**Assignment 2 Report – Nam Anh Truong**

1. **Schema**

|  |  |
| --- | --- |
| **TableA.csv (Transfermarkt data)** | **TableB.csv (FBref data)** |
| 1. ID 2. Name 3. position 4. age 5. nationality 6. market\_value | 1. ID 2. name 3. position 4. age 5. nationality 6. current\_club 7. matches\_played |

Schemas are almost identical, with the only difference being: "market\_value" in TableA vs "matches\_played" and “current\_club” in TableB.

1. **Set S of attributes**

Since schemas are almost identical, S would be:

1. ID
2. name
3. position
4. age
5. nationality
6. current\_club
7. market\_value
8. matches\_played
9. **Analyze table A:**

Handling missing values:

1. **ID**: If missing, generate sequential IDs.
2. **name**: Cannot be filled in automatically; may require manual research.
3. **position**: Could use the most common position for that player from other sources or leave as unknown.
4. **age**: Could impute with the mean or median age, or use regression based on other attributes.
5. **nationality**: Could use the most common nationality for players with the same name pattern or from the same club.
6. **current\_club**: Could use information from other sources or leave as unknown.
7. **market\_value**: Could impute with the mean or median value, or use regression based on other attributes like age and position.

* **Missing Values:**
  + Position: 0% missing (0/1000)
  + ID and name: 0% missing (0/1000)
  + Age and nationality: 0% missing (0/1000)
  + Market value: 0% missing (0/1000)
* **Position Analysis:**
  + Classification: Categorical
  + Unique position values: 13
  + ['Centre-Forward' 'Left Winger' 'Right Winger' 'Attacking Midfield' 'Central Midfield' 'Defensive Midfield' 'Centre-Back' 'Left-Back' 'Right-Back' 'Second Striker' 'Right Midfield' 'Goalkeeper' 'Left Midfield']
* **ID Analysis:**
  + Classification: Categorical (identifier)
  + Format: All IDs follow the format 't' followed by a number
* **Name Analysis:** A graph of different sizes of blue bars

  AI-generated content may be incorrect.
  + Classification: Textual
  + Average length: 13.14 characters
  + Range: 4 to 22 characters
  + Outliers: 26 outliers identified, including very short names (like "Gavi", "Pepê", "Igor", "Toti") and very long names (like "Trent Alexander-Arnold")
  + Note that "Toti" appears multiple times in the dataset (this suggests potential duplicate records)
* **Age Distribution:** A graph of a distribution of player ages

  AI-generated content may be incorrect.
  + Classification: Numeric
  + Mean: 23.60 years
  + Standard deviation: 3.28 years
  + Range: 17 to 33 years
  + No age outliers detected
  + The histogram shows a normal-like distribution with most players between 21-24 years old
* **Nationality Analysis:** A graph of blue bars

  AI-generated content may be incorrect.
  + Classification: Categorical
  + 59 unique nationalities
  + Top nationalities: England, Brazil, France, Spain, Portugal
  + The distribution is uneven with a few countries dominating
* **Market value Analysis:** A graph of a distribution of market values

  AI-generated content may be incorrect.
  + Classification: Numeric (with currency)
  + Mean market value: 30.25 million €
  + Standard deviation: 20.27 million €
  + Minimum value: 20.0 million €
  + Maximum value: 200.0 million €
  + Potential market value outliers: 17

**Software Tools Used**

* Python for data analysis
* Pandas for data manipulation and cleaning
* Matplotlib for creating histograms and visualizations
* NumPy for numerical operations and statistical analysis