**Comprehensive Report: Data Warehouse Project on Football Player Market Value Analysis**

**Project Objective:** The goal of this project is to analyze and explain the various factors influencing the market value of football players using a structured Data Warehouse and a series of analytical models. Through the development of several SQL views and Python-based statistical analyses, we investigate the role of age, performance metrics, league bias, positional trends, and seasonal dynamics on market value.

**1. Age vs Market Value (First Analysis)**

**Objective:** Evaluate how age impacts the market value of players with similar performances, highlighting market preferences for youth.

**Method:**

* Selected pairs of players with at least a 5-year age difference but with similar performance metrics.
* Compared the market value between the younger and older player in each pair.
* Computed:
  + **Age gap** = older\_age - young\_age
  + **Value gap** = young\_mv - older\_mv

**Visualization:**

* **Scatter Plot**: Showed age\_gap on the x-axis and value\_gap on the y-axis with a regression line.
* **Box Plot**: Compared overall distribution of market values between older and younger players.

**Findings:**

* Across all roles, younger players tend to have higher market values than older players with similar performance, confirming a market bias toward youth.
* The regression line typically showed a positive slope, indicating greater value differences as age gap increases.

**2. Metric Importance via R-Squared (Second Analysis)**

**Objective:** Identify which individual performance metrics most strongly explain market value for each player role.

**Method:**

* Created role-specific views with 5 relevant performance metrics.
* For each metric, performed simple linear regression:
  + **y** = market\_value\_m
  + **x** = metric (e.g., goals\_per\_90, tackles\_per90)
* Calculated **R²** to assess explanatory power.

**Visualization:**

* **Bar Chart** of R² scores for each metric.

**Findings:**

* Forwards: Goals\_per\_90 and npxg\_per90 had the highest impact.
* Midfielders: Assists\_per\_90 and progressive\_passes\_per90 were most predictive.
* Defenders: Tackles and clearances\_per90 explained the most value variance.
* This helped isolate which stats clubs prioritize when valuing players.

**3. League Bias Based on Equal Performance (Third Analysis)**

**Objective:** Examine if players with similar performances are valued differently depending on the league they play in.

**Method:**

* Built composite performance scores using z-scores across five metrics.
* Binned players into 5 performance quintiles.
* For each quintile:
  + Compared player market values across leagues (e.g., Premier League, Serie A).

**Visualization:**

* **Box Plots** for each league per performance quintile.
* **Regression Model** with league fixed effects to detect systematic over/undervaluation.

**Findings:**

* Players in Premier League and La Liga often valued higher even within the same performance band.
* Ligue 1 and Bundesliga showed more conservative valuations.
* Suggests visibility and financial power of leagues skew market prices.

**4. Temporal Evolution of Market Value (Fourth Analysis)**

**Objective:** Understand how player market value trends evolved over time, globally and by role.

**Method:**

* Used 2018–2024 player-season data.
* Computed average market value per season.
* Analyzed trends for:
  + All players
  + Midfielders
  + Forwards
  + Defenders

**Visualization:**

* **Box Plots**: Distribution of values by season.
* **Line Plot**: Season-wise average market value.

**Findings:**

* Clear upward trend in player values from 2018 to 2022, slight plateau after 2022.
* Forwards saw the sharpest increase in average value.
* Defenders lagged behind, consistent with prior analyses.

**5. Under- and Overvalued Players via Residual Analysis (Fifth Analysis)**

**Objective:** Identify players who are significantly over- or undervalued relative to their performance metrics.

**Method:**

* Built multivariate regression model using 5 metrics per role.
* Calculated:
  + **Predicted Value** (based on model)
  + **Residual** = actual - predicted
* Ranked players by residuals.

**Visualization:**

* **Scatter Plot**: Predicted vs Actual market value.
* **Bar Chart**: Top 10 overvalued and undervalued players per role.

**Findings:**

* Helped spotlight potential market inefficiencies.
* Useful for scouting: clubs could identify undervalued talents.
* Overvaluation often tied to reputation, brand, or league inflation.

**6. Role-Based Market Value Modeling (Sixth Analysis)**

**Objective:** Examine how much "positional bias" exists in the market and how well metrics explain value per role.

**Method:**

* Performed multivariate regression for each role:
  + Metrics → market\_value\_m
* Collected intercept and R².

**Visualization:**

* **Bar Chart** of Intercepts: Shows baseline value for each role.
* **Bar Chart** of R²: Measures how well metrics explain value.

**Findings:**

* Forwards had the highest baseline (intercept) → positional premium.
* Midfielders had the best R² → metrics explain their value well.
* Confirms market values some roles more favorably even before stats.

**7. Summary and Final Notes**

**Technologies Used:**

* PostgreSQL for star-schema data warehouse
* SQL Views for data preparation
* Python (pandas, matplotlib, seaborn, statsmodels) for statistical analysis

**Insights Achieved:**

* Confirmed market bias toward younger and attacking players.
* Quantified positional and league-based valuation differences.
* Detected inefficiencies and outliers in market pricing.

**Outcome:** This project not only demonstrates strong data warehousing and analysis skills, but also produces meaningful, interpretable insights for the football industry. The layered use of performance data, statistical modeling, and visualization meets the goal of uncovering **non-trivial patterns** and showing how **technology supports strategic decisions** in sports management.