**Sports Performance Database Report**

**Name:** Ans Riaz

**Student Id:** 23079633

**GitHub:** [**https://github.com/ansmalik67/sql\_assignment.git**](https://github.com/ansmalik67/sql_assignment.git)

**1. Introduction**

This report documents the development and implementation of a Sports Performance Database using SQLite. The database is designed to store structured data related to players, tournaments, matches, performance statistics, and sponsorships. The schema ensures data integrity, realistic constraints, and optimized query performance, making it highly suitable for sports analytics and management applications.

**2. Database Overview**

The database consists of five primary tables, each interconnected using primary keys (PKs) and foreign keys (FKs):

* Players: Stores information about athletes, including personal details and rankings.
* Tournaments: Tracks various tournaments and their attributes.
* Matches: Records details of individual matches played between players.
* Performance: Captures key performance metrics for players in each match.
* Sponsorships: Stores sponsorship deals associated with players.

This relational structure ensures efficient querying and data consistency while supporting various statistical analyses and performance tracking.

**3. Data Generation Approach**

**3.1 Players Table (Athlete Profiles)**

* **Number of Records:** 1,000+
* **Key Attributes:**
  + player\_id: Unique identifier (PK)
  + name: Player’s full name
  + age: Age (18 to 40)
  + nationality: Randomly selected from five countries
  + ranking: Integer ranking (assigned based on Gaussian distribution)
  + gender: Randomly assigned as Male or Female
  + experience\_level: Categorized as Beginner, Intermediate, or Professional
* **Realism Enhancements:**
  + Rankings follow a normal distribution to ensure realistic ranking distribution.
  + Some players intentionally have missing experience levels for realistic data variability.
  + 5% of players have duplicate entries to simulate data inconsistencies.

**3.2 Tournaments Table**

* **Number of Records:** 100+
* **Key Attributes:**
  + tournament\_id: Unique identifier (PK)
  + tournament\_name: Unique tournament name
  + location: Random city from major sports locations
  + year: Random year between 2000-2025
  + category: Regional, National, or International
* **Realism Enhancements:**
  + Ensures tournaments are **evenly distributed across years**.
  + Random selection of **real-world tournament locations**.

**3.3 Matches Table (Match Records)**

* **Number of Records:** 300+
* **Key Attributes:**
  + match\_id: Unique identifier (PK)
  + tournament\_id: FK linking to Tournaments
  + player1\_id, player2\_id: FKs linking to Players
  + round: Group Stage, Quarterfinal, Semifinal, Final
  + score: Random match score in a realistic format
  + match\_date: Randomized date between 2020-2025
* **Realism Enhancements:**
  + **Ensures realistic tournament match distributions**.
  + **25% of matches have missing scores** for real-world data issues.
  + **Players assigned based on realistic participation patterns**.

**3.4 Performance Table (Player Statistics)**

* **Number of Records:** 400+
* **Key Attributes:**
  + performance\_id: Unique identifier (PK)
  + player\_id: FK linking to Players
  + match\_id: FK linking to Matches
  + points\_scored: Integer between 0-100
  + fouls: Integer between 0-10
  + ranking\_change: Positive or negative ranking movement
* **Realism Enhancements:**
  + Players’ rankings update dynamically based on performance.
  + Negative and positive ranking changes assigned based on match outcomes.

**3.5 Sponsorships Table (Financial Agreements)**

* **Number of Records:** 50+
* **Key Attributes:**
  + sponsor\_id: Unique identifier (PK)
  + sponsor\_name: Company sponsoring the player
  + player\_id: FK linking to Players
  + contract\_value: Random value between $50,000-$500,000
  + duration\_years: Contract duration between 1-5 years
* **Realism Enhancements:**
  + Higher-ranked players receive larger sponsorship deals.
  + Ensures balanced sponsorship distributions across players.

**4. Entity-Relationship (ER) Design & Constraints**

* **Each table follows normalization principles** to minimize redundancy.
* **Foreign key relationships maintain data integrity**.
* **Compound keys** exist in Matches to prevent duplicate match entries.
* **Realistic constraints** applied (e.g., Age 18-40, Ranking > 0, Contract Value > 0).



**5. Data Quality Analysis & Fixes**

**5.1 Handling Duplicates**

**Issue:** 5% of players were duplicated for testing data quality.

**Fix:** Unique constraints can be applied to prevent accidental duplication.

**5.2 Handling Missing Data**

**Issue:** Some match scores and experience levels were deliberately set to NULL.

**Fix:** Missing values can be handled via imputation (e.g., median ranking for missing experience levels).

**6. Example SQL Queries for Analysis**

**6.1 Top 5 Players by Performance**

SELECT P.name, SUM(Pr.points\_scored) AS Total\_Points

FROM Players P

JOIN Performance Pr ON P.player\_id = Pr.player\_id

GROUP BY P.player\_id

ORDER BY Total\_Points DESC

LIMIT 5;

**6.2 Most Successful Tournament Locations**

SELECT T.location, COUNT(M.match\_id) AS Matches\_Held

FROM Tournaments T

JOIN Matches M ON T.tournament\_id = M.tournament\_id

GROUP BY T.location

ORDER BY Matches\_Held DESC;

**6.3 Sponsorship Distribution by Experience Level**

SELECT P.experience\_level, COUNT(S.sponsor\_id) AS Sponsorships

FROM Players P

JOIN Sponsorships S ON P.player\_id = S.player\_id

GROUP BY P.experience\_level;

**6.4 Retrieve Players with Sponsorship Details**

SELECT P.name, S.sponsor\_name, S.contract\_value, S.duration\_years

FROM Players P

JOIN Sponsorships S ON P.player\_id = S.player\_id

ORDER BY S.contract\_value DESC

LIMIT 5;

**6.5 Check Performance Stats For a Specific Player**

SELECT P.name, M.match\_id, Pr.points\_scored, Pr.fouls, Pr.ranking\_change

FROM Players P

JOIN Performance Pr ON P.player\_id = Pr.player\_id

JOIN Matches M ON Pr.match\_id = M.match\_id

ORDER BY Pr.points\_scored DESC

LIMIT 5;

**7. Ethical Considerations & Data Privacy**

**7.1 Data Privacy and Security**

* The database contains synthetic, randomly generated data, ensuring no real-world personal or sensitive data is stored or used.
* Personally identifiable information (PII), such as player addresses or financial account details, is not included, safeguarding against data breaches.
* Access control mechanisms (e.g., user authentication and role-based access) should be implemented in real-world applications to prevent unauthorized modifications.

**7.2 Handling Missing and Duplicate Data Ethically**

* Missing data (e.g., experience levels, match scores) is deliberately included to simulate real-world data gaps. This ensures statistical methods and machine learning models can be tested against incomplete datasets.
* Duplicate data (e.g., player records) is intentionally added to stress-test database integrity and improve data validation techniques.
* In a production environment, data deduplication techniques and imputation strategies should be applied responsibly to maintain fairness and accuracy in reporting.

**7.3 Transparency in Data Use**

* If this database were used in real-world applications, clear disclosure policies should be established regarding how player performance data is collected, stored, and used.
* Players and stakeholders should have visibility into the ranking algorithms and sponsorship allocation criteria to prevent unfair advantages or biases.
* Ethical data use policies should be aligned with GDPR, CCPA, and other relevant data protection laws when handling personal or financial data.

**7.4 Sustainability and Data Management**

* Efficient data management practices (e.g., regular database optimization, duplicate data removal, and data retention policies) should be enforced to reduce storage costs and carbon footprint.
* Archival policies should be established for old tournament data, ensuring only relevant and necessary data is stored long-term.
* Synthetic Data: No real-world player or sponsorship data is used.
* Missing Data Handling: Players with missing experience are marked for review.
* Duplicate Data Analysis: Used to test system resilience.

**8. Results:**

|  |  |
| --- | --- |
| **A screenshot of a game  AI-generated content may be incorrect.** | **A screenshot of a computer  AI-generated content may be incorrect.** |
| **A screenshot of a black screen  AI-generated content may be incorrect.** | **A screenshot of a computer  AI-generated content may be incorrect.** |
| A screenshot of a graph  AI-generated content may be incorrect. | |

**9. Conclusion**

This database provides a realistic, well-structured model for sports analytics, tournament tracking, and performance monitoring. The schema follows best practices in relational database design, ensuring efficient queries and high data integrity.

Future improvements can include:

* Integration with real-world API data for live tracking.
* More complex ranking algorithms to enhance player evaluation.

This structured approach ensures comprehensive performance analysis and seamless data integrity.