DataMiningHW

Sharad Kumar Singh

2022-09-11

My Objective is to simulate and try to predict the outcomes of the 48 Group Stage Matches at this year’s Fifa World Cup in Qatar.

Football (Soccer) is one of the hardest sports to predict due to the free-flowing nature of the game, unlike other sports which are more of a Stop-Start nature. This fluidity is what makes it quite unique and also quite difficult to predict. Nonetheless, I shall attempt to run simulations and try to come up with Predictions.

The most common way to go about this is to use past results as an indicator of future performances. However as we shall later discover that this method although has some merit to it, doesn’t always hold true. No one would’ve predicted coming into the 2010 world cup that Italy the World Cup winners of 2006 would fail to get past the group stages. A similar fate has fallen upon other World cup winners like Spain and Germany. Hence we cannot use past results as an indicator for future outcomes.

Therefore we need a metric that can fully encompass the events in a match and also be indicative of the eventual outcome of the game. As we know Scorelines can often be deceptive. We’ve all witnessed the classic “smash and grab” phenomenon where one team keeps on piling up the pressure on the opponent’s goal only to be undone by the counterattack. The scoreline in those games isn’t indicative of the team’s performance.

Fortunately for us, there exists such a metric that allows us to capture most of the nuances in a game and be used as a fair indicator of the teams’ performance. It’s called xG(Expected Goals).

This metric is calculated for each attempt on goal and takes into account various factors like Distance from Goal, the angle from which the shot is taken, and other similar things. The purpose of xG is not to predict right before a shot gets taken whether it’ll be a goal or not. The way in which it is supposed to be used is to assess the quality of a chance and thereby the quality of the team’s performance.

R Markdown

Unfortunately for us, there is no dataset that exists that has the xG statistics for International Teams, especially in their recent competitive fixtures. As we know that form coming into the world cup is very important, hence we need xG values for the most recent games.

Not all hope is lost yet, as we can create the required dataset ourselves by scraping the values from this website called footystats. (Many websites these days do not allow us to scrape values from them however having gone through the disclaimer notes of this site in particular I was able to find that they are fine with the scraping as long as it’s not used for commercial purposes.)

So we start at a page where we have links to the individual team pages and from there we can iterate over all the teams and extract the xG statistics for their last 10 games.

We need to fix a few issues before moving ahead with the analysis for instance USA and South Korea don’t have proper names.

```
#df<-edit(df)
```

We also have a few missing values which need to be addressed before proceeding. Unfortunately the xG values for Australia, Wales and Costa Rica weren’t present in our source so we need to add them manually.

Let’s take a look at our scraped xG data:

```
print(df)
```

```
##       TeamName xGFor xGAgainst
## 1    Argentina   1.7      1.09
## 2       Brazil  2.27      1.03
## 3  Netherlands  2.22      1.02
## 4        Spain  1.62      0.68
## 5         Iran  1.65      1.16
## 6       Serbia  1.71         1
## 7        Japan  2.01      0.99
## 8      Germany  2.68      1.07
## 9       Canada  1.68      1.19
## 10     Croatia  2.24      1.12
## 11     Denmark  1.62      1.29
## 12     Senegal  1.98         1
## 13    Cameroon  1.84      1.12
## 14     Morocco  2.15      1.06
## 15    Portugal  1.71       0.9
## 16       South  1.69      1.05
## 17      France  2.23      0.95
## 18      United  2.09      1.06
## 19      Mexico  1.87      0.79
## 20     England  2.02      0.92
## 21     Ecuador  1.67      1.25
## 22      Poland  1.59      1.44
## 23       Qatar  1.48      1.28
## 24     Uruguay  1.69      1.25
## 25     Belgium  1.82      1.29
## 26 Switzerland  1.84      1.07
## 27     Tunisia  1.53      1.16
## 28       Saudi  1.23      1.28
## 29       Ghana   1.6      1.25
## 30   Australia  1.45      1.31
## 31       Wales  1.48      1.57
## 32  Costa Rica  1.13      1.85
```

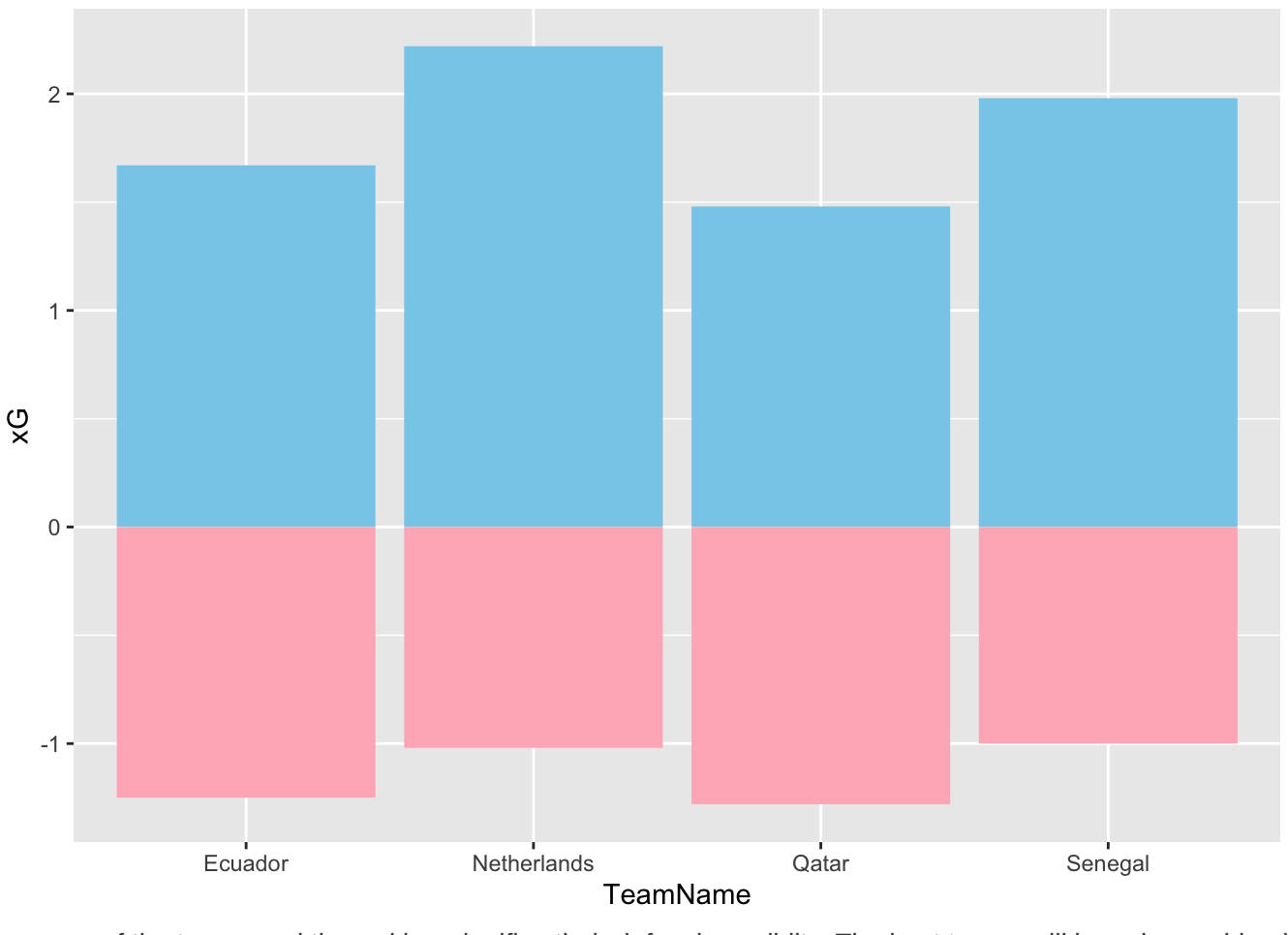
Now Let’s try to visualize some data

Some Pre-processing is required before Plotting

Let’s divide the teams according to their Groups

Let the Actual plotting begin:

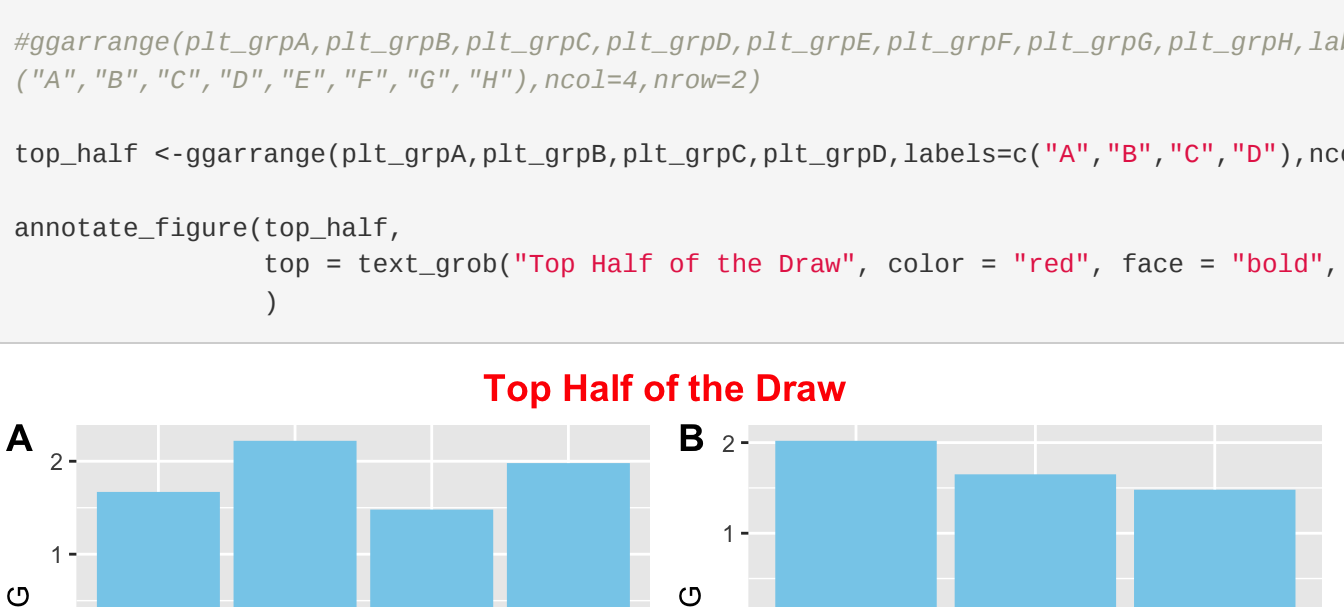
```
df_xg_vs_xgA_grpA <- df_xg_vs_xgA %>%
  filter(
    TeamName == "Qatar" |
      TeamName == "Ecuador" |
      TeamName == "Senegal" | TeamName == "Netherlands"
  )


plt_grpA<-ggplot(data = groupA, aes(x=TeamName, y=xG)) +
  geom_bar(
    data=df_xg_vs_xgA_grpA,
    aes(x=TeamName, y=xG_vs_xGA),
    fill=df_xg_vs_xgA_grpA$f_color,
    color= df_xg_vs_xgA_grpA$b_color,
    size=0,
    stat = "identity",
    show.legend = TRUE
    )
  # theme(plot.background = element_rect(fill = "slategrey"),
        # panel.background = element_rect(fill = "slategrey"))
plt_grpA
```



The Blue Bar indicates the attacking prowess of the teams and the red bar signifies their defensive solidity. The best teams will have longer blue bars and shorter red bars. As we can see from the above plot for Group A the standout Favorite to top the group is the Netherlands, which is in standing with their World Ranking. They seem to have the strongest attack and the strongest defense in the group. Qatar the tournament hosts are most likely to struggle as is visible by the diminutive nature of their blue bar. Netherlands and Senegal look to be the favorites to progress through this group as they stand head and shoulders above the rest.

lets do multiple plots to look at more than one groups at a time:

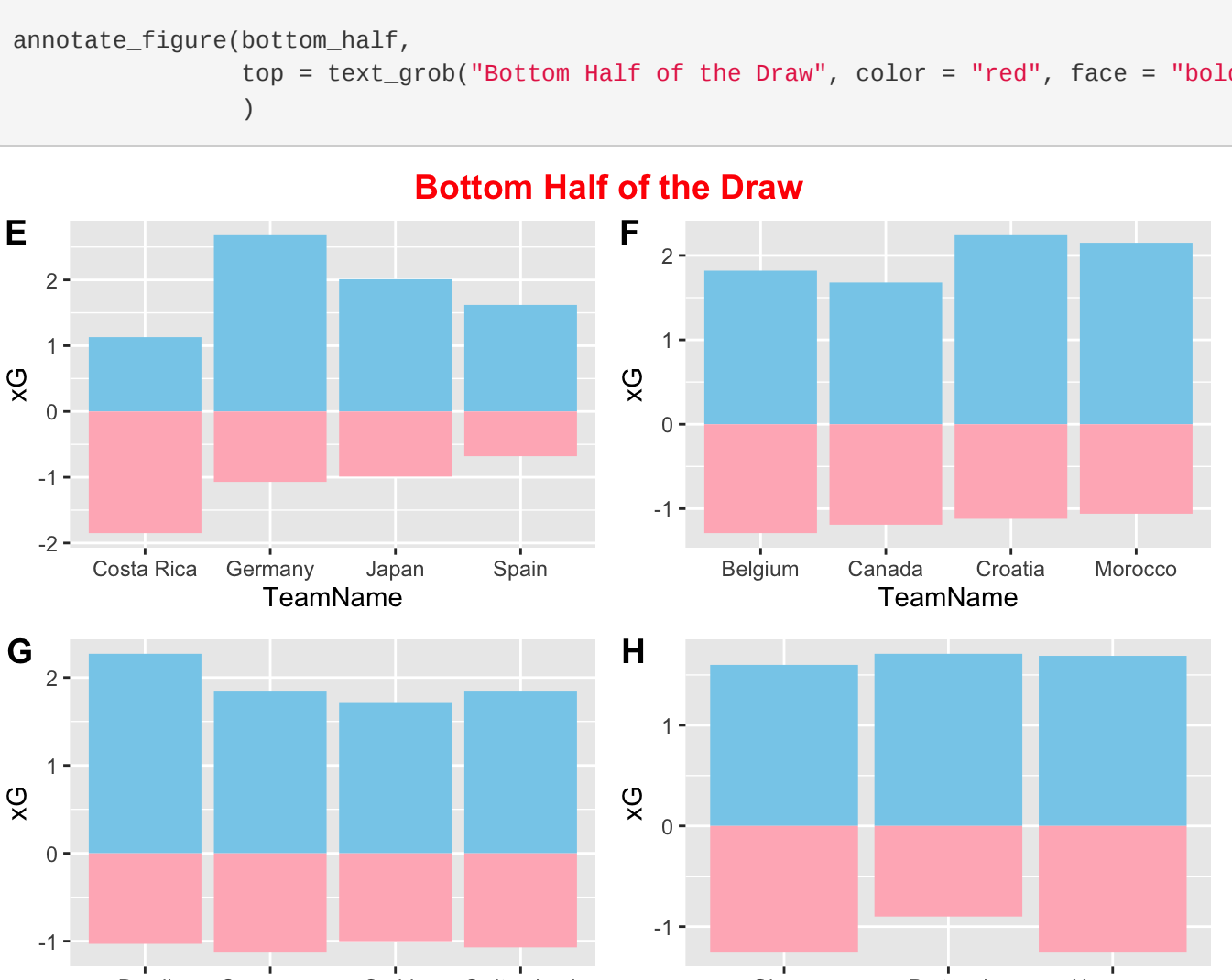
```
# Group B- England, IR Iran, USA and Wales####
groupB <- team_xGdiffxGA %>%
  filter(TeamName == "England" |
           TeamName == "Iran" | TeamName == "USMNT" | TeamName == "Wales")

df_xg_vs_xgA_grpB <- df_xg_vs_xgA %>%
  filter(TeamName == "England" |
           TeamName == "Iran" | TeamName == "USMNT" | TeamName == "Wales")


plt_grpB <- ggplot(data = groupB, aes(x = TeamName, y = xG)) +
  geom_bar(
    data = df_xg_vs_xgA_grpB,
    aes(x = TeamName, y = xG_vs_xGA),
    fill = df_xg_vs_xgA_grpB$f_color,
    color = df_xg_vs_xgA_grpB$b_color,
    size = 0,
    stat = "identity",
    show.legend = FALSE
  )

#Group C- Argentina, Saudi Arabia, Mexico, Poland ####
groupC <- team_xGdiffxGA %>%
  filter(TeamName == "Argentina" |
           TeamName == "Saudi" | TeamName == "Mexico" | TeamName == "Poland")

df_xg_vs_xgA_grpC <- df_xg_vs_xgA %>%
  filter(TeamName == "Argentina" |
           TeamName == "Saudi" | TeamName == "Mexico" | TeamName == "Poland")


plt_grpC <- ggplot(data = groupC, aes(x = TeamName, y = xG)) +
  geom_bar(
    data = df_xg_vs_xgA_grpC,
    aes(x = TeamName, y = xG_vs_xGA),
    fill = df_xg_vs_xgA_grpA$f_color,
    color = df_xg_vs_xgA_grpA$b_color,
    size = 0,
    stat = "identity",
    show.legend = FALSE
  )

#Group D- France, Australia, Denmark and Tunisia####
groupD <- team_xGdiffxGA %>%
  filter(TeamName == "France" |
           TeamName == "Australia" | TeamName == "Denmark" | TeamName == "Tunisia")

df_xg_vs_xgA_grpD <- df_xg_vs_xgA %>%
  filter(TeamName == "France" |
           TeamName == "Australia" | TeamName == "Denmark" | TeamName == "Tunisia")

plt_grpD <- ggplot(data = groupD, aes(x = TeamName, y = xG)) +
  geom_bar(
    data = df_xg_vs_xgA_grpD,
    aes(x = TeamName, y = xG_vs_xGA),
    fill = df_xg_vs_xgA_grpD$f_color,
    color = df_xg_vs_xgA_grpD$b_color,
    size = 0,
    stat = "identity",
    show.legend = FALSE
  )

#Group E- Spain, Costa Rica, Germany and Japan  ####
groupE <- team_xGdiffxGA %>%
  filter(TeamName == "Spain" |
           TeamName == "Costa Rica" | TeamName == "Germany" | TeamName == "Japan")

df_xg_vs_xgA_grpE <- df_xg_vs_xgA %>%
  filter(TeamName == "Spain" |
           TeamName == "Costa Rica" | TeamName == "Germany" | TeamName == "Japan")


plt_grpE <- ggplot(data = groupE, aes(x = TeamName, y = xG)) +
  geom_bar(
    data = df_xg_vs_xgA_grpE,
    aes(x = TeamName, y = xG_vs_xGA),
    fill = df_xg_vs_xgA_grpE$f_color,
    color = df_xg_vs_xgA_grpE$b_color,
    size = 0,
    stat = "identity",
    show.legend = FALSE
  )

#Group F- Belgium, Canada, Morocco and Croatia ####
groupF <- team_xGdiffxGA %>%
  filter(TeamName == "Belgium" |
           TeamName == "Canada" | TeamName == "Morocco" | TeamName == "Croatia")

df_xg_vs_xgA_grpF <- df_xg_vs_xgA %>%
  filter(TeamName == "Belgium" |
           TeamName == "Canada" | TeamName == "Morocco" | TeamName == "Croatia")


plt_grpF <- ggplot(data = groupF, aes(x = TeamName, y = xG)) +
  geom_bar(
    data = df_xg_vs_xgA_grpF,
    aes(x = TeamName, y = xG_vs_xGA),
    fill = df_xg_vs_xgA_grpF$f_color,
    color = df_xg_vs_xgA_grpF$b_color,
    size = 0,
    stat = "identity",
    show.legend = FALSE
  )


#Group G- Brazil, Serbia, Switzerland and Cameroon####
groupG <- team_xGdiffxGA %>%
  filter(TeamName == "Brazil" |
           TeamName == "Serbia" | TeamName == "Switzerland" | TeamName == "Cameroon")

df_xg_vs_xgA_grpG <- df_xg_vs_xgA %>%
  filter(TeamName == "Brazil" |
           TeamName == "Serbia" | TeamName == "Switzerland" | TeamName == "Cameroon")


plt_grpG <- ggplot(data = groupG, aes(x = TeamName, y = xG)) +
  geom_bar(
    data = df_xg_vs_xgA_grpG,
    aes(x = TeamName, y = xG_vs_xGA),
    fill = df_xg_vs_xgA_grpG$f_color,
    color = df_xg_vs_xgA_grpG$b_color,
    size = 0,
    stat = "identity",
    show.legend = FALSE
  )

#Group H- Portugal, Ghana, Uruguay and South Korea####
groupH <- team_xGdiffxGA %>%
  filter(TeamName == "Portugal" |
           TeamName == "Ghana" | TeamName == "Uruguay" | TeamName == "South Korea")

df_xg_vs_xgA_grpH <- df_xg_vs_xgA %>%
  filter(TeamName == "Portugal" |
           TeamName == "Ghana" | TeamName == "Uruguay" | TeamName == "South Korea")


plt_grpH <- ggplot(data = groupH, aes(x = TeamName, y = xG)) +
  geom_bar(
    data = df_xg_vs_xgA_grpH,
    aes(x = TeamName, y = xG_vs_xGA),
    fill = df_xg_vs_xgA_grpH$f_color,
    color = df_xg_vs_xgA_grpH$b_color,
    size = 0,
    stat = "identity",
    show.legend = FALSE
  )

#ggarrange(plt_grpA,plt_grpB,plt_grpC,plt_grpD,plt_grpE,plt_grpF,plt_grpG,plt_grpH,labels=c
("A","B","C","D","E","F","G","H"),ncol=4,nrow=2)

top_half <-ggarrange(plt_grpA,plt_grpB,plt_grpC,plt_grpD,labels=c("A","B","C","D"),ncol=2,nrow=2)

annotate_figure(top_half,
                top = text_grob("Top Half of the Draw", color = "red", face = "bold", size = 14),
                )
```



At a Glance, the favorites are quite literally jumping up to the eye. Mexico seems to be a strong favorite despite the presence of Argentina in that group. Their defensive solidity as indicated by their short red bar will hold them in good stead. Interestingly Poland is shown to have a leakier defense compared to Saudi Arabia although that could be attributed to the fact that Poland faces much tougher opponents in Europe compared to Saudi Arabia. In Group C as we can see England and USA seem the most likely teams to progress and Wales is tipped to struggle due to their defensive weaknesses. In Group D France rise above the rest and stand clear of the other teams and look poised to top the group. All other teams are well balanced and have equal chances of progressing through.

Let’s take a look at the Bottom Half of the Draw

```
bottom_half <-ggarrange(plt_grpE,plt_grpF,plt_grpG,plt_grpH,labels=c("E","F","G","H"),ncol=2,nrow=2)

annotate_figure(bottom_half,
                top = text_grob("Bottom Half of the Draw", color = "red", face = "bold", size = 14),
                )
```



Group H seems to be the most finely balanced of all the groups, although defensively Portugal seems to be quite solid compared to the rest. Every other Group seems to have two clear favorites rising above the rest. For Group E we can see That Spain’s defensive solidity should help them in progressing through to the knockout rounds. Germany also stands out due to their attacking prowess bolstered by stars like Serge Gnabry and Veterans like Thomas Mueller. Costa Rica’s defensive issues might flare up against their opponents and they seem to be in a tough spot and would find it very difficult to get even a point out of their fixtures. For Group G we can see that Brazil is the clear favorite to march on ahead. All the teams are finely poised from a defensive standpoint and it seems that Attack might after all be the best form of defense for the teams looking to progress further into the tournament. In group F we see Morocco to be quite solid at the back with the short red bar and also quite decent going forward. But this is where we might realize that there is more to this than meets the eye. Morroco being an African nation played most of their games against weaker opponents which could explain their inflated numbers compared to other teams like Belgium which is currently ranked number 2 in the world.

Hence we can see that we cannot solely rely on the xG statistic by itself to be an accurate representation. Therefore we might need to tweak it a little before proceeding with our simulations

It’s Prediction Time:

Let’s create the matchups first: For this we shall scrape the fixture list from SkySports official website:

```
page4<-"https://www.skysports.com/world-cup-fixtures"

sc_page3<-read_html(page4)
HomeTeams <- sc_page3 %>% html_nodes(".matches__participant--side1 .swap-text__target") %>% html_text() %>% as.ch
aracter()
AwayTeams <- sc_page3 %>% html_nodes(".matches__participant--side2 .swap-text__target") %>% html_text() %>% as.ch
aracter()

wc_df<- data.frame(HomeTeams,AwayTeams)
head(wc_df)
```

	HomeTeams <chr>	AwayTeams <chr>
1	Qatar	Ecuador
2	England	Iran
3	Senegal	Netherlands
4	United States of America	Wales
5	Argentina	Saudi Arabia
6	Denmark	Tunisia

6 rows

Time to clean up a little.

Let’s get started with the actual predictions:

Now we’ll be moving over to Python to continue with the analysis of the data we’ve accumulated so far.