Desafio – Analytics Engineer

Felipe Luís Teixeira

Criado em: 11/2022

Sumário

Introdução 3
Lista de Arquivos 4
Questão 1 – Python 5
Questão 2 – SQL 6
Questão 3 – SQL 7
3A – Limpeza 7
3B – Análise I – Proporção e Comparação Destro x Canhoto 8
3C – Análise II – País com o melhor saldo de gols e melhor média de gols por jogo 10
3D – Análise III – Times com melhor média de gols por jogo 11
3E – Análise IV – Média de gols por temporada de cada país 13
3F – Análise V – Estatísticas de cartões e faltas 14
Questão 4 – SQL - Relations 15
4A – Relations I – Relação Altura x Desempenho e Peso x Desempenho 15
4B – Relations II – Relação Idade x Desempenho (Análise adicional) 17
Questão 5 – SQL - CTE 18

Introdução

O objetivo deste Readme é documentar explicar, separado por questão, quais arquivos foram utilizados para a solução do desafio.

Na documentação de cada Questão, há uma tabela indicando o código-fonte utilizado e nome do arquivos de saída, assim como a análise do resultados, com um texto explicativo, tabelas e/ou gráficos.

Após a Introdução, apresento uma lista geral de todos os arquivos gerados par a solução deste desafio, sejam eles SQL, Py ou CSV.

As questões que pediam (ou permitiam) mais de uma análise, eu dividi em A, B, C, D, etc.

No código Python e nas queries disponibilizadas há comentários adicionais sobre o que foi desenvolvido.

Como os dados e o nome das colunas do banco de dados estão todos em inglês, optei por deixar as legendas e títulos dos gráficos no mesmo idioma.

Lista de Arquivos

Relação dos arquivos que foram enviados via Git.

Diretório	Arquivo
\Output	▪ Questão 2A - Player_Attributes_Modified.csv ▪ Questão 2B - Team_Attributes_Modified.csv ▪ Questão 2C - Match_Modified.csv ▪ Questao_3B_Analise_I_Destro_Canhoto.csv ▪ Questao_3C_Analise_II_Pais_Melhor_Media_Gols.csv ▪ Questao_3C_Analise_II_Pais_Melhor_Saldo_Gols.csv ▪ Questao_3D_Analise_III_Time_Media_Casa.csv ▪ Questao_3D_Analise_III_Time_Media_Geral.csv ▪ Questao_3E_Analise_III_Time_Media_Visitante.csv ▪ Questao_3E_Analise_IV_Media_Gols_Temporada_Pais.csv ▪ Questao_3F_Estatística de cartões e falta - Por Cartao.csv ▪ Questao_3F_Estatística de cartões e falta - Por Falta.csv ▪ Questao_3F_Estatística de cartões e falta - Por Porcentagem.csv ▪ Questao_4A_Relations_I_Altura_x_Desempenho.csv ▪ Questao_4A_Relations_I_Peso_x_ Desempenho.csv ▪ Questao_4B_Relations_II_Idade_x_Desempenho.csv ▪ Questao_5_Ouput.csv
\Python	▪ create_tables.py ▪ import_tables.py ▪ import_match_cards.py ▪ import_match_fouls.py
\Query	▪ Questão 2A - Create table Player_Attributes_Modified.sql ▪ Questão 2A - Insert into Player_Attributes_Modified.sql ▪ Questão 2B - Create table Team_Attributes_Modified.sql ▪ Questão 2B - Insert into Team_Attributes_Modified.sql ▪ Questão 2C - Create table Match_Modified.sql ▪ Questão 2C - Insert into Match_Modified.sql ▪ Questão 3A – Limpeza.sql ▪ Questão 3B - Análise I - Proporção e Comparação Destro x Canhoto.sql ▪ Questão 3C - Análise II - Saldo de Gols e Média de Gols por jogo por país.sql ▪ Questão 3D - Análise III - Time com melhor média de gols.sql ▪ Questão 3E - Análise IV - Média de gols por temporada.sql ▪ Questão 3F - Análise V - Estatísticas de cartões e faltas.sql ▪ Questão 4A - Relations I - Relação Altura-Peso x Desempenho.sql ▪ Questão 4B - Relations II - Relação Idade x Desempenho.sql ▪ Questão 5 - SQL CTE - Média de Gols por semana.sql

Questão 1 – Python

Para esta questão, o código foi divido em duas partes e deve ser executado na sequência abaixo:

Código Fonte	1. **create_tables.py** ○ Cria o database no SQLite **test_analytics_engineer.db** ○ Cria as tabelas referentes às planilhas CSV ○ Cria as tabelas adicionais (calculadas): Match_Cards e Match_Fouls **2. import_tables.py:** ○ Importa as planilhas CSV para o banco de dados. ○ Deixei o código genérico para poder importar qualquer planilha em massa (desde que a tabela exista na base), sem a necessidade de deixar o nome das colunas de cada tabela fixas no código. O programa irá ler as colunas do CSV para montar o INSERT; ○ É possível executar o código de duas formas: ▪ Sem parâmetros: *.\Python\import_tables.py* ➔Importará todas os CSV do diretório \Data ▪ Com parâmetro: *.\Python\import_tables.py Player* ➔ *Importará somente os dados do arquivo Player.csv para a tabela Player* **3. import_match_cards.py:** ○ Lê a coluna cards (XML) da tabela Match e gera dados para a tabela Match_Cards (para ser utilizado na resolução da Questão 3F) **4. import_match_fouls.py:** ○ Lê a coluna foulcommit (XML) da tabela Match e gera dados para a tabela Match_Fouls (para ser utilizado na resolução da Questão 3F)
Resultado	Tabelas criadas, arquivos CSV lidos e dados inseridos nas tabelas
Análise	-

Questão 2 – SQL

Para cada uma das tabelas a serem criadas, separei o script em dois: create table e o insert/select.

Código Fonte	• Questão 2A - Create table Player_Attributes_Modified.sql • Questão 2A - Insert into Player_Attributes_Modified.sql • Questão 2B- Create table Team_Attributes_Modified.sql • Questão 2B - Insert into Team_Attributes_Modified.sql • Questão 2C - Create table Match_Modified.sql • Questão 2C - Insert into Match_Modified.sql
Resultado	Tabelas criadas e dados inseridos. • Questão 2A - Player_Attributes_Modified.csv • Questão 2B - Team_Attributes_Modified.csv • Questão 2C - Match_Modified.csv
Análise	-

Questão 3 – SQL

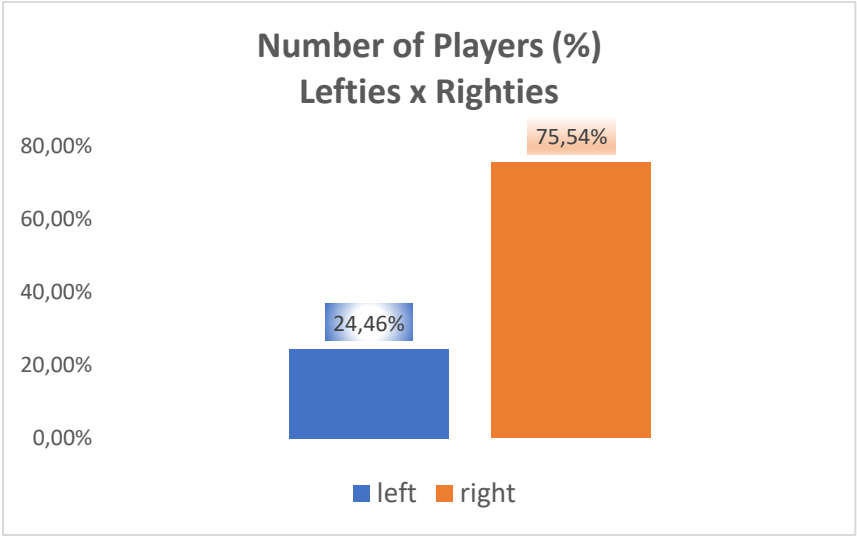
Como esta questão demanda uma análise livre, eu o dividi em cinco subtarefas (A, B, C, D e E), sendo a primeira para a limpeza das “sujeiras” encontradas, e as outras quatro para análises dos dados.

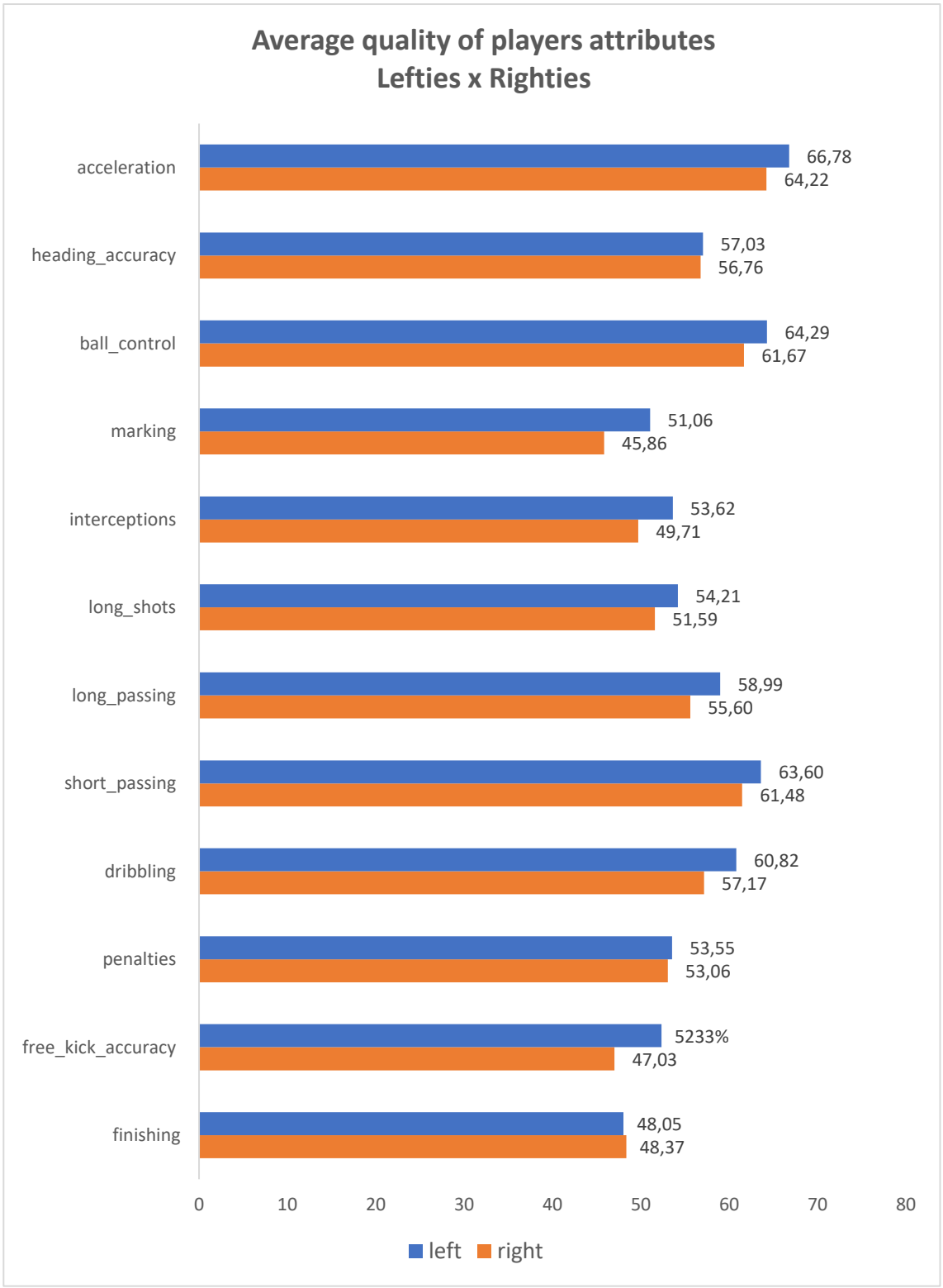
3A – Limpeza

Código Fonte	▪ Questão 3A – Limpeza.sql
Resultado	Dados apagados
Análise	▪ Identifiquei e apaguei países duplicados ou em branco. ▪ Identifiquei e apaguei Ligas sem o referente country_id. ▪ Identifiquei e apaguei Player_Attribues e Team_Attributes com JSON inválido, onde uma ou mais chaves possuíam valor “NaN” (attacking_work_rate). Entendi que isso poderia distorcer os dados e eliminei.

3B – Análise I – Proporção e Comparação Destro x Canhoto

Código Fonte	▪ Questão 3B - Análise I - Proporção e Comparação Destro x Canhoto.sql
Resultado	▪ Questao_3b_Analise_I_Destro_Canhoto.csv
Análise	▪ **24,45%** dos jogadores são canhotos e **75,54%** são destros ▪ Na média, os canhotos ficam à frente dos destros em todos os atributos analisados: finalização, chute livre, pênalti, drible, passe curto, passe longo, chute longo, interceptação, marcação, controle da bola, cabeçada e aceleração. ▪ A maior vantagem foi no atributo marcação, cuja diferença chegou a **5,2:** ○ Canhotos: média de **51,06** ○ Destros: média de **45,86**

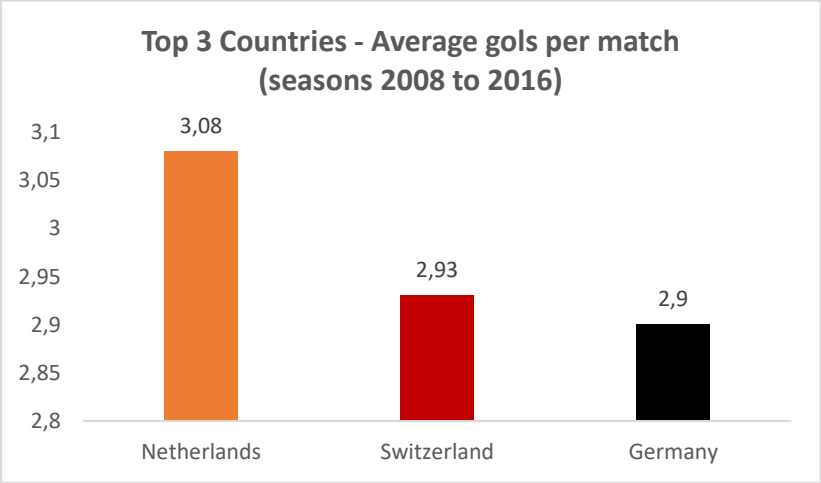


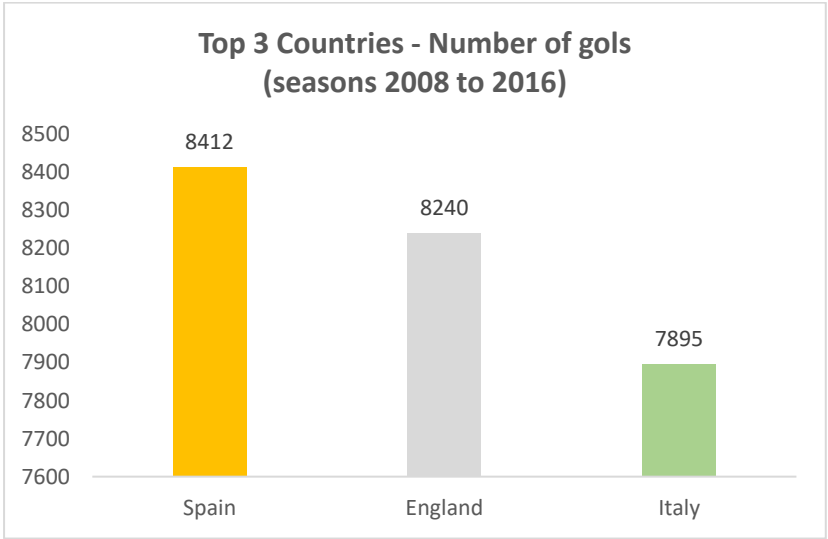

Average quality of players attributes
Lefties x Righties
acceleration
66,78
64,22
heading_accuracy
57,03
56,76
ball_control
64,29
61,67
marking
51,06
45,86
interceptions
53,62
49,71
long_shots
54,21
51,59
long_passing
58,99
55,60
short_passing
63,60
61,48
dribbling
60,82
57,17
penalties
53,55
53,06
free_kick_accuracy
5233%
47,03
finishing
48,05
48,37
0
10
20
30
40
50
60
70
80
left
right

3C – Análise II – País com o melhor saldo de gols e melhor média de gols por jogo

Como o número de gols pode estar relacionado ao número de partidas, eu dividi a análise em duas partes: uma mostrando os países com a melhor média de gols por partida e a outra com os países com mais gols marcados, considerando todas as temporadas.

Código Fonte	▪ Questão 3C - Análise II - Saldo de Gols e Média de Gols por jogo por país.sql
Resultado	▪ Questao_3C_Analise_II_Pais_Melhor_Media_Gols.csv ▪ Questao_3C_Analise_II_Pais_Melhor_Saldo_Gols.csv
Análise	▪ Os **Países Baixos** possuem a melhor média de gols por jogo: **3,08** ▪ A **Espanha** é o país com o melhor saldo: **8412** gols



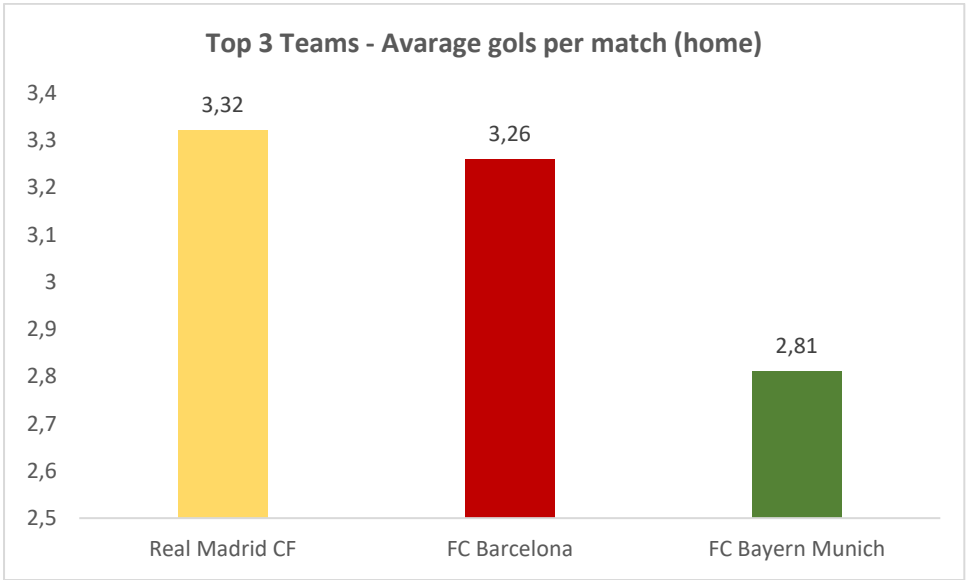


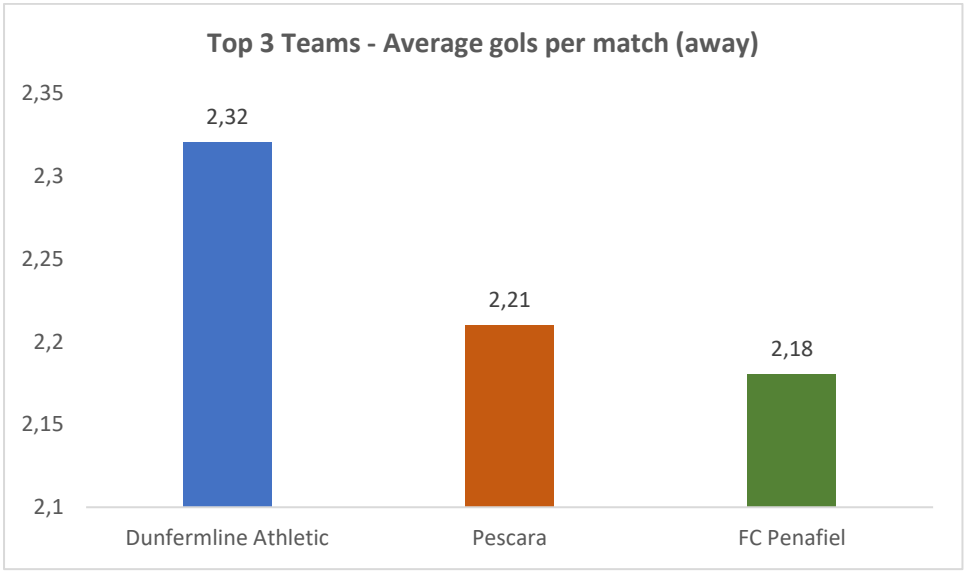
3D – Análise III – Times com melhor média de gols por jogo

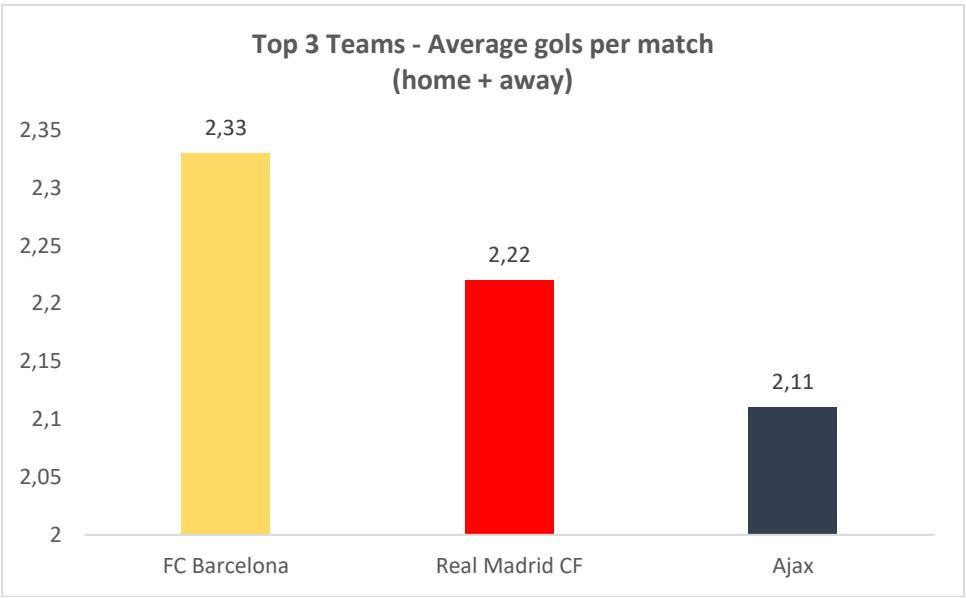
Separei a análise em três partes:

- Melhor média de gols jogando em casa;
- Melhor média de gols jogando como visitante;
- Melhor média de gols geral (em casa e como visitante0

Código Fonte	▪ Questão 3D - Análise III - Time com melhor média de gols.sql
Resultado	▪ Questao_3D_Analise_III_Time_Media_Casa.csv ▪ Questao_3D_Analise_III_Time_Media_Geral.csv ▪ Questao_3D_Analise_III_Time_Media_Visitante.csv
Análise	▪ O **Real Madrid** tem a melhor média de gols jogando em casa: **3,32 gols** por jogo ▪ **Dunfemline Athletic** tem a melhor média como visitante**:** **2,32 gols** por jogo ▪ O **Barcelona** possui a melhor média de gols considerando todos os jogos (em casa + visitante)**:** **2,33 gols** por jogo

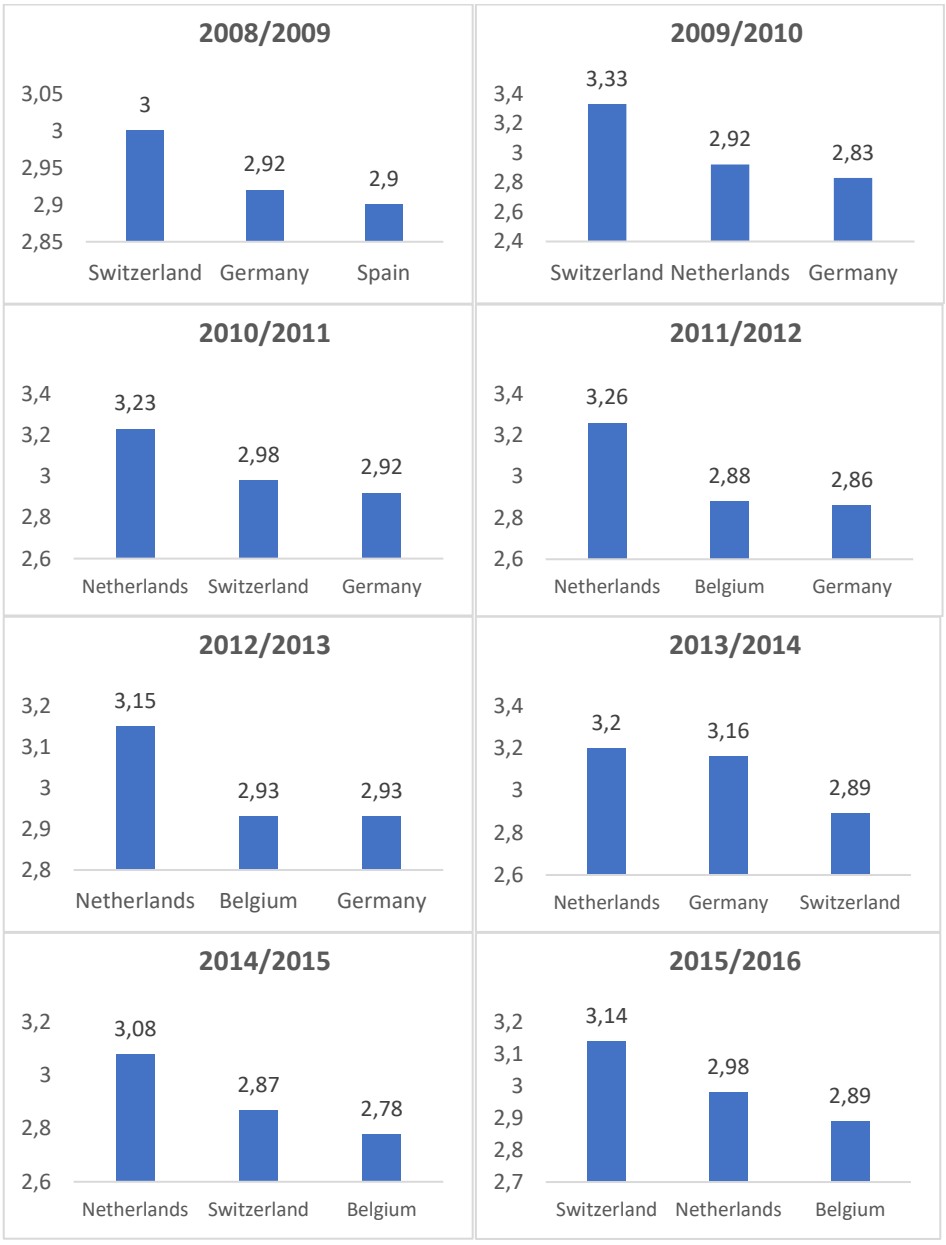

Top 3 Teams - Avarage gols per match (home)
3,4
3,3
3,2
3,1
3
2,9
2,8
2,7
2,6
2,5
3,32
3,26
2,81
Real Madrid CF
FC Barcelona
FC Bayern Munich


Top 3 Teams - Average gols per match (away)
2,35
2,3
2,25
2,2
2,15
2,1
2,32
2,21
2,18
Dunfermline Athletic
Pescara
FC Penafiel


Top 3 Teams - Average gols per match
(home + away)
2,35
2,3
2,25
2,2
2,15
2,1
2,05
2
2,33
2,22
2,11
FC Barcelona
Real Madrid CF
Ajax

3E – Análise IV – Média de gols por temporada de cada país

Código Fonte	▪ Questão 3E - Análise IV - Média de gols por temporada.sql
Resultado	▪ Questao_3E_Analise_IV_Media_Gols_Temporada_Pais.csv
Análise	▪ A **Suíça** deve a melhor média de gols nas temporadas **2008/2009**, **2009/2010 e 2015/2016** ▪ Em todas as outras temporadas (de **2010** a **2014**), a melhor média de gols foi sempre dos **Países Baixos**



3F – Análise V – Estatísticas de cartões e faltas

Extraí a lista dos dez times com maior número de faltas, maior número de cartões e maior porcentagem de cartões recebidos por faltas cometidas

Código Fonte	▪ Questão 3F - Análise V - Estatísticas de cartões e faltas.sql
Resultado	▪ Questao_3F_Estatística de cartões e falta - Por Cartao.csv ▪ Questao_3F_Estatística de cartões e falta - Por Falta.csv ▪ Questao_3F_Estatística de cartões e falta - Por Porcentagem.csv
Análise	▪ Os times com maior número de faltas, possuem uma porcentagem baixa de cartões recebidos, não chegando a 20% ▪ Em compensação, os times com maior número de cartões, a porcentagem fica sempre acima dos 30%. ▪ A terceira análise mostra os times com maior porcentagem de cartões recebidos por faltas cometidas, chegando a 82%.

Times com maior número de faltas

team	qty_fouls	qty_cards	percent
Stoke City	3513	580	16,51%
Manchester City	3357	525	15,64%
Aston Villa	3343	572	17,11%
Manchester United	3289	500	15,20%
Sunderland	3255	592	18,19%

Times com maior número de cartões

team	qty_fouls	qty_cards	percent
RCD Espanyol	2113	993	46,99%
Getafe CF	1888	972	51,48%
Valencia CF	3019	947	31,37%
Sevilla FC	2707	904	33,39%
Málaga CF	1901	871	45,82%

Times com porcentagem de cartões recebidos por número de faltas

team	qty_fouls	qty_cards	percent
CA Osasuna	788	653	82,87%
DSC Arminia Bielefeld	89	73	82,02%
Xerez Club Deportivo	140	111	79,29%
Catania	750	530	70,67%
Brescia	144	101	70,14%

Questão 4 – SQL - Relations

4A – Relations I – Relação Altura x Desempenho e Peso x Desempenho

Para esta análise eu crie três faixas de altura e peso.

Altura:
- 155 a 170 (cm)
- 171 a 185 (cm)
- 185 a 210 (cm)

Peso:
- 100 a 150 (lb.)
- 151 a 200 (lb.)
- 200 a 250 (lb.)

E gerei um Ranking com os seis (6) melhores jogadores de cada faixa, considerando a coluna **overall_rating** da tabela **Players_Attributes,** alimentando a tabela nova chamada **Relations**.

Código Fonte	▪ Questão 4A - Relations I - Relação Altura-Peso x Desempenho.sql
Resultado	▪ Questao_4A_Relations_I_Altura_x_Desempenho.csv ▪ Questao_4A_Relations_I_Peso_x_Desempenho.csv ▪ Tabela Relations
Análise	▪ **Melhor desempenho por Altura** ○ 155-170 → Leonel Messi (94) ○ 171-185 → Cristiano Ronaldo (93) ○ 185-210 → Manuel Neuer (90) ▪ **Melhor desempenho por Peso:** ○ 100-150 → Neymar (90) ○ 151-200 → Leonel Messi (94) ○ 200-250 → Manuel Neuer (90)

Ranking by Height

range	ranking	player_name	overall_rating
155-170	1	Lionel Messi	94
155-170	2	Andres Iniesta	88
155-170	3	Philipp Lahm	87
155-170	4	Alexis Sanchez	86
155-170	5	David Silva	86
155-170	6	Franck Ribery	86
171-185	1	Cristiano Ronaldo	93
171-185	2	Luis Suarez	90
171-185	3	Neymar	90
171-185	4	Arjen Robben	89
171-185	5	Eden Hazard	88
171-185	6	Mesut Oezil	88
185-210	1	Manuel Neuer	90
185-210	2	Zlatan Ibrahimovic	89
185-210	3	David De Gea	87
185-210	4	Jerome Boateng	87
185-210	5	Giorgio Chiellini	86
185-210	6	Karim Benzema	86

Ranking by Weight

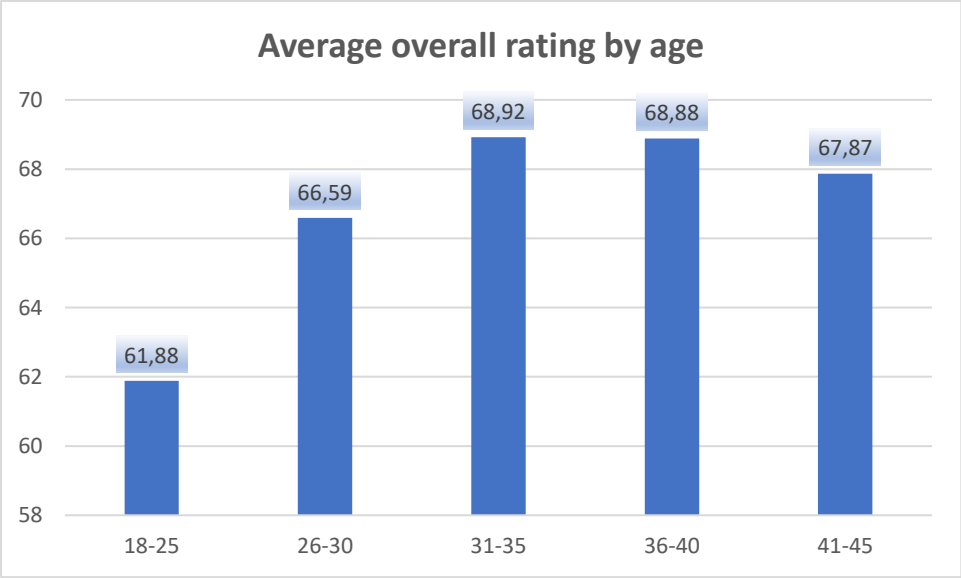
range	ranking	player_name	overall_rating
100-150	1	Neymar	90
100-150	2	Andres Iniesta	88
100-150	3	Luka Modric	87
100-150	4	Philipp Lahm	87
100-150	5	Alexis Sanchez	86
100-150	6	David Silva	86
151-200	1	Lionel Messi	94
151-200	2	Cristiano Ronaldo	93
151-200	3	Luis Suarez	90
151-200	4	Arjen Robben	89
151-200	5	Eden Hazard	88
151-200	6	Mesut Oezil	88
200-250	1	Manuel Neuer	90
200-250	2	Zlatan Ibrahimovic	89
200-250	3	Mats Hummels	86
200-250	4	Gianluigi Buffon	84
200-250	5	Joe Hart	84
200-250	6	Mehdi Benatia	83

4B – Relations II – Relação Idade x Desempenho (Análise adicional)

Fiz uma análise adicional, comparando a média de desempenho (**overall_rating)** dos jogadores por faixa de idade.

Código Fonte	▪ Questão 4B - Relations II - Relação Idade x Desempenho.sql
Resultado	▪ Questao_4B_Relations_ II_Idade_x_Desempenho.csv
Análise	▪ Nota-se um grande crescimento na média qualidade dos jogadores conforme a idade aumenta, com pico na faixa dos 31 a 35 anos, e uma pequena queda no fim da carreira.

age_range	overall_rating (average)
18-25	61,88
26-30	66,59
31-35	68,92
36-40	68,88
41-45	67,87



Questão 5 – SQL - CTE

Para esta análise criei uma query para retornar a média de gols semanal do time da casa e do time visitante, comparando a média da semana atual com a média da semana anterior, mostrando a porcentagem de crescimento.

A planilha de saída exibe o Ano e o Mês de referência. O cálculo semanal é feito sempre considerando as partidas realizadas de segunda a domingo (cuja data é exibida na planilha)

Código Fonte	▪ Questão 5 - SQL CTE - Média de Gols por semana.sql
Resultado	▪ Questao_5_Ouput.csv
Análise	▪ O arquivo CSV exibe o histórico semanal, trazendo o histórico semanal de 2008 a 2016. ▪ Abaixo, uma pequena amostra, com dados dos meses de março, abril e maio de 2016 (ordem decrescente), e o gráfico para ilustrar.

year	month	week	avg_goals_home	growth	avg_goals_away	growth
2016	5	29/05	1,6	-2,44%	1,4	-3,45%
2016	5	22/05	1,64	-18,00%	1,45	8,21%
2016	5	15/05	2	36,05%	1,34	-2,19%
2016	5	08/05	1,47	-3,92%	1,37	8,73%
2016	5	01/05	1,53	-14,53%	1,26	13,51%
2016	4	24/04	1,79	15,48%	1,11	-16,54%
2016	4	17/04	1,55	4,73%	1,33	33,00%
2016	4	10/04	1,48	-17,78%	1	-24,81%
2016	4	03/04	1,8	42,86%	1,33	2,31%
2016	3	20/03	1,26	-23,64%	1,3	14,04%
2016	3	13/03	1,65	10,74%	1,14	-1,72%
2016	3	06/03	1,49	4,20%	1,16	-2,52%