



UNIPAC - CENTRO UNIVERSITÁRIO PRESIDENTE ANTÔNIO CARLOS  
CAMPUS BARBACENA

Bacharelado em Ciência da Computação



---

# *Mineração de dados*

## **Material de Apoio**

*Parte IV – Análise exploratória dos dados*

*Parte Gráfica*

Prof. Felipe Roncalli de Paula Carneiro  
felipecarneiro@unipac.br

*2º sem / 2023*

*Material cedido pela Profª Livia e Prof Osvano*

# Percentis e quartis

---

- Os percentis e quartis são medidas separatrizes que dividem o conjunto de valores, ordenado de forma crescente, em partes tão iguais quanto possível.

# Percentis

---

- O percentil de ordem  $p$  determina os  $p\%$  menores valores contidos em  $v$ , e a posição  $i$ , que delimita o percentil de ordem  $p$  em  $v$  é dada por:

$$i = \frac{p(n+1)}{100} \quad (\text{arredondar para cima se necessário})$$

- Isso significa que os  $p\%$  menores valores em  $v$  estão abaixo do valor da posição  $i$ .

# Exemplo

---

Item	601	111	712	068	002	065	103	809	044
quant	25	25	29	30	32	65	65	65	90

- $p=50$ , indica que os 50% menores valores no conjunto ordenado estão abaixo do valor localizado na posição 5, pois

$$i = \frac{50(9+1)}{100}$$

- Assim, o valor 32 é o percentil de ordem 50 em  $v$ .

# Quartis

---

- Quartis são casos particulares de percentis.
- O primeiro quartil (Q1) é o percentil de ordem  $p=25$ , ou seja, o valor de  $v$  que separa os 25% valores menores que  $v_i$  dos 75% valores maiores que  $v_i$ .
- Já o segundo quartil (Q2) é o percentil de ordem  $p=50$ , e é equivalente à mediana do conjunto de valores.
- Já o terceiro quartil (Q3) é equivalente ao percentil de ordem  $p=75$ .

# Exemplo

---

Item	601	111	712	068	002	065	103	809	044
quant	25	25	29	30	32	65	65	65	90

- Nesse exemplo:

$$Q1 \Rightarrow v_3=29$$

$$Q2 \Rightarrow v_5=32$$

$$Q3 \Rightarrow v_8=65$$

## Exemplo

---

- Em termos práticos, pode-se dizer que o item 044 está acima do percentil de ordem 75 (ou acima do Q3), indicando que pelo menos 75% dos itens observados vendem menos que ele, evidenciando sua boa aceitação por parte dos clientes.

# Exemplo

---

- Os quartis fornecem uma indicação de centro, dispersão e forma da distribuição dos dados.
- A combinação dos quartis com valores de mínimo e máximo do conjunto de valores formam um conjunto de cinco medidas, chamado de resumo dos cinco números.
  - *Esse conjunto normalmente é visualizado em boxplot.*



# Distribuição de frequências

---

- Dados podem ser organizados de maneira a serem resumidos e visualizados por meios de gráficos ou tabelas.
- Uma organização possível é pela distribuição de frequências, que pode ser obtida distribuindo um conjunto de valores em faixas (intervalos ou bins) e fornecendo o número (ou porcentagem) de valores do conjunto que aparece em cada faixa (ou intervalo).

# Distribuição de frequências

---

- Assim, os valores de um conjunto são resumidos de maneira alternativa ao uso de medidas que resultam em valores únicos (como média, desvio padrão, etc), permitindo a construção de uma visualização compacta desse conjunto.
- A distribuição dos dados pela frequência pode ser feita de modo relativo ou acumulado.

# Frequência relativa

---

- A frequência relativa é a apresentação da frequência de valores que aparecem em cada uma das faixas, dividida pela frequência total de valores de um conjunto e, geralmente, é expressa em porcentagem.
- A frequência relativa é dada por:

$$\text{frequência relativa (faixa)} = \frac{(\text{frequência da faixa})}{(\text{frequência total})} * 100$$

# Frequência acumulada

---

- A frequência acumulada relativa, por sua vez, é a frequência relativa acumulada a cada faixa. No cálculo da frequência relativa acumulada de uma faixa consideram-se as frequências acumuladas das faixas anteriores.
- Normalmente, a representação gráfica sobre frequência é feita por meio de histogramas.

# Exemplo

---

Item	601	111	712	068	002	065	103	809	044
quant	25	25	29	30	32	65	65	65	90

Faixas	Frequências		Frequências acumuladas	
	absolutas	relativas (%)	absolutas	relativas (%)
1-10	0	0	0	0
11-20	0	0	0	0
21-30	4	44,44	4	44,44
31-40	1	11,11	5	55,56
41-50	0	0	5	55,56
51-60	0	0	5	55,56
61-70	3	33,33	8	88,88
71-80	0	0	8	88,88
81-90	1	11,11	9	100
91-100	0	0	9	100

# Exemplo

---

- Para estabelecer as frequências em cada faixa, verifica-se o número de vezes que os valores assumidos pela variável sob análise ocorrem dentro do intervalo que define a faixa. O resultado dessa contagem é a frequência absoluta.
- As frequências relativas podem também ser apresentadas de maneira acumulada, de forma a somar com as faixas anteriores.

# Análise de correlação

---

- As medidas de posição e dispersão se constituem como ferramentas de análise exploratória de um único conjunto de valores.
- A análise de correlação, por outro lado, permite estudar a relação entre dois conjuntos de valores.
- Nesse tipo de análise, quantifica-se o quanto um conjunto de valores (ou uma variável) está relacionado com outro, no sentido de determinar a intensidade e a direção dessa relação.

# Análise de correlação

---

- A correlação indica se, e com que intensidade, os valores de uma variável aumentam (ou diminuem) enquanto os valores de outra variável aumentam (ou diminuem).
- Coeficiente de correlação de Pearson (r):

$$r_{v_1, v_2} = \frac{\sum_{i=1}^n (v_{1i} - média(\mathbf{v}_1)) (v_{2i} - média(\mathbf{v}_2))}{\sqrt{\sum_{i=1}^n (v_{1i} - média(\mathbf{v}_1))^2} * \sqrt{\sum_{i=1}^n (v_{2i} - média(\mathbf{v}_2))^2}}$$

- *Para dois conjuntos de valores  $v_1$  e  $v_2$ .*



# Análise de correlação

---

- O resultado do coeficiente de correlação assumirá valores entre -1 e +1.
- O sinal do resultado indica a direção, se a correlação é positiva ou negativa, e o valor indica a intensidade da correlação.
- Valores acima de  $|0,70|$  indicam forte correlação.

# Exemplo

mês	nov	dez	jan	fev	mar	abr	mai	jun	jul
Quant I Feijoada	24	27	29	30	32	58	64	64	65
Quant II Capirinha	10	25	20	36	28	38	50	60	69

		A		B			
v1	v2	v1 – média(v1)	v2 – média(v2)	A*B	(v1 – média(v1))²	(v2 – média(v2))²	
24	10	-19,67	-27,33	537,56	386,78	747,11	
27	25	-16,67	-12,33	205,56	277,78	152,11	
29	20	-14,67	-17,33	254,22	215,11	300,44	
30	36	-13,67	-1,33	18,22	186,78	1,78	
32	28	-11,67	-9,33	108,89	136,11	87,11	
58	38	14,33	0,67	9,56	205,44	0,44	
64	50	20,33	12,67	257,56	413,44	160,44	
64	60	20,33	22,67	460,89	413,44	513,78	
65	69	21,33	31,67	675,56	455,11	1002,78	
somas	393	336			2690	2966	

- A correlação é 0,89

# Exemplo

---

- Essa correlação de 0,89 (alta e positiva) indica que o comportamento das vendas dos dois itens é similar à tendência de aumento (ou queda) de vendas.
  - Quando o prato feijoada tem mais procura no restaurante, a bebida caipirinha também tem o comportamento de venda alterado positivamente. Se a venda da feijoada cai, cai a também a venda da caipirinha.

# Análise de correlação

---

- Nesse exemplo, o uso da correlação permite interpretar a relação do comportamento dos valores em duas variáveis (análise exploratória dos dados).
- Essa análise pode ser mais útil ainda na fase de pré-processamento dos dados, especificamente para selecionar atributos mais interessantes a serem submetidos a algoritmos de classificação ou agrupamento.
  - *Ferramenta útil para redução de dimensionalidade, contribuindo para diminuir a complexidade de um problema de análise de dados.*

# Representações gráficas

---

- Representações gráficas auxiliam na visualização das características dos dados, o que é útil para os analistas, seja em relação a uma primeira ação de estudo dos dados ou à necessidade de comparação de resultados obtidos com a resolução de tarefas de mineração de dados.

# Representações gráficas

---

- Apesar de poderem ser usados em muitos contextos similares, gráficos de linhas devem ter preferência quando se tem a comparação de diversos tipos conjuntos de dados, uma vez que linhas podem facilmente se sobrepor. Comparações usando gráficos de barras são limitadas, uma vez que as barras diferentes devem colocadas lado a lado.
  - *Obs. de cunho pessoal: gráficos de pizza -> geralmente não são a melhor alternativa para a representação dos dados.*

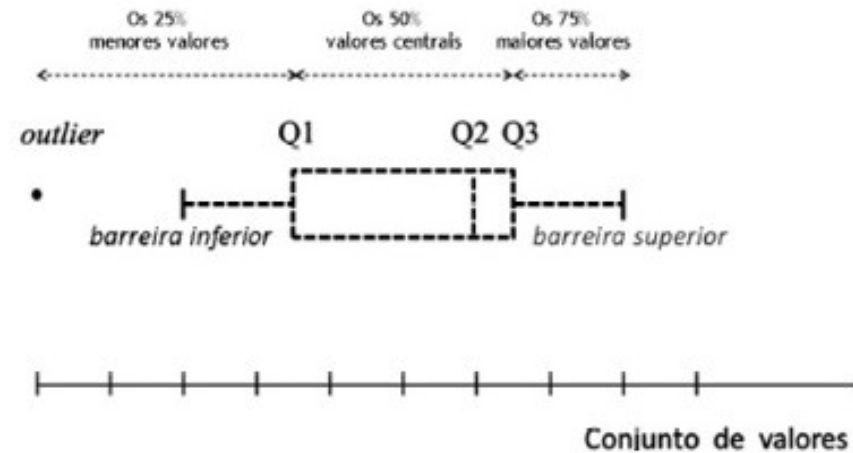
## Diagrama de caixa ou *boxplot*

---

- Boxplot é uma maneira popular de visualizar a distribuição dos dados. Incorpora o resumo dos cinco números, obtidos a partir da combinação dos quartis com os valores mínimo de máximo de um conjunto de valores.

# Diagrama de caixa ou *boxplot*

---



- Os pontos finais da caixa são os Q1 e Q3, sendo que o tamanho da caixa é a faixa interquartil
  - $IQR = Q3 - Q1$



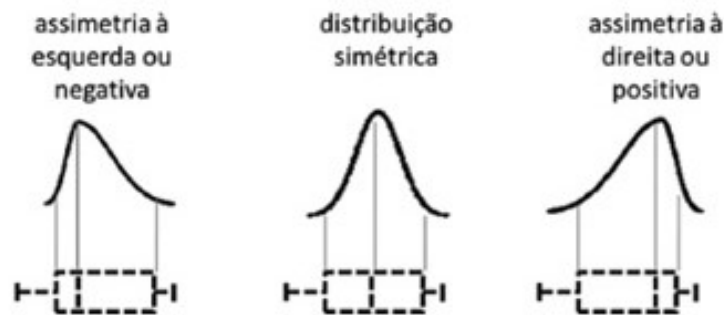
# Diagrama de caixa ou *boxplot*

---

- A mediana é marcada por uma linha dentro da caixa (Q2).
- Barreiras de outlier (traços transversais)
  - Barreira inferior:  $Q1 - (1,5 * IQR)$
  - Barreira superior:  $Q3 + (1,5 * IQR)$

# Diagrama de caixa ou *boxplot*

---



- Adicionalmente, é possível inferir conclusões a respeito da forma de distribuição dos dados, se o conjunto está mais concentrado ou mais disperso, em que região se encontra a maioria dos dados e como eles se dispersam.

# Histogramas

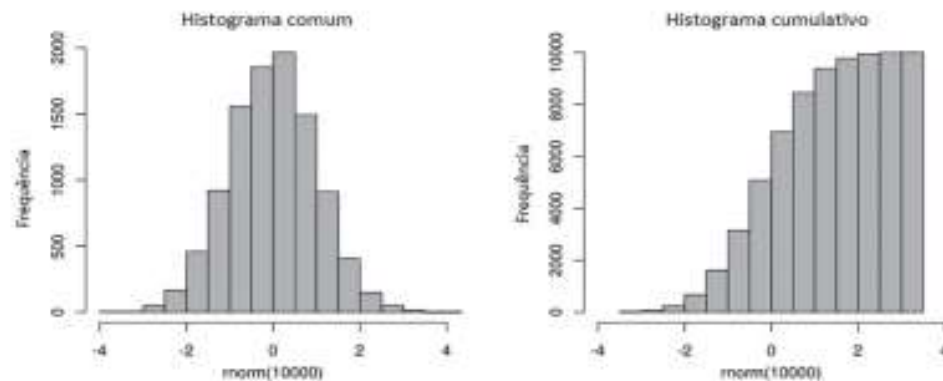
---

- Histograma é uma representação gráfica da distribuição de frequências.
- É construído alocando as faixas no eixo horizontal e as frequências (absolutas ou relativas, acumuladas ou não) no eixo vertical.
- Para cada faixa é atribuída uma barra, de forma que sua altura represente a frequência da faixa.

# Histogramas

---

- Em geral: no eixo x são apresentadas classes que podem representar conjuntos de valores. E no eixo y, é apresentada a frequência de cada classe.



# Histogramas

---

- Exemplo simples: uma base de dados de uma empresa de moda que deseja utilizar a idade de clientes para determinar qual tipo de roupas pode recomendar em uma propaganda.
- Uma excelente maneira de fazer isso: construir categorias que agrupam pessoas de idades próximas.
  - categoria 1: “crianças” (0–12 anos);
  - categoria 2: “adolescentes” (13–17 anos);
  - categoria 3: “jovens adultos” (18–25 anos);
  - categoria 4: “adultos” (25–60 anos);
  - categoria 5: “terceira idade” (>60 anos).

# Histogramas

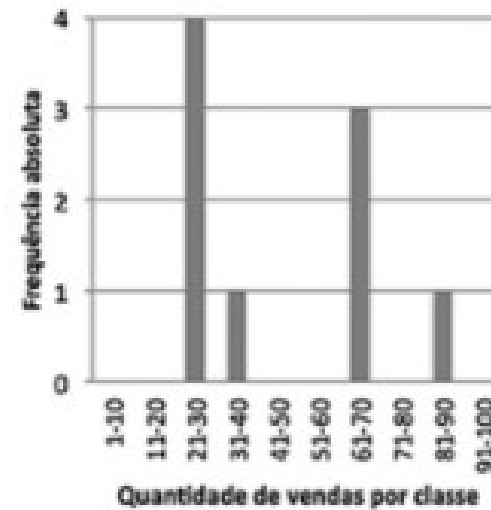
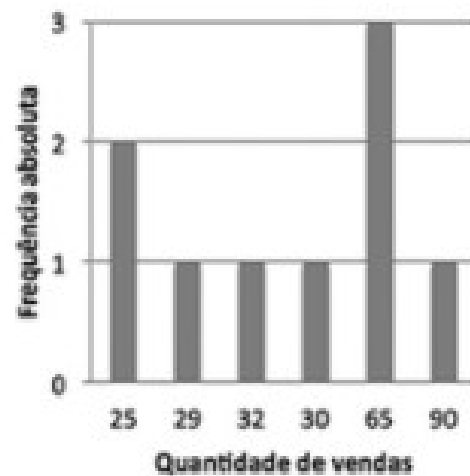
---

- Em um histograma que represente esses dados, haveria 5 barras (uma para cada categoria), e a altura de cada barra indicaria a quantidade de pessoas naquela faixa etária.
- Avaliar tal gráfico permitiria à empresa definir qual é a faixa etária predominante em seu público e, logo, qual deve ser priorizada ao investir em publicidade.

# Histogramas

- Exemplo:

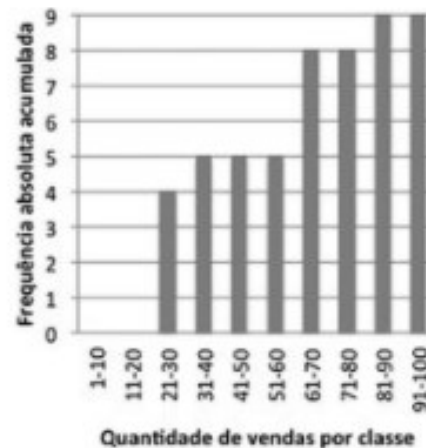
Item	601	111	712	068	002	065	103	809	044
quant	25	25	29	30	32	65	65	65	90



# Histogramas

---

- Exemplo de histograma com frequência acumulada



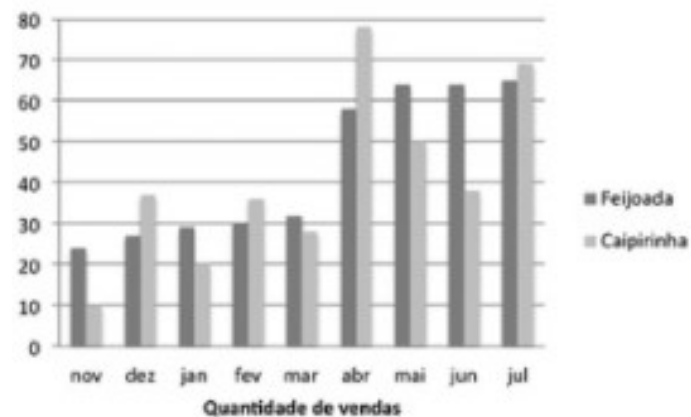


# Histogramas

---

- Outro exemplo:






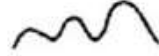
mês	nov	dez	jan	fev	mar	abr	mai	jun	jul
Quant I Feijoadá	24	27	29	30	32	58	64	64	65
Quant II Capirinha	10	25	20	36	28	38	50	60	69



# Histogramas

---

- Um histograma é capaz de indicar o tipo de curva criada pela distribuição das frequências.

Simétricas ou em forma de sino: indicam que valores equidistantes do valor modal têm a mesma frequência. Um exemplo é a curva normal.	
Assimétricas: a cauda da curva é mais longa em um dos lados. Se a parte mais alongada fica à direita, a curva é dita desviada para a direita, ou de assimetria positiva; se ocorre o inverso, diz-se que a curva é desviada para a esquerda, ou de assimetria negativa.	
Na curva em forma de J, ou em J invertido, o valor máximo ocorre em uma das extremidades.	
Uma curva de frequência em forma de U tem os valores máximos nas extremidades.	
Uma curva de frequência bimodal tem dois valores máximos.	
Uma curva de frequência multimodal tem mais de dois valores máximos.	

# Gráfico de setores

---

- O gráfico de setores (ou gráfico pizza) apresenta uma visualização diferente da disposição dos valores de frequência; porém essa visualização não permite observar informação referente a curvas de frequência.

# Diagrama de Pareto

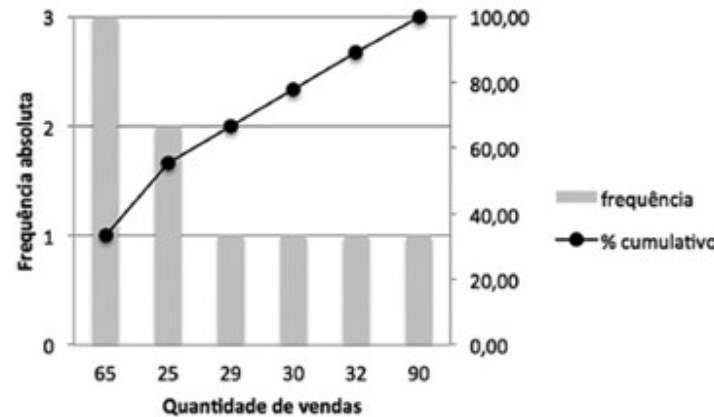
---

- Diagrama de Pareto é uma variação do histograma a partir da ordenação das frequências das faixas, de maior para menor, sendo bastante utilizado como visualização de causas ou prioridades em resolução de problemas.
- Usualmente, uma curva de frequências acumuladas (relativas ou não) acompanha o gráfico, a fim de apoiar a interpretação.

# Diagrama de Pareto

---

- Exemplo:

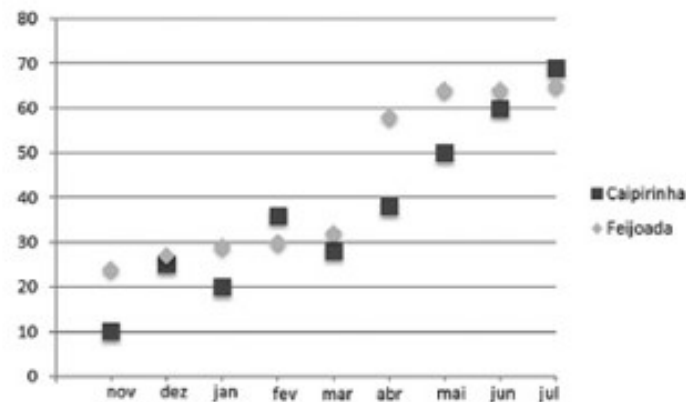


- Percebe-se que vender 65 unidades de um item é a situação mais frequente (ocorre um pouco mais de 30% das medições de desempenho de vendas).

# Diagrama de dispersão (*scatter plot*)

---

- *Scatter plot* é um gráfico no qual os valores assumidos pelas variáveis podem ser representados simultaneamente, permitindo observar a relação existente entre elas.



# Referências

---

- RUSSEL, S., NORVIG, P. *Inteligência Artificial*, Editora Campus, 2ª. edição.
- OLIVEIRA, S. R. M. *Introdução à mineração de dados*, Material para aulas, 2012.
- ZARATE, L.E. *Descoberta de Conhecimento em Banco de Dados e Data Mining*, Material para aulas, 2008.
- FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. *From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, MIT, 1996.
- SILVA, L. A. *Introdução à mineração de dados com aplicações em R*, Elsevier, 1ª ed, 2016.