UNIVERSIDADE DE VASSOURAS – CAMPUS MARICÁ

Cristiano César Xavier Marinho
Mat. 202212004

Leandro Loffeu Pereira Costa
Mat. 202212089

Lucas Cerqueira Ferreira Carneiro
Mat. 202211189

Nilton Cezar Marins Brum Junior
Mat. 202211166

RELATÓRIO DO ESTUDO DA BASE DADOS HISTÓRICOS DAS OLIMPÍADAS

Maricá - RJ
2024

“Tenha em mente que tudo que você aprende na escola é trabalho de muitas gerações. Receba essa herança, honre-a, acrescente a ela e, um dia, fielmente, deposite-a nas mãos de seus filhos”.

Albert Einstein

Maricá - RJ

2024

SUMÁRIO

Maricá - RJ
2024

Introdução

O código Python realiza uma análise exploratória de dados relacionados a competições olímpicas, oferecendo uma interface gráfica intuitiva que permite ao usuário interagir com os dados e visualizar os resultados de forma eficaz. A aplicação emprega bibliotecas como Pandas para a manipulação de dados, Matplotlib e Seaborn para a visualização e Tkinter para a criação da interface gráfica.

Funcionalidades Principais

Carregamento de Dados: Esta funcionalidade possibilita ao usuário escolher um arquivo CSV que contém os dados das Olimpíadas. Exibição de países: Apresenta uma relação abrangente de todos os países registrados na base de dados.

Análise Exploratória: Conduz uma série de análises estatísticas e produz gráficos que permitem visualizar a distribuição de altura e peso, a relação entre essas variáveis, a contagem de medalhas por país e a matriz de correlação.

Filtragem de dados: possibilita ao usuário realizar uma segmentação dos dados por ano e país, promovendo uma análise mais aprofundada.

Remoção de outliers: Este método emprega a abordagem interquartil para identificar e eliminar outliers dos dados, o que resulta em uma análise de maior qualidade.

Interface Gráfica: A interface, desenvolvida com Tkinter, proporciona uma experiência de uso simplificada, permitindo ao usuário interagir com os dados de maneira intuitiva.

Maricá - RJ
2024

Estrutura dos Códigos

Os códigos estão estruturados em diversas funções, cada uma desempenhando uma responsabilidade específica:

Listar países: Esta função é responsável por carregar os dados provenientes do arquivo CSV e exibir uma lista de todos os países contidos na base de dados. selecionar arquivo: Possibilita ao usuário a seleção do arquivo CSV a ser analisado.

Processar_dados: Conduz a análise primária, abrangendo a filtragem, o tratamento de dados, o cálculo de estatísticas e a criação de gráficos.

Remove_outliers: Uma função auxiliar projetada para eliminar outliers dos dados.

Bibliotecas Utilizadas

Pandas: Utilizado para a manipulação e análise de dados em formato tabular.

Matplotlib: Biblioteca dedicada à geração de gráficos estáticos, como histogramas, gráficos de dispersão e mapas de calor.

Seaborn: Uma biblioteca de alto nível, fundamentada no Matplotlib, que otimiza a criação de gráficos sofisticados e visualmente atraentes.

Tkinter: Biblioteca padrão do Python para a criação de interfaces gráficas.

os: Biblioteca destinada à interação com o sistema operacional, permitindo, por exemplo, a verificação da existência de arquivos.

Cálculos Estatísticos Básicos

Os cálculos estatísticos mais comuns utilizados na análise exploratória de dados, como a realizada no código fornecido, são:

Média: Representa o valor central de um conjunto de dados. É calculado somando todos os valores e dividindo pelo número total de valores. No código, a média da altura é calculada por:

Python: media_altura = dados_filtrados['altura'].mean()

Desvio padrão: Mede a dispersão dos dados em relação à média. Valores altos indicam uma maior variabilidade nos dados. É calculado como a raiz quadrada da variância.

Python: desvio_padrao_altura = dados_filtrados['altura'].std()

Quartis: Dividem os dados em quatro partes iguais. O primeiro quartil (Q1) separa os 25% menores valores, o segundo quartil (Q2, ou mediana) separa os 50% menores valores, e o terceiro quartil (Q3) separa os 75% menores valores.

Intervalo interquartil (IQR): É a diferença entre o terceiro e o primeiro quartil. É utilizado para identificar outliers.

Remoção de Outliers:

Método interquartil: Um valor é considerado outlier se estiver abaixo de Q1 - 1.5*IQR ou acima de Q3 + 1.5*IQR.

Outliers podem distorcer os resultados da análise, especialmente quando se calculam estatísticas como a média e o desvio padrão. Removê-los ajuda a obter resultados mais representativos da maioria dos dados.

Visualização de Dados:

Histograma: Mostra a distribuição de frequência de uma variável contínua (como a altura), dividindo os dados em intervalos e contando o número de observações em cada intervalo.

Gráfico de dispersão: Mostra a relação entre duas variáveis numéricas (como peso e altura), permitindo identificar padrões e correlações.

Gráfico de contagem: Mostra a frequência de cada categoria de uma variável categórica (como o país), utilizando barras.

Matriz de correlação: Mostra a correlação entre todas as variáveis numéricas de um conjunto de dados, permitindo identificar quais variáveis estão mais relacionadas entre si.

Por que esses cálculos e visualizações?

Média e desvio padrão: Fornecem uma visão geral da distribuição dos dados.

Quartis e IQR: Ajudam a identificar outliers e a entender a dispersão dos dados.

Histograma: Visualiza a forma da distribuição dos dados, permitindo identificar se ela é normal, assimétrica, etc.

Gráfico de dispersão: Permite identificar se existe uma relação linear ou não linear entre duas variáveis.

Gráfico de contagem: Mostra a frequência de cada categoria, permitindo comparar diferentes grupos.

Matriz de correlação: Ajuda a identificar quais variáveis estão relacionadas e qual a força dessa relação.

Conclusão

A análise de dados nas competições olímpicas, através das bibliotecas Python, possibilita uma exploração única de padrões, tendências e insights relacionados ao desempenho atlético ao longo da história.

Essa abordagem não só enriquece nossa compreensão dos jogos, mas também fornece ferramentas para melhorar o futuro do esporte. Por meio da aplicação de técnicas estatísticas e visualização de dados, como o cálculo de estatísticas descritivas e a elaboração de diferentes tipos de gráficos, torna-se viável responder a questões fundamentais, como quais países lideram determinadas modalidades e de que forma o desempenho dos atletas evoluiu ao longo do tempo.

Os códigos apresentados neste trabalho demonstram a viabilidade de construir uma ferramenta para a análise exploratória de dados olímpicos. Utilizando bibliotecas como Pandas, NumPy, Matplotlib e Seaborn, é possível manipular e visualizar grandes volumes de dados de forma eficiente. No entanto, há ainda um grande potencial para aprimorar essa ferramenta, incorporando modelos de machine learning para realizar previsões, criando uma interface web mais intuitiva e integrando outras fontes de dados.

Com as devidas adaptações, essa ferramenta pode ser utilizada por pesquisadores, jornalistas e admiradores de esportes para aprofundar seus conhecimentos sobre as Olimpíadas, identificar novas oportunidades de pesquisa e até mesmo auxiliar na tomada de decisões estratégicas por parte de atletas, comitês olímpicos e patrocinadores.

Maricá - RJ
2024