

**Exploratory Analysis of the FIFA World Cup 2022**

Report by: Farrukh Nizam Arain  
Designation: Data Analyst  
Date: 25th February 2023

**Introduction:**

FIFA was established on May 21, 1904, by seven national associations — Belgium, Denmark, France, Netherlands, Spain, Sweden and Switzerland — to “promote the game of Association Football (as opposed to rugby or American football), to foster friendly relations among National Associations, Confederations, and their officials and players, by promoting the organization of football matches at all levels, and to control every type of association football by taking steps as shall be deemed necessary or advisable.”  
  
FIFA’s birth was a result of the growing number of international games shortly after the dawn of the 20th century. Soccer leaders in Europe felt that such expanded competition required a governing body, and under the leadership of Robert Guerin, a French journalist, the seven founding members gathered in Paris to shape the future of the sport. Guerin, FIFA’s first president, presided over the organization from 1904 to 1906. Seven other men have also served as FIFA president, including Jules Rimet for 33 years from 1921 to 1954.  
  
Currently, Italy’s Gianni Infantino serves as FIFA president, having been elected in 2016’s Extraordinary Congress held in the wake of corruption allegations against numerous FIFA Officials that resulted in former president Joseph “Sepp” Blatter stepping aside and then being banned from FIFA by its Ethics Committee. FIFA’s general secretary since 2009, Infantino will serve a three-year term as FIFA president.  
  
The Qatar 2022, FIFA World Cup was the 22nd FIFA World Cup. It has been hosted in Qatar from 20 November to 18 December 2022. It was the first World Cup hosted in the Arab world, and the second to be hosted fully in Asia. This was the last World Cup with 32 teams, the next World Cups are going to have 48 teams. The tournament was played in November and December because Qatar is a very hot country. This was the first World Cup that isn't played in May, June or July. The previous champions are France.

**Problem Statement:**

The purpose of my assessment in this project is to answer the questions below by analyzing the different aspects of participating countries in the football world cups. Some of the important questions are:

1. How have host countries performed in World Cups over time? Did Qatar follow a similar path?
2. Based on recent form and historical dominance, which countries underperformed, overachieved, and showed expected performances in the last tournament?
3. Which clubs had the most players who participated in the world cup?
4. How would you rate the World Cup Performances of the Pakistani Football Team
5. Which Teams has the most Wins & percentage of wins in all the matches other than world cups?

**Preparation of the Datasets:**

Public datasets were downloaded at the following [link](https://www.mavenanalytics.io/data-playground) provided by Maven Analytics under this [license](https://www.mavenanalytics.io/terms-for-individuals). No issues with bias and credibility were found with the data through the methodology of ROCCC.

Following are the list of files (in CSV format) and the descriptions:

|  |  |
| --- | --- |
| Filename | Description |
| 2022\_world\_cup\_groups.csv | Groups of teams participating in the world cup |
| 2022\_world\_cup\_matches.csv | Matches scheduled in the tournament |
| 2022\_world\_cup\_squads.csv | Players participated in the international teams |
| international\_matches.csv | All the international matches held in the past |
| world\_cup\_matches.csv | All the world cup matches held in history |
| world\_cups.csv | Results of the past world cups around the world |

After inspecting the datasets mentioned above, it was quickly observed that the data inside the tables was not only incorrect, but also redundancy (duplicate data) and irrelevant data was present. As I already knew that such datasets are most often than not the sole reason of issues stemming from database modifications such as insertions, deletions, and updates. So, I decided to process and normalize the datasets as much as I could for preventing anomalies and error-free data analysis and visualization.

Some of the reasons of Data Normalization are:

* Making the database more efficient.
* Preventing the same data from being stored in more than one place (called an “insert anomaly”)
* Preventing updates being made to some data but not others (called an “update anomaly”)
* Preventing data not being deleted when it is supposed to be, or from data being lost when it is not supposed to be (called a “delete anomaly”)
* Ensuring the data is accurate.
* Reducing the storage space that a database takes up.
* Ensuring the queries on a database run as fast as possible.

**Processing of the Datasets:**

The tools that were used for the data processing process were:

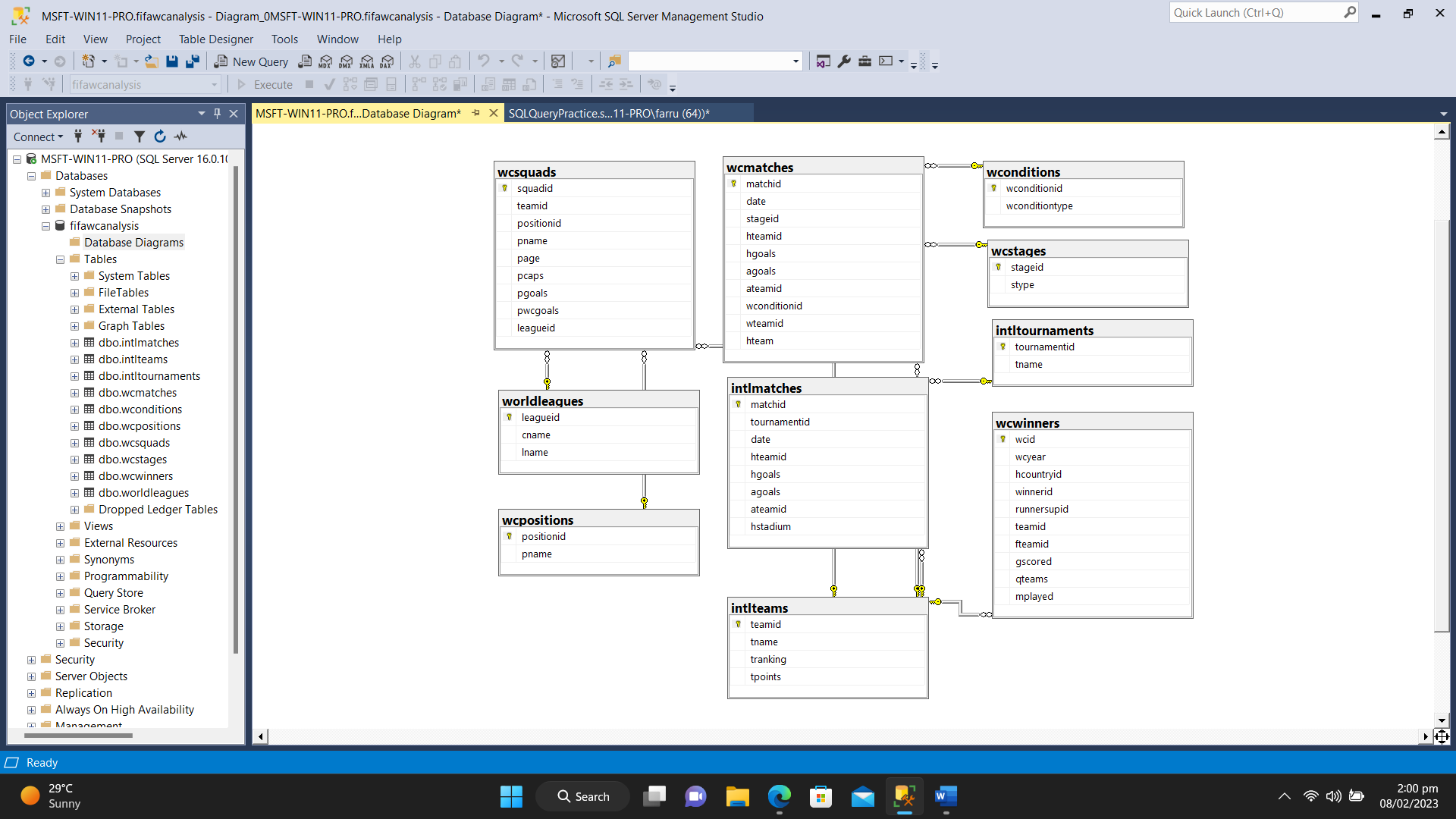
* Microsoft Excel 365
* Microsoft SQL Server 2022
* SQL Server Management Studio 19

As the dataset was not to be considered as bigdata and it could’ve been easily cleaned in Microsoft Excel 365, so I went ahead with using a spreadsheet software to do most of the cleaning and modelling purposes.

Additionally, I transformed the datasets to comply with the normalization standards to eliminate the anomalies mentioned below:

* First Normal Form:
  + The combination of all columns made a unique row every single time.
  + A field was used to uniquely identify the rows.
* Second Normal Form:
  + Fulfilled the requirements of the first normal form.
  + Each non-key attribute was functionally dependent on the primary key.
* Third Normal Form:
  + Fulfilled the requirements of the second normal form.
  + Had no transitive functional dependency.

After executing all the changes and modelling the datasets, mentioned below was the final database diagram which was used for analysis and visualization purposes. As you can see from the diagram that the database was generated to be efficient enough to not have anomalies and other errors.



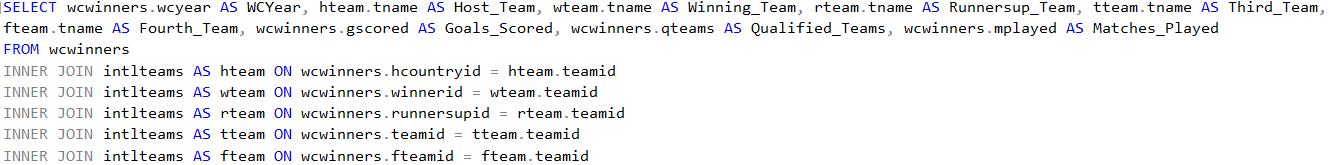
In addition to the database structure, I also developed the data dictionary to make one understands the fields of the database and which primary and foreign keys were used in the database’s structure.



**Analysing the Datasets through Querying:**

After importing the data from the dataset, I used every query and functions possible to extract patterns through exploratory analysis.

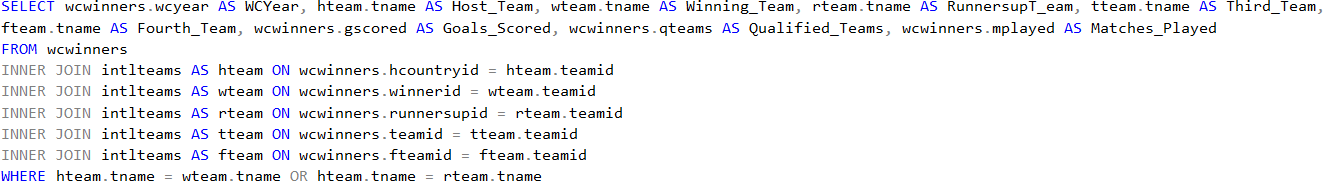
**How have host countries performed in World Cups over time? Did Qatar follow a similar path?**  
First, I queried the *‘wcinners’* table to know which host countries have won any of the medals or even a trophy in the past.



After executing the query mentioned above, I observed that total 22 world cups which took from 1930 till 2022 world cup. As the data was time consuming to analyze and pass on the judgment according to it, I filtered the horizon to know the performances of host teams.



By running the query described below, I filtered out data for those instances where host country either won the world cup or were placed as a runner up to know their performances.



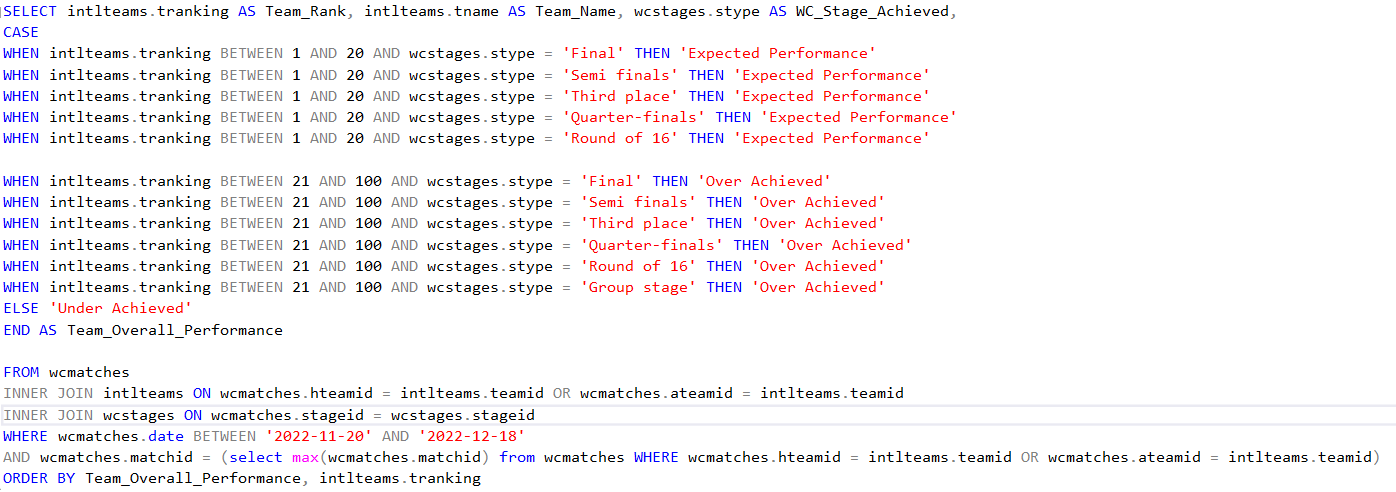
Quite astonishingly, only **six** times host countries won the world up and only **two** times in history they were in second place.



So, according to the tabular data, it is quite evident that Qatar followed the similar path of host countries not performing well during the world cup, even though they were playing in their own turfs.

**Based on recent form and historical dominance, which countries underperformed in 2022 tournament? Which countries overachieved?**

The analysis for this question was a bit tricky one as I had to go through heaps of data about recent performances and compared it with the performances with the last world cup for a quick and easy analysis.



I distributed the ranks of the teams into different leagues and compared the stages that they reached and using the CASE statements within the query, I designated an identifier that either their performances were ‘Expected Performance’, ‘Overachieved’, or ‘Under Achieved.’

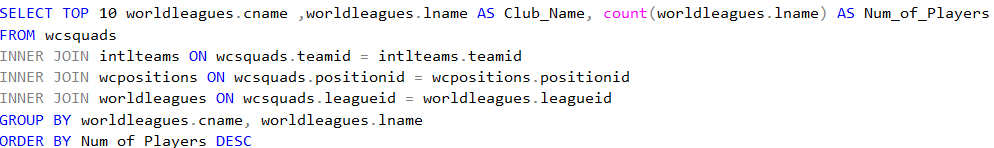


As per the tabular data, most of the teams overachieved with their performances during the last world cup, along with the usual high ranked teams. However, there were some teams like Belgium, Germany, Mexico, Uruguay, and Denmark which didn’t qualify past the group stages, and they underachieved.

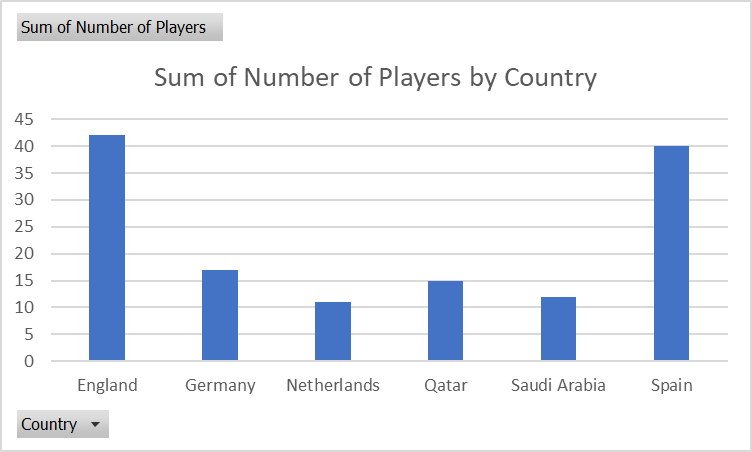
**Which clubs had the most players who participated in the world cup?**

World cup is an event which gives an opportunity for players from around the world to join them teams and play for their country. Due to the popularity of these tournaments, major domestic leagues also big for the players and they contribute to the well-being of the players as well.

When it comes to analysis of the relevant data, I ran an inner join query to extract the top countries and clubs which hires the most international players. Additionally, it displayed the popularity of football in those countries as leagues invest millions in hiring them.

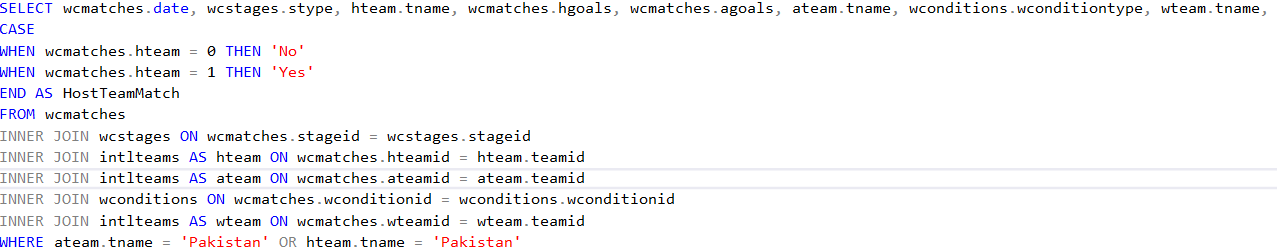


Most of the players are hired by those leagues where football is very popular, except for Qatar’s Al-Sadd where football is not that much popular in comparison of other sports played there.

**World Cup Performance of Pakistan**

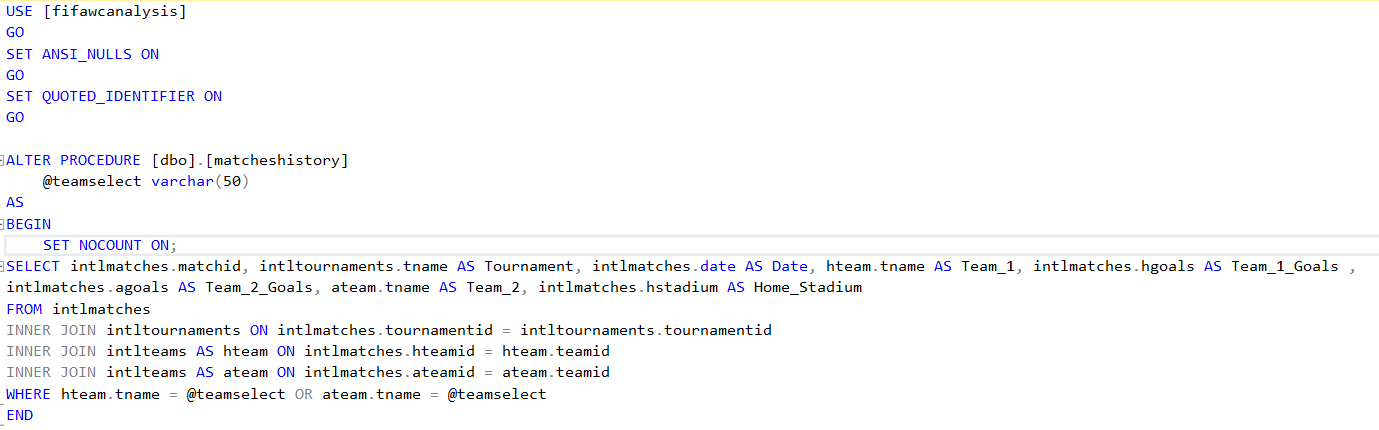
Even though I am an avid Pakistani football fan, hardly anyone plays or follow it in comparison of cricket, which is an unofficial national sport over here. To prove my point, I conducted analysis and unsurprisingly the abysmal performances are shown from the past. Not even once that they have entered the group stages of any world cup.





If we talk about international matches played by Pakistani Football Team, I observed that the record has been abysmal and out of 14 times, Pakistan has won only two times.

For analysis, first I created stored procedure and named it dbo.matcheshistory



Then I executed it by inserting a variable value of ‘Pakistan’





**Conclusion of the Analysis:**

It was a great learning experience for me during this professional project, as it helped me not only to learn new ways of analysis through SQL queries and using the functions of SQL Server 2022, but also to find interesting patterns and findings of the biggest sporting event of football that I love the most.