

BUILDING A FOOTBALL DATASET

Noâm BOUSSOUF

Efstathios CHATZILOZOS

Emmarius DELAR

PRESENTATION OVERVIEW

I. Objective dataset and motivations

II. Data acquisition and extraction

III. Post-processing

IV. Data storing

V. Conclusion

I. OBJECTIVE DATASET AND MOTIVATIONS

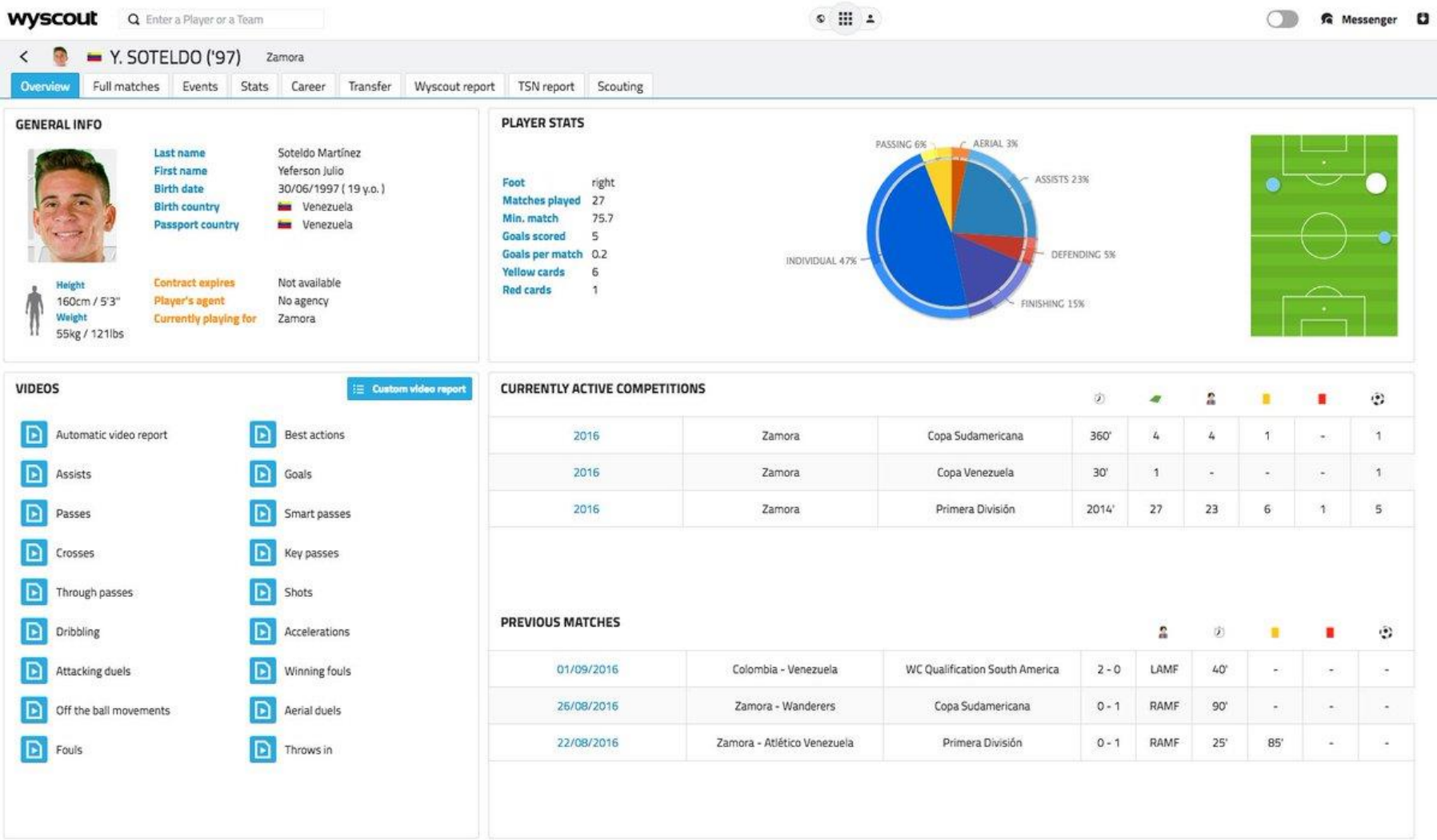
Data Science and analytics in football :

Analyze match performances and dynamics of a team, of players ...

Analyze a player's profile (scouting and recruitment, training optimization)

| | | |
|---|--|--|
|  1 Team & Match Performance With advanced analytics, teams leverage match player performance data, video footage, expected goals and assists, and data signals to understand and optimize their performance deeply. This is usually a more reliable indicator of how well a team performed than the score itself. |  2 Player Performance By using metrics calculated from event data like passes, tackles, and saves, teams can identify over or underperforming players, pinpoint areas for their improvement, and select the best players possible for a given gameplan or playstyle. |  3 Set-piece Optimization Improving the performance of corners and free kicks provides the quickest way to increase the number of goals scored. Spatial analysis can help determine the best approach for pieces, for example, in-swinging corners vs. out-swinging corners. |
|  4 Player Recruitment Choosing the next player signing is one of the biggest decisions for a team. With data, teams can identify players that the market undervalues and that have the exact skillset they're looking for. This in turn helps smaller teams take on well-funded bigger teams. |  5 Training Optimization Data can be used to decide where to concentrate training time and monitor player fitness. This helps keep players injury-free, avoid overtraining, and optimize training sessions to improve a player's weakness. |  6 Gameplanning & Strategy Using the same techniques to analyze match and player performance, teams can also analyze their opponent's performances and identify weaknesses. Data can be used as a tool to deliver tailored strategies and game plans for teams to use. |

I. OBJECTIVE DATASET AND MOTIVATIONS



II. WEB CRAWLING

Using scrapy:

- **Competitions Spider**
 - ✓ Extract detailed information on **various competitions (leagues)** across Europe.
 - ✓ Mapping of country codes to country names.

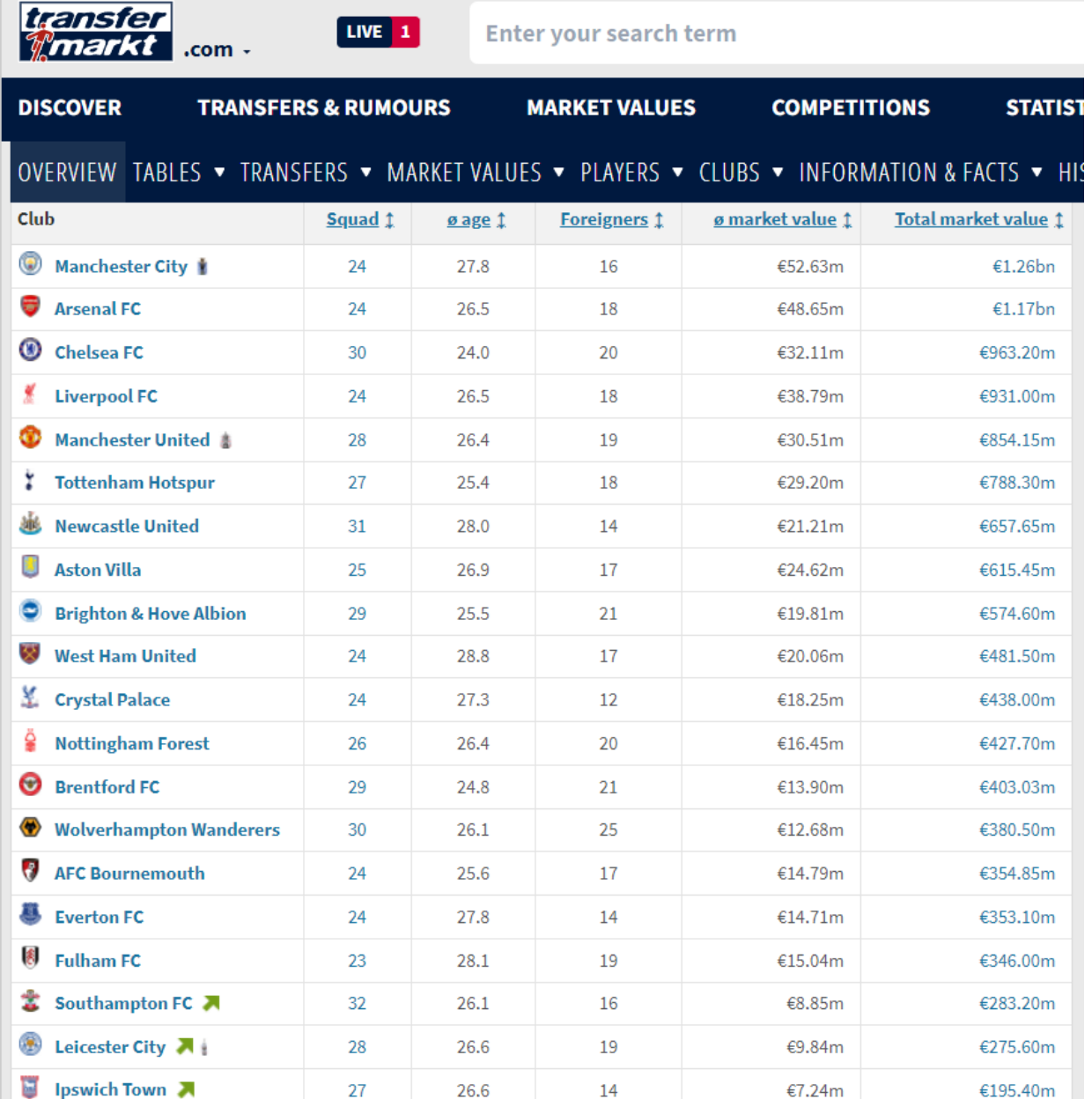
| transfermarkt.com | | | | | | | |
|--------------------|---------|---------------------|----------|---------------|--------------|--------------|---------------|
| DISCOVER | | TRANSFERS & RUMOURS | | MARKET VALUES | | COMPETITIONS | |
| Competition | Country | Clubs | Player ↑ | Avg. age ↓ | Foreigners ↑ | Forum | Total value ↑ |
| First Tier | | | | | | | |
| Premier League | | 20 | 533 | 26.5 | 66.6 % | | €11.75bn |
| LaLiga | | 20 | 512 | 27.1 | 41.2 % | | €5.48bn |
| Serie A | | 20 | 581 | 26.3 | 65.4 % | | €4.95bn |
| Bundesliga | | 18 | 505 | 26.2 | 52.9 % | | €4.48bn |
| Ligue 1 | | 18 | 489 | 25.7 | 61.3 % | | €3.59bn |
| Liga Portugal | | 18 | 492 | 25.8 | 62.6 % | | €1.62bn |
| Eredivisie | | 18 | 504 | 24.8 | 50.2 % | | €1.32bn |
| Süper Lig | | 19 | 562 | 26.5 | 46.4 % | | €1.15bn |
| Jupiler Pro League | | 16 | 434 | 24.8 | 56.9 % | | €975.25m |
| Premier Liga | | 16 | 439 | 26.0 | 40.1 % | | €914.05m |
| Super League 1 | | 14 | 429 | 27.1 | 59.7 % | | €515.63m |
| Bundesliga | | 12 | 347 | 24.6 | 44.1 % | | €446.68m |
| Premier Liga | | 16 | 427 | 25.4 | 21.1 % | | €390.85m |
| Super League | | 12 | 355 | 25.1 | 58.9 % | | €378.40m |
| Superliga | | 12 | 325 | 24.9 | 48.9 % | | €373.88m |
| Chance Liga | | 16 | 460 | 25.9 | 30.4 % | | €347.30m |
| Premiership | | 12 | 320 | 26.2 | 63.8 % | | €326.20m |
| Eliteserien | | 16 | 440 | 25.2 | 28.6 % | | €285.30m |
| Allsvenskan | | 16 | 466 | 25.4 | 37.1 % | | €275.54m |
| Super liga Srbije | | 16 | 488 | 25.0 | 30.3 % | | €267.48m |
| Ekstraklasa | | 18 | 527 | 25.4 | 40.2 % | | €260.73m |
| SuperLiga | | 16 | 475 | 26.5 | 40.0 % | | €232.87m |
| SuperSport HNL | | 10 | 312 | 24.9 | 35.9 % | | €232.74m |
| efbet Liga | | 16 | 402 | 26.0 | 49.3 % | | €183.46m |

Figure1. Competitions Page (European Competitions)

II. WEB CRAWLING

Using scrapy:

- **Competitions Spider**
 - ✓ Extract detailed information on **various competitions (leagues)** across Europe.
 - ✓ Mapping of country codes to country names.
- **Clubs Spider**
 - ✓ Crawl each **competition** crawled by the competitions spider.
 - ✓ Gather information **about each club** of a competition.



| Club | Squad ↑ | Age ↑ | Foreigners ↑ | Market value ↑ | Total market value ↑ |
|-------------------------|---------|-------|--------------|----------------|----------------------|
| Manchester City | 24 | 27.8 | 16 | €52.63m | €1.26bn |
| Arsenal FC | 24 | 26.5 | 18 | €48.65m | €1.17bn |
| Chelsea FC | 30 | 24.0 | 20 | €32.11m | €963.20m |
| Liverpool FC | 24 | 26.5 | 18 | €38.79m | €931.00m |
| Manchester United | 28 | 26.4 | 19 | €30.51m | €854.15m |
| Tottenham Hotspur | 27 | 25.4 | 18 | €29.20m | €788.30m |
| Newcastle United | 31 | 28.0 | 14 | €21.21m | €657.65m |
| Aston Villa | 25 | 26.9 | 17 | €24.62m | €615.45m |
| Brighton & Hove Albion | 29 | 25.5 | 21 | €19.81m | €574.60m |
| West Ham United | 24 | 28.8 | 17 | €20.06m | €481.50m |
| Crystal Palace | 24 | 27.3 | 12 | €18.25m | €438.00m |
| Nottingham Forest | 26 | 26.4 | 20 | €16.45m | €427.70m |
| Brentford FC | 29 | 24.8 | 21 | €13.90m | €403.03m |
| Wolverhampton Wanderers | 30 | 26.1 | 25 | €12.68m | €380.50m |
| AFC Bournemouth | 24 | 25.6 | 17 | €14.79m | €354.85m |
| Everton FC | 24 | 27.8 | 14 | €14.71m | €353.10m |
| Fulham FC | 23 | 28.1 | 19 | €15.04m | €346.00m |
| Southampton FC | 32 | 26.1 | 16 | €8.85m | €283.20m |
| Leicester City | 28 | 26.6 | 19 | €9.84m | €275.60m |
| Ipswich Town | 27 | 26.6 | 14 | €7.24m | €195.40m |

Figure 2. Clubs Page (Premier League)

II. WEB CRAWLING

Using scrapy:

- **Competitions Spider**
 - ✓ Extract detailed information on **various competitions (leagues)** across Europe.
 - ✓ Mapping of country codes to country names.
- **Clubs Spider**
 - ✓ Crawl each **competition** crawled by the competitions spider.
 - ✓ Gather information **about each club** of a competition.
- **Players Spider**
 - ✓ Crawl each **club** crawled by the clubs spider.
 - ✓ Collect information **about each player** of a club.

































| transfermarkt.com | | | | |
|---|--|---------------------|--|----------------|
| LIVE 1 | | | | |
| Enter your search term | | | | |
| DISCOVER TRANSFERS & RUMOURS MARKET VALUES COMPETITIONS STATISTICS | | | | |
| OVERVIEW SQUAD ▾ FIXTURES ▾ TRANSFERS & RUMOURS ▾ INFORMATION & FACTS ▾ STADIUM ▾ HISTORY | | | | |
| # ↑ | Player ↓ | Date of birth/Age ↑ | Nat. | Market value ↓ |
| 1 |  Alisson Goalkeeper | Oct 2, 1992 (32) |  | €28.00m |
| 62 |  Caoimhín Kelleher Goalkeeper | Nov 23, 1998 (26) |  | €20.00m |
| 56 |  Vitezslav Jaros Goalkeeper | Jul 23, 2001 (23) |  | €5.00m |
| 5 |  Ibrahima Konaté  | May 25, 1999 (25) |   | €45.00m |
| 4 |  Virgil van Dijk  | Jul 8, 1991 (33) |   | €30.00m |
| 2 |  Joe Gomez Centre-Back | May 23, 1997 (27) |   | €28.00m |
| 78 |  Jarell Quansah Centre-Back | Jan 29, 2003 (21) |   | €22.00m |
| 26 |  Andrew Robertson Left-Back | Mar 11, 1994 (30) |  | €30.00m |
| 21 |  Konstantinos Tsimikas Left-Back | May 12, 1996 (28) |  | €22.00m |
| 66 |  Trent Alexander-Arnold Right-Back | Oct 7, 1998 (26) |  | €70.00m |
| 84 |  Conor Bradley  | Jul 9, 2003 (21) |  | €15.00m |
| 38 |  Ryan Gravenberch Defensive Midfield | May 16, 2002 (22) |   | €40.00m |

Figure 3. Players Page (Liverpool FC)

II. PLAYER DATA ENRICHMENT - WIKIPEDIA

Enrich players with Wikipedia Bios:

- For each player:
 - Construct the Wikipedia URL based on the player's name
 - Use Beautiful Soup Python Library to parse HTML content
 - Extract the 1st meaningful paragraph (up to 50 words).
 - Append the fetched bio to the rest of the player's data

Cole Palmer

[51 languages](#)[Article](#) [Talk](#)[Read](#) [View source](#) [View history](#) [Tools](#)

From Wikipedia, the free encyclopedia

Cole Jermaine Palmer (born 6 May 2002) is an English professional [footballer](#) who plays as an [attacking midfielder](#) or [winger](#) for [Premier League](#) club [Chelsea](#) and the [England national team](#). Known for his dribbling, playmaking and composure, he is regarded as one of the best young players in the world.^[9]

An [academy graduate](#) of [Manchester City](#), Palmer made his senior debut for the club in 2020, and was later part of their squad that won a [continental treble](#) of the Premier League, [FA Cup](#), and [UEFA Champions League](#) in 2023. He signed for Chelsea in 2023 for a fee of £40 million and had a breakout debut season in which he was rewarded with several awards, including both the [PFA Fans' Player of the Year](#) and the [Young Player of the Year](#).^[10]

Palmer has represented England across [various youth levels](#), including winning the [2023 UEFA European Under-21 Championship](#), before making his senior debut in the same year. He represented his country at [UEFA Euro 2024](#), scoring the equalising goal in [the final](#).

Early and personal life

Cole Palmer



Palmer playing for [Manchester City](#) in 2023

Personal information

Figure 4. Player's Wikipedia (Cole Palmer) – Selection in red rectangle

II. DATASETS

1. Players Table

| name | team | age | position | country | number | value | href | bio |
|-------------------|------------------|------|------------------|---------|--------|------------|---|---|
| Azzedine Ounahi | Panathinaikos FC | 24.0 | Central Midfield | Morocco | 8.0 | 12000000.0 | /olympique-marseille/kadernaechstesaison/verei... | Azzedine Ounahi(Arabic:أوناحي عز الدين,pronoun... |
| Tetê | Panathinaikos FC | 24.0 | Right Winger | Brazil | 10.0 | 10000000.0 | /galatasaray/startseite/verein/141/saison_id/2024 | Mateus Cardoso Lemos Martins(born 15 February ... |
| Facundo Pellistri | Panathinaikos FC | 22.0 | Right Winger | Uruguay | 28.0 | 10000000.0 | /manchester-united/startseite/verein/985/saiso... | Facundo Pellistri Rebollo(Spanish pronunciatio... |
| Fotis Ioannidis | Panathinaikos FC | 24.0 | Centre-Forward | Greece | 7.0 | 18000000.0 | /fotis-ioannidis/profil/spieler/532444 | Fotis Ioannidis(Greek: Φώτης Ιωαννίδης; born 1... |

2. Clubs Table

| league | name | href | value |
|----------------|-----------------|---|-------------|
| Premier League | Manchester City | https://www.transfermarkt.com/manchester-city/... | 12600000000 |
| Premier League | Arsenal FC | https://www.transfermarkt.com/arsenal-fc/kader... | 11700000000 |
| Premier League | Chelsea FC | https://www.transfermarkt.com/chelsea-fc/kader... | 9632000000 |
| Premier League | Liverpool FC | https://www.transfermarkt.com/liverpool-fc/kad... | 9310000000 |

3. Competitions Table

| competition_type | competition_name | country | number_of_clubs | number_of_players | average_age | foreigners_percentage | total_market_value | competition_url |
|-------------------------|------------------|----------|-----------------|-------------------|-------------|-----------------------|--------------------|---|
| European leagues & cups | LaLiga | Spain | 20 | 510 | 27.1 | 41.0 | 5480000000 | https://www.transfermarkt.com/laliga/startseit... |
| European leagues & cups | Serie A | Italy | 20 | 581 | 26.3 | 65.4 | 4950000000 | https://www.transfermarkt.com/serie-a/startsei... |
| European leagues & cups | Bundesliga | Germany | 18 | 505 | 26.1 | 53.1 | 4480000000 | https://www.transfermarkt.com/bundesliga/start... |
| European leagues & cups | Ligue 1 | France | 18 | 492 | 25.6 | 61.2 | 3580000000 | https://www.transfermarkt.com/ligue-1/startsei... |
| European leagues & cups | Liga Portugal | Portugal | 18 | 493 | 25.8 | 62.5 | 1620000000 | https://www.transfermarkt.com/liga-portugal/st... |

III. POST-PROCESSING

- Correct inconsistencies : Standardize the formatting (e.g : "€250k" and "€1.17bn") and spelling ("Ligue1"/"Ligue 1") of data across sources → Normalization (e.g 250 000 and 1 170 000 000)
- Correct data types issues, handle duplicates
- Handle missing values and anomalies in the data (especially the data crawled from the Wikipedia API)

| name | team | age | position | country | number | value | href | bio |
|---------------|---------------------|-----|--------------|------------|--------|-------|--------------------------------------|---|
| Igor Ivanovic | Tobol Kostanay | 27 | Right Winger | Serbia | 10 | €500k | /igor-ivanovic/profil/spieler/327611 | Igor Ivanović(Serbian Cyrillic: Игор Ивановић;... |
| Igor Ivanovic | Buducnost Podgorica | 34 | Right Winger | Montenegro | 7 | €275k | /igor-ivanovic/profil/spieler/192417 | Igor Ivanović(Serbian Cyrillic: Игор Ивановић;... |

| name | team | age | position | country | number | value | href | bio |
|------------|---------------|-----|------------|---------|--------|-------|-----------------------------------|--------------------|
| Vaná Alves | Aris Limassol | 33 | Goalkeeper | Brazil | 1 | €400k | /vana-alves/profil/spieler/212198 | No wiki bio found. |

III. POST-PROCESSING

- Correct the bio obtained using Wikipedia API

2 approaches :

- Using library "re" : add spaces after punctuation, capitalize the first letter of each sentence, ensure there is space before brackets, remove citations...

Original : Connor Lambert Goldson(born 18 December 1992) is an English professionalfootballerwho plays forAris Limassol. His preferred position is atcentre-back, although he has also been utilised atright-back,[4]and as acentral midfielder.[5][6]

Output : Connor Lambert Goldson (born 18 December 1992) is an English professionalfootballerwho plays for Aris Limassol. His preferred position is atcentre-back, although he has also been utilised atright-back, and as acentral midfielder.

III. POST-PROCESSING

- Correct the bio obtained using Wikipedia API

Second approach :

- Using a pretrained LLM (T5 loaded from Hugging Face) :

Original : Connor Lambert Goldson(born 18 December 1992) is an English professionalfootballerwho plays forAris Limassol. His preferred position is atcentre-back, although he has also been utilised atright-back,[4]and as acentral midfielder.[5][6]

Output : Connor Lambert Goldson (born 18 December 1992) is an english professional footballer who plays for Aris Limassol . his preferred position is at centre-back, although he has also been utilised as acentral midfielder .

Original : Steeve Farid Yago(born 16 December 1992) is a professionalfootballerwho plays as adefenderforCypriot First DivisionclubAris Limassol. Born in France, he represents theBurkina Faso national team.[2]

Output : Born in France, he represents theBurkina Faso national team.

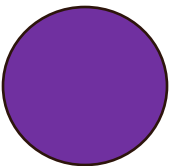
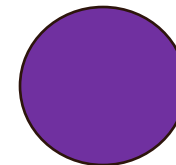
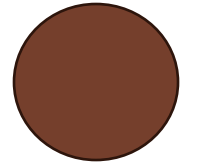
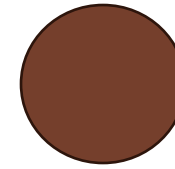
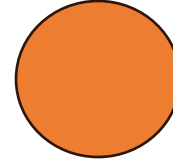
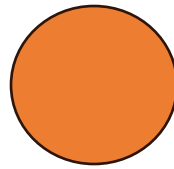
IV. DATA STORING

- A - Relational approach
- B - Practical example
- C - Additional Considerations

IV. DATA STORING

A – RELATIONAL APPROACH

- **Eco-system**
- Manage workflow
- Query examples



IV. DATA STORING

A – RELATIONAL APPROACH

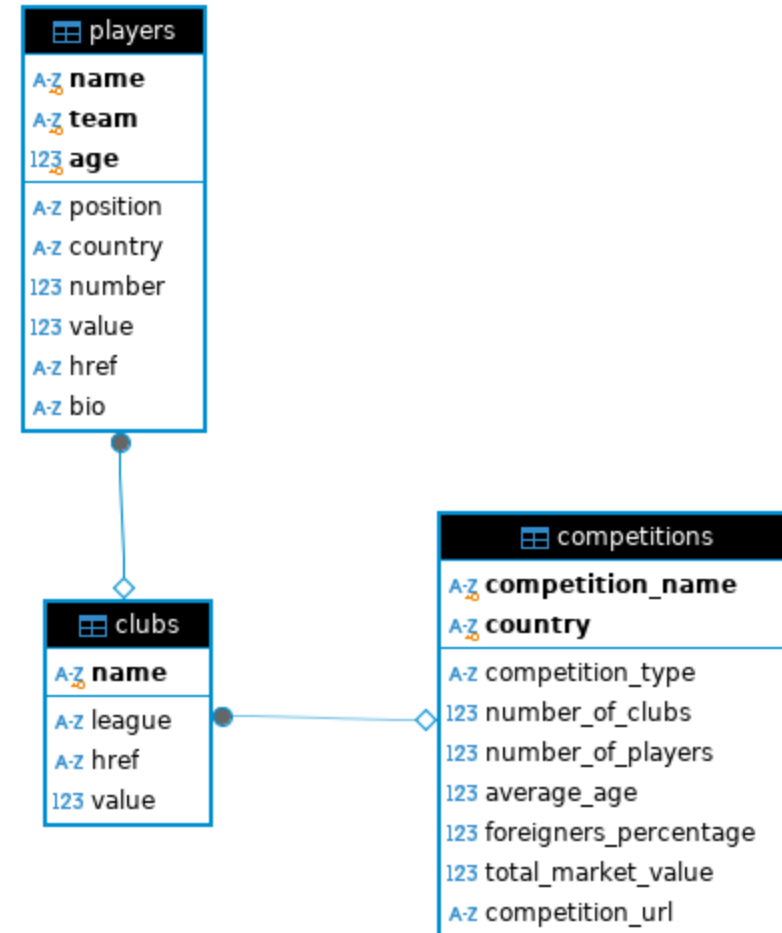
- **Eco-system**
- Manage workflow
- Query examples



IV. DATA STORING

A – RELATIONAL APPROACH

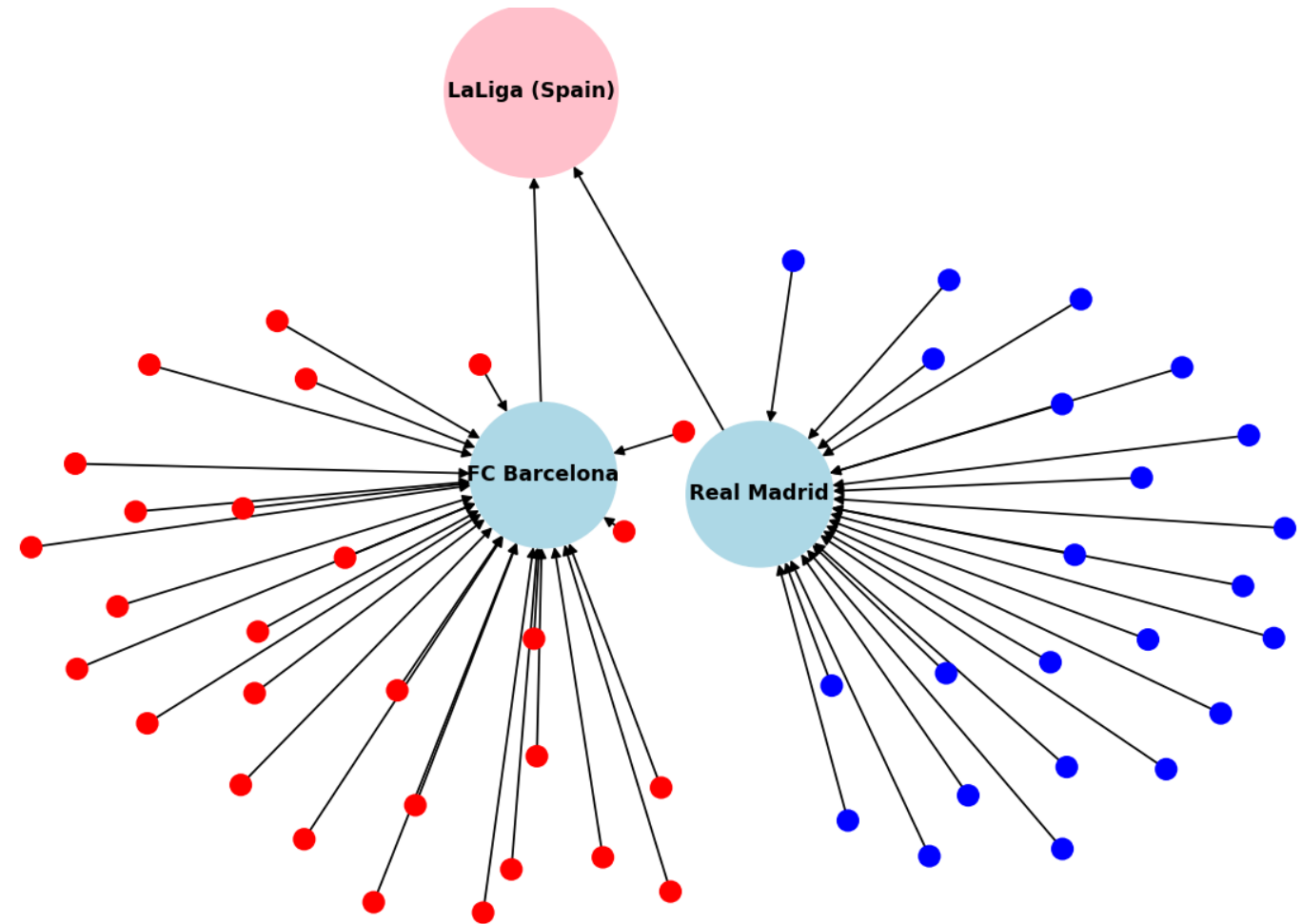
- **Eco-system**
- Manage workflow
- Query examples



IV. DATA STORING

A – RELATIONAL APPROACH

- Eco-system
- **Manage workflow**
- Complex queries



IV. DATA STORING

A – RELATIONAL APPROACH

- Eco-system
- Manage workflow
- **Query examples**

IV. DATA STORING

B – SQL EXAMPLE

inloop.github.io/sqlite-viewer/



SQLite Viewer

view sqlite file online



Drop file here to load content or click on this box to open file dialog.

clubs (683 rows)



Export ▾

```
SELECT players.name, (competitions.number_of_players - 1) AS Nb_concurrent
FROM 'players'
  JOIN 'clubs' ON players.team = clubs.name
  LEFT JOIN 'competitions' ON clubs.league = competitions.competition_name
WHERE players.name = 'Cole Palmer'
```

Execute


| name | Nb_concurrent |
|-------------|---------------|
| Cole Palmer | 532 |

Q: How many competitors does Cole Palmer have in his league?

IV. DATA STORING

B – SQL EXAMPLE

inloop.github.io/sqlite-viewer/

 **SQLite Viewer**
view sqlite file online

Drop file here to load content or click on this box to open file dialog.

clubs (683 rows) Export

```
SELECT clubs.league, SUM(players.value) AS total_MOC_value
FROM 'players' JOIN 'clubs' ON players.team = clubs.name
WHERE players.position = 'Attacking Midfield'
GROUP BY clubs.league
ORDER BY total_MOC_value DESC
LIMIT 6
```

Execute

| league | total_MOC_value |
|----------------|-----------------|
| Premier League | 1123000000 |
| Bundesliga | 680400000 |
| LaLiga | 531200000 |
| Serie A | 292825000 |
| Ligue 1 | 235350000 |
| Premier Liga | 197225000 |

Q: Rank the top 6 competitions by the total value of all 'Attacking Midfielders'

IV. DATA STORING

ADDITIONAL CONSIDERATIONS

Primary Keys Choice:

- Competitions Table (**competition_name, country**)
- Clubs Table (**name**)
- Players Table (**name, team, age**)

*The Primary Keys were chosen only after verifying that they ensure record uniqueness through our notebook **duplicates_and_unique_ids.ipynb**.*

Nullable Fields:

- Player's **value** is set as nullable to accommodate undisclosed or low market values.
- Player's Wikipedia **bio** is set as nullable to handle cases of non-existent or non-found Wikipedia pages.

Quality Assesment of the final Dataset:

- **Duplicate Checking:** No duplicate entries assurances through our notebook **duplicates_and_unique_ids.ipynb**.
- **Data Correctness:** Manual Sampling methods to verify the accuracy of our data.

THANK YOU FOR YOUR
ATTENTION

ANY QUESTIONS ?