

Trabalho - Agrupamento - (5 pontos)

No contexto de produção animal, seja a seguinte base de dados contendo o consumo diário de água pelo grupo de animais (chamado de lote):

	lot_day	season	age	per_capita	lot_id	prod_id
0	3	inv	23.0	0.346664	615812	n_3
1	4	inv	24.0	0.708991	615812	n_3
2	5	inv	25.0	0.936537	615812	n_3
3	6	inv	26.0	0.722595	615812	n_3
4	7	inv	27.0	0.744303	615812	n_3

(base water_flow_by_day.csv)

Para cada dia de alojamento (lot_day), temos, na ordem, a estação do ano a qual se refere o dia, a idade dos animais do lote, o consumo médio de água por cabeça, um número identificador do lote e um número identificador do produtor/fazenda que produziu o lote. Uma fazenda pode produzir mais de um lote.

A base contém o consumo diário entre os dias 3 e 43 de alojamento dos animais. Como mostra a imagem.

A base pode conter dados faltantes, você deve lidar com isso. Perceba que o consumo de água aumenta ao longo dos dias, então talvez preencher com valor médio não seja a solução ideal (por exemplo, o dia 3, poderia ter valor muito alto, se usarmos a média de todos os dias para preencher seu dado). Uma média local pode ser mais adequada. Além disso, descartar lotes com dados faltantes pode resultar em uma base muito pequena, então cuidado ao usar essa estratégia.

Objetivos:

1. Você deve usar o consumo diário como variáveis para agrupar os lotes. Isto é, seu
x1=consumo de água por cabeça no dia 3 de alojamento
x2=consumo de água por cabeça no dia 4 de alojamento
...
xn=consumo de água por cabeça no dia m de alojamento
 - a. Deve escolher quais métricas para comparação usar
 - b. Qual número de grupos
 - c. Qual algoritmo ou variação do algoritmo
 - d. Como limpar ou lidar com os dados faltantes
 - e. Para cada grupo, deve mostrar as curvas de consumo pertencentes a um grupo e seu centróide

- Comparar o perfil dos lotes de cada grupo encontrado. Mostrar as comparações em gráficos e comentar sobre as diferenças ou semelhanças. Para isso, uma base de dados auxiliar é dada, com informações extras referente ao lote:

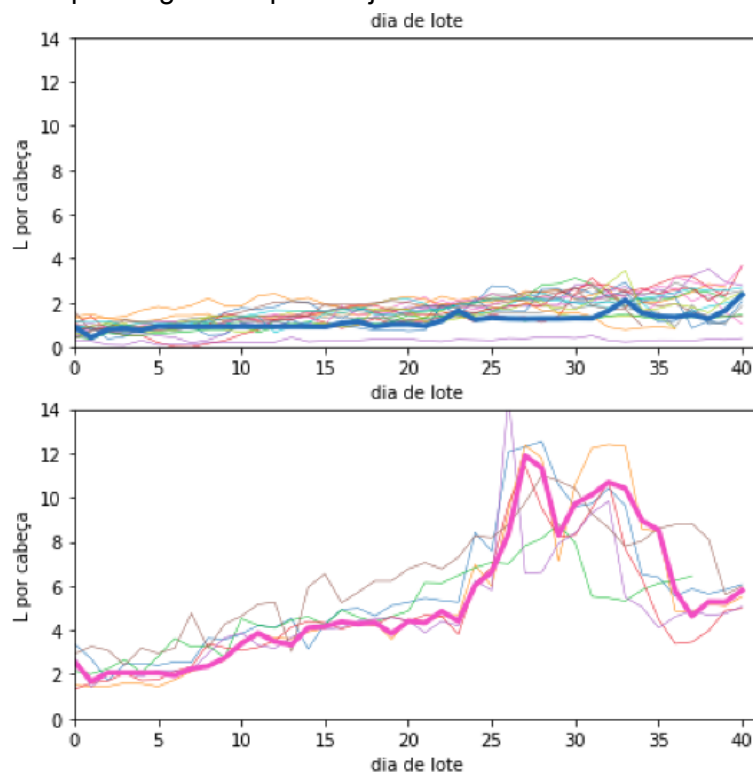
	lot_id	total_input	mortality	input_age	output_age	input_weight	output_weight	daily_weight_gain	daily_feed_per_animal	adjusted_feed_conversion_ratio	days_in_stage
0	615812	10536	2.107062	21.0	67.0	7.640801	22.516912	0.323394	0.459359	1.445263	47
1	542538	6243	3.315714	21.0	86.0	7.655272	24.382962	0.257349	0.287582	1.123649	66
2	542539	5926	3.307459	21.0	80.0	7.582211	24.420199	0.285390	0.661379	2.323259	60
3	562033	5931	8.497724	21.0	84.0	7.329978	23.202733	0.251948	0.000351	0.019366	64

(base lots_filtered.csv)

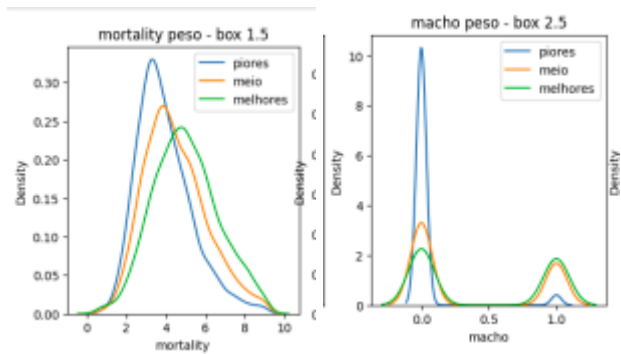
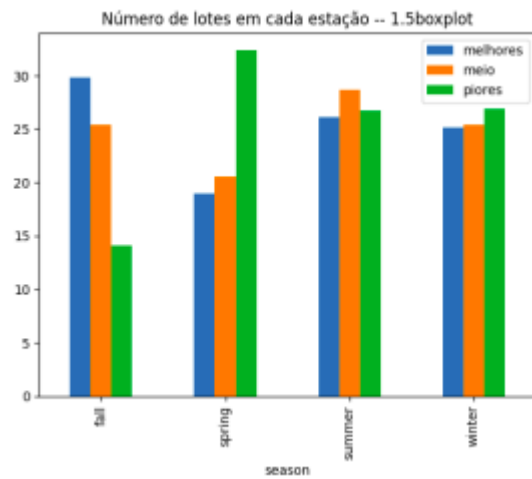
Em ordem: número identificador do lote, número de animais alojados, mortalidade em percentual, idade dos animais no momento no início do lote, idade dos animais no final do lote, peso médio dos animais no início do lote, peso médio dos animais no final do lote, ganho de peso médio diário, consumo médio diário de ração, taxa de conversão alimentar, dias de alojamento.

Da mesma maneira, podem haver outliers ou dados faltantes. Você não precisa usar todas as variáveis, pode usar as mais relevantes na comparação dos grupos.

Exemplo de gráficos para objetivo 1:



Exemplos de gráficos para objetivo 2:



Use a criatividade, são só inspirações.