Projeto de Análise e Limpeza de Dados

Enunciado:

Chegou o momento! O primeiro mini projeto que você, trainee, terá que realizar é uma limpeza e análise de dados de um dataset. Para isso, você utilizará todos os conhecimentos que adquiriu com o Turing Academy e com os cursos vistos para construir sua própria análise do zero, mas claro, sempre com ajuda e apoio de seus mentores!

Problema:

Em 2021, aconteceu mais uma edição das Olimpíadas. Esse evento ocorre normalmente uma vez a cada quatro anos e reúne grandes atletas de diferentes países para competirem entre si em diversas modalidades. Aproveitando a grande mídia cobrindo os jogos da edição deste ano, o Comitê Olímpico Internacional (IOC) decidiu fazer **um estudo em cima de todos os dados que eles possuíam de edições passadas a fim de obter insights interessantes** para divulgarem. Mais especificamente, o comitê busca através dessa análise **encontrar possíveis padrões** entre os **ganhadores de medalhas** que já participaram das edições dos jogos.

Para isso, sabendo da sua grande reputação como um grande cientista de dados, eles requisitaram **você** para executar essa tarefa. Sua missão é **limpar e organizar os dados** que eles te enviaram e **analisá-los**, anotando e descrevendo cada um dos seus passos e descobertas e, ao final, elaborar uma conclusão a respeito delas. Para isso, utilize os conhecimentos adquiridos nas aulas a respeito das bibliotecas de manipulação e análise de dados.

Informações sobre o dataset

Você pode baixar o dataset para o projeto aqui.

Esse dataset possui informações a respeito dos atletas que já participaram de edições passadas dos jogos olímpicos. Abaixo os detalhes de cada coluna:

1. **ID** - Um número de identificação único de cada atleta
2. **Name** - Nome do atleta
3. **Sex** - Gênero do atleta: M (masculino) ou F (feminino)
4. **Age** - Idade
5. **Height** - Altura em centímetros
6. **Weight** - Peso em kg
7. **Team** - Nome do time ao qual o atleta pertence
8. **NOC** - Nome do comitê olímpico nacional ao qual o atleta pertence, sempre será um código de 3 letras (BRA para Brasil, USA para Estados Unidos, etc.)
9. **Games** - Ano e época dos jogos
10. **Year** - Ano da edição que o atleta participou
11. **Season** - Estação na qual ocorreu os jogos Summer (verão) ou Winter (inverno)
12. **City** - Cidade onde ocorreu a edição dos jogos
13. **Sport** - Esporte do atleta
14. **Event** - Especificação a respeito da categoria do esporte (Ex. Futebol masculino, vôlei feminino, corrida 500m, etc.)
15. **Medal** - Medalha ganha pelo atleta: Gold (ouro), Silver (prata), Bronze, ou NA (nenhuma medalha)

Materiais de apoio

Caso seja necessário, não hesite em rever as aulas dadas no Turing Academy ou os cursos do Datacamp sobre os assuntos, e também não se esqueça de recorrer ao seu mentor sempre que tiver dúvidas ou precisar de ajuda!

Dicas

- ➔ Em uma análise é sempre necessário comentar todos os passos e descobertas, mesmo as que parecem “óbvias”, pois é importante para o leitor entender cada etapa
- ➔ Não se esqueça de sempre observar se as colunas estão com os data types corretos

- ➔ Para colunas categóricas verifique se os valores únicos fazem sentido
- ➔ Verifique também sempre se os valores mínimos e máximos das variáveis numéricas fazem sentido para cada coluna
- ➔ Os NaNs devem ser tratados apropriadamente, e nem sempre removê-los é a melhor solução
- ➔ Embora a limpeza seja importante, não foque apenas nela, divida bem seu tempo para fazer também uma análise bem feita
- ➔ Explore todos os tipos de gráficos possíveis na hora da análise e teste suas hipóteses com eles
- ➔ Caso precise de inspiração, veja análises de outras pessoas (o Kaggle é um bom lugar para isso)
- ➔ Por fim, lembre-se: O objetivo desse projeto é explicar o dataset! Foque em mostrar o que ele tem de interessante, pense em plots que podem tirar boas informações (Plots por ano/medalha, por exemplo, podem ser um bom começo) e em métricas interessantes de serem mostradas: Variância dos dados? Top maiores medalhistas? Entre outras