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 provided by Maven Analytics under this license. 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:

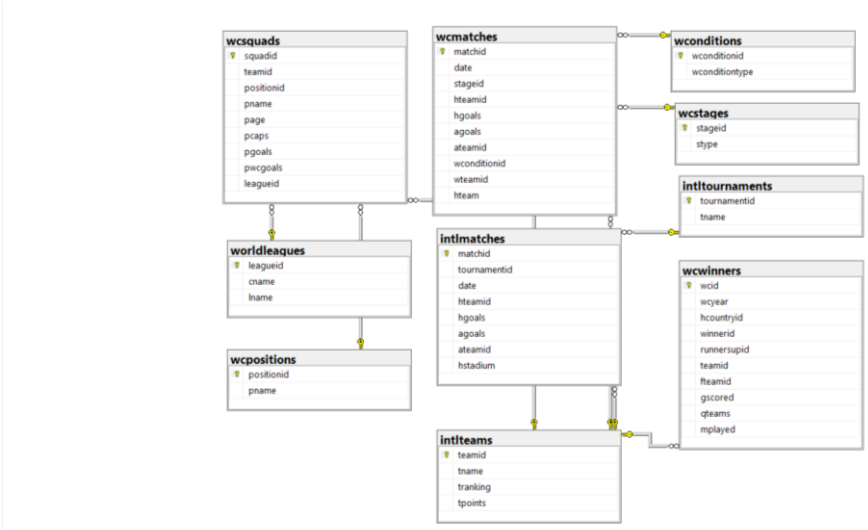
- Microsoft Excel 365
- Microsoft SQL Server 2022
- SQL Server Management Studio 19

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.

Table	Field	Description
intlmatches	matchid	unique identifier (PK) for the match
intlmatches	tournamentid	unique identifier (FK) for the tournament
intlmatches	date	date of the match
intlmatches	hteamid	unique identifier (FK) for first team of the match
intlmatches	hgoals	goals scored by the first team
intlmatches	agoals	goals scored by the second team
intlmatches	ateamid	unique identifier (FK) for second team of the matc
intlmatches	hstadium	match in the host stadium or on a neutral venue
intlteams	teamid	unique identifier (PK) for the team
intlteams	tname	team name
intlteams	tranking	team in the FIFA ranking system
intlteams	tpoints	total team points
intltournaments	tournamentid	unique identifier (PK) for the tournament
intltournaments	tname	tournament name
wcmatches	matchid	unique identifier (PK) of world cup matches
wcmatches	date	date of the match
wcmatches	stageid	unique identifier (FK) of world cup's stage
wcmatches	hteamid	unique identifier (FK) of the first team
wcmatches	hgoals	goals scored by the first team
wcmatches	agoals	goals scord by the second team
wcmatches	ateamid	unique identifier (FK) of the second team
wcmatches	wconditionid	unique identifier (PK) of type of win
wcmatches	wteamid	unique identifer (FK) of the winning team
wcmatches	hteam	host venue / stadium or not
wconditions	wconditionid	unique identifier (PK) of winning condition
wconditions	wconditiontyp	type of win of the match
wcpositions	positionid	unique identifier (PK) of position of the player
wcpositions	pname	position name of the player
wcsquads	squadid	unique identifier (PK) of world cup squad
wcsquads	teamid	unique identifier (FK) of world cup team
wcsquads	positionid	unique identifer (FK) of the player's position
wcsquads	pname	player name
wcsquads	page	player's age
wcsquads	pcaps	player's caps
wcsquads	pgoals	player's goals
wcsquads	pwcgoals	player's goals in the world cups
wcsquads	leagueid	unique identifier (FK) of leagues the player are in
wcstages	stageid	unique identifier (PK) of world cup stage
wcstages	stype	stage name of the world cups
wcwinners	wcid	unique identifier (PK) of world cup
wcwinners	wcyear	world cup year
wcwinners	hcountryid	unique identifier (FK) of host county
wcwinners	winnerid	unique identifier (FK) of winning team
wcwinners	runnersupid	unique identifier (FK) of runner's up team
wcwinners	teamid	unique identifier (FK) of 3rd place team
wcwinners	fteamid	unique identifier (FK) of 4th position team
wcwinners	gscored	goals scored in the world cup
wcwinners	qteams	qualified teams in the world cup
wcwinners	mplayed	matches played in the world cup
worldleagues	leagueid	unique identifier (PK) of the domestic league
worldleagues	cname	country name
worldleagues	lname	league name

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.

```
SELECT wcwinners.wcyear AS WCYear, hteam.tname AS Host_Team, wteam.tname AS Winning_Team, rteam.tname AS Runnersup_Team, tteam.tname AS Third_Team,
fteam.tname AS Fourth_Team, wcwinners.gscored AS Goals_Scored, wcwinners.qteams AS Qualified_Teams, wcwinners.mplayed AS Matches_Played
FROM wcwinners
INNER JOIN intlteams AS hteam ON wcwinners.hcountryid = hteam.teamid
INNER JOIN intlteams AS wteam ON wcwinners.winnerid = wteam.teamid
INNER JOIN intlteams AS rteam ON wcwinners.runnersupid = rteam.teamid
INNER JOIN intlteams AS tteam ON wcwinners.teamid = tteam.teamid
INNER JOIN intlteams AS fteam ON wcwinners.fteamid = fteam.teamid
```

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.

WC Year	Host Team	Winning Team	Runnersup Team	Third Team	Fourth Team	Goals Scored	Qualified Teams	Matches Played
1930	Uruguay	Uruguay	Argentina	United States	Yugoslavia	70	13	18
1934	Italy	Italy	Czechoslovakia	Germany	Austria	70	16	17
1938	France	Italy	Hungary	Brazil	Sweden	84	15	18
1950	Brazil	Uruguay	Brazil	Sweden	Spain	88	13	22
1954	Switzerland	Germany	Hungary	Austria	Uruguay	140	16	26
1958	Sweden	Brazil	Sweden	France	Germany	126	16	35
1962	Chile	Brazil	Czechoslovakia	Chile	Yugoslavia	89	16	32
1966	England	England	Germany	Portugal	Russia	89	16	32
1970	Mexico	Brazil	Italy	Germany	Uruguay	95	16	32
1974	Germany	Germany	Netherlands	Poland	Brazil	97	16	38
1978	Argentina	Argentina	Netherlands	Brazil	Italy	102	16	38
1982	Spain	Italy	Germany	Poland	France	146	24	52
1986	Mexico	Argentina	Germany	France	Belgium	132	24	52
1990	Italy	Germany	Argentina	Italy	England	115	24	52
1994	Uzbekistan	Brazil	Italy	Sweden	Bulgaria	141	24	52
1998	France	France	Brazil	Croatia	Netherlands	171	32	64
2002	Kosovo	Brazil	Germany	Turkey	South Korea	161	32	64
2006	Germany	Italy	France	Germany	Portugal	147	32	64
2010	South Africa	Spain	Netherlands	Germany	Uruguay	145	32	64
2014	Brazil	Germany	Argentina	Netherlands	Brazil	171	32	64
2018	Russia	France	Croatia	Belgium	England	169	32	64
2022	Qatar	Argentina	France	Croatia	Morocco	172	32	64

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.

```
SELECT wcwinners.wcyear AS WCYear, hteam.tname AS Host_Team, wteam.tname AS Winning_Team, rteam.tname AS RunnersupT_eam, tteam.tname AS Third_Team,
fteam.tname AS Fourth_Team, wcwinners.gscored AS Goals_Scored, wcwinners.qteams AS Qualified_Teams, wcwinners.mplayed AS Matches_Played
FROM wcwinners
INNER JOIN intlteams AS hteam ON wcwinners.hcountryid = hteam.teamid
INNER JOIN intlteams AS wteam ON wcwinners.winnerid = wteam.teamid
INNER JOIN intlteams AS rteam ON wcwinners.runnersupid = rteam.teamid
INNER JOIN intlteams AS tteam ON wcwinners.teamid = tteam.teamid
INNER JOIN intlteams AS fteam ON wcwinners.fteamid = fteam.teamid
WHERE hteam.tname = wteam.tname OR hteam.tname = rteam.tname
```

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

WCYear	Host_Team	Winning_Team	RunnersupT_eam	Third_Team	Fourth_Team	Goals_Scored	Qualified_Teams	Matches_Played
1930	Uruguay	Uruguay	Argentina	United States	Yugoslavia	70	13	18
1934	Italy	Italy	Czechoslovakia	Germany	Austria	70	16	17
1950	Brazil	Uruguay	Brazil	Sweden	Spain	88	13	22
1958	Sweden	Brazil	Sweden	France	Germany	126	16	35
1966	England	England	Germany	Portugal	Russia	89	16	32
1974	Germany	Germany	Netherlands	Poland	Brazil	97	16	38
1978	Argentina	Argentina	Netherlands	Brazil	Italy	102	16	38
1998	France	France	Brazil	Croatia	Netherlands	171	32	64

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.

```
SELECT intlteams.tranking AS Team_Rank, intlteams.tname AS Team_Name, wcstages.stype AS WC_Stage_Achieved,
CASE
WHEN intlteams.tranking BETWEEN 1 AND 20 AND wcstages.stype = 'Final' THEN 'Expected Performance'
WHEN intlteams.tranking BETWEEN 1 AND 20 AND wcstages.stype = 'Semi finals' THEN 'Expected Performance'
WHEN intlteams.tranking BETWEEN 1 AND 20 AND wcstages.stype = 'Third place' THEN 'Expected Performance'
WHEN intlteams.tranking BETWEEN 1 AND 20 AND wcstages.stype = 'Quarter-finals' THEN 'Expected Performance'
WHEN intlteams.tranking BETWEEN 1 AND 20 AND wcstages.stype = 'Round of 16' THEN 'Expected Performance'

WHEN intlteams.tranking BETWEEN 21 AND 100 AND wcstages.stype = 'Final' THEN 'Over Achieved'
WHEN intlteams.tranking BETWEEN 21 AND 100 AND wcstages.stype = 'Semi finals' THEN 'Over Achieved'
WHEN intlteams.tranking BETWEEN 21 AND 100 AND wcstages.stype = 'Third place' THEN 'Over Achieved'
WHEN intlteams.tranking BETWEEN 21 AND 100 AND wcstages.stype = 'Quarter-finals' THEN 'Over Achieved'
WHEN intlteams.tranking BETWEEN 21 AND 100 AND wcstages.stype = 'Round of 16' THEN 'Over Achieved'
WHEN intlteams.tranking BETWEEN 21 AND 100 AND wcstages.stype = 'Group stage' THEN 'Over Achieved'
ELSE 'Under Achieved'
END AS Team_Overall_Performance

FROM wcmatches
INNER JOIN intlteams ON wcmatches.hteamid = intlteams.teamid OR wcmatches.ateamid = intlteams.teamid
INNER JOIN wcstages ON wcmatches.stageid = wcstages.stageid
WHERE wcmatches.date BETWEEN '2022-11-20' AND '2022-12-18'
AND wcmatches.matchid = (select max(wcmatches.matchid) from wcmatches WHERE wcmatches.hteamid = intlteams.teamid OR wcmatches.ateamid = intlteams.teamid)
ORDER BY Team_Overall_Performance, intlteams.tranking
```

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.’

Team Rank	Team Name	WC Stage Achieved	Team Overall Performance
1	Brazil	Quarter-finals	Expected Performance
2	Argentina	Final	Expected Performance
3	France	Final	Expected Performance
5	England	Quarter-finals	Expected Performance
6	Netherlands	Quarter-finals	Expected Performance
7	Croatia	Third place	Expected Performance
9	Portugal	Quarter-finals	Expected Performance
10	Spain	Round of 16	Expected Performance
11	Morocco	Third place	Expected Performance
12	Switzerland	Round of 16	Expected Performance
13	United States	Round of 16	Expected Performance
19	Senegal	Round of 16	Expected Performance
20	Japan	Round of 16	Expected Performance
22	Poland	Round of 16	Over Achieved
24	Iran	Group stage	Over Achieved
27	Australia	Round of 16	Over Achieved
28	Wales	Group stage	Over Achieved
29	Serbia	Group stage	Over Achieved
30	Tunisia	Group stage	Over Achieved
32	Costa Rica	Group stage	Over Achieved
33	Cameroon	Group stage	Over Achieved
41	Ecuador	Group stage	Over Achieved
49	Saudi Arabia	Group stage	Over Achieved
53	Canada	Group stage	Over Achieved
58	Ghana	Group stage	Over Achieved
60	Qatar	Group stage	Over Achieved
4	Belgium	Group stage	Under Achieved
14	Germany	Group stage	Under Achieved
15	Mexico	Group stage	Under Achieved
16	Uruguay	Group stage	Under Achieved
18	Denmark	Group stage	Under Achieved
999	South Korea	Round of 16	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?

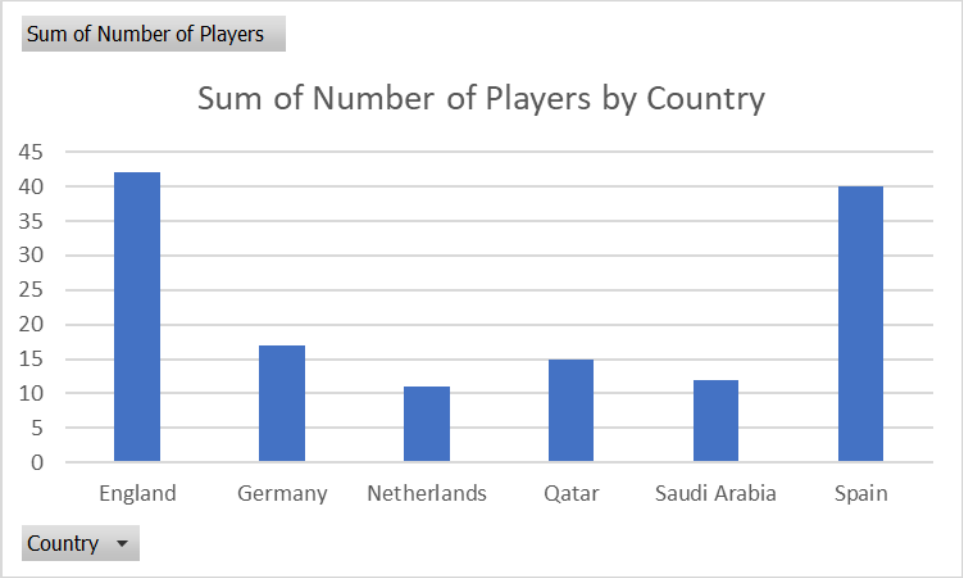
World cup is an event which gives an opportunity for players from around the world to join them teams and play for their country. Due to the popularity of these tournaments, major domestic leagues also big for the players and they contribute to the well-being of the players as well.

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.

```sql
SELECT TOP 10 worldleagues.cname ,worldleagues.lname AS Club_Name, count(worldleagues.lname) AS Num_of_Players
FROM wcsquads
INNER JOIN intlteams ON wcsquads.teamid = intlteams.teamid
INNER JOIN wcpositions ON wcsquads.positionid = wcpositions.positionid
INNER JOIN worldleagues ON wcsquads.leagueid = worldleagues.leagueid
GROUP BY worldleagues.cname, worldleagues.lname
ORDER BY Num of Players DESC
```

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.

Country	Club Name	Number of Players
Germany	Bayern Munich	17
Spain	Barcelona	16
England	Manchester City	16
Qatar	Al-Sadd	15
England	Manchester United	14
Spain	Real Madrid	13
Saudi Arabia	Al-Hilal	12
England	Chelsea	12
Netherlands	Ajax	11
Spain	Atltico Madrid	11



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.

```sql
SELECT wcmatches.date, wcstages.stype, hteam.tname, wcmatches.hgoals, wcmatches.agoals, ateam.tname, wconditions.wconditiontype, wteam.tname,
CASE
WHEN wcmatches.hteam = 0 THEN 'No'
WHEN wcmatches.hteam = 1 THEN 'Yes'
END AS HostTeamMatch
FROM wcmatches
INNER JOIN wcstages ON wcmatches.stageid = wcstages.stageid
INNER JOIN intlteams AS hteam ON wcmatches.hteamid = hteam.teamid
INNER JOIN intlteams AS ateam ON wcmatches.ateamid = ateam.teamid
INNER JOIN wconditions ON wcmatches.wconditionid = wconditions.wconditionid
INNER JOIN intlteams AS wteam ON wcmatches.wteamid = wteam.teamid
WHERE ateam.tname = 'Pakistan' OR hteam.tname = 'Pakistan'
```

date	stype	tname	hgoals	agoals	tname	wconditiontype	tname	HostTeamMatch

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

```
USE [fifawcanalysis]
GO
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO

ALTER PROCEDURE [dbo].[matcheshistory]
    @teamselect varchar(50)
AS
BEGIN
    SET NOCOUNT ON;
SELECT intlmatches.matchid, intltournaments.tname AS Tournament, intlmatches.date AS Date, hteam.tname AS Team_1, intlmatches.hgoals AS Team_1_Goals ,
intlmatches.agoals AS Team_2_Goals, ateam.tname AS Team_2, intlmatches.hstadium AS Home_Stadium
FROM intlmatches
INNER JOIN intltournaments ON intlmatches.tournamentid = intltournaments.tournamentid
INNER JOIN intlteams AS hteam ON intlmatches.hteamid = hteam.teamid
INNER JOIN intlteams AS ateam ON intlmatches.ateamid = ateam.teamid
WHERE hteam.tname = @teamselect OR ateam.tname = @teamselect
END
```

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

```
EXEC matcheshistory @teamselect = 'Pakistan'
```

Date	Tournament	Team 1	Team 1 Goals	Team 2 Goals	Team 2	Home Stadium
10/08/1960 0:00	Merdeka Tournament	Pakistan	3	1	Japan	FALSE
12/09/1962 0:00	Merdeka Tournament	Japan	1	1	Pakistan	FALSE
23/07/1965 0:00	Friendly	Iran	4	1	Pakistan	TRUE
12/03/1969 0:00	Friendly	Iran	9	1	Pakistan	TRUE
13/09/1969 0:00	Friendly	Iran	4	2	Pakistan	FALSE
06/09/1970 0:00	Friendly	Iran	7	0	Pakistan	TRUE
17/01/1974 0:00	Friendly	Pakistan	1	2	Iran	TRUE
13/10/1984 0:00	AFC Asian Cup qualification	South Korea	6	0	Pakistan	FALSE
16/02/1986 0:00	Friendly	Iran	2	0	Pakistan	TRUE
26/04/1986 0:00	Friendly	Pakistan	1	0	South Korea	TRUE
18/04/1988 0:00	AFC Asian Cup qualification	Japan	4	1	Pakistan	FALSE
11/05/1992 0:00	AFC Asian Cup qualification	Iran	7	0	Pakistan	FALSE
06/06/1993 0:00	Friendly	Iran	5	0	Pakistan	TRUE
04/04/2000 0:00	AFC Asian Cup qualification	Qatar	5	0	Pakistan	TRUE

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.