INSTITUTO UNIVERSITÁRIO DE LISBOA

PROJETO APLICADO EM CIÊNCIA DE DADOS I

DATA UNDERSTANDING AND 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

25 de abril de 2023

Verificação das variáveis

Estrutura da base de dados

- 1308835 linhas e 16 colunas –base de dados ATP
- 37367 linhas e 23 colunas – base de dados ATP Brasil

Investigação do conteúdo

Verificação da qualidade (base de dados ATP Brasil)

- Dados omissos:
  - **RankPlayer**- 2620
  - **BornCountry**- 11305
  - **Height**- 11432
  - **Prize-** 367
  - **RankOpponent**- 3511
- Dados duplicados: 0

Data Preparation – Passos tomados

1. Uniformização da variável Prize;
2. Criação das variáveis DateStart e DateEnd;
3. Valores omissos;
4. Imputação dos valores omissos na variável DateEnd;
5. Substituição de valores omissos na variável Height;
6. Substituição de valores omissos na variável GameRank;
7. Renomear a variável GameRank;
8. Criação da função que associa o jogador ao seu rank por ano;
9. Criação da função que busca o valor do ranking e da coluna RankPlayer;
10. Uniformização do Score;
11. Criação da coluna NumberSets;
12. Tratamento do Campo _id;
13. Criação da nova variável City a apartir da variável Location;
14. Importação do dataset Worldcities;
15. Duplicados de cidades em países diferentes;
16. Valores omissos em Country;
17. Países com escritas diferentes;
18. Formas diferentes de escrita de Brasil;
19. Uniformização da escrita para “Brasill”;
20. Valores Únicos
21. Investigação da qualidade dos dados
22. Imputação

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

Medidas descritivas das variáveis numéricas

Statistic	RankPlayer	Height	Prize	RankOpponent	NumberSets
count	34747.0	25935.0	37000.0	33856.0	37367.0
mean	526.6	179.0	72149.6	515.1	2.3
std	402.9	30.8	215879.9	400.1	0.5
min	1.0	0.0	10000.0	1.0	1.0
25%	221.0	178.0	10000.0	220.0	2.0
50%	407.0	183.0	25000.0	391.0	2.0
75%	744.0	188.0	50000.0	709.0	3.0
max	2243.0	510.0	1786690.0	2243.0	5.0

Variáveis únicas: *dataset* filtrado para o Brasil

Variável	Valores únicos
Country	1
WL	3
Ground	4
L_OR_R	4
NumberSets	5
Hand	8
GameRound	11
Height	21
Prize	40
City	51
Location	57
Tournament	148
BornCountry	206
DateEnd	577
DateStart	595
Date	599
Born	779
RankPlayer	1422
RankOpponent	1825
LinkPlayer	1900
PlayerName	1900
Score	2102
Oponent	2655