Estudo dirigido: Tratamento de dados

João Vitor[2], Lucas Lima[2], Paulo Mascarenhas[1], Romildo Andrade[1] e Tiago Avelino[1].
Engenharia Elétrica[1], Ciência da Computação[2]
MAT236-T07
Professora: Renata Esquivel

Resumo: O trabalho a seguir, visa iniciar aos alunos de Métodos Estatísticos o tratamento inicial de dados utilizando *softwares* para mais rápido manejo das informações. A priori, é disponibilizada uma base de dados que relaciona variáveis de diferentes naturezas, como quantitativas discretas ou contínuas, qualitativas nominais ou ordinais. Diante disso, é feito primeiramente uma classificação das variáveis a serem tratadas, posteriormente, uma vez estabelecido um intervalo de confiança, podemos estimar intervalos numéricos para média e proporção das respectivas variáveis. Dessa forma, é feita ainda uma análise exploratória dos dados, detectando desvios padrões amostrais, pontos "fora da curva", mediana, etc. Assim, as consequentes interpretações surgem de maneira mais natural e facilitada, concluindo o objetivo do trabalho de introduzir ao aluno o ambiente computacional para análise de dados, em especial na linguagem R, através da IDE RStudio.

1. Classificação das variáveis de análise

Primeiramente, foi fornecido para nós um conjunto de dados que, basicamente, relaciona A seguir, segue-se uma imagem das linhas iniciais de base de dados a ser trabalhada:

Tabela 1: Linhas iniciais da base de dados.

Família	Bairro	Programa Social	Instrução Chefe	Nº Moradores	Renda Familiar (sm)	Idade Chefe	Estado Civil
1	Monte Verde	Não usa	Médio	4	10.3	26	Solteiro
2	Monte Verde	Não usa	Médio	4	15.4	32	Casado
3	Monte Verde	Usa	Fundamental	4	9.6	36	Casado

4	Monte Verde	Não usa	Fundamental	5	5.5	25	Solteiro
5	Monte Verde	Usa	Médio	4	9	41	Solteiro
6	Monte Verde	Usa	Nenhum	1	2.4	28	Casado
7	Monte Verde	Não usa	Médio	2	4.1	41	Solteiro
8	Monte Verde	Usa	Médio	3	8.4	43	Solteiro
9	Monte Verde	Usa	Médio	6	10.3	34	Casado
10	Monte Verde	Usa	Fundamental	4	4.6	29	Solteiro
11	Monte Verde	Não usa	Fundamental	6	18.6	34	Casado
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

FONTE: Base disponibilizada pela Professora.

A fim de simplificar a visualização da base de dados, fizemos uma tabela mais objetiva, somando as variáveis quantitativas.

Tabela 2: Representação objetiva da base de dados.

Bairro	Programa Social	Contagem Família	Soma Renda familiar (sm)	Soma Idade Chefe
Encosta do Morro	Não Usa	12	86.7	750
	Usa	25	99.1	1518
Encosta do Morro TOTAL		37	185.8	2268
Monte Verde	Não Usa	18	170.3	666
	Usa	22	153.5	820
Monte Verde TOTAL		40	323.8	1486
Parque da Figueira	Não Usa	12	64.8	509
	Usa	31	180.1	1734
Parque da Figueira TOTAL		43	244.9	2243

TOTAL GERAL	120	754.5	5997

Fonte: Própria.

Notamos que as variáveis a serem classificadas são: Família (que possui um número inteiro para endereçar de quais famílias os dados se referem), Programa social (que afere se a família em questão desfruta ou não de um programa social) , Instrução chefe (variável que averigua o grau de instrução do líder da família), número de moradores (variável que ilustra o tamanho da família), renda familiar (mede a renda total da família em salários mínimos), idade chefe (afere a idade do líder da família) e estado civil do chefe (averigua se o líder da família está casado, solteiro ou viúvo). Dessa forma, temos as seguintes relações:

Tabela 3: Variáveis e suas respectivas naturezas.

VARIÁVEL	NATUREZA
Família	Quantitativa discreta
Bairro	Qualitativa nominal
Programa social	Qualitativa nominal
Instrução chefe	Qualitativa ordinal
Número de moradores	Quantitativa discreta
Renda familiar	Quantitativa contínua
Idade chefe	Quantitativa contínua
Estado civil chefe	Qualitativa nominal

Fonte: Própria.

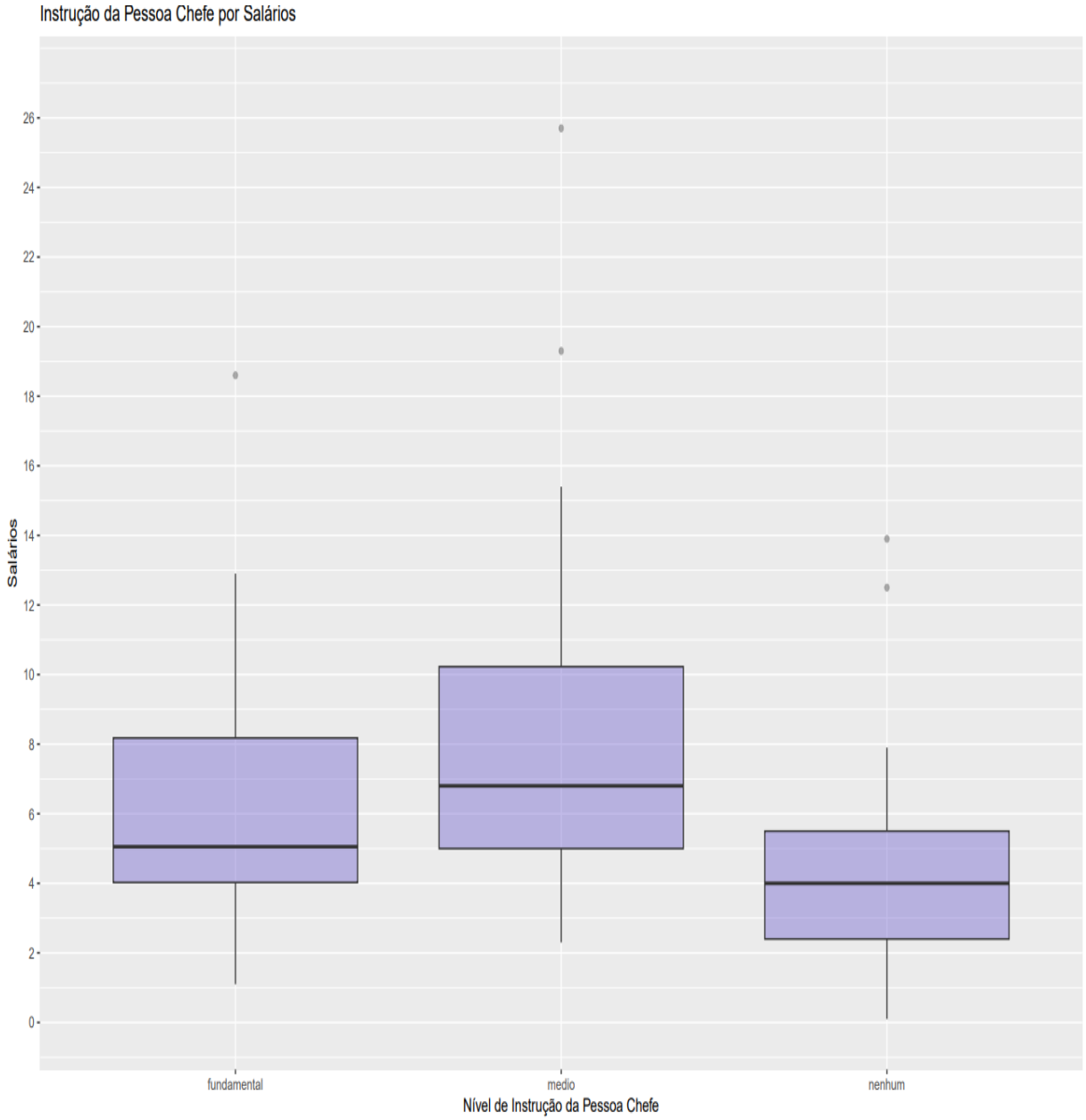
2. Análise exploratória de dados

Para análise exploratória de dados, estudaremos a relação entre variáveis duas a duas, para verificarmos linhas de tendência e validar hipóteses. A priori algumas hipóteses imediatas podem ser feitas, tais como:

1: Quanto maior o grau de instrução do chefe, maior a renda familiar;
2: Uma família de casados tende a ter uma renda maior do que a de um solteiro;
3: Famílias maiores, tendem a ter grau de instrução mais baixo;
4: Bairros mais nobres, comportam famílias com maior renda.

Para a primeira hipótese temos temos o seguinte gráfico que em que os pontos apontam a renda para determinado grau de instrução:

Figura 1: Pontos de renda para grau de instrução.

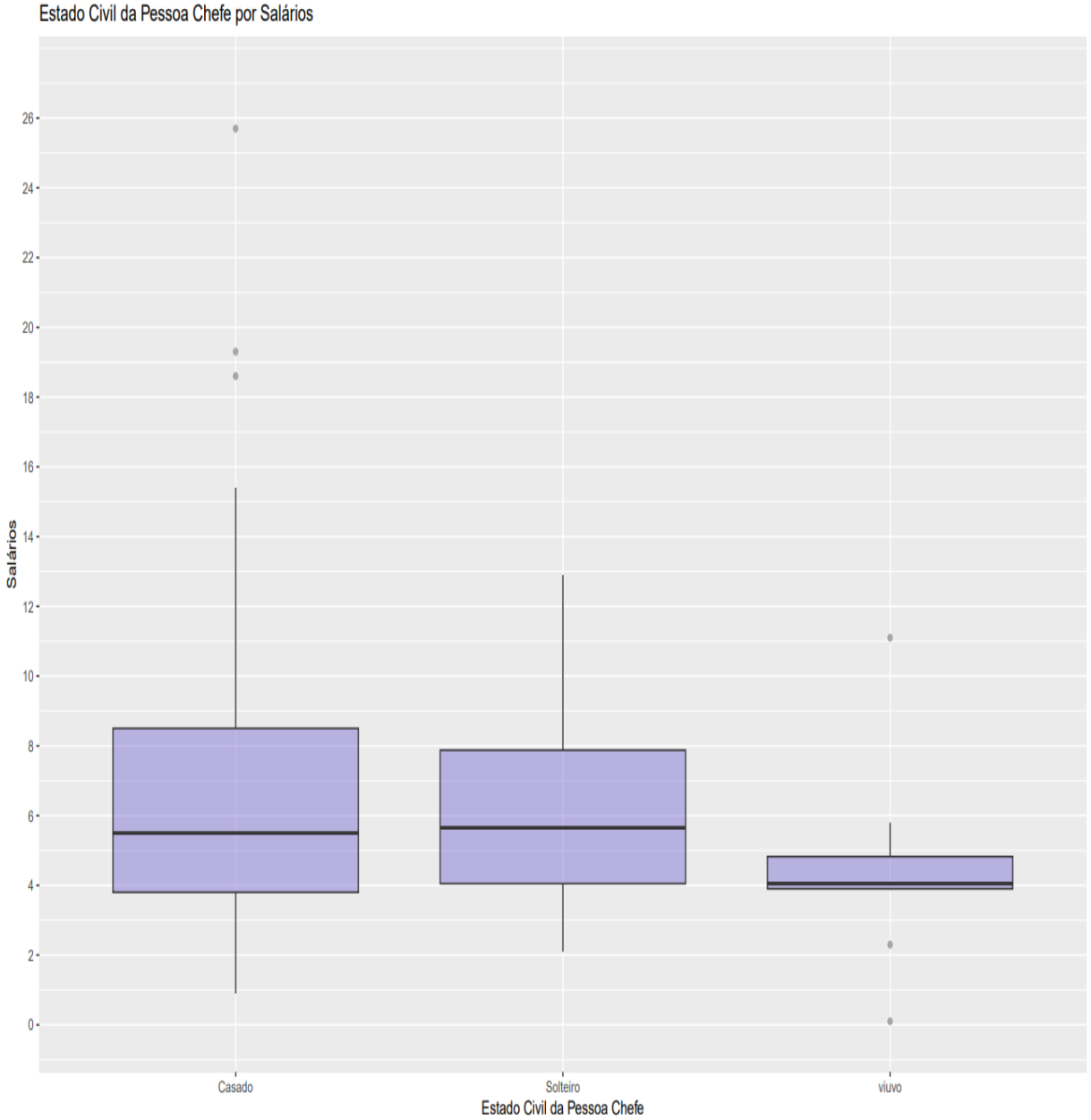


FONTE: Própria.

Como esperado, as rendas mais baixas se concentram em famílias com menor grau de instrução. Porém nota-se também que famílias com grau fundamental e médio possuem distribuição similar, o que pode apontar desvalorização dos empregos.

Para a segunda hipótese o seguinte gráfico:

Figura 2: Renda para estado civil.

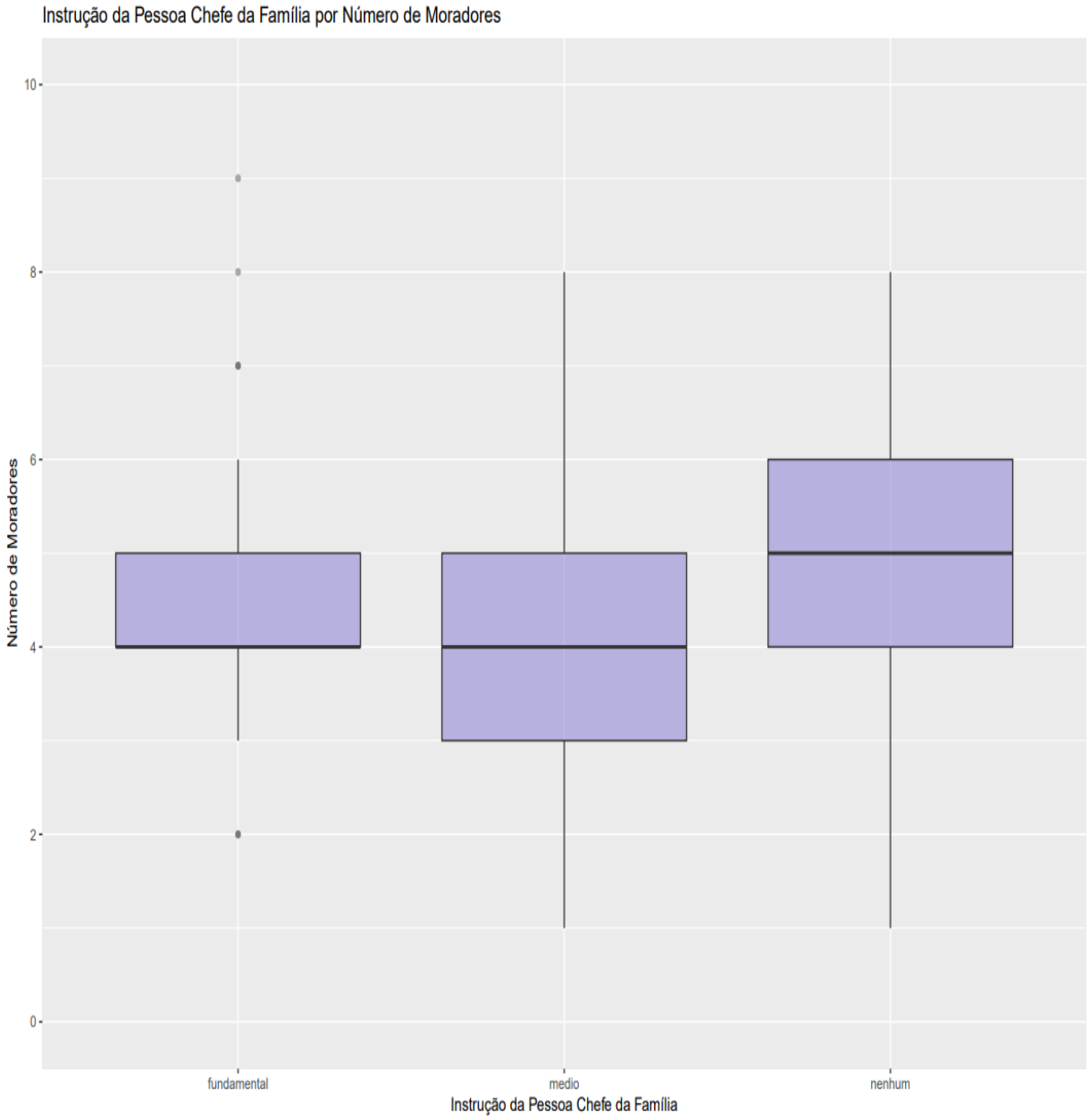


FONTE: Própria.

De fato, ocorre um maior número de famílias com maior renda para o estado civil casado. Porém percebe-se ainda que há um número muito grande de estado civil casado em relação aos demais estados, o que pode flexibilizar a força dessa hipótese.

Da terceira hipótese temos o seguinte gráfico:

Figura 3: Número de moradores por escolaridade



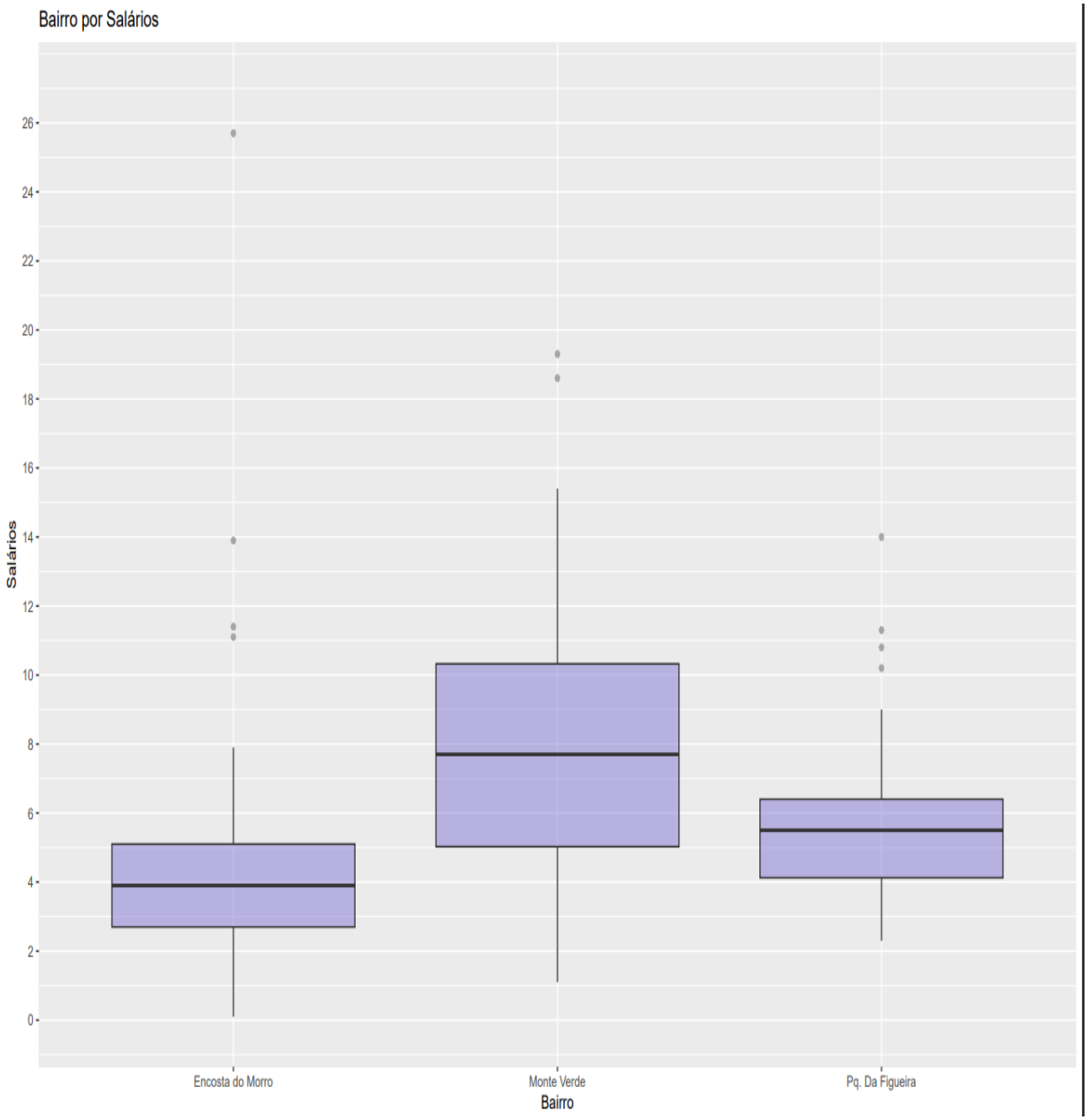
FONTE: Própria

É importante apontar que na primeira caixa - da instrução “Fundamental” - o primeiro quartil está bem próximo do segundo quartil. Devido ao zoom do gráfico, é difícil de visualizar a diferença entre esses quartis.

Conforme esperado, famílias muito grandes apresentam escolaridade mais baixa. Isto está intrinsecamente relacionado ao baixo grau de instrução e apoio que essas famílias recebem acerca de métodos contraceptivos e planejamento familiar.

Da quarta hipótese, temos o seguinte gráfico:

Figura 4: Renda por bairro.



FONTE: Própria.

Conforme esperado, os bairros mais valorizados como Monte Verde e Parque da Figueira apresentam concentrações de renda mais alta.

Para facilitar a visualização e interpretação das características do conjunto de dados, montamos uma tabela com aspectos fundamentais para análise das variáveis quantitativas. Ademais, é importante ressaltar que para as variáveis qualitativas, podemos apenas inferir a

moda de cada uma: Bairro (“Parque da Figueira”), Programa Social (“Usa”), Instrução (“Médio”), Estado Civil (“Casado”).

Tabela 4: Análise exploratória das variáveis quantitativas.

		Nº Moradores	Renda Familiar (sm)	Idade Chefe
Média		4.4917	6.2875	49.975
Mediana		4	5.35	50
Moda		4	3.9	26
Máximo		9	25.7	79
Mínimo		1	0	25
Desvio Padrão		1.5284	4.0568	17.0118
Quartis	**Q1**	3	3.9	34
	Q2	4	5.3	50
	Q3	5	7.7	65
Amplitude Interquartílica		2	3.8	31
Limite Superior		8	13.4	111.5
Limite Inferior		0	-1.8	-12.5

Fonte: Própria.

3. Estimar Parâmetros

Nesta etapa, uma vez obtidos os dados e verificadas as hipóteses, é necessário estimar intervalos de confiança para avançar nas análises estatísticas. Para este problema, utilizaremos um intervalo de confiança de 97%, que corresponde a um Z crítico de 2.17 na distribuição normal (como a base de dados é grande o suficiente, podemos aproximar o desvio amostral para o desvio populacional e não ser necessário utilizar a distribuição t-student, pois $n > 30$). A seguir, realizaremos o cálculo para intervalo de confiança da média e da proporção das variáveis estudadas. Adiante, é realizado o equacionamento para o intervalo da média (1) e para o intervalo da proporção (2):

$$I.C = \bar{x} \pm \frac{Z_c \sigma}{\sqrt{n}} \quad (1)$$

$$I.C = \hat{p} \pm \frac{Z_c \sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \quad (2)$$

Onde “I.C.” é o intervalo de confiança, “$\bar{x}$” é a média amostral, “σ” é o desvio padrão, “n” é o tamanho da amostra, “Z_C” é o parâmetro Z da normal e “$\hat{p}$” é a proporção.

Utilizando os parâmetros obtidos, registrados na tabela 4, iremos calcular os intervalos de confiança para 97% de confiabilidade, para cada uma das variáveis quantitativas. Para essa análise, abordaremos proporções condizentes com a natureza da base de dados e do problema. Para isso, buscaremos a proporção estimada das famílias que possuem de 3 a 4 integrantes, (51 famílias), renda familiar de 1 a 3 salários mínimos (19 famílias) e para chefes de família que possuem idade de 27 a 29 anos (8 famílias). Assim, temos:

Tabela 5: Intervalos de confiança para as variáveis quantitativas.

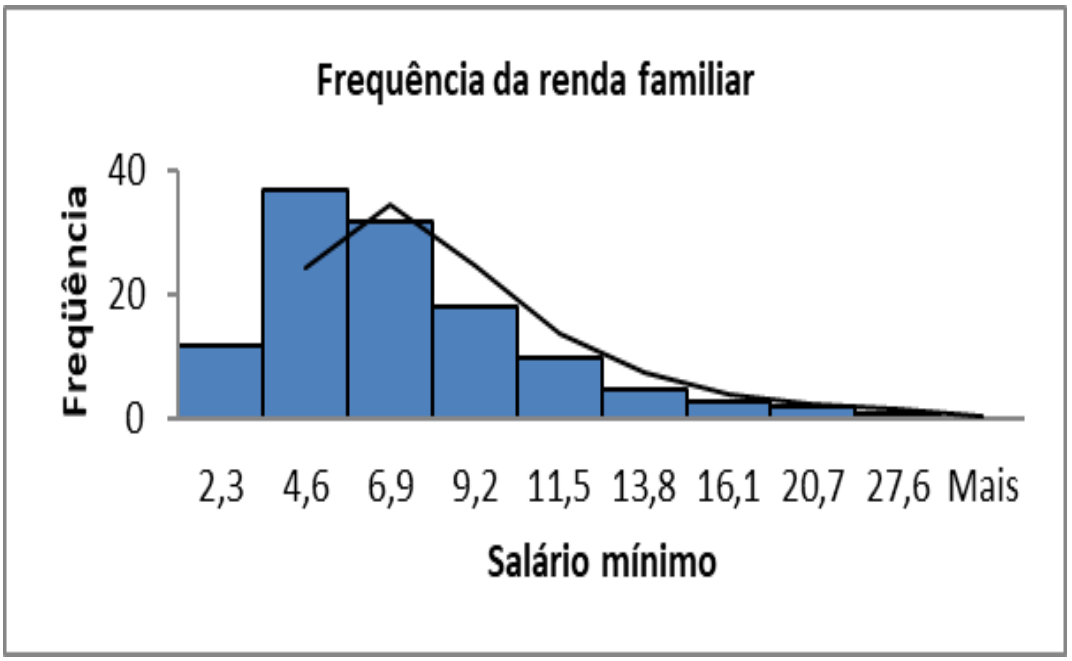
	Intervalo de Confiança	
	Média	**Proporção**
Nº de Moradores	[4.21; 4.77]	[0.327; 0.523]
Renda Familiar (sm)	[5.5; 7.06]	[0.086; 0.23]
Idade Chefe	[46.61; 53.33]	[0.017; 0.115]

FONTE: Própria.

4. Apresentar resultado das análises

A variável quantitativa contínua “renda familiar” é uma das principais variáveis de estudo na base de dados, sendo o objetivo analisar, por exemplo, uma renda familiar baixa, quantos moradores há na residência, a instrução do chefe de família, bairro da família, etc . A seguir, temos um histograma de frequência da renda familiar:

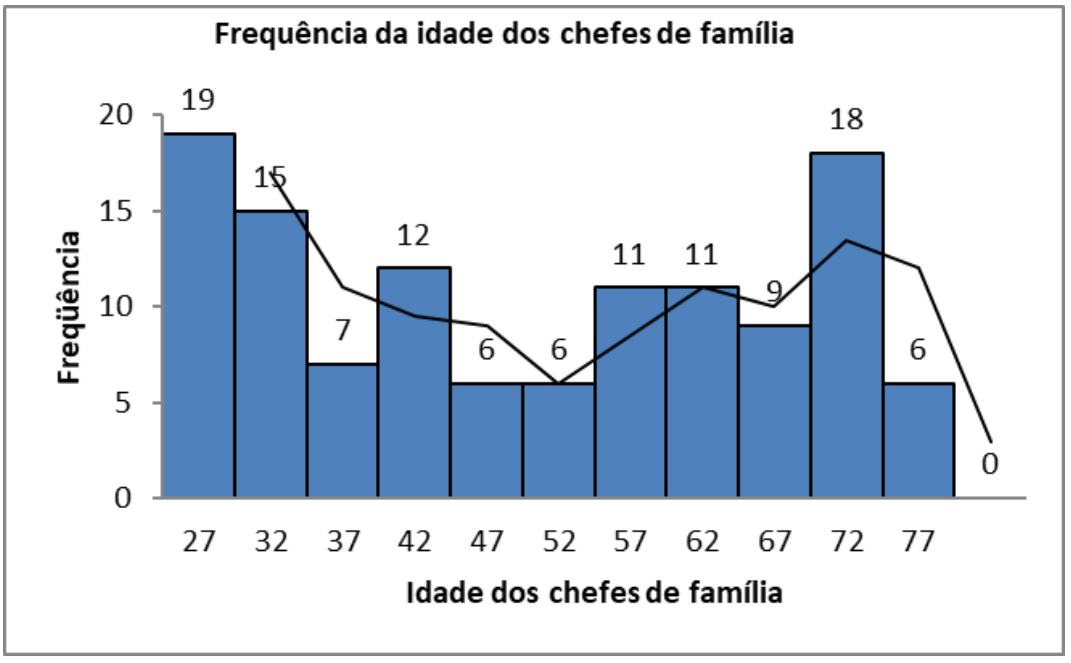
Figura 5: Histograma de frequência da renda familiar.



FONTE: Própria.

Primeiramente foi avaliado a renda familiar como uma variável unilateral. Fazendo a análise inicial a partir do gráfico é possível analisar uma distribuição assimétrica à direita da renda familiar. Das 120 famílias analisadas, conseguimos observar pelo gráfico histograma que a maioria das famílias ganham entre 4,6 e 9,2 salários mínimos. Foi realizada uma análise bivariada para entender quais as características dessas famílias. Agora, segue um histograma da frequência da idade dos(as) chefes de família:

Figura 6: Histograma da frequência da idade dos(as) chefes de família:



FONTE: Própria.

Com o histograma da frequência da idade dos(as) chefes de família, podemos verificar que há uma simetria da idade dos chefes de família, visto que a média é aproximadamente igual ao valor da mediana (Média da variável idade = 49.975; Mediana da Variável Idade = 50).

5. Auto avaliação

A seguir, temos a tabela de auto avaliação, requisitada pela professora. Nesta tabela, cada participante do grupo deverá avaliar a si mesmo e aos seus colegas de acordo com sua percepção da participação dos mesmos na concepção e realização do trabalho. Essa avaliação será feita numa escala de zero a cinco, e cada linha representa o conjunto de notas que um dos integrantes do grupo atribuiu para os demais.

Tabela 6: Tabela de auto avaliação.

	Avaliados				
Avaliadores	**Paulo**	**Tiago**	**Romildo**	**João**	**Lucas**
Paulo	10	10	10	10	10
Tiago	10	10	10	10	10
Romildo	10	10	10	10	10
João	10	10	10	10	10
Lucas	10	10	10	10	10

Fonte: Elaborada pelos autores.

6. Conclusão

A partir do exposto, é imperioso que etapas importantes numa análise estatística sejam feitas. Assim, foi feito tratamento dos dados, análise exploratória, levantamento de hipóteses, validação de hipóteses e seus devidos ajustes, gráficos para melhor interpretação e avaliação de medidas de posição e dispersão utilizando a linguagem R. Dessa maneira, é claro que o objetivo da atividade foi alcançado com êxito e a equipe saiu com um ferramental mais poderoso para os futuros problemas estatísticos abordados no futuro.