SCC5836/SCC0252Visualização Computacional
M. Cristina
2023

LISTA DE EXERCÍCIOS 2

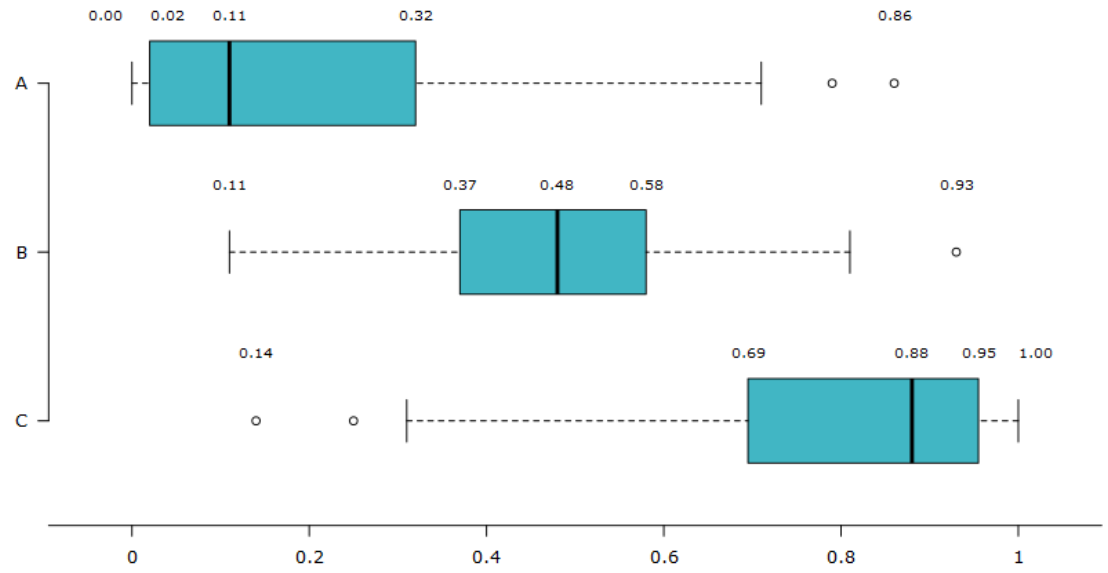
1. Considere o conjunto de dados a seguir: {1, 11, 15, 19, 20, 24, 28, 34, 37, 47, 50, 57}

   (a) Calcule a mediana, quartis inferior e superior.
   (b) Desenhe um *box plot* para esses dados.

2. Quais estatísticas você usaria para sumarizar o seguinte conjunto de dados, que descreve os preços de 11 produtos vendidos em uma loja popular? Desenhe um box plot para esses dados.
   {10, 10, 15, 5, 10, 10, 10, 10, 10, 10, 200}

3. Considere o gráfico abaixo, que sumariza por *box plots* 3 distribuições. Pergunta-se:
   Quais os valores mínimo, máximo, quartil inferior, superior e mediana de cada distribuição?
   Qual o IQR (Intervalo inter-quartil) de cada distribuição?
   Quais distribuições incluem outliers, e quais são os valores?
   O que se pode afirmar sobre a simetria das distribuições?

Chart 4.5.2.1
Box and whisker plots and five-number summaries of distributions A, B and C



4. Dê exemplo de uma visualização que pode ser distorcida pela ocorrência de *outliers* nos dados.

5. Considere o conjunto de dados da tabela abaixo. Apresente a tabela dos dados (a) após um pré-processamento de normalização (min-max) de todos os atributos numéricos; (b) após um processamento de *standardization* dos mesmos atributos; (c) Calcule as distâncias entre os pares de itens P10 e P15; P10 e P18; P10 e P22, usando a distância Euclideana e a similaridade do coseno; (d) qual destes produtos é o mais similar ao produto P10, segundo essas distâncias?

Produto	**Preço**	**Peso**	**Volume**	**Quantidade**
P10	10	70	100	8
P15	12	20	30	18
P03	30	15	30	20
P20	30	10	200	10
P07	10	60	50	8
P22	9	30	40	20
P31	15	30	35	30
P18	10	70	100	5
P30	20	20	30	3

6. Suponha que você coletou dados demográficos de estudantes do ICMC por meio de um questionário, em que pediu para os estudantes informarem: idade, altura e peso. Ao verificar os dados, observou que há várias ocorrências de valores ausentes ou errôneos para as três variáveis. Você gostaria de considerar o maior número possível de respostas. (a) Como você trataria os dados ausentes? Discuta. (b) Descreva duas estratégias que poderiam ser utilizadas para verificar a presença de outliers nesses mesmos dados.

7. Considere o conjunto de dados abaixo, que mostra dados de consumo de diferentes produtos em diferentes países do Reino Unido.

```
Food,England,Wales,Scotland,N  Ireland
Cheese,105,103,103,66
Carcass  meat,245,227,242,267
Other  meat,685,803,750,586
Fish,147,160,122,93
Fats  and  oils,193,235,184,209
Sugars,156,175,147,139
Fresh  potatoes,720,874,566,1033
Fresh  Veg,253,265,171,143
Other  Veg,488,570,418,355
Processed  potatoes,198,203,220,187
Processed  Veg,360,365,337,334
Fresh  fruit,1102,1137,957,674
Cereals,1472,1582,1462,1494
Beverages,57,73,53,47
Soft  drinks,1374,1256,1572,1506
Alcoholic  drinks,375,475,458,135
Confectionery,54,64,62,41
```

(a) Gere uma matriz de correlação entre os países: é possível identificar algum padrão?

(b) Aplique o PCA para obter os componentes principais (você pode usar o *sklearn*), e faça o gráfico PC1 x PC2 (chamado *score plot*): é possível observar algum padrão? Qual parcela da variabilidade desses dados é explicada por esses dois componentes? (não esqueça de aplicar a normalização de escala aos seus dados!)

8. Suponha que você não tem a função do sklearn para computer o PCA: escreva o código Python para computer o PCA utilizando a matriz de covariância/correlação (o sklearn utiliza SVD – Singular Value Decomposition). Verifique a correção comparando os resultados obtidos no conjunto de dados da questão anterior.