ANALYSIS OF INVESTMENT OPPORTUNITIES IN EUROPEAN SOCCER LEAGUES

An In-depth Study Using FIFA 2022 Dataset

GROUP 5

BAUDOT, CAQUILALA, DINESH, GUPTA, THAKUR

AGENDA

- Data Source and Population
- EDA
- Literature Review
- Clustering Analysis & Classification
- Study Objectives
- Conclusion

DATA SOURCE AND POPULATION

SCENARIO & STUDY OBJECTIVES

Scenario: An investor in soccer teams wants to know which of the top 4 leagues in Europe (Spain, England, Italy, France) to invest in. They would like to analyze the performance of teams based on statistics recorded and make a decision on the league to invest in.

Study Objectives:

1. Using clustering, are players grouped based on their overall ratings, skill ratings, physical characteristics, and wages? What is the distribution across the clusters in terms of league, club, position, nationality, and reputation?
2. Can we identify the team with the best worth in each league for an affordable investment?
3. Predict the wage of a player in each league based on their abilities, physical characteristics, and nationality.
4. Can we determine the international player's reputation from their contribution to their club?

DATA SOURCE AND POPULATION

- The FIFA dataset was obtained from Kaggle (https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset).

- The original dataset comprises data from 2014 to 2022, with every file ranging in rows of 16155 and 110 columns.

- The population is made up of 19,239 male soccer players from 55 soccer leagues, with player data taken from the Career Mode of the FIFA 22 computer game. For the sample dataset, 880 players from the 80 clubs in the 1st level leagues in England, France, Italy, and Spain for season 2022 were considered. All players from the FIFA game are selected except the Substitutes and Reserves. Each team consists of 11 players playing eleven for their respective team.

- The new column created "attack_position" is from the club_position variable in the dataset and ranges in 4 values Attack, Midfield, Defense, GoalKeeper.

VARIABLES AND POSSIBLE BIAS

CATEGORICAL VARIABLES

Variable	Measure	Description
short_name	character	Player's name
league_name	4 leagues: England, France, Italy, Spain	League the player's club team competes in
club_name	~20 clubs per league	Club team the player is currently playing for
club_position	25 different positions categorized into Attack, Midfield, Defense, and Goalkeeper	Player's position in the club’s formation
nationality_name	character	Nationality of the player
international_reputation	rating 1-5	Player's international reputation rating

NUMERICAL VARIABLES

Variable	Measure	Description
age	years	Player's age
height_cm	centimeters	Player's height
weight_kg	kilograms	Player's weight
value_eur	amount in Euros	Player's market value
wage_eur	amount in Euros	Player's monthly wage
overall	rating 1-100	Overall rating of the player's skills and performance
potential	rating 1-100	Potential rating representing the maximum overall rating the player can achieve
pace, shooting, passing, dribbling, defending, physic	rating 1-100	Rating of player's individual abilities

POSSIBLE BIAS

- Sampling Bias: By focusing on the 880 players from the first-level leagues in England, France, Italy, and Spain, the analysis might not account for emerging talents or high-performing players in lower leagues or other countries. This could impact the generalizability of the findings.

- Game-Based Ratings: SoFIFA's subjective player ratings may not fully match real-world performances.

- Market Value Misalignment: In-game values and wages might not reflect real market conditions accurately.

- Potential Bias from Overlooking Substitutes and Reserves: By excluding substitutes and reserves, the dataset misses out on capturing the complete strength and strategy of a team, as these players contribute significantly to performance, injury management, and match tactics. This could lead to an incomplete understanding of team capabilities.

POSSIBLE BIAS

01

Sampling Bias:

Focuses only on players from top-tier leagues in select countries, potentially missing talents from lower leagues or other nations.

02

Market Value Misalignment:

In-game values and wages might not accurately reflect actual market conditions and player worth.

03

Game-Based Ratings:

Relies on subjective player ratings from SoFIFA, which may not fully align with real-world performances.

04

Potential Bias from Overlooking Substitutes and Reserves:

Excludes substitutes and reserves, potentially overlooking their significant contributions to team performance and strategy.

LITERATURE REVIEW

CLUSTERING OF FOOTBALL PLAYERS BASED ON PERFORMANCE DATA AND AGGREGATED CLUSTERING VALIDITY INDEXES

Authors: Serhat Emre Akhanli and Christian Hennig
Published: May 2023

Clustering Objectives:

Analyze data containing 1501 football players from the 2014-2015 season in 8 major league (with 107 variables):

1. Partition data set into major groups of different, easily interpretable player types ***(low k)***
2. Partition data set into many small clusters of players with very similar profiles ***(high k)***

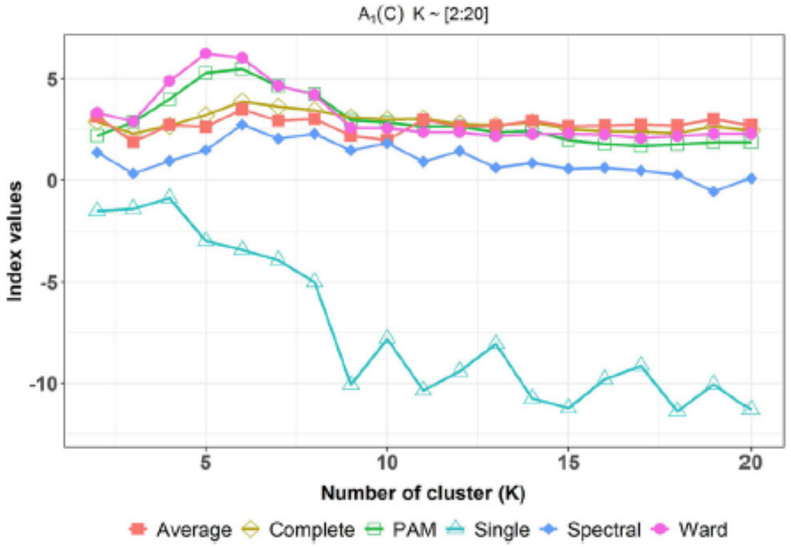
Methods:

1. **Clustering methods**: k-medoids, hierarchical (single, average, complete, Ward's linkage), spectral
2. **Validity index**: Define a suitable validation index as a weighted average of calibrated individual indexes measuring the desirable features
    - separation, average within-cluster dissimilarities, Pearson, entropy, stability

CLUSTERING OF FOOTBALL PLAYERS BASED ON PERFORMANCE DATA AND AGGREGATED CLUSTERING VALIDITY INDEXES

Key Findings:

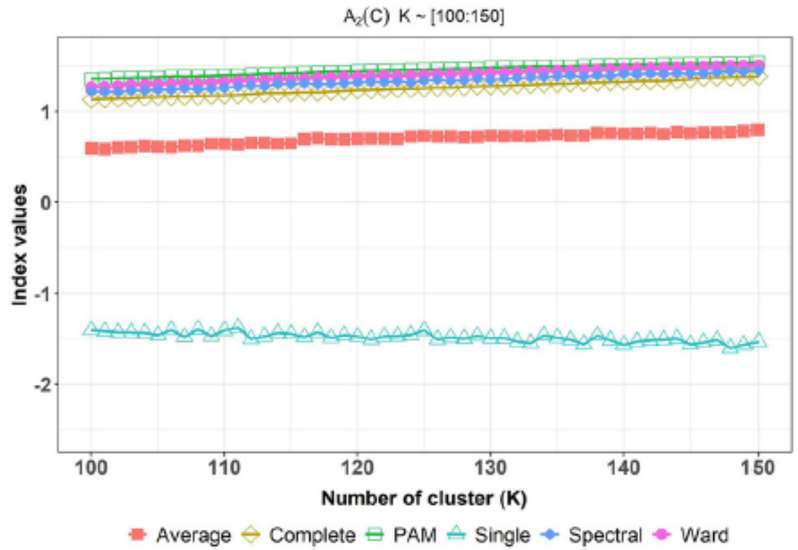
Objective 1 - inherent grouping structure (k=5)



Cluster predominantly based on a combination of position and technical skills:

Cluster 1: midfielders / tackles and short passes
Cluster 2: full backs / blocks and cross passes
Cluster 3: center backs / defence
Cluster 4: attacking midfielders / dribbles, assists, crosses
Cluster 5: forwards / shots and goals

Objective 2 - smaller homogeneous clusters (k=150)



Cluster predominantly based on a combination of position:

Cluster 127: Messi, Ronaldo, Neymar, Robben (forwards)
Cluster 12: defenders
Cluster 11: strikers
Cluster 7: midfielders

EDA

MISSING VALUES

Pace	Shooting	Passing	Dribbling	Defending	Physic
80	80	80	80	80	80

Missing values for Goal Keepers only

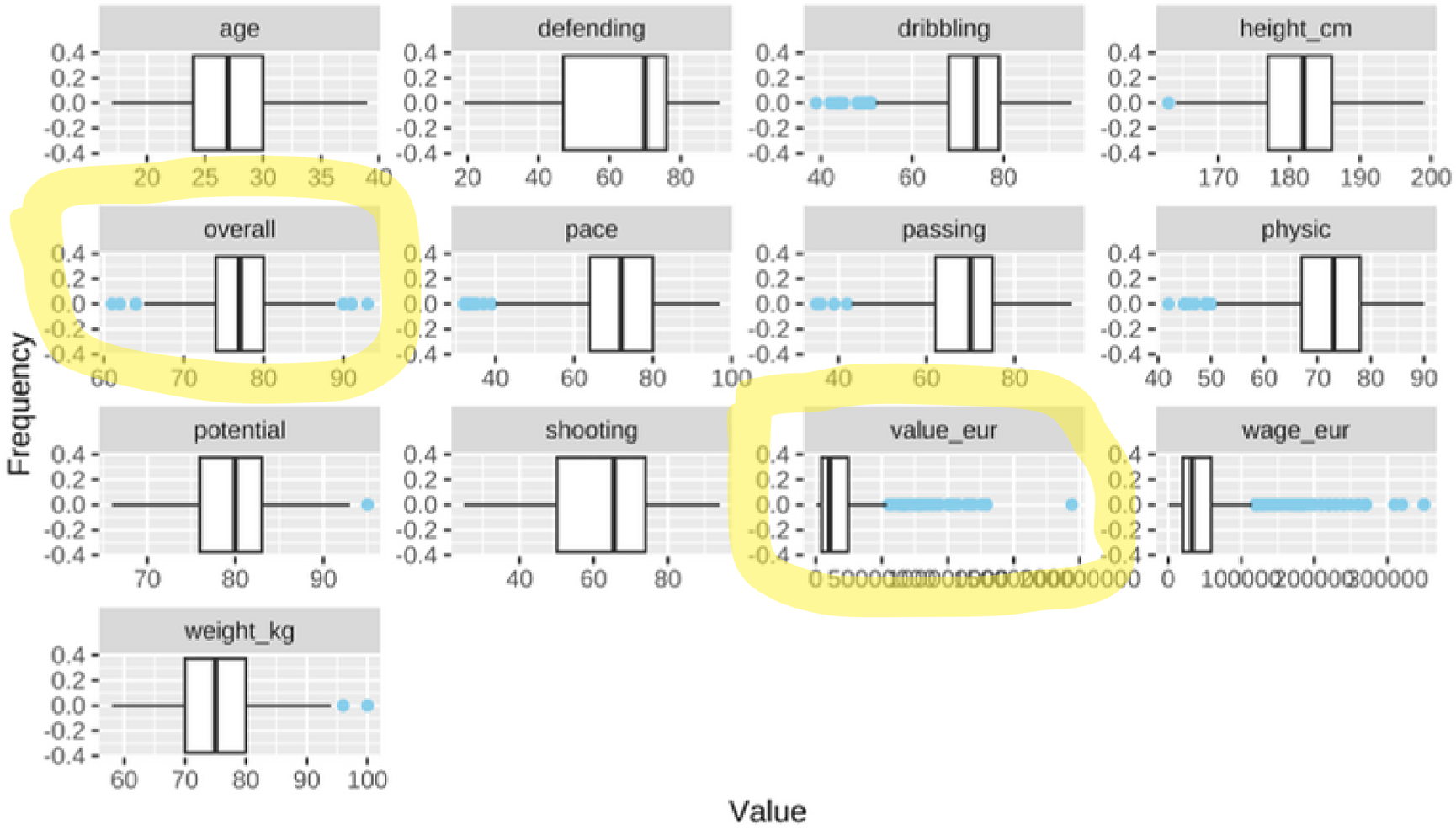
Including GK

exclude individual skills variables

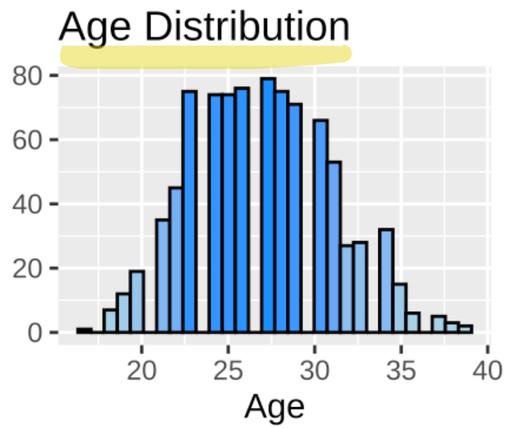
Excluding GK

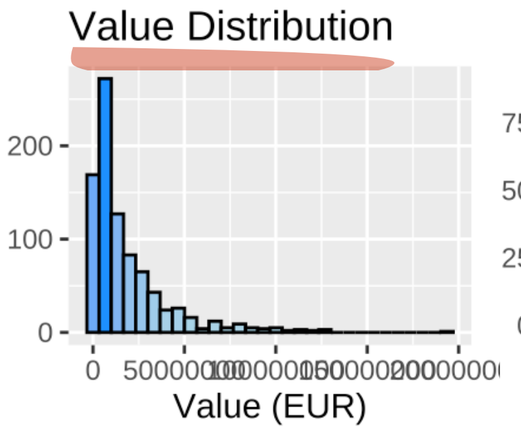
include individual skills variables

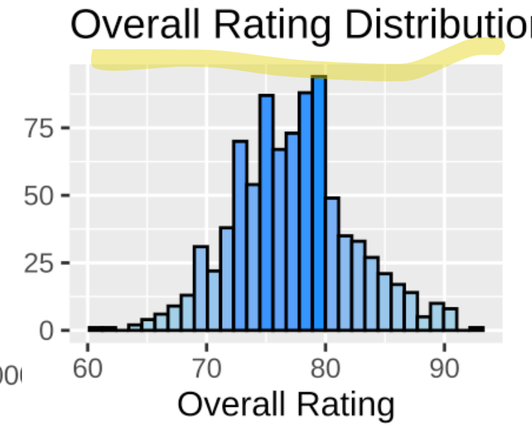
OUTLIERS

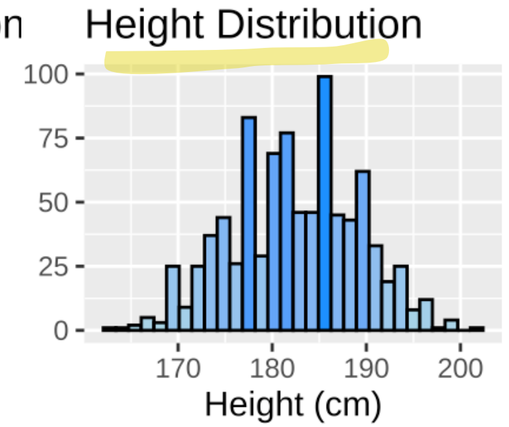

age
defending
dribbling
height_cm
overall
pace
passing
physic
potential
shooting
value_eur
wage_eur
weight_kg
Frequency
Value

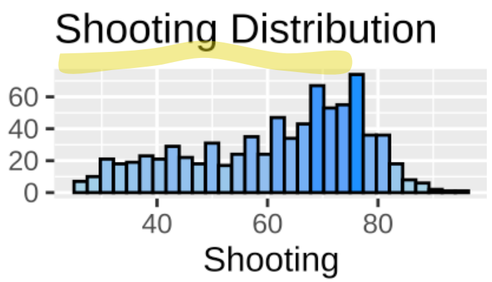
HISTOGRAM

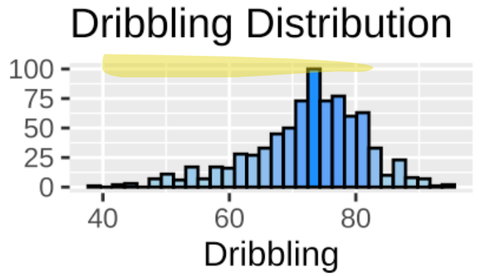

Age Distribution
Age

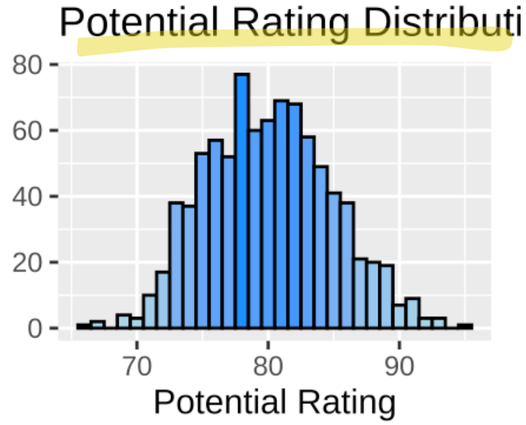

Value Distribution
Value (EUR)

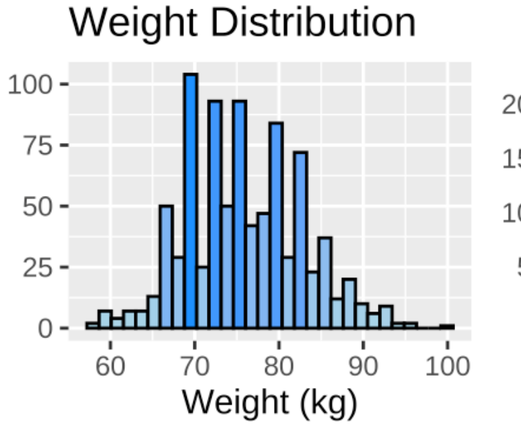

Overall Rating Distribution
Overall Rating

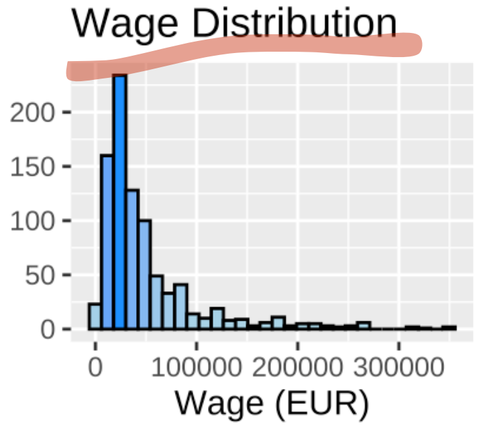

Height Distribution
Height (cm)

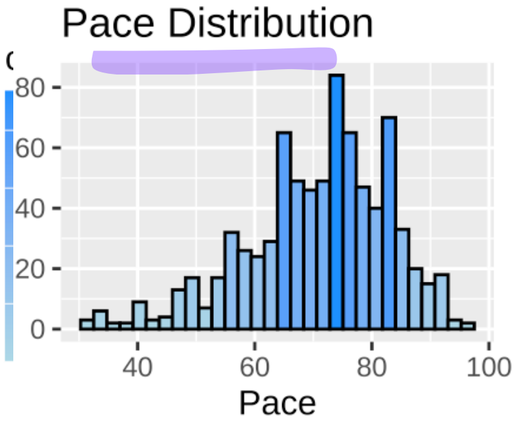

Shooting Distribution
Shooting

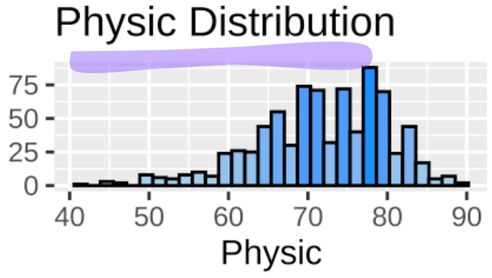

Dribbling Distribution
Dribbling

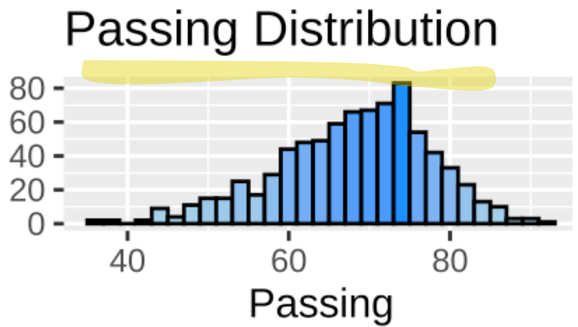

Potential Rating Distributi
Potential Rating

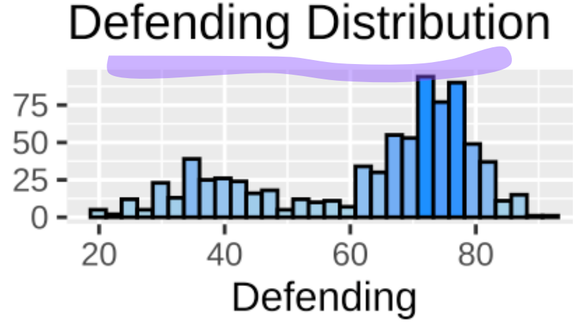

Weight Distribution
Weight (kg)


Wage Distribution
Wage (EUR)

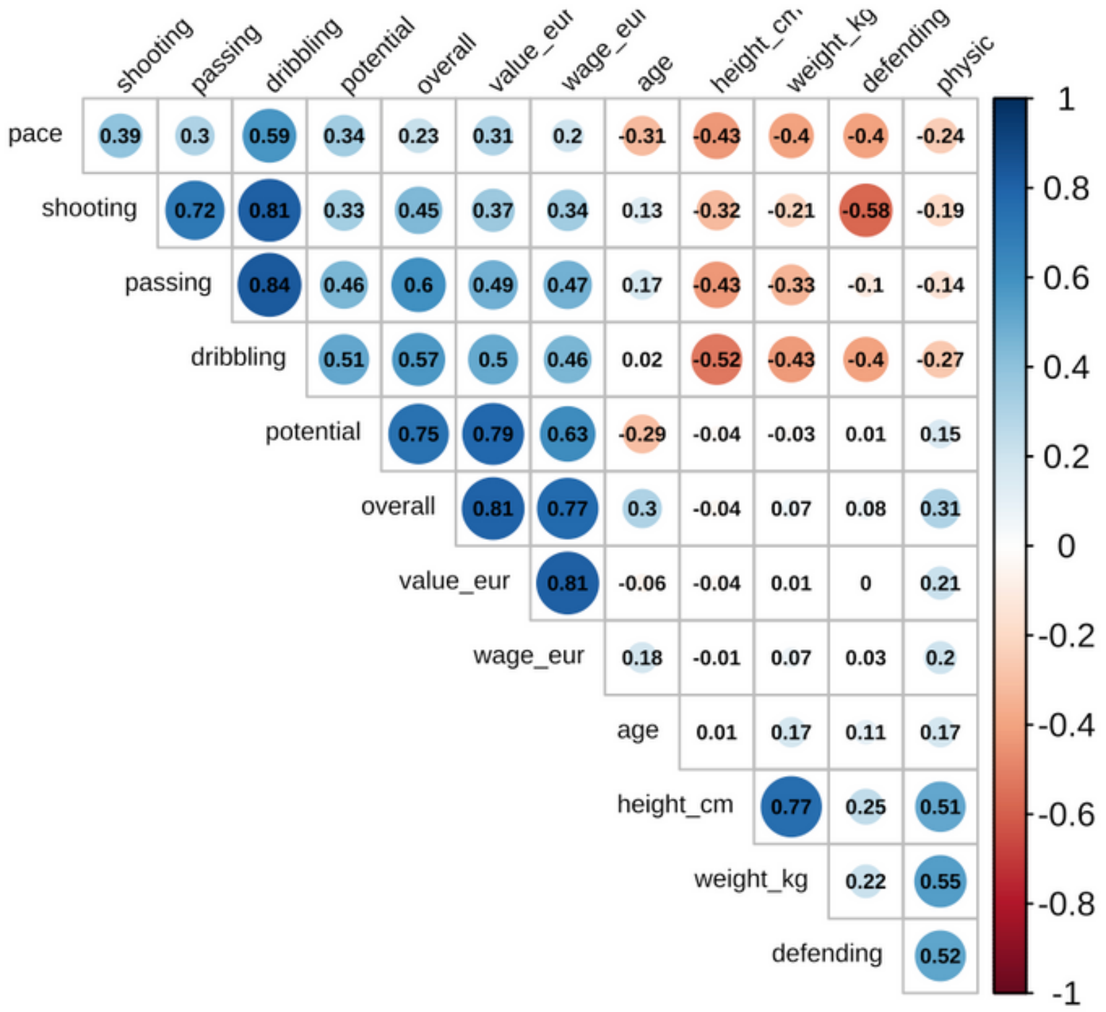

Pace Distribution
Pace


Physic Distribution
Physic

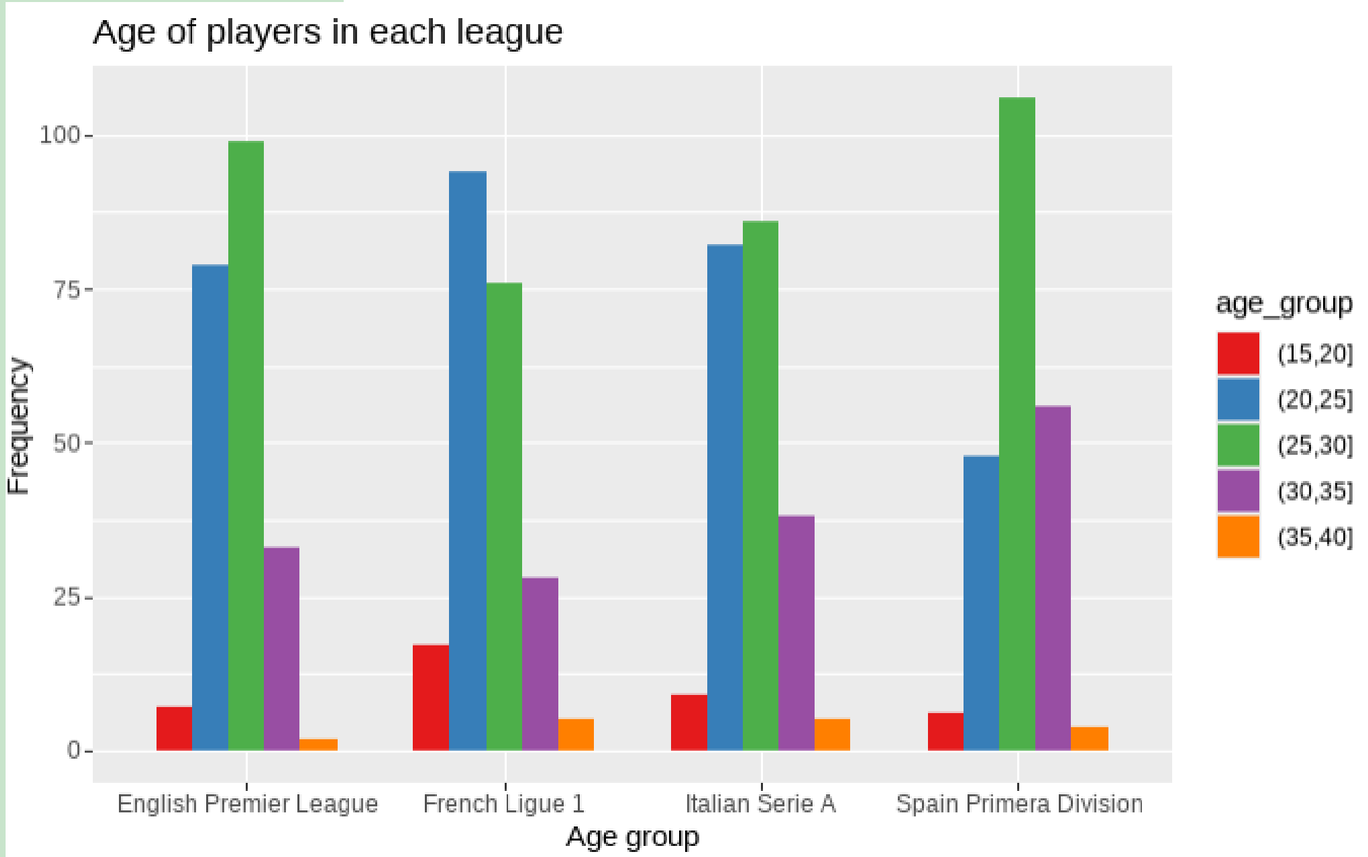

Passing Distribution
Passing


Defending Distribution
Defending

CORRELATION

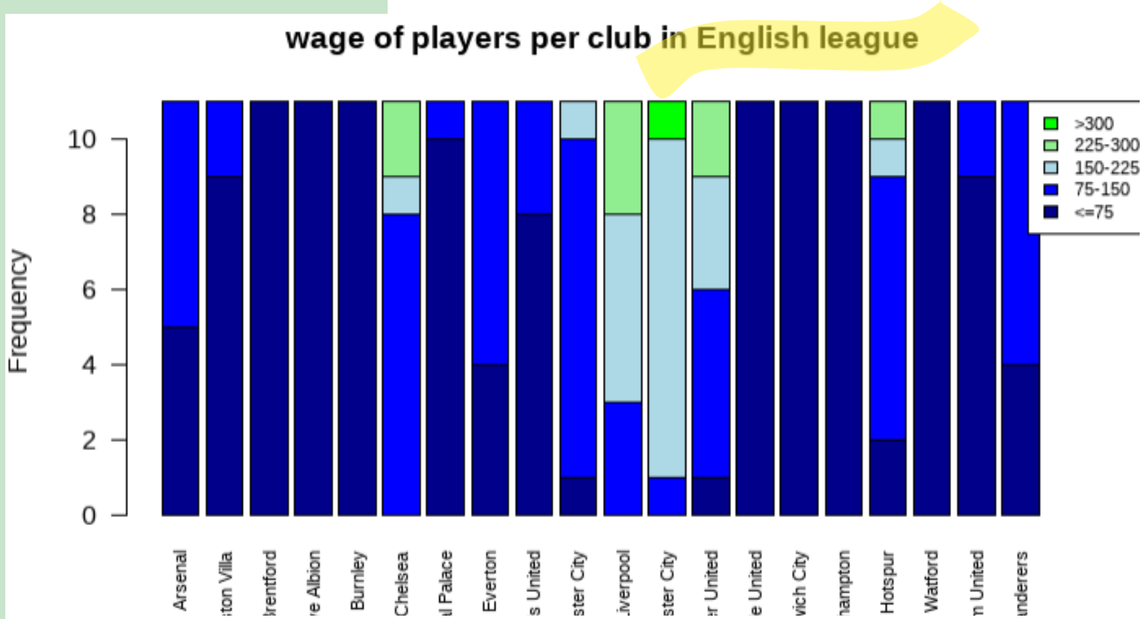


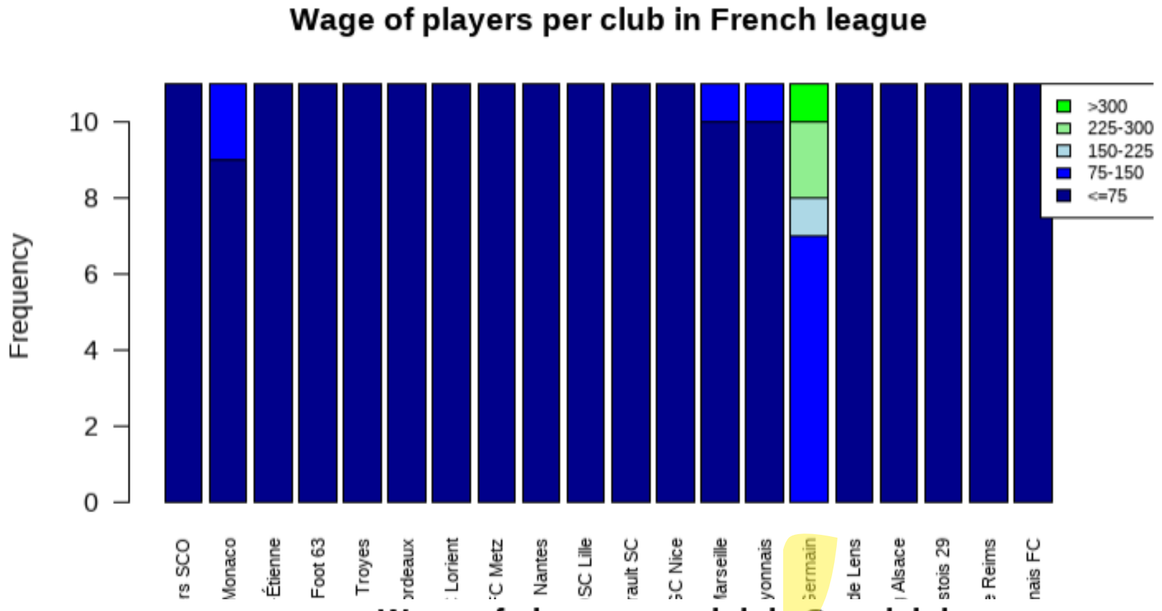
FREQUENCY OF AGE GROUPS PER LEAGUE

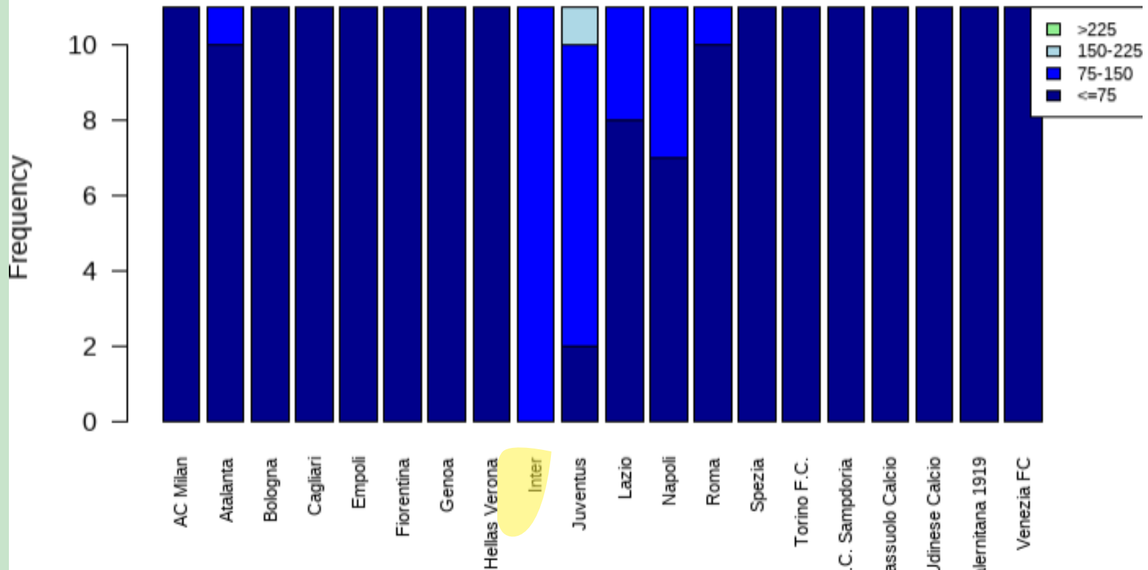


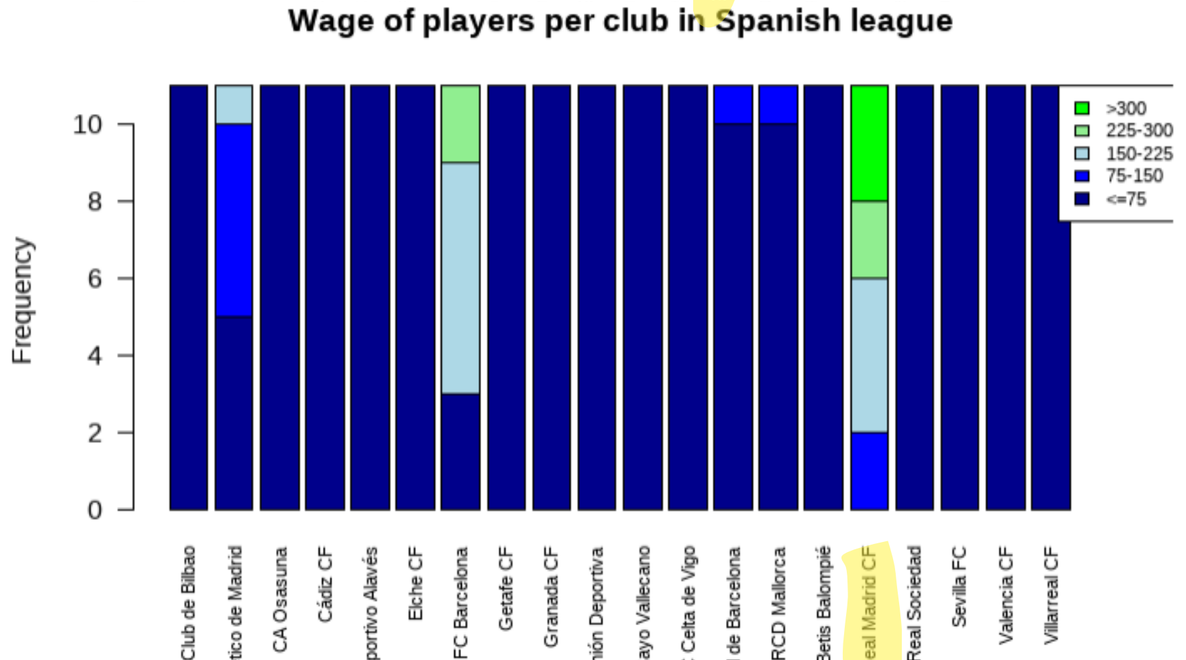
Younger players in French league, Older in Spanish league

SALARY RANGE PER CLUB IN EACH LEAGUE

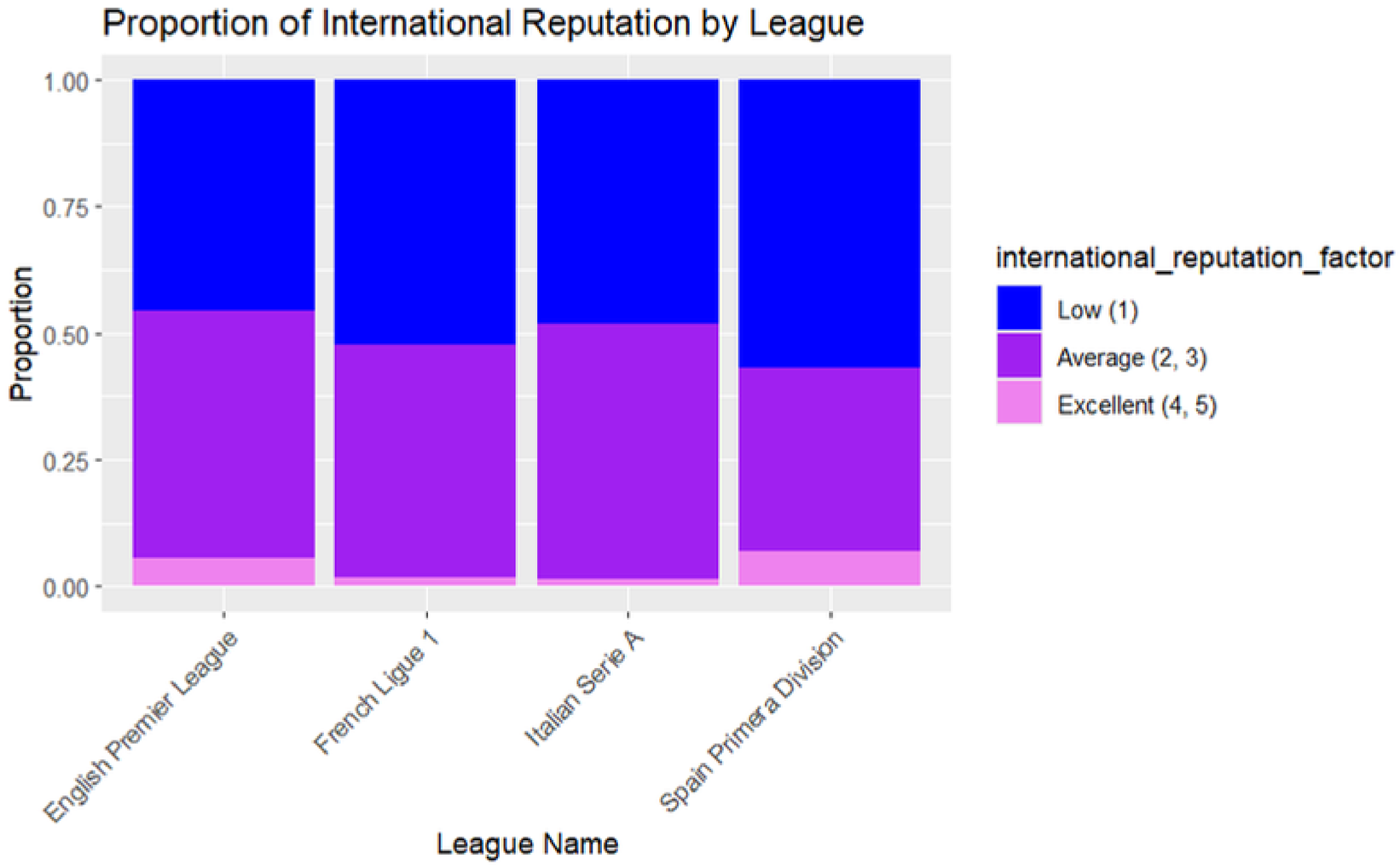






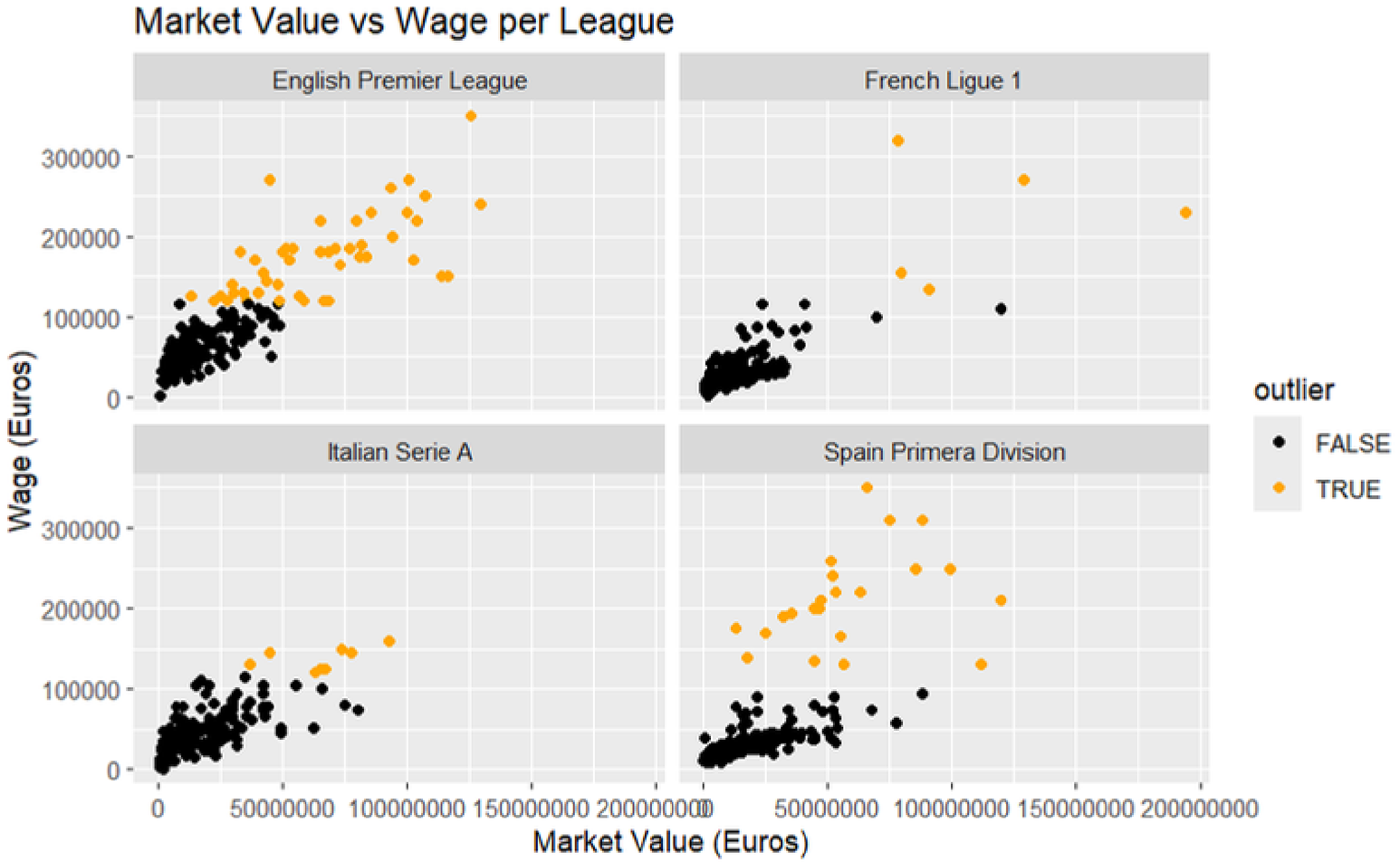


PROPORTION OF INTERNATIONAL REPUTATION BY LEAGUE



Player Reputation Skew Predominantly 'Low' across leagues; 'Excellent' is rare

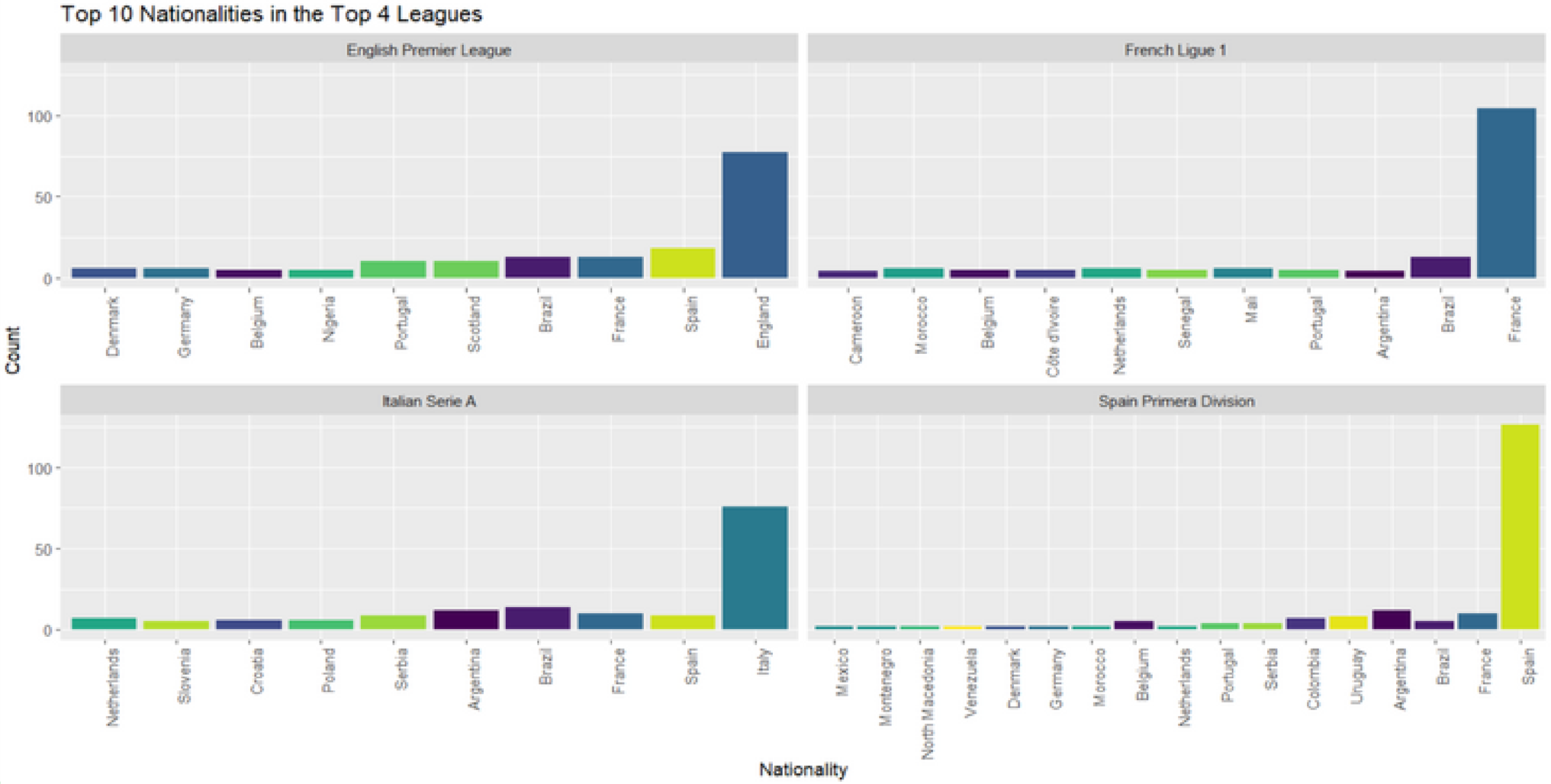
RELATIONSHIP BETWEEN PLAYERS' MARKET VALUE AND WAGES WITHIN THE LEAGUES



Correlation and Outliers in Player Valuation and Wages

Strong market value-to-wage correlation, esp in **English Premier League,** with notable wage outliers indicating additional factors at

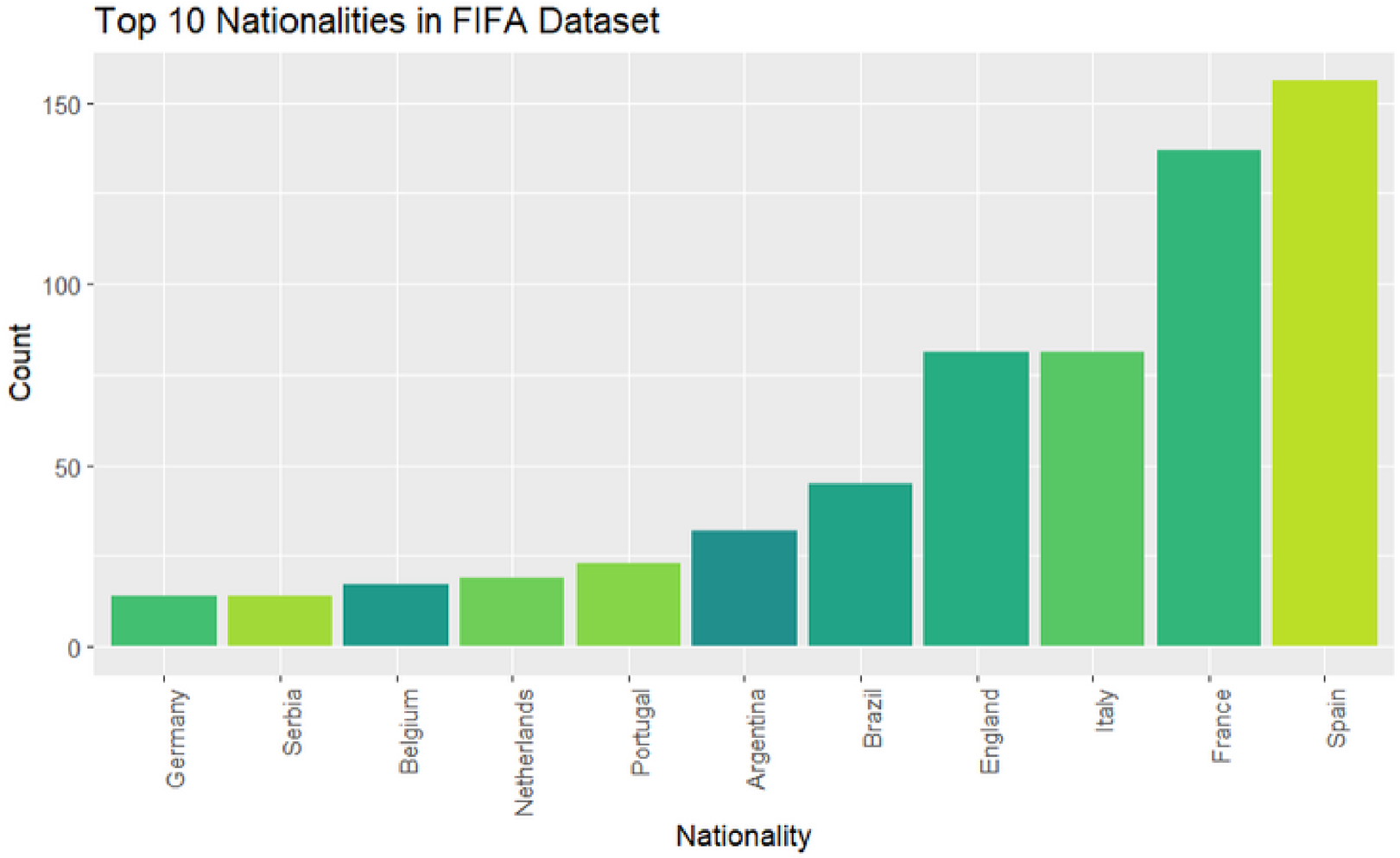
NATIONALITY IN EACH 4 LEAGUES: ENGLAND, FRANCE, ITALY, SPAIN



Top Leagues' Nationality Mix

Data highlights a blend of local and international players, with a high presence of **Spanish and French** nationals.

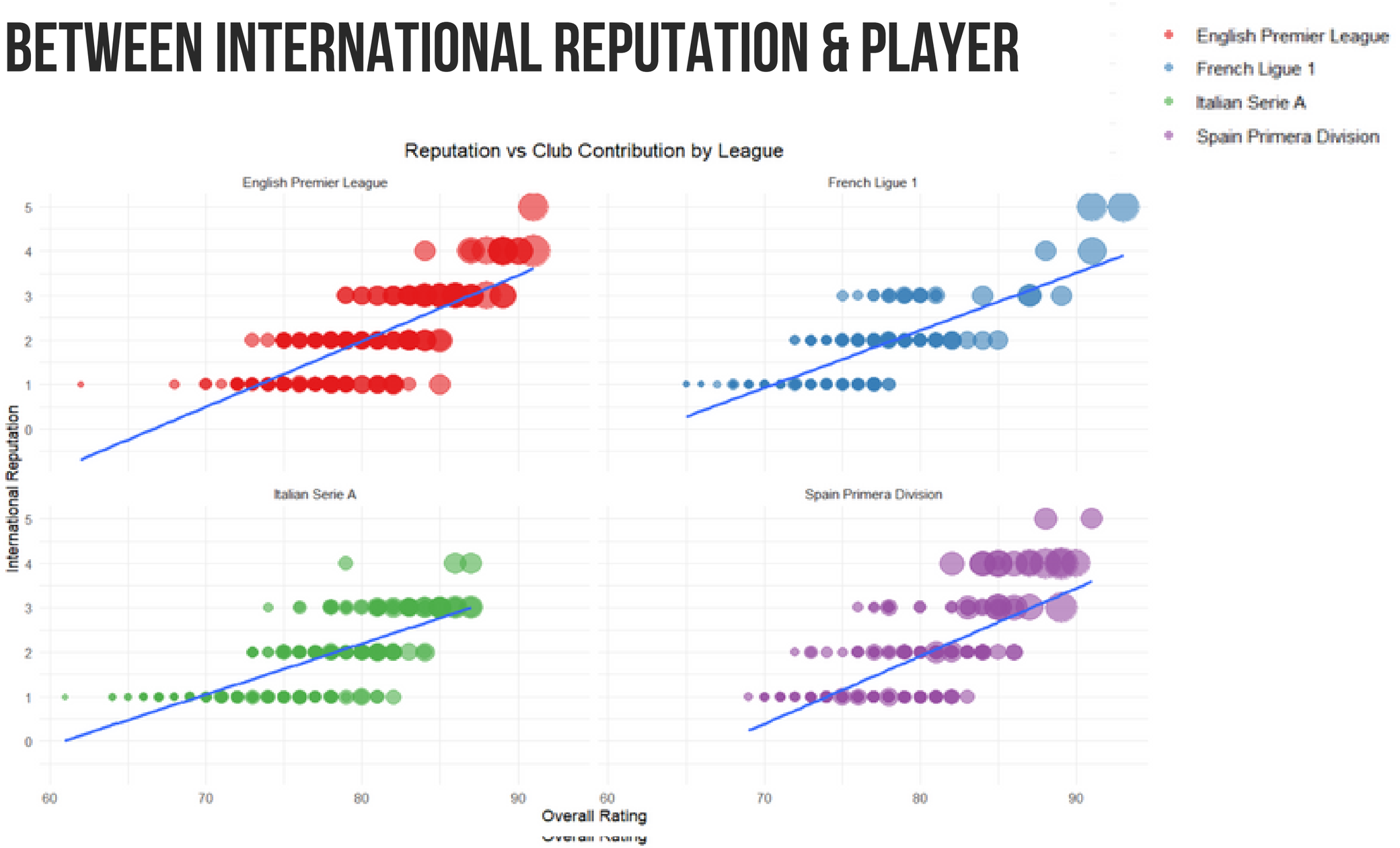
NATIONALITY IN EACH 4 LEAGUES: ENGLAND, FRANCE, ITALY, SPAIN**



Add a main point

Briefly elaborate on what you want to discuss.

RELATIONSHIP BETWEEN INTERNATIONAL REPUTATION & PLAYER CONTRIBUTION



Rating-Reputation-Wage Correlation

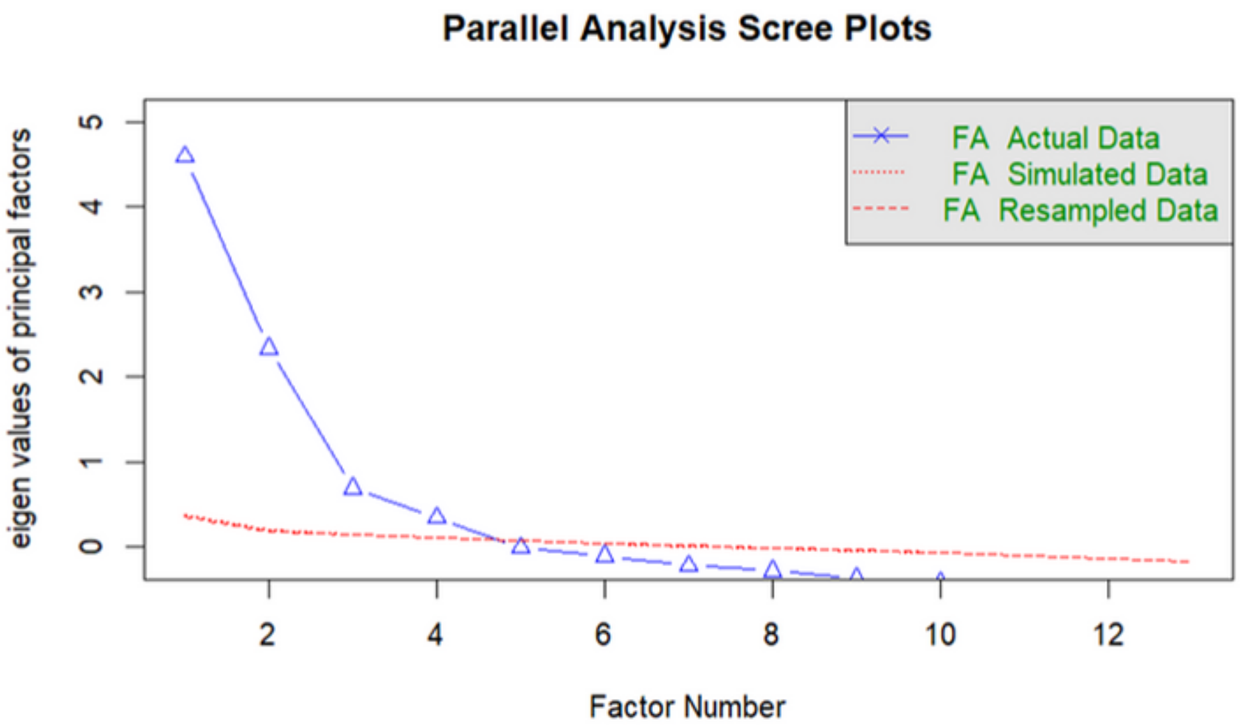
Higher player ratings & reputations align with higher wages, with the **Premier League and La Liga** showing high wages for some with moderate reputations.

FACTOR ANALYSIS

Measure	Result
KMO Measure of Sampling Adequacy	0.73
Bartlett's Test of Sphericity	
Chi-Square	9528.147
P-Value	0
Degrees of Freedom	78

High MSA for 'wage_eur' indicated factor analysis suitability

FACTOR ANALYSIS



Analysis Method	Factors
Parallel Analysis	4
Kaiser Criterion(>0.7)	4
Scree Plot	3 or 4

No. of factors chosen for the analysis: 4

Factor loadings:-

Varimax rotation;
cutt off = 0.3

- **MR1:** Market and performance metrics: "Overall," "Potential," "Value," and "Wage."
-
- **MR2:** Physical dimensions:" Weight" and "Height."
-
- **MR3:** Age & maturity experience:"Age"
-
- **MR4:** Skill-related dynamics: "Pace" inversely, and "Defending" inversely, "Shooting" positively.

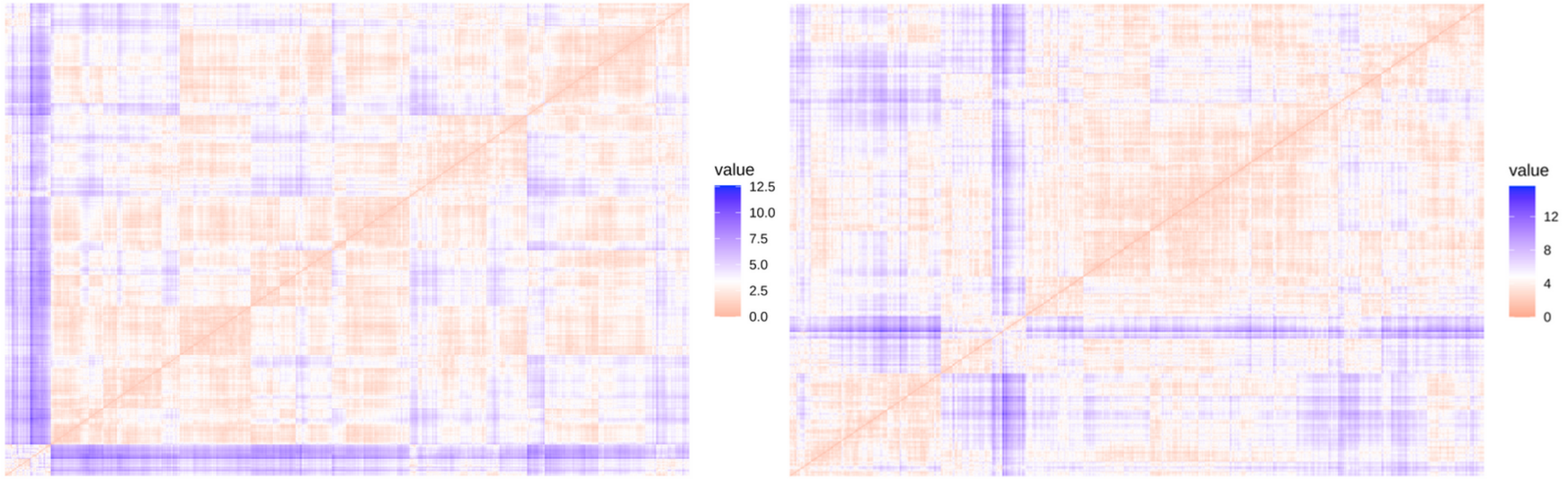
Service Aspect	MR1	MR2	MR3	MR4
Overall	0.934		-0.348	
Potential	0.851			
Value	0.897			
Wage	0.801			
Passing	0.6			
Dribbling	0.6			
weight		0.9		
height		0.9		
physical			0.5	
pace				-0.4
defending				-0.8
shooting				0.8
age			0.9	

CLUSTERING ANALYSIS

& CLASSIFICATION

ASSESSING CLUSTERING TENDENCY

Hopkins Statistics	
Including Goalkeepers: **0.86**	Excluding Goalkeepers: **0.81**



CLUSTERING DISTANCE MEASURES

Including goalkeepers

	Euclidean	Manhattan	Canberra	Minkowski	Cosine
Euclidean	1.0	0.98	0.56	1.0	-0.61
Manhattan	0.98	1.0	0.64	0.98	-0.61
Canberra	0.56	0.64	1.0	0.56	-0.79
Minkowski	1.0	0.98	0.56	1.0	-0.61
Cosine	-0.61	-0.61	-0.79	-0.61	1.0

Excluding goalkeepers

	Euclidean	Manhattan	Canberra	Minkowski	Cosine
Euclidean	1.0	0.98	0.62	0.99	0.66
Manhattan	0.98	1.0	0.69	0.95	0.68
Canberra	0.62	0.69	1.0	0.57	0.85
Minkowski	0.99	0.95	0.57	1.0	0.64
Cosine	0.66	0.68	0.85	0.64	1.0

OPTIMAL NUMBER OF CLUSTERS

Dataset	Elbow Method	Silhouette Method	NbClust (30 indices)
Including Goalkeepers	2 to 6	2	2 to 5
Excluding Goalkeepers	2 to 4	2 to 3	2 to 5

Clustering methods: *k-means, k-medoids, hierarchical (ward, complete, average)*

Clustering distances: *Euclidean, Manhattan, Canberra*

CHOOSING THE BEST CLUSTERING ALGORITHM

INTERNAL MEASURES

Dataset	Validation Measures	Optimal Scores	Method	Clusters
Including Goalkeepers	Connectivity	24.44	hierarchical	2
	Dunn	0.13	hierarchical	2
	Silhouette	0.49	hierarchical	2
Excluding Goalkeepers	Connectivity	2.93	hierarchical	2
	Dunn	0.28	hierarchical	2
	Silhouette	0.54	hierarchical	2

Hierarchical clustering with two clusters performs the best in each case for connectivity, Dunn, and Silhouette measures.

CHOOSING THE BEST CLUSTERING ALGORITHM

Dataset	Method	Number of Clusters		
		Connectivity	Dunn	Silhouette
Including Goalkeepers	k-means	2	2	2
	k-medoids	2	2	2
	hierarchical	2	2	2

The three measures suggest two as the optimal number of clusters for all methods.

Excluding Goalkeepers	k-means	2	6	3
	k-medoids	2	5 or 6	2
	hierarchical	2	2	2

The three measures suggest varying optimal number of clusters depending on the method.

CHOOSING THE BEST CLUSTERING ALGORITHM

STABILITY MEASURES

Dataset	Validation Measures	Optimal Scores	Method	Clusters
Including Goalkeepers	APN	0.02	hierarchical	2
	AD	2.49	k-means	6
	ADM	0.23	k-means	2
	FOM	0.78	k-medoids	6
Excluding Goalkeepers	APN	0.01	hierarchical	2
	AD	3.64	k-medoids	6
	ADM	0.09	hierarchical	2
	FOM	0.78	k-means	6

STUDY OBJECTIVE #1

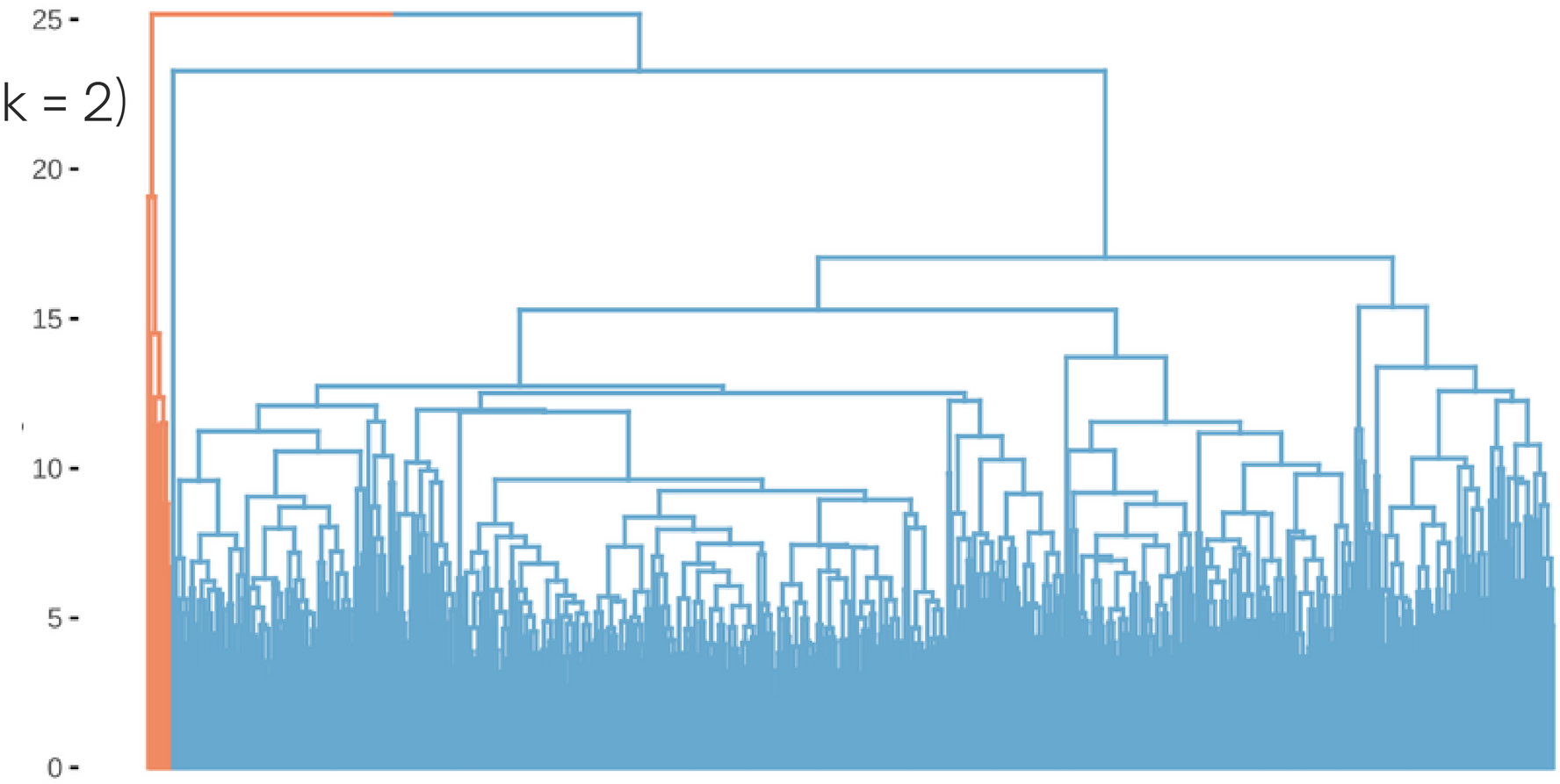
Using clustering, are players grouped based on their overall ratings, skill ratings, physical characteristics, and wages?

What is the distribution across the clusters in terms of league, club, position, nationality, and reputation?

CLUSTER RESULTS

DENDROGRAM

Excluding goalkeepers (k = 2)



CLUSTER DISTRIBUTION

CLUSTER MEANS

Excluding goalkeepers (k = 2)

Cluster	Size	Overall	Potential	Value	Wage	Age	Height	Weight
1	14	89	90	101,357,143	260,357	29	180	76
2	786	77	80	17,176,877	47,228	27	182	75

Cluster	Pace	Shooting	Passing	Dribbling	Defending	Physic
1	83	86	84	88	45	73
2	71	61	68	72	63	72

CLUSTER DISTRIBUTION

ATTACK POSITION

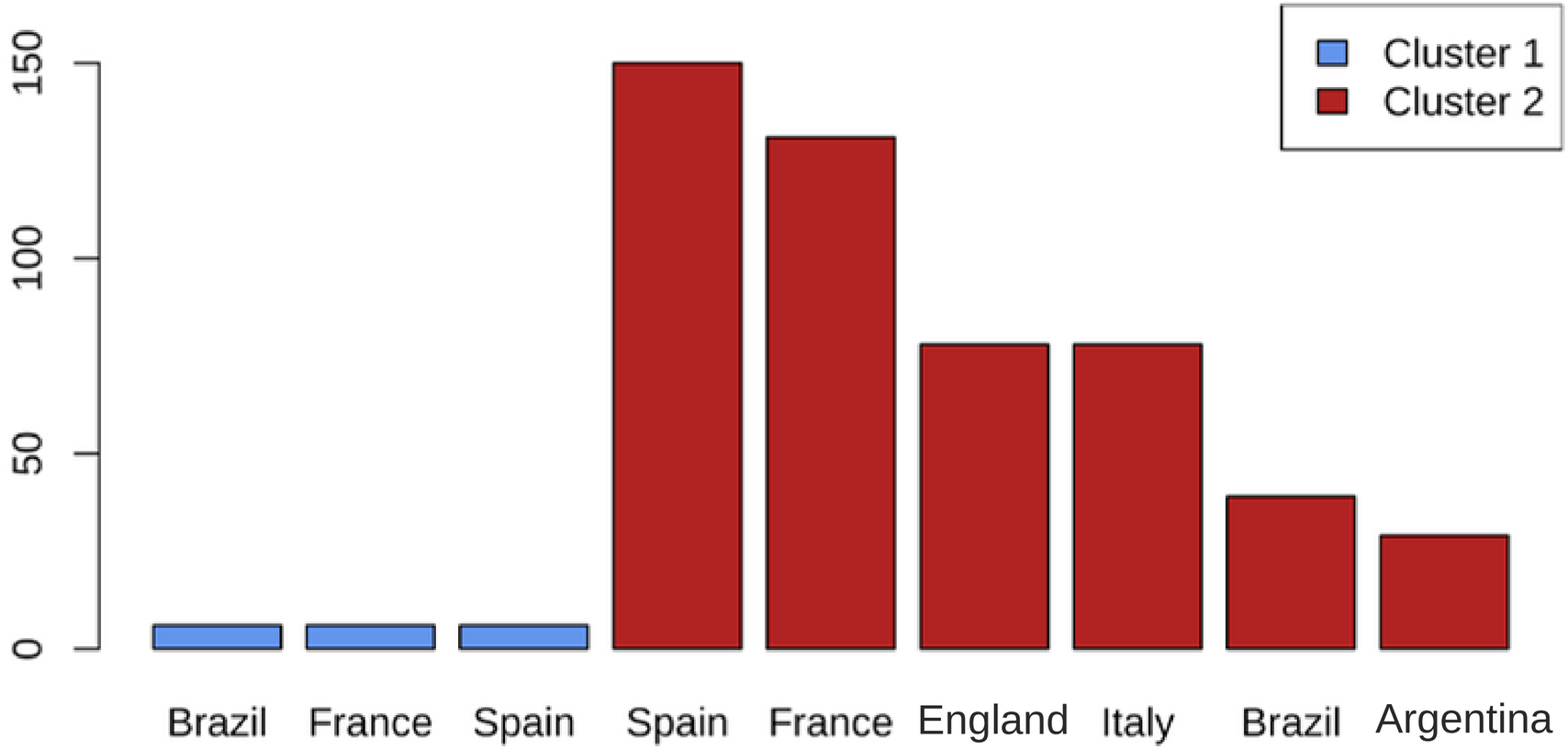
Cluster	Attack	Defense	Midfield
1	11	0	3
2	147	317	322

INTERNATIONAL REPUTATION

Cluster	1	2	3	4	5
1	0	0	1	10	3
2	407	251	112	15	1

CLUSTER DISTRIBUTION

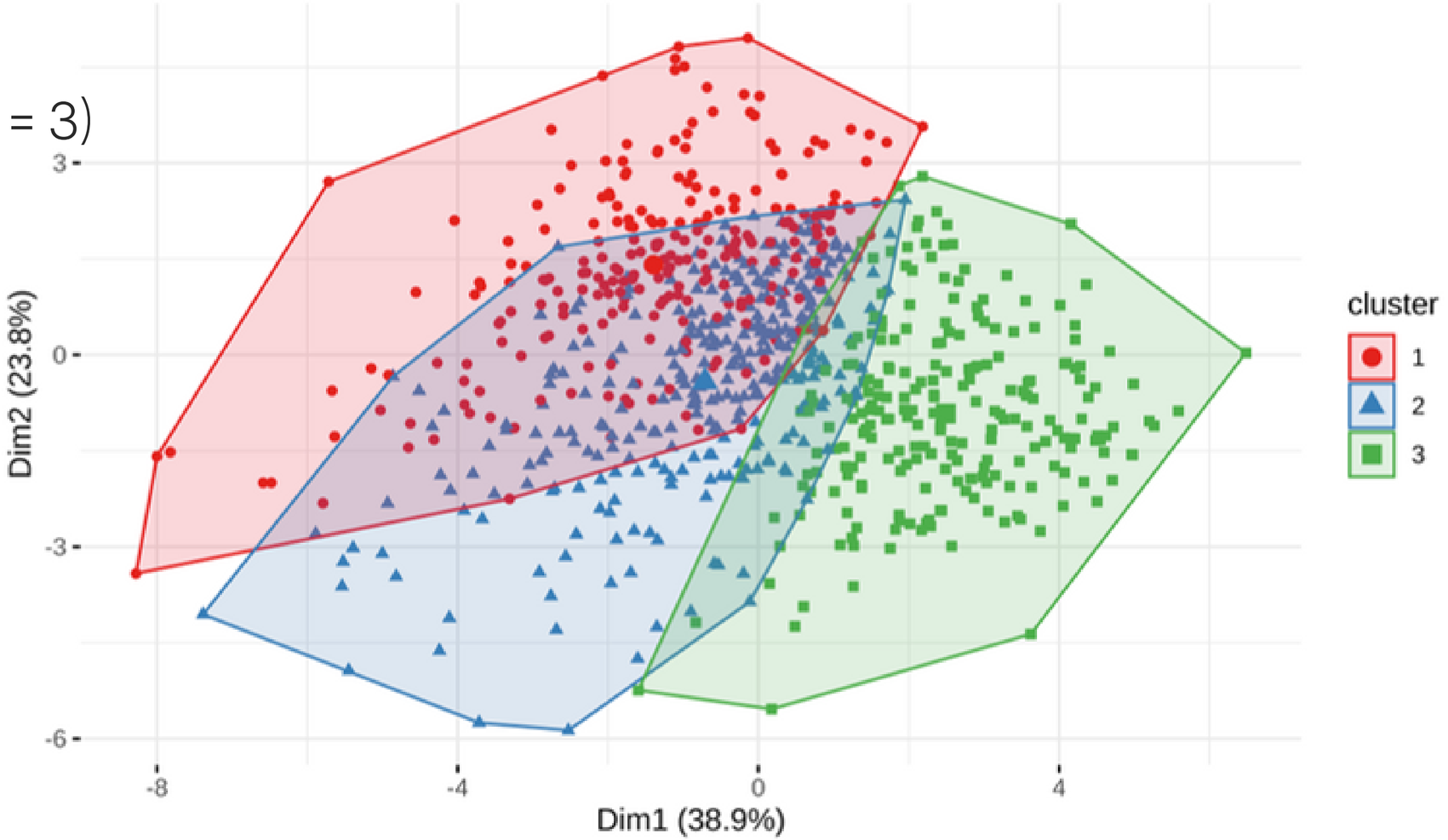
NATIONALITY



CLUSTER RESULTS

CLUSTERS

Excluding goalkeepers (k = 3)



CLUSTER DISTRIBUTION

CLUSTER MEANS

Excluding goalkeepers (k = 3)

Cluster	Size	Overall	Potential	Value	Wage	Age	Height	Weight
1	180	84	86	49,119,444	118,439	27	181	75
2	234	75	78	9,290,278	31,350	27	187	80
3	386	75	79	10,115,544	31,377	27	178	72

Cluster	Pace	Shooting	Passing	Dribbling	Defending	Physic
1	77	72	77	81	65	75
2	61	45	59	62	73	76
3	74	67	70	75	56	67

CLUSTER DISTRIBUTION

ATTACK POSITION

Cluster	Attack	Defense	Midfield
1	50	44	86
2	12	187	35
3	96	86	204

INTERNATIONAL REPUTATION

Cluster	1	2	3	4	5
1	18	64	71	23	4
2	153	66	15	0	0
3	236	121	27	2	0

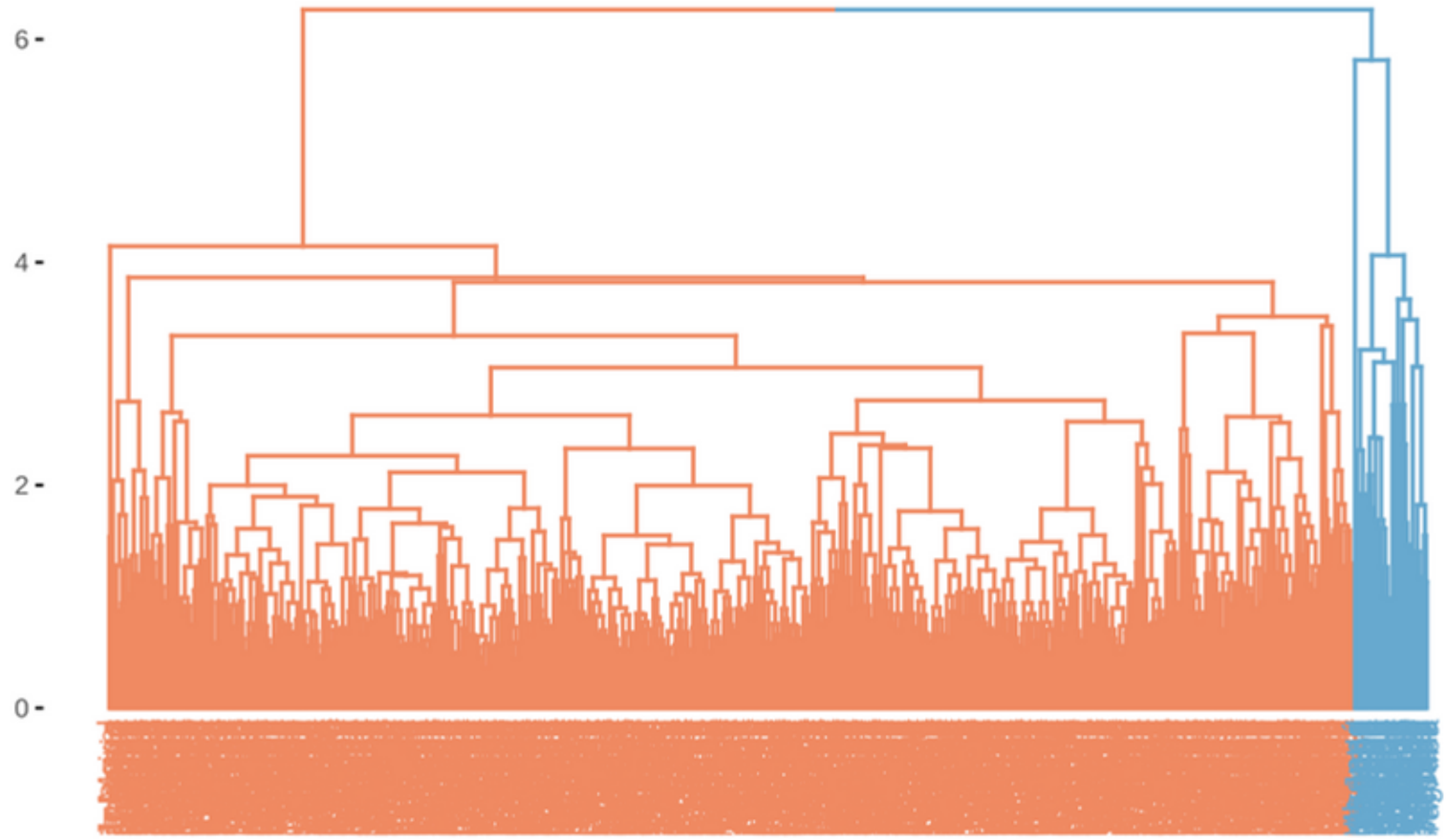
CLUSTER RESULTS

DENDROGRAM

Including goalkeepers (k = 2)

Notable players in Cluster 1:

- Messi
- Ronaldo
- Neymar
- Mbappé
- Benzema
- Modrić
- De Bruyne
- Courtois (GK)



CLUSTER DISTRIBUTION

CLUSTER MEANS

Including goalkeepers (k = 2)

Cluster	Size	Overall	Potential	Value	Wage	Age	Height	Weight
1	49	88	89	86,602,041	210,531	28	182	77
2	831	77	80	14,627,046	41,073	27	182	76

**Results from hierarchical clustering*

CLUSTER DISTRIBUTION

ATTACK POSITION

Cluster	Attack	Defense	GoalKeeper	Midfield
1	16	12	6	15
2	142	305	74	310

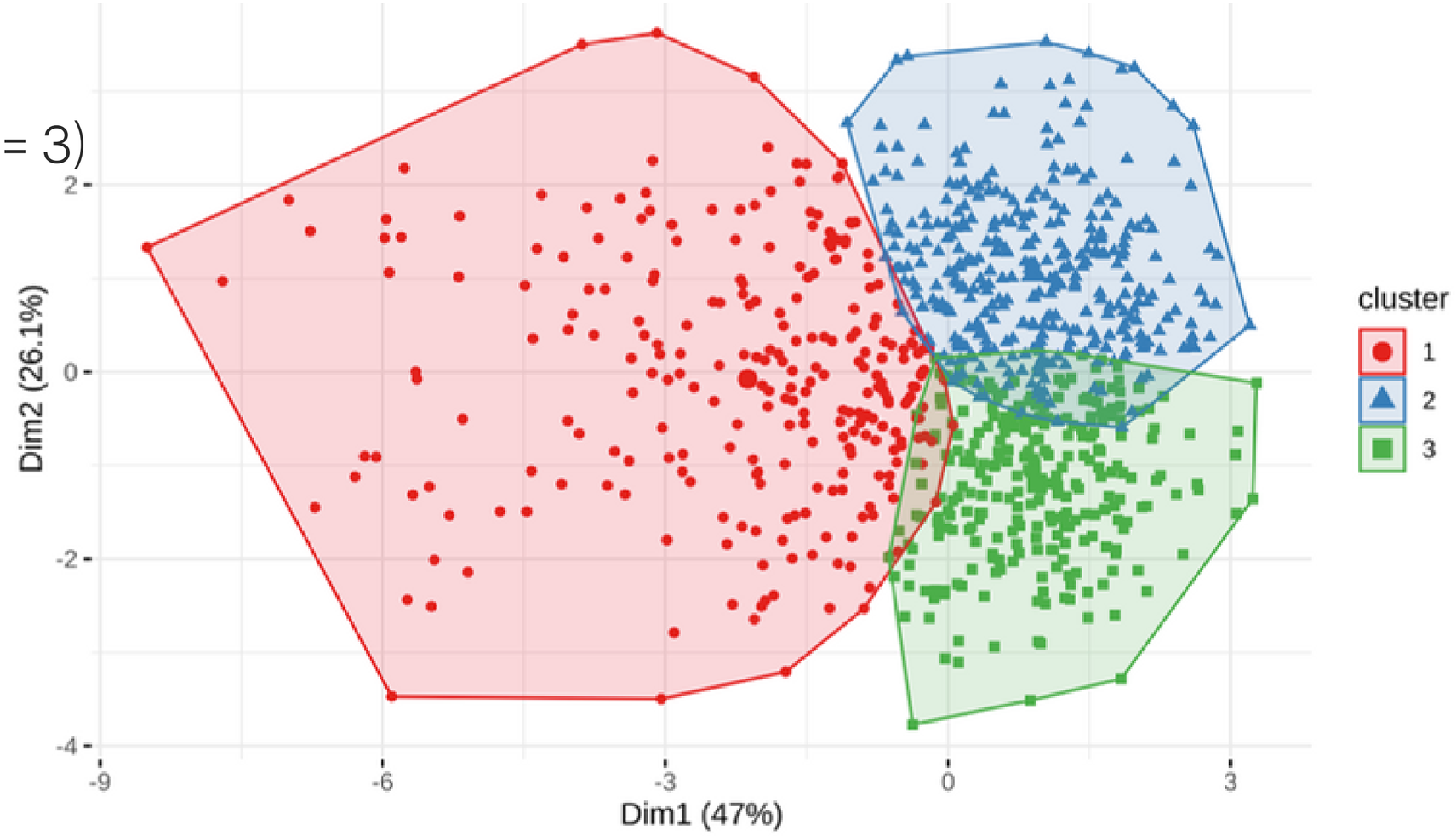
INTERNATIONAL REPUTATION

Cluster	1	2	3	4	5
1	0	0	26	19	4
2	446	272	102	10	1

CLUSTER RESULTS

CLUSTERS

Including goalkeepers (k = 3)



CLUSTER DISTRIBUTION

CLUSTER MEANS

Including goalkeepers (k = 3)

Cluster	Size	Overall	Potential	Value	Wage	Age	Height	Weight
1	267	83	85	431,516,85	100,843	27	183	77
2	334	75	79	9,172,530	28,400	25	178	71
3	279	75	77	6,499,821	28,806	29	187	81

**Results from k-medoids clustering*

CLUSTER DISTRIBUTION

ATTACK POSITION

Cluster	Attack	Defense	GoalKeeper	Midfield
1	64	78	29	96
2	55	110	4	165
3	39	129	47	64

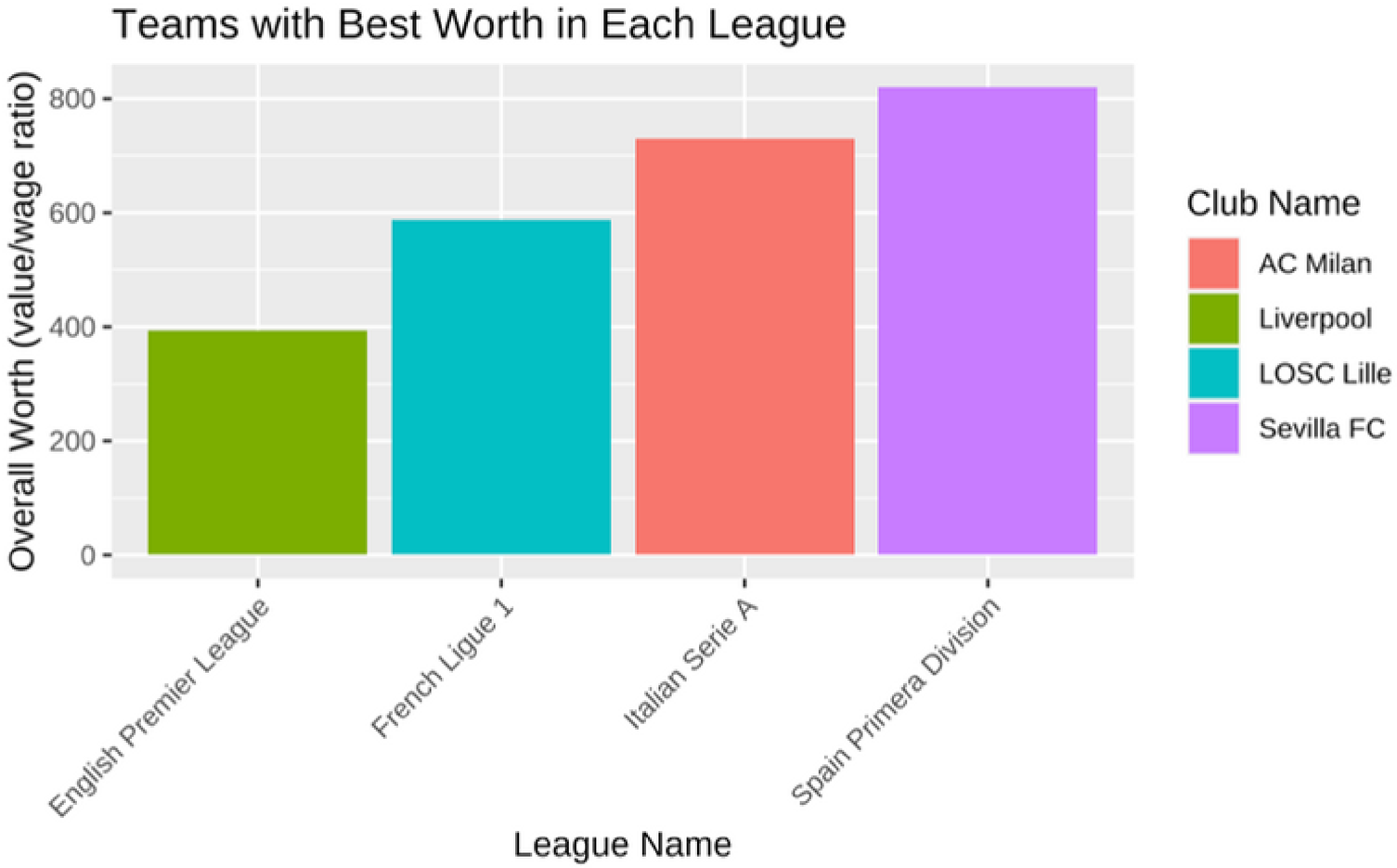
INTERNATIONAL REPUTATION

Cluster	1	2	3	4	5
1	48	97	89	28	5
2	238	77	18	1	0
3	160	98	21	0	0

STUDY OBJECTIVE #2

CAN WE IDENTIFY THE TEAM WITH THE BEST WORTH IN EACH LEAGUE FOR AN AFFORDABLE INVESTMENT?

IDENTIFYING HIGH-VALUE TEAMS: LEAGUE LEADERS IN EFFICIENT INVESTMENT



STUDY OBJECTIVE #3

Predict the wage of a player in each league based on their abilities, physical characteristics, and nationality.

STEPWISE

TO PREDICT WAGE

forward & backward elimination

Including GK

- club_name
- international_reputation
- value_eur
- age
- club_position
- overall
- body_type
- potential

Excluding GK

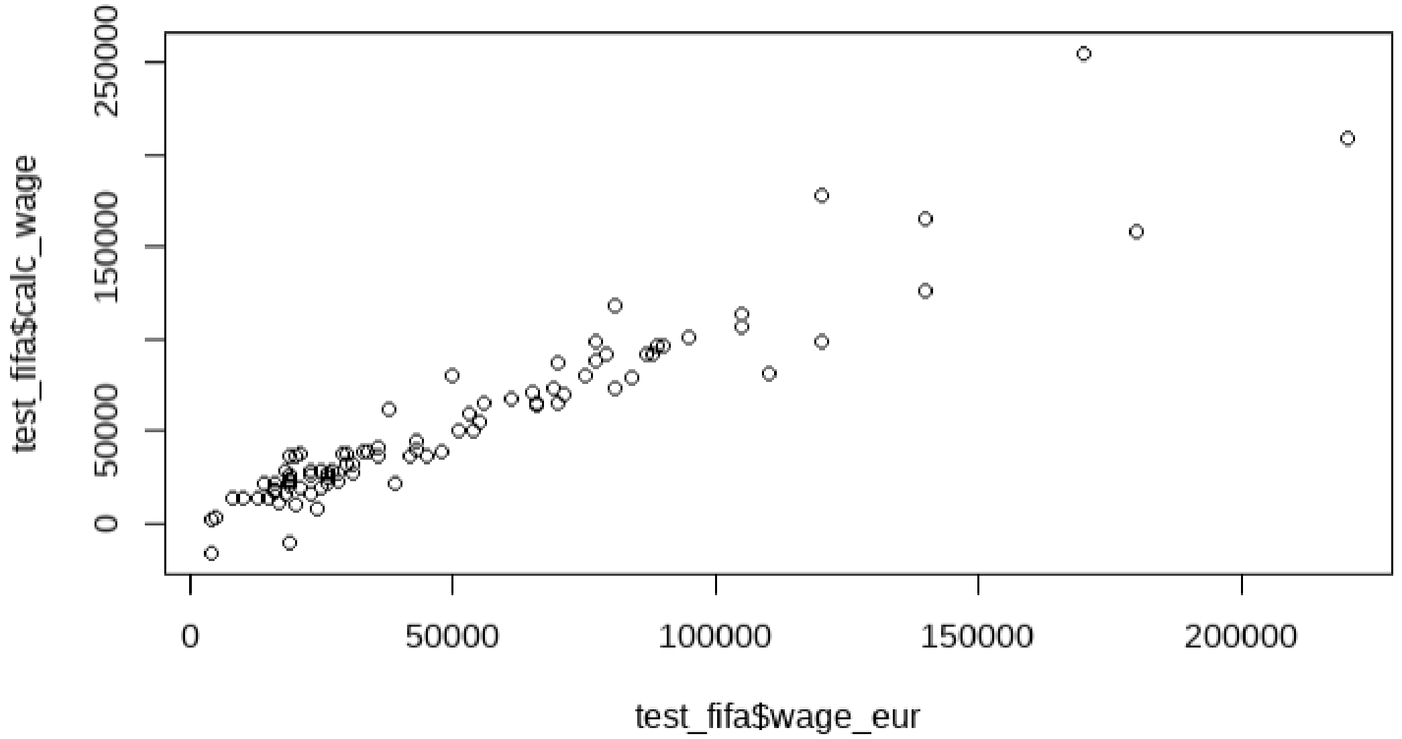
- club_name
- international_reputation
- value_eur
- age
- club_position
- overall
- body_type
- potential
- shooting

STEPWISE TO PREDICT WAGE

small test sample (10%)

Including GK

- club_name
- international_reputation
- value_eur
- age
- club_position
- overall
- body_type
- potential



Avg pct error is 0.78% , median is 5.70%

player_name	club_name	league_name	nationality_name	international_reputation	attack_position	wage_eur	calc_wage	pct_error
Salva Ferrer	Spezia	Italian Serie A	Spain	1	Midfield	€ 5,000	€ 4,027	-19%
É. Mendy	Chelsea	English Premier League	Senegal	2	GoalKeeper	€ 105,000	€ 107,325	2%
J. Ikoné	LOSC Lille	French Ligue 1	France	3	Midfield	€ 36,000	€ 33,507	-7%
G. Bale	Real Madrid CF	Spain Primera Division	Wales	4	Attack	€ 170,000	€ 234,077	38%

STUDY OBJECTIVE #4

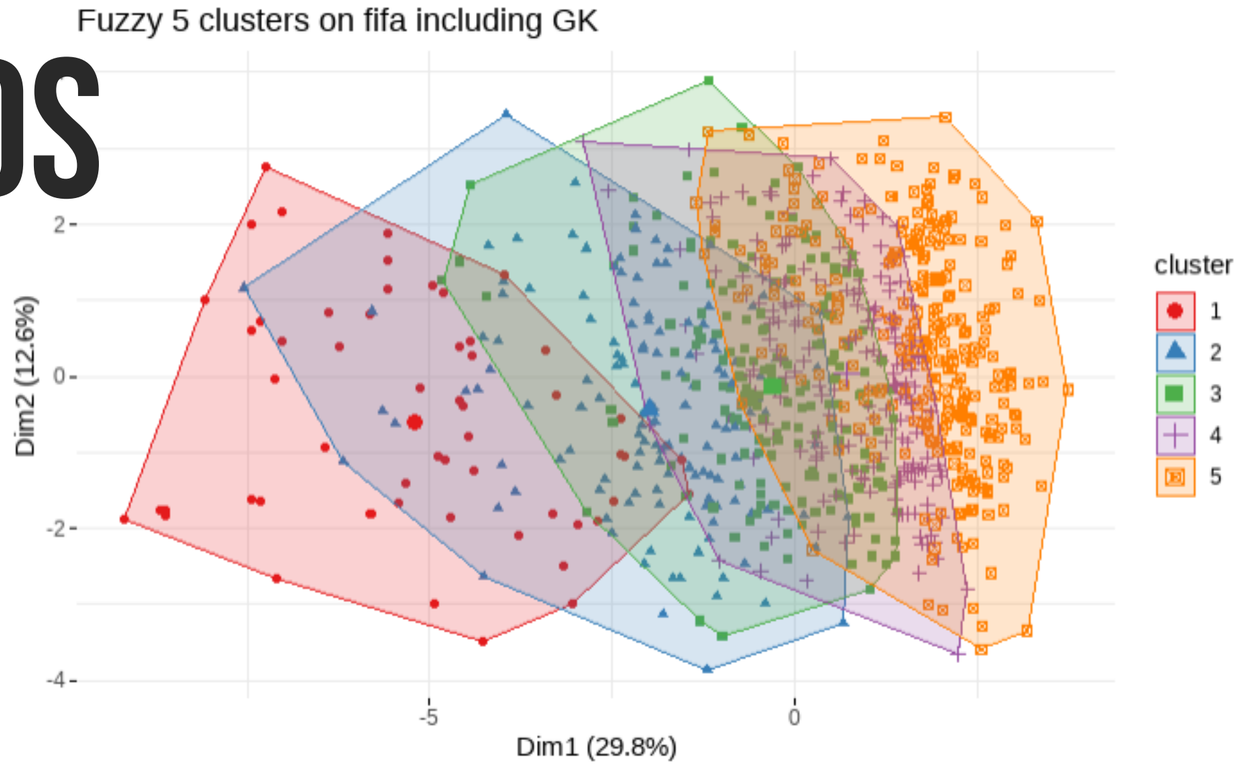
Can we determine the international player’s reputation from their contribution to their club?

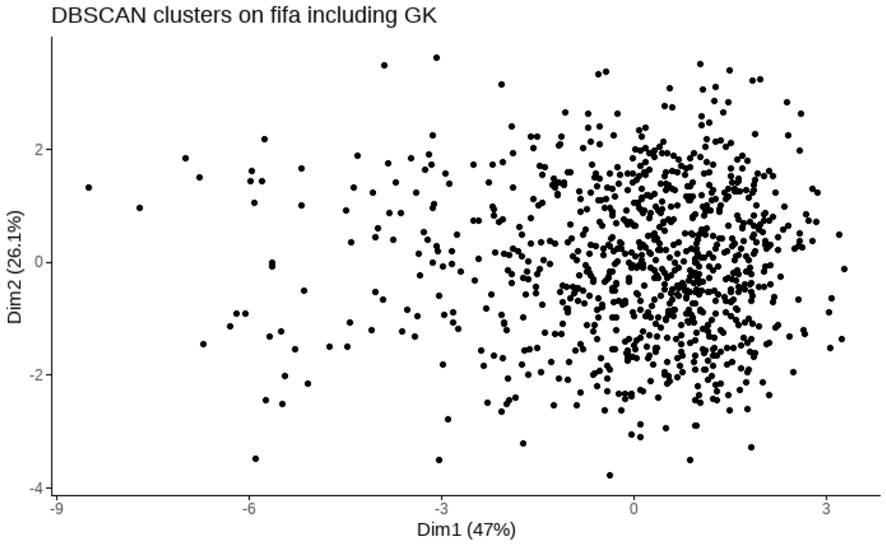
CLUSTERING METHODS

Including GK, excluding GK

- k-means,
- k-medoids,
- hierarchical with various linkage,
- fuzzy,
- dbscan

- 3 clusters
- 5 clusters





External validation statistics :
Rand index = 0.1 or less
Meila's VI = 2.6 or less

CONCLUSION

SUMMARY

Q1: clustering the dataset

- hierarchical clustering for 2 clusters
  - cluster 1 top international players
  - cluster 2 league players
- k-medoids for 3 clusters
  - cluster 1 top international players
  - cluster 2 younger generation with potential
  - cluster 3 older players at the end of their career

Q2: investing in soccer league

Best worth for investment:
Sevilla FC (Spain Primera Division)

Q3: predict players wage

- Stepwise elimination on train/test dataset
- Good prediction but impacted by outliers

Q4: clustering international reputation

- Unsuccessful with clustering
- Clusters are representing something else

DISCUSSION

Clustering exploration

Two clusters were clearly identified by many methods.

Many methods and many indices give different results, knowledge is key.

Clustering is not usable for everything.

Bias

Goal keepers have different variables to define their abilities.

Possible improvement by including more of the skills variables present in the original dataset, and more players (substitutes, reserve).

Real life

Players can easily move to a different league in Europe.

Wage is not the only remuneration (sponsoring).

Value of a player is not only monetary and physical skills (image of their club).

QUESTIONS