

Previsão da renda média familiar da região nordeste.



Analise de dados e datamining



Nome do Aluno:
Emanuel da Silva Santos

Coordenadores:
Profª Drª Alessandra de Ávila Montini
Profª Dr. Adolpho Walter Pimazoni Canton

Agenda

1. Objetivo do Trabalho
2. Contextualização do Problema
3. Base de Dados
4. Exploratória da base
5. Estatística tradicional
6. Métodos de inteligência artificial
7. Métricas
8. Análise complementar
9. Conclusões e sugestões para o futuro
-
-
-

1. Objetivo do Trabalho

O objetivo deste trabalho é prever o **valor médio familiar** das famílias cadastradas no sistema Cad.Unico na região **nordeste** no período de **2018**.

Nesta predição, foi utilizado os dados da família cadastra e das pessoas pertencentes a essas famílias onde os dados foram compilados em uma única tabela contendo características financeiras, moradia, abastecimento e educação, podendo assim, ser utilizados em modelos de **regressão linear múltipla** e outras técnicas de **inteligência artificial**.

Os resultados obtidos poderão dar insumo em decisões aos entes federativos (estados e municípios) daquela região, bem como entidades que tenham interesses e desejem um modelo que preveja o valor médio familiar para triagem em programas locais.



2. Contextualização do Problema

O **Cad.Único** é um sistema federal que busca concentrar os dados das famílias que estão relacionadas em algum programa social; além disso, também é utilizado para mapear grupos de população carente ou que fazem parte de algum grupo étnico ou ligado a natureza ou aqueles que tiveram seus empreendimentos afetados por obras governamentais.

Os dados contidos nesse sistema podem embasar estudos técnicos dos estados e municípios, e proporcionar novos programas para a região, seja para uma pequena parte como um bairro ou município, ou até mesmo o estado ou a região nordeste como um todo. Para novos programas sociais, há a possibilidade do órgão responsável optar por priorizar as pessoas com **rendas inferiores**, no entanto, nem sempre as regiões que farão esse controle terão acesso ao Cad.Unico, dessa forma o script de **predição** da renda média família pode apoiar nesse cenário, baseando-se nos dados cadastrais focado tanto na rendas, como nas características de moradia, abastecimento e ensino.

Para o cenário do ente privado (**CNPJ**), que também oferece programas sociais e desejam utilizar algum algoritmo de predição, estes também podem optar por usar os resultados desse estudo em suas triagens.

Nesses cenários, os códigos e os resultados finais, poderão ser utilizados tanto pelo poder público que necessite e não tenha estrutura de acesso ao **cad.unico**, como pelo pessoa jurídica, formando assim sua política de seleção, seja por **priorizar** por valor previsto, como utilizar um valor de **corte** em cima dós resultados previstos.

4. Exploratória



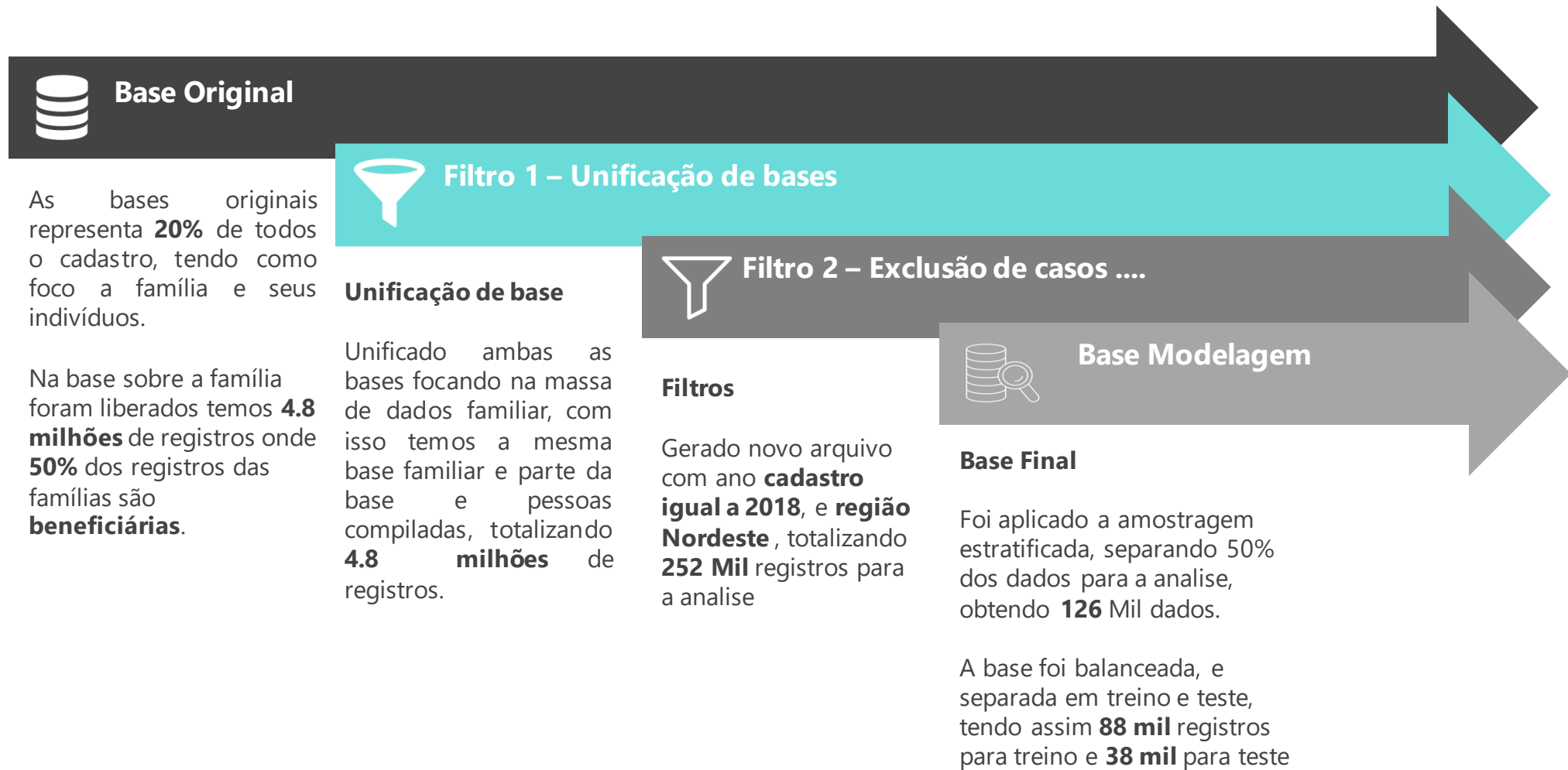
3. Base de dados original

CAD ÚNICO

- ✓ Os dados foram obtidos no site do ministério do desenvolvimento social (MDS - <https://aplicacoes.mds.gov.br/sagi/portal/index.php?grupo=212>)
- ✓ Foi obtido um total de dois arquivos
- ✓ O tópico para download é o **Microdados Dez/2018 - Cadastro Único e PBF**
- ✓ Foi necessário carrega-los em banco de dados Microsoft SQL Server para posterior tratamento da massa de dados.
- ✓ O dicionário de dados original consta no site acima o que ajudou na analise



3.ii. Filtros



3.iii. Variáveis (ABT)



Variáveis cadastrais

- Ano cadastro
 - Identificação família
 - Sexo do responsável
 - Município
 - Mesorregião
 - Local de domicílio
 - Espécie domicílio
 - Material Piso
 - Material Construção
 - Água encanada
 - Abastecimento de água
 - Tem banheiro?
 - Escoamento Sanitário
 - Destino Lixo
 - Iluminação
 - Calçamento
 - Família
 - Classe da cidade
- Tem Analfabeto
 - Ensino superior?
 - Menor estuda?
 - Cor
 - Ensino Família



Variáveis Quantitativas

- Nº Comodo
- Nº Dormitórios
- Quantidade de pessoas
- Quantidade de homem
- Quantidade de mulheres
- Quantidade de menores
- Média de idade



Variáveis financeiras

- **Renda média familiar**
- Renda de trabalho registrado
- Renda nos últimos 12 meses
- Renda aposentadoria
- Rendas diversas
- Período máximo de trabalho
- Valores recebido pelo menor



Benefício

- Participa do bolsa família



4.Exploratória – Sobre a Família



Variáveis associadas a família:

Sexo Responsável

Cor

Media de Idade

Estado de residência

MESORREGIAO

Família

- As famílias são representadas por mulheres em **66%** dos registros
- A cor / raça declarada na família é bem representado pela parda, com **67%** do total
- Na visão geral dos registros da base, a média de idade familiar fica em torno de **36 anos**, o que indica uma população adulta; há de se destacar as famílias que são outliers com idade de **100 a 116 anos**, o que se considerou normal.
- Os três estados com mais inscritos foram Bahia com **24%**, Pernambuco com **18%** e Ceará com **13%**
- A mesorregião foram agrupadas seguindo a semelhança para quantidade de famílias, neste agrupamento a mesorregião três possui o maior grupo familiar com **22%**.



A variável município não foi analisada devido a alta granularidade



4. Exploratória – Moradia



Variáveis associada a moradia

Local de domicilio

Material do Piso






Material da construção

Calçamento

Classe da cidade

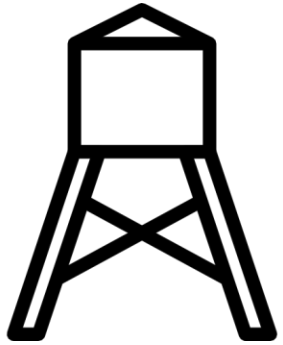
Quantidade de cômodos

MORADIA

-  76% da população reside na região urbana
-  53% das residências tem o calçamento total e 40% não possuem calçamento, as demais residências não possuem calçamento
-  A maioria das famílias não residem nas capitais e regiões metropolitanas (60%), as famílias que residem em capitais e regiões são representadas por 40% das famílias.
-  76% das residências são construídas com alvenaria e / ou tijolo com revestimento;
-  Já com relação ao piso das residências, podemos observar que as duas categorias predominantes são cimento, e cerâmica e/ou lajota e/ou pedra, cada uma com 44%



4. Exploratória – Saneamento e relacionados



Variáveis associada a saneamento e relacionados

Abastecimento (água)

Escoamento sanitário

Destino lixo

SANEAMENTO

70% das moradias possuem água encanada, outra forma de possui aguar na residência é por poço, cisterna ou nascente (16%), seguido pelos demais meios.

Com relação ao escoamento sanitário, as formas estão consideravelmente distribuída sendo: a rede coletora (38%) seguida por fossa rudimentar (29%) e fossa séptica (16%), neste ponto, os órgãos públicos devem dar atenção, até mesmo por questões de saúde.

O lixo nessas residências e coletado diretamente em 72% das residências, mas também deve ser observado os outros meios que podem não ser benefício à saúde e meio ambiente



4. Exploratória – referente aos estudos



Variáveis associada ao Ensino

Ensino do menor de idade

Ensino da família

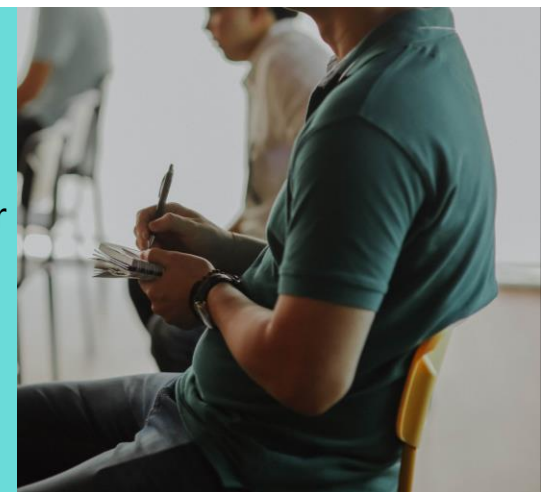
Existe analfabeto

ENSINO

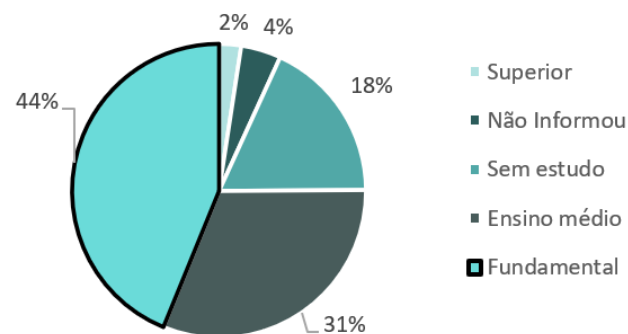
68% das famílias não tem menores em seus membros, já 14% das famílias tem seus menores estudando seguidos por 10% com os menores não estudando ou sem idade de estar na escola. Apenas 6% das famílias os menores com idade escolar não estudam.

22% das residências possui uma pessoa analfabeta.

É possível observar que boa parte da população tem como maior nível de escolaridade o ensino fundamental e médio (juntos possuem 75% das famílias), para o ensino superior podemos observar apenas 2% da base



Ensino Família



4. Exploratória – referente a renda






Variáveis associada Renda

Outras rendas

Renda Proveniente

Período máximo de
trabalho registrado

Rendas

-  80% das famílias não receberam valores de rendas diversas
-  Grande parte das famílias (até 75%) receberam até R\$ 250,00 como renda proveniente de trabalho registrado, por outro lado quando observamos o período, esta mais faixa de famílias não tem um mês completo, neste cenário devemos observar as regiões, e se houve condições de trabalhos à essas famílias ou se passam curtos períodos em Trabalhos temporários.
-  Por outro lado uma parcela da população completaram o ciclo de todo ano de 2018 trabalhando, e alguns e alguns salários bem superiores à media (poucas famílias)



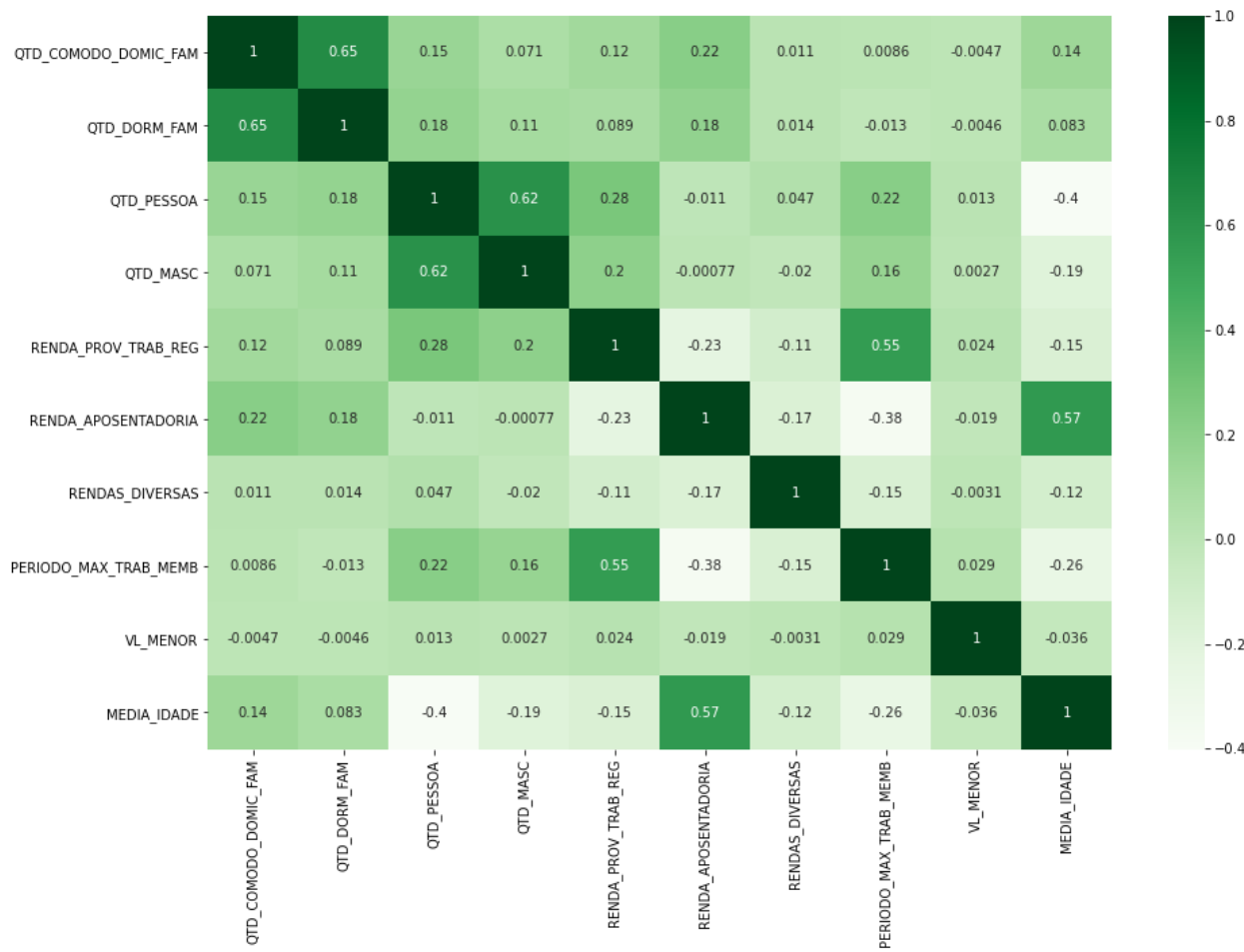


Sobre a modelagem

1. Após tratativas, os **126** Mil registros foram separados em outras duas sub-bases, utilizados para treinar o modelo e posteriormente testa-los. Abaixo podemos observar a distribuição de ambas as bases:

Treino: **88** Mil

Teste: **38** Mil



Correlação:

Buscamos aqui neste gráfico de heatmap, localizar as variáveis quantitativa que explicam o comportamento de outra variável quantitativa, pois procuramos diversidade e não ambiguidade.



Dicionário de variáveis

5. Estatística Tradicional

Modelos:

Regressão Linear múltipla

Com esse algoritmo, será avaliado as variáveis onde, as categóricas passaram por um processo para transforma cada uma em um dado / coluna numérico (0 ou 1, sim ou não, é ou não é aquela categoria), com isso o modelo de Regressão Linear Múltipla não tem problemas em utilizar esse tipo de variável.

Os dados numéricos por padrão foram tratados alguns que tinham outliers muito discrepante, para que estes dados não influenciasse nos cálculos, prejudicando os resultados finais.

A partir disso, o modelo considerou importante utilizar as variáveis a seguir, atribuindo os valores que constam planilhados para melhor compreensão.



5. Regressão Linear Múltipla



Variável	Valor/Obs	Coeficiente	Interpretação
(Intercept)		366.69	
Sexo do responsável	Home	-17.26	Quando homem o valor reduz
Estado	Bahia	-9.11	Quando um dos estados o valor reduz
Estado	Pernambuco	-17.74	conforme lista
Mesorregião	MES3	-8.11	
Mesorregião	MES4	-11.24	Quando um dos grupos da mesorregião, o
Mesorregião	MES5	-8.59	valor previsto reduz conforme lista.
Mesorregião	MES6	-22.15	
Mesorregião	MES7	-54.05	
Local Domicilio	Urbana	-7.17	Se local de domicilio urbana, então o valor médio reduz



5. Regressão Linear Múltipla

Variável	Valor / Obs	Coefficiente Interpretação
Material Piso	Cimento	-11.87 Há queda se a familiar residir em domicílio com
	Outro Material	-28.42 essas características
Material Construção Abastecimento	Categoria A	-9.21 Se está categoria, o valor reduz
	Rede geral	10.96 Se rede geral, a valor aumenta
Escoamento sanitário	Fossa séptica	-5.77 Se essas categorias o valor irá reduzir
	Rede coletora ou pluvial	-6.68
Destino do lixo	Queimado ou enterrado no local	10.43 Se essas categorias o valor aumenta
Classe da cidade	Outros	48.01
	Região metropolitana	31.71 Se metropolitana o valor aumenta
Analfabeto	Sim	8.24 Se analfabeto o valor aumenta
	Todos estudam	43.40
Menor estuda	Todos não estudam	38.23 O valor irá aumentar conforme a categoria
	Sem idade ou algum não estuda	110.15
Bolsa Familia	Tem	-398.54 Se possui bolsa família a renda média diminui
Cor	Mais de uma	-18.49 Se essas categorias o valor diminui
	Parda	-10.60
	Médio	16.10
Ensino	Não informou	42.40 A renda média aumenta conforme as categorias de
	Nenhum	24.78 ensino da família
	Superior	107.68



5. Regressão Linear Múltipla



Variável	Valor/ Obs	Coeficiente	Interpretação
Rendas diversas	sim	-120.823	Se há recebimento de rendas diversas o valor diminui
Quantidade de cômodos		8.8846	A cada cômodo ha o crescimento
Quantidade de pessoas		-44.2804	A cada pessoa a renda diminui este valor
Quantidade de homens		-17.8318	A cada homem a renda diminui este valor
Renda proveniente do trabalho registrado		0.1599	A cada real ganho, o valor é acrescido na renda média
Período máximo de trabalho		-11.1717	A cada mês trabalhado o valor diminui
Média de idade		6.3368	A cada ano médio familiar, o valor é acrescido



Modelos:

- ❖ Árvore de Decisão Linear
- ❖ Random Forest Regressor
- ❖ Gradient Boosting Regressor
- ❖ RN – Redes Neurais

Para a modelagem com os **três primeiros modelos**, faremos uso de técnicas de seleção de variáveis, que busca achar as mais importantes para posterior uso da Árvore de regressão Linear, Random Forest Regressor e Gradient Boosting Regressor.

Com **Redes Neurais**, utilizamos as funções e otimizadores para simulação de neurônios capazes de realizar previsões aproximadas.

Todos os modelos são técnicas **avançadas**, e, em muitos casos demonstrando uma eficácia maior que os modelos tradicionais.

Nota.: A árvore de decisão linear pode ser considerada um modelo tradicional, no entanto, nesta apresentação apresentaremos como um modelo avançado devido a como ela foi construída, utilizando codificação avançada.

6. Técnicas de inteligencia

Árvore de decisão linear

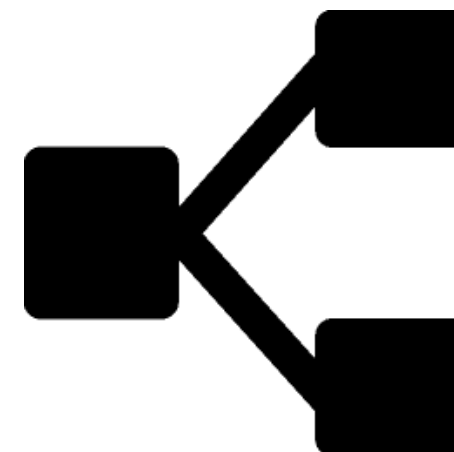
- ✓ Este modelo foi construído com as seguintes características
 - ✓ Critério MAE
 - ✓ Profundidade de 10 ramos

Random Forest Regressor

- ✓ O modelo foi construído com as seguintes características
 - ✓ Estimação de até cem árvores randômicas, com critério do tipo MAE (Mean Absolute Error) e cada árvore com até 5 ramos de profundidade

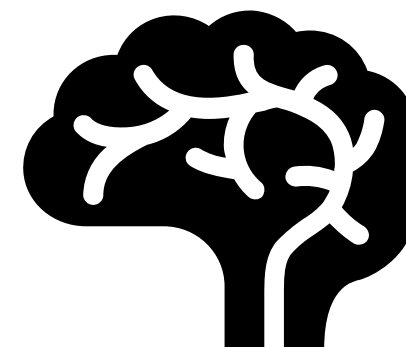
Gradient Boosting Regressor

- ✓ O modelo foi construído com as seguintes características
 - ✓ Estimação de até cem árvores, com critério do tipo MAE (Mean Absolute Error) e cada árvore com até 10 ramos de profundidade

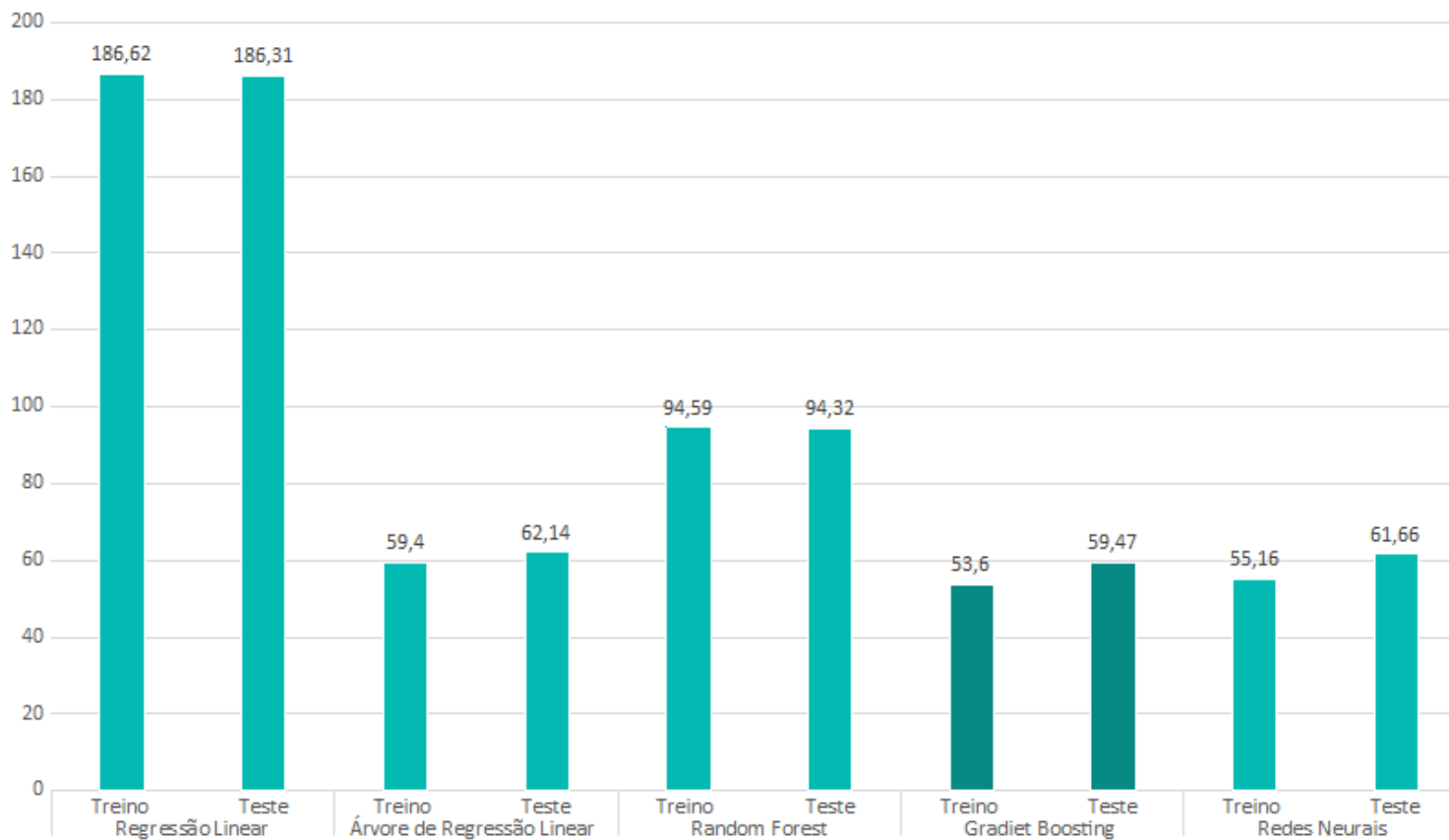


RN – Redes Neurais

- ✓ Este modelo busca a partir de neurónios artificiais, achar os valores para cada parâmetro da base que somados formam a previsão aproximada ou exata a depender do cenário.
- ✓ Foi construído com 7 camadas de neuronios onde:
 - ✓ A primeira camada possuía 55 neurônios, uma para cada valor que foi passado
 - ✓ 5 camadas com 100 neurônios cada
 - ✓ A última camada com 1 neurônio de saída
- ✓ a função de ativação foi utilizada a Relu – também foi testado a Linear em conjunto com a Relu e sozinha porem não houve melhoria na métrica.
- ✓ Otimizador Adam - não foram feito testes com outros otimizadores.
- ✓ E métrica MAE (mean absoluty error)
- ✓ Também foi utilizado um parâmetro de stop, onde, após 13 interações onde houvesse a ausência de queda no loss, o modelo encerrava seu treinamento. Esse parâmetro foi utilizado para se evitar o overfitting



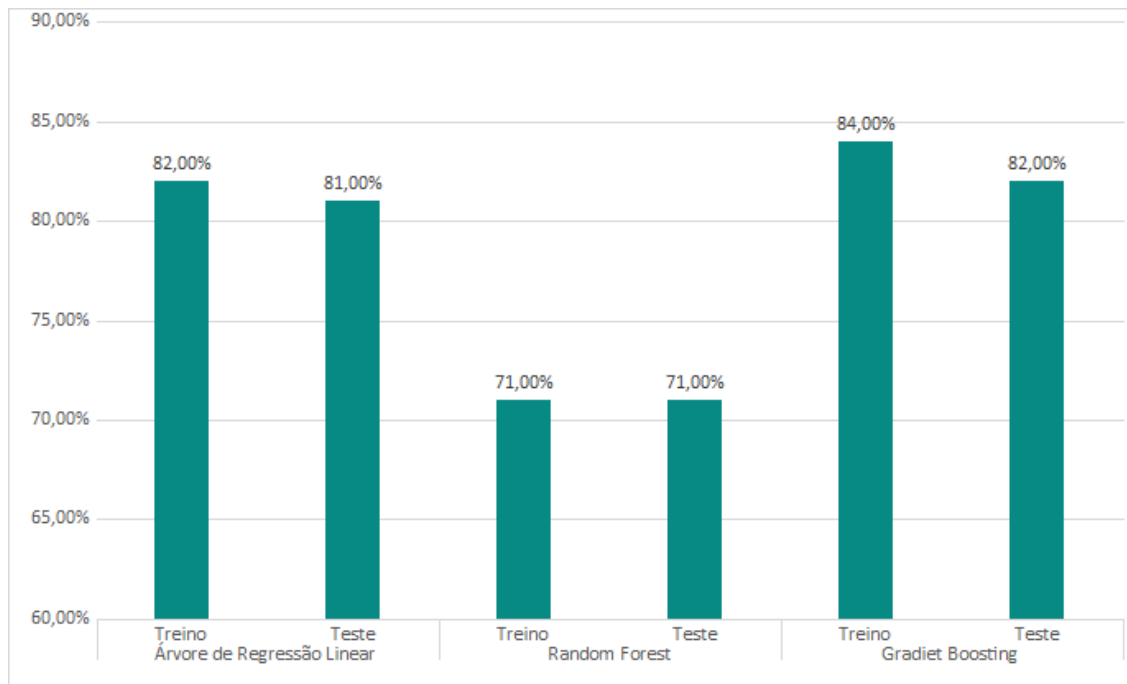
7. Métricas – MAE



Com os resultados dos modelos, podemos observar que o Gradiente Boosting possui o melhor MAE entre treino e teste, a regressão linear múltipla ficou com um MAE relativamente alto comparado aos demais.



7. Métricas - R^2 ajustado e acurácia



Regressão Linear Múltipla :
 R^2 ajustado: **0,61**

* Não foi possível calcular o MAPE, uma vez que o valor real da base continha dados com zero.

Analizando as métricas fornecidas pelos scripts, novamente o Gradiente Boosting mostrou bons resultados, podendo ser a técnica selecionada para implementação.



7. Métricas - Considerações



Todos os modelos empregados, estão relacionados diretamente ao objetivo desta apresentação. Era esperado um valor aproximado entre a árvore de regressão e a regressão linear múltipla, no entanto, podemos observar que a árvore teve um performance muito superior ao modelo de regressão linear – de **186** contra **59** de MAE em base treino ,e **186** contra **62** em teste. Parte disso pode ser explicada em como os modelos foram codificados, em ambos os casos houve a seleção de variáveis, no entanto para a árvore foi empregado técnicas superiores onde é avaliado as variáveis que serão melhores no modelos, já na regressão, o critério foi o tradicional, seleção baseada no P-Valor

Randon Forest e o Gradient Boosting, são evoluções de modelo de árvore. É possível observar que os MAE estão muito **próximo** entre todos os **modelos de árvore**, e por um intervalo muito pequeno, temos um modelo vencedor o **Gradient Boosting** (treino: 53 , teste: 59) .

A **Rede Neural** também teve o resultado muito satisfatório, no entanto, mesmo com a ideia de simulação de neurônios e o modo como opera, não teve uma performance melhor que o **Gradiente Boosting**.

É importante mencionar que para treino do modelo de uma rede neural, o processo foi muito mais rápido que o modelo de árvore Gradient Boosting e Randon Forest, uma possível mescla entre ambos poderia (talvez sim, talvez não) gerar métricas melhores, porem essa hipótese não foi explorada.



7. Métricas - R^2 ajustado e acurácia



Análise complementar



Conclusões e sugestões para o futuro



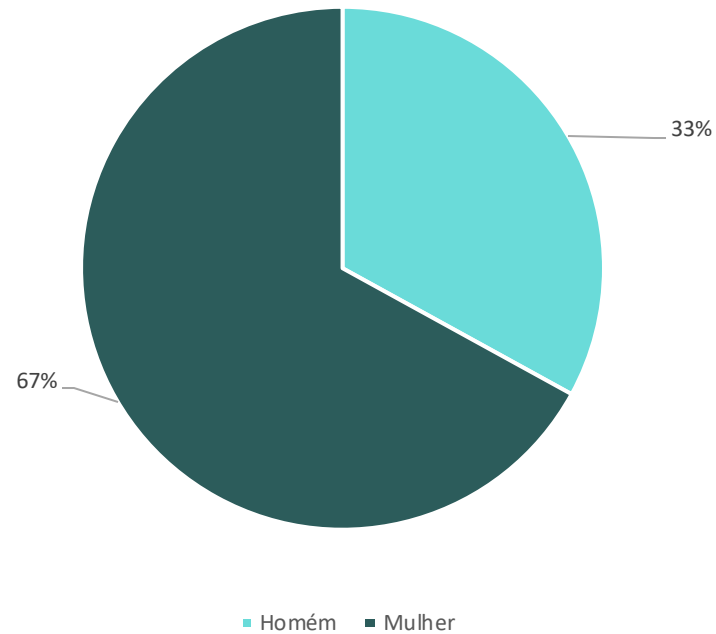
3.iii. Principais variáveis



Analises complementares



Responsável

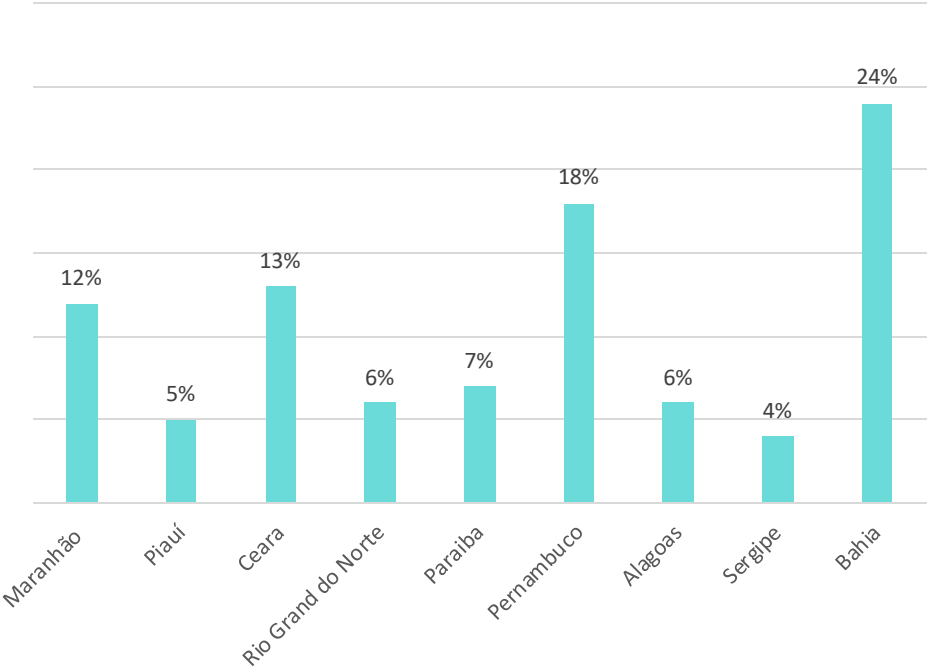


Categoria	Qtd	%
Homém	41890	33%
Mulher	84364	67%

É possível observar que as mulheres majoritariamente representa a família, com, seu percentual sobre o total cadastrado é d 67%



Estados



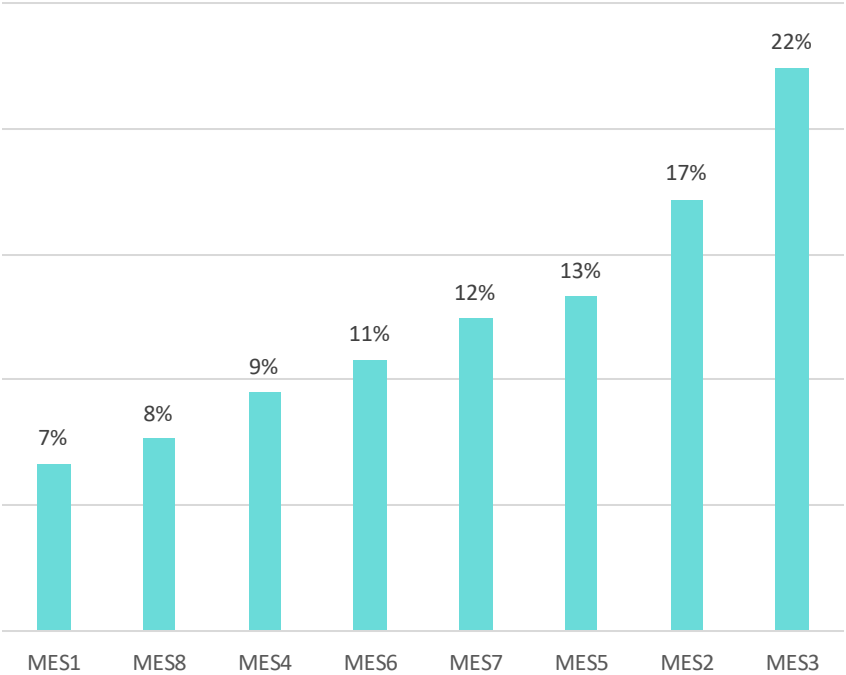
Categoria	Qtd	%
Maranhão	16409	13%
Piauí	7041	5%
Ceara	17448	14%
Rio Grand do Norte	8024	6%
Paraíba	9179	7%
Pernambuco	23181	18%
Alagoas	7796	6%
Sergipe	6112	5%
Bahia	31064	25%

O estado da Bahia é o de maior representatividade com 25% da população cadastrada em 2018



Mesorregião

Mesorregião Agrupado

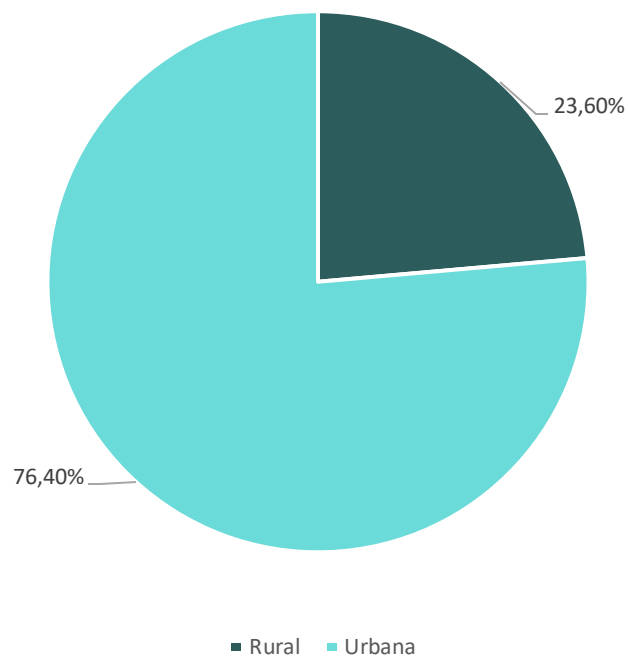


Categoria	Total	%
MES1	8413	7%
MES8	9660	8%
MES4	11985	9%
MES6	13670	11%
MES7	15688	12%
MES5	16804	13%
MES2	21707	17%
MES3	28327	22%

Devido a alta granularidade das mesorregiões; as mesorregiões foram agrupadas conforme as características populacionais semelhantes. As mesorregiões agrupadas podem estar em diferentes estados.



Residencia – Local de domicílio

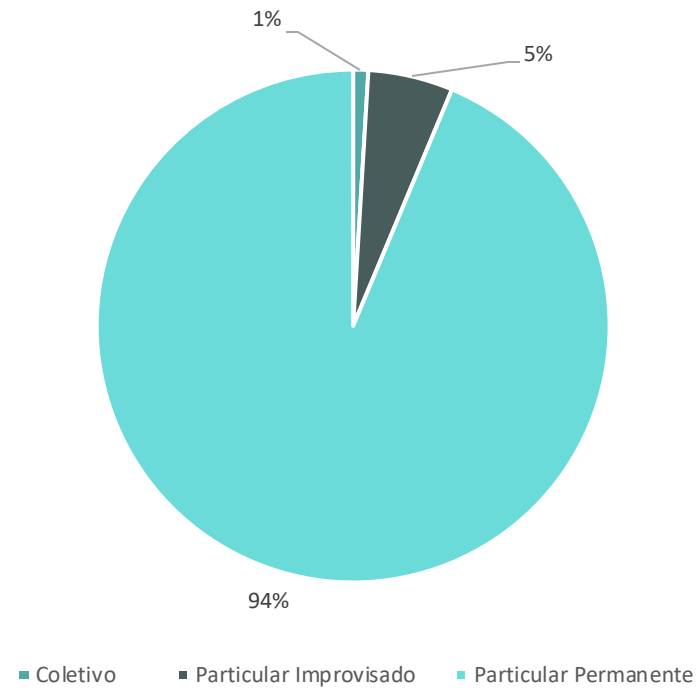


Categoria	Qtd.	%
Rural	29726	24%
Urbana	96528	76%

Grande parte da população vive em região urbana



Residencia – Tipo de residencia

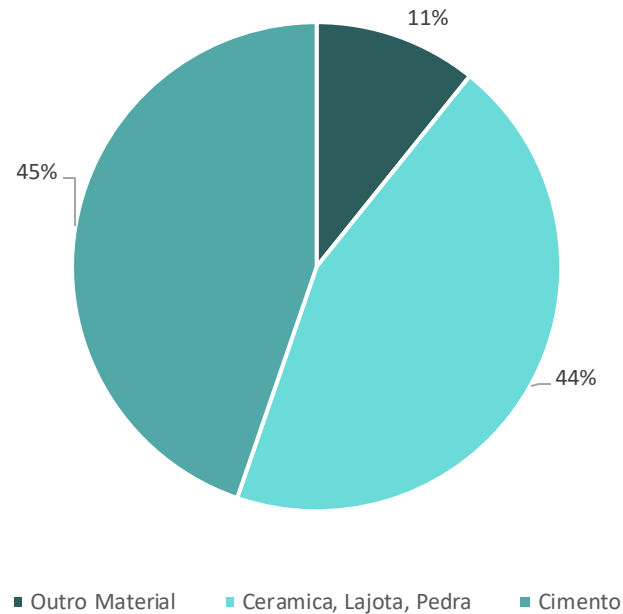


Categoria	Qtd	%
Coletivo	1200	1%
Particular Permanente	6756	94%
Particular improvisado	118298	5%

Quase toda população vive em algum imóvel permanente, com um total **94%**. Esta variável não foi incluída em modelo de regressão linear pois poderia prever um único tipo de população



Residencia – Material do piso da residencia

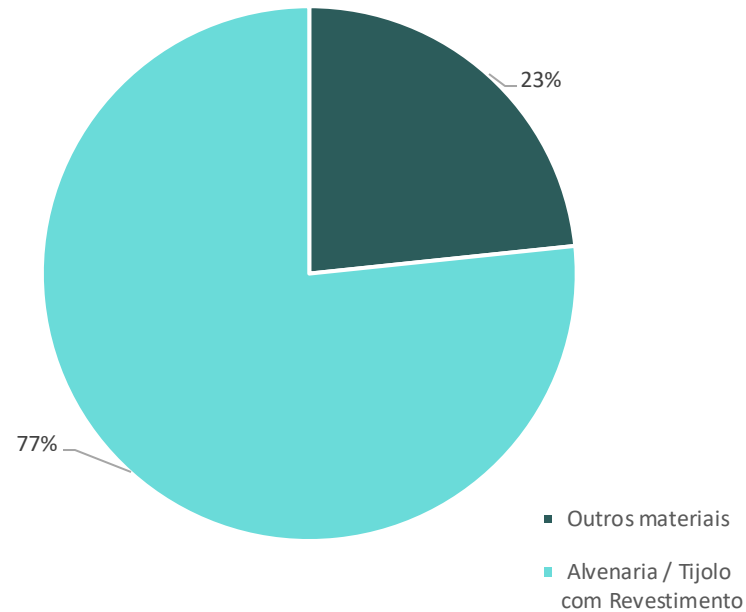


Categoria	Qtd	%
Outro Material	13599	11%
Ceramica, Lajota, Pedra	56181	44%
Cimento	56474	45%

45% das residências possuem piso de cimento, outros 44 possuem piso de cerâmica, lajota e pedra, embora sejam materiais diferentes, demonstra uma chão de residência previsto. 11% é de outros materiais.



Residencia – Material de construção

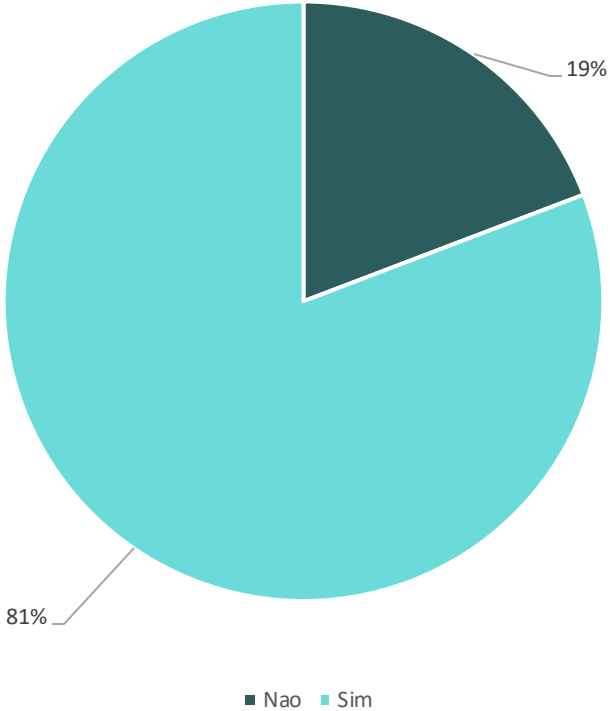


Categoria	Qtd	%
Outros materiais	29475	23%
Alvenaria / Tijolo com Revestimento	96779	77%

As construções são realizadas com materiais padrões, 77% faz uso d alvenaria/ tijolo com revestimento, também podemos considerar que poucas famílias usam materiais mais simples



Abastecimento – Água encanada

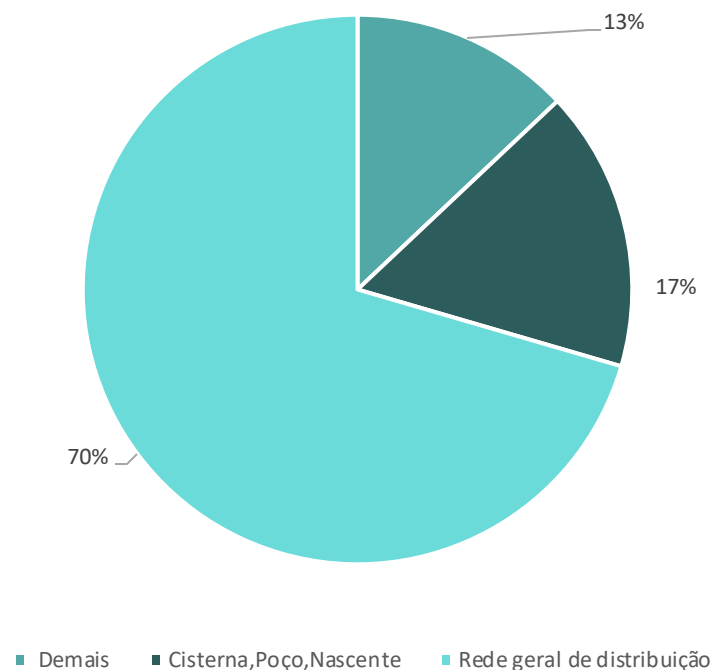


Categoria	Qtd	%
Nao	24256	19%
Sim	101998	81%

81% das residências possuem água encanada, no entanto, devemos observar o próximo item para avaliar se é de rede geral de distribuição ou essa água é recepcionada de outros locais que não o reservatório.



Abastecimento – Fornecimento de água

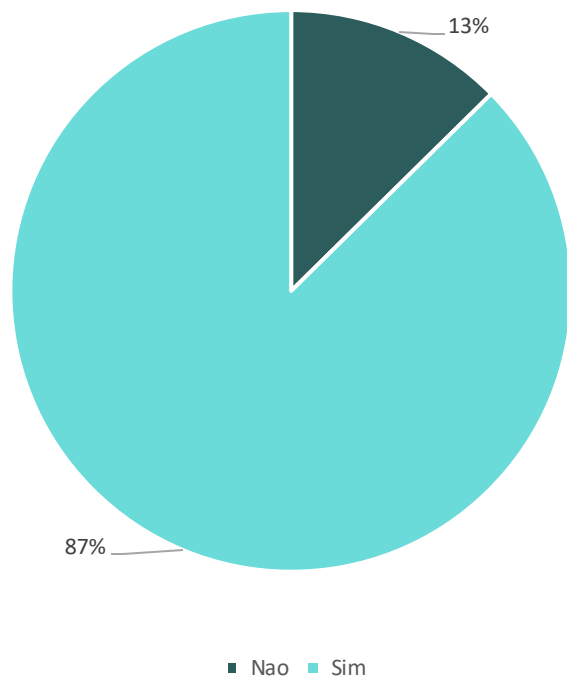


Categoria	Qtd	%
Demais	16380	13%
Cisterna,Poço,Nascente	20876	17%
Rede geral de distribuição	88998	70%

A distribuição por rede geral, tem se destacado com 70% da população, porem ainda podemos ver um número alto que não possui o fornecimento por empresas com esse objetivo (30%)



Saneamento – banheiro

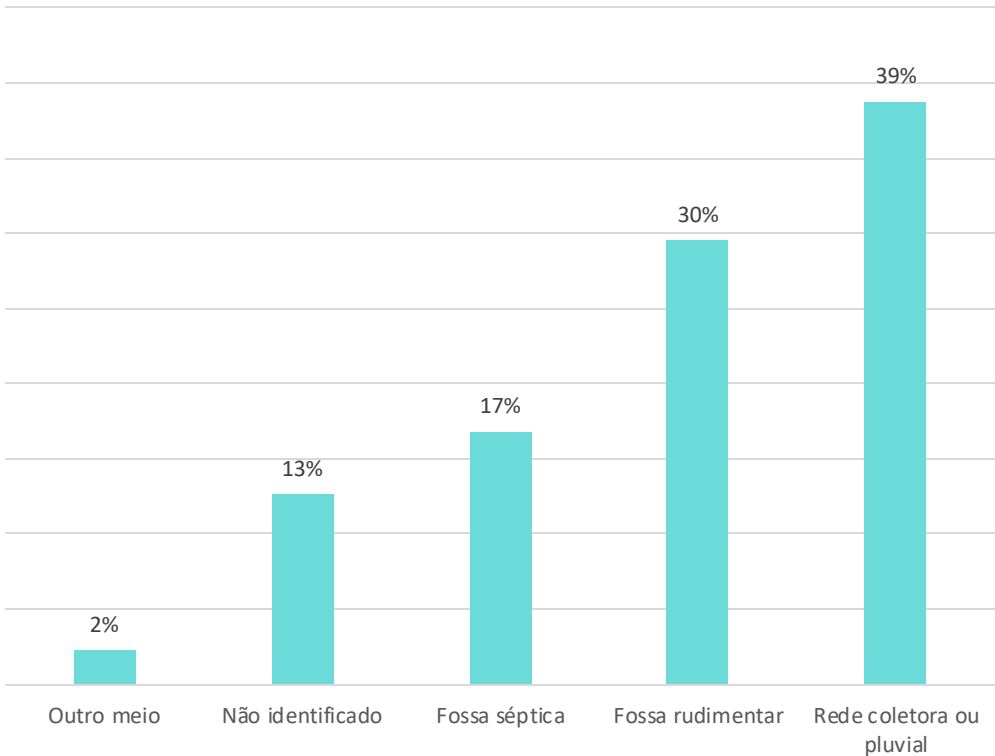


Categoria	Qtd	%
Sim	110317	87%
Não	15937	13%

87% das residências possuem banheiro, no entanto é necessário atenção básica aos 13% que não possuem.



Saneamento – coleta de esgoto

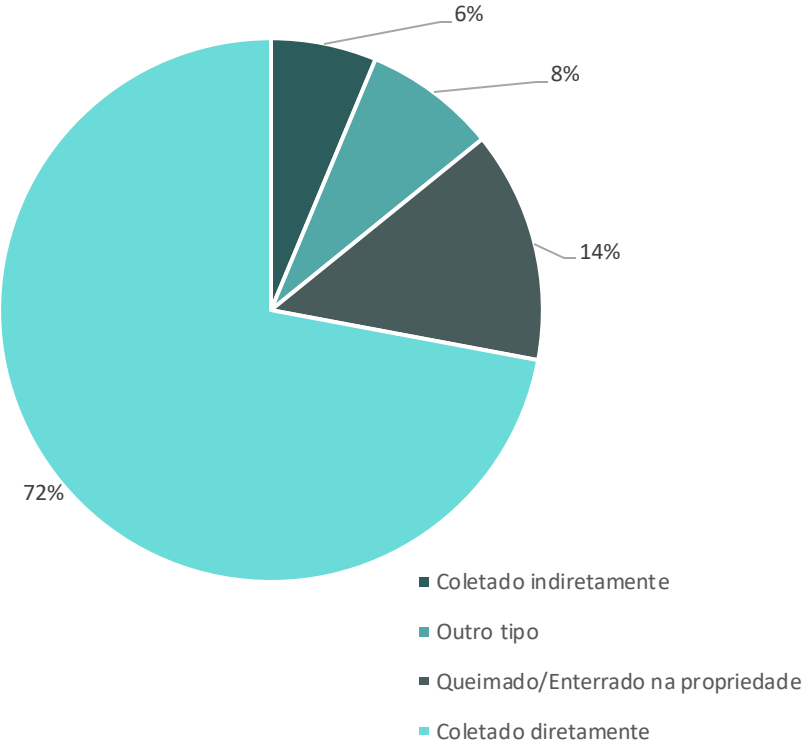


Categoria	Qtd	%
Outro meio	2873	2%
Não identificado	15937	13%
Fossa séptica	21227	17%
Fossa rudimentar	37294	30%
Rede coletora ou pluvial	48923	39%

Embora a rede coletora tenha um percentual significativo (39%), podemos observar outros meios que não são adequados ou melhores para um higiene familiar.



coleta de lixo

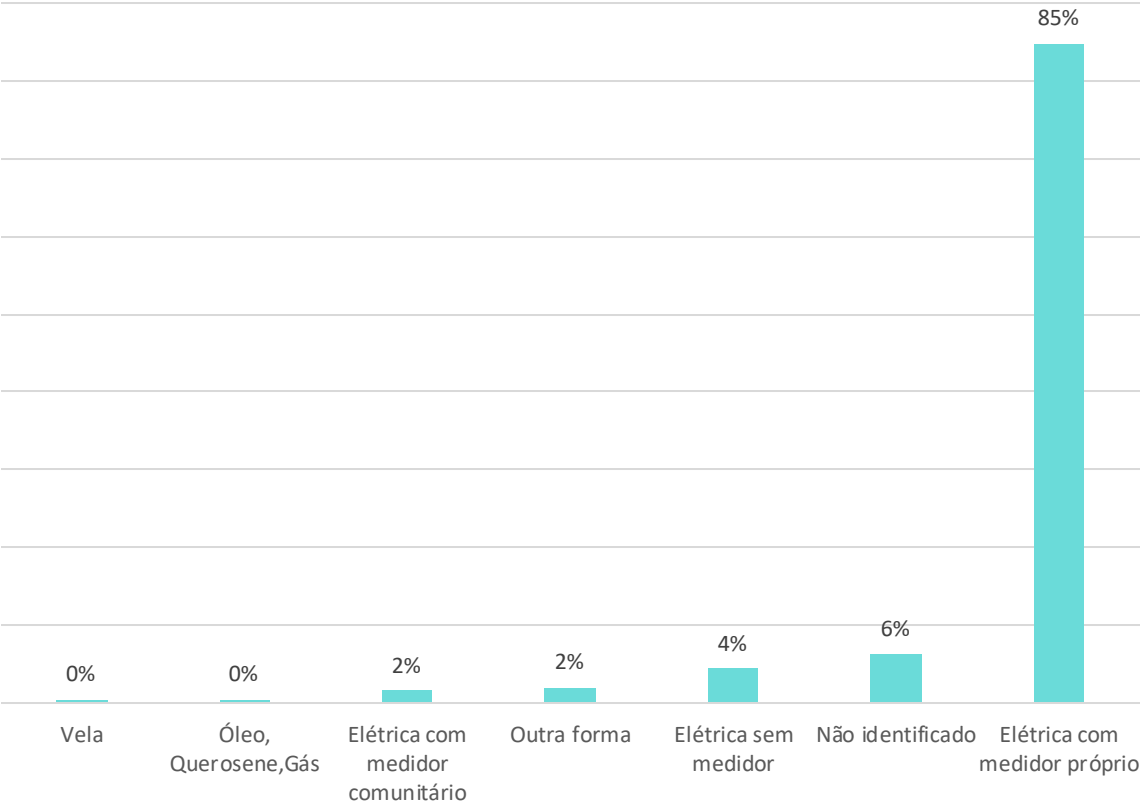


Categoria	Qtd	%
Coletado indiretamente	7935	6%
Outro tipo	9953	8%
Queimado/Enterrado na propriedade	17397	14%
Coletado diretamente	90969	72%

Grande parte da população tem o lixo coletado pelo município ou empresas com essa função



Iluminação

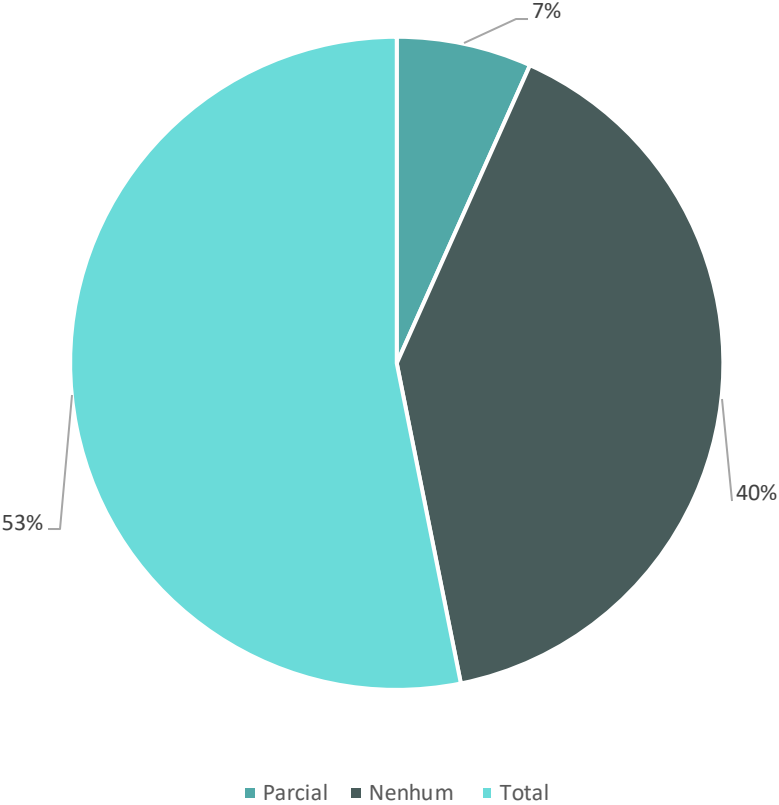


Categoria	Qtd	%
Vela	584	0%
Óleo, Querosene, Gás	605	0%
Elétrica com medidor comunitário	1970	2%
Outra forma	2548	2%
Elétrica sem medidor	5566	4%
Não identificado	7956	6%
Elétrica com medidor próprio	107025	85%

Grande parte da população possui energia com medidor próprio, isso é representado por 84% das famílias.



Calçamento

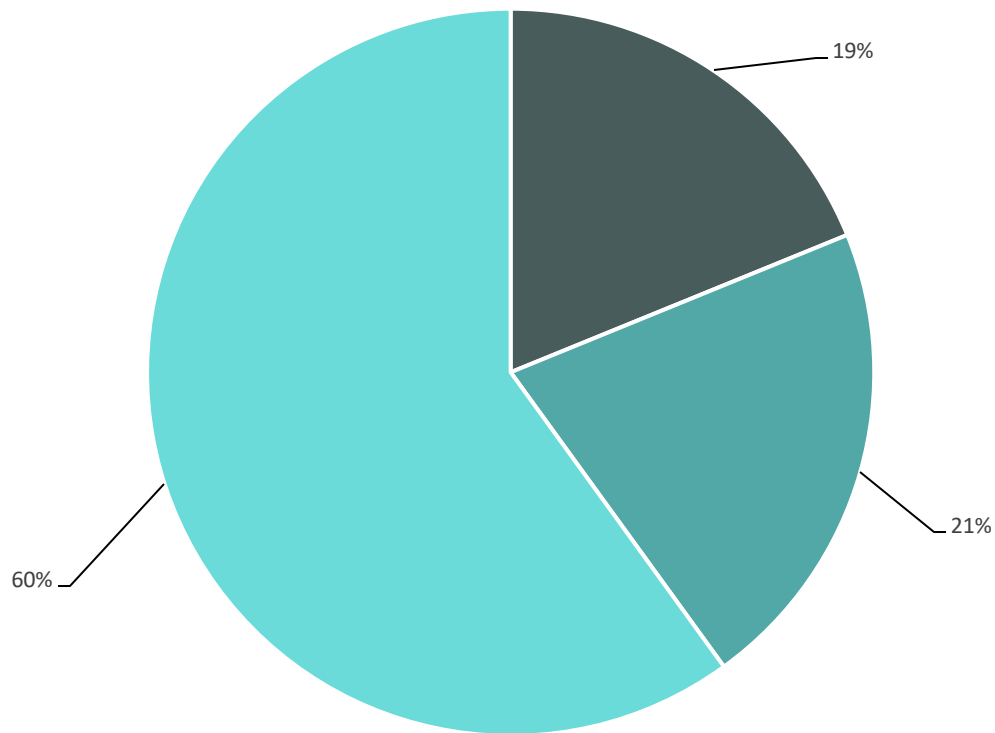


Categoria	Qtd	%
Parcial	8446	7%
Nenhum	50708	40%
Total	67100	53%

Do mesmo modo que grande parte da população vive com ruas calçadas (53%) , podemos observar que quase a outra metade da população vive próximas a ruas e estrada não calçadas (40%)



Classe Cidade



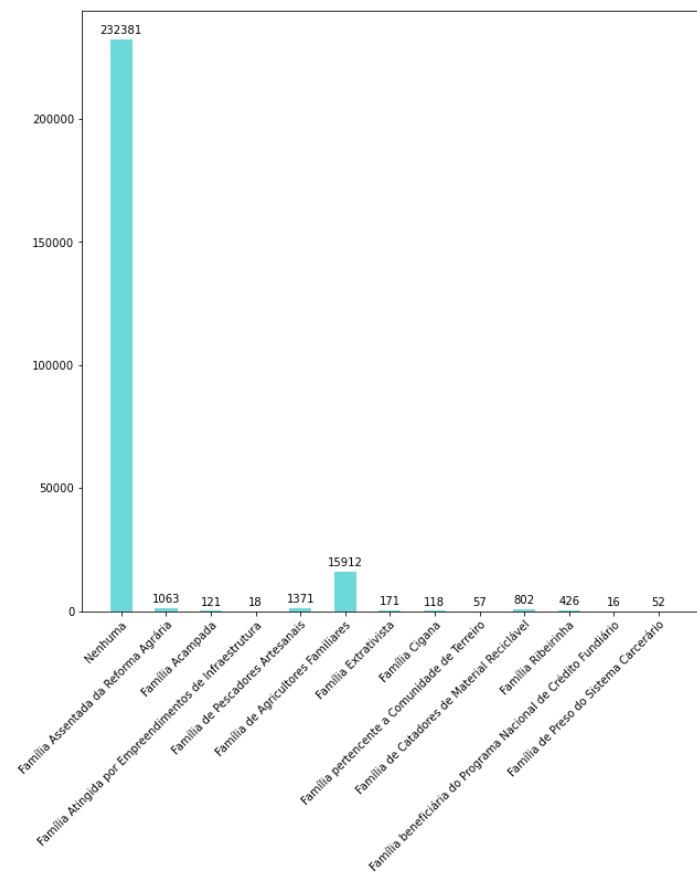
Categoria	Qtd	%
Capital	23779	19%
Região Metropolitana	26749	21%
Outros	75726	60%

■ Capital ■ Região Metropolitana ■ Outros

Um pouco mais da metade das famílias vivem no interior dos estados e relacionado (60%), enquanto a outra parte vivem na capital ou região metropolitana



Família

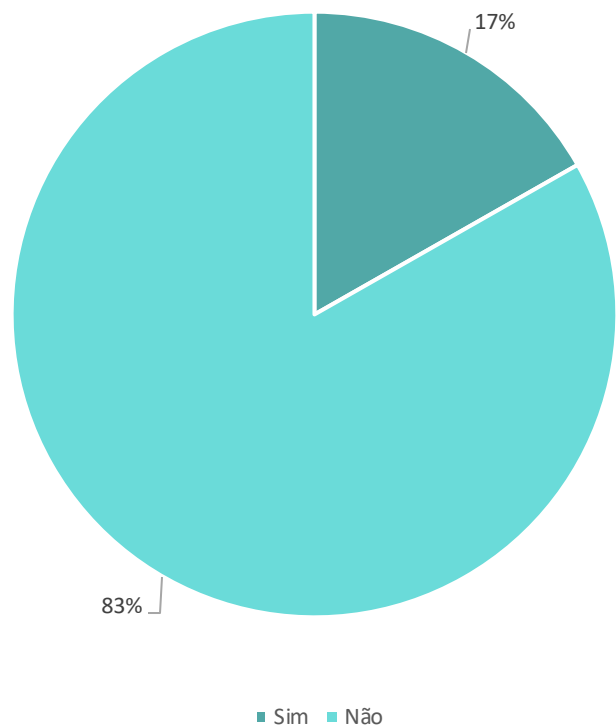


Categoria	Qtd	%
Nenhuma	22381	92%
Outros	168724	8%

92% das famílias não fazem partes de grupos diferenciados em nossa base de dados



Saúde – se tem deficiente

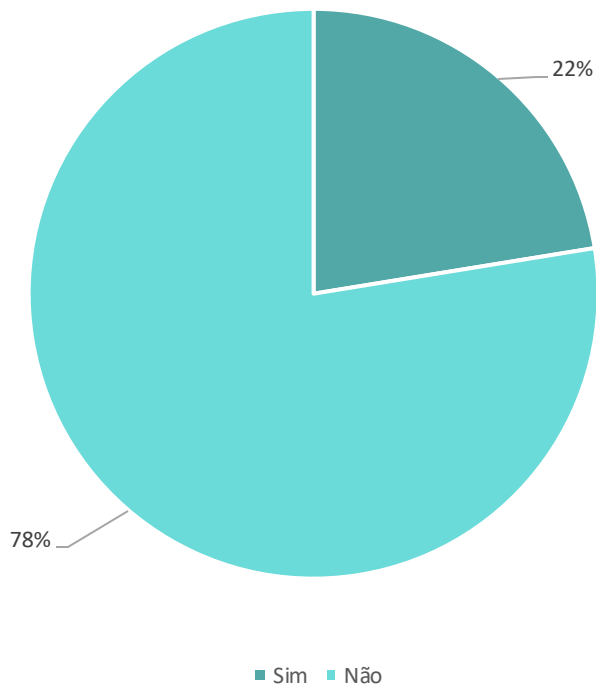


Categoria	Qtd	%
Sim	21174	17%
Nao	105080	83%

Majoritariamente, não há deficientes junto às famílias



Educação - Analfabeto

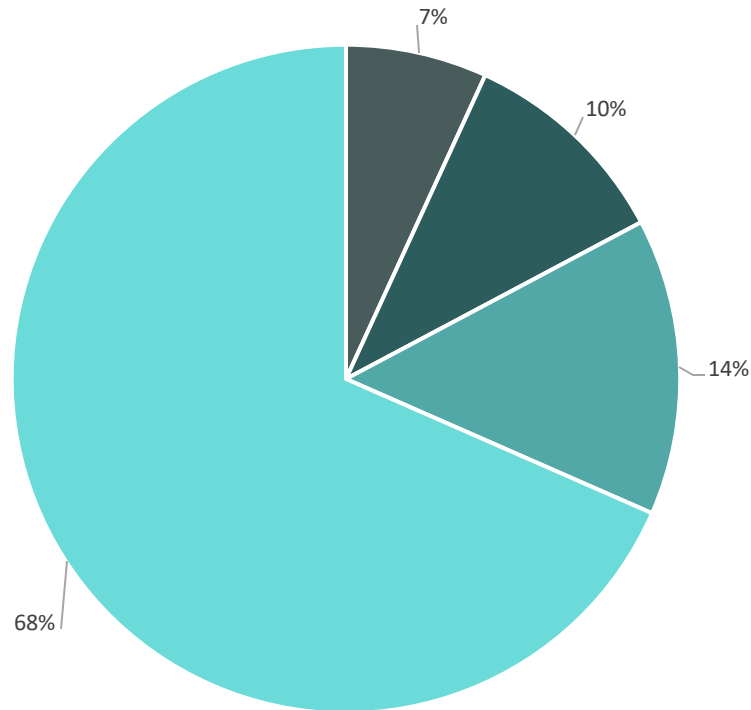


Categoria	Qtd	%
Sim	28327	22%
Não	97927	78%

Esta variável mostra as famílias que possuem ao menos um adulto que não sabe ler, 22% é um número alto para a população e merece um olhar mais próximo do poder público.



Educação - menores



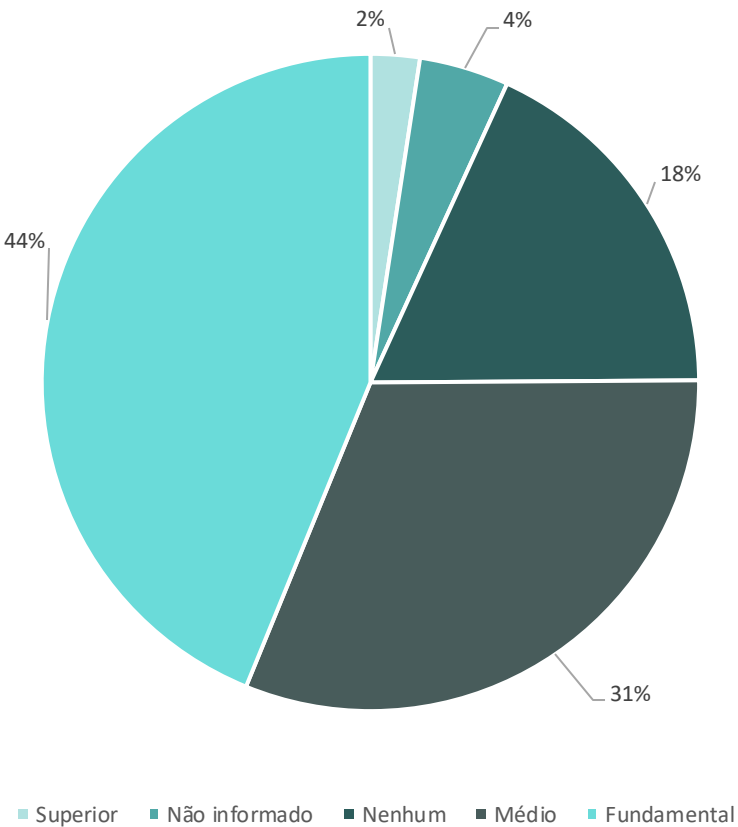
Categoria	Qtd	%
Menores não estudam	8640	7%
Sem idade ou algum não estuda	13114	10%
Todos estudam	18117	14%
Não tem menor	86383	68%

■ Menores não estudam ■ Sem idade ou algum não estuda ■ Todos estudam ■ Não tem menor

Grande parte das famílias não tem menores, no entanto podemos observar que daqueles que possuem boa parte ou está na escola (14%) ou não tem idade – incluir também as famílias que um menor não estudam - (10%), uma atenção básica deve ser dada as famílias que o menor não estuda.



Educação - família

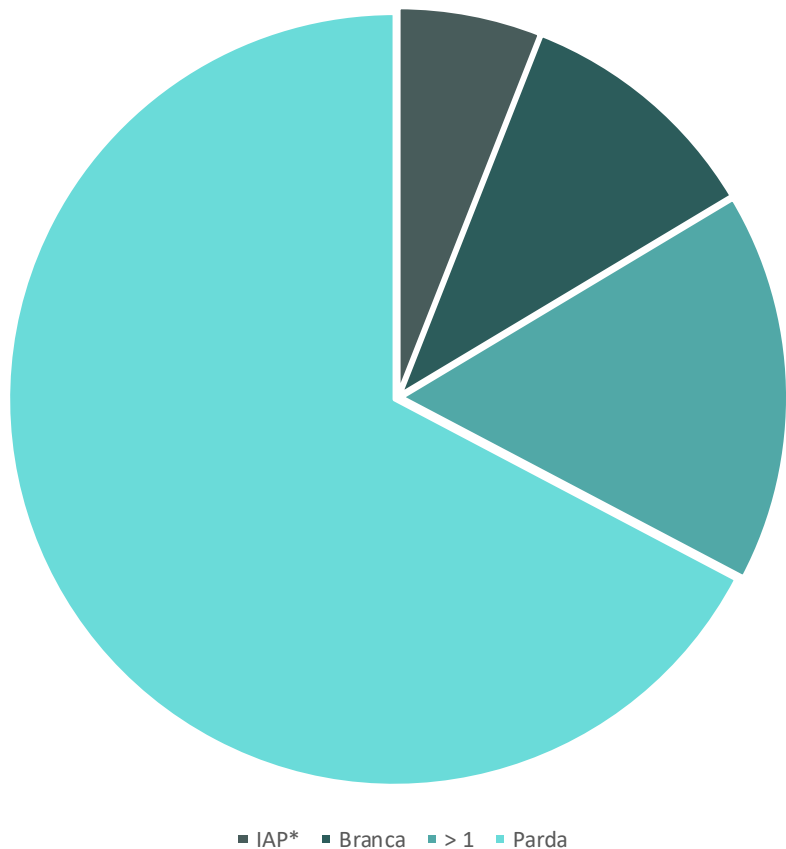


Categoria	Qtd	%
Superior	3055	2%
Não informado	5568	4%
Nenhum	22809	18%
Médio	39500	31%
Fundamental	55322	44%

Na família, o grau mais alto em destaque é o relacionado ao ensino fundamental representado por 44%, no entanto esse cenário pode mudar naturalmente, dado que os menores ao se formarem estarão em nível educacional maior



Cor / Raça



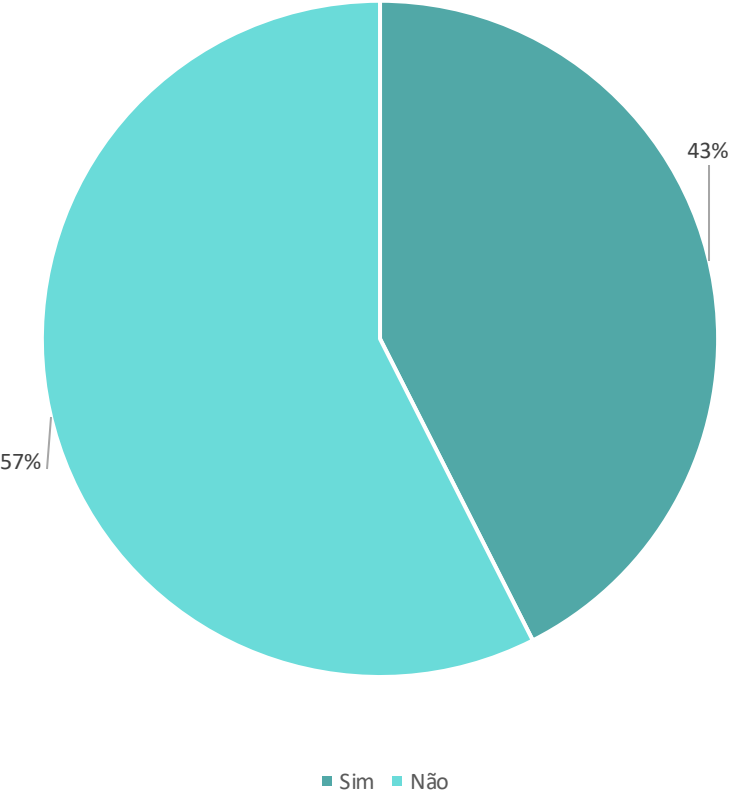
Categoria	Qtd	%
IAP *	7530	6%
Branca	13186	10%
> 1 **	20589	16%
Parda	84949	67%

A população nordestina neste período, vem sendo representa pela população parda com 67%

* Indígena, amarela, preta
** Mais de uma



Beneficiário do bolsa família

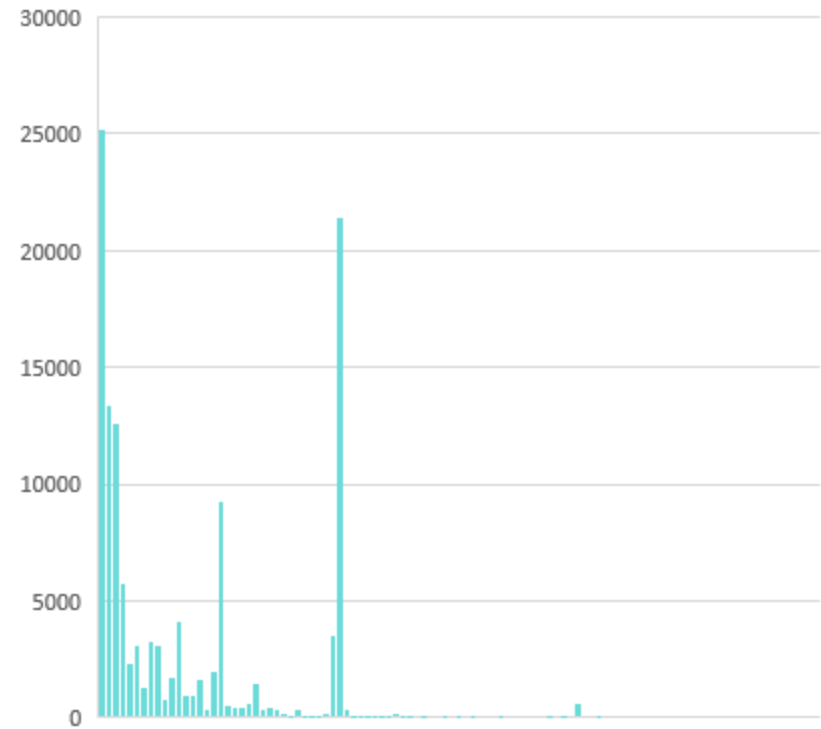
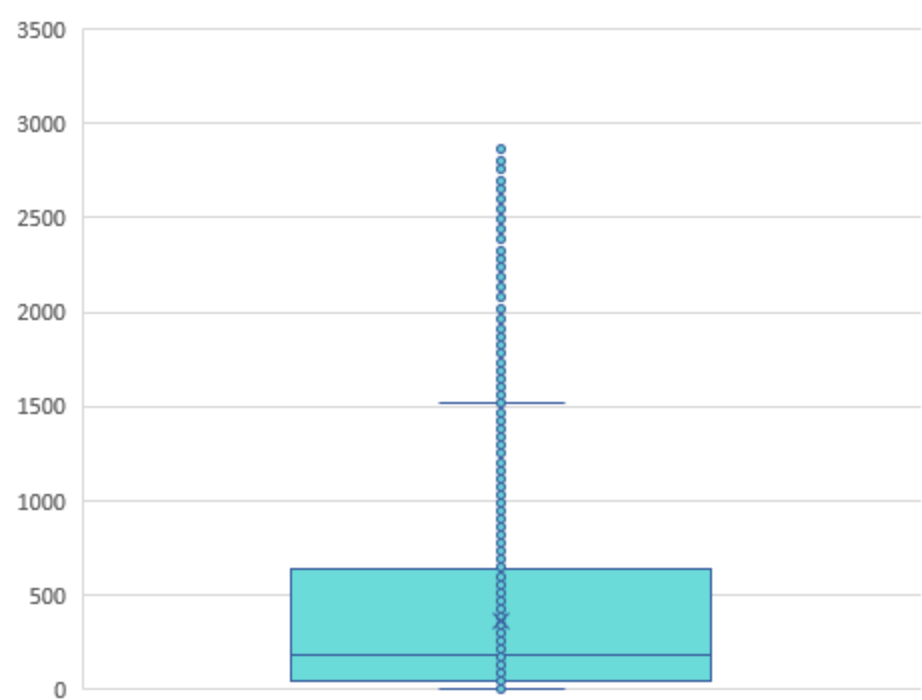


Categoria	Qtd	%
Sim	53706	43%
Não	72548	57%

57% das famílias não estão envolvidas com o bolsa família



Renda Media Familiar

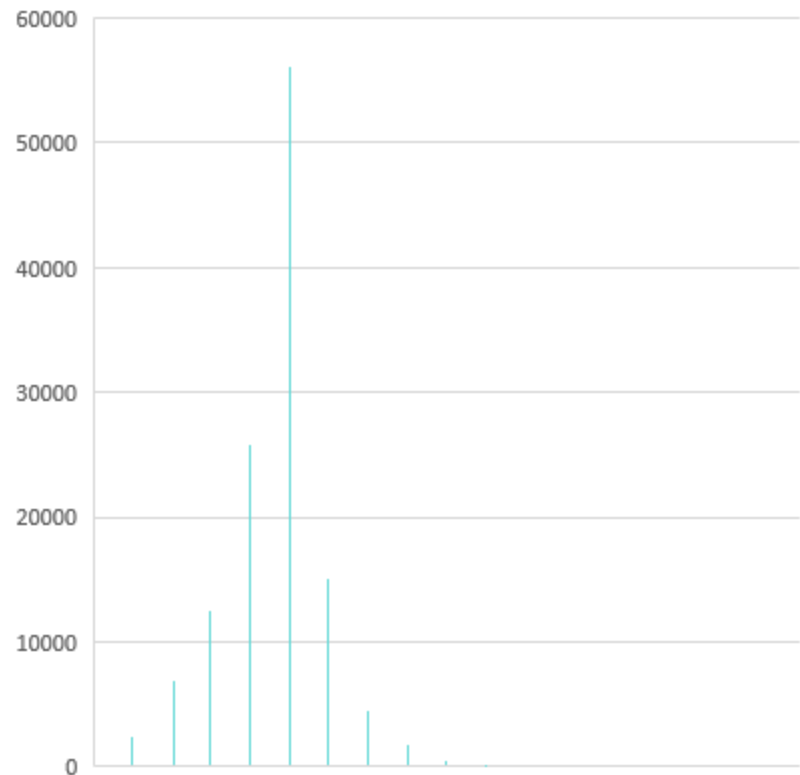
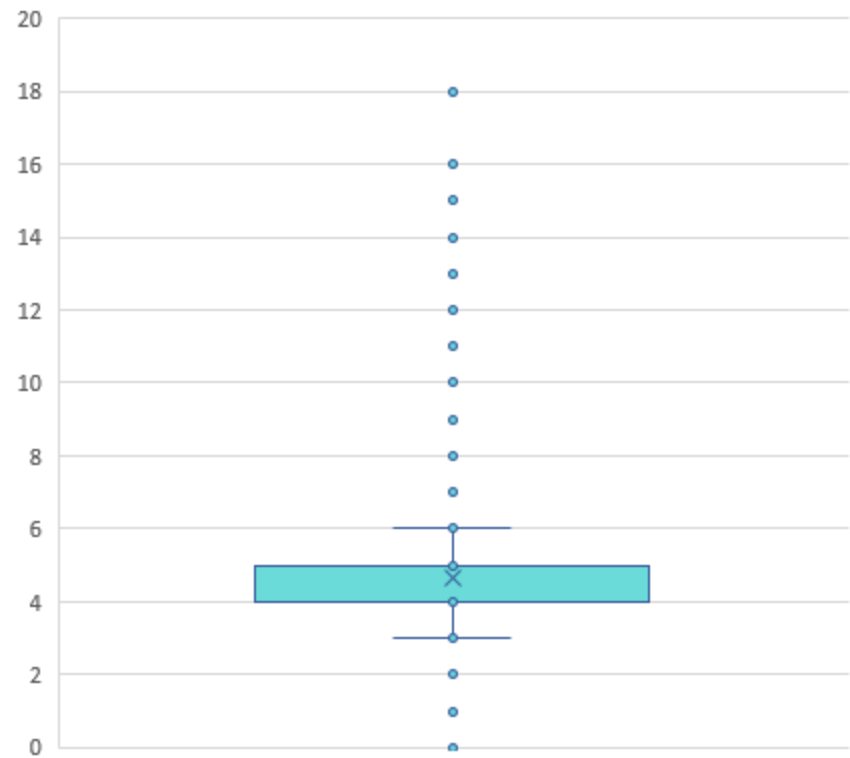


Medida	Valor
Média	364,04
Desvio	400,25
1 Quartil	45
2 Quartil	183
3 Quartil	636
Coef. Variação	1.10

A média da renda familiar é de 364,04 reais, o desvio padrão é muito alto, sendo de 400,25, no entanto pode se justificar devido aos outliers. 25% da família recebe até 45,00 reais, 50% das famílias recebem até 180 reais, e os demais familiares recebem a partir de 636,00 reais. Podemos observar a presença de muitos outliers.



Qtd comodo na família

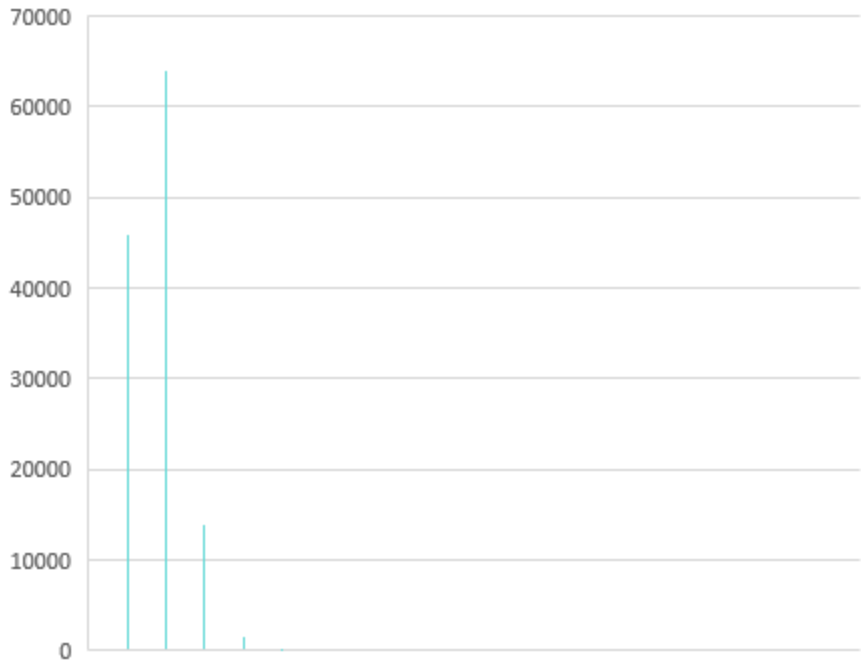
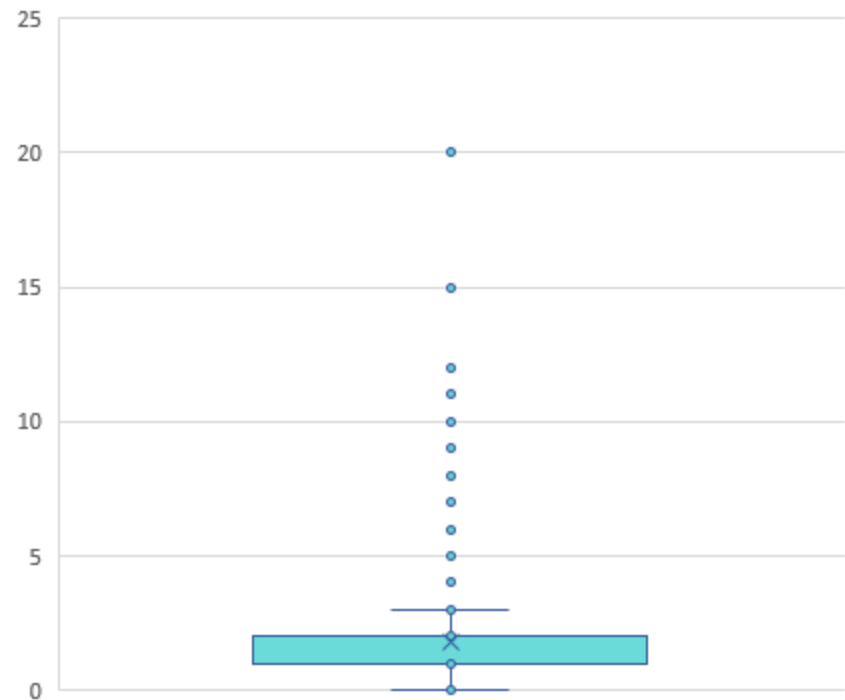


Medida	Vlr
Média	4.6
Desvio	1,35
1 Quartil	4
2 Quartil	5
3 Quartil	5
Coef. Variação	0.29

As residências, possuem em média 5 cômodos, 25% da população reside em casas com até 4 cômodos, metade das famílias residem em casa com até 5 cômodos, podemos observar a presença de outliers.



Qtd Dormitório na família

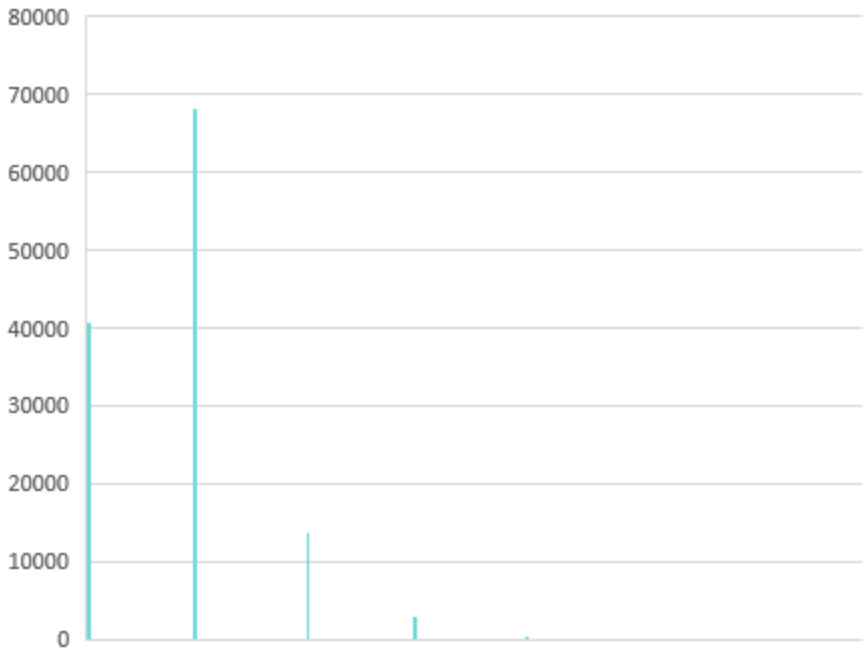
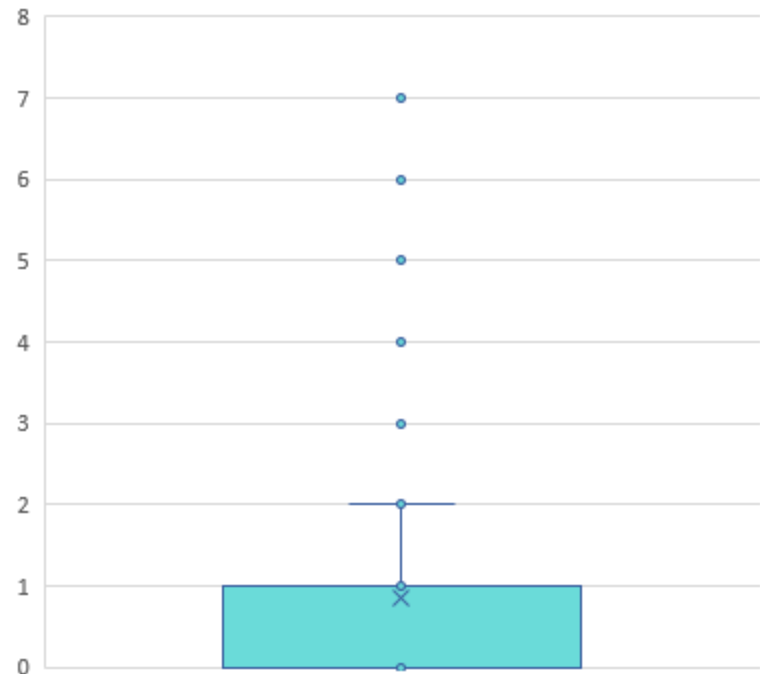


Medida	Vlr
Média	2
Desvio	0,73
1 Quartil	1
2 Quartil	2
3 Quartil	2
Coef. Variação	0,41

As residências, possuem em média 2 dormitórios , 25% das famílias possuem em sua residência 1 dormitório, metade das famílias possuem em sua residência com 2 dormitórios, o desvio padrão demonstra que não há uma mudança abrupta nesses valores.



Homens

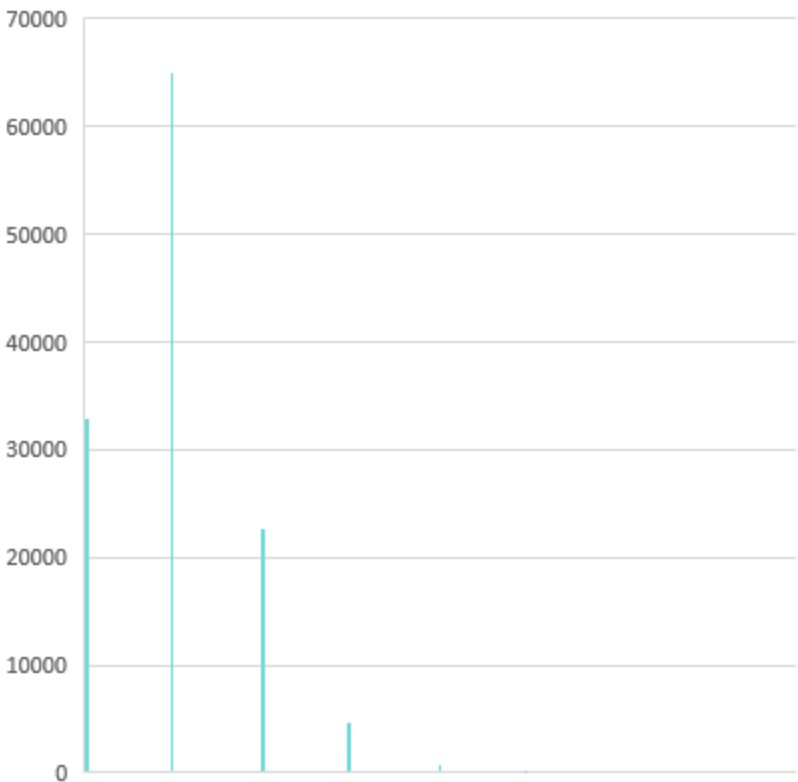
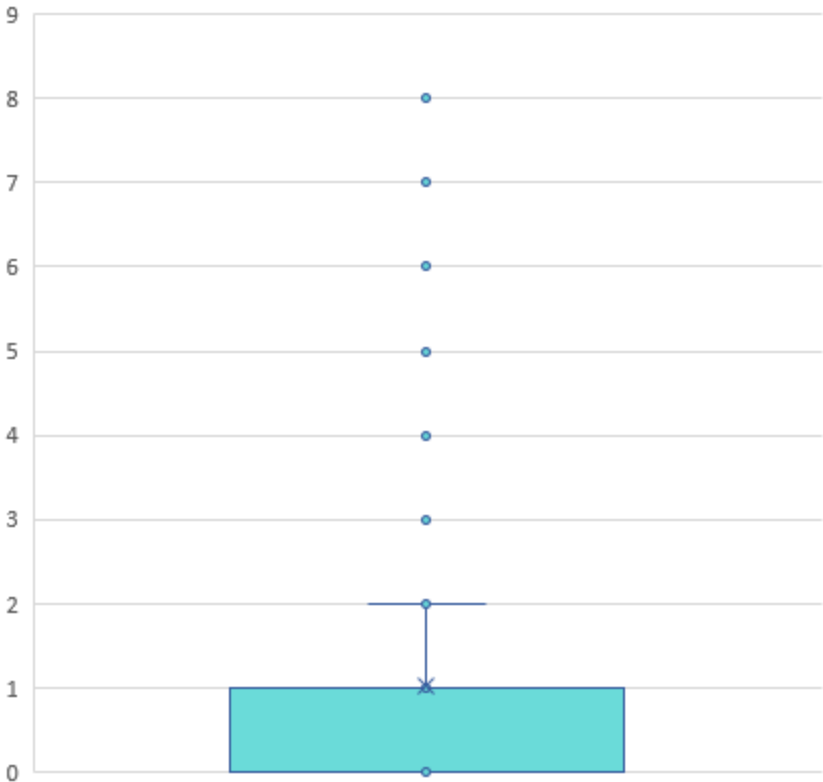


Medida	Vlr
Média	1
Desvio	0,73
1 Quartil	0
2 Quartil	1
3 Quartil	1
Coef. Variação	0.87

A media de homens na família é de 1, no entanto podemos ver que até 25% das famílias não possuem homens, e até 75% das famílias possuem somente um homem, o desvio padrão não é alto, logo indica que o número de homens permanece baixo em boa parte das famílias



Mulheres

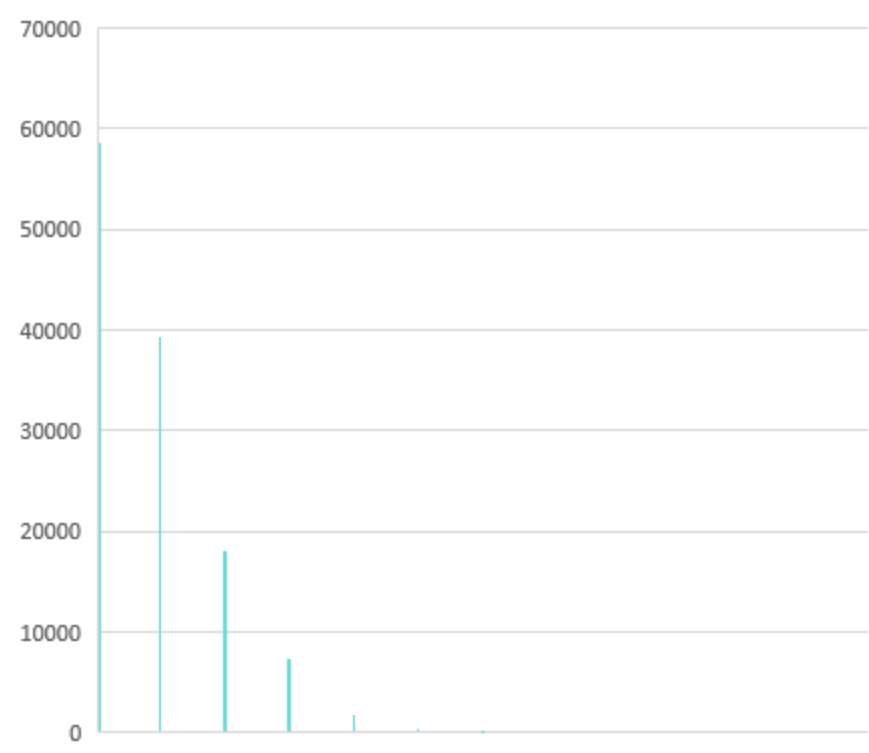
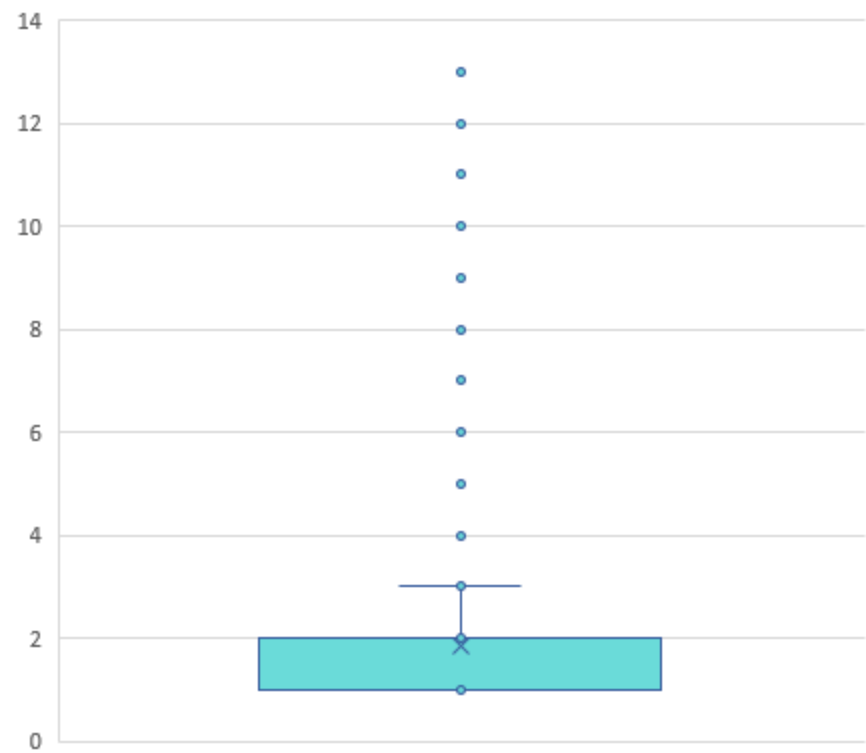


Medida	Vlr
Média	1
Desvio	0,81
1 Quartil	0
2 Quartil	1
3 Quartil	1
Coef. Variação	0.80

A presença de mulheres nas famílias tem média igual a um, o desvio padrão é maior que a dos homens, logo teremos mais chances de ver mulheres na família, e até 75 das famílias possuem apenas uma mulher



Pessoas

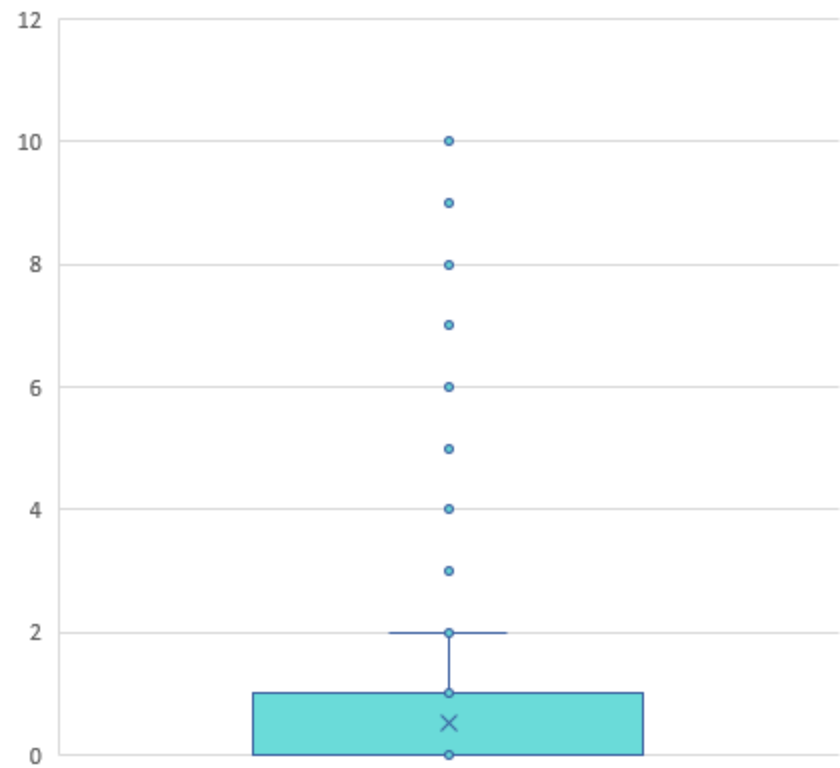


Medida	Vlr
Média	1,8
Desvio	1,03
1 Quartil	1
2 Quartil	2
3 Quartil	2
Coef. Variação	0.55

A média de indivíduos na família é um, com um desvio semelhante a esses valores, possivelmente a uma variabilidade alta, é possível observar que temos 75% das famílias formadas por até duas pessoas.



Menor de idade

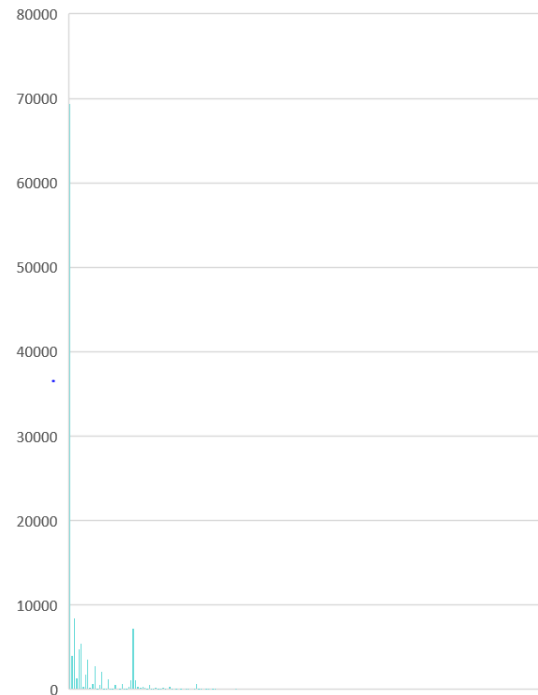
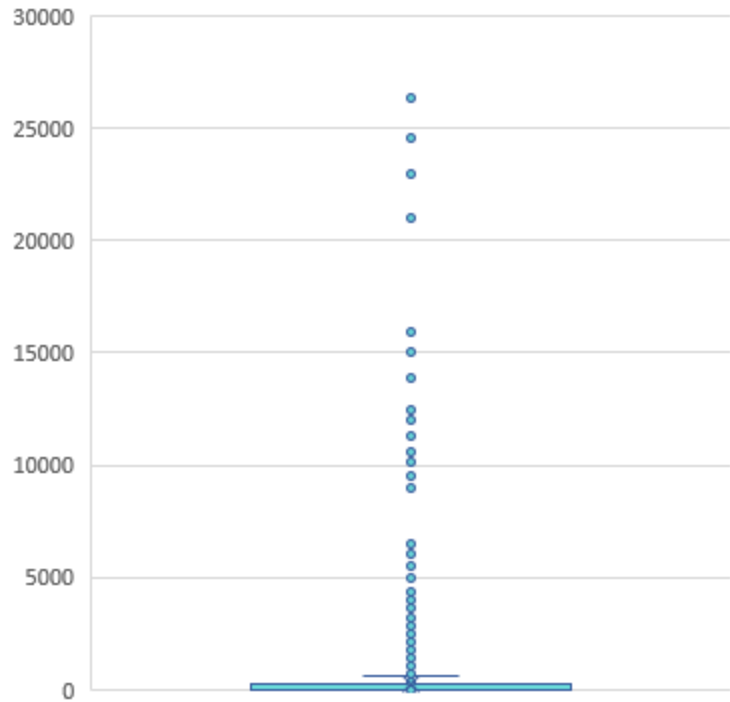


Medida	Vlr
Média	0.5
Desvio	0,81
1 Quartil	0
2 Quartil	0
3 Quartil	1
Coef. Variação	1.57

A presença de crianças na população familiar em geral é muito baixa, o desvio padrão é maior que a média, o que indica que existe famílias com números alto de crianças puxando o desvio para cima. Se analisarmos outras variáveis podemos ver que a maioria das famílias não possuem crianças



Renda de trabalho registrado

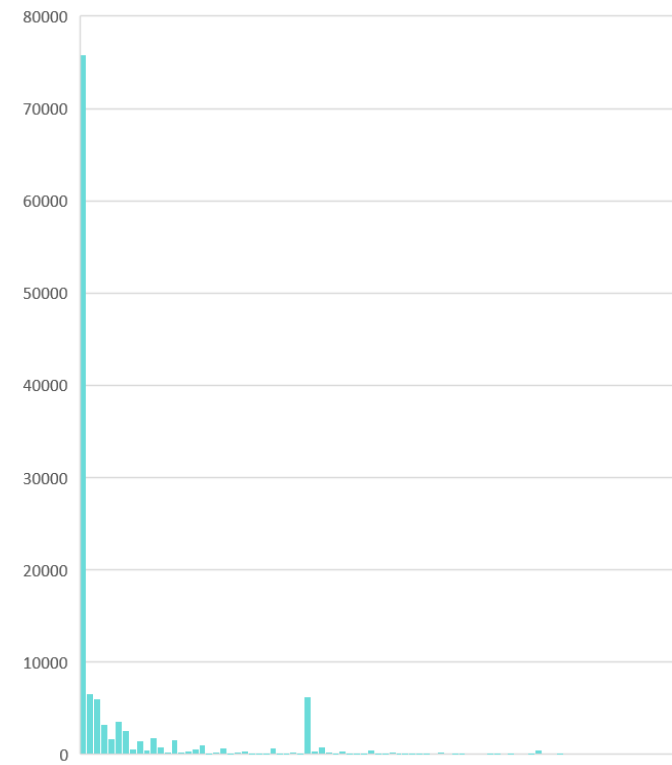
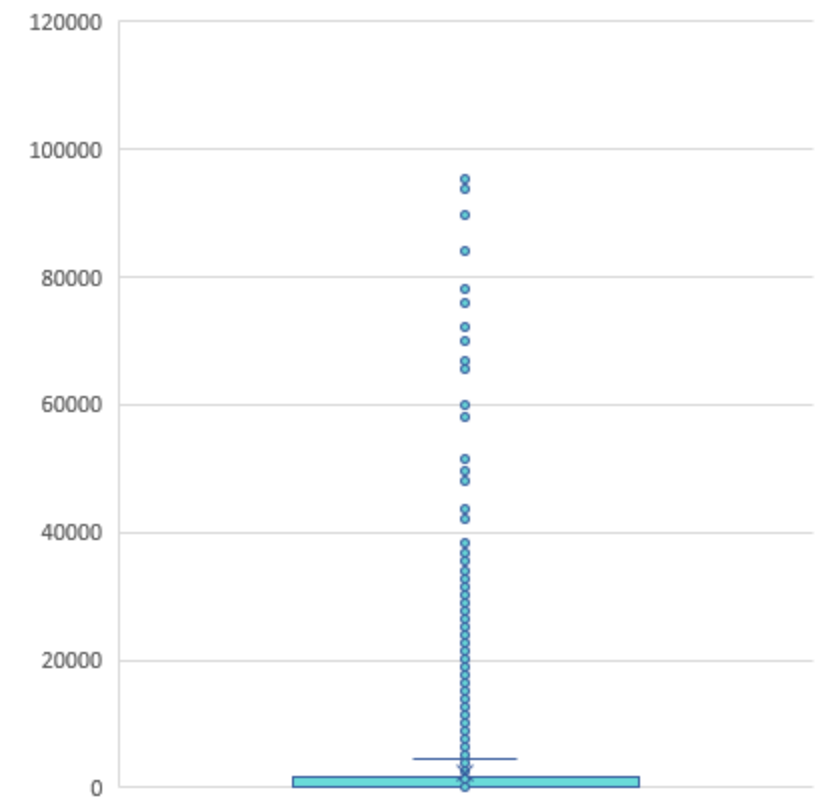


Medida	Vlr
Média	242,60
Desvio	493,4
1 Quartil	0
2 Quartil	0
3 Quartil	250
Coef. Variação	2.03

A renda média é baixa, podemos ver que até a maioria das famílias não tem renda por trabalho registrado, ou essa renda é muito baixa. 25% das família que mais ganharam algum valor, receberam R\$ 250,00 ou mais.



Soma de renda (12 Meses)

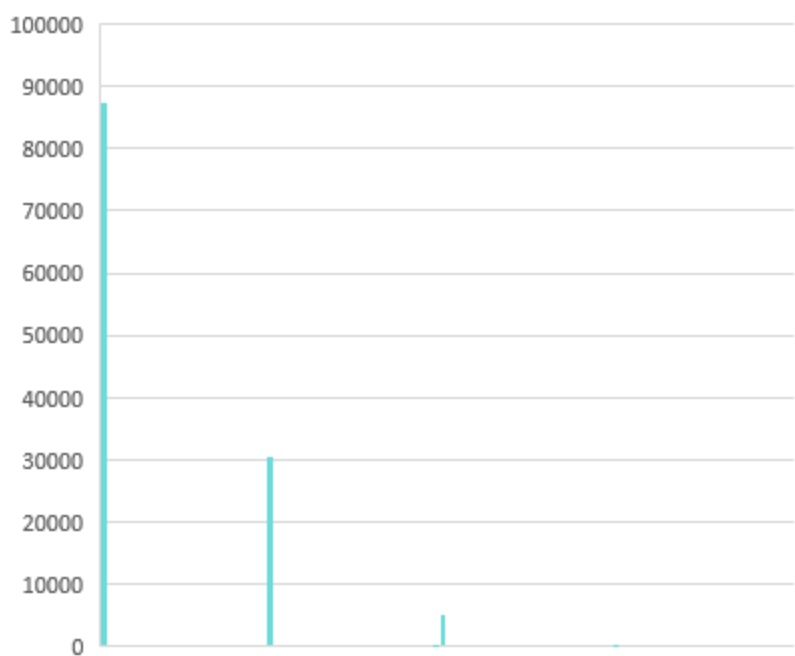
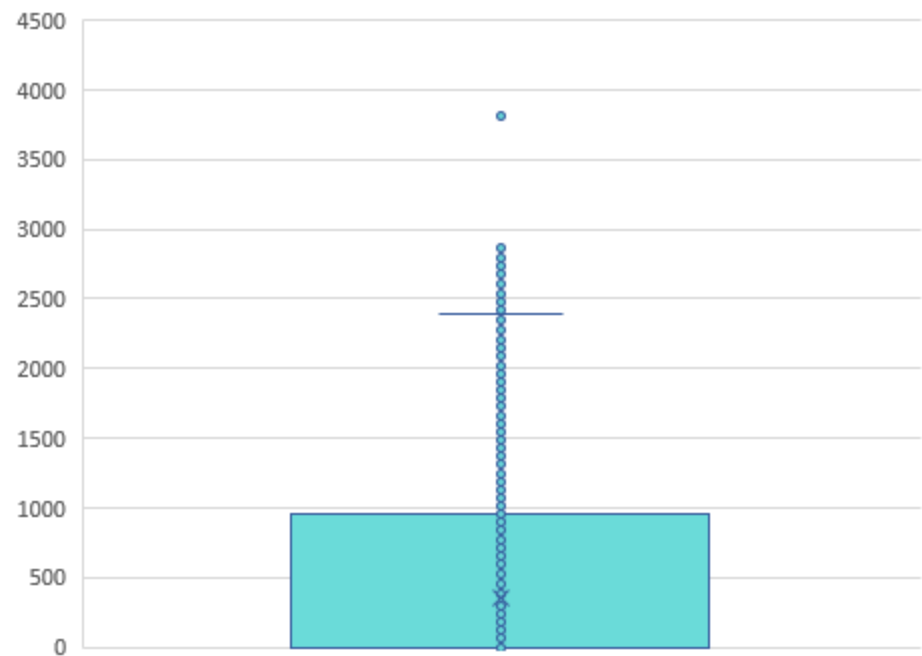


Medida	Vlr
Média	2389,31
Desvio	5066,93
1 Quartil	0
2 Quartil	0
3 Quartil	1800
Coef. Variação	2.12

Observamos um alto desvio alto elevado pelos outliers, com uma média 2 Mil reais. Mas podemos observar que a maioria das famílias não teve renda de trabalho registrado, dos 25% que receberam por meio de trabalho acumularam no ano um valor igual ou maior a 1.8 Mil



Aposentadoria

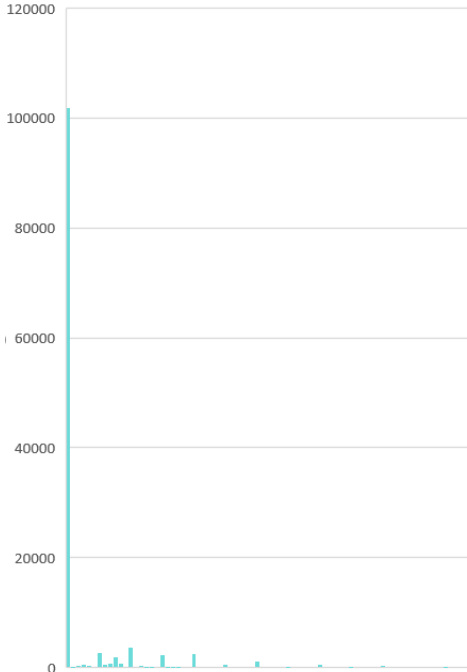


Medida	Vlr
Média	345,83
Desvio	568,34
1 Quartil	0
2 Quartil	0
3 Quartil	954
Coef. Variação	1,64

Até 50% das famílias não recebem, e 25% das famílias recebem 954 Reais ou mais. O desvio padrão é elevado devido a grande quantidade de famílias sem aposentados. O distribuição assimétrica acentuada a direita evidência uma média baixa. Notamos alguns outliers com um em específico em destaque como podemos notar no gráfico.



Outras rendas

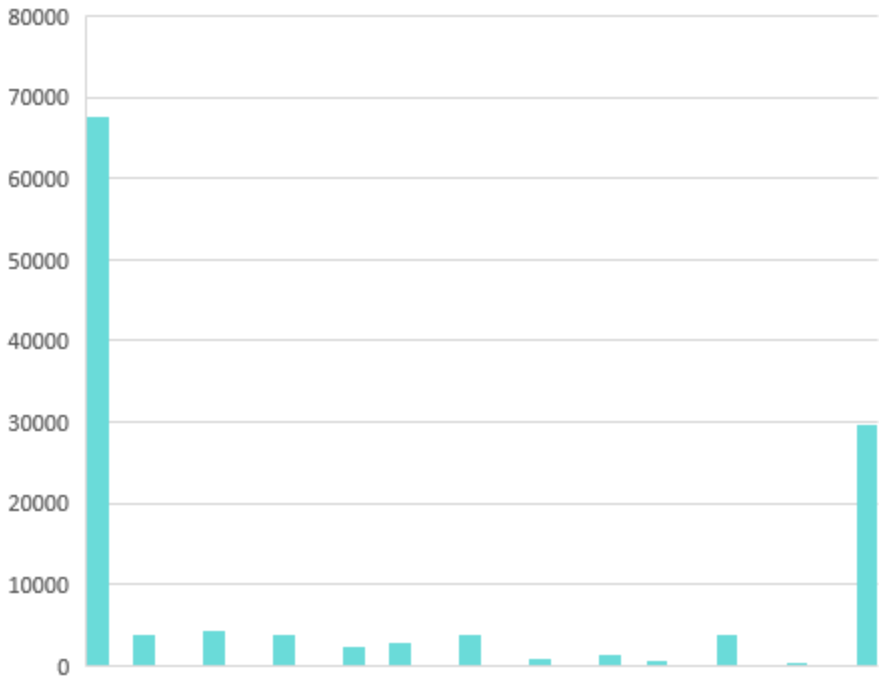
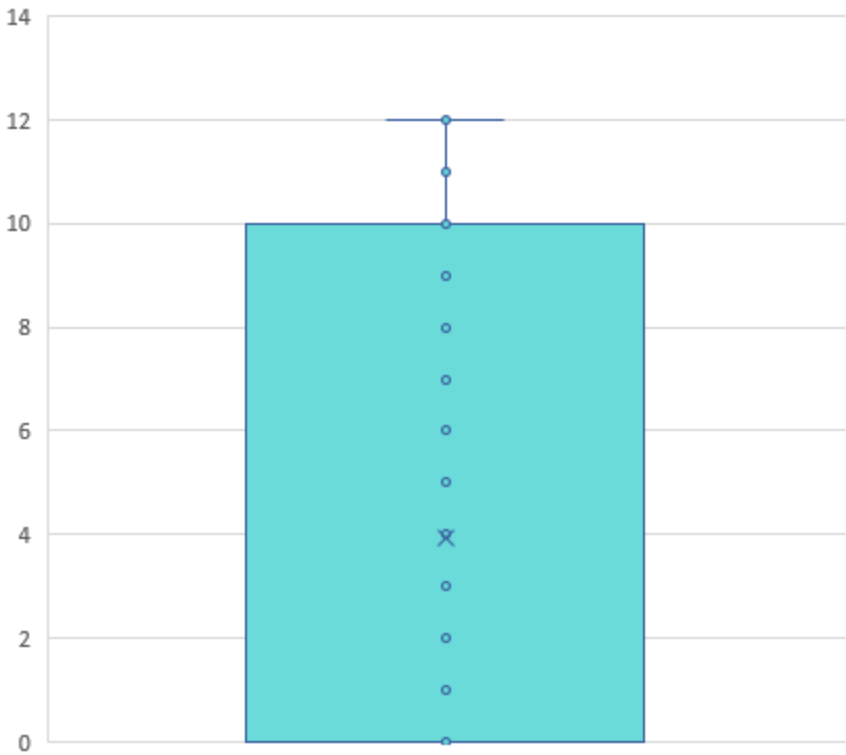


Medida	Vlr
Média	35,88
Desvio	118,30
1 Quartil	0
2 Quartil	0
3 Quartil	0
Coef. Variação	3,30

Variável formada por outliers, com isso temos um desvio padrão alto, e uma média baixa, podemos ver que quase toda a população não tem essa renda diversas, no modelo essa variável foi categorizada.



Periodo de trabalho

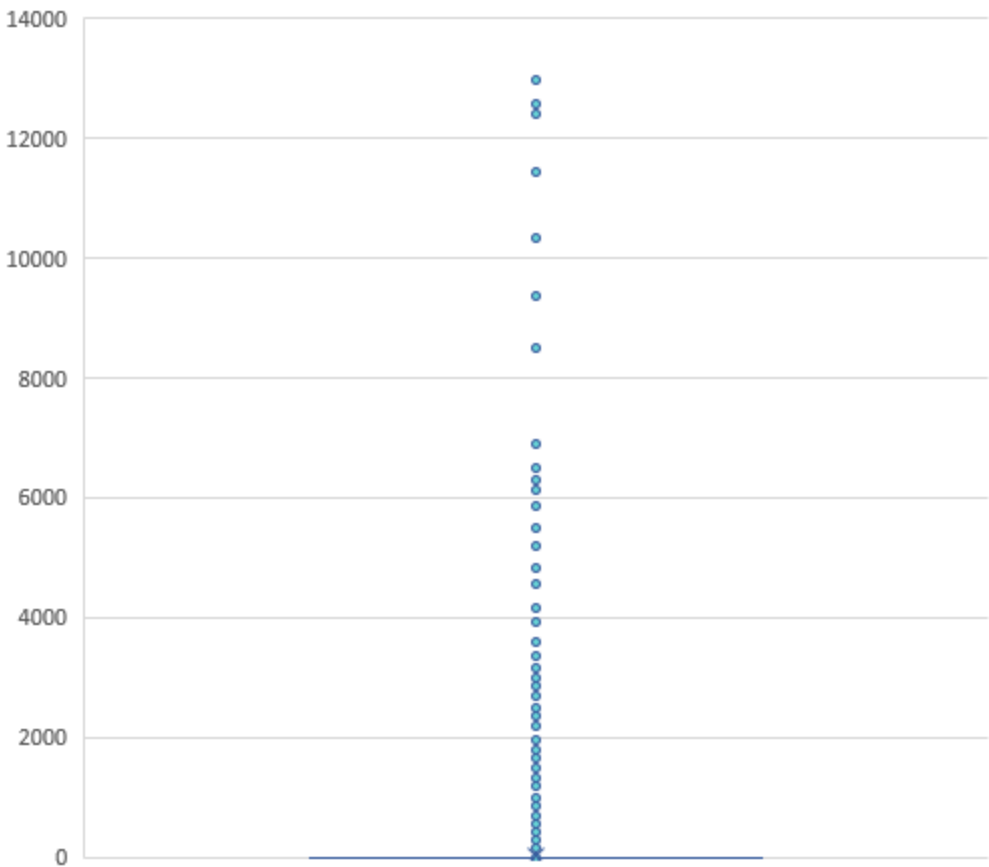


Medida	Vlr
Média	3,93
Desvio	5.09
1 Quartil	0
2 Quartil	0
3 Quartil	10
Coef. Variação	1.30

A média é de três meses de trabalho, com 50% das famílias ficando sem trabalhar no último ano. A média de meses trabalhado é influenciada para baixo devido as famílias sem trabalho.



Renda menor de idade

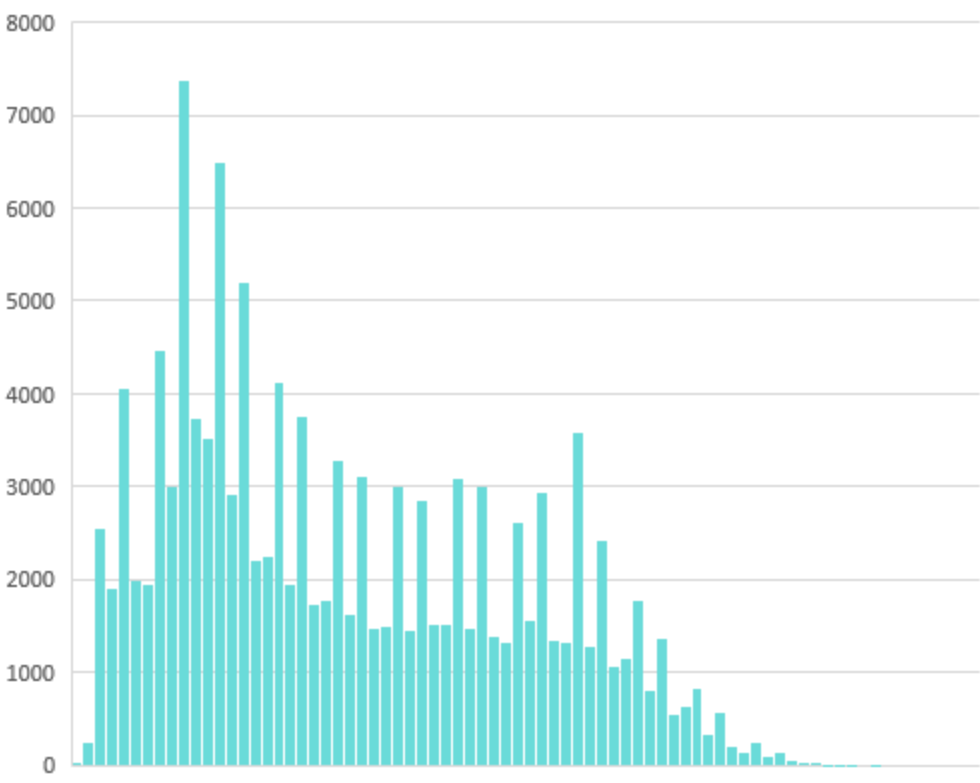
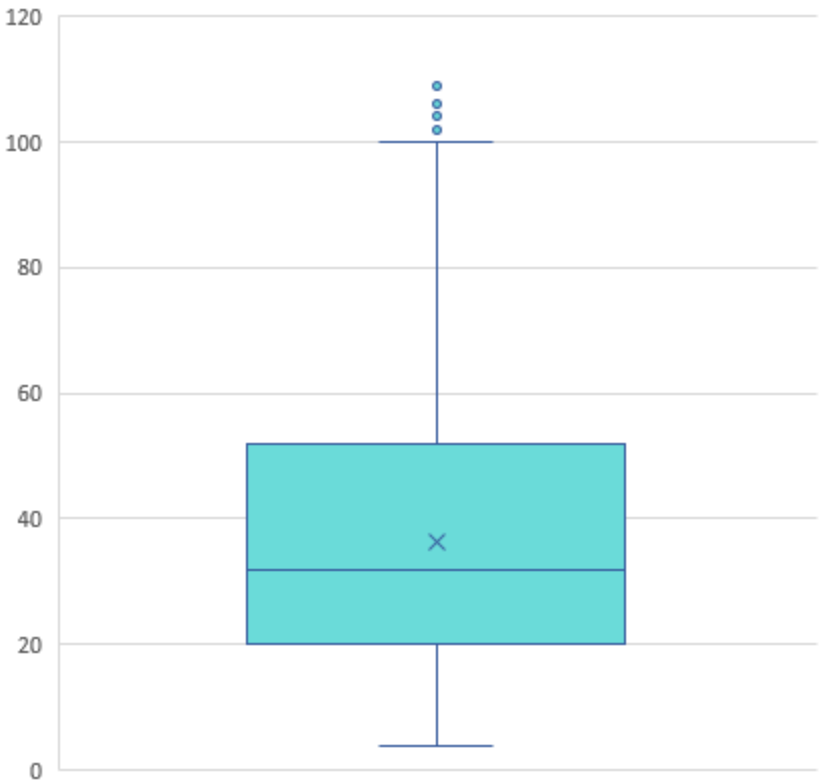


Medida	Vlr
Média	4,88
Desvio	136,64
1 Quartil	0
2 Quartil	0
3 Quartil	0
Coef. Variação	27,98

75% da população não tem trabalho infantil ou relacionado, á média é de 4,88 reais com desvio e 136 reais elevado pelos outliers. No modelo essa variável foi categorizada.



Idade média



Medida	Vlr
Média	36,46
Desvio	19,84
1 Quartil	20
2 Quartil	32
3 Quartil	52
Coef. Variação	0,54

Temos uma média de idade relativamente baixa, com 36 anos. 50% da população possui até 32 anos.



Dicionário de dados



Variável	Descrição
ID_FAMILIA	Identificação da família
SEXO_RESP	Sexo do responsável familiar
UF	Estado de residência familiar
MESORREGIAO	Mesorregião de residência familiar
VL_MED_FAM	Renda média Familiar
LOCAL_DOM_FAM	Local de domínio familiar, se urbana ou rural
ESPEC_DOM_FAM	Especie de domínio familiar
QTD_COMODO_DOMIC_FAM	Quantidade de quartos na residência da família
QTD_DORM_FAM	Quantidade de dormitórios na residência da família
MAT_PISO_FAMILIA	Material utilizado no piso da residência da família
MATERIA_CONSTRUCAO	Material de construção da residência familiar
AGUA_ENCANADA	Se a água é encanada (sim / não)
ABASTECIMENTO	Origem da água da residência familiar
TEM_BANHEIRO	Se a casa tem banheiro (sim / não)
ESCOAMENTO_SANITARIO	Como é o escoamento sanitário
DESTINO_LIXO	Destino do lixo



Dicionário de dados



Variável	Descrição
ILUMINACAO	Forma de entrega de eletricidade e iluminação
CALCAMENTO	Calçamento
FAMILIA	Grupo familiar
CLASSE_CIDADE	Classe da cidade de residencia
QTD_PESSOA	Quantidade de pessoas na família
QTD_MASC	Quantidade de homens na família
QTD_FEM	Quantidade de mulheres na família
QTD_MENOR18	Quantidade de menores na família
TEM_DEFICIENTE	Se há deficiente na família (sim / não)
TEM_ANALFABETO2	Se há analfabeto na família (sim / não)
MENOR_ESTUDA	Se os menores na família estudam
REND_APROV_TRAB_REG	Renda proveniente de trabalho registrado
REND_12MESES_REG	Soma de renda de trabalho registrado
REND_APOSENTADORIA	Soma da aposentadoria familiar



Dicionário de dados



Variável	Descrição
RENDAS_DIVERSAS	Soma das rendas diversas (difere de trabalho registrado)
PERIODO_MAX_TRAB_MEMB	O periodo máximo que alguém da família trabalhou nos últimos doze meses ao cadastro
VL_MENOR	Valores recebidos por trabalho de menores
MEDIA_IDADE	Média de idade familiar
TM_BF	Se a família é beneficiária do bolsa 0 - Não, é beneficiário 1 – Sim, é beneficiário
COR	Raça/Cor declarada na família
ENSINO	Maior grau de ensino da família



9 - Conclusões e sugestões para o futuro



CONCLUSÕES

Observando os modelos desenvolvidos, é sugerido o uso do **gradient boosting** para que o poder público e as instituições independentes a ele, que desejem prever o valor médio familiar possam usar em suas rotinas. Este modelo apresentou a menor taxa de erro dentre os demais, o que para o objetivo deste trabalho é necessário tentar minimizar.



Já com relação a população, podemos observar a partir da análise dos dados, que a questão de saneamento e rendas precisam ser acompanhadas de perto. Programas de ensino e incentivo a escola para adultos precisam de atuação o mais breve, o número de famílias com nenhuma escolaridade nessa região é relevante.

SUGESTÕES

Os próximos analistas poderão fazer a análise desses dados, ou até mesmo de outro conjunto de dados com essas variáveis, usando métodos de clusterização.

Também pode se aprofundar em mesclar redes neurais e Random Forest ou Gradient Boosting para verificar se terá melhores resultados



OBRIGADO

