

City Football Group

Scouting Intelligence Platform

A comprehensive data-driven approach to player identification, quality assurance, and actionable insights for global talent acquisition

Author: Daniel Carreño



Introduction and Objectives

This submission resolves the CFG Pre Task by following the exact requirements of each block, demonstrating a complete workflow from data quality to actionable insights.

01

Task 1.1: Data Cleaning

Explore Players.csv, identify problems, deduplicate the data, document before and after with metrics, and deliver a final cleaned player table.

02

Task 1.2: System Integration

Explain how to manage PlayerID merges or updates across multiple related tables, maintain integrity, and define a consistent repeatable process.

03

Task 2: Insights & Reporting

Analyse ReportingInsight.csv to surface patterns and insights, present them through visuals, justify the tool choice, and provide screenshots.

The overall approach was to secure data quality and integrity first (Tasks 1.1 and 1.2), and then convert the data into actionable insights (Task 2).

Task 1.1: Players.csv Problem and Solution

Problem Detected

The Players.csv dataset showed common raw export issues that threatened data reliability:

- Duplicates, including non-trivial duplicates
- Inconsistencies in identifying fields (naming)
- Missing or incomplete values in relevant attributes
- Same player appearing with different PlayerID values

This created a clear risk: if the same player appears with different PlayerID values, the table is not reliable for analysis or for linking to other tables.

Solution Implemented

The cleaning work was organised in four systematic steps:

1. **Exploration and audit** of data state, comparing total rows versus unique players
2. **Player identity rule** using normalized name, date of birth, and nationality
3. **Record consolidation** to select one canonical PlayerID per player
4. **Before and after summary** documenting impact with clear metrics

Task 1.2: PlayerID Updates Across Related Tables



Problem

After deduplication, some PlayerID values become obsolete or must be merged. In a real environment, PlayerID exists in multiple tables: Reports, Contracts, Appearances. If only the Players table is updated, relationships break and referential integrity issues appear.



Solution

A standard repeatable process was designed to maintain data integrity across the entire system while keeping full traceability of changes.



Identify Affected Tables

Create an inventory of all tables where PlayerID exists



Create ID Mapping Table

Build correspondence table: Old PlayerID to New canonical PlayerID



Apply Cascading Updates

Update every related table using the mapping so all references point to correct PlayerID



Post-Update Validation

Confirm no references remain to obsolete PlayerID values and relationships remain correct



Consistency & Repeatability

Process designed to repeat with future ingestions and automate in database environment

Task 2: Streamlit Interactive Exploration

Problem

Task 2 required analysing ReportingInsight.csv to find meaningful patterns and insights. Exploration requires fast filtering and comparisons, plus the ability to validate insights interactively when focusing on specific players, scouts, or markets.

Why Streamlit?

A Streamlit application was built as an interactive exploration layer focused on:

- **Exploration:** Navigate the dataset and discover patterns quickly
- **Validation:** Test insights when filtering by position, age, country
- **Depth:** Drill down into specific players, scouts, and markets

Key App Components

1. Basic login for access control and privacy (simulating real environment where scouting data is sensitive)
2. Home page to navigate to different analysis modules with clear structure
3. Consistent sidebar filter panel: position, age band, country, current team, preferred foot, report type, performance grade, potential grade, date range, and Top N
4. KPI cards to summarise filtered subset: total reports, unique players, average performance, high potential share, top performers share
5. Interactive visuals for rankings, distributions, performance vs potential, analysis by country, position, age, and foot
6. Excel export to share filtered subsets and support offline work

Username / Password for login: CityGroup

Link: https://citygroupanalytics.streamlit.app/?page=dashboard&tab=player_analysis

Task 2: Tableau Executive Dashboard

Problem

Task 2 requires tool choice, explanation of key insights and why visuals were selected, plus screenshots. From a role perspective, Tableau is a key reporting tool for executive consumption and cross-functional alignment across scouting and decision making.

Solution

Tableau was used as the final reporting layer, with visuals selected to answer core scouting questions and support strategic decisions.



Coverage by Country Map

Understand market coverage and identify under-scouted regions



Talent Matrix

Performance versus potential to prioritise targets across current level and future ceiling



Position Intelligence Grid

Age and position analysis to understand focus areas and spot coverage gaps by profile



Scout Efficiency

Volume versus quality to compare scouting productivity and effectiveness



Preferred Foot Distribution

A tactical variable that supports specific profile needs



KPI Cards

Provide immediate context for interpreting the dashboard

[Click here to access the dashboard](#)

GitHub Creation, Structure, and Rationale

Problem

The submission needed to be more than results. It needed to be easy to review, understand, and reproduce, with clear ownership of where each component lives.



Solution: Structured Repository

The repository is organised to separate concerns clearly:

- **data/raw:** Original data
- **data/processed:** Cleaned data ready for analysis
- **notebooks:** Notebooks by task
- **app:** Streamlit application
- **src:** Reusable code
- **documentation:** README and supporting notes

This structure makes it easy to see what comes in (raw), what comes out (processed), where each task is solved, and how to run and reproduce the work.

[Click here to access the repository](#)

Conclusions: Aligned to the Role

This Pre Task was delivered as a complete realistic workflow, demonstrating the core capabilities expected for the role.

Task 1.1: Reliable Player Table

Dataset explored, issues identified, deduplication performed using player identity rule, impact documented with before and after metrics, and final cleaned table delivered.

Task 1.2: Real Data Ecosystem

Method defined to merge or update PlayerID while keeping relationships correct and the process repeatable across the entire system.

Task 2: Data into Decisions

Streamlit supports interactive exploration and validation. Tableau supports executive reporting. Submission includes tool justification and screenshots as required.

3

Core Capabilities

Ensuring data quality and integrity, designing scalable repeatable processes, and converting scouting data into consumable insights

Overall, the submission demonstrates ensuring data quality and integrity, designing scalable repeatable processes across related tables, and converting scouting data into insights that are consumable by both technical and sporting stakeholders.