Internship Projects

UEFA Intelligence Center

Dogan Parlak

Projects Overview

- Player Workload Analysis - Intelligence Center
- OPTA Data Exploration - Intelligence Center
- Technical Analysis - Football Team

PLAYER WORKLOAD

Content

Feature Introductions

Pairwise Relations

Player Workload Rates Over Seasons

Machine Learning Analysis

Introduction

- Player workload analysis using the TOP 500 players, ranked according to Transfermarkt values based on the year 2022.

- Investigate change of player workload rates over seasons for TOP 500 players (2011-2022).

- Examine relevant features by plotting marginal distributions and highlight significant correlations using pairwise plots. Apply machine learning analysis and assess the performance.

- Datasets used:
  - OPTA (Domestic league & cups)
  - FAME (UEFA competitions)
  - TFM (National team matches)

FEATURES

Fundamental Features

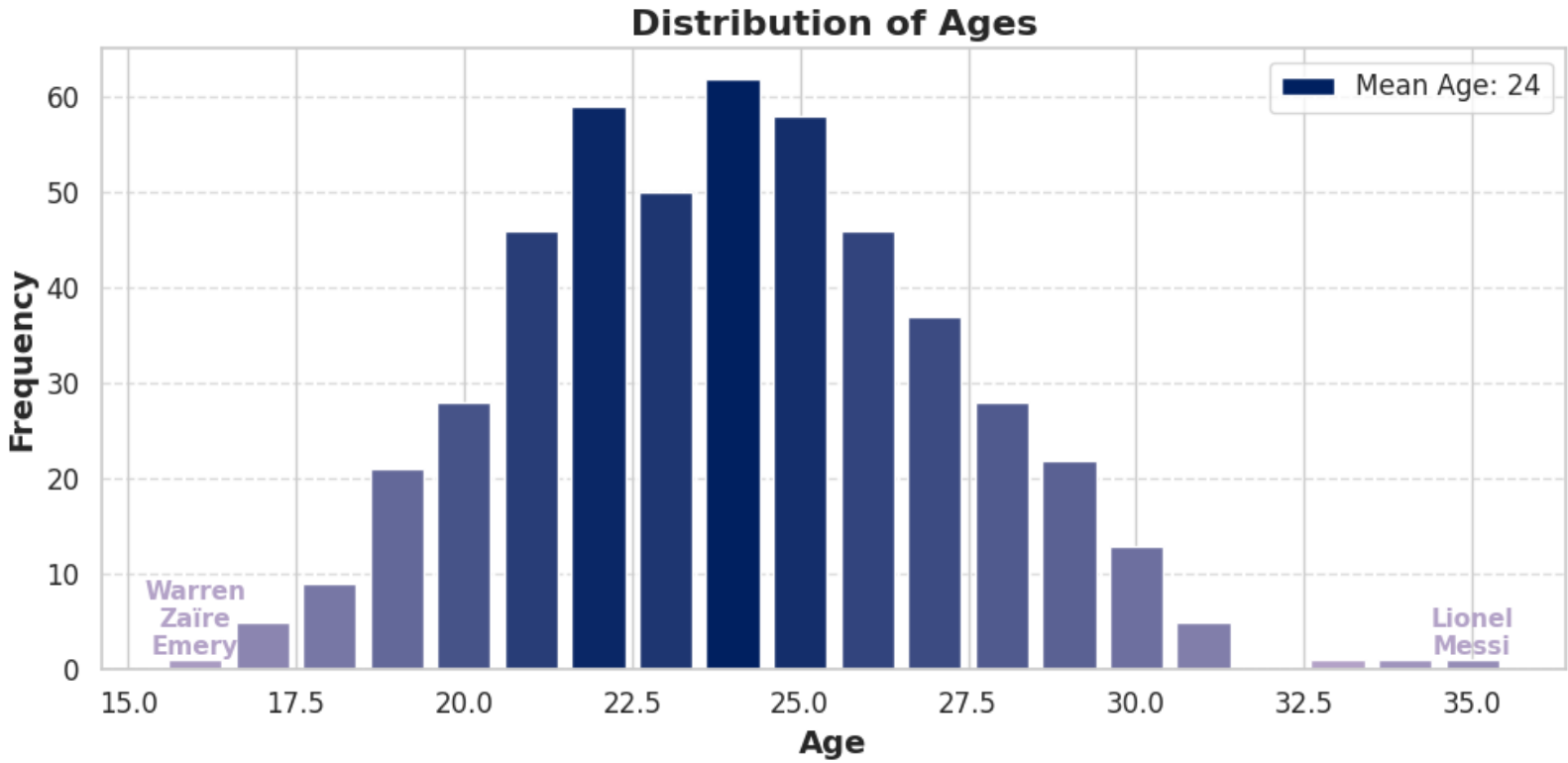
- Age
- Nationality
- Market value
- Player position
- Competition type

Rates

- Appearance
- Minutes played
- Starting XI
- Subs-on
- Subs-off

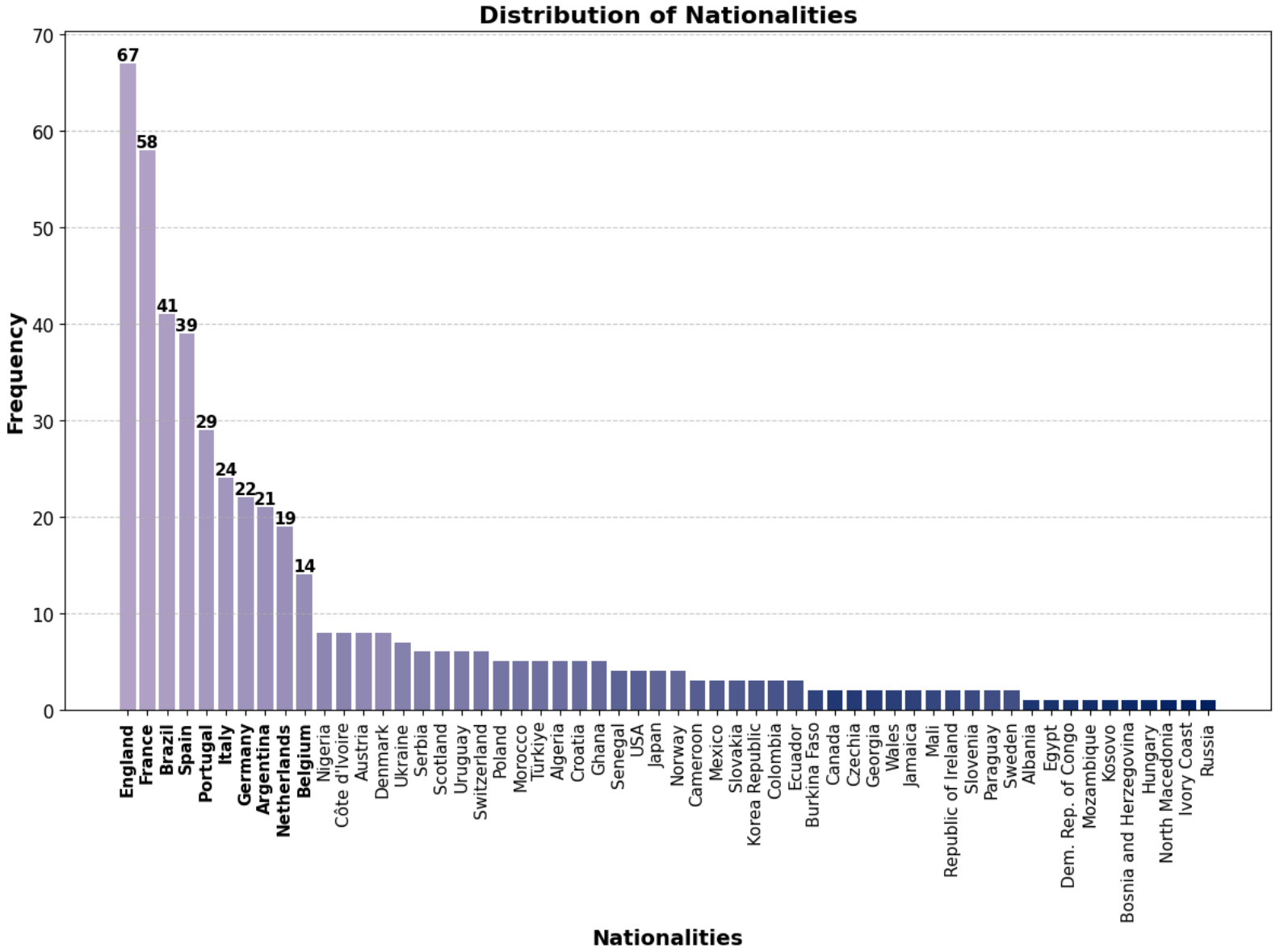
Fundamental Features - Age

- Normally distributed
- Minimum age: 16
- Maximum age: 35



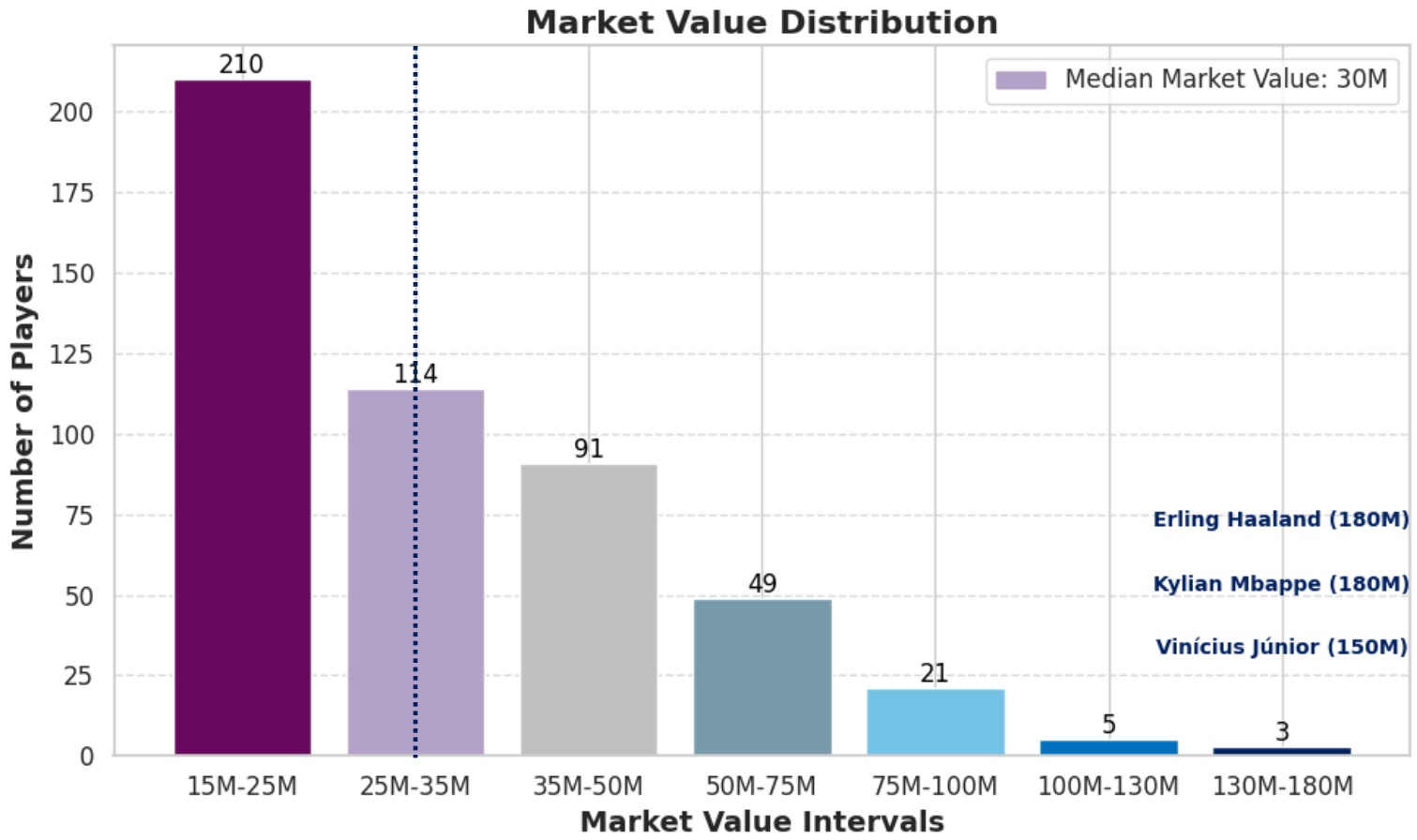
Fundamental Features - Nationality

- 56 distinct nationalities
- Elbow point after Belgium
- Countries hosting the top 5 football leagues were among the 10 most frequent
  - England
  - France
  - Spain
  - Italy
  - Germany
- The only non-European countries in top 10 were Argentina and Brazil



Fundamental Features - Market Value

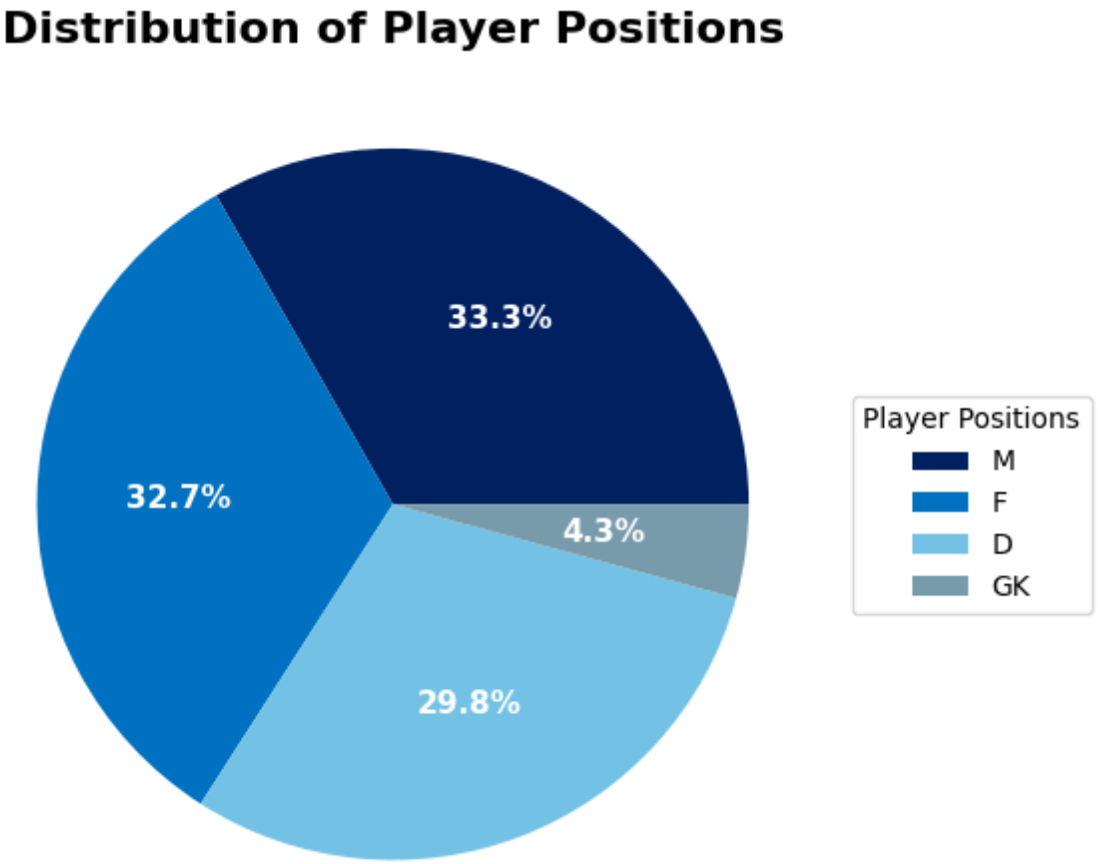
- Grouped market values
- Exponentially decreasing
- Outliers in the high-end



Fundamental Features - Player Position

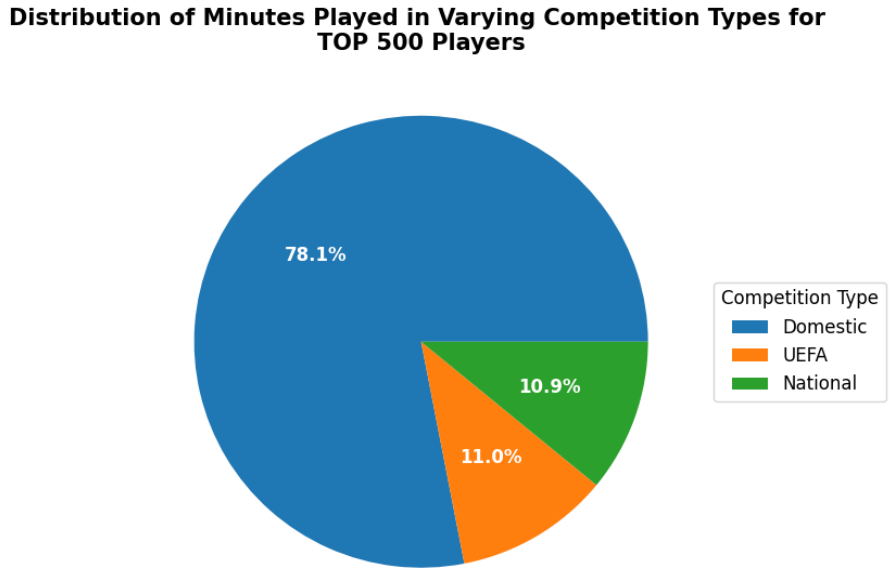
- Basic player positions
  - Midfield
  - Forward
  - Defense
  - Goalkeeper

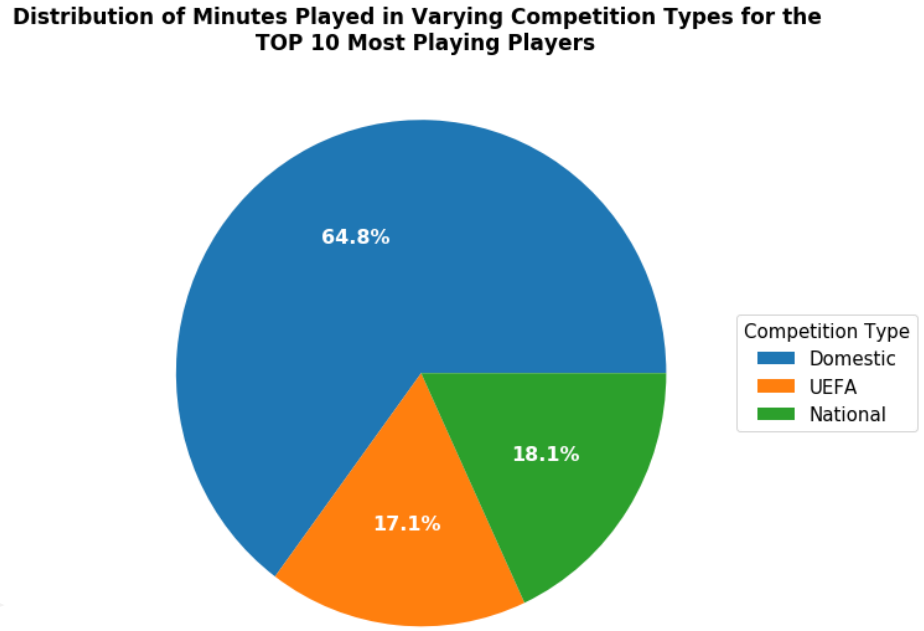
- Goalkeepers under-represented (1/11 = 9,1%, compared to 4,3% in top 500)



Fundamental Features - Competition Type

- Competition types
  - Domestic matches
    - League
    - Cups
  - UEFA competitions
    - Champions League
    - Europe League
    - Conference League
  - National team matches





Player Workload Features Summary in 2022

Feature	Minimum	Maximum	Average
Minutes Played	26	6147 (Bruno Fernandes)	3076
#Appearance	2	72 (Bruno Fernandes)	43
#Starting-XI	0	70 (Bruno Fernandes)	35
#Sub-on	0	37 (Ansu Fati)	8
#Sub-off	0	40 (Federico Dimarco)	13

Rates

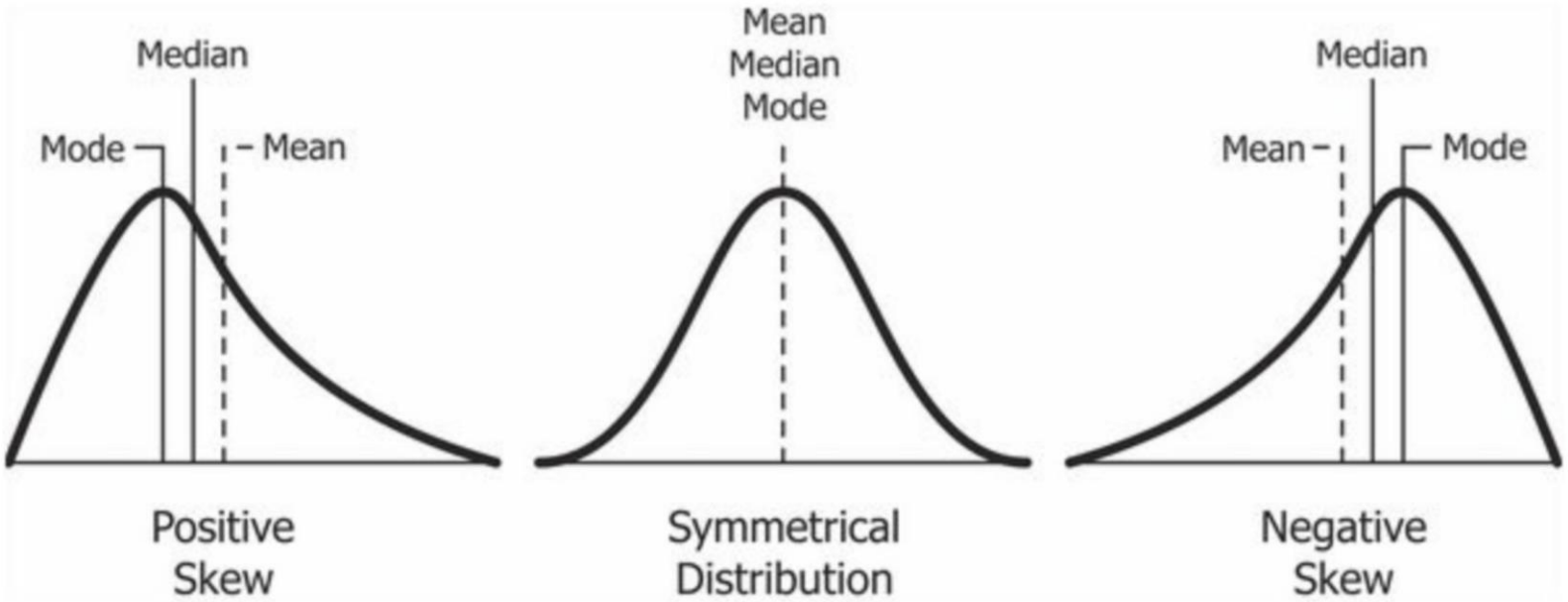
- Within a year: 10 months (on-season) + 2 months (off-season)
- On-season:
  - MaxGames = 100 games (max amount within a year)
  - MaxMins = 100 x 90 = 9000 mins (max minutes that can be played within a year)
- Appearance rate: $\frac{\#Appearances}{MaxGames}$
- Minutes played rate: $\frac{Minutes\ played}{MaxMins}$
- Start-XI rate: $\frac{\#StartXI}{MaxGames}$
- Sub-on rate: $\frac{\#Sub-on}{MaxGames}$
- Sub-on rate: $\frac{\#Sub-off}{MaxGames}$

Players with Highest Minutes Played Rate in 2022

Name	Age	Nationality	Position	Market Value	Minutes Played Rate
Bruno Fernandes	27	Portugal	M	75M	0.72
Gianluigi Donnarumma	23	Italy	GK	45M	0.65
David Hancko	24	Slovakia	D	25M	0.63
Rodri	26	Spain	M	90M	0.62
Stanislav Lobotka	27	Slovakia	M	40M	0.61
Vinicius Junior	21	Brazil	F	150M	0.59
Federico Valverde	23	Uruguay	M	100M	0.59
Bryan Cristante	27	Italy	M	20M	0.58
Pierre-Emile Hojbjerg	26	Denmark	M	45M	0.58
Granit Xhaka	29	Switzerland	M	20M	0.58

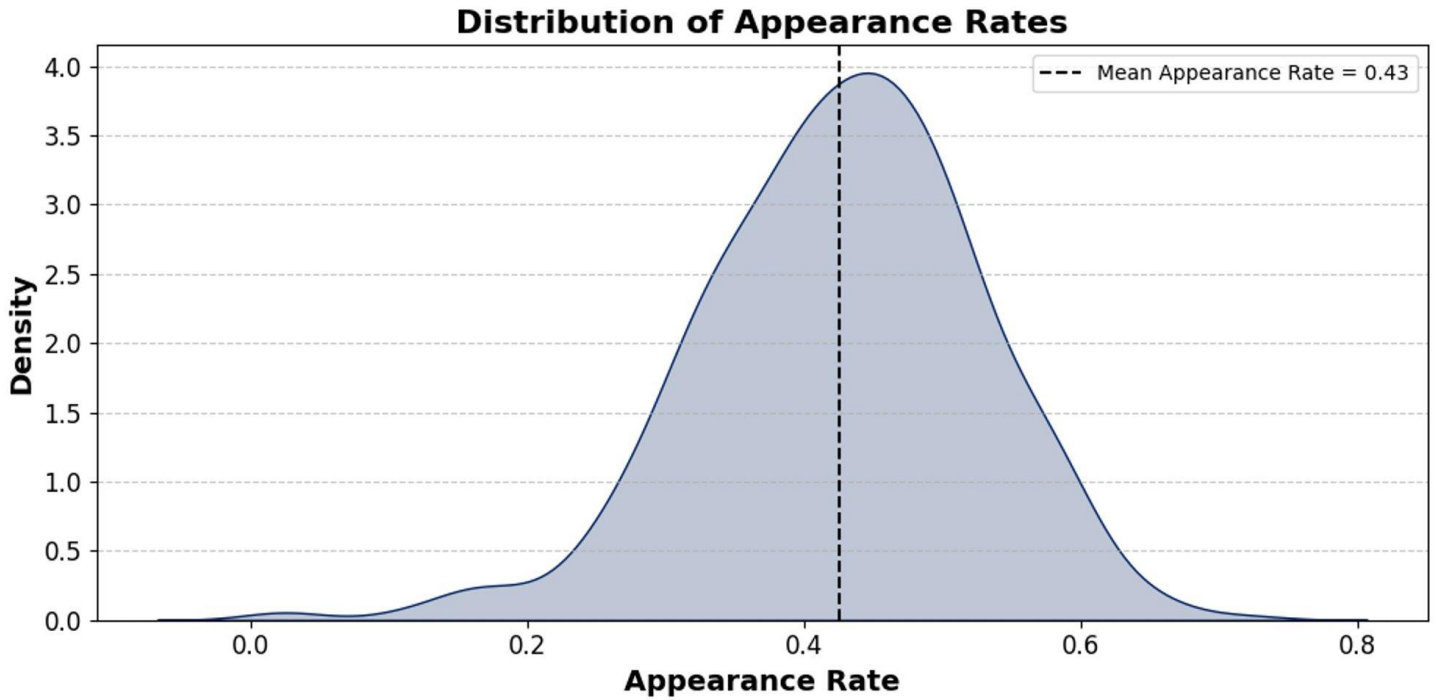
Kernel Density Estimate (KDE) Plots

- Non-parametric way to estimate the probability density function
- X-axis: Values of the variables (i.e., player workload rates)
- Y-axis: Density of the corresponding variables. Higher points indicated higher frequency of the rates.



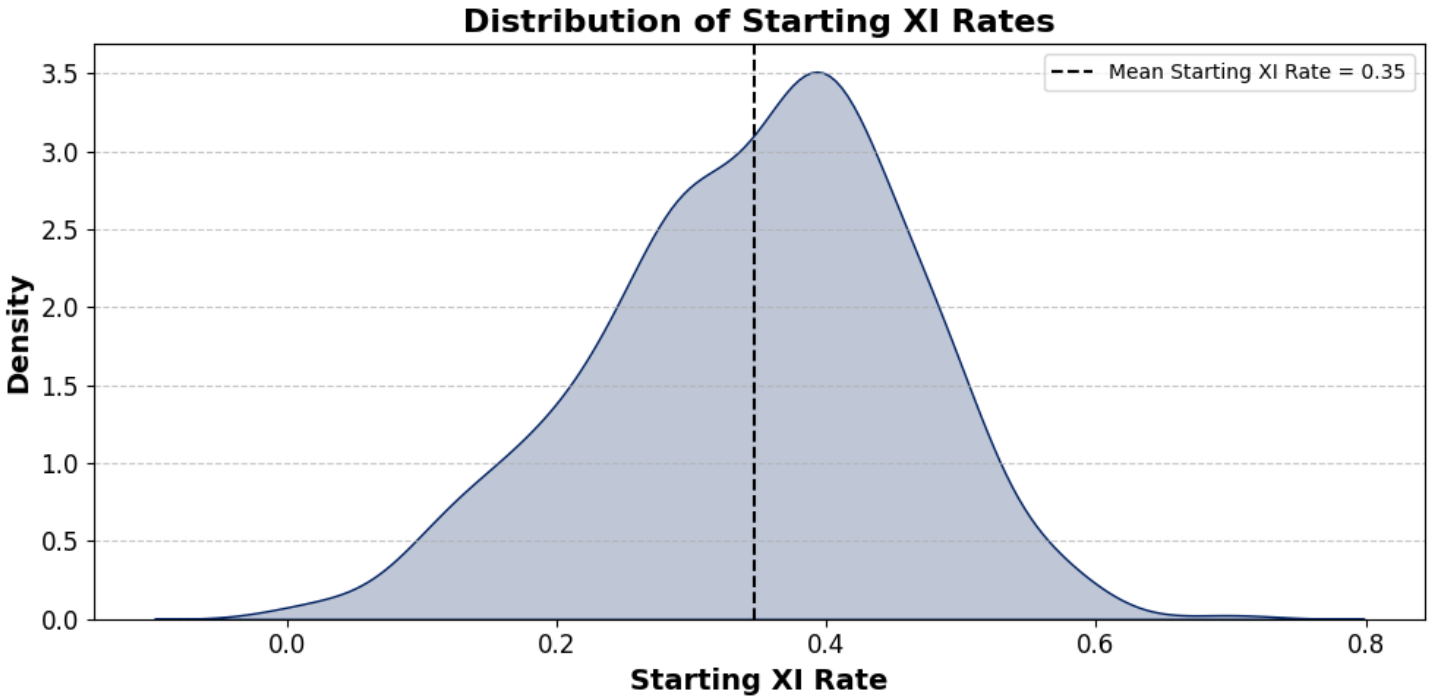
Appearance Rate

- Normally distributed
- Slightly negative skewed
- Peaked top, indicated concentration across a narrower range



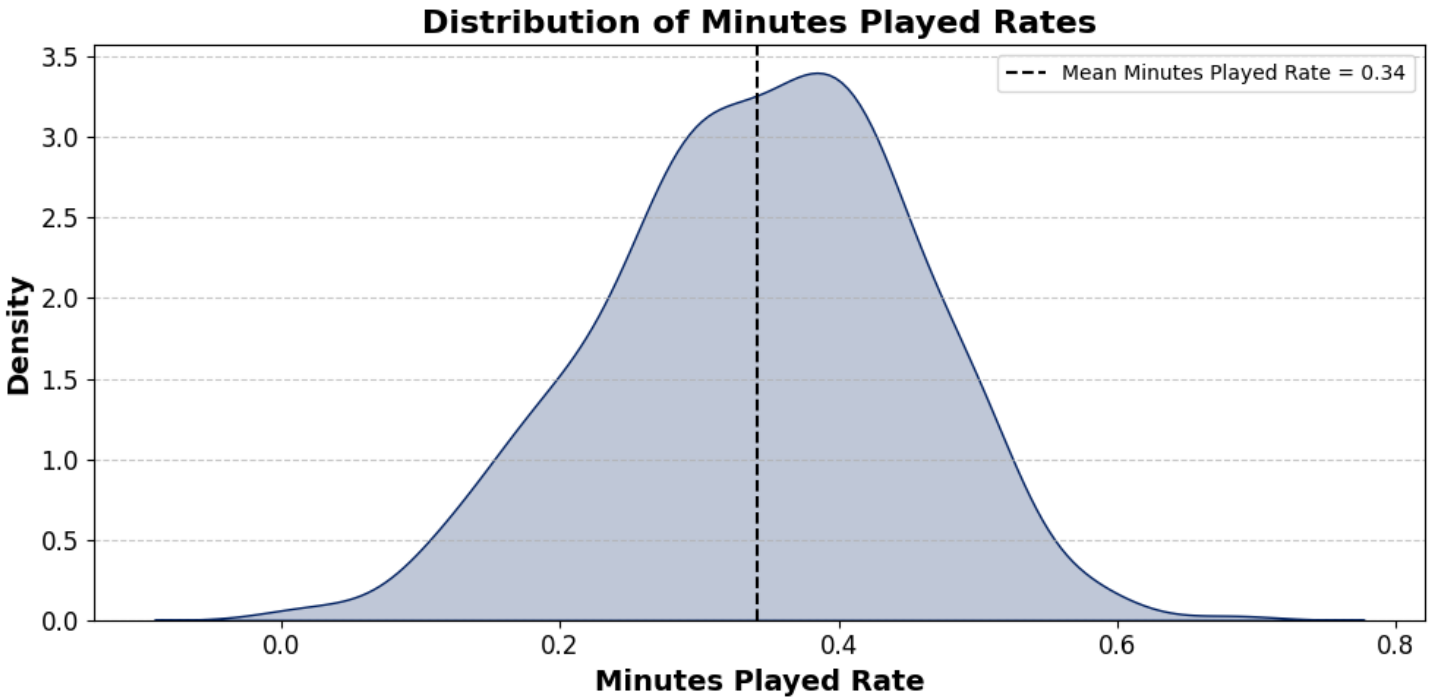
Starting XI Rate

- Normally distributed
- Slightly negative skewed
- Peaked top
- With a mean appearance rate of 0.43, players in the top 500 tended to start in the starting XI



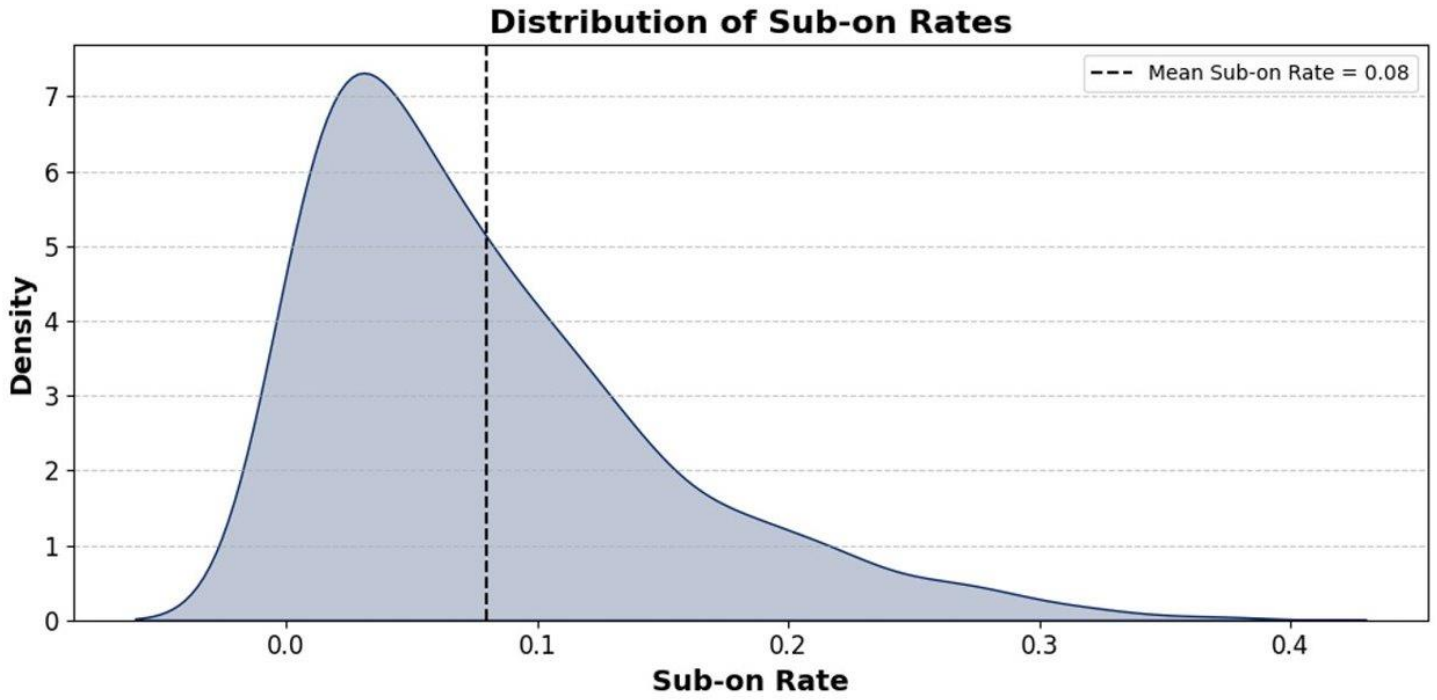
Minutes Played Rate

- Normally distributed
- Symmetric
- Flat top,
  indicated concentration
  across a wider range



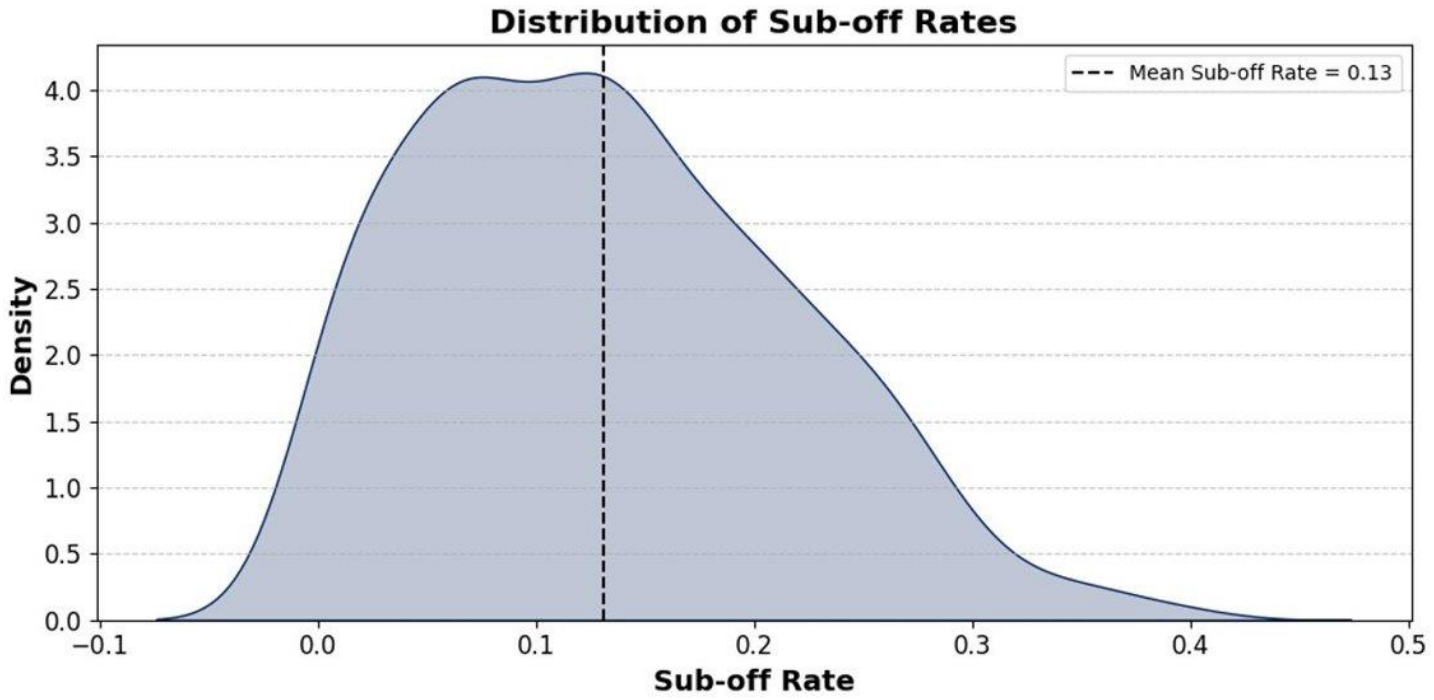
Subs-on Rate

- Positive skewed distribution
- Peaked top
- With a mean appearance rate of 0.43 and start XI rate of 0.35, top 500 players tended to have fewer sub-ons



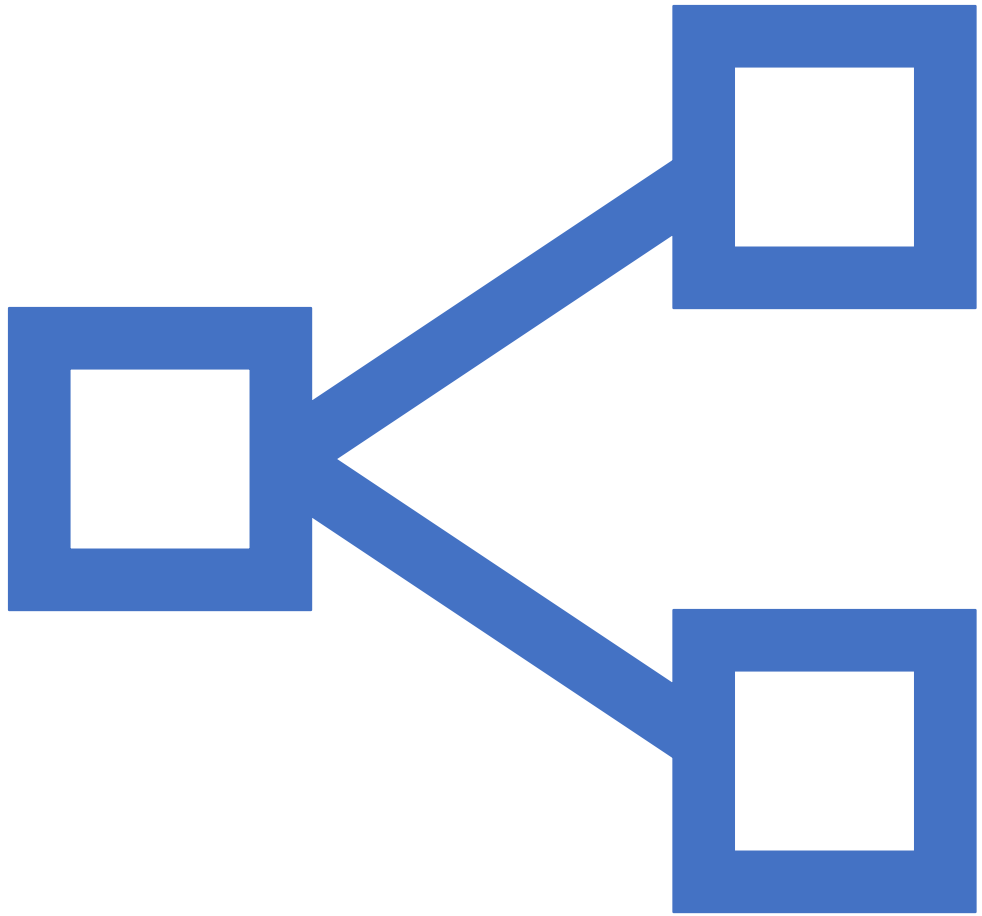
Subs-off Rate

- Positive skewed distribution
- Flat top
- Presence of the outliers in the higher end of the distribution
- Given the magnitude of the mean sub-off rate and outlier locations, top 500 players tended to sub-off less frequently



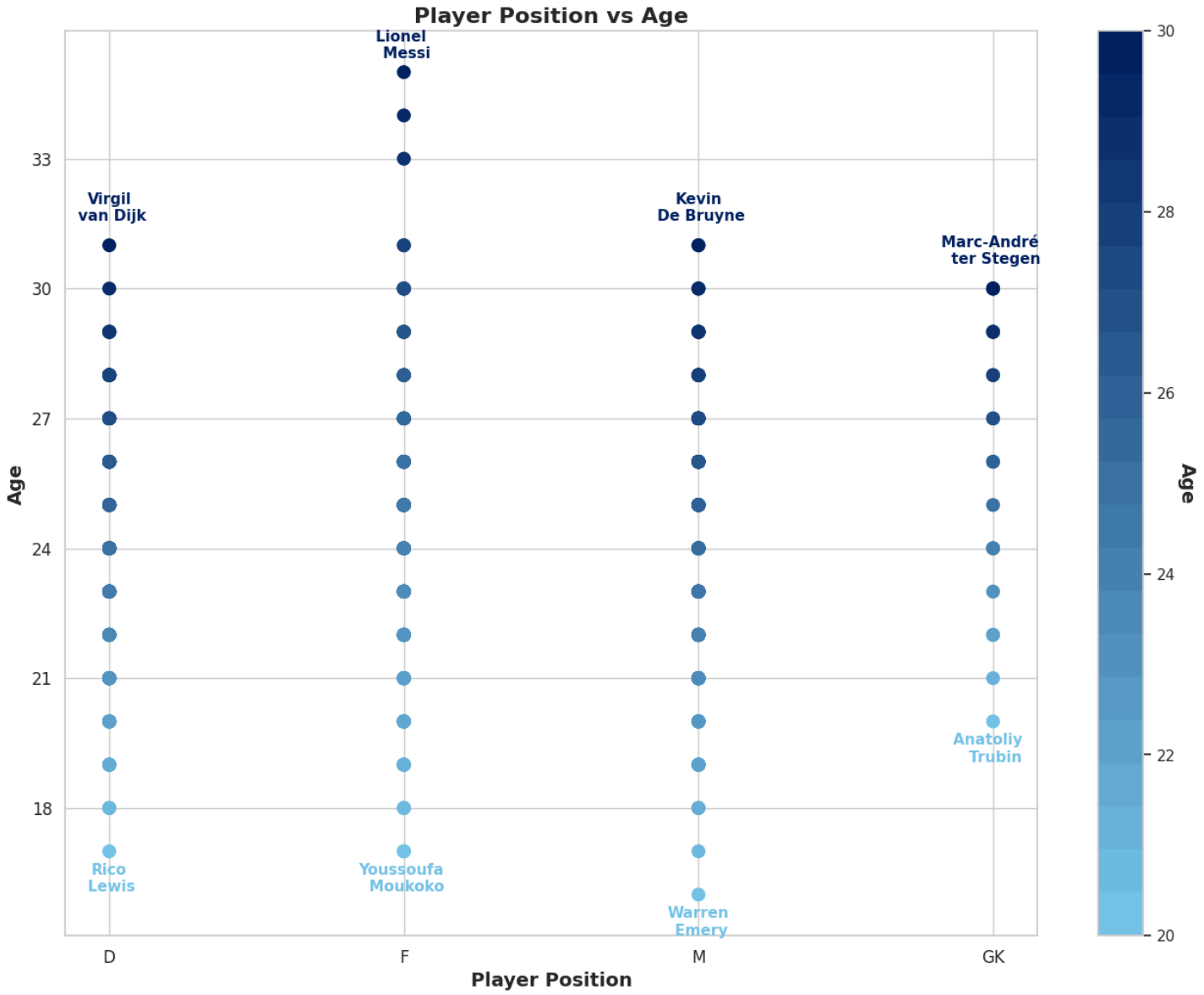
Pairwise Relations

- Relationship between features, especially between fundamental features and rates



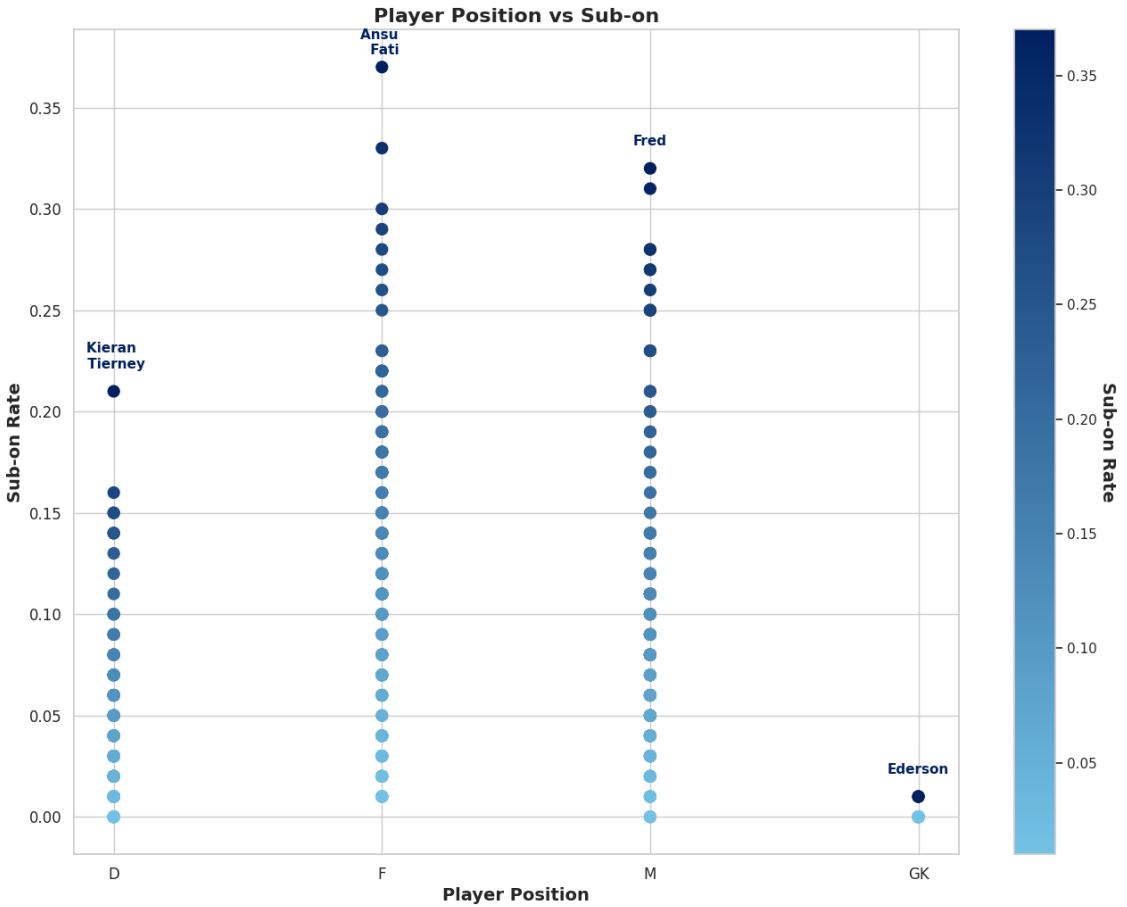
Player Position - Age

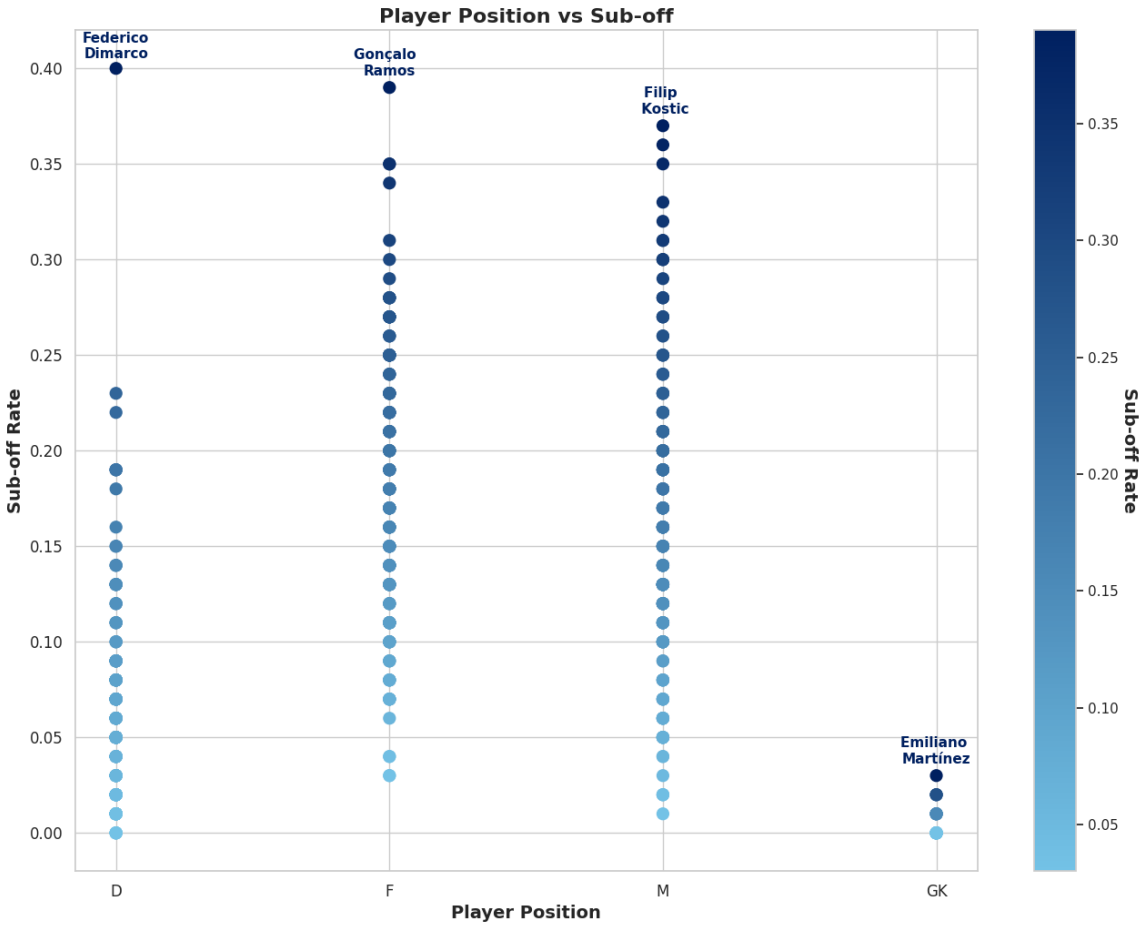
- Largest age range in forward position
- Smallest age range in goalkeeper position
- Similar age ranges for defense and midfield



Position – Substitution Rates

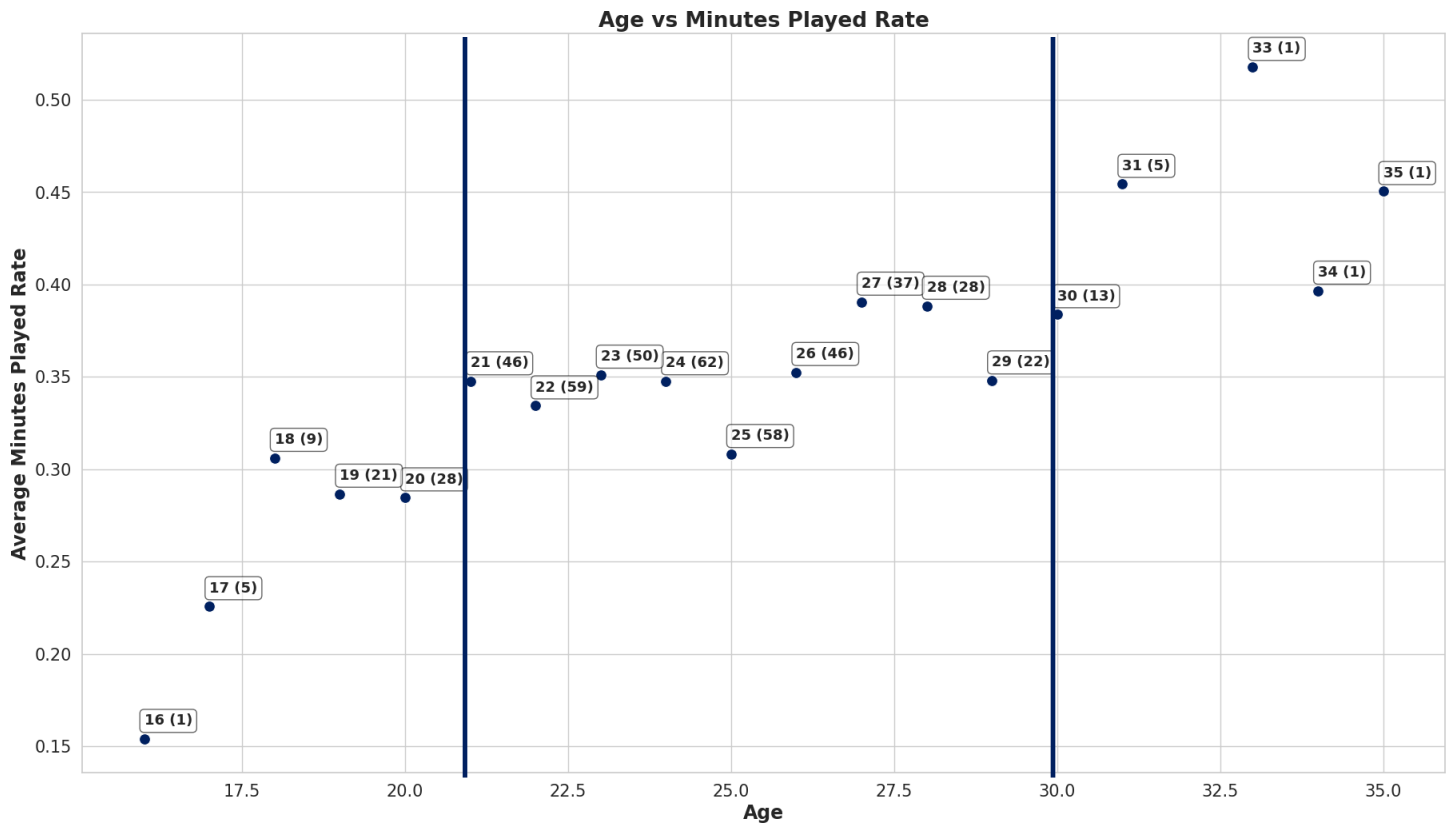
- Similar characteristics for both sub-on and sub-off rates
- Midfield and forward players had the highest rate of substitution





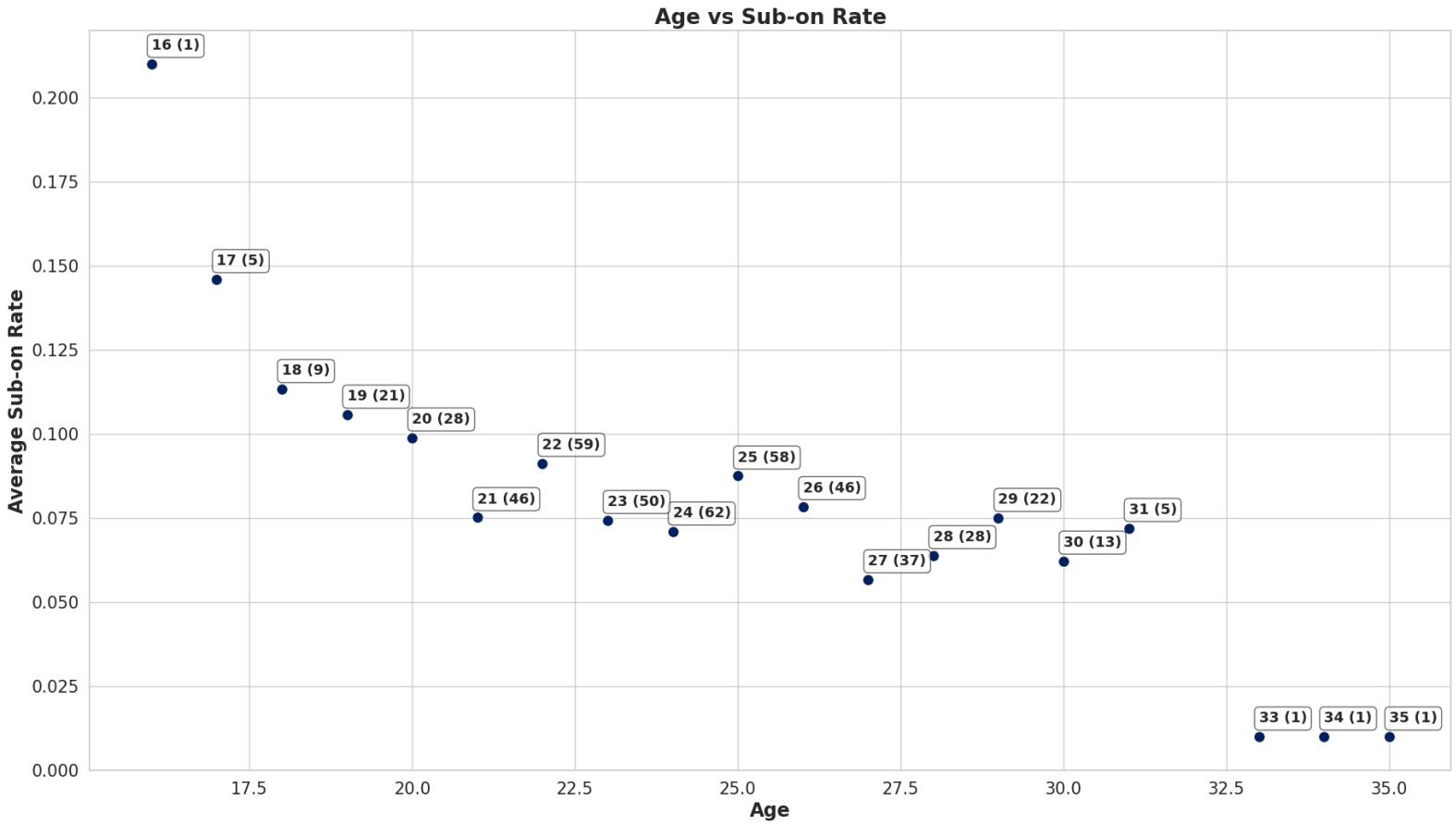
Age - Minutes Played Rate

- Positive correlation between age and minutes played rate
- Trends:
  - Steep from 16-20 and 30-35
  - Gradual from 21-29
- Older top 500 players demonstrated ongoing value to their teams, secured more playing opportunities



Age - Sub-on Rate

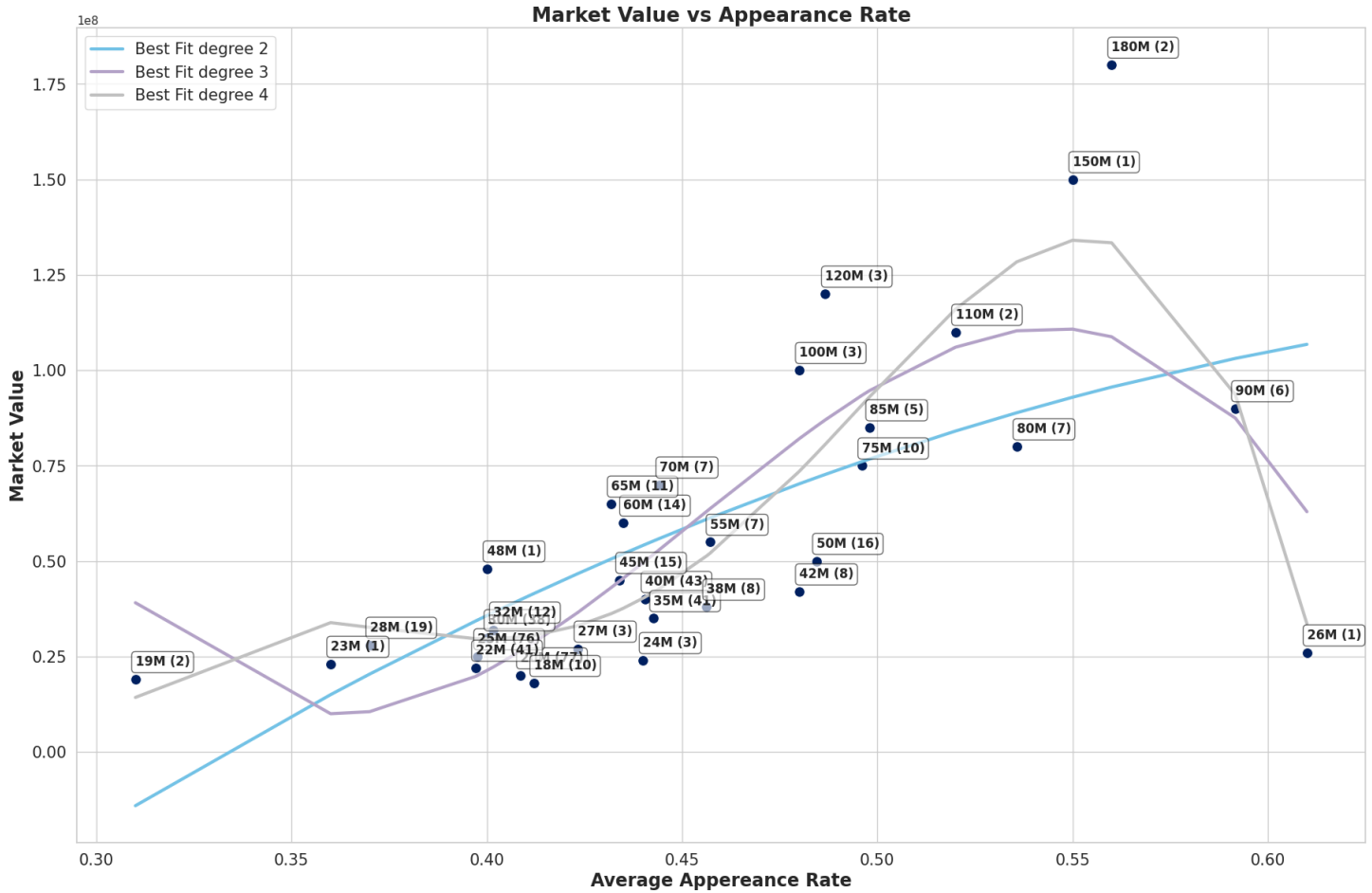
- Negative correlation between age and sub-on rate
  - Exponentially decreasing relationship
- Younger top 500 players had more sub-on opportunities than their older counterparts



Market Value - Appearance Rate

- In general, players with close market values exhibit similar appearance rates

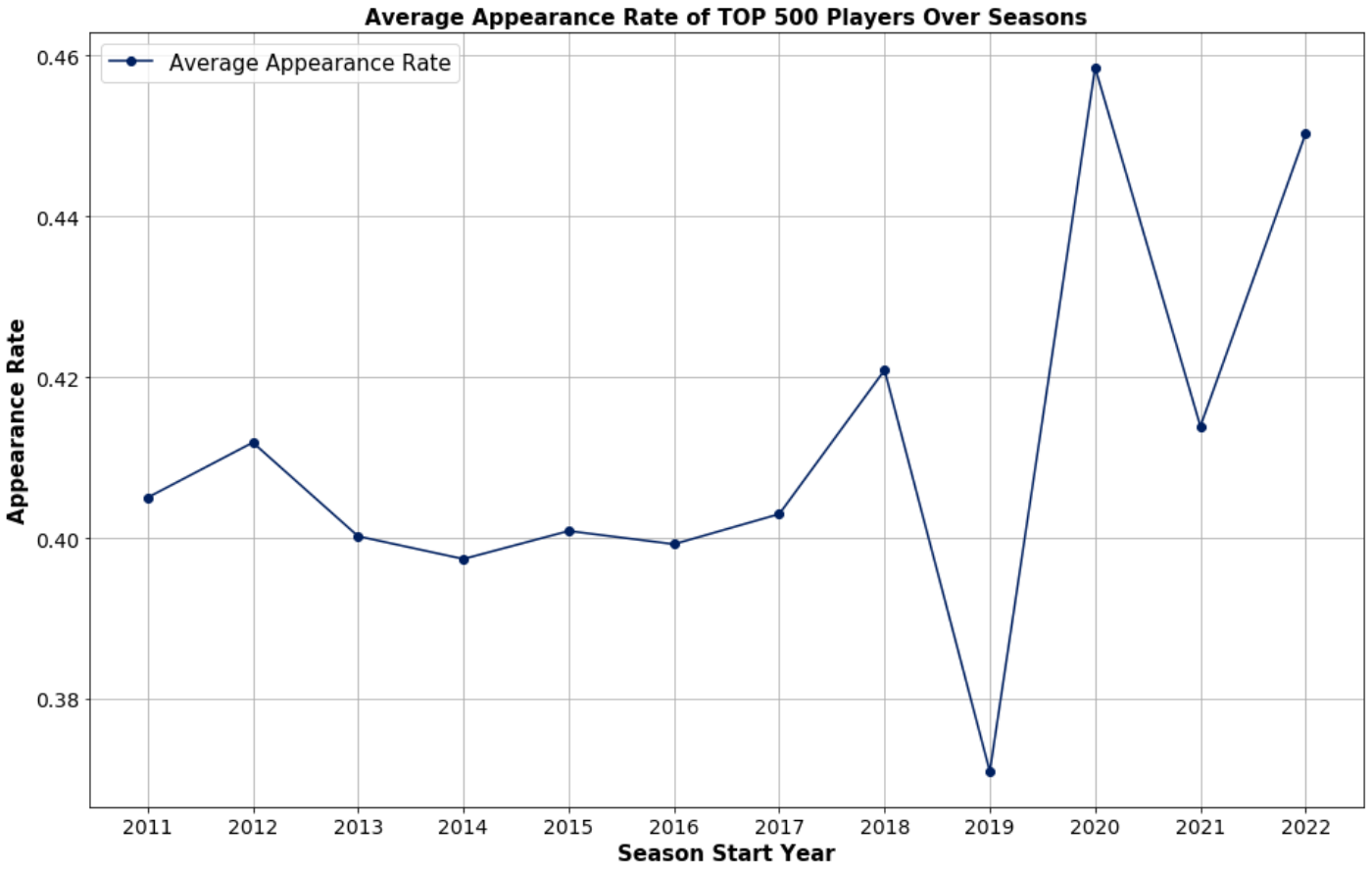
- A large variance revealed a
  - Linear-like relationship for the best-fit polynomial of degree 2
  - Declining S shape for the best-fit polynomials of degree 3 and 4



Players with the Highest Rates Over the Years

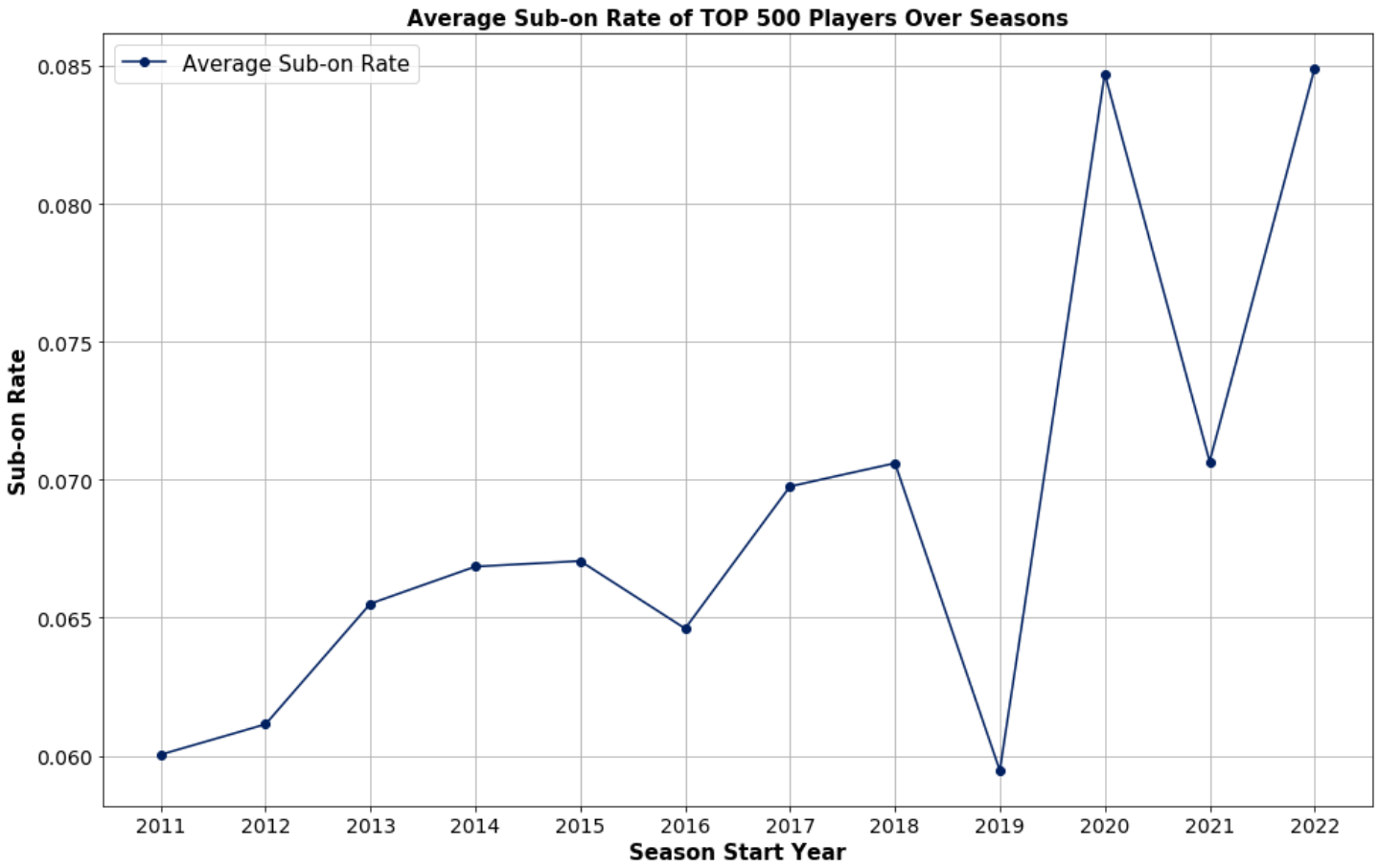
	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Appeareances	Lionel Messi	Oscar	Alexis Sanchez	Carlos Tevez	Javier Mascherano	Bernardo Silva	Paulinho	Daley Bling	Riyad Mahrez	Bruno Fernandes	Sadio Mane	Bruno Fernandes
Minutes Played	Lionel Messi	Petr Cech	Thibaut Courtois	Lionel Messi	Nicolas Otamendi	Everton Ribeiro	Rui Patricio	Ruben Dias	Djene	Gianluigi Donnaruma	Edouard Mendy	Bruno Fernandes
Start XI	Lionel Messi	Petr Cech	Thibaut Courtois	Lionel Messi	Javier Mascherano	Antoine Griezmann	Rui Patricio	Ruben Dias	Djene	Gianluigi Donnaruma	Edouard Mendy	Bruno Fernandes
Subs-on	Aritz Aduriz	Julian Schieber	Andre Schurrle	Pedro	Dries Mertens	Lucas Vazquez	Raul Jimenez	Gabriel Jesus	Christian Eriksen	Trincao	Eduardo Camavinga	Ansu Fati
Subs-off	Jadson	Juan Mata	Ezequiel Lavezzi	Willian	Antoine Griezmann	Yannick Carrasco	Andres Iniesta	Pizzi	Lautaro Martinez	Alexander Isak	Taiwo Awoniyi	Federico Dimarco

Appearance Rate Over Seasons



- Stable-like trend between 2011-2018
- Covid-19 in 2019
- The number of allowed substitutions changed from 3 to 5 after 2019, except for Premier league, which agreed to this change in the upcoming seasons.
- Dramatically increasing trend after the change of substitution rule

Subtitution Rate Over Seasons



- Gradually increasing trend between 2011-2018
- Covid-19 in 2019
- Dramatically increasing trend after the change of substitution rule

Machine Learning

Regression task

- Market Value prediction

Features

- Age
- Nationality
- Market value
- Player position (GK, D, M, F)
- Competition type (Domestic league & Cup, UEFA competition, National matches)
- Player workload features (#Selections, #Appearances, Minutes played, #Starting XI, #Sub-on, #Sub-off)

Models

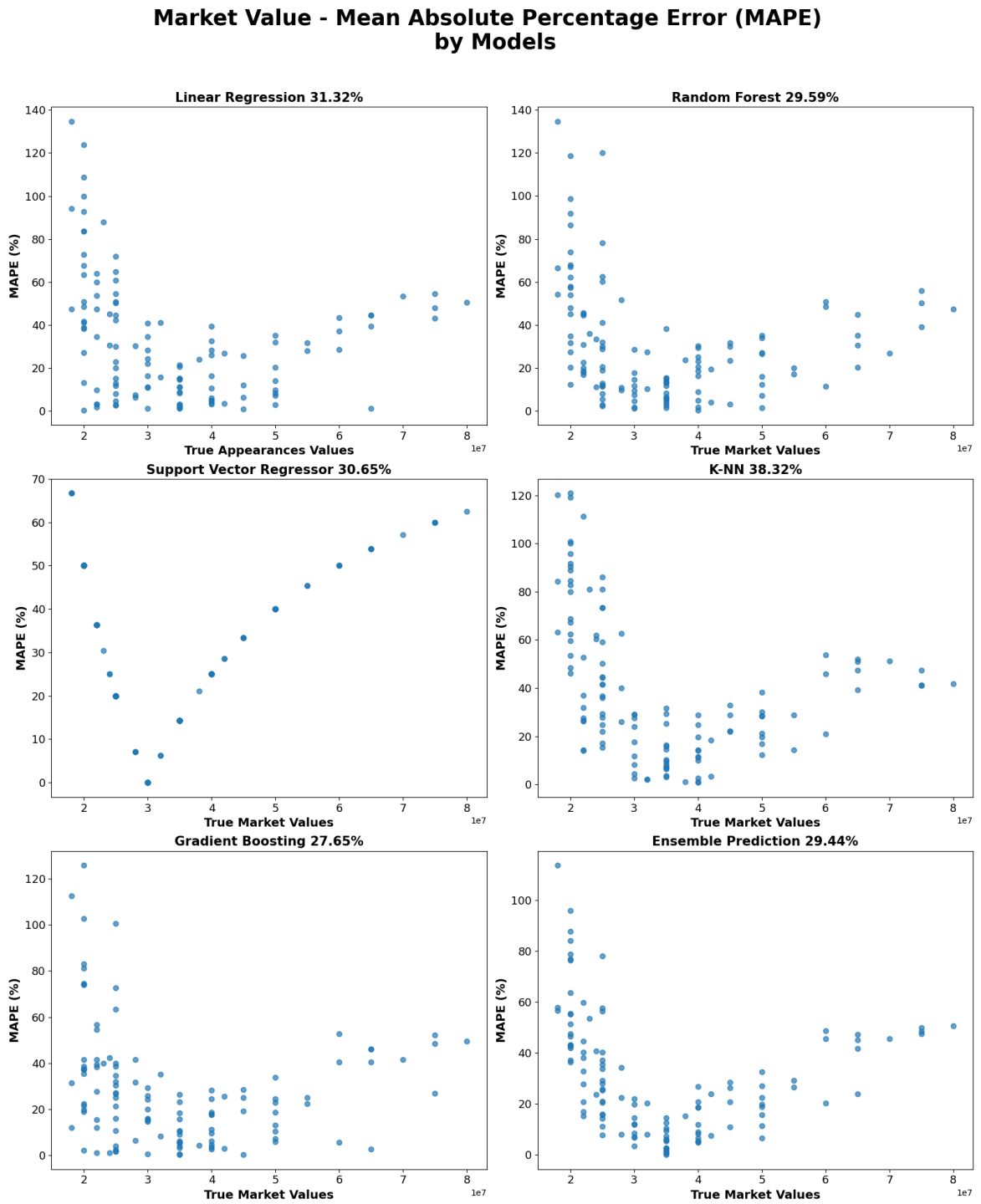
- Linear Regressor
- Support Vector Machine
- Random Forest
- Gradient Boosting
- K-NN
- Ensemble (Combines successful model's predictions)

Evaluation Methodology

- Mean Average Percent Error (MAPE):
  - Average percentage difference between predicted and actual values
  - $MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|$ x 100
    - Y_t: Actual value
    - $\hat{Y}_t$: Predicted value
    - N: Number of observations

Market Value Prediction

- All models, except for k-NN, exhibited similar performance.
- The plots display that each model performed better in predicting market values around 30M, as most top 500 players fall within that range.
- Ensemble prediction yields a stable output.
- Balancing and expanding the dataset for league-level analysis can assist identifying overrated and underrated leagues.



Conclusion

- Fundamental features and player workload rates were introduced
  - Countries hosting the top 5 football leagues were among the 10 most frequent
  - Player positions are equally distributed, except for goalkeepers
  - Exponentially decreasing market value distribution
  - Normally distributed: age, appearance rate, starting XI rate, minutes played rate
  - Positive skewed: sub-on and sub-off rates

- Marginal distributions and pairwise relationships were examined
  - Highest rate of substitution for midfield and forward players
  - Positive correlation between age and minutes played rate
  - Negative correlation between age and sub-on rate
  - In general, players with close market values exhibit similar appearance rates

- Despite the limited dataset, market value prediction revealed strong correlation with player workload features

OPTA DATA EXPLORATION

Content

- Introduction and Filtering of Datasets
- Coverage Analysis
- Stage Analysis
- OPTA vs TFM Data Comparison

Datasets

- OPTA Data
  - Team Stats
  - Player Stats
  - Domestic League Matches
  - Domestic Cup Matches

- TFM Data
  - Player Stats in Domestic League Matches

Dataset Filtering

- The aim was to keep only relevant columns for analysis and filter out non-informative ones.
  - Reduce dimensionality to enhance computation time.
  - Prevent memory issues and facilitate the execution of recipes.

- General characteristics of columns to keep:
  - Team stats:
    - Goals, assists, cards, substutitions, possession control, passes, formations etc.
  - Match stats:
    - Length, tier, stage, country, week, calendar type, season etc.
  - Covarage level:
    - Indicator of data availability level.
  - IDs:
    - Team, competition, match, country etc.

Team-Level Datasets

- Team – Level
  - Team Stats (OPTA), Domestic League Matches (OPTA), Domestic Cup Matches (OPTA)
  - Team Stats = Domestic League Matches + Domestic Cup Matches
  - For each match played, there are two entries, each representing the stats of a team.

- Includes the number of matches played for each:
  - Country
  - Season
    - 22/23 and 23/24
  - Gender
    - Men or Women
  - Tier and Competition (Dataset dependent)
    - Domestic League Matches: First Tier & Second Tier
    - Domestuc Cup

Player-Level Datasets

- Player – Level
  - Player Stats (OPTA) , Player Stats (TFM)
  - Player Stats (OPTA)
    - Domestic League Matches + Domestic Cup Matches
    - Each player and their corresponding matches have an entry in the dataset.
  - Player Stats (TFM)
    - Domestic League Matches
    - One entry per player, including aggregated stats over the matches.

- Similar to the Team Stats dataset in terms of both data content and structure. The stats now focus more on individual player performance rather than team performance.
  - Player workload stats: Mintes Played, Substutitions etc.
  - Performance stats: Goals, Assists, Cards etc.

Coverage Analysis

- Coverage: The level indicates data availability within range [0,15]. As the level increases, more detailed data is provided.
- Utilized the Team Stats dataset to analyze the coverage level of competitions.
- Grouped matches based on season start year, gender, country, tier, and competition name.
- Calculated the percentage (%) of coverage levels for each combination.

	SeasonStartYear	Gender	Country	Tier	CompetitionName	MatchCount	1.0	2.0	3.0	4.0	...	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0	14.0	15.0
0	2022	men	England	First Tier	Premier League	380	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.0	100.00
1	2022	men	England	Second Tier	Championship	557	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.0	100.00
2	2022	men	England	Domestic Cup	FA Cup	144	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	53.47	0.69	34.03	0.0	11.81

Stage Analysis

- Utilized OPTA Datasets:
  - Team Stats
  - Player Stats
  - Domestic League Matches
  - Domestic Cup Matches

- Stage:
  - Specific phase of a competition
  - Regular season, Play-offs, Last 16, Quarter-Finals, Semi-Finals, Final etc.

- Each dataset was used to gather information on season, gender, country, tier, and competition combinations.

- For each combination, the number of matches played and teams involved were computed, stages were assessed.

- As the Player-Stats dataset was exclusively at the player-level, appropriate aggregation was performed to transform it into team-level data.

Stage Analysis - Team Stats

- Stage analysis conducted using the OPTA Team Stats dataset.

SeasonStartYear	Gender	Country	Tier	CompetitionName	MatchCount	ParticipantCount	Regular Season	Relegation Round	1st Round	...	2nd Round Replays
2022	men	Italy	First Tier	Serie A	381	20	380	0	0	...	0
2022	men	Italy	Second Tier	Serie B	390	20	380	0	0	...	0
2022	men	Italy	Domestic Cup	Coppa Italia	45	44	0	0	16	...	0
2022	women	Italy	First Tier	Serie A Women	130	10	90	20	0	...	0
2023	men	Italy	First Tier	Serie A	30	20	30	0	0	...	0
2023	men	Italy	Second Tier	Serie B	34	20	34	0	0	...	0
2023	men	Italy	Domestic Cup	Coppa Italia	20	36	0	0	16	...	0

Stage Analysis - Player Stats

- Stage analysis conducted using the OPTA Player Stats dataset.

SeasonStartYear	Gender	Country	Tier	CompetitionName	MatchCount	ParticipantCount	Regular Season	1st Round	Relegation Round	...	Promotion Play-offs - 1st Round
2022	men	Spain	First Tier	Primera División	380	20	380	0	0	...	0
2022	men	Spain	Second Tier	Segunda División	468	22	462	0	0	...	0
2022	men	Spain	Domestic Cup	Copa del Rey	116	115	0	55	0	...	0
2022	women	Spain	First Tier	Primera División Femenina	240	16	240	0	0	...	0
2022	women	Spain	Domestic Cup	Supercopa Femenina	3	4	0	0	0	...	0
2023	men	Spain	First Tier	Primera División	39	20	39	0	0	...	0
2023	men	Spain	Second Tier	Segunda División	44	22	44	0	0	...	0
2023	women	Spain	Domestic Cup	Supercopa Femenina	3	4	0	0	0	...	0

Stage Analysis - Domestic League Matches

- Stage analysis conducted using the OPTA Domestic League Matches dataset.

SeasonStartYear	Gender	Country	Tier	CompetitionName	MatchCount	ParticipantCount	Regular Season
bigint	string	string	string	string	bigint	bigint	bigint
Integer	Text	Country	Text	Text	Integer	Integer	Integer
2023	men	Germany	Second Tier	2. Bundesliga	126	18	126
2023	women	Germany	First Tier	Frauen Bundesliga	48	12	48
2022	men	Germany	Second Tier	2. Bundesliga	306	18	306
2022	women	Germany	First Tier	Frauen Bundesliga	132	12	132
2022	men	Germany	First Tier	Bundesliga	306	18	306
2023	women	Germany	Second Tier	2. Bundesliga Women	77	14	77
2022	women	Germany	Second Tier	2. Bundesliga Women	182	14	182
2023	men	Germany	First Tier	Bundesliga	108	18	108

Stage Analysis - Domestic Cup Matches

- Stage analysis conducted using the OPTA Domestic Cup Matches dataset.

SeasonStartYear	Gender	Country	Tier	CompetitionName	MatchCount	ParticipantCount	8th Finals	16th Finals	Quarter-finals	7th Round	Semi-finals	32nd Finals	Final	8th Round
bigint	string	string	string	string	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint
Integer	Text	Country	Text	Text	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer	Integer
2022	men	France	Domestic Cup	Coupe de France	194	196	8	16	4	88	2	32	1	43
2023	men	France	Domestic Cup	Coupe de France	87	174	0	0	0	87	0	0	0	0

Stage Analysis Takeovers - Consistents

- Considering the season starting in 2022, data completeness was examined for the first tier of men's domestic league matches, based on league structures.

- Participant counts are complete for all countries.

Complete for all stages and participants
Croatia
England
France
Poland
Portugal
Republic of Ireland
Russia
Spain
Switzerland
Turkiye
Ukraine

Stage Analysis Takeovers - Inconsistents

	Austria	Belgium	Czech Republic	Denmark	Greece	Hungary	Israel	Italy
Regular Season						LS - 2	LS - 3	
Other Stages	SD - 63	SD - 24	SD - 36	SD - 61	SD - 58		SD - 57	SD - 1

	Netherlands	Norway	Romania	Scotland	Serbia	Slovakia	Greece	Sweden
Regular Season		LS - 29		*classified 1st Phase 2nd Phase				LS - 21
Other Stages	SD - 6		SD - 77	SD - 30	SD - 56	SD - 61	SD - 63	

OPTA vs TFM Comparison

- Utilized OPTA and TFM Player Stats for information consistency in first-tier men's competitions starting in 2022, focusing on leagues with coverage levels > 8 and no missing regular season matches.
- Compared features:
  - Goals
  - Assists
  - Yellow Cards
  - Red Cards
  - Substitutions On/Off
  - Minutes Played

- Comparison Methodology: For each feature, generated a new column by applying the following calculation.
  - Stat Difference = #OPTA Stats - #TFM Stats
    - Stat Difference = 0, then consistent
    - Stat Difference > 0, then OPTA stats are greater
    - Stat Difference < 0, then TFM stats are greater

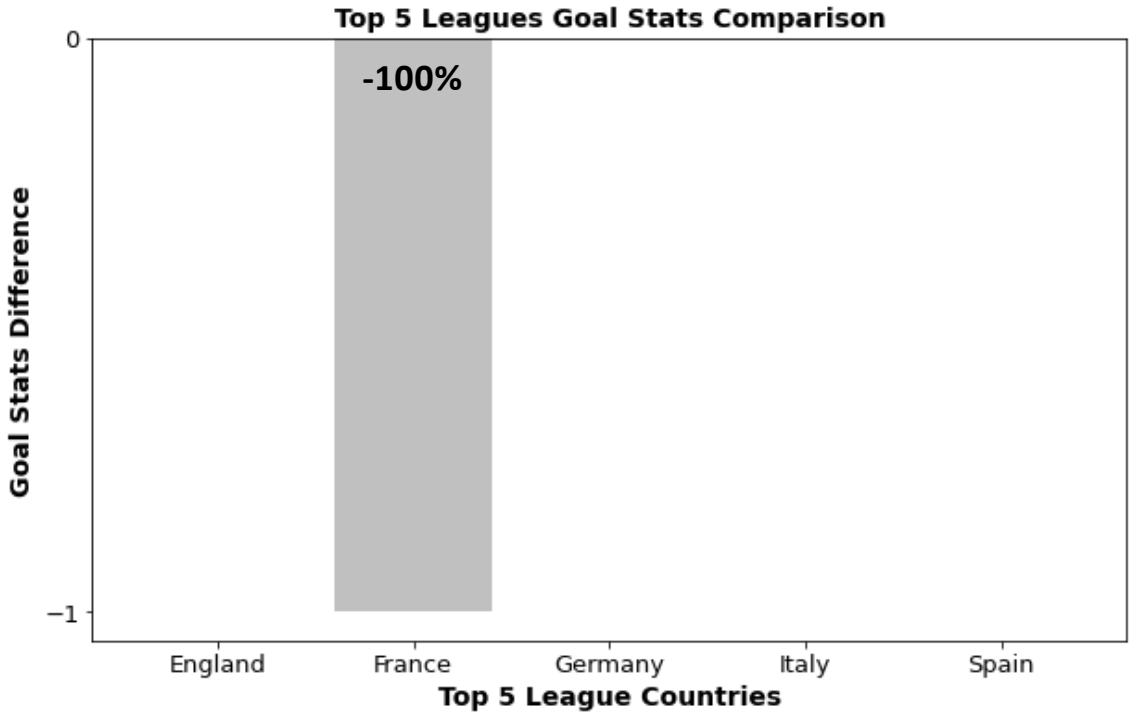
OPTA vs TFM Comparison - TOP 5

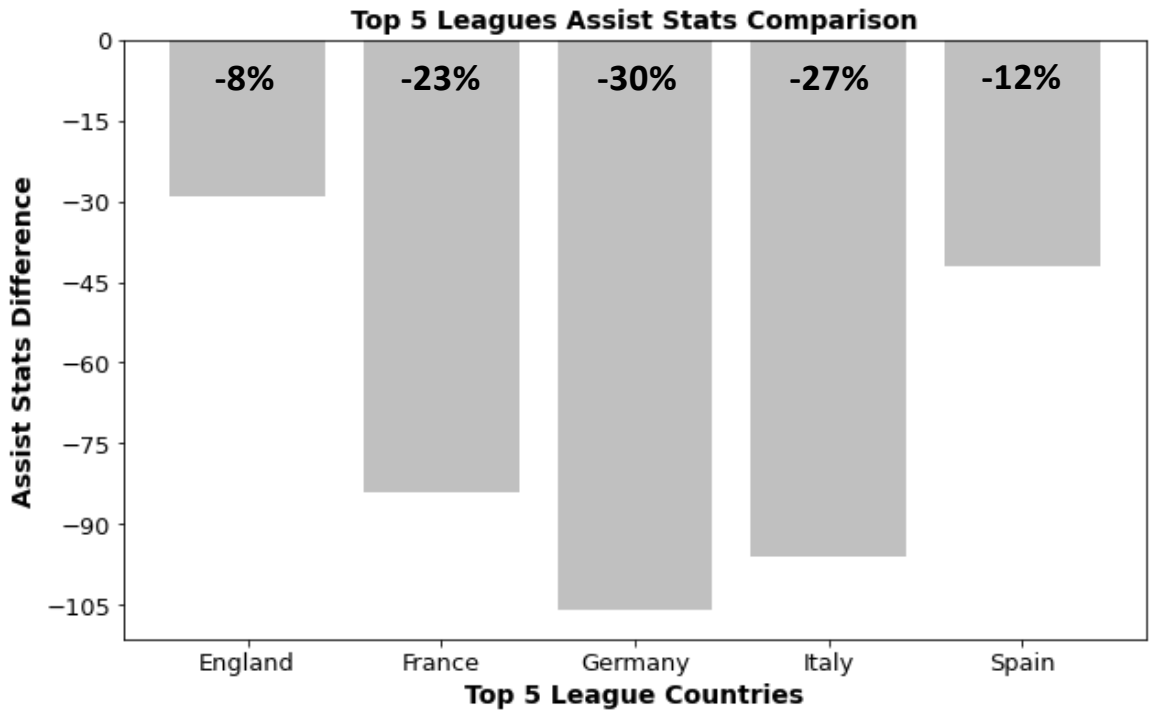
- OPTA and TFM Player Stats aggregated in league-level.
- Season: 2022
- Type: Men
- Tier: First Tier (Domestic League Matches)

	country	goals_diff	assists_diff	subs_off_diff	subs_on_diff	minutes_played_diff	yellow_cards_fin_diff	red_cards_fin_diff
0	England	0.0	-29.0	0.0	0.0	193.0	10.0	-1.0
1	France	-1.0	-84.0	0.0	-1.0	101.0	6.0	0.0
2	Germany	0.0	-106.0	-2.0	0.0	233.0	1.0	0.0
3	Italy	0.0	-96.0	-2.0	0.0	272.0	5.0	0.0
4	Spain	0.0	-42.0	-2.0	0.0	314.0	3.0	-1.0

OPTA vs TFM Comparison - TOP 5 Goal Contribution

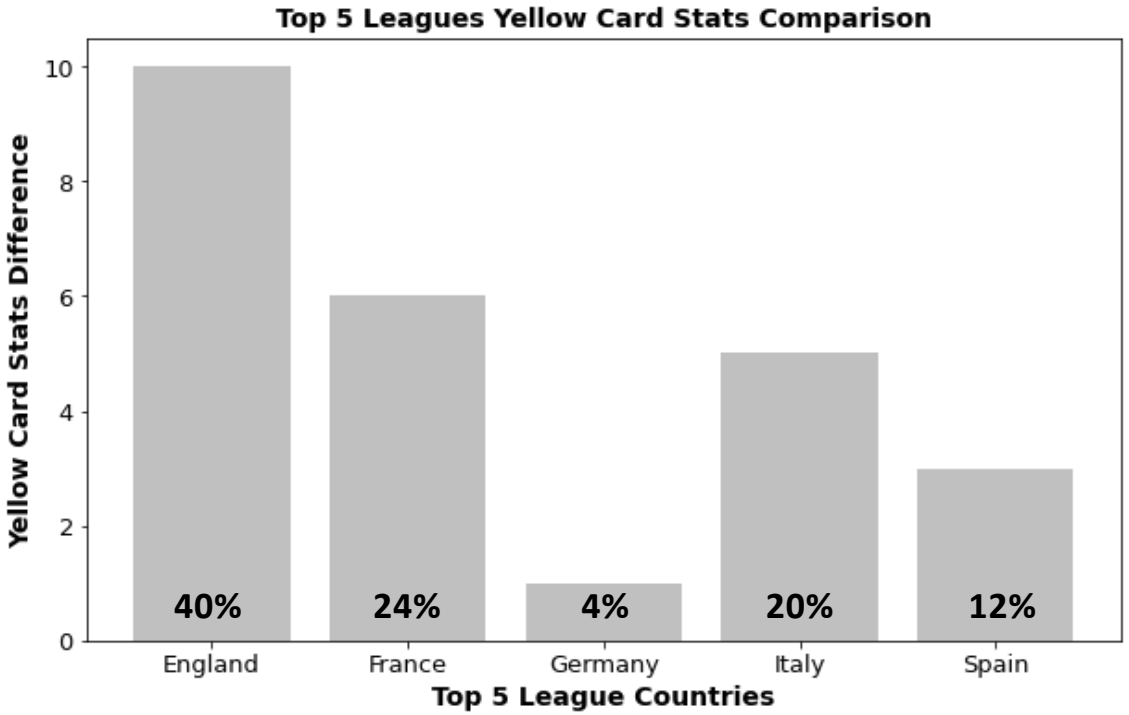
- Goal Contribution = Goals + Assists
- Goal stats are consistent in general.
- TFM stats include more Assists than OPTA stats for all Top 5 league.

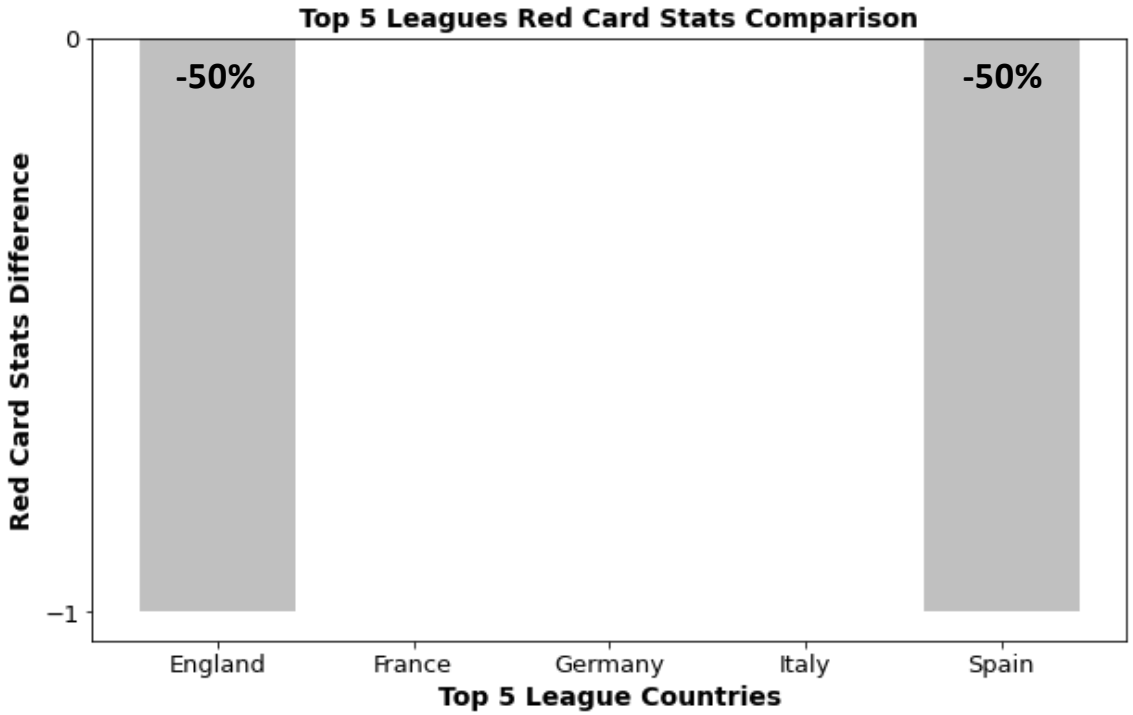




OPTA vs TFM Comparison - TOP 5 Cards

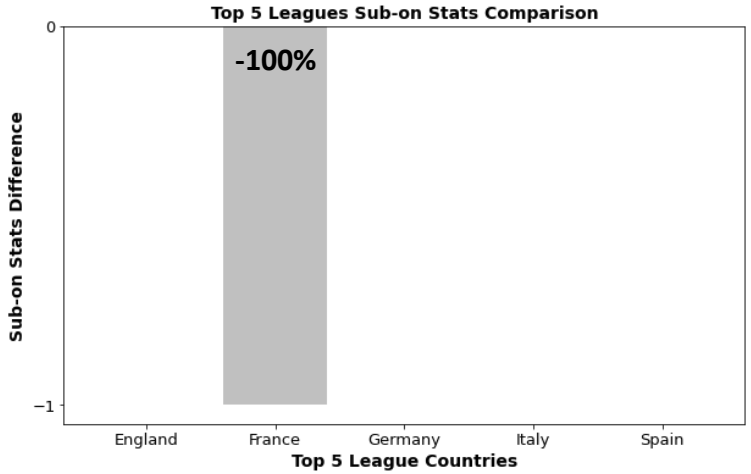
- Yellow Cards = Yellow Cards + Second Yellow Cards
- Red Card stats are consistent in general.
- OPTA stats include more Yellow Cards than TFM stats for all Top 5 league.

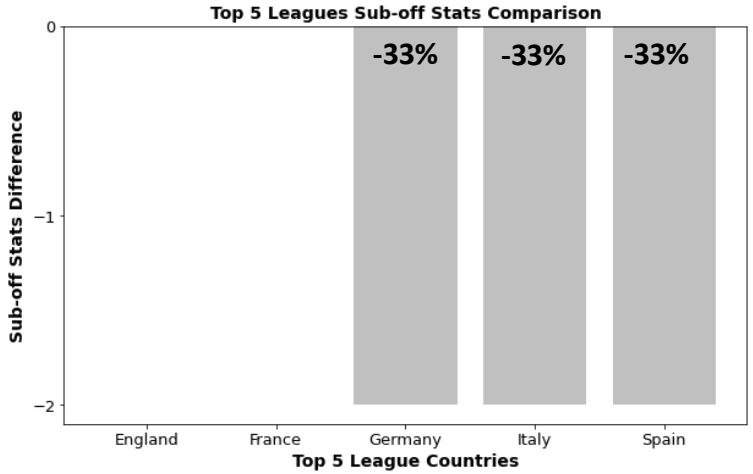


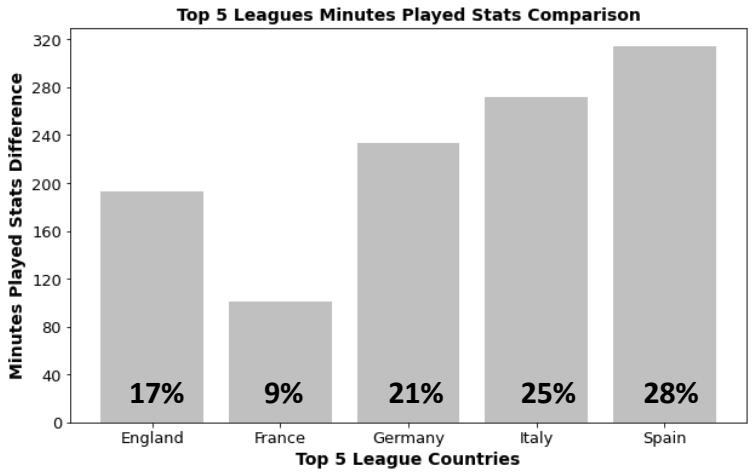


OPTA vs TFM Comparison - TOP 5 Player Workload Stats

- Player Workload Stats = Sub-on+ Sub-off+ Minutes Played
- Substitution stats are consistent in general.
- OPTA stats include more Minutes Played stats than TFM stats for all Top 5 leagues.







Conclusion

- Coverage analysis was conducted to assess the completeness of data at the league-level.
  - Top 5 leagues had full coverage (i.e., level 15).

- Stage analysis aided in comprehending the distinct phases along with the corresponding number of matches played in each. Differences were observed in comparison to league structures for the year 2022.
  - In general, stages other than regular season were also included in the OPTA dataset.

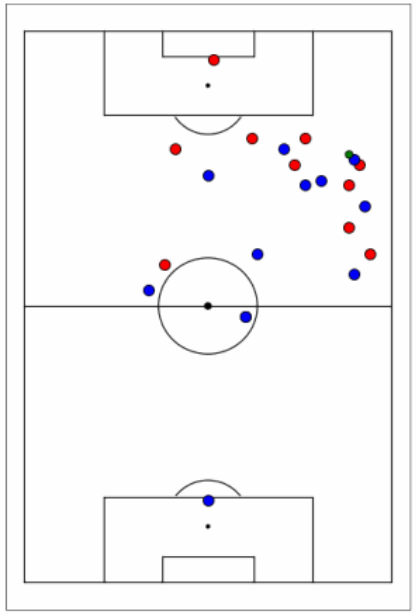
- OPTA and TFM datasets were compared at the league-level and differences were investigated for goal contributions, cards and player workload stats, focusing specifically on the top 5 leagues.
  - Goals, Red Cards and Substutitons were consistent in general.
  - TFM stats include more Assists than OPTA stats for all Top 5 leagues.
  - OPTA stats include more Minutes Played and Yellow Cards than TFM stats for all Top 5 leagues.

TECHNICAL ANALYSIS

Concepts

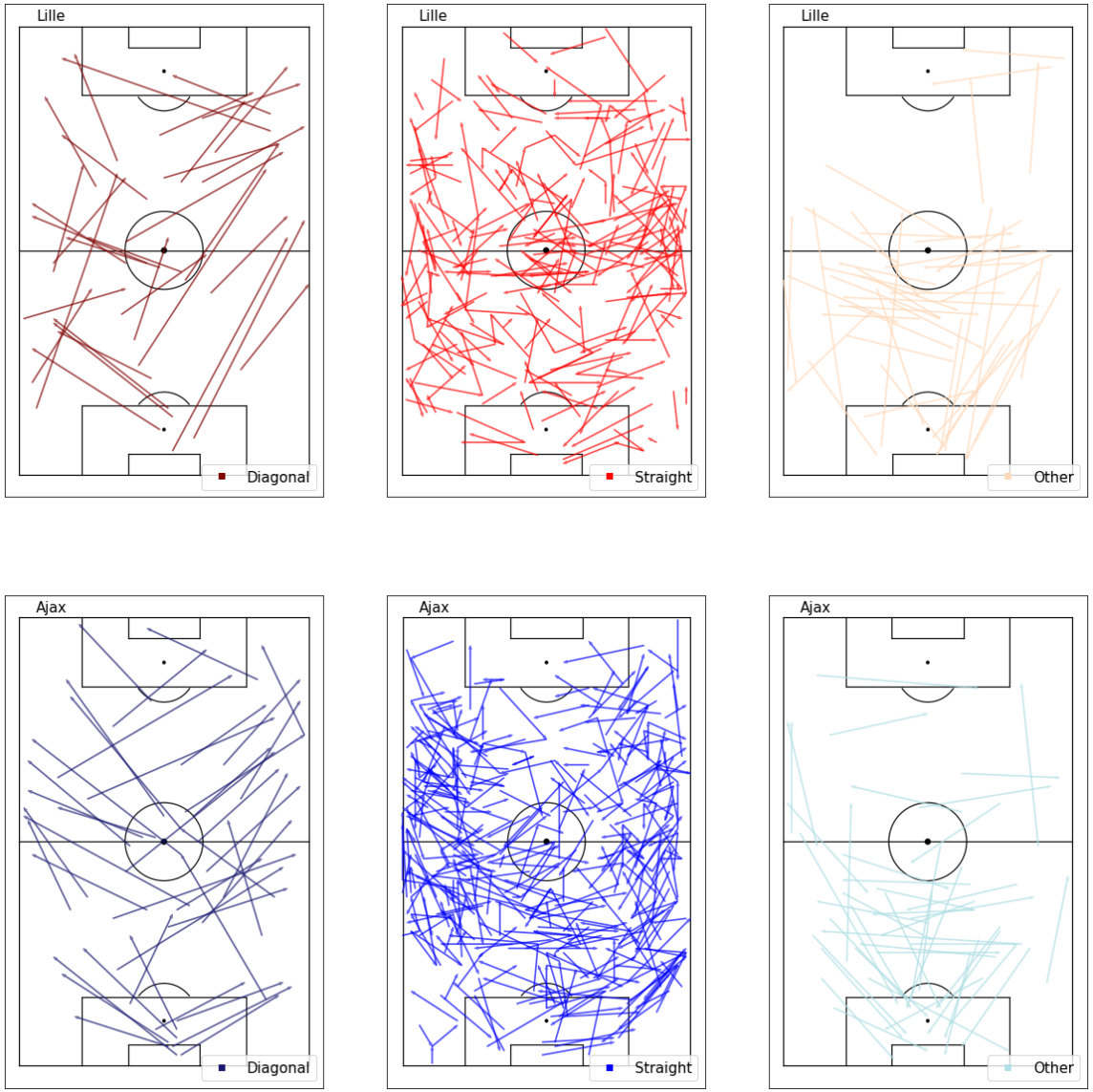
- **Ball Location**
  1) In which third is the ball? (penalty box. final third, midfield third, defensive third)
- **Press Classification**
  2) What kind of press is being employed? (press type e.g., direct, indirect)
- **Running Total of Effective Time**
  3) Possession (%)
- **Running Total of Ball Out of Play**
  4) Ball out of Play (%)
- **Progression Classification**
  5) Progression from defensive third to midfield third (% & totals)
  6) Unsuccessful progressions from defensive third to midfield third (% & totals)
  7) Progression from midfield third to final third (% & totals)
  8) Unsuccessful progressions from midfield third to final third (% & totals)
- **Pass Classification**
  9) Straight passes
  10) Diagonal passes

First Goal
Ajax



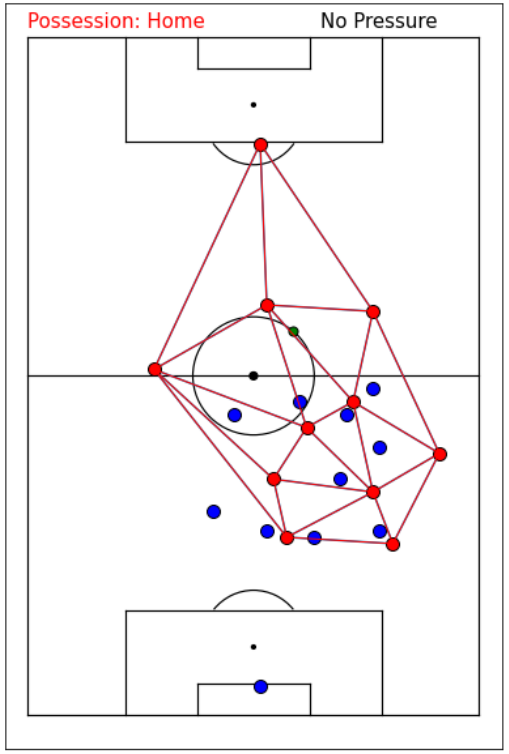
Example Outputs

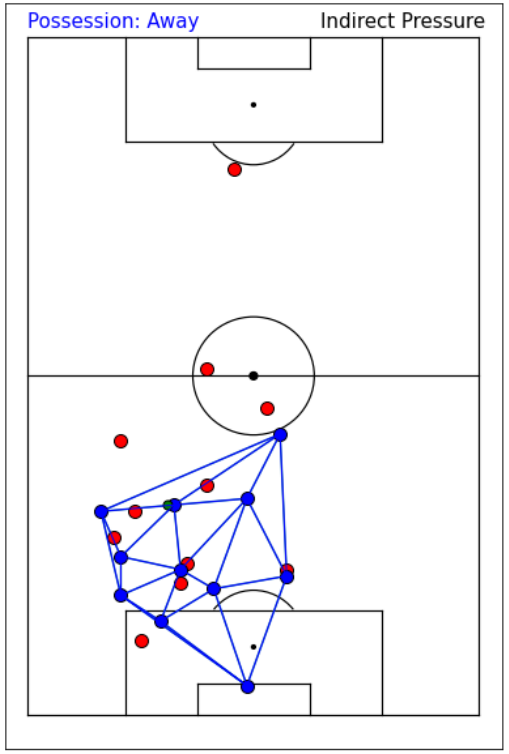
- Output obtained from the implementation of pass classification

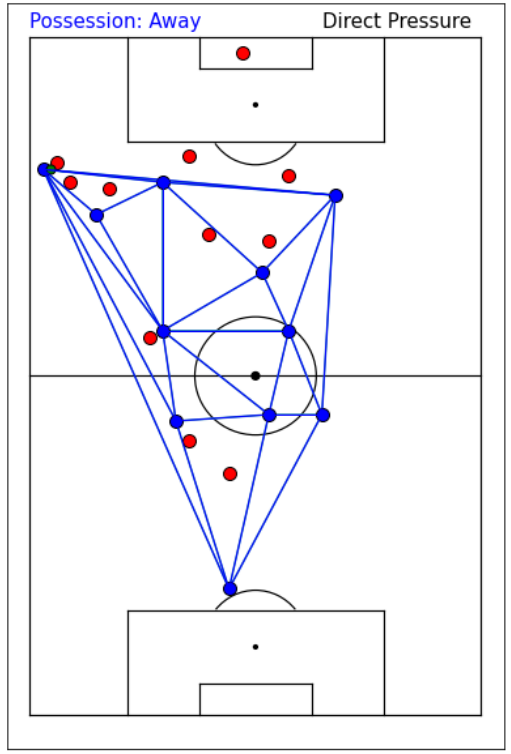


Example Outputs

- Output obtained from the implementation of press classification







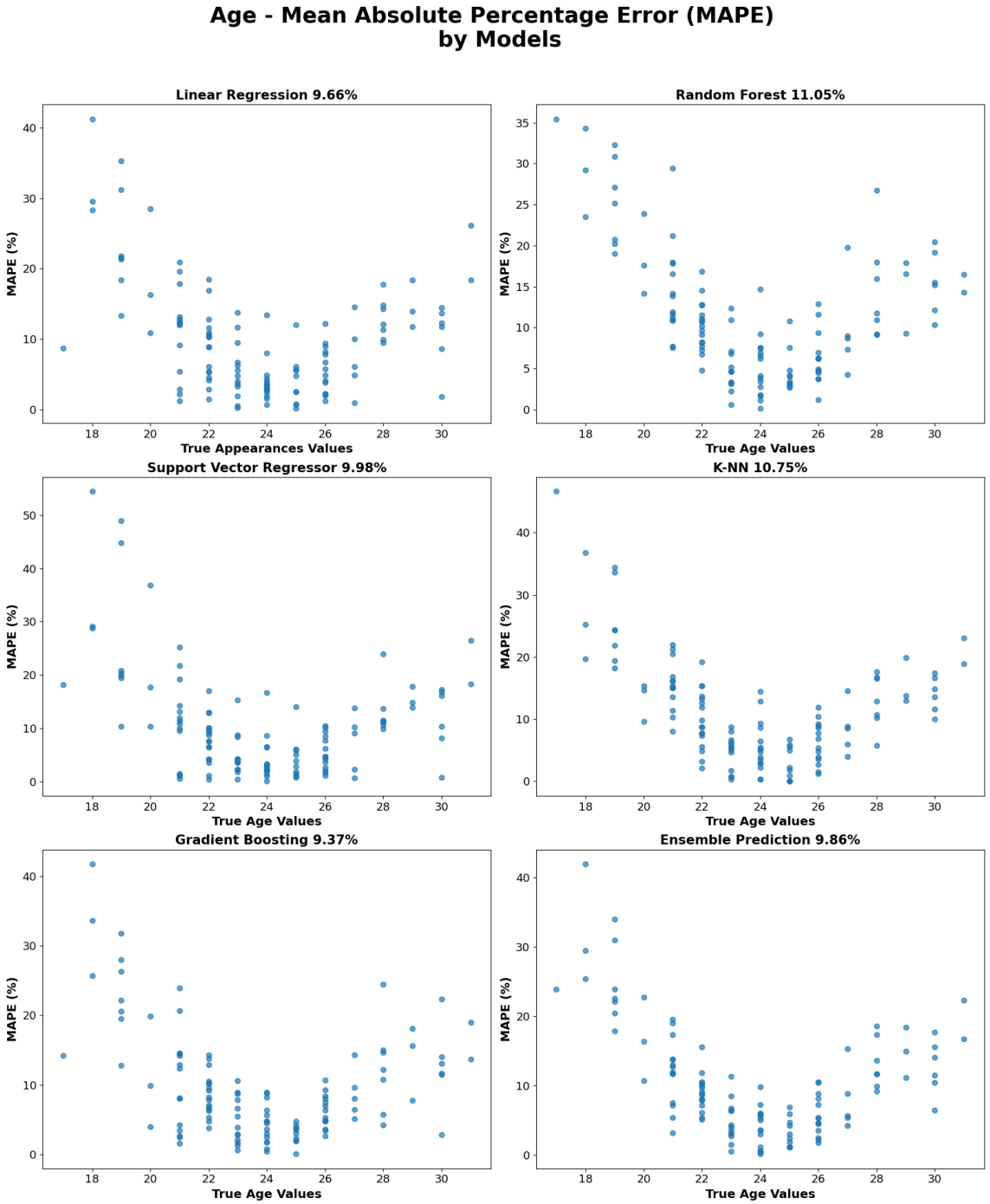
Q&A

APPENDIX

Appearance Prediction

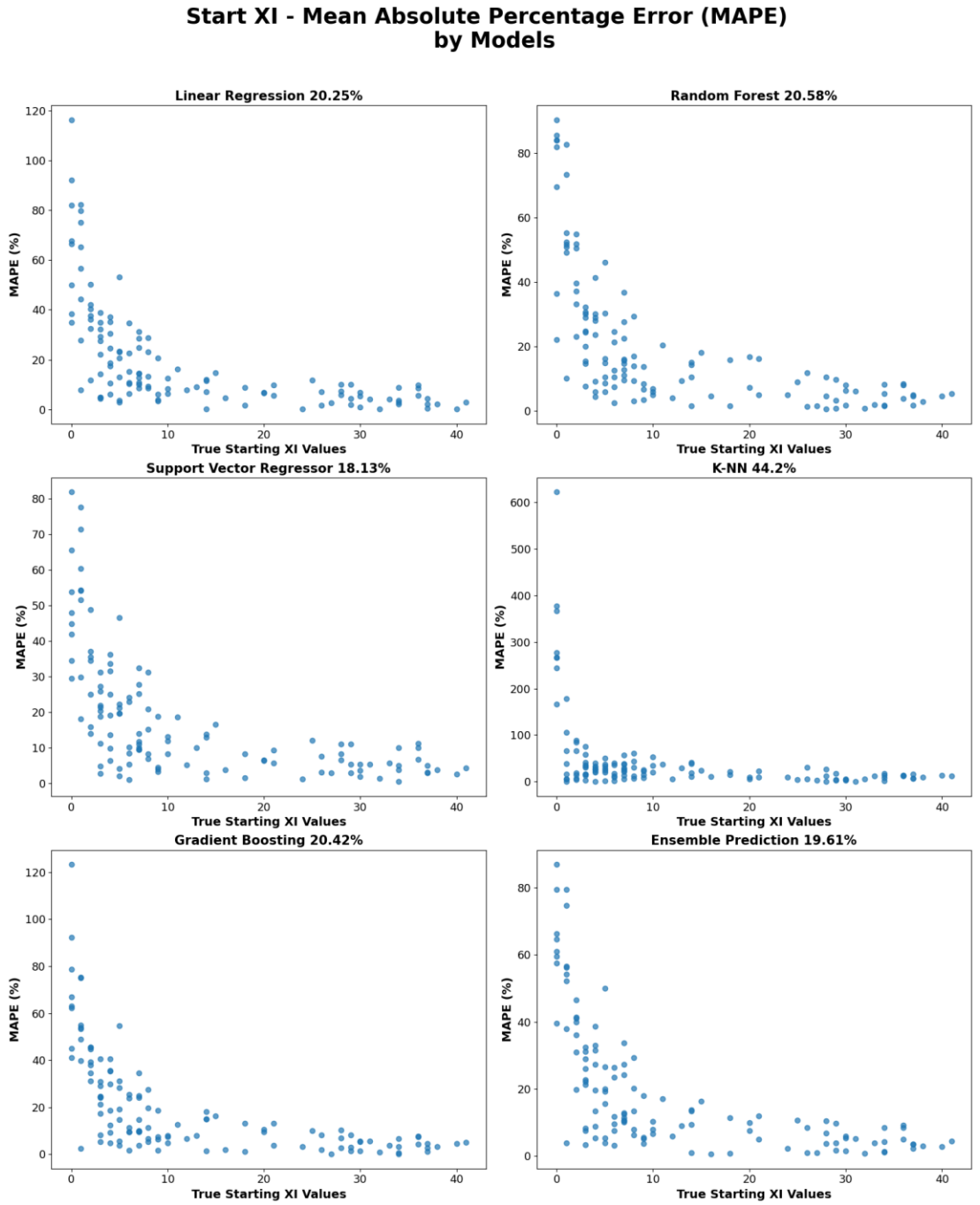
- On average, models achieved a 10% Mean Absolute Percentage Error, indicating a 10% average deviation from the true ages.

- The plots indicated that each model performed better in predicting the age group around 24, given that most top 500 players fall within that age range.



Starting XI Prediction

- All models, except for k-NN, exhibited similar performance
- MAPE decreased exponentially with increasing age
- Although some outliers with low starting XI rates influenced high MAPE values, overall, the models demonstrated stable and consistent performance across varying starting XI rates



Player Position Prediction

- A random choice of models would yield 25% accuracy, given the presence of four types of positions
- SVC and Gradient Boosting Classifier outperformed the other two models and were chosen for the final prediction
- Accurate predictions were observed for defense and forward players, as evidenced by the diagonal values