Estatística e Futebol: Aspectos do Futebol Europeu

ME315 - Manipulação de Banco de Dados

Esther Cleveston - RA 250709
Felipe Basilio da Silva Santos - RA 238550
Leonardo Rodrigues Pinheiro - RA 225212
Marcos Vinicius de Carvalho Araujo - RA 242711
Nicole Fredericci do Amaral - RA 204355
Rafaela da Silva Barril - RA 204368

Campinas, 28 de março de 2025

Conteúdo

1 Introdução

O presente trabalho tem como objetivo o desenvolvimento de uma análise estatística utilizando dados de milhares de partidas de futebol, disponíveis no link `https://drive.google.com/file/d/14MfEJkNnCrjQf27t4mZTjuBFtzV8WeyK/view`. Para tal, foram feitas limpezas e integrações nas tabelas contidas no banco disponibilizado, e ao final de sua análise descritiva definiu-se três perguntas principais que poderiam ser respondidas a partir das observações coletadas:

1. Qual o gasto médio com os capitães dos times ao longo do tempo? Isso muda de acordo com sua posição?
2. Qual o desempenho dos jogadores de acordo com sua nacionalidade? Como o percentual de jogadores estrangeiros afeta o time?
3. Qual o tempo médio que um jogador permanece no mesmo time? Isso muda de acordo com seu valor de mercado?

Além de métodos estatísticos e técnicas de manipulação de bancos de dados, foram criados gráficos para auxiliar o entendimento da conclusão obtida a partir da análise de cada pergunta estabelecida.

2 Materiais e Métodos

2.1 Descrição do Banco

O banco disponibilizado contém 10 tabelas com dados de jogos de futebol, as quais abrangem sinteticamente as seguintes informações:

- Appearances: estatísticas dos jogadores em diferentes partidas (gols, cartões, minutos jogados).
- Club Games: estatísticas dos clubes em diferentes partidas (gols, oponentes, vitórias).
- Game Lineups: formação dos times por partida jogada.
- Games: dados dos oponentes e dos jogos.
- Transfers: data e valor das transferências de jogadores entre clubes.
- Players: informações sobre os jogadores.
- Player Valuations: valor de mercado dos jogadores.
- Game Events: acontecimentos importantes durante as partidas.
- Competitions: tipo e local das competições.
- Clubs: dados sobre os estádios, jogadores e o financeiro dos clubes.

2.2 Google Colab

Para o desenvolvimento da atividade, o Google Colab foi utilizado como ambiente de execução. Essa plataforma online permite a criação de projetos no navegador, processamento em nuvem e possibilita que diversos usuários editem o mesmo arquivo. Para a utilização desse recurso as configurações padrões foram mantidas e o tipo de ambiente de execução foi alterado para "R".

2.3 Linguagem R

A Linguagem R consiste em uma linguagem de programação amplamente utilizada para a realização de análises estatísticas, de modo que seus pacotes permitem acrescentar funcionalidades úteis para análises mais complexas e melhorias na visualização de gráficos. Nesse contexto, para o trabalho em questão, utilizamos os seguintes pacotes com os seguintes objetivos:

- **readr:** importação e leitura dos bancos de dados;

- **tidyverse:** criação de gráficos personalizados utilizando o ggplot - pacote pertencente ao tidyverse;
- **sf:** desenvolvimento de análises com dados geoespaciais com integração com outros pacotes.
- **maps:** criação e visualização de mapas geográficos com os contornos das regiões.
- **hrbrthemes:** personalização de gráficos feitos pelo ggplot com temas e estilos.
- **countrycode:** padronização de nomes de países para um mesmo idioma.
- **RSQLite**: conexão e manipulação dos bancos de dados no SQLite. Essa ferramenta permite trabalhar com os dados sem ocupar a memória do computador, dado que o RSQLite implementa um mecanismo de banco de dados leve e sem servidor para a manipulação dos dados.

2.4 Pré-processamento de Dados

Inicialmente, para agilizar o processo de instalação de bibliotecas no Google Colab, utilizamos um código fornecido em aula pelo professor que reduz significativamente o tempo de instalação de pacotes do R. Em seguida, conectamos o Google Colab ao Google Drive, onde os arquivos estão armazenados.

Dessa forma, com os arquivos acessíveis diretamente no Google Drive, removemos a compactação do arquivo zip, carregamos os dados no formato CSV e, por fim, conectamos esses dados a um banco de dados SQLite para o gerenciamento e a consulta das informações.

2.5 Manipulação dos Dados

2.5.1 Parte 1

Para realizar a análise da primeira pergunta, algumas manipulações foram necessárias nos bancos de dados. Inicialmente, identificamos que as informações relevantes para a análise eram: valor de mercado em euros, ano e posição do capitão do time - dados presentes no banco "Player Valuations" e "Game Lineups". Dessa forma, como queríamos as informações tanto para as posições no geral quanto para casos mais específicos, optamos, a princípio, por elaborar os códigos de forma separada.

Durante o processo, surgiram ajustes necessários: ao converter o CSV para o RSQLite, as datas não são preservadas e sofrem alterações, assim, para corrigir, selecionamos o valor apresentado no RSQLite, dividimos por 365, somamos 1970 e extraímos os 4 primeiros dígitos do resultado obtido já que nossa variável de interesse era o ano em questão. Dado

que as variáveis de interesse estavam presentes em bancos diferentes, as tabelas "Player Valuations"e "Game Lineups" foram unidas no código SQL pela variável "player_id"em comum. Ademais, usamos funcionalidades do SQL para selecionar os jogadores que eram capitães dos times e remover os dados faltantes nas categorias "valor de mercado", "data" e "posição". Em seguida, agrupamos os dados nesse mesmo código por ano para responder à pergunta de interesse.

É importante mencionar que, em relação às partes feitas em códigos separados, a primeira abrangia todas as posições, logo todas foram denominadas como "Geral". No segundo caso, dado que existiam muitas posições, foi realizado um agrupamento para melhorar a visualização, assim, as posições "Centre-Back", "Left-Back", "Right-Back", "Defender"e "Sweeper" foram classificadas como "Defesa", "Defensive Midfield", "Centre Midfield", "Central Midfield", "Attacking Midfield", "Left Midfield", "Right Midfield"e "Midfield"como "Meio de Campo", "Second Striker", "Attack"e "Centre-Forward"como "Atacante", "Right Winger"e "Left Winger" como "Lateral"e "Goalkeeper" como "Goleiro".

Por fim, as duas tabelas foram unidas, pequenas conversões foram realizadas para que as variáveis estivessem na classe adequada e identificou-se que a elaboração de um gráfico de linhas se mostrava bastante apropriada para mostrar o comportamento do valor do mercado ao longo do tempo.

2.5.2 Parte 2

Para a segunda pergunta, a análise foi divida em duas partes: o desempenho dos jogadores por nacionalidade e o desempenho dos clubes de acordo com a porcentagem de jogadores estrangeiros na composição de seus times.

Em primeira instância, as tabelas "Appearances" e "Players" foram unidas pelo código de identificação dos jogadores e criou-se um sistema de pontuação para classificar o desempenho dos mesmos em uma escala passível de comparação. Para tanto, cada minuto em campo era somado como 1 ponto e cada gol como 50 pontos; cartões amarelos subtraíam 10 pontos e vermelhos, 20 pontos. Em seguida, percebeu-se que os nomes dos países por vezes não seguiam o mesmo padrão como, por exemplo, o nome "Turquia"que aparecia como "Turkey"e "Türkiye". A fim de uniformizar os nomes, as seguintes alterações foram realizadas: "Crimea"= "Ukraine"; "CSSR"= "Czech Republic"; "England"= "United Kingdom"; "Jugoslawien (SFR)"= "Yugoslavia"; "Neukaledonien"= "New Caledonia"; "Northern Ireland"= "United Kingdom"; "Scotland"= "United Kingdom"; "UdSSR"= "Russia"; "Wales"= "United Kingdom"; "United States"= "USA"; "United Kingdom"= "UK". As demais padronizações foram feitas automaticamente pelo pacote "countrycode".

Ao banco de dados, unidos pelo código de identificação dos países, foram adicionadas duas novas tabelas: a primeira contendo as latitudes e longitudes de cada país, e a segunda o nome do continente ao qual pertencem. Por fim, foram selecionadas as obser-

vações cujos nomes do país de nascimento dos jogadores não era nulo, e adicionou-se a restrição de no mínimo 2000 minutos jogados, acumulados durante as temporadas registradas. As observações foram então agrupadas por país e utilizou-se o comando "left join"para combiná-las à estrutura de um mapa, que foi posteriormente personalizada com funções do pacote "ggplot2".

Para responder à segunda parte da pergunta, as tabelas "Clubs"e "Appearances" foram unidas pelo código de identificação dos clubes e da junção de suas observações foram selecionadas aquelas cujas porcentagens de jogadores estrangeiros não era nula. Aqui, as variáveis selecionadas para a análise foram os nomes dos clubes, a porcentagem de jogadores estrangeiros nos times, o número de gols feitos e levados, e o número de partidas ganhas, sendo as últimas três informações referentes ao intervalo de 2003 a 2024. Como as porcentagens foram retiradas do próprio banco, que não acompanhava nenhuma explicação acerca da coleta dos dados, inferiu-se que jogadores considerados estrangeiros eram aqueles que nasceram em países diferentes daquele em que o clube se situa.

O resultado foi então agrupado pelo nome e ID dos clubes e ordenado de forma decrescente pela quantidade de vitórias, que foi o critério principal escolhido para definir o desempenho dos clubes. Enfim, os 5 clubes com mais vitórias foram tabelados e as conclusões tiradas desta análise foram discutidas na seção seguinte.

2.5.3 Parte 3

Para a terceira pergunta foram criadas 3 tabelas, cada uma com uma variável relevante diferente para a resposta da pergunta: valor médio das transferências de um clube para outro, número de clubes diferentes jogados e o tempo médio de permanência em um clube para cada jogador, de forma que pudéssemos observar como cada uma destas variáveis se comportam isoladamente. Não foram considerados jogadores que apresentaram valor de transferência ou data de transferência nulas.

Nas primeira e na segunda tabela calculamos, respectivamente, o preço médio das transferências e o número de transferências de cada jogador. Para calcular o tempo médio em que o jogador ficou em um clube, medimos a distância em dias entre as transferências consecutivas entre clubes e calculamos a média. Em sequência, as três tabelas foram unidas utilizando o "player_id" como elemento em comum para possibilitar a observação do comportamento das variáveis de interesse.

Dessa forma, observamos que gráficos de pontos poderiam proporcionar boas visualizações para o caso em questão. Entretanto, os outliers comprometiam a clareza da visualização gráfica da situação apresentada. Assim, para tornar a interpretação mais explícita, aplicou-se um ’zoom’ no gráfico, limitando o eixo referente ao valor de mercado a 110.000.000. Além disso, foram consideradas até 17 transações, pois, embora o número máximo de transações registradas tenha sido 35, notou-se uma redução significativa na

quantidade de informações conforme o número de transações aumentava, nesse sentido, em alguns casos, apenas uma observação foi registrada.

3 Análises

3.1 Pergunta 1: Qual o gasto médio com os capitães dos times ao longo do tempo? Isso muda de acordo com sua posição?

A Figura 1 apresenta os valores de mercado dos capitães dos times ao longo dos anos. Nele, considera-se que o gasto com um capitão corresponde ao seu valor de mercado, permitindo uma análise da valorização desses jogadores ao longo do tempo.

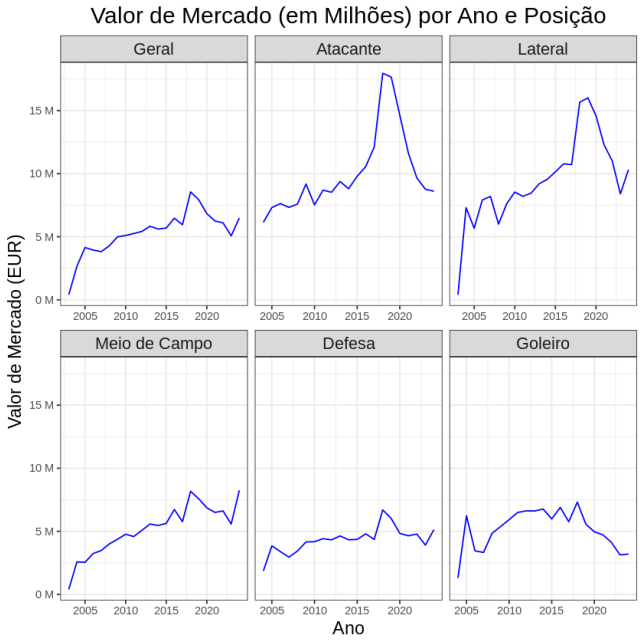

Figura 1: Gasto médio, em milhões, com capitães dos times de acordo com as posições ao longo dos anos.

Dessa forma, é possível observar que, no caso geral, houve um aumento significativo entre os anos 2002 e 2005, em sequência, houveram tanto aumentos quanto declínios sutis de modo que foi possível observar um aumento ao longo tempo, mas de forma equilibrada.

Em relação aos Atacantes, observa-se que esse é o tipo de posição mais bem paga com dados iniciais em 2003 registrando gastos superiores a 5 milhões, além de um crescimento bastante significativo ao longo dos anos, registrando seu pico em 2018 com gastos superiores a 17,5 milhões e, em seguida, apresentando um declínio bastante perceptível, contudo, até o momento, ainda se mantendo acima dos 7,5 milhões como uma das posições com maior valor de mercado.

Em relação aos capitães cuja posição é lateral, observa-se que os gastos inicialmente foram baixos, mas cresceram rapidamente ao longo do tempo atingindo seu pico em 2019

com gastos superiores a 15 milhões e, nos anos seguintes, apesar de ainda apresentar um valor de mercado elevado, é possível observar um declínio significativo.

No que se refere aos meio-campistas, observa-se um comportamento crescente dos gastos ao longo do tempo, contudo, se mantendo abaixo de 7,5 milhões e apenas pontualmente ultrapassando esse valor.

Os gastos com as posições de defesa em campo foram os que apresentaram comportamento mais constante ao longo dos anos, de modo que se mantiveram próximos de 5 milhões na maior parte do tempo.

Por fim, a posição de goleiro também registrou crescimentos e declínios ao longo do tempo de modo que de 2006 a 2020 tiveram gastos entre 5 milhões e 7,5 milhões, enquanto entre 2020 e 2023 os gastos oscilaram entre 2,5 milhões e 5 milhões.

3.2 Pergunta 2: Qual o desempenho dos jogadores de acordo com sua nacionalidade? Como o percentual de jogadores estrangeiros afeta o time?

Ao analisar a tabela "Clubs", foi visto que os clubes possuíam porcentagens distintas de jogadores estrangeiros na composição de seus times, e que a tabela "Players" continha o nome do país no qual cada jogador nasceu. A partir dessas informações, decidiu-se estudar o desempenho dos clubes e jogadores com relação à porcentagem de jogadores estrangeiros no time e sua nacionalidade, respectivamente.

Primeiramente, definiu-se o sistema de pontuação explicado na metodologia para classificar o desempenho em uma escala finita de modo que os resultados obtidos fossem passíveis de comparação entre si. Aqui, foram considerados apenas os jogadores que acumularam mais de 2000 minutos em campo de modo a não enviesar os resultados, ou seja, não permitir que um número reduzido de estatísticas obtido em pouco tempo de jogo criasse uma tendência nas conclusões. Tendo o desempenho individual calculado, os jogadores foram agrupados de acordo com o país em que nasceram, caso essa informação não fosse nula. A escala desses dados variou entre 4, correspondente à pontuação do Sudão, e 91,7 pontos, correspondente ao Paquistão.

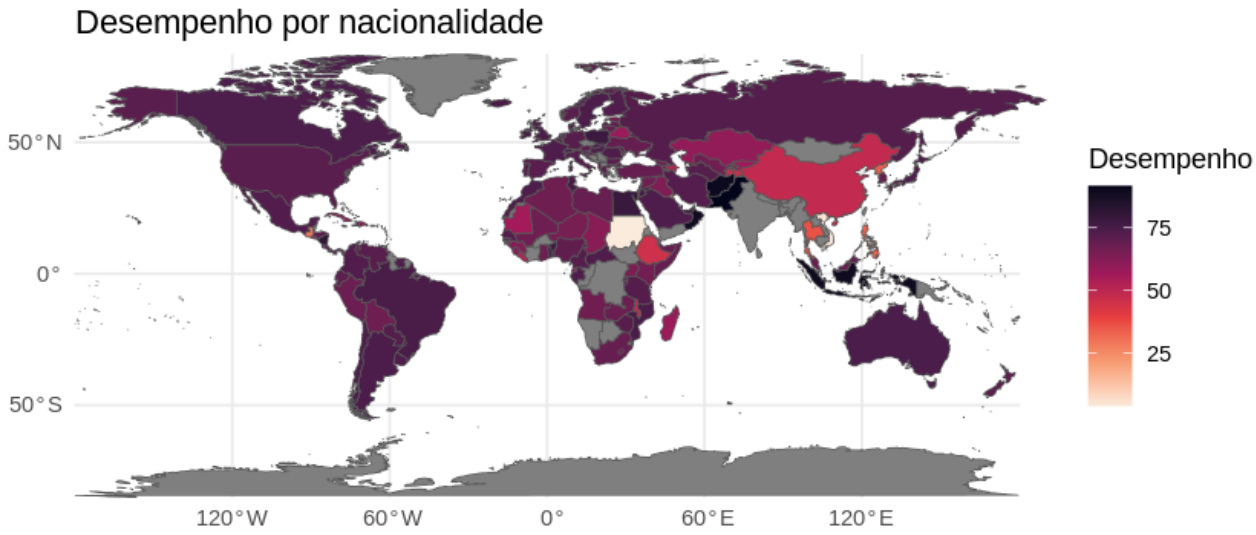

Figura 2: Mapa-múndi do desempenho dos jogadores por nacionalidade.

Por fim, os nomes dos países foram padronizados de forma a possibilitar a criação de um mapa, ilustrado na Figura 2, no qual a divergência entre os desempenhos dos países é explicitada pela escala de cores. É possível notar o destaque dos países Europeus e Sul-Americanos, além da Austrália e Nova Zelândia, com relação à pontuação que seus jogadores de origem desempenharam durante as temporadas recentes no futebol. Como

adendo, ressalta-se que os países em cinza são aqueles em que não foram registrados jogadores no banco de dados.

Em adição, analisou-se os dados referentes às partidas de cada clube de modo a medir seu desempenho durante as temporadas registradas. Para tanto, foram considerados a quantidade de gols feitos e levados, e a quantidade de jogos ganhos pelos clubes. Aqui, foram selecionados apenas os times cuja porcentagem de jogadores estrangeiros era maior que 0% para responder à pergunta elaborada, e as estatísticas acumuladas se referem aos dados somados desde 2003 até o último jogo abrangido pelo banco em 2024.

Estatísticas sumárias dos maiores times

Clubes	Porcentagem	Vitórias	Gols Feitos	Gols Levados
Real Madrid Club de Fútbol	77,3	12835	44211	18907
Futbol Club Barcelona	28,0	12721	44732	18434
FC Bayern München	60,0	12655	46045	16425
Manchester City Football Club	68,0	12094	41902	16813
Juventus Football Club	67,9	11766	33035	15270

Tabela 1: Tabela com os 5 clubes que mais acumularam vitórias entre as temporadas de 2003 e 2024

Os 5 clubes com mais vitórias acumuladas, listados na Tabela 1, apresentavam no mínimo 60% da composição de seus times compreendida por jogadores estrangeiros, com exceção do Barcelona Futebol Clube. Além disso, com exceção do Real Madrid Futebol Clube, os clubes com maiores porcentagens levaram menos gols e dentre todos, o Bayern München Futebol Clube foi o que mais acumulou gols.

3.3 Pergunta 3: Qual o tempo médio que um jogador permanece no mesmo time? Isso muda de acordo com seu valor de mercado?

A figura 3 apresenta uma representação gráfica do tempo médio que um jogador permanece em um mesmo time, comparado ao seu valor de mercado, para diferentes quantidades de transferências (de 2 a 17).

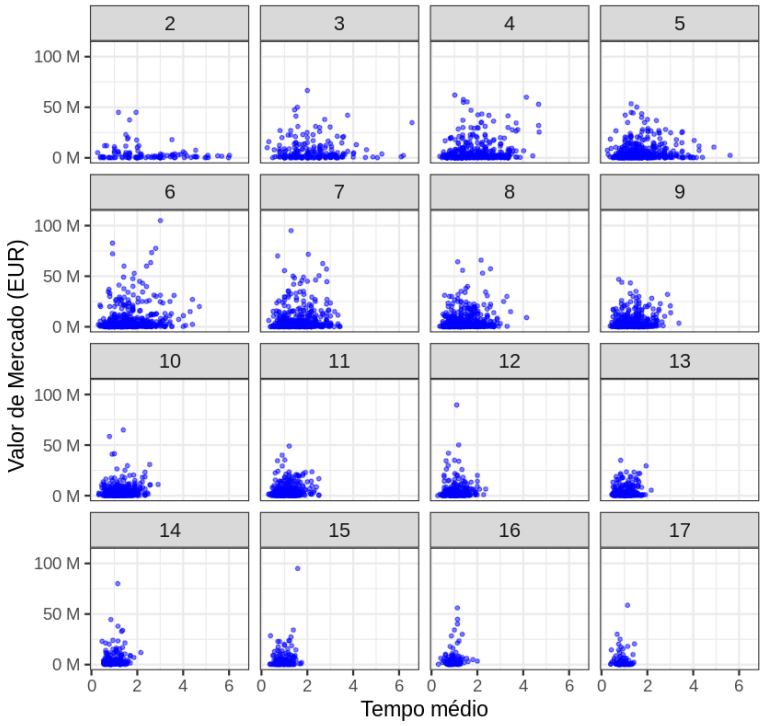

Figura 3: Tempo médio, em anos, e valor de mercado, em milhões de euros, pelo número de transferências.

Com base nos padrões visuais do gráfico, pode-se perceber que conforme o número de transferências aumenta, o padrão se mantém similar, com a maioria dos jogadores apresentando um valor de mercado baixo e tempo médio relativamente curto.

Além desse fato, a maioria dos pontos está concentrada em um intervalo de tempo médio entre 0 e 2 anos, independentemente do valor de mercado. Isso sugere que, para a maioria dos jogadores, independentemente de seu valor de mercado, o tempo médio no mesmo time é geralmente baixo.

4 Conclusão

No que se refere à primeira pergunta, foi possível observar que o gasto médio com capitães dos times - caracterizado nesse contexto pelo valor de mercado - sofreu oscilações no decorrer dos anos e que, de fato, o valor muda de acordo com a posição do capitão. Nesse contexto, observamos que a posição de Defesa manteve o comportamento mais constante, tal como foi a que apresentou menores gastos. Em contrapartida, as posições de Atacante e Lateral são as que apresentam, nos dados analisados, os gastos mais elevados.

Com relação à segunda pergunta, foi possível concluir que jogadores da Oceania, da América do Sul e da Europa se destacam em termos de desempenho, medido a partir do sistema de pontos estabelecido, quando comparados a jogadores que nasceram na África e na Ásia. Assim, ratifica-se o senso comum acerca da diferenciação do futebol sul-americano e do europeu, cujos países são os que mais acumularam grandes premiações como Copas do Mundo. Além disso, constatou-se que os times que mais acumularam vitórias entre 2003 e 2024 possuem mais de 60% de seus times compostos por jogadores estrangeiros, com exceção de um. Portanto, infere-se que o aumento do percentual de jogadores estrangeiros afeta positivamente o desempenho de seus times.

Por fim, para a terceira pergunta, observa-se que, independentemente da quantidade de transações realizadas e do valor de mercado do jogador no momento da transação, existe uma tendência de permanecer pouco tempo no time, sendo que, no geral, os jogadores costumam permanecer até dois 2 anos.

5 Referências

CARVALHO, Benilton; BENAGLIA, Tatiana. ME315 - UNICAMP, 2024. Disponível em: <http://me315-unicamp.github.io/>. Acesso em: 13 de nov. de 2024.

SQLITE TUTORIAL. SQLite Tutorial, 2024. Disponível em: <https://www.sqlitetutorial.net/>. Acesso em: 13 de nov. de 2024.

AMBIENTAL PRO DEV. Como fazer mapas no R com o pacote GGPLOT2. Disponível em: <https://www.youtube.com/watch?v=45x6BTaZcxc>. Acesso em: 13 nov. 2024.

6 Anexo

Os códigos elaborados para o desenvolvimento do trabalho podem ser acessados pelo link: `https://colab.research.google.com/drive/1Hr2pYMN8u6x1M1eSQql7pr7pcz-CKAIF?usp=sharing`.