**Scenario**

I chose to work on this simply out of interest. I grew up playing soccer my entire life, so I have a deep love for the sport. My goal is to create an all-star team based on how the players performed in the world cup. I plan on using SQL and Python to clean the data and track the performance of each player. Eventually, I’ll be creating a dashboard that displays all of this information.

**The Data**

Historical data leading up to the 2022 FIFA World Cup, including all the matches from the previous world cups, all international matches for the qualified countries, and the groups, squads, and matches for the upcoming tournament.I found my datasets on Kaggle, GitHub, and Maven analytics.

**Key Findings**

* The highest goal scorers were Kylian Mbappe – France, and Lionel Messi- Argentina
* The players with the most assists were Lionel Messi(3),Antoine Griezmann(3), Ivan Perisic(3),Harry Kane(3), Bruno Fernandez(3)
* The players with the most appearances played for Argentina (Emiliano Martinez, Lionel Messi, Nicholas Otamendi, Rodrigo De Paul, and Nahuel Molina) with 7 appearances.
* The defender with the most successful tackles was Achraf Hakimi from Morocco with 17 successful tackles.
* Kylian Mbappe had the most successful dribbles with a total of 25
* The Midfielder with the most completed passes was Rodri(Spain) with a total of 642. He had a 93% pass completed rate
* Manchester City and Bayern Munich had the most International Players
* The players with the most minutes played were from Croatia and Argentina
* Lionel Messi was fouled the most times with a total of 22
* Denzel Dumfries (Netherlands) caused the most fouls with a total of 17
* Argentina players were fouled the most with a total of 115 times
* Argentina also committed the most fouls with a total of 100 times
* France had the most shots with a total of 100
* Argentina had the most shots on target with a total of 41

**Step by step process**:

1. I began by getting all the datasets that I needed. I’m working on a world cup dataset and will be doing analysis on the 2022 world cup.
2. First I started by creating my database on MySQL, then created a schema and named it world cup. I then inserted the csv files into the workbench and checked all 6 of the tables to make sure they have the appropriate data types.
3. One dataset was unable to load onto SQL so I will be doing analysis through excel. I began by eliminating rows that I would not need. I filtered data, and checked for any blanks and what I could possibly insert in those cells. I got rid of rows that I did not need in my opinion. I looked for duplicates and removed made sure to remove only unnecessary duplicates. I made sure every column had the appropriate data type and that they all followed same structure. (ex. Made sure all dates were in the same format rather than one saying march 9th, 2012 and another saying 03/12/2012).
4. Wrote a list of over 20 questions
5. Did data cleaning and changed various rows and columns so it can be easier to work with
6. I manually matched over 600 rows/players names with a unique id number in 2 different tables so I can join them In SQL.
7. After I had the clean data ready, I uploaded them onto SQL and python and began answering all the questions I came up with (see ipynb file and PDF document)
8. The plan is to eventually create a dashboard that will display the top 11 players who performed the best in the worldcup