Part 1.2 Data Quality Assessment Summary

In this project, we analyzed a unified collegiate athletics performance database that integrates data from three primary systems: Hawkins (force plates), Kinexon (GPS/accelerometry), and Vald (Strength testing). Using our `part1_exploration.py` script, we conducted an initial data quality assessment to understand the structure, coverage, and potential issues in the dataset.

Across the full table (`research_experiment_refactor_test`), there are **1,287 unique athletes** represented. These athletes come from **92 different teams/sports**, indicating that the dataset spans a wide range of programs rather than being restricted to a single sport or squad. The **timestamp range** extends from **October 15, 2018** to **October 21, 2025**, providing nearly seven years of longitudinal performance and workload information. This time span supports not only cross-sectional comparisons between teams but also long-term trend analyses at the athlete and team level.

In terms of data sources, **Kinexon** accounts for the largest share of records with **4,073,754 rows**, followed by **Hawkins** with **2,492,372 rows**, and **Vald** with **51,300 rows**. This distribution suggests that GPS/accelerometry-based external load monitoring (Kinexon) is the most frequently collected data type, while force-plate testing (Hawkins) is also highly represented. Vald strength testing is present but at a smaller scale, which is still sufficient for meaningful strength and asymmetry analyses but may require more careful handling of sample sizes and missingness.

We also evaluated basic data quality issues around athlete identifiers. Using simple rules (flagging any `playername` that is NULL, an empty string, or equal to “UNKNOWN”), our query found **no athletes with missing or invalid names**. This is encouraging from a data integrity perspective, as it means that records can be reliably linked back to individual anonymized players, for example `PLAYER_001`, without needing to resolve ambiguous IDs.

Finally, we examined cross-system coverage by counting how many data sources each athlete appears in. There are **541 athletes** with data from **two or more systems** (Hawkins, Kinexon, and/or Vald), and many of these appear in all three. This multi-source overlap is particularly valuable for downstream analyses, because it enables integrated questions such as how force-plate metrics relate to GPS-derived workload or how strength testing interacts with performance and load over time.

Overall, the data quality assessment indicates that the database is **large, multi-sport, and longitudinal**, with good identifier integrity and substantial overlap across systems. The main considerations going forward will be: (1) uneven record counts between sources, and (2) ensuring that our selected metrics and research questions leverage the subset of athletes who have robust multi-system data. These findings informed our metric selection and research question for the remainder of the project.