

Rutvi Bheda – rb1859  
Mohammed Raeesul Irfan – mr6248  
Niharika Ahirekar – na2572

## CSCI 620 Introduction to Big Data Project

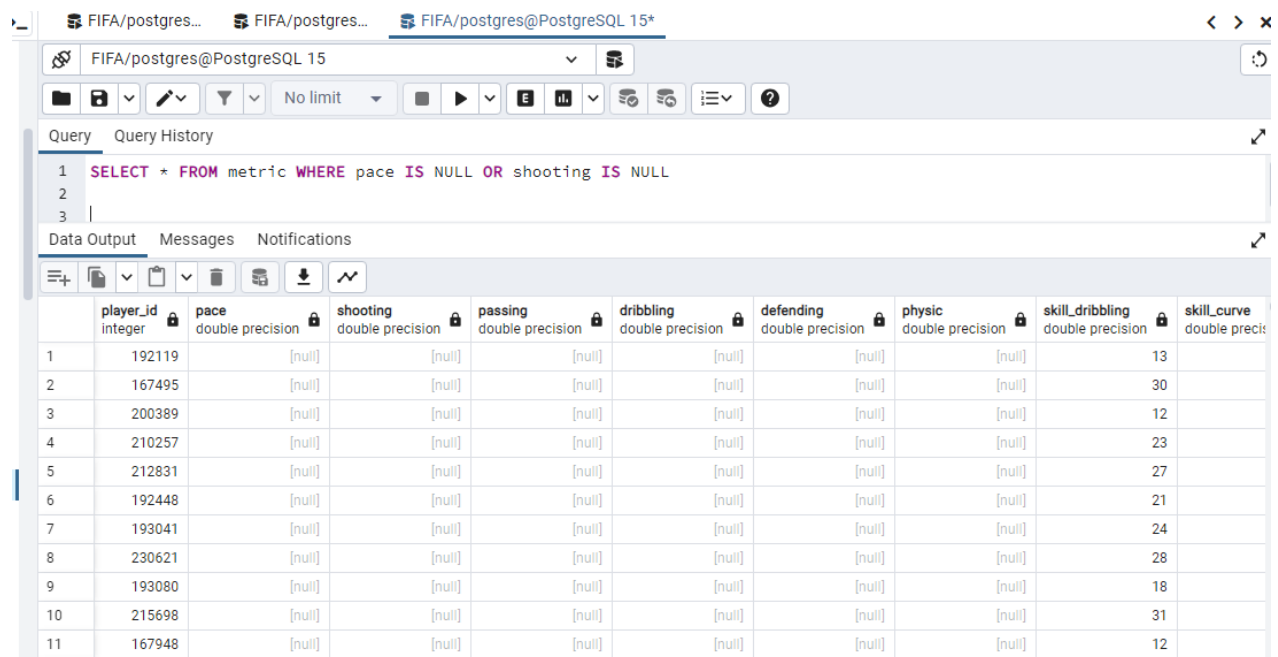
### Group 7

### Phase III

#### Data Cleaning:

Data cleaning in SQL involves manipulating and modifying the data in a database to handle issues such as missing or inconsistent values. It aims to enhance the reliability and accuracy of datasets.

We had some null values in some of the column, hence we would be eliminating those.



The screenshot shows a PostgreSQL query editor with the following query:

```
1 SELECT * FROM metric WHERE pace IS NULL OR shooting IS NULL
2
3
```

The query results are displayed in a table with 11 rows. The columns are: player\_id, pace, shooting, passing, dribbling, defending, physic, skill\_dribbling, and skill\_curve. The values for pace and shooting are all null, while the other columns contain numerical values.

|    | player_id<br>integer | pace<br>double precision | shooting<br>double precision | passing<br>double precision | dribbling<br>double precision | defending<br>double precision | physic<br>double precision | skill_dribbling<br>double precision | skill_curve<br>double prec |
|----|----------------------|--------------------------|------------------------------|-----------------------------|-------------------------------|-------------------------------|----------------------------|-------------------------------------|----------------------------|
| 1  | 192119               | [null]                   | [null]                       | [null]                      | [null]                        | [null]                        | [null]                     | [null]                              | 13                         |
| 2  | 167495               | [null]                   | [null]                       | [null]                      | [null]                        | [null]                        | [null]                     | [null]                              | 30                         |
| 3  | 200389               | [null]                   | [null]                       | [null]                      | [null]                        | [null]                        | [null]                     | [null]                              | 12                         |
| 4  | 210257               | [null]                   | [null]                       | [null]                      | [null]                        | [null]                        | [null]                     | [null]                              | 23                         |
| 5  | 212831               | [null]                   | [null]                       | [null]                      | [null]                        | [null]                        | [null]                     | [null]                              | 27                         |
| 6  | 192448               | [null]                   | [null]                       | [null]                      | [null]                        | [null]                        | [null]                     | [null]                              | 21                         |
| 7  | 193041               | [null]                   | [null]                       | [null]                      | [null]                        | [null]                        | [null]                     | [null]                              | 24                         |
| 8  | 230621               | [null]                   | [null]                       | [null]                      | [null]                        | [null]                        | [null]                     | [null]                              | 28                         |
| 9  | 193080               | [null]                   | [null]                       | [null]                      | [null]                        | [null]                        | [null]                     | [null]                              | 18                         |
| 10 | 215698               | [null]                   | [null]                       | [null]                      | [null]                        | [null]                        | [null]                     | [null]                              | 31                         |
| 11 | 167948               | [null]                   | [null]                       | [null]                      | [null]                        | [null]                        | [null]                     | [null]                              | 12                         |

We have some entries that have null values hence we will be eliminating those records.

</

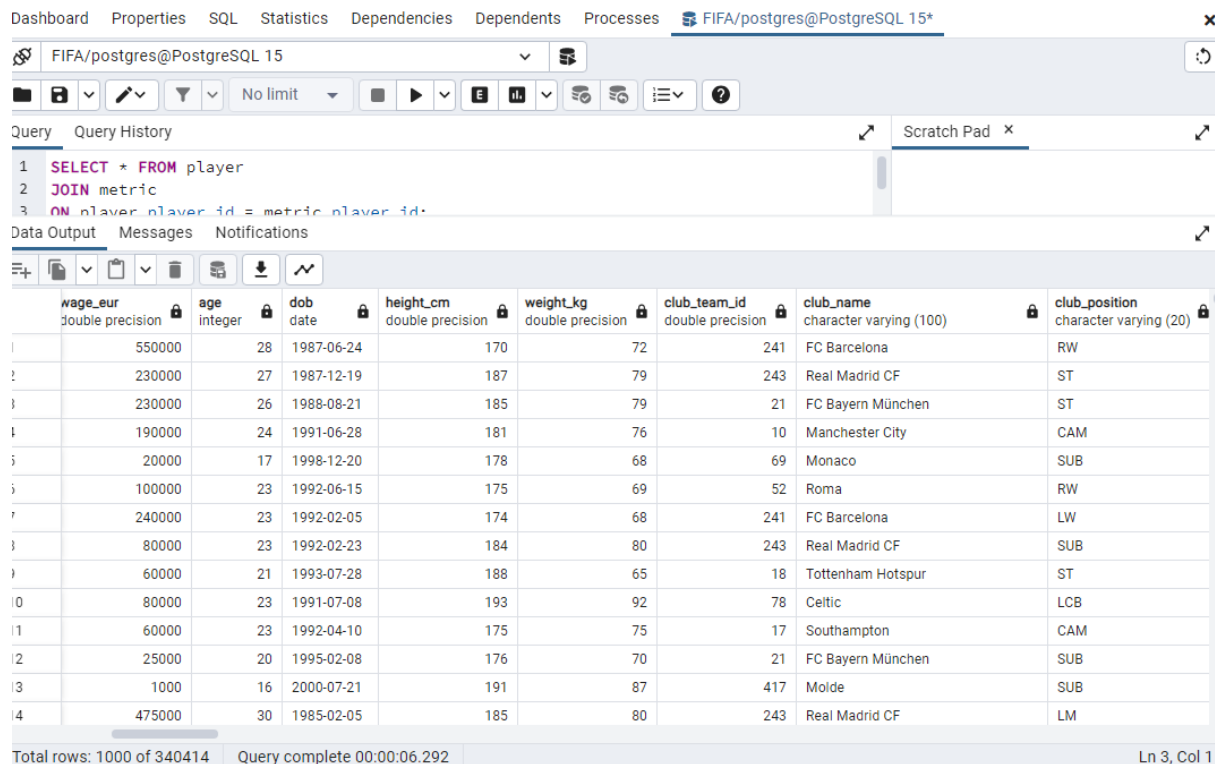
FIFA/postgres@PostgreSQL 15

As seen now on deleting those records have been removed and we have now have the cleaned data with no null values.

Hence, data cleaning has been performed.

## Data Integrity:

Maintaining data integrity in SQL, especially when joining tables, involves enforcing constraints, relationships, and actions that ensure the accuracy and consistency of the data. Hence we are integrating some tables in our database for ease of access and getting accuracy.



The screenshot shows a PostgreSQL query editor interface. The query being executed is:

```
1 SELECT * FROM player
2 JOIN metric
3 ON player.player_id = metric.player_id;
```

The results are displayed in a table with the following columns and data:

|    | wage_eur<br>double precision | age<br>integer | dob<br>date | height_cm<br>double precision | weight_kg<br>double precision | club_team_id<br>double precision | club_name<br>character varying (100) | club_position<br>character varying (20) |
|----|------------------------------|----------------|-------------|-------------------------------|-------------------------------|----------------------------------|--------------------------------------|---|
| 1  | 550000                       | 28             | 1987-06-24  | 170                           | 72                            | 241                              | FC Barcelona                         | RW                                      |
| 2  | 230000                       | 27             | 1987-12-19  | 187                           | 79                            | 243                              | Real Madrid CF                       | ST                                      |
| 3  | 230000                       | 26             | 1988-08-21  | 185                           | 79                            | 21                               | FC Bayern München                    | ST                                      |
| 4  | 190000                       | 24             | 1991-06-28  | 181                           | 76                            | 10                               | Manchester City                      | CAM                                     |
| 5  | 20000                        | 17             | 1998-12-20  | 178                           | 68                            | 69                               | Monaco                               | SUB                                     |
| 6  | 100000                       | 23             | 1992-06-15  | 175                           | 69                            | 52                               | Roma                                 | RW                                      |
| 7  | 240000                       | 23             | 1992-02-05  | 174                           | 68                            | 241                              | FC Barcelona                         | LW                                      |
| 8  | 80000                        | 23             | 1992-02-23  | 184                           | 80                            | 243                              | Real Madrid CF                       | SUB                                     |
| 9  | 60000                        | 21             | 1993-07-28  | 188                           | 65                            | 18                               | Tottenham Hotspur                    | ST                                      |
| 10 | 80000                        | 23             | 1991-07-08  | 193                           | 92                            | 78                               | Celtic                               | LCB                                     |
| 11 | 60000                        | 23             | 1992-04-10  | 175                           | 75                            | 17                               | Southampton                          | CAM                                     |
| 12 | 25000                        | 20             | 1995-02-08  | 176                           | 70                            | 21                               | FC Bayern München                    | SUB                                     |
| 13 | 1000                         | 16             | 2000-07-21  | 191                           | 87                            | 417                              | Molde                                | SUB                                     |
| 14 | 475000                       | 30             | 1985-02-05  | 185                           | 80                            | 243                              | Real Madrid CF                       | LM                                      |

Total rows: 1000 of 340414    Query complete 00:00:06.292    Ln 3, Col 1

## Relational Model vs Document Oriented Model:

Document Oriented model on the other hand, stores data in flexible, JSON-like documents. Each document contains key-value pairs, and collections of documents are used to represent entities. The structure within a document can be hierarchical and nested. It has a dynamic or schema-less structure, allowing for more flexibility. Each document in a collection can have different fields, and new fields can be added without affecting existing documents.

The relational model includes setting data into tables with rows and columns. Each row represents a record, and each column represents a specific attribute of

that record. Relationships between tables are established through keys. It however has a rigid schema and changes to the schema may be needed to alter the existing data. Relational databases keep data integrity through constraints such as primary keys, foreign keys, and unique constraints. This ensures that data is accurate and consistent, reducing the risk of errors hence it is more preferred over document-oriented model. Additionally, the relational databases follow the ACID properties which help in maintaining reliability and integrity of the database. Hence, for all these reasons, usage of relational database would be better for itemset mining hence we will be proceeding it for the querying.

```

20         s+=f"JOIN {table} p{i+1} ON p{i+1}.{clubname}=p{i}.{clubname} AND p{i+1}.playe
21     return s
22
23 def create_lattice(max_level=5):
24     level=1
25     lastresult=None
26     while level<=max_level:
27         query=f"""
28             select {getPlayers(level)},COUNT(DISTINCT p1.year) AS count from {table} p1
29             {getJoin(level)}
30             GROUP BY {",".join([f"p{i}.{player}" for i in range(1,level+1)])}
31             HAVING COUNT(DISTINCT p1.year) >= {support};
32         """
33

```

PROBLEMS TERMINAL OUTPUT PORTS DEBUG CONSOLE

```

GROUP BY p1.player,p2.player,p3.player
HAVING COUNT(DISTINCT p1.year) >= 2;

L4 - 7

select p1.player as player1,p2.player as player2,p3.player as player3,p4.player as player4,COUNT(DISTIN
CT p1.year) AS count from player p1
JOIN player p2 ON p2.club_name=p1.club_name AND p2.player > p1.player
JOIN player p3 ON p3.club_name=p2.club_name AND p3.player > p2.player
JOIN player p4 ON p4.club_name=p3.club_name AND p4.player > p3.player

GROUP BY p1.player,p2.player,p3.player,p4.player

```

The association rules we have observed is that there are players who have been loyal to the same club for an extended period of time. Upon implementation of itemset mining, we have received the output of 5 lattices. Lattice 1 displays 5132 combinations of players who have been in the same club for a long time. Lattice 2 displays 741 combinations of players. Lattice 3 exhibits 39 combinations of players. Lattice 4 exhibits 7 combinations of players - so there are 7 combinations of 4 players. And finally lattice 5 has no combinations of players.

Query

Query History

1

```
select p1.player as player1,p2.player as player2,p3.player as player3,p4.player as player4,
2     COUNT(DISTINCT p1.year) AS count from player p1
3     JOIN player p2 ON p2.club_name=p1.club_name AND p2.player > p1.player
4     JOIN player p3 ON p3.club_name=p2.club_name AND p3.player > p2.player
5     JOIN player p4 ON p4.club_name=p3.club_name AND p4.player > p3.player
6
7     GROUP BY p1.player,p2.player,p3.player,p4.player
8     HAVING COUNT(DISTINCT p1.year) >= 2;
```

Data Output

Messages

Notifications

≡+

📄

▼

📋

▼

🗑️

🗑️

📥

⬇️

📈

|   | player1<br>integer | player2<br>integer | player3<br>integer | player4<br>integer |
|---|--------------------|--------------------|--------------------|--------------------|
| 1 | 188545             | 192448             | 228702             | 189332             |

This is the data at the last level of the lattice. The players referenced by these ids are Robert Lewandowski, Mark Stagen, Frenkie de Jong and Jordi Ramos. This means that these four players have played in the same club for the longest period of time.