**Q2. Document describing the subset you chose for MongoDB, your design and your validation.**

**Subset Chosen for MongoDB:**

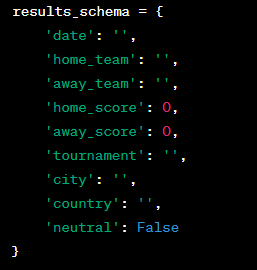
For the MongoDB assignment, I chose to focus on the following three datasets from the provided football use case:

1. "results.csv" dataset containing information about international football matches.
2. "shootouts.csv" dataset containing information about penalty shootouts in international football matches.
3. "goalscorers.csv" dataset containing information about goalscorers in international football matches.

**Design of MongoDB Collections:**

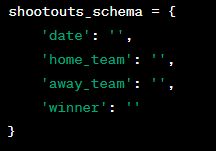
**1. Matches Collection:**

I designed the MongoDB collection named "matches" to store data from the "results.csv" dataset. The structure of the collection is based on the provided results\_schema:



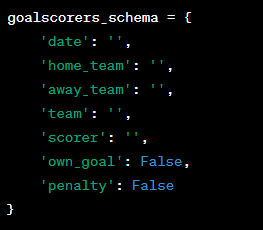
**2. Shootouts Collection:**

For the "shootouts.csv" dataset, I designed the MongoDB collection named "shootouts" using the provided shootouts\_schema:



**3. Goalscorers Collection:**

For the "goalscorers.csv" dataset, I designed the MongoDB collection named "goalscorers" based on the provided goalscorers\_schema:



**Validation Approach:**

For all three collections, I implemented basic validation rules to ensure data quality:

* String fields (`home\_team`, `away\_team`, `tournament`, `city`, `country`, `team`, `scorer`): Verify that the fields are non-empty strings.
* Numeric fields (`home\_score`, `away\_score`): Validate that the scores are non-negative integers.
* Boolean fields (`neutral`, `own\_goal`, `penalty`): Ensure accurate representation of boolean values.
* Date field (`date`): Validate that the date follows a valid date format.

**Inserting Data:**

I used the respective CSV files to insert data into the MongoDB collections using the `*pymongo*` library in Python. The insertion code reads the CSV files, extracts relevant information, and inserts documents into the collections.and to extract data from file *csv* library will help us

**Queries:**

I implemented querying capabilities for each collection. I can retrieve documents, select specific records based on criteria like team names, dates, or winners, and perform aggregation queries to analyse the data.