# Dados de Empréstimos da Prosper - Análise Exploratória de Dados (EDA)

*Marcio Ozorio de Jesus*

E-mail: marcio.ozorio@gmail.com (mailto:marcio.ozorio@gmail.com)
LinkedIn: https://www.linkedin.com/in/marciojesus/ (https://www.linkedin.com/in/marciojesus/)

## INTRODUÇÃO

Este estudo faz parte do curso de Análise de Dados da Udacity (http://udacity.com) onde será utilizada a base de dados da empresa de Prosper (https://www.prosper.com). A Prosper foi fundada em 2005 e foi a primeira plataforma de serviço de empréstimos ponto a ponto (peer-to_peer) dos Estados Unidos com mais de US $ 7 bilhões em empréstimos financiados.

Os mutuários (recebedores do empréstimo) solicitam empréstimos através da Prosper e os investidores (pessoas Físicas ou Jurídicas) podem financiar quantias entre US $ 1.000 a US $ 35.000.

Os mutuários fazem a solicitações de empréstimo, a Prosper verifica a identidade, seleciona dados pessoais e executa um algoritimo que analisa o risco de crédito e retorna a taxa a ser utilizada. Os investidores analisam o relatório de crédito do mutuário com a taxa de juros calculada pela Prosper e decidem se irão ou não conceder o empréstimo. A Prosper lida com o serviço de coleta e distribui o pagamento de mutuários juntamente com os juros de volta aos investidores dos empréstimos, gerenciando todas as etapas do processo. A Prosper gera receita cobrando uma taxa sobre os empréstimos financiados.

Obs.: Irei trabalhar com a lingua padrão da base de dados que é a lingua inglesa e adicionarei comentários em português.

## INFORMAÇÕES DA BASE DE DADOS

Temos um total 81 colunas ou variáveis e 113.937 observações (registros).

Abaixo segue o nome de todas as colunas de nossa base:

```
##  [1] "ListingKey"
##  [2] "ListingNumber"
##  [3] "ListingCreationDate"
##  [4] "CreditGrade"
##  [5] "Term"
##  [6] "LoanStatus"
##  [7] "ClosedDate"
##  [8] "BorrowerAPR"
##  [9] "BorrowerRate"
## [10] "LenderYield"
## [11] "EstimatedEffectiveYield"
## [12] "EstimatedLoss"
## [13] "EstimatedReturn"
## [14] "ProsperRating..numeric."
## [15] "ProsperRating..Alpha."
## [16] "ProsperScore"
## [17] "ListingCategory..numeric."
## [18] "BorrowerState"
## [19] "Occupation"
## [20] "EmploymentStatus"
## [21] "EmploymentStatusDuration"
## [22] "IsBorrowerHomeowner"
## [23] "CurrentlyInGroup"
## [24] "GroupKey"
## [25] "DateCreditPulled"
## [26] "CreditScoreRangeLower"
## [27] "CreditScoreRangeUpper"
## [28] "FirstRecordedCreditLine"
## [29] "CurrentCreditLines"
## [30] "OpenCreditLines"
## [31] "TotalCreditLinespast7years"
## [32] "OpenRevolvingAccounts"
## [33] "OpenRevolvingMonthlyPayment"
## [34] "InquiriesLast6Months"
## [35] "TotalInquiries"
## [36] "CurrentDelinquencies"
## [37] "AmountDelinquent"
## [38] "DelinquenciesLast7Years"
## [39] "PublicRecordsLast10Years"
## [40] "PublicRecordsLast12Months"
## [41] "RevolvingCreditBalance"
## [42] "BankcardUtilization"
## [43] "AvailableBankcardCredit"
## [44] "TotalTrades"
## [45] "TradesNeverDelinquent..percentage."
## [46] "TradesOpenedLast6Months"
## [47] "DebtToIncomeRatio"
## [48] "IncomeRange"
## [49] "IncomeVerifiable"
## [50] "StatedMonthlyIncome"
## [51] "LoanKey"
## [52] "TotalProsperLoans"
## [53] "TotalProsperPaymentsBilled"
## [54] "OnTimeProsperPayments"
## [55] "ProsperPaymentsLessThanOneMonthLate"
## [56] "ProsperPaymentsOneMonthPlusLate"
## [57] "ProsperPrincipalBorrowed"
```

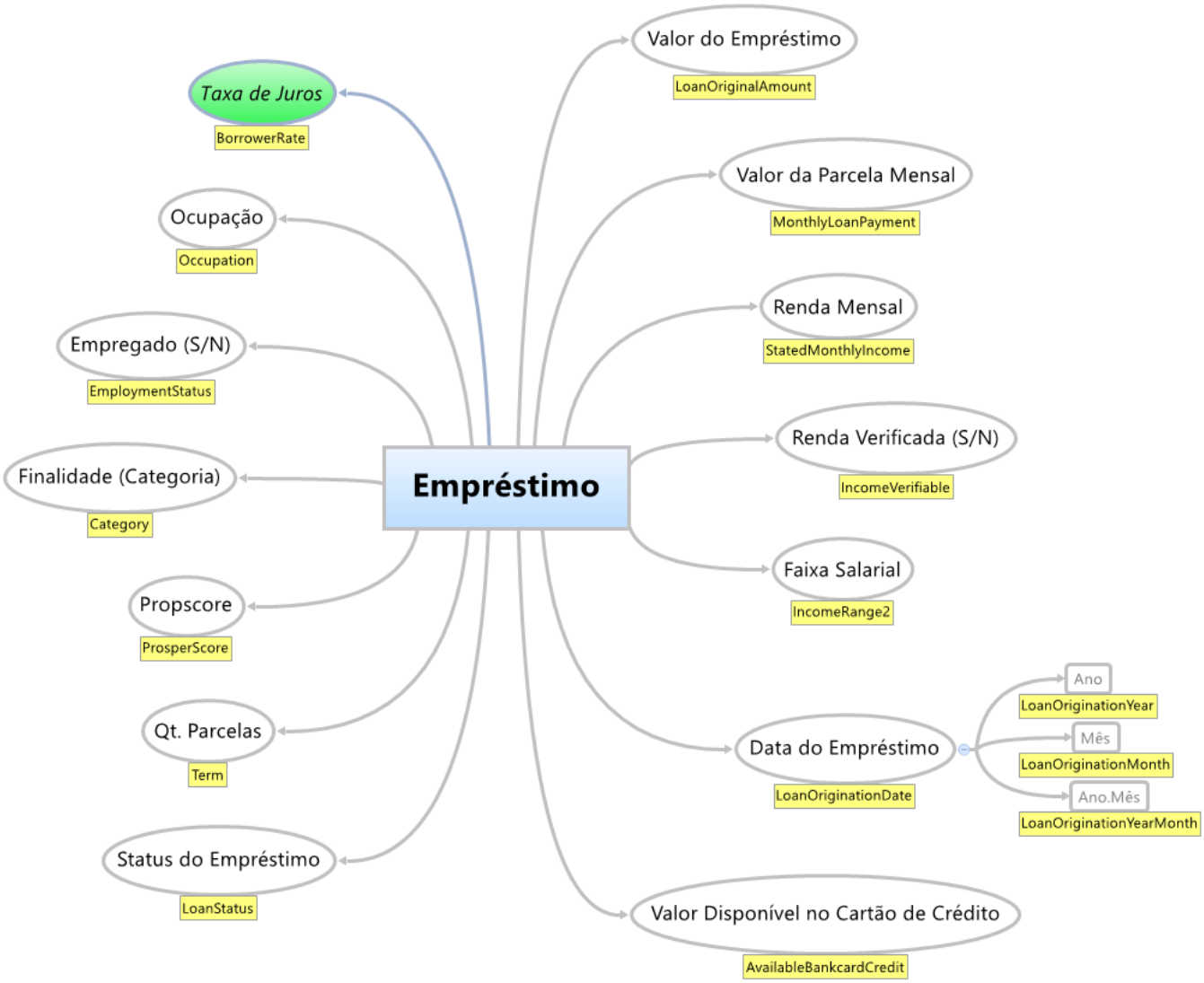
```
## [58] "ProsperPrincipalOutstanding"
## [59] "ScorexChangeAtTimeOfListing"
## [60] "LoanCurrentDaysDelinquent"
## [61] "LoanFirstDefaultedCycleNumber"
## [62] "LoanMonthsSinceOrigination"
## [63] "LoanNumber"
## [64] "LoanOriginalAmount"
## [65] "LoanOriginationDate"
## [66] "LoanOriginationQuarter"
## [67] "MemberKey"
## [68] "MonthlyLoanPayment"
## [69] "LP_CustomerPayments"
## [70] "LP_CustomerPrincipalPayments"
## [71] "LP_InterestandFees"
## [72] "LP_ServiceFees"
## [73] "LP_CollectionFees"
## [74] "LP_GrossPrincipalLoss"
## [75] "LP_NetPrincipalLoss"
## [76] "LP_NonPrincipalRecoverypayments"
## [77] "PercentFunded"
## [78] "Recommendations"
## [79] "InvestmentFromFriendsCount"
## [80] "InvestmentFromFriendsAmount"
## [81] "Investors"
```

Dicionário de dados com maiores detalhes:
https://docs.google.com/spreadsheets/d/1gDyi_L4UvIrLTEC6Wri5nbaMmkGmLQBk-Yx3z0XDEtI/edit#gid=0
(https://docs.google.com/spreadsheets/d/1gDyi_L4UvIrLTEC6Wri5nbaMmkGmLQBk-Yx3z0XDEtI/edit#gid=0)

Analisando as colunas de nossa base de dados, encontrou-se duas variáveis com valores nulos. As quantidades estão indicadas abaixo (de um total de 113.932). Apesar disso, os valores nulos não irão afetar a nossa Análise.

```
##   isnull                variable
## 4  29084            ProsperScore
## 7   7544 AvailableBankcardCredit
```

Conforme as instruções da Udacity, iremos selecionar entre 10 a 15 variáveis para este estudo. O processo de Análise das informações mais relevantes resultou no seguinte mapa mental contendo colunas já existentes, novas colunas que serão criadas e a Taxa de Juros que esta destacada na cor verde, pois terá um papel importante em nossa Análise.




Portanto, foi seleciondo 13 colunas da nossa base de dados. E com base nestas informaçoes foram adcionados mais 4 variáveis.

Segue uma tabela com todas as vaiáveis que iremos trabalhar:

| **Variáveis** | **Descrição** |
|---|---|
| Term | O prazo do empréstimo expresso em meses. |
| LoanStatus | O status atual do empréstimo: cancelado, carregado, concluído, atual, inadimplente, pagamento final em progresso, Vencido. O status Vencido será acompanhado por uma faixa de inadimplência. |
| BorrowerRate | A taxa de juros do mutuário para este empréstimo. |
| ProsperScore | Uma pontuação de risco personalizada construída usando dados históricos Prosper. A pontuação varia de 1-10, sendo 10 o melhor, ou menor pontuação de risco. Aplicável para empréstimos originados após julho de 2009. |
| Occupation | A Ocupação selecionada pelo mutuário no momento em que eles criaram a listagem. |

| Variáveis | Descrição |
|---|---|
| Category | 0 - Nao disponível, 1 - Dívida, 2 - Melhoria domiciliar, 3 - Negócios, 4 - Empréstimo pessoal, 5 - Uso do aluno, 6 - Auto, 7- Outro 8 - Bebê e Adopção, 9 - Barco, 10 - Procedimento Cosmético, 11 - Anel de Noivado, 12 - empréstimos Verdes, 13 - Despesas Domiciliares, 14 - Grandes Compras, 15 - Médico / Dental, 16 - Motocicleta, 17 - RV, 18 - Impostos, 19 - Férias, 20 - empréstimos de casamento |
| EmploymentStatus | O status de emprego do mutuário no momento em que publicaram a listagem. |
| AvailableBankcardCredit | O crédito total disponível via cartão bancário no momento em que o perfil de crédito foi puxado. |
| IncomeVerifiable | O mutuário indicou que eles têm a documentação necessária para suportar sua renda. |
| StatedMonthlyIncome | A renda mensal que o mutuário declarou no momento em que a lista foi criada. |
| LoanOriginalAmount | O montante original do empréstimo. |
| LoanOriginationDate | A data em que o empréstimo foi originado. |
| MonthlyLoanPayment | O pagamento mensal programado do empréstimo. |
| LoanOriginationYear | Ano da data em que o empréstimo foi originado. |
| LoanOriginationMonth | Mês da data em que o empréstimo foi originado. |
| LoanOriginationYearMonth | Ano e mês da data em que o empréstimo foi originado (separdo por ponto). |
| IncomeRange | A faixa salarial do mutuário no momento em que a lista foi criada. |

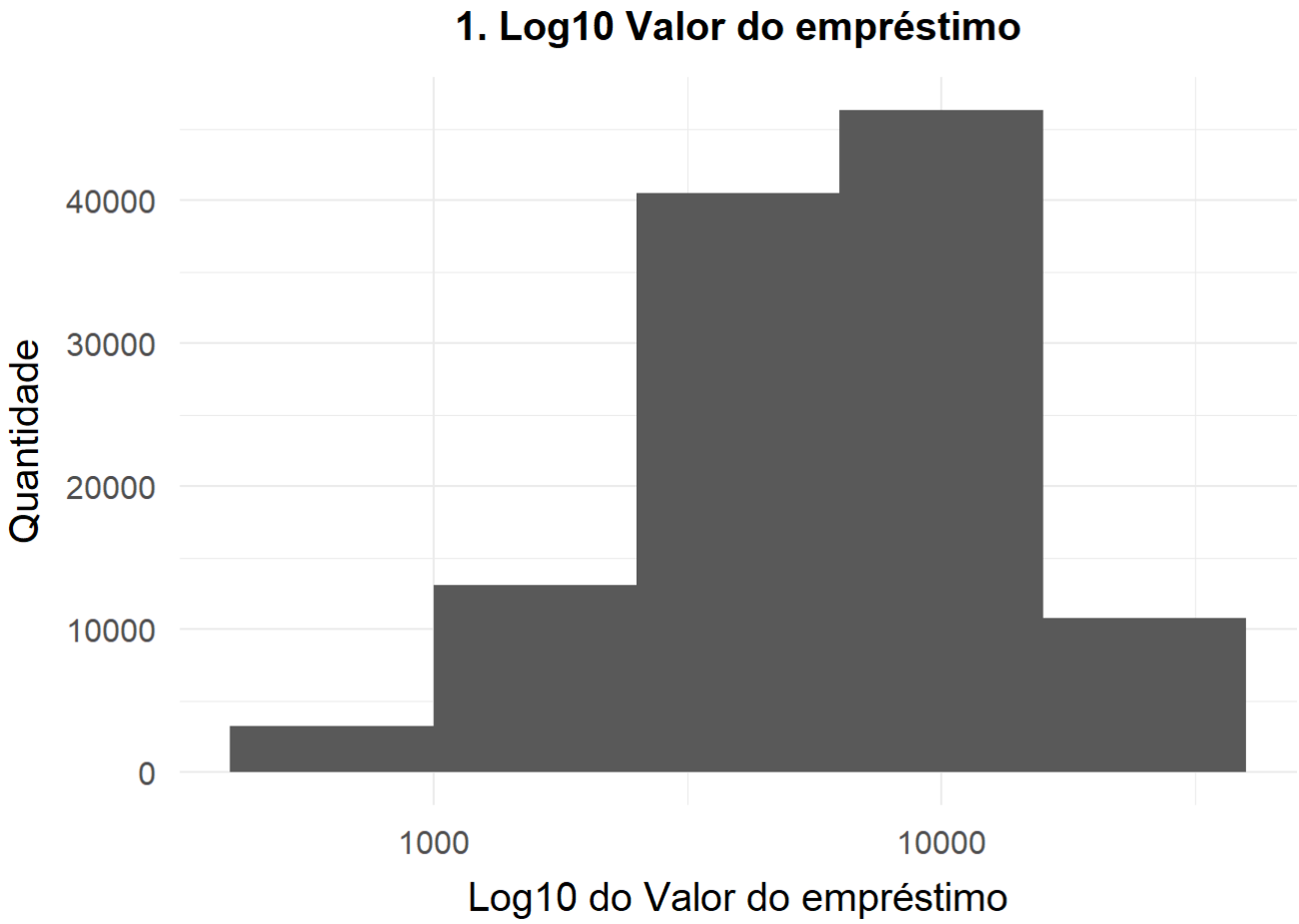
# QUESTÕES A SEREM EXPLORADAS

Com base nas variáveis selecionadas, seguem abaixo algumas perguntas que iremos tentar responder:

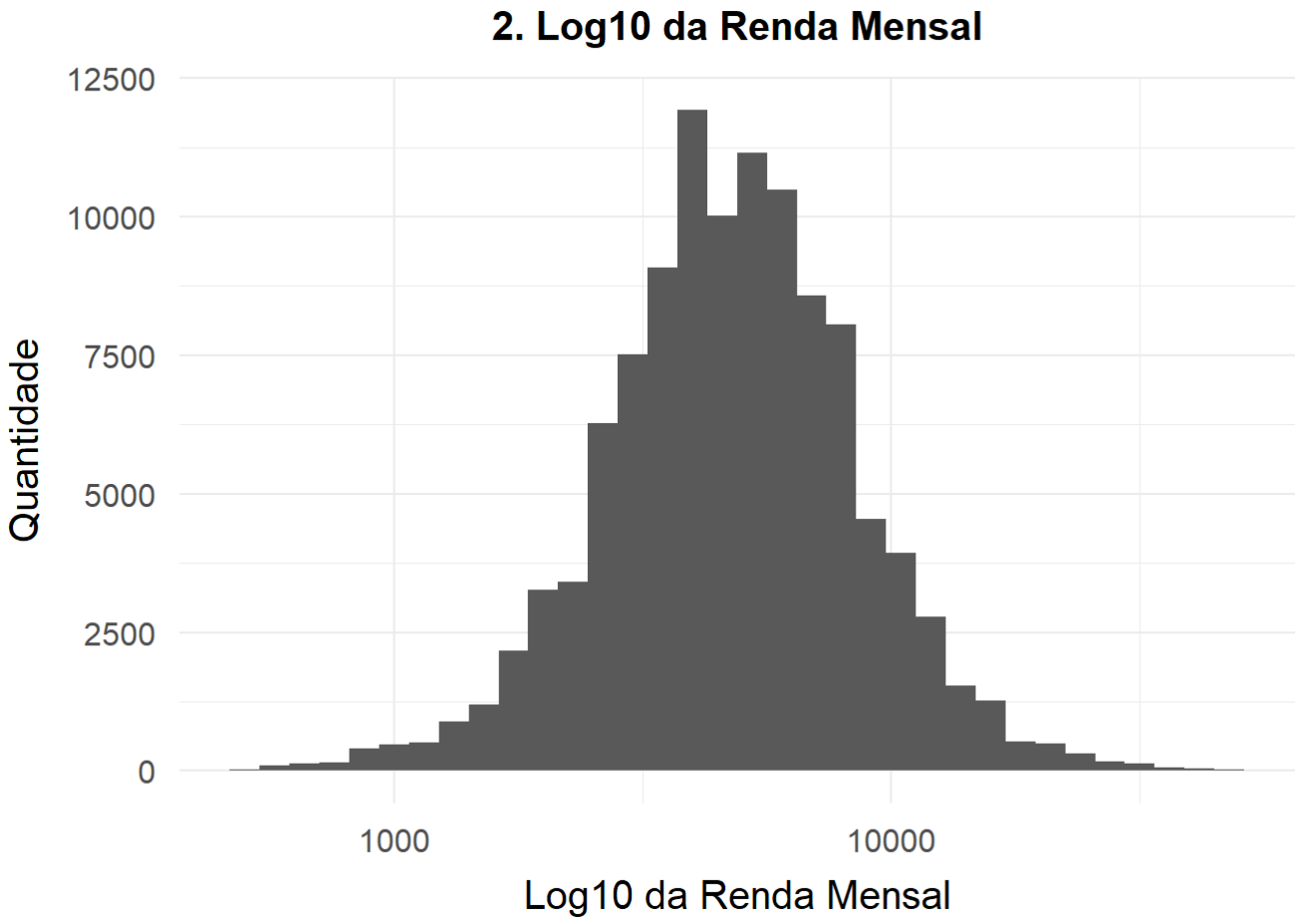
1. Qual é o maior motivo para a realização de empréstimos?
2. A taxa de juros sofre alguma variação de acordo com a finalidade (categoria) do empréstimo?
3. Quem tem uma renda maior, paga uma parcela maior?
4. Quem tem uma renda maior, faz empréstimos mais altos?
5. Mutuários que estão desempregados tem taxa de juros maiores que as que estão empregadas?
6. Mutuários que estão empregados e que também comprovaram renda, tem juros menor que as que estão empregadas mas não comprovaram renda?
7. Com base nas informações e histórico, a Prosper apura uma pontuação chamada ProsperScore para os mutuarios, onde quanto maior a pontuação, maior a chance de ser um bom pagador (Consequentemente menor risco para o investidor). Esse score realmente tem relação com uma maior ou menor taxa de juros?
8. A renda da pessoa tem alguma influência sobre a taxa de juros utilizada no empréstimo?

# REALIZANDO A ANÁLISE

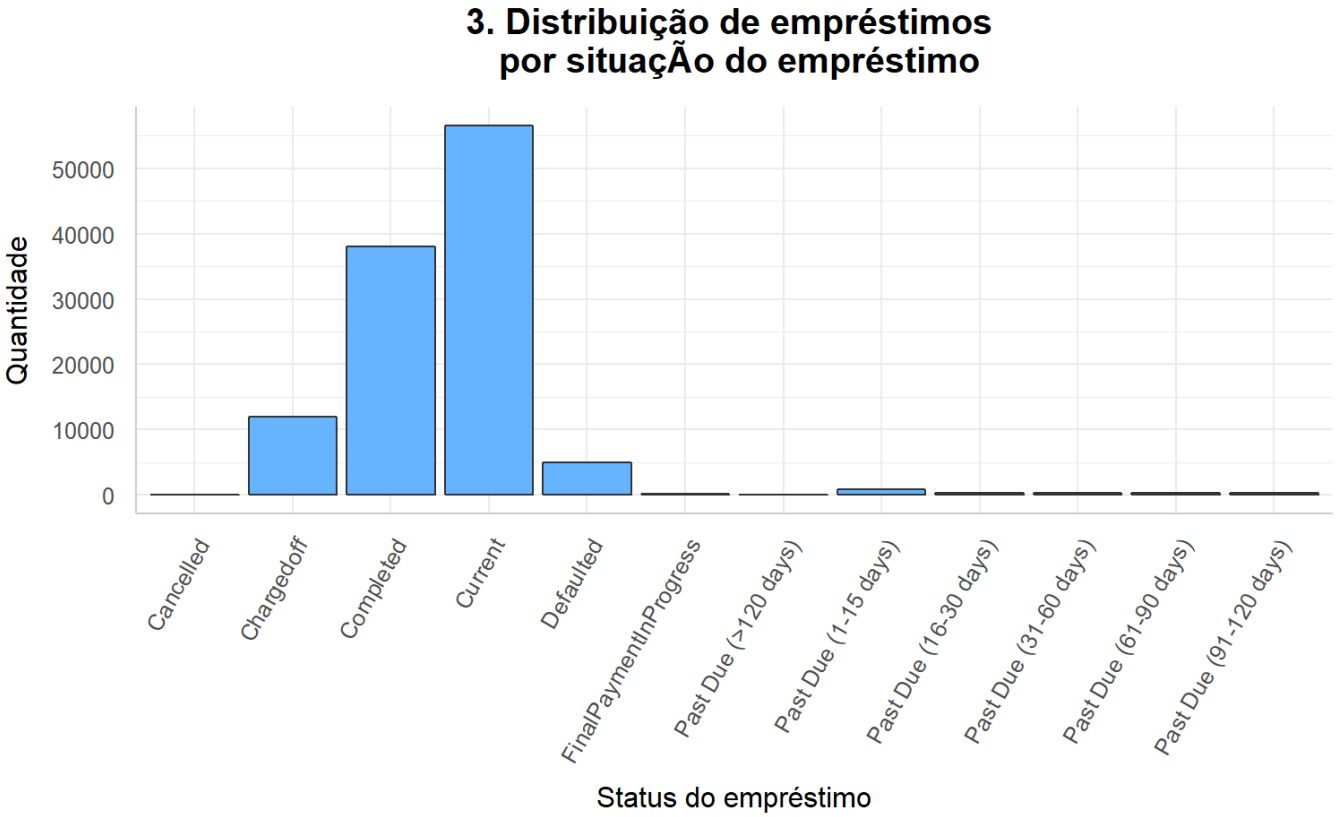
Primeiro vamos começar realizando uma Análise das distibuições para termos uma idéia geral da base de dados.




Usando a escala de log10 para o valor de empréstimo, não foi possivel identificar exatamente de qual distribuição se trata. Parece ser uma distribuição assimétrica negativa.



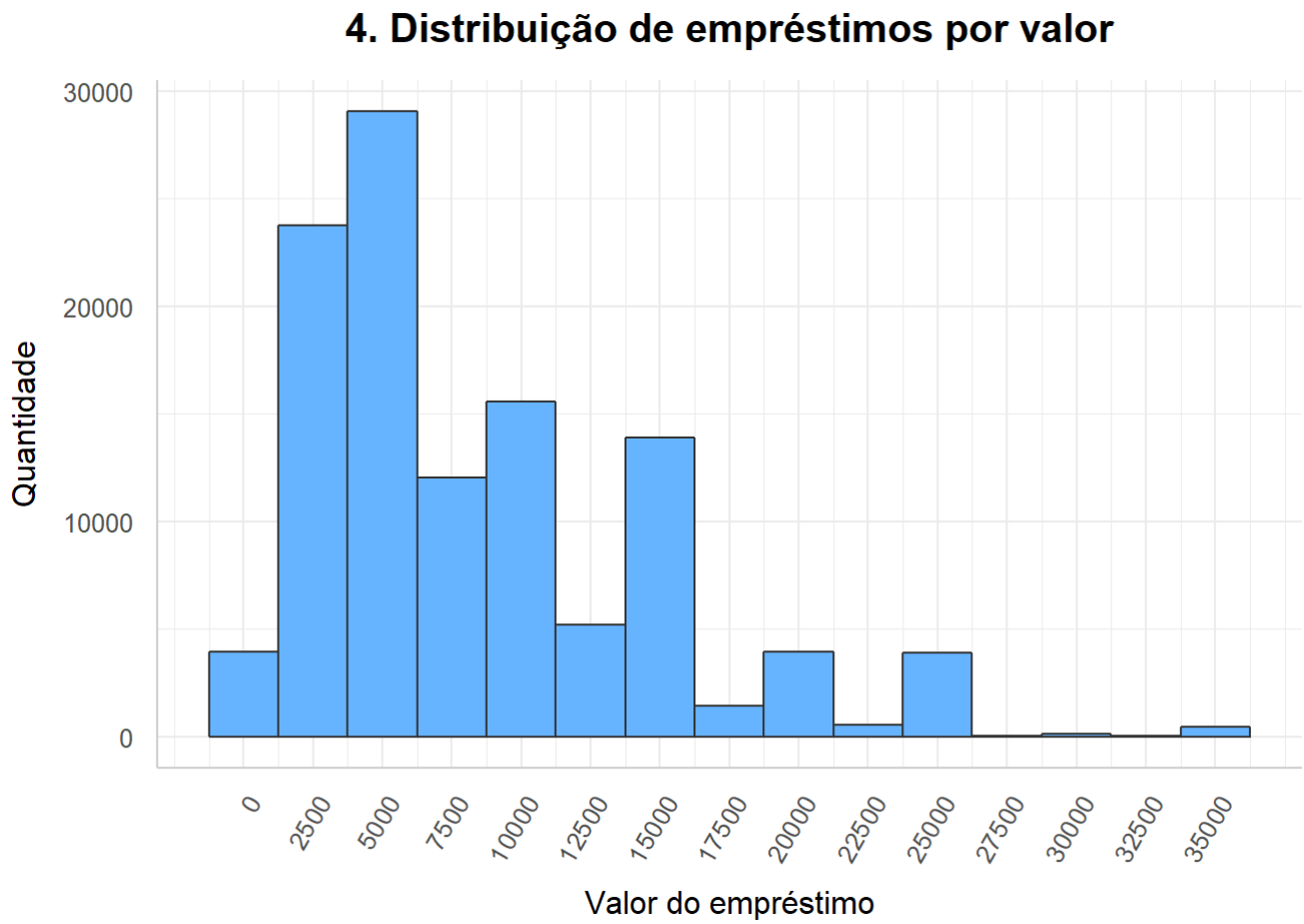

A distribuição da renda mensal é uma distribuição normal, também gerada usando uma escala de log10.




Note acima que a maioria dos empréstimos estão em andamento (Current). Existem uma série de status de empréstimo indicando a quantidade de dias que estão vencidos.

| | **Tradução** |
|---|---|
| Cancelled | Cancelado |
| Chargedoff | Cobrados fora |
| Completed | Concluído |
| Current | Em andamento |
| Defaulted | Padronizados |
| Final Payment In Progress | Pagamento final em andamento |
| Past Due (>120 days) | Vencido (> 120 dias) |
| Past Due (1-15 days) | Vencido (1-15 dias) |
| Past Due (16-30 days) | Vencido (16 a 30 dias) |
| Past Due (31-60 days) | Vencido (31-60 dias) |
| Past Due (61-90 days) | Vencido (61-90 dias) |
| Past Due (91-120 days) | Vencido (91-120 dias) |

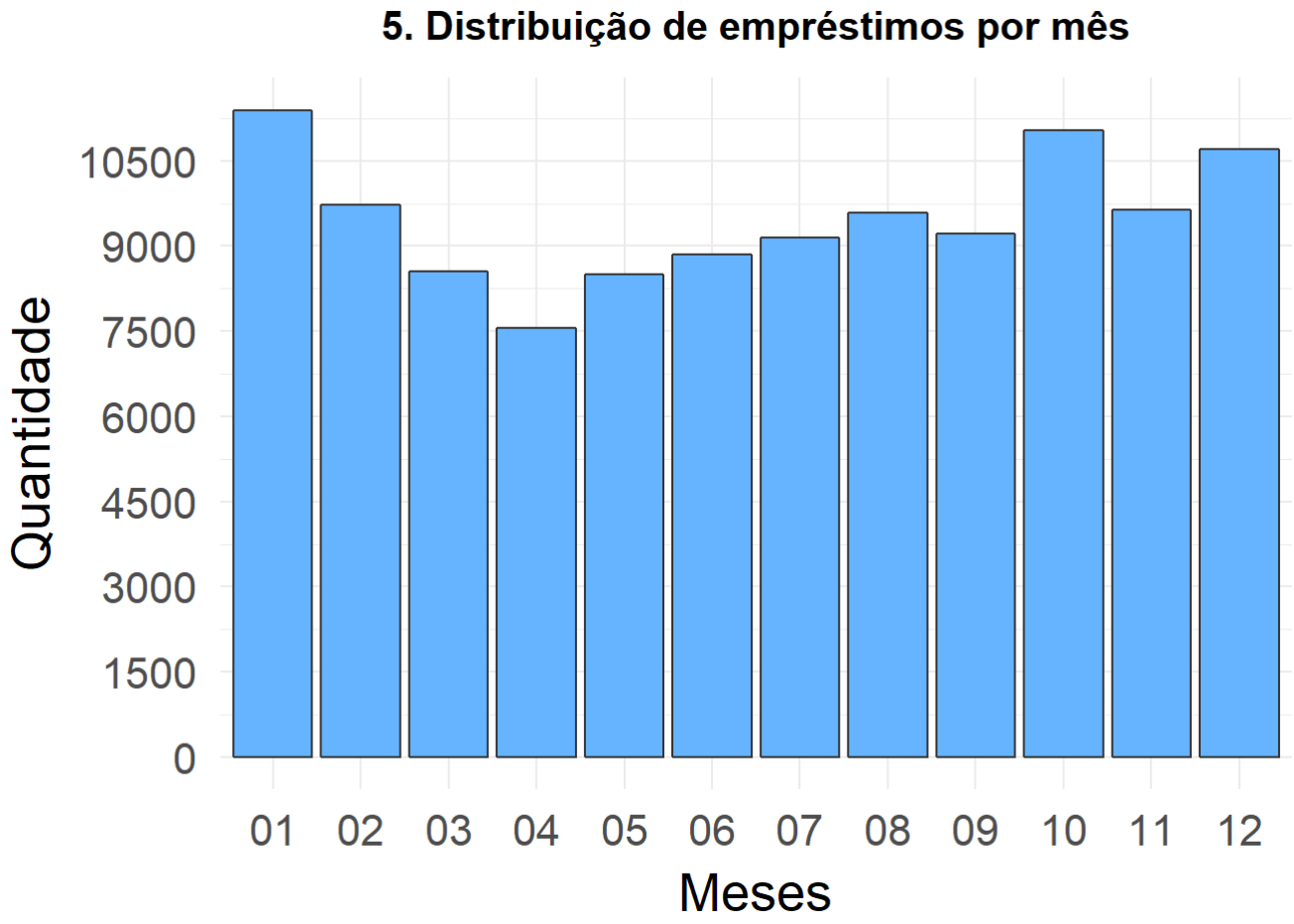
> Iremos desconsidearr as observações com status do empréstimo igual a “Cancelled” (são apenas 5)



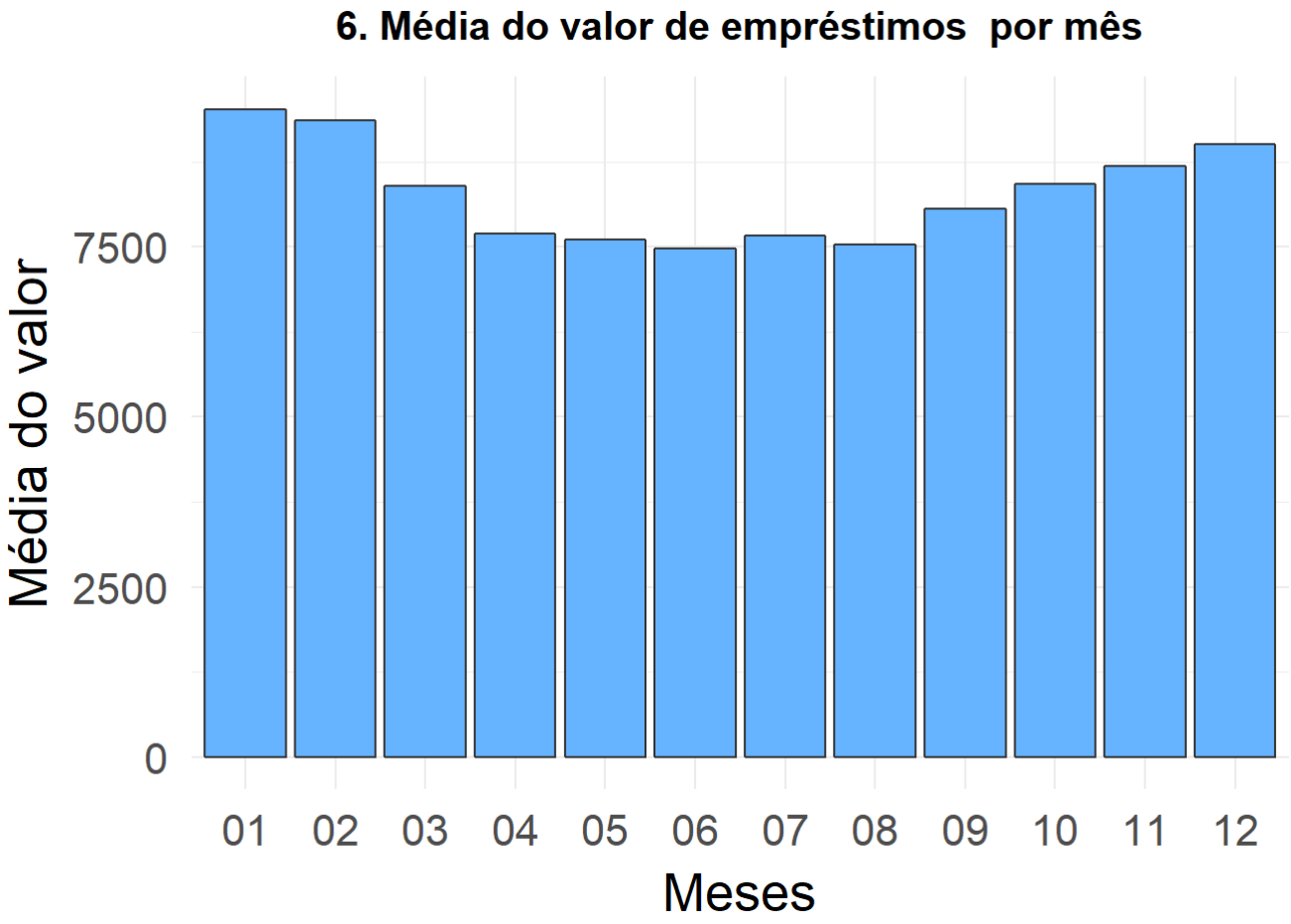

A maioria dos mutuários realizam emprestimos aproximadamente entre US$ 5000 a US$ 7.000.

As estatisticas abaixo indicam que a média de valor dos emprestimos é de US$ 8337 e que do total de empéstimos 75% de até US$ 12000.

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

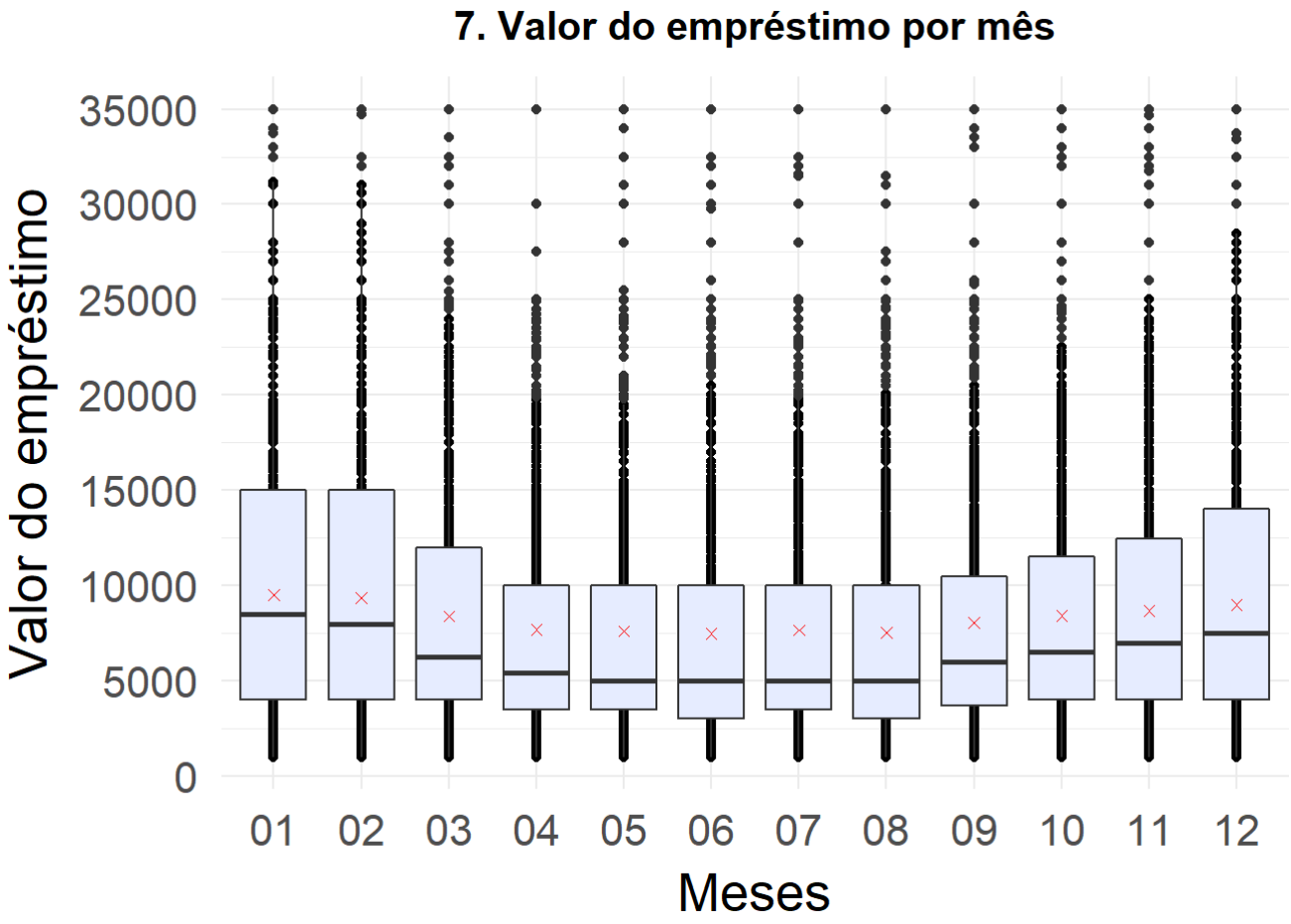



Note que os meses que tiveram maiores quantidades de empréstimos foram os meses do final e inicio do ano e também o mês de outubro.

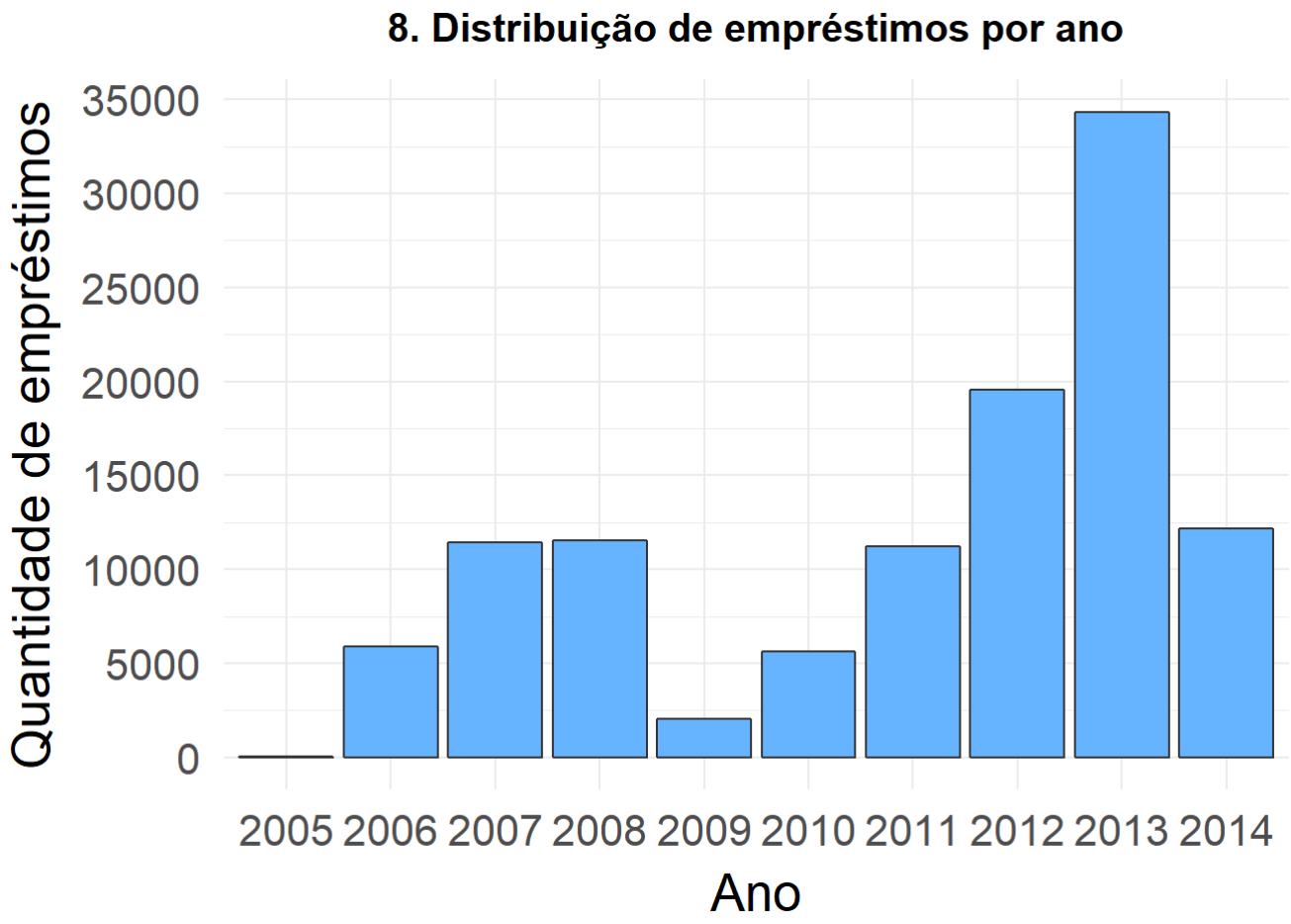



Agora considerando a média de valores emprestados, os meses com maior média foram janeiro, fevereiro seguido de dezembro.

Analisando os valores de empréstimos por mês




Neste gráfico odemos ver a concentração dos empréstimos de acordo com os valores. O traço na horizontal no meio do retângulo indica a mediana. O “x” na cor vermelha indica a posição da média. Os pontos pretos indicam os outliers (exceções).



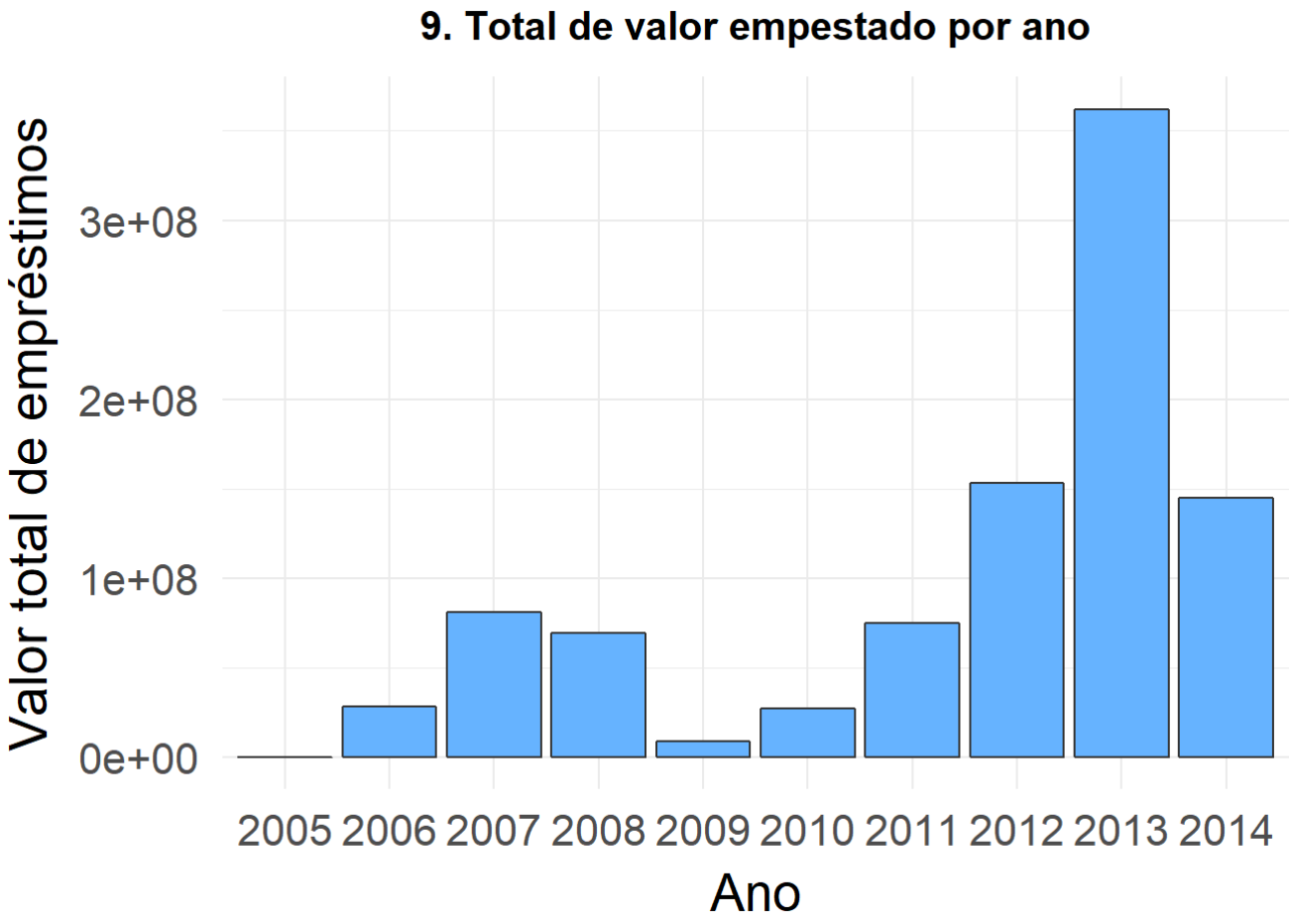

Nossa base de dados contém dados de 2005 até 2014. Notamos que em 2009 houve uma grande queda. Sabe-se que entre 2007 e 2008 houve uma crise financeira mundial. Talvez isto pode ter alguma relação com a diminuição dos empréstimos nestes anos.

Após 2009 a quantidade de emprestimos tornou a crescer, no entnato, em 2014 o valor volta a cair. Podemos observar também uma pequena quantidade de empréstimos em 2005. Vamos analisar mais de perto esses dois anos:

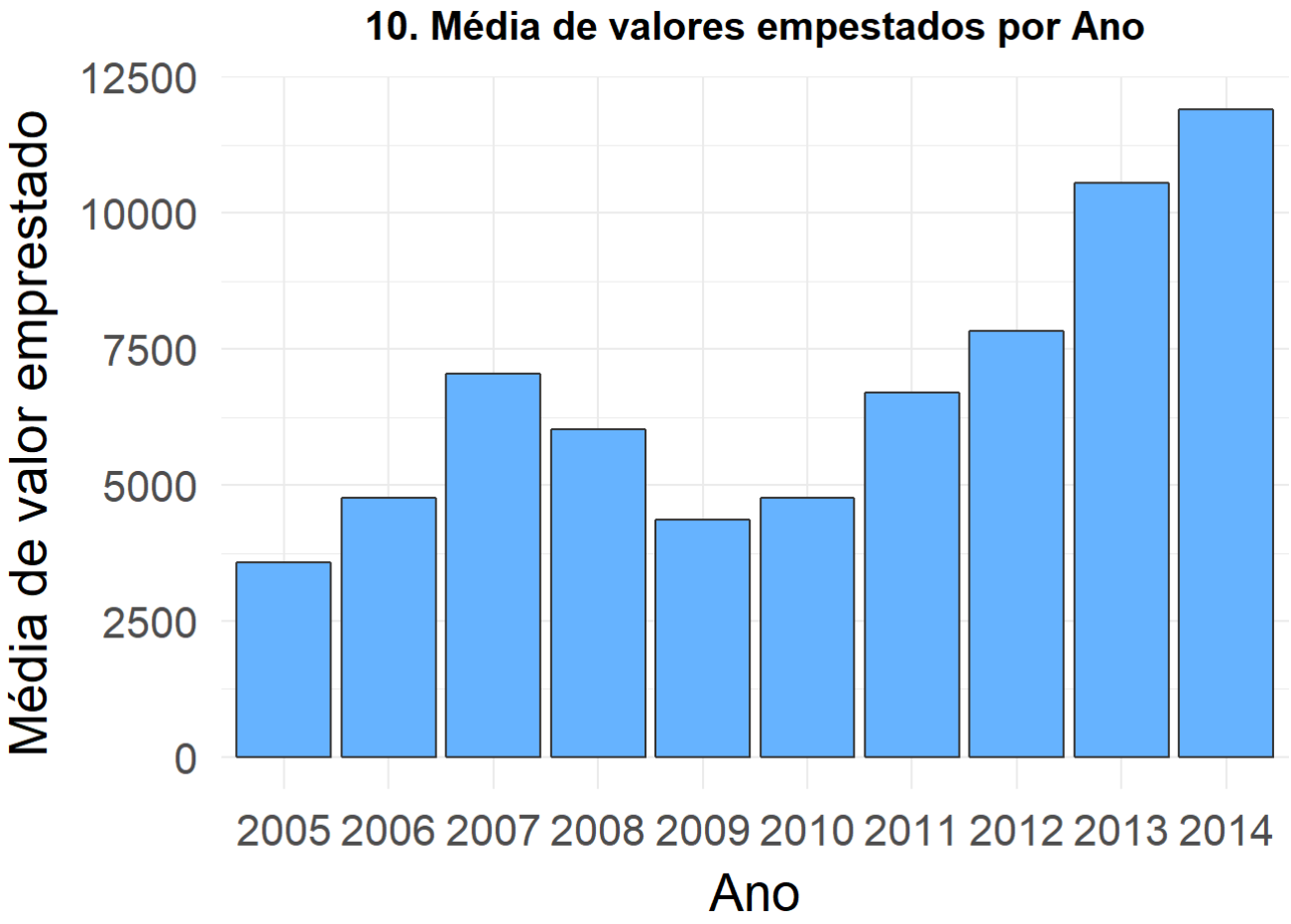
**Quantidade de empréstimos realizados em 2005 e 2014 por mês:**

```
## # A tibble: 5 x 2
##   LoanOriginationYearMonth     n
##   <chr>                    <int>
## 1 2005.11                     13
## 2 2005.12                      9
## 3 2014.01                   5865
## 4 2014.02                   4485
## 5 2014.03                   1822
```

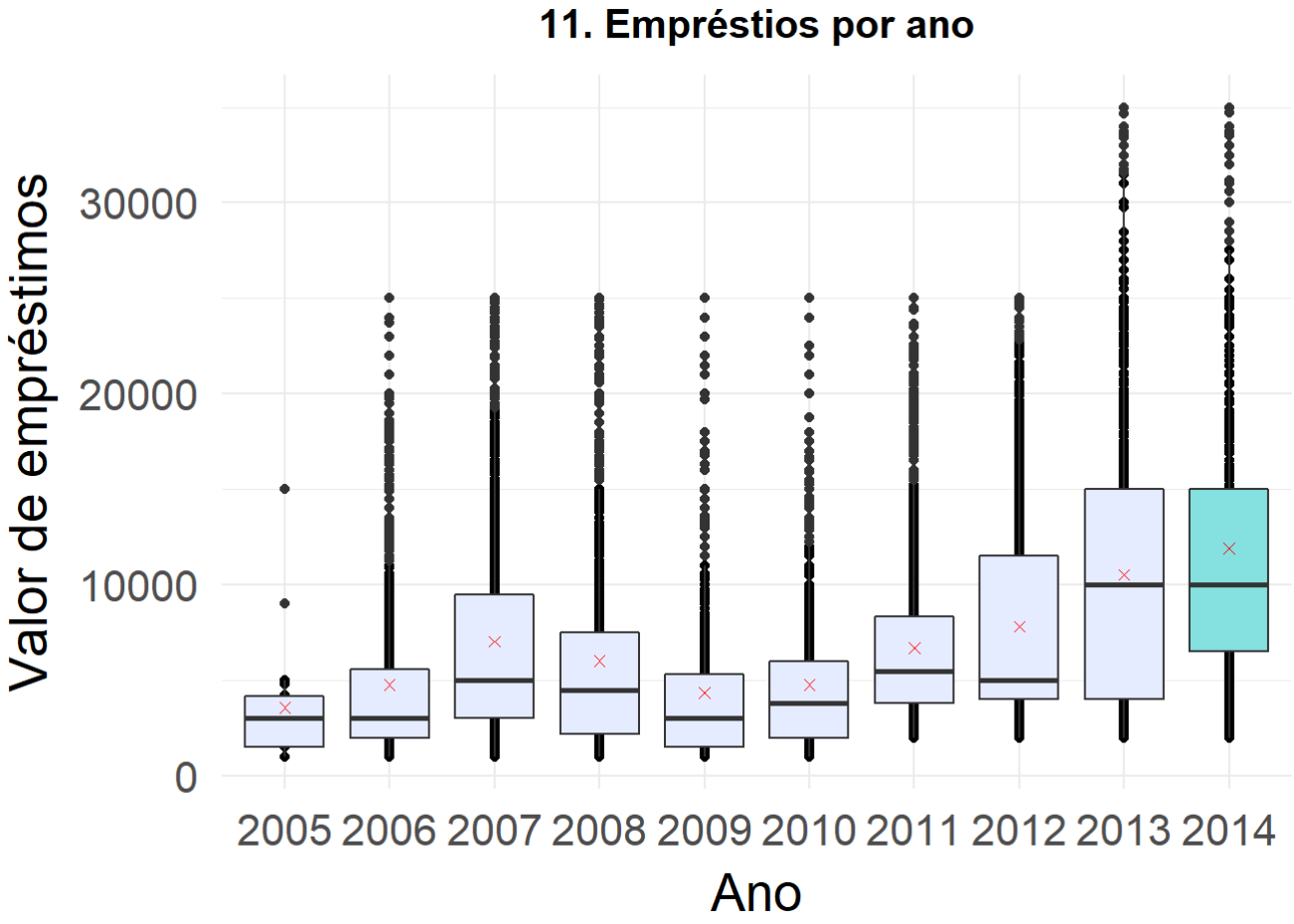
Veja que o ano de 2005 tiveram somente dois meses e a quantidade de emprestimos foram somente 22. Já para o ano de 2014, tivemos somente três meses registrados com emprestimos, mas apesar disso, 2014 tem uma contagem bastante alta, somando 12.172 emprestimos. Vamos continuar a análise.




Apesar de 2014 ter somente três meses de empréstimos realizados, em valor ele já tem a mesma quantia que o ano inteiro de 2012. Isso indica que 2014 provavelmente será um ano com a maior número de emprestimos realizados.




Aqui podemos cofirmar uma tendência. Em 2014 houve uma quantidade pequena de empréstimos em relação ao ano de 2013 conforme vimos no gráfico 8, no entanto, a média dos valores emprestados foi a mais alta já registrada. Há que considera ainda que em 2014 temos apenas três meses.



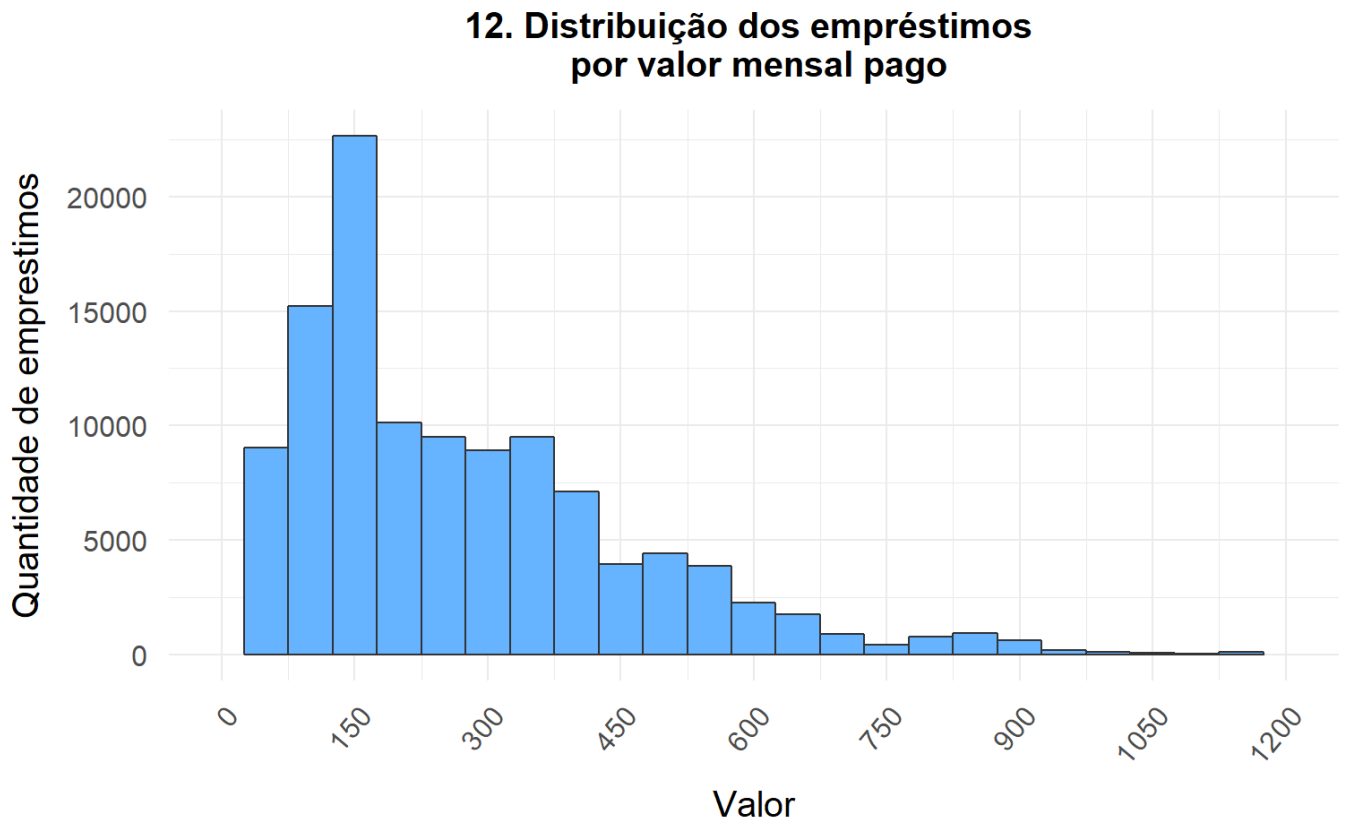

O “x” vermelho indica a média do valor do empréstimo. O traço horizontal no meio do retângulo é a mediana. Aqui podemos ver informações estatésticas sobre Distribuição dos emprestimos ao longo dos anos. Para a maioria dos anos a média de empréstimos vai até US$ 10.000. De 2010 em diante podemos notar um aumento crescente dos valores a cada ano. Em 2014 vêmos novamente a maior média de todos os anos.

** Tentar mexer no visual, usar X ao invés de O

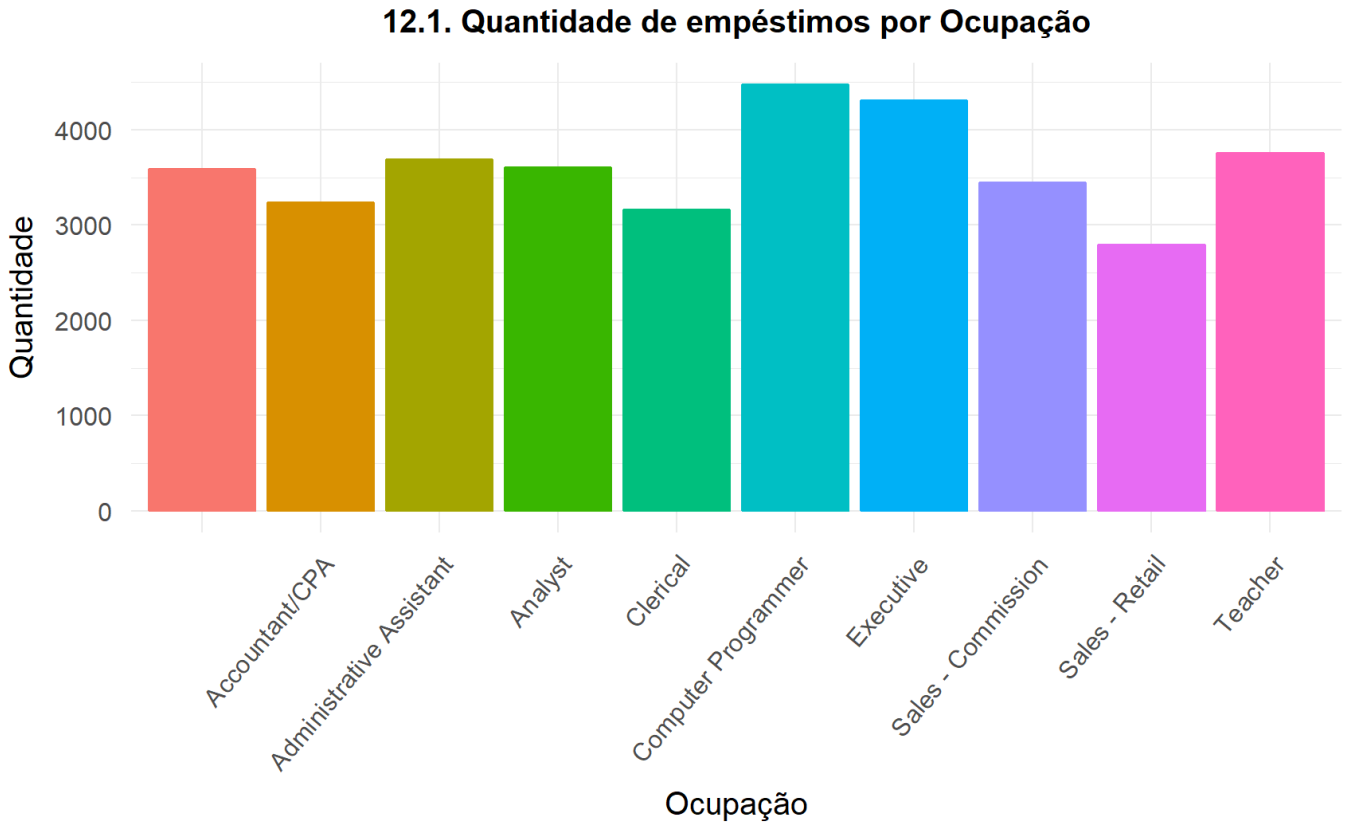
No entanto temos que considerar que em 2014 há somente três meses. Para compararmos com os demais meses e sabermos se realmente pode ser uma tendência iremos calcular abaixo a média somente os três primeiros meses de cada ano:

```
## df_tree_months$LoanOriginationYear: 2006
## [1] 4959.831
## ------------------------------------------------------------
## df_tree_months$LoanOriginationYear: 2007
## [1] 6472.595
## ------------------------------------------------------------
## df_tree_months$LoanOriginationYear: 2008
## [1] 6658.935
## ------------------------------------------------------------
## df_tree_months$LoanOriginationYear: 2010
## [1] 4773.685
## ------------------------------------------------------------
## df_tree_months$LoanOriginationYear: 2011
## [1] 6577.909
## ------------------------------------------------------------
## df_tree_months$LoanOriginationYear: 2012
## [1] 7455.283
## ------------------------------------------------------------
## df_tree_months$LoanOriginationYear: 2013
## [1] 9386.756
## ------------------------------------------------------------
## df_tree_months$LoanOriginationYear: 2014
## [1] 11912.22
```

Podemos ver que para o ano de 2014 realmente o valor médio está bem maior que os demais anos, ou seja, há uma grande possibilidade de 2014 ser o ano com maior valor de empréstimos de todos os tempos.

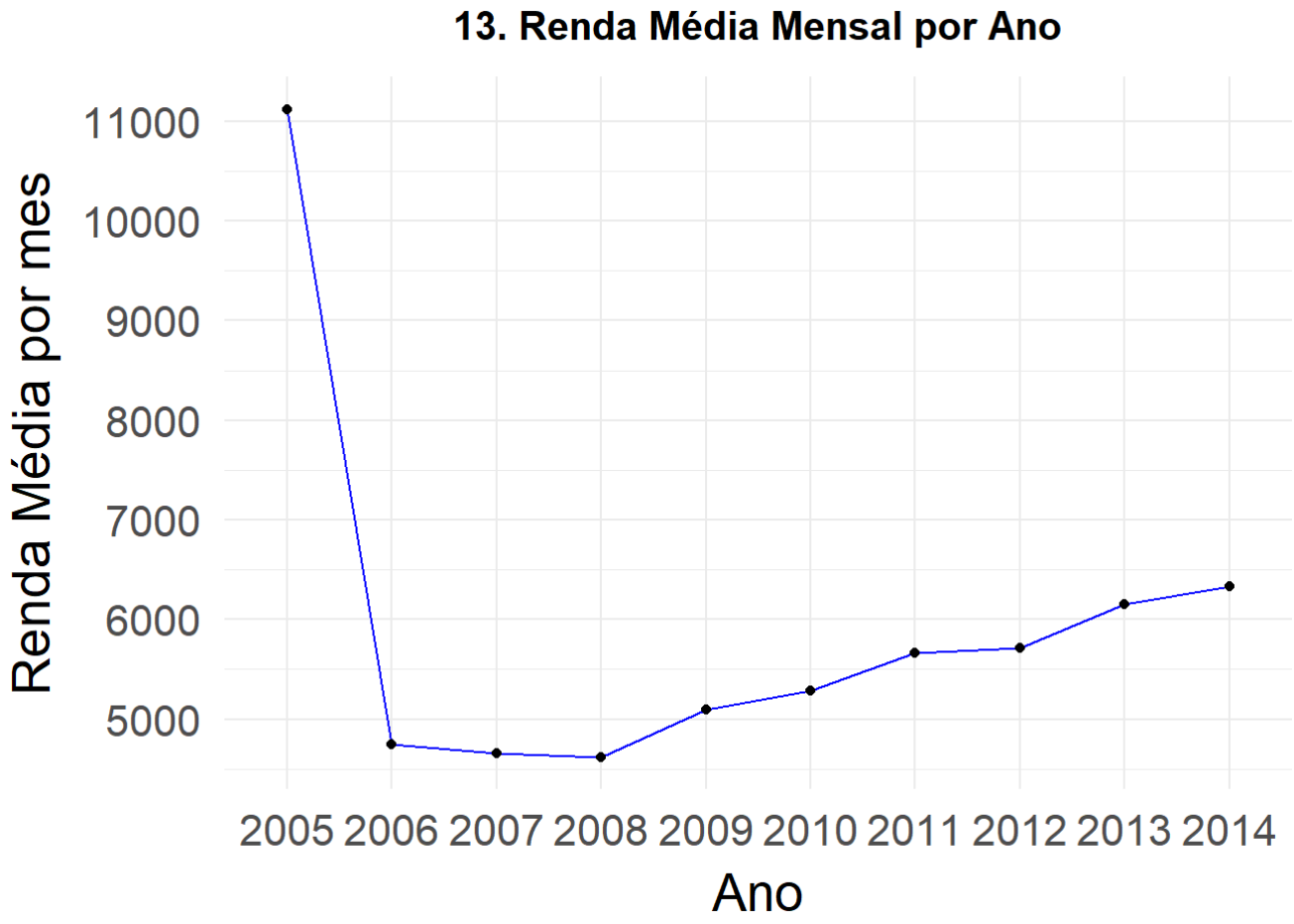



A maioria das parcelas são de aproximadamente US$ 150.



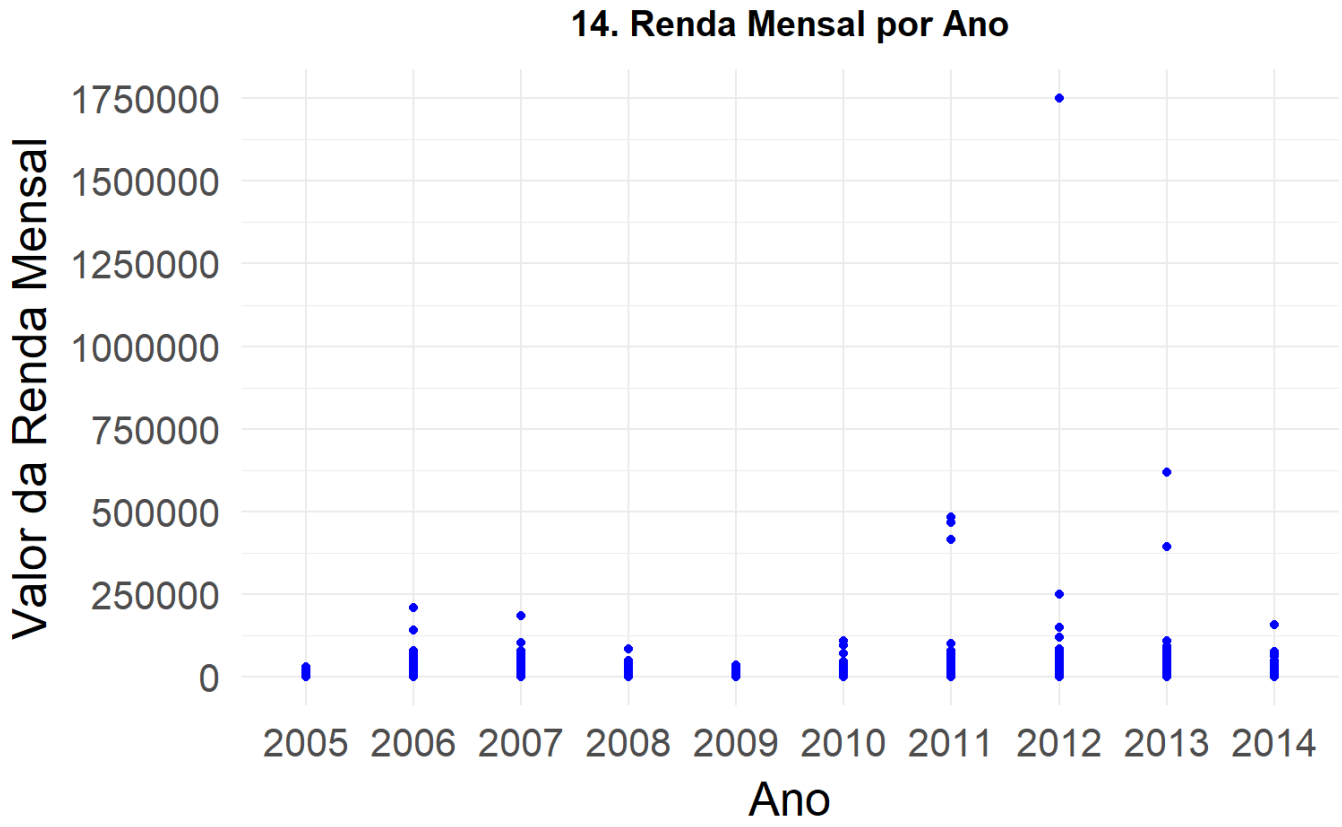

As ocupações que tem o maior número de emprestimos são: Programadores de computador, Executivos e Professores.

Abaixo estamos vendo ao contrário do que foi questionado anteriormente. 2005 foi o ano com a maior renda média de todos os anos.



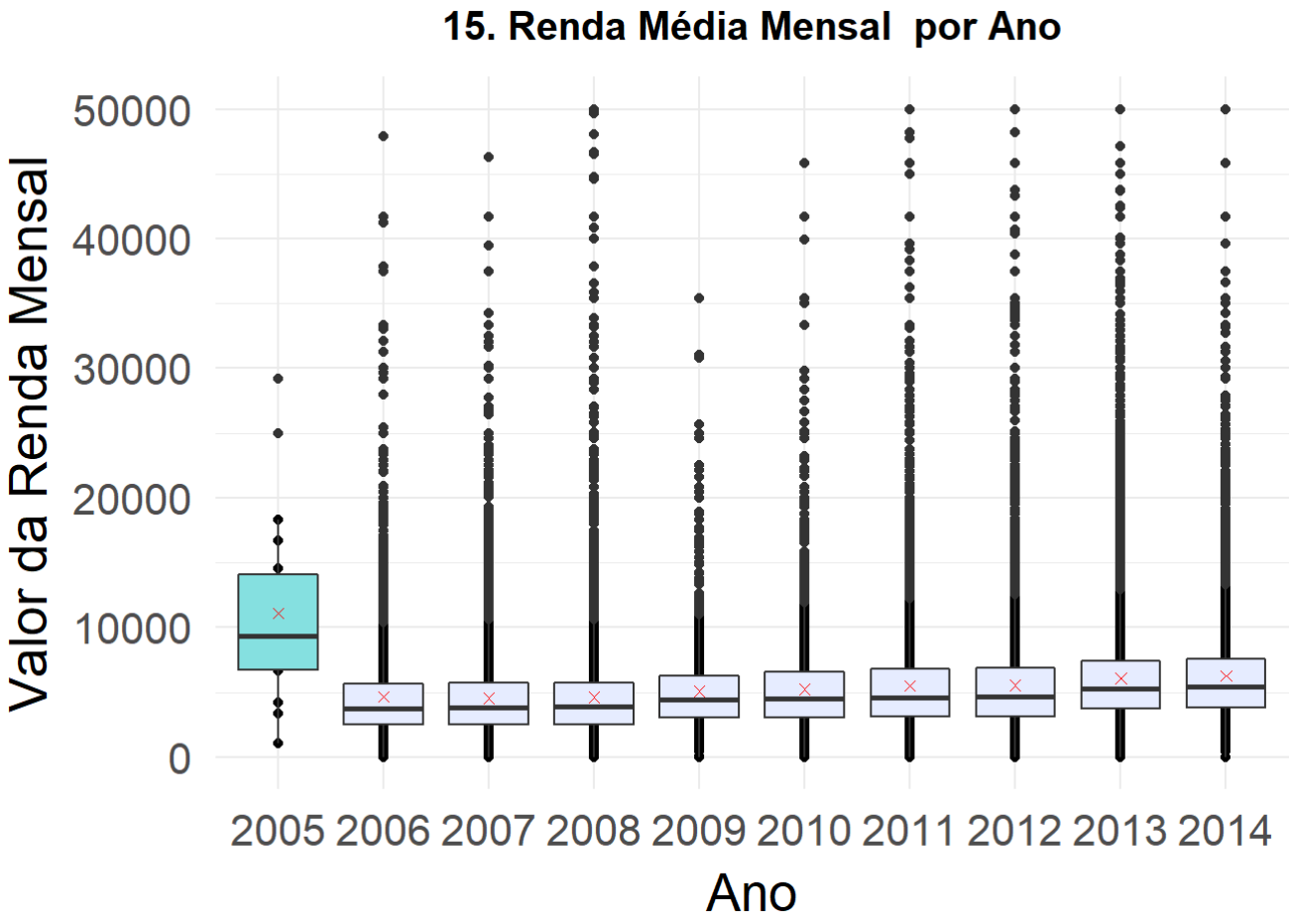

Enquanto os demais anos a média tem uma variação aproximadamente entre US$ 4500 a US$ 6500, o ano de 2005 a média foi de aproximadamente US$ 11.000. Há que se considerar que para este ano existem poucos empréstimos. É comum que acontecer de qualquer valor acima do comum torna-se um outlayer influenciando a média.

Vamos analisar a dispersão da renda ao longo dos anos:



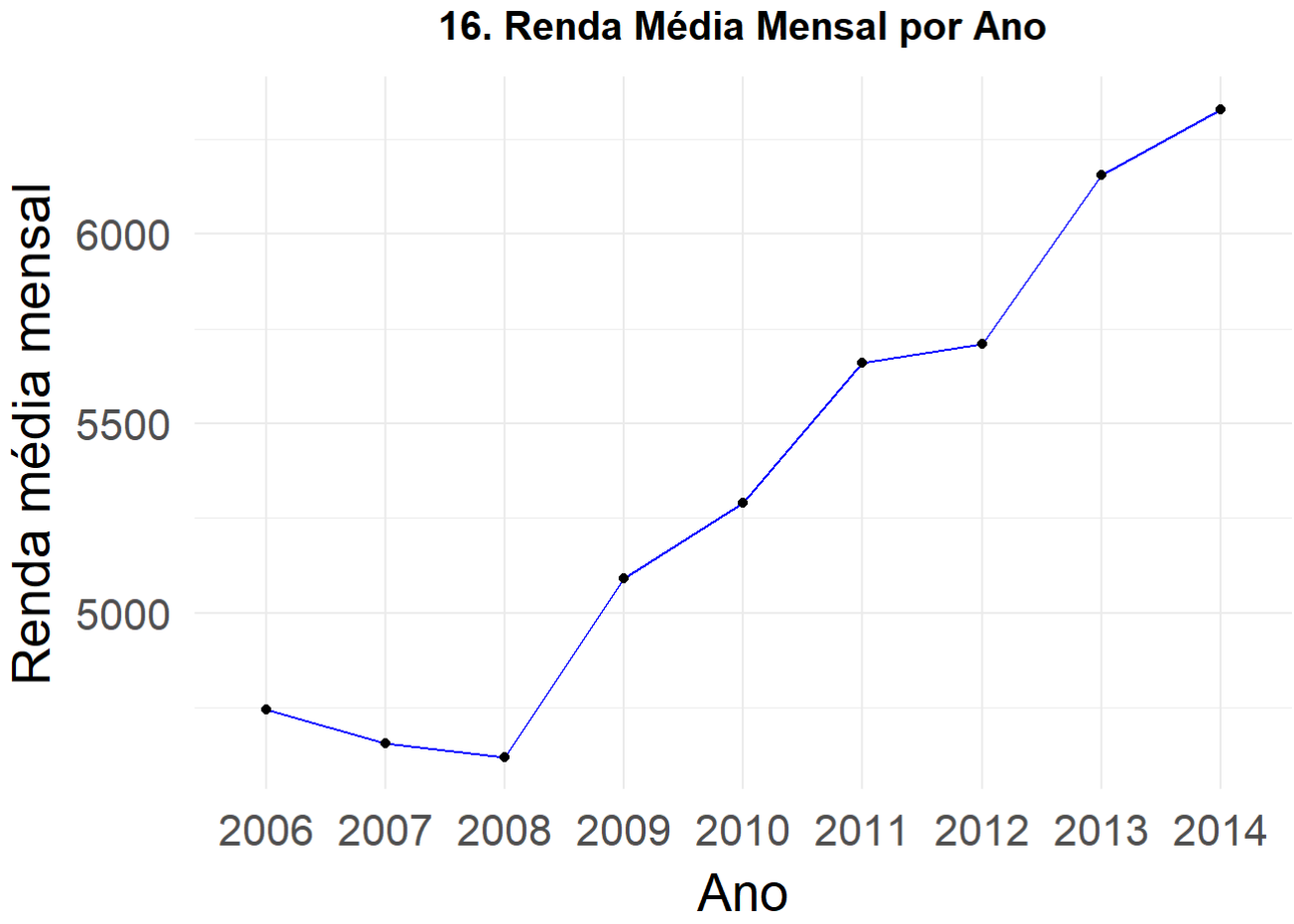

Podemos notar que em 2012 existe um outlayer. Analisando os detalhes, foi constatado que este mutuário tem a renda mensal de US$ 1750000, mas nao tem a renda comprovada.

Vamos agora visualizar as informações retirando os outliers, ou seja, vamos filtrar somente os mutuários com renda mensal abaxo de US$ 50000.

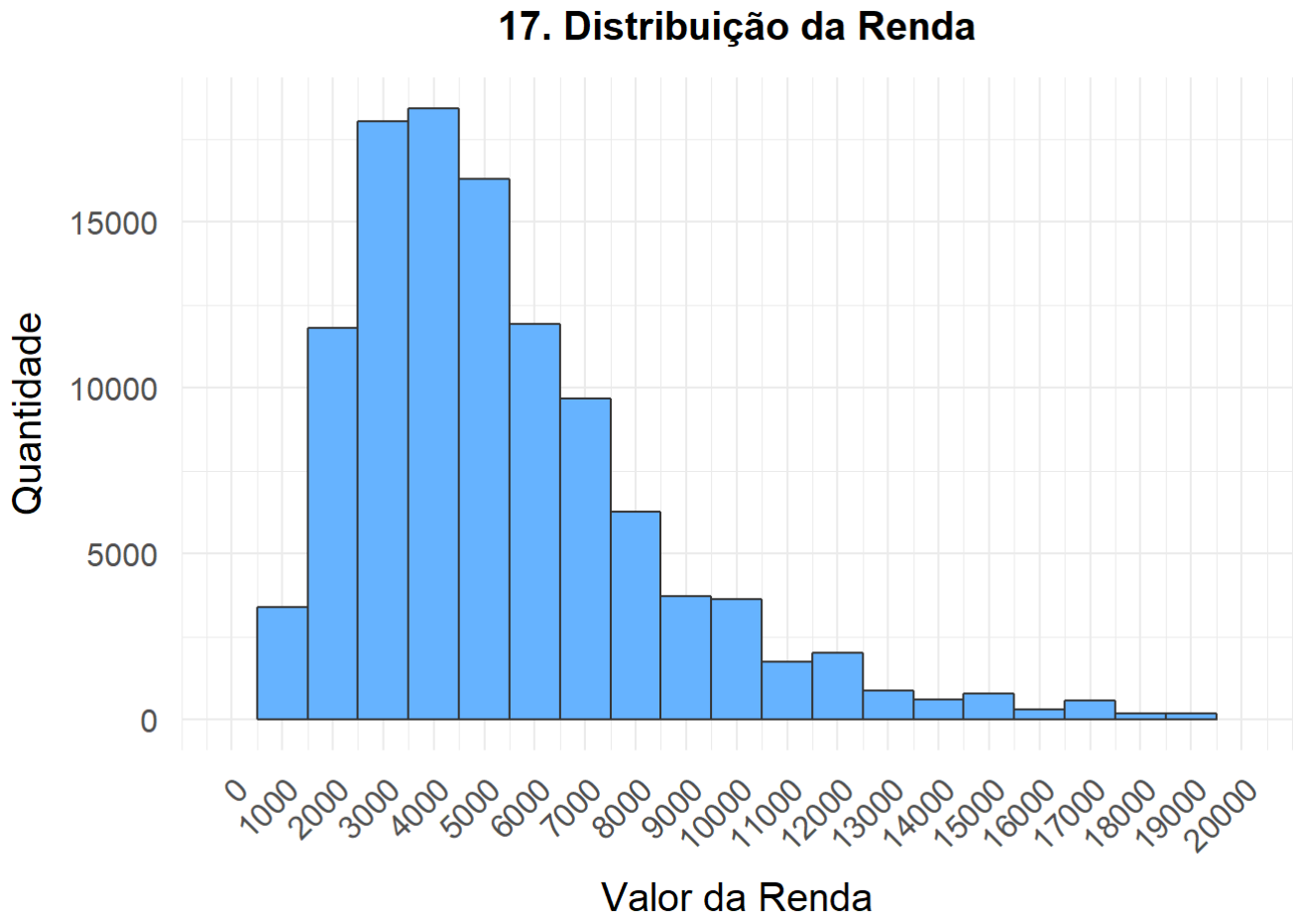



Realmente no ano de 2005 (destacado na cor verde), foi o ano com menor montante de valores emprestados (conforme vimos no gráfico 7), mas também foi um ano onde os mutuários tiveram a maior renda.

Vamos retirar o ano de 2005 para melhor visualizarmos a evolução da média de renda por ano:

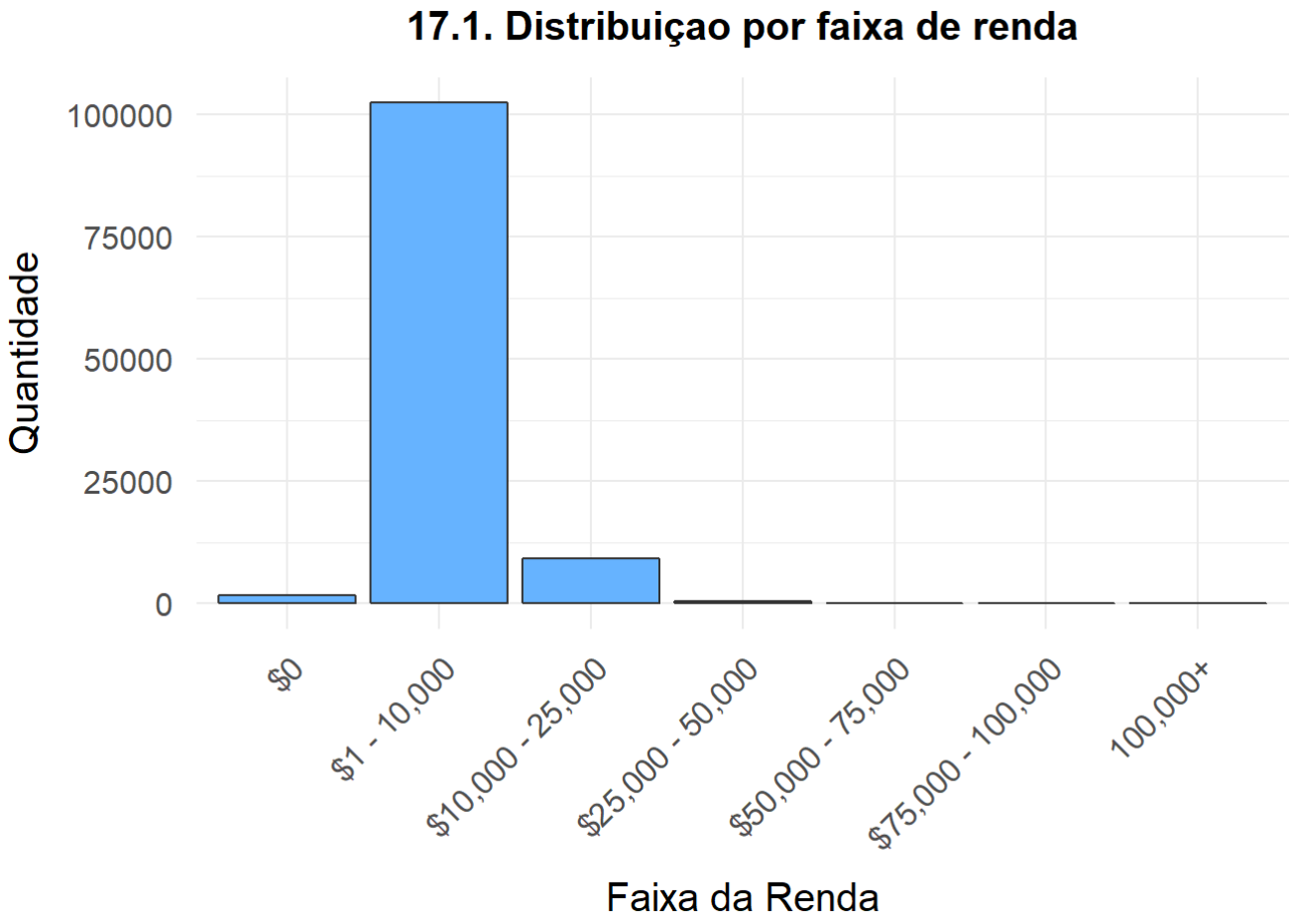



Entre o anos de 2006 a 2008, houve uma ligeira queda. De 2009 em diante, a renda vem aumentando a cada ano.

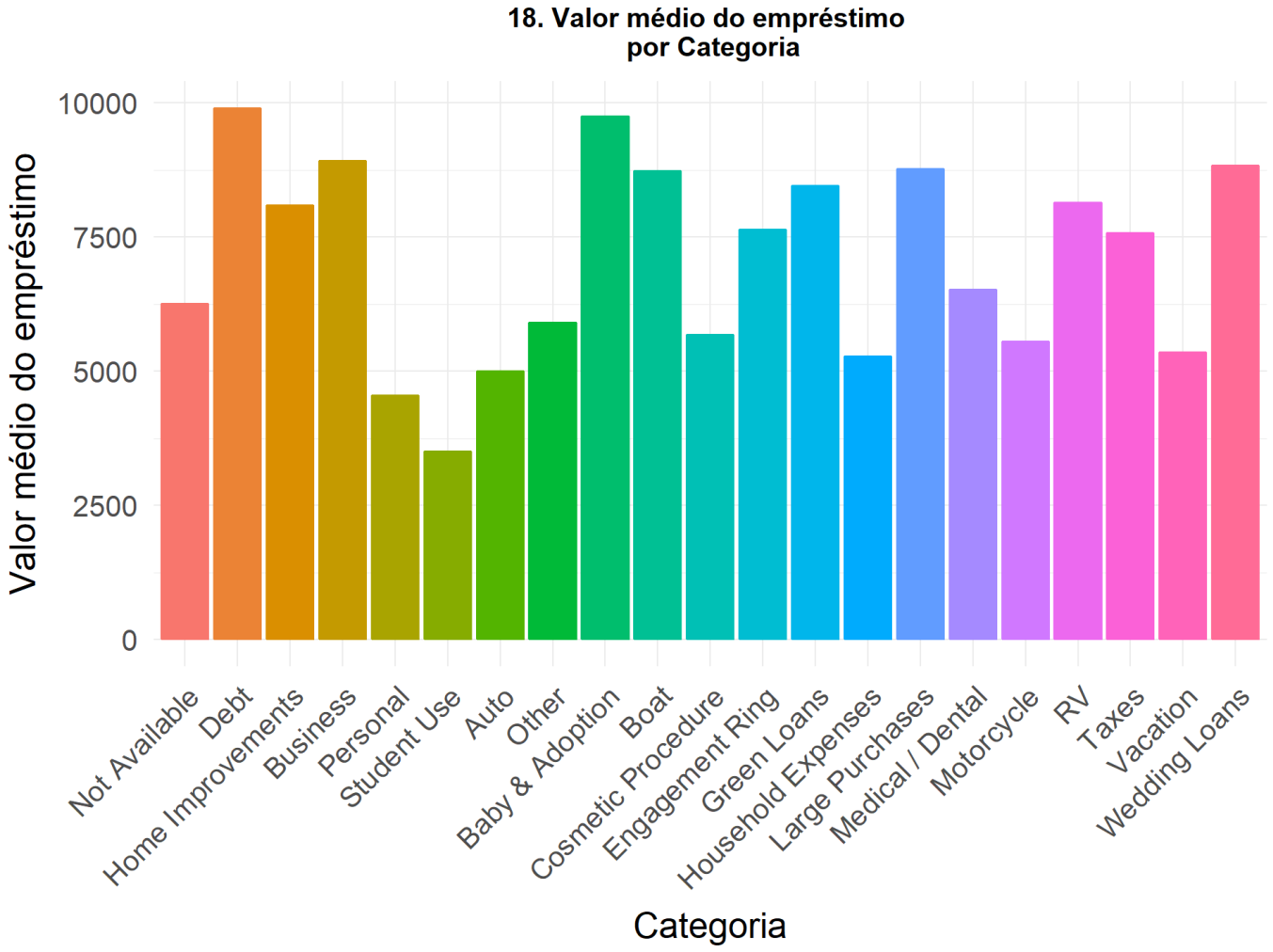



A maioria dos mutuários tem uma renda entre US$ 3000 a US$ 5000 mensais.

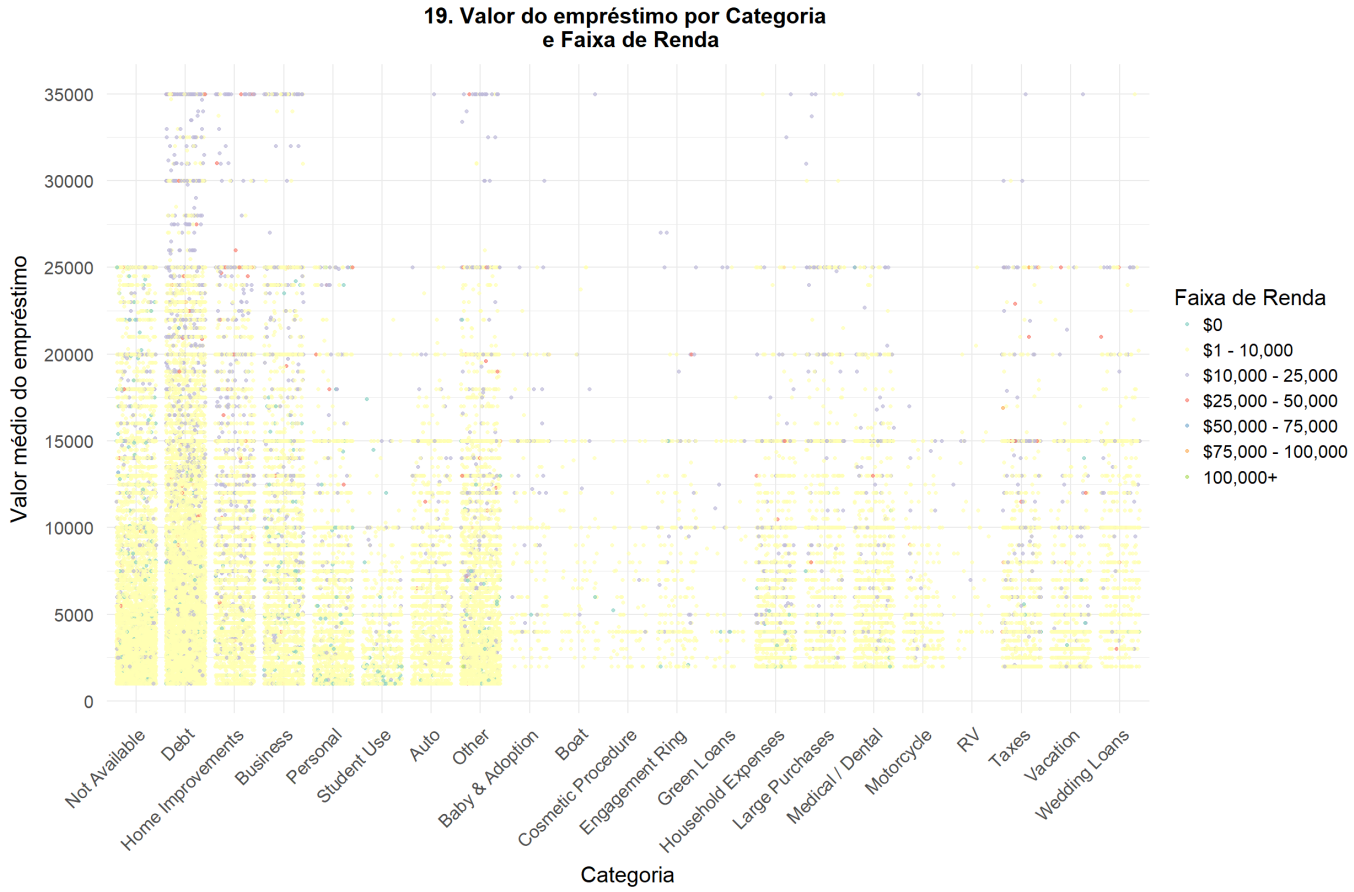
Obs.: Foi utilizado um limite para diminuir a calda do gráfico onde esta sendo exibidas somente as rendas mensais de até US$ 20.000.




Como verificado no gráfico anterior (17) a maior concentração de renda está entre US$ 3000 e 5000, é obvio que o gráfico acima apresentará ocorrência maior na faixa de renda de US$ 1 a 10000.

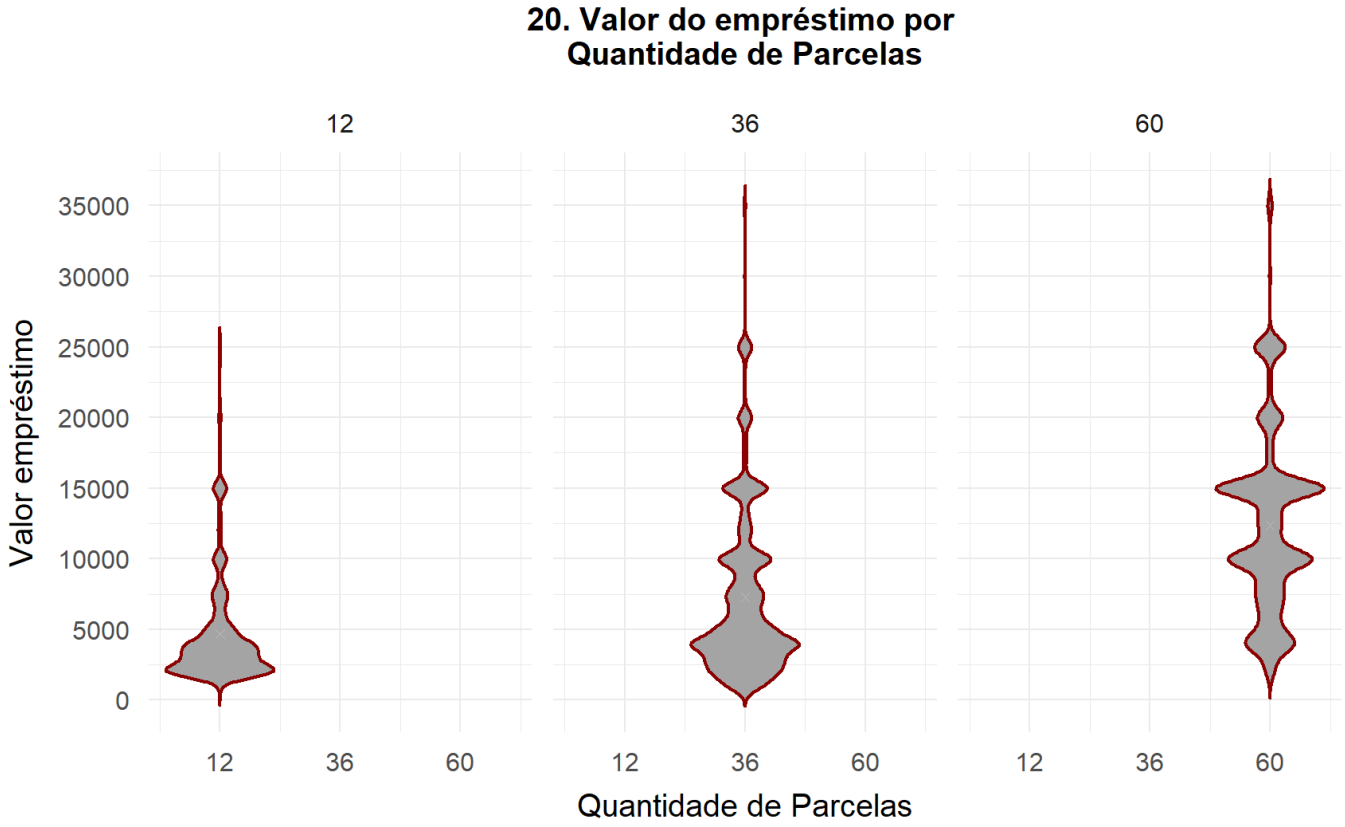



Respondendo à pergunta número 1, neste gráfico é possível identificar que em em média o maior motivo para a realização do empréstimo são para pagamento de dividas. Seguido de Bebê e Adoção, Negócios e Empréstimos para Casamento.



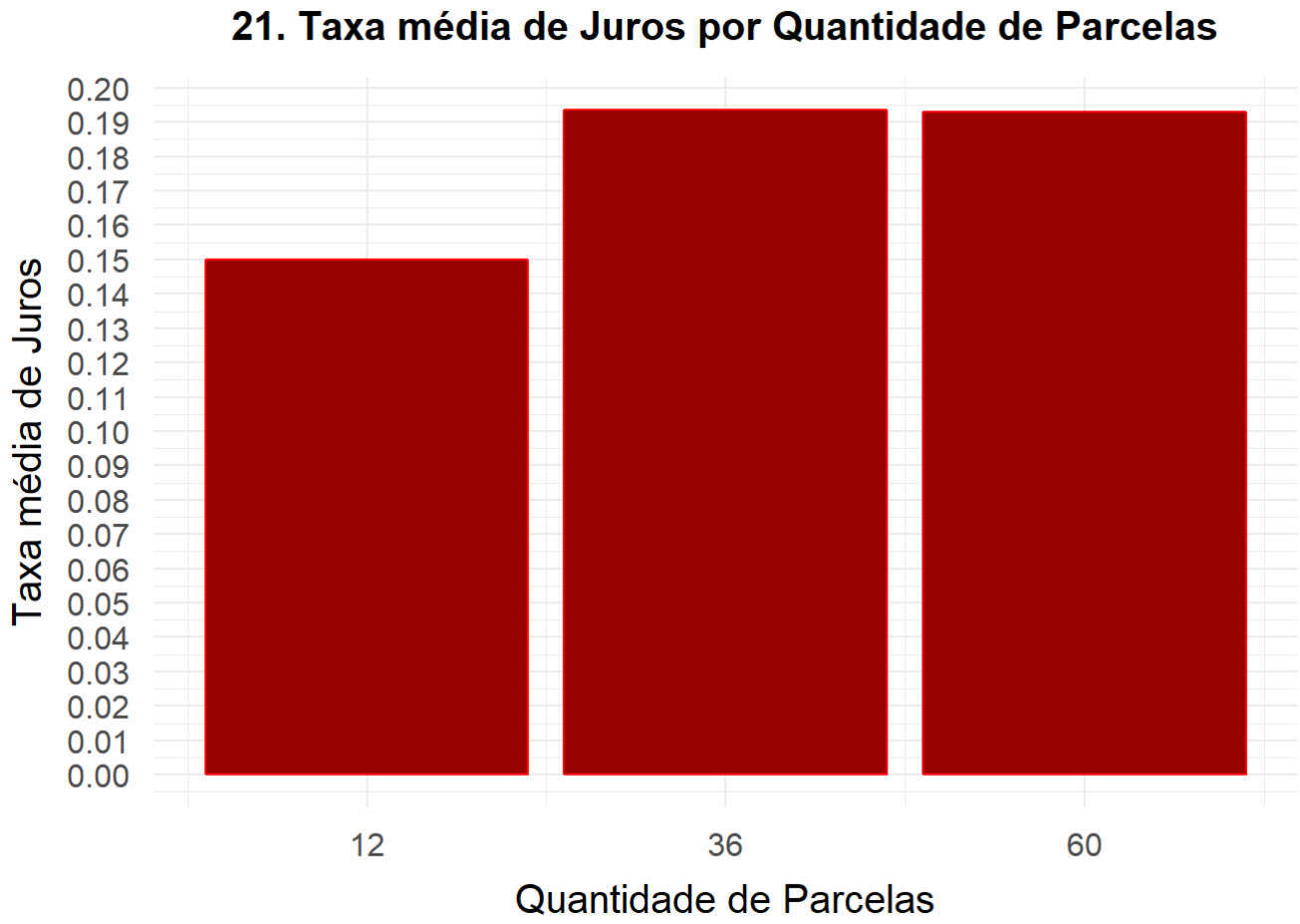

O gráfico acima deixa claro que o perfil mais comum de renda de quem realiza empréstimos são mutuários com faixas de renda entre $ 1 a 25000 e os valores dos emprestimos em geral vão somente até US$ 250000 onde a categoria mais utilizada é a para pagamento de dívidas (debt) A faixa de renda com maior quantidade de emprestimos com valor acima de US 25000 é de 10.000 a 25.000, e bastante utilizado para pagamento de dívidas (debt). As categorias com maiores concentrações de empréstimos são: Dívidas (debt), Outros (Others), Melhorias domiciliares (Home Improvements), Negócios (Business) e não disponível (Not available).

Segue um gráfico de violino com a Distribuição do valor por quantidade de parcelas

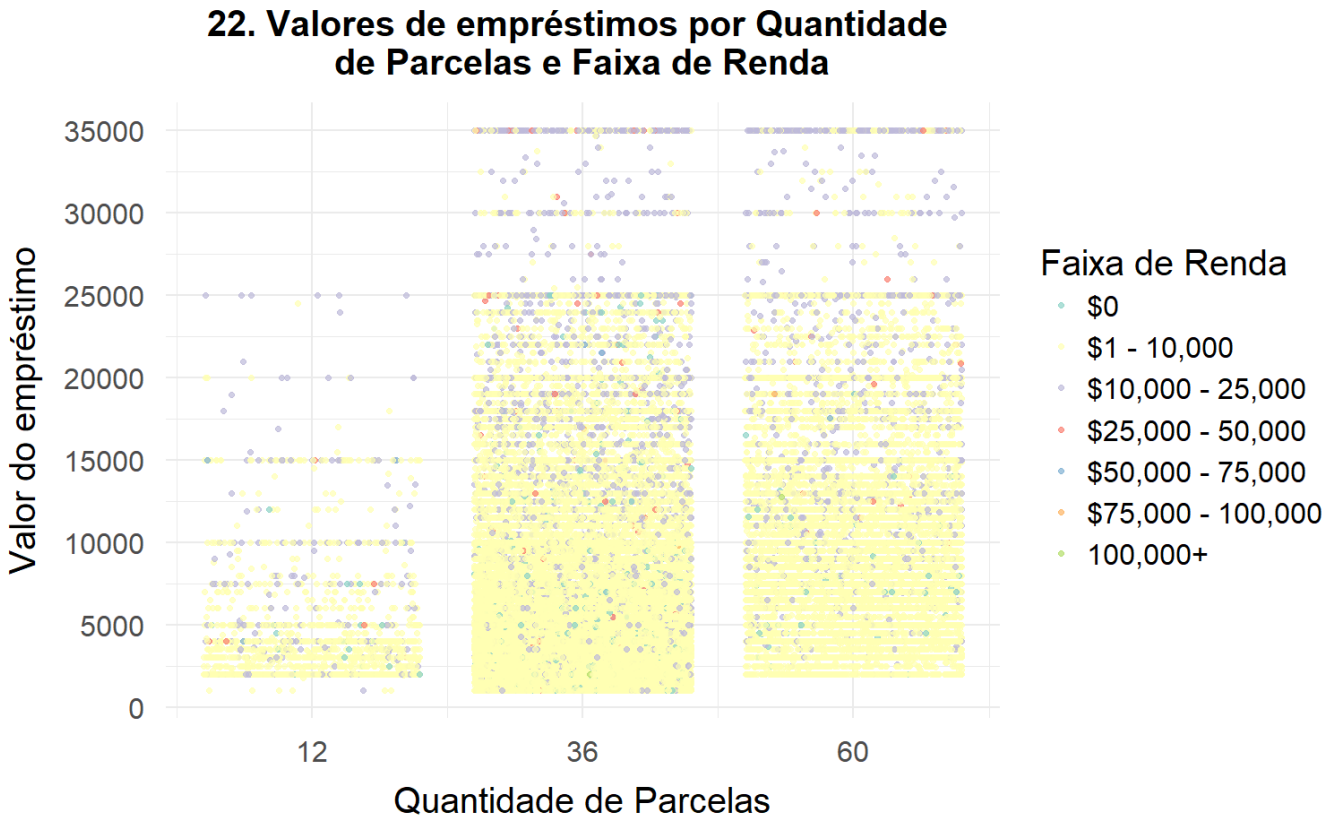



Podemos identificar algumas informações da relação entre o valor do empréstimo e em quantas parcelas eles geralmente são realizados. Note que a maior concentração de empréstimos para 12 parcelas são com valor de até US$ 5000. Para parcelamento em 36 meses, a maioria dos emrpéstimos são de até 10000, com um pequeno pico em 15000. Já para os empréstimos para pagamento em 60 meses, os valores são maiores. Podemos notar que os maiores picos estão em 10000 e 15000.

Em resumo a maioria dos parcelamento realizados em 12 vezes são de valores até US$ 5000, provavelmente para não gerar um empréstimo com valor alto de parcela. Valores baixos a intermediários, são parcelados em 36 vezes. Já para valores médios ou altos, geralmente são parcelados em 60 vezes.




A taxa de juros para quem faz o emprestimo em 12 parcelas é bem menor.

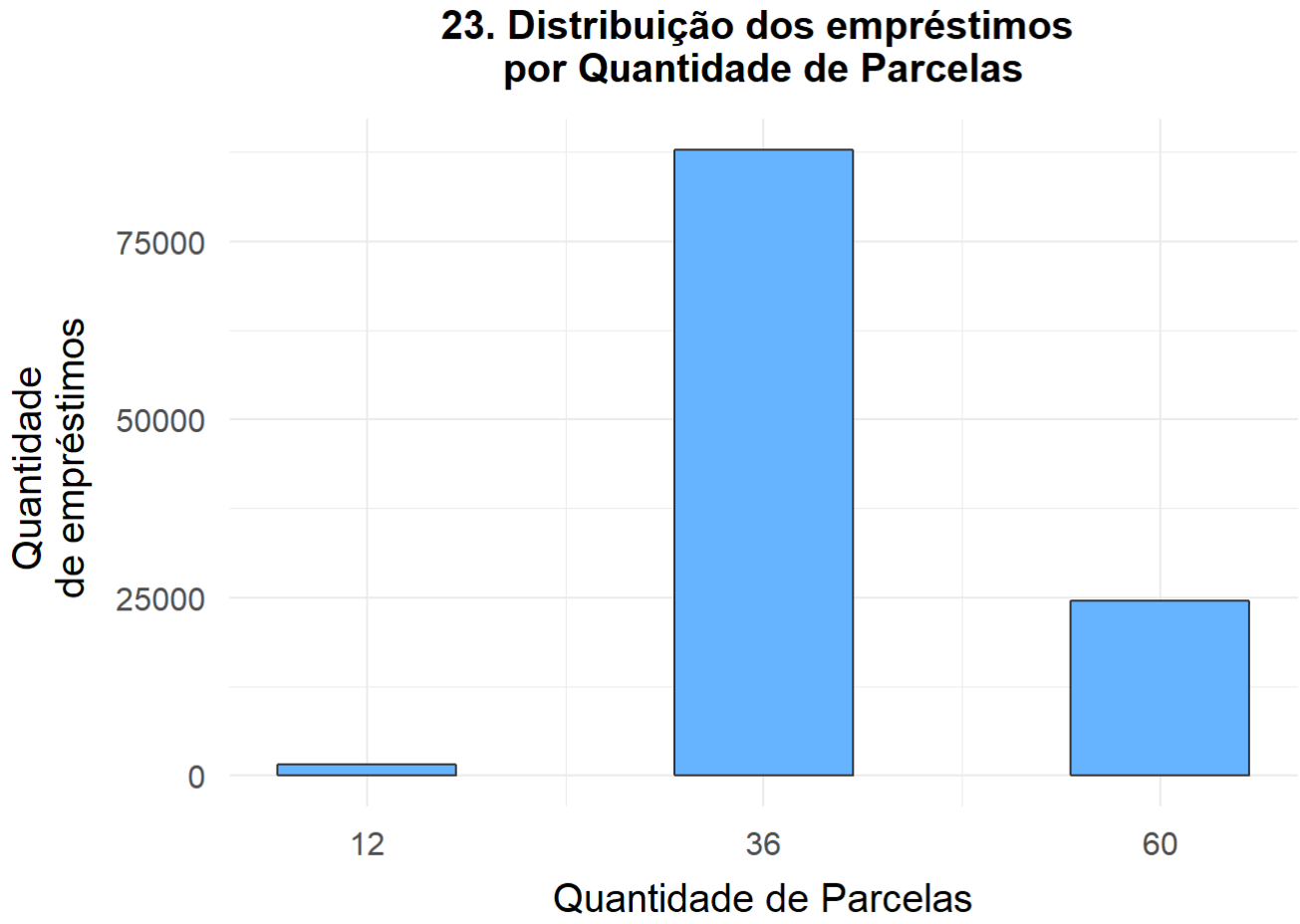


A quantidade de empréstimos para pagamento em 12 meses é bem menor que para 36 e 60 meses. Em todos os casos é nítido que empréstimos com valores maiores, em geral são mais realizados por quem tem a renda mais maior (para empréstimos com pagamento em 12 parcelas - ver ocorrências acima de 15000 e para 36 e 60 meses ver ocorrências acima de 25000).

Vamos aprofundar e verificar a média de valor de empéstimos por faixa de renda:

Considerando que a maioria dos mutuários tem renda entre US$ 1 e 10000, a resposta para a pergunta 4 é sim. Em geral, quem tem maior renda, realiza empréstimos de maior valor.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



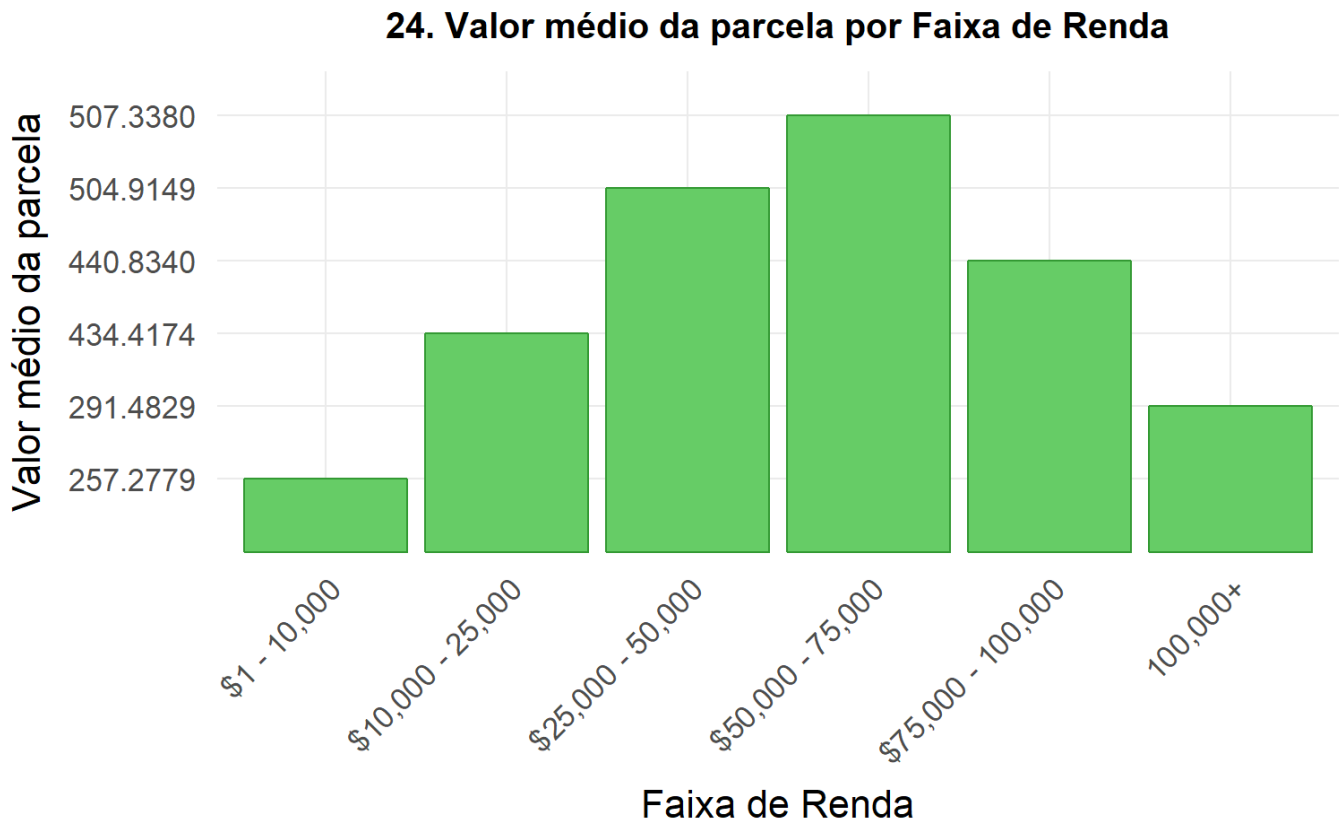

Apensar da opção de parcelamento em 12 vezes ser a opção com a menor taxa de juros, a maioria dos mutuários optam pelo pagamento em 36 meses. Isso se deve provavelmente ao valor da parcela.

Vamos analisar a média de valor por número de parcelas:

```
## df$Term: 12
## [1] 383.9334
## ------------------------------------------------------------
## df$Term: 36
## [1] 258.4879
## ------------------------------------------------------------
## df$Term: 60
## [1] 315.2104
```

Realmente, o menor valor médio de parcelas são para parcelamentos até 36 vezes. E média do valor mensal para empréstimos de 12 vezes é bem maior que para as demais formas de parcelamento.

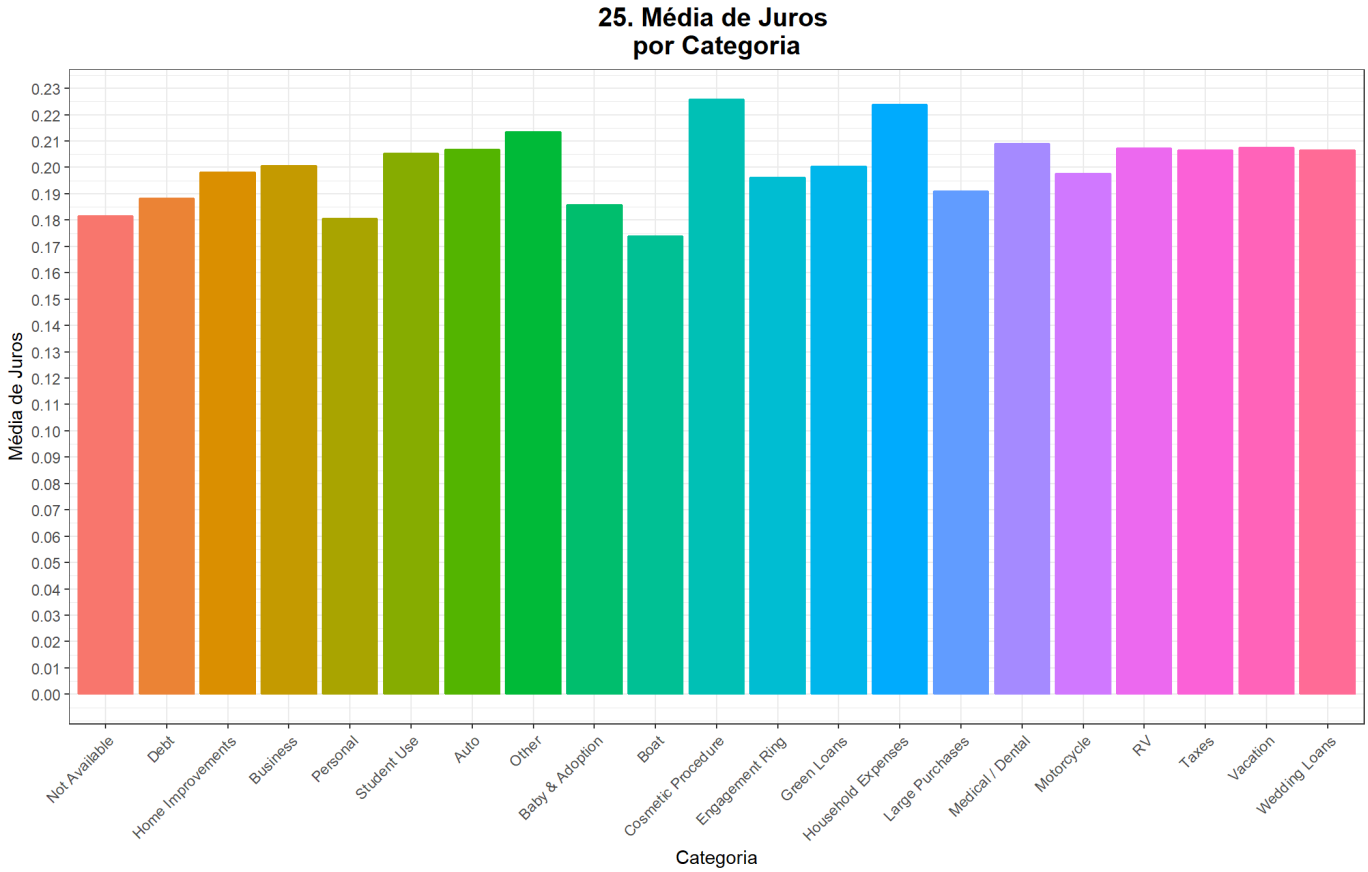
VALOR médio MENSAL DA PARCELA POR FAIXA DE SALÀRIO




A faixa de renda que paga o maior valor de parcela é de US$ 75 a 100000, seguido da faixa de US$ 25000 a 50000.

Considerando que o maior número de empréstimos são de mutuários com rendas entre 1 a 10000, podemos responder que sim para a pergunta número 3.

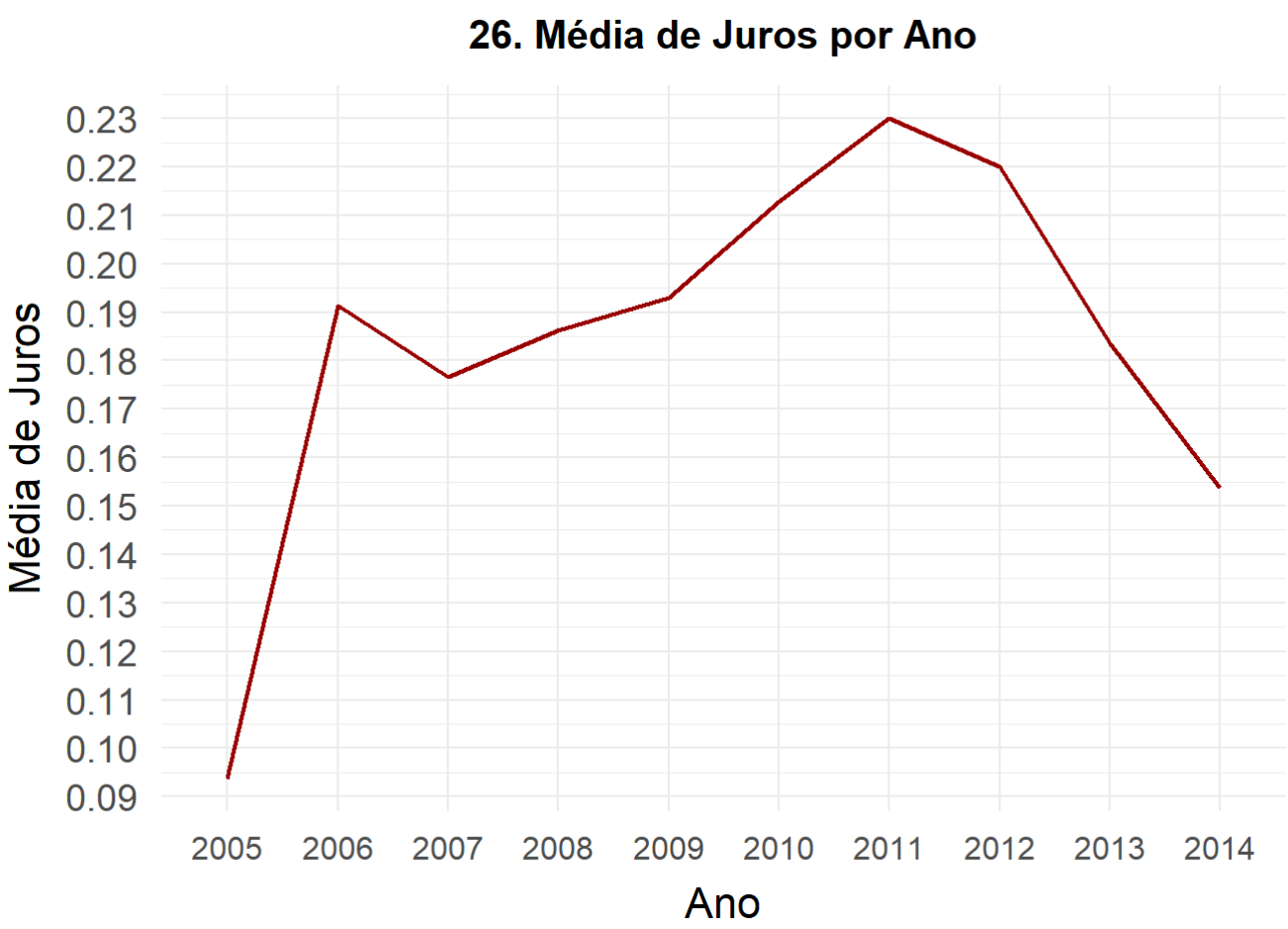
Levando em conta que a maior parte dos empréstimos são realizadas por mutuários com faixa de renda entre 1 a 10000 (conforme gráfico 17.1) e esta faixa de renda tem o valor médio da parcela de aproximadamente US$ 260 conforme o gráfico acima, confirmamos então o que o comentário do gráfico 23 apresenta, onde o parcelamento em 36 vezes é o mais utilizado com uma parcela média aproximada justamente de US$ 260.




Respondendo a segunda pergunta sim, a taxa de juros também sofre variações de acordo com a categoria do empréstimo. A categorias com maior taxa de juros foi de Procedimentos cosméticos, seguido de Despesas com a Familia e Outros. Já as categorias com menor taxa foram: Barco, Pessoal e a categoria Não disponível.

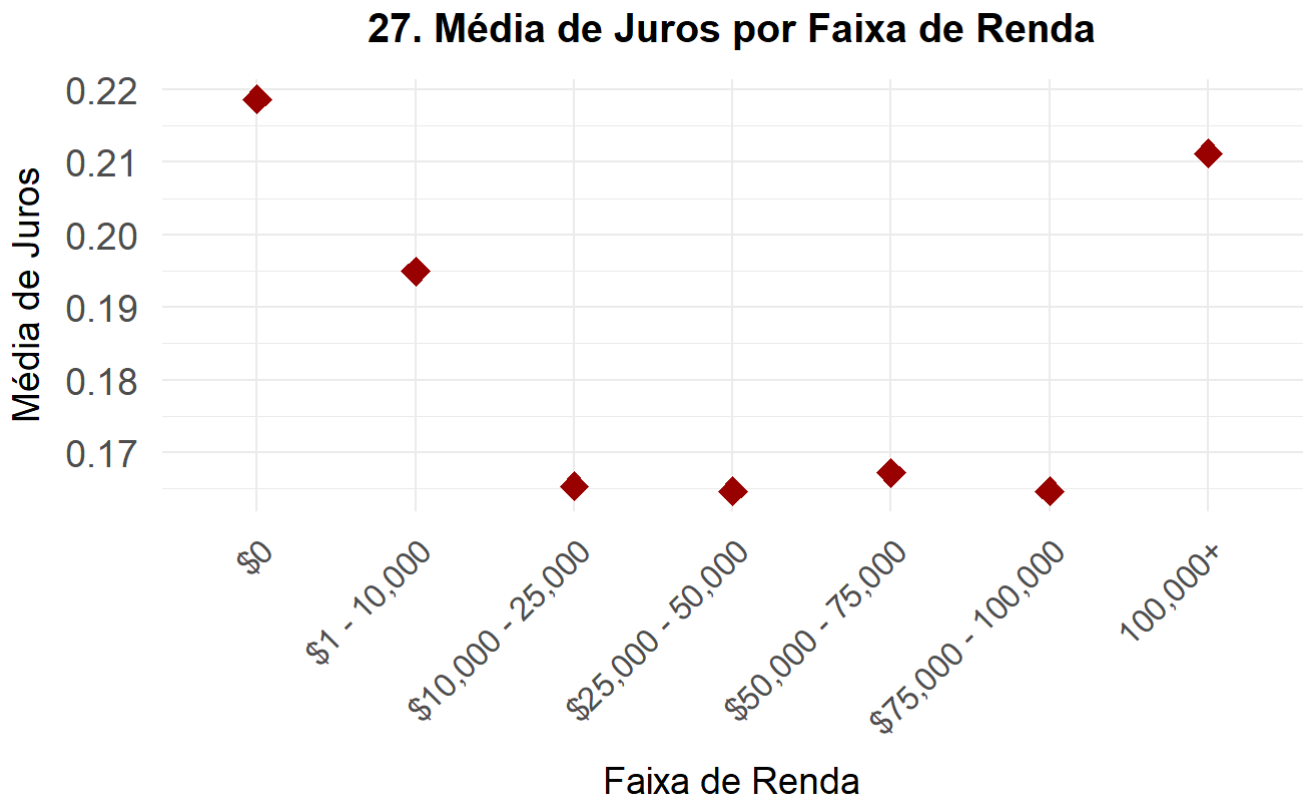
MÉDIA DA TAXA DE JUROS POR CATEGORIA

Vamos ver a variação da taxa de juros ao longo dos anos:




A taxa de juros iniciou em 2005 baixissima, no entanto foram realizados somente 22 empréstimos neste ano. A partir de 2006 houveram grandes aumentos até chegar em 2011, onde se deu inicio de quedas sucessivas da taxa de juros até chegar em 2014 onde iniciou com o juros bem mais baixo com relação aos demais anos.

Vamos analisar a taxa de juros conforme a renda do mutuário:

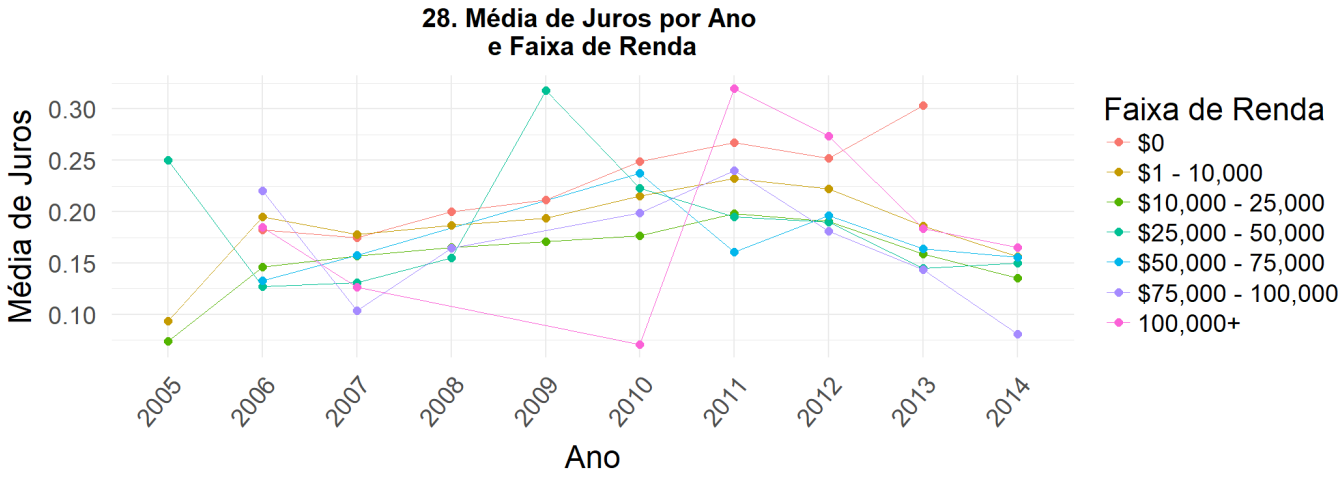


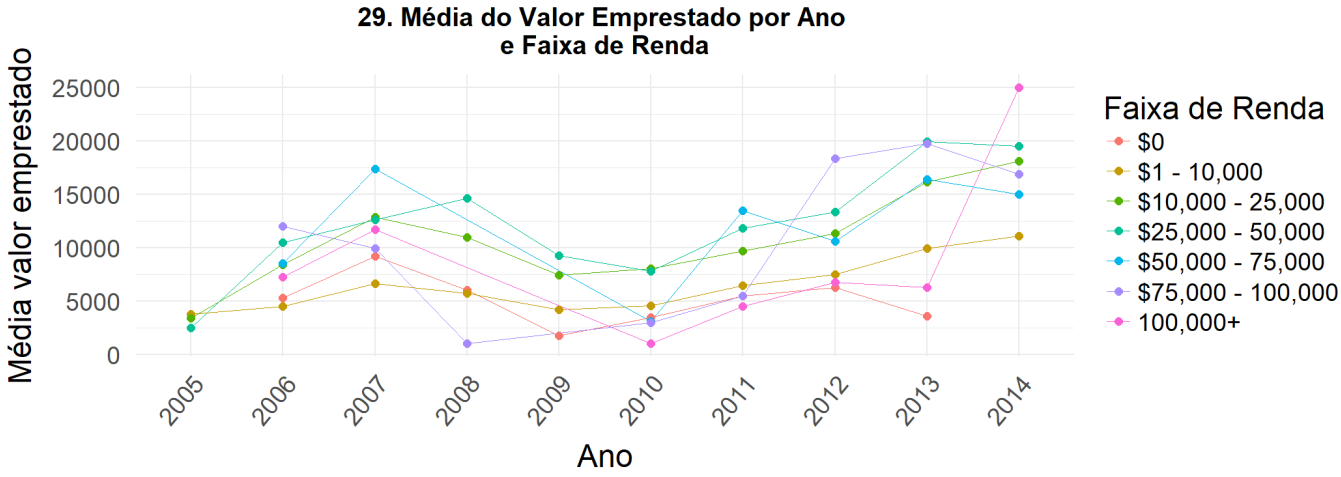

Podemos ver que quem não tem renda, paga em média a maior taxa de juros. Quem tem a renda entre US$ 1 a 10000, paga uma taxa de juros intermediária, com média entre 0,195%. As faixas que vão de 10000 a 100000 pagam as menores taxas de juros. Algo que ainda não foi possível explicar foi o fato de que a taxa de juros para quem tem renda acima de $ 100000 é uma das mais altas.

Analisando as informações acima, podmeos responder que sim para a pergunta 8, em geral, mutuários com maior renda tem juros mais baixos.

MÉDIA DE JUROS E DE VALOR DE EMPRESTIMOS POR FAIXA DE RENDA AO LONGO DOS ANOS

Nos gráficos abaixo podemos analisar o comportamento tanto da média da taxa de juros quanto a média do valor emprestado por faixa de renda. Talve até identificar pontos de um que reflete no outro. Segue:



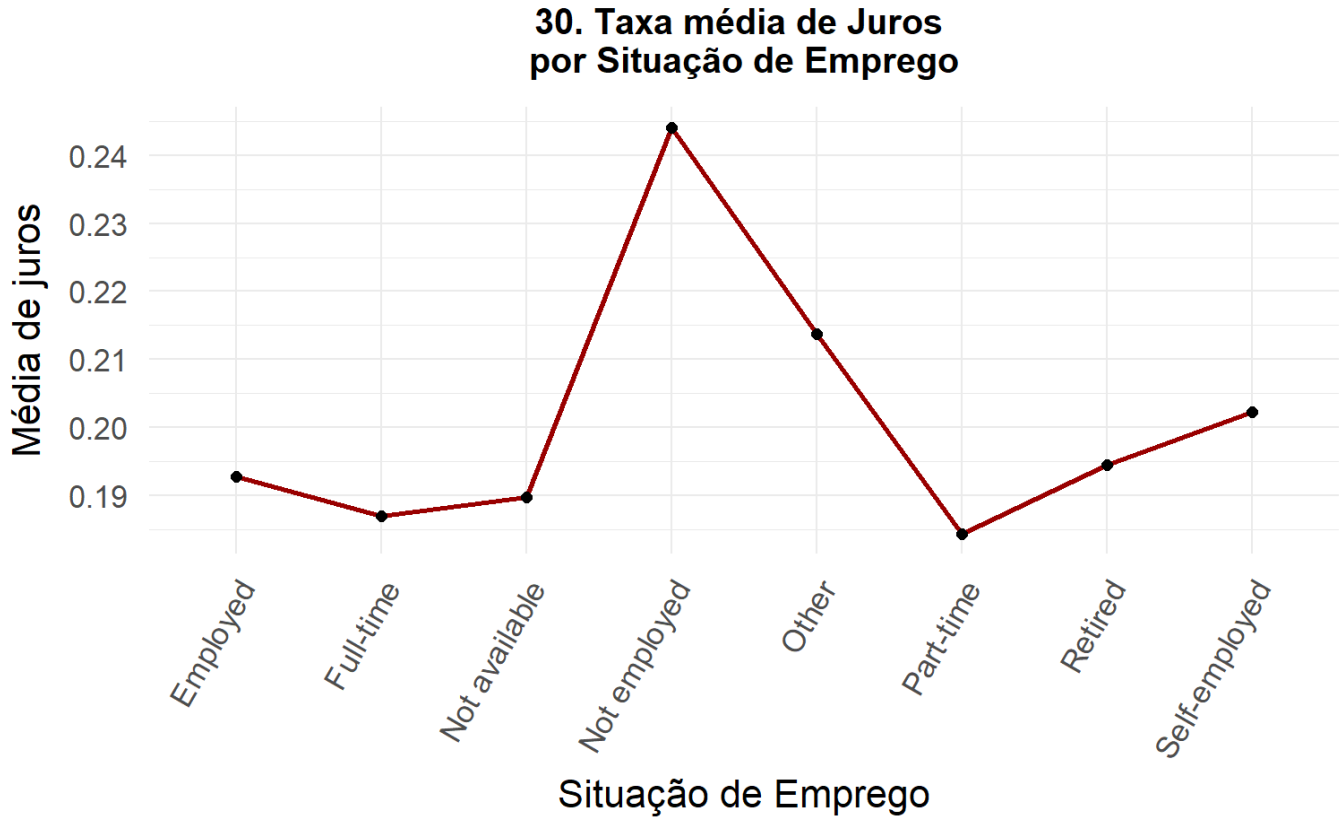

Para analisar os gráficos acima, é necessário investigar caso a caso. Vamos ver algumas descobertas:

28. Média de Juros por Ano e Faixa de Renda As faixas que vão de US$ 1 a 25000 tem uma evolução estável.

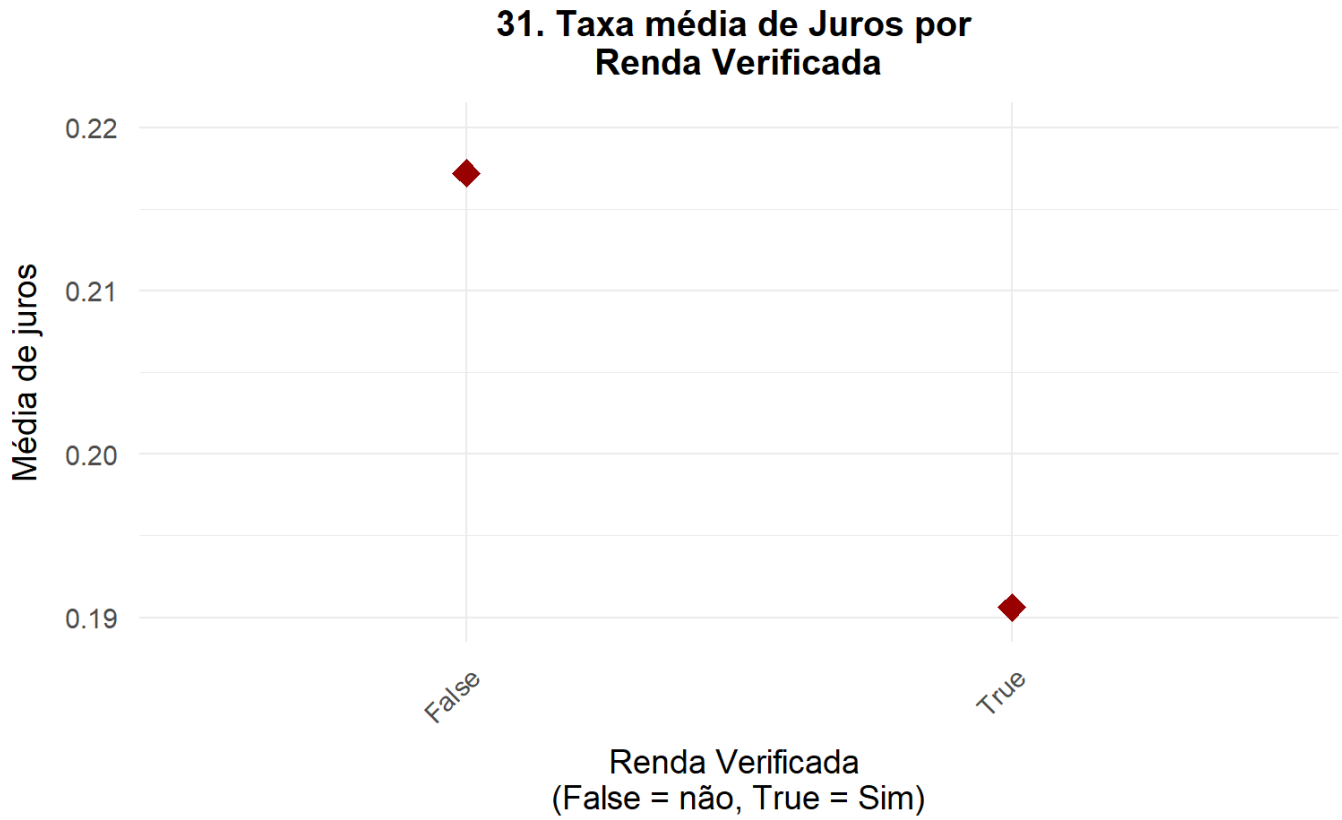
- Dá pra ver que quem tem zero renda (linha em vermelho), teve as maiores taxas de juros na maioria dos anos.
- A faixa de renda entre US$ 1 a 10.000 também ficou entre as maiores, é onde estão concentradas a maior quantidade de empréstimos.
- Quem tem salário acima de US$ 100.000 até o ano de 2010 teve uma taxa de juros baixissima. No entanto, ela aumentou muito tornando-se uma das maiores de 2011 em diante.
- Analisar as outras três em separado.
- As faixas de US$ 25000 a 50000 e acima de 100000 foram as faixas com maior variação de juros entre os anos.

29. Média do Valor Emprestado por Ano e Faixa de Renda A idéia era tentar achar relação de aumento ou diminuição de empréstimos com aumento ou diminuição da taxa de juros.

- Para a faixa de 75 a 1000 houve um grande aumento de empréstimos a partir de 2012 onde a taxa de juros caiu bastante (ver gráfico anterior)
- 2010 apesar de ser o ano com menor taxa de juros para quem tem renda acima de US$ 100000, foi um ano com pouquíssimos emprestimos para essa faixa de renda. (continuar)
- Renda na faixa de US$ 50000 a 75000 (linha azul) tende a realizar mais empréstimos sempre que a taxa de juros cai.
- A partir de 2011 conforme vimos no gráfico 26, a taxa de juros começou a cair a cada ano. Podemos ver isso no gráfico 28 e o reflexo do aumento de valores empréstados a partir deste ano.




Respondendo a pergunta 5 sim, mutuários que estão dempregados tem uma taxa de juros bem maior do que os que estão empregados.




A resposta para a pergunta 6 também é afirmativa, mutuários que estão empregados e comprovam sua renda tem juros menores do que as que não comprovam a renda. Vale ressaltar que os mutuários que não comprovaram renda ainda tem uma taxa de juros média menor do que quem tem a situação de emprego como “Não Empregado” (Not employed).

# ANÁLISE DA CORRELAÇÃO

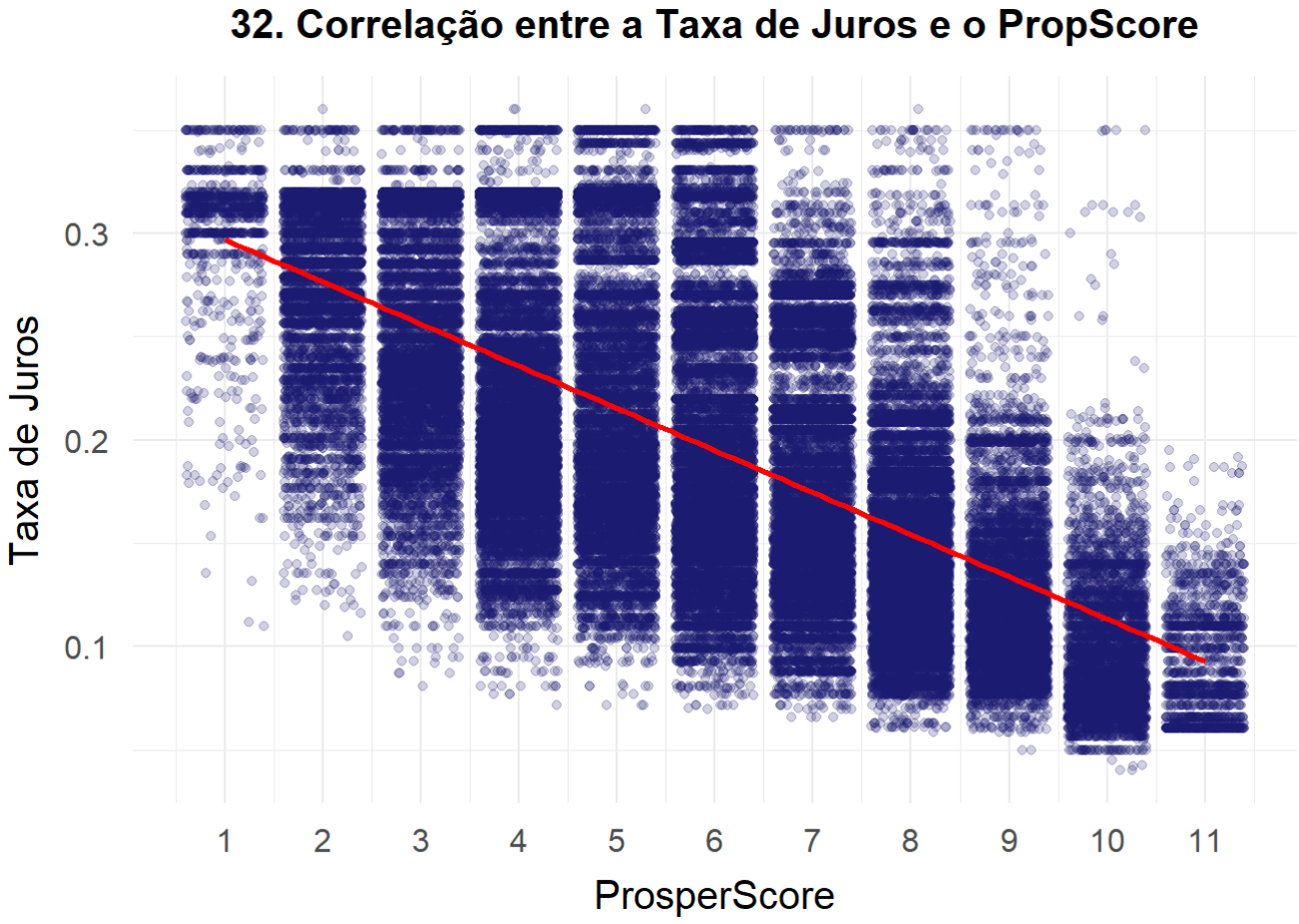
| Range coeficiênte | Força da correlação |
|---|---|
| 0,2 < r < 0,4 | Correlação fraca |
| 0,4 < r < 0,7 | Correlação moderada |
| 0,7 < r < 0,9 | Correlação forte |

*Onde r = coeficiênte de correlação linear*

Vamos calcular o coeficiênte de correlação usando o método de Pearson para encontrar a correlação entre o PropScore e a Taxa de Juros:

> Para análise do PropScore, iremos considerar somente as observações onde esta variável é diferente de nulo.

```
##
##  Pearson's product-moment correlation
##
## data:  df_score$ProsperScore and df_score$BorrowerRate
## t = -248.98, df = 84851, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6536072 -0.6458311
## sample estimates:
##        cor
## -0.6497361
```



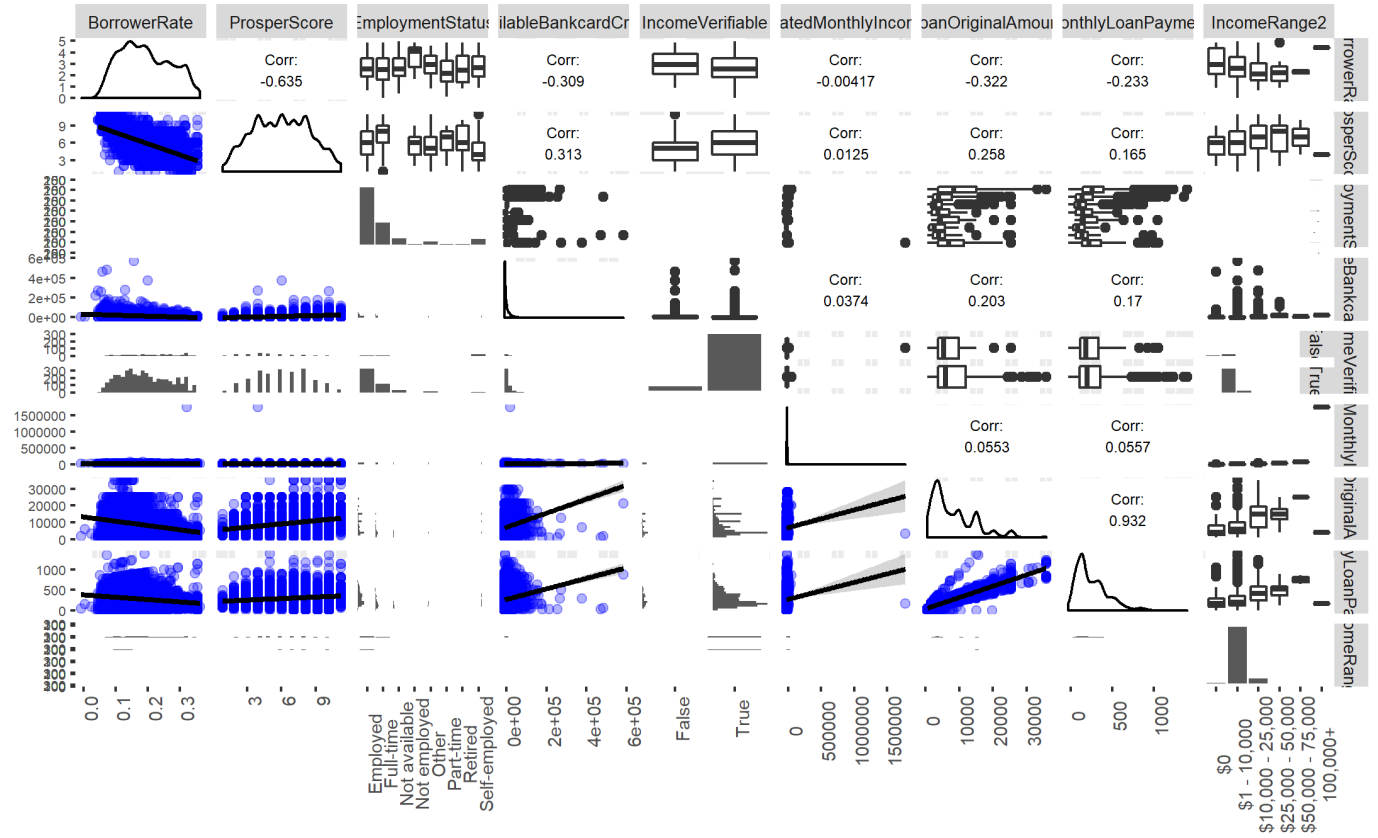

Como vimos, o coeficiênte encontrado foi -0,64. Podemos observar que existe uma correlaçao fraca, conforme a referência informada na tabela. Significa que a Taxa de Juros tem uma correlação fraca com com o Propscore. A taxa de juros provavelmente é influenciado também por outras variáveis, como por exemplo a quantidade de percelas, etc.

Existem outras variáveis que provavelmente tem uma correlação entre si. Ao invés de realizarmos suposições, vamos gerar uma matriz de correlação para facilitar nossa Análise.

**33. MATRIZ DE CORrelação**

Vamos agora selecionar somente alguns campos da nossa base:

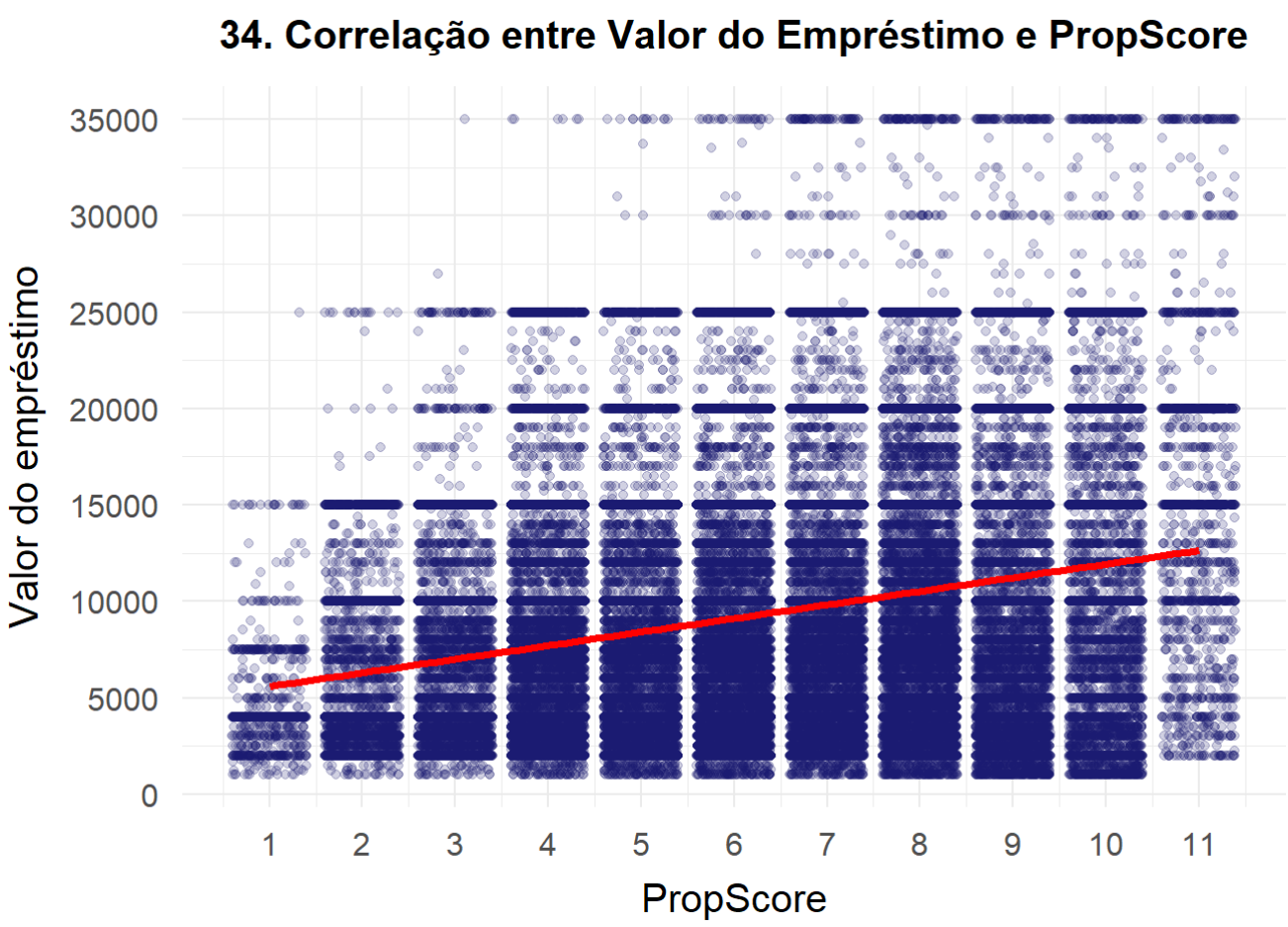
> Obs.: Para possibiitar a geração da matriz de correlação é utilizada uma amostra da nossa base de dados. Significa que pode gerar pequenas diferenças de resultados se comparado com o cálculo da base toda.




A análise deste gráfico é encontrando a intercessão entre a coluna (ver a variável no título da coluna) e a linha (ultima informação ao lado direito da linha). O ponto onde eles se encontram irá mostrar as informações de correlação. Por exemplo, o título da segunda posição PropScore, se relaciona com a linha 1 que é o BorrowRate (Taxa de Juros). Onde os dois se cruzam esta sendo exibido o valor -0.653, o que quer dizer que quando o propscore aumenta a taxa de juros tende a cair. Para analisarmos o gráfico desta mesma correlação, basta inverter a análise. Procure agora primeiro a informação BorrowRate como coluna e depois a informação Propscore como linha. Você verá o gráfico desta correlação.

Algumas correlações identificadas são óbvias, e por isso não poderemos considerá-las, como por exemplo o valor mensal a ser pago com o valor total emprestado. Esta relação quase sempre estará relacinoada. Neste caso, o coeficiente de correlação foi bastante alto: 0,93.

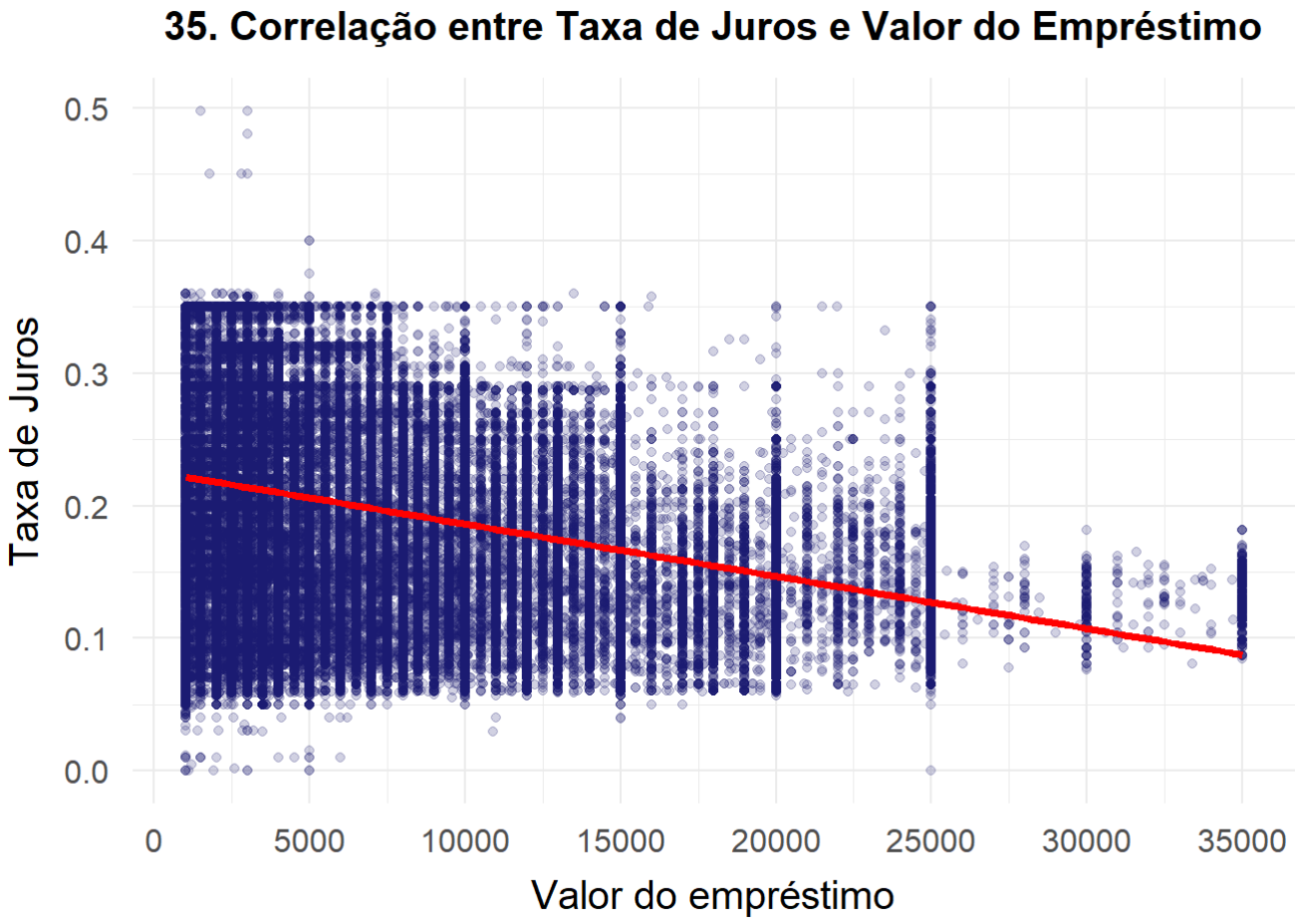
Vamos analisar a correlação entre o Propscore e o Valor emprestado.




Apesar de haver uma tendência como pode-se ver os valores mais altos a partir do PropScore 7, a correlação identifica é considerada fraca.

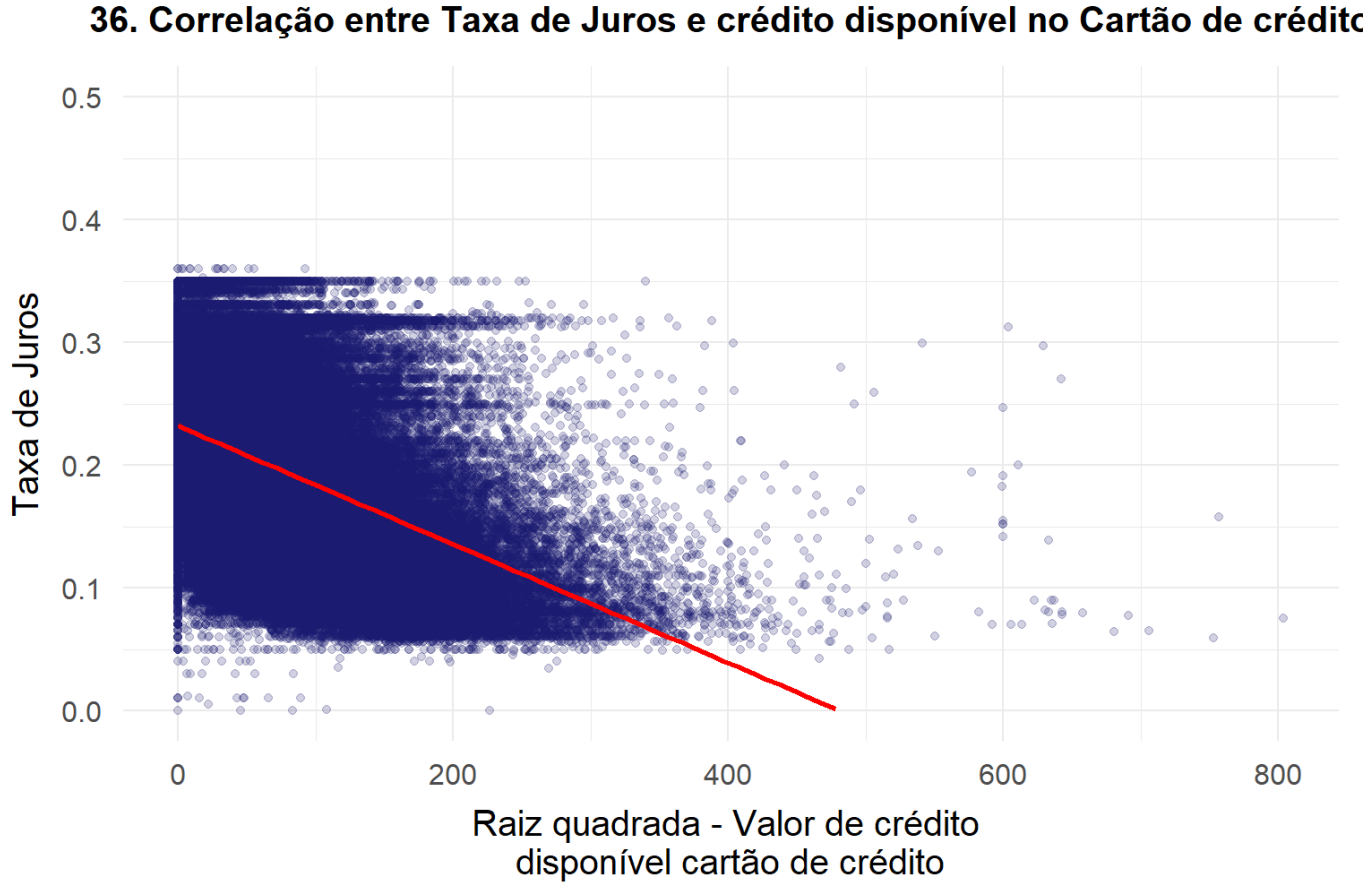
Será que quanto maior o valor, menor é a taxa de juros?

A matriz de correlação indica uma correlação de -0.319 entre a Taxa de Juros e o Valor do empréstimo. Vamos ver o gráfico:



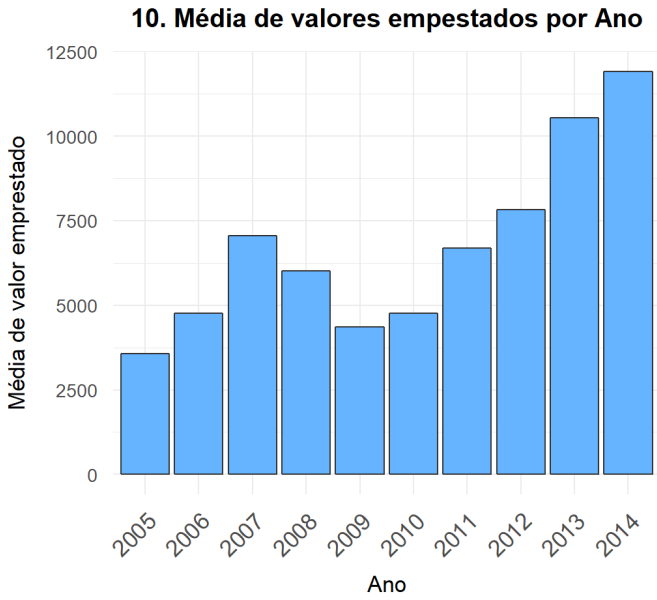

Também é uma correlação fraca.

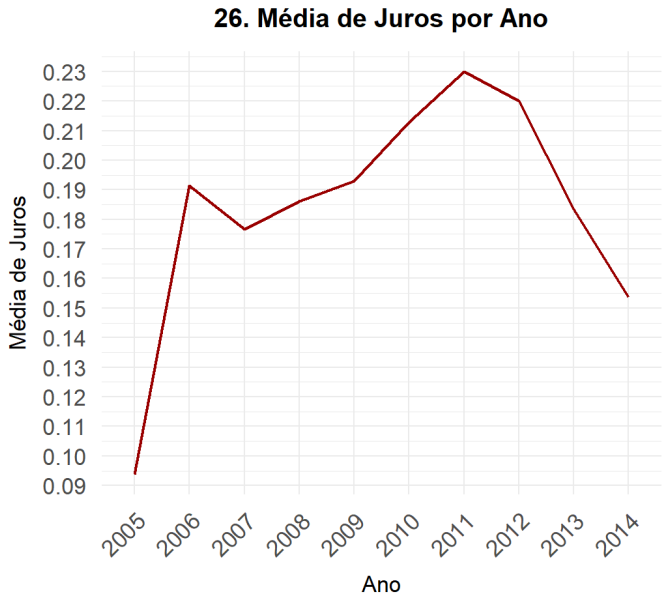
Devido à grande diverença entre os valores disponíveis, tive que usar a funão sqrt para calcularmos a raiz quadrada do valor disponível do cartão de credito e assim deixar o gráfico melhor para visualização:




Podemos ver no gráfico uma ligeira tendência que indica que quanto maior o valor disponível no cartão de crédito, menor é a taxa de juros. O coeficiênte de correlação neste caso foi -0.357.

# GRÁFICOS FINAIS E RESUMO

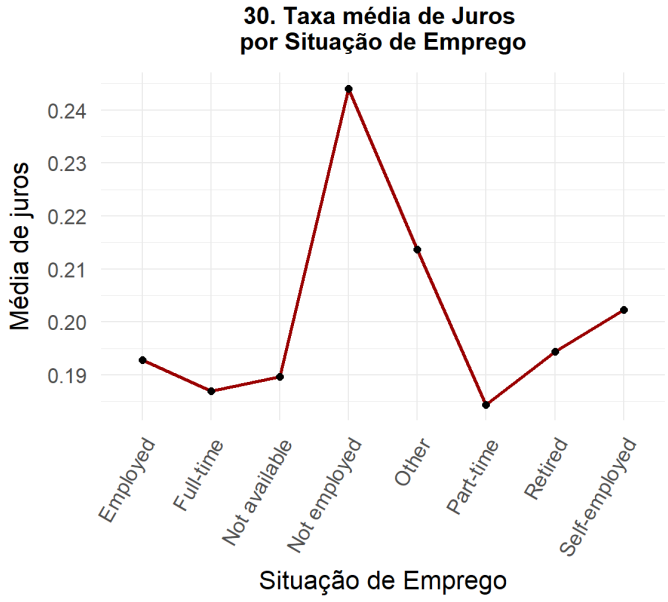


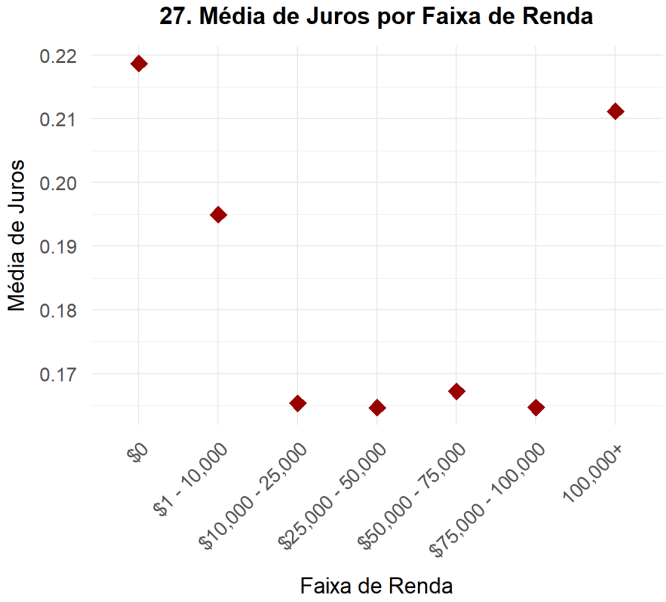




As análises mostram uma tendência que indica que em 2014 será o ano com maior valor de empréstimo de todos os anos. O gráfico de média de juros, indica uma queda ascentuada da taxa de juros a partir do ano de 2012. Com exceção de 2005, onde foi um ano atípico, o ano de 2014 até agora é o ano com menor taxa de juros.

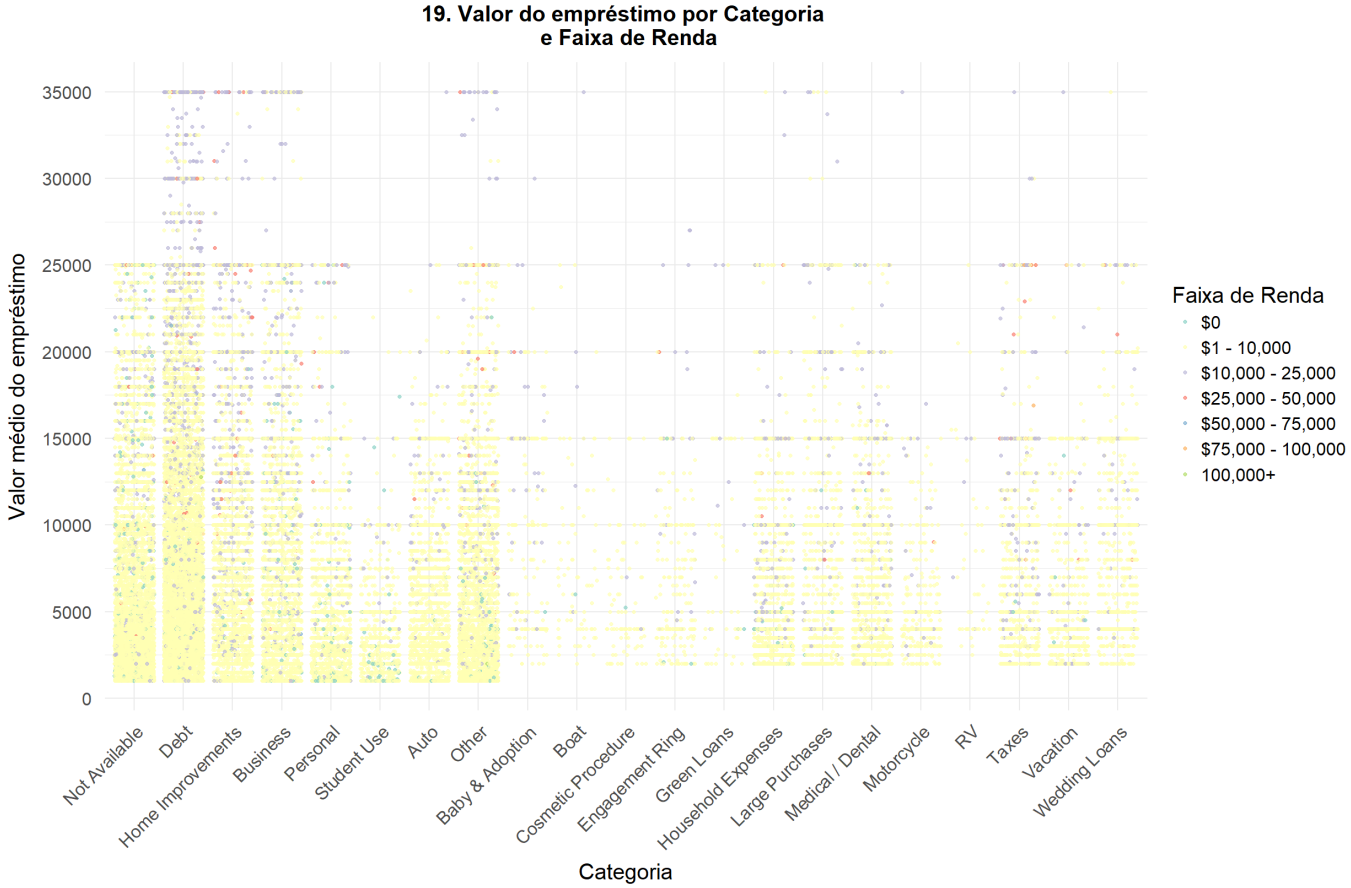
Sabe-se que a taxa de juros está intimamente ligada ao risco. Quanto maior o risco do investidor em não receber o pagamento pelo empréstimo, maior será a taxa de juros.

Os gráficos abaixo indicam fatores que podem apresentar risco:




Segue abaixo algumas descobertas do estudo: * Quem esta desempregado, tem a taxa de juros muito maior do que quem está empregado. * Vale ressaltar que dos mutuários que estão empregados, os que comprovaram sua renda tem juros muito menores do que os que não comprovaram. * Em geral, quem tem maior renda paga menos juros para o empréstimo (Com exceção de quem ganha acima de US$ 100.000. Para estes casos, uma hipótese que poderia justificar o alto * percentual de juros talvez possa ser a de que o mutuário não administra bem os seus recursos).



Podemos visualizar a dispersão dos empréstimos visualizando os valores separados por categoria e faixa de renda. Fica claro que maior quantidade de empréstimos é realizada por mutuários com renda entre US$ 1 a 10000. Para empéstimos com valor a partir de US$ 25000, essa faixa quase não tem empéstimos. Provavelmente não são aprovados valores tão altos para esta faixa de renda. Acima de 25000 é mais comum empréstimos para quem tem renda entre US$ 10000 a 25000.

# REFLEXÃO

Esse estudo foi realizado analisando uma série de informações da base de dados, procurando conhecer as informações e uma maneira geral e direcionando o estudo para responder perguntas sobre a taxa de juros aplicada nos diferentes empréstimos.

Apesar de termos analisado uma série de variáveis, pode-se dizer que para o cálculo da taxa de juros é considerado não só um fator, e sim uma série de fatores. Alguns influenciam mais, outros nem tanto. O PropScore é o índice que mais teve influência com no percentual de juros utilizado no empréstimo. Ele provavelmente é calculado considerando uma série de características, como por exemplo o fato da pessoa estar trabalhando ou não, sua faixa de renda, se a renda foi ou não foi comprovada, etc. Outro fator importante é a quantidade de parcelas para pagamento do empréstimo. Em 12 vezes, o juros aplicado é bem menor do que parcelamentos em 36 e 60 meses.

Esta é uma base de dados bastante rica, e que possibilita uma série de insights. A maior dificuldade na realização deste estudo foi em meio a 81 variáveis identificar as com maior relevância. Apesar de haver um dicionário de dados, a forma exata de como cada variável funciona muitas vezes não está clara. A criação do mapa mental ajudou bastante nesse processo de eliminação de variáveis que não faziam tanto sentido, me ajudou a focar no objetivo e a identificar mais questões a serem respondidas. A partir daí, pouco a pouco fui me familiarizando com os comandos do R e a medida que eu fui avançando, mais insights foram acontecendo.

Um próximo passo para esse estudo seria analisar as características dos empréstimos que estão inadimplentes. Identificar os diferentes padrões destes com relação aos empréstimos com pagamento em dia. Para realizar esta identificação, poderia também serem utilizadas técnicas de machine learning, como por exemplo classificação ou clusterização. Uma outra sugestão seria analisar as demais variáveis da base de dados disponíveis na base de dados juntamente com informações mais atualizadas, contendo dados de anos mais atuais (considerando que o ultimo ano que temos é o de 2014).

Apesar do propósito deste projeto ter sido cumprido, tenho certeza que com mais tempo e mais prática, poderia criar gráficos mais sofisticados, realizando a junção de visões facilitando ainda mais a compreensão e prosseguindo com a identificação de novas descobertas.

Até o próximo projeto!

## REFERÊNCIAS:


- Prosper
- https://www.prosper.com/ (https://www.prosper.com/)
- https://en.wikipedia.org/wiki/Prosper_Marketplace (https://en.wikipedia.org/wiki/Prosper_Marketplace)

- Otimização do tamanho das imagens
- http://optipng.sourceforge.net/ (http://optipng.sourceforge.net/)

- https://www.zevross.com/blog/2017/06/19/tips-and-tricks-for-working-with-images-and-figures-in-r-markdown-documents/ (https://www.zevross.com/blog/2017/06/19/tips-and-tricks-for-working-with-images-and-figures-in-r-markdown-documents/)
- https://www.linuxhelp.com/install-jpegoptim-optipng-linux/ (https://www.linuxhelp.com/install-jpegoptim-optipng-linux/)

- Coeficiente de correlação
- http://w3.ufsm.br/adriano/aulas/coreg/Aula%2001%20Correla%E7ao%20Linear.pdf (http://w3.ufsm.br/adriano/aulas/coreg/Aula%2001%20Correla%E7ao%20Linear.pdf)

- Formatação R Markdown
- https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf)

- Visualização
- http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html (http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html)
- https://stackoverflow.com/questions/39709745/decreasing-the-line-thickness-and-corr-font-size-in-ggpairs-plot (https://stackoverflow.com/questions/39709745/decreasing-the-line-thickness-and-corr-font-size-in-ggpairs-plot)
- https://github.com/ggobi/ggally/issues/6 (https://github.com/ggobi/ggally/issues/6)