EST0133 - Projeto I

Marcus Nunes

10 de Novembro de 2021

Instruções

- A data limite de entrega é 22/11/2021, às 23:59, via SIGAA
- O R é o único software permitido para coleta, limpeza e análise dos dados
- O trabalho deve ser feito em R Markdown, utilizando o arquivo `modelo.Rmd`
- Renomeie-o para `NomeSobrenome.Rmd`, em que `Nome` é o seu primeiro nome e `Sobrenome` é um de seus sobrenomes
- Envie conjuntamente seus arquivos `NomeSobrenome.Rmd` e `NomeSobrenome.pdf` para avaliação em um arquivo chamado `NomeSobrenome.zip`
- Respostas em arquivos que não estejam nos formatos Rmd e pdf não serão consideradas
- Identifique corretamente os eixos dos gráficos produzidos
- Respostas numeradas incorretamente não serão corrigidas
- Respostas com códigos supérfluos para a sua resolução, como pacotes desnecessários para as análises realizadas, terão pontuação descontada
- Não é permitido reportar resultados como capturas de tela (*screenshots*), a menos que sejam figuras

Parte I: Clusterização

O objetivo da primeira análise é identificar padrões em um conjunto de dados de jogadores de futebol presentes no game FIFA 2022. O arquivo `fifa.csv` possui informacões a respeito de 505 jogadores selecionados aleatoriamente. A descrição das colunas presentes no conjunto de dados está a seguir. A partir da coluna `crossing`, os valores possíveis para os atributos variam de 1 a 100[1]

- `id`: código de identificação do jogador (aleatório)
- `age`: idade (anos)
- `height`: altura (cm)
- `weight`: peso (kg)
- `club_number`: número da camisa no clube
- `national_number`: número da camisa na seleção
- `crossing`: cruzamento
- `finishing`: finalização
- `heading_accuracy`: precisão no cabeceio
- `short_passing`: qualidade do passe curto
- `volleys`: voleio
- `dribbling`: drible
- `curve`: efeito na bola ao chutar
- `fk_accuracy`: precisão na cobrança de faltas
- `long_passing`: qualidade do passe longo
- `ball_control`: controle da bola
- `acceleration`: aceleração
- `sprint_speed`: velocidade de corrida
- `agility`: agilidade
- `reactions`: reflexos
- `balance`: equilíbrio
- `shot_power`: força no chute
- `jumping`: pulo
- `stamina`: vigor físico
- `strength`: força física
- `long_shots`: qualidade dos chutes longos
- `aggression`: disposição para marcação
- `interceptions`: interceptação de passes
- `positioning`: posicionamento no campo
- `vision`: visão de jogo
- `penalties`: precisão na cobrança de pênaltis
- `composure`: habilidade de se manter calmo durante a partida
- `marking`: marcação
- `standing_tackle`: roubar a bola em pé
- `sliding_tackle`: roubar a bola com um carrinho
- `gk_diving`: habilidade do goleiro em pular para realizar uma defesa
- `gk_handling`: habilidade do goleiro em manusear e segurar a bola
- `gk_kicking`: habilidade do goleiro em chutar a bola
- `gk_positioning`: posicionamento do goleiro
- `gk_reflexes`: reflexos do goleiro

O objetivo desta análise é separar estes 505 jogadores em grupos similares, a partir das suas características físicas.

[1] Isto é, estas colunas podem, teoricamente, atingir quaisquer valores entre 1 e 100, mas não necessariamente algum jogador os tenha atingido nesta versão do jogo.

Questão 1

(*10 pontos*) Leia o conjunto de dados dentro do R em um objeto chamado `fifa` e verifique suas variáveis através dos comandos `head` e `summary`. Em um primeiro momento, sem realizar análise quantitativa alguma, é possível afirmar se existem uma ou mais variáveis deste conjunto de dados que deveriam ficar de fora da análise? Em particular, quais variáveis não irão acrescentar informação válida ao objetivo que desejamos completar? Justifique as suas escolhas. Em seguida, retire esta(s) variável(is) do conjunto de dados antes de prosseguir com a sua análise.

Questão 2

(*10 pontos*) Faça a PCA para este conjunto de dados, interpretando o resultado obtido. Essa interpretação deve ser numérica e visual.

Questão 3

(*10 pontos*) Os jogadores presentes no conjunto de dados foram selecionados a partir de um determinado número de posições no campo. Estas posições foram simplificadas a ponto de serem apenas Goleiro (G), Defesa (D), Meio-Campo (M) e Ataque (A). Não necessariamente todas elas estão representadas no conjunto de dados fornecido. Qual é o número de clusters que os métodos do cotovelo e da silhueta sugerem para estes dados?

Questão 4

(*10 pontos*) Combine os resultados das questões 2 e 3, isto é, visualize o resultado da PCA colorindo os pontos de acordo com a divisão em clusters pelo k-means com 2, 3, 4 e 5 grupos, sem inserir os loadings. Estas visualizações alteram aquilo concluído na Questão 3?

Questão 5

(*10 pontos*) Coloque as informações sobre a clusterização final no conjunto de dados original, sem as transformações aplicadas. Calcule a média, por cluster, das variáveis consideradas no estudo. Determine qual cluster equivale a cada posição no campo. Lembre-se que as posições no campo disponíveis para clusterizar os jogadores são

- Goleiro (G)
- Defesa (D)
- Meio-Campo (M)
- Ataque (A)

Não é necessário utilizar todos as posições na sua resposta. Algumas delas podem sobrar (ou não). Justifique a sua resposta.

Parte II: Webscraping

A página *List of countries and dependencies by population* da Wikipedia - https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population - exibe uma lista de países de acordo com a sua população atual ou estimativa mais recente. Além disso, a página *List of countries by population in 2015* - https://en.wikipedia.org/wiki/List_of_countries_by_population_in_2015 - exibe uma lista análoga, referente ao ano de 2015. O objetivo deste problema é juntar os dados destas duas páginas e analisar o crescimento populacional de acordo com as informações coletadas.

Questão 6

(*10 pontos*) Baixe a tabela principal disponível na página *List of countries and dependencies by population*. Mantenha apenas as colunas referentes ao país, região e população no final, renomeando-as como `pais`, `regiao` e `populacao`. Deixe a tabela limpa e pronta para ser analisada.

Questão 7

(*10 pontos*) Baixe a tabela principal disponível na página *List of countries by population in 2015*. Mantenha apenas as colunas referentes ao país, area e população no final, renomeando-as como `pais`, `area` e `populacao`. Deixe a tabela limpa e pronta para ser analisada.

Questão 8

(*10 pontos*) A quantidade de países disponível na tabela mais atual é diferente da quantidade disponível na tabela de 2015. Com isso em mente, crie um objeto chamado `tabela_final`, juntando os dados das duas tabelas e mantendo na tabela final apenas os países com dados de população disponíveis para os dois anos. Note que a tabela final deverá ter 5 colunas.

Questão 9

(*10 pontos*) Crie gráficos adequados para comparar a evolução da população atual em relação à população de 2015 por região. Os devem ser fáceis de interpretar, tanto em relação à utilização da área gráfica quanto às legendas. Descreva o que é possível perceber nesta visualização.

Questão 10

(*10 pontos*) Calcule a variação percentual entre os dados mais atuais em relação a 2015. Coloque em ordem decrescente a média destas variações por região, informando qual região teve os países com populações que mais cresceram, em média, durante este intervalo de tempo. Qual foi a variação média entre os países desta região?