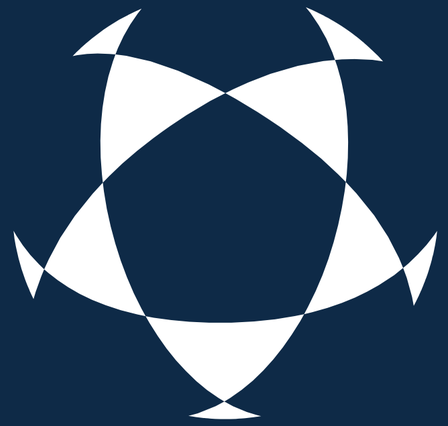
FIFA

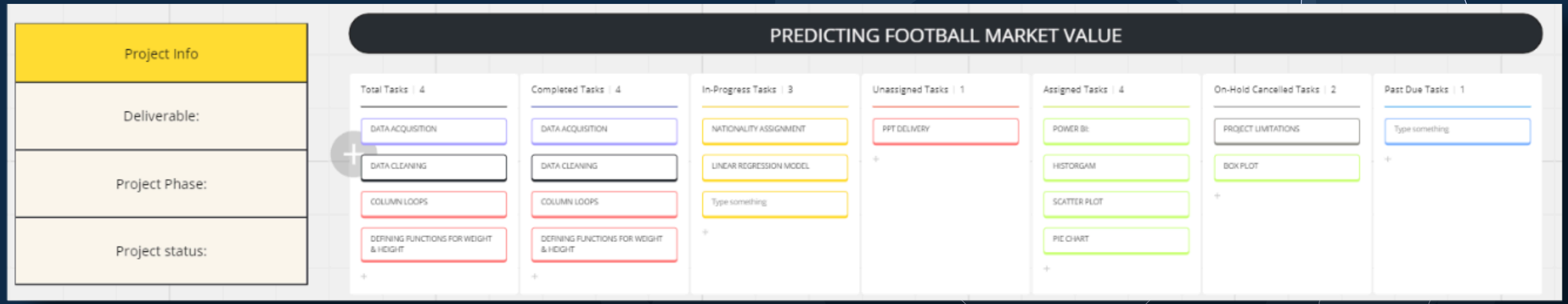
FIFA
OFFICIAL
LICENSED
PRODUCT

PREDICTING FOOTBALL PLAYER MARKET VALUE

CONTENT TABLE

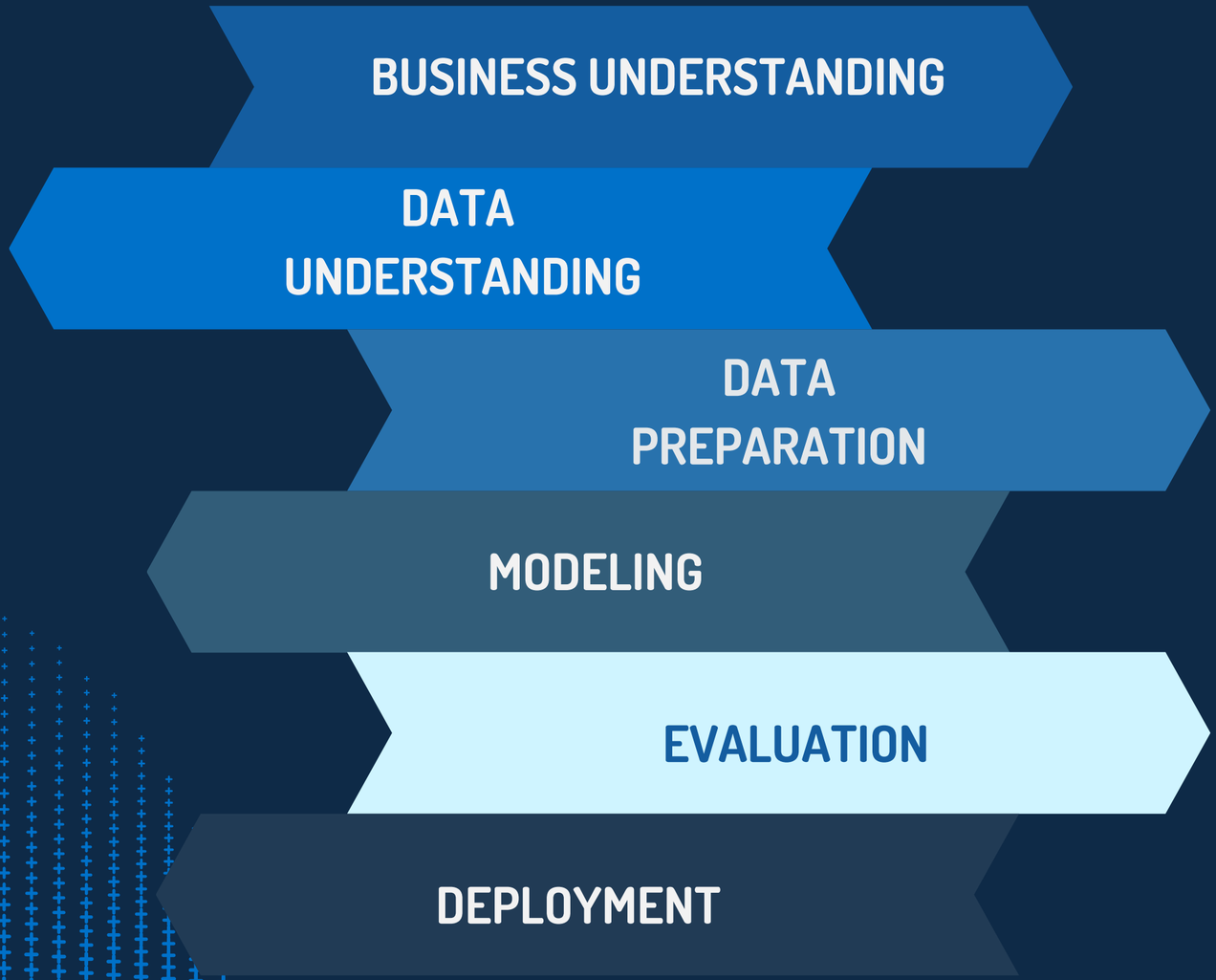


MIRO BOARD: DEVELOPMENT STRATEGY



WHO WE ARE

OBJECTIVES AT HAND: THE DATA SCIENCE METHODOLOGY



BUSINESS UNDERSTANDING : OUR OBJECTIVES

CHALLENGES

Decision-Making for Stakeholders

Clubs, agents, and investors rely on accurate valuations for negotiations, transfers, and investments. Incorrect valuations based on FIFA stats can lead to poor financial decisions.

Overvalued and Undervalued Players

Identifying discrepancies between a player's market price and their actual performance and potential as indicated by their FIFA stats. This can lead to inefficient transfer spending or missed opportunities.

OBJECTIVES

By leveraging machine learning algorithms and player characteristics data from sources, the project aims to develop predictive models that can estimate players' market values based on their attributes.

Additionally, the project seeks to identify factors that contribute most significantly to a player's market value and assess the effectiveness of different regression techniques in predicting these values accurately. Therefore, the problem encompasses data collection, cleansing, modeling, evaluation, and interpretation to provide actionable insights for stakeholders in the football industry.

+Add we we're using a game as it provides an accurate database with all the stats

MODEL EVOLUTION: AN EXPLORATORY ANALYSIS

TOP 5 VALUABLE PLAYERS (BY MARKET VALUE)

Name	Age	Team	Market Value
E. Haaland	22	Manchester City	185000000
K. Mbappé	24	Paris Saint Germain	181500000
Vini Jr.	22	Real Madrid	158500000
J. Musiala	20	FC Bayern München	134500000
F. Valverde	24	Real Madrid	130500000

TOP 5 CLUBS (BY TOTAL)

Team	Market Value
Manchester City	1151030000
Real Madrid	1118075000
Paris Saint Germain	1022800000
FC Bayern München	961475000
Arsenal	960725000

TOP 5 CLUBS (BY MEAN)

Team	Market Value
FC Bayern München	30046093.75
Paris Saint Germain	30082352.94
Real Madrid	30218243.24
Wolfsburg W	35500000.00
Manchester City	37130000.00

TOP 5 POSITION

Best Position	Best Overall
CF	73.43
RW	68.79
CM	68.71
LW	68.70
CDM	66.58

TOP 5 RATED PLAYERS (BY BEST OVERALL)

Name	Age	Team	Best Overall
L. Messi	36	Inter Miami	91
T. Courtois	31	Real Madrid	90
M. ter Stegen	31	FC Barcelona	89
Ederson	29	Manchester City	88
J. Oblak	30	Atlético Madrid	88
Casemiro	31	Manchester United	86
R. De Paul	29	Atlético Madrid	84
R. Lukaku	30	Roma	84

AGE, HEIGHT & WEIGHT

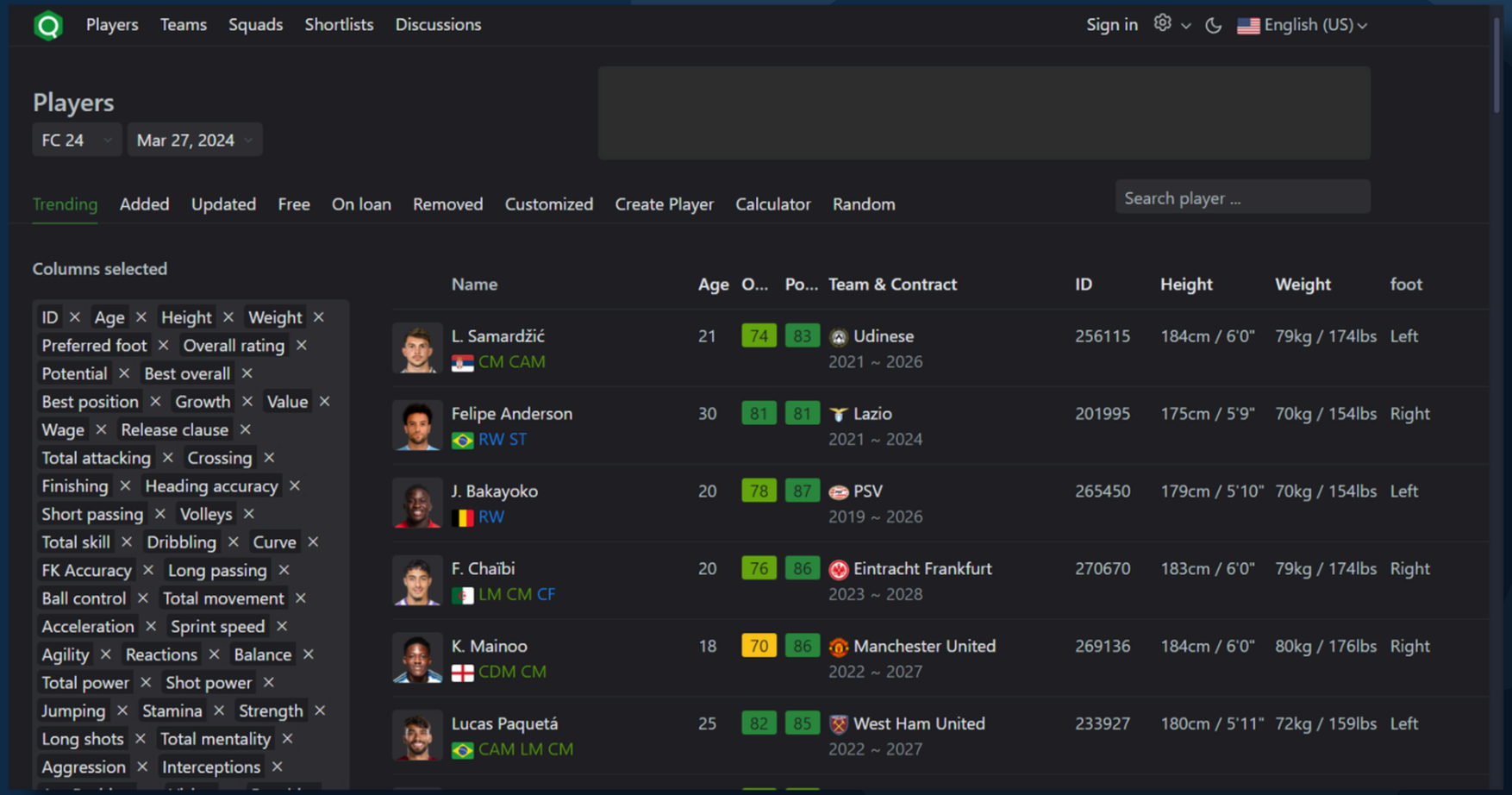
Statistic	Age	Height	Weight
25%	20.00	176.00	70.00
75%	26.00	186.00	79.00
Max	42.00	206.00	103.00
Mean	23.08	180.89	74.08
Min	15.00	149.00	45.00

I

DATA ACQUISITION

DATA ACQUISITION

Sofifa Dataset



https://sofifa.com/

DATA ACQUISITION

1. Extracted data from Sofifa using Beautiful Soup and store the raw dataset in **FootballPlayers.xlsx** file

```
player_list = []

i = 0
while i < 36000:
    url = "https://sofifa.com/?&showCol%5B%5D=pi&showCol%5B%5D=ae&showCol%5B%5D=hi&showCol%5B%5D=wi&showCol%5B%5D=pf&showCol%5B%5D=oa&showCol%
    headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/109.0.0.0 Safari/537.36',
}
    response = requests.get(url, headers=headers)
    page = response.text
    #print(page)
    soup = bts(page,"html.parser")
    #print(soup.prettify)
    rows = soup.find_all("tr")
    #print(rows)
    for row in rows:
        cells = row.find_all('td')
        #print(cells)
        cells_to_string = str(cells)
        cells_sub = (re.sub(re.compile('<.*?>'),'',cells_to_string))
        #print(cells_sub)
        player_list.append(cells_sub)
    i += 60

result = pd.DataFrame(player_list)
```

	0
0	[]
1	[\n, \nA. Jashari CDM CM\n, 20, 72, 84, \n\n\n...
2	[\n, \nM. Bard LB\n, 22, 77, 82, \n\n\n\nNice\...
3	[\n, \nN. Irankunda RM LM\n, 17, 64, 85, \n\n\...
4	[\n, \nM. Lacroix CB\n, 23, 76, 84, \n\n\n\nVf...
...	...
20178	[\n, \nG. Valle GK\n, 27, 66, 69, \n\n\n\nLDU ...
20179	[\n, \n22 J. Jiménez CB RB\n, 26, 62, 67, \n\n...
20180	[\n, \n23 W. Vargas RM RB RWB\n, 25, 64, 67, \...
20181	[\n, \n23 K. Becerra CB\n, 26, 66, 69, \n\n\n\...
20237	[\n\n, \nJ. Kiwior CB LB\n, 23, 76+2, 83, \n\n...

19426 rows × 1 columns

```
file_name = "FootballPlayers.xlsx"

result.to_excel(file_name)
```

II

DATA CLEANING

DATA CLEANING

1. Created a new column called “Contract Start Date” and “Contract End Date”
2. Created a new column called “Contract Duration” which is the difference of Contract End Date and Contract Start Date
3. Removed “null” values
4. Ensured all the columns are of correct types, i.e, int, string, etc.
5. Converted the Height and Weight columns to cm’s and kg’s respectively
6. Converted the Value, Wage, Release Clause columns to proper integer types
7. Created a new column called “Best Position Category” which is just assigning unique numbers to each position
8. Created a new column called “Foot Category” which is just assigning unique numbers to each foot

1

COMPLEXITY OF MARKET VALUATIONS

2

MODEL OVERFITTING

3

PRIVACY CONCERNS

4

DATA QUALITY AND AVAILABILITY

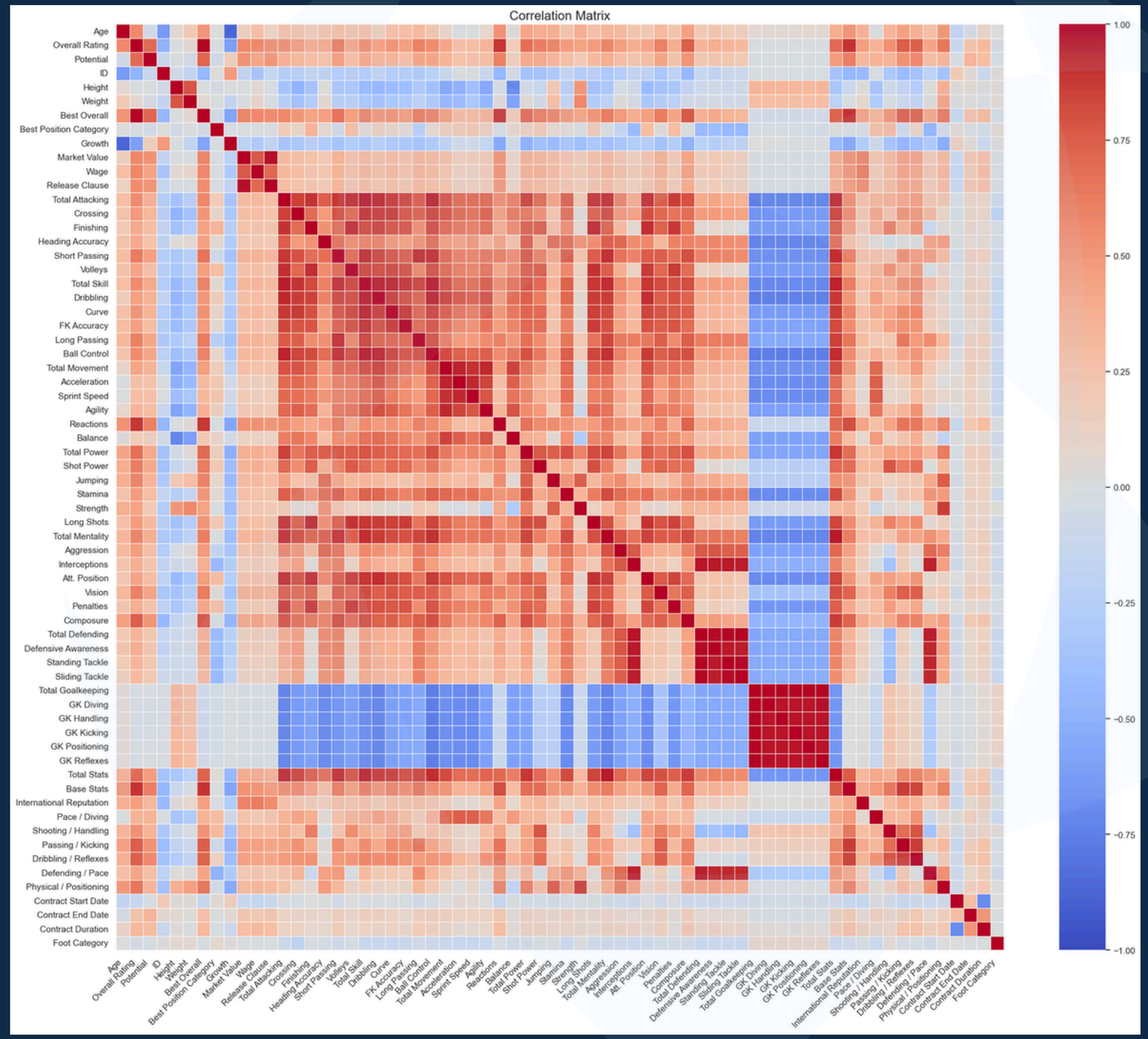
POSSIBLE LIMITATIONS

IV

LOOKING AT OUR DATA ATTRIBUTES

CORRELATION MATRIX

- The color scale on the right ranges from blue to red.
- Red indicates a strong positive correlation (closer to +1).
- Blue indicates a strong negative correlation (closer to -1).
- White or light colors indicate little to no correlation (closer to 0).



Due to the complexity of the previous analysis, we identified the key features that are highly correlated with Market Value and those that are not

	Market Value
Market Value	1.000000
Release Clause	0.975560
Wage	0.766693
International Reputation	0.591019
Best Overall	0.584640
Overall Rating	0.578025
Reactions	0.525931
Potential	0.523294
Base Stats	0.518266
Dribbling / Reflexes	0.447635
Passing / Kicking	0.445594
Composure	0.432036
Total Stats	0.401420

	Market Value
Growth	-0.205567
ID	-0.178767
Contract Start Date	-0.122120
Best Position Category	-0.032176
Height	-0.029647
Foot Category	-0.021102
GK Handling	-0.018457
GK Diving	-0.017709
Weight	-0.016603
Total Goalkeeping	-0.016534
GK Reflexes	-0.015223
GK Kicking	-0.015164

NARROWING DOWN ATTRIBUTES: A GLOSSARY

Overall Rating: A rating assigned to a player based on all the attributes they have in the game. A score out of 100.

Release Clause: An amount a team would need to pay the club to get the player.

Best Overall: The highest overall the player has had in the current edition of the game. A score out of 100 which fluctuates throughout the year based on real-life performances.

International Reputation: How well known and respected the player is internationally. A score out of 5.

Potential: The potential future rating of the player based on expected growth.

Base Stats: Basic statistics of the player.

Total Stats: A sum of all the attributes related to the player such as Shooting, Speed, Defending, etc.

Wage: The wage or salary of the player.

Height: Height of the player in cm's.

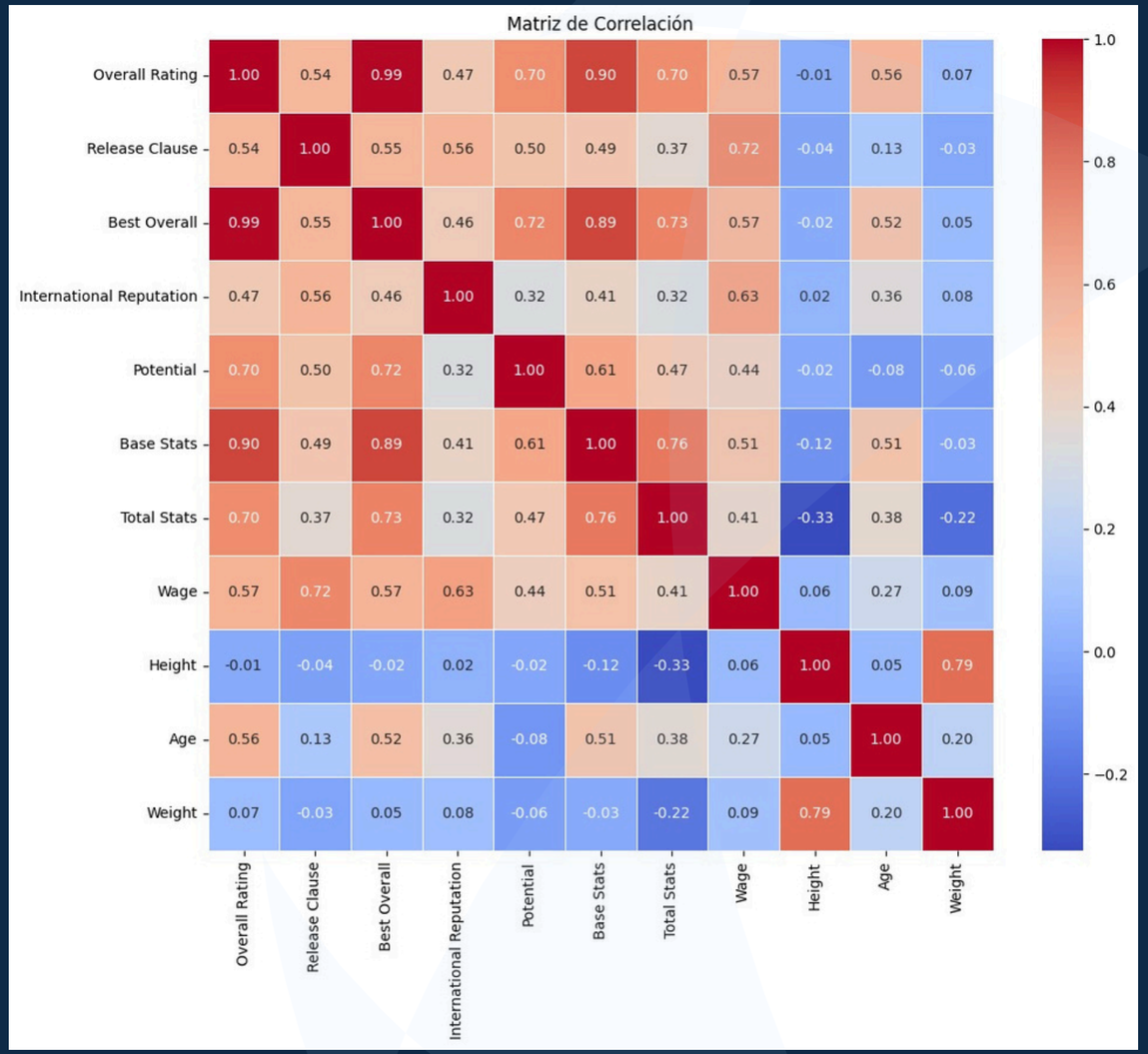
Age: Age of the player.

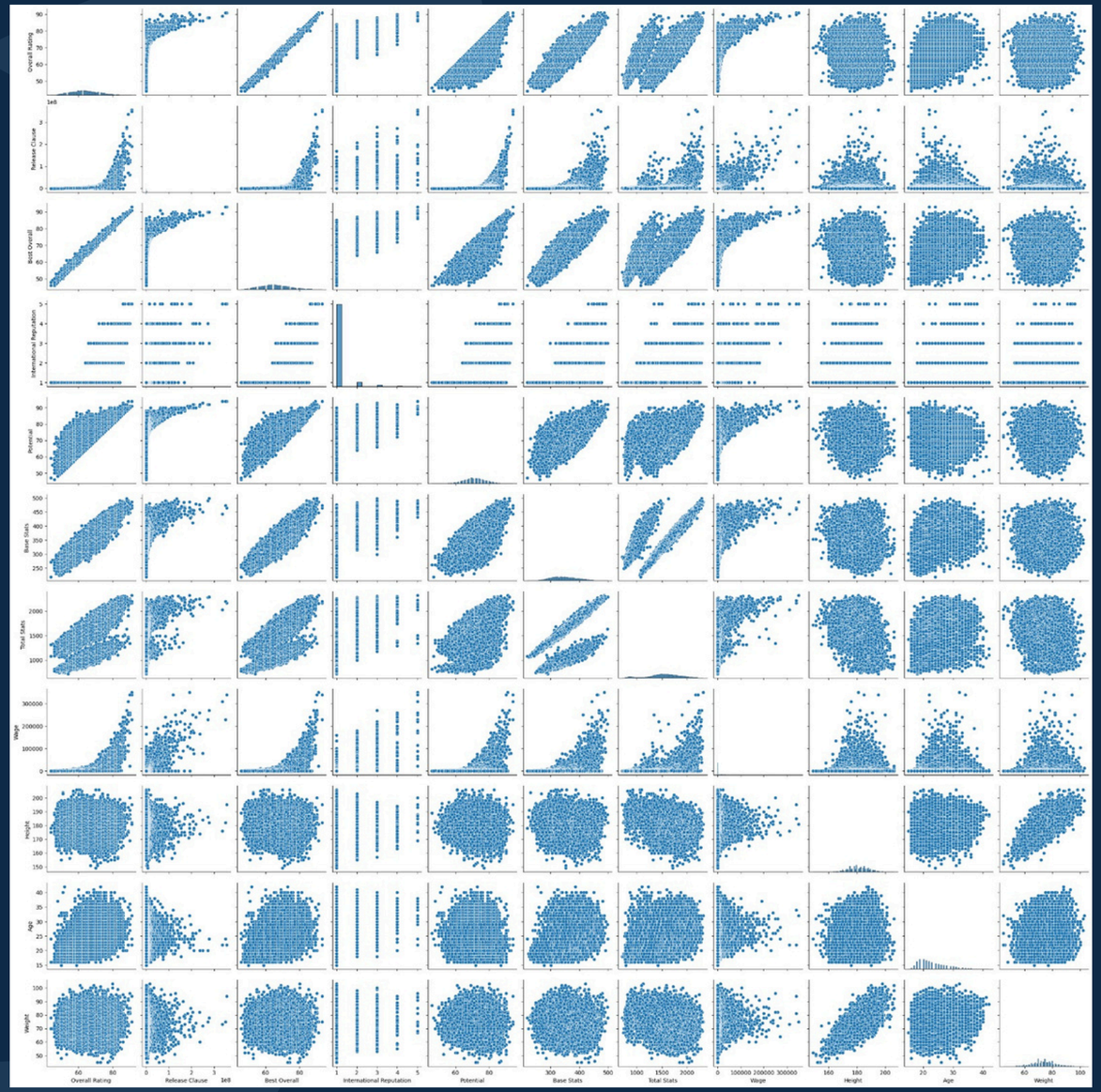
Weight: Weight of the player in kg's.

Market Value: The current market value of the player.

V

BUILDING THE MODEL





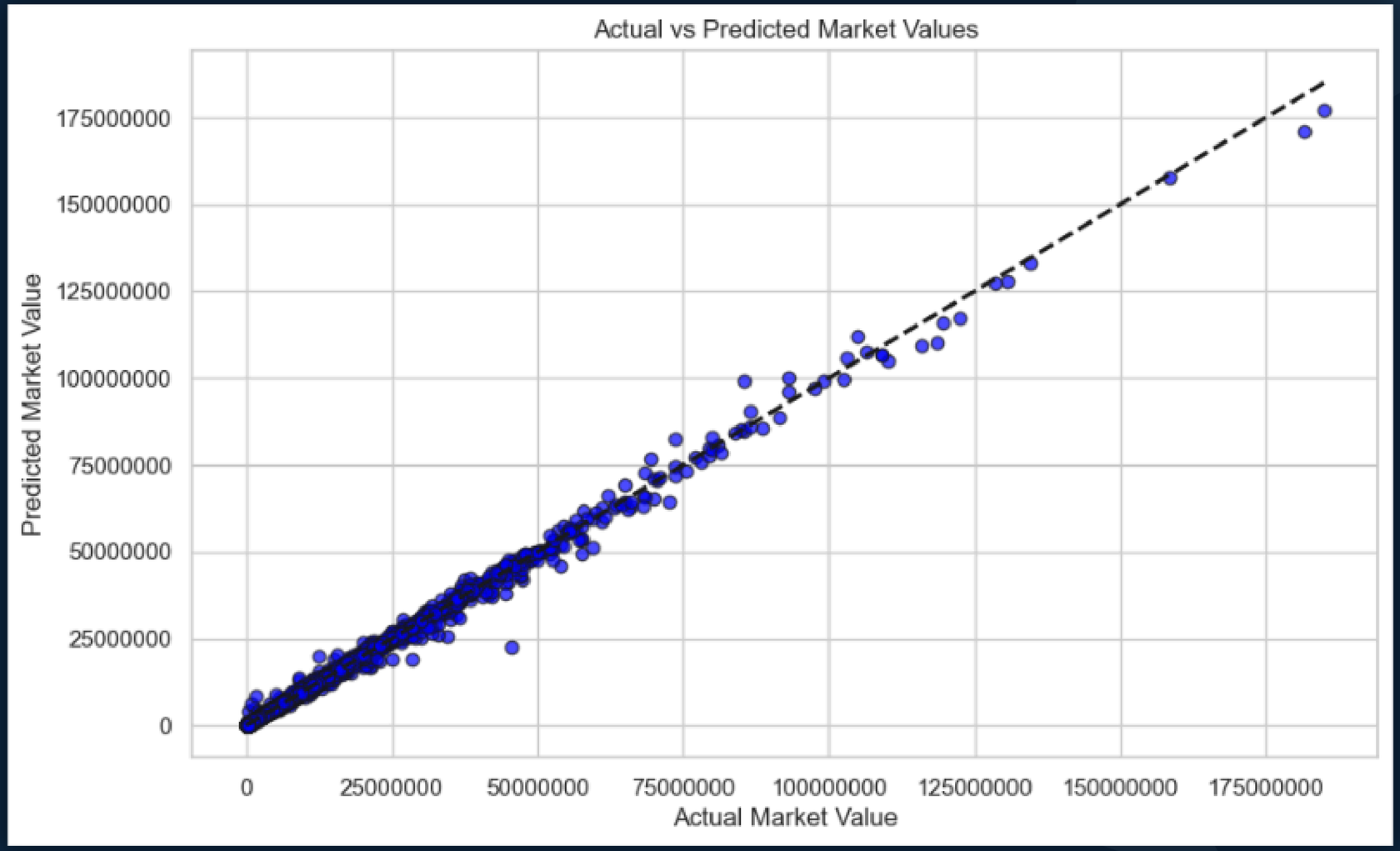
BRIDGING THE LINKS BETWEEN ATTRIBUTES & CORRELATION

MEAN SQUARED ERROR AND R^2 SCORE AFTER SPLITTING, TRAINING AND TESTING THE DATA FOR LINEAR REGRESSION AND RANDOM FOREST REGRESSION

```
R² Score LG: 0.9554098648431475
Mean Squared Error LG: 2962877207329.086
Cross-Validation R² Scores LG: [0.9475888  0.96457697 0.98530477 0.9434087  0.84150235]
Coefficients LG: [ 6.86452039e+05  6.86787120e+06 -5.96047273e+03  2.62699716e+05
 -2.67118623e+05  3.10957854e+04  2.31538317e+04  9.80264710e+05
  5.21666725e+03 -4.83176921e+05  2.87860330e+04]
```

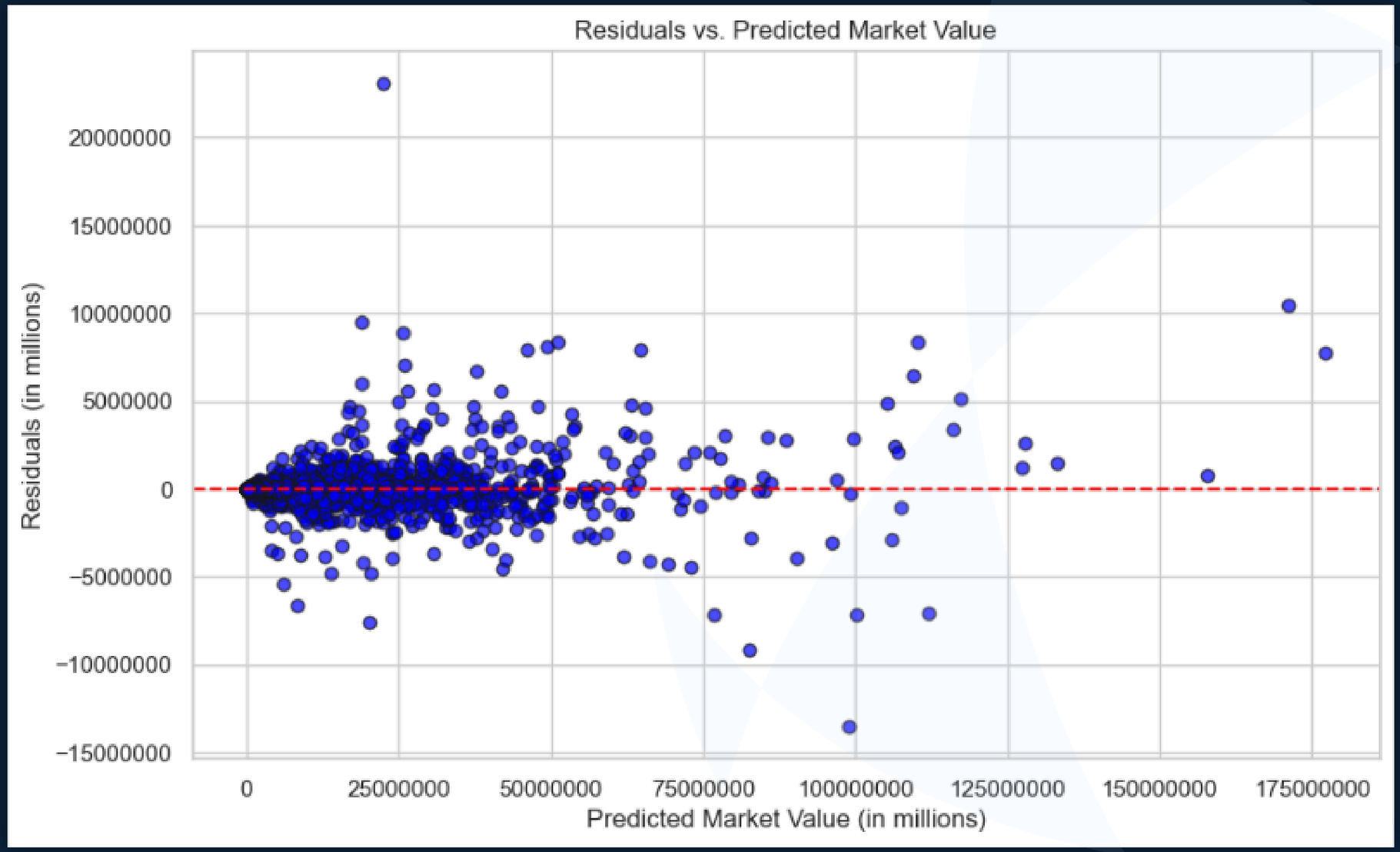
```
R² Score RFG: 0.9899598516610172
Mean Squared Error RFG: 667136947827.8801
Cross-Validation R² Scores RFG: [0.97594843 0.98728009 0.99298581 0.98631157 0.99529891]
Root Mean Squared Error RFG: 816784.5173776742
```

SCATTER PLOT OF ACTUAL MARKET VALUE VS PREDICTED MARKET VALUE



- X-axis: Represents the actual market values of the players
- Y-axis: Represents the predicted market values of the players.

SCATTER PLOT OF RESIDUALS VS PREDICTED MARKET VALUE



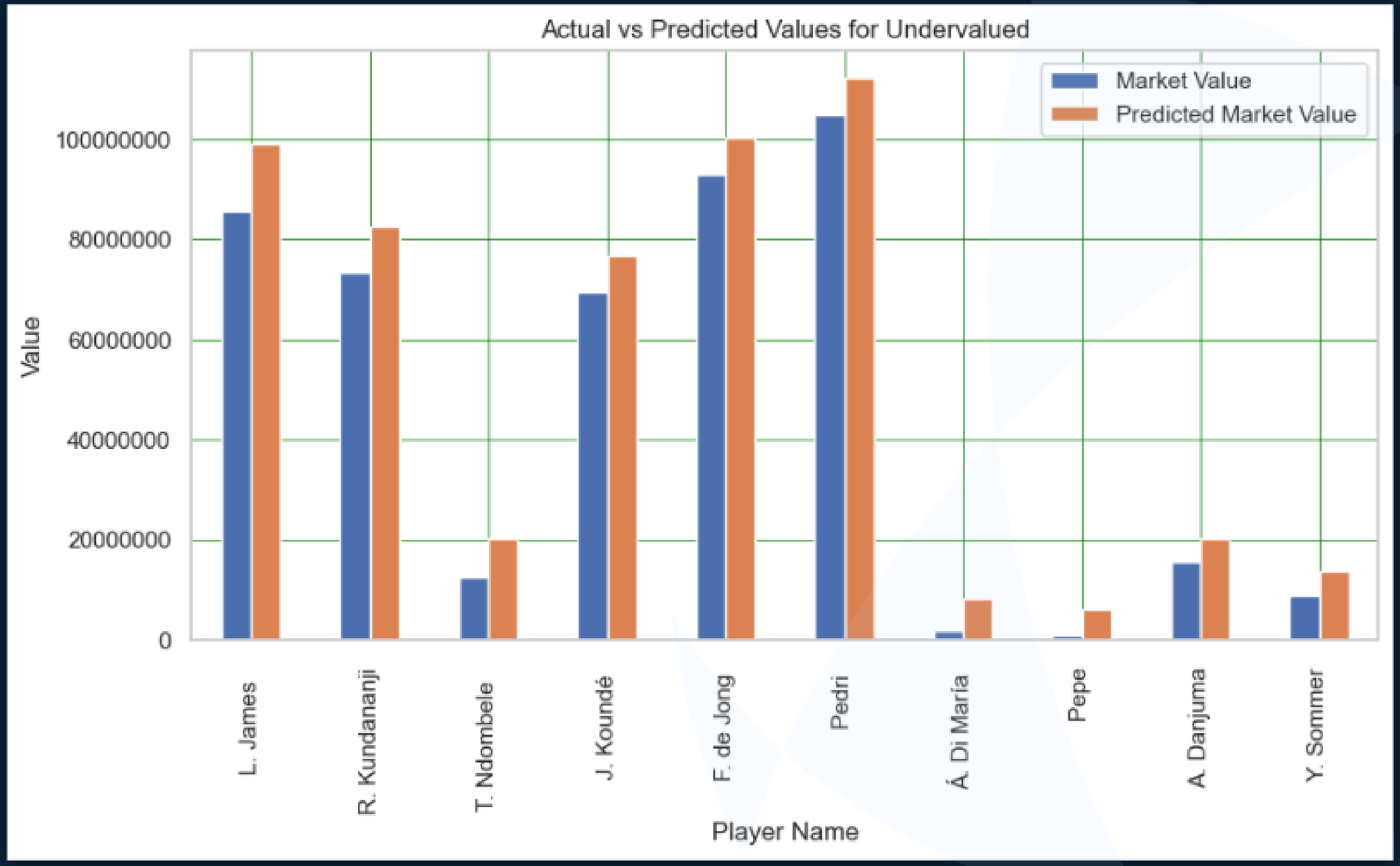
- Presents the majority of residuals clustered around the zero line

CREATE 2 NEW DATAFRAMES: UNDERVALUED_DF AND OVERVALUED_DF

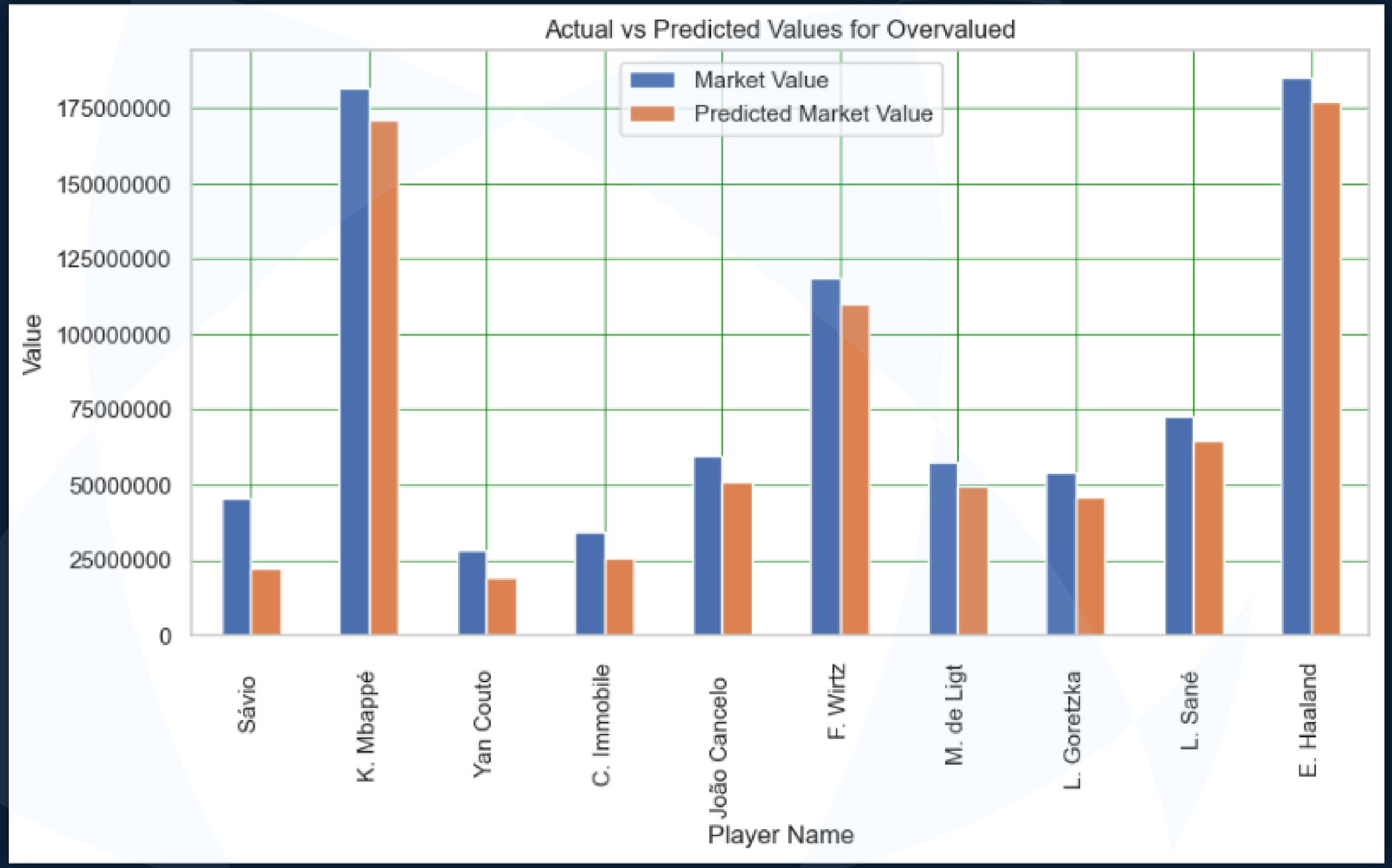
e	Overall Rating	Potential	Team	ID	Height	Weight	Foot	Best Overall	Best Position	...	Passing / Kicking	Dribbling / Reflexes	Defending / Pace	Physical / Positioning	Contract Start Date	Contract End Date	Contract Duration	Foot Category	Predicted Market Value	Difference
9	86	86	FC Barcelona	210514	182	74	Right	86	RB	...	85	84	80	72	NaN	NaN	NaN	1	12402544	47097456
9	80	88	Girona	270409	176	71	Left	81	CAM	...	75	83	30	53	NaN	NaN	NaN	0	1709868	43790132
3	81	87	FC Barcelona	242444	181	70	Right	85	CAM	...	78	85	40	67	NaN	NaN	NaN	1	5526983	35473017
3	81	85	Inter	239807	178	74	Right	83	CAM	...	78	81	70	71	NaN	NaN	NaN	1	2341508	34158492
0	84	84	Roma	192505	191	94	Left	84	ST	...	74	76	38	82	NaN	NaN	NaN	0	11366221	30133779
3	81	85	Borussia Dortmund	233049	180	76	Right	85	CAM	...	80	87	31	42	NaN	NaN	NaN	1	8169655	28830345
6	81	83	AFC Bournemouth	223197	187	78	Right	83	ST	...	70	81	40	80	NaN	NaN	NaN	1	3755594	28744406
0	82	82	Bayer 04 Leverkusen	224179	187	86	Right	82	ST	...	66	76	40	82	NaN	NaN	NaN	1	3322967	26677033
7	82	84	Arsenal	220901	183	75	Right	82	GK	...	86	82	57	81	NaN	NaN	NaN	1	3765645	24234355
1	78	85	Getafe	246147	181	70	Left	80	CAM	...	74	80	36	60	NaN	NaN	NaN	0	4875932	23624068

ge	Overall Rating	Potential	Team	ID	Height	Weight	Foot	Best Overall	Best Position	...	Passing / Kicking	Dribbling / Reflexes	Defending / Pace	Physical / Positioning	Contract Start Date	Contract End Date	Contract Duration	Foot Category	Predicted Market Value	Difference
26	87	90	FC Barcelona	228702	181	74	Right	89	CM	...	86	87	77	78	2019	2026	7	1	110302769	-17302769
35	84	84	Benfica	183898	180	69	Left	85	CAM	...	86	89	48	66	2023	2024	1	0	18268375	-16518375
34	90	90	FC Barcelona	188545	185	81	Right	90	ST	...	80	87	44	84	2022	2026	4	1	72088671	-14088671
31	86	86	Manchester United	200145	185	84	Right	86	CDM	...	78	72	87	87	2022	2026	4	1	50727727	-13227727
37	87	87	Real Madrid	177003	172	66	Right	87	CM	...	89	87	72	66	2012	2024	12	1	34049611	-12549611
30	84	84	Bayer 04 Leverkusen	199503	185	82	Left	84	CDM	...	82	74	76	82	2023	2028	5	0	31951212	-11951212
22	85	91	Real Madrid	243812	174	64	Right	87	CAM	...	79	86	31	64	2019	2028	9	1	97306996	-10806996
31	90	90	Real Madrid	192119	200	96	Left	90	GK	...	76	93	46	90	2018	2026	8	0	72199354	-9199354
31	85	85	Real Madrid	197445	180	78	Left	85	CB	...	83	80	85	77	2021	2026	5	0	45765442	-8765442
31	89	89	FC Barcelona	192448	187	85	Right	89	GK	...	89	91	47	86	2014	2028	14	1	63029200	-8529200

BAR CHART FOR UNDERVALUED PLAYERS



BAR CHART FOR OVERVALUED PLAYERS



WHAT WE COULD HAVE DONE DIFFERENTLY

INCORPORATE **REAL- TIME DATA**

REGULAR MODEL UPDATES: IMPLEMENT A PIPELINE TO PERIODICALLY UPDATE (RETRAIN MODEL?)

PERFORMANCE HISTORY: UTILISE DATA FROM PREVIOUS FIFA GAMES TO MEASURE LONG-TERM PERFORMANCE TRAJECTORY