Organização e Visualização de Dados e Resultados

Prof. Flávio Varejão
Informática
Universidade Federal do Espírito Santo

Sumário

- Importância
- Tabelas
  - Ordenadas, Destacadas, Multisegmentadas
- Gráficos
  - Linhas, Barras, Histogramas
- Conclusão

Importância

- Oceano de Dados e Resultados
  - Dificuldade para entender a natureza dos dados e interpretar os resultados
- Tabelas
  - Estruturação e Organização
  - Ordenação
  - Destaque para valores importantes
- Gráficos
  - Visualização
  - Sumarização
  - Interpretação

Tabelas

- Organização de dados em linhas e colunas
- Usadas quando valores são mais relevantes
  - Visão micro
- Ordenada
  - Apresentar dados ordenados por alguma característica
  - Por uma linha ou por uma coluna
- Destacada
  - Destacar valores específicos na tabela
  - Máximo e/ou mínimo geral
  - Máximos e/ou mínimos de linhas e colunas

Tabelas

- Multisegmentada
  - Tabelas dentro da tabela
  - Usadas com dados hierárquicos
- Com Média e Desvio Padrão
  - Não é uma categoria de tabela
  - Uso muito frequente e importante
  - Idéia de Intervalo de Confiança
- Pareada
  - Informação par a par
  - Antissimétrica permite dois tipos de informação par a par

Tabela Simples

Relation name	Attributes	Cardinality	Instances	Labels	Diversity
20NG	1006.0	1.029	19300.0	20.0	0.003
3sources_bbc1000	1000.0	1.125	352.0	6.0	0.234
3sources_guardian1000	1000.0	1.126	302.0	6.0	0.219
3sources_inter3000	3000.0	1.142	169.0	6.0	0.172
CAL500	68.0	26.044	502.0	174.0	1.000
Emotions	72.0	1.868	593.0	6.0	0.422
Enron	1001.0	3.378	1702.0	53.0	0.442
Genbase	1186.0	1.252	662.0	27.0	0.048
GnegativeGO	1717.0	1.046	1392.0	8.0	0.074
GpositiveGO	912.0	1.008	519.0	4.0	0.438
HumanGO	9844.0	1.185	3106.0	14.0	0.027
Image	294.0	1.236	2000.0	5.0	0.625
Langlog	1004.0	1.180	1460.0	75.0	0.208
Medical	1449.0	1.245	978.0	45.0	0.096
Reuters-K500	500.0	1.462	6000.0	103.0	0.135
Scene	294.0	1.074	2407.0	6.0	0.234
Stackex_chemistry	540.0	2.109	6961.0	175.0	0.436
VirusGO	749.0	1.217	207.0	6.0	0.266
Water-quality	16.0	5.073	1060.0	14.0	0.778
Yeast	103.0	4.237	2417.0	14.0	0.082

Tabela Ordenada

	Ordered frequencies		Cumulative ordered frequencies	
Ordered Class	Economics	Engineering	Economics	Engineering
(1)	48.5	50.2	48.5	50.2
(2)	38.2	43.3	86.7	93.5
(3)	8.2	4.5	94.9	98
(4)	1.7	2.0	96.6	100
(5)	1.7	0.0	98.3	100
(6)	1.7	0.0	100	100
(7)	0.0	0.0	100	100

Tabela Destacada

Learner	Accuracy	Precision	Recall	Subset Accuracy	F-measure	Hamming Loss
DTECC	3.184211	3.026316	4.342105	3.368421	3.157895	2.684211
DTECCd	**2.710526**	3.105263	3.026316	3.000000	**2.684211**	3.157895
DTECCf	4.500000	4.550000	**2.050000**	4.750000	4.300000	5.150000
MEECC	3.157895	**2.842105**	5.052632	**2.421053**	3.157895	**2.289474**
MVECC	3.105263	3.236842	3.842105	2.789474	3.105263	3.026316
STACKECC	4.000000	3.900000	2.550000	4.300000	4.250000	4.300000

Tabela Multisegmentada

Model	Hyperparameters	Range
KNN	Number of neighbors	$\{1, 3, 5\}$
	Distance metric	{Euclidean, Weighted}
SVM	C-SVM C, $\log_{10}(.)$	$\{-3, -2, -1, 0, 1\}$
	RBF kernel γ $\log_{10}(.)$	$\{-3, -2, -1, 0\}$
RF	Number of trees (estimators)	$\{10, 20, 50\}$
	Number of features for split	$\{1, \ldots, 8\}$
MLP	Hidden nodes	$\{2, 3, 4, 5, 10, 15, 20\}$
	Hidden layer activation	{Relu, Sigmoid}
	Max iter	{100, 1000}

Tabela Multisegmentada com Média e Desvio Padrão

DATASETS	QTY	PROTRAS			BIRCHSCAN			DENDIS			DIDES			R*		
		μ	σ	○	μ	σ	○	μ	σ	○	μ	σ	○	μ	σ	○
breast	1000	0,06	2,00	4	**97,08**	4,79	**1**	0,06	2,00	4	0,06	2,02	4	94,52	0,03	2
leaves	553	**77,02**	28,68	**1**	71,4	10,06	2	-	-	5	0	0,05	4	29,76	13,55	3
wdbc	1000	**99,86**	2,37	**1**	95,95	11,81	3	-	-	5	53,96	29,86	4	99,43	4,45	2
iris	1000	54,55	43,11	4	**74,83**	23,40	**1**	32,05	31,20	5	62,94	44,52	2	60,25	42,66	3
wine	553	35,29	39,80	4	**83**	15,54	**1**	9,87	11,77	5	38,7	39,74	3	38,71	39,74	2
yeast	741	83,26	24,33	2	74,2	9,87	3	57,94	16,46	4	**96,86**	3,06	**1**	52,50	26,62	5
thyroid	827	63,82	36,66	4	70,48	18,07	3	42,58	28,62	5	**86,83**	22,21	**1**	76,44	19,21	2
glass	917	73,61	33,84	4	79,52	8,63	3	59,67	33,30	5	**92,11**	9,92	**1**	83,03	11,62	2
r3	1000	**80,57**	16,44	**1**	68,74	19,69	2	67,45	24,89	3	60,86	27,68	4	47,80	29,46	5
r4	844	87,68	17,87	2	34,79	17,12	5	72,85	22,92	3	**98,61**	11,34	**1**	65,20	33,59	4
r7	728	72,40	30,02	2	62,06	20,32	3	**81,86**	19,97	5	7,82	19,25	2	0,54	1,79	**1**
r8	788	73,44	28,30	4	78,07	29,81	3	46,29	22,58	5	85,16	33,18	2	**86,49**	33,43	**1**
Abalone	952	94,25	17,89	2	88,95	9,25	4	**96,42**	11,46	**1**	88,55	28,55	5	92,13	19,36	3
Cadata	580	65,51	16,22	3	**80,42**	10,64	**1**	75,72	13,16	2	7,46	7,52	5	7,47	10,97	4
Letter	341	5,63	4,87	3	**59,67**	13,00	**1**	12,66	9,58	2	0.00	0,00	4	-1,91	9,89	5
Mushrooms	429	-0,29	1,39	4	60,47	34,46	2	-	-	5	0,78	1,49	3	**73,73**	37,91	**1**
Pendigits	611	-1,21	3,95	5	**67,02**	19,73	**1**	4,51	7,71	3	-1,19	2,86	4	5,79	5,42	2
Sensorless	701	24,07	25,04	3	**78,85** (625)	16,85	**1**	-	-	5	17,04	25,90	4	67,97	39,01	2
Shuttle	274	35,75	37,41	3	78,64	13,14	2	**89,98**	11,90	**1**	-0,45	0,52	4	-1,11	0,75	5
Average		53,96	21,59	2	73,63	16,11	**1**	49,99	17,83	5	41,90	16,30	4	51,51	19,97	3
Victory				3			**7**			3			4			2

Tabela Pareada

p-value for F1-Score \ p-value for accuracy			
KNN	0.7169	0.2849	0.5420
0.3166	SVM	0.4259	0.1739
0.9244	0.5709	RF	0.8893
0.4666	0.0170	0.5093	CNN

Gráficos

- Informação relativa é mais importante do que valores absolutos
- Permitem condensar os dados
  - Visão macro
  - Sumarização
  - Foco no que é relevante
- Facilitam comparações e interpretações

Gráfico de Linhas

- Ordenação
  - Comportamento observado segundo alguma característica ordenada
- Série temporal
  - Variação cambial
  - Valor de ações na bolsa
  - Variação de Temperatura de Paciente
- Relações entre Variáveis não necessita ser cronológica
  - Processo físico
    - Temperatura x Pressão

Gráfico de Linhas

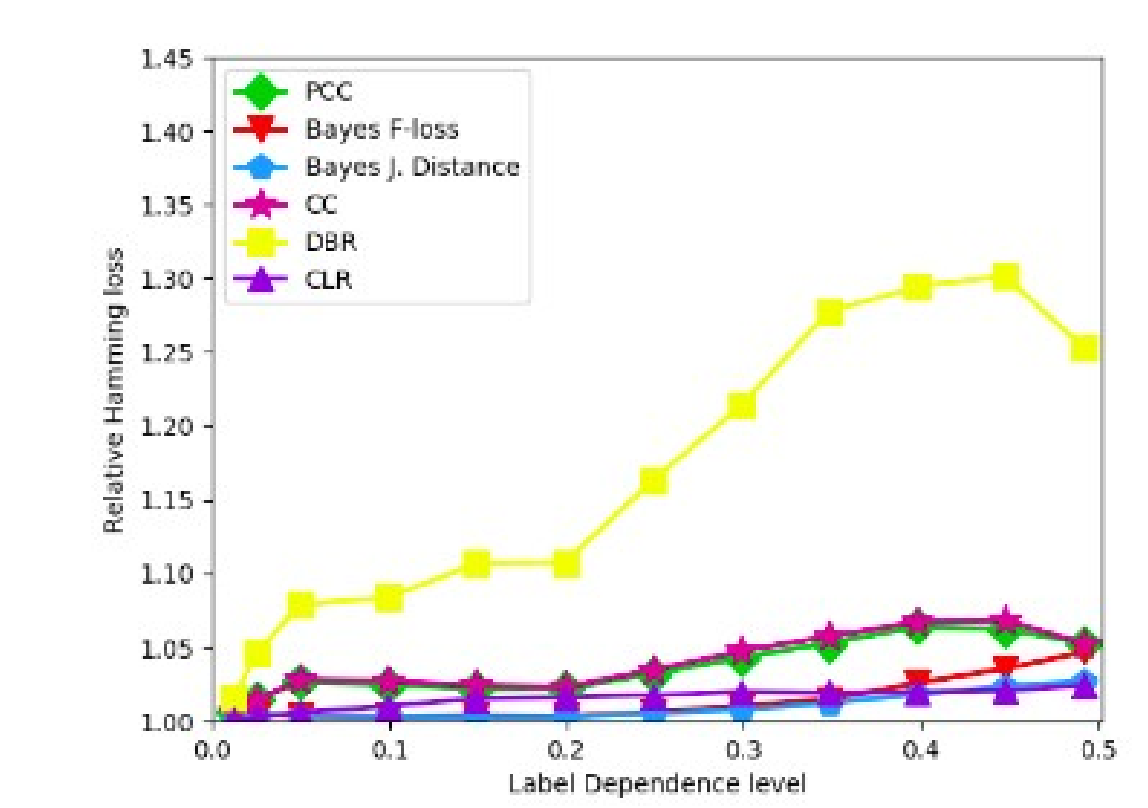


Gráfico de Barras

- Pode não haver relação de ordem entre as unidades medidas
- Vendas de Produtos
  - Faturamento com venda de veículos no ano

Gráfico de Barras

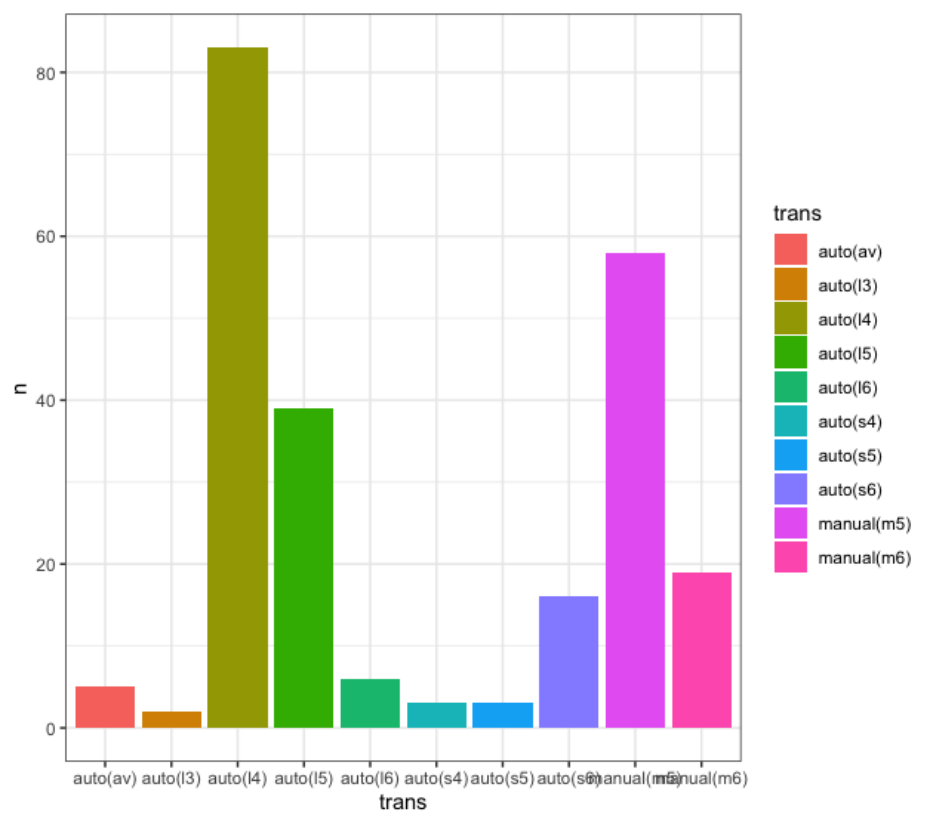


Gráfico de Barras com Média e Intervalo de Confiança

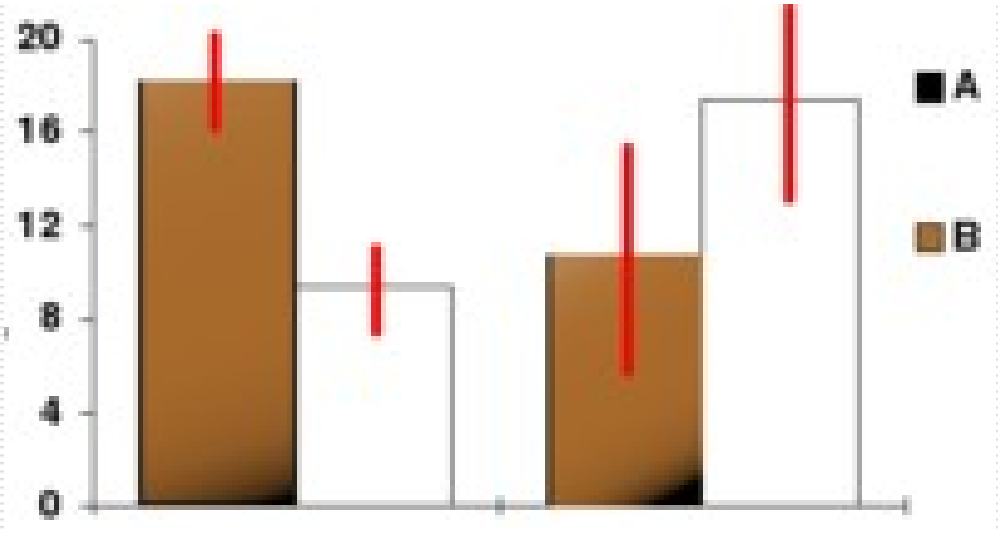


Gráfico de Barras Empilhadas e Segmentadas

- Permite representação de mais uma dimensão
- Empilhadas
  - Foco na Composição
- Segmentadas
  - Foco na Comparação

Gráfico de Barras Empilhadas

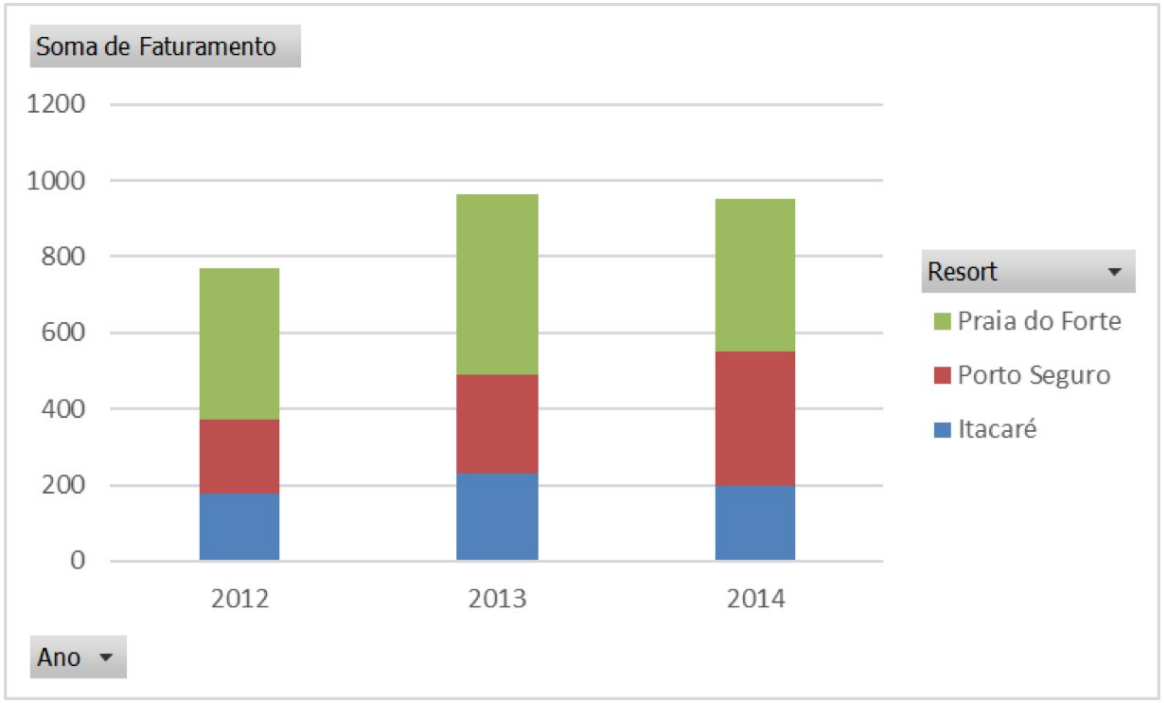
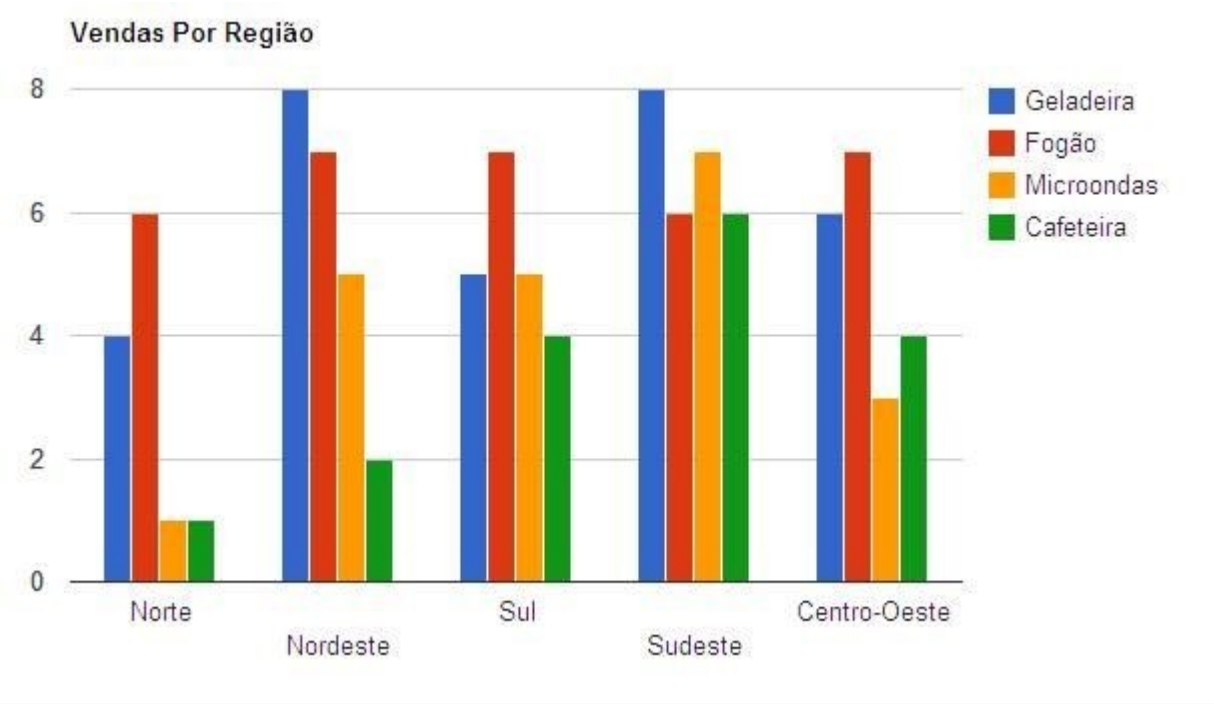


Gráfico de Barras Segmentado



Histograma

- Foco em Comparação
- Contagem de Ocorrências

Histograma

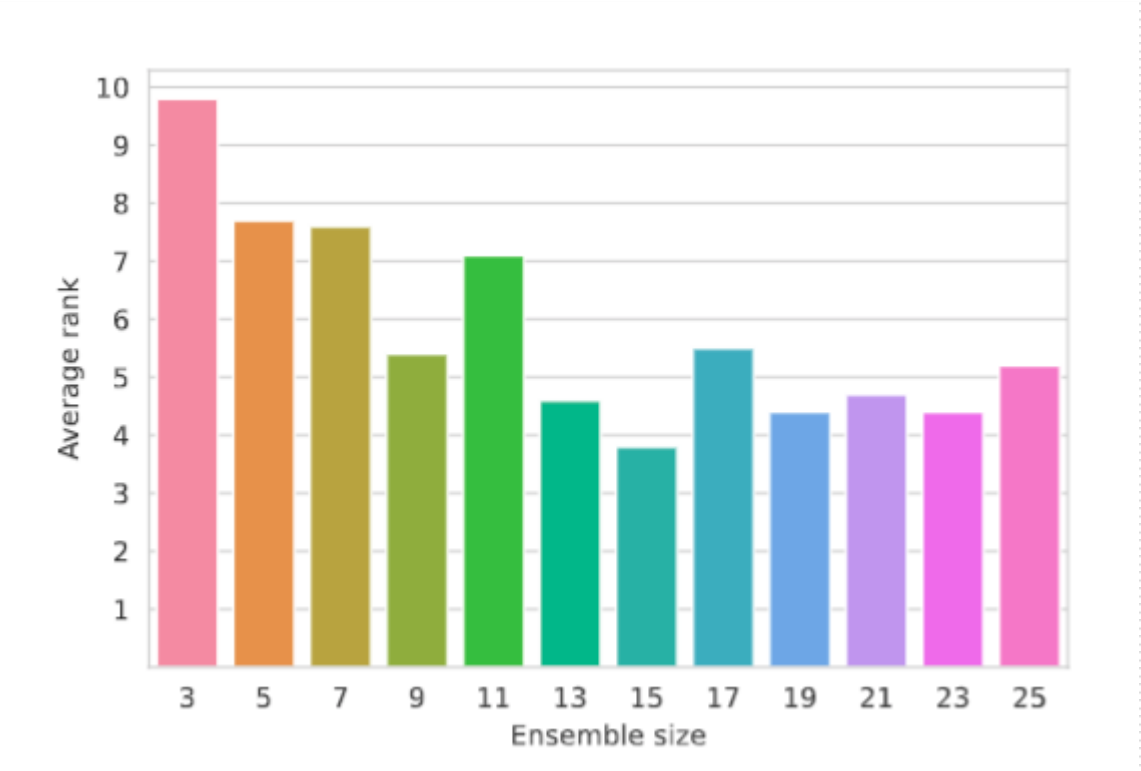
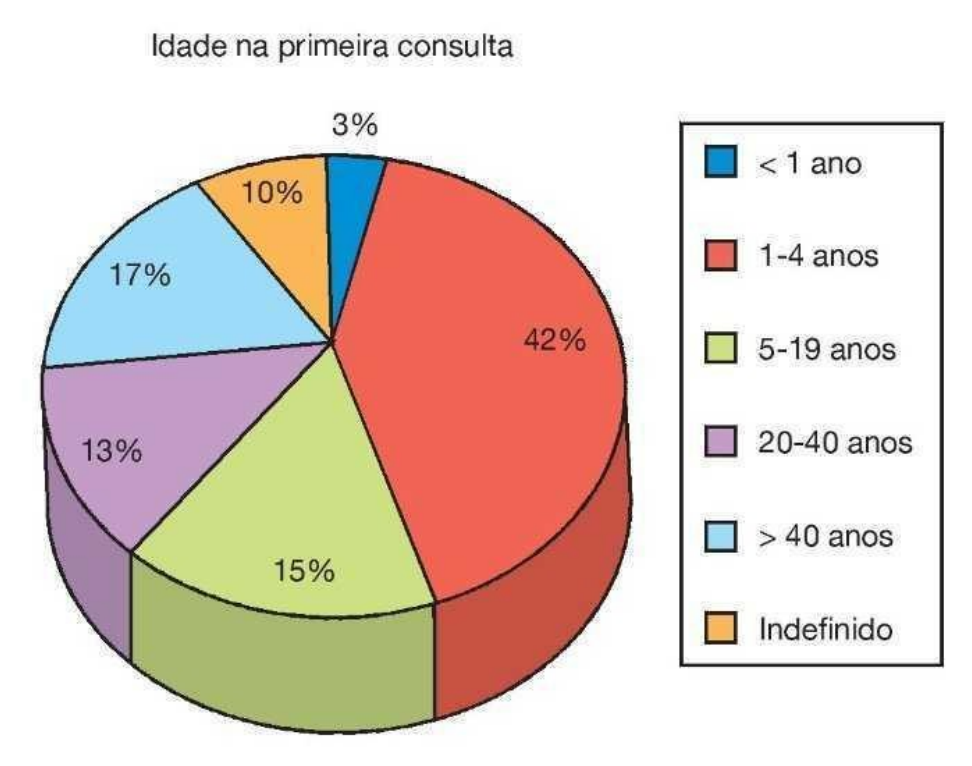

Average rank
10
9
8
7
6
5
4
3
2
1
3
5
7
9
11
13
15
17
19
21
23
25
Ensemble size

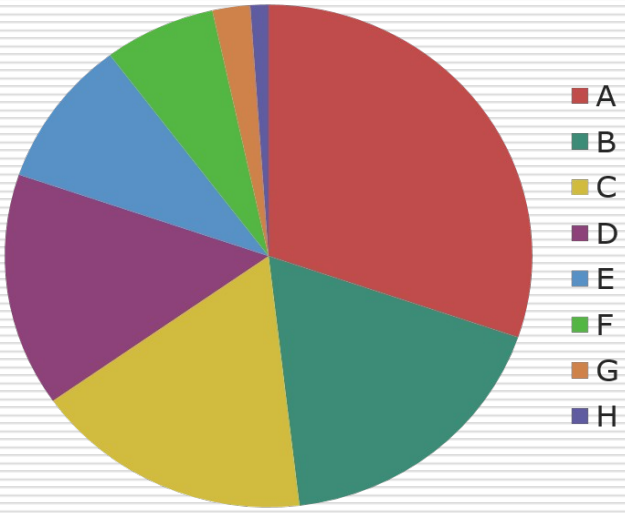
Gráfico de Pizza

- Foca na Composição
  - Partes de um todo
  - Dá mais noção do tamanho da fatia no todo
- Representação de Percentagens ou Ocorrências

Gráfico de Pizza



Histograma x Pizza



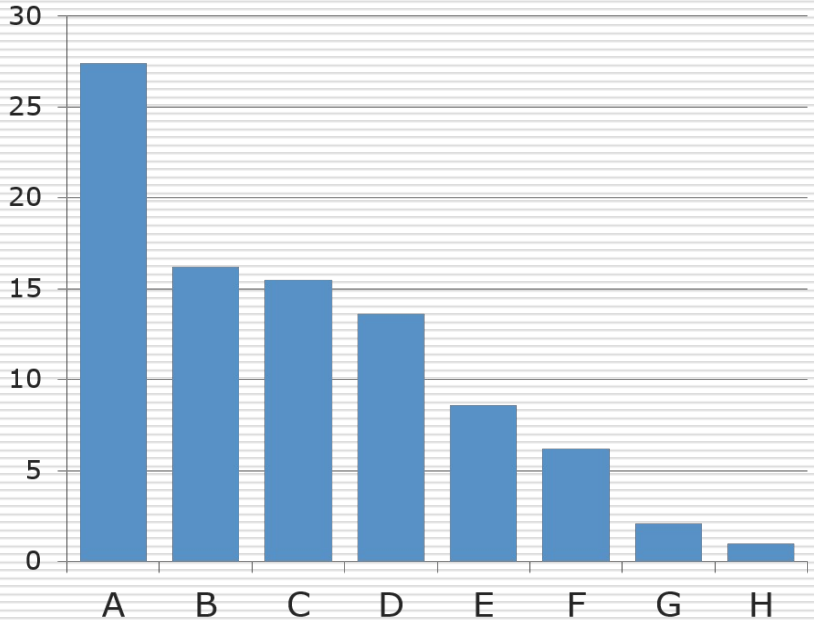


Gráfico de Dispersão (Scatter Plot)

- Mostra como indivíduos se apresentam no espaço dimensional
- Noção de Dispersão e Agrupamento
  - Perfil de produtos ou clientes

Gráfico de Dispersão (Scatter Plot)

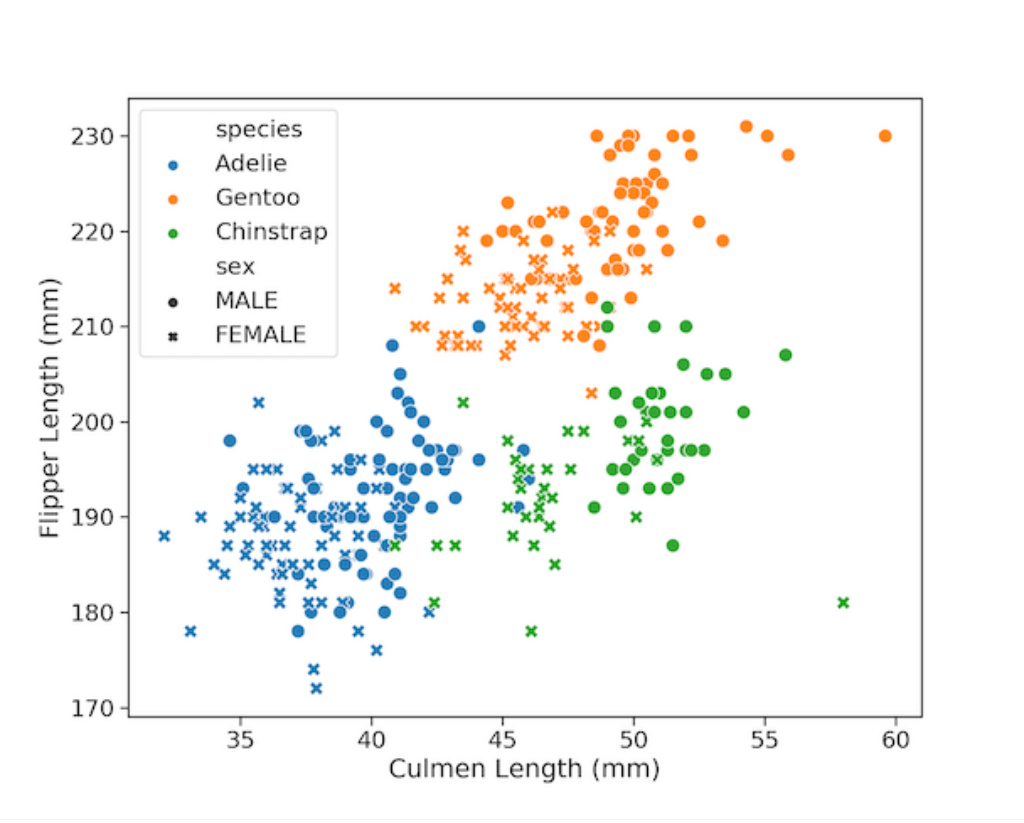


Gráfico de Caixas (Boxplot)

- Foca na Dispersão
- Noção de Distribuição de Dados
- Interessante porque mostra graficamente medidas estatísticas
  - Mediana
  - Quartis
  - Média
- Mostra presença de ruídos (outliers)
  - Limite Inferior = Primeiro Quartil – 1,5 * (Terceiro Quartil – Primeiro Quartil)
  - Limite Superior = Terceiro Quartil + 1,5 * (Terceiro Quartil – Primeiro Quartil)

Gráfico de Caixas (Boxplot)

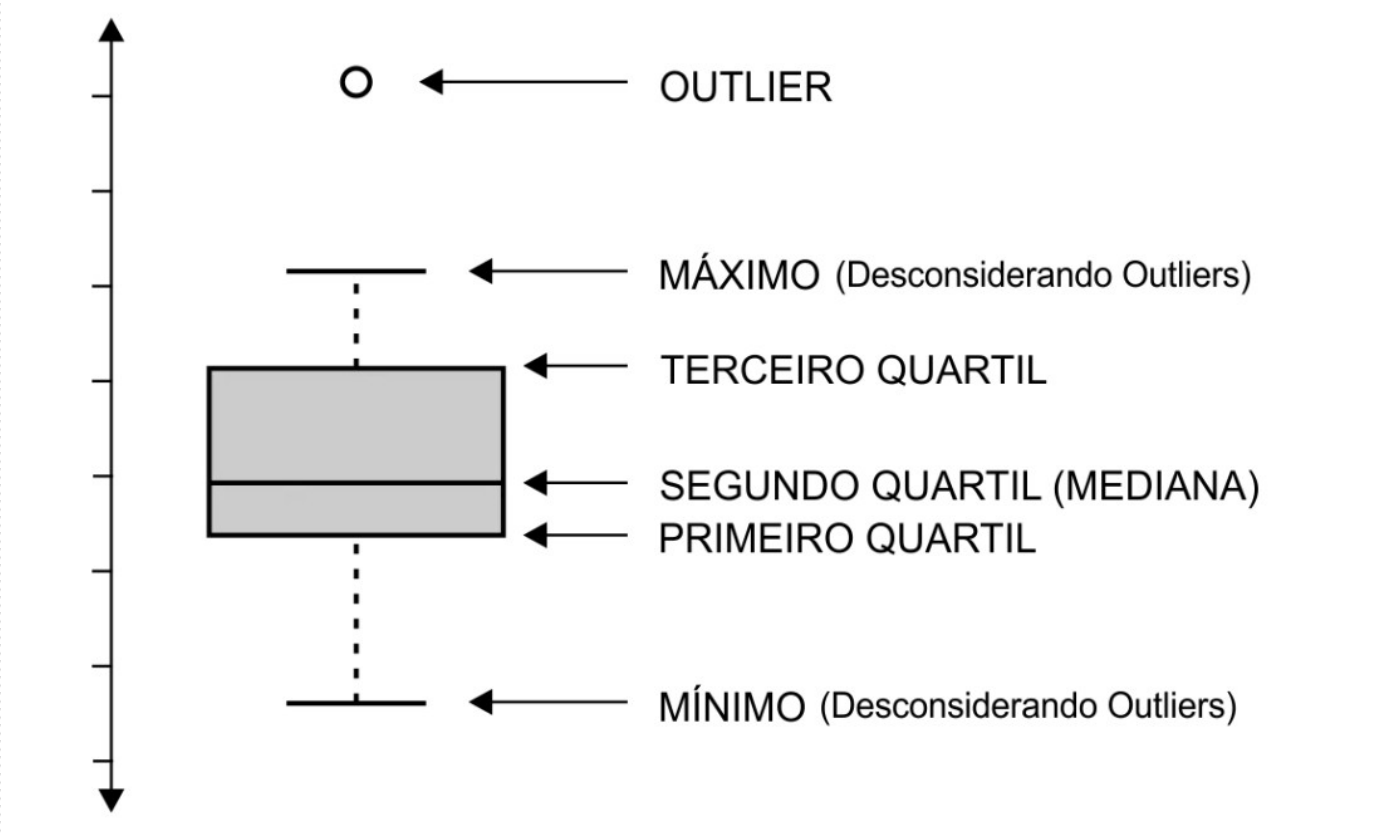


Gráfico de Caixas (Boxplot)

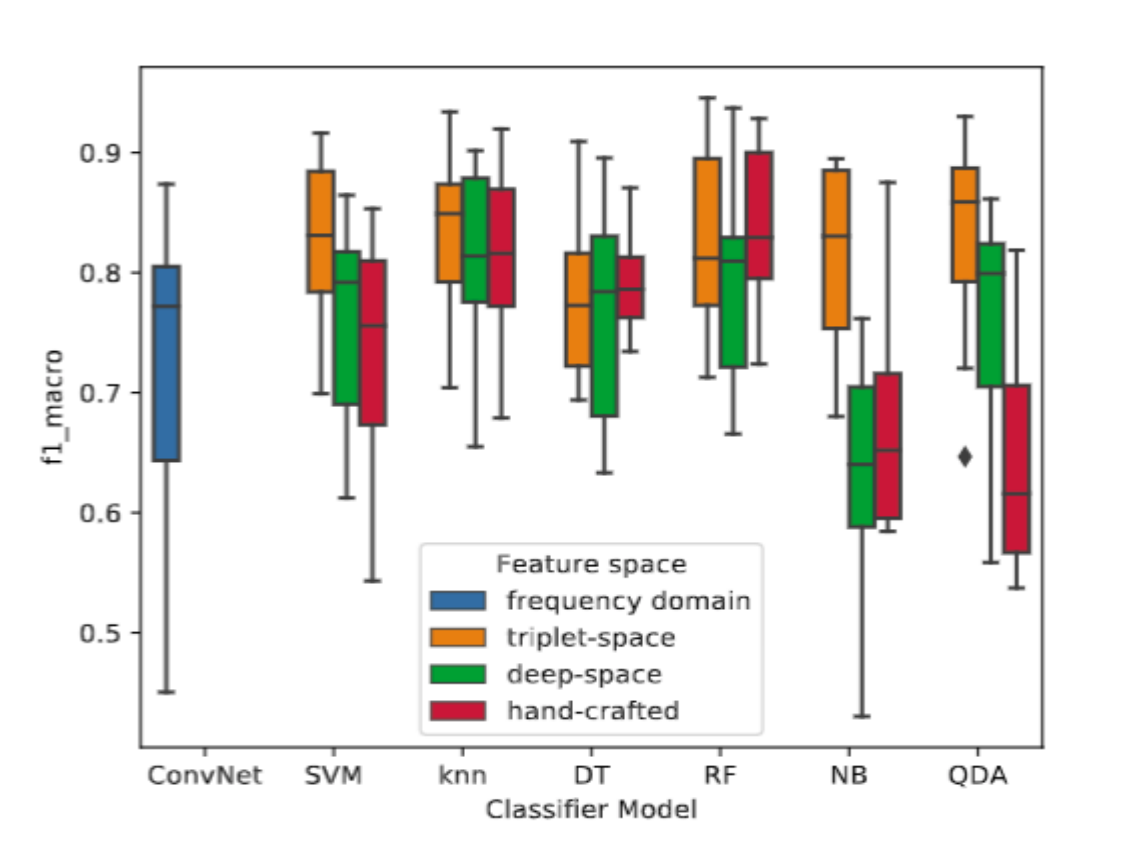


Gráfico 3D

- Permite visualizar 4 dimensões
  - Com uso de cor
  - Mais dificil interpretar
- Mais impacto visual do que representacional

Histograma 3D

Gráfico de Dispersão 3D

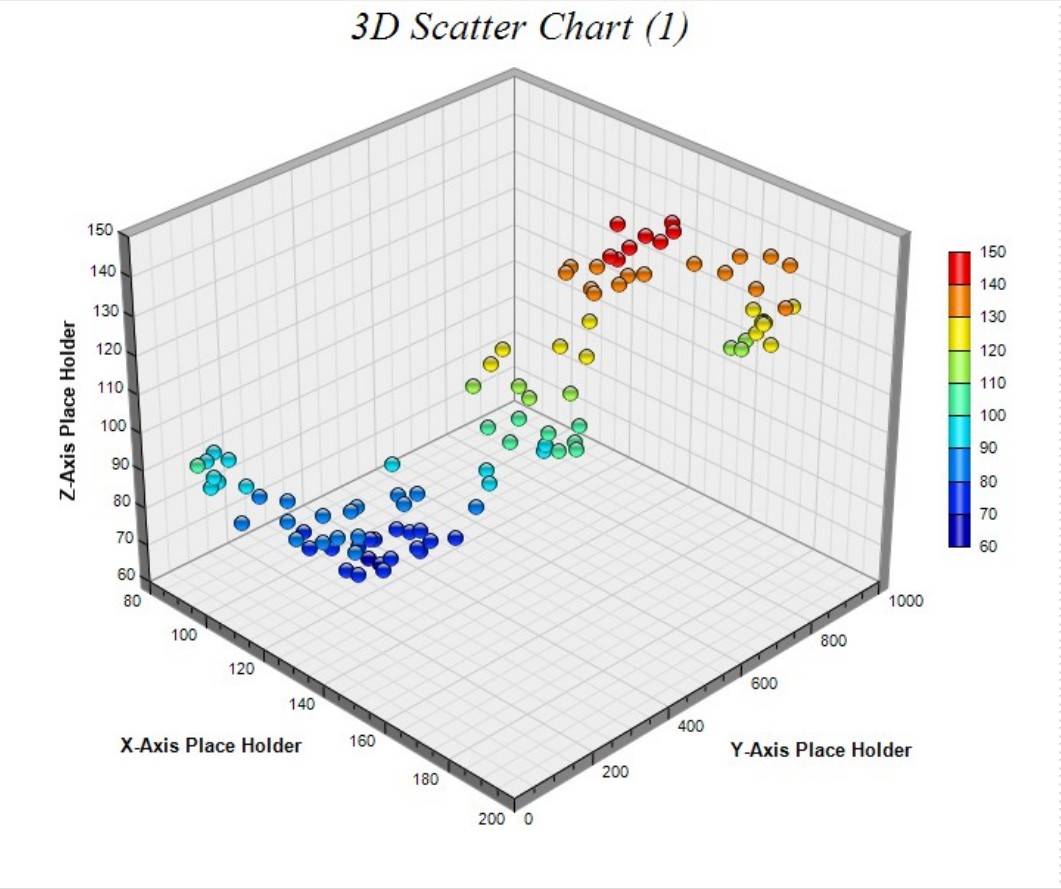


Gráfico de Linhas 3D