IA PARA A ESTIMAÇÃO DE JOGADORES DE FUTEBOL

SCC0230 - Inteligência Artificial (2023)

Rogério Lopes Lübe

Bernardo Maia Coelho

João Gabriel Sasseron Roberto Amorim

João Pedro Buzzo Silva

Pedro Guilherme dos Reis Teixeira

Marcos Patrício Nogueira Filho

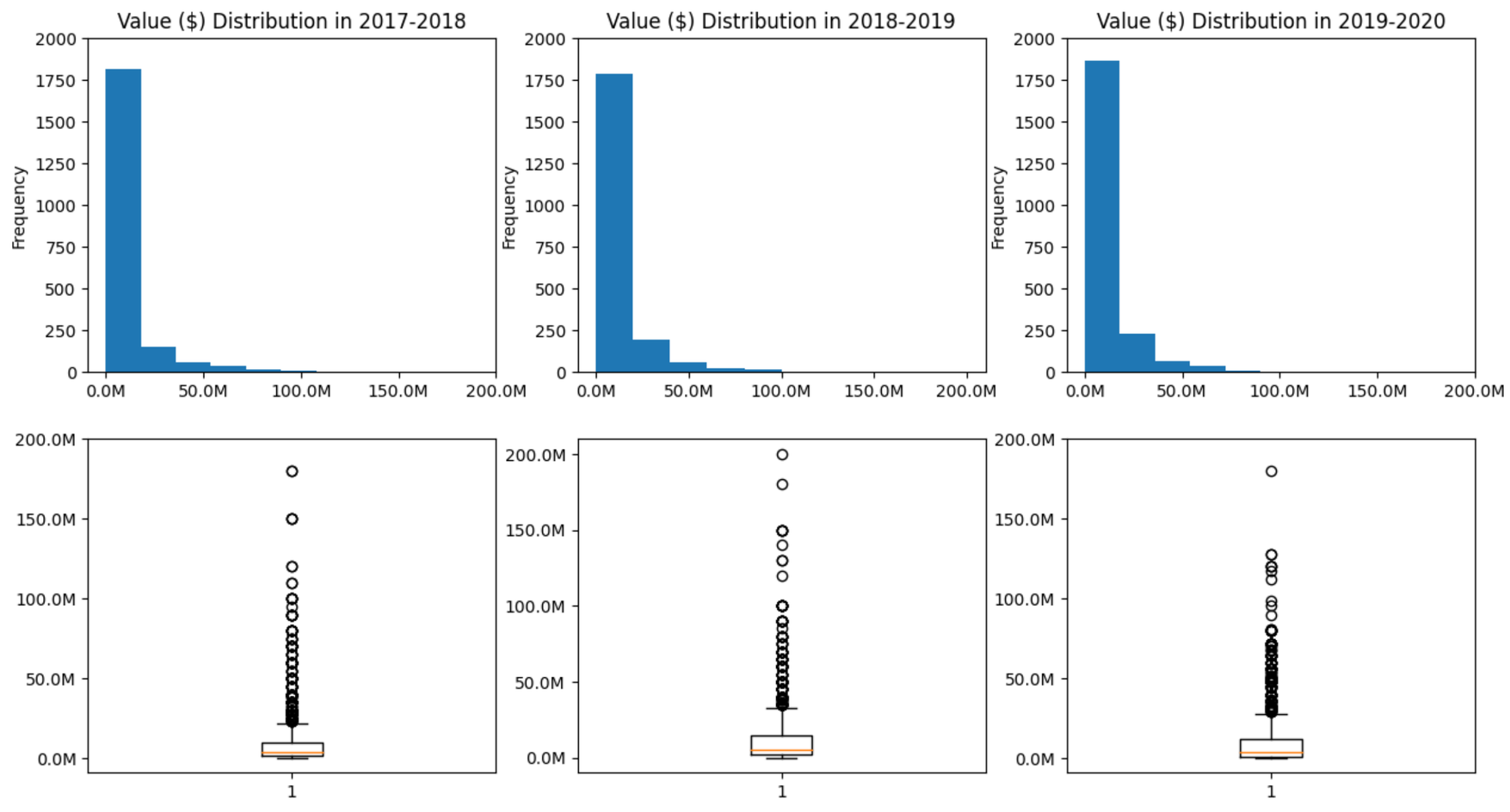
ICMC USP SÃO CARLOS

TÓPICOS

- Exploração dos Dados
  - Apresentação Gráfica
- Pre-Processamento
- Aplicação de Classificações

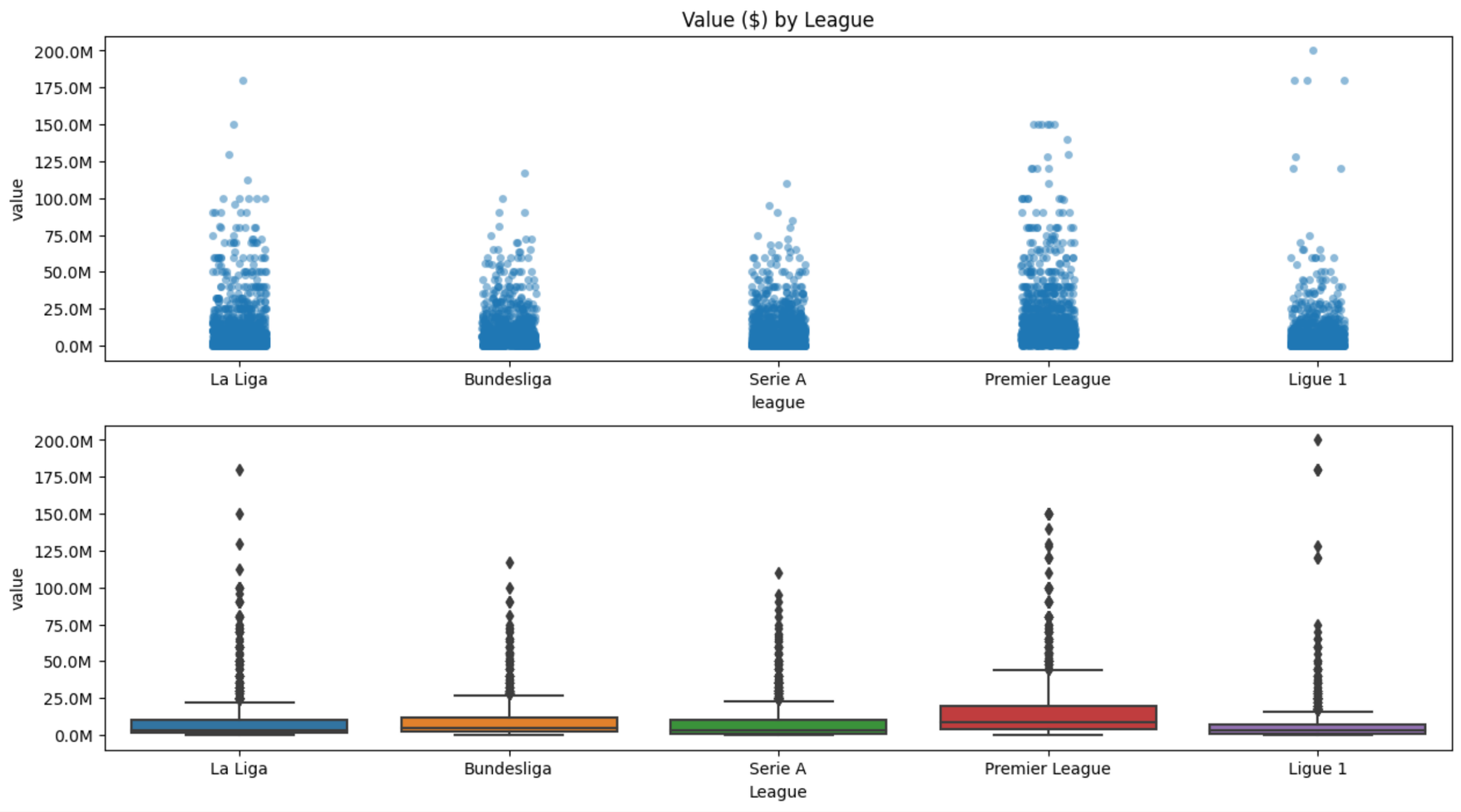
VALOR POR TEMPORADA

Maioria entre 0 e 20 milhões, poucos chegam a 100 milhões.



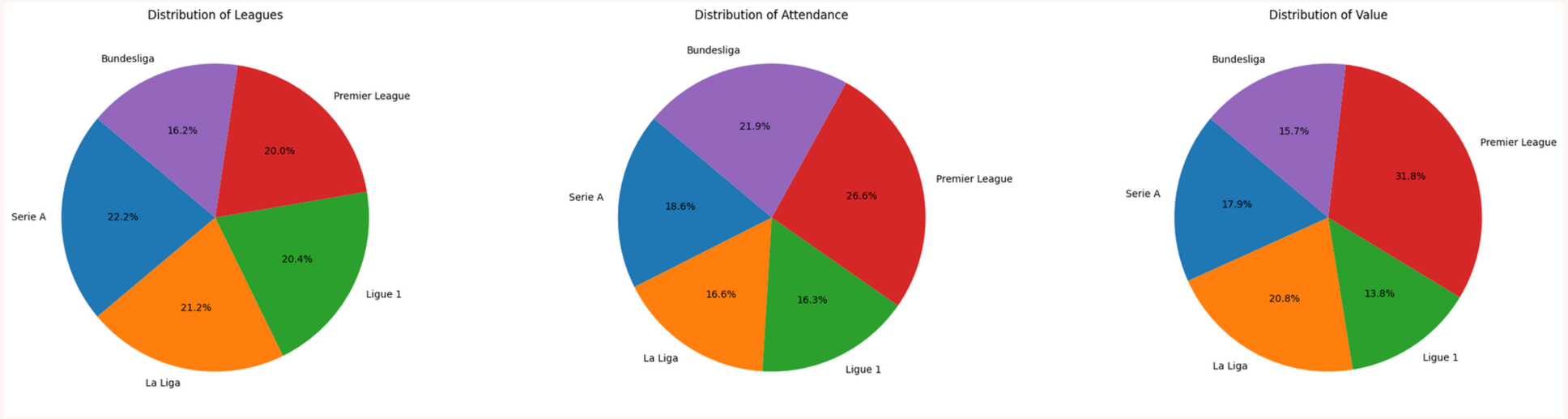
PREÇO JOGADOR X LIGA

Premier League maior valor médio, Ligue 1 menor valor médio porém com outliers significantes como o Neymar.



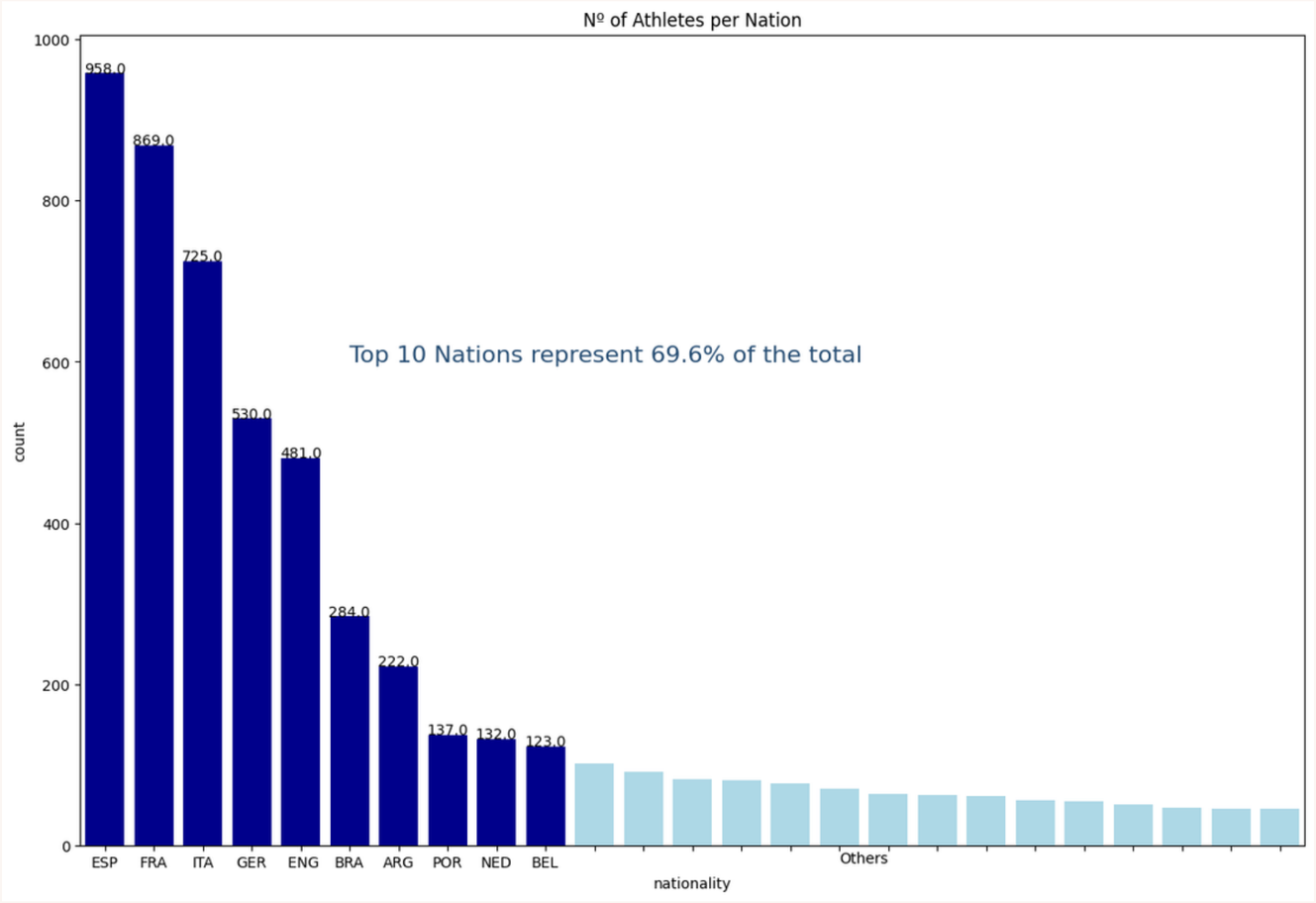
DISTRIBUIÇÃO DE JOGADORES, VALUE E ATTENDANCE

Premier League segunda menor quantidade de jogadores no dataset, ainda com a maior attendance e value totais, o ultímo por uma margem significativa.



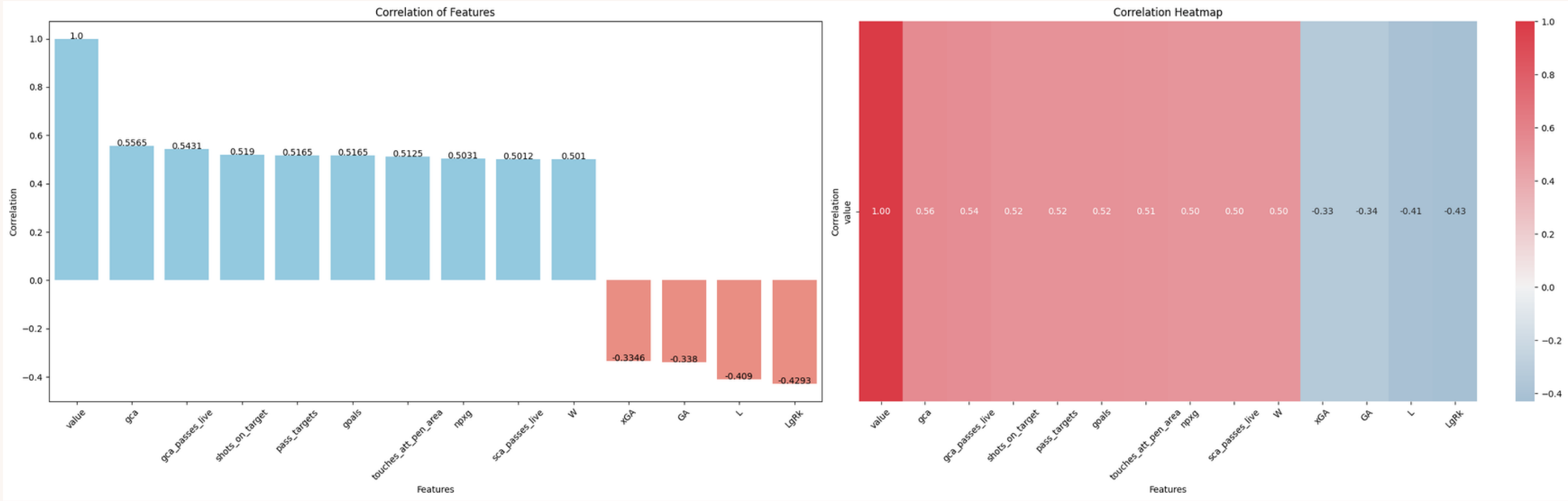
JOGADORES POR NACIONALIDADE

Como esperado a grande maioria dos jogadores no dataset são europeus.



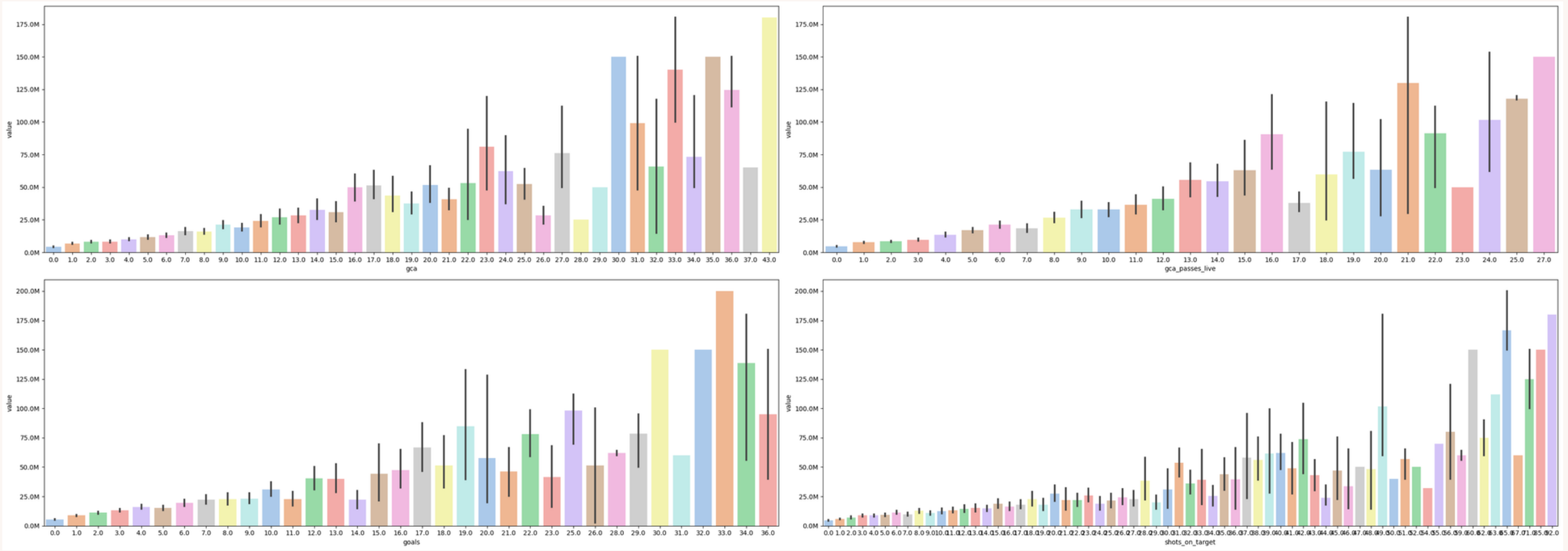
CORRELAÇÃO DE DADOS COM RELAÇÃO A VALOR

Maior correlação é com gca que representa a presença do jogador no campo, L que representa número de derrotas do jogador
é um dos exemplos de correlação negativa.



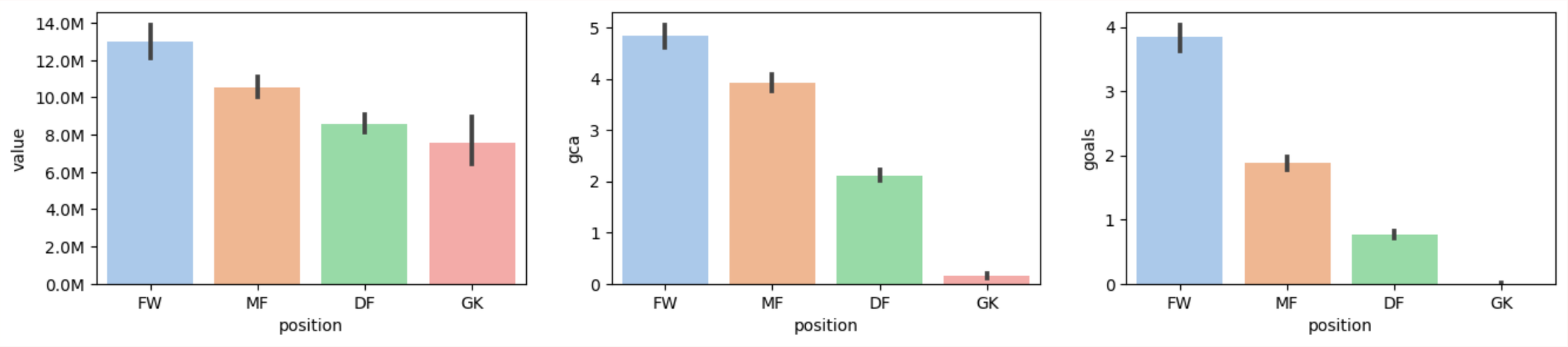
DISTRIBUIÇÃO DE DADOS ALTAMENTE CORRELACIONADOS COM VALOR

A relação calculada é apresentada evidentemente nas distribuições, sendo que o incremento das métricas o valor médio também tendo a aumentar.



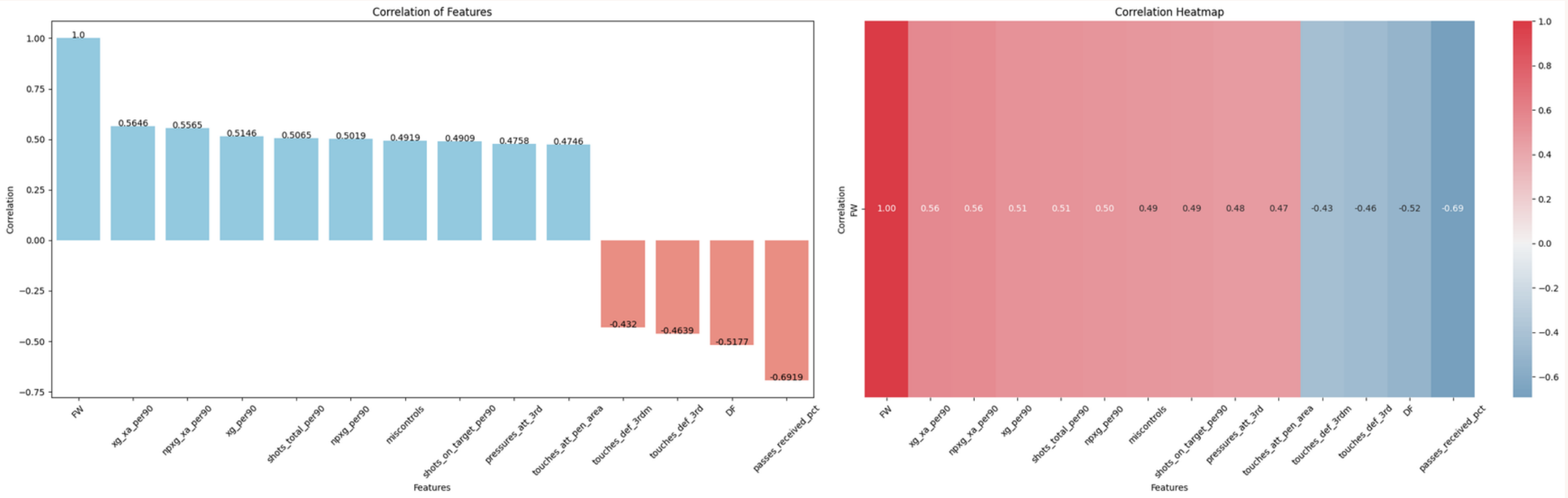
DISTRIBUIÇÃO DE POSIÇÕES COM RELAÇÃO A VALOR, GCA E NÚMERO DE GOLS

Atacantes possuem maior valor médio, goleiros possuem menor. Como é esperado goleiros em médio não fazem nenhum gol.

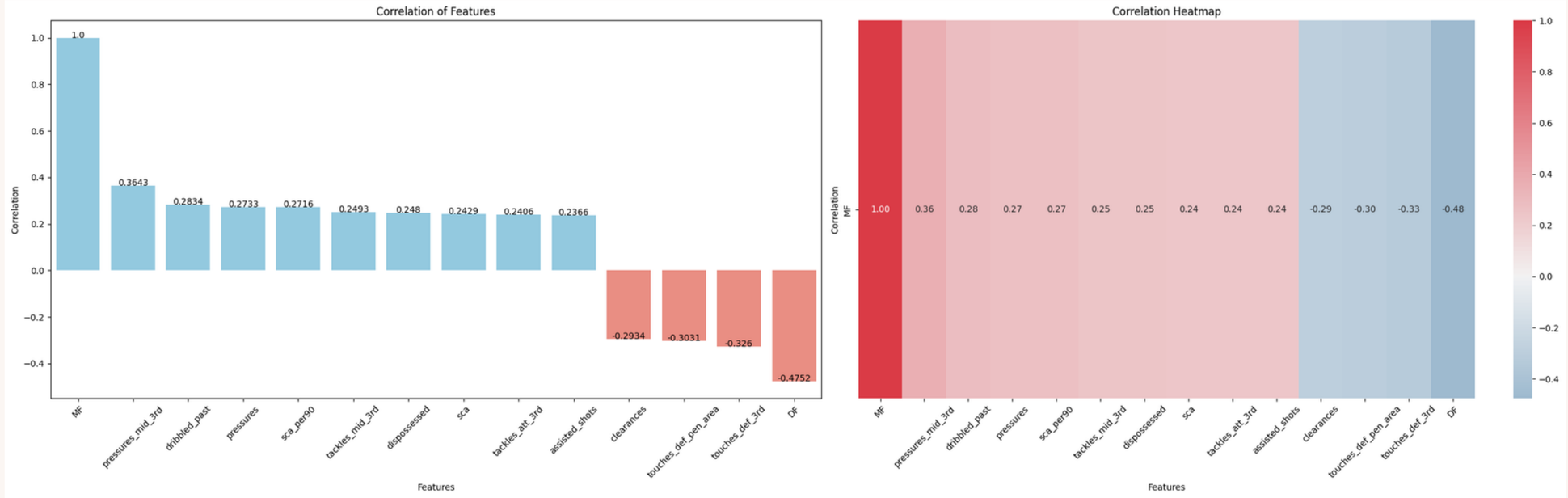


CORRELAÇÃO DE POSIÇÕES COM VALORES

Atacante

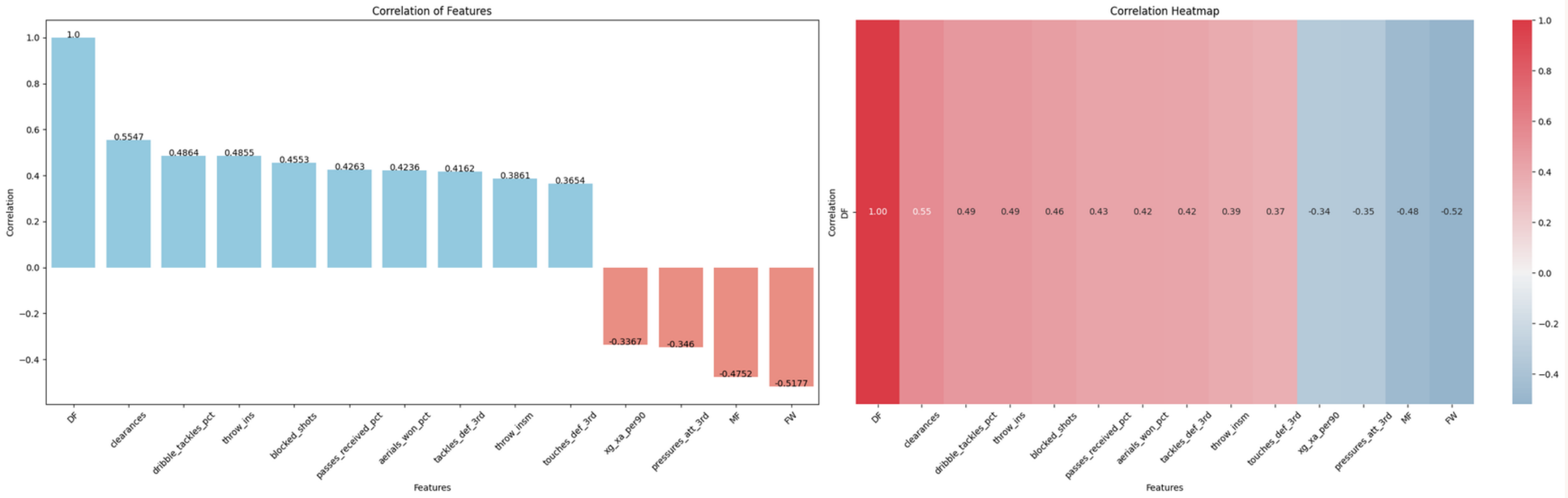


Meio de Campo

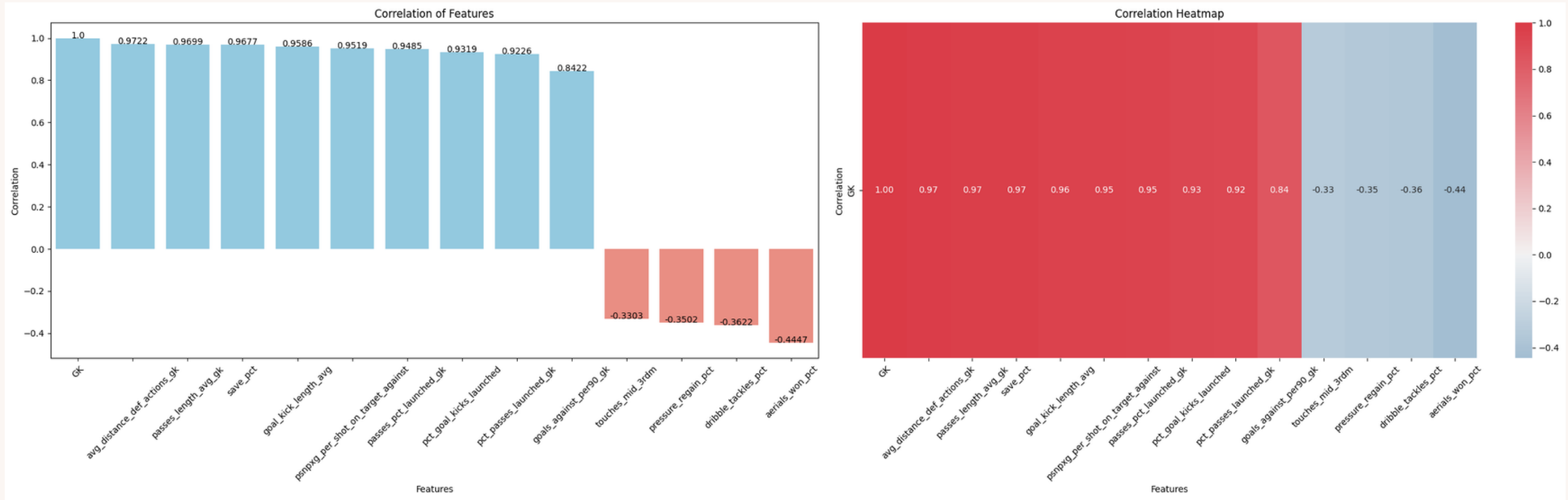


CORRELAÇÃO DE POSIÇÕES COM VALORES

Defensor

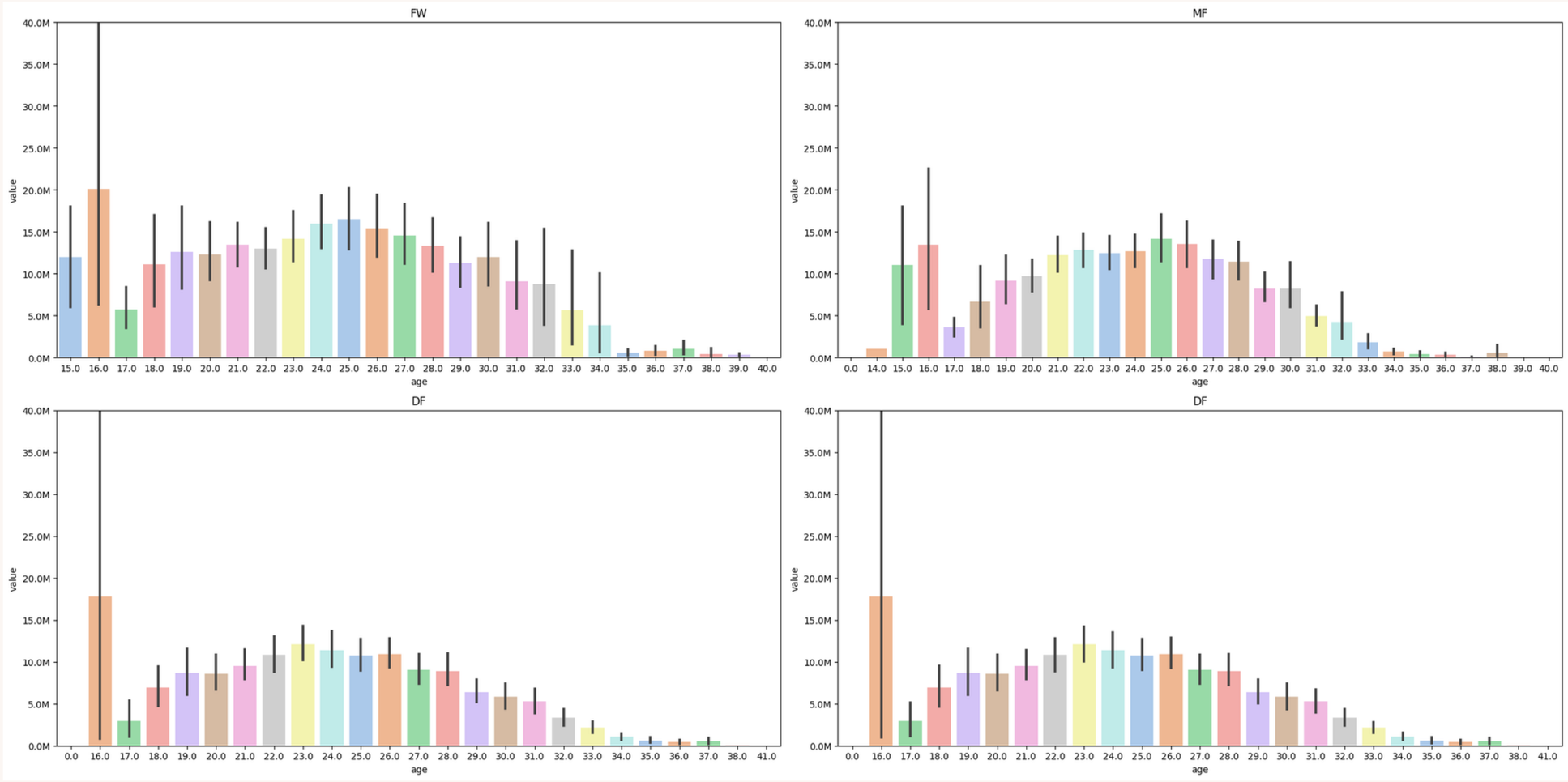


Goleiro



DISTRIBUIÇÃO DE VALOR COM RELAÇÃO A IDADE PARA CADA POSIÇÃO

Traço comum é o curto pico aos 16 anos, seguidos por uma queda, crescimento e estabilização de valor durante o período dos 20 anos.



ETAPAS DE PRE-PROCESSAMENTO

- Primariamente **foram removidos dados desnecessários** para o treinamento do modelo. No caso a coluna que representava o nome dos jogadores foi retirada.
- Em sequência foram **retirados todos os valores nulos** do dataset evitando problemas na execução do algoritmo.

ETAPAS DE PRE-PROCESSAMENTO

- Considerando a presença de vários valores categoricos como por exemplo as **posições dos jogadores** (FW, MF, DF, GK) foi implementado o **hot enconding** de todos os dados que se encaixam. Assim varias novas colunas que representam cada categoria foram adicionadas, sendo que uma linha possui 1 em uma coluna se pertencer a categoria e 0 se não.

ETAPAS DE PRE-PROCESSAMENTO

- A próxima etapa foi a refatoração dos valores da coluna ‘Attendance’, **trocando seu tipo de string para float**.
- Por fim todas as colunas representadas por valores numéricos foram exploradas a **procura de outliers**, os encontrados não foram retirados do dataset em receio de serem fundamentais para o aprendizado porém estão armazenados para análises caso necessário.

ETAPAS DE PROCESSAMENTO

- Em primeiro lugar os **dados** pré-processados devem ser **aleatoriamente distribuídos** em conjuntos de treino e de teste.
- Os **dados são então redimensionados** a fim de garantir uma influência uniforme dos dados durante o treinamento dos modelos preditivos.

ETAPAS DE PROCESSAMENTO

- Foram utilizados **diversos modelos de previsão**: modelos de regressão como Linear Regression, Ridge e Lasso; modelos baseados em árvores de decisão como Random Forest e Gradient Boosting; modelos baseados em redes neurais como MLPRegressor; e outros métodos como SVR e LGBMRegressor.

ETAPAS DE PROCESSAMENTO

- Os modelos citados passam então por uma **Validação Cruzada de 5 folds** a fim de encontrar configurações ótimas de parâmetros para que os modelos se ajustem aos dados.

AVALIAÇÃO

- O algoritmo que apresentou o **melhor desempenho** foi o **LGBMRegressor** com:
  - **Parâmetros**:
    - **learning_rate**: 0.1;
    - **n_estimators**: 200.
  - **Best Score** (RMSE): 8.782069e+06

AVALIAÇÃO

- **Métricas** que serão utilizadas para a avaliação:
  - Mean Absolute Error (**MAE**)
  - Mean Squared Error (**MSE**)
  - Root Mean Squared Error (**RMSE**)
  - R^2 Score (**R-squared**)

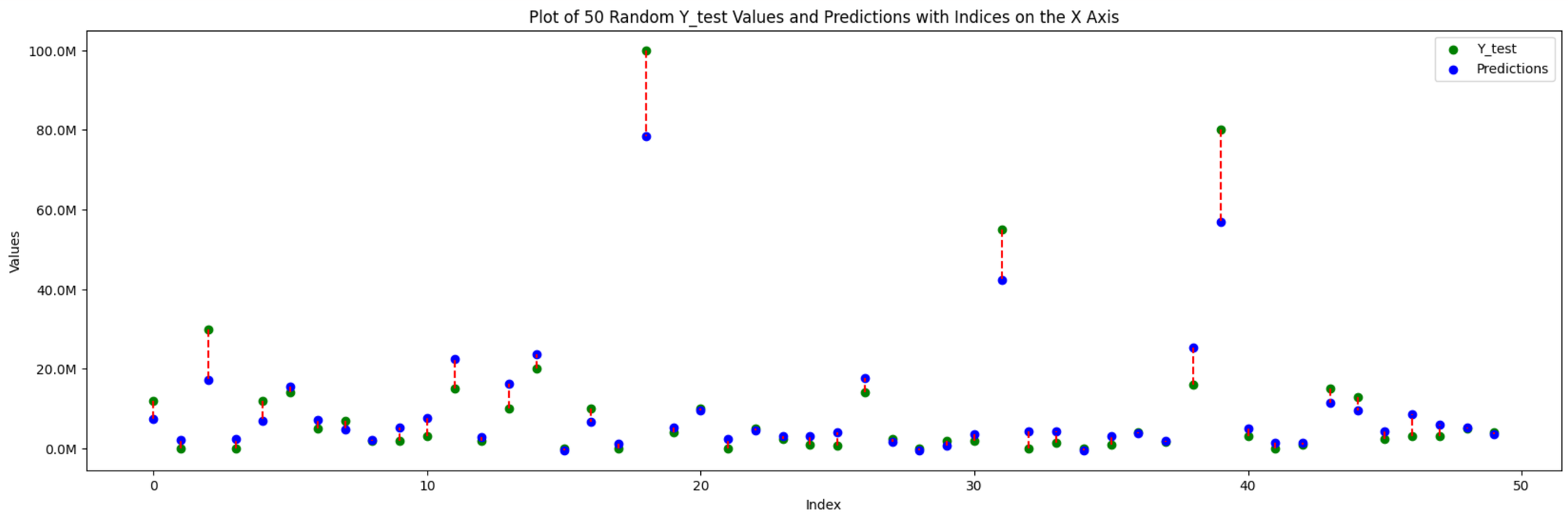
AVALIAÇÃO

- Apresentação dos **resultados** do **melhor algoritmo**:

```
################################ LGBMRegressor ################################
Mean Absolute Error (MAE):      4.232256e+06
Mean Squared Error (MSE):       5.416315e+13
Root Mean Squared Error (RMSE): 7.359562e+06
R² Score (R-squared):   0.8114547460702778
```

AVALIAÇÃO

- Visualização gráfica entre predições e valores esperados:



OBRIGADO PELA ATENÇÃO