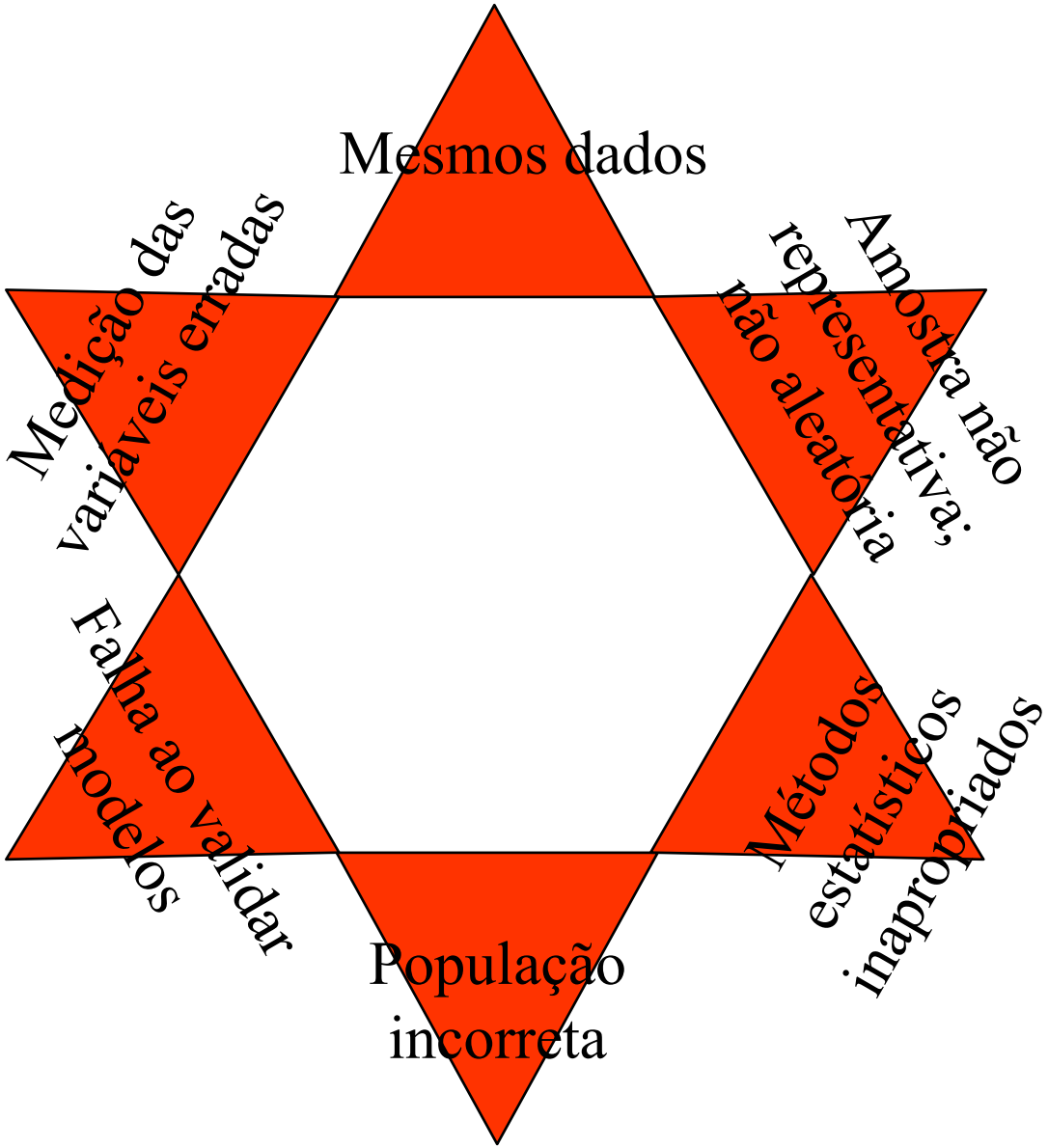
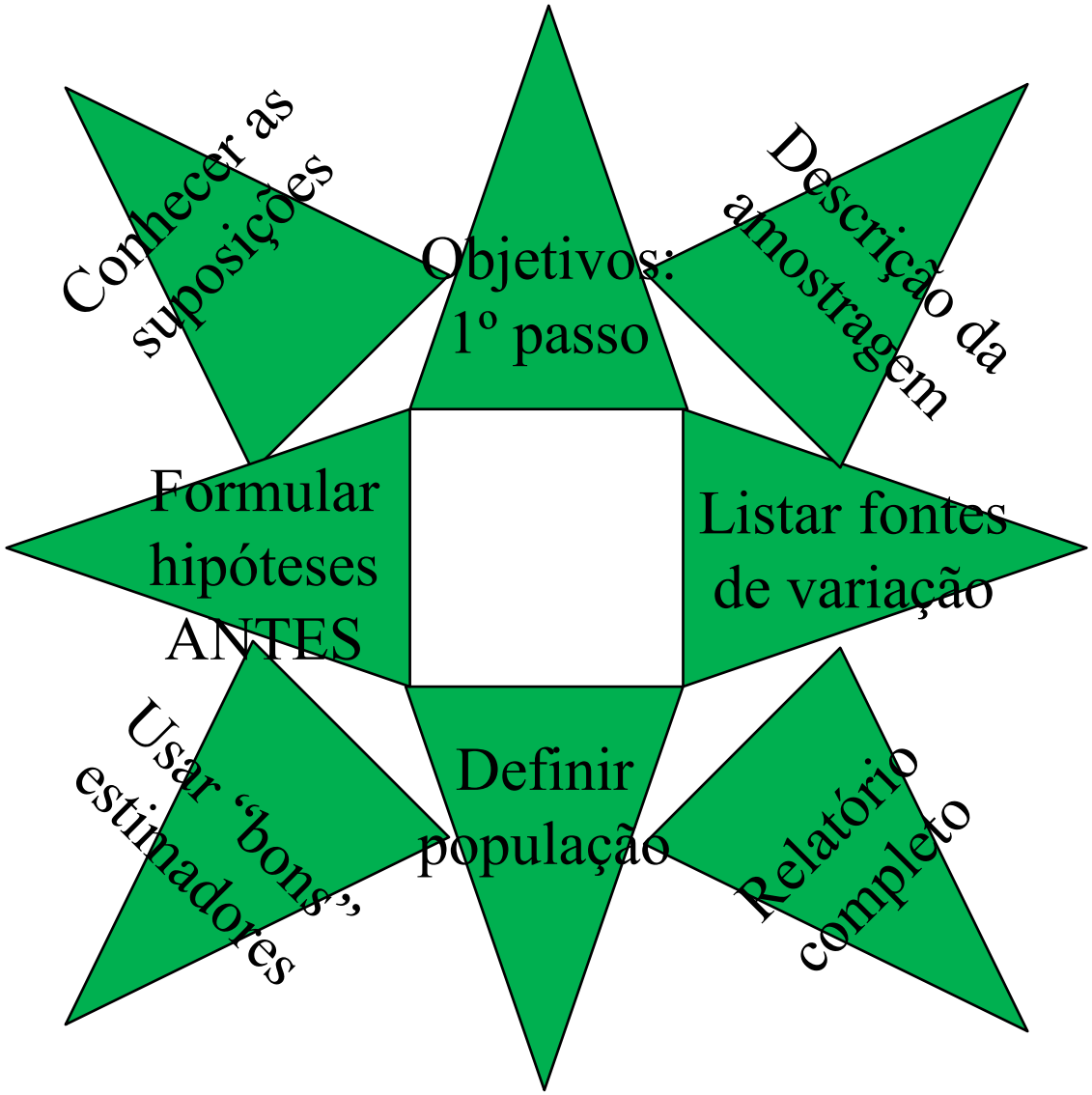
FONTES DE ERRO, HIPÓTESES, ANÁLISE EXPLORATÓRIA DE DADOS – 1ª PARTE

Fontes de erro



"Prescrição"


Conhecer as suposições
Objetivos: 1º passo
Descrição da amostragem
Formular hipóteses ANTES
Listar fontes de variação
Usar "bons" estimadores
Definir população
Relatório completo

Hipóteses de Pesquisa

- “Para homens acima de 40 anos com hipertensão crônica, uma dose diária de 100 mg desta nova droga reduzirá em 10 mm de mercúrio (em média) a pressão sanguínea diastólica “.
- “Para homens acima de 40 anos com hipertensão crônica, uma dose diária de 100 mg desta nova droga reduzirá em 10 mm de mercúrio (em média) a pressão sanguínea diastólica comparada a uma dose equivalente de metropolol“.
- “Esta nova variedade de tijolo refratário apresentará um ponto de fusão 200º C maior (em média) do que a variedade atualmente usada”.

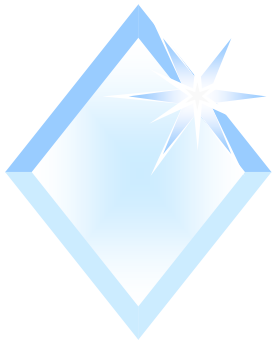
AED - Conceito

	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ
1	Sexo	Veículo	Equipamer	Seguro	Cor	Sinistro	Renda	Idade	Potência	Anos
2	Masculino	Automóvel	Alarme	Simples	Metálica	Hospitalare	5,470	34	88,00	7
3	Feminino	Automóvel	Alarme	Simples	Metálica	Materiais	5,080	33	123,00	21
4	Masculino	Automóvel	Alarme	Total	Vermelha	Hospitalare	5,290	31	100,00	5
5	Feminino	Automóvel	Alarme	Total	Sólida	Materiais	5,350	30	101,00	12
6	Masculino	Automóvel	Alarme	Simpl	Vermelha	Materiais	5,710	31	108,00	12
7	Feminino	Automóvel	Nenhum	Simples	Sólida	Materiais	5,050	34	93,00	12
8	Feminino	Automóvel	Nenhum	Simples	Metálica	Materiais	4,330		96,00	15
9	Masculino	Esportivo	Alarme	Terceiros	Metálica	Materiais	3,270	29	94,00	26
10	Masculino	Automóvel	Air-bag	Terceiros	Metálica	Materiais	3,300	31	119,00	14
11	Masculino	Automóvel	Alarme	Terceiros	Metálica	Hospitalare	5,950	30	92,00	11
12	Masculino	Automóvel	Alarme	Total	Metálica	Hospitalare	5,800	30	97,00	12
13	Masculino	Automóvel	Nenhum	Simples	Metálica	Terceiros	6,460	27	94,00	13
14	Feminino	Automóvel	Alarme	Terceiros	Metálica	Materiais	4,930	29	107,00	14
15	Feminino	Automóvel	Nenhum	Si	Metálica	Materiais	4,600	32	100,00	19
16	Feminino	Automóvel	Alarme	Terceiros	Perolizada	Hospitalare	6,820	30	106,00	14
17	Masculino	Automóvel	Alarme	Terceiros	Perolizada	Materiais	5,200	32	95,00	27
18	Masculino	Automóvel	Nenhum	Simples	Metálica	Materiais	3,630	30	104,00	13
19	Masculino	Automóvel	Nenhum	Simples	Metálica	Terceiros	7,350	25	78,00	17
20	Masculino	Esportivo	Alarme	Total	Vermelha	PerdaTotal	9,240	26	95,00	12
21	Feminino	Automóvel	Nenhum	Terceiros	Metálica	Materiais	4,930	32	87,00	21

Necessário resumi-los!

Necessário organizá-los!

Interpretação e tomada de decisões.



Objetivo

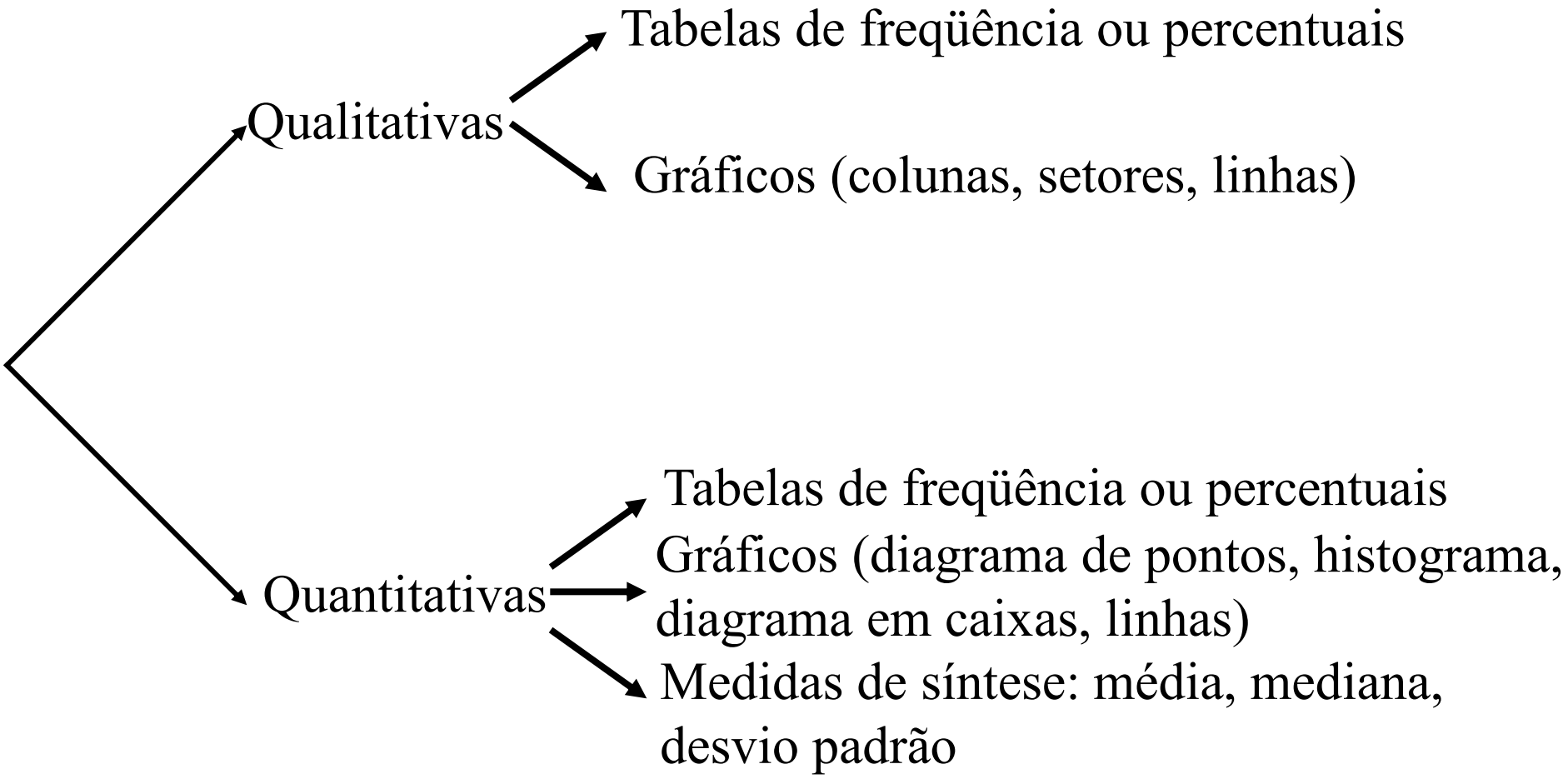
Estudar comportamento INDIVIDUAL das variáveis.

Estudar RELACIONAMENTO entre as variáveis.

Escolha das técnicas de AED

- Nível de mensuração das variáveis.
- Objetivo da análise:
  - Comportamento individual da variável.
  - Comportamento da variável em função de uma ou mais variáveis (ferramentas múltiplas).
- Número de variáveis envolvidas.
- Tamanho do conjunto de dados.
- Tempo disponível para a apresentação dos resultados.
- Grau de conhecimento estatístico do público alvo.

Nível de mensuração



Nível de mensuração

- Variáveis QUANTITATIVAS:
  - **Discretas** - lista finita (geralmente, números inteiros).
    - **Exemplo**: quantidade de máquinas ligadas.
  - **Contínuas** - infinitos resultados possíveis (um intervalo dos números reais).
    - **Exemplo**: tempo de resposta (em segundos).

Mensuração de variáveis

- Como medir satisfação com o trabalho?
  - classificar: "satisfeito" / "não satisfeito"
  - grau de satisfação: escala de 0 a 10
  - grau de satisfação: escala de 1 a 5 associada a adjetivos
  - grau de satisfação: escala construída com vários itens de um questionário

Mensuração de variáveis

- Como medir qualidade de um algoritmo?
  - Medir tempo de processamento (comparando com algoritmos existentes).
  - Registrar percentual de "acertos" (comparando com algoritmos existentes).
    - Como definir/medir "acertos".

Variáveis intervenientes

- Refletir sobre quais variáveis podem influenciar a variável de resposta.
- “Pressão arterial diastólica” pode ser influenciada por...
  - Sexo do paciente?
  - Idade do paciente?
  - Hábitos alimentares?
  - Hábitos de atividade física?
  - Outras condições médicas pré-existentes?

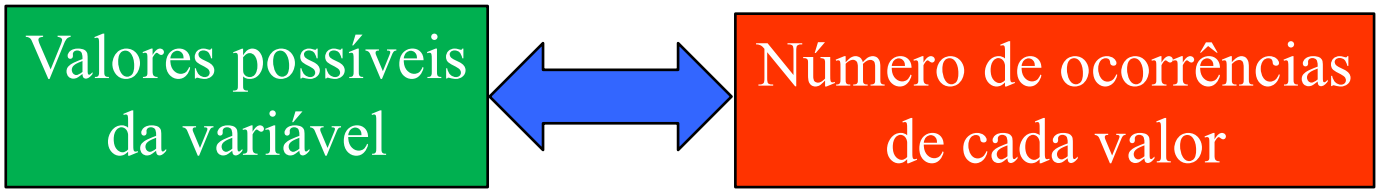
Pré-análise dos dados

- Dados perdidos: não foram registrados para um ou mais dos integrantes do conjunto.
  - Até 5% aceitável.
- Erros de registro: problemas de ortografia, digitação (facilmente identificáveis), valores discrepantes (quando resultante de erros).
- Inconsistências: sua identificação já faz parte da análise dos dados.
- Importante para **mineração de dados**.

Recodificação e Transformação

- Criar novas variáveis usando condições fixadas.
- Recodificação:
  - Qualitativa para qualitativa.
  - Quantitativa para qualitativa (categorização).
  - Quantitativa contínua para classes (agrupamento em classes)
- Transformação:
  - Quantitativa para quantitativa (operação matemática).

Distribuição de frequências



Distribuição de frequências - variáveis qualitativas

Tipo de seguro contratado em 5000 sinistros

Tipo de seguro	Frequência	%
	4	0,08%
Simples	2465	49,30%
Terceiros	1517	30,34%
Total	1014	20,28%
Total Geral	**5000**	**100,00%**

Fonte: hipotética

Gráfico de colunas

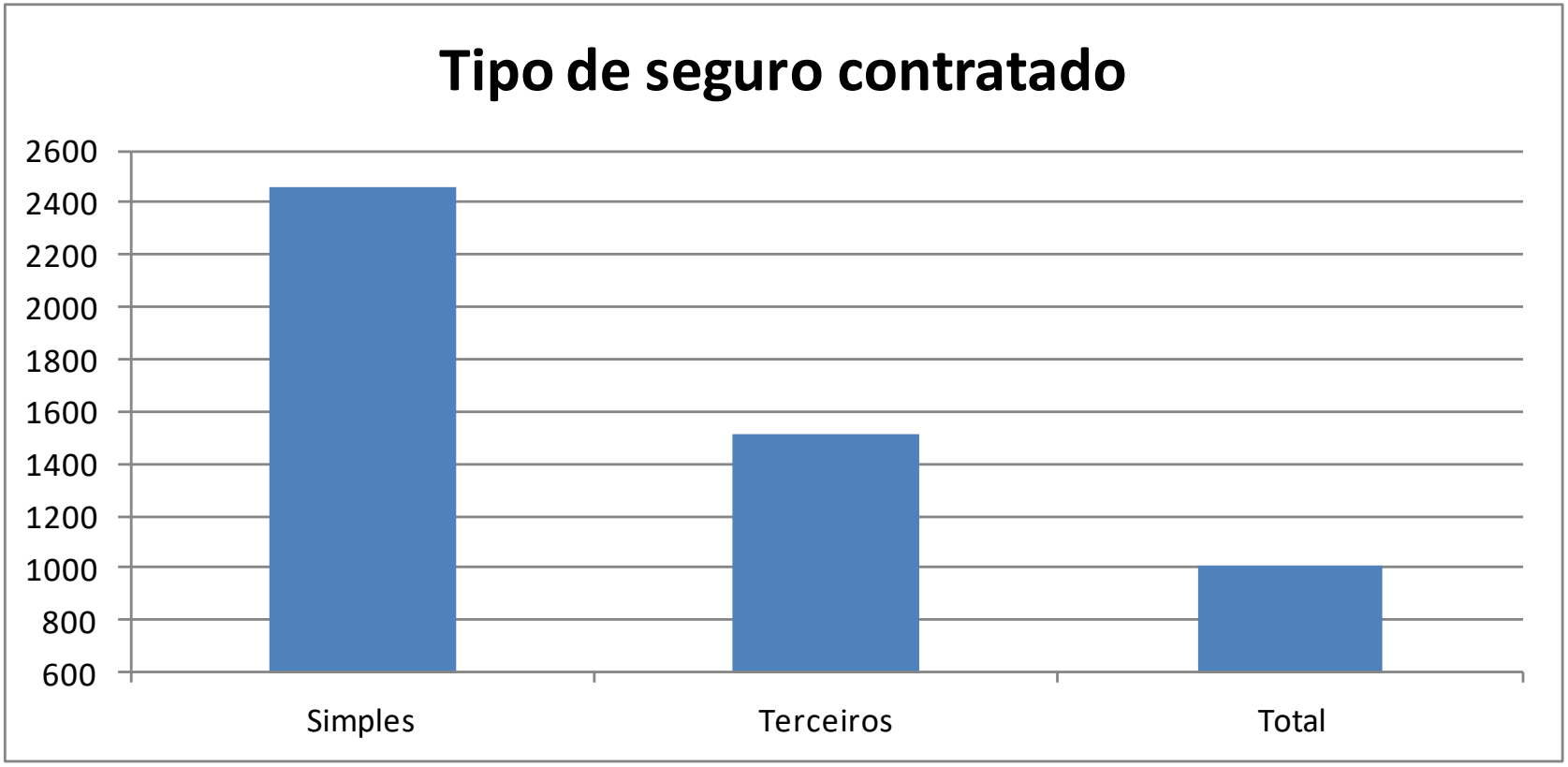


Gráfico de colunas

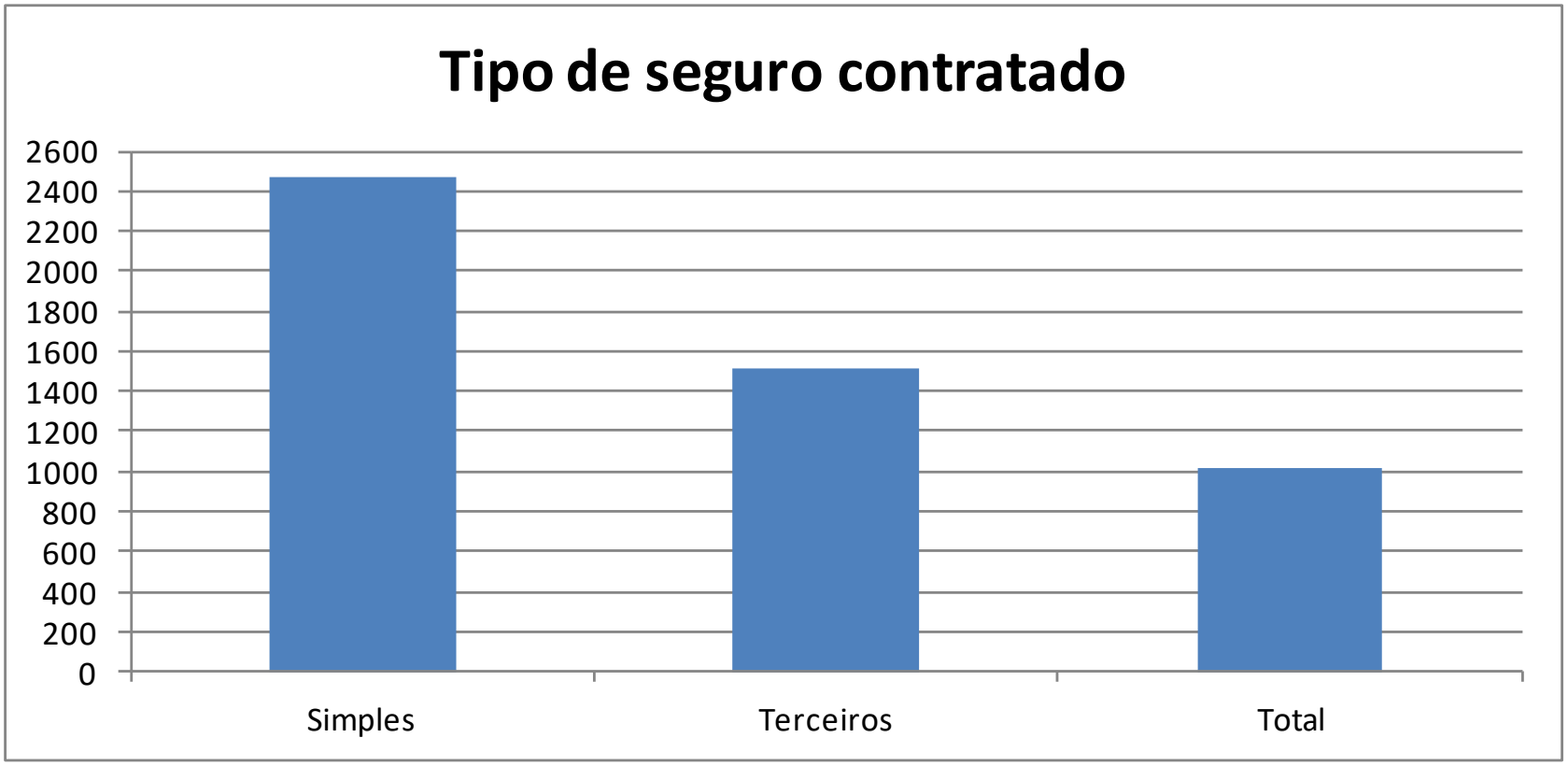
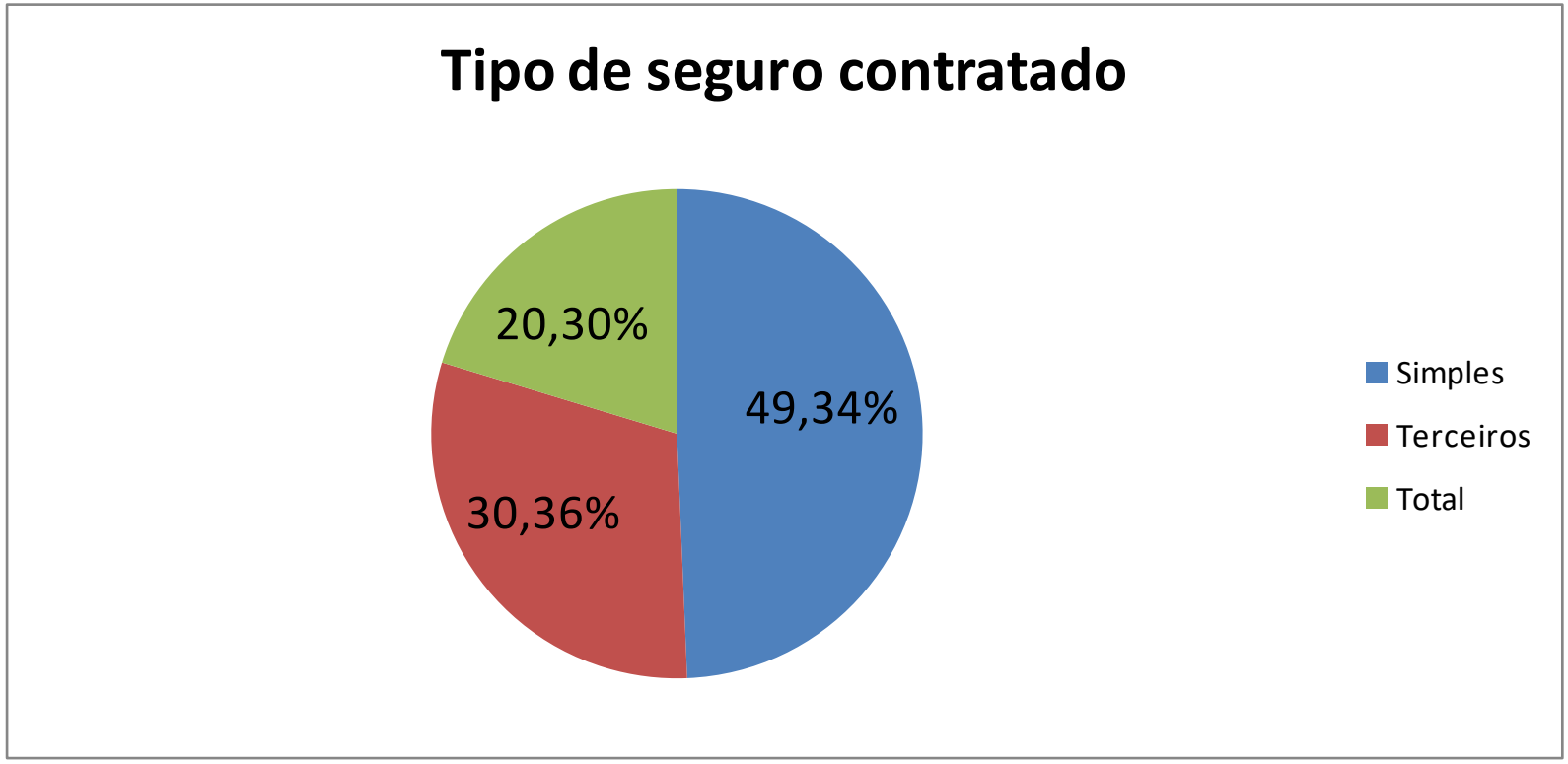


Gráfico em setores (circular ou pizza)



Distribuição de frequências - variáveis quantitativas

- Nível de mensuração da variável quantitativa:
  - DISCRETA: semelhante às variáveis qualitativas.
    - Tabela de frequências e histograma para dados não agrupados.
  - CONTÍNUA: necessário agrupar os dados para possibilitar o resumo do conjunto e melhor visualização.
    - Tabelas de frequências e histograma para dados agrupados, diagramas em caixa.

Tabela de frequências: dados não agrupados

Máquinas em uso	Frequência (absoluta)	Proporção
20	2	0,10 (10%)
21	4	0,20 (20%)
22	6	0,30 (30%)
23	5	0,25 (25%)
24	2	0,10 (10%)
25	0	0
26	1	0,05 (5%)
Total	20	1 (100%)

Histograma

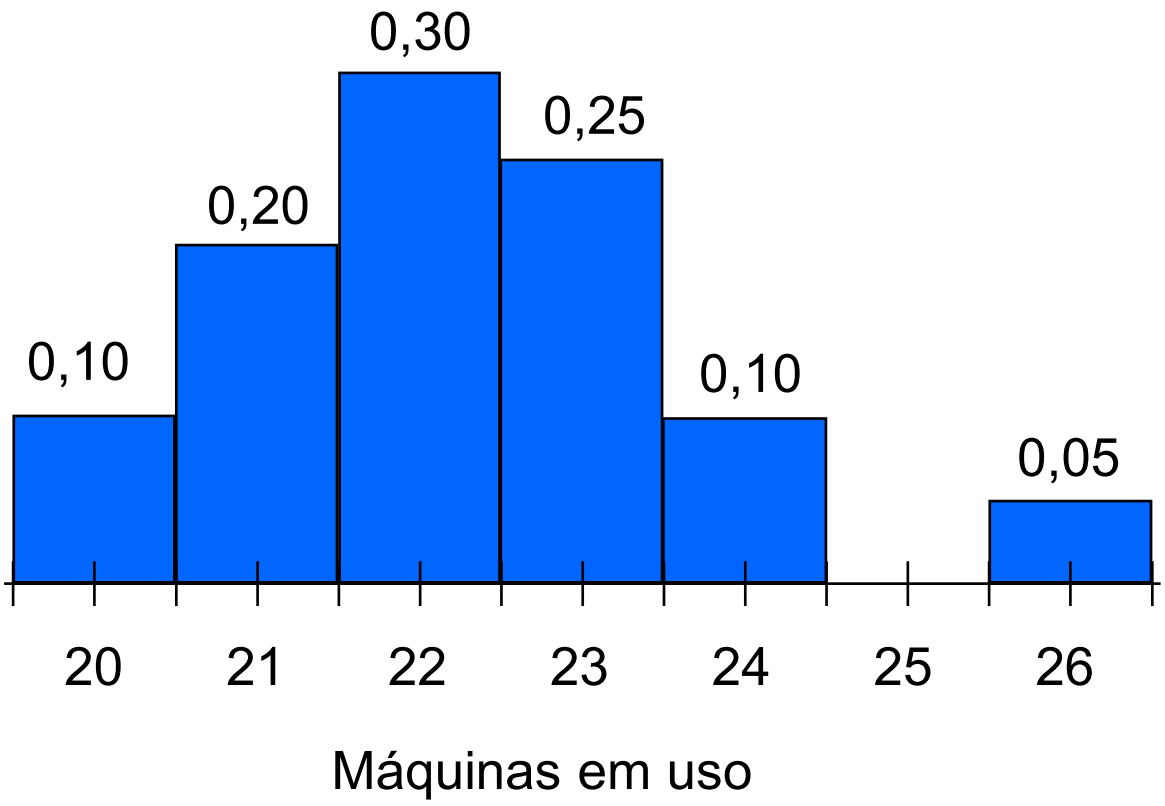


Tabela de frequências para dados agrupados

- Recomendável para grande conjuntos de variáveis QUANTITATIVAS.
- **PERDE-SE** informação sobre o conjunto original para obter sua compactação.

Tabela de frequências para dados agrupados

- Passos para construção:
  - Determinar o intervalo do conjunto.
  - Dividir o intervalo em k classes: $k = 5 \times \log_{10} n$ (para $n > 100$)
  - Obter limites das classes.
  - Contar frequências dentro das classes.
  - Renda de uma amostra de clientes de uma seguradora (5000 observações):
    - $k = 5 \times \log_{10} 5000 = 18{,}49485 \Rightarrow k = 18$
    - Mínimo = 3 salários mínimos; Máximo = 47,88 salários mínimos
    - Amplitude classes = (47,88 – 3)/2,49333 => 2,50

Renda dos clientes de uma seguradora (salários mínimos)

Limite Inferior	Limite superior	Frequência	%	Freq. Acumulada	% acumulado
3	5,50	**2723**	**54,50%**	2723	54,50%
5,50	8,00	1601	32,05%	4324	86,55%
8,00	10,50	318	6,37%	4642	92,91%
10,50	13,00	160	3,20%	4802	96,12%
13,00	15,50	82	1,64%	4884	97,76%
15,50	18,00	51	1,02%	4935	98,78%
18,00	20,50	28	0,56%	4963	99,34%
20,50	23,00	15	0,30%	4978	99,64%
23,00	25,50	11	0,22%	4989	99,86%
25,50	28,00	3	0,06%	4992	99,92%
28,00	30,50	2	0,04%	4994	99,96%
30,50	33,00	0	0,00%	4994	99,96%
33,00	35,50	1	0,02%	4995	99,98%
35,50	38,00	0	0,00%	4995	99,98%
38,00	40,50	0	0,00%	4995	99,98%
40,50	43,00	0	0,00%	4995	99,98%
43,00	45,50	0	0,00%	4995	99,98%
45,50	48,00	1	0,02%	4996	100%
Total		4996	100%	-	-

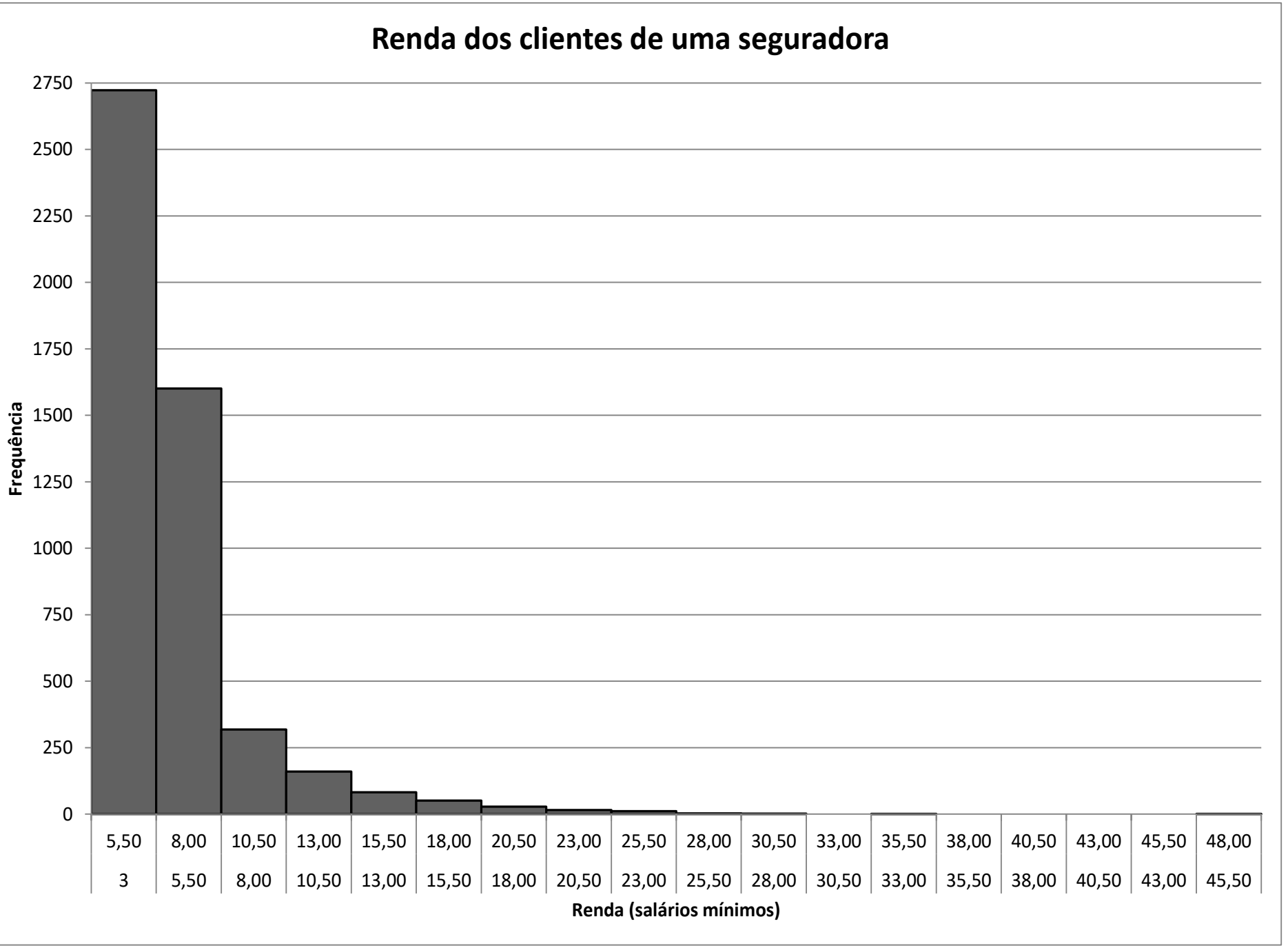

Renda dos clientes de uma seguradora
Frequência
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
5,50 8,00 10,50 13,00 15,50 18,00 20,50 23,00 25,50 28,00 30,50 33,00 35,50 38,00 40,50 43,00 45,50 48,00
3 5,50 8,00 10,50 13,00 15,50 18,00 20,50 23,00 25,50 28,00 30,50 33,00 35,50 38,00 40,50 43,00 45,50
Renda (salários mínimos)

Diagrama de pontos

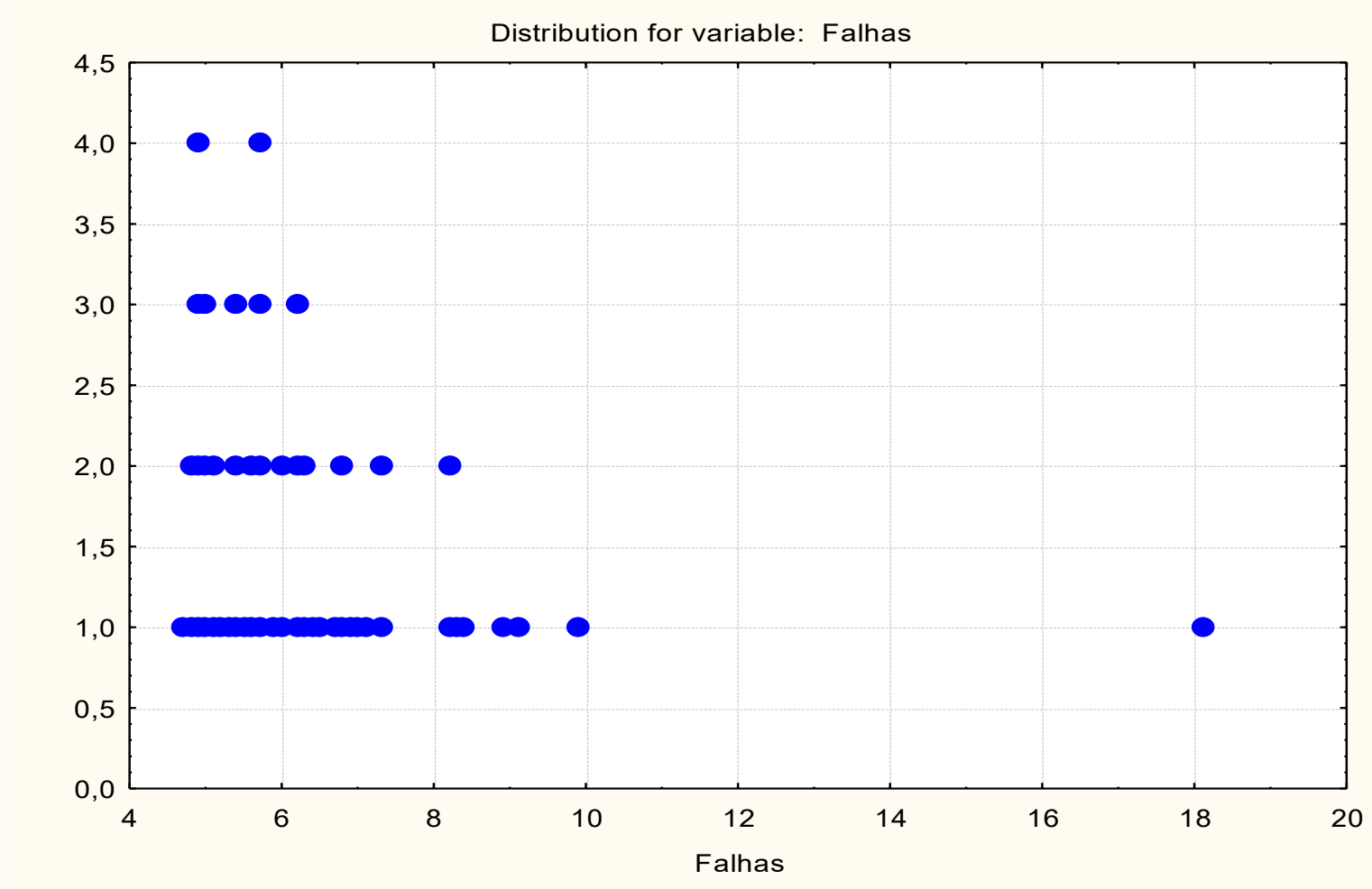


Gráfico de linhas

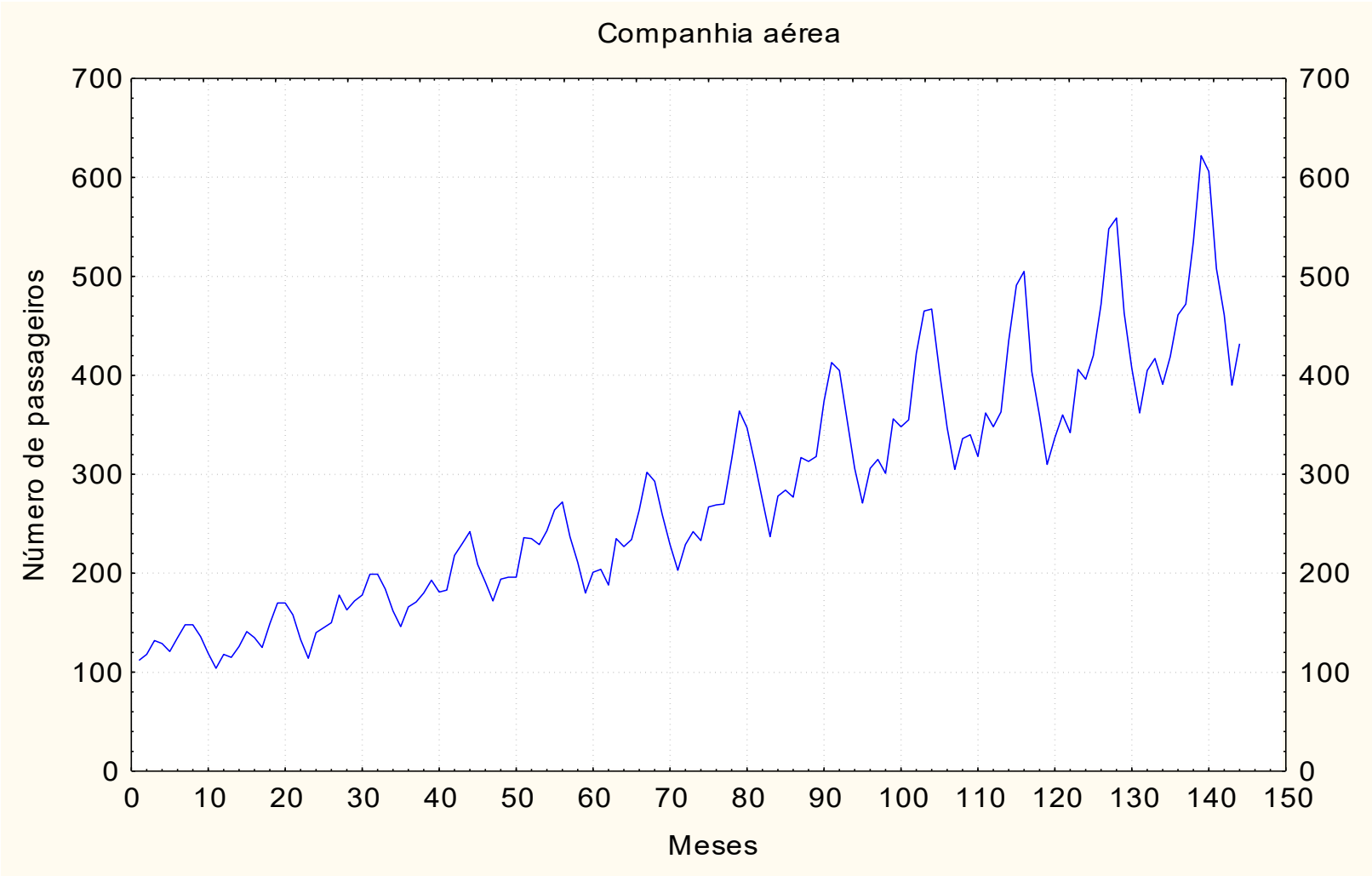
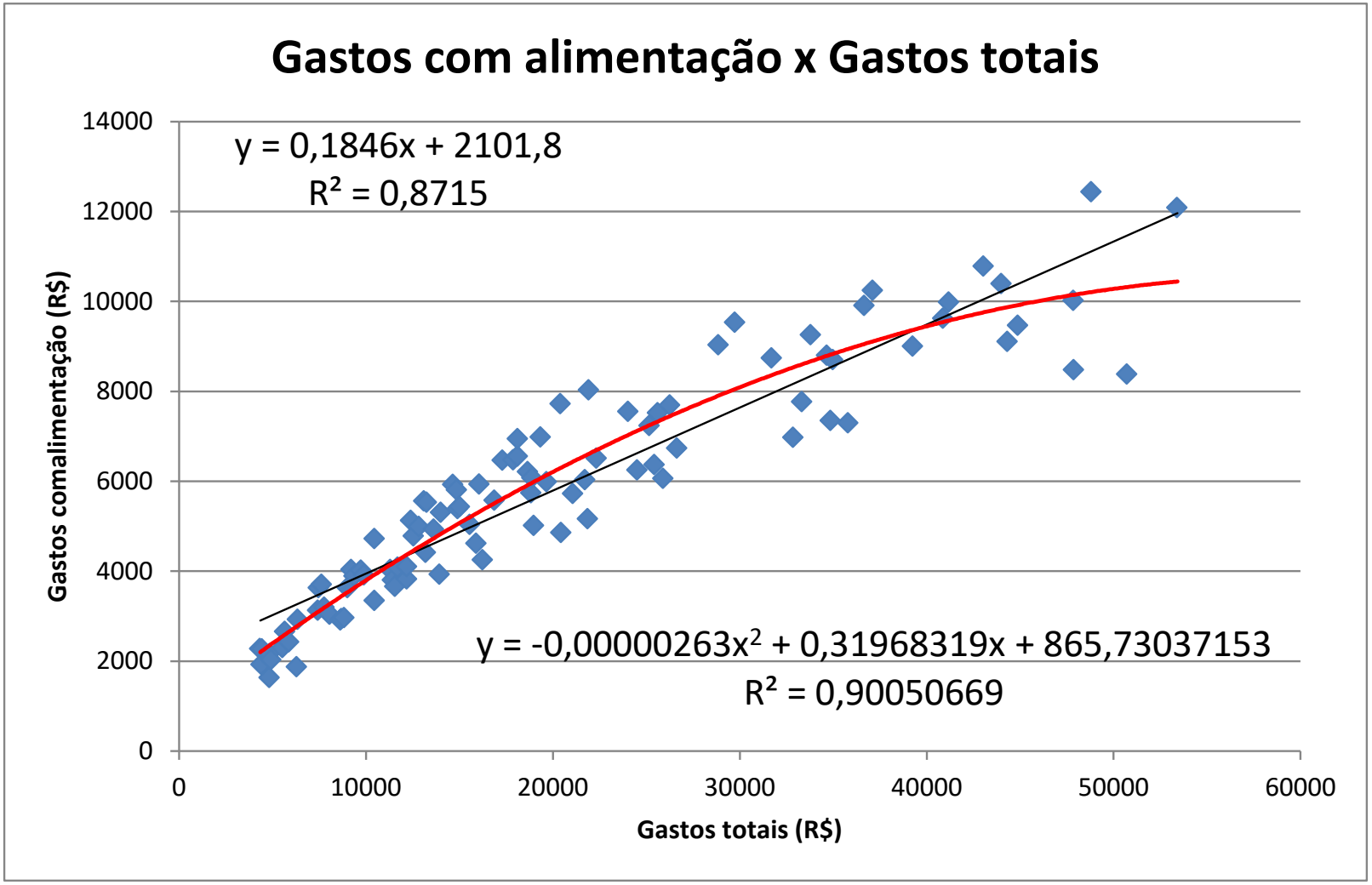


Diagrama de Dispersão



Distribuição de frequência múltipla

Tabulação Cruzada

Dupla Classificação

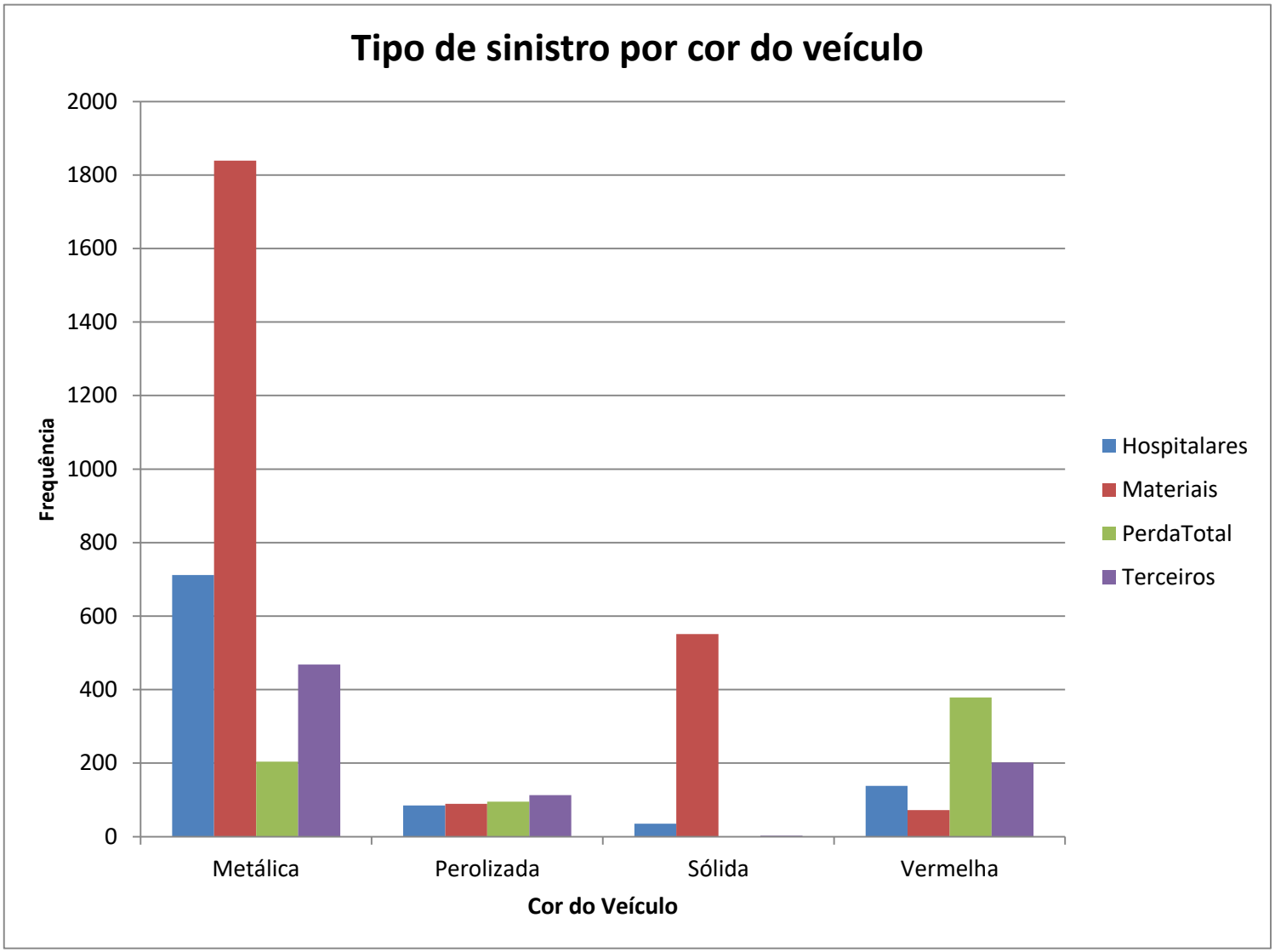
Tabela de Contingências

	Valores variável 2
Valores variável 1	Frequências cruzamentos

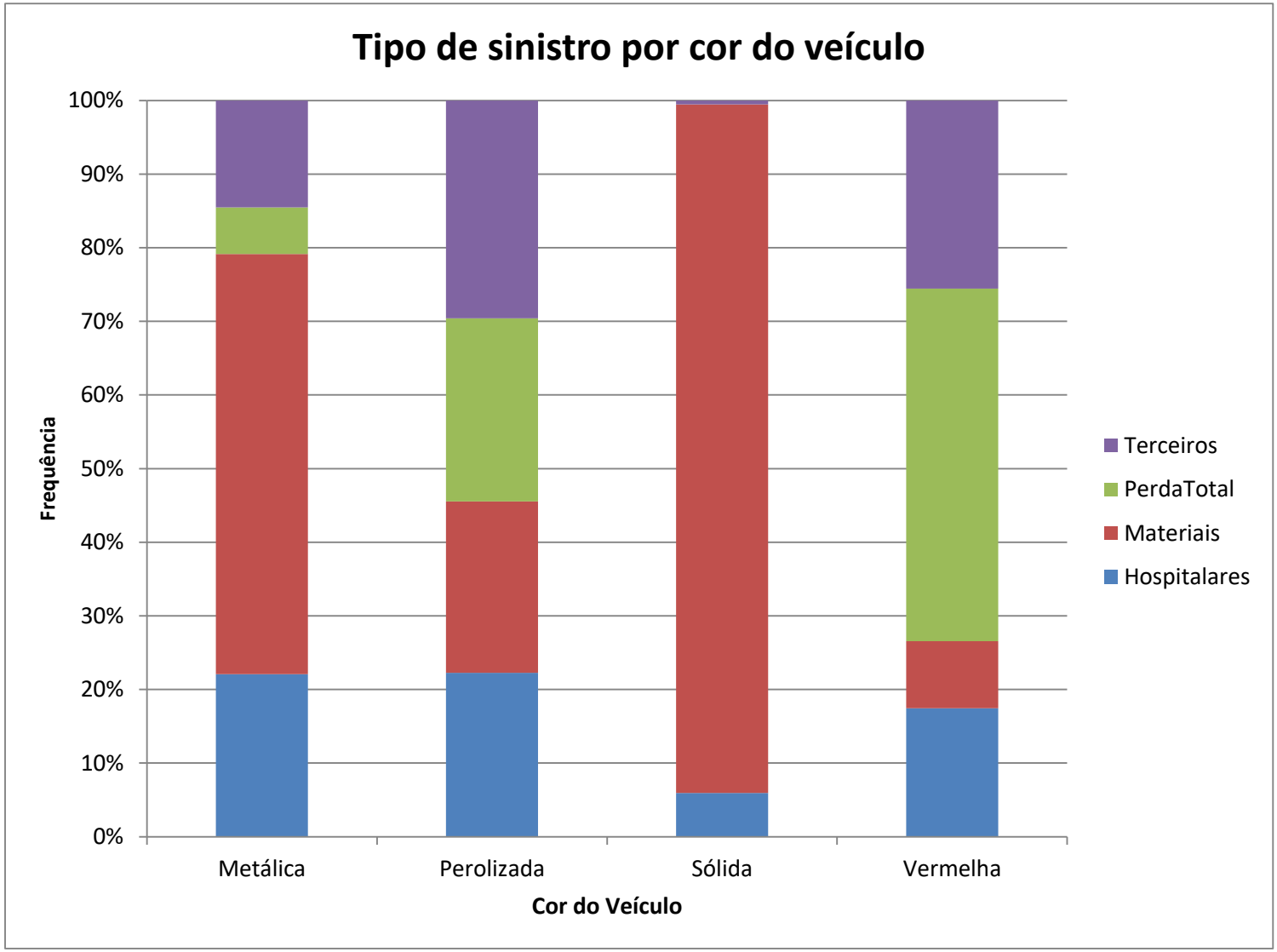
Tabela de contingências

	Sinistro				
Cor do veículo	**Hospitalares**	**Materiais**	**PerdaTotal**	**Terceiros**	**Total Geral**
Metálica					
Frequência	712	1839	204	468	3223
% linha	22,09%	57,06%	6,33%	14,52%	100,00%
% coluna	73,40%	72,09%	30,09%	59,54%	64,65%
Perolizada					
Frequência	85	89	95	113	382
% linha	22,25%	23,30%	24,87%	29,58%	100,00%
% coluna	8,76%	3,49%	14,01%	14,38%	7,66%
Sólida					
Frequência	35	551		3	589
% linha	5,94%	93,55%	0,00%	0,51%	100,00%
% coluna	3,61%	21,60%	0,00%	0,38%	11,82%
Vermelha					
Frequência	138	72	379	202	791
% linha	17,45%	9,10%	47,91%	25,54%	100,00%
% coluna	14,23%	2,82%	55,90%	25,70%	15,87%
Total Frequência	**970**	**2551**	**678**	**786**	**4985**
Total % linha	**19,46%**	**51,17%**	**13,60%**	**15,77%**	**100,00%**
Total % coluna	**100,00%**	**100,00%**	**100,00%**	**100,00%**	**100,00%**

Apresentação gráfica



Apresentação gráfica



Apresentação gráfica

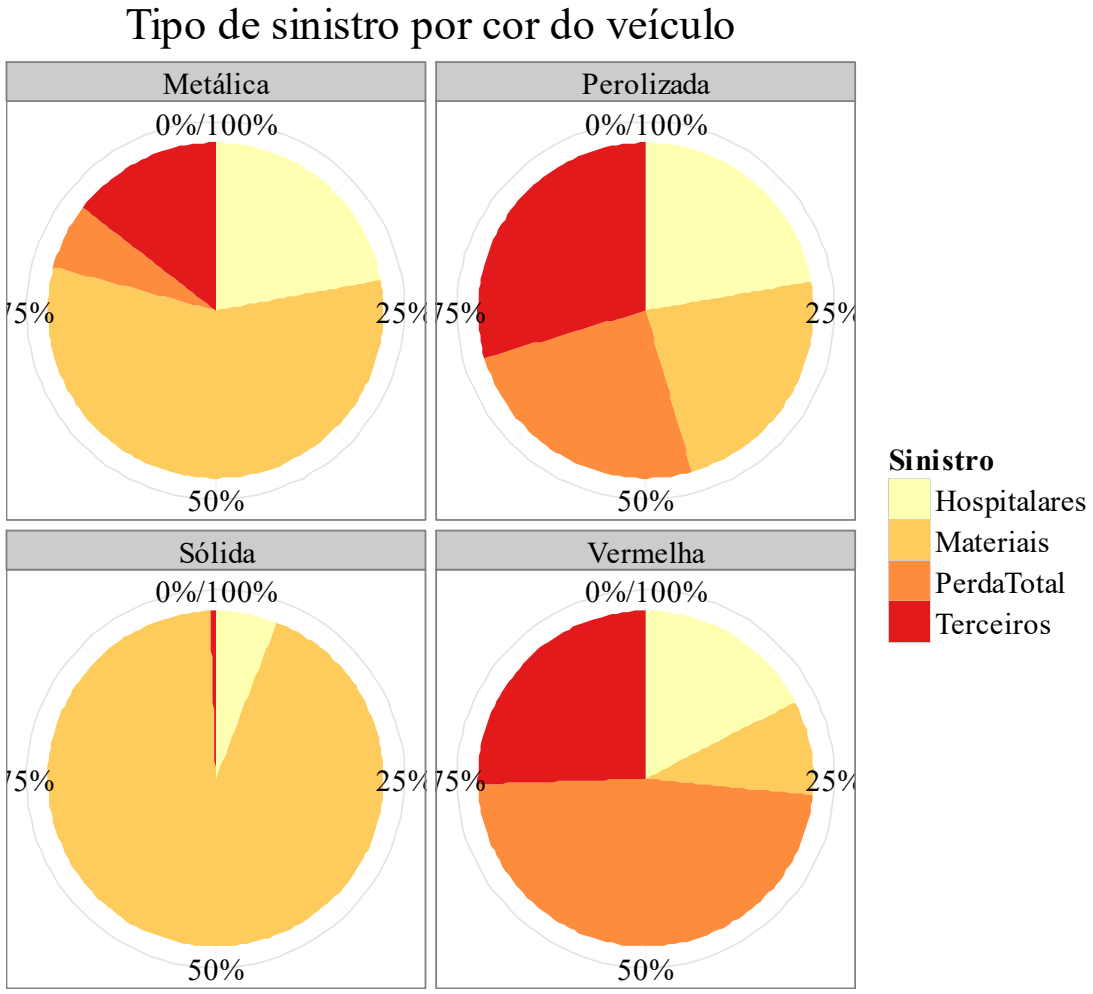


Tabela com 3 variáveis

Sexo	Feminino				
	Sinistro				
Cor do veículo	Hospitalares	Materiais	PerdaTotal	Terceiros	Total Geral
Metálica					
Frequência	168	1403	2	18	1591
% linha	10,56%	88,18%	0,13%	1,13%	100,00%
% coluna	84,85%	70,08%	100,00%	75,00%	71,47%
Perolizada					
Frequência	6	51		3	60
% linha	10,00%	85,00%	0,00%	5,00%	100,00%
% coluna	3,03%	2,55%	0,00%	12,50%	2,70%
Sólida					
Frequência	16	525			541
% linha	2,96%	97,04%	0,00%	0,00%	100,00%
% coluna	8,08%	26,22%	0,00%	0,00%	24,30%
Vermelha					
Frequência	8	23		3	34
% linha	23,53%	67,65%	0,00%	8,82%	100,00%
% coluna	4,04%	1,15%	0,00%	12,50%	1,53%
Total Frequência	**198**	**2002**	**2**	**24**	**2226**
Total % linha	**8,89%**	**89,94%**	**0,09%**	**1,08%**	**100,00%**
Total % coluna	**100,00%**	**100,00%**	**100,00%**	**100,00%**	**100,00%**

Sexo	Masculino				
	Sinistro				
Cor do veículo	Hospitalares	Materiais	PerdaTotal	Terceiros	Total Geral
Metálica					
Frequência	544	434	202	450	1630
% linha	33,37%	26,63%	12,39%	27,61%	100,00%
% coluna	70,47%	79,49%	30,01%	59,13%	59,23%
Perolizada					
Frequência	79	38	94	110	321
% linha	24,61%	11,84%	29,28%	34,27%	100,00%
% coluna	10,23%	6,96%	13,97%	14,45%	11,66%
Sólida					
Frequência	19	26		3	48
% linha	39,58%	54,17%	0,00%	6,25%	100,00%
% coluna	2,46%	4,76%	0,00%	0,39%	1,74%
Vermelha					
Frequência	130	48	377	198	753
% linha	17,26%	6,37%	50,07%	26,29%	100,00%
% coluna	16,84%	8,79%	56,02%	26,02%	27,36%
Total Frequência	**772**	**546**	**673**	**761**	**2752**
Total % linha	**28,05%**	**19,84%**	**24,45%**	**27,65%**	**100,00%**
Total % coluna	**100,00%**	**100,00%**	**100,00%**	**100,00%**	**100,00%**

Gráficos com 3 variáveis