ENTREGA 2 - CMI104

Mariana Marques Cabral - GRR20205954

Murilo Stellfeld de Oliveira Poloi - GRR20185705
UFPR - Matemática Industrial

Introdução

Em 2022 tivemos a vigésima segunda edição da Copa do Mundo FIFA. A Copa do Mundo é sem dúvida um dos eventos mais notáveis no mundo esportivo e possui muita importância para a economia do país que sedia o torneio.

Apesar de ser um grande desafio para os times, é um desafio ainda maior para estatísticos e matemáticos pelo fato de que é díficil prever exatamente o que acontecerá em uma partida de futebol, ou de qualquer esporte. Até mesmo em partidas entre um time considerado muito fraco e um time considerado muito forte, ainda há a possibilidade da equipe favorita não sair vitoriosa, quem dirá prever placar, quem fará o(s) gol(s), entre outros.

Neste trabalho, vamos fazer uma abordagem básica, considerando apenas resultados e placares de todas as edições anteriores da Copa do Mundo, de 1930 à 2018 (as edições de 1942 e 1946 não ocorreram devido à segunda guerra mundial) e utilizaremos conceitos de matemática, estatística e especificamente aprendizagem de máquina para simular cada fase da edição de 2022 e verificar que o esporte é, na maior parte do tempo, imprevisível. Também iremos comentar sobre como tratamos a base de dados afim de fazer uma análise dos mesmos afim de obter alguma conclusão sobre qual ou quais times estão mais aptos a ganhar com base em número de gols, número de vitórias, entre outros.

Vale ressaltar que nem mesmo modelos mais complexos como modelos feitos por casas de apostas ou cientistas com bases de dados maiores conseguem prever os desfechos das partidas com tanta confiança, considerando que a previsão foi feita antes do ínicio do torneio.

Análise de Dados

Para a análise de dados das bases escolhidas, decidimos verificar as seguintes relações:

- Gols por edição;
- Gols por partida em cada edição;
- Partidas por edição;
- Gols% por time;
- Número de vitórias, derrotas e empates por time;
- Top 10 times que fizeram mais gols;
- Top 10 times que sofreram mais gols;
- Total de público por edição;
- Médias, mínimos e máximos gerais.

As análises feitas sobre os gráficos estarão no fim desta seção.

Gols por edição:

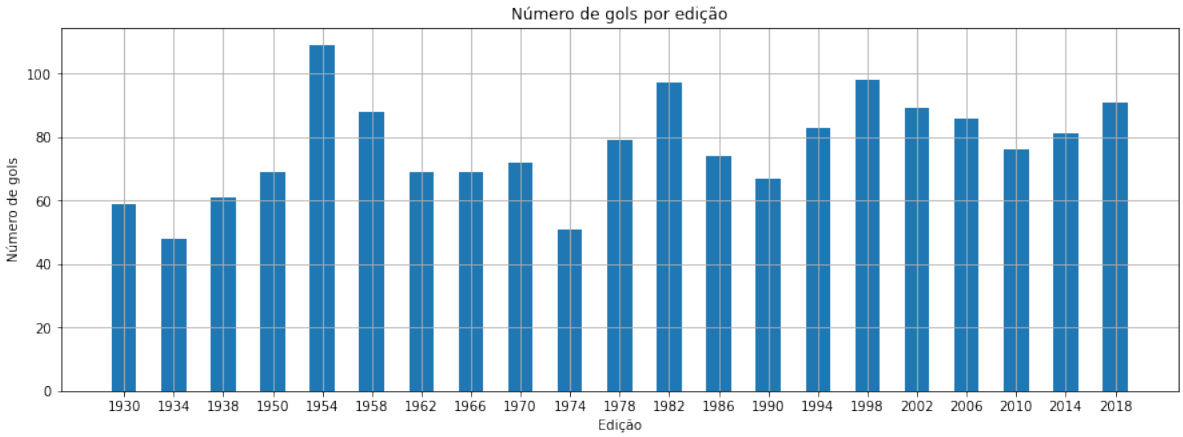

Figura 1: Número de gols por edição.

Gols por partida em cada edição:

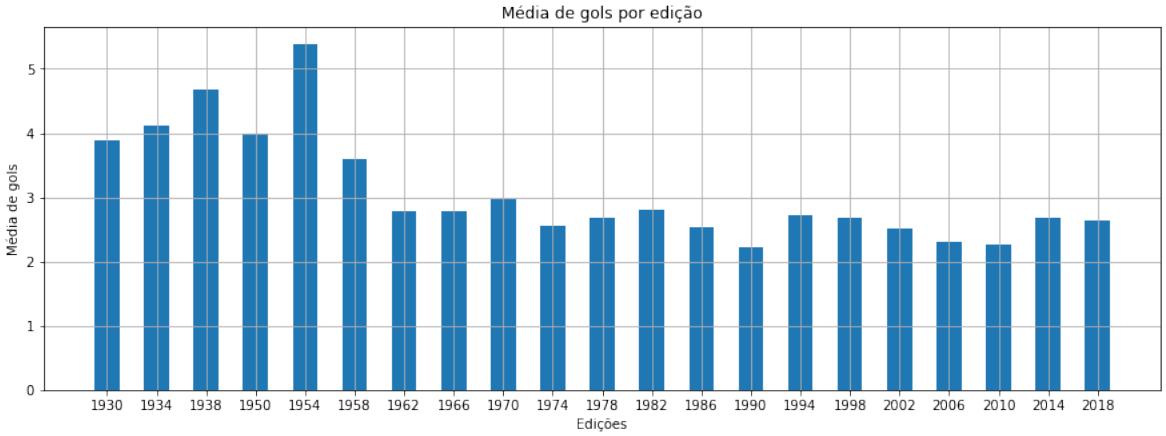

Figura 2: Média de gols por partida de cada edição.

Partidas por edição:

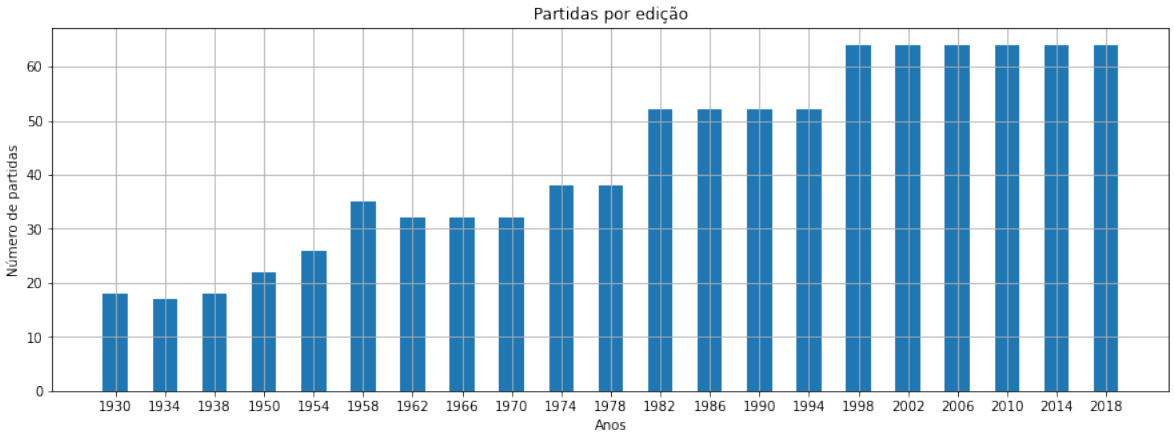

Figura 3: Número de partidas por edição.

Gols% por time:

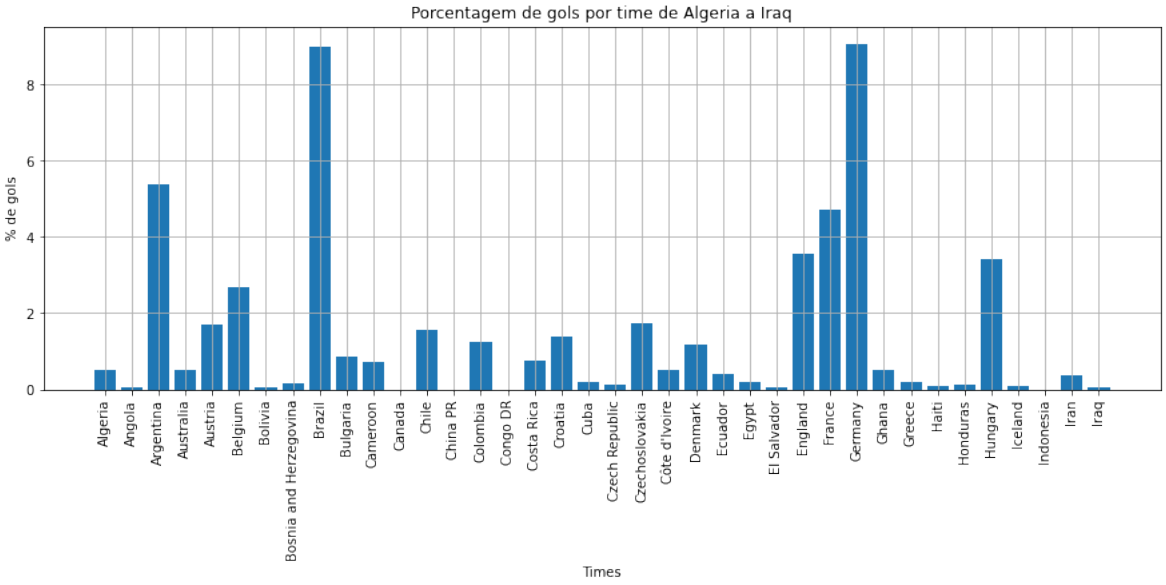

Figura 4: Porcentagem de gols por time de Algeria a Iraq.

Número de vitórias, derrotas e empates por time:

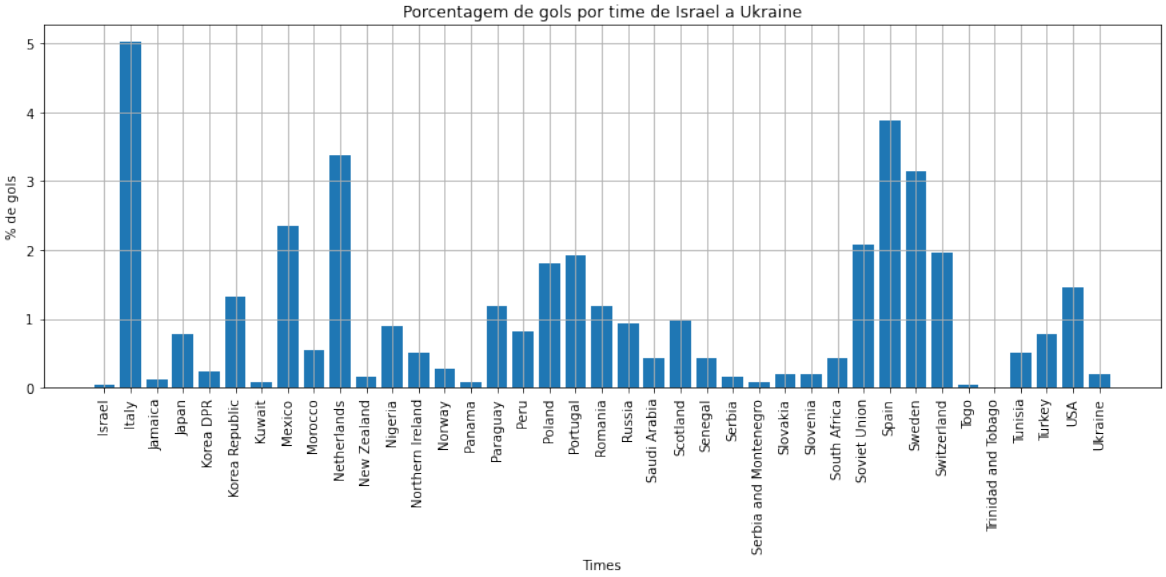

Figura 5: Porcentagem de gols por time de Israel a Ukraine.

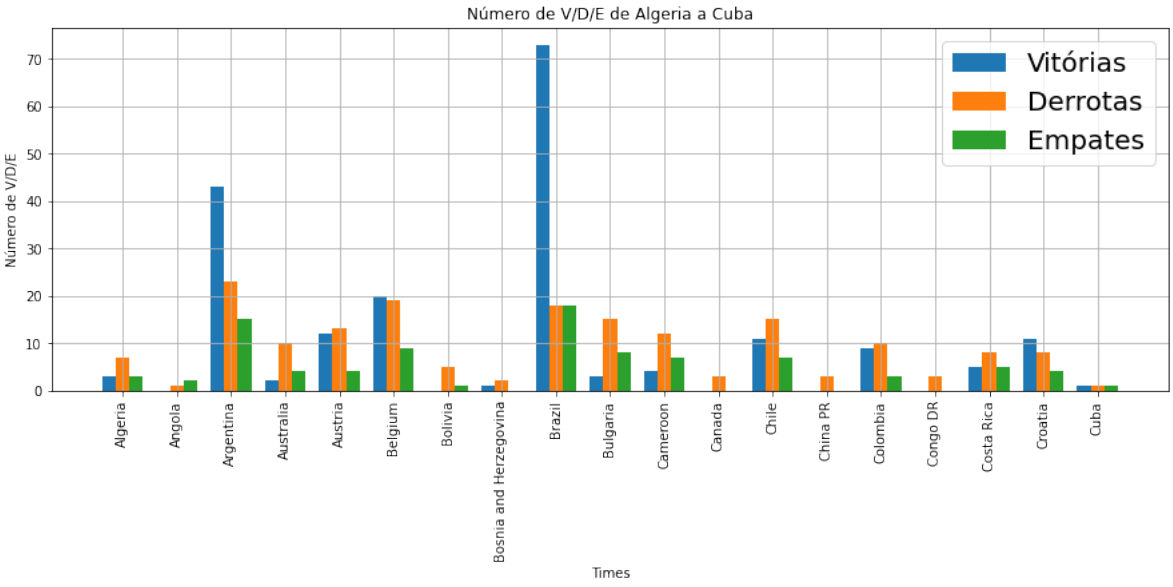

Figura 6: Número de vitórias/empates/derrotas por time de Algeria a Cuba.

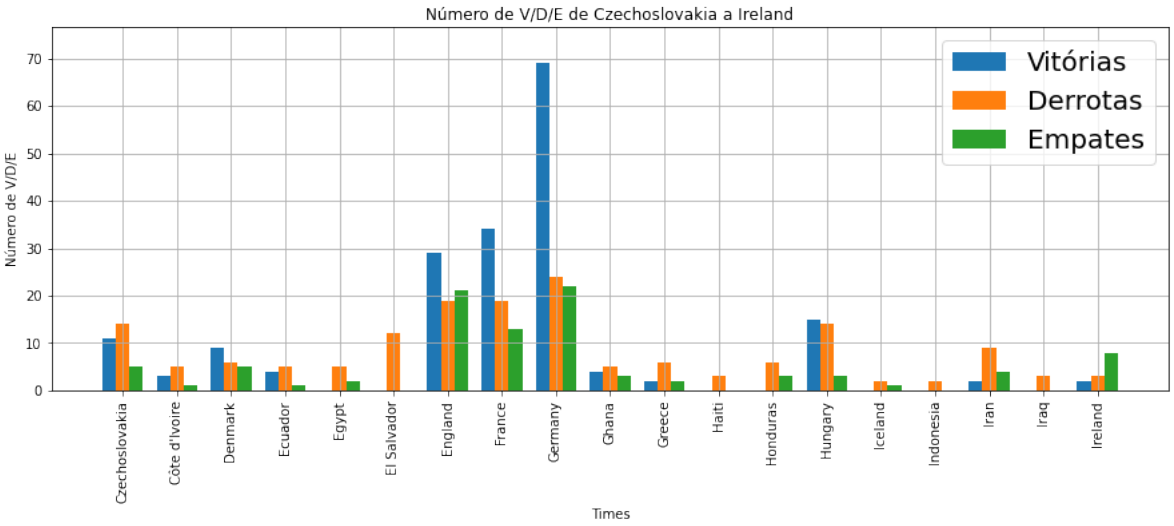

Figura 7: Número de vitórias/empates/derrotas por time de Czechoslovakia a Ireland.

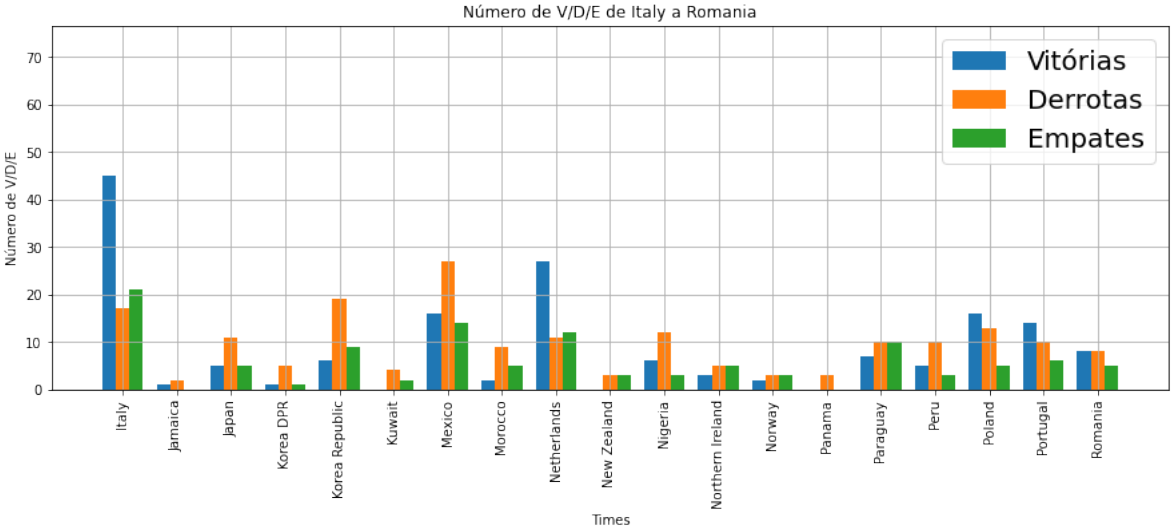

Figura 8: Número de vitórias/empates/derrotas por time de Italy a Romania.

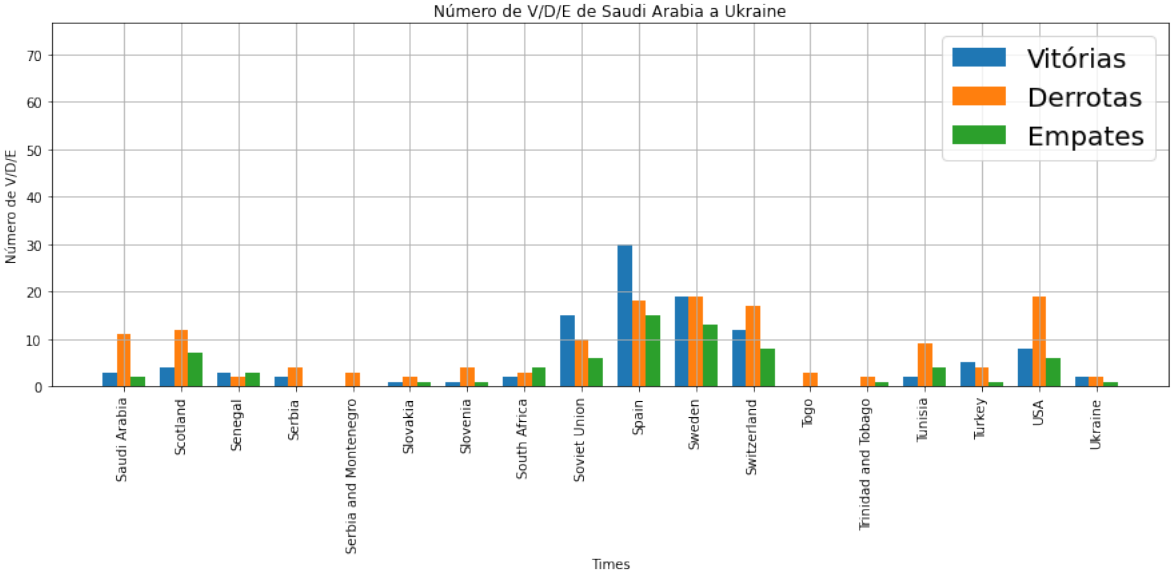

Figura 9: Número de vitórias/empates/derrotas por time de Saudi Arabia a Ukraine.

Top 10 times com mais gols:

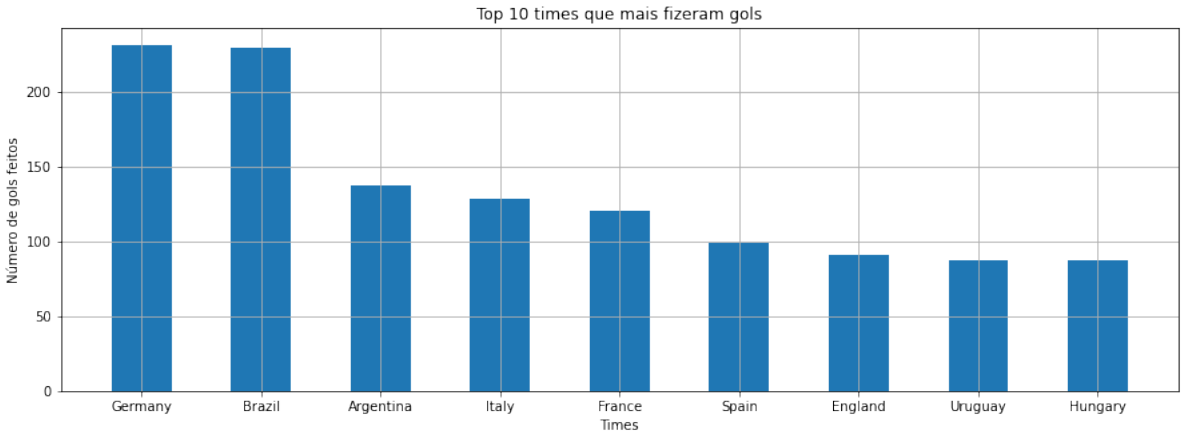

Figura 10: Os dez times que mais fizeram gols.

Top 10 times que sofreram mais gols

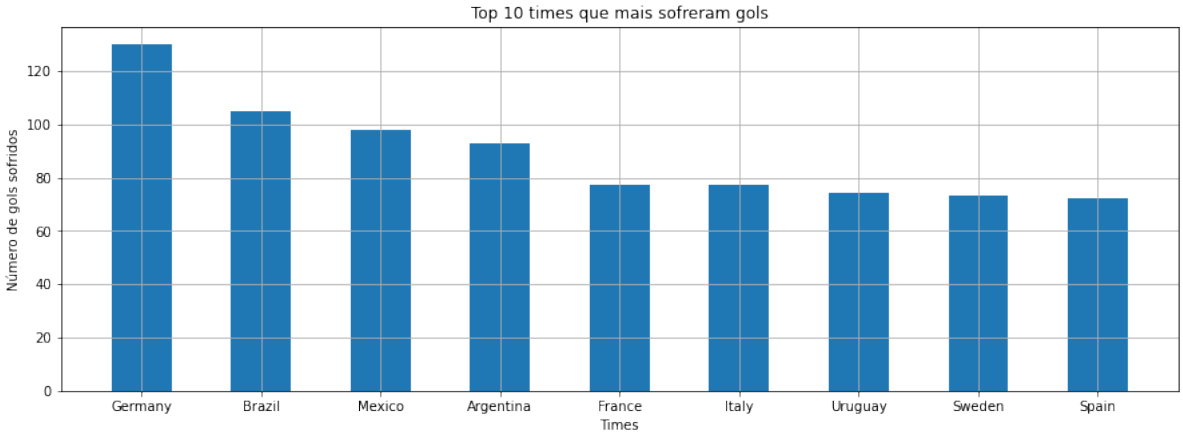

Figura 11: Os dez times que mais sofreram gols.

Total de público por edição:

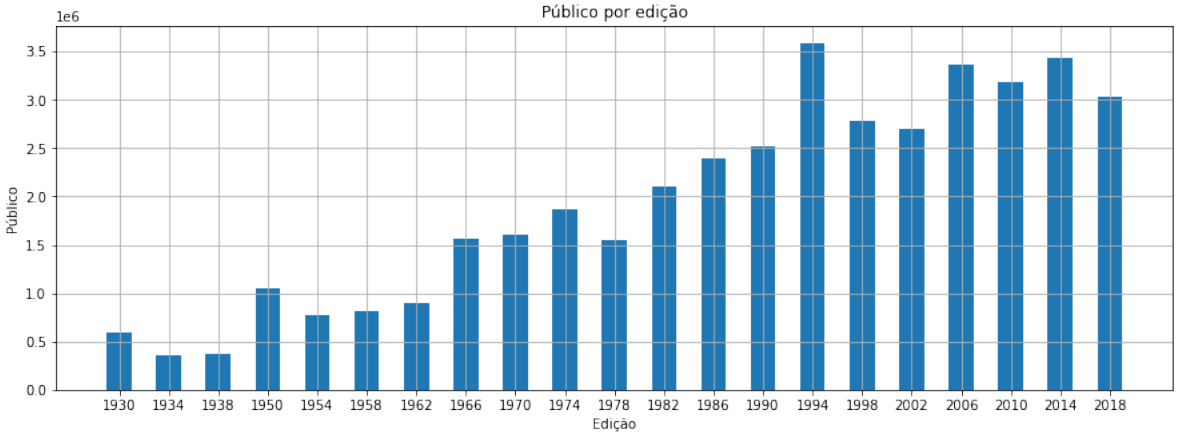

Figura 12: Total de público em cada edição.

Médias, mínimos e máximos gerais:

	Year	Home Team Goals	Away Team Goals	Attendance
mean	1986.916	1.796	1.036	45036.087
min	1930.000	0.000	0.000	2000.000
max	2018.000	10.000	7.000	173850.000

Figura 13: Descrição dos dados usando média, minímos e máximos.

É visível que o edição de 1954 foi o ano com mais gols e maior média de gols, mas o ano em questão foi apenas o quinto menor em termo de número de partidas por edição. Com isso, é possível dizer que havia um certo desnível entre as equipes: algumas tinham facilidade de fazer gol, enquanto outras tinham facilidade de tomar gol. Nas útlimas edições, viemos tendo seleções que sequer era esperado que passassem da fase de grupos chegando em fases eliminatórias e avançando nelas.

Em termos de quantidade ou média percentual de gols, o time da Alemanha ganha do Brasil por pouco. É notável que, por motivos políticos, a Alemanha foi dividida entre ocidental e oriental até o ano de 1990, mas só começou a participar como o país unificado que conhecemos hoje em 1994. Apesar de so ter tido uma participação em 1974, a alemanha oriental fez um total de 5 gols, o que foi suficiente para ficar a frente do Brasil.

O número de partidas tendeu a se normalizar, a partir de 1982 foi criado um sistema mais constante de jogos e o número de partidas somente aumentou a partir de 1998 devido ao aumento de equipes participantes.

Países como Brasil, Alemanha, França, Inglaterra, Itália, Espanha e Argentina tem, em geral mais de trinta vitórias cada, são países que num geral tem uma média de gols por time alta também, ou seja, é plausível dizer que partidas envolvem estes times há boas chances de haver(em) gol(s). É possível fazer a conexão também do número de vitórias baixo, ou de derrotas alto, com times que tiveram poucas participações e/ou gols.

Os países que mais fizeram gols costumam aparecer nos que mais sofreram também. Esse fato, provavelmente se dá também pelo número de participações de cada seleção.

A edição com mais espectadores nos estádios foi a de 1994, que ocorreu nos Estados Unidos. O país sede era o terceiro mais populoso na época.

Sobre as médias, minímos e máximos, é notável que no geral, os times que tem ”mando de campo” tendem a fazer mais gols que os times ”visitantes”. Podemos ver também o menor e maior público até 2018, assim como a média de público geral.

Pré-processamento e engenharia de características da base

Como mencionado na Entrega 1, a base de dados utilizada possuía somente dados até a edição de 2014 da Copa do Mundo, assim, foi necessário agregar os dados da edição de 2018. Como ambas bases não eram tão grandes, esse ajuste foi feito via planilha eletrônica.

Para tratar e manipular os dados utilizados neste projeto foi necessário remover colunas com informações julgadas desnecessárias do *dataset*, como por exemplo, data e horário, estádio, cidade, árbitros, entre outros.

Foi necessário arrumar os nomes de alguns dos times, pois estes continham erros de codificação (como foi o caso da Costa do Marfim), ou ainda, fundir ou alterar, por decisão própia, o nome de alguns países, por exemplo, ”Zaire” passou a ser chamado de ”Congo DR”, ”Dutch East Indies” de ”Indonesia” e ”Germany FR” e ”German DR” passaram a se chamar ”Germany”.

A base de dados utilizada e os tratamentos feitos nela, assim como a parte de análise de dados pode ser vista em `https://github.com/murlopoloi/CMI104/blob/main/Entrega2/Entrega2_Mariana_Murilo_CMI104.ipynb`.