ATP Players

Apresentação Semanal [2]

UC | Projeto Aplicado a Ciência de Dados I
Docentes | Diana Mendes **&** Sérgio Moro

Grupo 2

André Silvestre Nº104532
Diogo Catarino Nº104745
Francisco Gomes Nº104944
Rita Matos Nº104936

CDB1

3 | Data Preparation

- Eliminar observações duplicadas – Foram afetadas **2830 linhas** (0.2%)
- Identificar valores omissos **NaN** (observações s/ caracteres, **`0`**, **`null`**)
- Limpar os **Jogos**, restringindo ao país atribuído (**Suécia**)
- Limpar os **Jogadores**

1 | Location

countries_net.csv [1]
Lista com as designações dos 249 países

Das **2512 designações distintas** verificámos as designações que já estavam corretas

Depois de uniformizar, restringir o *dataset* aos **12 277** jogos da **Suécia**

Restaram **213 designações** que tiveram de ser limpas manualmente

	Mal(Original)	Bem	Dicionário {desginação a alterar : uniformizada}
2	's-Hertogenbosch	Netherlands	"'s-Hertogenbosch":"Netherlands",
3	Abidjan	Côte d'Ivoire	"Abidjan":"Côte d'Ivoire",
4	Calgary, Alberta	Canada	"Calgary, Alberta":"Canada",
5	Angleur - Liege	Belgium	"Angleur - Liege":"Belgium",
6	Antalya, Antalya	Turkey	"Antalya, Antalya":"Turkey",
7	Fayetteville, Fayetteville, AR	United States	"Fayetteville, Fayetteville, AR":"United States",

	Tournament	Location
0	Botswana F1	TBA
1	Egypt F1	TBA
2	Egypt F2	TBA
3	Egypt F3	TBA
4	France F13	TBA
5	Great Britain F10	TBC
6	Great Britain F7	TBA
7	Great Britain F9	TBC

Casos Particulares

- Existem terminologias como:
  - **TBA** (*To Be Announced*),
  - **TBC** (*To Be Confirmed*)
  - **TBD** (*To Be Determined*)
- Escolhas do país a considerar:
  - **Bolivia/Chile**
  - **Czechoslovakia**
  - **Yuguslavia**

Optámos por uniformizar em **Bolivia**, **Slovakia** e **Serbia**, respetivamente.

[1] *https://datahub.io/core/country-list#data*

2

Tournament

Renomeação de casos específicos pelo nome do torneiro.

18 designações distintas.

3

Date

Para resolver o formato `aaaa.mm.dd - aaaa.mm.dd` criámos uma coluna para a data inicial e outra para a data final do torneio.

Intervalo temporal dos jogos em estudo
[1969 ; 2022]

4

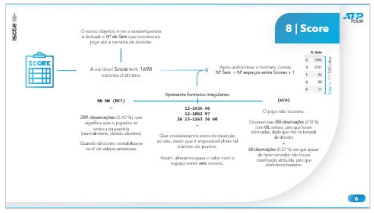
Ground

Já se encontra corretamente limpa.

Ground	
Hard	6582
Clay	4696
Carpet	999

8

Score



7

WL

Eliminámos 269 valores omissos (2.19 %) dado que não houve resultado final, nem *Score* atribuído.

6

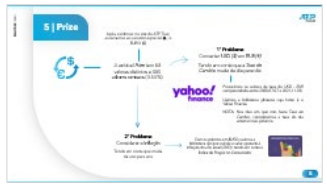
GameRound

Já está preparado a ser codificado.

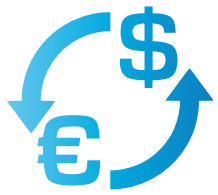
10 categorias distintos

5

Prize



5 | Prize



Após confirmar no site do ATP Tour, associamos ao caracter especial **?** , o EURO (€)

A variável *Prize* tem 63 valores distintos e **434 valores omissos** (3.54 %)

1° Problema
Converter USD *($)* em EUR *(€)*

Tendo em conta que a *Taxa de Cambio* muda de dia para dia

Procurámos os valores da taxa do *USD - EUR* compreendidos entre **2008-07-07** e **2021-11-08**

Usámos a biblioteca **yfinance** cuja fonte é o *Yahoo Finance*.

NOTA: Nos dias em que não havia *Taxa de Cambio*, considerámos a taxa do dia anterior mais próximo.

2° Problema
Considerar a **Inflação**

Tendo em conta que muda de ano para ano

Com os prémios em *$USD*, usámos a biblioteca **cpi** que calcula o valor ajustado à inflação do ano atual (2023) tendo em conta o *Índice de Preços no Consumidor*

8 | Score

O nosso objetivo é ter a variável pronta a deduzir o *Nº de Sets* que ocorreu no jogo até à tomada de decisão

A variável *Score* tem 1698 valores distintos

Após uniformizar o formato, contar
Nº Sets = Nº espaços entre Scores + 1

N_Sets	
2	7895
3	3791
1	95
4	88
5	71

Total = 11 940 obs.

Apresenta formatos irregulares:

NN NN (RET)
…

289 observações (2.42 %) que significa que o jogador se retirou da partida (normalmente, devido a lesões).

Quando tal ocorre, contabiliza-se os nº de valores anteriores.

12-1426 46
12-1062 97
26 13-1163 36 60
…

Que considerámos erros de inserção no site, dado que é impossível obter tal número de pontos.

Assim, alterámos para o valor com o espaço entre *sets* correto.

(W/o)

O jogo não ocorreu.

Ocorrem nas 269 observações (2.19 %) com WL omisso, pelo que foram eliminadas, dado que não há tomada de decisão.

+

68 observações (0.57 %) em que apesar de haver vencedor não houve classificação atribuída, pelo que eliminámos também.