Data Analysis Report – Player Popularity Project (PSV Eindhoven)

By Georgi Fidanov OLS3

Table of Contents

Data Analysis Report – Player Popularity Project (PSV Eindhoven)..... 0
Table of Contents..... 1
Version Control..... 2
Introduction..... 3
Research Question..... 4
Data Preparation..... 5
Sources..... 5
Data Cleaning..... 6
Files Ignored..... 7
Data Analysis..... 8
Market Value vs Age..... 8
Goals & Assists by Position..... 9
Competitor Followers..... 10
News Sentiment Over Time..... 11
Distribution of Likes on Posts..... 11
Engagement by Source..... 12
Gaps & Recommendations..... 14
Missing Data..... 14
Recommendations..... 14
What We Might Still Need..... 15
1. Player Metadata..... 15
2. Performance Data (on-field)..... 15
3. Popularity Signals (off-field)..... 15
4. News & Media Coverage..... 15
5. Sponsorship & Brand Links..... 15
Conclusion..... 16

Version Control

Version	Date	Author	Amendments	Status
0.1	02.10	Georgi Fidanov	Initial research	Finished
0.2	06.10	Georgi Fidanov	Document restructuring	Finished

Introduction

This project is a collaboration between **Citrics Labs** (data partner) and **PSV Eindhoven** (client). PSV wants to better understand and forecast player popularity. Popularity is influenced by performance, social media, news sentiment, and external events. Without structured data analysis, PSV risks missing opportunities in **sponsorships, marketing, and fan engagement**.

My task in this project was to:

- Analyze the datasets provided by Citrics Labs.
- Identify quality issues.
- Prepare cleaned and structured data.
- Provide insights through visualizations.
- Suggest what additional data is needed for reliable forecasting.

Research Question

Main Question:
How can PSV measure and forecast a player's popularity using data-driven methods?

Sub-questions:

1. What datasets are available, and what quality issues exist?
2. How can the data be cleaned and structured for analysis?
3. What insights can be drawn from existing visualizations?
4. Which additional data would improve the accuracy of popularity forecasting?

Data Preparation

Sources

The datasets were provided by **Citrics Labs** and contained both **structured** and **unstructured** files. These covered a wide range of information such as player performance, social media activity, news sentiment, and fan engagement.

Structured sources included:

- **Transfermarket.csv** – player-level data including *Name, Position, Age, Market Value (EUR), Goals, Assists, Performance Score*.
- **combined_competitor_profiles.csv** – competitor club profile statistics across platforms.
- **news_topics-sentiment-actions.csv** – news headlines, topics, actionable insights, and sentiment percentages.
- **socials_topics-sentiment-actions.csv** – social media comments with sentiment scoring and suggested actions.
- **socials-posts-overview.csv** – post-level social media statistics such as hashtags, likes, comments, and shares.

Unstructured sources included:

- **facebook-comments.csv** – Facebook user comment data linked to PSV posts.
- **facebook-competitor-profiles.csv** – competitor club pages on Facebook with metadata such as followers, likes, verification status.
- **facebook-posts_overview.csv** – overview of PSV's Facebook posts.
- **instagram-comments.csv** – Instagram comment data linked to posts.
- **instagram-posts_overview.csv** – Instagram posts with metadata such as location, content type, followers, engagement.
- **tiktok-comments.csv** – TikTok comment data linked to PSV's videos.

- **tiktok-posts_overview.csv** – TikTok video metadata including digg counts, shares, play counts, and creator information.
- **Google_headlines.csv** – PSV-related headlines, URLs, keywords, and publication data.
- **Eventregistry_headlines.csv** – event-driven headlines with topics, summaries, and external URLs.
- **youtube-comments.csv** – unstructured fan comments from YouTube videos.

Data Cleaning

The data was cleaned using a **Jupyter Notebook** workflow that:

- Removed redundant or empty rows.
- Validated data types (e.g., dates, numeric values for engagement, booleans for verification).
- Merged duplicate label entries (e.g., “PSV” vs “psv”).
- Normalized fields (e.g., Market Value into euros).
- Checked that every row contained the **minimum required fields** for analysis.

This process ensured that the datasets could later be **visualized and modeled reliably**.

Files Ignored

Not all files were suitable for analysis. The following were excluded because they were either too raw, incomplete, or inefficient for structured use:

- **youtube-posts-overview.csv**
- **youtube-posts.csv**
- **Youtube-competitor-profiles.csv**
- **twitter-competitor-profiles.csv**
- **tiktok-competitor-profiles.csv**
- **instagram-competitor-profiles.csv**

Additionally, some files were **duplicates** or redundant:

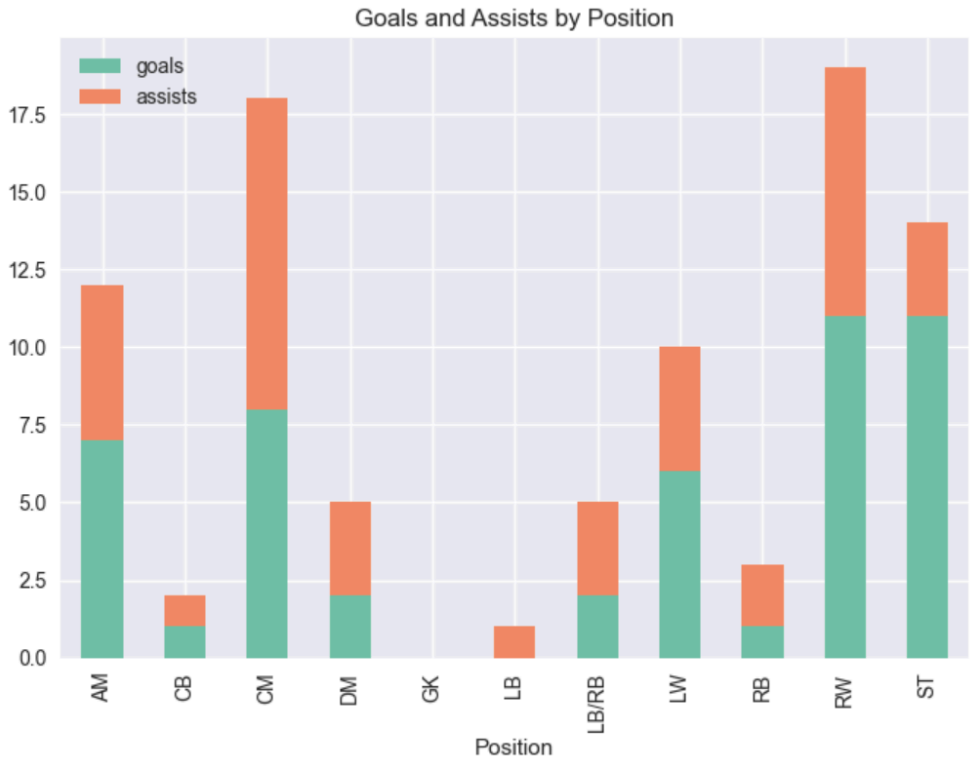
- **Transfermarket.csv** (appeared twice, one copy removed).

Data Analysis

Market Value vs Age

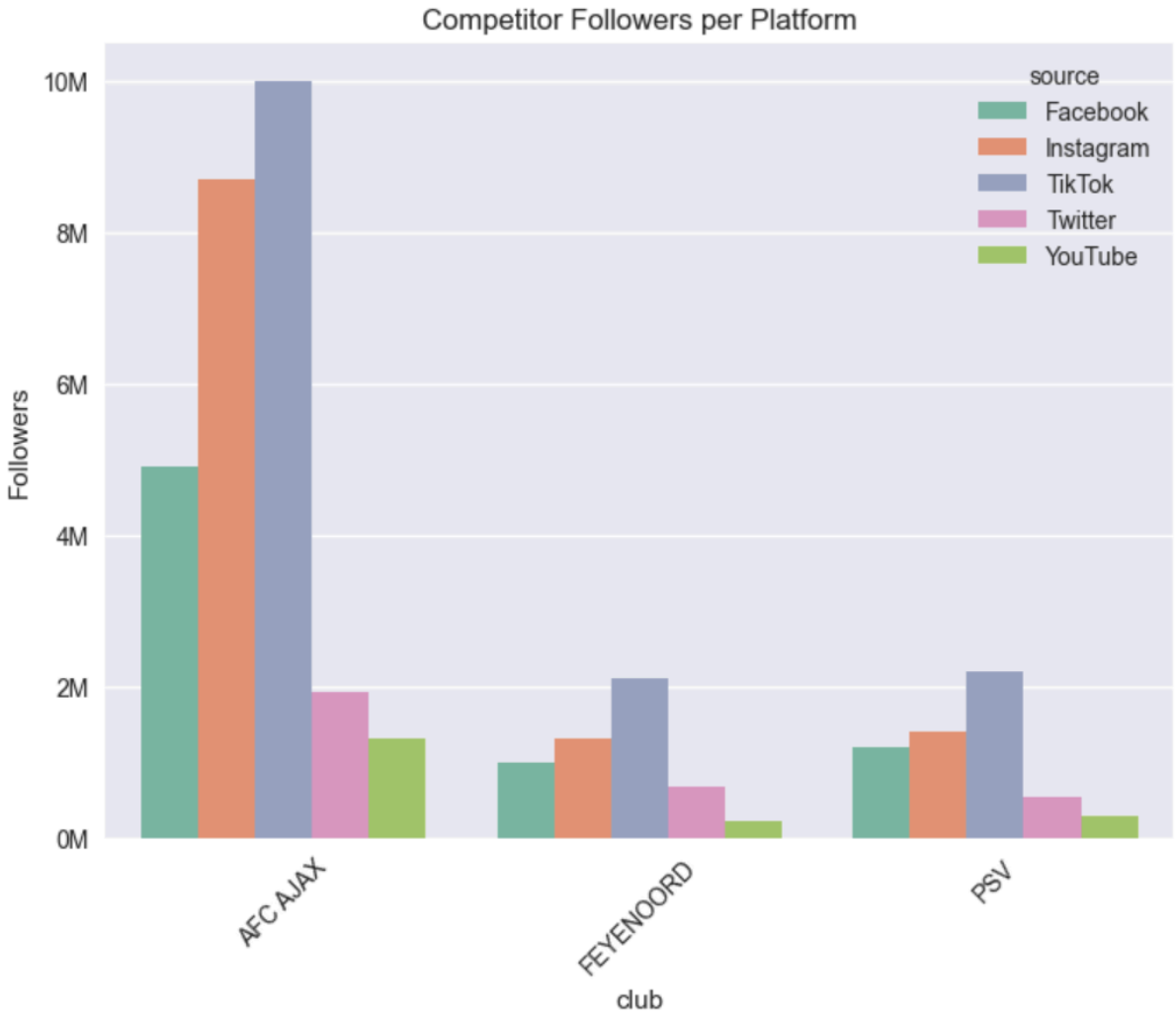
- Insight: younger players (<26) show steep growth in value.

Goals & Assists by Position



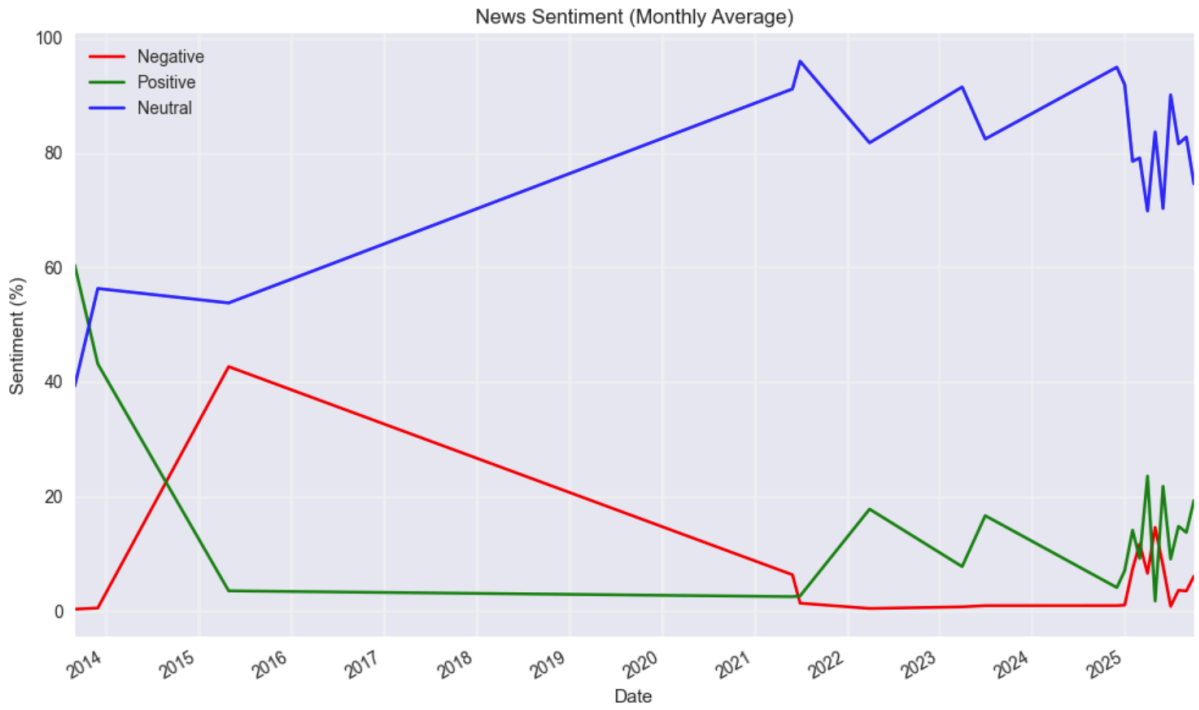
- Insight: the players' position and interaction ratio are in check

Competitor Followers



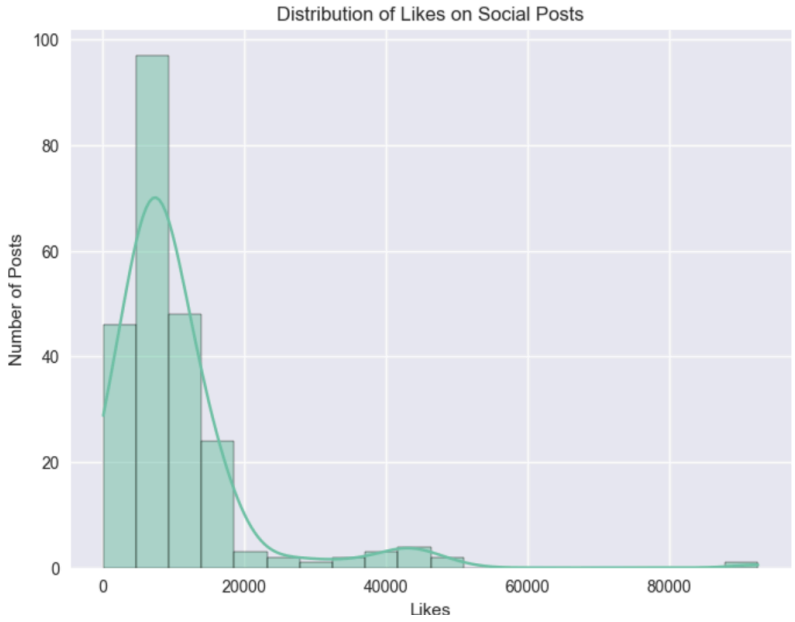
- Insight: PSV has almost the same media reach as FEYENOORD ROTTERDAM.

News Sentiment Over Time



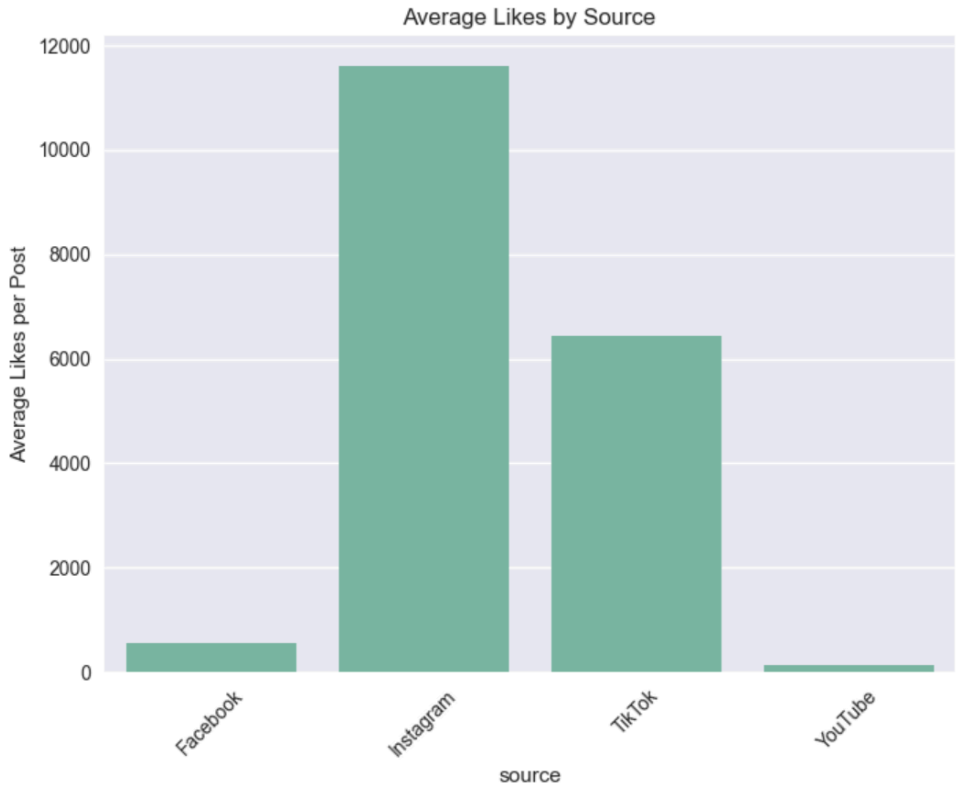
- Insight: most of PSV recent media traction is quite neutral.

Distribution of Likes on Posts

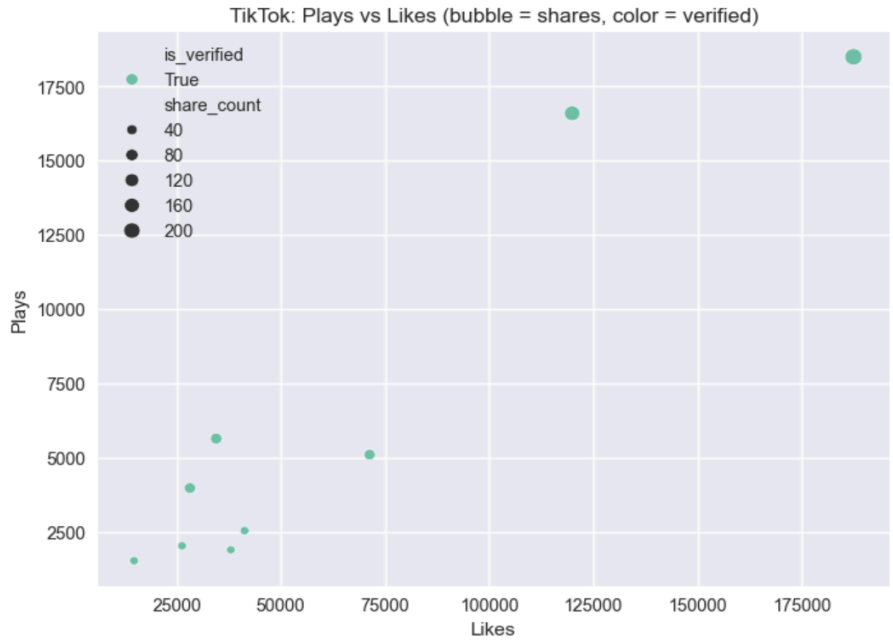


- Insight: most PSV posts cluster around 10k likes, with a few viral outliers at 50k+.

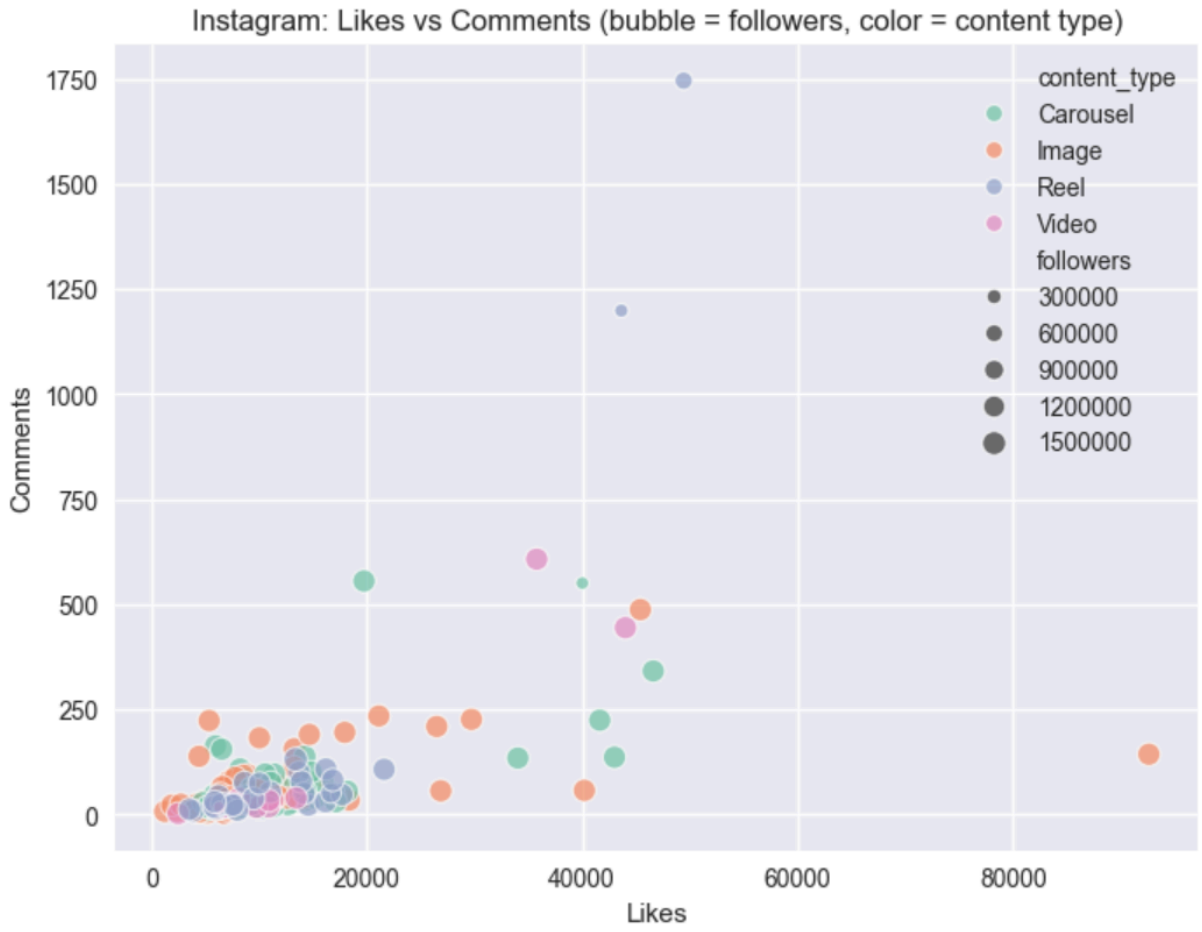
Engagement by Source



- Insight: Instagram = key platform for fan interaction.



- Insight: there is a need for more TikTok content



- Insight: Well crafted posts such as Carousels and Images tend to have a better and consistent Like/Comment ratio, while Reels are shown to be underperforming.

Gaps & Recommendations

Missing Data

- **Player contract status & injuries** (affect availability and market value).
- **Fan profile data** (age, region, influence level).
- **Team performance metrics** (wins/losses amplify or reduce popularity).

Recommendations

- Build a **Player Popularity Index (PPI)** combining 4 pillars:
    1. Performance (market value, goals, assists).
    2. Media sentiment.
    3. Fan sentiment.
    4. Social engagement.
- Create a **Team Popularity Index (PTPI)** to measure PSV's overall brand strength.
- Use cleaned data to train forecasting models (e.g., LSTM, **Prophet**).

What We Might Still Need

To **successfully measure & forecast "social value"**, you should enrich with:

1. Player Metadata

- Height, nationality, shirt number (helps fan recognition).
- Contract length (shorter = more transfer talk → media buzz).
- Injury/availability history (absence affects visibility).

2. Performance Data (on-field)

- Minutes played per game.
- Goals, assists, key passes, defensive stats (context matters by position).
- Match ratings (from sources like WhoScored, Sofascore).
- Team results (winning streaks = more attention).

3. Popularity Signals (off-field)

- Google Trends / search volume index.

4. News & Media Coverage

- PR events (interviews, sponsorship announcements).

5. Sponsorship & Brand Links

- Current sponsorships (Nike, Adidas, Puma, etc.).
- Merch sales / shirt sales (if accessible).

Conclusion

Through structured data analysis, I transformed a large, fragmented dataset from Citrics Labs into a clean and consistent foundation for further research. The datasets, collected from multiple structured and unstructured sources, were standardized, validated, and merged to eliminate redundancy and inconsistencies.

Using this cleaned data, I explored relationships between player performance, market value, media sentiment, and social media engagement. The analysis revealed that player popularity is influenced by both on-field performance and off-field digital visibility. However, the lack of player metadata (contracts, injuries), team results, and detailed fan demographics currently limits accurate forecasting.

To move toward predictive modeling, I recommend creating a **Player Popularity Index (PPI)** that integrates four data pillars:

1. **Performance** (goals, assists, market value),
2. **Media sentiment**,
3. **Fan sentiment**, and
4. **Social engagement metrics**.

This composite index would enable PSV to measure and forecast player popularity transparently, supporting marketing, sponsorship, and talent management decisions. With additional data (e.g., team results, fan profiles, and player contracts), future work can apply forecasting tools such as **Facebook Prophet** or **LSTM neural networks** to predict changes in popularity and market value over time.