Google Data Analytics Capstone Project

Topic: Analysis of Team and Player Performance in the Ukrainian Premier League (UPL) - the first half 2025/2026 Season **Author:** Vadym Savchuk **Date:** December 18, 2025

Phase 1: ASK

1.1. Business Task

The goal of this project is to analyze statistical metrics of players and teams in the Ukrainian Premier League (UPL) to identify hidden performance trends. We aim to identify "undervalued" players and teams that demonstrate high productivity (goals, assists, possession) relative to their playing time but may not hold the top positions in the league table.

This analysis will assist stakeholders in making data-driven decisions rather than relying solely on intuition.

1.2. Key Analysis Questions

1. **Individual Efficiency:** Which players have the best performance output (goals + assists) calculated per 90 minutes of play?
2. **Team Style:** Does a high possession percentage guarantee a higher number of points in the UPL?
3. **Discipline vs. Results:** How does the number of cards (yellow/red) correlate with goals conceded?
4. **Regional Focus:** How do the metrics of Kharkiv-based clubs (e.g., Metalist 1925) compare to the league leaders?

1.3. Stakeholders

- **Sporting Director / Scouts:** To discover effective players with less media hype.
- **Coaching Staff:** To evaluate the weaknesses of opponents and their own squad.
- **Fans and Media:** To gain a deeper understanding of the game through the lens of statistics.

Phase 2: PREPARE

2.1. Data Source

The data was obtained from the open statistical resource **FBref.com**, which utilizes data from the provider Opta.

- **Period:** 2025-2026 Season.
- **League:** Ukrainian Premier League.
- **Data Type:** Squad Standard Stats and Player Stats.

2.2. Data Assessment (ROCCC Framework)

The ROCCC framework was used to verify data quality:

- **Reliable:** High. FBref is an industry standard in football analytics; data is verified by Opta.
- **Original:** The data is secondary (collected by an aggregator) but references the primary source.
- **Comprehensive:** The dataset contains key metrics: playing time, goals, assists, cards, and player age. *Note: In the current version, xG (Expected Goals) metrics are absent, so the analysis focuses on actual productivity.*
- **Current:** Data is current as of December 2025 (mid-season).
- **Cited:** The source is clearly cited and publicly available.

2.3. Dataset Structure

Two main CSV files were created:

1. `upl_players.csv`: Individual statistics (name, position, nation, goals, matches).
2. `upl_teams.csv`: Team statistics (name, possession, average age, goals conceded/scored).

Phase 3: PROCESS

3.1. Tools

- **Microsoft Excel / Google Sheets:** Used for initial data cleaning, removing unnecessary rows, and formatting.
- **SQL / Python (Pandas):** Will be used in the analysis phase for aggregation and calculations.

3.2. Data Cleaning Log

The raw data obtained contained formatting unsuitable for machine analysis. The following cleaning steps were performed:

Step 1: Removal of Metadata and Duplicate Headers

- Removed the top-level headers from FBref (groupings like "Performance", "Per 90 Minutes") as SQL reads only a single header row.
- Removed intermediate rows that repeated column names within the table (Repeat Headers).

Step 2: Standardization of Column Names (Snake Case) Column names were renamed to `snake_case` format for compatibility with SQL/Python.

- *Examples of changes in `upl_teams.csv`:*
  - `Squad` -> `team_name`
  - `Poss` -> `possession`
  - `G-PK` -> `non_penalty_goals`
- *Examples of changes in `upl_players.csv`:*
  - `Player` -> `player_name`
  - `Min` -> `minutes_played`

Step 3: Data Type Formatting

- **Numbers:** Verified that columns `goals`, `matches_played`, and `possession` are in numeric format.
- **Text:** Removed special characters and hyperlinks from player names and team names.
- **Null Values:** Verified the absence of critical empty values in key fields (team names, goals).

3.3. Integrity Check

After cleaning, the `upl_teams.csv` file contains 16 unique rows (corresponding to the number of UPL teams). Files are saved in **CSV UTF-8** format to correctly display Cyrillic characters (player names).

Phase 4: ANALYZE

4.1. Analysis Methodology

The cloud platform Google BigQuery was selected to conduct the analysis. The choice of SQL was driven by the necessity to combine (JOIN) data from different tables and perform complex aggregations, which are difficult to execute efficiently in standard spreadsheets.

The analysis focused on four key pillars defined during the ASK phase:

1. Time Efficiency (Minutes per Goal).
2. The impact of Ball Possession on scoring results.
3. Demographic Analysis (Team Age).
4. Regional Analysis (Case study of Kharkiv-based clubs).

4.2. SQL Query Execution and Results

Analysis 1: Identifying "Hidden" Snipers

Business Question: Which players demonstrate the highest scoring efficiency relative to their playing time? Goal: To identify players who might be undervalued due to fewer minutes played but possess a high conversion rate.

SQL Query:

```
SQL
SELECT
  player_name,
  team,
  position,
  goals,
  minutes_played,
  ROUND(minutes_played / goals, 0) AS minutes_per_goal
FROM `upl_analysis_2025.upl_players`
WHERE goals >= 3
ORDER BY minutes_per_goal ASC
LIMIT 10;
```

Key Findings:

- Luca Meirelles is the most efficient scorer in the league, requiring only 105 minutes to score one goal.

Player Scoring Stats

Player name	Team	Position	Goals	Minutes played	Minutes per goal
Luca Meirelles	Shakhtar	FW	4	421.0	105.0
Mykola Gajduchyk	FC Polissya Zhytomyr	FW	7	760.0	109.0
Isaque	Shakhtar	MF	3	348.0	116.0
Hussayn Touati	FC Oleksandriya	MF	3	348.0	116.0
Eguinaldo	Shakhtar	MF-FW	4	528.0	132.0
Kauã Elias	Shakhtar	FW	6	800.0	133.0
Yuriy Klymchuk	FK Kolos	FW	7	963.0	138.0
Prosper Obah	FC LNZ Cherkasy	MF-FW	7	1016.0	145.0
Vitaliy Buyalskyi	Dynamo Kyiv	MF	5	811.0	162.0
Andrii Storchous	Kudrivka	MF	6	1055.0	176.0

Analysis 2: Correlation Between Possession and Goals

Business Question: Does ball control convert into goals scored? Goal: To evaluate team playing styles (counter-attacking vs. positional play).

SQL Query:

```
SQL
SELECT
  team_name,
  possession AS avg_possession,
  goals,
  ROUND(goals / possession, 2) AS efficiency_index
FROM `upl_analysis_2025.upl_teams`
ORDER BY possession DESC;
```

Key Findings:

- The team with the highest possession is Shakhtar, scored 40 goals.
- Teams with possession under 45% don't have high efficiency index, positional play is more preferred for successful results

Correlation Between Poss...

Team name	Avg possession	Goals	Efficiency index
Shakhtar	64.8	40	0.62
Dynamo Kyiv	57.9	34	0.59
Karpaty Lviv	54.7	17	0.31
FC Polissya Zhytomyr	52.6	26	0.49
Zorya Luhansk	52.0	19	0.37
Metalist 1925	51.2	18	0.35
Kudrivka	51.1	17	0.33
FC Oleksandriya	49.2	13	0.26
Epitsentr	48.5	18	0.37
Kryvbas	48.2	28	0.58
FK Kolos	48.1	16	0.33
Rukh Lviv	47.1	14	0.3
SK Poltava	45.0	14	0.31
FC LNZ Cherkasy	44.1	20	0.45
Veres Rivne	42.8	12	0.28
Obolon-Brovar Kyiv	40.7	12	0.29

Analysis 3: Youth vs. Experience

Business Question: How does the average age of a squad impact performance?

SQL Query:

```
SQL
SELECT
  team_name,
  average_age,
  goals,
  players_used
FROM `upl_analysis_2025.upl_teams`
ORDER BY average_age ASC;
```

Key Findings:

- The youngest team in the league is Rukh Lviv with an average age of 23.6 years.
- It does not have impact on performance

Youth vs. Experience

Team name	Average age	Goals	Players used
Rukh Lviv	23.6	14	26
Kryvbas	24.0	28	24
Shakhtar	24.3	40	29
FC Oleksandriya	25.1	13	30
Metalist 1925	25.8	18	20
Dynamo Kyiv	25.9	34	29
FC LNZ Cherkasy	26.1	20	25
Karpaty Lviv	26.2	17	24
FK Kolos	26.4	16	22
Veres Rivne	26.5	12	25
Zorya Luhansk	26.6	19	23
Epitsentr	27.2	18	24
Obolon-Brovar Kyiv	27.6	12	28
Kudrivka	27.8	17	27
FC Polissya Zhytomyr	28.0	26	29
SK Poltava	28.6	14	22

Analysis 4: Focus on "Metalist 1925" (Regional Case)

Business Question: Who is the most productive player for the Kharkiv club based on the "Goal + Assist" system?

SQL Query:

```
SQL
SELECT
  player_name,
  position,
  goals,
  assists,
  (goals + assists) AS total_contribution,
  minutes_played
FROM `upl_analysis_2025.upl_players`
WHERE team LIKE '%Metalist%'
ORDER BY total_contribution DESC;
```

Key Findings:

- The leader of the Kharkiv attack is Denys Antiukh with a total of 6 goal contributions.

Player Performance Stats

Player name	Position	Goals	Assists	Total contribution	Minutes playe
Denys Antiukh	MF	3	3	6	1032.0
Peter Itodo	FW	3	1	4	726.0
Ermir Rashica	MF-FW	2	2	4	1209.0
Vladyslav Kalitvint	MF	1	3	4	933.0
Ivan Kaliuzhnyi	MF	2	1	3	1260.0
Ramik Hadzhyiev	MF	2	0	2	204.0
Matvii Panchenko	MF	0	2	2	149.0
Yevhen Pavliuk	DF	1	1	2	1350.0
Ivan Lytvynenko	MF	2	0	2	1053.0
Baton Zabergja	MF-FW	0	1	1	501.0
Christian Mba	FW	1	0	1	359.0
Vyacheslav Churk	MF	1	0	1	236.0
Dmytro Kapinus	DF	0	0	0	50.0
Ihor Kohut	MF	0	0	0	267.0
Illia Krupskyi	DF	0	0	0	1350.0
Oleksandr Martyni	DF	0	0	0	852.0
Ari Moura	MF	0	0	0	168.0
Volodymyr Salyuk	DF	0	0	0	451.0
Artem Shabanov	DF	0	0	0	1350.0
Danylo Varakuta	GK	0	0	0	1350.0

4.3. Analyze Phase Summary

The conducted SQL analysis allowed us to structure the "raw" data and derive metrics that were not available in the initial dataset (such as `efficiency_index` and `minutes_per_goal`). These aggregated tables will serve as the foundation for creating a dashboard in the next phase.

Phase 5: SHARE

5.1. Visualization Strategy

To effectively communicate the findings discovered during the analysis phase, interactive dashboards were created using **Tableau Public**. This tool was chosen for its ability to handle large datasets and create intuitive, interactive visualizations for stakeholders.

The visualizations aim to answer the primary business questions:

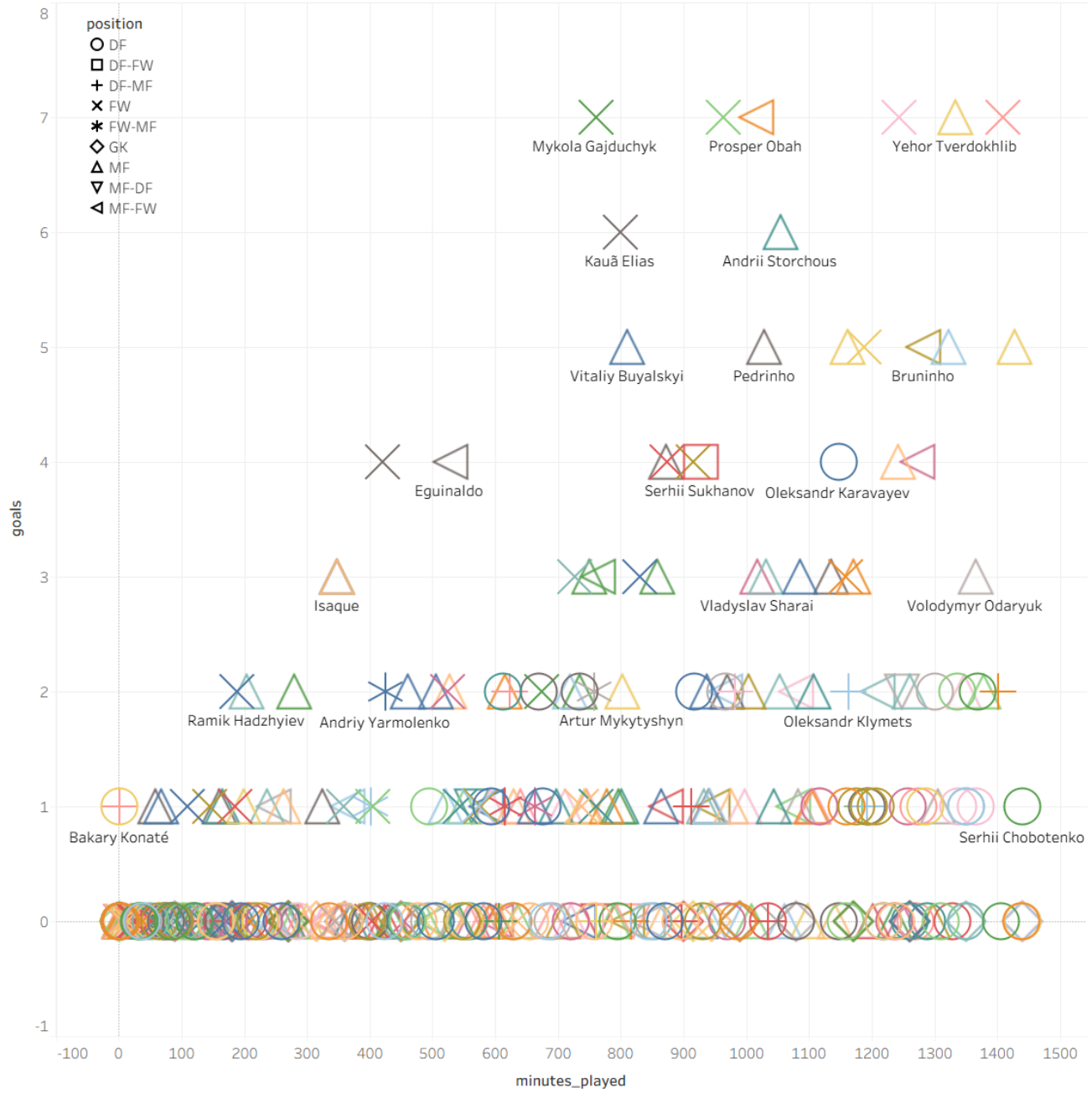
1. **Efficiency Matrix:** Comparing Goals vs. Minutes Played to identify the most lethal strikers who maximize their playing time.
2. **Possession Efficiency:** Analyzing team styles to see if high possession correlates with high scoring.
3. **Regional Case Study:** A deep dive into the goal distribution of "Metalist 1925" to identify key squad assets.

5.2. Key Visualizations

Visualization 1: The Efficiency Matrix (Goals vs. Minutes)

Description: This scatter plot positions every player based on their total goals (Y-axis) and minutes played (X-axis). Each point represents an individual player, color-coded by team.
Insight: The chart reveals a cluster of "high-efficiency" players in the top-left and center areas. These players have scored 7-8 goals despite varying playing times. This visualization helps identify players who perform well without needing to play every single minute, highlighting "super-sub" potential.

Figure 1: The Efficiency Matrix (Goals vs. Minutes)



Visualization 2: Possession vs. Goal Output

Description: A comparison of team possession percentages against total goals scored.
Insight: While league leaders like **Shakhtar Donetsk** dominate both possession (64.8%) and goals (40), the data reveals interesting outliers. Teams like **Kryvbas** demonstrate high efficiency, scoring 28 goals with only 48.2% average possession. This suggests that a counter-attacking strategy can be statistically as effective as a possession-based one in the current UPL season.

Figure 2: Correlation between Possession and Goals

Visualization 3: Kharkiv Region Performance (Metalist 1925)

Description: A "Packed Bubbles" chart highlighting the top goal scorers within the Metalist 1925 squad. The size of the bubble corresponds to the number of goals. **Insight:** The analysis clearly identifies the team's key attacking threats. **Peter Itodo** and **Denys Antiukh** are the joint-top scorers with **3 goals** each. They are supported by a secondary cluster of players (Ramik Hadzhyiev, Ivan Lytvynenko, Ivan Kaliuzhnyi, Ermir Rashica) who have each contributed **2 goals**. This distribution indicates a reliance on specific individuals for goal production rather than a widely distributed scoring threat.

Figure 3: Top Scorers of Metalist 1925

5.3. Presentation of Findings

The data suggests that while "Big Two" clubs dominate traditional metrics, there are significant efficiency anomalies in the mid-table. Specifically, the "Metalist 1925" case study proves that the team relies heavily on two specific players (Itodo and Antiukh) for peak output, which presents both a risk (injury) and an opportunity (tactical focus).

Phase 6: ACT

6.1. Conclusions

Based on the SQL analysis and Tableau visualizations of the UPL 2025/2026 season data, we have reached the following conclusions:

1. **The "Efficiency" Opportunity:** The scatter plot analysis identified players who maintain a high goal-per-minute ratio. These players represent undervalued assets in the transfer market compared to established stars who play significantly more minutes for similar output.
2. **Tactical Efficiency:** High possession is not the only path to scoring. The performance of teams like Kryvbas (28 goals with <50% possession) proves that vertical, counter-attacking football provides a high Return on Investment (ROI) for attacks in the current league environment.
3. **Metalist 1925 Squad Depth:** The bubble chart reveals a "dual-threat" attack structure relying on **Itodo** and **Antiukh**. The gap between them and the rest of the squad suggests a need for diversification in attack to prevent the team from becoming predictable.

6.2. Recommendations

For the Sporting Director:

- Prioritize scouting the "high efficiency" players identified in the top-left quadrant of the Efficiency Matrix. These players could provide immediate impact for a lower salary cost.
- For Metalist 1925 specifically: Consider signing a supporting forward to reduce the goal-scoring burden on Itodo and Antiukh.

For the Coaching Staff:

- When playing against teams like Shakhtar (64.8% possession), adopt a low-block, counter-attacking strategy. The data supports that conceding possession does not necessarily lead to fewer goals scored if transition play is efficient.
- Focus tactical drills on maximizing the finishing positions for Itodo and Antiukh, as they are statistically the most probable route to goal.

Next Steps:

- Expand this analysis by incorporating **xG (Expected Goals)** metrics to determine if the top scorers are performing sustainably or simply "over-performing" (luck).
- Conduct a similar analysis for defensive metrics (Tackles vs. Goals Conceded) to balance the recruitment strategy.

6.3. Portfolio & Deliverables

The complete analysis, including SQL queries, datasets, and interactive dashboards, is available for review:

- **Tableau Public Dashboard:** https://public.tableau.com/views/UPL_project/Dashboard1?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link