

Relatório do Trabalho Final de Análise de Dados

Diego V. G. Dias, Gustavo H. S. Pinto, Matheus F. da Silva, Kauã. J. L. Melo

Instituto de Computação – Universidade Federal do Rio de Janeiro (UFRJ)

Caixa Postal 68.530 – 21.941-909 – Rio de Janeiro – RJ – Brazil

diegovgd@ic.ufrj.br, gustavohsp@ic.ufrj.br, mfelix@ic.ufrj.br, kauajlm@ic.ufrj.br,

Abstract. *This work presents the analysis and visualization of a public dataset, "International football results from 1872 to 2024," obtained from Kaggle. Using Python and libraries such as Pandas, Matplotlib, Wordcloud, Streamlit, and Datetime, we processed and explored the data to uncover trends and insights.*

Resumo. *Este trabalho apresenta a análise e visualização de um dataset público, "International football results from 1872 to 2024," obtido na plataforma Kaggle. Com o uso da linguagem Python e bibliotecas como Pandas, Matplotlib, Wordcloud, Streamlit e Datetime, foram realizadas a exploração e o processamento dos dados para identificar tendências e gerar insights.*

1. Introdução

Neste trabalho, um dataset público foi analisado de forma a gerar a visualização dos seus dados. O dataset **International football results from 1872 to 2024**, obtido por meio da plataforma **Kaggle**, foi escolhido para a realização da tarefa por apresentar uma grande variedade de informações disponíveis e um ótimo volume de dados para a análise. Esse dataset é, na realidade, formado por três arquivos independentes com informações sobre futebol: um primeiro arquivo sobre resultado de partidas, um segundo sobre os gols realizados por jogadores em diversas partidas e, finalmente, o último arquivo contém informações sobre disputa de pênaltis pelos países.

1.1. Ferramentas

Para a conclusão da tarefa, foi empregada a linguagem de programação Python, e suas bibliotecas:

- Pandas, uma ferramenta leve, flexível, gratuita e de código aberto que facilita a manipulação e análise de dados utilizando Python;
- Matplotlib, que é capaz de gerar visualização de dados em python de forma simplificada;
- Wordcloud, utilizado para gerar a visualização de ocorrências de dados em uma determinada coluna, evidenciando aqueles mais presentes;
- Streamlit, uma ferramenta utilizada para a criação de webapps e dashboards; e
- Datetime, biblioteca utilizada para lidar com datas.

2. Análise de Dados

2.1. Leitura e Exploração Inicial

✓
0s

```
[3] # Carregamento dos Arquivos
results = pd.read_csv('results.csv')
scorers = pd.read_csv('goalscorers.csv')
shoots = pd.read_csv('shootouts.csv')
```

```
[6] # Apresentação inicial do dataset
print(results.head())
print(scorers.head())
print(shoots.head())
```



Mostrar saída oculta

```
[9] # Entendendo o dado contido de uma forma geral
print(results.describe())
print(scorers.describe())
print(shoots.describe())
```



Mostrar saída oculta

✓
0s

```
[12] # Procurando valores nulos
print('\nResults\n', results.isnull().sum())
print('\nGoalcorers\n', scorers.isnull().sum())
print('\nShootouts\n', shoots.isnull().sum())
```



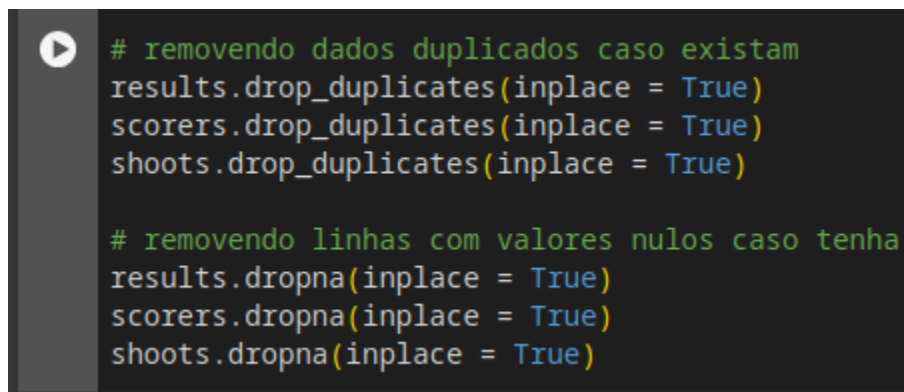
```
Results
  date      0
home_team  0
away_team  0
home_score  0
away_score  0
tournament  0
city        0
country     0
neutral     0
dtype: int64
```

O primeiro passo foi realizar o carregamento dos datasets com a função **read_csv()** do Pandas. Em seguida, foi realizada uma análise exploratória dos datasets, a fim de entender as principais métricas, como média, mediana e variância dos dados contidos. Após a leitura

dos arquivos, utilizou-se a função **head()**, a qual exibe uma parcela dos valores contidos no dataset. A utilização dessa função auxilia a compreensão do conteúdo do arquivo e dá um direcionamento sobre quais dados podem gerar insights interessantes. Nessa fase, também foram obtidas estatísticas sobre o dataset como as métricas de média, mediana e variância e a quantidade de valores nulos.

2.2. Limpeza e Tratamento

A limpeza de dados foi realizada para tratar valores ausentes e inconsistências. As etapas de limpeza envolveram a remoção de linhas com valores faltantes e a remoção de dados duplicados.

A code block with a dark background and light green text. It contains two sections of code, each preceded by a comment. The first section removes duplicates from 'results', 'scorers', and 'shoots' datasets. The second section removes rows with null values from the same datasets. The code uses pandas methods: drop_duplicates and dropna, both with inplace = True.

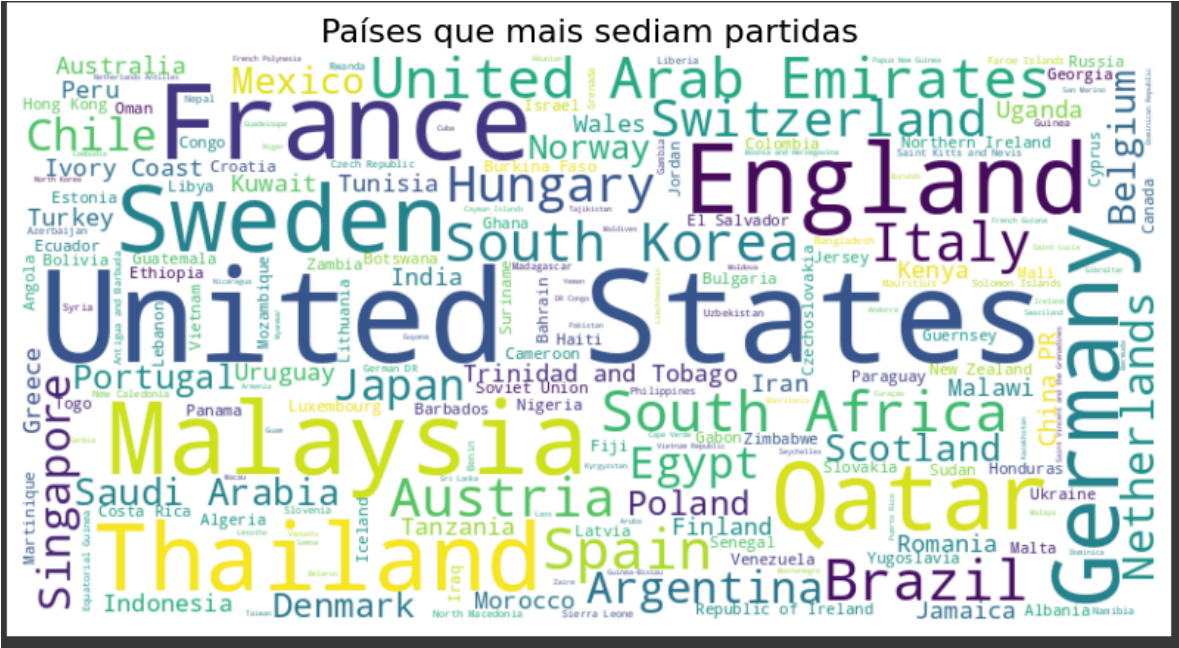
```
# removendo dados duplicados caso existam
results.drop_duplicates(inplace = True)
scorers.drop_duplicates(inplace = True)
shoots.drop_duplicates(inplace = True)

# removendo linhas com valores nulos caso tenha
results.dropna(inplace = True)
scorers.dropna(inplace = True)
shoots.dropna(inplace = True)
```

2.3. Análise Específica

Finalmente, com os dados limpos, foi possível se conduzir análises específicas a fim de identificar tendências. Nessa etapa do desenvolvimento, foram utilizados recursos gráficos, como tabelas, imagens e gráficos, que permitem a um usuário ter a visualização dos dados contidos no dataset.

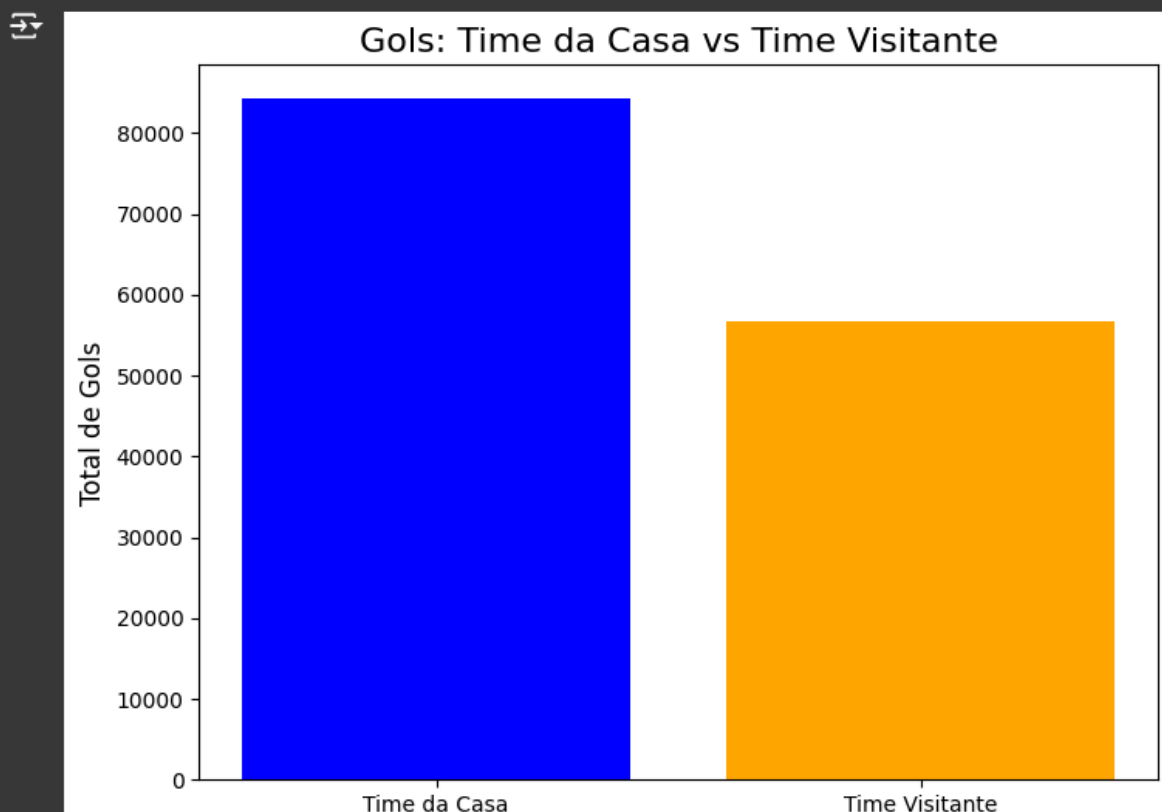
Países que mais sediam partidas



```
[ ] # Somar os gols do time da casa e do time adversário
    home_goals = results['home_score'].sum()
    away_goals = results['away_score'].sum()

    # Criar uma tabela com as somas
    goals_summary = pd.DataFrame({
        'Categoria': ['Time da Casa', 'Time Visitante'],
        'Gols': [home_goals, away_goals]
    })

    # Plotar o gráfico de barras
    plt.figure(figsize=(8, 6))
    plt.bar(goals_summary['Categoria'], goals_summary['Gols'], color=['blue', 'orange'])
    plt.title("Gols: Time da Casa vs Time Visitante", fontsize=16)
    plt.ylabel("Total de Gols", fontsize=12)
    plt.show()
```



Top 10 Times com Mais Gols e seu Artilheiro

Time	Gols	Vitórias	%	Maio Goleador	Qtd de Gols	% de Gols
England	2350	614.0	57.06	Harry Kane	58.0	2.47
Brazil	2278	665.0	63.58	Ronaldo	39.0	1.71
Germany	2268	588.0	57.99	Miroslav Klose	48.0	2.12
Sweden	2130	535.0	49.49	Zlatan Ibrahimović	44.0	2.07
Argentina	1990	579.0	54.99	Lionel Messi	55.0	2.76
Hungary	1988	465.0	46.92	Ferenc Bene	25.0	1.26
Netherlands	1779	440.0	51.34	Memphis Depay	35.0	1.97
South Korea	1773	531.0	53.31	Son Heung-min	25.0	1.41
Mexico	1732	501.0	51.12	Jared Borgetti	37.0	2.14
France	1658	463.0	50.66	Kylian Mbappé	38.0	2.29

3. Dashboard

O dashboard apresenta 5 formas de interação dinâmica com os dados contidos no dataset. Na barra lateral esquerda, é possível alterar o filtro da data para todos os datasets: qualquer informação buscada descartará todas as linhas com as datas indesejadas. Também é possível realizar baixar os datasets como um arquivo .zip a qualquer momento.

Início

1872/01/01

Fim

2024/12/31

Baixar os Datasets

A primeira visualização de dados oferecida, o “Grande Hall da Fama”, permite ao usuário descobrir quais são os países que mais venceram partidas. É possível filtrar por torneios específicos ou simplesmente ver todo o histórico. Por conta da grande quantidade de países que participam de certos torneios, uma seção foi adicionada para mostrar apenas os melhores, ainda podendo visualizar todos, caso desejado. A partir dessa análise, é possível se obter a quantidade de vitórias de um país ao longo do tempo.

Grande Hall da Fama

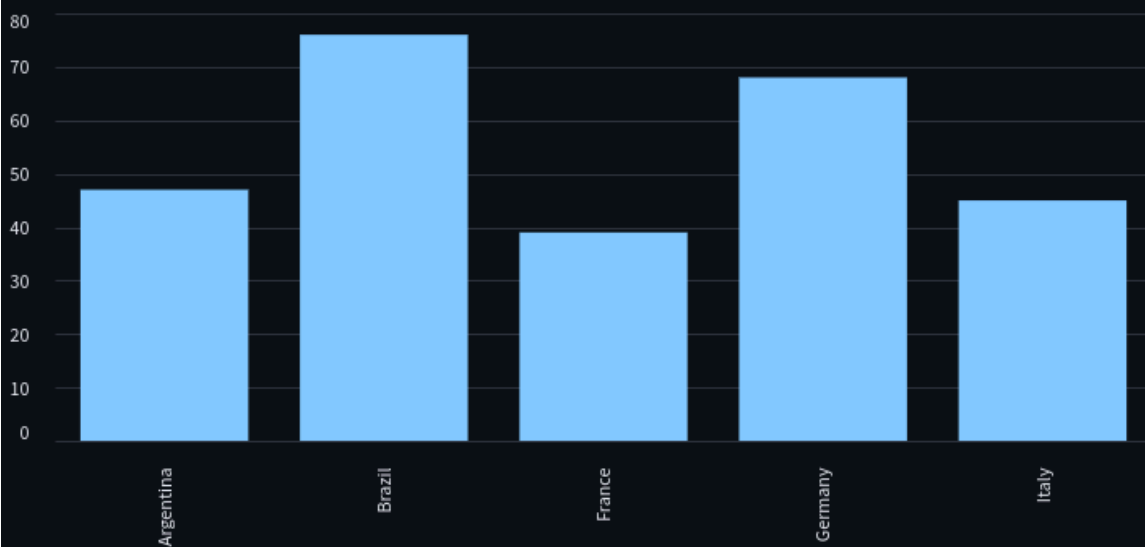
Descubra quais são os países com mais vitórias em cada torneio.

Torneio

FIFA World Cup

Países

5



A segunda parte do dashboard exibe informações mais detalhadas sobre cada país: é possível obter a quantidade de partidas jogadas, seu desempenho total, dentro ou fora de casa. O gráfico de pizza adicionado é gerado a partir da biblioteca Matplotlib. A interpretação desses dados pode ser relevante para cada seleção. Combinado com o filtro de data, pode-se obter o desempenho atual da seleção.

Análise de País

Utilize essa seção para obter um relatório sobre cada país.

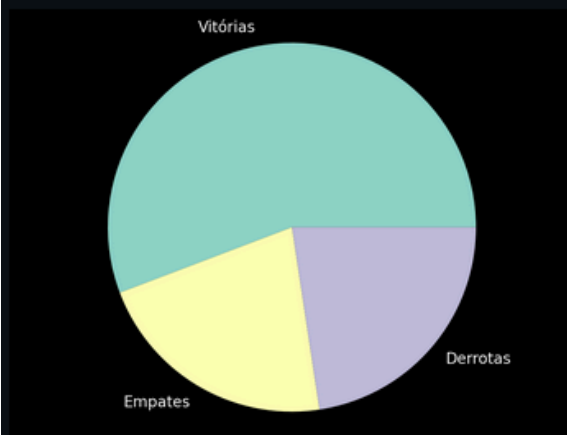
País

Scotland



Scotland tem um total de 1670 partidas jogadas. Sendo 392 vitórias, 181 empates e 262 derrotas.

Scotland em Casa



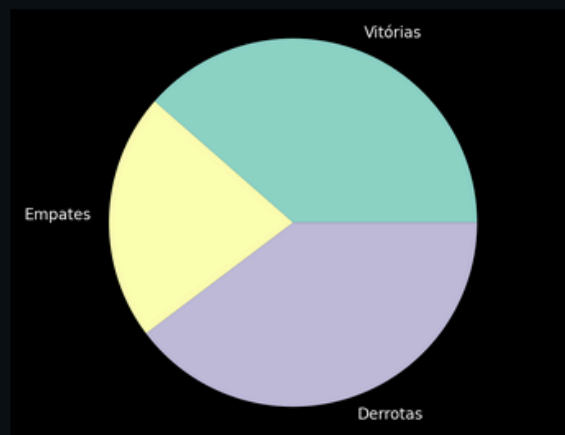
Scotland tem 407 partidas jogadas em casa.

Vitórias: 227

Empates: 88

Derrotas: 92

Scotland como Visitante



Scotland tem 428 partidas jogadas como visitante.

Vitórias: 165

Empates: 93

Derrotas: 170

Em terceiro lugar, a seção de artilheiros é capaz de mostrar os jogadores com mais gols registrados. É possível filtrar por países e limitar a quantidade de jogadores exibidos, ou simplesmente ver todo o histórico. A interpretação desses dados é relevante principalmente para equipes que buscam ótimos jogadores.

Artilheiros

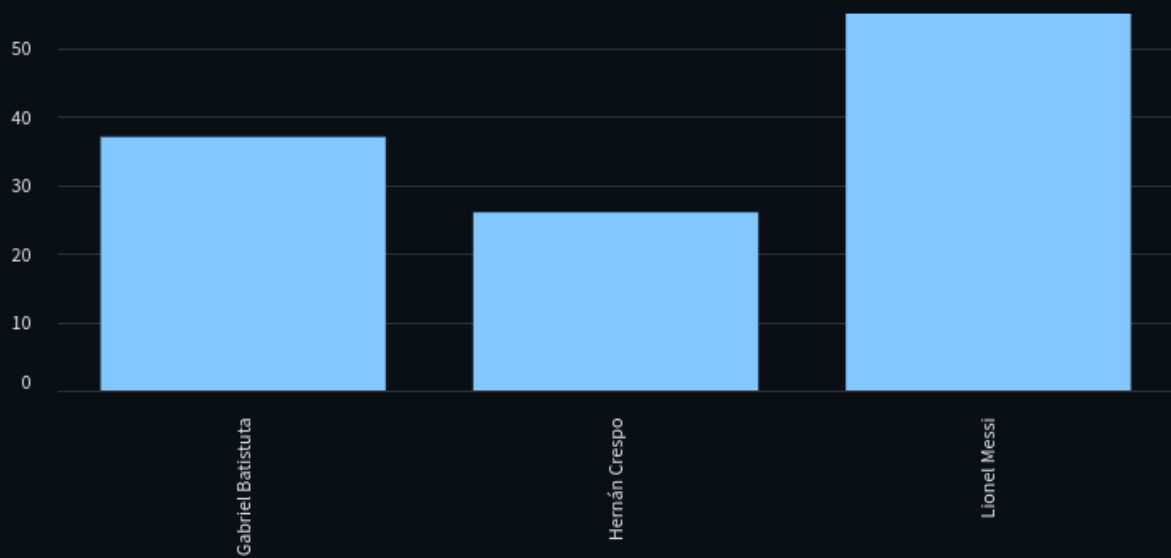
Os jogadores com mais gols em cada partida.

País

Argentina

Jogadores

3



Em seguida, a quarta seção disponibiliza o relatório dos gols de um jogador ao longo da sua carreira. É possível visualizar a distribuição dos gols ao longo da carreira dos jogadores.

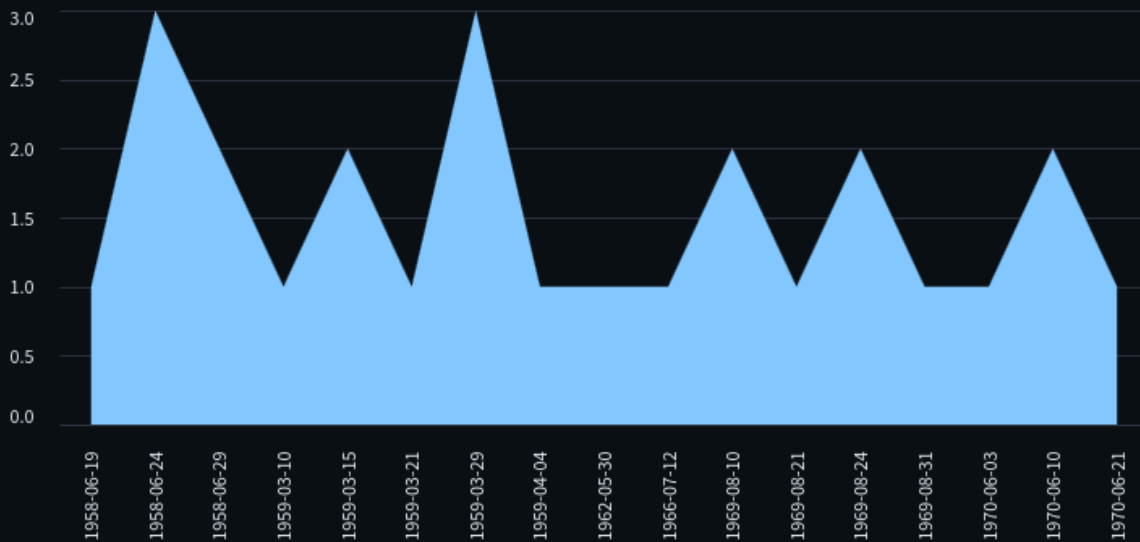
Avaliação do Jogador

Jogador

Pelé



Pelé tem um total de 26 gols em sua carreira



Finalmente, a última seção permite ao usuário visualizar as informações sobre decisões por pênalti de cada país. Novamente, um gráfico de pizza é utilizado para mostrar vitórias e derrotas dentro e fora de casa.

Decisões por Pênalti

Quantidade

1

Países

Brazil x

Brazil jogou decisões por pênalti 16 vezes.

Vitórias: 9

Derrotas: 7



4. Conclusão

Para concluir, é importante observar que os comentários são extremamente úteis na análise de dados, como pode ser percebido nos códigos apresentados. Além disso, compreender o tipo de dado é fundamental para evitar operações incorretas. Saber escolher o tipo de gráfico mais adequado para cada conjunto de dados é um processo que envolve criatividade e é essencial para proporcionar uma melhor compreensão das visualizações.