Lista 1

Davi Wentrick Feijó

2023-04-11

Exercicio de leitura:

Fazer um resumo de três parágrafos e não mais de uma página indicando semelhanças e diferenças entre Estatística, Mineração de Dados e Ciência de Dados. Você pode incluir referência a textos mais recentes sobre estes temas.

Estatística, Mineração de Dados e Ciência de Dados são áreas relacionadas ao processamento e análise de dados, mas com focos diferentes. A Estatística é uma ciência que envolve a coleta, análise e interpretação de dados para tomar decisões informadas e fazer inferências sobre populações a partir de amostras. A Estatística tem uma abordagem mais formal e matemática do que a Mineração de Dados e a Ciência de Dados, com ênfase em métodos estatísticos inferenciais, modelagem e análise de dados.

A Mineração de Dados é uma disciplina que envolve o processo de descoberta de padrões e relacionamentos interessantes em grandes conjuntos de dados. A Mineração de Dados tem um foco mais técnico e prático do que a Estatística e a Ciência de Dados, utilizando técnicas de aprendizado de máquina, reconhecimento de padrões e análise exploratória de dados para identificar informações úteis e insights.

A Ciência de Dados é uma área interdisciplinar que combina conhecimentos de Estatística, Mineração de Dados, Inteligência Artificial e outras disciplinas relacionadas para extrair conhecimentos e insights de grandes volumes de dados. A Ciência de Dados tem uma abordagem mais abrangente do que a Estatística e a Mineração de Dados, envolvendo todo o ciclo de vida de dados, desde a coleta e armazenamento até a análise e interpretação de dados. A Ciência de Dados é uma área que tem crescido muito nos últimos anos, com aplicações em diversas áreas, como finanças, marketing, saúde e tecnologia.

Em resumo, enquanto a Estatística tem uma abordagem mais matemática e formal, a Mineração de Dados é mais técnica e prática, e a Ciência de Dados envolve uma abordagem mais abrangente e interdisciplinar para análise e interpretação de dados. Todas as três áreas são importantes e complementares, e são usadas para resolver problemas e tomar decisões informadas em diversos campos.

Escolha uma área de pesquisa de interesse. Pesquise artigos publicados em revista indexadas e descreva (resumidamente) um exemplo incando o tipo de problema (ou problemas) entre os listados abaixo. Inclua referência bibliográfica e indique as características dos dados e estudo que relacionam ao tipo de problema (ou problemas) indicado.

(a) Análise multivariada clássica (n < p).
(b) Mineração de dados (Data Mining) (n elevado).
(c) Aprendizado estatístico (Statistical Learning) (p elevado).
(d) Reconhecimento de padrões.
(e) Data Science.

Descreva cada um dos problemas encontrados na análise multivariada, encontre um exemplo de caso real e indique uma falha nas técnicas estatísticas tradicionais.

(a) Mining (mineração, n muito elevado).

O grande número de observações (n) pode causar problemas nos testes de hipótese, que acabam sempre apresentando resultados significativos. Por exemplo, na mineração de dados imobiliários, ao coletar todas as variáveis de anúncios de imóveis em sites, pode-se obter um grande banco de observações. Entretanto, se for necessário realizar um teste de hipótese, o resultado sempre seria significativo.

(b) Scalability (escalabilidade).

(c) High Dimensional Data (dados em alta dimensão, $n > p$).

(d) Pequenas amostras.

A obtenção de boas estatísticas é mais difícil quando se possui pequenas amostras, devido ao grande erro amostral. Esse problema é comum em laboratórios, como no caso de análises de DNA, que possuem muitas categorias, mas poucas observações.

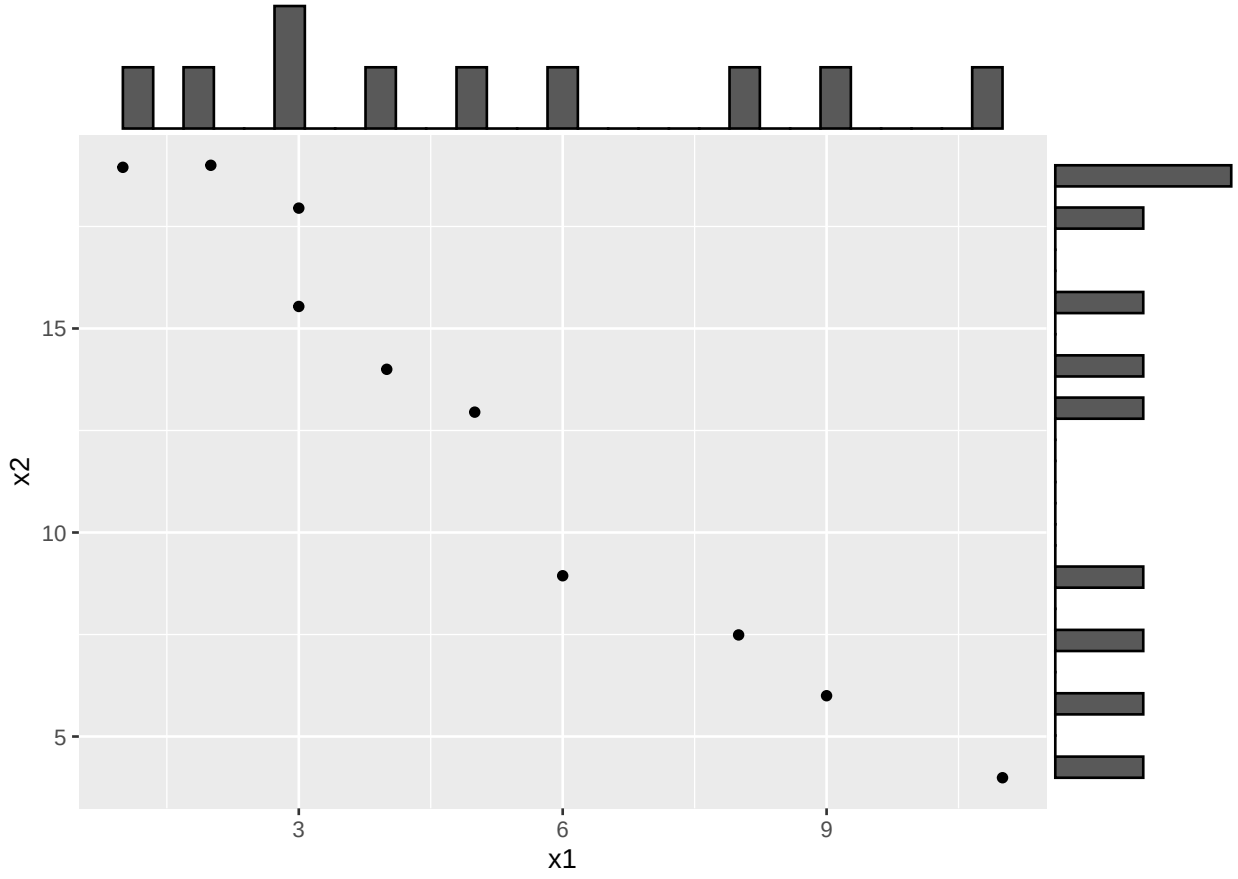
(e) Curse of Dimensionality (Problema de dimensionalidade).

Fazer os seguintes exercícios do capítulo 1 de Johnson e Wichern: 1.2, 1.6, 1.14 e 1.22.

1.2)

x1	x2
1	18.95
2	19.00
3	17.95
3	15.54
4	14.00
5	12.95
6	8.94
8	7.49
9	6.00
11	3.99

a) Construir um Scatter plot com graficos marginais



b) inferir o sinal da covariância da amostra a partir do gráfico de dispersão Podemos percebe que o sinal será negativo

c) Calcule a media amostral X1 e X2, as variancias amostrais S11 e S22. Calcule a covariancia amostras S12 e o coeficiente de correlacao da amostra r12. Interprete os resultados

```
c(mean(x1),mean(x2))
c(sd(x1),sd(x2))
print(cov(x1, x2, method = "spearman"))
cor(x1, x2, method = c("pearson"))
```

d) Mostre os vetores da média amostral, variancia-covariancia e correlacao da amostra.

```
## [1]  5.200 12.481
```

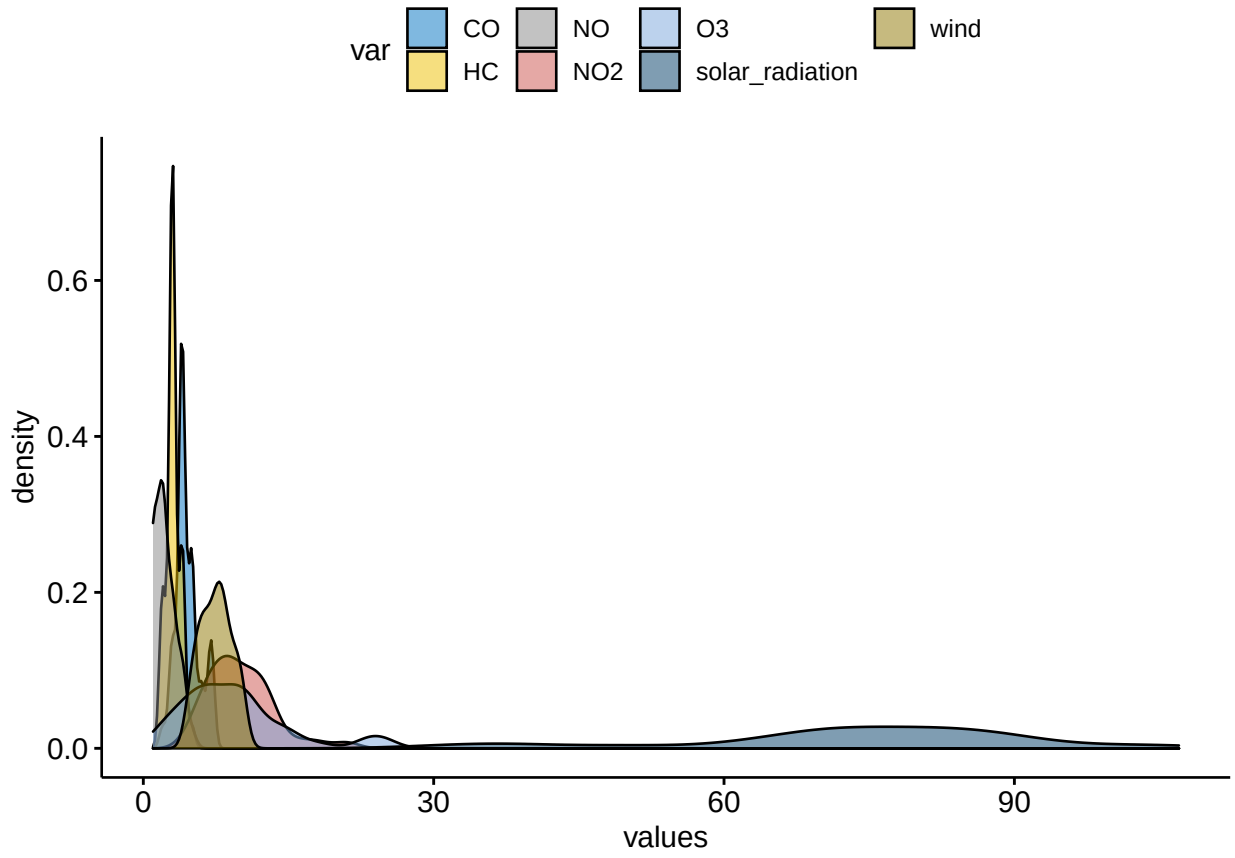
```
## [1] 3.259175 5.554671
```

```
## [1] -9
```

```
## [1] -0.9782684
```

1.6)

a) Construir um Scatter plot com graficos marginais para todas as variaveis

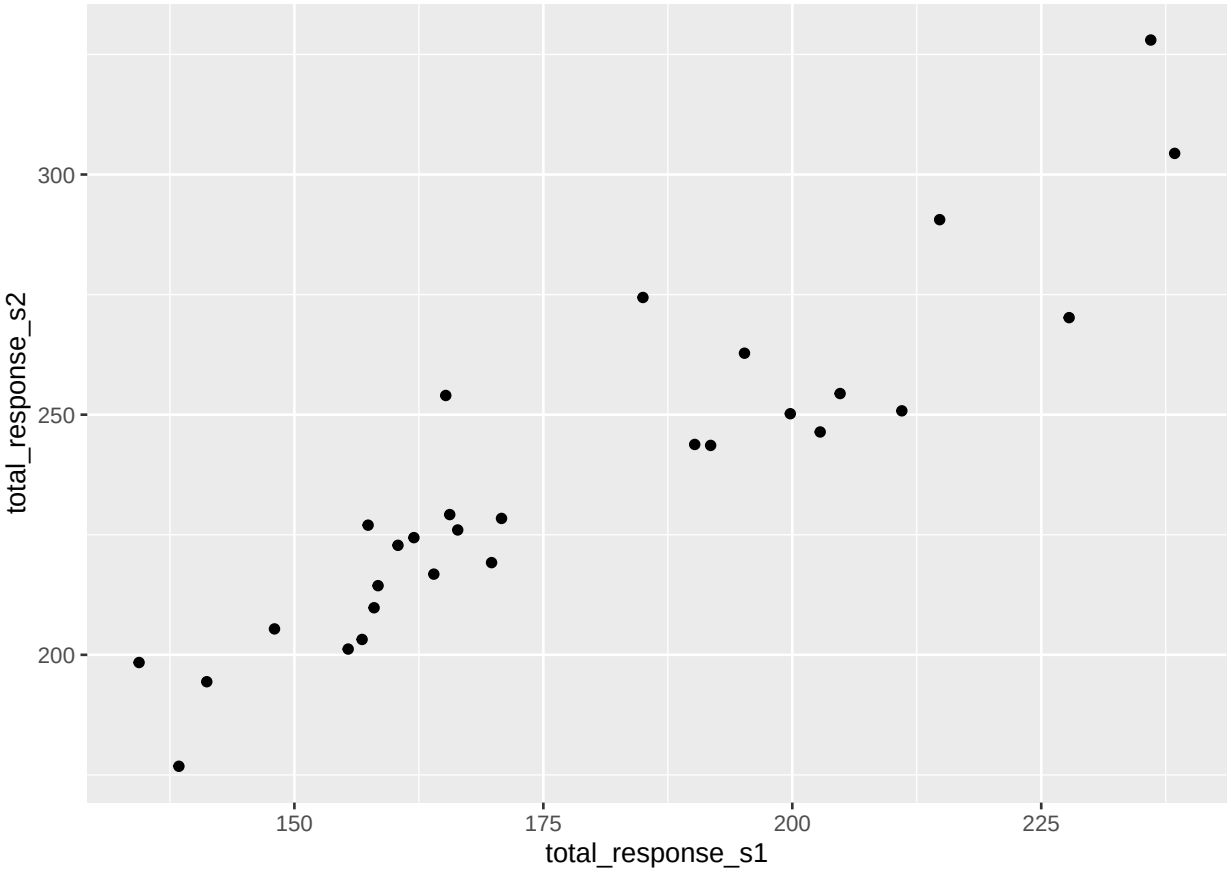


b) Construir os vetores da média amostra, variancia-covariancia e correlacao da amostra

var	mean	sd
CO	4.55	1.23
HC	3.10	0.69
NO	2.19	1.09
NO2	10.05	3.37
O3	9.40	5.57
solar_radiation	73.86	17.34
wind	7.50	1.58

1.14)

a) Construa um Scatter plot bidimensional para as variaveis x2 e x4 para o grupo de sclerose-multipla. Comente a aparencia do grafico



b) Construir os vetores da média amostra, variancia-covariancia e correlacao da amostra. Interprete a "pairwise correlation"

group	var	mean	sd
0	delta_s1	1.56	1.34
0	delta_s2	1.62	1.53
0	idade	37.99	16.66
0	total_response_s1	147.29	10.60
0	total_response_s2	195.60	13.61
1	delta_s1	12.28	17.81
1	delta_s2	13.08	18.74
1	idade	42.07	11.01
1	total_response_s1	178.27	29.06
1	total_response_s2	236.93	34.35

1.22)