Projetos em Ciência de Dados

Gianlucca Devigili e Maisa O. Fraiz

1.
Preparação dos Dados

Tratamento de Dados

- Dados salvos em **.pkl**
  - Redução do tempo de carga de **28.4s** para **5.34s**
  - Redução do uso de memória RAM
  - Redução do espaço em disco de **3.8GB** para **3.4GB**

Higienização e Separação dos dados

- Padronização dos nomes de colunas
- Remoção de colunas que representam o mesmo dado (e.g. nome e sigla do time)
- Redução do tamanho dos arquivos
- Encapsulamento dos dados

Higienização e Separação dos dados


awards.pkl


games.pkl


playerBoxScores.pkl


players.pkl


playerTwitterFollowers.pkl


rosters.pkl


standings.pkl


targets.pkl

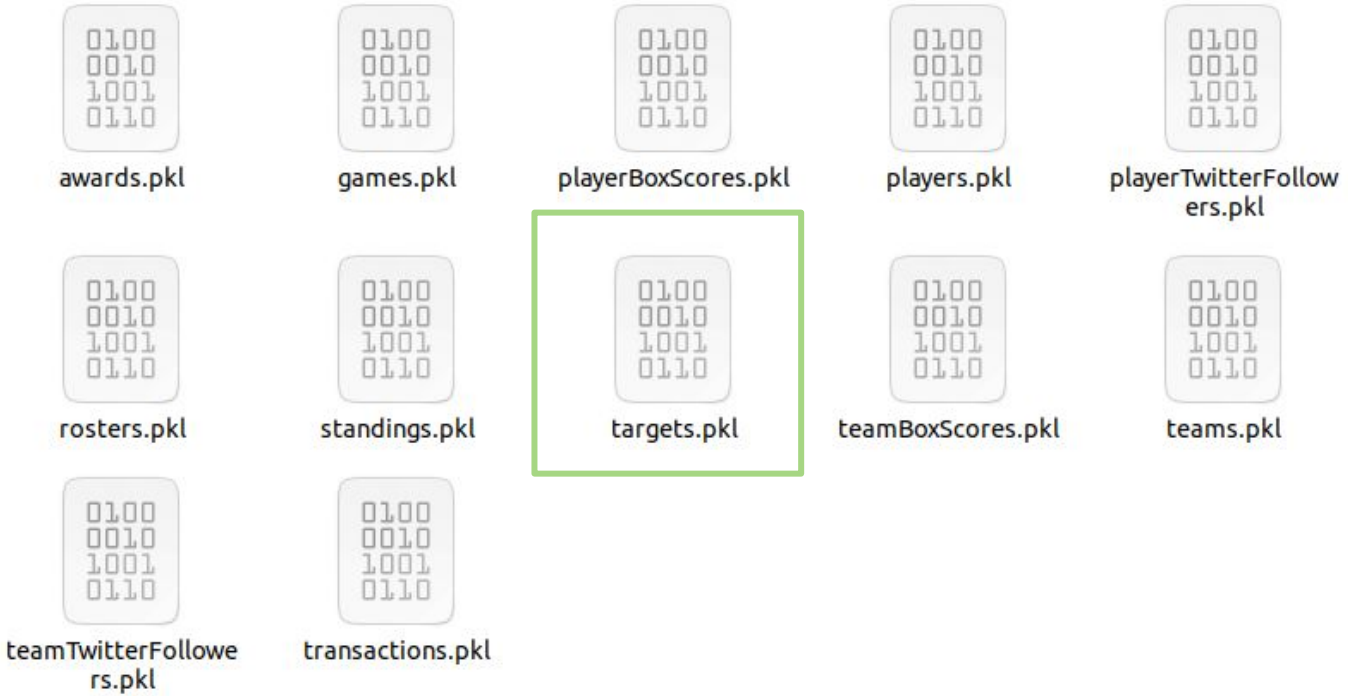

teamBoxScores.pkl


teams.pkl

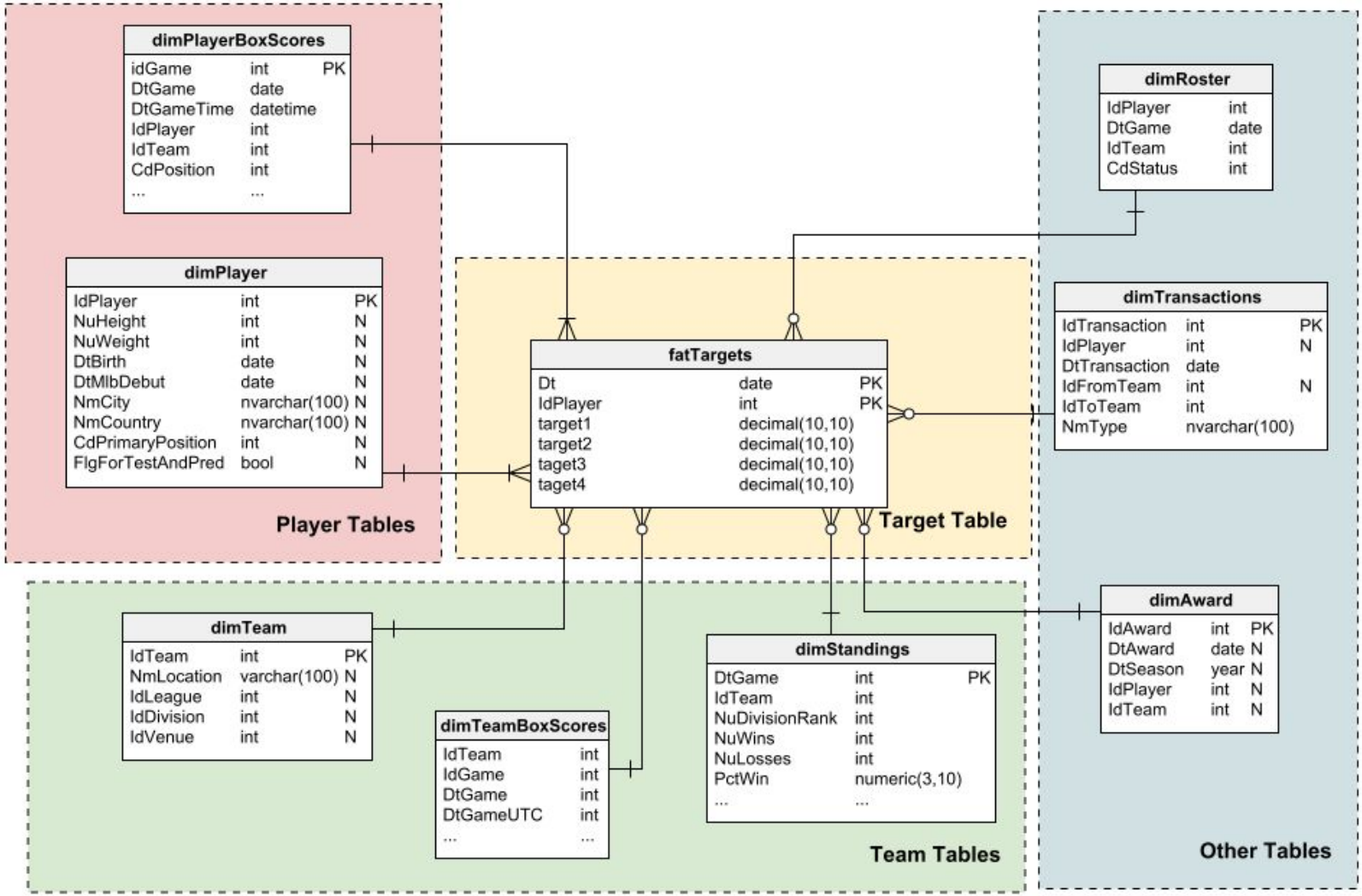

teamTwitterFollowers.pkl


transactions.pkl

Higienização e Separação dos dados



Modelo do Data Warehouse



2.
Análise Exploratória

Estatísticas de Resumo dos Targets

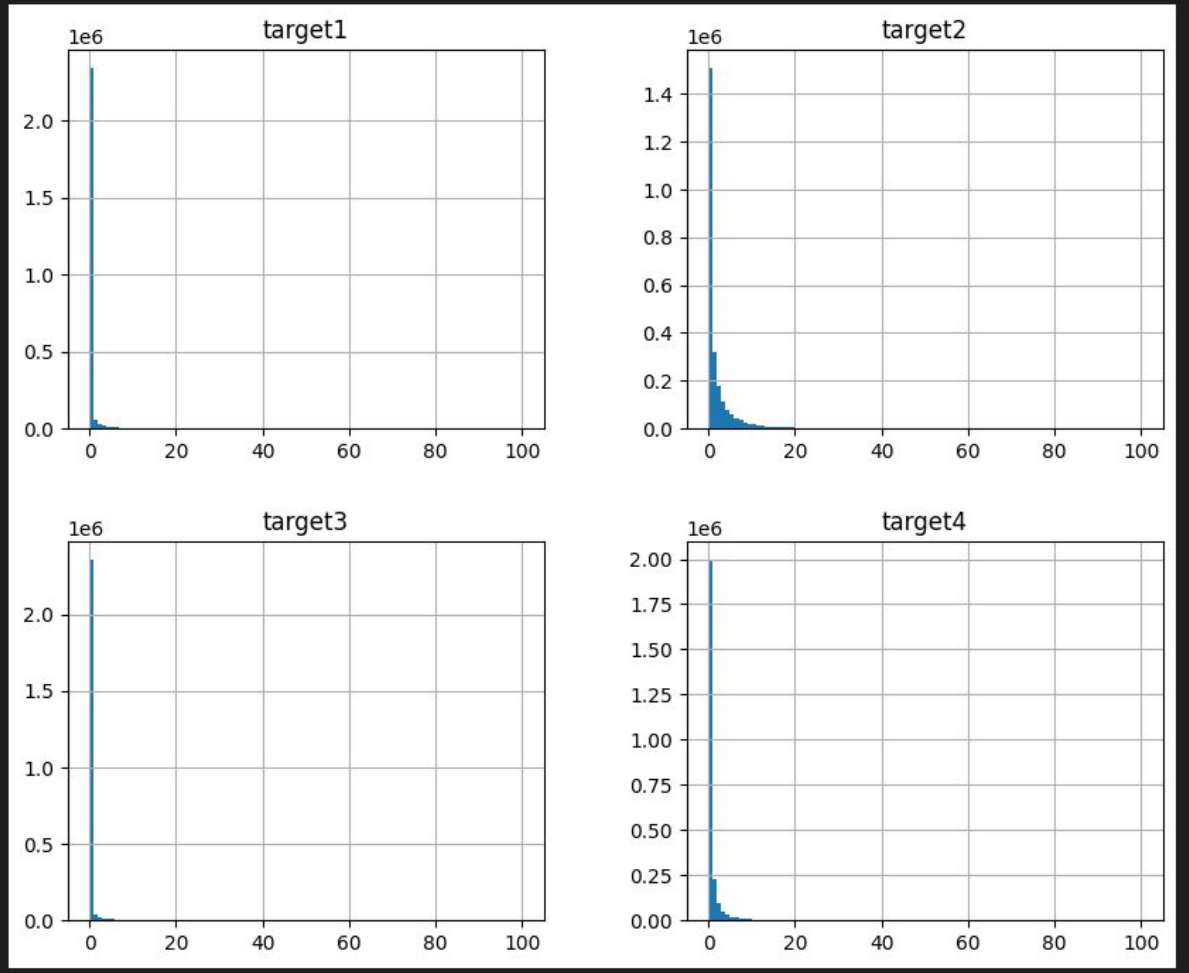
```
# describe the targets
trg_cols = ['target1', 'target2', 'target3', 'target4']
round(targets[trg_cols].describe(), 2)
```

	target1	target2	target3	target4
count	2506176.00	2506176.00	2506176.00	2506176.00
mean	0.57	2.46	0.69	1.14
std	4.17	6.23	5.07	4.23
min	0.00	0.00	0.00	0.00
25%	0.00	0.08	0.00	0.05
50%	0.00	0.56	0.00	0.22
75%	0.02	2.24	0.02	0.76
max	100.00	100.00	100.00	100.00

+ 2 milhões de registros
+ Os targets variam de 0 a 100
+ Médias baixas
+ Maioria dos valores 0 ou próximos de 0

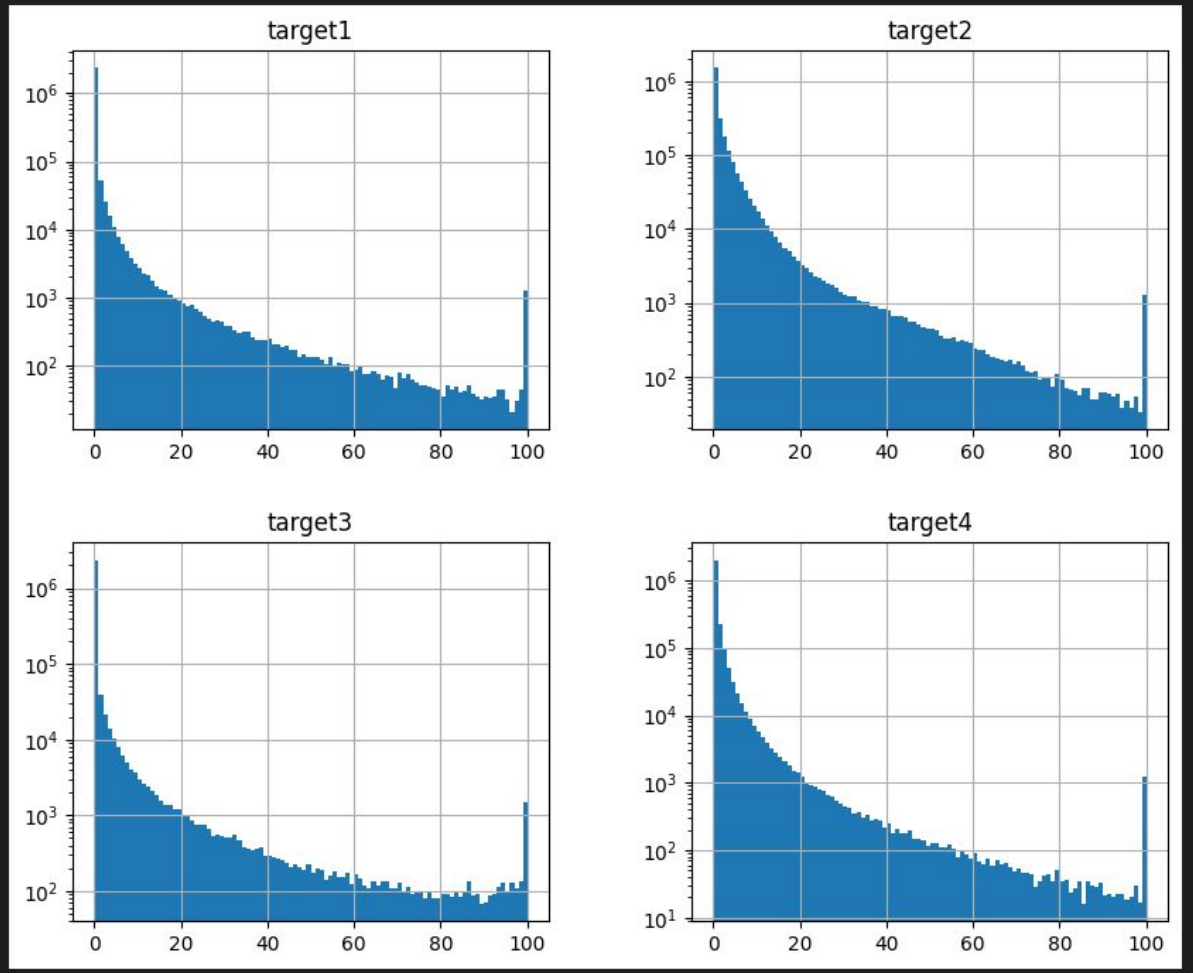
Histograma dos targets

```
targets[trg_cols].hist(figsize=(10, 8), bins=100);
```



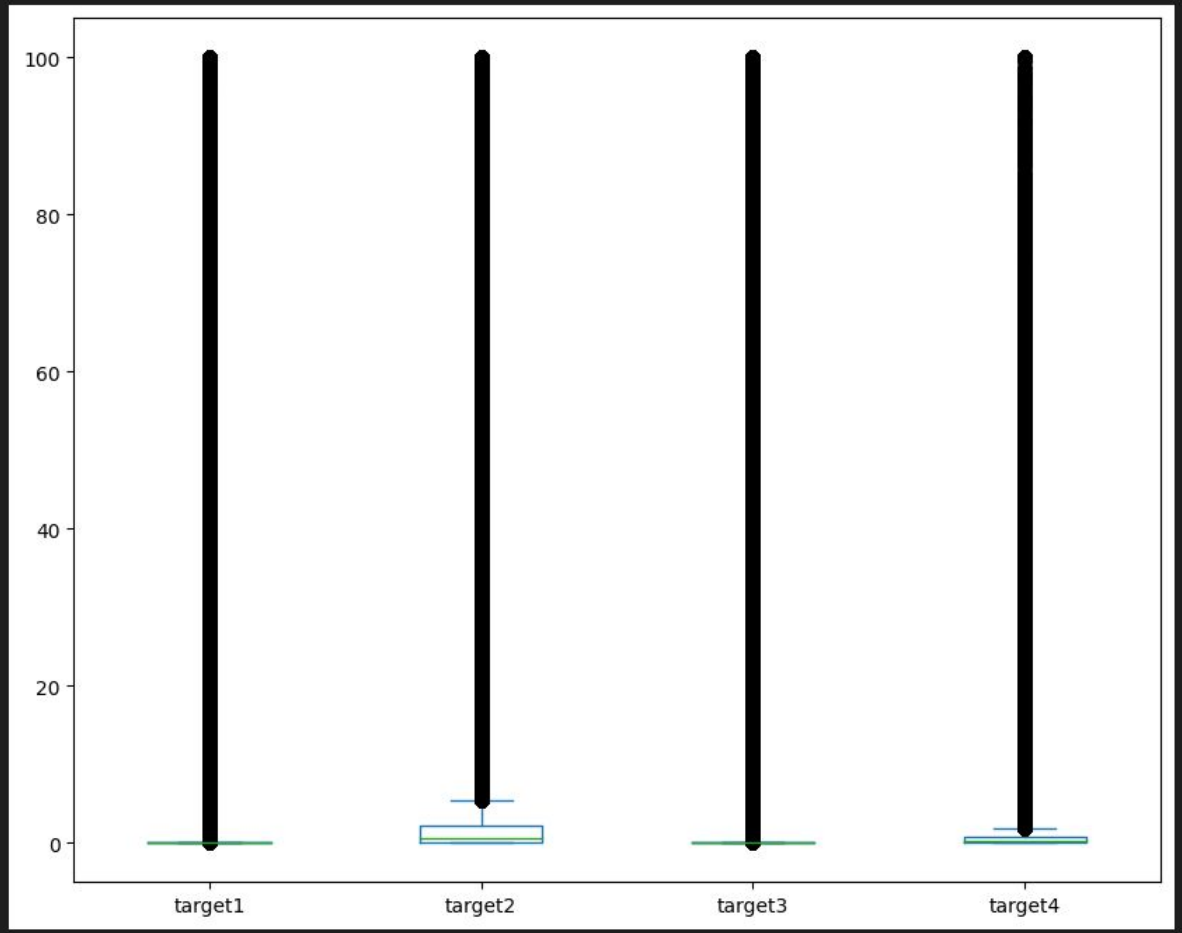
Histograma dos targets em escala logarítmica

```
targets[trg_cols].hist(figsize=(10, 8), bins=100, log=True);
```



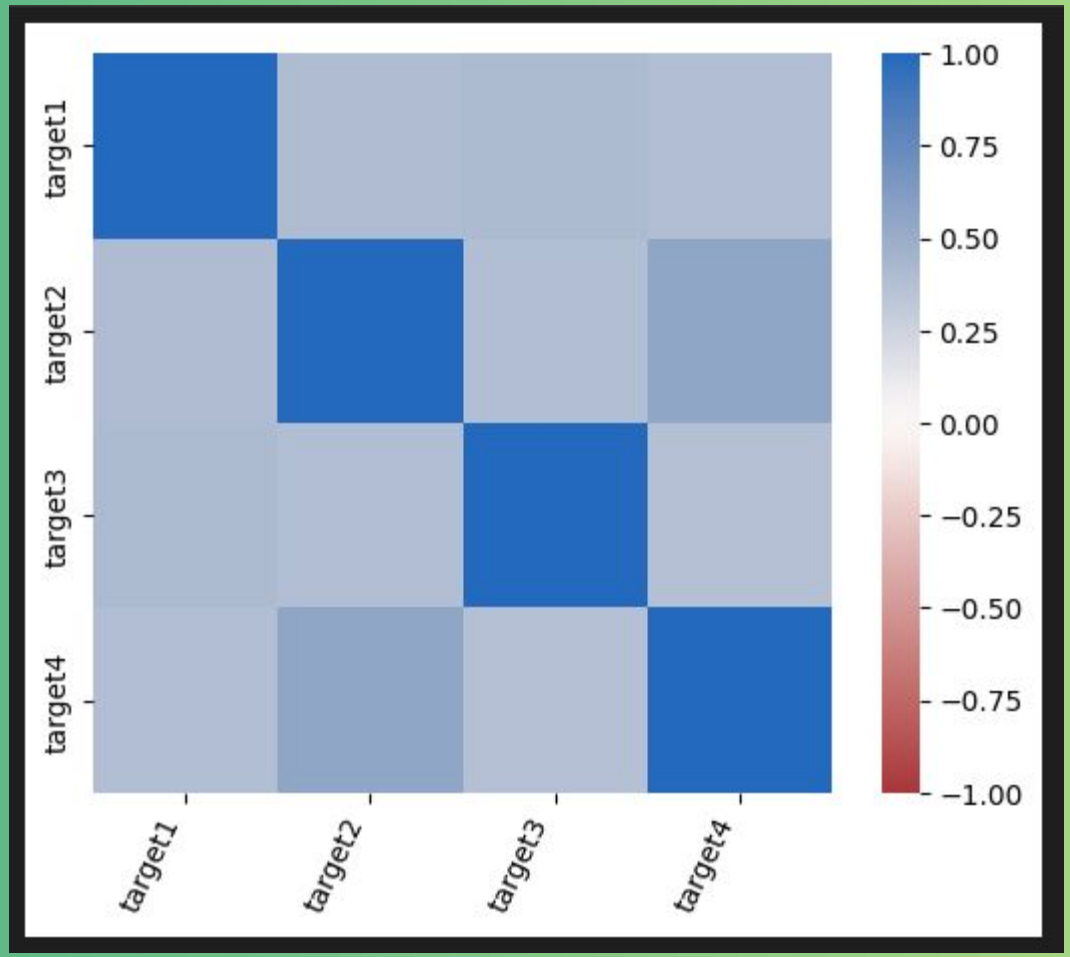
Boxplot dos Targets

```
targets[trg_cols].plot(kind='box', figsize=(10, 8));
```

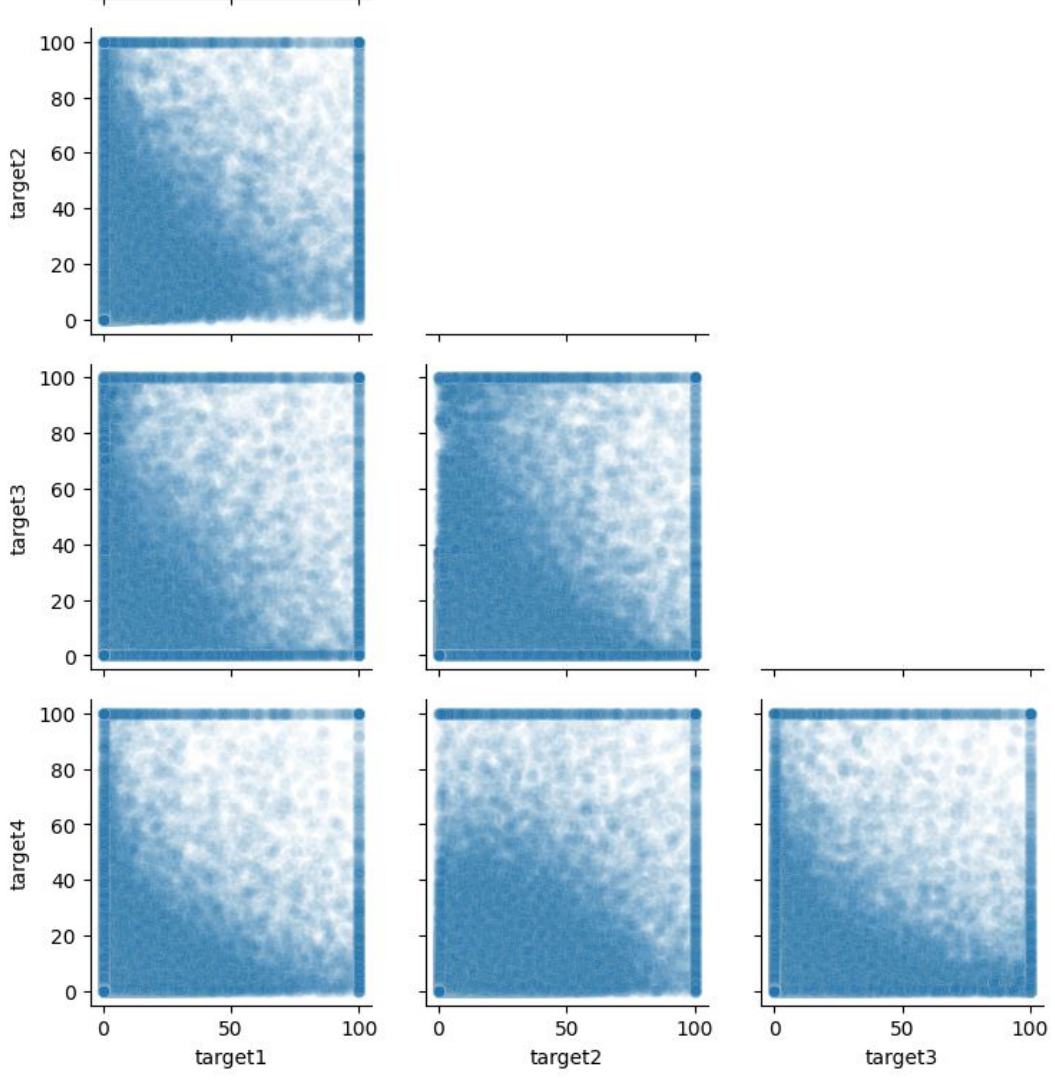


Correlação Entre os Targets

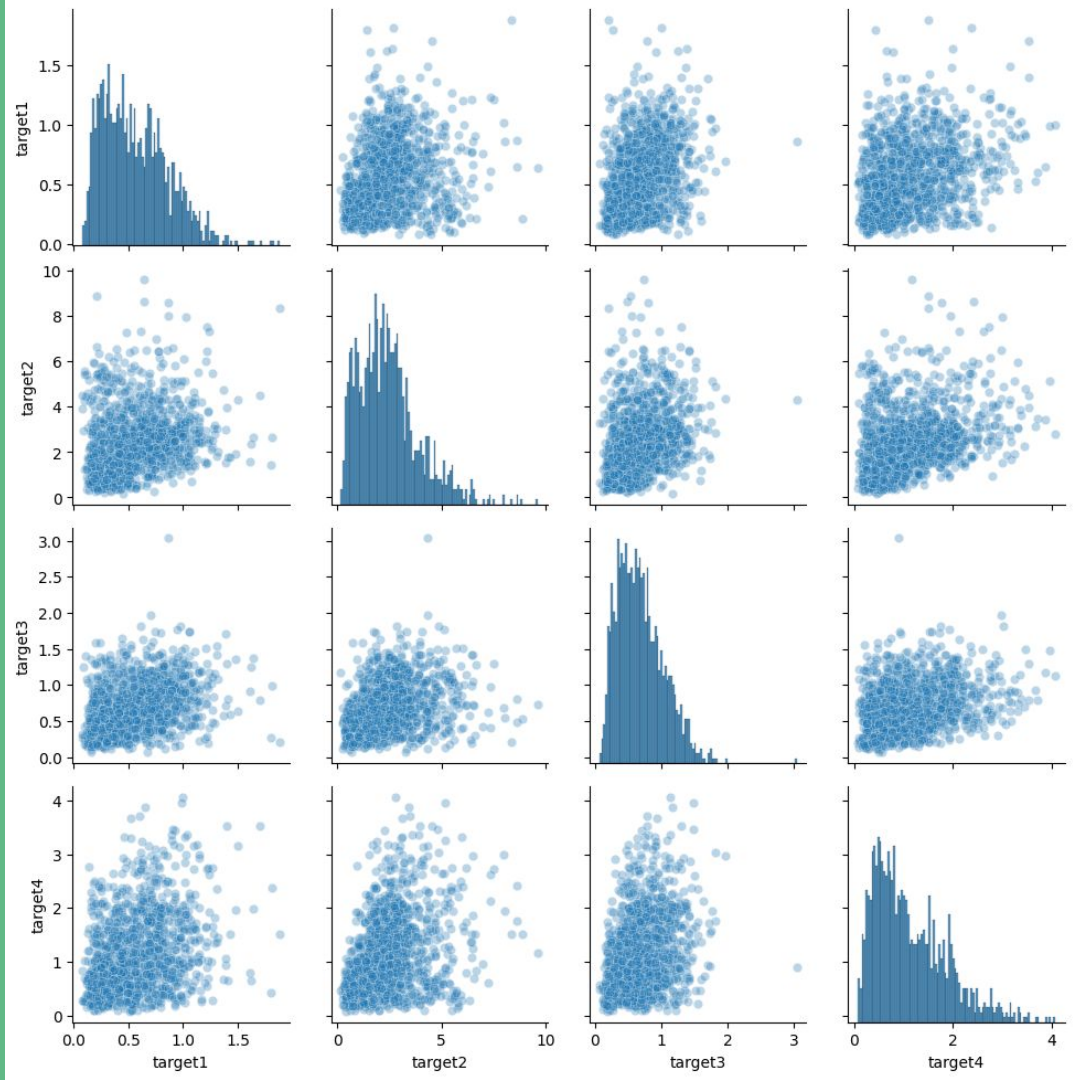
	target1	target2	target3	target4
target1	1.000000	0.404532	0.411024	0.384962
target2	0.404532	1.000000	0.388134	0.548991
target3	0.411024	0.388134	1.000000	0.370333
target4	0.384962	0.548991	0.370333	1.000000



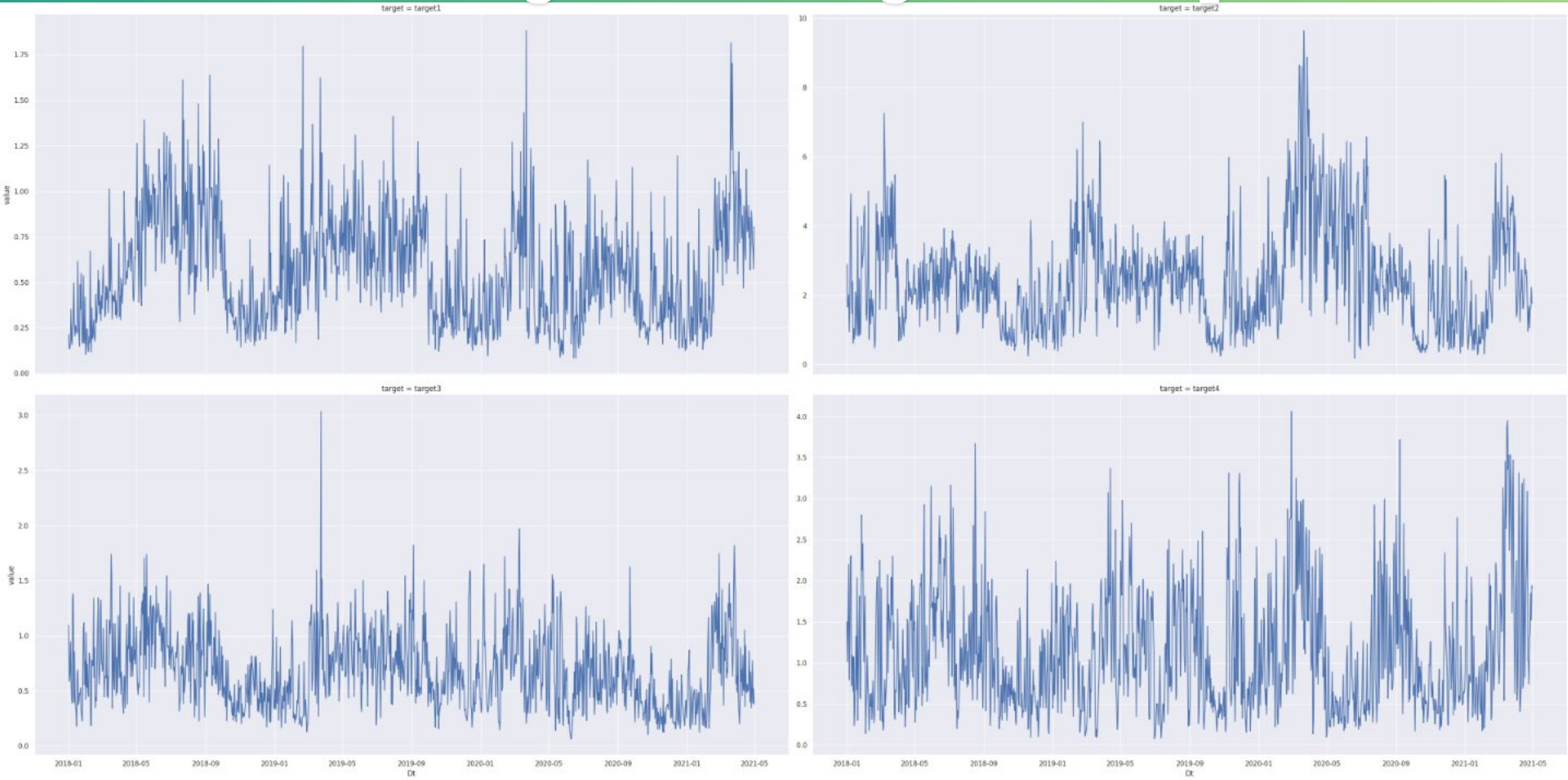
Scatterplot dos Targets



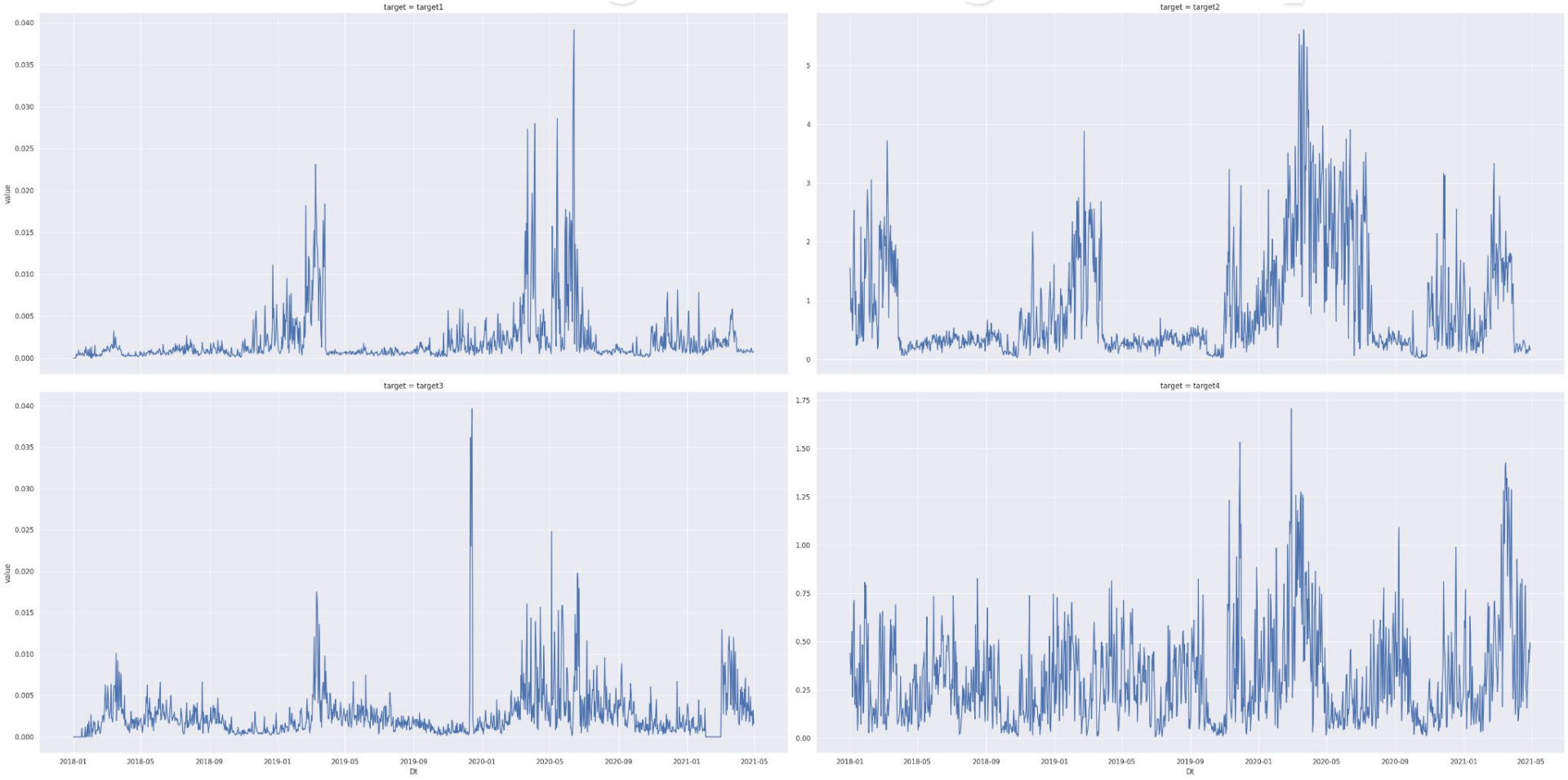
Média diária dos Targets

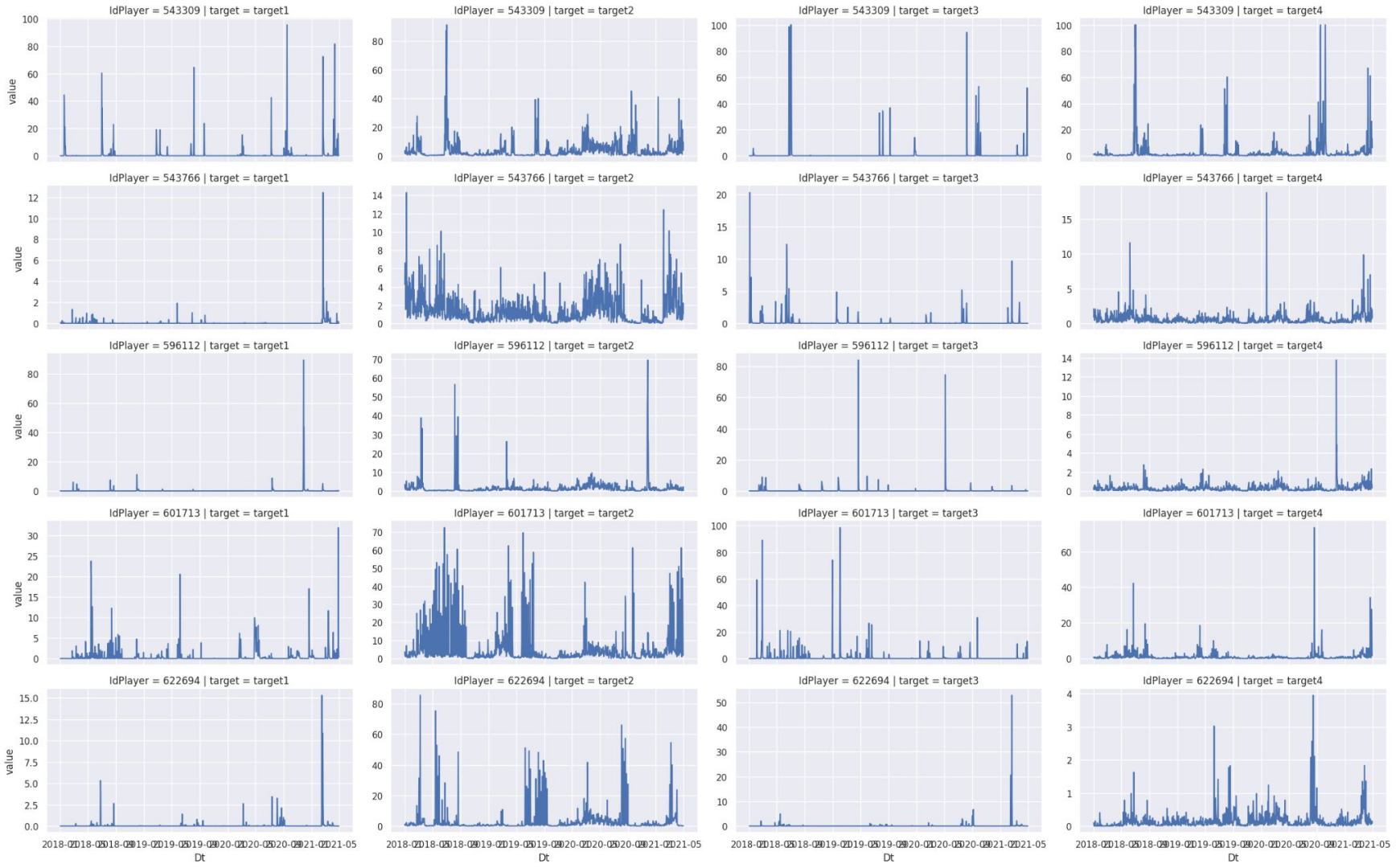


Média dos Targets ao longo do tempo

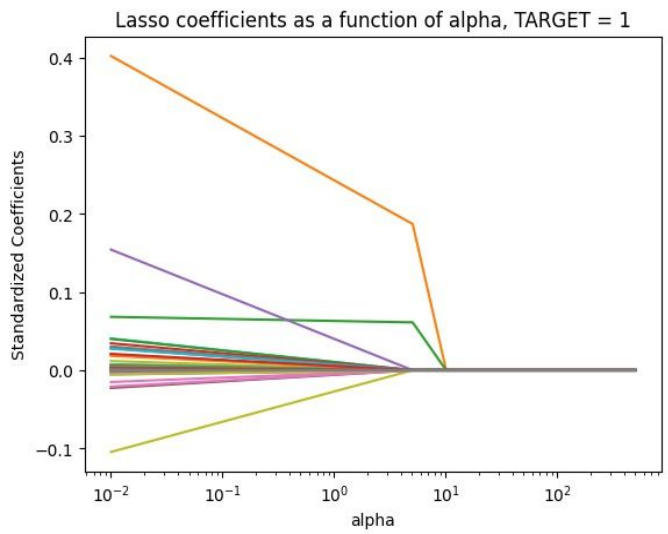


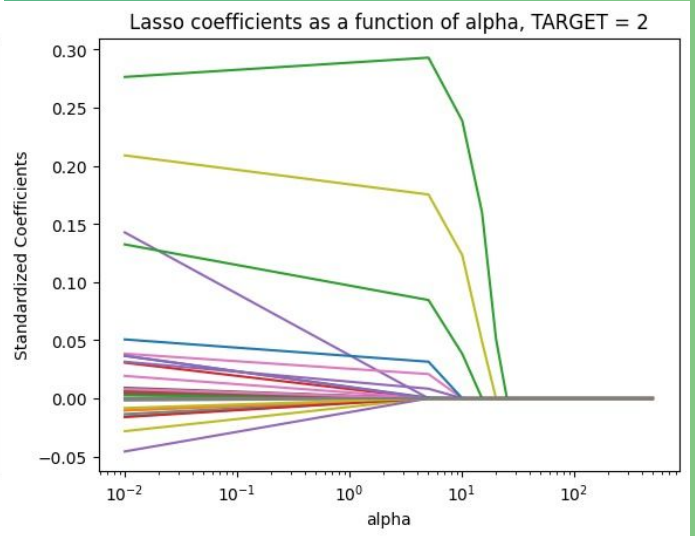
Mediana dos Targets ao longo do tempo

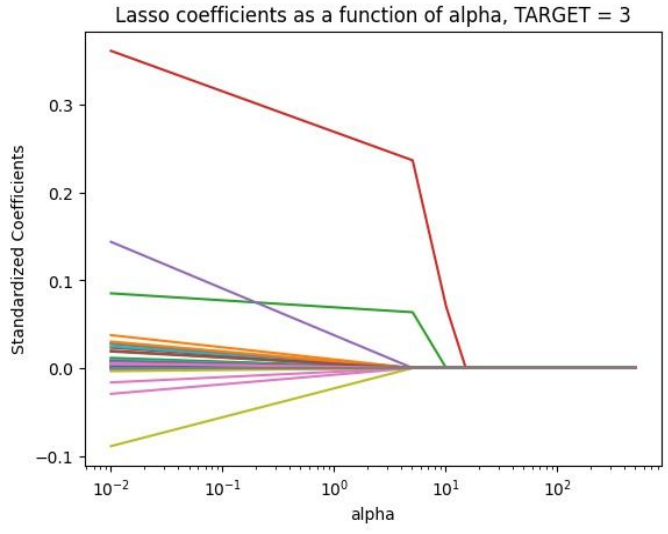


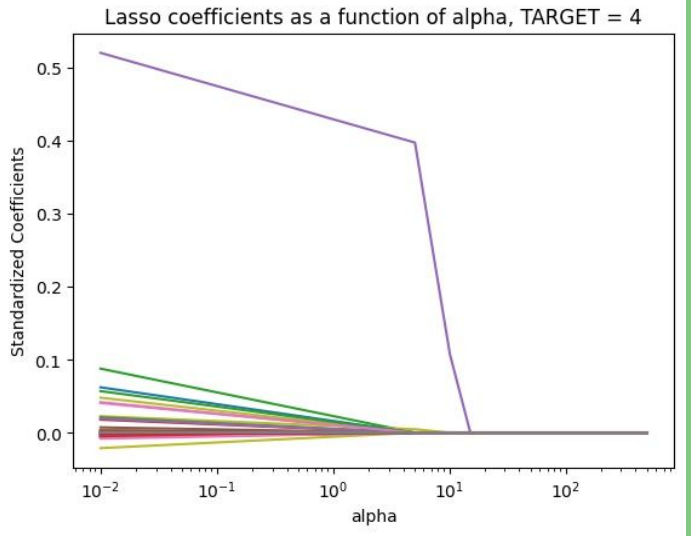

IdPlayer = 543309 | target = target1
IdPlayer = 543309 | target = target2
IdPlayer = 543309 | target = target3
IdPlayer = 543309 | target = target4
IdPlayer = 543766 | target = target1
IdPlayer = 543766 | target = target2
IdPlayer = 543766 | target = target3
IdPlayer = 543766 | target = target4
IdPlayer = 596112 | target = target1
IdPlayer = 596112 | target = target2
IdPlayer = 596112 | target = target3
IdPlayer = 596112 | target = target4
IdPlayer = 601713 | target = target1
IdPlayer = 601713 | target = target2
IdPlayer = 601713 | target = target3
IdPlayer = 601713 | target = target4
IdPlayer = 622694 | target = target1
IdPlayer = 622694 | target = target2
IdPlayer = 622694 | target = target3
IdPlayer = 622694 | target = target4
value
Dt


IdPlayer = 445197 | target = target1
IdPlayer = 445197 | target = target2
IdPlayer = 445197 | target = target3
IdPlayer = 445197 | target = target4
IdPlayer = 489295 | target = target1
IdPlayer = 489295 | target = target2
IdPlayer = 489295 | target = target3
IdPlayer = 489295 | target = target4
IdPlayer = 542364 | target = target1
IdPlayer = 542364 | target = target2
IdPlayer = 542364 | target = target3
IdPlayer = 542364 | target = target4
IdPlayer = 570663 | target = target1
IdPlayer = 570663 | target = target2
IdPlayer = 570663 | target = target3
IdPlayer = 570663 | target = target4
IdPlayer = 606192 | target = target1
IdPlayer = 606192 | target = target2
IdPlayer = 606192 | target = target3
IdPlayer = 606192 | target = target4
value
Dt


Lasso coefficients as a function of alpha, TARGET = 1
Standardized Coefficients
alpha


Lasso coefficients as a function of alpha, TARGET = 2
Standardized Coefficients
alpha


Lasso coefficients as a function of alpha, TARGET = 3
Standardized Coefficients
alpha


Lasso coefficients as a function of alpha, TARGET = 4
Standardized Coefficients
alpha

Projetos em Ciência de Dados

Gianlucca Devigili e Maisa O. Fraiz