Final Presentation link:
https://docs.google.com/presentation/d/1o7rwn--xvYRLLX1RlcDE4rYXeYjEUwxukOP5f2nPYfk/edit?usp=sharing

INFO 432 Project Report
Project Group 4: Richardson Chhin, Kevin Shi, Kathryn Swatek
Dataset: FIFA 22 complete player dataset
(https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset?select=players_22.csv)

Problem 1: Project Proposal & EDA
This section of the report will describe the statistical hypothesis testing performed as a part of our EDA. To see more EDA (boxplots, correlation matrix, histograms, summary statistics), refer to our code.

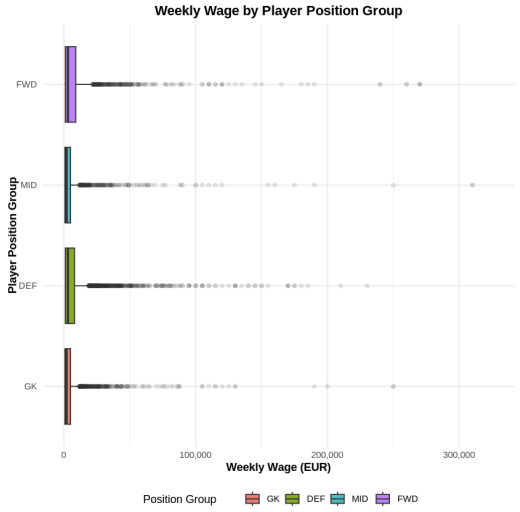
Dataset overview:

Original dataset: 19,239 x 110

Cleaned dataset: 17,107 x 48

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.



term	df	sumsq	meansq	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
position_group	3	1.852968e+10	6176561425	17.35705	3.099546e-11
Residuals	9394	3.342884e+12	355853129	NA	NA

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = wage_eur ~ position_group, data = df_anova)

$position_group
               diff        lwr        upr     p adj
FWD-DEF  1158.7480  -191.5466  2509.0426 0.1218957
GK-DEF  -2384.5071 -3708.0518 -1060.9624 0.0000221
MID-DEF -2228.2100 -3669.8124  -786.6077 0.0004181
GK-FWD  -3543.2551 -5054.4824 -2032.0278 0.0000000
MID-FWD -3386.9580 -5002.5862 -1771.3298 0.0000004
MID-GK    156.2971 -1437.0421  1749.6362 0.9943832
```

ANOVA results:

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

Tukey post-hoc results:

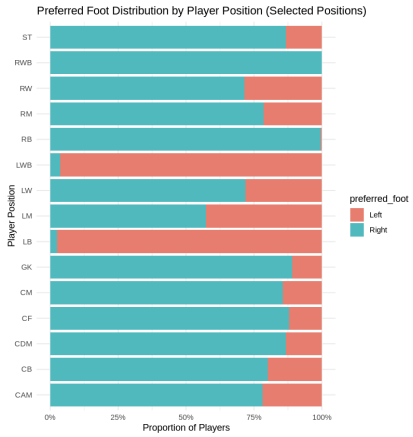
- FWD-DEF: Not statistically significant (p = 0.12)
- GK vs DEF: GK wages significantly lower than DEF (p = 2.2 x 10^-5)
- MID vs DEF: MID wages significantly lower than DEF (p = 4.18 x 10^-4)
- GK-FWD: GK wages significantly lower than FWD (p is basically 0)
- MID-FWD: MID wages significantly lower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

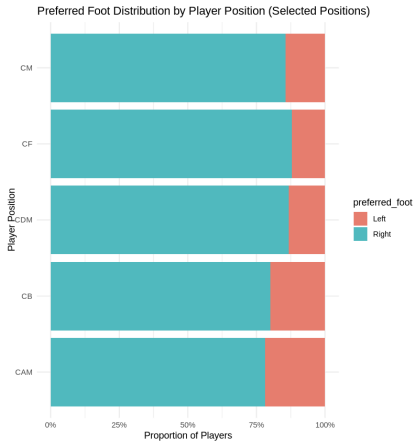
- MID-GK: Not statistically significant (p=0.994)

Since the p-value=3.10*10^-11 is less than alpha=0.05, we reject the null hypothesis. We conclude that there is strong evidence that at least one player position has a different average wage than the others. 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 #2:

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



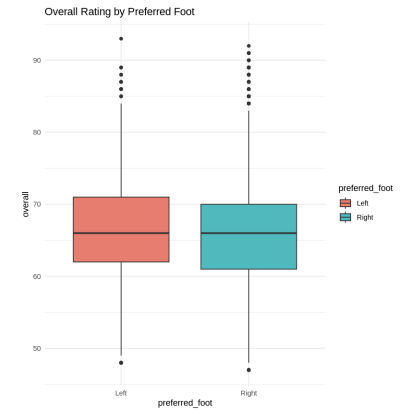


Chi-squared test results: p-value = 0.007686

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 #3:

- H0: The average overall rating of players with preferred foot of right is equal to that of players with preferred foot of left.
- Ha: The average overall rating of players with preferred foot of right is not equal to that of players with preferred foot of left.



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.

Problem 2: Dimensional Reduction

Choose 2 out of these 3 methods (PCA, CCA, Factor Analysis) and find the 'best fitting' model for the data. Explain which original features contribute to the new low-dimensional features and what these new features may be capturing about the dataset.

Principal Component Analysis (PCA):
Variance Explained:
PC1: 33.0% variance, PC2: 19.0% variance
First two components: 52.0% of total variance
First 10 components: 79.7% of total variance

PC1 (33.0%) - "Technical Skill Mastery" Top contributing features:
Dribbling (-0.239), Skill Dribbling (-0.228), Ball Control (-0.225)
Passing (-0.223), Shooting (-0.222), Vision (-0.219)
Long Shots (-0.217), Positioning (-0.216), Curve (-0.213), Finishing (-0.204)

What it captures: Overall technical ability and offensive skill. Higher PC1 scores indicate players with superior ball-handling, decision-making, and attacking capabilities.

PC2 (19.0%) - "Defensive vs Offensive Specialization" Top contributing features:
Defending (0.290), Interceptions (0.279), Marking Awareness (0.278)
Standing Tackle (0.271), Physique (0.267), Sliding Tackle (0.263)
Aggression (0.263), Strength (0.220), Heading Accuracy (0.204), Overall (0.189)

What it captures: Separates defensive specialists (positive scores) from offensive specialists (negative scores). This dimension captures playing style rather than skill level.

Factor Analysis (FA)

Model Performance:

- **4 factors** explain 64.1% of total variance
- **Excellent model fit**: RMSR = 0.03

Factor 1 (PA1 - 30.5%) - "Technical/Offensive Mastery" High loadings (>0.70):

- Ball Control (0.88), Dribbling (0.89), Passing (0.92), Shooting (0.87)
- Vision (0.88), Positioning (0.81), Curve (0.84), Long Shots (0.87)
- Short Passing (0.80), Finishing (0.78), FK Accuracy (0.78)

What it captures: Core attacking and technical skills essential for creative and offensive players.

Factor 2 (PA2 - 14.6%) - "Defensive Ability" High loadings (>0.90):

- Defending (0.95), Standing Tackle (0.94), Sliding Tackle (0.94)
- Marking Awareness (0.92), Interceptions (0.93)

What it captures: Pure defensive capabilities and tactical awareness for defensive specialists.

Factor 3 (PA3 - 10.9%) - "Physical Attributes" High loadings:

- Pace (0.93), Sprint Speed (0.88), Acceleration (0.86), Strength (0.89)
- Physique (0.84), Height (0.70), Weight (0.74), Heading Accuracy (0.78)

What it captures: Physical prowess, speed, and size characteristics that define athletic ability.

Factor 4 (PA4 - 8.0%) - "Agility/Movement" Key loadings:

- Agility (0.57), Balance (0.40), various movement attributes

What it captures: Fine motor skills, mobility, and movement control for technical maneuverability.

Best Fitting Model Determination

PCA: Superior for variance retention (79.7% with 10 components) and clear two-dimensional interpretation **Factor Analysis**: Better for theoretical interpretation with distinct, meaningful skill domains

Conclusion: Both methods are complementary - PCA is optimal for clustering applications due to variance retention, while Factor Analysis provides a cleaner theoretical understanding of latent player skill constructs. The consistent identification of technical vs. defensive dimensions across both methods validates the underlying structure of football player abilities.

Problem 3: Fitting Unsupervised Clustering

Fit two clustering models of your choosing to both the original variables and to the low-dimensional representations you learned in the prior problem. Which feature sets provide the best clustering structure? How easy are the clustering results to explain? Develop a two-page narrative about your experience in working with this data set, and if you were able to confirm or refute any of the hypotheses developed in problem 1.

3-1: Fitting Unsupervised Clustering

We will detail the applications of k-means clustering to the original player data and the low-dimensional representations derived from PCA and Factor Analysis. The goal is to determine which feature set provides the most meaningful and well-structured clusters corresponding to real world player positions.

To determine the best clustering structure, we evaluated the models based on their Silhouette and Cluster Purity scores. The results clearly indicate that the low dimensional feature sets provided a superior clustering structure compared to the original, high dimensional data

Clustering Performance Metrics Table

Feature Set	Average Silhouette Score	Cluster Purity
Original Variables	0.138	N/A
PCA Scores	0.187	0.501
Factor Scores	0.185	0.501

The PCA scores yielded the best clustering structure. It achieved the highest Silhouette Score (0.187), indicating that its 4 clusters were the most dense and well separated. While both PCA and Factor Analysis models achieved an identical Cluster Purity of 50.1%, the superior silhouette score makes the PCA based model the best choice.

Below are the contingency tables for both the PCA and Factor Analysis models:

PCA Clusters vs. Football Positions

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
DEF	1164	264	341	1888
FWD	638	157	203	1002
MID	587	206	175	673

Factor Analysis Clusters vs. Football Positions

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
DEF	679	586	945	1447
FWD	393	311	549	747
MID	353	262	449	577

In both models, the algorithm successfully identified patterns in the data, as shown by the non-random distribution of players. In the PCA model, Cluster 1 appears to have captured a significant portion of defensive players (1164) and attackers (638), suggesting it may represent a "core outfield player" profile.

Cluster 4 is the largest and most mixed group, likely representing generalists or players who didn't fit a highly specialized profile. The algorithm did not produce perfectly clean clusters (Ex: 1 cluster for each position), which is expected given the tactical flexibility of modern players. However, the 50.1% purity score confirms that the underlying skill dimensions are strongly related to a player's primary role on the field

3-2: Modeling Player Value & Wage

We will address our hypothesis that a player's market value and weekly wage are significantly driven by age, potential, overall ability, and latent skill factors. We built a series of multiple linear regression models to test this, using the outputs from both PCA and Factor Analysis as predictors

The results from the 8 regression models are summarized below. Table 1 shows the models built using PCs and Table 2 shows the models built using Factor Scores.

Table 1:

Predictor	Value (All)	Value (<24)	Wage (All)	Wage (<24)
age	0.281	0.336	0.333	1.491
age^2	-0.006	-0.008	-0.006	-0.033
potential	0.048	0.038	0.038	0.039
overall	0.133	0.138	0.097	0.080
PC1	-0.019	-0.011	-0.038	-0.028
PC2	Not sig.	Not sig.	0.029	Not Sign.
Adj R-squared	0.976	0.984	0.602	0.566

Table 2:

Predictor	Value (All)	Value (<24)	Wage (All)	Wage (<24)
age	0.272	0.330	0.320	1.465
age^2	-0.006	-0.008	-0.006	-0.033
potential	0.052	0.039	0.044	0.046
overall	0.139	0.141	0.111	0.088
Factor1	0.029	0.017	0.053	0.041
Factor2	-0.019	-0.010	-0.014	-0.029

Adj R-squared	0.975	0.984	0.599	0.563

Predictor Analysis

Across the models, age, potential, and overall were highly significant predictors of both value and wage. The squared term for age consistently had a negative coefficient, correctly modeling that a player's value and wage potential peak and then decline with age. The latent factors from both PCA and Factor Analysis were also significant, though their impact was smaller compared to the core metrics.

Hypothesis Validation: Potential vs. Overall for Younger Players

Our initial hypothesis stated that for younger players (under 24), potential would be a stronger driver of value and wages than their current overall ability. The model results refute this hypothesis.

- For Market Value (<24): The coefficient for overall (0.138) is substantially larger than the coefficient for potential (0.038)
- For Weekly Wage (<24): The coefficient for overall (0.080) is more than double the coefficient for potential (0.039)

This surprising result was consistent across both PCA and Factor Analysis models, which indicates that when determining financial worth, the market values a young player's current demonstrated skill far more heavily than their projected future talent

PCA vs. FA Model Comparison & Overall Performance

Both sets of models performed nearly identically, confirming that either dimensional reduction technique is valid for this analysis.

- Predicting Value: The models are extremely effective at predicting a player's market value, with Adjusted R-squared values around 0.975-0.984. This means the models can explain approximately 98% of the variation in player value, which is an exceptionally strong fit
- Predicting Wages: The models are less effective at predicting wages, with Adjusted R-squared values around 0.56-0.60. While this is still a moderately strong fit, it suggests that nearly 40% of the variation in wages is due to factors not included in this dataset, such as the specific league, club wealth, or individual contract negotiation