

Analyzing Football Player Market Value Using FIFA 23 Data

1. Introduction

Player market value in professional football reflects a complex economic process shaped by performance indicators, physical attributes, and expectations of future development. Clubs do not only invest based on current ability, but also on a player's projected growth and suitability within modern football dynamics.

The initial objective of this project was to evaluate the **consistency of player valuation across multiple data sources**, specifically FIFA, Football Manager, and Transfermarkt. The aim was to assess whether different independent systems assign value to players in a consistent manner.

However, during the data preparation phase, significant inconsistencies in player naming conventions across datasets prevented reliable and automated data merging within the project timeline. Due to this limitation, the scope of the project was refined. The final analysis focuses on **a single, rich dataset (FIFA 23)**, allowing for a deeper methodological exploration through exploratory data analysis, hypothesis testing, and machine learning models.

This shift enabled a structured investigation of **internal relationships within FIFA's valuation framework**, rather than cross-source verification.

2. Dataset and Variables

The dataset used in this study is the **FIFA 23 Complete Player Dataset**, obtained from Kaggle. It contains detailed information on professional football players, including demographic, physical, and rating-based attributes, as well as estimated market value.

The main variables used in the analysis are:

- **Age**: Player age
- **Height (cm)**: Player height
- **Weight (kg)**: Player weight
- **Overall**: Current performance rating
- **Potential**: Estimated future performance level (constructed FIFA metric)
- **Physic**: Physical strength rating
- **Market Value**: Estimated player market value in euros

3. Data Cleaning and Feature Engineering

Prior to analysis, the dataset was cleaned to ensure data quality. Observations with missing market value or invalid physical measurements were removed. Only numeric and measurable attributes were retained.

To enrich the analysis, two derived variables were created:

- **Physical Index** = Height × Weight
- **Body Mass Index (BMI)** = Weight / Height²

These features were introduced to better represent physical profiles beyond raw height and weight measurements.

(Buraya grafik koyabilirsiniz: Boy histogramı)

(Buraya grafik koyabilirsiniz: Kilo histogramı)

Figure explanation:

The distributions of height and weight after data cleaning show realistic ranges, confirming that invalid or extreme measurements have been successfully removed.

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the distributional properties of the variables and their relationships with market value.

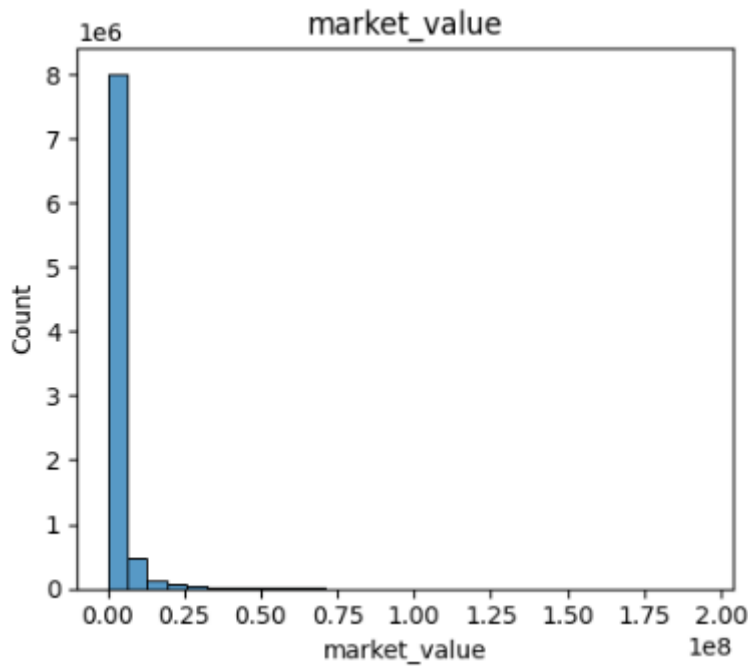


Figure explanation:

The market value distribution is highly right-skewed, with a small number of players accounting for extremely high values. This motivates the use of rank-based statistical tests and non-linear modeling approaches.

4.1 Relationships with Market Value

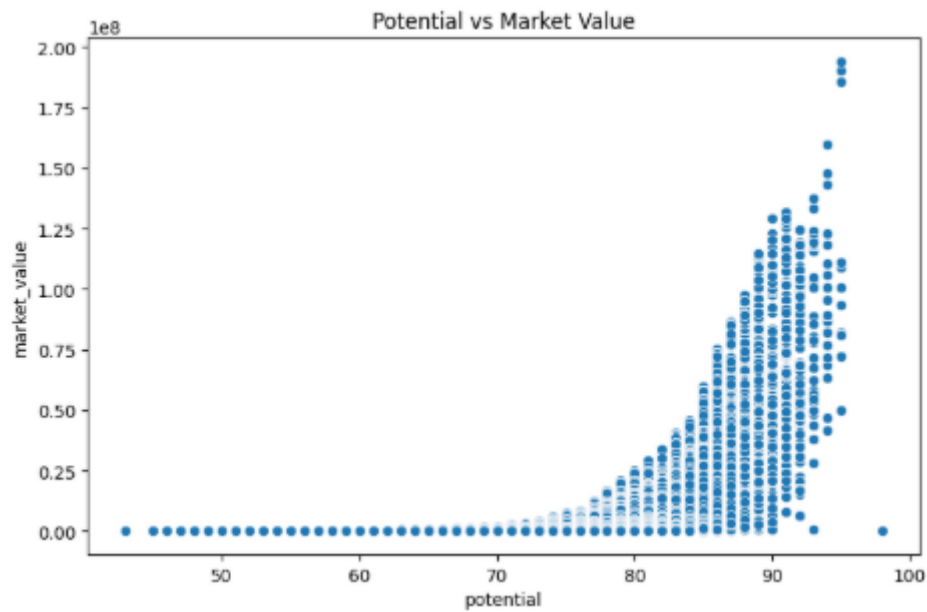


Figure explanation:

A strong, non-linear increasing relationship is observed between player potential and market value, suggesting that future performance expectations play a central role in valuation.

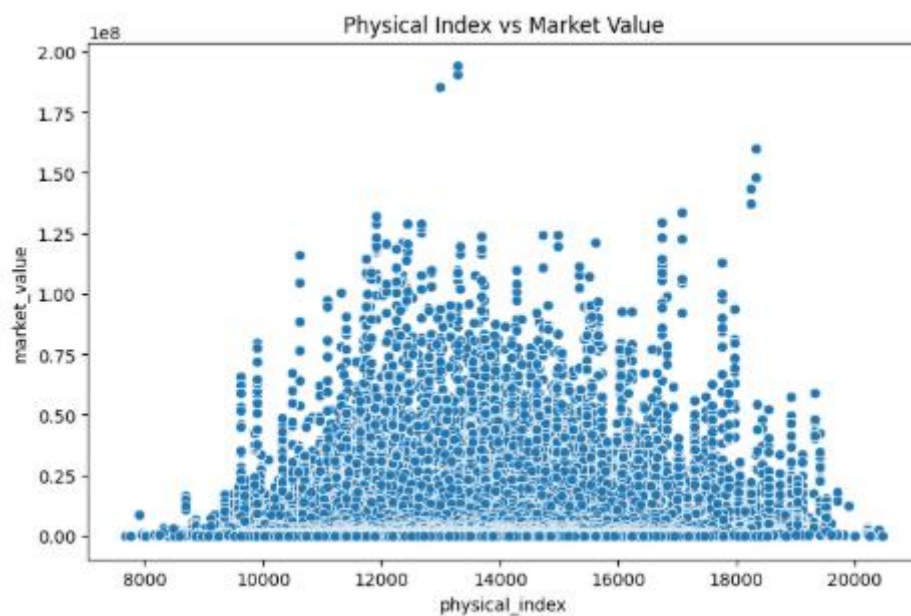


Figure explanation:

The Physical Index displays a limited but observable relationship with market value, indicating that physical strength contributes to valuation but is not a dominant factor.

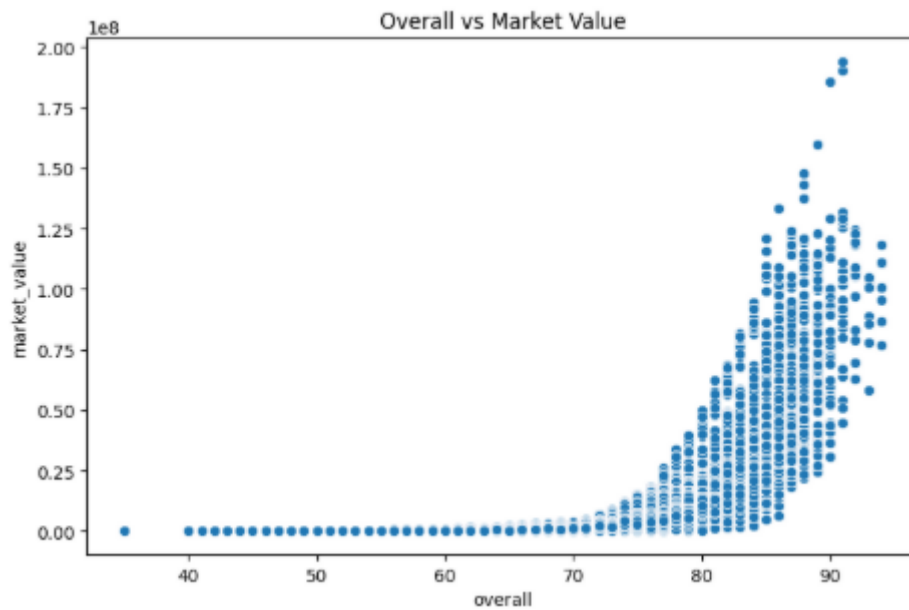


Figure explanation:

Overall rating shows a positive relationship with market value; however, the association appears weaker compared to potential.

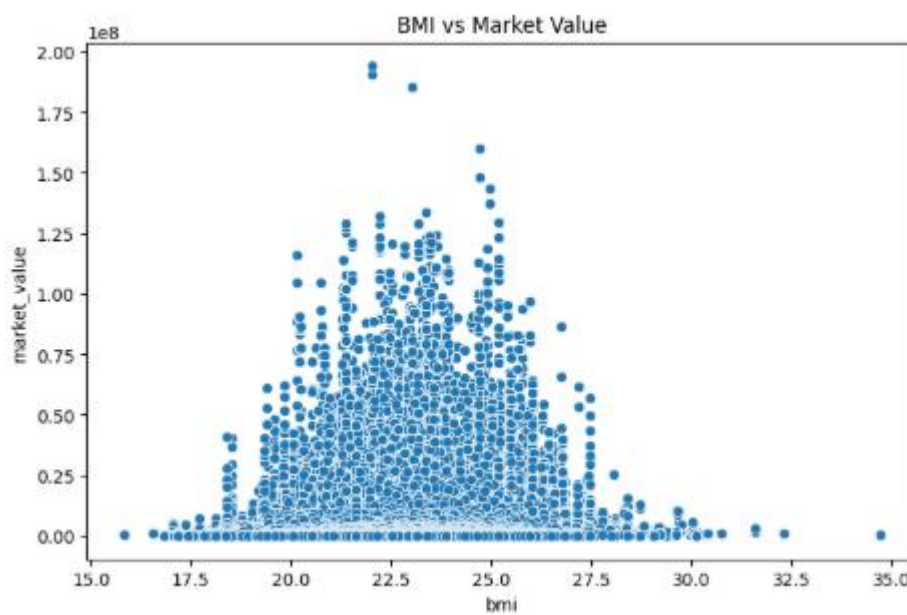


Figure explanation:

BMI shows little to no clear association with market value, suggesting that raw body composition alone is not a strong determinant of player valuation

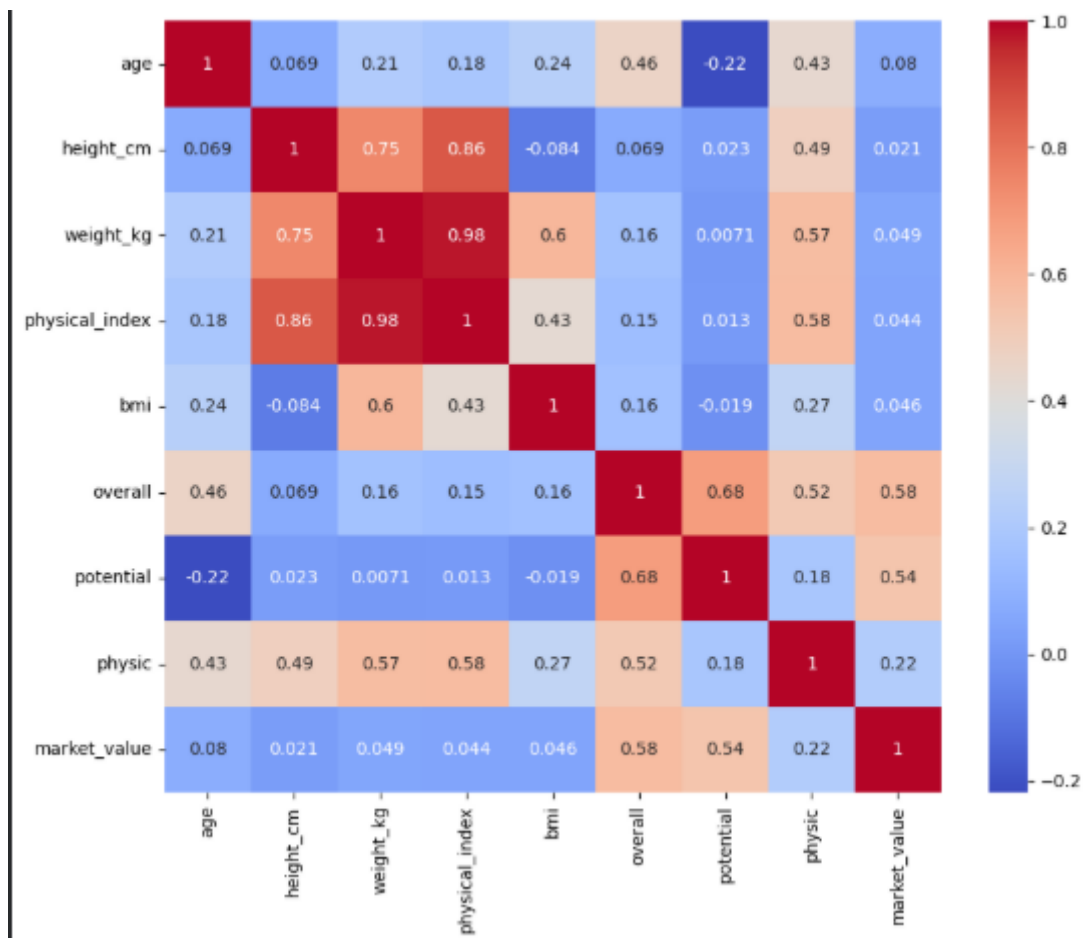


Figure explanation:

The correlation heatmap highlights potential as the variable most strongly associated with market value among the analyzed features.

5. Hypothesis Testing

5.1 Research Question

Which player attributes are most strongly associated with market value?

5.2 Hypothesis

H₀: There is no statistically significant relationship between player attributes and market value.

H₁: Player potential exhibits a stronger association with market value than physical attributes such as physic rating and BMI.

5.3 Methodology

Due to the skewed distribution of market value and the presence of outliers, **Spearman rank correlation** was used. This method is suitable for capturing monotonic but non-linear relationships.

5.4 Results

Variable	Spearman ρ	p-value
----------	-----------------	---------

Potential	0.776	< 0.001
-----------	-------	---------

Physic	0.412	< 0.001
--------	-------	---------

BMI	0.092	< 0.001
-----	-------	---------

These results indicate that **player potential has the strongest and most statistically significant relationship with market value**, while physic shows a moderate relationship and BMI a very weak one.

6. Machine Learning Analysis

6.1 Model Setup

A supervised learning approach was applied to predict player market value. The dataset was split into training (80%) and testing (20%) sets, and all features were standardized.

6.2 Models Applied

- **Linear Regression** (interpretable baseline model)
- **K-Nearest Neighbors (KNN) Regression** with $K = 5$

6.3 Linear Regression Results

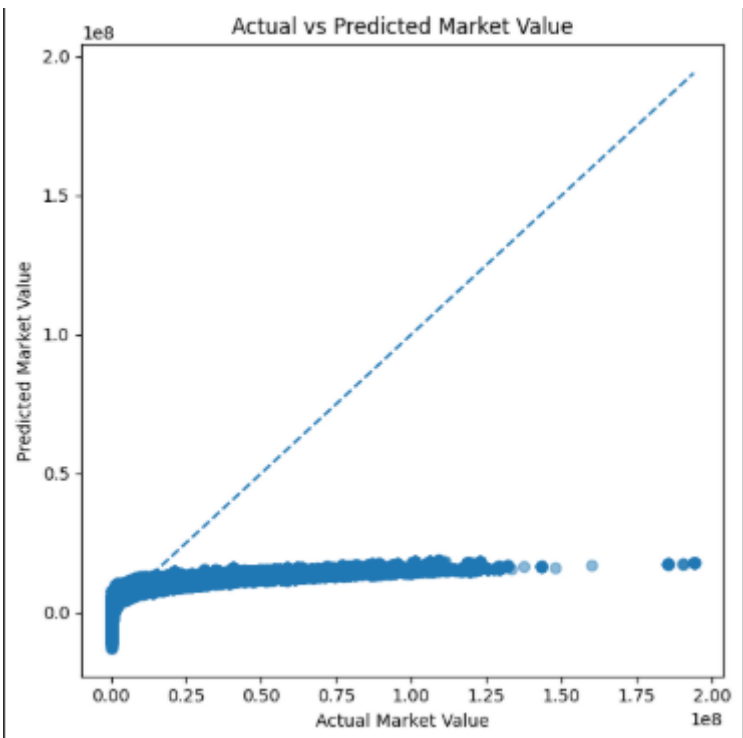


Figure explanation:

Linear Regression captures the overall trend but struggles with extreme market values.

6.4 KNN Results

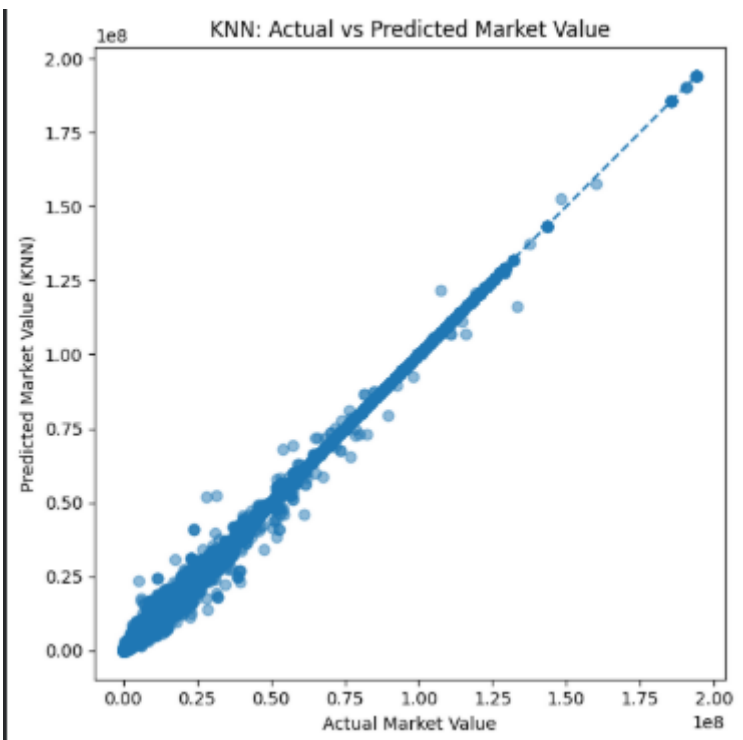


Figure explanation:

KNN achieves near-perfect predictions on the test set by leveraging local similarity between players. This suggests strong memorization effects rather than global generalization.

7. Discussion

Results from EDA, hypothesis testing, and machine learning models consistently indicate that **potential is the dominant driver of market value within FIFA's valuation framework**. The stark performance difference between Linear Regression and KNN highlights that FIFA market values are closely aligned with discrete rating-based player profiles rather than a smooth economic function.

8. Limitations and Future Work

This study relies on a single data source, reflecting FIFA's internal valuation logic rather than real-world market mechanisms. Player potential is treated as a constructed rating rather than a causal economic variable.

Future work may involve integrating Football Manager and Transfermarkt datasets using advanced name-matching techniques to enable external validation and cross-source consistency analysis.

9. Conclusion

This project demonstrates that within FIFA 23 data, player potential exhibits the strongest association with market value compared to physical attributes. By combining exploratory analysis, statistical hypothesis testing, and machine learning models, the study provides a structured understanding of how valuation patterns emerge within a rating-based system.