FIFA 22 Player Analysis

INFO 432 Final Project

Kevin Shi, Richardson Chhin, Kathryn Swatek

1. Introduction + Goals
2. Dataset Overview & Data Cleaning
3. Exploratory Data Analysis (EDA)
4. Hypothesis Testing
5. Dimensional Reduction
6. Clustering Results
7. Data Modeling
8. Key Findings
9. Lessons Learned
10. Q&A

Agenda

- Inspired by our shared passion for soccer
- Timely with the World Cup coming to the US next summer
- Analyzed FIFA 22 dataset

- **Project Goals**:
  - Derive insights into player characteristics
  - Compare elite vs average players
  - Evaluate hypotheses about positions, values, and wages
  - Learn more about top players and our personal favorites

Introduction + Goals

Dataset Overview + Data Cleaning

Dataset is from Kaggle (FIFA 22 complete player dataset)

Original: 19,239 players x 110 features

Cleaned: 17,107 players x 48 features

Feature Groups

Core Metrics & Vitals: Primary indicators of a player's quality and status

- overall: The player's current rating (1-99)
- potential: The player's predicted peak rating (1-99)
- value_eur: Estimated market value in Euros
- wage_eur: Weekly wage in Euros
- age: Player's age in years

Physical & Demographic Attributes: Basic information about the player

- height_cm: Player's height in centimeters
- weight_kg: Player's weight in kilograms
- preferred_foot: Player's dominant foot (Ex: Right, Left)
- work_rate: Player's work effort (Ex: High/Medium)
- body_type: Player's physique (Ex: Lean, Stocky)

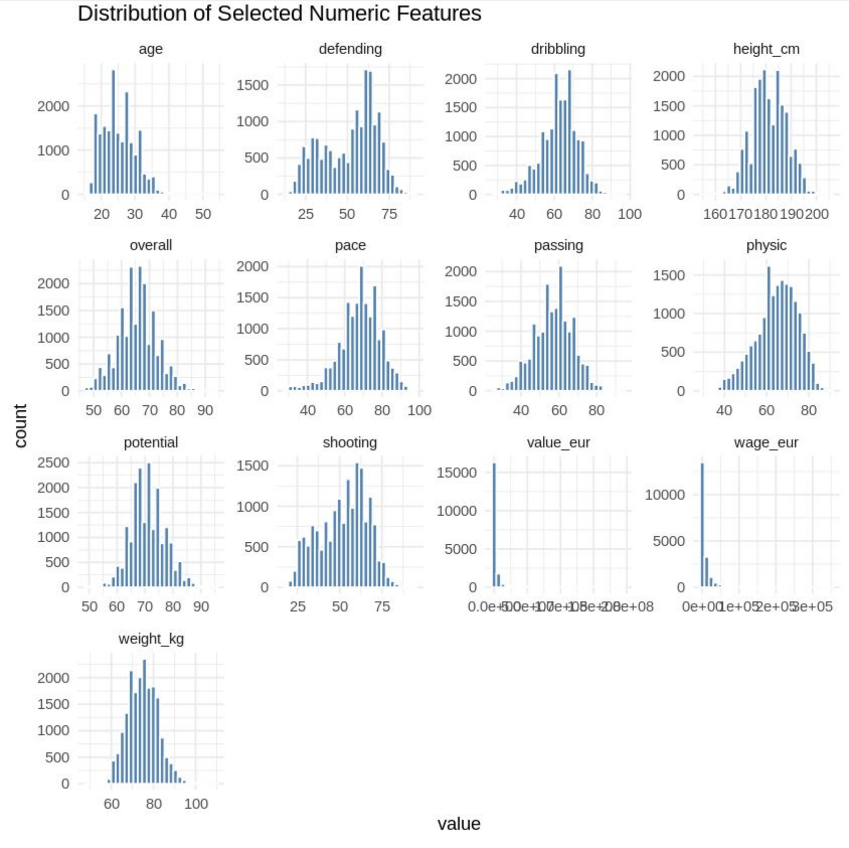
Positional & Summary Skills:

- player_positions: The player's primary listed position(s)
- pace
- shooting
- passing
- dribbling
- defending
- physic

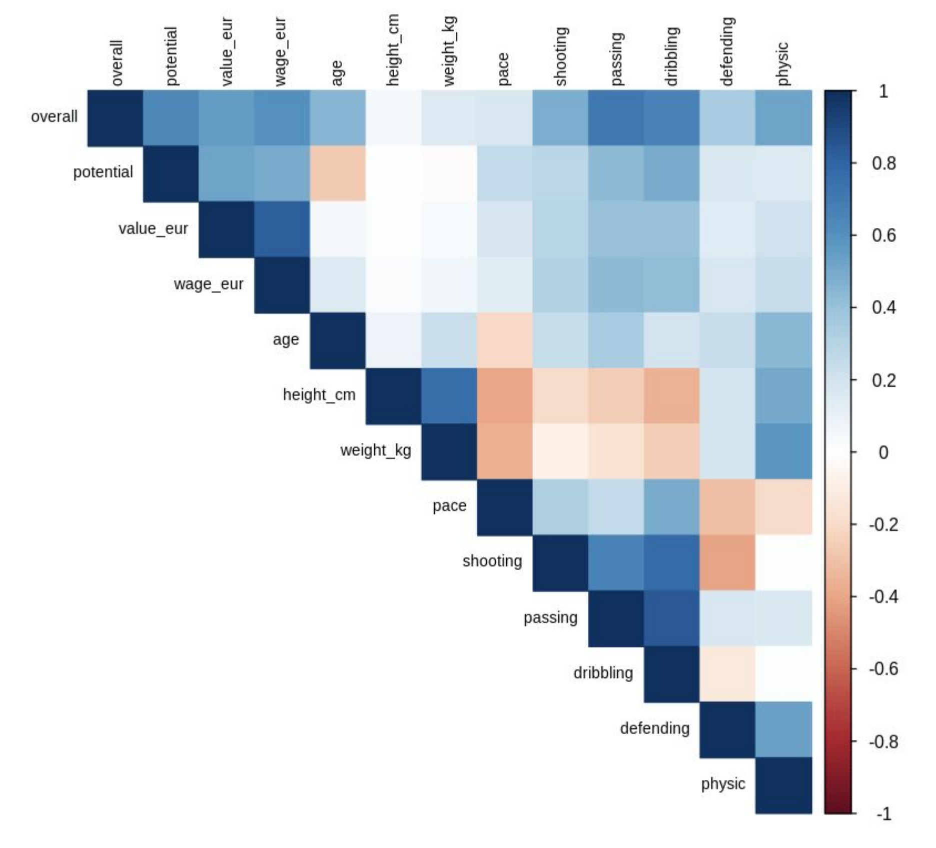
Detailed Skill Attributes: Skill ratings (1-99) that are important to this analysis

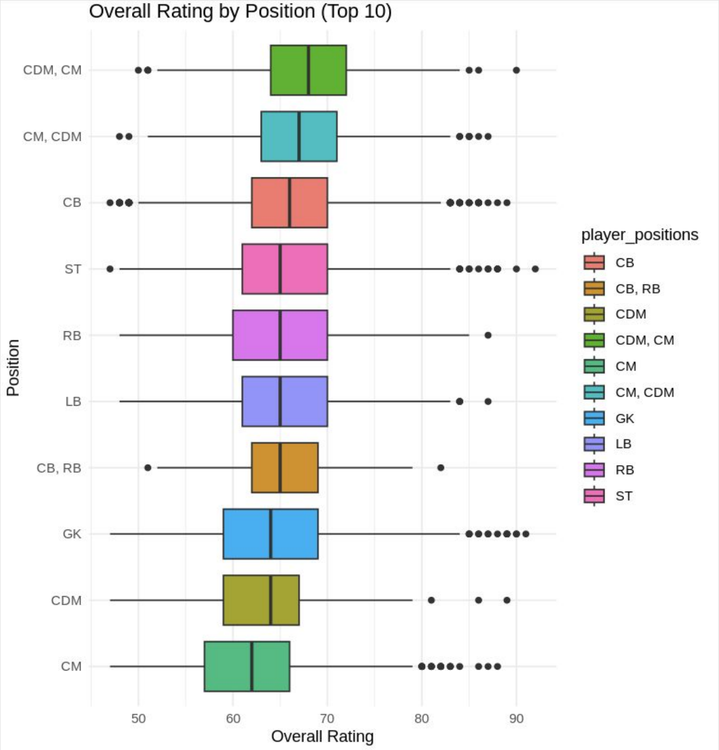
- Attacking: attacking_crossing, attacking_finishing, attacking_heading_accuracy, attacking_short_passing, attacking_volleys
- Skill: skill_dribbling, skill_curve, skill_fk_accuracy, skill_long_passing, skill_ball_control
- Movement: movement_acceleration, movement_sprint_speed, movement_agility, movement_reactions, movement_balance
- Power: power_shot_power, power_jumping, power_stamina, power_strength, power_long_shots
- Mentality: mentality_aggression, mentality_interceptions, mentality_positioning, mentality_vision, mentality_penalties, mentality_composure
- Defending: defending_marking_awareness, defending_standing_tackle, defending_sliding_tackle
- Goalkeeping: goalkeeping_diving, goalkeeping_handling, goalkeeping_kicking, goalkeeping_positioning, goalkeeping_reflexes

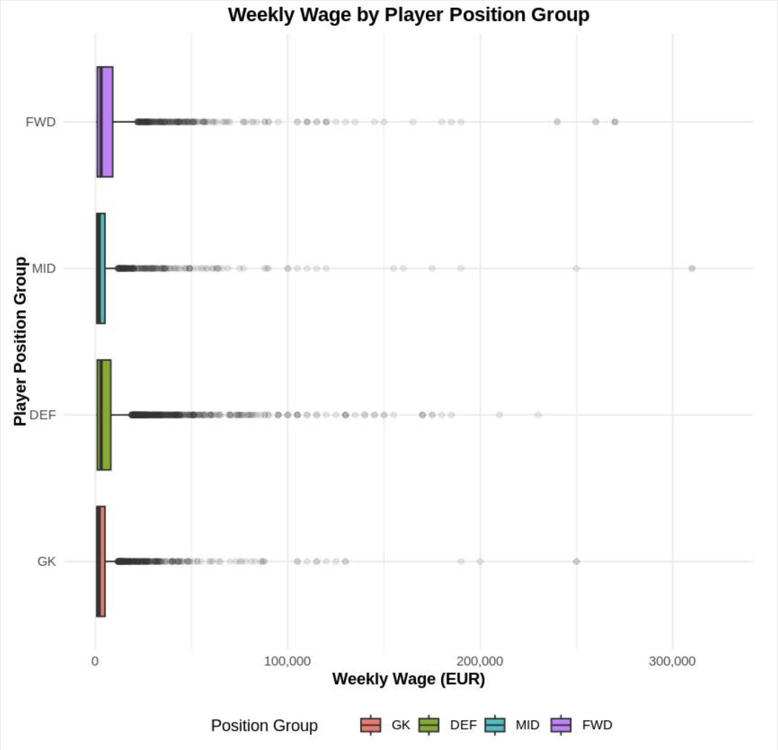
EDA


Distribution of Selected Numeric Features
age
defending
dribbling
height_cm
overall
pace
passing
physic
potential
shooting
value_eur
wage_eur
weight_kg
count
value

EDA


overall
potential
value_eur
wage_eur
age
height_cm
weight_kg
pace
shooting
passing
dribbling
defending
physic
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Overall Rating by Position (Top 10)
CDM, CM
CM, CDM
CB
ST
RB
LB
CB, RB
GK
CDM
CM
Position
50
60
70
80
90
Overall Rating
player_positions
CB
CB, RB
CDM
CDM, CM
CM
CM, CDM
GK
LB
RB
ST


Weekly Wage by Player Position Group
FWD
MID
DEF
GK
Player Position Group
0
100,000
200,000
300,000
Weekly Wage (EUR)
Position Group
GK
DEF
MID
FWD

EDA

Hypotheses

Hypothesis #1:

H0: There is no significant difference in average wage across different player positions (i.e., GK, DEF, MID, FWD).

Ha: At least one player position has a different average wage than the others.

Hypothesis #2:

H0: Preferred foot (left vs. right) is independent of player position.

Ha: Preferred foot is dependent on player position.

Hypothesis #3:

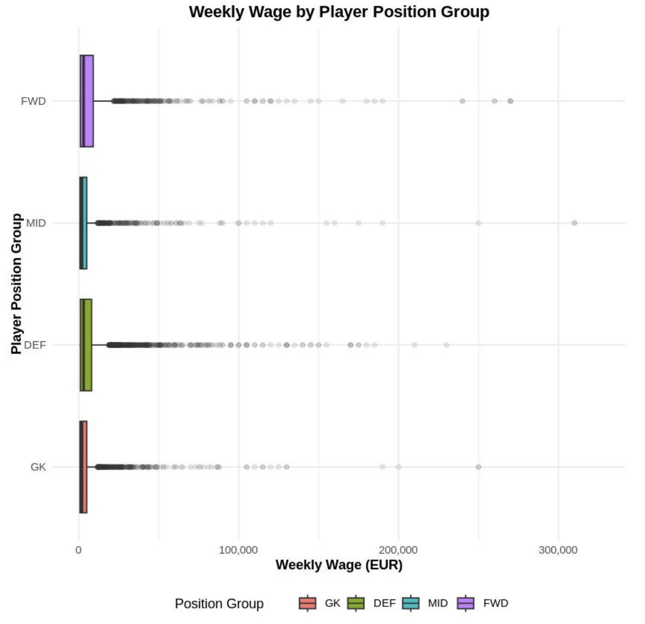
H0: A player’s overall rating is the same regardless of their preferred foot.

Ha: A player’s overall rating is different depending on their preferred foot.

We tested these hypotheses using:

- ANOVA
- Chi-squared test
- Two-sample t-test

H0: There is no significant difference in average wage across different player positions (i.e., GK, DEF, MID, FWD).
Ha: At least one player position has a different average wage than the others.



<u>ANOVA results:</u>

- F-statistic: 17.36
- p-value: 3.10 x 10^-11

<u>Tukey post-hoc results:</u>

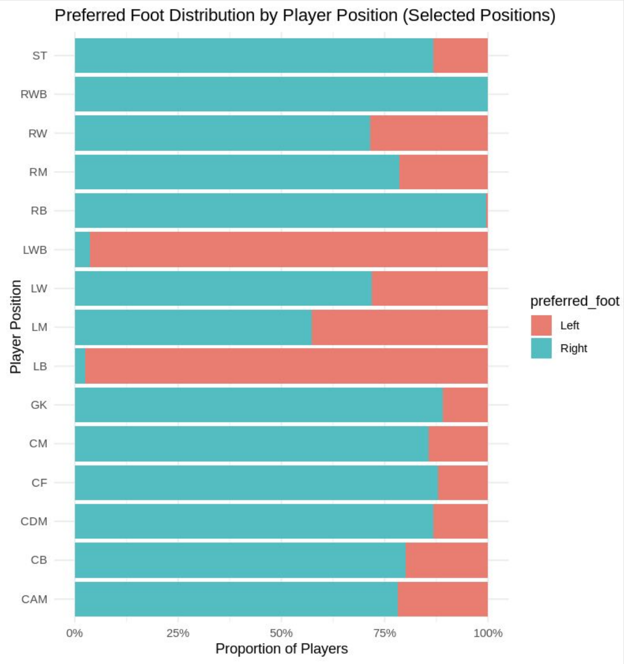
- FWD-DEF: Not statistically significant (p = 0.12)
- GK vs DEF: GK wages significantly lower than DEF (p = 2.2 x 10^-5)
- MID vs DEF: MID wages significantly lower than DEF (p = 4.18 x 10^-4)
- GK-FWD: GK wages significantly lower than FWD (p is basically 0)
- MID-FWD: MID wages significantly lower than FWD (p = 0.0000004)
- MID-GK: Not statistically significant (p=0.994)

Since the p-value=3.10*10^-11 is less than alpha=0.05, we reject the null hypothesis.

Overall, wages vary by position, mainly because forwards earn more and GKs/MIDs earn less compared to certain groups. DEF and FWD are relatively closer in pay.

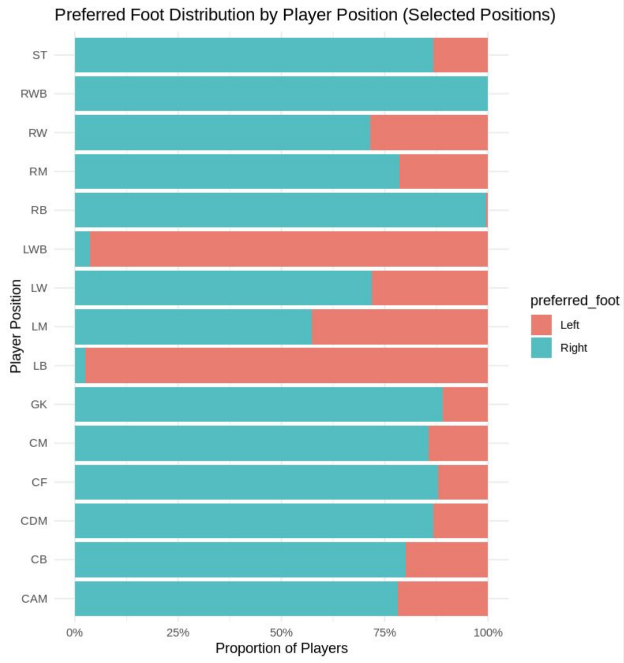
Hypothesis #1

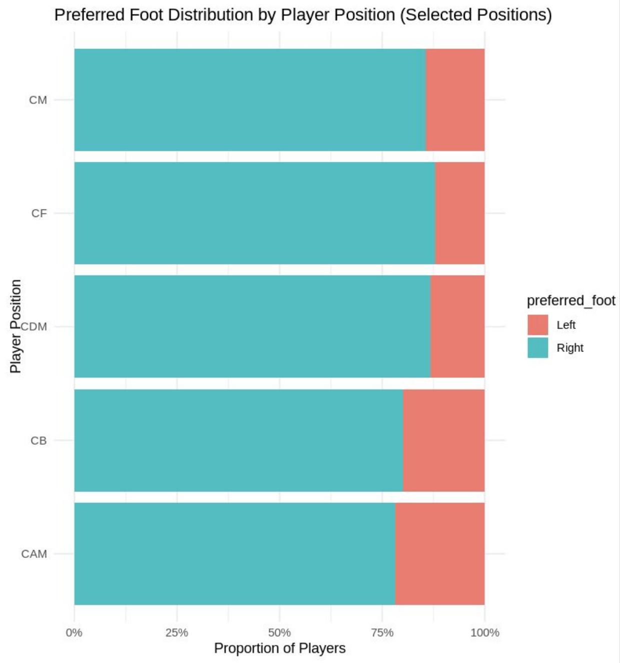
H0: Preferred foot (left vs. right) is independent of player position.
Ha: Preferred foot is dependent on player position.



Hypothesis #2 - Slide 1

H0: Preferred foot (left vs. right) is independent of player position.
Ha: Preferred foot is dependent on player position.





<u>Chi-squared test results:</u>
p-value = 0.007686

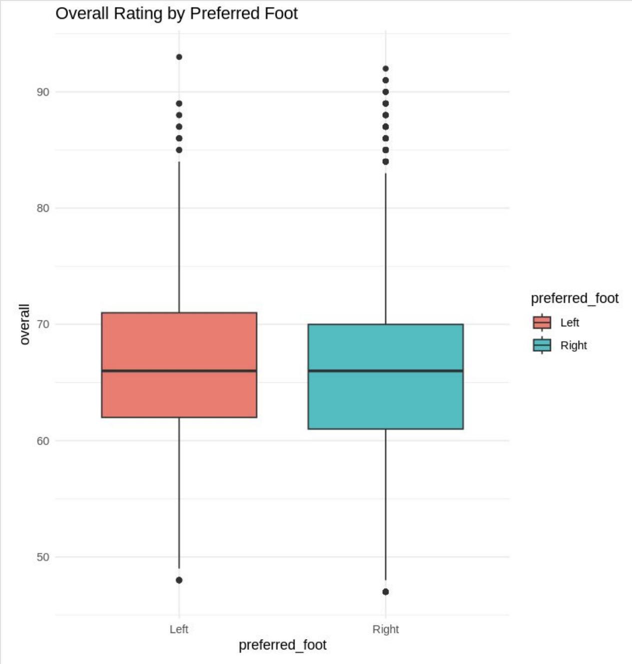
The p-value≈0.0077 is less than alpha=0.05, so we reject H0.

We conclude that there is strong evidence that preferred foot is dependent on the player positions tested.

Certain positions tend to have a higher proportion of left-footed or right-footed players.

Hypothesis #2 - Slide 2

H0: A player’s overall rating is the same regardless of their preferred foot.
Ha: A player’s overall rating is different depending on their preferred foot.



Two Sample t-test results:

- t: 2.26
- p-value: 0.0261
- Mean (Left): 66.52
- Mean (Right): 63.70

Since the p-value=0.0261 is less than alpha=0.05, we reject H0.

We conclude that there is strong evidence that the average overall rating of players with preferred foot of right is not equal to that of players with preferred foot of left.

It seems as though left-footed players have a slightly higher average overall rating than right-footed players.

Hypothesis #3

Principal Component Analysis (PCA):
PC1 (33.0% of variance): "Technical Skill Mastery" - Captures overall technical and offensive skill
PC2 (19.0% of variance): "Defensive vs. Offensive Specialization" - Separates defensive players from offensive players
Finding: PCA is useful for clustering players with similar skill sets

Factor Analysis (FA)
Identified 4 key factors explaining 64.1% of the variance:

1. Technical/Offensive Mastery
2. Defensive Ability
3. Physical Attributes
4. Agility/Movement

Finding: Factor Analysis provides a clearer theoretical interpretation of the underlying player attributes

Conclusion: Both methods are complementary. PCA is better for clustering, while FA offers a more interpretable model

Dimensional Reduction

Goal: To group players into distinct clusters based on their attributes.

Used K-Means clustering on the principal components from our PCA.
The "elbow method" suggested that 4 clusters was the optimal number.

Cluster Profiles:

- Cluster 1: "Elite Forwards/Attackers" - Highest ratings in overall, potential, and offensive skills
- Cluster 2: "Defensive Specialists" - High ratings in defensive and physical attributes
- Cluster 3: "Well-Rounded Midfielders" - Balanced ratings across the board
- Cluster 4: "Developing Young Talent" - Lower overall ratings but high potential

Clustering Results: K-Means

Goal: To build regression models to predict a player's market value and weekly wage based on their attributes

Key Question: For young players (under 24), is potential or overall a stronger predictor of their financial worth?

Findings:

- Current Ability Trumps Potential: For players under 24, their current overall rating is a much stronger predictor of both market value and weekly wage than their potential
- Predicting Value: Our models were extremely effective at predicting market value, explaining about 98% of the variation (Adjusted R-squared of ~0.98)
- Predicting Wages: The models were less effective at predicting wages, explaining about 56-60% of the variation. This suggests that other factors not in our dataset, like league or club wealth, play a significant role in determining wages

Modeling Player Value & Wage

Wages & Position: There's a significant difference in wages across positions. Forwards tend to earn more, while Goalkeepers and Midfielders earn less

Preferred Foot:

- A player's preferred foot is not independent of their position. Certain positions have a higher proportion of left or right footed players
- Left-footed players have a slightly higher average overall rating than right footed players.

Player Attributes: We can boil down player skills into a few key dimensions: technical skill, defensive ability, physical strength, and agility

Player Value: For young players, current ability (overall) is a much better predictor of market value and wage than future potential

Key Findings

Importance of Data Cleaning:

- The cleaned data had 2,132 less players and 62 less features

EDA Guided Hypotheses:

- Statistical tests confirmed/refuted assumptions

Power of Dimensional Reduction:

- PCA and FA both revealed meaningful underlying skill structures

Lessons Learned

Include External Data:

- Club finances, league differences, transfer history, contract details

Test Other Clustering Methods:

- Hierarchical, DBSCAN, or Gaussian

Explore Predictive Modeling:

- Use ML techniques (Random Forest or Naive Bayes)

Future Steps/ With More Time

Q&A