Modelling Football Players Valuation

DATA SCIENCE PROJECT

HTTPS://GITHUB.COM/MARCOWONG96/CAPSTONE-PROJECT

Marco Wong

PRESENTATION OUTLINE

INSPIRATION

PROJECT OVERVIEW

Create a model that predicts a player value, given the players statistics in the previous season. The player must be in one of the top 5 leagues.

Test against historical transfer fees to validate accuracy.

STAKEHOLDERS

Objective n° 1

To offer clubs a reliable method to assess player value. Will provide the ability to value potential signings and existing players.

Objective n° 2

To provide additional insight into what makes a player valuable. Are certain traits currently undervalued or overvalued?

PROJECT PROGRESSION

- DATA COLLECTION
- DATA PREPROCESSING
- DATA CLEANING
- EXPLORATORY DATA ANALYSIS
- FEATURE ENGINEERING
- MODEL SELECTION/ TRAINING
- MODEL TUNING/ TESTING

DATASET OVERVIEW

2022 – 2023 Player Statistics
2689 rows x 124 columns

2021 – 2022 Player Statistics
2921 rows x 143 columns

Player Valuation
440643 rows x 9 columns

Player List
30302 rows x 23 columns

Transfers in 2022 and 2023 (Scraped Data)

Player Statistics Datasets

CATEGORICAL VARIABLES

- **Player:** *Player name*
- **Nation:** *Player Nationality*
- **Pos:** *Position*
- **Squad:** *Player Team*
- **Comp:** *Player Competition*

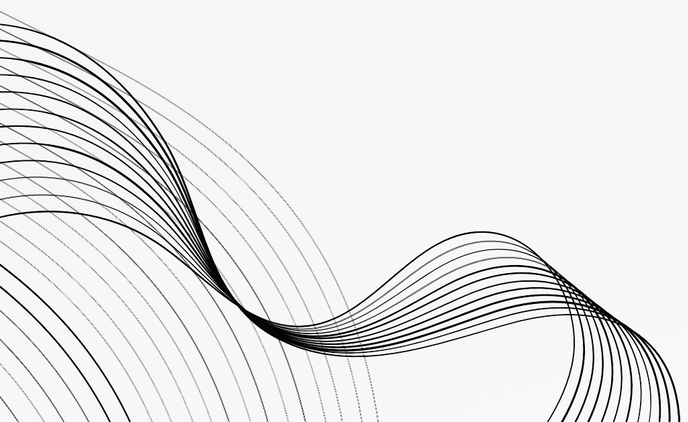
EXAMPLE NUMERIC VARIABLES

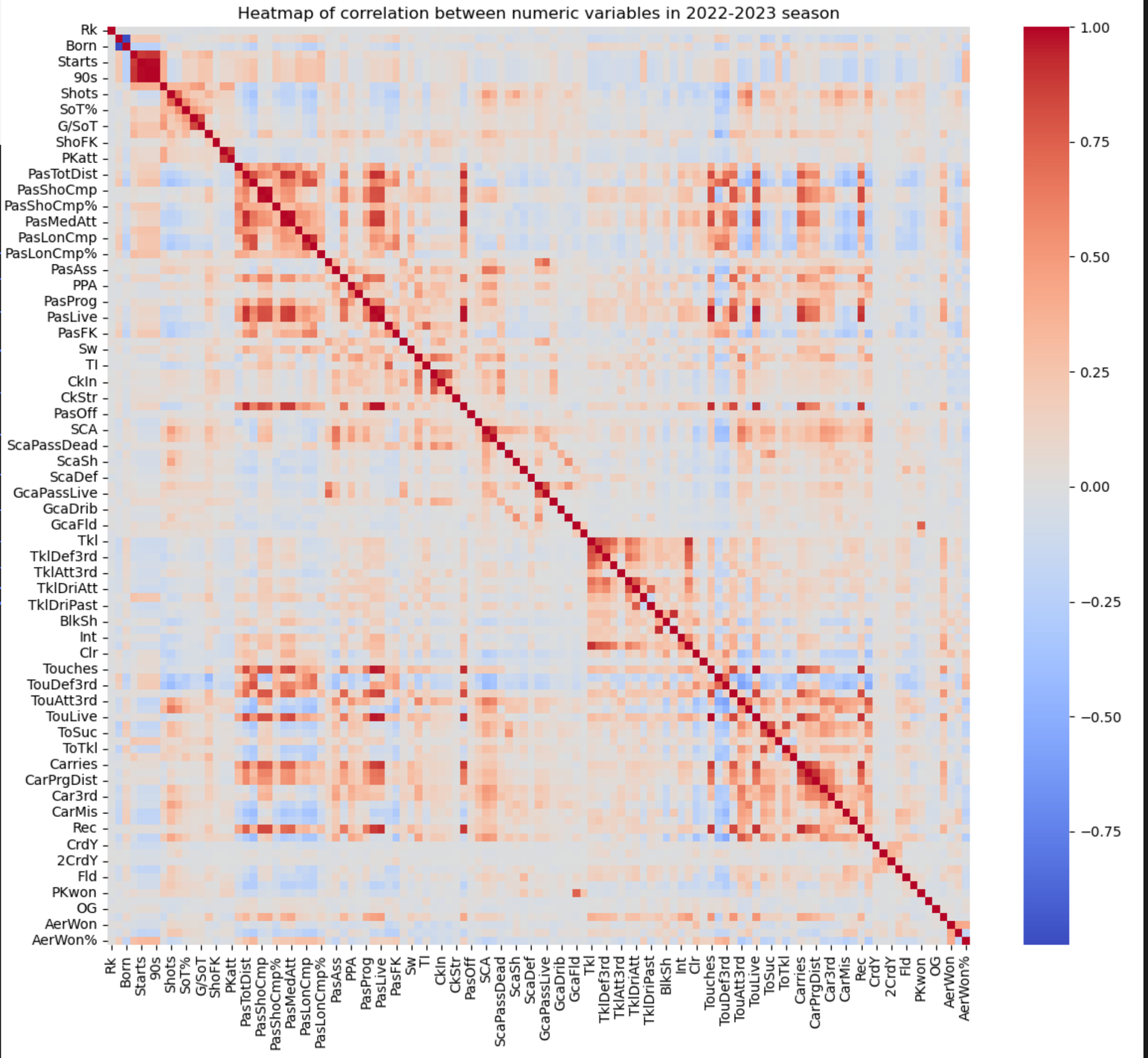
- **PasAss:** *Passes that directly lead to a shot (assisted shots)*
- **TklDriAtt:** *Number of times dribbled past plus number of tackles*
- **Blocks:** *Number of times blocking the ball by standing in its path*
- **Touches:** *Number of times a player touched the ball.*
- **TouDefPen:** *Touches in defensive penalty area*

Player Valuation Datasets

The player valuation dataset is determined by TransferMarkt, a third party company. This will be used as the training data.

The Transfer Fees dataset, which will be scraped, will be used as the test data. Ultimately, this is what we are trying to measure.

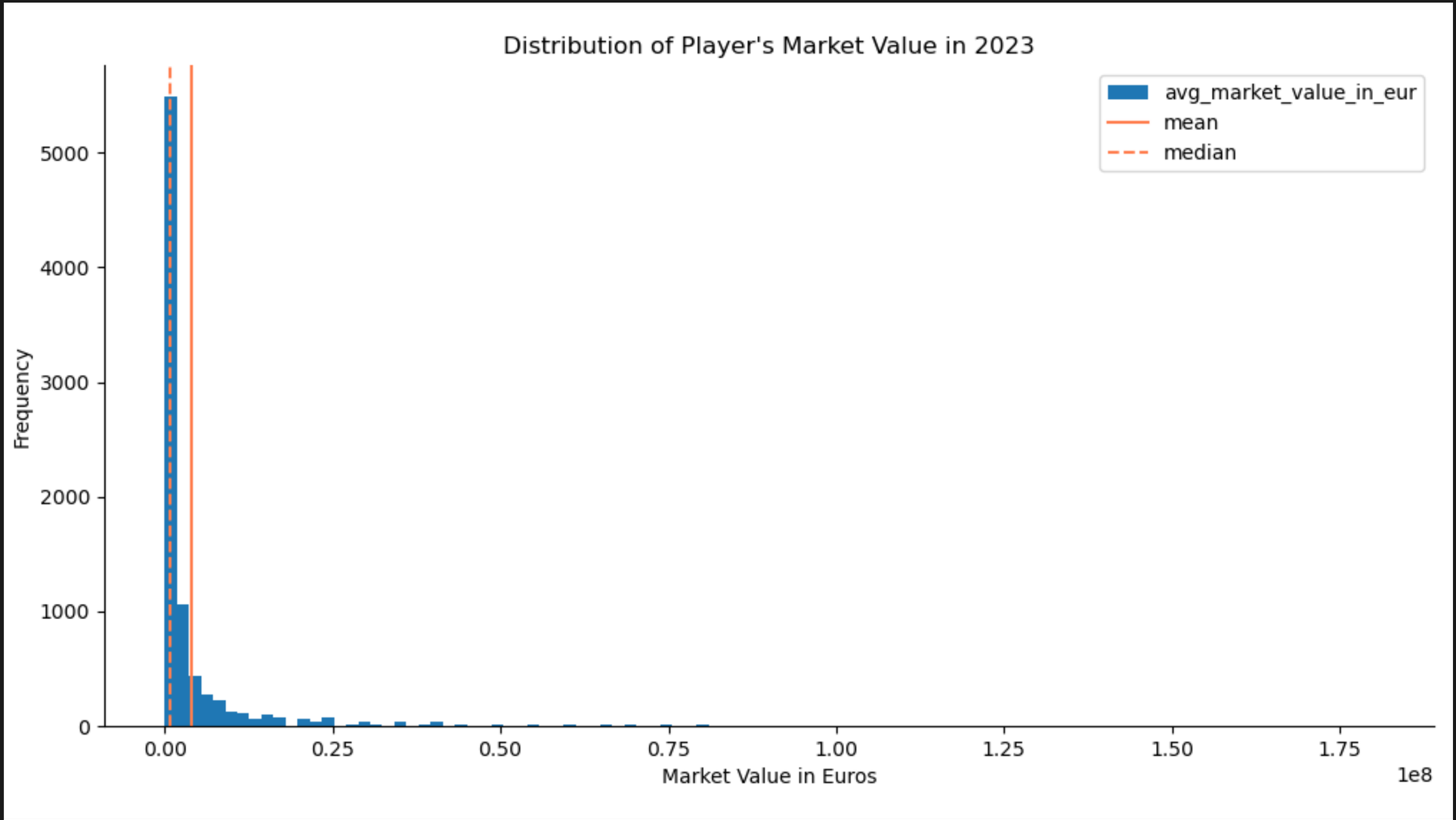




STATISTICS

- Clusters of highly correlated points
- Smaller clusters of negatively correlated groups
- Majority of points have low correlation with each other

VALUATION DISTRIBUTION



- Mean: 4,032,582
- Median: 800,000
- Max: 180,000,000

- Heavy skew to the right

- Biased skew from dataset

NEXT STEPS

Fully clean the data and merge the datasets together. Proceed to remove highly correlated and irrelevant columns.

CLEAN THE DATA

Apply a baseline model. Adapt to results and test different models. Iterate, iterate, iterate.

MODEL THE DATA

Scrape transfer data and use this as test data. Perform hyperparameter tuning to maximize results.

TEST THE DATA

THANK'S FOR WATCHING