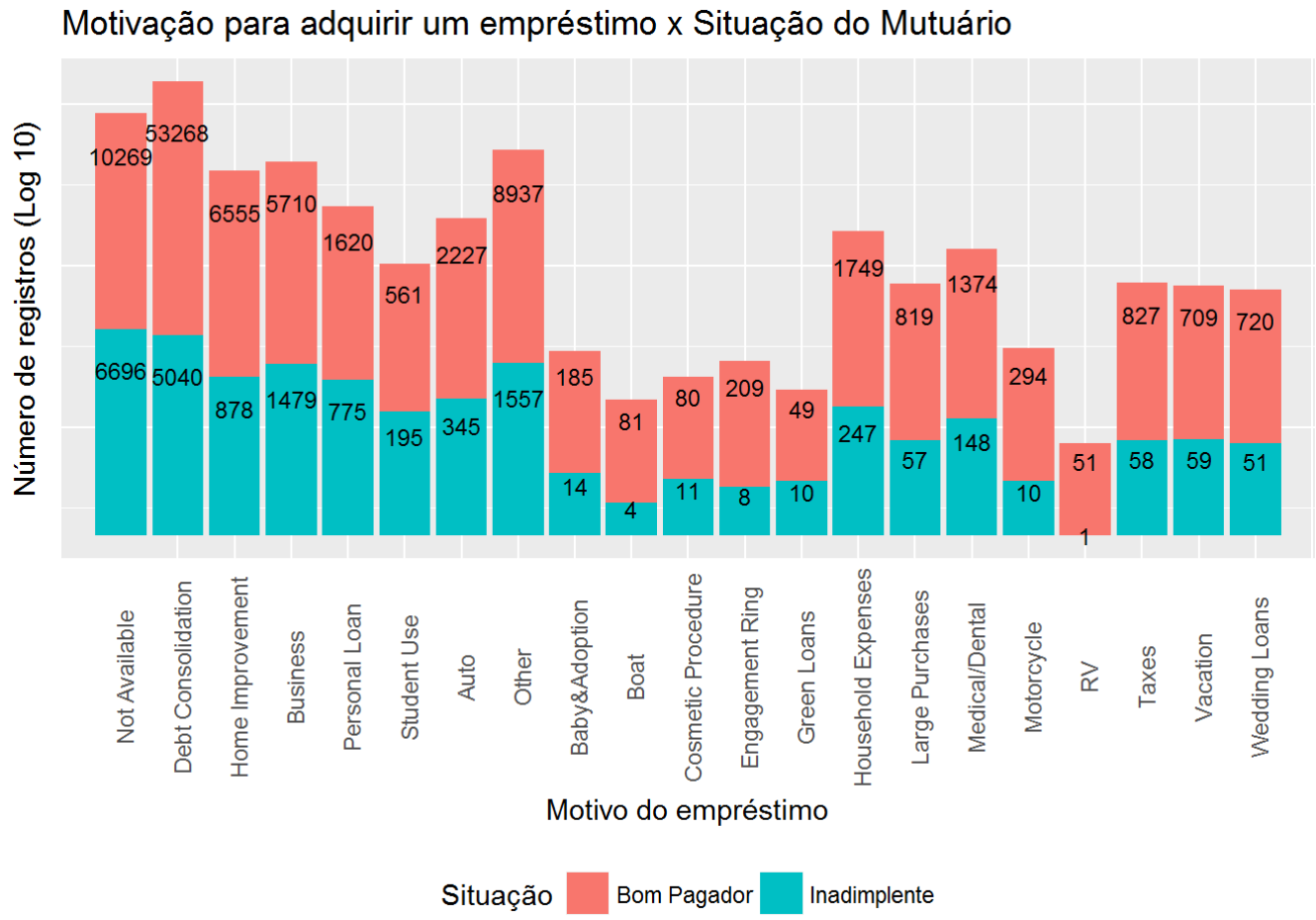
Gráficos finais e sumário.

Primeira Visualização



Na análise univariada visualizamos um mapa demonstrando a concentração mutuários inadimplentes e bons pagadores em seus respectivos estados. Isso me deu curiosidade em saber como seria os rendimentos dos credores separados por estado, para isso, criei a visualização acima. Visualizamos um crescimento constante dos rendimentos na maioria dos estados entre 2006 e 2010, com exceção de alguns estados como Montana(MT) e Novo México(NM), onde estavam diminuindo os rendimentos e depois voltaram a subi em 2010. A explicação desse pico em 2010, acredito estar relacionado com a crise mundial de 2008, onde muitos mutuários ficaram endividados/desempregados e recorreram aos empréstimos, sendo categorizados como mutuários de alto risco. A queda após 2010 deve estar relacionado com a leve melhora da economia, onde muitos investidores encontravam os mutuários já em uma situação financeira/emprego mais estável, o que levava a uma categorização de menor risco e consequentemente menor rendimento.

Segunda visualização

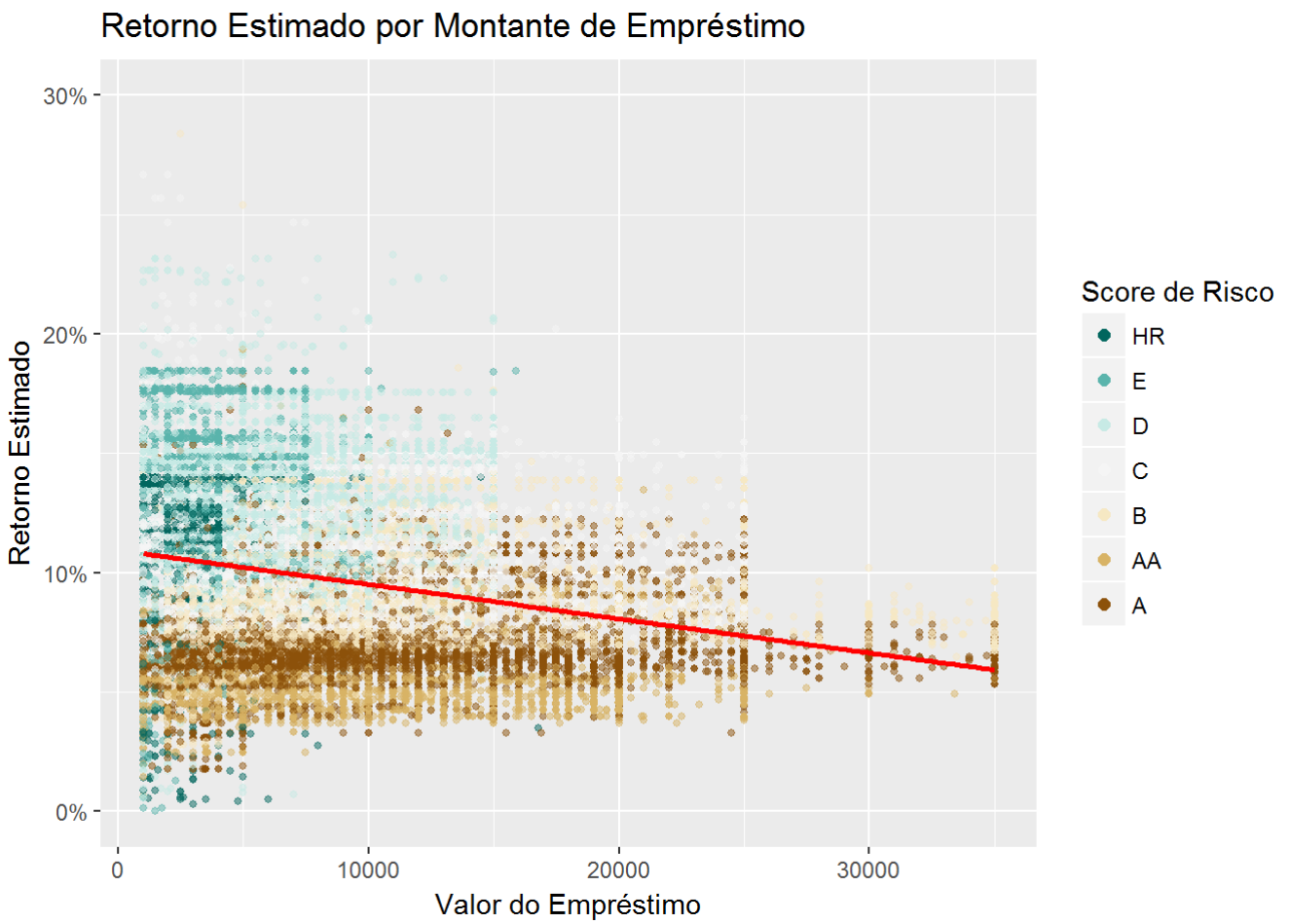



```
## [1] "Situação de emprego dos mutuários que solicitaram empréstivo para RV"
```

```
##
##      Employed     Full-time Not available  Not employed         Other
##            42             2             0             0             6
##     Part-time       Retired Self-employed
##             0             0             2
```

Aqui o propósito é identificar a situação do mutuário para cada categoria de solicitação do empréstimo que vimos anteriormente. Eu imaginava encontrar uma categoria que tivesse mais inadimplentes que bons pagadores, mas percebemos que não existe tal situação. O que me chamou mais atenção foi que emprestar dinheiro para (RV) e (boat) é um bom negócio. Com uma pesquisada rápida no google, descobri que “RV” são aqueles “ônibus casa”, isso faz muito sentido, porque normalmente quem compra esse tipo de veículo está com uma vida financeira estável, o que foi devidamente comprovado na segunda visualização. O item que mais me chamava atenção era “Baby&Adoption”, porque é um momento em que as despesas crescem muito na vida de um casal, mas pelo que vimos, os mutuários dessa categoria são, em grande maioria, bons pagadores.

Terceira visualização




```
##
##  Pearson's product-moment correlation
##
## data:  LoanOriginalAmount and EstimatedReturn
## t = -86.98, df = 84851, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2922833 -0.2799279
## sample estimates:
##        cor
## -0.2861175
```

Como tinha suspeitado na análise bivariada, nossa correlação negativa está diretamente relacionada com a concentração de transações de alto risco próximo ao início do eixo X, o que gera um maior retorno para os investidores. Está relação também se explica na análise univariada, quando vimos que a moda do valor original do emprestimo fica próximo de $ 5.000. Outro fator que pode ser levado em conta é o fato dos investidores não puderem emprestar mais do que 10% do seu patrimônio líquido.

Reflexão

Analisar o dataset da Prosper foi algo muito interessante. Navegar entre os 113.937 empréstimos com 81 variáveis foi algo desafiador, por isso separei 32 variáveis de interesse e mesmo assim não consegui utilizar todas como deveria. O maior desafio aqui, foi entender todas essas informações sendo um leigo no assunto da área de empréstimos peer-to-peer. Me concentrei em tentar identificar padrões entre mutuários inadimplentes/bons pagadores e os padrões que afetam o rendimento dos credores.

Tive sucesso ao identificar o motivo da queda no retorno estimado dos credores, devido aos investimentos iniciais estarem diretamente relacionado a empréstimos de alto risco. Outra grande descoberta foi um crescimento nos pagamentos em dia durante a crise financeira de 2008, principalmente para mutuários que declararam renda de $0 e desempregados. A maior surpresa foi quando descobri que um mutuário com uma renda mensal de quase 2 milhões, solicitou um empréstimo de 4 mil dolares. Além dessas grandes descobertas, fiquei um pouco decepcionado com a relação entre o número investidores e valor do empréstimo. Acreditava que encontraria um crescimento linear, mas achei um crescimento polinomial.

Acredito que o trabalho foi interessante, com grandes descobertas em varios pontos, mas sei que ficaram alguns pontos sem explicação devido a quantidade de variáveis analisadas, pouco tempo e falta de informações importantes como: estado civil, sexo, situação de emprego do cônjugue, ou seja, informações mais pessoais dos mutuários. Para trabalhos futuros, penso em fechar o estudo com as variáveis que não foram exploradas, a fim de desenvolver alguns modelos de regressão para tentar prever quando uma solicitação de empréstimo é um bom investimento.

Referências

- https://stackoverflow.com/questions/45107302/error-in-seq-lennrowdata-1-argument-must-be-coercible-to-non-negative-in?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa (https://stackoverflow.com/questions/45107302/error-in-seq-lennrowdata-1-argument-must-be-coercible-to-non-negative-in?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa)
- https://stackoverflow.com/questions/22458970/how-to-reverse-legend-labels-and-color-so-high-value-starts-downstairs?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa (https://stackoverflow.com/questions/22458970/how-to-reverse-legend-labels-and-color-so-high-value-starts-downstairs?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa)
- https://stackoverflow.com/questions/20936840/r-get-longitude-latitude-data-for-cities-and-add-it-to-my-dataframe?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa (https://stackoverflow.com/questions/20936840/r-get-longitude-latitude-data-for-cities-and-add-it-to-my-dataframe?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa)
- https://s3.amazonaws.com/udacity-hosted-downloads/ud651/GeographyOfAmericanMusic.html (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/GeographyOfAmericanMusic.html)
- https://s3.amazonaws.com/udacity-hosted-downloads/ud651/AtlanticHurricaneTracking.html (https://s3.amazonaws.com/udacity-hosted-downloads/ud651/AtlanticHurricaneTracking.html)
- https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/diamondsExample_2016-05.html (https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/diamondsExample_2016-05.html)
- http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#Treemap (http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html#Treemap)
- https://stackoverflow.com/questions/5411979/state-name-to-abbreviation-in-r?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa (https://stackoverflow.com/questions/5411979/state-name-to-abbreviation-in-r?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa)
- https://stackoverflow.com/questions/6644997/showing-data-values-on-stacked-bar-chart-in-ggplot2?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa (https://stackoverflow.com/questions/6644997/showing-data-values-on-stacked-bar-chart-in-ggplot2?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa)
- https://www.prosper.com/plp/legal/compliance/ (https://www.prosper.com/plp/legal/compliance/)
- https://stackoverflow.com/questions/37329074/geom-smooth-and-exponential-fits (https://stackoverflow.com/questions/37329074/geom-smooth-and-exponential-fits)
- https://www.r-bloggers.com/multiple-legends-for-the-same-aesthetic-2/ (https://www.r-bloggers.com/multiple-legends-for-the-same-aesthetic-2/)