Data Understanding e Data Preparation (início) Ténis na Bélgica

Grupo 4:

ALLAN KARDEC, Nº 103380, TURMA CDB1
DIOGO FREITAS, Nº104841, TURMA CDB1
JOÃO BOTAS, Nº104782, TURMA CDB1
RICARDO ÂNGELO, Nº 104826, TURMA CDB1

Data Understanding

- Proporção da Bélgica na base de dados -> ≈0.84%

- Feature selection (LinkPlayer)

- Valores omissos nas variáveis

- Variáveis qualitativas e quantitativas

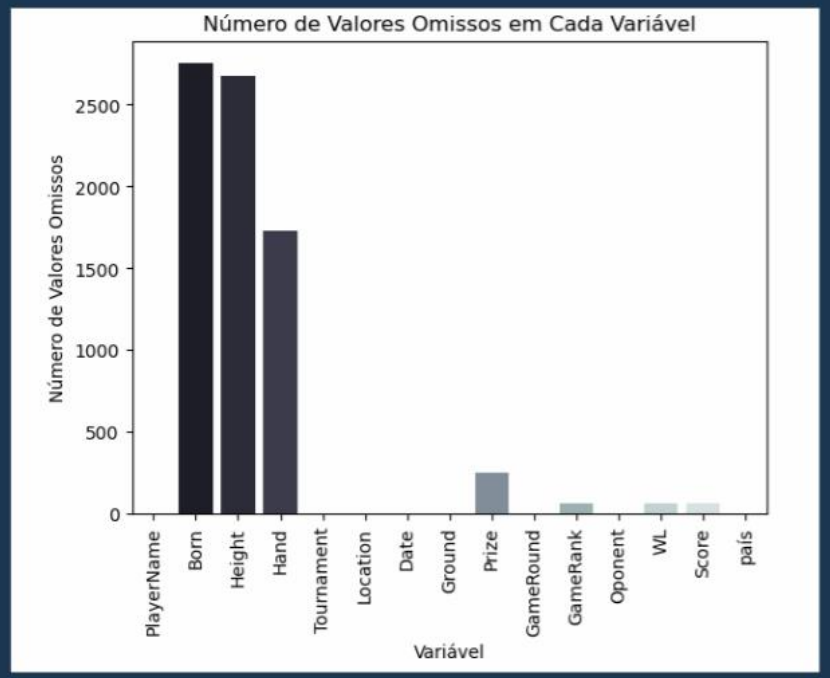
- Deteção de outliers

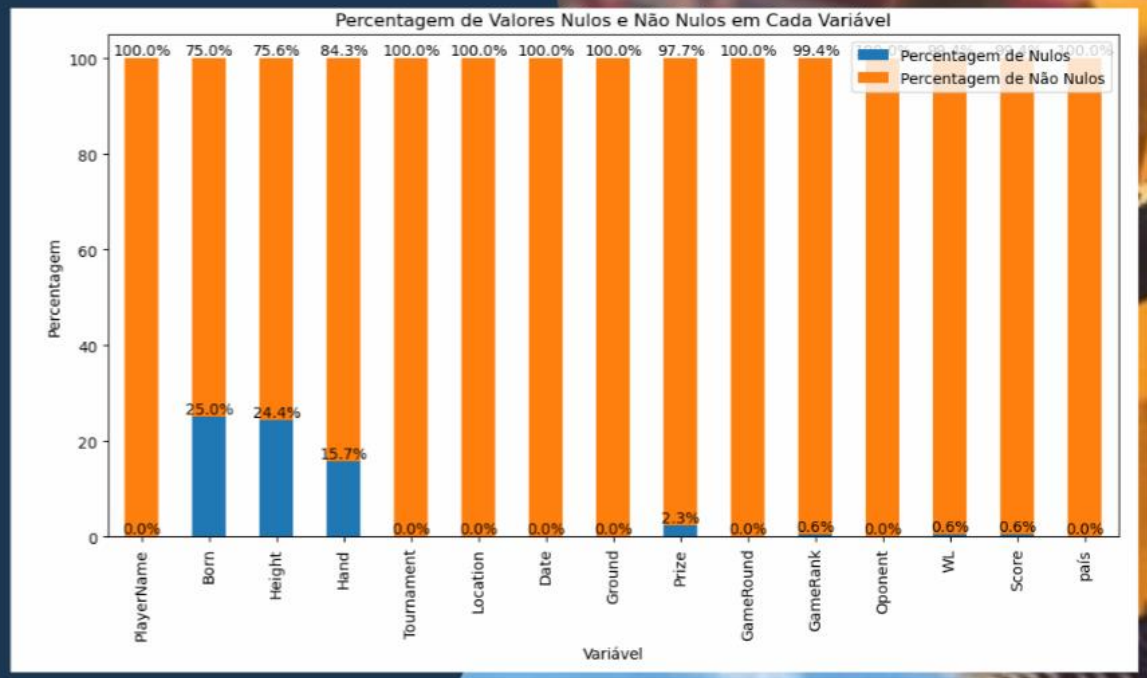
Variáveis da Base de dados

PLAYERNAME	Nome do jogador.
BORN	O local onde o Jogador nasceu.
HEIGHT	A altura, em centímetros, do jogador.
HAND	Como é que o jogador joga e qual a sua mão principal.
LINKPLAYER	Um link que nos leva direto á página web do jogador, tendo todas as características e todo o seu histórico de jogos.
TOURNAMENT	O nome do torneio que foi decorrido o jogo.
LOCATION	A localização do torneio (serão só torneios localizados na Bélgica).
DATE	A data em que foi decorrido o torneio. Varia entre 1968 e 2021 nos nossos dados, referentes à Bélgica.
GROUND	O material do chão do campo, podendo ser um dos seguintes: • Clay • Hard • Carpet
PRIZE	O prémio que o jogador ganhou no torneio.
GAMEROUND	A ronda em que o jogo está a decorrer. As rondas existentes em torneios na Bélgica, ordenados por ordem decrescente. • Round of 32 • Round of 16 • Quarter-Finals • Semi-Finals • Finals • 1st Round Qualifying • Round Robin • 2nd Round Qualifying • 3rd Round Qualifying • Round of 64
GAMERANK	O *Rank* atribuído à partida, tendo como análise o torneio, os jogadores, o prémio, a localização do torneio, e várias outras características.
OPONENT	O jogador oponente do jogador apresentado.
WL	Se o jogador apresentado venceu/perdeu o jogo.
SCORE	A pontuação do jogador neste jogo, em sets de jogos.

Gráficos e Informações relevantes

- Valores Omissos

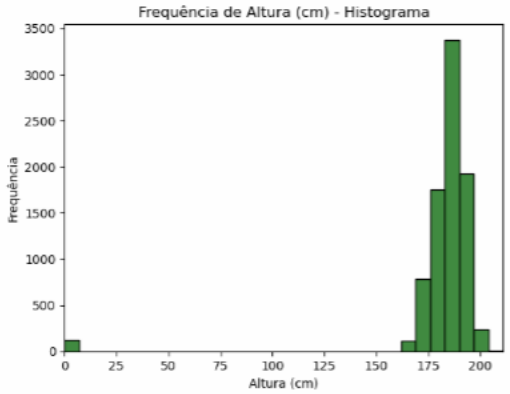


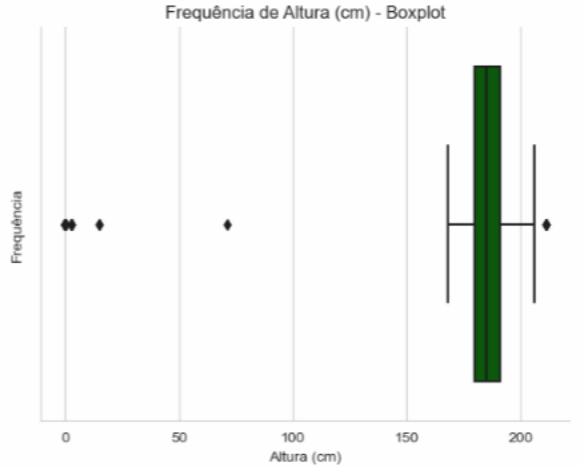


Variáveis quantitativas (Distribuição)

Existem três variávies quantitativas:

- Height (Altura do jogador)





- GameRank

```
belgica_raw["GameRank"].value_counts().sort_values(ascending=False)
```

```
-       899
410      34
1        33
150      29
15       28
        ...
1862      1
1138      1
1712      1
878       1
1687      1
Name: GameRank, Length: 1610, dtype: int64
```

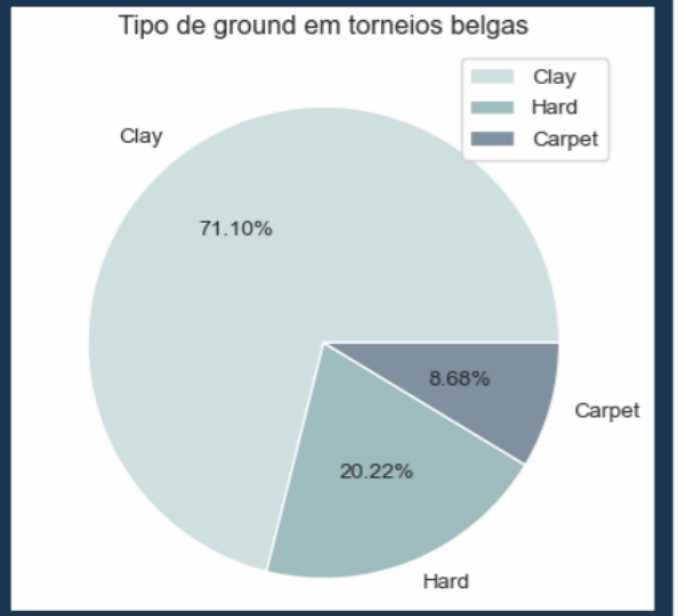
- Prize

```
belgica_raw['Prize'].value_counts().sort_values(ascending=False)
```

```
$10,000       4124
$15,000       2300
$25,000       1166
€106,500       968
$75,000        322
$125,000       247
$875,000       186
$250,000       175
$50,000        136
$465,000       120
€42,500         97
€394,800        84
€635,750        84
€508,600        82
€612,755        82
€566,525        79
€589,185        78
$1,085,000      62
$100,000        62
$1,100,000      62
$665,000        62
$1,000,000      62
$372,500        54
$210,000        32
$175,000        16
Name: Prize, dtype: int64
```

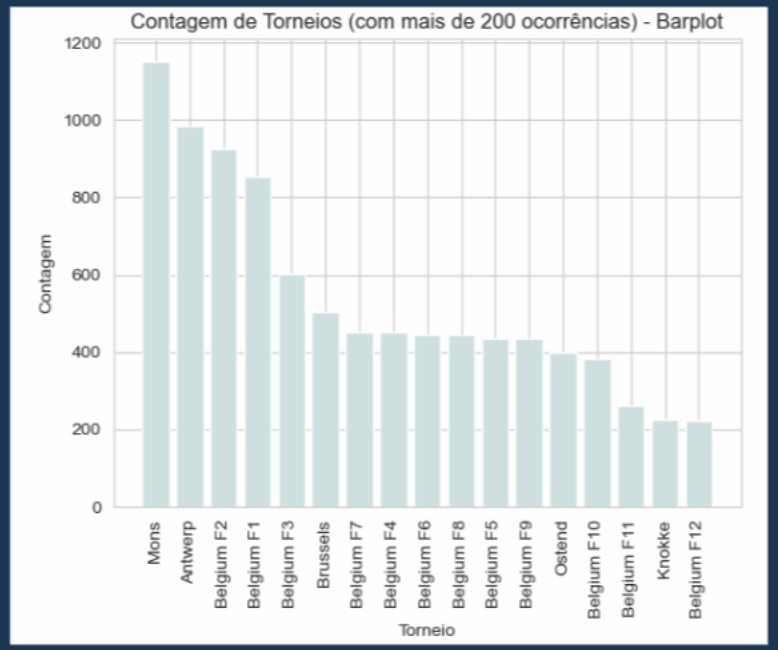
Variáveis Qualitativas (Pontos importantes)

- Ground



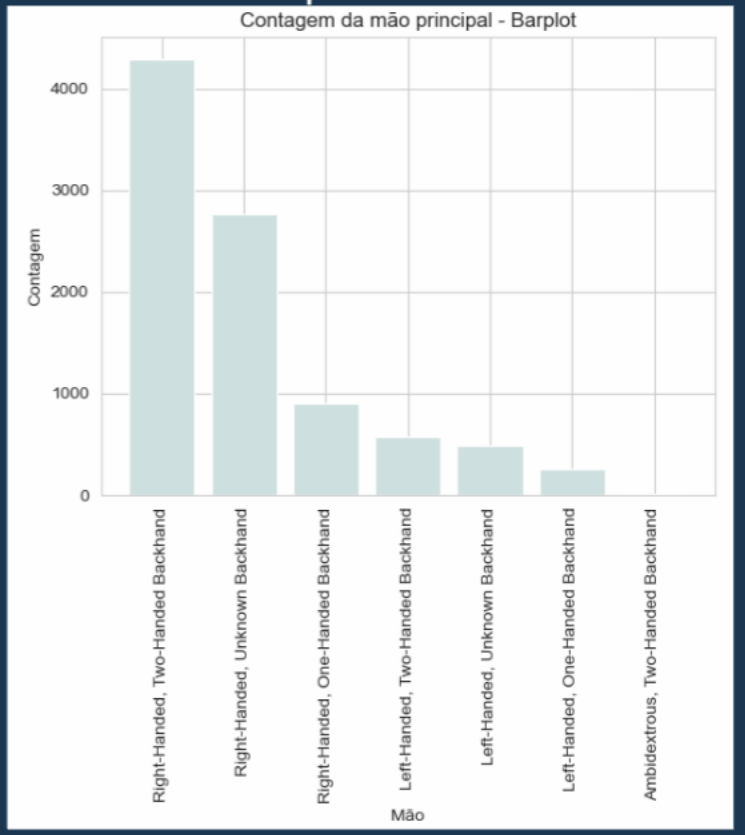
- Não existem jogos em relva, na Bélgica

- Tournament



- Torneios mais jogados são de Mons e Antwerp

- Mão Principal



- Há mais destros que esquerdinos na "**main hand**", e quase nenhum ambidestro

Nota: apenas algumas variáveis qualitativas estão representadas

Início do Data Preparation

- Webscrapping

- Campos com dados não uniformes (**Price**, por exemplo)

- Discussão sobre W/O e RET nos sets

- Tipo de dados

Para a próxima semana...

- Continuar/Concluir o Data Preparation

- Ideias...
    - Feature Engineering;
    - Gráficos com dados limpos;