worldcup

April 28, 2023

1 2022 Qatar World Cup Analysis

The World Cup is one of the, if not the largest, sporting event in the world. The World Cup is a tournament consisting of some of the best qualifying nations in international football. Countries are divided into groups, where the top two teams from each group move on to the 'knockout rounds', effectively dwindling down the competition until one team is left standing.The World Cup is a global event that transcends borders, cultures, and languages, captivating audiences with the sheer talent, passion, and dedication of the world's top footballers. The World Cup also serves as a platform for countries to showcase their national identity and pride, and to unite their citizens around a common goal. Whether you are a casual fan or a die-hard supporter, the World Cup is an unforgettable experience that captures the imagination and inspires a sense of unity that extends far beyond the pitch. And for the players, it is considered the highest honor in the sport; one that demands a lot of passion and excellence in order to persevere and bring home glory to their country.

The 2022 World Cup in Qatar has recently concluded with an Argentinian extra-time victory in the final over France; a dramatic victory that perfectly captures the essence of the tournament. The dataset that we will be utilizing in our project focuses on each individual match from the tournament, and contains a large amount of data about the match itself, specifically for each team; possession, shots attempted, shots on goal, total passes, etc. We are hoping to showcase some trends, and insights that can summarize and visualize the tournament effectively through data.

1.1 Importing packages

1.2 Loading, cleaning, and exploring datasets

We have three main datasets which we will be working with in our project; our main dataset is the `matches` dataset which contains information about the matches played in the 2022 Qatar World Cup. The second dataset, `countries`, relates to the individual country statistics, which we will be aggregating from our information about each match in order to get a better understanding of the teams and their performance. Lastly, we are going to take a look at a `players` dataset, which contains much information about the actual participants in the tournament, and their performance in the matches.

1.2.1 Matches Dataset - loading and cleaning

	team1	team2	possession team1	possession team2	\
0	QATAR	ECUADOR	42%	50%	
1	ENGLAND	IRAN	72%	19%	
2	SENEGAL	NETHERLANDS	44%	45%	

```
3  UNITED STATES         WALES              51%              39%
4      ARGENTINA  SAUDI ARABIA              64%              24%

  possession in contest  number of goals team1  number of goals team2  \
0                    8%                      0                      2
1                    9%                      6                      2
2                   11%                      0                      2
3                   10%                      1                      1
4                   12%                      1                      2

          date     hour category  …  penalties scored team1  \
0  20 NOV 2022  17 : 00  Group A  …                       0
1  21 NOV 2022  14 : 00  Group B  …                       0
2  21 NOV 2022  17 : 00  Group A  …                       0
3  21 NOV 2022  20 : 00  Group B  …                       0
4  22 NOV 2022  11 : 00  Group C  …                       1

   penalties scored team2  goal preventions team1  goal preventions team2  \
0                       1                       6                       5
1                       1                       8                      13
2                       0                       9                      15
3                       1                       7                       7
4                       0                       4                      14

   own goals team1  own goals team2  forced turnovers team1  \
0                0                0                      52
1                0                0                      63
2                0                0                      63
3                0                0                      81
4                0                0                      65

   forced turnovers team2  defensive pressures applied team1  \
0                      72                               256
1                      72                               139
2                      73                               263
3                      72                               242
4                      80                               163

   defensive pressures applied team2
0                                279
1                                416
2                                251
3                                292
4                                361

[5 rows x 88 columns]
```

```
          team1         team2  possession team1  possession team2  \
0         Qatar       Ecuador                42                50
1       England          Iran                72                19
2       Senegal   Netherlands                44                45
3  United States         Wales                51                39
4     Argentina  Saudi Arabia                64                24

   possession in contest  number of goals team1  number of goals team2  \
0                      8                      0                      2
1                      9                      6                      2
2                     11                      0                      2
3                     10                      1                      1
4                     12                      1                      2

  category  total attempts team1  total attempts team2  …  free kicks team2  \
0    Group                    5                     6  …                17
1    Group                   13                     8  …                10
2    Group                   14                     9  …                14
3    Group                    6                     7  …                15
4    Group                   14                     3  …                16

   penalties scored team1  penalties scored team2  goal preventions team1  \
0                       0                       1                       6
1                       0                       1                       8
2                       0                       0                       9
3                       0                       1                       7
4                       1                       0                       4

   goal preventions team2  own goals team1  own goals team2  \
0                       5                0                0
1                      13                0                0
2                      15                0                0
3                       7                0                0
4                      14                0                0

   forced turnovers team1  forced turnovers team2 GroupID
0                      52                      72       A
1                      63                      72       B
2                      63                      73       A
3                      81                      72       B
4                      65                      80       C

[5 rows x 47 columns]

array(['Group', 'Round of 16', 'Quarter-final', 'Semi-final',
       'Play-off for third place', 'Final'], dtype=object)
```

1.2.2 Countries Dataset - loading and cleaning

Nation	Rank	Points	World Cup Wins
Brazil	1	1841.30	5
Belgium	2	1816.71	0
Argentina	3	1773.88	3
France	4	1759.78	2
England	5	1728.47	1

Now that we have our `countries` dataset prepared for data to enter it, we need to start to modify our `matches` dataset, so that any redundant information is discarded, and all the information we need is correctly represented.

We are now going to write some code which will allow us to clean up the way some of this data looks. Ideally, we want to observe these statistics based on country, while we have it here as `team1` or `team2`, which isn't really helpful if we want to get a context of a particular country. So, using the `countries` dataset that we have introduced earlier that just contains their FIFA rank at the time of the World Cup, we will be adding each countries individual statistics to the dataset.

Nation	Rank	Points	World Cup Wins	Possession in Group	\
Brazil	1	1841.30	5	160	
Belgium	2	1816.71	0	149	
Argentina	3	1773.88	3	181	
France	4	1759.78	2	156	
England	5	1728.47	1	181	

Nation	Possession in Round of 16	Possession in Quarter-final	\
Brazil	47	45	
Belgium	0	0	
Argentina	53	44	
France	48	36	
England	54	54	

Nation	Possession in Semi-final	Possession in Play-off for third place	\
Brazil	0	0	
Belgium	0	0	
Argentina	34	0	
France	34	0	
England	0	0	

Nation	Possession in Final	Total Passes in Group	…	\
			…	
Brazil	0	1698	…	
Belgium	0	1779	…	

```
Argentina                   46                  2005  …
France                      40                  1873  …
England                      0                  1947  …

           Forced Turnovers in Group  Forced Turnovers in Round of 16  \
Nation
Brazil                           211                               73
Belgium                          180                                0
Argentina                        176                               67
France                           223                               71
England                          171                               60

           Forced Turnovers in Quarter-final  Forced Turnovers in Semi-final  \
Nation
Brazil                                    77                               0
Belgium                                    0                               0
Argentina                                 79                              85
France                                    54                              72
England                                   49                               0

           Forced Turnovers in Play-off for third place  \
Nation
Brazil                                                0
Belgium                                               0
Argentina                                             0
France                                                0
England                                               0

           Forced Turnovers in Final  GroupID  Average Possession in Group  \
Nation
Brazil                             0        G                        53.33
Belgium                            0        F                        49.67
Argentina                         87        C                        60.33
France                           104        D                        52.00
England                            0        B                        60.33

           Goals Per Game in Group  Total Goals
Nation
Brazil                        1.00            8
Belgium                       0.33            1
Argentina                     1.67           15
France                        2.00           16
England                       3.00           13

[5 rows x 121 columns]

Group A average team ranking: 30.0
```

```
Group B average team ranking: 15.0
Group C average team ranking: 23.25
Group D average team ranking: 20.5
Group E average team ranking: 18.25
Group F average team ranking: 19.25
Group G average team ranking: 20.0
Group H average team ranking: 28.0
```

It seems that based on the mean ranks of our groups, Group B has the highest average rank, which coins this group 'The Group of Death', a term designated to the toughest group in the tournament. On the other hand, Group A seems to have the lowest average rank. Let's take a look at the two groups, and see what that's about.

Nation	Rank
England	5
United States	16
Wales	19
Iran	20

Looking at the group of death, we can see that all 4 teams are ranked in the top 20, with 3 of the teams being very close to each other in rank.

Nation	Rank
Netherlands	8
Senegal	18
Ecuador	44
Qatar	50

This group is much different than Group B; while we can see that we have 2 top 20 nations here, but two of the lower ranked countries in the tournaments as well, in Ecuador and Qatar (the host country)

1.2.3 Players Dataset - loading and cleaning

	player	club	position	age	team	\
0	Aaron Mooy	Celtic	MF	32	Australia	
1	Aaron Ramsey	Nice	MF	31	Wales	
2	Abdelhamid Sabiri	Sampdoria	MF	26	Morocco	
3	Abdelkarim Hassan	Al Sadd SC	DF	29	Qatar	
4	Abderrazak Hamdallah	Al-Ittihad	FW	32	Morocco	
..	…	…	… …	…		
675	Ángel Di María	Juventus	MF	34	Argentina	
676	Ángelo Preciado	Genk	DF	24	Ecuador	
677	Éder Militão	Real Madrid	DF	24	Brazil	
678	Óscar Duarte	Al-Wehda	DF	33	Costa Rica	
679	İlkay Gündoğan	Manchester City	MF	32	Germany	

```
     birth_year  minutes_90s  tackles  tackles_won  tackles_def_3rd  …  \
0          1990          4.0      9.0            6              4.0  …
1          1990          3.0      2.0            0              0.0  …
2          1996          2.0      3.0            1              1.0  …
3          1993          3.0      7.0            3              5.0  …
4          1990          0.8      0.0            0              0.0  …
..          …           …        …             …              …  …
675        1988          3.2      3.0            1              2.0  …
676        1998          2.9      7.0            5              3.0  …
677        1998          3.9      7.0            6              4.0  …
678        1989          3.0      4.0            2              4.0  …
679        1990          2.1      3.0            1              1.0  …

     touches_att_3rd  touches_att_pen_area  touches_live_ball  \
0               26.0                   0.0              255.0
1               42.0                   5.0              147.0
2               13.0                   1.0               86.0
3               17.0                   2.0              193.0
4               12.0                   5.0               28.0
..               …                     …                 …
675            132.0                  17.0              201.0
676             46.0                   3.0              162.0
677             55.0                   6.0              306.0
678              4.0                   1.0              132.0
679             55.0                   6.0              186.0

     dribbles_completed  dribbles  dribbles_completed_pct  miscontrols  \
0                   2.0       3.0                    66.7          5.0
1                   2.0       8.0                    25.0          9.0
2                   0.0       3.0                     0.0          0.0
3                   1.0       5.0                    20.0          2.0
4                   2.0       3.0                    66.7          4.0
..                  …        …                      …            …
675                13.0      25.0                    52.0         10.0
676                 0.0       4.0                     0.0          6.0
677                 0.0       0.0                     NaN          6.0
678                 0.0       0.0                     NaN          1.0
679                 2.0       2.0                   100.0          4.0

     dispossessed  passes_received  progressive_passes_received
0             4.0            152.0                          1.0
1             4.0             98.0                          7.0
2             3.0             54.0                          0.0
3             0.0            138.0                          1.0
4             3.0             18.0                          3.0
..            …               …                             …
```

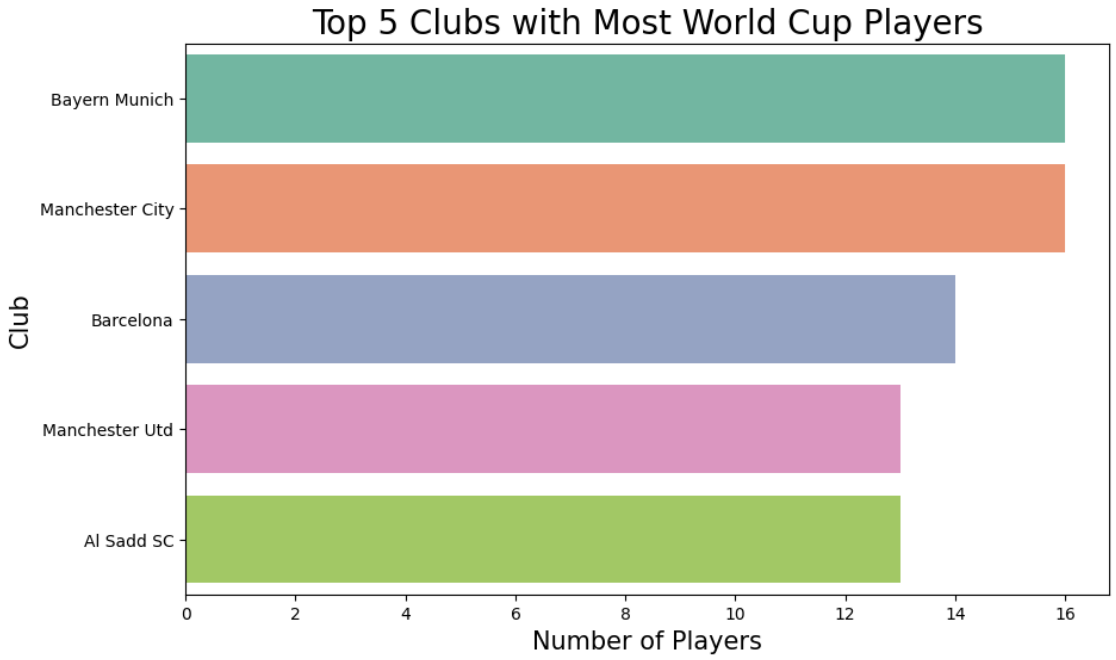
```
675          6.0          163.0                    24.0
676          2.0           81.0                     1.0
677          1.0          217.0                     5.0
678          0.0           70.0                     0.0
679          1.0          142.0                     6.0

[680 rows x 54 columns]
```

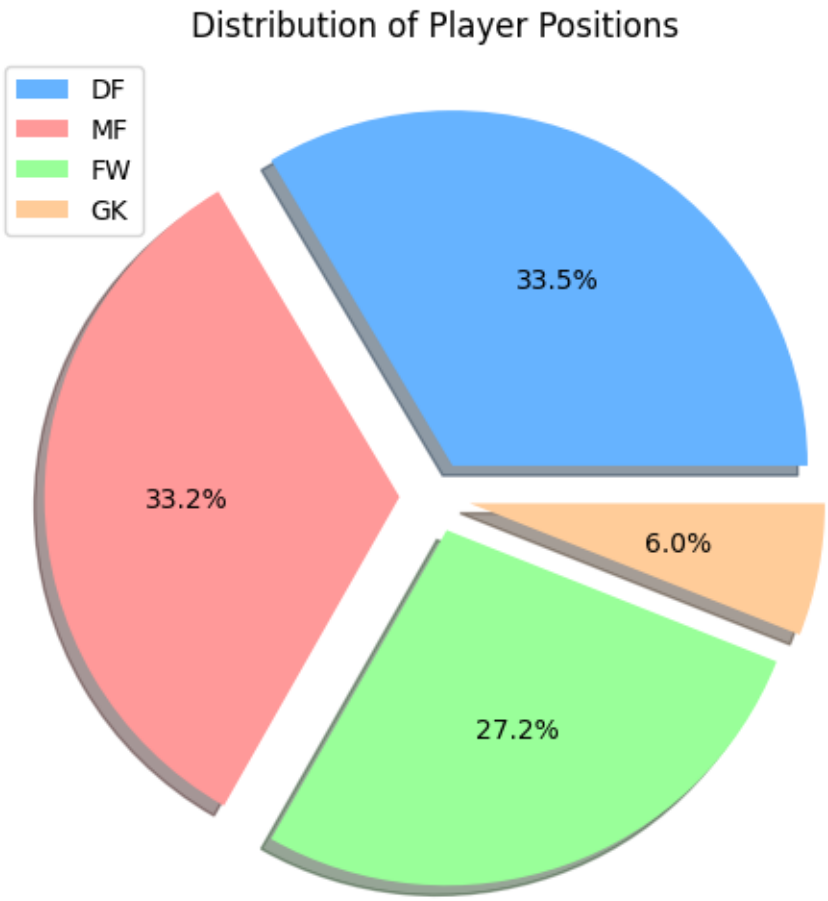
1.3 Visualizing the players, and their performances

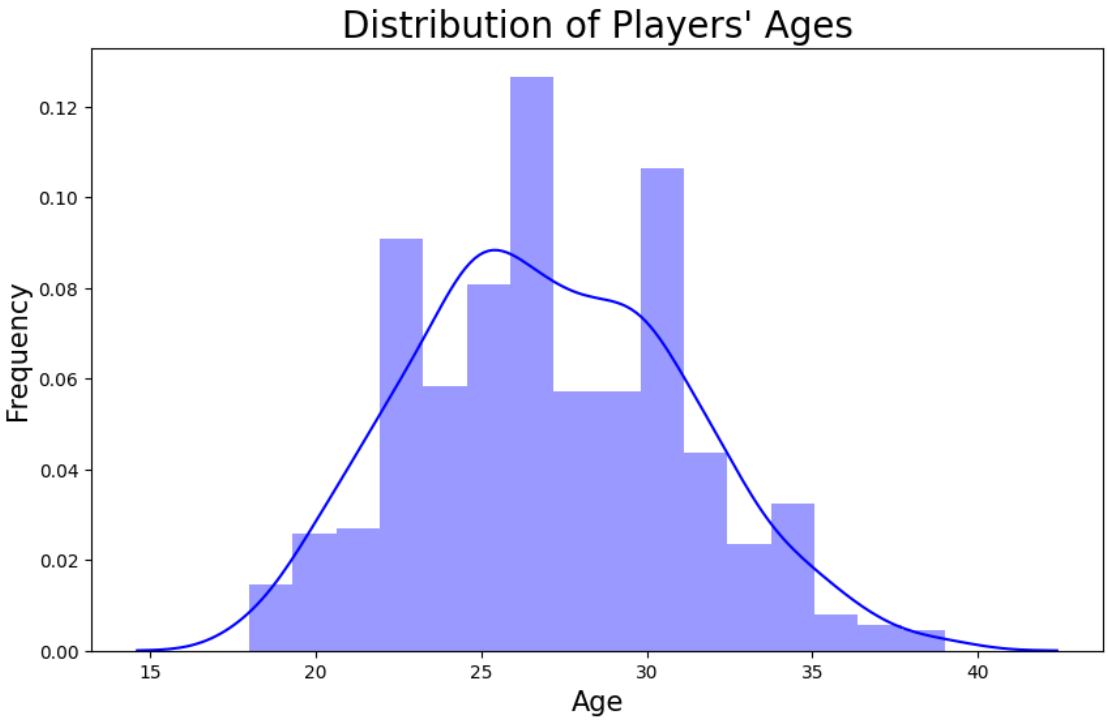
Some things we want to see with our players:

- Most represented clubs
- Distribution of positions
- Average age of players
- Top goal scorers
- Top assisters
- Top players under 21 goals + assists (g/a)



Al Sadd SC is an interesting outlier because all of these other clubs are very well known and highly regarded around the world, but Al Sadd SC does not fit the mold of the rest of these historic clubs. The reason behind the strong showing from Al Saad SC is because it is a club in Qatar (who is the host nation of this World Cup), and the majority of the players from the Qatar national team actually play for that club.


Distribution of Player Positions
DF
MF
FW
GK
33.5%
33.2%
6.0%
27.2%



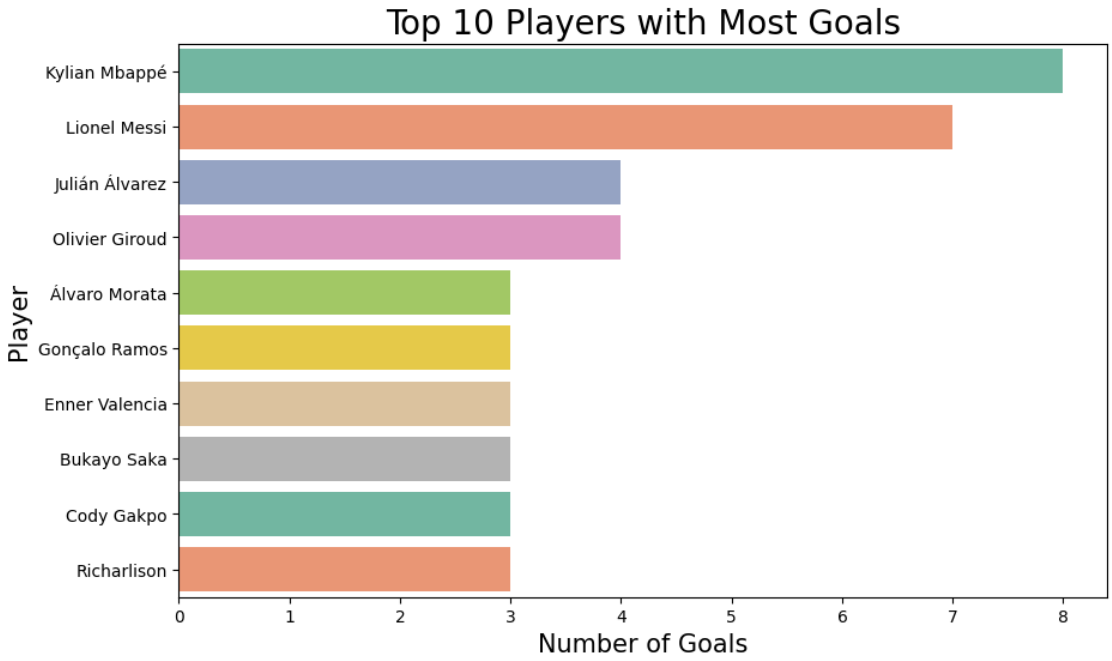

```
27.054411764705883
```

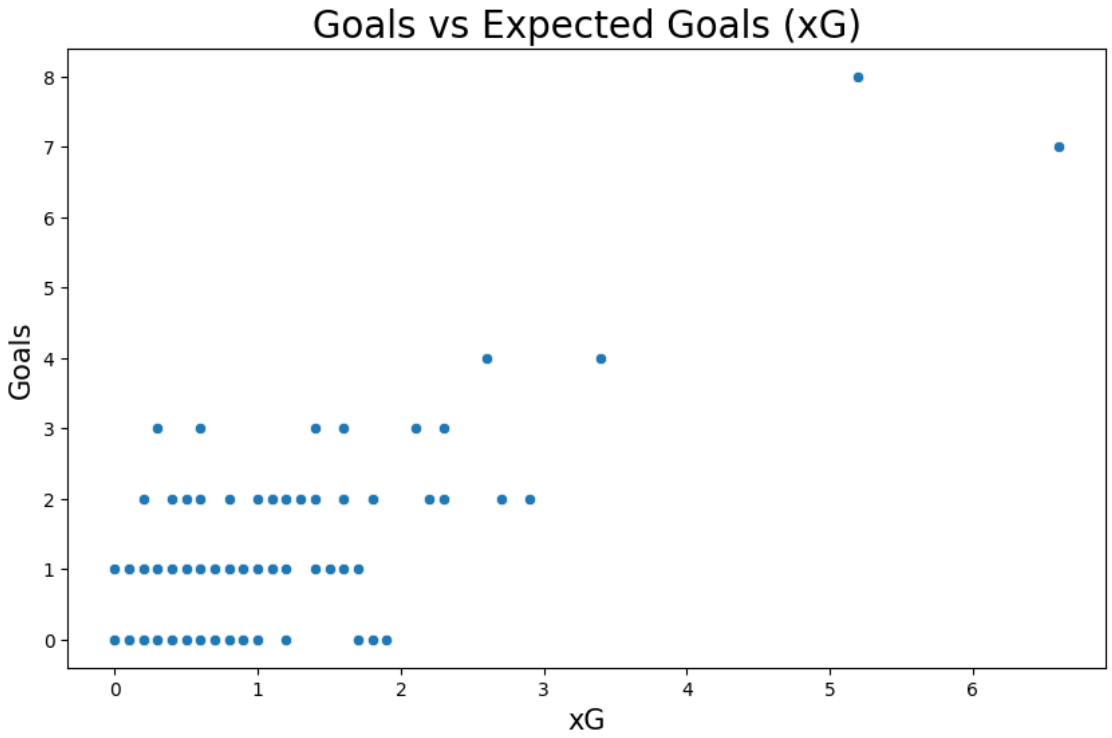
Based on the visual, and with our `.mean()` as further proof, we can see that there are a lot of players in the 27 yera old range, with other spikes at around the 30-32 range, and at the 21-23 range as well. This makes us interested, though, in who the oldest and youngest players were in the tournament.

```
              player  age        team position  goals
519              Pepe   39    Portugal       DF      1
142       Dani Alves    39      Brazil       DF      0
75   Atiba Hutchinson   39      Canada       MF      0
616      Thiago Silva   38      Brazil       DF      0
108        Bryan Ruiz   37  Costa Rica       MF      0
```

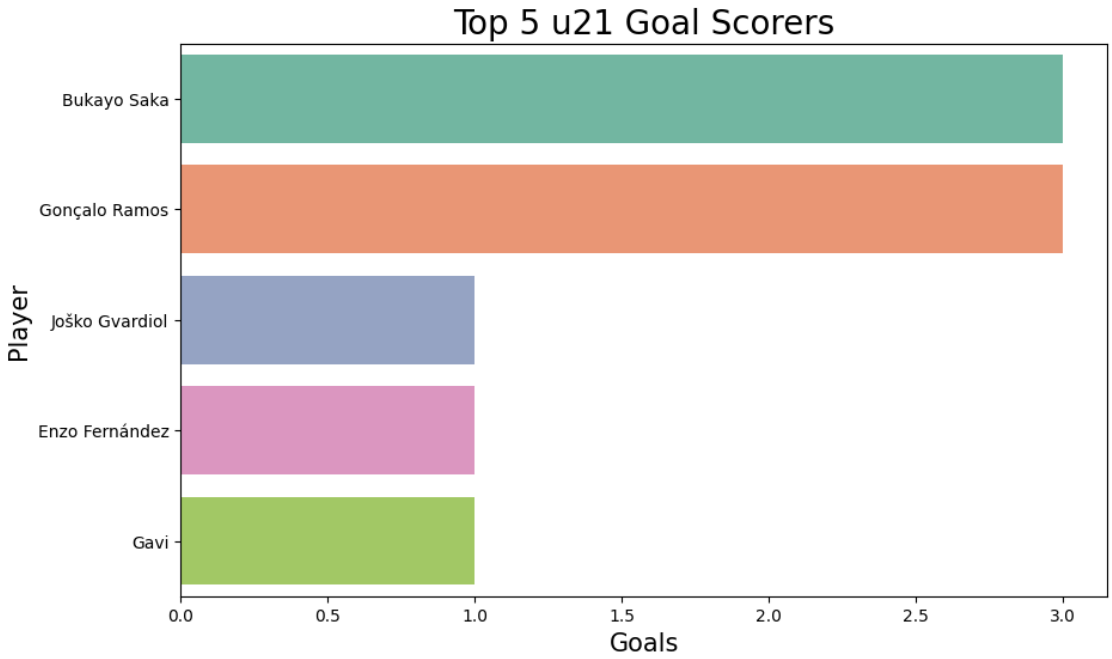
```
                   player  age       team position  goals
217                  Gavi   18      Spain       MF      1
215          Garang Kuol    18  Australia       FW      0
666      Youssoufa Moukoko   18    Germany       FW      0
94       Bilal El Khannous   18    Morocco       MF      0
7    Abdul Fatawu Issahaku   18      Ghana       FW      0
```

It is interesting that there are no players in the tournament over the age of 40, and that there were no players under the age of 18.

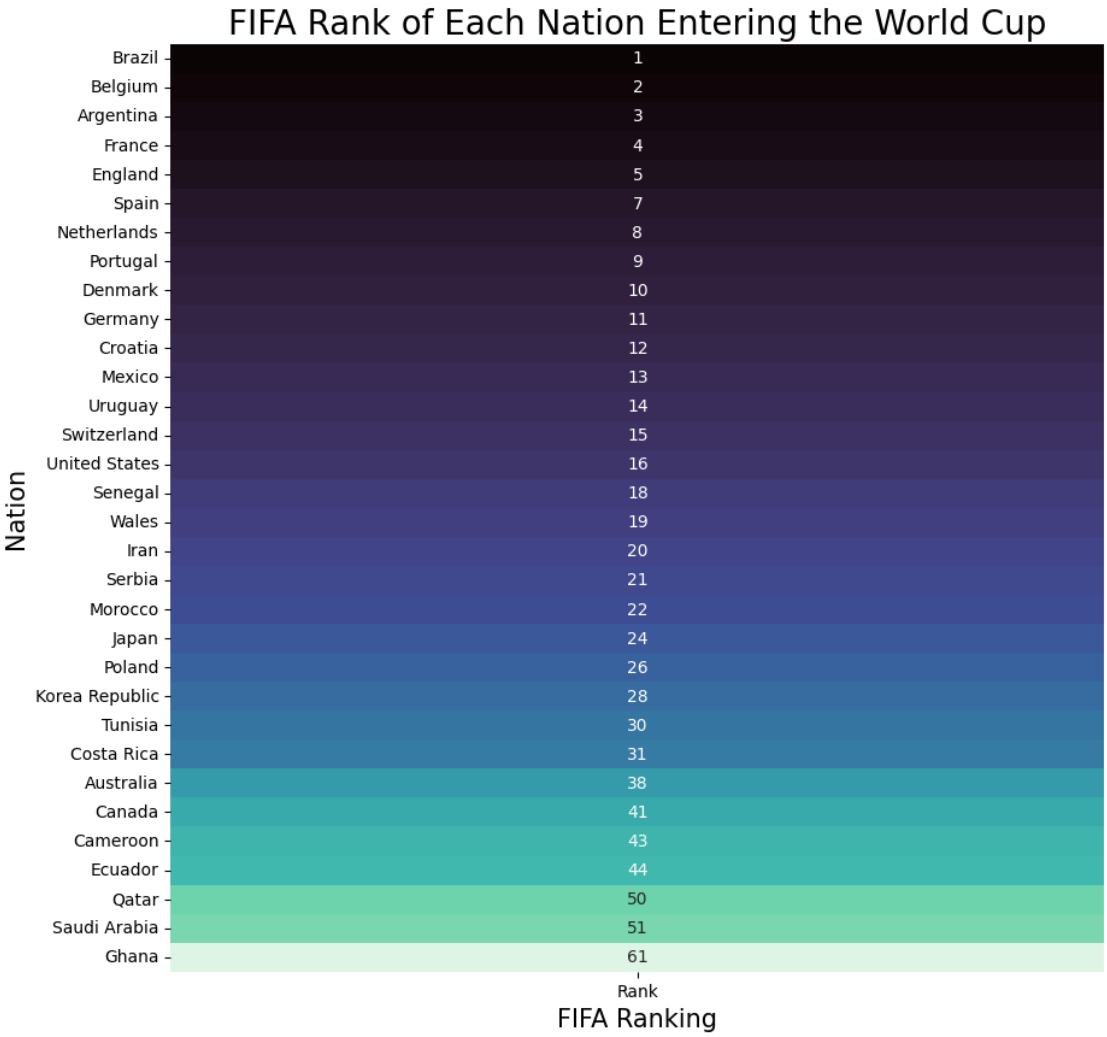


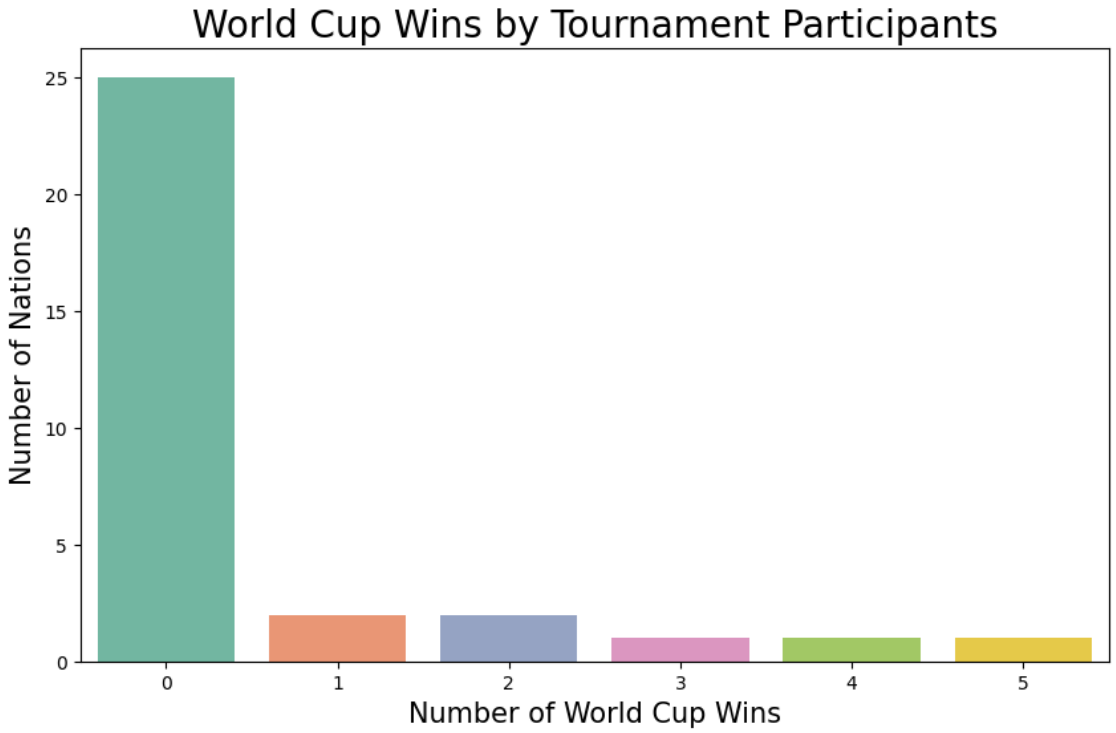



```
Text(0, 0.5, 'Player')
```


Top 5 u21 Goal Scorers
Bukayo Saka
Gonçalo Ramos
Joško Gvardiol
Enzo Fernández
Gavi
Player
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Goals

1.4 Visualizing the Countries





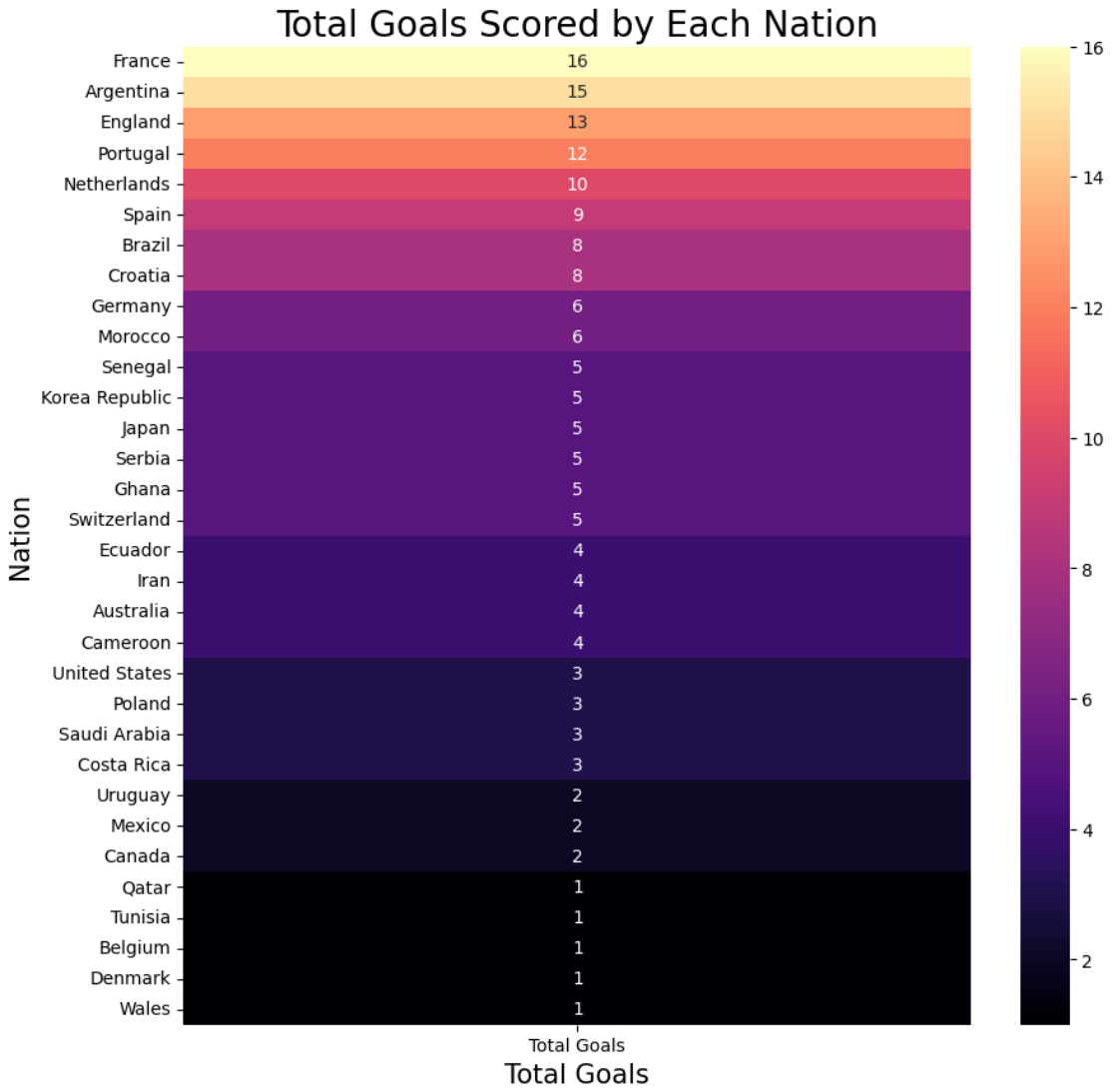
It is easy to see that most of the countries have never won a World Cup, and a few countries have won one, or two. There are three countries however that have won 3 or more, so let's see who they are.

```
           World Cup Wins
Nation
Brazil                  5
Germany                 4
Argentina               3

4
```

This is notable because the total number of World Cups played is 22 (counting this one), and we have 18 of the winners represented here. That would mean that 4 World Cup winners did not qualify for this World Cup, which is a very interesting statistic. The other 4 missing World Cups were won by Italy, who failed to qualify to this World Cup due to a surprise 1-0 upset loss in a World Cup qualifying match to North Macedonia.

1.5 Visualizing the Group Stage




```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
Cell In[28], line 3
      1 # create a heatmap using the Total Goals column
      2 plt.figure(figsize=(10, 10))
----> 3 sns.heatmap(countries[['Total Assists']].sort_values(by='Total Assists'
 ↪ascending=False), annot=True, cmap='magma')
      4 plt.title('Total Goals Scored by Each Nation', fontsize=20)
      5 plt.xlabel('Total Goals', fontsize=15)

File ~/opt/anaconda3/envs/master/lib/python3.9/site-packages/pandas/core/frame.
 ↪py:3811, in DataFrame.__getitem__(self, key)
```

```
   3809     if is_iterator(key):
   3810         key = list(key)
-> 3811     indexer = self.columns._get_indexer_strict(key, "columns")[1]
   3813 # take() does not accept boolean indexers
   3814 if getattr(indexer, "dtype", None) == bool:

File ~/opt/anaconda3/envs/master/lib/python3.9/site-packages/pandas/core/indexe /
 ↪base.py:6113, in Index._get_indexer_strict(self, key, axis_name)
   6110 else:
   6111     keyarr, indexer, new_indexer = self._reindex_non_unique(keyarr)
-> 6113 self._raise_if_missing(keyarr, indexer, axis_name)
   6115 keyarr = self.take(indexer)
   6116 if isinstance(key, Index):
   6117     # GH 42790 - Preserve name from an Index

File ~/opt/anaconda3/envs/master/lib/python3.9/site-packages/pandas/core/indexe /
 ↪base.py:6173, in Index._raise_if_missing(self, key, indexer, axis_name)
   6171     if use_interval_msg:
   6172         key = list(key)
-> 6173     raise KeyError(f"None of [{key}] are in the [{axis_name}]")
   6175 not_found = list(ensure_index(key)[missing_mask.nonzero()[0]].unique())
   6176 raise KeyError(f"{not_found} not in index")

KeyError: "None of [Index(['Total Assists'], dtype='object')] are in the␣
 ↪[columns]"
```

```
<Figure size 1000x1000 with 0 Axes>
```

It makes sense that France and Argentina were the two top scorers tournament wide, because they were the two finalists, and went all the way.

It appears that this particular visualization gives us many different groups in tiers in which they found goals in the group stages. England and Spain were king among them with 3 goals averaged per match (who were also both in the top 3 of possession as well), while the likes of Tunisia, Belgium, Qatar, Denmark, and Wales were among the lowest with under 0.5 goals per match. Let's look further at these two kings of the Group Stage to see what else they have in common.

We can see that there were a lot of yellow cards distributed during the group stage, but not as many red cards. Wales and Cameroon were the only nations to have a player get sent off during the group stages, and England was the only nation not to pick up a single card during the round. Saudia Arabia, followed by Serbia were by far the most carded nations.