

Relatório IC

Fernanda Buzza Alves Barros

___ de ___ de ___

INTRODUÇÃO

Problemas de dados faltantes em pesquisa são recorrentes em bancos de dados. Para a solução desses problemas existem vários métodos que podem ser utilizados. Entretanto, todos os métodos possuem uma questão principal: como inferir os valores não observados?

Para a resposta dessa pergunta, temos que o ideal seria ter os dados, porém na falta deles temos que utilizar o método que melhor se ajusta a distribuição dos dados.

Nessa pesquisa utilizaremos o método proposto por Rubin (1987), Van Buuren e Groothuis-Oudshoorn (2011), que é conhecido como Imputação Múltipla.

METODOLOGIA

A Imputação Múltipla consiste em gerar valores (m vezes) para os dados faltantes, ela cria uma matriz com todas as M imputações. Para gerar essas imputações existem alguns métodos, como por exemplo *Predictive Mean Matching (pmm)* e *Unconditional Mean Imputation (mean)*, que serão os métodos utilizados nesse estudo.

Predictive Mean Matching (pmm)

Unconditional Mean Imputation (mean)

RESULTADOS

Banco de dados

Para realizar a imputação dos dados utilizamos o banco de dados *US Term Life insurance* do pacote *CASdatasets* disponível no software R. As imputações e os resultados foram obtidos utilizando esse mesmo software estatístico. O banco de dados possui 18 variáveis com 500 observações, como pode ser visto abaixo.

```
## 'data.frame':   500 obs. of  18 variables:
##  $ Gender       : int  1 1 1 1 1 1 0 1 1 1 ...
##  $ Age          : int  30 50 39 43 61 34 75 29 35 70 ...
##  $ MarStat      : int  1 1 1 1 1 2 0 1 2 1 ...
##  $ Education    : int  16 9 16 17 15 11 8 16 4 17 ...
##  $ Ethnicity    : int  3 3 1 1 1 2 1 1 3 1 ...
##  $ SmarStat     : int  2 1 2 1 2 1 0 2 1 2 ...
##  $ Sgender      : int  2 2 2 2 2 2 0 2 2 2 ...
##  $ Sage         : int  27 47 38 35 59 31 0 31 45 74 ...
##  $ Seducation   : int  16 8 16 14 12 14 0 17 9 16 ...
##  $ NumHH        : int  3 3 5 4 2 4 1 3 2 2 ...
##  $ Income       : int  43000 12000 120000 40000 25000 28000 2500 100000 20000 101000 ...
##  $ TotIncome    : int  43000 0 90000 40000 1020000 0 0 84000 0 6510000 ...
```

```
## $ Charity      : int  0 0 500 0 500 0 0 0 0 284000 ...
## $ Face         : int  20000 130000 1500000 50000 0 220000 0 600000 0 0 ...
## $ FaceCVLifePol : int  0 0 0 75000 7000000 0 14000 0 0 2350000 ...
## $ CashCVLifePol : int  0 0 0 0 300000 0 5000 0 0 0 ...
## $ BorrowCVLifePol: int  0 0 0 5 5 0 5 0 0 5 ...
## $ NetValue      : int  0 0 0 0 0 0 0 0 0 0 ...
```

Porém selecionamos as seguintes variáveis para realizar a pesquisa: Gênero (gênero do entrevistado); Idade (idade do entrevistado); Estado Civil (estado civil do entrevistado); Escolaridade (número de anos de escolaridade do entrevistado); Etnia (etnia do entrevistado); Renda (renda anual da família do entrevistado).

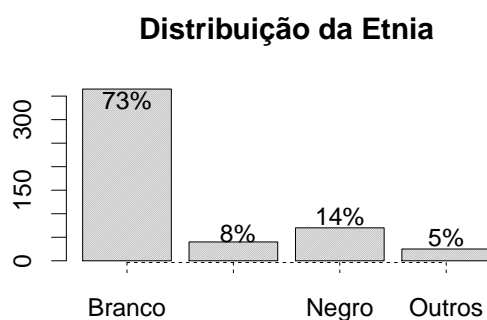
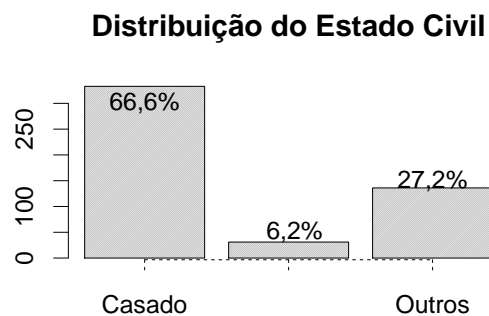
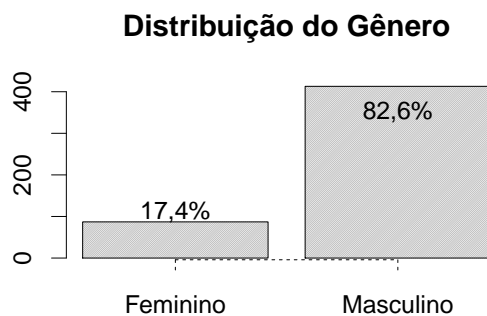
Primeiras observações do banco de dados original:

```
##   Gender Age MarStat Education Ethnicity Income
## 1      1  30      1      16          3  43000
## 2      1  50      1       9          3  12000
## 3      1  39      1      16          1 120000
## 4      1  43      1      17          1  40000
## 5      1  61      1      15          1  25000
## 6      1  34      2      11          2  28000
```

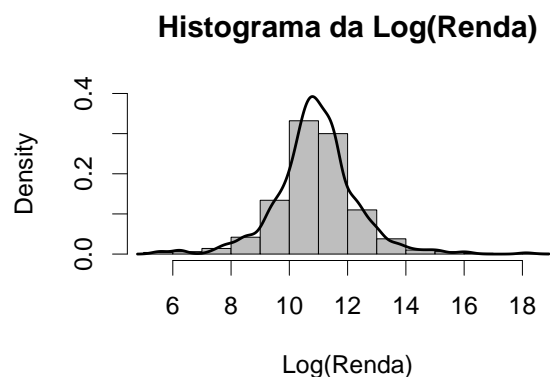
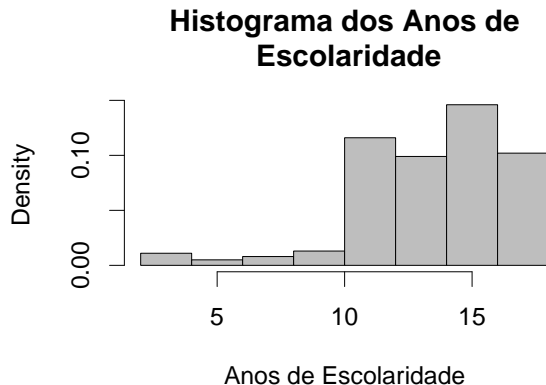
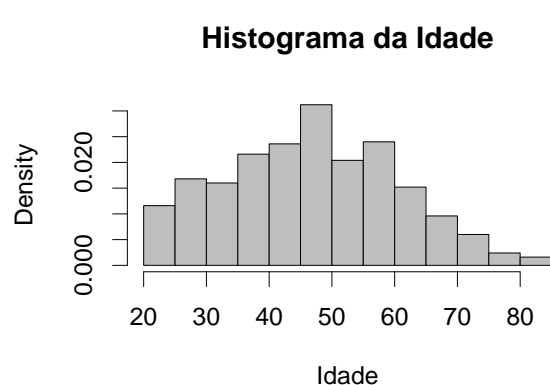
Análise Descritiva

Após a escolha das variáveis para esse estudo, iremos realizar uma análise descritiva de cada uma delas, e assim avaliar as relações existentes entre a variável resposta e as covariáveis do banco de dados. Ao final realizaremos um dos principais objetivos dessa pesquisa, que é verificar os possíveis questionamentos sobre a Renda a partir das outras variáveis.

Primeiramente analisaremos as variáveis individualmente, com interesse em suas distribuições e comportamentos. Pelos dados observamos que as variáveis contínuas são: Renda, Idade e Escolaridade, e as variáveis discretas são: Gênero, Estado Civil e Etnia.



Para as variáveis discretas temos que o banco de dados possui uma quantidade maior de entrevistados do sexo masculino (413 entrevistados) do que do sexo feminino (87 entrevistadas); para o estado civil temos uma concentração maior de respostas para os entrevistados Casados (333 entrevistados) e a menor quantidade de entrevistados pertence ao estado civil de Morando Juntos (31 entrevistados), sendo que o estado civil Outros possui 136 entrevistados; por fim para a etnia temos uma maior quantidade de entrevistados que possuem etnia Branco (365 entrevistados), sendo que as outras etnias possuem valores menores de entrevistados no banco de dados: hispânico (40 entrevistados), Negro (70 entrevistados) e Outros (25 entrevistados).



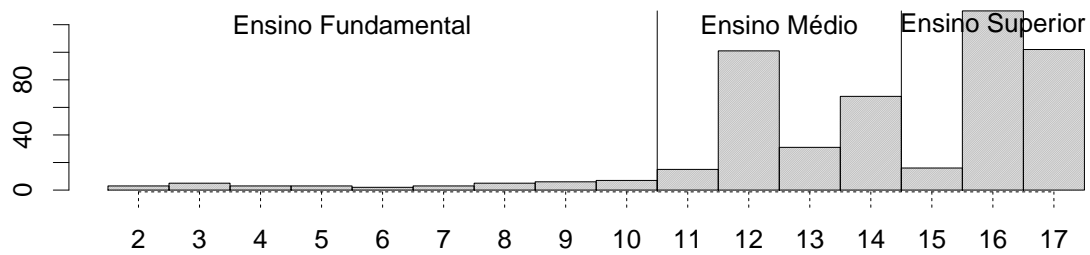
Para as variáveis contínuas temos que a variável Idade está bastante distribuída entre os 20 anos e 70 anos, após 70 anos vemos poucos entrevistados no banco de dados, sendo também que a idade máxima é 85 anos e a idade mínima é 20 anos. A média é 47.164 anos e a mediana 47 anos. O primeiro quantil é de 37 anos, representando a idade que deixa 25% das observações abaixo e 75% acima dessa idade. E o terceiro quantil é de 58 anos, representando a idade que possui 75% das observações abaixo dela e 25% das observações acima dela.

A distribuição da variável Escolaridade possui maior concentração de entrevistados após 10 anos de escolaridade.

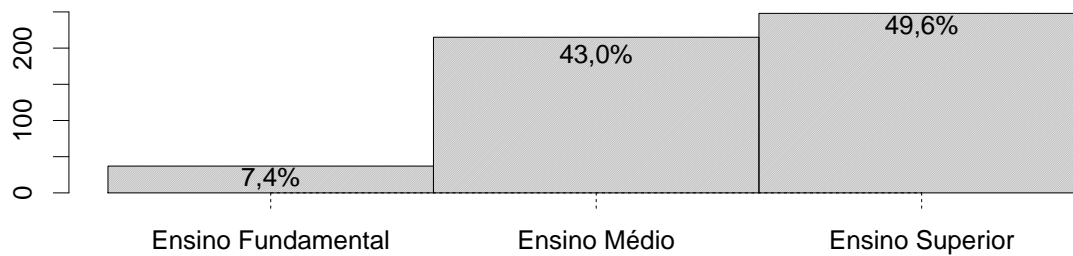
E para a variável Renda temos que a renda mínima anual é 260 dólares e a renda máxima anual é 75000000 dólares. A mediana e a média são 54000 dólares e 321021 dólares, respectivamente. O primeiro quantil é de 28000 dólares, representando o valor de renda anual que deixa 25% das observações abaixo dela e 75% acima dela. E o terceiro quantil é de 106000 dólares, representando a renda anual que possui 75% das observações abaixo dela e 25% das observações acima dela. Aplicamos o logarítmo para melhor visualização da distribuição da renda anual através do histograma e percebemos uma aparência com a distribuição normal.

Além da análise da variável Escolaridade em anos, foi realizada também a análise dos anos de escolaridade divididos pelos tipos de ensino existentes, que são: 2-10 anos de escolaridade é o Ensino Fundamental, 11-14 anos de escolaridade é o Ensino Médio e de 15-17 anos de escolaridade é o Ensino Superior, assim obtemos:

Distribuição dos Anos de Escolaridade

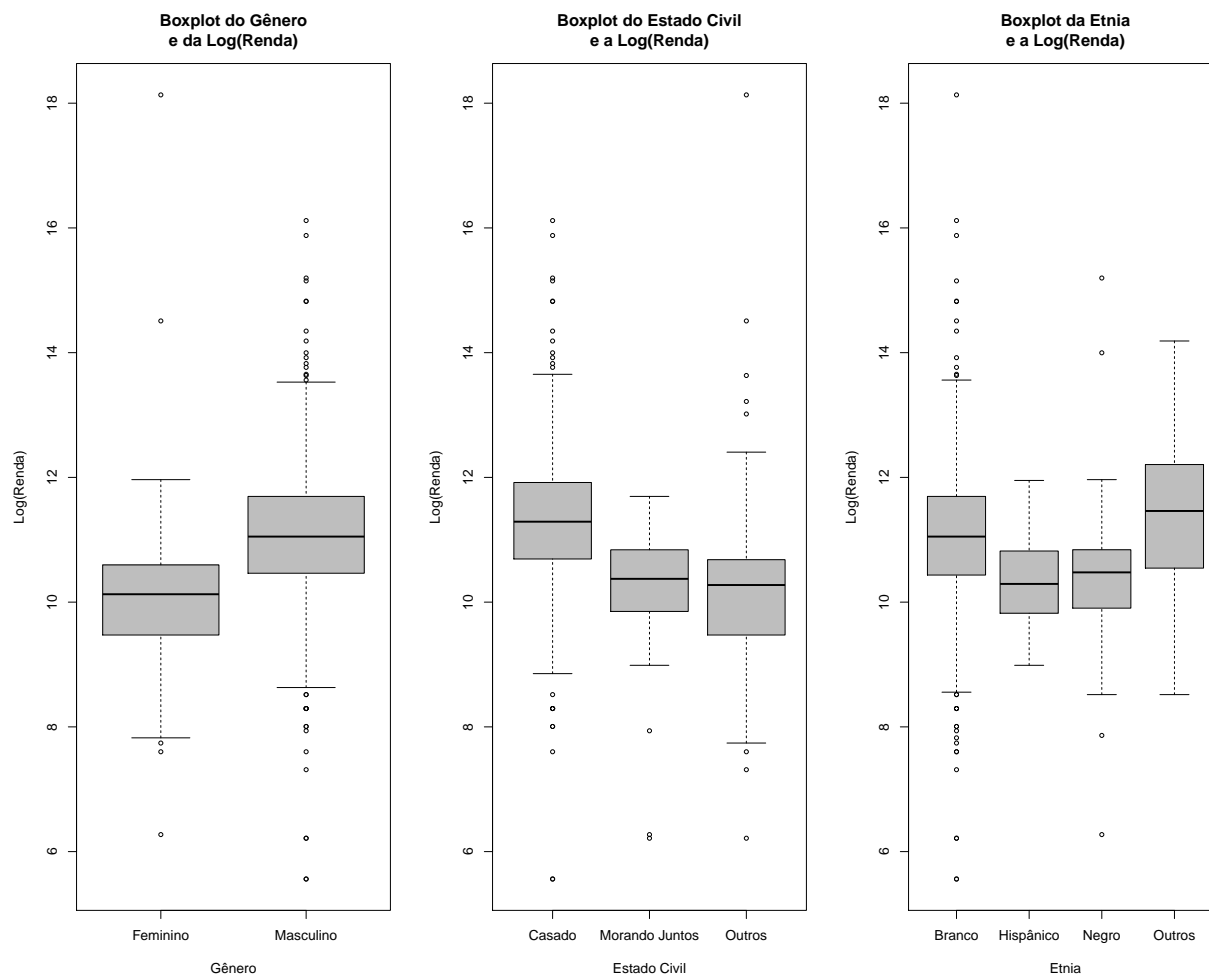


Distribuição dos Tipos de Ensino



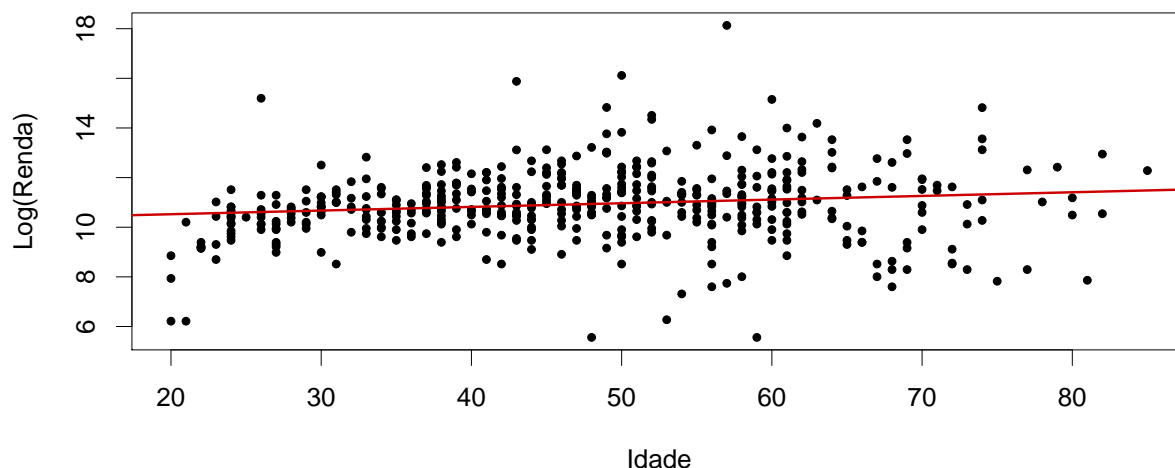
Avaliando os tipos de ensino percebemos uma maior concentração de entrevistados que possuem o ensino médio e o ensino superior, sendo que quase a metade dos entrevistados possuem ensino superior, esse valor corresponde a 248 entrevistados.

Analizamos também a relação entre a variável resposta (Renda) e as covariáveis presentes no banco de dados escolhido. Assim obtivemos os seguintes resultados:



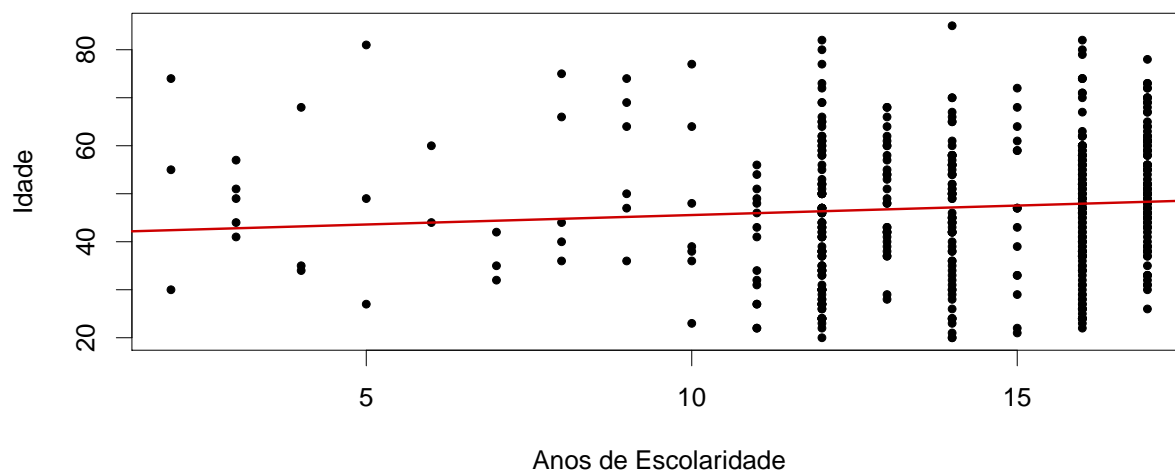
Para as variáveis discretas temos os boxplots da Log(Renda) com cada uma das variáveis separadamente. Para o gênero observamos um valor de renda maior para o sexo masculino, comparando com o sexo feminino; já a variável Estado Civil os entrevistados casados possuem uma renda maior, sendo que a mediana da renda dos que moram juntos com o parceiro(a) e outros estão bastante próximas, porém são inferiores aos valores de renda dos entrevistados casados. Na comparação entre a Log(Renda) e a Etnia percebemos uma amplitude da renda maior para as etnias Branco e Outros, entretanto, como foi observado anteriormente, essas etnias correspondem a 73% e 5%, respectivamente, do total do banco de dados.

Gráfico da Idade e a Log(Renda)



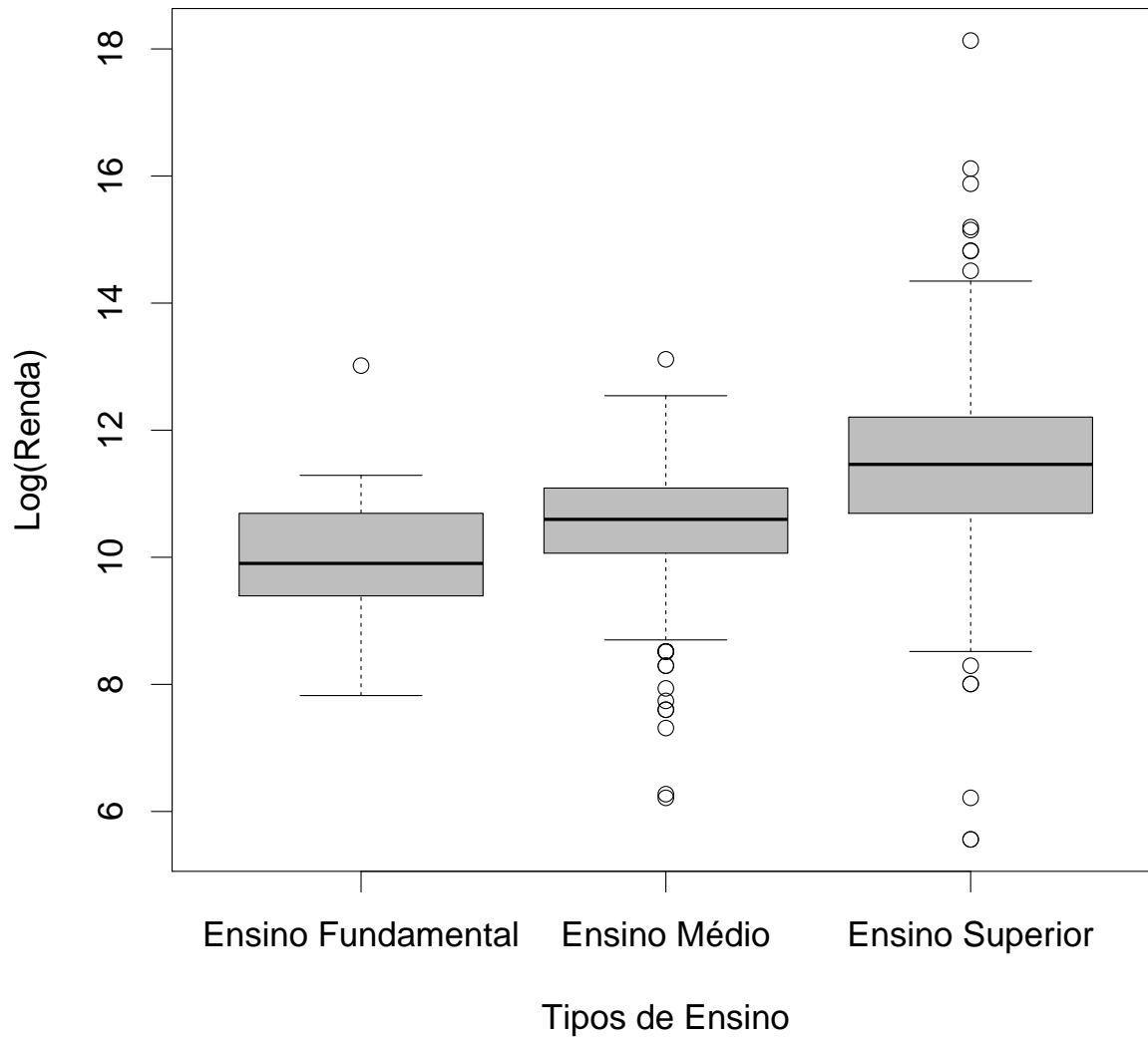
Para a relação entre as variáveis Log(Renda) e a Idade temos o gráfico de dispersão acima, nele percebemos uma pequena inclinação no ajuste da curva quando ocorre o aumento das idades dos entrevistados, o que indica um possível ganho de renda anual maior para os entrevistados a medida que aumenta a idade.

Gráfico dos Anos de Escolaridade e a Idade



Analisando as variáveis Anos de Escolaridade e a Idade, percebemos uma inclinação quando aumenta os anos de escolaridade e as idades dos entrevistados, ou seja, os entrevistados possuem maior anos de escolaridade à medida que aumentam as idade, o que condiz com a realidade dos ensinos visto anteriormente. Embora quantidades de anos de escolaridade maior possuem grande concentração de pessoas em diversas idades, para os anos de escolaridade entre 2-10 percebemos a presença de variabilidade.

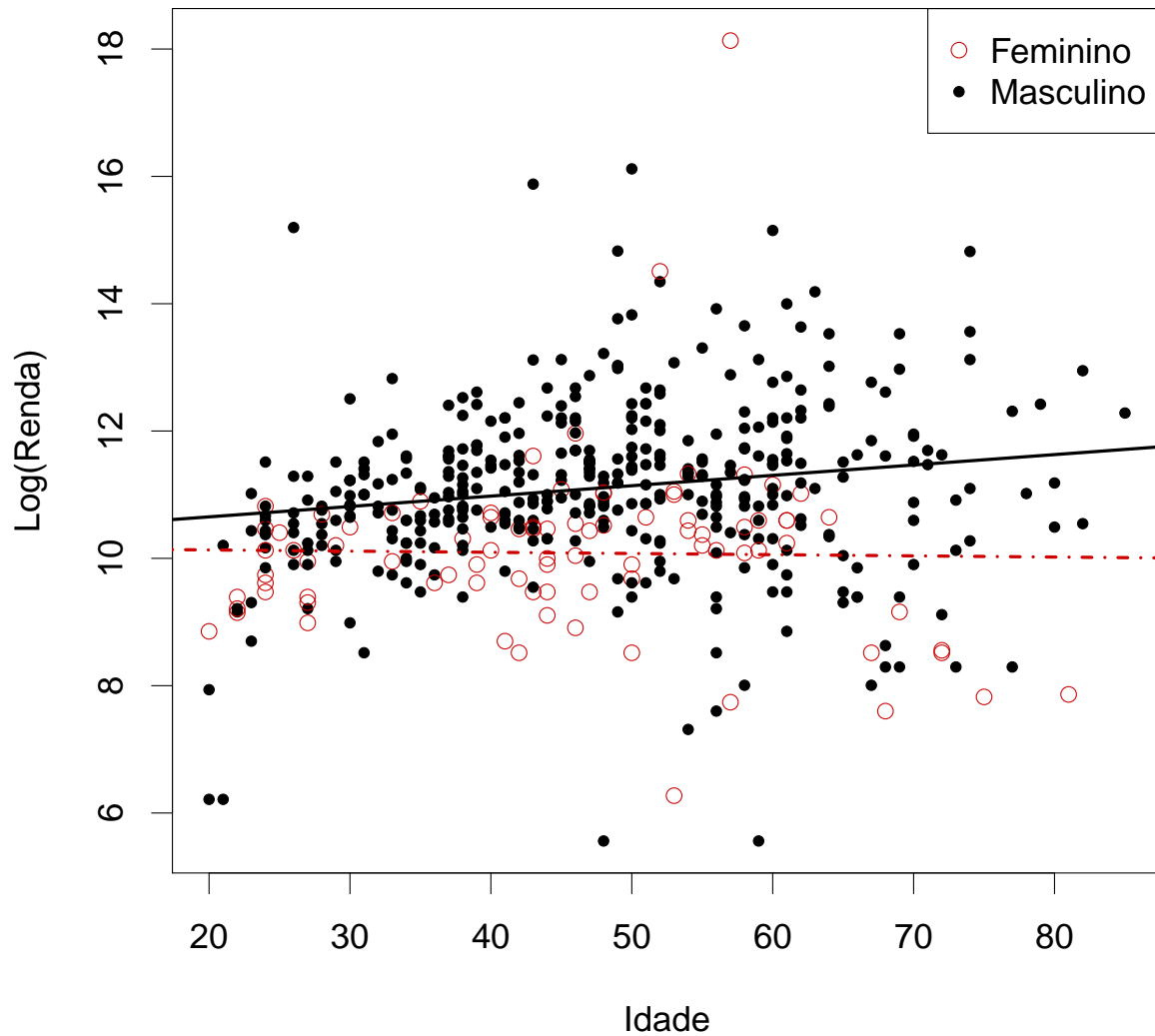
Boxplot dos Tipos de Ensino e a Log(Renda)



Avaliando a relação entre as variáveis Log(Renda) e a recodificação da variável Anos de Escolaridade, variável separada em tipos de ensino para melhor visualização da relação existente, temos que o tipo de ensino influencia na renda dos entrevistados. Assim observamos valores de renda maiores para o Ensino Superior, que possui de 15-17 anos de escolaridade.

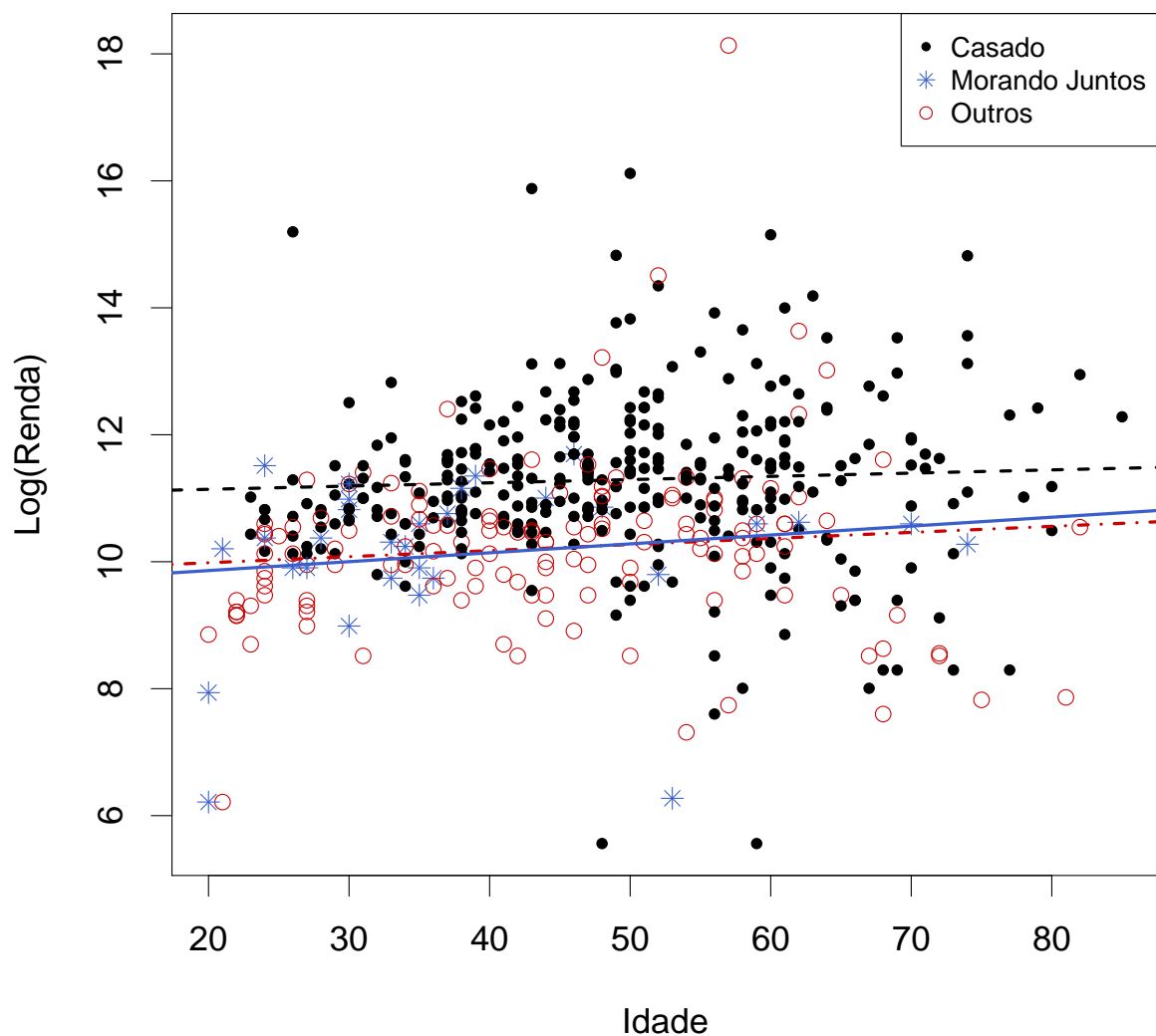
Após a apresentação das variáveis individualmente e em pares com a variável de interesse renda anual, realizamos a análise das variáveis em trios, como por exemplo a Log(Renda), Idade e o Gênero. As análises da relação dessas variáveis permitem fazer suposições sobre os modelos a serem estudados e verificar a influência de cada covariável na variável resposta.

Gráfico da Idade e a Log(Renda)

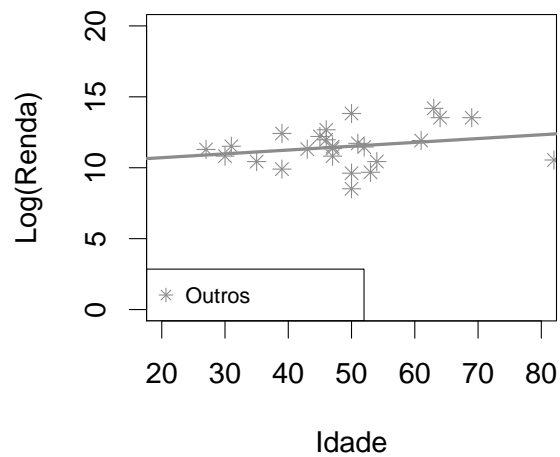
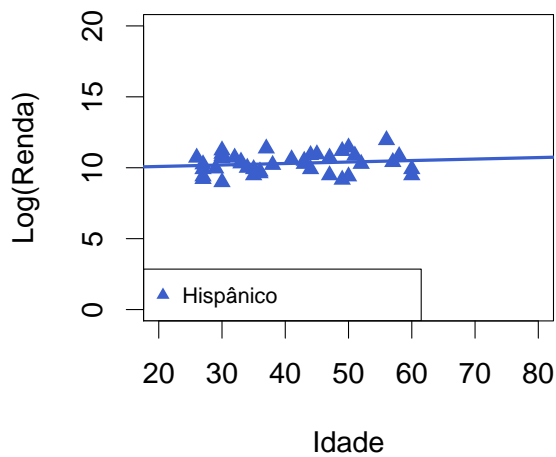
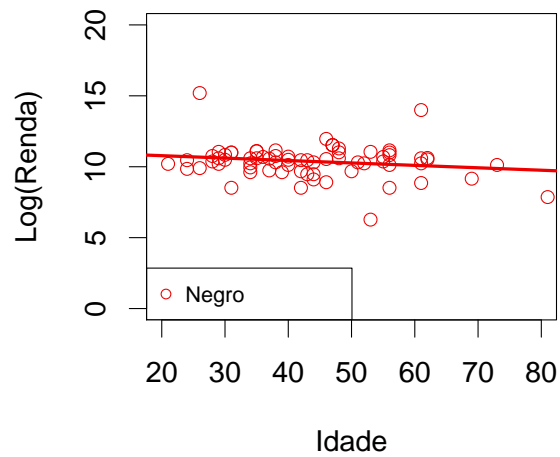
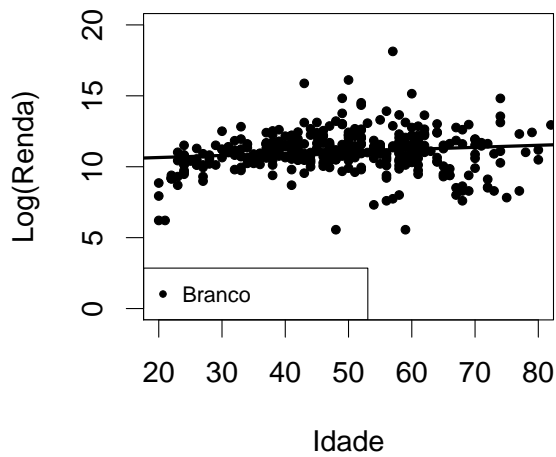


Analisando a relação entre as variáveis Log(Renda), Idade e Gênero percebemos o quanto a idade e o gênero influenciam nos valores da renda anual. O gráfico possui o ajuste das retas e mostra claramente o que foi discutido anteriormente, sobre os valores de renda anual para o sexo masculino serem maiores que o do sexo feminino; podemos perceber também uma inclinação na reta, à medida que aumenta a idade, para o sexo masculino, entretanto para as mulheres essa inclinação é muito pequena ou inexistente.

Gráfico da Idade e a Log(Renda)

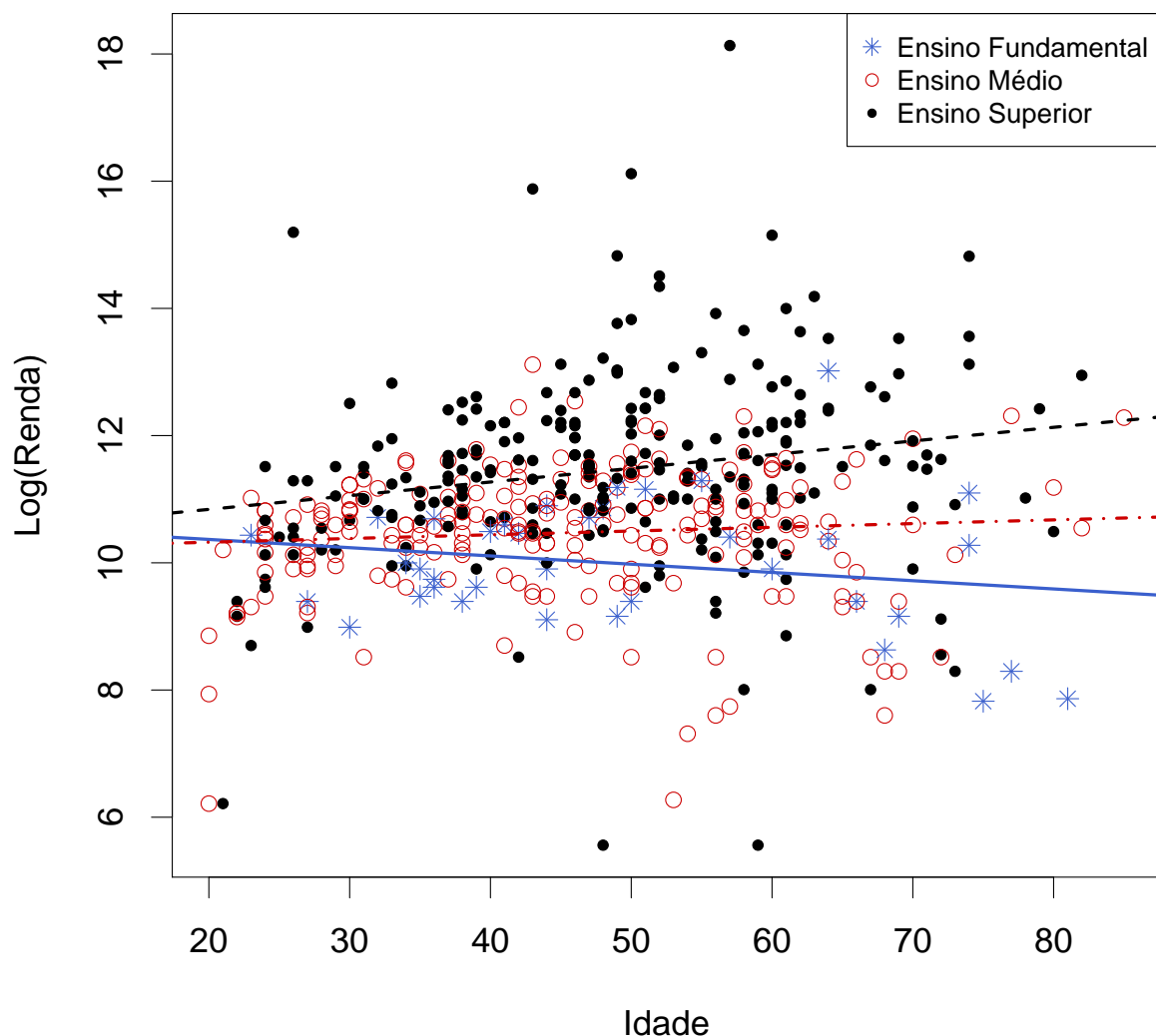


Para a análise entre as variáveis Log(Renda), Idade e Estado Civil, podemos observar que os entrevistados casados possuem uma quantidade maior de renda e através da inclinação da reta podemos concluir que a renda aumenta através da idade, já para os entrevistados que estão morando junto e os outros tipos de estado civil as retas e a inclinação estão quase juntas, sendo que possuem pequenas diferenças em algumas idades; o gráfico acima também mostra como está a distribuição por estado civil dos entrevistados.



Analisando a relação entre as variáveis Log(Renda) , Idade e a Etnia , vemos que as retas pertencentes as etnias Branco, Negro e Outros partem quase do mesmo valor de renda, entretanto, ao longo das idades, possuem comportamentos diferentes para a inclinação da reta, sendo que para a etnia Negro a renda decresce a medida que aumenta a idade. Observamos maior renda para a etnia Outros, os Hispânicos, que começam a reta abaixo das outras etnias, intercepta a etnia Negro entre as idades 40-50 anos. Pelo gráfico da distribuição percebemos que a etnia Hispânico estão concentrados em torno do Log(Renda) igual a 10, e que a etnia Negro possui valores de renda bastante dispersos a medida que aumenta a idade.

Gráfico da Idade e a Log(Renda)



Para a análise entre as variáveis Log(Renda) , Idade e os tipos de Ensino, confirmamos a conclusão anterior sobre o Ensino Superior possuir renda maior que os outros tipos de ensino. Sendo que nas idades mais jovens a renda para o Ensino Fundamental e o Ensino Médio estão muito próximas quando analisamos a inclinação da reta, entretanto a partir dos 30 anos as retas desses dois tipos de ensino começam a se distanciar; assim há um aumento de renda para o Ensino Médio enquanto o Ensino Fundamental apresenta declínio.

Imputação

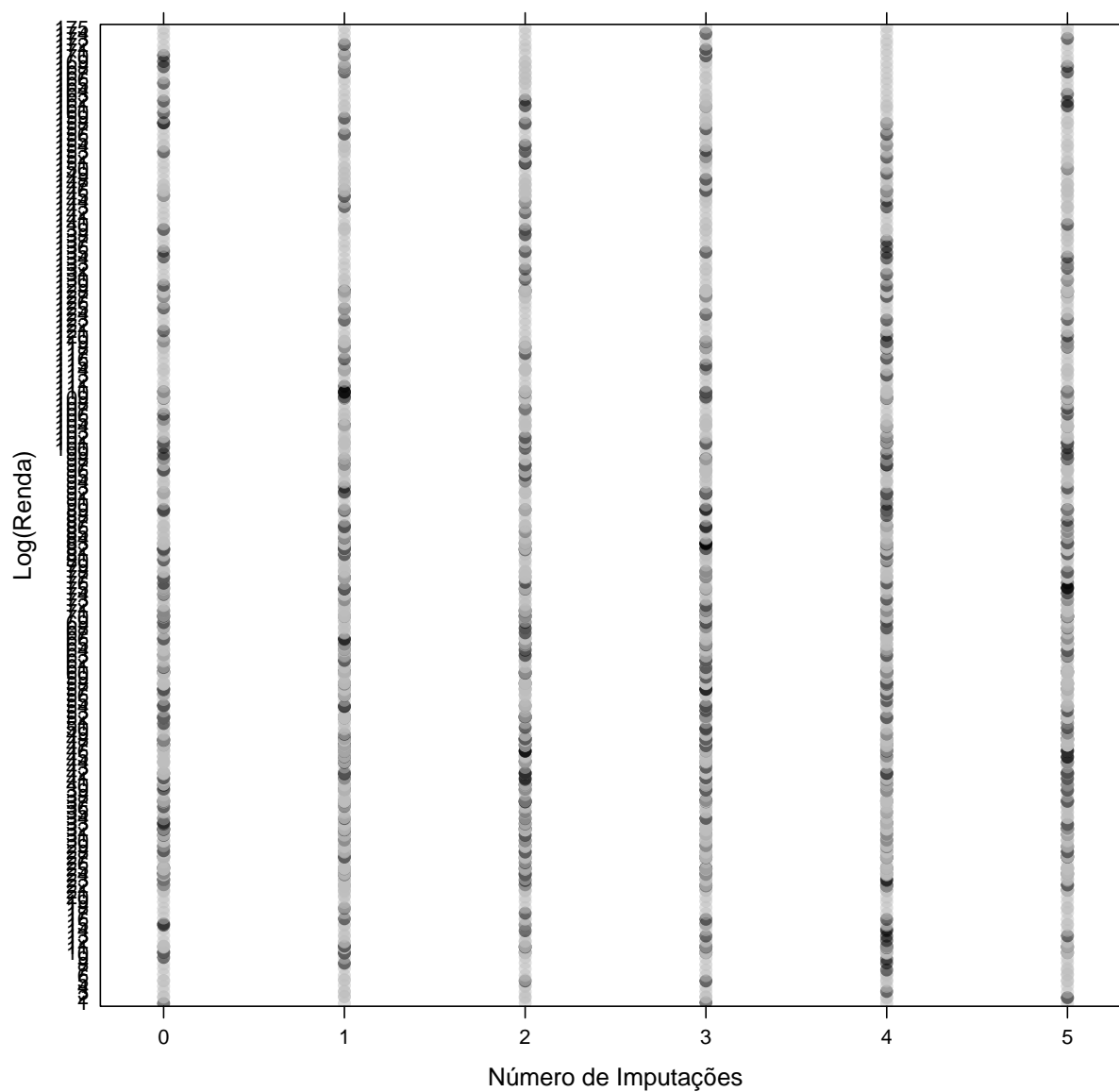
O banco de dados não possui dados faltantes, portanto para avaliar a Renda (variável de interesse) foi necessário gerar os dados faltantes. Sendo assim utilizamos uma distribuição binomial com probabilidade de sucesso de 0.2 para a criar dos dados faltantes na variável Renda, e utilizamos a função de fixa a semente ao gerar os números aleatórios.

Primeiras observações do banco de dados com dados faltantes na variável Renda:

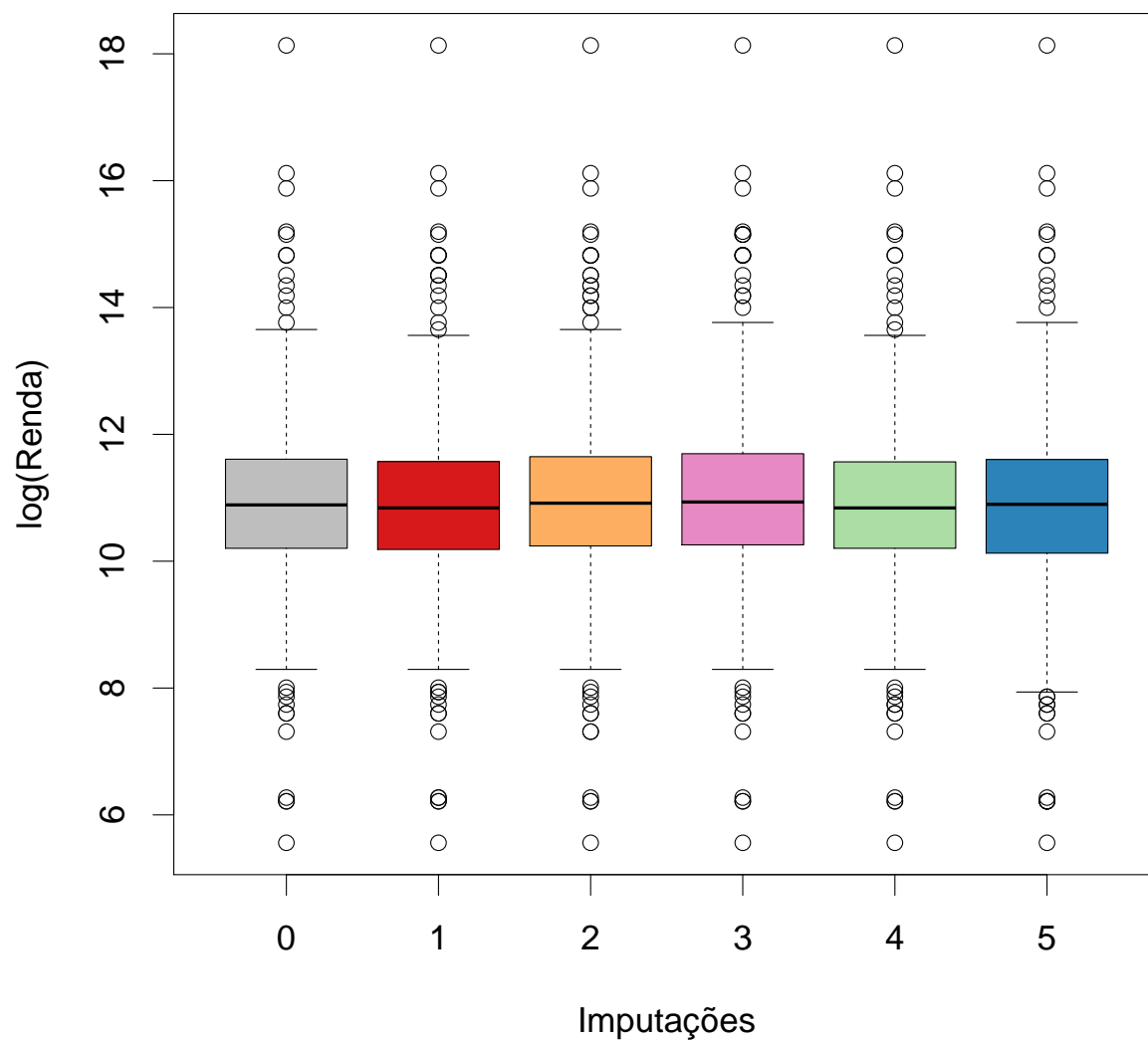
Assim para realizar a imputação utilizamos o pacote *Multivariate Imputation With Chained Equations (MICE)*. A função que realiza a imputação chama-se *mice*, e nesse estudo realizamos a imputação 5 vezes ($m=5$) tanto para o método da *PMM* e da *Mean* da função para comparar os resultados.

Missing Complete at Random

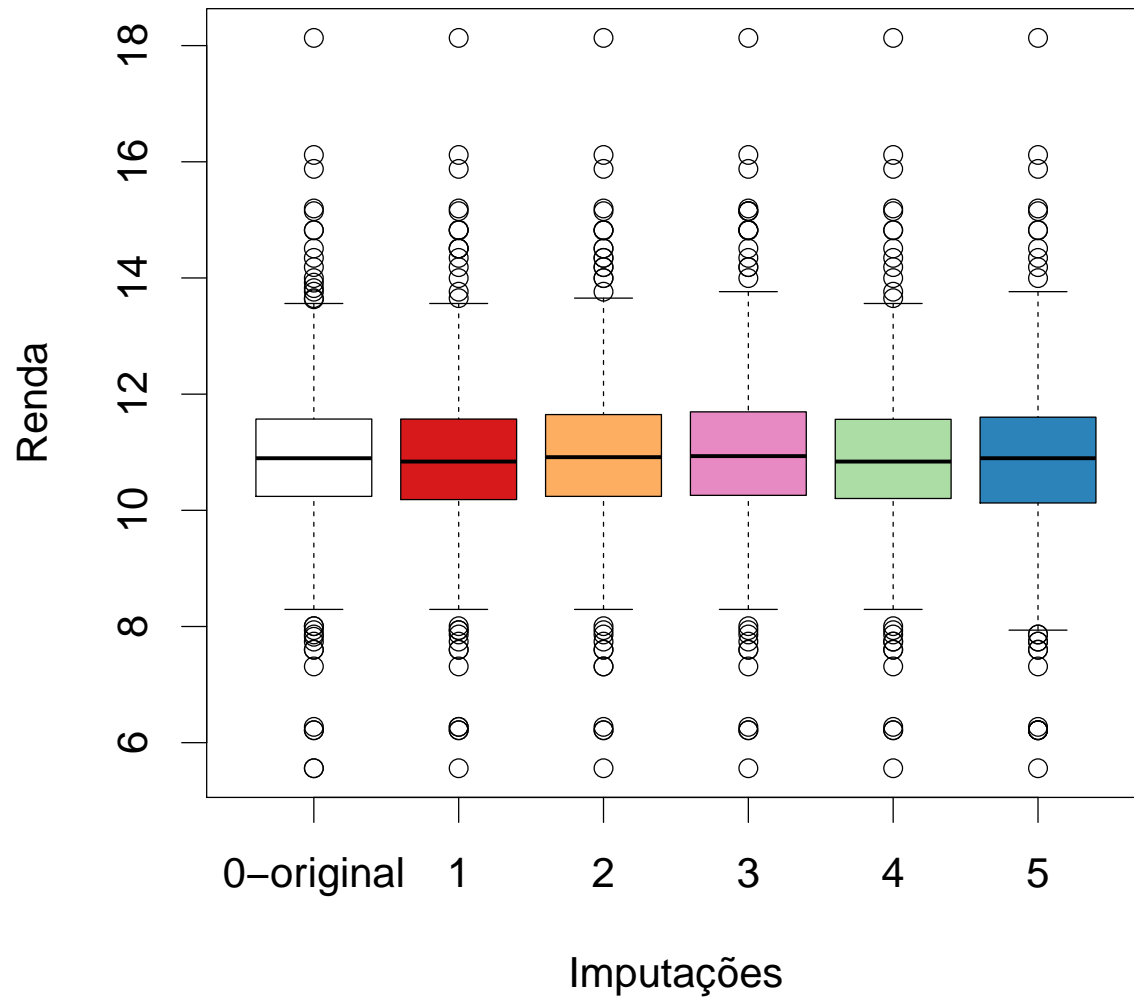
Gráfico com as distribuições dos valores imputados



Box-plots da variável Renda com os dados ausentes e as imputações



Box-plots dos dados originais e das imputações



QQ-plot das imputações

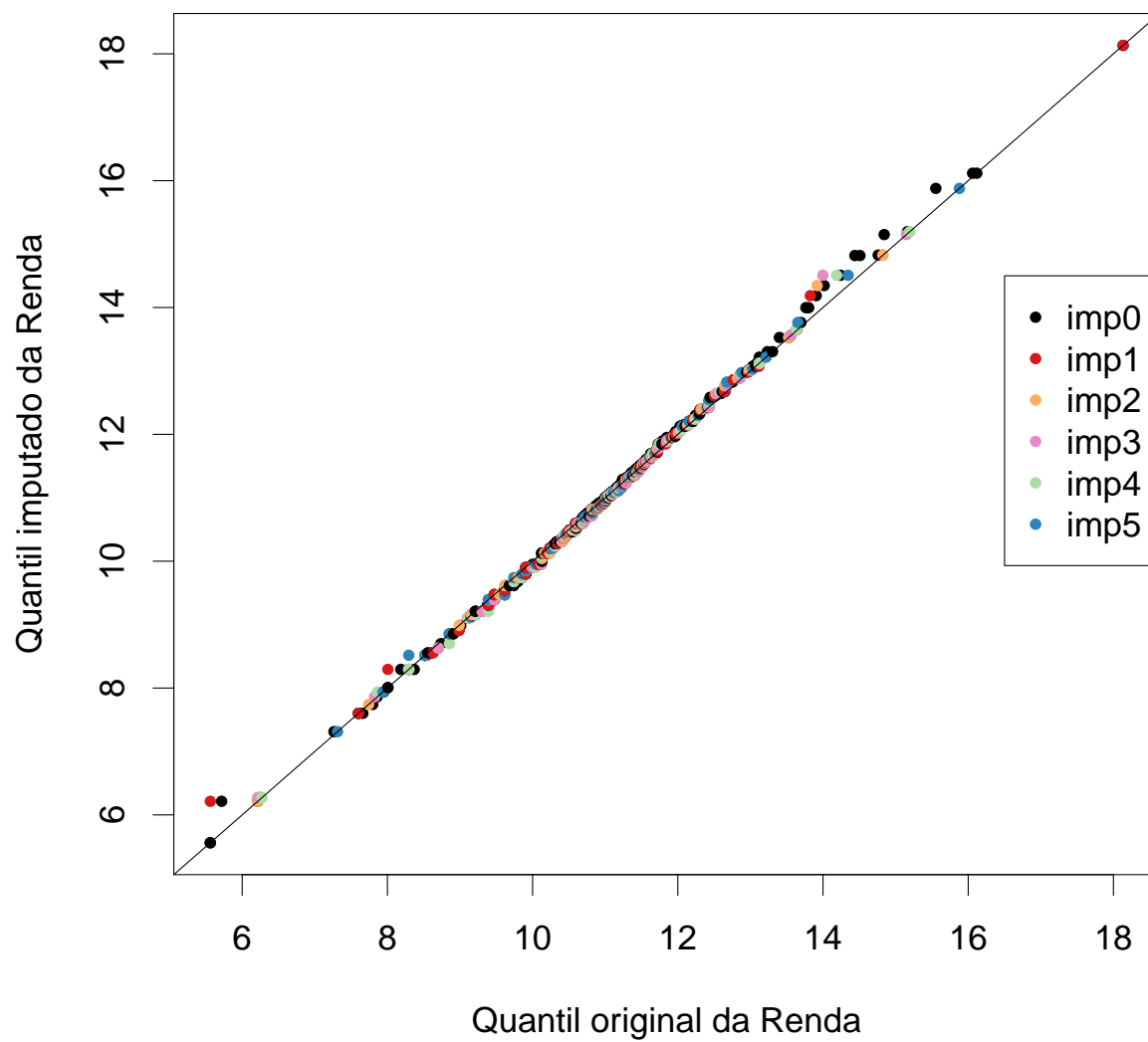
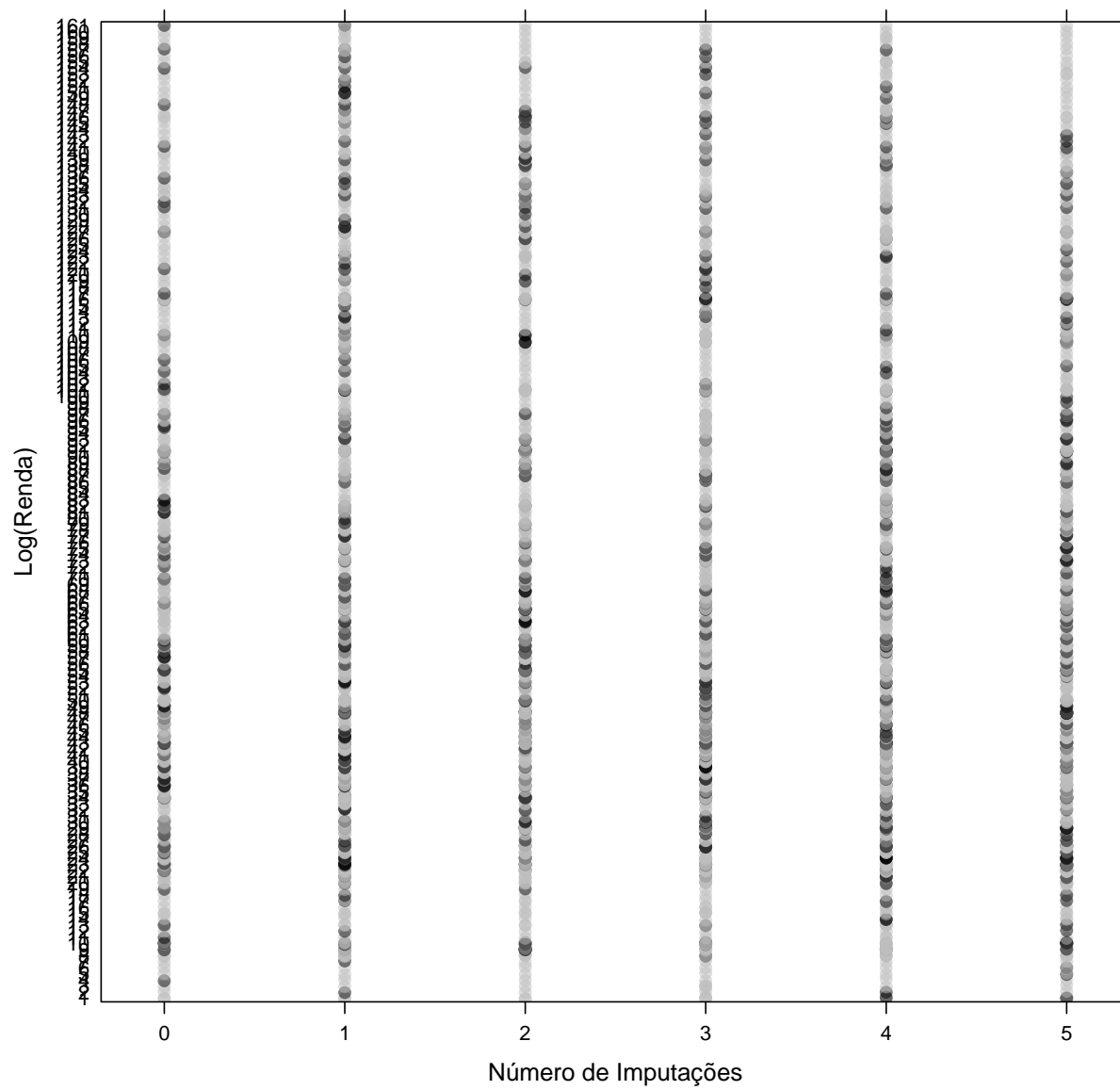
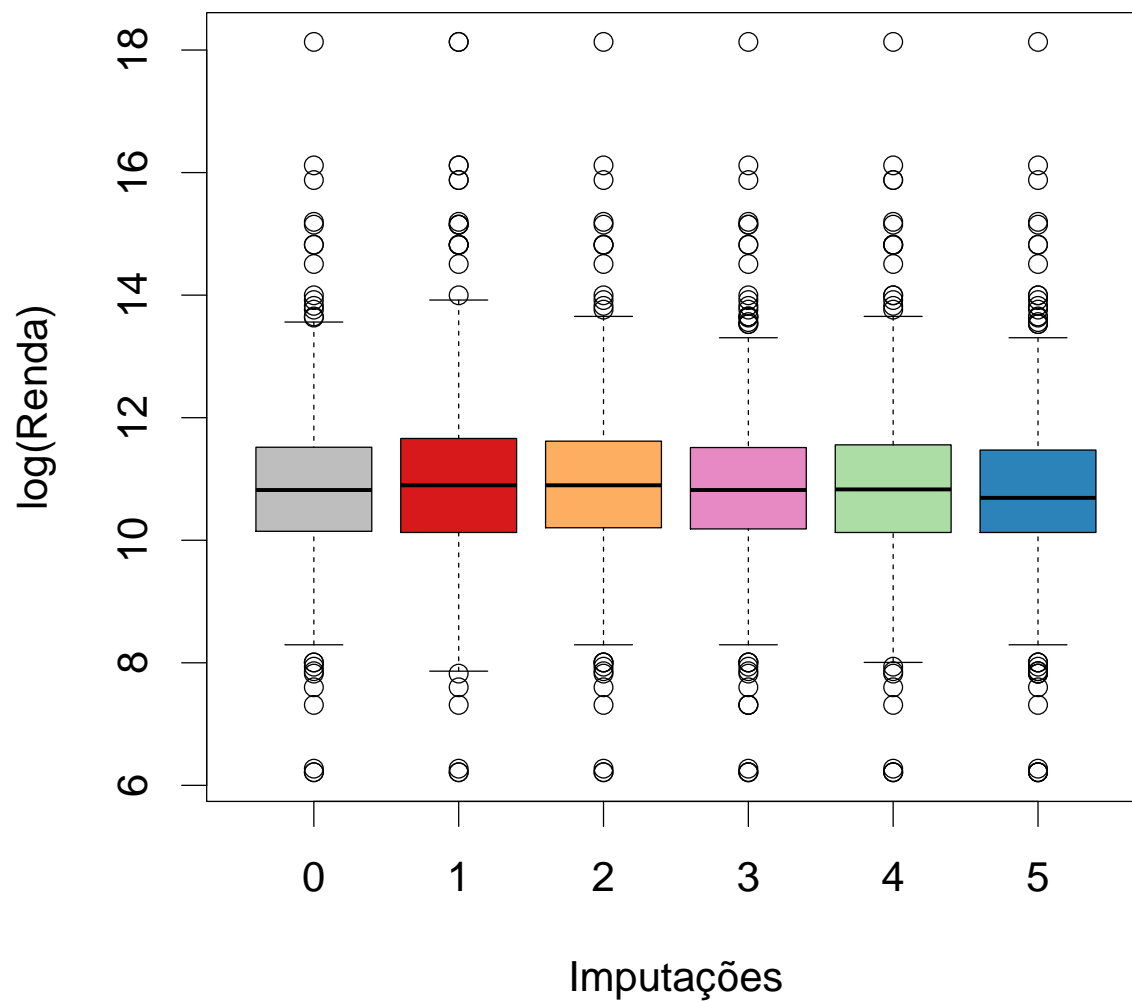


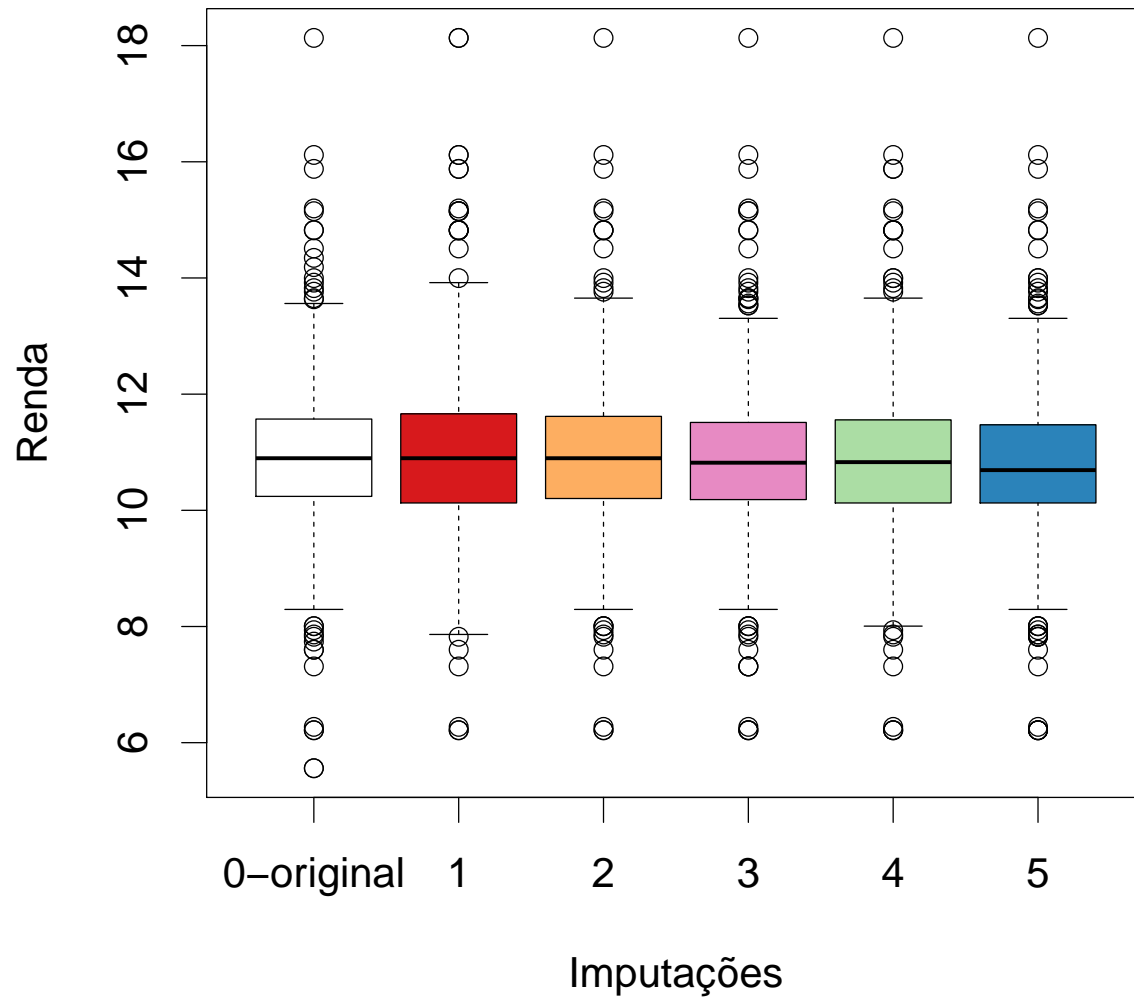
Gráfico com as distribuições dos valores imputados



Box-plots da variável Renda com os dados ausentes e as imputações



Box-plots dos dados originais e das imputações



QQ-plot das imputações

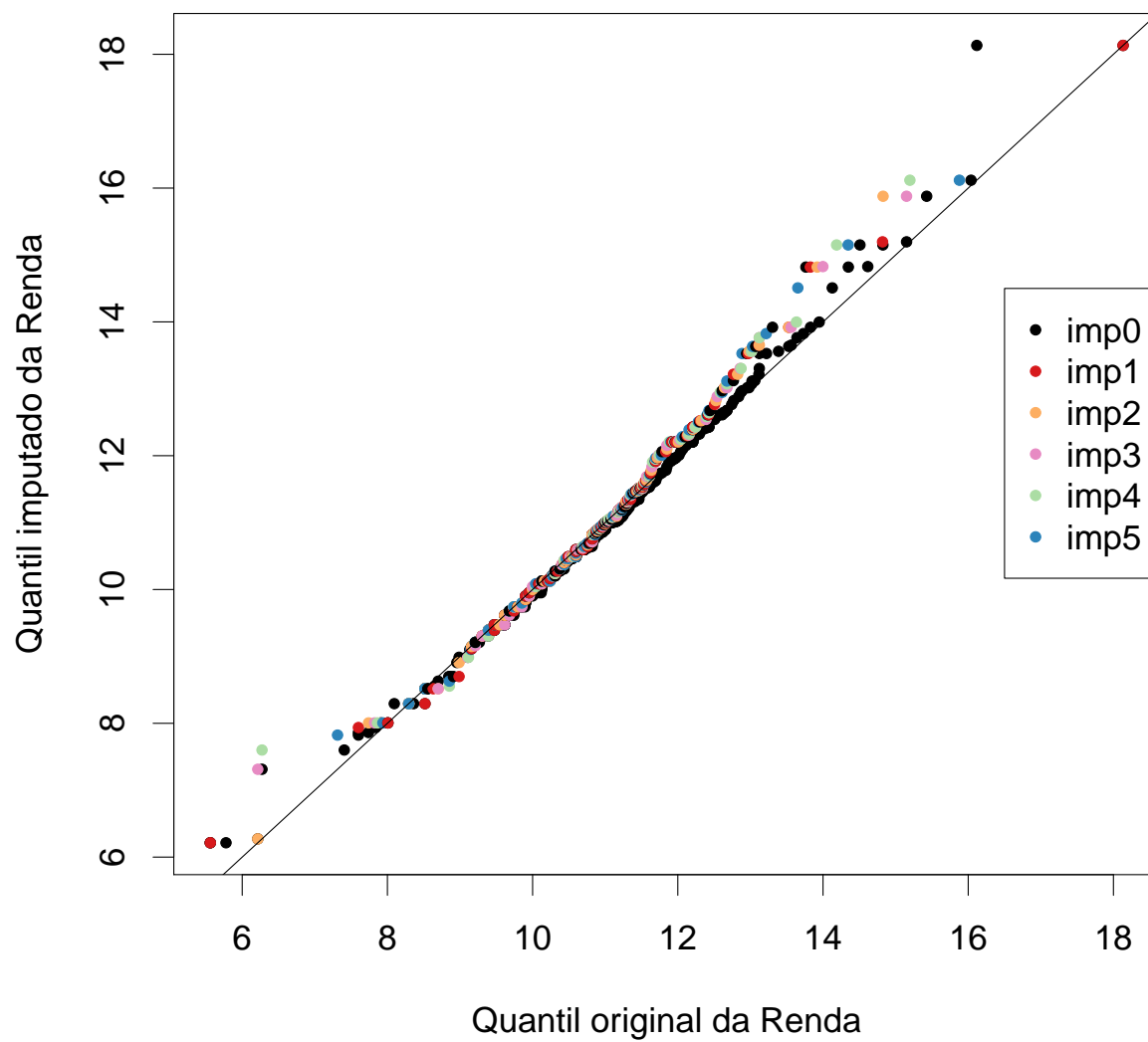
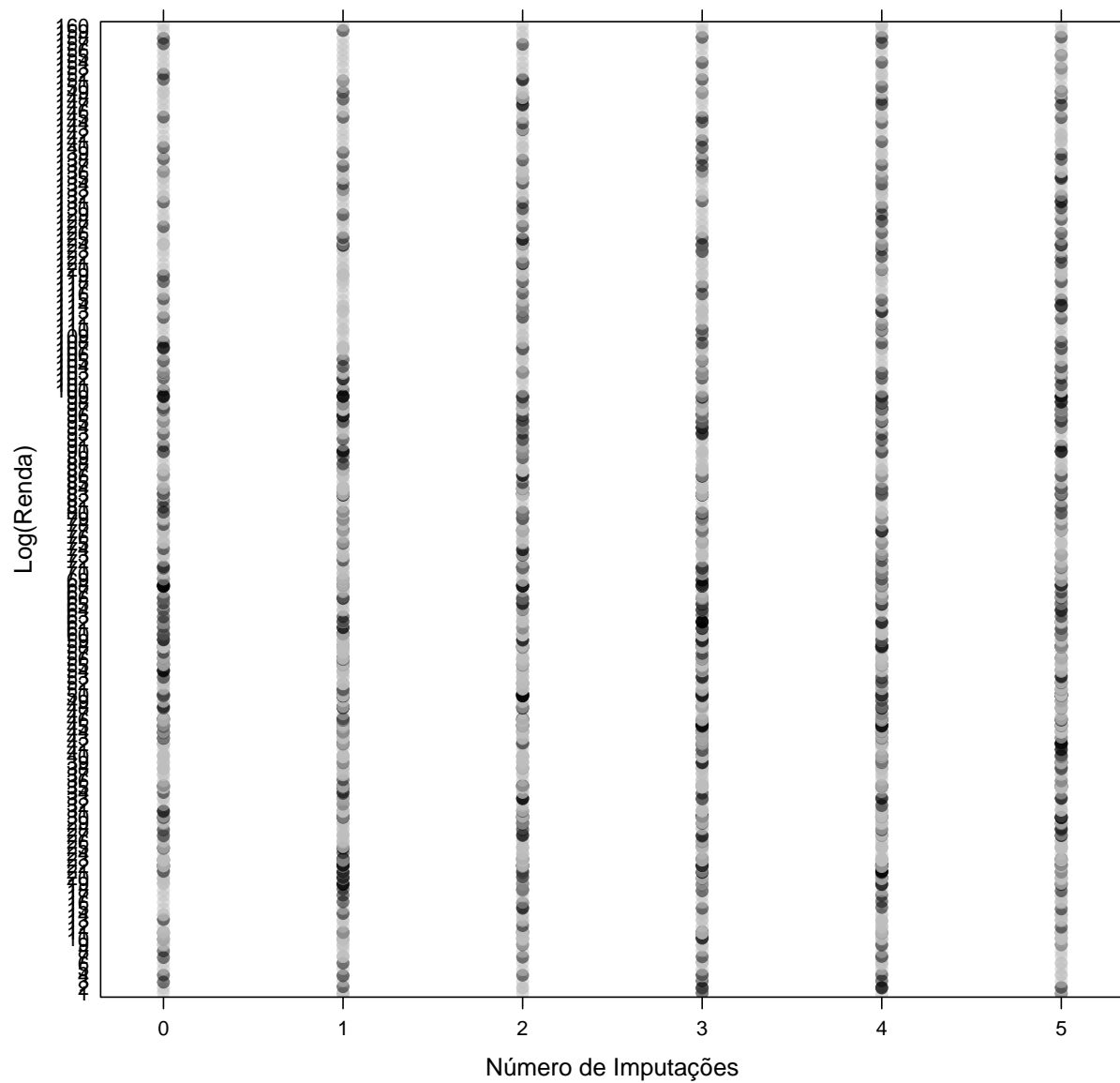
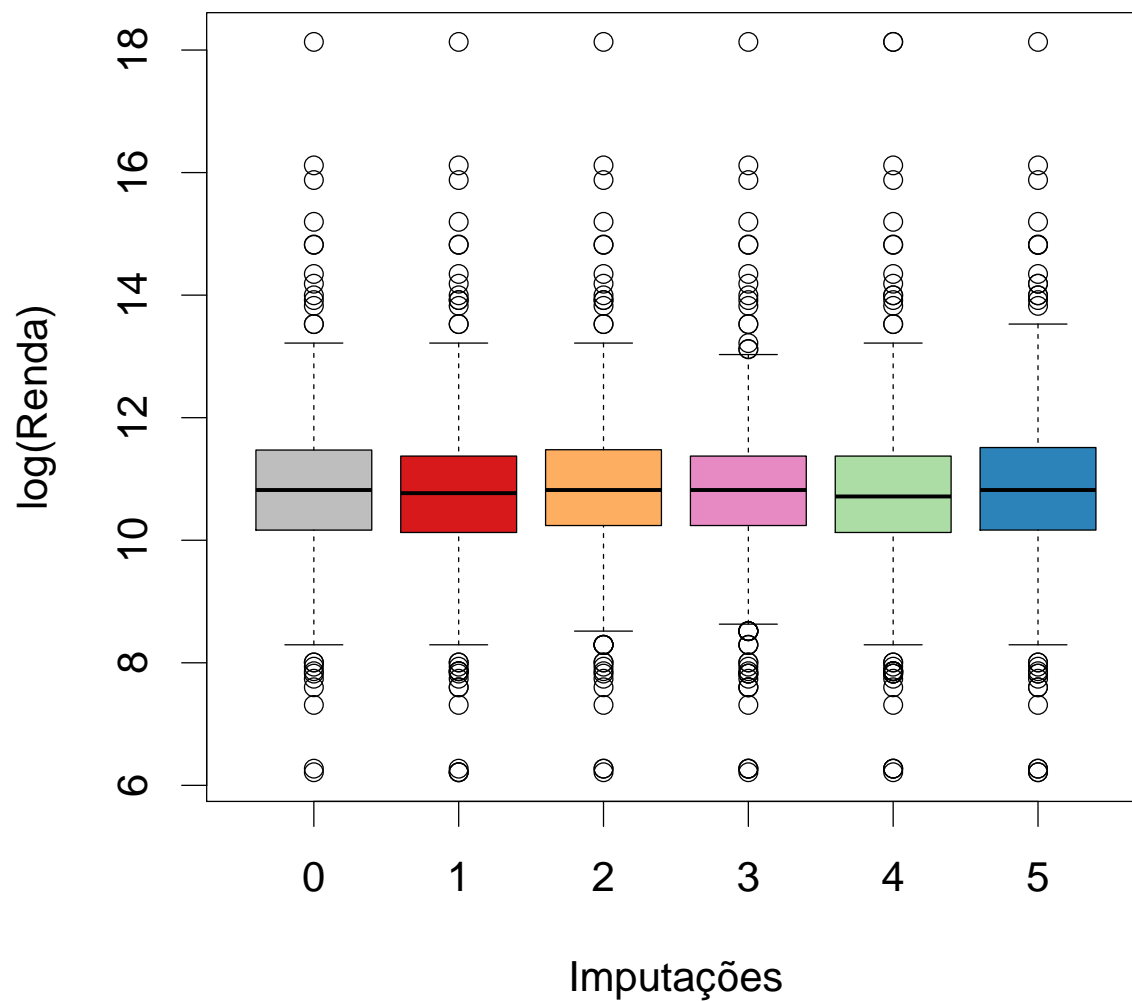


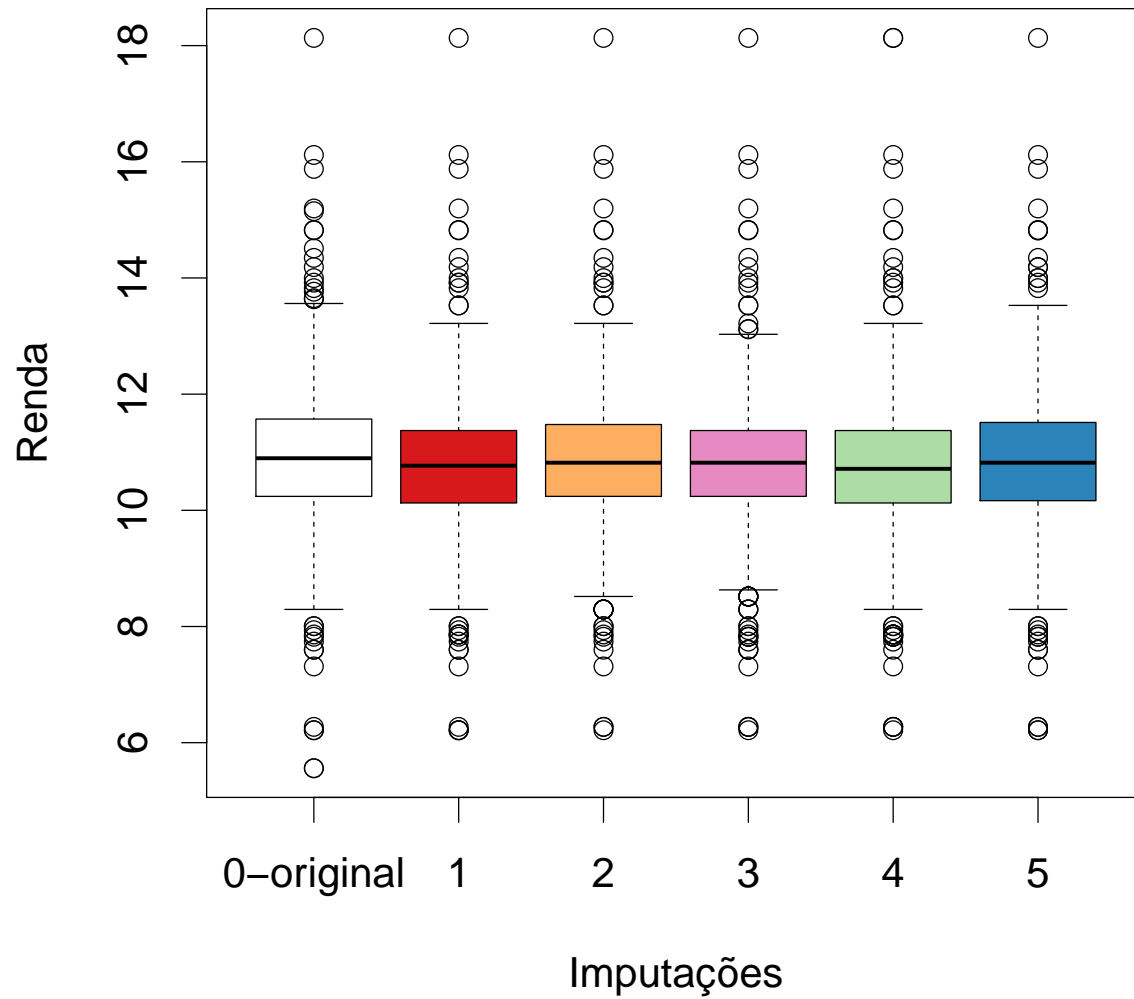
Gráfico com as distribuições dos valores imputados



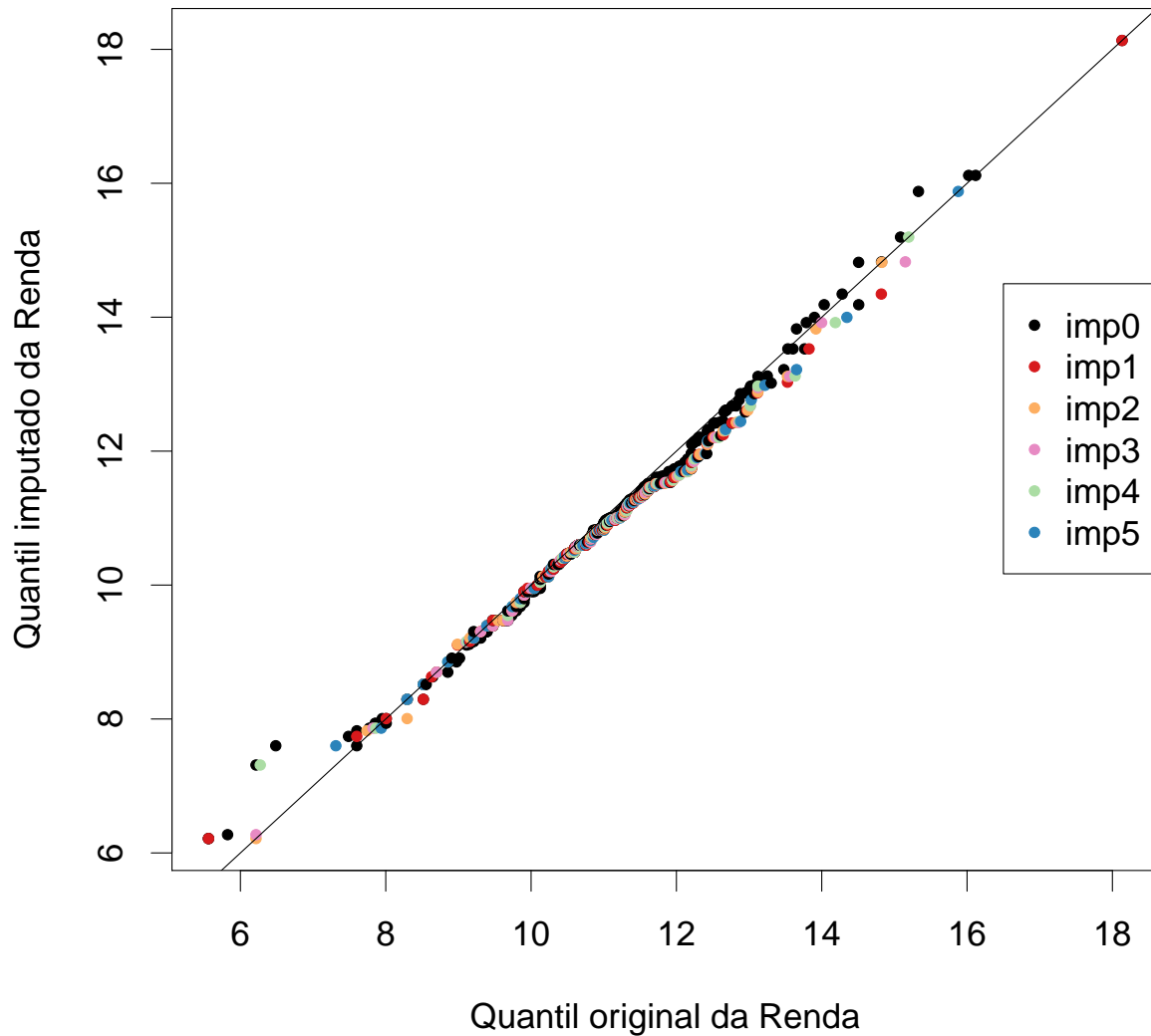
Box-plots da variável Renda com os dados ausentes e as imputações



Box-plots dos dados originais e das imputações



QQ-plot das imputações



CONCLUSÃO

REFERÊNCIAS BIBLIOGRÁFICAS

Rubin (1987)

Van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. [linked phrase](#)

Morris TP, White IR, Royston P (2015). Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med Res Methodol. ;14:75.

Frees, E.W. (2011). Regression Modeling with Actuarial and Financial Applications, Cambridge University Press.