FIFA WORLD CUP
Qatar2022

2022 Soccer World Cup Prediction

07.01.2023

—

Ancy Joseph

Introduction

Context:

Children First Soccer (CFS) Ltd. (non-profit that financially supports underprivileged children to enter into the professional world of soccer) wants to promote soccer in the developing markets across Asia (e.g. Laos, Cambodia, Myanmar etc.) leading up to the 2026 FIFA World Cup in USA, Canada and Mexico. CFS would like to understand which competitive teams should drive promotional campaigns within these markets in order to create excitement and passion for the game. CFS believes this will set up the foundation for children (both boys & girls), specifically in the age group of 6-15, to be coached, trained and presumably selected for the 2027 FIFA U-17 Boys World Cup in China & Girls World Cup in New Zealand respectively.

Criteria for Success:

CFS needs to understand the countries who have a likelihood to qualify into the Round of 16, Quarter Finals, Semi Finals and Finals specifically for the 2022 FIFA World Cup in Qatar in order to drive a 5 year strategy in the developing Asia markets leading up to the U-17 World Cups in 2027. The 5 year roadmap is segmented into 2 Phases:

> **Phase 1:** Focused on creating excitement in the markets and;
> **Phase 2:** Drive training programs across priority markets/ age groups.

This proposal is focused on *Phase 1 only.*

Scope of Solution:

Prediction will be based only on teams who have participated in FIFA Soccer World Cups till date and their performance in international matches.

Constraints:

Factors such as venue, host country weather, timing of the tournament, referee judgment, Video Assistant Referee (VAR) interventions, squad formation, in-game tactical switches, and player concentration and stamina all play a huge role in predicting the final outcome.

These elements are relatively new to sports science and unsure about how to apply them as influential statistical factors in an algorithm.

Stakeholders:

1. CFS CMO (Chief Marketing Officer)
2. CFS COO (Chief Operating Officer)
3. CFS APAC Regional Head
4. Ministry of Sports (APAC Developing Markets)

Key data Sources:

1. World Cups
2. 2022 World Cup Groups
3. 2022 World Cup Matches
4. World Cup Matches Stats
5. International Matches Stats
6. FIFA World Cup Ranking
7. FIFA International Matches

Data

Data Files:

All of the data utilized for this project is online user-entered and prone to discrepancies which made it a particularly challenging dataset to clean. The dataset contains 7 tables in CSV format:

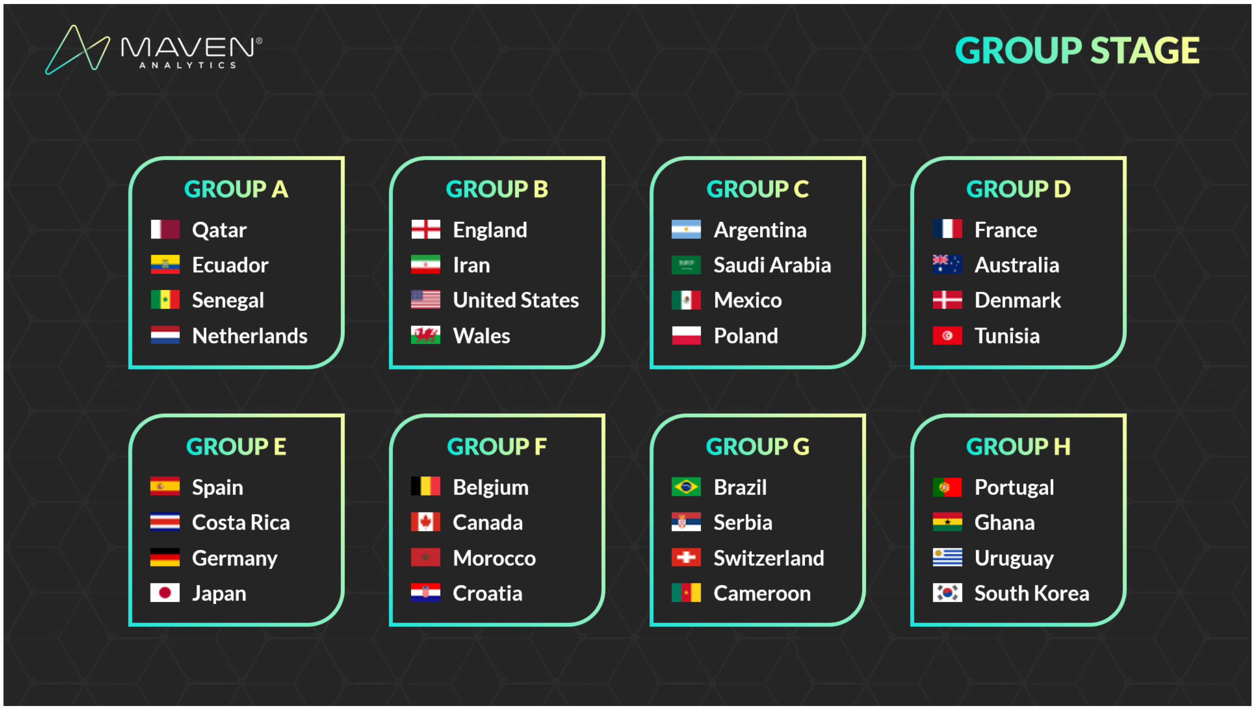
1. The `World_Cups` table outlines each World Cup in history, including the year, host country, and winner
2. The `2022_World_Cup_Groups` table includes all the qualified countries for this year's World Cup, the group they were drawn to, and their FIFA Ranking
3. The `2022_World_Cup_Matches` table contains the date and opponents for each of the 64 matches scheduled for the 2022 FIFA World Cup
4. The `World_Cup_Matches` table contains all the results from the previous editions of the World Cup
5. The `International_Matches` table contains all the results from every international match in history outside of the World Cup for each qualified country

6. The FIFA World Cup Ranking table contains ranking of all countries from 1992 to 2022
7. The FIFAallMatchBoxData contains information of international matches played between 2012 to 2017 with data on home & away team goals, shots on target, possession percentage etc.

In order to properly clean these tables, I needed to delete specific extraneous information that would not serve the analysis.

About the 2022 FIFA World Cup

Here's some context you may need into how the tournament plays out:

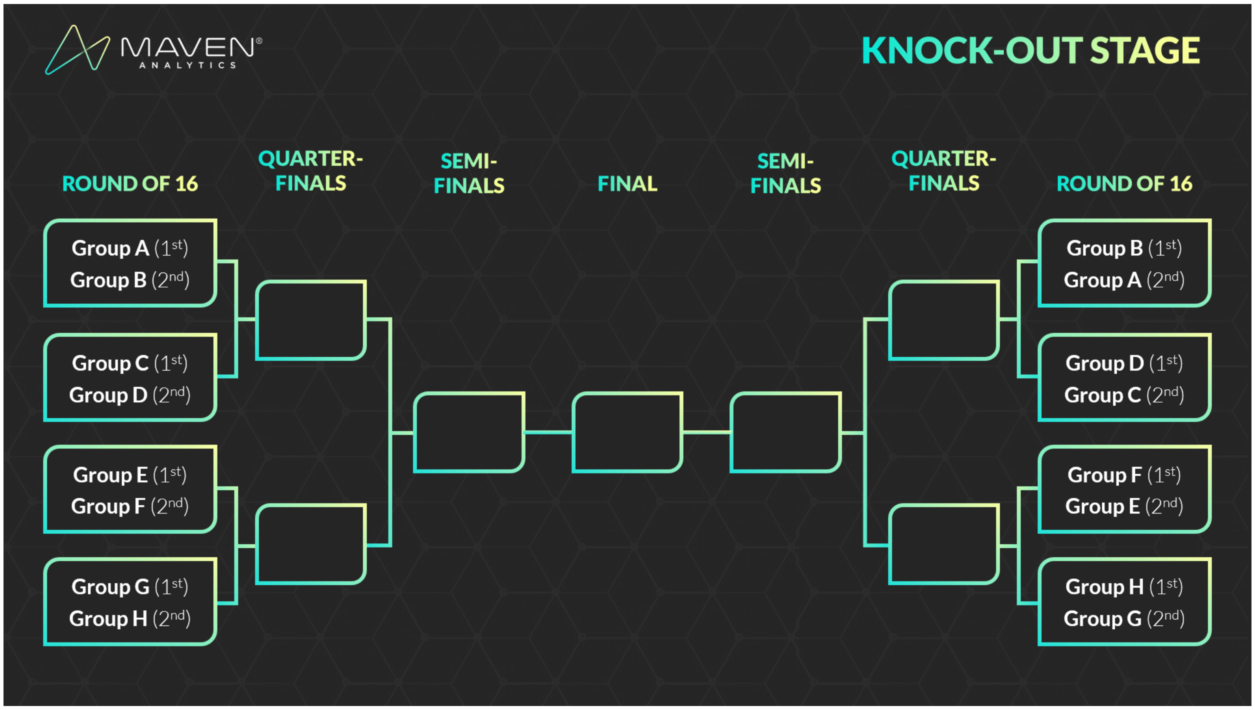

MAVEN ANALYTICS

GROUP STAGE

GROUP A	GROUP B	GROUP C	GROUP D
Qatar	England	Argentina	France
Ecuador	Iran	Saudi Arabia	Australia
Senegal	United States	Mexico	Denmark
Netherlands	Wales	Poland	Tunisia

GROUP E	GROUP F	GROUP G	GROUP H
Spain	Belgium	Brazil	Portugal
Costa Rica	Canada	Serbia	Ghana
Germany	Morocco	Switzerland	Uruguay
Japan	Croatia	Cameroon	South Korea

The 32 qualified countries are drawn into 8 groups with 4 teams each. For the initial group stage, each country plays the others in their group once (3 matches) and gets 3 points for a win, 1 for a draw, and 0 for a loss. Once the group stage ends, the two countries with the most points from each group advance to the round of 16.


MAVEN ANALYTICS
KNOCK-OUT STAGE
ROUND OF 16
QUARTER-FINALS
SEMI-FINALS
FINAL
SEMI-FINALS
QUARTER-FINALS
ROUND OF 16
Group A (1st) Group B (2nd)
Group C (1st) Group D (2nd)
Group E (1st) Group F (2nd)
Group G (1st) Group H (2nd)
Group B (1st) Group A (2nd)
Group D (1st) Group C (2nd)
Group F (1st) Group E (2nd)
Group H (1st) Group G (2nd)

The bracket for the round of 16, and subsequent stages, is determined by splitting up the winner from each group and facing them off against the runner-up from another group.

The team that wins each match advances to the next round until a winner is crowned!

Data Wrangling

All 7 datasets in scope were cleaned.

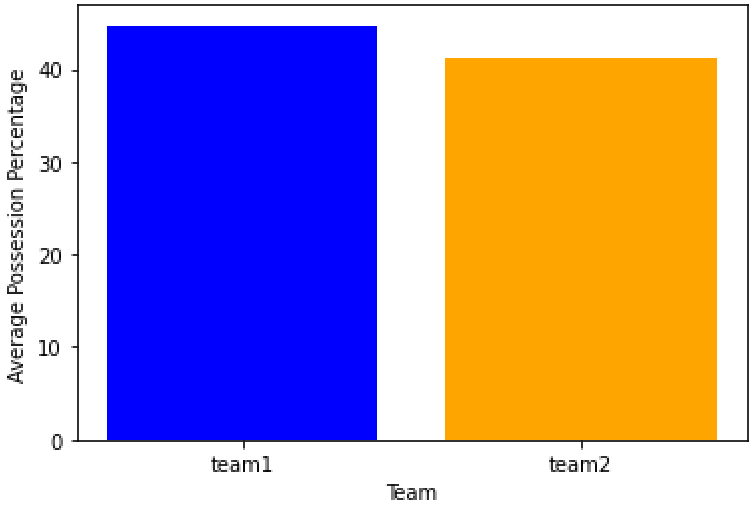
The data types across all 7 tables were consistent; however the "Win Condition" column in both `World Cup Matches` and `International Matches` tables had to be dropped due to null values and this was also considered to be a low priority feature to be analyzed.

The "*Winning Team*" & "*Losing Team*" columns are expected to have null values when the match is a draw.

Exploratory Data Analysis

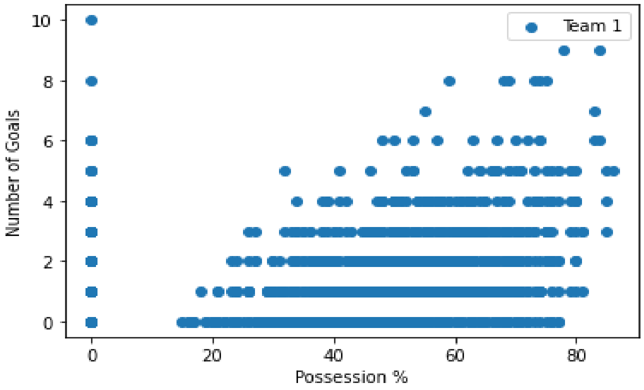
Analysis was focused on match results, player statistics, and overall team performance utilizing data from ***international matches*** played from 2012 to 2017. This was done utilizing the `FIFAallMatchBoxData` table. Data visualization techniques were utilized to explore patterns and trends in the data, and statistical methods to analyze and make inferences about the dataset.

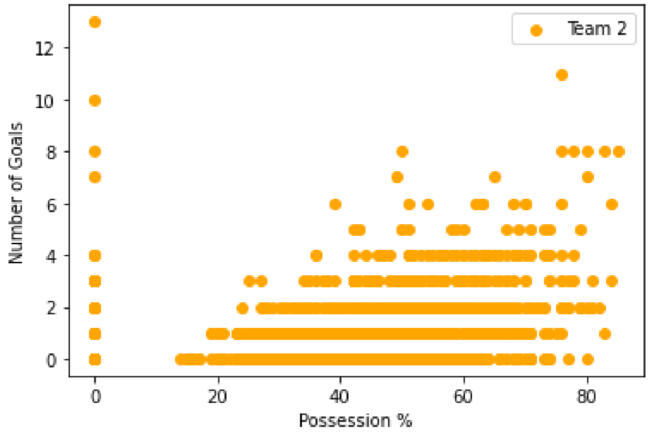
Possession Comparison:



On average, team1 has a better possession % than team2.

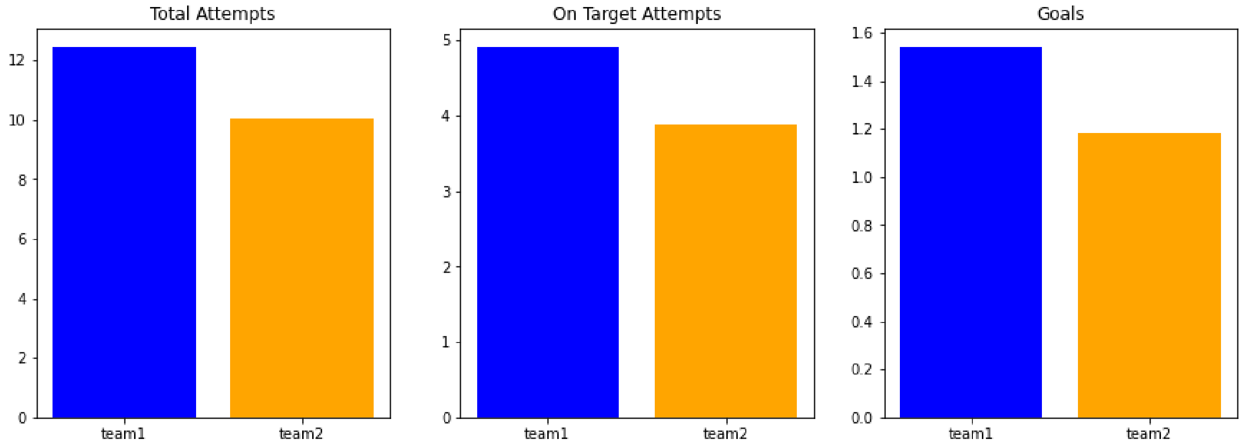
Possession versus Goals scored:





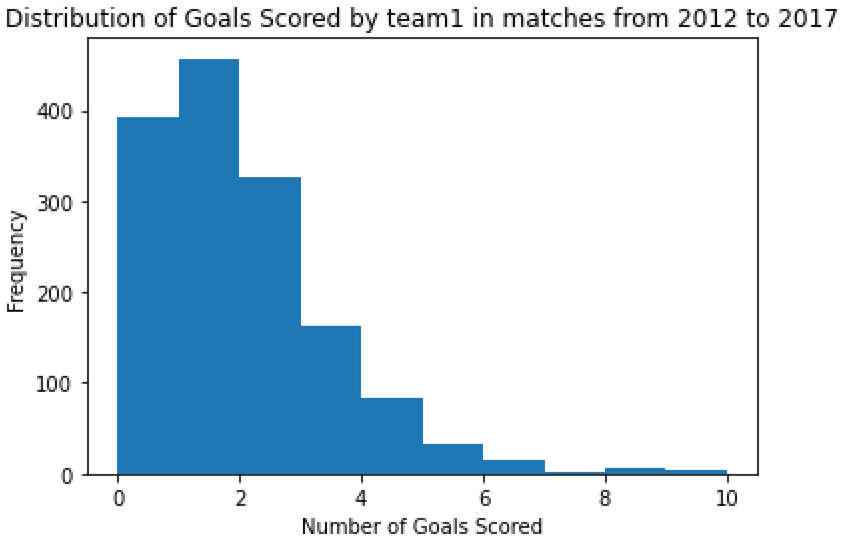
Both team1 and team2 scored more goals when they had more ball possession with team1 scoring more than team2.

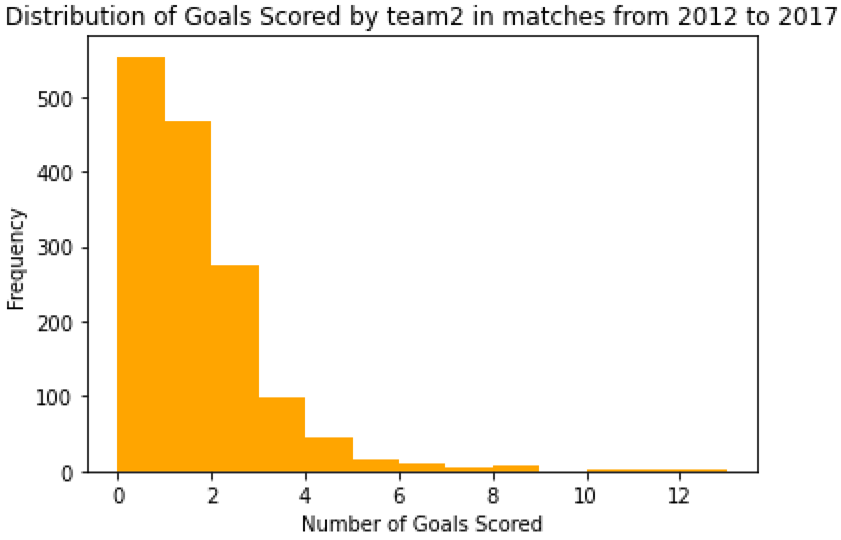
Comparison of Total Attempts, On Target Attempts, and Goals scored:



Team1 had better stats across Total Attempts, On Target Attempts as well as Goals Scored compared to team2.

Distribution of Goals:





Team2 had a better distribution of goals scored compared to team1.

CONCLUSION: It can be inferred that possession and attempts are crucial factors in determining the number of goals scored in a match, however, it is not the only factor, other factors like team strategies, player skills, and luck also play a role.

Featured Engineering & ML

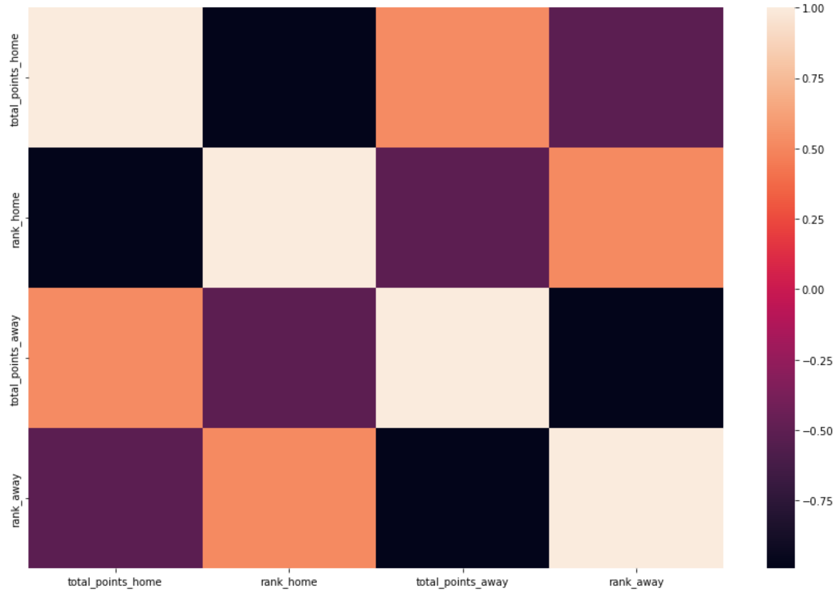
The objective was to prepare the data to apply feature engineering methods that will create the database in order to build Machine Learning algorithms to predict the winner of the 2022 Soccer World Cup.

Create Features:

The datasets were filtered to only utilize data from 2018 leading up to the 2022 World Cup to make it relevant. The idea here is to create possible features that have an impact on predicting football games. By intuition, we say that features that impact could be:

- Past game points made
- Past goals scored and suffered
- The importance of game (friendly or not)
- Rank of the teams
- Rank increment of the teams
- Goals made and suffered by ranking faced

So, the first thing to do is create a feature that says which team won and how many points they made in the game. The game points were assigned as 3 for win, 1 for draw and 0 for lose and which are different from the FIFA rank points that are already in the database. Also, it's supposed that FIFA Rank points and FIFA Ranking of the same team are negatively correlated, and we should use only one of them to create new features. This supposition is checked below:



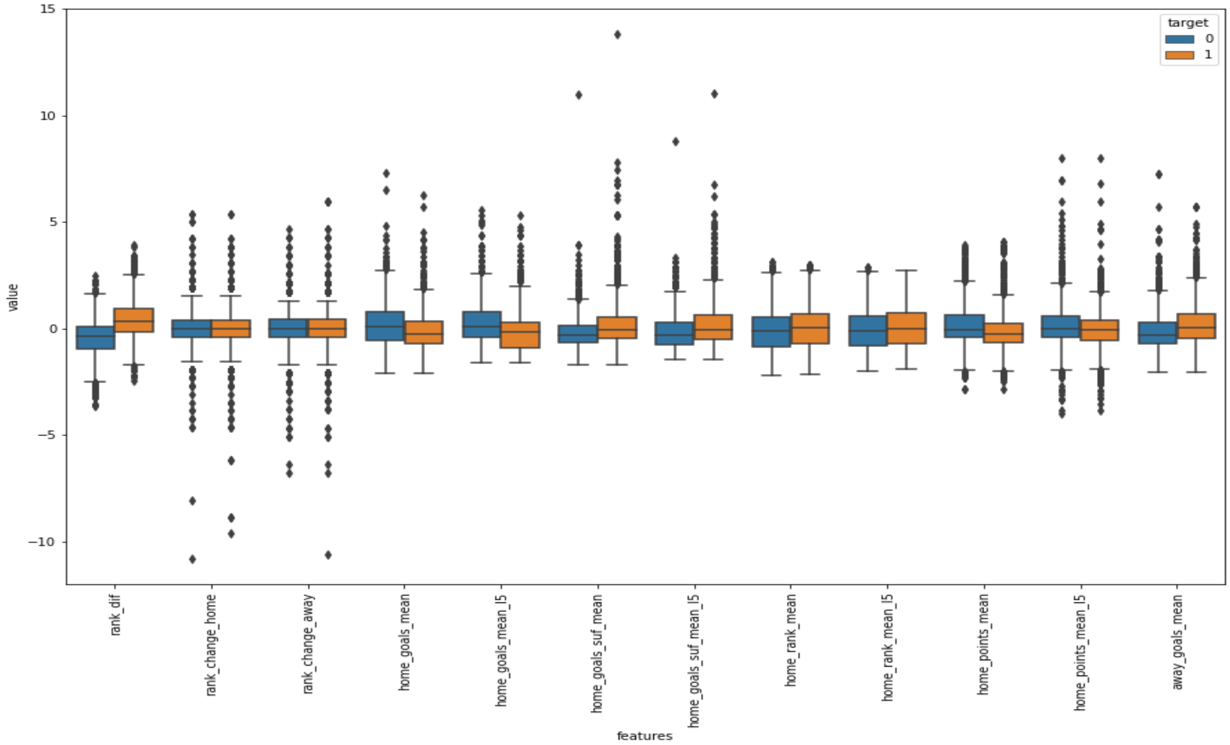
Feature Significance:

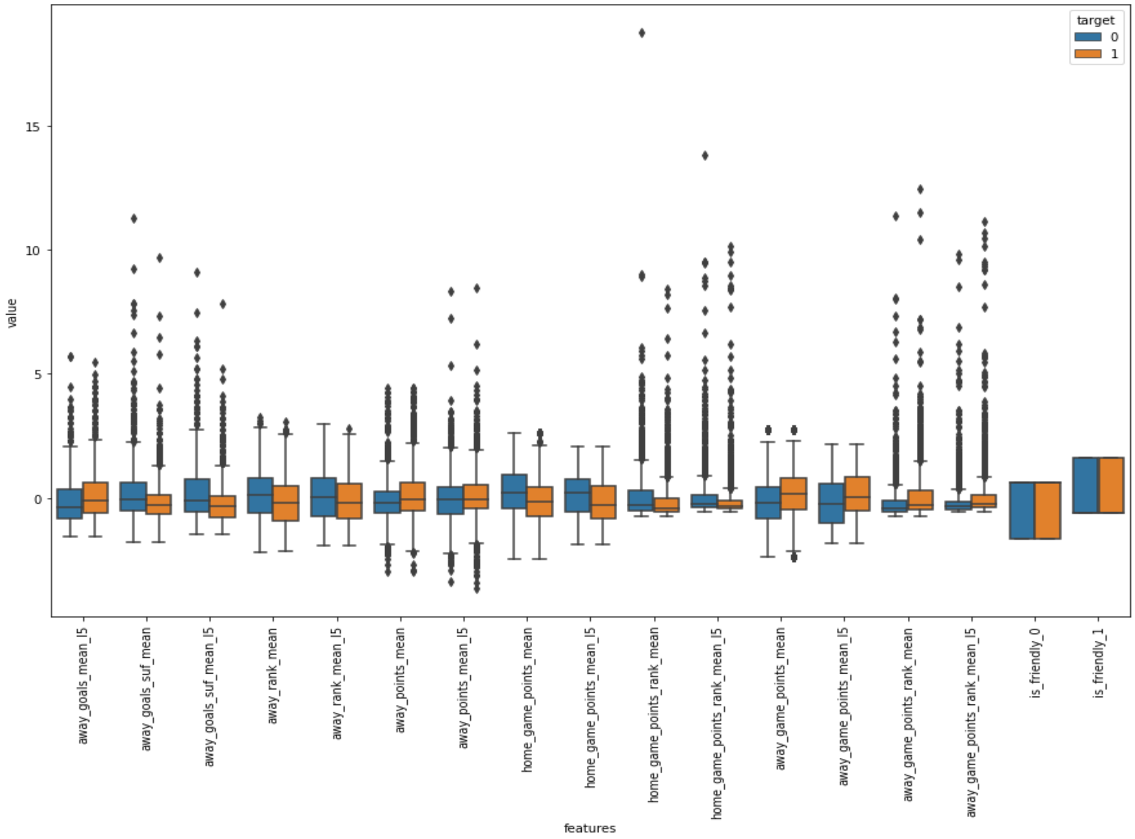
Data Scientists who have analyzed the game for many years have typically focused on 3 broad feature categories namely: Goals, Points & Rank. Within these categories, there are many different variables that have to be considered in order to identify further nuances on how each feature influences the outcome of a game. Some examples used in this study are:

1. **Goals** - # of goals scored by the team, scored against by other teams, scored in home games, scored in away games etc.
2. **Points** - # of points won in home games, # of points won in away games etc.
3. **Rank** - FIFA Ranking

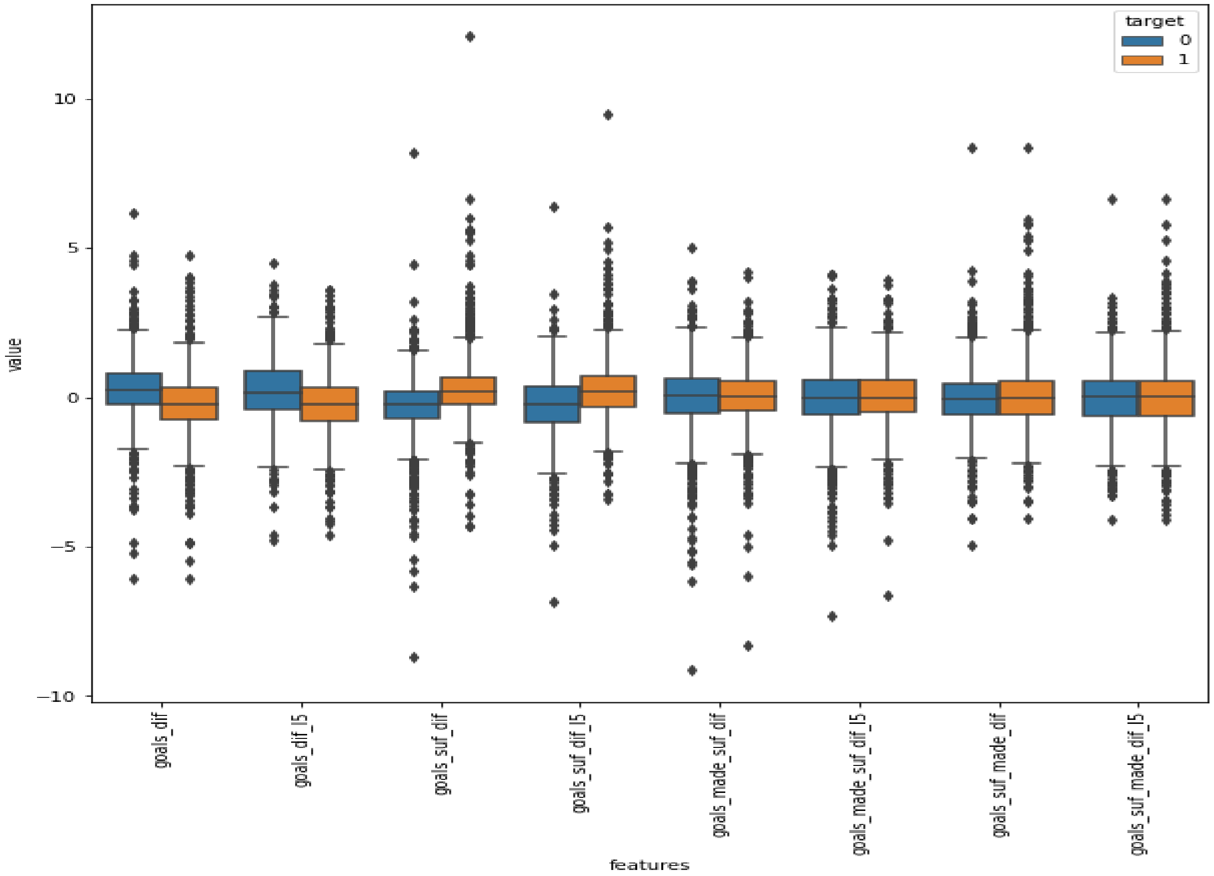
In the modern era, Data Scientists have access to even more granular data such as heat map of players during the game, effectiveness of formations etc. which I've not considered for this analysis as these datasets were not available in the public domain.

Feature Analysis:





Based on the boxplots, "*rank difference*" and "is_friendly" are the only good separators of data. But, we can create new features that differentiates between home and away teams and analyze if they are good at separating the data.

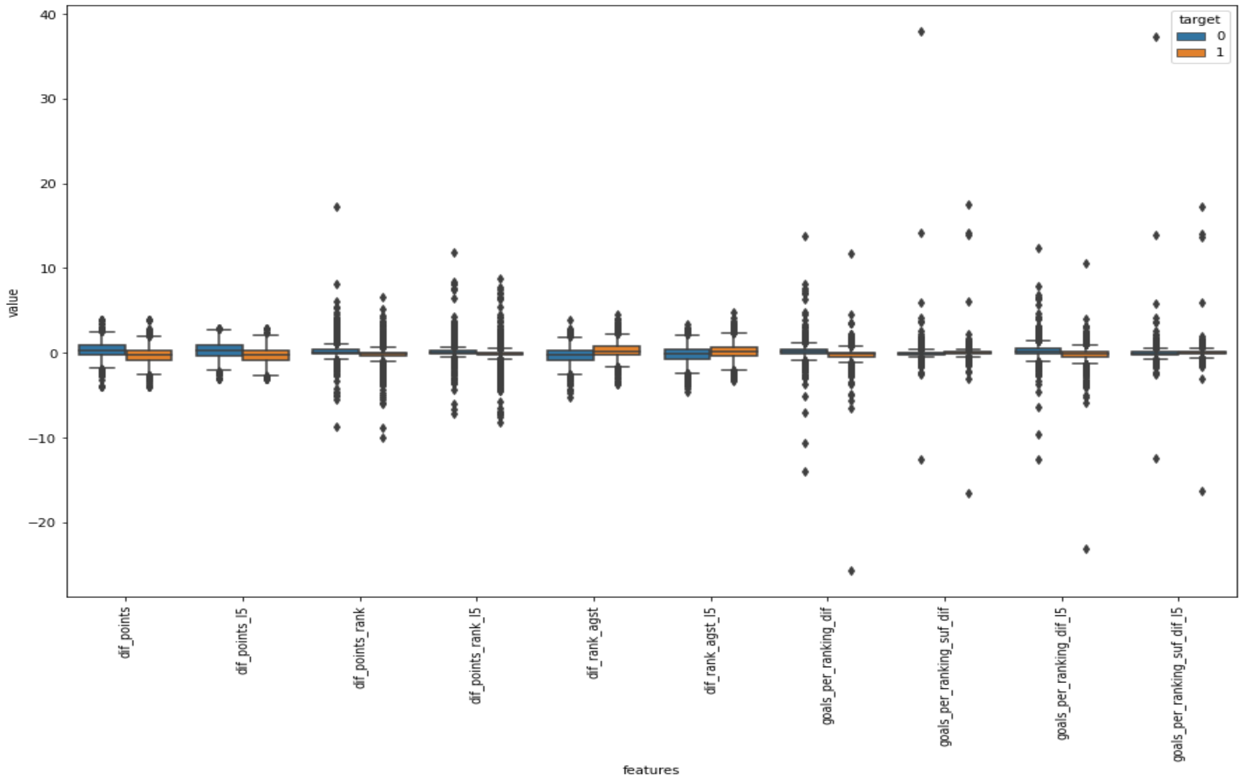


With that plot, we see that "*goal differences*" (full and last 5 games) and "*goals suffered*" (full and last 5 games) are good separators.

Now, we have 6 features:

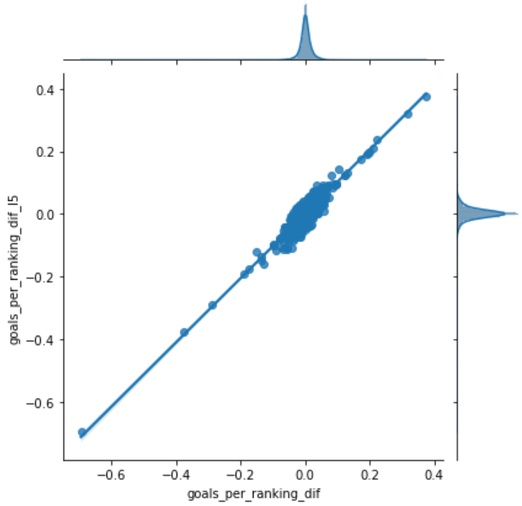
1. rank_dif
2. goals_dif
3. goals_dif_l5
4. goals_suf_dif
5. goals_suf_dif_l5
6. is_friendly

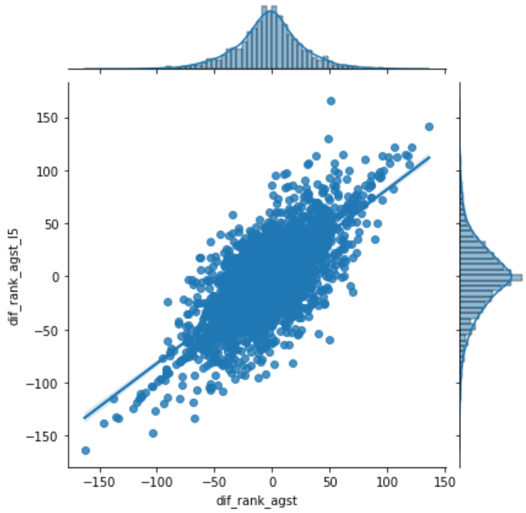
We can create other features, like differences of points made, differences of points made by rank faced and differences of rank faced.

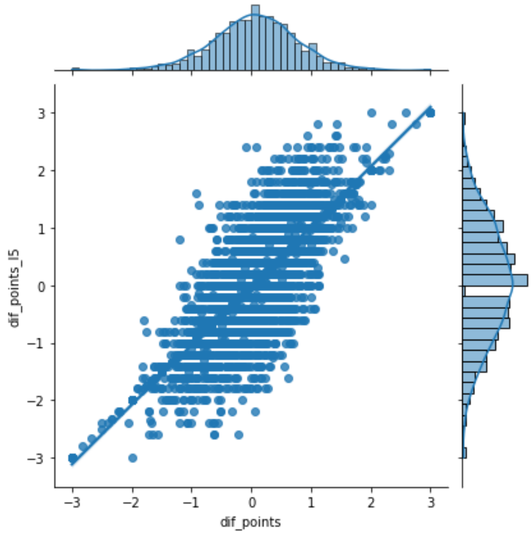


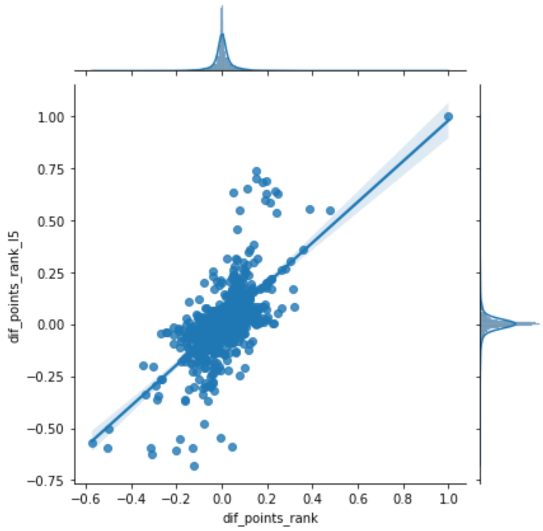
"*Difference of points*" (full and last 5 games), "*difference of points by ranking faced"* (full and last 5 games) and "*difference of rank faced"* (full and last 5 games) are good features.

Some of the generated features have very similar distributions which were analyzed using the following scatterplots:




150
100
50
0
−50
−100
−150
dif_rank_agst_l5
−150
−100
−50
0
50
100
150
dif_rank_agst


3
2
1
0
−1
−2
−3
dif_points_l5
−3
−2
−1
0
1
2
3
dif_points

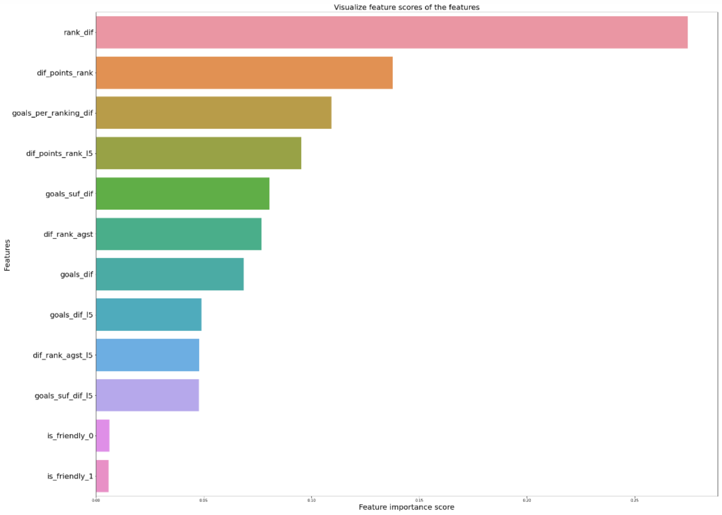

1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
dif_points_rank_l5
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
dif_points_rank

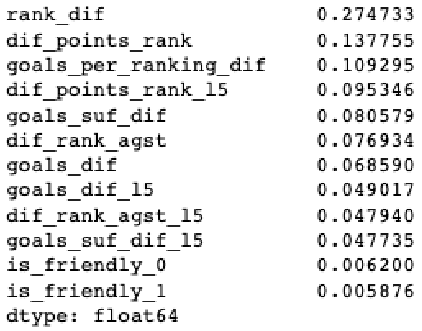
“*Goals difference by ranking faced*” and its last 5 games version has very similar distributions. So, we will use only the full version (goals_per_ranking_dif). For “*differences of rank faced*”, “*game points by rank faced*” and “*mean game points by rank faced*”, the two versions (full and 5 games) are not so similar. So, we decided to use both.

Final features that were selected were:

1. rank_dif
2. goals_dif
3. goals_dif_l5
4. goals_suf_dif
5. goals_suf_dif_l5
6. dif_rank_agst
7. dif_rank_agst_l5
8. goals_per_ranking_dif
9. dif_points_rank
10. dif_points_rank_l5
11. is_friendly_0
12. is_friendly_1

Feature Importance:





```
rank_dif                 0.274733
dif_points_rank          0.137755
goals_per_ranking_dif    0.109295
dif_points_rank_l5       0.095346
goals_suf_dif            0.080579
dif_rank_agst            0.076934
goals_dif                0.068590
goals_dif_l5             0.049017
dif_rank_agst_l5         0.047940
goals_suf_dif_l5         0.047735
is_friendly_0            0.006200
is_friendly_1            0.005876
dtype: float64
```

1. “*rank_dif*” is the most important feature and both “*is_friendly*” features are the lowest in importance.

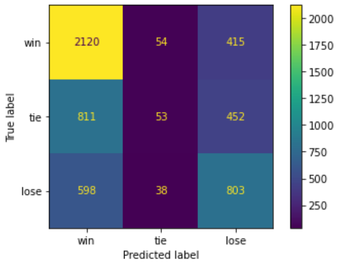
2. Since friendly games leading up to a world cup tournament is pretty significant from a match preparation and tournament readiness standpoint, both “*is_friendly*” features were included in the predictive model selection process.

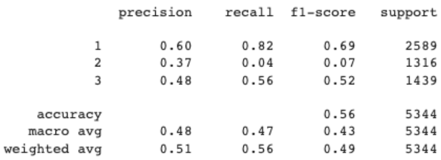
Machine Learning (Modeling Results):

Utilizing the 12 features selected, the objective of this segment is to identify & run multiple machine learning algorithms and select the most appropriate model to predict the winner of the 2022 Soccer World Cup.

3 of the following models were analyzed for the right fit:

1. Decision Tree



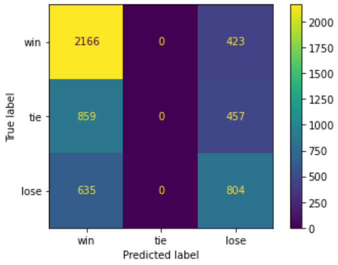


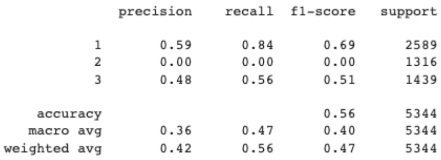
```
              precision    recall  f1-score   support

           1       0.60      0.82      0.69      2589
           2       0.37      0.04      0.07      1316
           3       0.48      0.56      0.52      1439

    accuracy                           0.56      5344
   macro avg       0.48      0.47      0.43      5344
weighted avg       0.51      0.56      0.49      5344
```

2. Logistic Regression



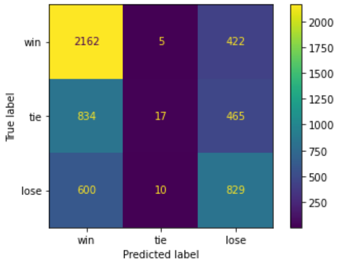


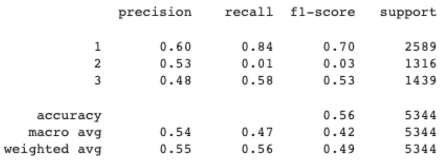
```
              precision    recall  f1-score   support

           1       0.59      0.84      0.69      2589
           2       0.00      0.00      0.00      1316
           3       0.48      0.56      0.51      1439

    accuracy                           0.56      5344
   macro avg       0.36      0.47      0.40      5344
weighted avg       0.42      0.56      0.47      5344
```

3. Random Forest





```
              precision    recall  f1-score   support

           1       0.60      0.84      0.70      2589
           2       0.53      0.01      0.03      1316
           3       0.48      0.58      0.53      1439

    accuracy                           0.56      5344
   macro avg       0.54      0.47      0.42      5344
weighted avg       0.55      0.56      0.49      5344
```

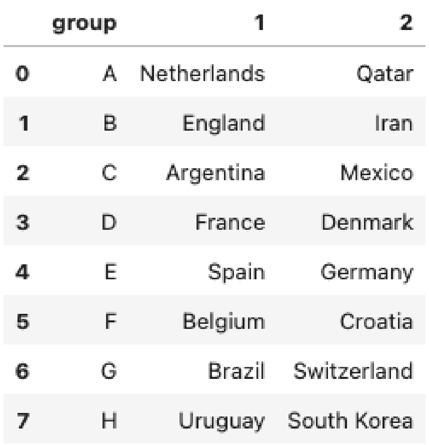
Random forest was selected as the most appropriate model.

Predictions

Random forest model was utilized to predict the winners across the tournament. The teams who qualified for the 2022 Soccer World Cup were identified utilizing the `2022 World Cup Groups` and `2022 World Cup Matches` datasets.

Group Stage:

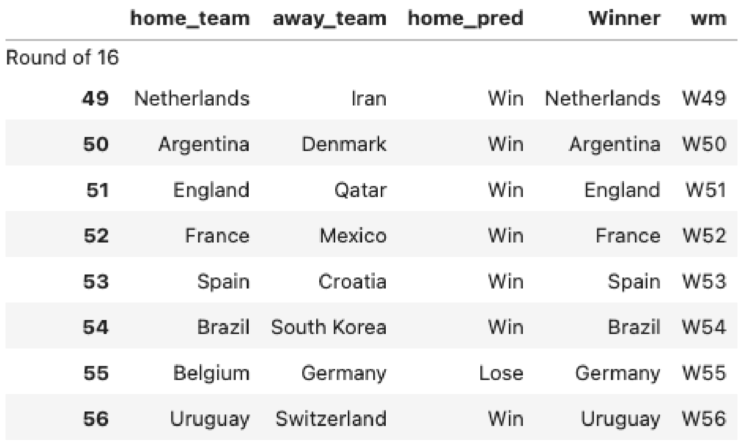
Following are the Group Stage winners:



	group	1	2
0	A	Netherlands	Qatar
1	B	England	Iran
2	C	Argentina	Mexico
3	D	France	Denmark
4	E	Spain	Germany
5	F	Belgium	Croatia
6	G	Brazil	Switzerland
7	H	Uruguay	South Korea

Round of 16:

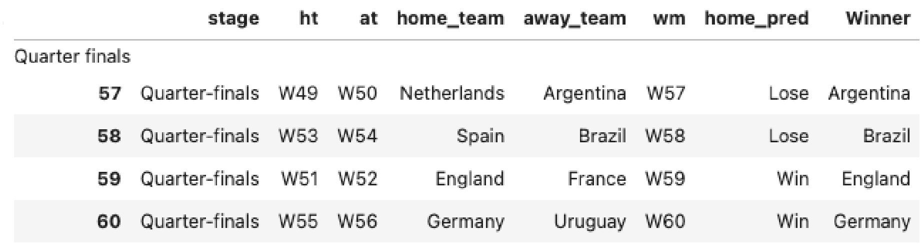
Following are the Round of 16 winners:



	home_team	away_team	home_pred	Winner	wm
Round of 16					
49	Netherlands	Iran	Win	Netherlands	W49
50	Argentina	Denmark	Win	Argentina	W50
51	England	Qatar	Win	England	W51
52	France	Mexico	Win	France	W52
53	Spain	Croatia	Win	Spain	W53
54	Brazil	South Korea	Win	Brazil	W54
55	Belgium	Germany	Lose	Germany	W55
56	Uruguay	Switzerland	Win	Uruguay	W56

Quarter Finals:

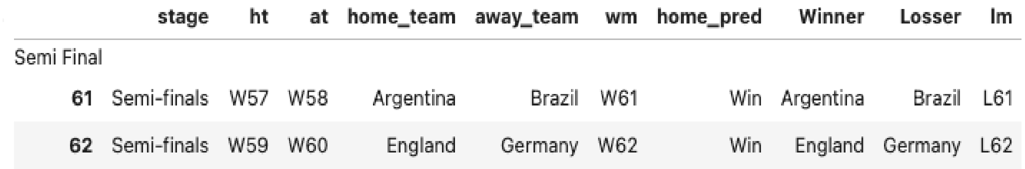
Following are the Quarter Final winners:



	stage	ht	at	home_team	away_team	wm	home_pred	Winner
Quarter finals								
57	Quarter-finals	W49	W50	Netherlands	Argentina	W57	Lose	Argentina
58	Quarter-finals	W53	W54	Spain	Brazil	W58	Lose	Brazil
59	Quarter-finals	W51	W52	England	France	W59	Win	England
60	Quarter-finals	W55	W56	Germany	Uruguay	W60	Win	Germany

Semi Finals:

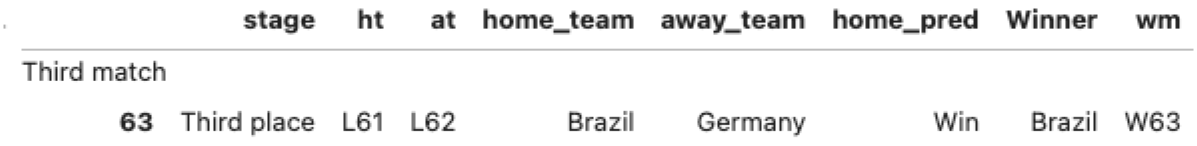
Following are the Semi Final winners:



	stage	ht	at	home_team	away_team	wm	home_pred	Winner	Losser	lm
Semi Final										
61	Semi-finals	W57	W58	Argentina	Brazil	W61	Win	Argentina	Brazil	L61
62	Semi-finals	W59	W60	England	Germany	W62	Win	England	Germany	L62

Third Place Match:

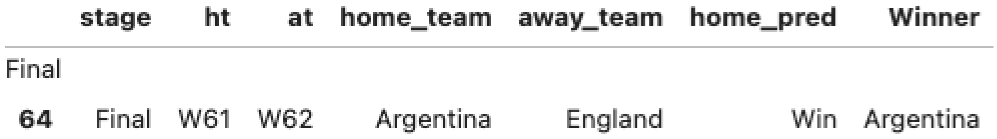
Brazil is the Third Place match winner!



	stage	ht	at	home_team	away_team	home_pred	Winner	wm
Third match								
63	Third place	L61	L62	Brazil	Germany	Win	Brazil	W63

Finals:

Argentina is the winner of the 2022 Soccer World Cup!



	stage	ht	at	home_team	away_team	home_pred	Winner
Final							
64	Final	W61	W62	Argentina	England	Win	Argentina

Summary & Recommendations

As per the predictive analysis completed and to meet the Phase 1 requirements of the proposal, the recommendation is for CFS Ltd. to focus their promotional campaigns on the **Top 14 countries** that were predicted to reach the Round of 16, with a special emphasis on South Korea as they were the only Asian country predicted to make it to the next round. This should also be supported by highlighting marquee players like Lionel Messi, Cristiano

Ronaldo as well as young and upcoming talent like Kylian Mbappe, Cody Gakpo etc. and their corresponding league and international impact (statistics) to create additional excitement and fan following.

Additional Considerations:

Analysis specifically focused on Asian datasets e.g. AFC (Asian Football Confederation) Cup stats, Asian players stats from European leagues etc. will definitely boost the participation rate in the focused markets however these datasets were not available in the public domain for analysis & deeper insights.

Credits

Thanks to **Daniel Wu** for his relentless support and thought leadership all throughout the project as my Springboard mentor, **Kenneth Gil-Pasquel** for troubleshooting and resolving my GitHub, Jupyter notebook & Google collab queries and **DJ Sarkar** for quick and timely responses to subject matter related doubts.