INSTITUTO UNIVERSITÁRIO DE LISBOA

PROJETO APLICADO EM CIÊNCIA DE DADOS I

DATA PREPARATION

LICENCIATURA EM CIÊNCIA DE DADOS

Base de Dados ATP – Brasil

Grupo 5: nº 103303, nº 110451, nº 104716, nº 99239

Docentes: Diana Aldea Mendes e Sérgio Moro

3 de maio de 2023

Data Preparation – Novas variáveis

BornOpponent: informação acerca do país e cidade de nascimento do jogador oponente;

BornCityOpponent: cidade de nascimento do jogador oponente;

BornCountryOpponent: país de nascimento do jogador oponente;

HeightOpponent: altura do jogador oponente, em cm;

HandOpponent: mão dominante do jogador oponente;

L_OR_R_Opponent: informa se o jogador é destro ou esquerdino e corresponde à primeira parte da variável HandOpponent;

NamesSorted: contém um array com os nomes do jogador principal e do seu oponente, ordenados por ordem alfabética;

PlayerComb: contém o nome do jogador principal vs o nome do jogador oponente;

Birthday: dia de nascimento do jogador principal;

BirthdayOpponent: dia de nascimento do jogador oponente.

Alterações feitas na base de dados

- Eliminação de jogos espelhados:

    1. Criação das variáveis NamesSorted e PlayerComb;
    2. Comparação entre o array criado e o nome do jogador e do oponente;
    3. Fazer o drop das linhas que continham os campos “torneio”, “data” e “PlayerComb” iguais.

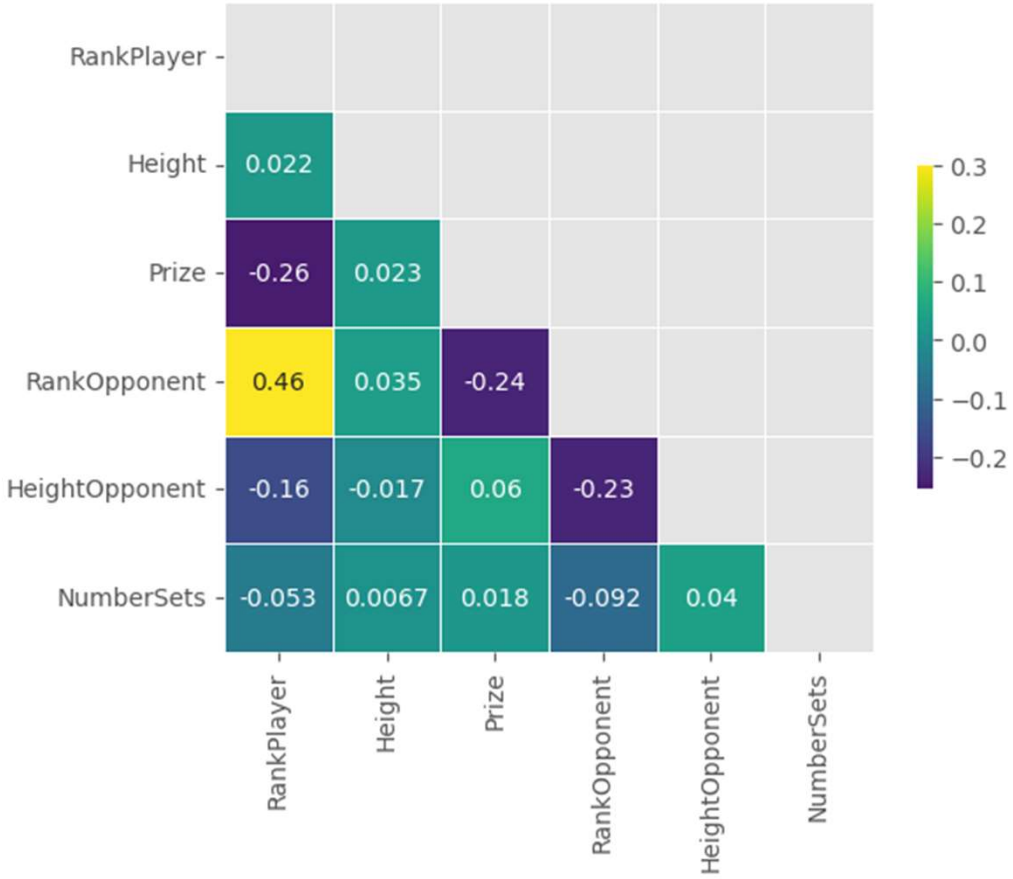
O número de jogos espelhados é 17,229, que representam 46.1% do dataset original do Brasil, lembrando que a base de dados df_brasil, tinha um total de 37,367 linhas.

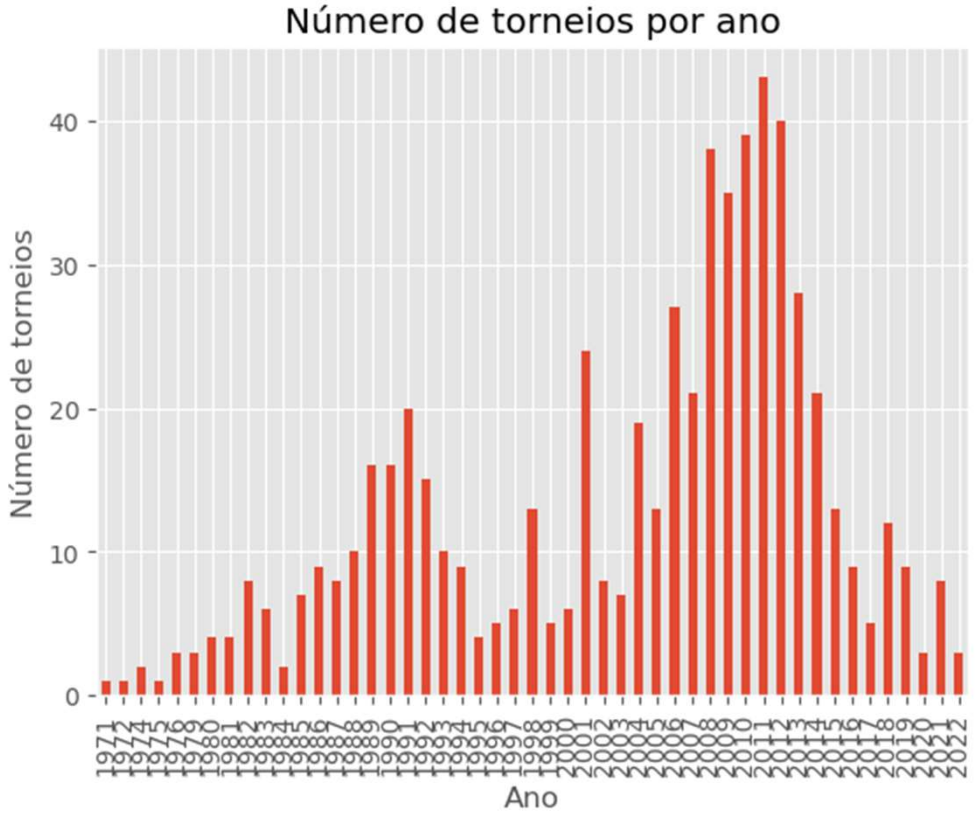
Variáveis do dataset

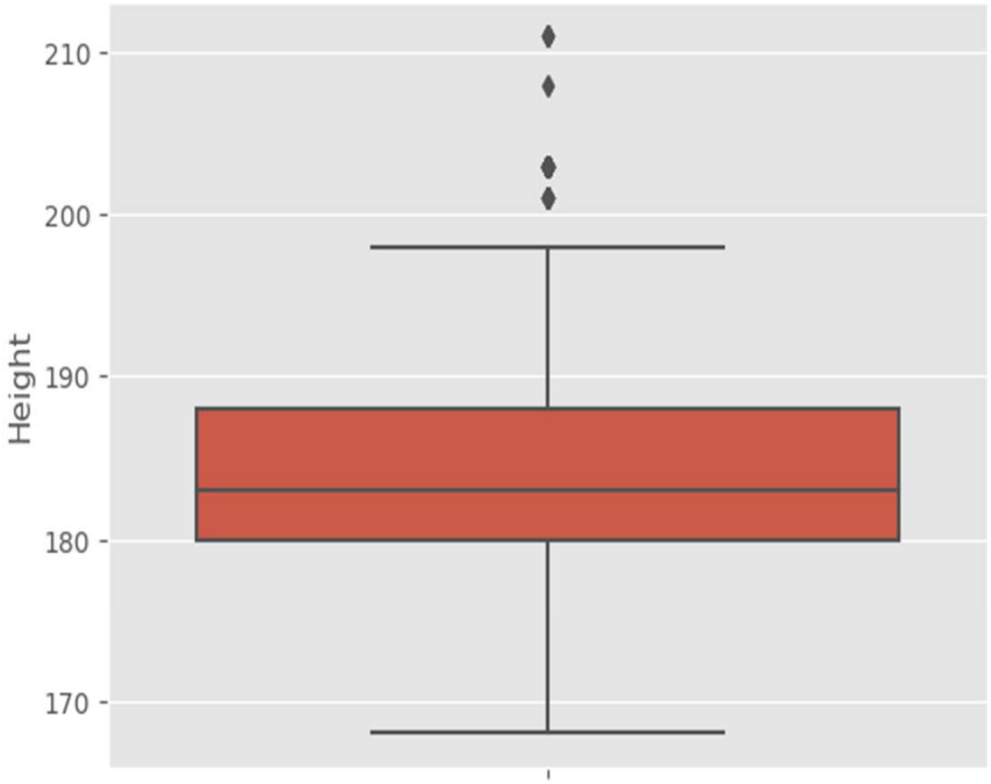
Variáveis originais	Descrição
_id	Identificador único de cada linha, que foi eliminado
PlayerName	Principal jogador da partida
Born	País e /ou cidade que o jogador nasceu
Height	Altura dos jogadores em cm
Hand	Mão dominante do jogador e a que usou
LinkPlayer	Link que nos remete para o perfil detalhado do jogador
Tournament	Nome do torneio
Location	Onde determinado torneio se realizou
Date	Indica as datas de início e fim do torneio
Ground	Tipo de terreno em que o torneio foi jogado
Prize	Prémio monetário
GameRound	Fase do torneio a que pertence o jogo
GameRank	Apelidada de rankopponent, rank do opponent
Oponent	Nome do adversário
WL	W - se ganhou - ou L - se perdeu
Score	Resultados do jogo por sets

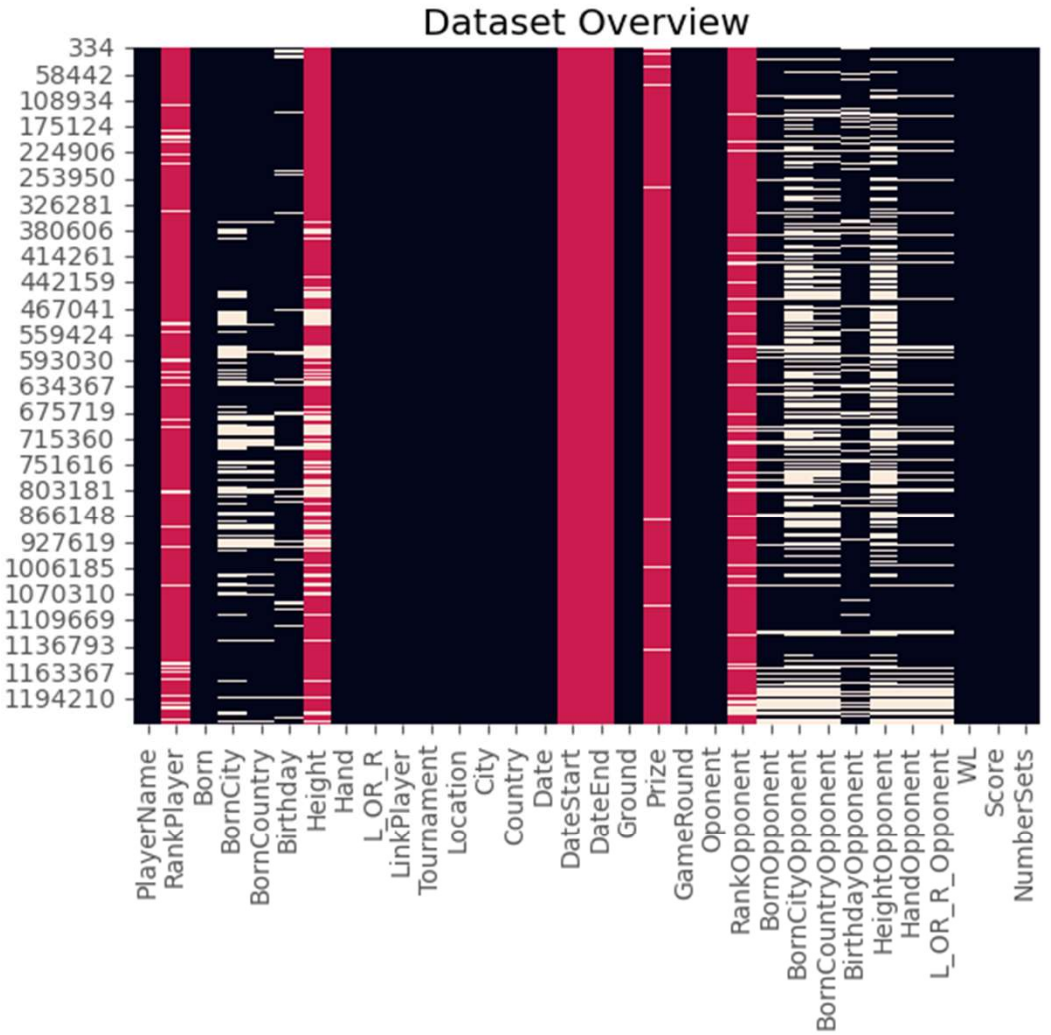
Variáveis criadas	Descrição
RankPlayer	Ranking do jogador principal num determinado ano
City	Cidade presente na variável location
DateStart	Data de início do torneio
DateEnd	Data de fim do torneio
NumberSets	Contagem do número de pares presentes na variável score
Country	Segunda parte da variável location, corresponde ao país
BornCountry	País de naturalidade do jogador, corresponde à segunda parte da variável born
L_OR_R	Mão dominante do jogador
BornOpponent	Informação acerca do país e cidade de nascimento do jogador oponente
BornCityOpponent	Cidade de nascimento do jogador oponente
BornCountryOpponent	País de nascimento do jogador oponente
HeightOpponent	Altura do jogador oponente, em cm
HandOpponent	Mão dominante do jogador oponente
L_OR_R_Opponent	Informa se o oponente é destro ou esquerdino
NamesSorted	Contém um array com os nomes do jogador principal e do seu oponente, ordenados por ordem alfabética
PlayerComb	Contém o nome do jogador principal vs o nome do jogador oponente
Birthday	Dia de nascimento do jogador principal
BirthdayOpponent	Dia de nascimento do jogador oponente

Algumas estatísticas






Height
210
200
190
180
170


Dataset Overview
334
58442
108934
175124
224906
253950
326281
380606
414261
442159
467041
559424
593030
634367
675719
715360
751616
803181
866148
927619
1006185
1070310
1109669
1136793
1163367
1194210
PlayerName
RankPlayer
Born
BornCity
BornCountry
Birthday
Height
Hand
L_OR_R
LinkPlayer
Tournament
Location
City
Country
Date
DateStart
DateEnd
Ground
Prize
GameRound
Oponent
RankOpponent
BornOpponent
BornCityOpponent
BornCountryOpponent
BirthdayOpponent
HeightOpponent
HandOpponent
L_OR_R_Opponent
WL
Score
NumberSets