Estatística Básica e Introdução ao R

Profª. Dra. Natalia Giordani

1.2.2 Análise exploratória de uma variável quantitativa

- Medidas resumo
  - Medidas de posição (ou localização central)
  - Medidas de dispersão (ou escala ou variabilidade)
  - Medidas de forma
  - Distribuição dos dados

1.2.2 Análise exploratória de uma variável quantitativa

- Medidas de posição

1. Média (média aritmética) = soma de todos os valores / número de observações

   - Ex.: idade, em anos completos, de 10 estudantes da turma

Estudante	Idade
1	17
2	22
3	22
4	25
5	25
6	26
7	27
8	28
9	28
10	60

$$Média = \bar{x} = \frac{17+22+22+25+25+26+27+28+28+60}{10} = \frac{280}{10} = 28{,}0 \text{ anos}$$

1.2.2 Análise exploratória de uma variável quantitativa

2. Mediana = estatística de **ordem**, valor que ocupa a posição central dos dados

- Número par de observações:

Estudante	1	2	3	4	5	6	7	8	9	10
Idade	17	22	22	25	25	26	27	28	28	60

$$Mediana = \frac{25 + 26}{2} = \frac{51}{2} = 25{,}5 \text{ anos}$$

- Número ímpar de observações:

Estudante	1	2	3	4	5	6	7	8	9
Idade	17	22	22	25	25	26	27	28	28

$$Mediana = 25{,}0 \text{ anos}$$

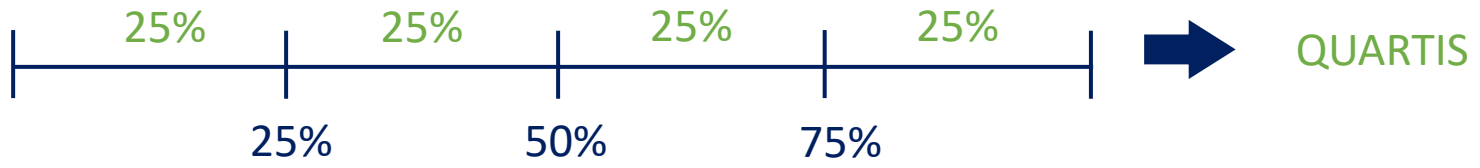
1.2.2 Análise exploratória de uma variável quantitativa

- Média ou Mediana?
  - Média é bastante afetada por valores extremos
  - Mediana é mais robusta
  - Nosso exemplo (idades):
    - Média = 28,0 anos
    - Mediana = 25,5 anos

1.2.2 Análise exploratória de uma variável quantitativa

3. Quantis (Q)

- Útil para indicar posições não centrais dos dados
- Quantil de ordem *p* (0 < p < 1) corresponde ao valor da variável que deixa 100*p% das observações à sua esquerda



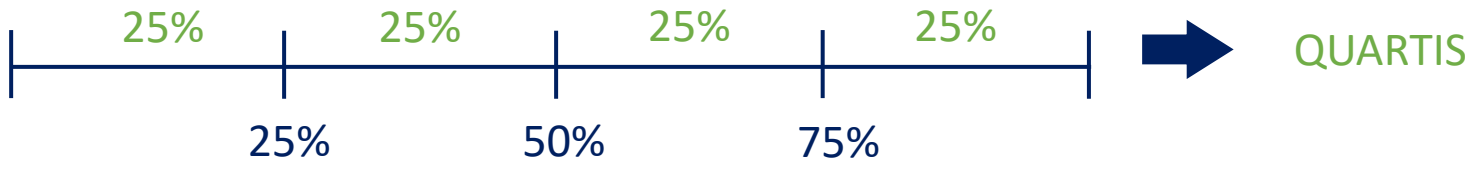
Estudante	1	2	3	4	5	6	7	8	9
Idade	17	22	22	25	25	26	27	28	28

50% = Q2 = mediana = 25 anos

1.2.2 Análise exploratória de uma variável quantitativa

3. Quantis (Q)

- Útil para indicar posições não centrais dos dados
- Quantil de ordem *p* (0 < p < 1) corresponde ao valor da variável que deixa 100*p% das observações à sua esquerda



QUARTIS

Estudante	1	2	3	4	6	7	8	9
Idade	17	22	22	25	26	27	28	28

25% = Q1 = (22 + 22)/2 = 22

75% = Q3 = (27 + 28)/2 = 27,5

1.2.2 Análise exploratória de uma variável quantitativa

- Medidas de dispersão

1. Variância amostral: valor baseado nas diferenças (desvios) entre a média (medida de localização) e o dado observado.

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Note que a unidade de medida da variância é o quadrado da unidade de medida da variável observada!

2. Desvio padrão:

$$s = \sqrt{s^2}$$

1.2.2 Análise exploratória de uma variável quantitativa

3. Distância interquartis (ou amplitude interquartis ou intervalo interquartílico):

$$d_Q = Q_3 - Q_1$$

4. Amplitude:

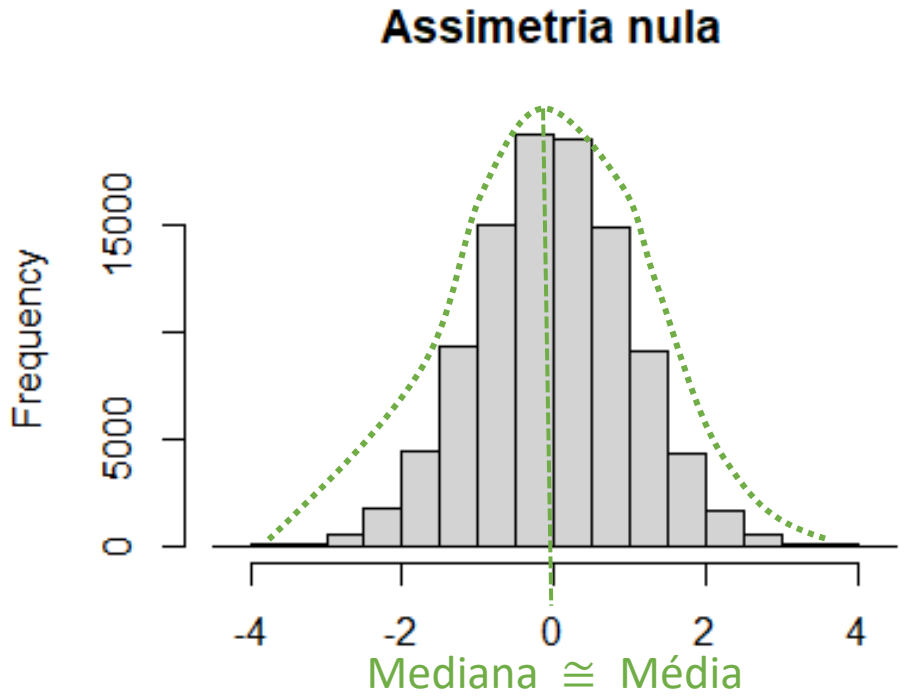
$$amplitude = max(x_1, \dots, x_n) - min(x_1, \dots, x_n)$$

1.2.2 Análise exploratória de uma variável quantitativa

- Medidas de forma

  1. Assimetria

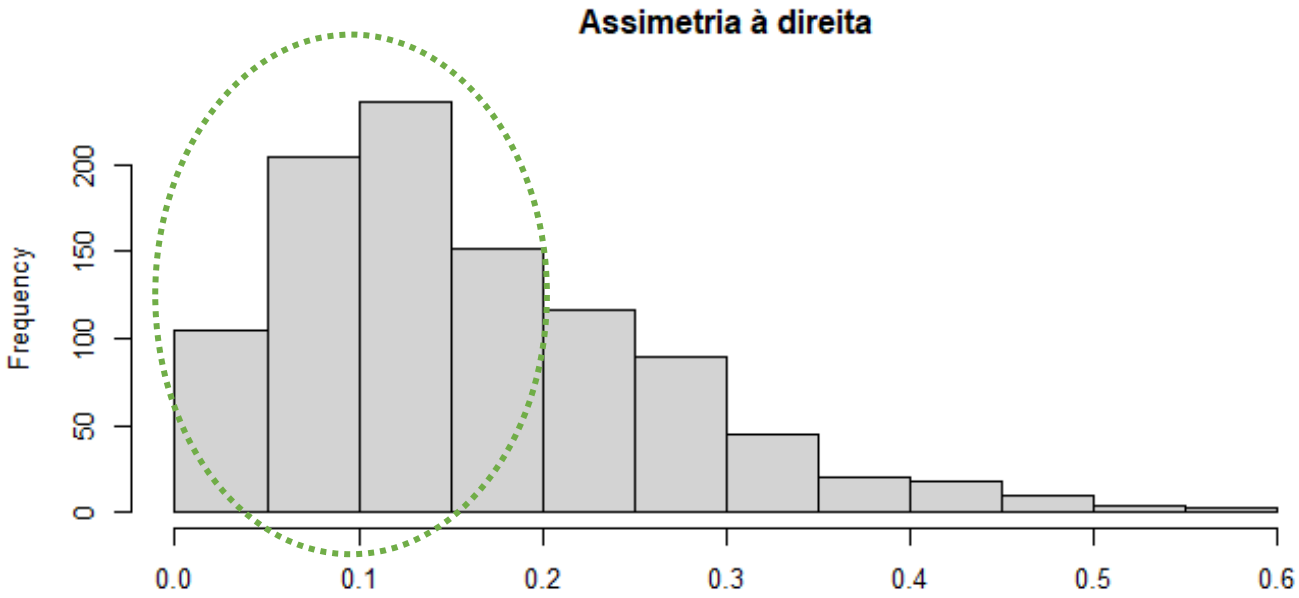
  - Histograma



1.2.2 Análise exploratória de uma variável quantitativa

1. Assimetria

- Histograma

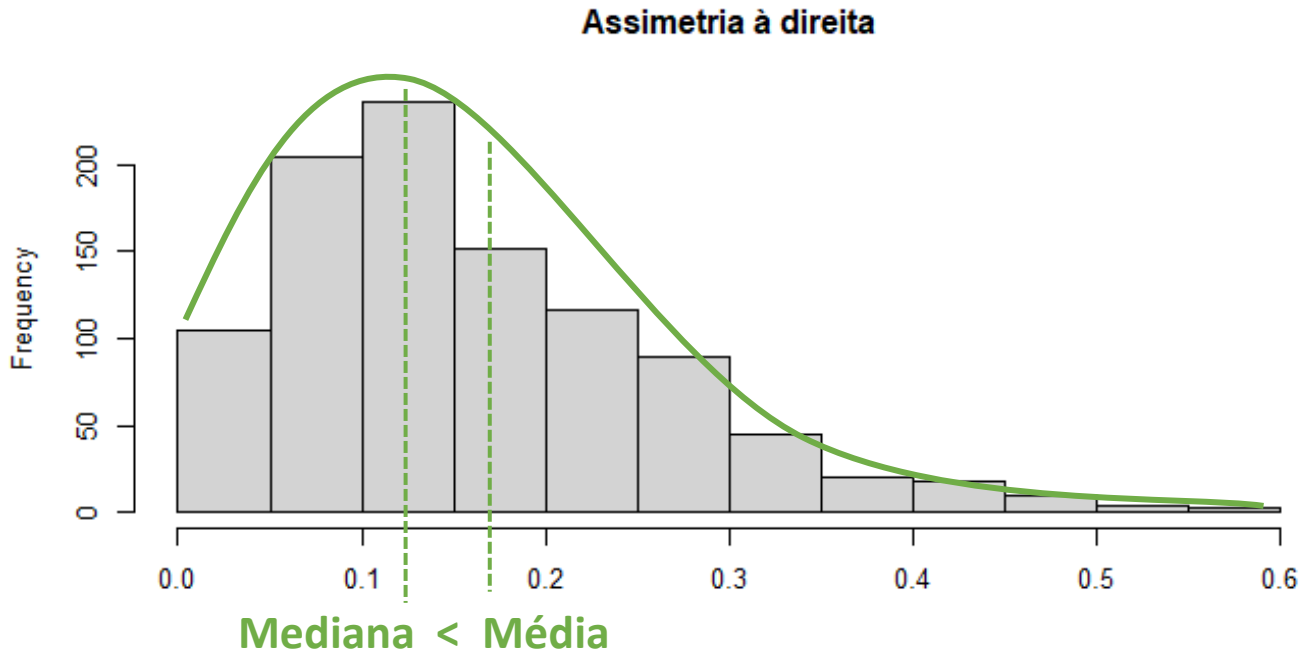


Maior concentração de valores na extremidade inferior da escala

1.2.2 Análise exploratória de uma variável quantitativa

1. Assimetria

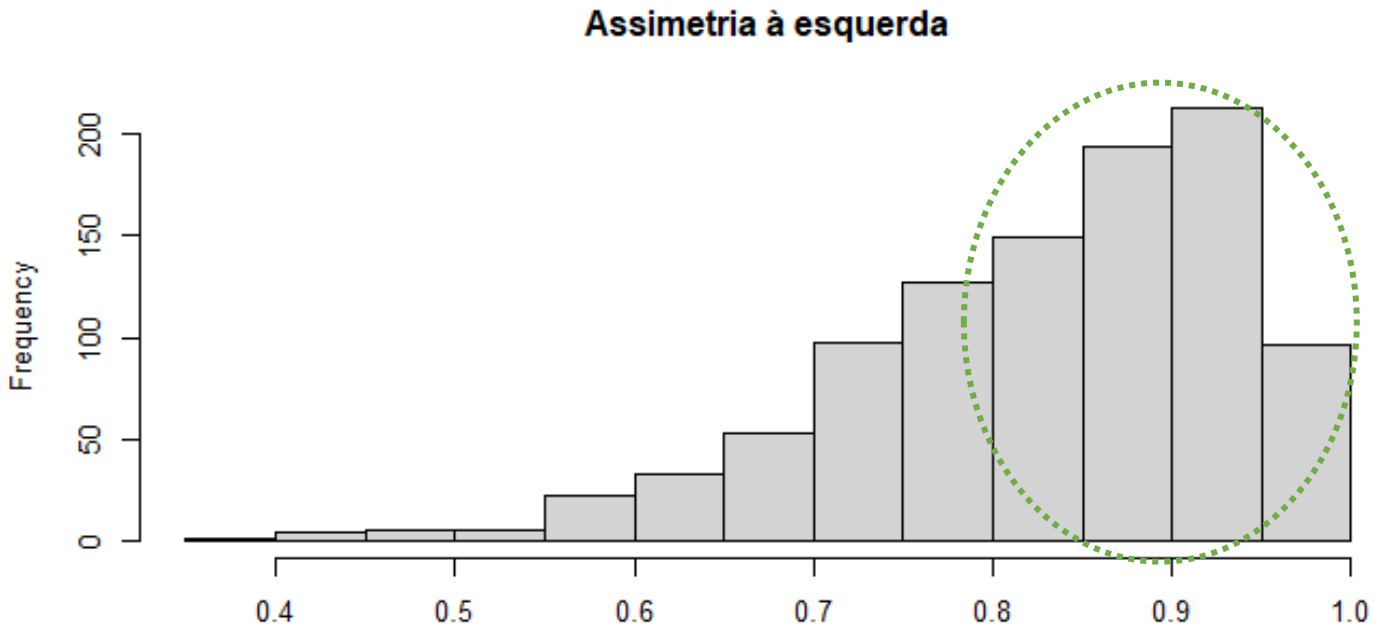
- Histograma



1.2.2 Análise exploratória de uma variável quantitativa

1. Assimetria

- Histograma

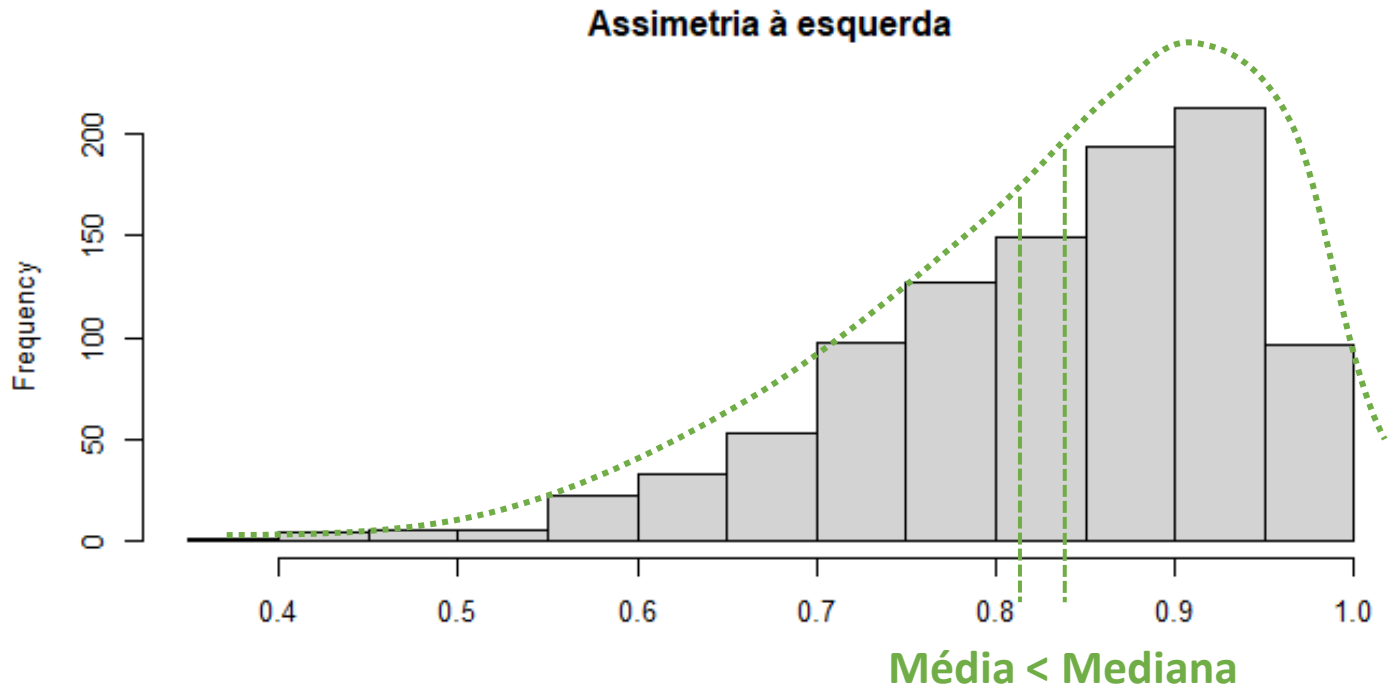


Maior concentração de valores na extremidade superior da escala

1.2.2 Análise exploratória de uma variável quantitativa

1. Assimetria

- Histograma



1.2.2 Análise exploratória de uma variável quantitativa

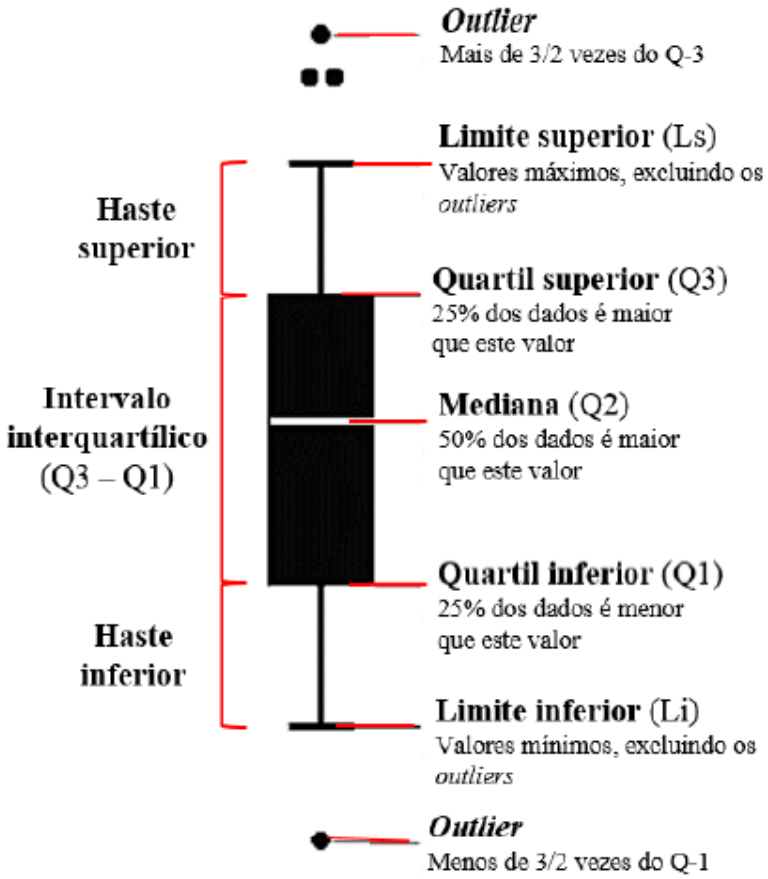
1. Assimetria

- Explorando um pouco mais...

1.2.2 Análise exploratória de uma variável quantitativa

- Distribuição dos dados

1. Boxplot



Retirado de "Boxplot: Um recurso gráfico para a análise e interpretação de dados quantitativos"

1.2.2 Análise exploratória de uma variável quantitativa

- Vamos praticar?
  - Análise exploratória para cada tipo de variável no R

1.3 Análise exploratória de duas variáveis

- Análise descritiva da **relação** entre duas variáveis
  - Duas variáveis qualitativas
  - Duas variáveis quantitativas
  - Uma variável qualitativas e uma variável quantitativa

1.3.1 Análise exploratória de duas variáveis qualitativas

1. Tabelas de contingência ou dupla entrada

- Linhas: categorias de uma variável
- Colunas: categorias da outra variável

Exemplo: distribuição conjunta das variáveis gênero e clique na funcionalidade

Gênero	Clicou na nova funcionalidade		Total
	Sim	Não	
Masculino	1200	426	1626
Feminino	158	2005	2163
Outro	100	41	141
Total	1458	2472	3930

1.3.1 Análise exploratória de duas variáveis qualitativas

Exemplo: porcentagem em relação ao total geral

Gênero	Clicou na nova funcionalidade		Total
	Sim	Não	
Masculino	30,5%	10,8%	41,4%
Feminino	4,0%	51,0%	55,0%
Outro	2,5%	1,0%	3,6%
Total	37,1%	62,9%	100,0%

- 37% dos usuários clicaram; 63% não
- 31% dos usuários clicaram e se identificam com o gênero masculino; 11% dos usuários não clicaram e se identificam com o gênero masculino

Exemplo: porcentagem em relação aos totais nas colunas

Gênero	Clicou na nova funcionalidade		Total
	Sim	Não	
Masculino	82,3%	17,2%	41,4%
Feminino	10,8%	81,1%	55,0%
Outro	6,9%	1,7%	3,6%
Total	100,0%	100,0%	100,0%

- Dentre os usuários que clicaram, 82% se identificam com o gênero Masculino; 11% com o Feminino; e 7% com Outro.

1.3.1 Análise exploratória de duas variáveis qualitativas

2. Avaliação de testes diagnósticos

- Origem: Medicina

Verdadeiro status	**Resultado do teste**		**Total**
	Positivo (**T+**)	Negativo (**T-**)	
Doente (**D**)	n_{11}	n_{12}	n_{1+}
Não doente (**ND**)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	**n**

n_{ij} = quantidade de indivíduos com o i-ésimo status da doença (i = 1 para doente; i = 2 para não doente) e j-ésimo status do teste (j = 1 para positivo; j = 2 para negativo)

$n_{i+} = n_{i1} + n_{i2}$

$n_{+j} = n_{1j} + n_{2j}$

1.3.1 Análise exploratória de duas variáveis qualitativas

2. Avaliação de testes diagnósticos

- **Sensibilidade:** probabilidade de resultado + em D
  - $s = n_{11}/n_{+1}$
  - Capacidade do teste detectar a doença
- **Especificidade:** probabilidade de resultado - em ND
  - $e = n_{22}/n_{2+}$
  - Capacidade de identificar os que não tem a doença

Verdadeiro status	**Resultado do teste**		**Total**
	Positivo (**T+**)	Negativo (**T-**)	
Doente (**D**)	n_{11}	n_{12}	n_{1+}
Não doente (**ND**)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	**n**

1.3.1 Análise exploratória de duas variáveis qualitativas

2. Avaliação de testes diagnósticos

- **Falso positivo:** probabilidade de resultado + em ND
  - $fp = n_{21}/n_{2+}$
- **Falso negativo**: probabilidade de resultado - em D
  - $fn = n_{12}/n_{1+}$
- **Acurácia:** probabilidade de resultado correto
  - $ac = (n_{11} + n_{22})/n$

Verdadeiro status	**Resultado do teste**		**Total**
	Positivo (**T+**)	Negativo (**T-**)	
Doente (**D**)	n_{11}	n_{12}	n_{1+}
Não doente (**ND**)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	**n**

1.3.1 Análise exploratória de duas variáveis qualitativas

2. Avaliação de testes diagnósticos

- Exemplo

Verdadeiro status	**Resultado do teste**		**Total**
	Positivo (**T+**)	Negativo (**T-**)	
Doente (**D**)	40	20	60
Não doente (**ND**)	66	74	140
Total	106	94	**200**

Sensibilidade = Capacidade do teste detectar a doença = 40/60 = 67%

Especificidade = Capacidade de identificar os que não tem a doença = 74/140 = 53%

Falso positivo = Resultado + em ND = 66/140 = 47%

Falso negativo = Resultado - em D = 20/60 = 33%

Acurácia = Resultado correto = (40+74)/200 = 57%

1.3.2 Análise exploratória de duas variáveis quantitativas

1. Gráficos de dispersão

- Eixo das abcissas (x) representa uma variável; eixo das ordenadas (y) outra; cada ponto do gráfico corresponde a uma observação
- **Exemplo:** número de anos de serviço e número de clientes de agentes de uma companhia de seguro

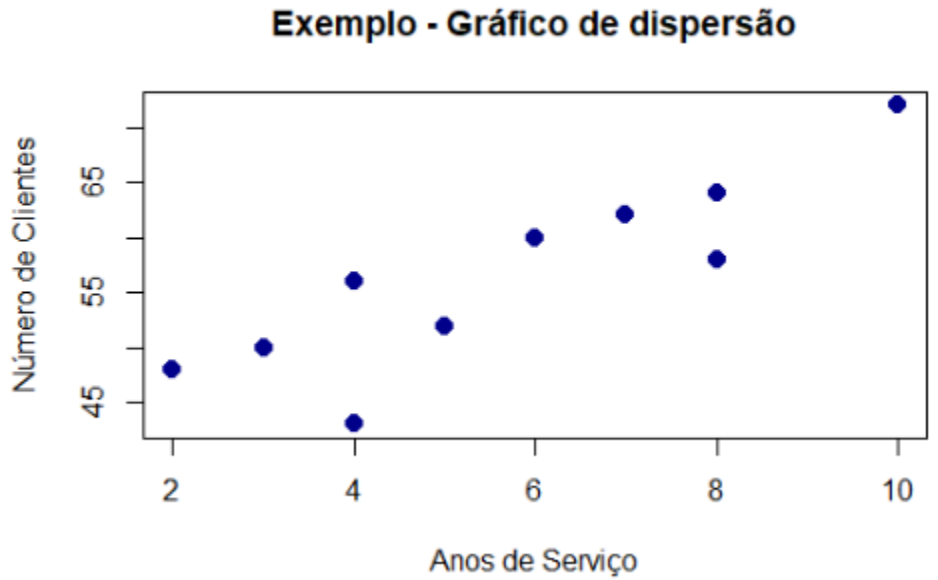
agente	anos_serviço	n_clientes
A	2	48
B	3	50
C	4	56
⋮	⋮	⋮
J	10	72

1.3.2 Análise exploratória de duas variáveis quantitativas

1. Gráficos de dispersão

- **Exemplo:** número de anos de serviço por número de clientes de agentes de uma companhia de seguro

agente	anos_servico	n_clientes
A	2	48
B	3	50
C	4	56
⋮	⋮	⋮
J	10	72



1.3.2 Análise exploratória de duas variáveis quantitativas

2. Coeficiente de correlação (linear)

- Métrica para medir o nível em que as variáveis numéricas estão relacionadas umas às outras
- Varia de -1 a 1
  - Valores próximos de -1 ou de +1: variáveis fortemente associadas ou (linearmente) correlacionadas
  - Valores próximos de 0: variáveis não são correlacionadas
- Quanto mais próximos de uma reta estiverem os pontos: maior a intensidade da correlação (linear) entre elas

1.3.2 Análise exploratória de duas variáveis quantitativas

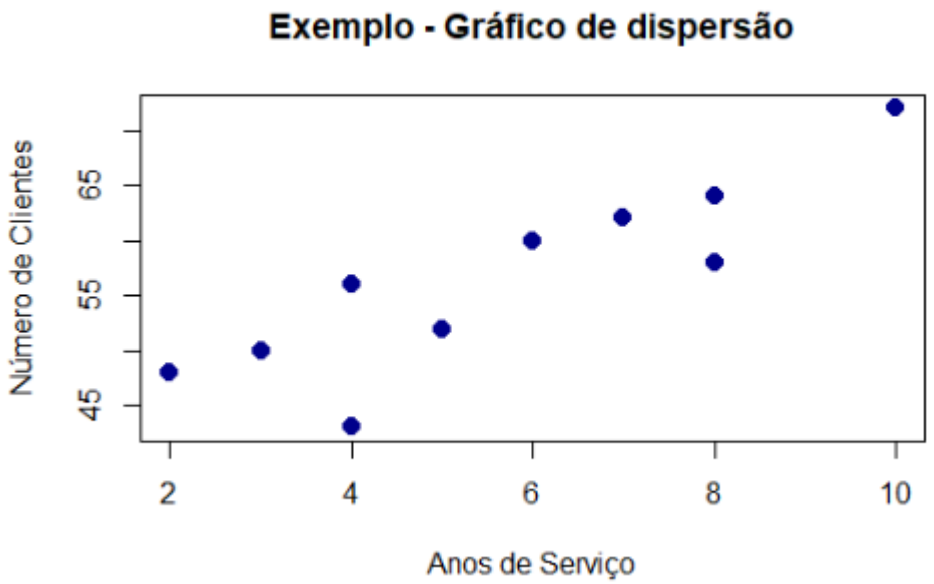
2. Coeficiente de correlação (linear)

- Coeficiente de **correlação de Pearson (r_P)**
  - Multiplicamos os desvios da média de x pelos desvios da média de y e dividimos pelo produto do desvio padrão
  - Método não robusto a outliers
- Coeficiente de **correlação de Spearman (r_S)**
  - São avaliados os desvios em relação aos postos (índice que corresponde à sua ordem)
  - Método robusto a outliers

1.3.2 Análise exploratória de duas variáveis quantitativas

2. Coeficiente de correlação (linear)

- **Exemplo:** número de anos de serviço por número de clientes de agentes de uma companhia de seguro



$r_P = 0,877$
$r_S = 0,878$

|r|	A correlação é dita ...
0	Nula
0 – 0,3	Fraca
0,3 |– 0,6	Regular
0,6 |– 0,9	Forte
0,9 |– 1	Muito forte
1	Perfeita

Fonte: Callegari-Jacques, SM. Bioestatística: princípios e aplicações. Porto Alegre: Artmed, 2003.

1.3.2 Análise exploratória de duas variáveis quantitativas

2. Coeficiente de correlação (linear)

- Explorando um pouco mais...
- Correlações espúrias...
- Atenção: **Correlação não** implica em **causalidade**

1.3.3 Análise exploratória de duas variáveis: uma quali e uma quantitativa

1. Comparação das distribuições por nível da variável qualitativa

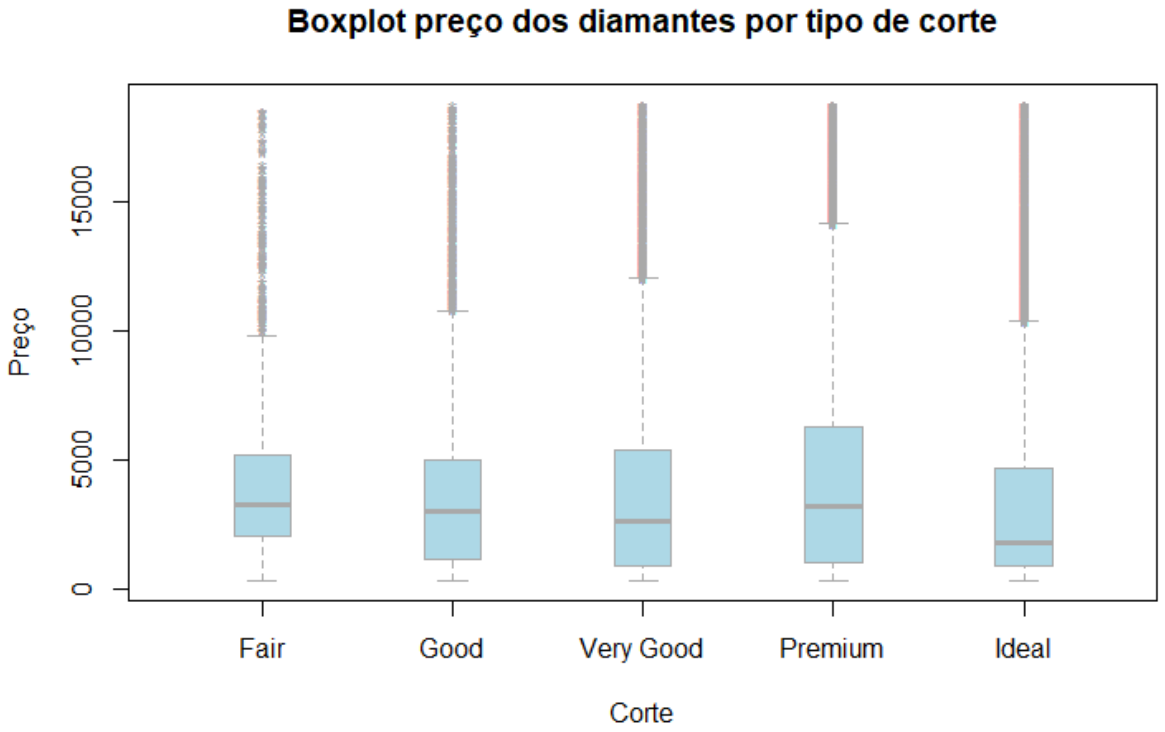
- **Exemplo:** voltando ao exemplo dos diamantes, vamos avaliar se existe associação entre a nota de corte do diamante e seu preço.
  - **Comparação das medidas resumo**

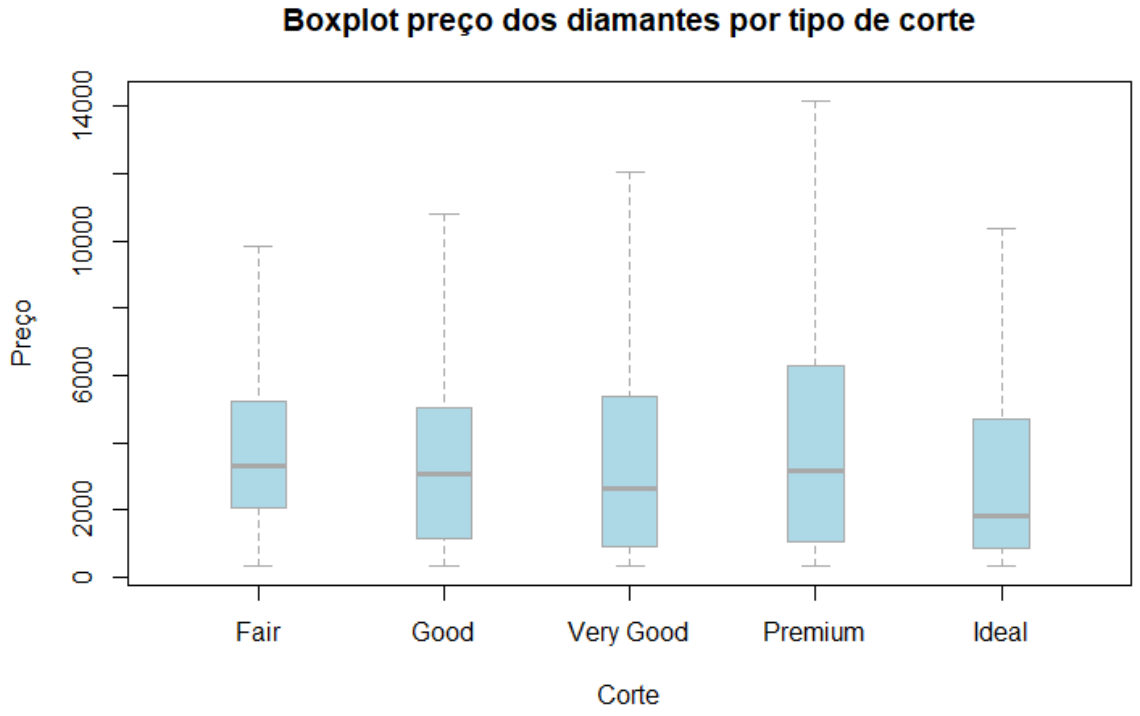
Corte	Média	DP	Mín	Q1	Q2	Q3	Máx
Razoável	$ 4.359,0	$ 3560,4	$ 337,0	$ 2.050,0	$ 3.282,0	$ 5.206,0	$ 18.574,0
Bom	$ 3.929,0	$ 3681,6	$ 327,0	$ 1.145,0	$ 3.050,0	$ 5.028,0	$ 18.788,0
Muito bom	$ 3.982,0	$ 3935,9	$ 336,0	$ 912,0	$ 2.648,0	$ 5.373,0	$ 18.818,0
Premium	$ 4.584,0	$ 4349,2	$ 326,0	$ 1.046,0	$ 3.185,0	$ 6.296,0	$ 18.823,0
Perfeita	$ 3.458,0	$ 3808,4	$ 326,0	$ 878,0	$ 1.810,0	$ 4.678,0	$ 18.806,0

1.3.3 Análise exploratória de duas variáveis: uma quali e uma quantitativa

1. Comparação das distribuições por nível da variável qualitativa

- **Boxplot**





1.3.3 Análise exploratória de duas variáveis: uma quali e uma quantitativa

1. Comparação das distribuições por nível da variável qualitativa

- **Exemplo:** verificar como o % de atrasos de voos varia entre companhias.
  - **Comparação das medidas resumo (%)**

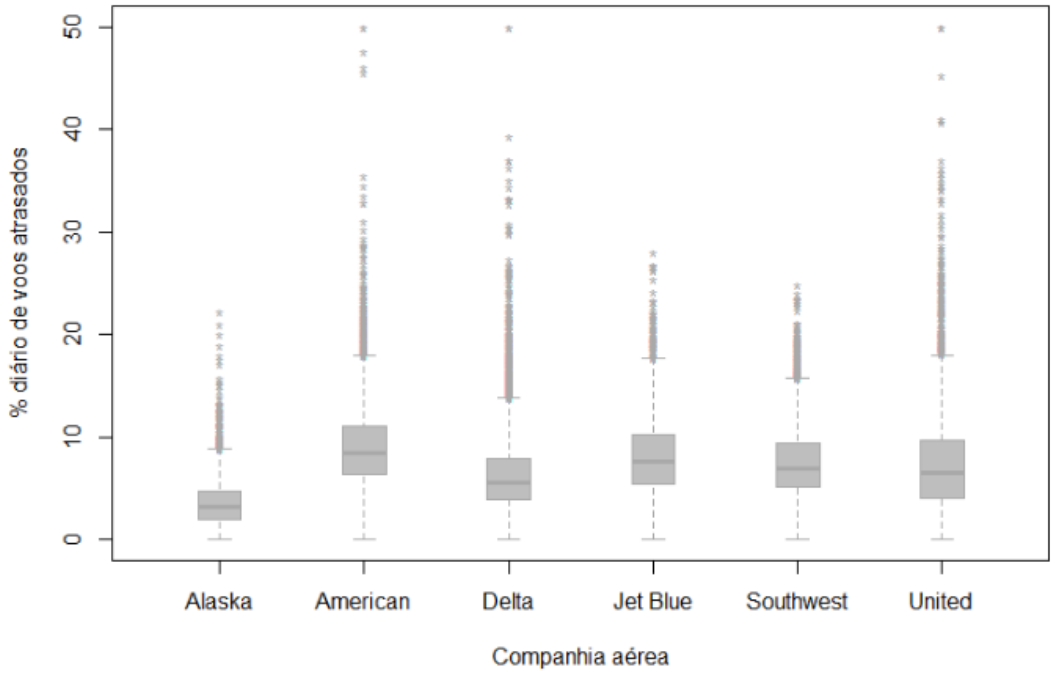
Companhia	Média	DP	Mín	Q1	Q2	Q3	Máx
Alaska	3,52	2,48	0,00	1,93	3,23	4,69	22,29
American	9,04	4,14	0,00	6,34	8,43	10,99	50,00
Delta	6,33	4,70	0,00	3,81	5,55	7,82	100,00
Jet Blue	8,08	3,80	0,00	5,34	7,66	10,28	28,00
Southwest	7,52	3,35	0,00	5,07	6,96	9,35	24,80
United	7,40	5,37	0,00	4,04	6,45	9,63	100,00

Retirado de: Estatística Prática para Cientistas de Dados

1.3.3 Análise exploratória de duas variáveis: uma quali e uma quantitativa

1. Comparação das distribuições por nível da variável qualitativa

- **Boxplot**



- Alaska se destaca por menos atrasos
- American tem a maior mediana
  - Q1 – American > Q3 Alaska

1.4 Análise exploratória de várias variáveis

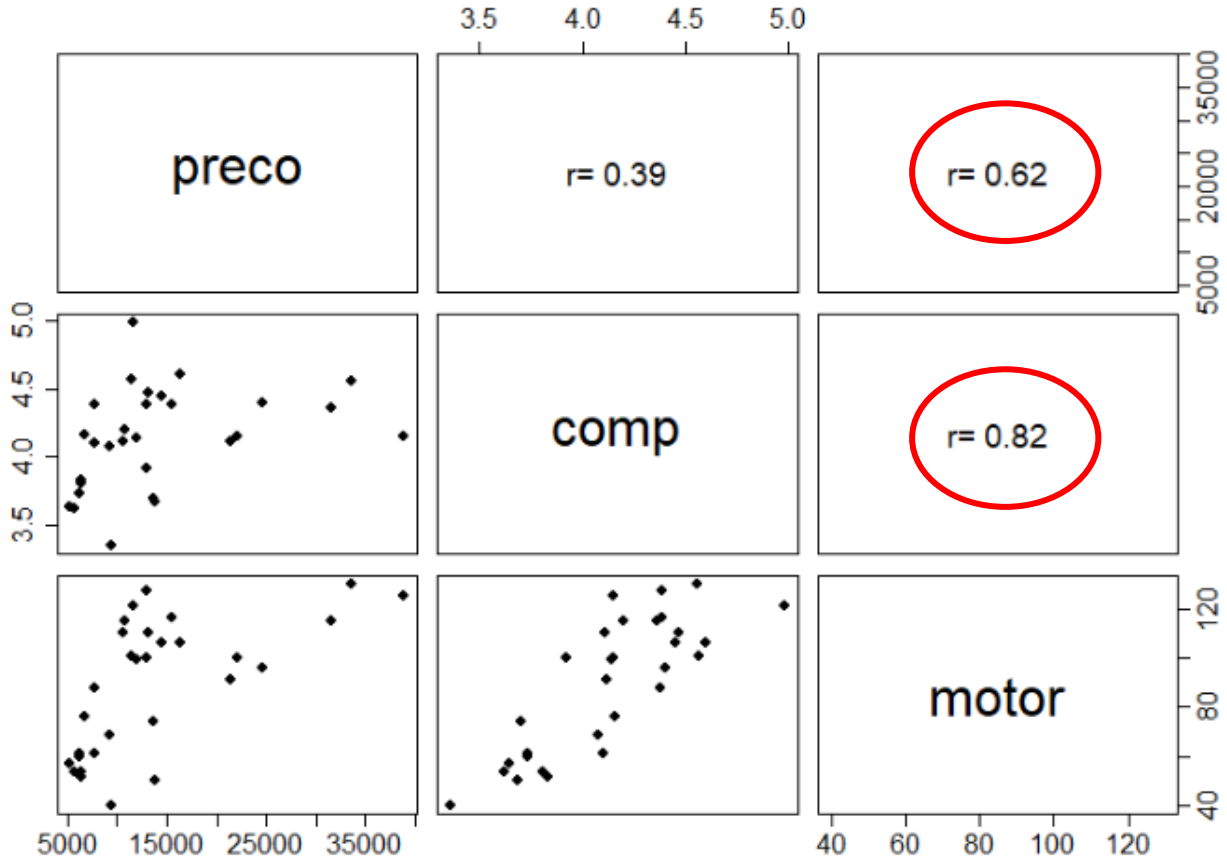
1. Gráficos

- Gráfico do desenhista (Draftsman's display)
  - Matriz cujos elementos são painéis com gráficos de dispersão para cada par de variáveis
  - Exemplo: amostra de 30 veículos onde, para cada um, foram observados preço, comprimento, potência do motor e procedência.

	veiculo	preco	comp	motor	proc
1	Asia Towner	9440	3.36	40	Importado
2	Audi A3	38850	4.15	125	Importado
3	Chevrolet Astra	10532	4.11	110	Nacional
4	Chevrolet Blazer	16346	4.60	106	Nacional
5	Chevrolet Corsa	6176	3.73	60	Nacional

1.4 Análise exploratória de várias variáveis

- Gráfico do desenhista (Draftsman's display)



- Maiores potências de motor associadas a maiores preços e maiores comprimentos
- Relação não tão forte entre preço e comprimento

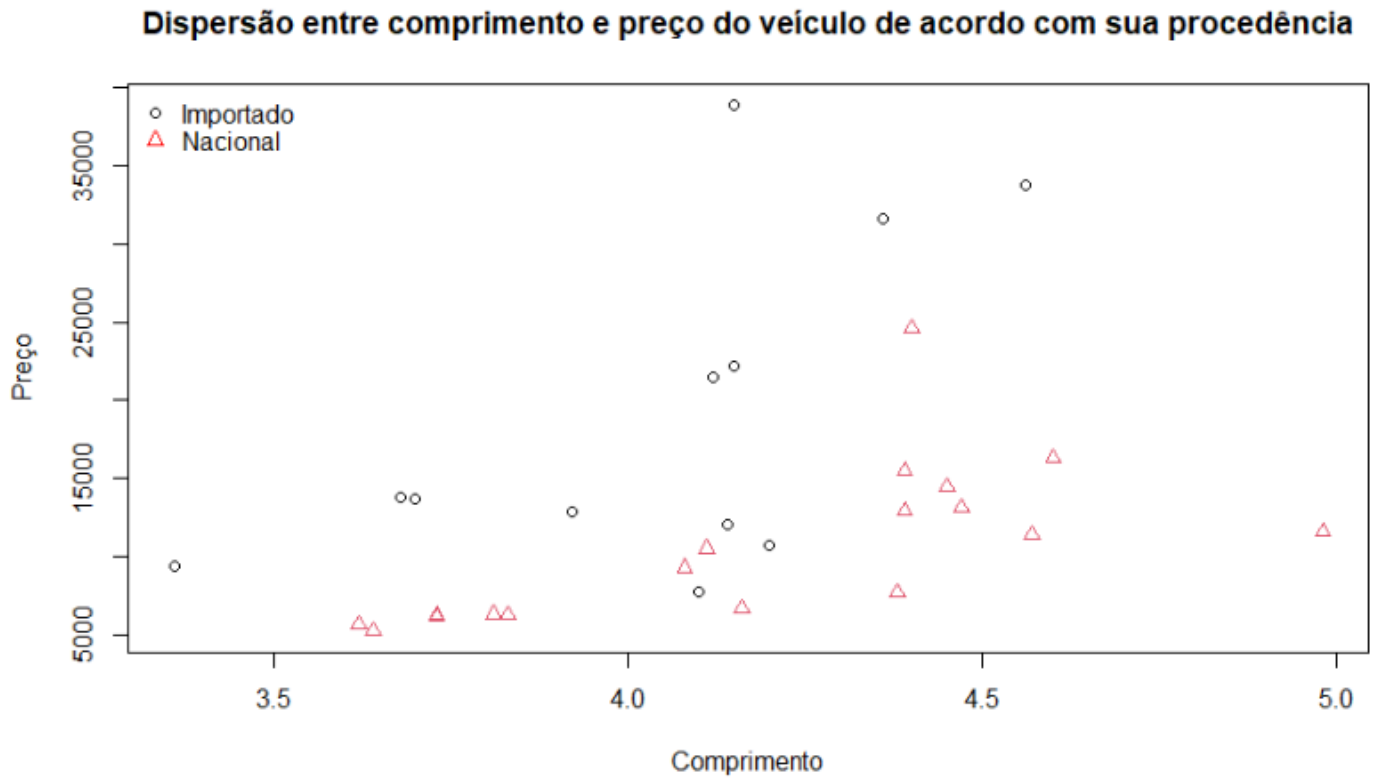
1.4 Análise exploratória de várias variáveis

1. Gráficos

- Gráfico de dispersão simbólico ou estético (*aesthetic*)
  - Gráficos de dispersão para mais de duas variáveis, que se distinguem por diferentes símbolos, cores ou forma dos pontos

1.4 Análise exploratória de várias variáveis

- Gráfico de dispersão simbólico ou estético (*aesthetic*)



- Maiores preços: carros importados

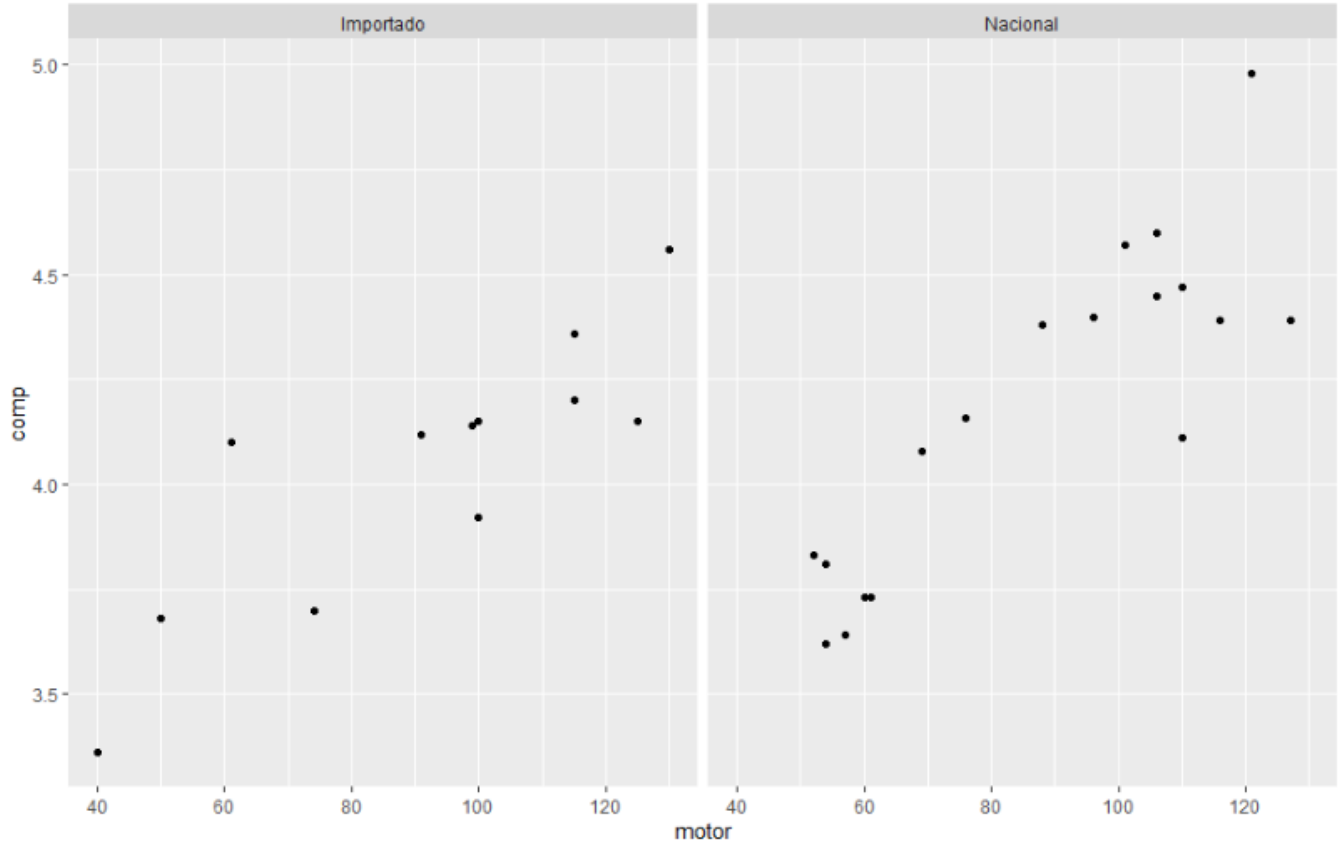
1.4 Análise exploratória de várias variáveis

1. Gráficos

- Partição e janelamento
  - Dividir as observações em subconjuntos (facetas) de acordo com valores de uma das variáveis e criar gráficos de dispersão para cada subconjunto
  - Exemplo: motor e comprimento do veículo de acordo com sua procedência

1.4 Análise exploratória de várias variáveis

- Partição e janelamento



- Janelamento para as variáveis motor e comprimento, categorizado pela variável procedência do veículo
- Indicativo de associação linear positiva entre potência do motor e seu comprimento, independente da procedência

1.4 Análise exploratória de várias variáveis

2. Tabelas

- Tabelas com medidas resumo

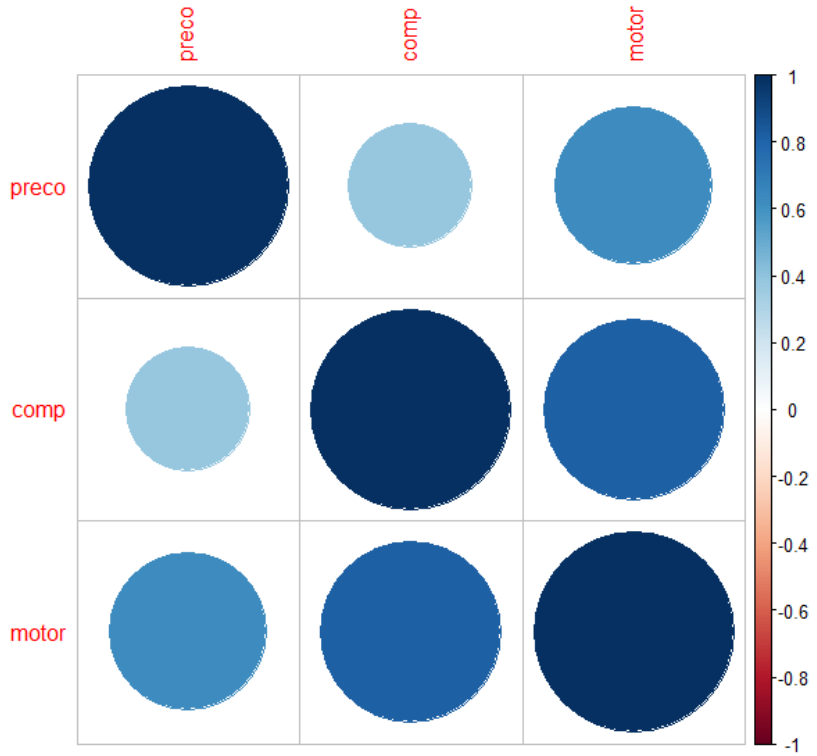
Procedência	Característica	Q1	Q2	Q3
Nacional	Motor (cv)	60,3	92,0	109,0
	Comprimento (m)	3,8	4,3	4,4
Importada	Motor (cv)	70,8	99,5	115,0
	Comprimento (m)	3,9	4,1	4,2

1.4 Análise exploratória de várias variáveis

2. Tabelas

- Matriz de correlação

	Preço (R$)	Comprimento (m)	Motor (cv)
Preço (R$)	1,00	0,39	0,62
Comprimento (m)	0,39	1,00	0,82
Motor (cv)	0,62	0,82	1,00



1.4 Análise exploratória de várias variáveis

2. Tabelas

- Tabelas de contingência de múltiplas entradas
  - Análise de três ou mais variáveis qualitativas (ou quantitativas categorizadas) pode ser realizada usando as mesmas abordagens vistas para análise de duas variáveis quali: tabelas de frequências absoluta e relativa
  - Exemplo: avaliação conjunta da qualidade do corte, cor e claridade do diamante

1.4 Análise exploratória de várias variáveis

- Tabelas de contingência de múltiplas entradas

Frequências absolutas

		D	E	F	G	H	I	J
Fair	I1	4	9	35	53	52	34	23
	SI2	56	78	89	80	91	45	27
	SI1	58	65	83	69	75	30	28
	VS2	25	42	53	45	41	32	23
	VS1	5	14	33	45	32	25	16
	VVS2	9	13	10	17	11	8	1
	VVS1	3	3	5	3	1	1	1
	IF	3	0	4	2	0	0	0
Good	I1	8	23	19	19	14	9	4
	SI2	223	202	201	163	158	81	53
	SI1	237	355	273	207	235	165	88
	VS2	104	160	184	192	138	110	90
	VS1	43	89	132	152	77	103	52
	VVS2	25	52	50	75	45	26	13
	VVS1	13	43	35	41	31	22	1

Frequências relativas

		D	E	F	G	H	I	J
Fair	I1	1.90	4.29	16.67	25.24	24.76	16.19	10.95
	SI2	12.02	16.74	19.10	17.17	19.53	9.66	5.79
	SI1	14.22	15.93	20.34	16.91	18.38	7.35	6.86
	VS2	9.58	16.09	20.31	17.24	15.71	12.26	8.81
	VS1	2.94	8.24	19.41	26.47	18.82	14.71	9.41
	VVS2	13.04	18.84	14.49	24.64	15.94	11.59	1.45
	VVS1	17.65	17.65	29.41	17.65	5.88	5.88	5.88
	IF	33.33	0.00	44.44	22.22	0.00	0.00	0.00
Good	I1	8.33	23.96	19.79	19.79	14.58	9.38	4.17
	SI2	20.63	18.69	18.59	15.08	14.62	7.49	4.90
	SI1	15.19	22.76	17.50	13.27	15.06	10.58	5.64
	VS2	10.63	16.36	18.81	19.63	14.11	11.25	9.20
	VS1	6.64	13.73	20.37	23.46	11.88	15.90	8.02
	VVS2	8.74	18.18	17.48	26.22	15.73	9.09	4.55
	VVS1	6.99	23.12	18.82	22.04	16.67	11.83	0.54

1.4 Análise exploratória de várias variáveis

- Vamos praticar?
  - Análise exploratória para duas ou mais variáveis

Leitura complementar sugerida

- Livro “Estatística: o que é, para que serve, como funciona”
  - 3 primeiros capítulos

- Livro “Storytelling com dados: um guia sobre visualização de dados para profissionais de negócios”
  - Capítulo 2