Analysis on European Soccer League

James Ding DATA 604

Introduction

The project explored the relationship between soccer players' attributes and their overall rating.

I chose to work on this project due to my interest in sports betting and my aspiration to become a sports analyst. The project is important because sports betting is growing rapidly alongside data analytics, as evidenced by major partnerships between sports leagues and gaming operators like FanDuel and the NBA/WNBA. This intersection of sports and data analysis signifies a promising future for sports analytics professionals. For this report, the focus is on analyzing the attributes of 2016 European Soccer League players that contribute to their overall ratings.

Individual Datasets

I worked on this project individually, utilizing the 2008-2016 European Soccer League Database from Kaggle. Among the various datasets available, I selected the player and player_attributes datasets for analysis. These datasets were chosen because they offer insights into players' attributes and overall ratings, which are crucial for understanding player performance in soccer.

The European Soccer League data we obtained from Kaggle was recorded via live matches. According to Premier League, “Live data is collected by a three-person team covering each match. Two highly trained analysts use a proprietary video-based collection system to gather information on what happens every time a player touches the ball, which player it was and where on the pitch the action occurred.” (premierleague.com) In addition, players’ biometrics were obtained through routine medical assessments conducted by club medical staff and administrative processes during player registration.

During the initial stages of the project, I planned to explore both overall and potential ratings of players. However, it was realized that analyzing overall and potential ratings essentially repeats the same analysis, leading to a focus solely on factors influencing overall ratings for a more detailed examination.

Here are the variables of my player_attribute table: (Note, everything after defensive_work_rate is measured from a 1-100 scale, and this table has 39530 rows and 43 columns)

1. **player_api_id**: Another unique identifier for each player within the dataset.
2. **date**: The date when the player's attributes were recorded.
3. **overall_rating**: The overall skill rating of the player, typically based on their performance in various aspects of the game. It represents the player's overall ability on a scale of 1 to 100.
4. **potential**: This represents the player's potential skill level, indicating how much they can improve in the future. It's also rated on a scale of 1 to 100.
5. **preferred_foot**: Indicates whether the player is left-footed or right-footed.
6. **attacking_work_rate**: The level of effort the player puts into attacking plays, rated as low, medium, or high.
7. **defensive_work_rate**: The level of effort the player puts into defensive plays, also rated as low, medium, or high.
8. **crossing**: The ability of the player to deliver accurate crosses into the penalty area.
9. **finishing**: How proficient the player is at scoring goals.
10. **heading_accuracy**: The accuracy and power of the player's headers.
11. **short_passing**: The accuracy of short passes made by the player.
12. **volleys**: The skill level of the player when striking the ball in the air without it touching the ground.
13. **dribbling**: The player's ability to maneuver the ball while running.
14. **curve**: The amount of curve the player can apply to the ball when passing or shooting.
15. **free_kick_accuracy**: The accuracy of the player's free kicks.
16. **long_passing**: The accuracy of long-range passes made by the player.
17. **ball_control**: The player's ability to control the ball effectively.
18. **acceleration**: How quickly the player can reach their maximum speed.
19. **sprint_speed**: The player's top speed over short distances.
20. **agility**: The player's agility and ability to change direction quickly.
21. **reactions**: How quickly the player reacts to changes on the field.
22. **balance**: The player's ability to maintain balance while performing various movements.
23. **shot_power**: The power behind the player's shots.

24. **jumping**: The player's ability to jump vertically.
25. **stamina**: The player's endurance and ability to maintain high levels of performance throughout a match.
26. **strength**: The physical strength of the player, affecting their ability to win challenges and hold off opponents.
27. **long_shots**: The player's proficiency at shooting from long distances.
28. **aggression**: How aggressively the player approaches challenges and tackles.
29. **interceptions**: The player's ability to intercept passes and disrupt opponents' plays.
30. **positioning**: How well the player positions themselves on the field, both offensively and defensively.
31. **vision**: The player's ability to see and exploit spaces on the field, as well as anticipate teammates' movements.
32. **penalties**: The player's accuracy when taking penalty kicks.
33. **marking**: The player's ability to mark opponents closely and prevent them from receiving passes or making plays.
34. **standing_tackle**: The player's ability to make tackles while on their feet.
35. **sliding_tackle**: The player's ability to make sliding tackles to dispossess opponents.
36. **gk_diving**: Goalkeeper's diving ability to make saves.
37. **gk_handling**: Goalkeeper's ability to handle the ball securely after making a save.
38. **gk_kicking**: Goalkeeper's ability to kick the ball effectively, particularly in terms of distribution.
39. **gk_positioning**: Goalkeeper's positioning to cover the goal effectively.
40. **gk_reflexes**: Goalkeeper's reflexes in reacting to shots and making saves.

As for my individual milestone, I focused on the player_attribute table. I employed basic selection and filtering methods to identify top players with overall ratings of 85 or above. This initial step provided a good overview of the attributes associated with high-performing players in the league.

1. **Foot Dominance Analysis:** Utilized aggregation with groupby to find the average overall rating of left-footed and right-footed players, exploring whether dominant foot affects player ratings.

2. **Speed Analysis:** Employed filtering and ordering techniques to identify the fastest players in the league, recognizing the importance of speed as a critical attribute influencing overall player ratings.

3. **Performance Comparison:** Used subqueries and aggregation to identify high and low performers compared to the league average, analyzing common high scores on attributes contributing to higher overall ratings.

4. **Attack Work Rate and Soccer Skills:** Leveraged simple and specific selection queries to identify players with 'high' attack work rates and those with exceptional soccer skills such as dribbling and finishing above 80. These attributes were selected due to their significant impact on overall player performance and ratings.

Data Exploration

Data Preparation For my data, I first collected it from Kaggle in SQlite 3 format. I ran it in my jupyter notebook, and I picked out 2 tables named players and player_attribute. Then I converted them both into CSV files. Then I exported them to University of Calgary's database and proceeds to perform queries on MySQL Workbench. From there, I performed the SQL query analysis to understand the and eexplore ways of joining the table to my guiding questions, which was my individual milestone assignment listed above. After attempting many different type of joins between my 2 tables, I ultimately decided on inner join via the primary key `player_api_id`, because I was concerned with the completeness of my data and I wanted to focus on the common record. If I were to employ left outer join, I wouldn't be eliminating null values thus leaving me with many rows of incomplete data.

After joining the tables, I noticed that there were 5+ rows per player. Therefore, I have opted to retain only the most recent observation for each player within the timeframe of 2008 to 2016. This decision was made to mitigate the impact of the number of observations on individual players, ensuring that each player is represented by a single, latest data point. By doing so, I aim to focus solely on the most recent attributes of each player while disregarding any potential bias introduced by multiple observations over time.

I also made sure to remove any rows containing null values, and I created a new variable `age` by subtracting player's birthdate from date of observation. After that, I removed unnecessary columns that had no relevance in my project such as `id`, `fifa_player_id`, `date` and `birthday` (since I used these 2 to create `age` already). In the end, my dataset contained 2301 rows and 48 columns.

Guiding Question

1. How do common last names correlate with overall ratings?

2. What is the relationship between height/weight & overall ratings?

3. How do various goalkeeping ratings affect goalies' overall ratings?

4. Is there a significant difference in overall ratings between left-footed & right-footed players?

5. How do attacking/defensive work rate affect a player's overall ratings?

6. What is the correlation between game-IQ attributes (vision, marking, positioning) and players' overall ratings?

7. What about physical attributes (strength, stamina, speed)?

8. What about hard skills? (crossing, finishing, dribbling...)

9. Is there a prime age for playing soccer?

In order to answer my guiding questions, I have to rely on SQL queries to find out the information I wanted.

1. How do common last names correlate with overall ratings?

```
SELECT SUBSTRING_INDEX(player_name, ' ', -1) AS last_name,
       COUNT(*) AS name_count,
       AVG(overall_rating) AS avg_overall_rating
FROM COMBINED_TABLE
GROUP BY last_name
HAVING name_count >= 15
ORDER BY name_count DESC
Limit 15;
```

This query returned a table with players' last name, name counts, and their average overall rating. From there, I looked up each last name's country of origin and created another column and manually put it in.

surname	name_counts	origin	avg_overall_rating
Silva	45	Portugal, Spain, Brazil	70.1428571428571 4
Garcia	32	Spain	68.85714285714286
Rodriguez	28	Spain	70
Santos	27	Portugal, Spain, Brazil	69.66666666666667
Lopez	26	Spain	69.85714285714286
Fernandez	24	Spain	70.22222222222223
Gonzalez	22	Spain	70.625
Traore	20	Mali, Senegal	69.14285714285714
Costa	20	Portugal, Spain, Brazil	67.2
Perez	18	Spain	66.5
Martinez	16	Spain	71.66666666666667
Gomez	15	Spain	69.75
Smith	15	English, Irish	61.285714285714285

As we can see, Spanish origin is very common these European Soccer League players, I did a little calculation comparing the overall average rating of Spanish descent vs overall average in the league and it was 69.5 vs 67.9. This indicates Spanish descents are on average, the most talented players in the league in the year 2016.

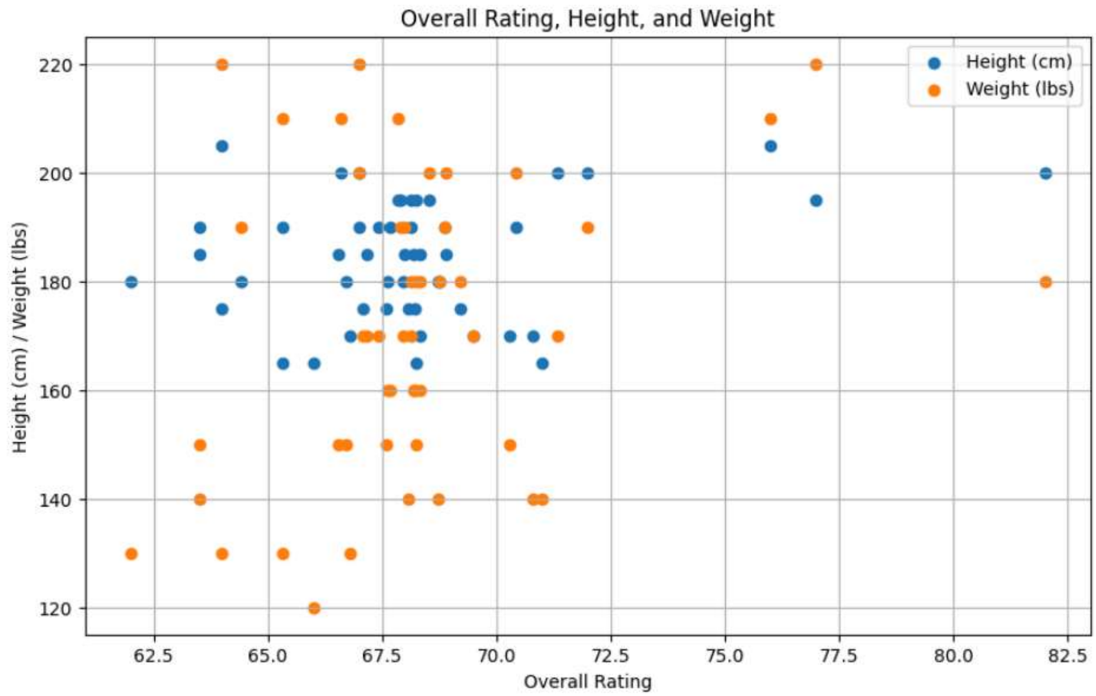
2. What is the relationship between height/weight & overall ratings?

For this question, we will look at players' height/weight and determine if there is a correlation between them.

```
SELECT
    ROUND(height / 5) * 5 AS rounded_height,
    ROUND(weight / 10) * 10 AS rounded_weight,
    AVG(overall_rating) AS avg_overall_rating
FROM Combined_table
GROUP BY rounded_height, rounded_weight
ORDER BY avg_overall_rating DESC;
```

Instead of comparing separately like in my minor deliverable, I decided it is better to put them together and create a scatter graph for visualization purpose. I also rounded the height to nearest 5cm and weight to nearest 10 lbs.

rounded_height	rounded_weight	avg_overall_rating
200	180	82
195	220	77
205	210	76
200	190	72
200	170	71.3333333333333
165	140	71
170	140	70.8125
190	200	70.4285714285714 3
170	150	70.2926829268292 7
170	170	69.5
175	180	69.2
185	200	68.9090909090909
190	190	68.8644067796610 1
180	180	68.7692307692307 7
180	140	68.7333333333333
195	200	68.5161290322580 6
185	180	68.3350253807106 6
170	160	68.3333333333333
165	150	68.25
195	180	68.24
175	160	68.21
185	160	68.1779661016949 2
195	170	68.1333333333333 4
190	180	68.1240875912408 8
175	140	68.0810810810810 8
185	190	67.9855072463768 1
180	170	67.9581395348837 3
195	190	67.9117647058823 5
195	210	67.8571428571428 6
190	160	67.6666666666666 7
180	160	67.6288659793814 4
175	150	67.5960264900662 2
190	170	67.4155844155844 2
185	170	67.1525423728813 6
175	170	67.0681818181818 1
190	220	67
200	200	67
170	130	66.7857142857142 9
180	150	66.6992481203007 5
200	210	66.6
185	150	66.5454545454545 5
165	120	66
165	130	65.3333333333333
190	210	65.3333333333333
180	190	64.4
175	130	64
205	220	64
185	140	63.5
190	150	63.5
180	130	62



As we can see from this graph, there are no correlations between height/weight and overall rating. We can make an argument for weight as it does somewhat positively correlate with overall rating, indicating holding your ground is very important in the game. However, solely based on the graph, I conclude that height/weight have no correlations with overall rating.

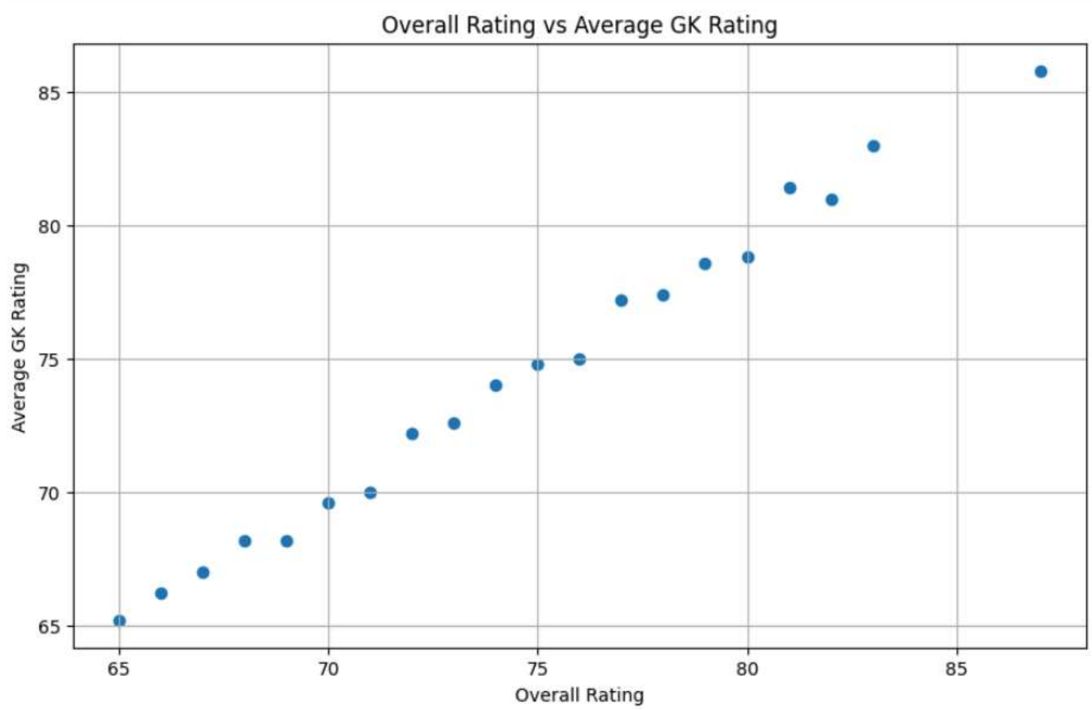
3. How do various goalkeeping ratings affect goalies' overall ratings?

This questions require we do a little filtering finding out who the goalkeepers are since we don't have a `position` column in our dataset. To do so, I set different benchmarks and then google players' names to find out what position they played. I found that by having all goal keeping attributes above 60 is the perfect separator betweeen a goalie and a non-goalie. There were 5 variables with the 'gk' prefix: `gk_diving`, `gk_handling`, `gk_kicking`, `gkk_positioning`, and `gk_reflexes`. By taking the average of these variables of an observation, we can compare it to their overall rating and see if there is a positive relationship.

```sql
SELECT
    overall_rating,
    (gk_diving + gk_handling + gk_kicking + gk_positioning +
gk_reflexes)/5 AS avg_gk_rating,
    player_name
FROM
    Combined_table
WHERE
    gk_diving > 60
    AND gk_handling > 60
    AND gk_kicking > 60
    AND gk_positioning > 60
    AND gk_reflexes > 60
```

```
GROUP BY
    overall_rating
ORDER BY
    overall_rating DESC
LIMIT 20;
```

overall_rating	avg_gk_rating	player_name
87	85.8	David De Gea
83	83	Bernd Leno
82	81	Anthony Lopes
81	81.4	Benoit Costil
80	78.8	Alphonse Areola
79	78.6	Ben Foster
78	77.4	Adrian
77	77.2	Anders Lindegaard
76	75	Apoula Edima Edel
75	74.8	Albano Benjamin Bizzarri
74	74	Allan McGregor
73	72.6	Adam Bogdan
72	72.2	Andres Palop
71	70	Alexandros Tzorvas
70	69.6	Adriano Facchini
69	68.2	Abdoulaye Diallo
68	68.2	Alan Mannus
67	67	Alberto Garcia
66	66.2	Anthony Mfa Meuzi
65	65.2	Anthony Moris



As we can see, we get a beautiful positive linear relationship here. Therefore, we can conclude that the average of all goal keeping statistics have a direct positive linear relationship with their overall rating, given the player is a goalie.

4. Is there a significant difference in overall ratings between left-footed & right-footed players?

```
SELECT
    preferred_foot,
    COUNT(*) AS player_count,
    AVG(overall_rating) AS avg_overall_rating
FROM
    Combined_table
GROUP BY
    preferred_foot;
```

preferred_foot	player_count	avg_overall_rating
left	563	67.93072824156306
right	1775	67.90422535211268

For this question, we examined which foot is these players' preferred foot and averaged out the players' overall rating by the preferred foot. It turned out that the average overall rating between the two is essentially the same. In conclusion, there are no significant difference in overall ratings between left-footed and right-footed players.

5. How do attacking/defensive work rate affect a player's overall ratings?

```
SELECT
    CASE
        WHEN attacking_work_rate IN ('low', 'medium', 'high') THEN
attacking_work_rate
        ELSE 'other'
    END AS attacking_work_rate_category,
    CASE
        WHEN defensive_work_rate IN ('low', 'medium', 'high') THEN
defensive_work_rate
        ELSE 'other'
    END AS defensive_work_rate_category,
    AVG(overall_rating) AS avg_overall_rating
FROM
    Combined_table
GROUP BY
    attacking_work_rate_category,
    defensive_work_rate_category;
```

attacking_work_rate_category	defensive_work_rate_category	avg_overall_rating
high	high	72.79245283018868
high	low	70.78260869565217
high	medium	69.82294264339153
low	high	68.95652173913044
low	low	69.5
low	medium	67.63414634146342
medium	high	69.05084745762711
medium	low	68.67333333333333
medium	medium	66.76103896103896
other	other	64.30281690140845

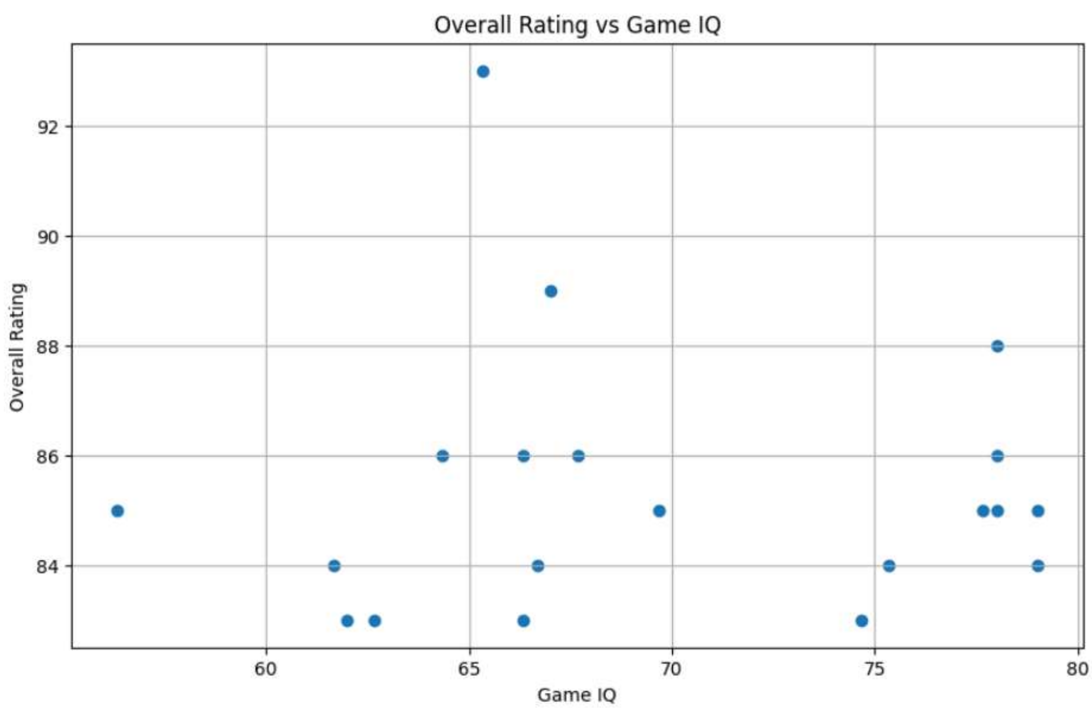
For this one, I took a different approach since `attacking_work_rate` and `defensive_work_rate` are the only variables in the data that are categorical. I queried for the combination of these 2 attributes and noticed something interesting. The low-low combination of players actually sitting in the middle in terms of average overall rating. This came as a surprise to me, it was kind of hard to believe if you are a lazy player on both end, you are still better than players that gave medium efforts on both end. Therefore, I concluded that `attacking_work_rate` and `defensive_work_rate` do not affect a player's overall rating.

6. What is the correlation between game-IQ attributes (vision, marking, positioning) and players' overall ratings?

For the next 3 guiding questions, I am taking a similar approach to all of them but I believe each section contributes important statistics to my project. First, I averaged out vision, marking, and positioning and compare it to players' overall rating to look for a positive relationship. For these queries, we are excluding goalies in case their overall rating is high.

```sql
SELECT
    overall_rating,
    (vision + marking + positioning) / 3 AS Game_IQ
FROM
    Combined_table
ORDER BY
    overall_rating DESC
LIMIT 20;
```

overall_rating	Game_IQ
93	65.3333333333333
89	67
88	78
86	64.3333333333333
86	67.6666666666667
86	66.3333333333333
86	78
85	79
85	78
85	77.6666666666667
85	69.6666666666667
85	56.333333333333336
84	61.666666666666664
84	66.6666666666667
84	79
84	75.3333333333333
83	66.3333333333333
83	74.6666666666667
83	62
83	62.666666666666664



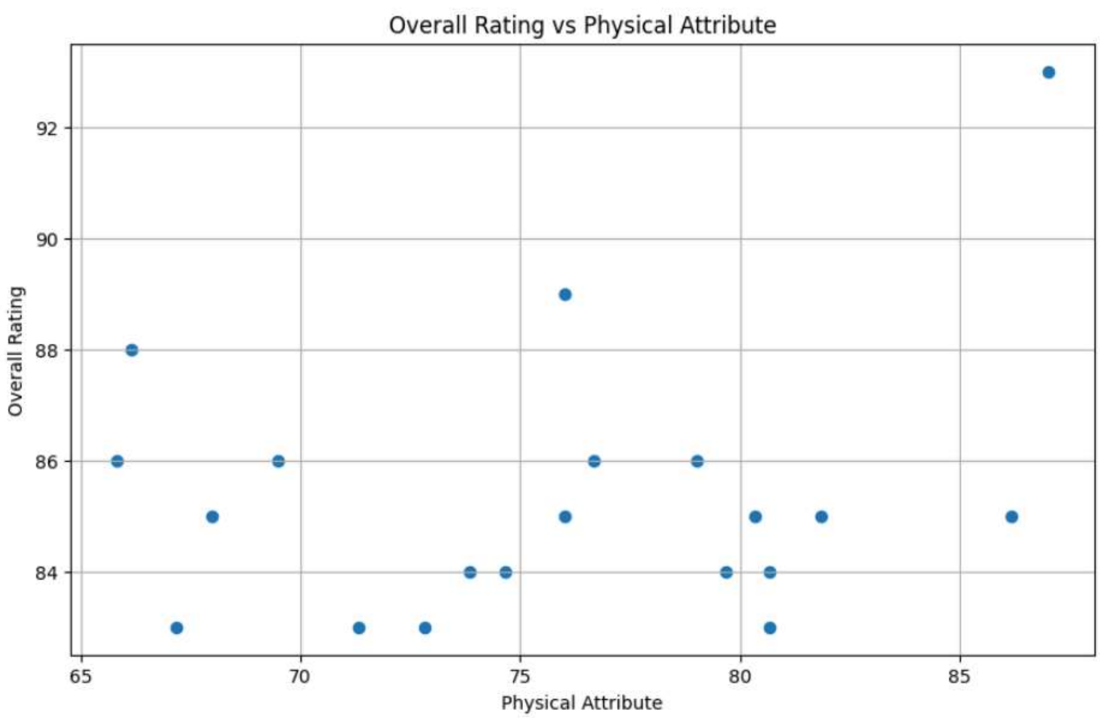
As we can observe from the data and the graph, I conclude that there is no apparent pattern between game IQ and overall rating. I suspect that these attributes are used for rating a defensive player, and top overall rating players may have been mainly forwards. I will continue on to research on this part after I finish this project.

7. What about physical attributes (strength, stamina, speed)?

```
SELECT
    overall_rating,
    (stamina + strength + (acceleration+sprint_speed) /2)/ 3 AS Physical
FROM
```

```sql
    Combined_table
WHERE
    gk_diving < 60
    AND gk_handling < 60
    AND gk_kicking < 60
    AND gk_positioning < 60
    AND gk_reflexes< 60
ORDER BY
    overall_rating DESC
LIMIT 20;
```

overall_rating	Physical
93	87
89	76
88	66.1666666666667
86	79
86	76.6666666666667
86	65.8333333333333
86	69.5
85	80.3333333333333
85	68
85	81.8333333333333
85	86.1666666666667
85	76
84	74.6666666666667
84	73.8333333333333
84	79.6666666666667
84	80.6666666666667
83	71.3333333333333
83	72.8333333333333
83	80.6666666666667
83	67.1666666666667

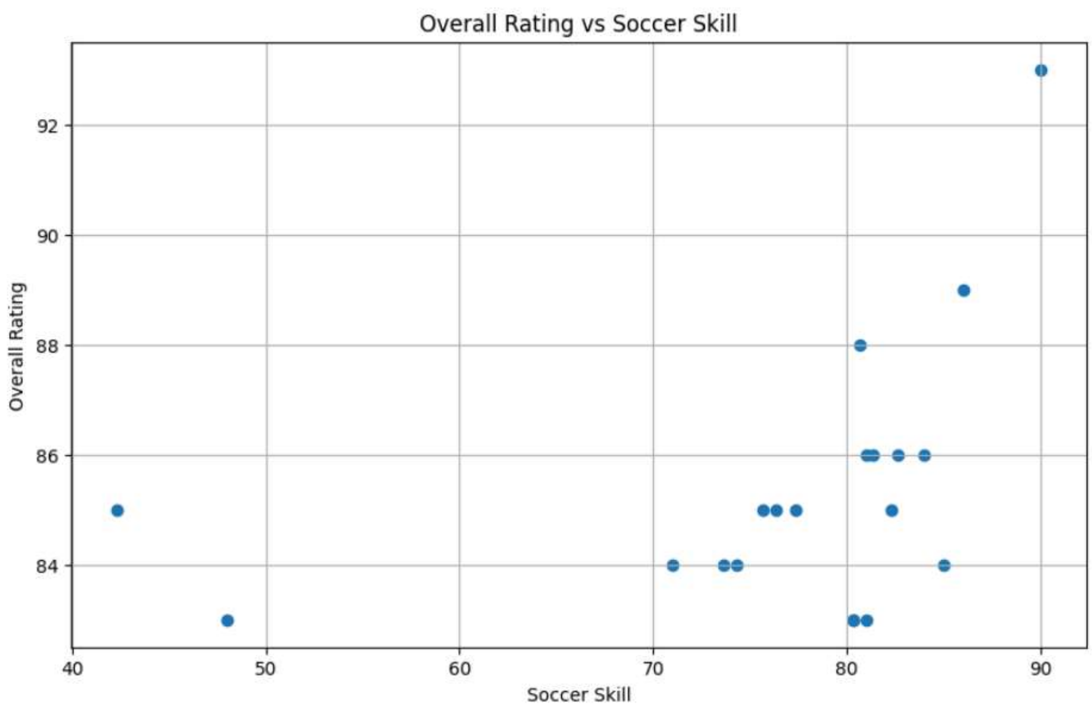


From the graphs above, we can concclude that most players lie between 73 to 81 without overall rating around 86. However, there is no apparent pattern in the graphs. However, if we start looking at 70 (physical attribute), we discovered that there is indeed a positive linear relationship. In practical, it makes sense that it takes at least a very talented soccer player to enter the league barring a few outliers. Therefore, I conclude that there is a positive linear relationship between physical and overall rating.

8. What about hard skills? (crossing, finishing, dribbling...)

```
SELECT
    overall_rating,
    (crossing+finishing+dribbling)/ 3 AS soccer_skill
FROM
    Combined_table
WHERE
    gk_diving < 60
    AND gk_handling < 60
    AND gk_kicking < 60
    AND gk_positioning < 60
    AND gk_reflexes< 60
ORDER BY
    overall_rating DESC
LIMIT 20;
```

overall_rating	soccer_skill
93	90
89	86
88	80.6666666666667
86	82.6666666666667
86	84
86	81.3333333333333
86	81
85	75.6666666666667
85	76.3333333333333
85	77.3333333333333
85	82.3333333333333
85	42.333333333333336
84	85
84	73.6666666666667
84	74.3333333333333
84	71
83	81
83	80.3333333333333
83	80.3333333333333
83	48



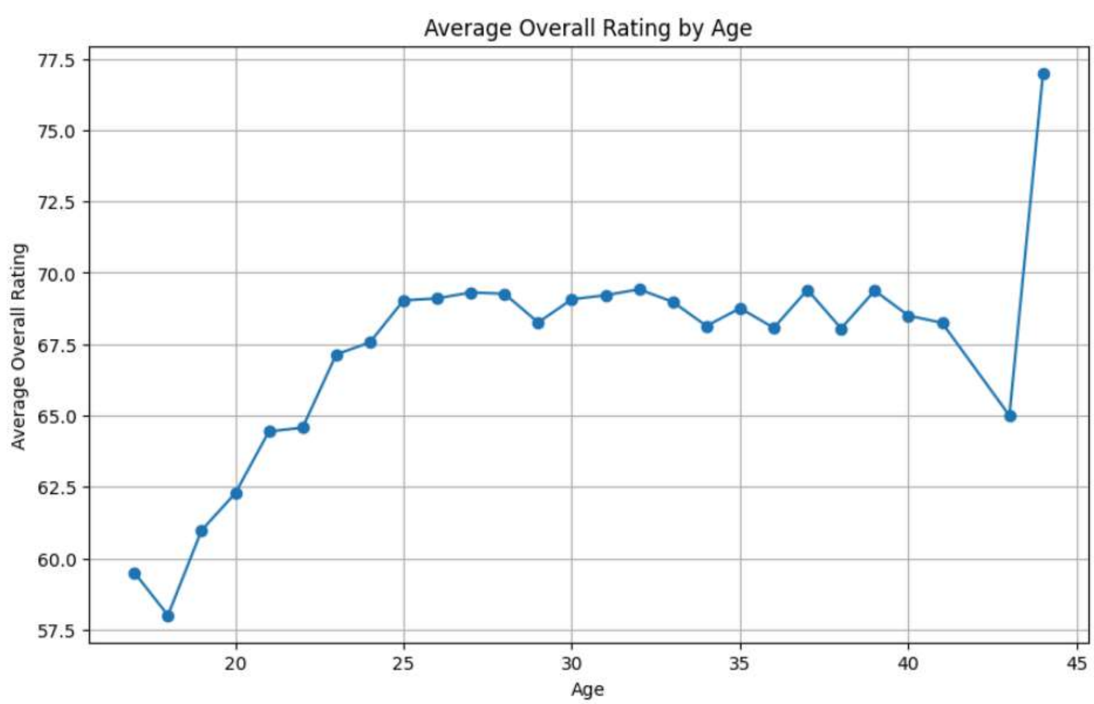
These graphs clearly depict starting at 70 soccer skills, a very clear linear positive relationship is present. Therefore, I conclude that there is a positive relationship between soccer skills and overall rating.

9. Is there a prime age for playing soccer?

```
SELECT
    Age,
    AVG(Overall_rating) AS AverageOverallRating
FROM
```

```sql
    combined_table
GROUP BY
    Age
ORDER BY
    Age DESC;
```

Age	AverageOverallRating
44	77
43	65
41	68.25
40	68.5
39	69.375
38	68.05263157894737
37	69.40625
36	68.07246376811594
35	68.76041666666667
34	68.13761467889908
33	68.98275862068965
32	69.42276422764228
31	69.21383647798743
30	69.07361963190183
29	68.2603550295858
28	69.26704545454545
27	69.3132530120482
26	69.10493827160494
25	69.03496503496504
24	67.5625
23	67.14018691588785
22	64.57462686567165
21	64.44086021505376
20	62.273809523809526
19	60.97959183673469
18	58
17	59.5



As we can observe from the graph, there is a steep learning curve when a player first enter the league, then they hit their prime from age 25 to age 40. Therefore, I conclude that age 25-40 is the pime age for any player in the European Soccer League.

Discussion

What's next? Exploring attributes like height, weight, and their relevance to specific positions in the game could unveil valuable insights. Similarly, delving deeper into the significance of attacking and defensive work rates may shed light on their impact, especially since it's counterintuitive for players displaying less effort to perform as well as those exerting themselves. It's plausible that players with different work rates might be compensating with varied skill sets or tactics. Moreover, investigating how these factors intersect with overall ratings could reveal nuanced relationships, contributing to a more comprehensive understanding of player performance in the league.

Peer Feedback Part C of the project's minor deliverable was exceptionally insightful. It prompted valuable feedback from my peers, which enriched my understanding and direction for the project. Two particular questions resonated with me deeply. Firstly, the observation that Brazilians excel in dribbling, Italians in defense, and Spaniards in passing was eye-opening. Further investigation, focusing on two exemplary players, confirmed this notion. For instance, considering the case of a player named Alexandre Silva, whose last name suggests a multi-country origin, his above-average overall rating of 67 was primarily attributed to his exceptional passing ability, which stood at an impressive 74. This evidence underscores the validity of the observation, suggesting a correlation between players' nationalities and their proficiency in specific skills. I googled him, he is Brazilian.

player_name	short_passing	long_passing	overall_rating	dribbling
David Silva	92	85	86	87
Adrien Silva	82	84	80	80
Bernardo Silva	80	76	79	83
Antonio da Silva	75	74	69	72
Andre Ramalho Silva	70	66	71	59
Damien Da Silva	69	51	75	42
Alex Silva	67	60	72	41
Alexandre Silva	67	48	69	74
Bruno Silva	64	63	66	66
Daniel De Silva	63	59	60	63
Augusto Da Silva	60	56	62	68
Adriano Pereira da ...	56	53	67	44
Andre Silva	49	40	70	66
Dany da Silva	22	24	55	25

Secondly, there was a question about carrying over my analysis to other sports. Exploring skill specialization across different sports reveals a diverse landscape where athletes excel in various abilities tailored to their respective games. In basketball, proficiency in agility, jumping, and shooting accuracy is paramount, with players like Stephen Curry exemplifying unparalleled shooting prowess. American football demands a multifaceted skill set, with quarterbacks like Tom Brady showcasing precision passing and strategic acumen. In baseball, hand-eye coordination and hitting power are crucial, exemplified by players such as Mike Trout. Meanwhile, racket sports like tennis prioritize agility, endurance, and precision, with athletes like Roger Federer showcasing mastery in serving, groundstrokes, and net play. While each sport presents unique challenges and skill requirements, the pursuit of excellence remains a common thread among athletes across disciplines.

In summary, in order to understand the attributes of a player in that particular sport, we must understand the sport. This question also raised more questions such as coaching, team strategy favoring their 'star' player. In addition, it raised questions about external factor such as play time, play time when a starter is injured, player needing rest (popular in NBA), first few games coming back from injury, emotional factor (such as bereavement, divorce (I suppose)), and many more. Finally, there are questions like what if a diamond in the rough begins to shine with given opportunities, can we get ahead of the curve and bet on that rising star before Vegas gets more information on him. The questions are endless, that's why I really enjoy this field and hope I can go further with this project.

Conclusion

In conclusion, examining the impact of player attributes on overall ratings reveals several noteworthy trends. Firstly, players of Spanish origin tend to exhibit higher overall ratings. Secondly, goalkeepers with superior ratings in goalkeeping skills often boast higher overall ratings. Additionally, there's a discernible pattern indicating that players aged between 25 and 40 typically reach the pinnacle of their careers, correlating with higher overall ratings. Finally, there's a positive correlation between a player's game intelligence, physical prowess, soccer skills, and their overall rating, suggesting that heightened proficiency in these areas contributes to an elevated overall rating.

Overall, I believe the analysis is fairly accurate. The part that was noteworthy is the attacking work rate and defensive work rate to be really unbelievable. We were always taught that effort pays off, but to have data telling us otherwise raises the debate of talent versus effort. In addition, I am surprised that people can still maintaining top form in tthe later stage of their 30s in the league. Finally, I believe there is more to be done to research in the biometrics of a player, although soccer may not be the sport, but weight and height directly correlates to muscle mass and that has to have an effect on the overall rating of a player.

Citation

1. Kaggle Dataset:
   - Hugo Mathien. "European Soccer Database." Kaggle, 2016, https://www.kaggle.com/hugomathien/soccer
2. Football-Data.co.uk:

- "Football Data." Football-Data.co.uk, https://www.football-data.co.uk/

3. Sofifa:

- "Sofifa." Sofifa.com, https://sofifa.com/

4. Premier League:

- "Premier League Stats." Premierleague.com, https://www.premierleague.com/stats/

5. Football Data API:

- "Football Data API." Football-data.mx-api.enetscores.com, http://football-data.mx-a

6. Google, for googling players' position

7. Chatgpt, for fixing my grammatical errors in the reporti.enetscores.com/
i.enetscores.com/

```
In [ ]:
```

```
In [ ]:
```