Exploring Football Player Market Value Trends with Age

BY: Alex Ekka

Exploratory Data Analysis Project

2024

mswordcoverpages.com

❖ Contents

❖ Introduction

Welcome to the documentation for the project titled "Exploring Player Market Value Trends with Age." In this study, we delve into the fascinating intersection of statistical metrics and market values in the soccer industry. The primary objective is to identify undervalued players based on their on-field performance relative to market expectations. Additionally, we investigate the influence of player age and position on their market value.

➢ Project Overview

In the dynamic and competitive world of soccer, understanding player valuations is crucial for clubs, scouts, and enthusiasts alike. By leveraging web scraping techniques, we have gathered extensive data on player statistics and market values to perform a comprehensive analysis. Our goal is to provide actionable insights that can benefit clubs in player recruitment, scouts in talent identification, and enthusiasts in gaining a deeper understanding of the factors influencing market valuations.

➢ Key Objectives

1. **Identifying Undervalued Players:** Utilizing statistical metrics, we aim to pinpoint players whose on-field performance exceeds market expectations, potentially uncovering hidden gems for clubs and scouts.

2. **Analyzing Age and Position Impact:** We explore how player age and position correlate with market values, shedding light on the dynamics of age-related valuation trends and positional considerations.

➢ Scope and Significance

By providing a data-driven exploration of player market value trends, this study contributes valuable insights to the soccer community. The findings aim to empower decision-makers with the knowledge to make informed choices in player recruitment, ultimately enhancing team performance and strategic planning.

➢ How to Use This Documentation

This documentation is structured to guide you through the various aspects of our web scraping project, data analysis, and the presentation of results. Whether you are a seasoned professional or a soccer enthusiast, you'll find valuable information to deepen your understanding of player market dynamics.

Thank you for exploring the exciting world of player market value trends with us. Let's dive into the details and uncover the stories hidden in the data!

❖ Data Collection

➢ Source Website

The data for this project was collected from two Websites

1.Fc rating

2.Transfermarket

➢ Data Categories

1. **Player Statistics:** Information such as Overall, Potential, Skills, and other performance metrics.

2. **Market Values:** Player valuation figures based on current market expectations.

➢ Web Scraping Process

To gather the necessary data, we employed the following libraries:

1. **Beautiful Soup:** A Python library for pulling data out of HTML and XML files, making it essential for parsing the website's HTML structure.

2. **Requests:** Used to send HTTP requests and retrieve the HTML content of the web pages from the source website.

➢ Data Collection Workflow

1. **URL Identification:** We identified the specific URLs containing the relevant player data on the source website.

2. **Web Scraping Script:** A Python script was developed to navigate through the website, extract player information, and store it for further analysis.

➢ Raw Data Samples

Below are snippets of the raw data obtained from the web scraping process:

	Player	Position	Age	Nationality1	Nationality2	Club	MarketValue(m)
0	Erling Haaland	Centre-Forward	22	Norway	NaN	Manchester City	180
1	Kylian Mbappé	Centre-Forward	24	France	Cameroon	Paris Saint-Germain	180
2	Vinicius Junior	Left Winger	22	Brazil	Spain	Real Madrid	150
3	Jude Bellingham	Central Midfield	20	England	NaN	Real Madrid	120
4	Bukayo Saka	Right Winger	21	England	Nigeria	Arsenal FC	120
...	...	...	...	...	...	...	...
495	Adrien Truffert	Left-Back	21	France	NaN	Stade Rennais FC	18
496	Odilon Kossounou	Centre-Back	22	Cote d'Ivoire	NaN	Bayer 04 Leverkusen	18
497	Wilfried Gnonto	Left Winger	19	Italy	Cote d'Ivoire	Leeds United	18
498	Iliman Ndiaye	Centre-Forward	23	Senegal	France	Sheffield United	18
499	Vitinha	Centre-Forward	23	Portugal	NaN	Olympique Marseille	18

500 rows × 7 columns

	rank	Player	OVA	POT	ATT	SKI	MOV	POW	MEN	DEF	GK	STATS
0	1.	Kylian MbappéST|Paris Saint-Germain	91	94	82	81	92	82	74	31	8	2,204
1	2.	Erling HaalandST|Manchester City	91	94	76	70	82	86	75	42	10	2,156
2	3.	Kevin De BruyneCM|Manchester City	91	91	82	89	79	82	83	62	11	2,317
3	4.	Lionel MessiCAM|Inter Miami CF	90	90	85	93	88	77	74	26	11	2,166
4	5.	Karim BenzemaCF|Al Ittihad	90	90	87	82	80	82	76	28	8	2,152
...	...	...	...	...	...	...	...	...	...	...	...	...
104	96.	Paul PogbaCM|Juventus	84	84	78	87	69	81	79	61	4	2,167
105	97.	CanalesRM|Real Betis Balompié	84	84	72	84	83	70	80	67	14	2,215
106	98.	F. KessiéCDM|FC Barcelona	84	86	70	70	77	83	82	82	11	2,206
107	99.	WevertonGK|Palmeiras	84	84	21	23	47	52	28	14	83	1,282
108	100.	G. De ArrascaetaCAM|Free Agency	84	84	76	83	83	69	65	43	13	2,069

109 rows × 12 columns

❖ . Data Cleaning

In this section, we walk through the data cleaning process for the collected datasets, focusing on standardizing column names and extracting relevant information.

➢ Dataframes Overview

Dataframe 1: No Cleaning Required

Which is stored in dfmarket variable

The first dataset is well-structured, and no data cleaning is necessary. Here are the initial column names:

Resulting Dataframe

	Player	Position	Age	Nationality1	Nationality2	Club	MarketValue(m)
0	Erling Haaland	Centre-Forward	22	Norway	NaN	Manchester City	180
1	Kylian Mbappé	Centre-Forward	24	France	Cameroon	Paris Saint-Germain	180
2	Vinicius Junior	Left Winger	22	Brazil	Spain	Real Madrid	150
3	Jude Bellingham	Central Midfield	20	England	NaN	Real Madrid	120
4	Bukayo Saka	Right Winger	21	England	Nigeria	Arsenal FC	120
...	...	...	...	...	...	...	...
495	Adrien Truffert	Left-Back	21	France	NaN	Stade Rennais FC	18
496	Odilon Kossounou	Centre-Back	22	Cote d'Ivoire	NaN	Bayer 04 Leverkusen	18
497	Wilfried Gnonto	Left Winger	19	Italy	Cote d'Ivoire	Leeds United	18
498	Iliman Ndiaye	Centre-Forward	23	Senegal	France	Sheffield United	18
499	Vitinha	Centre-Forward	23	Portugal	NaN	Olympique Marseille	18

500 rows × 7 columns

Dataframe 2: Column Renaming and Club Extraction

Which is stored in dfplayer

The second dataset required some cleaning steps, including renaming columns for better clarity and extracting information from the "PLAYER" column.

- Column Renaming

We renamed the columns to improve readability and consistency:

Index(['rank', 'Player', 'Overall', 'Potential', 'Attacking', 'Skills',

'Movement', 'Power', 'Mentality', 'Defending', 'Goalkeeping', 'STATS', 'club'],

dtype='object')

- Player and Club Extraction

We further processed the "PLAYER" column to extract player names and club information. The resulting dataframe includes a new column called "CLUB" containing the extracted club names.

Resulting Dataframe

	rank	Player	Overall	Potential	Attacking	Skills	Movement	Power	Mentality	Defending	Goalkeeping	STATS	club
0	1.0	Kylian MbappéST	91.0	94.0	82.0	81.0	92.0	82.0	74.0	31.0	8.0	2,204	Paris Saint-Germain
1	2.0	Erling HaalandST	91.0	94.0	76.0	70.0	82.0	86.0	75.0	42.0	10.0	2,156	Manchester City
2	3.0	Kevin De BruyneCM	91.0	91.0	82.0	89.0	79.0	82.0	83.0	62.0	11.0	2,317	Manchester City
3	4.0	Lionel MessiCAM	90.0	90.0	85.0	93.0	88.0	77.0	74.0	26.0	11.0	2,166	Inter Miami CF
4	5.0	Karim BenzemaCF	90.0	90.0	87.0	82.0	80.0	82.0	76.0	28.0	8.0	2,152	Al Ittihad
...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	96.0	Paul PogbaCM	84.0	84.0	78.0	87.0	69.0	81.0	79.0	61.0	4.0	2,167	Juventus
96	97.0	CanalesRM	84.0	84.0	72.0	84.0	83.0	70.0	80.0	67.0	14.0	2,215	Real Betis Balompié
97	98.0	F. KessiéCDM	84.0	86.0	70.0	70.0	77.0	83.0	82.0	82.0	11.0	2,206	FC Barcelona
98	99.0	WevertonGK	84.0	84.0	21.0	23.0	47.0	52.0	28.0	14.0	83.0	1,282	Palmeiras
99	100.0	G. De ArrascaetaCAM	84.0	84.0	76.0	83.0	83.0	69.0	65.0	43.0	13.0	2,069	Free Agency

100 rows × 13 columns

❖ Data Description

In this section, we provide a detailed description of the columns in the two datasets, highlighting key information and providing a brief overview of each variable.

- Dataframe 1

1) Columns Overview

1. **Player:** The name of the soccer player.
2. **Position:** The playing position of the player.
3. **Age:** The age of the player.
4. **Nationality1:** The primary nationality of the player.
5. **Nationality2:** The secondary nationality of the player (if applicable).
6. **Club:** The current club of the player.
7. **MarketValue(m):** The market value of the player in millions.

2) Key Insights

- The dataset provides comprehensive information about individual players, including their positions, ages, nationalities, current clubs, and market values.

- Dual nationality is considered for certain players, providing a nuanced perspective on the diverse backgrounds of the athletes.

Dataframe 2

1**) Columns Overview**

1. **rank:** The ranking of the player.

2. **Player:** The name of the soccer player.

3. **Overall:** This is a general assessment of the player's overall skill and effectiveness on the field..

4. **Potential:** This indicates the player's perceived future potential for development and improvement.

5. **Attacking:** This attribute focuses on the player's skills and contributions in attacking situations, such as dribbling, passing, shooting, and creating goal-scoring opportunities.

6. **Skills:** This is a broader category encompassing various technical abilities like passing, ball control, and receiving.

7. **Movement:** This refers to the player's ability to move fluidly, change direction quickly, and position themselves effectively on the field.

8. **Power:** This encompasses the player's physical strength, stamina, and ability to exert force in actions like tackling, shooting, and shielding the ball

9. **Mentality:** This attribute focuses on the player's psychological aspects, including their mental toughness, decision-making, leadership, and focus during the game..

10. **Defending:** This attribute highlights the player's effectiveness in defensive situations, including tackling, marking opponents, and preventing goals.

11. **Goalkeeping:** This attribute is specifically relevant for goalkeepers and assesses their skills in shot-stopping, aerial presence, communication, and command of the penalty area.

12. **STATS:** his could represent various statistical data associated with the player's performance, such as goals scored, assists made, tackles won, pass completion percentage, etc.13. **club:** The current club of the player.

2**) Key Insights**

- The dataset focuses on player rankings and detailed attributes, providing a granular analysis of each player's skills, potential, and overall performance.

- The "STATS" column includes additional statistical information, contributing to a comprehensive player profile.

- The "club" column identifies the current club affiliation of each player.

These datasets, with their distinct focuses, offer valuable insights for various analyses in the realm of soccer player evaluation and market trends. Proceed to the following sections for in-depth analysis and visualizations.

- summary of the DataFrame

it provides a concise summary of the DataFrame, including information about the data types, non-null values, and memory usage.

```
: dfplayer.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 109 entries, 0 to 108
Data columns (total 12 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   rank    100 non-null    float64
 1   Player  100 non-null    object 
 2   OVA     100 non-null    float64
 3   POT     100 non-null    float64
 4   ATT     100 non-null    float64
 5   SKI     100 non-null    float64
 6   MOV     100 non-null    float64
 7   POW     100 non-null    float64
 8   MEN     100 non-null    float64
 9   DEF     100 non-null    float64
 10  GK      100 non-null    float64
 11  STATS   100 non-null    object 
dtypes: float64(10), object(2)
memory usage: 10.3+ KB
```

```
: dfplayer.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   rank         100 non-null    float64
 1   Player       100 non-null    object
 2   Overall      100 non-null    float64
 3   Potential    100 non-null    float64
 4   Attacking    100 non-null    float64
 5   Skills       100 non-null    float64
 6   Movement     100 non-null    float64
 7   Power        100 non-null    float64
 8   Mentality    100 non-null    float64
 9   Defending    100 non-null    float64
 10  Goalkeeping  100 non-null    float64
 11  STATS        100 non-null    object
 12  club         100 non-null    object
dtypes: float64(10), object(3)
memory usage: 10.3+ KB
```

- Descriptive statistics DataFrame

it provides descriptive statistics of the numerical columns in the DataFrame

```
dfmarket.describe()
```

	Age	MarketValue(m)
count	500.000000	500.000000
mean	24.884000	36.770000
std	3.167113	21.760978
min	16.000000	18.000000
25%	23.000000	22.000000
50%	25.000000	30.000000
75%	27.000000	40.000000
max	36.000000	180.000000

```
dfplayer.describe()
```

	rank	Overall	Potential	Attacking	Skills	Movement	Power	Mentality	Defending	Goalkeeping
count	100.000000	100.000000	100.000000	100.000000	100.00000	100.000000	100.000000	100.000000	100.000000	100.000000
mean	50.500000	86.460000	87.910000	64.530000	68.83000	75.260000	71.440000	67.570000	53.600000	23.070000
std	29.011492	1.794605	2.558784	21.009596	20.73452	11.058723	10.687782	15.472527	26.346947	28.123916
min	1.000000	84.000000	84.000000	15.000000	14.00000	45.000000	47.000000	28.000000	13.000000	4.000000
25%	25.750000	85.000000	86.000000	62.750000	64.00000	69.000000	67.000000	66.000000	30.250000	10.000000
50%	50.500000	86.000000	88.000000	72.000000	77.50000	79.000000	73.500000	73.000000	55.000000	11.000000
75%	75.250000	87.250000	90.000000	77.250000	82.00000	83.000000	80.000000	77.000000	80.250000	12.000000
max	100.000000	91.000000	94.000000	87.000000	93.00000	92.000000	87.000000	83.000000	90.000000	89.000000

Here's what each part of the output represents:

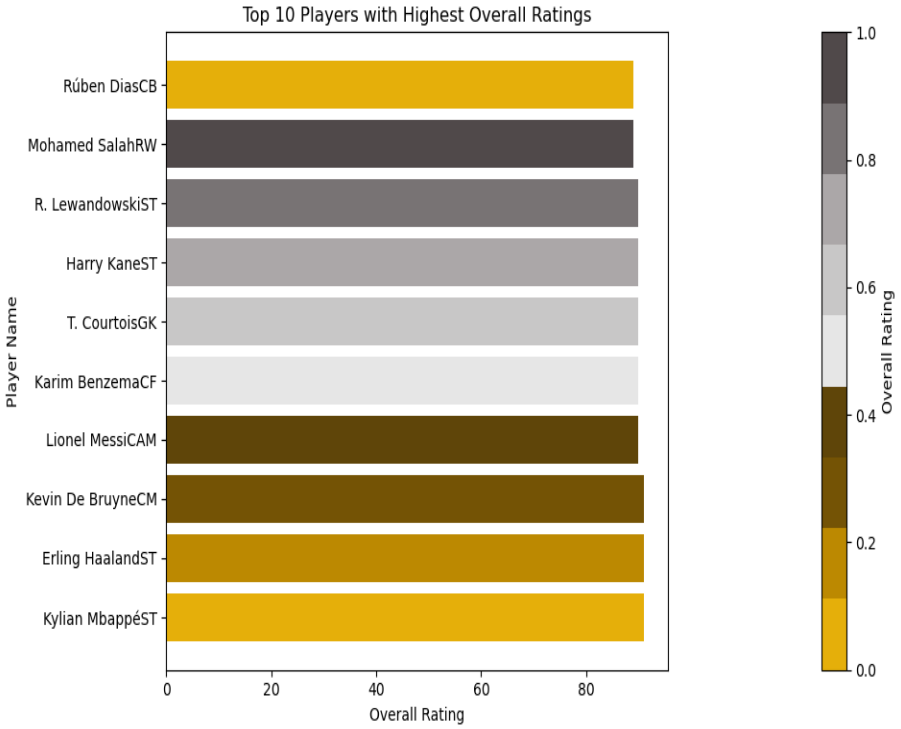
count: Number of non-null values in each column.
mean: Mean (average) value of each column.
std: Standard deviation, a measure of the amount of variation or dispersion in each column.
min: Minimum value in each column.
25%: 25th percentile (first quartile).
50%: Median (50th percentile or second quartile).
75%: 75th percentile (third quartile).
max: Maximum value in each column.

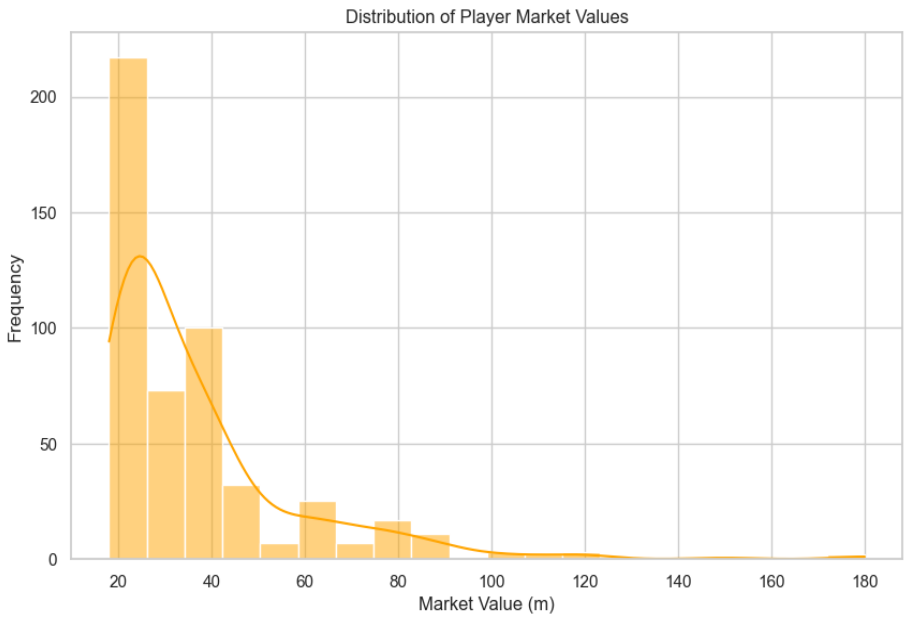
The output provides a quick overview of the central tendency, dispersion, and distribution of the numerical data in your DataFrame. It helps in identifying potential outliers, understanding the range of values, and getting a sense of the overall distribution. Note that this method only considers numerical columns by default and excludes non-numeric columns.

❖ Exploratory Data Analysis (EDA)

In this section, we conduct an Exploratory Data Analysis (EDA) on the collected datasets. EDA is a crucial step in understanding the characteristics of the data, identifying patterns, and deriving meaningful insights.

Exploring Player Market Value Trends with Age

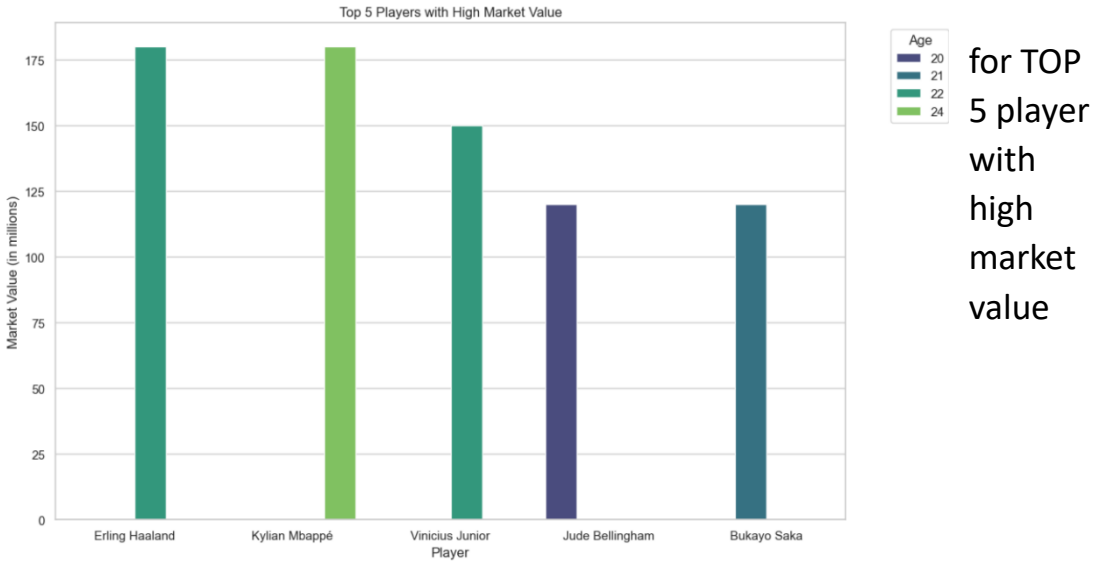




➢ Market Value Summary

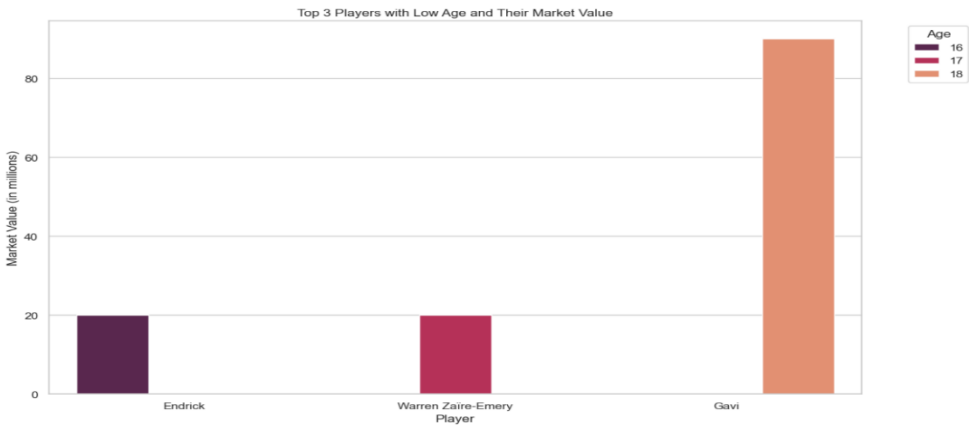
For the dataset related to player market values (`dfmarket`):

- **Count:** 500 players are included in the dataset.
- **Mean:** The average market value is approximately $36.77 million.
- **Standard Deviation:** Market values vary with a standard deviation of $21.76 million.
- **Minimum Value:** The lowest market value recorded is $18 million.
- **25th Percentile:** 25% of players have a market value of $22 million or less.
- **Median (50th Percentile):** The median market value is $30 million.
- **75th Percentile:** 75% of players have a market value of $40 million or less.
- **Maximum Value:** The highest market value observed is $180 million.



for TOP 5 player with high market value

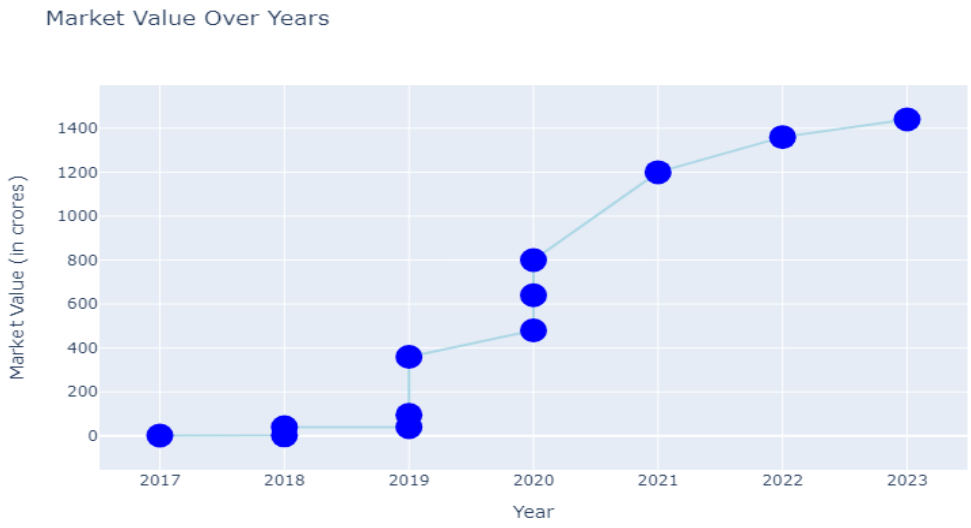
for Top 3 Players with Low Age and Their Market Value



➢ Longitudinal Analysis

In this section, we explore the market value trends of two prominent soccer players, Kylian Mbappe and Erling Haaland, over the years. Analyzing their market values provides insights into the dynamic nature of player valuations in the soccer industry.

Erling Haaland



Kylian Mbappe

Unveiling Undervalued Players

In this section, we delve into the analysis of identifying undervalued players in the soccer industry. The goal is to uncover players whose on-field performance surpasses market expectations, presenting potential opportunities for clubs and scouts seeking valuable talent.

➢ Analysis Approach

1) Fuzzy Matching

To find the best match for each player in the `dfplayer` DataFrame within the `dfmarket` DataFrame, we utilized the `fuzzywuzzy` library. The matching process considered player names, and a match score was assigned based on the similarity.

```python
from fuzzywuzzy import process
```

```python
from fuzzywuzzy import process

# Function to find the best match for a player in the market DataFrame
def find_best_match(player_name, market_players):
    return process.extractOne(player_name, market_players)

# Create a mapping between players in dfplayer and their best match in dfmarket
dfplayer['Best_Match'] = dfplayer['Player'].apply(lambda x: find_best_match(x, dfmarket['Player']))
dfplayer['Best_Match_Player'] = dfplayer['Best_Match'].apply(lambda x: x[0] if x else None)
dfplayer['Match_Score'] = dfplayer['Best_Match'].apply(lambda x: x[1] if x else None)

# Filter rows where the match score is above a certain threshold (adjust as needed)
threshold = 80
dfplayer_filtered = dfplayer[dfplayer['Match_Score'] >= threshold]

# Merge based on the best-matched player
merged_df = pd.merge(dfplayer_filtered, dfmarket, left_on='Best_Match_Player', right_on='Player', how='inner')

# Selecting relevant columns for analysis
selected_columns = ['Player_x', 'Overall', 'Potential', 'Attacking', 'Skills', 'Movement', 'Power',
                    'Mentality', 'Defending', 'Goalkeeping', 'STATS', 'MarketValue(m)']
df_analysis = merged_df[selected_columns].rename(columns={'Player_x': 'Player'})
```

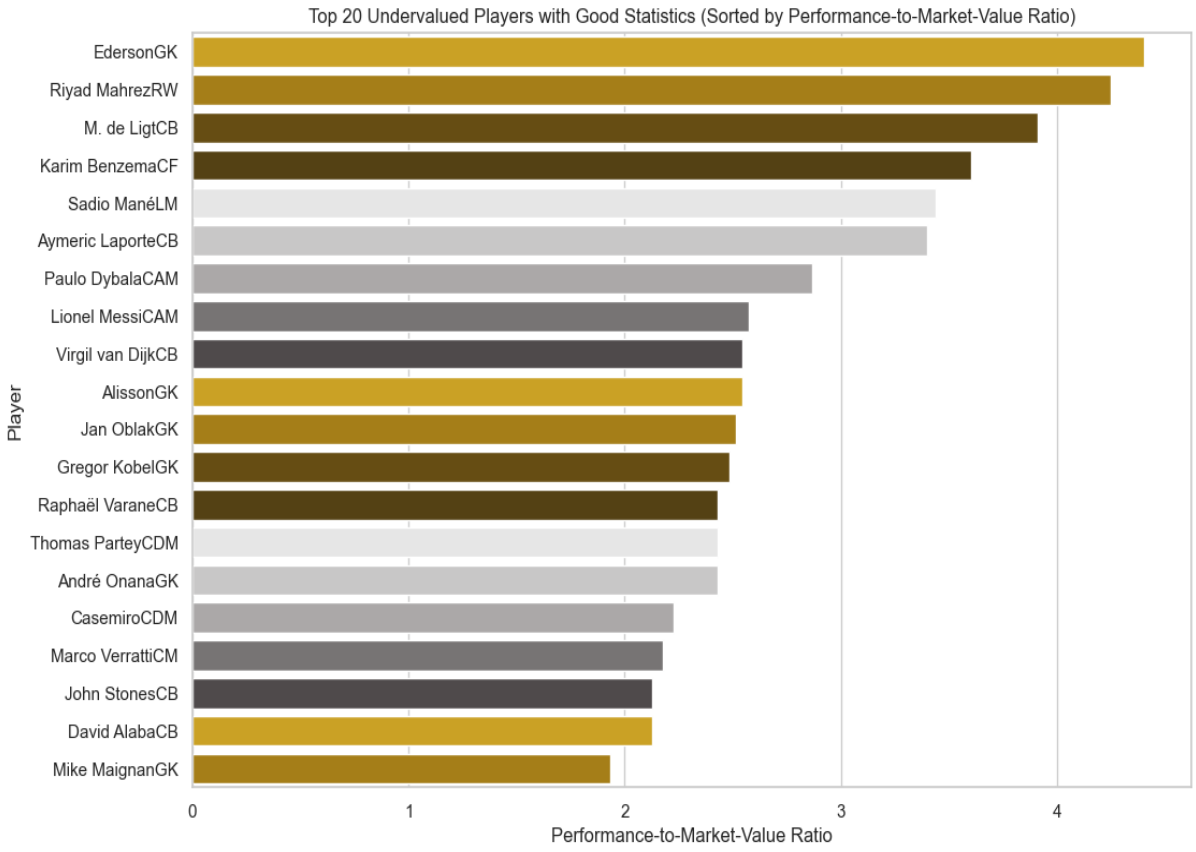
2) Filtering and Analysis

Players with a match score above a specified threshold (here set at 80) were considered for further analysis. The relevant columns were selected, and a new DataFrame (df_analysis) was created by merging the matched players from both DataFrames.

3) Performance-to-Market-Value Ratio

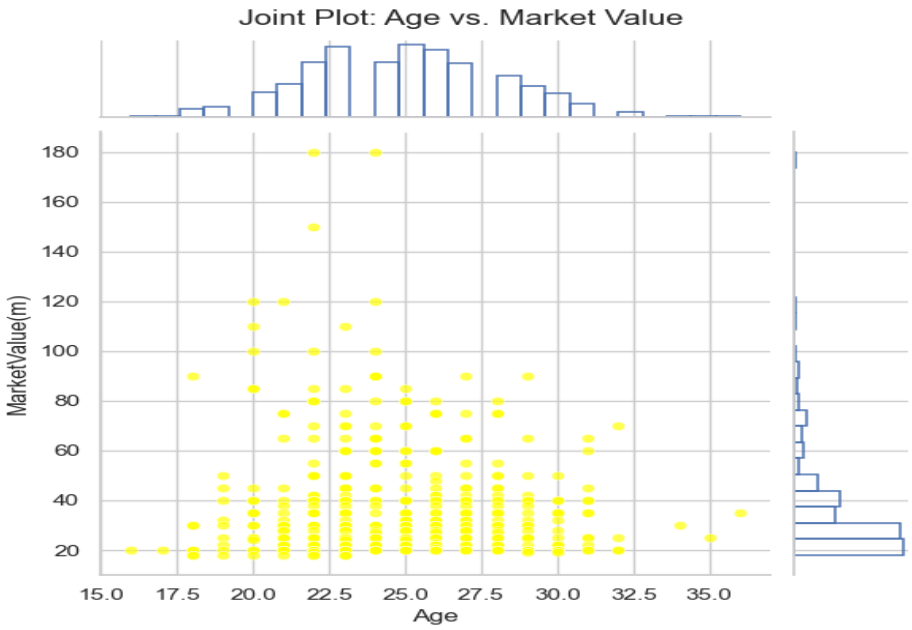
To identify undervalued players, a threshold ratio was applied, and players with a performance-to-market-value ratio exceeding the threshold were considered undervalued.

- Finally, the top 20 undervalued players were selected based on the calculated performance-to-market-value ratio.



These top undervalued players are presented below, showcasing their performance, market value, and the calculated performance-to-market-value ratio. This analysis provides a strategic approach to uncovering undervalued players in the soccer industry, helping clubs and scouts identify potentially valuable talent.

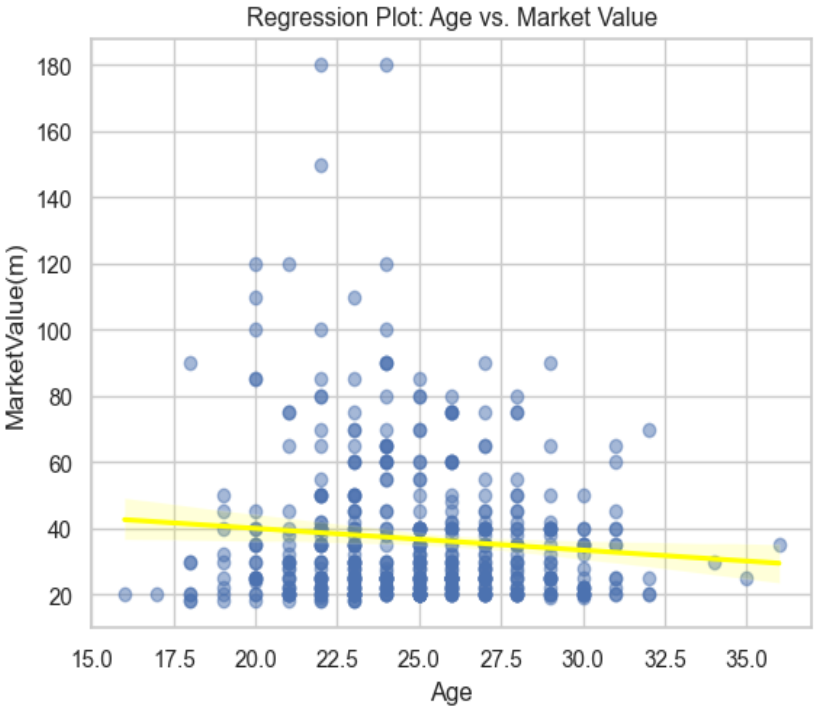
➢ Age and Market Value



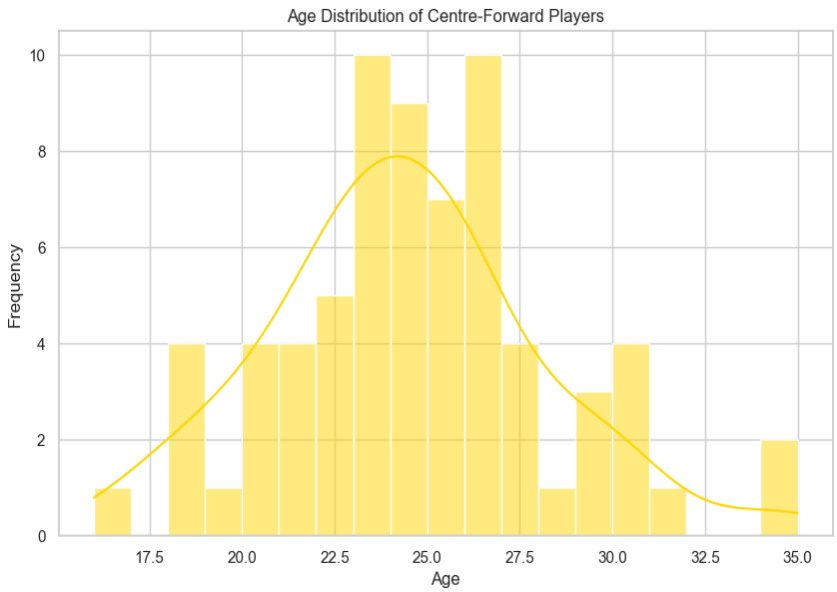
The scatter plot reveals a significant concentration of data points between the ages of 20 and 27.

Negative Slope (Downward from left to right):

- Indicates a negative correlation between the two variables. As one variable increases, the other variable tends to decrease.



➢ Insights by Position



➢ **Age Distribution:**

➢ The majority of Centre-Forward players fall within the age range of 17 to 36 years.

➢ **Peak Age Range:**

➢ The peak of the density curve indicates the most common age range for Centre-Forward players.

➢ **Youthful and Experienced Players:**

➢ The concentration of players between 17 to 35 suggests a mix of both younger, potentially emerging talents, and experienced players in this position.

➢ **Scouting and Recruitment:**

➢ For talent scouts and recruitment strategies, focusing on players within this age range might be more fruitful, considering the density of players in this segment

Exploring Relationships: Age, Market Value, and Position

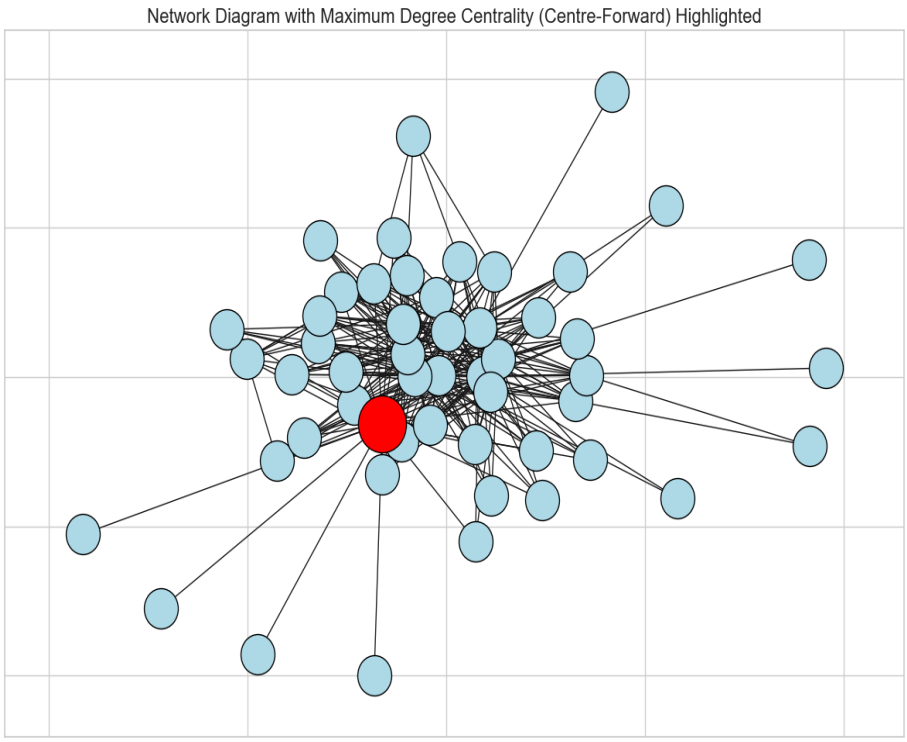
In this section, we employ a network diagram to visualize the intricate relationships between player age, market value, and playing positions. This graphical representation allows for a comprehensive understanding of how these variables are interconnected within the dataset.

➢ Network Diagram

The network diagram provides a visual representation of the relationships between age, market value, and position for soccer players.

Each node in the diagram represents a player, and the edges indicate the connections and correlations between the variables.

![Network Diagram: Age, Market Value, Position]



Key Observations

Age and Market Value Relationship

- Nodes clustered in specific age groups may reveal patterns in market value trends.
- Identify outliers that defy the typical age-to-market-value expectations.

- Market Value and Position Relationship

- Observe if certain playing positions correlate with higher or lower market values.
- Explore how positional roles influence market valuations in the soccer industry.

- Age and Position Relationship

- Identify age distributions within different playing positions.
- Uncover any age-related patterns specific to certain positions.

- Conclusion

Centre-Forward

Centre-Forward" position is highly connected with various ages ranging from 16 to 35 and market values ranging from 18 to 120. This suggests that players in the "Centre-Forward" position exhibit a diverse range of ages and market values in

The network diagram offers a holistic view of the relationships between age, market value, and position in the soccer player dataset. This visual exploration serves as a foundation for more in-depth analyses, allowing for targeted investigations into the factors influencing player market values across different age groups and positions.

Proceed to the next sections for additional visualizations and insights derived from the interconnected nature of these variables.

❖ Conclusion and Key Takeaways

- **Strategic Approach**

Employing a strategic model to identify undervalued players

- **Continuous Evaluation**

Emphasizing the need for regular re-evaluation of player valuation

- **Enhanced Decision Making**

Utilizing data-driven analysis for improved player selection.