FIFA: *World Cup Data Analysis*

Prepared by:
Robert H. | Antoine M. | Ryan T. and Zain M.

Table of Contents

Executive Summary

The focus of this project is to analyze the dataset ‘fifa_world_cup.csv’. This .csv formatted data file contains a complete overview of all international fútbol matches played since the 90’s. On top of this, the strength of each team is explained through the incorporation of actual FIFA rankings as well as player strengths based on EA Sports’ FIFA video game. In order to properly analyze the dataset, we decided to formally use Python as our primary programming language for this project. The use of Python gave us the opportunities to grasp a better comprehension of the data as well as the ‘story’ hidden inside the frame.

The aim of our analysis revolved around a group of 5 tasks or questions which ultimately gave us the insight to predict the biggest question of this tournament; That is who will win the 2022 World Cup? Overall, the data allowed us to discover various insights to further grasp a more focused perception of the 2022 FIFA World Cup.

Introduction + Pre-processing

The FIFA World Cup is the most prestigious association fútbol tournament in the world. It is also the most viewed and followed sporting event in the world with roughly over 800 million viewers tuning in globally. To better enhance one’s perspective of the World Cup, the Super Bowl is the most watched sporting event in the United States with only 200 million viewers nationwide. The World Cup is played once every 4 years with a total of 32 countries competing every tournament. Only 1 of the 32 countries within the tournament are granted the gift to host such an exciting worldwide event.

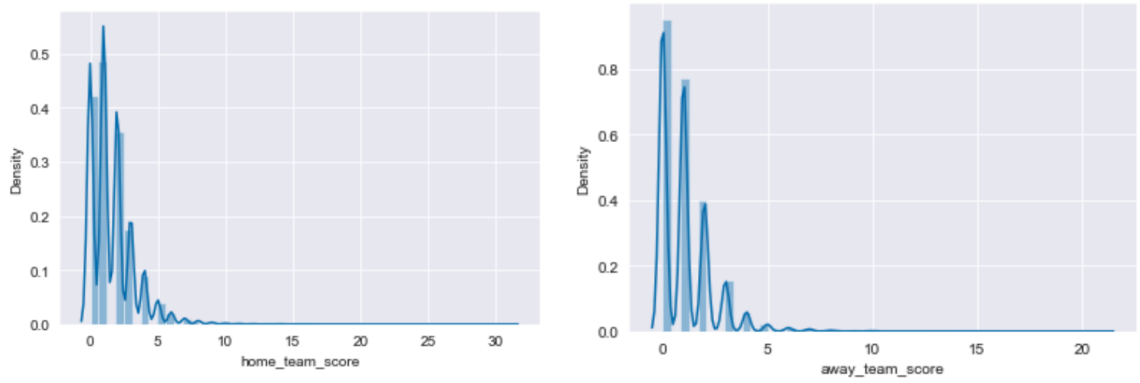
- **Data + N/A Report:**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23921 entries, 0 to 23920
Data columns (total 25 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   date                            23921 non-null  object
 1   home_team                       23921 non-null  object
 2   away_team                       23921 non-null  object
 3   home_team_continent             23921 non-null  object
 4   away_team_continent             23921 non-null  object
 5   home_team_fifa_rank             23921 non-null  int64
 6   away_team_fifa_rank             23921 non-null  int64
 7   home_team_total_fifa_points     23921 non-null  int64
 8   away_team_total_fifa_points     23921 non-null  int64
 9   home_team_score                 23921 non-null  int64
 10  away_team_score                 23921 non-null  int64
 11  tournament                      23921 non-null  object
 12  city                            23921 non-null  object
 13  country                         23921 non-null  object
 14  neutral_location                23921 non-null  bool
 15  shoot_out                       23921 non-null  object
 16  home_team_result                23921 non-null  object
 17  home_team_goalkeeper_score      8379 non-null   float64
 18  away_team_goalkeeper_score      8095 non-null   float64
 19  home_team_mean_defense_score    7787 non-null   float64
 20  home_team_mean_offense_score    8510 non-null   float64
 21  home_team_mean_midfield_score   8162 non-null   float64
 22  away_team_mean_defense_score    7564 non-null   float64
 23  away_team_mean_offense_score    8312 non-null   float64
 24  away_team_mean_midfield_score   7979 non-null   float64
dtypes: bool(1), float64(8), int64(6), object(10)
memory usage: 4.4+ MB
```

```
home_team_goalkeeper_score
15542

away_team_goalkeeper_score
15826

home_team_mean_defense_score
16134

home_team_mean_offense_score
15411

home_team_mean_midfield_score
15759

away_team_mean_defense_score
16357

away_team_mean_offense_score
15609

away_team_mean_midfield_score
15942
```

The CSV formatted data file for this project contains 25 columns of data with nearly 24,000 rows. Of the 25 columns: 10 are listed as objects, 8 are floats, 6 are integers, and 1 is a boolean (True or False). The dataset also contains numerous null values (NA), all of which resort from variables that take FIFA rankings into account. This is significant as it helps us acknowledge what values are missing and what columns need to be cleaned or replaced

- **Distributions:**



From the histogram above, we can check the distributions for columns to identify any

outliers that may coexist in our dataframe.

- **Dropping:**

```
['FIFA World Cup qualification' 'Friendly'
 'African Cup of Nations qualification' 'Amílcar Cabral Cup'
 'CFU Caribbean Cup qualification'
 'United Arab Emirates Friendship Tournament'
 'Malta International Tournament' 'Lunar New Year Cup'
 'African Cup of Nations' 'CFU Caribbean Cup' 'UEFA Euro qualification'
 'Kirin Cup' 'FIFA World Cup' 'Oceania Nations Cup qualification'
 'Baltic Cup' 'Gulf Cup' 'Simba Tournament' 'CECAFA Cup'
 'Confederations Cup' 'Dynasty Cup' "King's Cup" 'Nehru Cup' 'SAFF Cup'
 'Copa Paz del Chaco' 'Korea Cup' 'USA Cup' 'Copa América'
 'Merdeka Tournament' 'South Pacific Games' 'UNCAF Cup'
 'Oceania Nations Cup' 'Windward Islands Tournament' 'Gold Cup'
 'AFC Asian Cup qualification' 'UEFA Euro' 'AFF Championship'
 'AFC Asian Cup' 'King Hassan II Tournament'
 'Cyprus International Tournament' 'Dunhill Cup'
 'COSAFA Cup qualification' 'COSAFA Cup' 'Tournoi de France'
 'Gold Cup qualification' 'SKN Football Festival' 'Arab Cup qualification'
 'Arab Cup' 'UNIFFAC Cup' 'Nordic Championship' 'WAFF Championship'
 'Millennium Cup' 'Cup of Ancient Civilizations' "Prime Minister's Cup"
 'EAFF Championship' 'TIFOCO Tournament' 'Afro-Asian Games'
 'AFC Challenge Cup' 'Copa del Pacífico' 'AFC Challenge Cup qualification'
 'African Nations Championship' 'VFF Cup' 'Dragon Cup'
 'Nile Basin Tournament' 'Nations Cup' 'Copa Confraternidad'
 'Pacific Games' 'Superclásico de las Américas' 'ABCS Tournament'
 'Kirin Challenge Cup' 'OSN Cup' 'Copa América qualification'
 'Pacific Mini Games' 'Intercontinental Cup'
 'AFF Championship qualification' 'UEFA Nations League'
 'CONCACAF Nations League qualification'
 'African Nations Championship qualification' 'CONCACAF Nations League'
 'Three Nations Cup' 'Mahinda Rajapaksa Cup' 'Navruz Cup'
 'CONMEBOL–UEFA Cup of Champions']
```

The last step in our data cleansing process was to eliminate any non World Cup related data in our frame. From the image above, you can see from the ‘tournament’ column we had a lot of different types of data on games we deemed were unnecessary for our project analysis.

Task Overview

To attain a better understanding of the 2022 FIFA World Cup we wanted to ask ourselves specific questions. All of these questions revolve solely in the environment of the World Cup. Our analysis and findings lead us to answer these questions in which we split into tasks:

1. Which countries’ have the strongest defense? Midfield? Offense?
2. Is there such a thing called ‘Home-Team Advantage’?
3. Country with the longest winning streak?
4. Which country has the most wins?

5. Is it better to have a strong defense, midfield, or offense in terms of winning the World Cup?

All these questions are significant to ultimately leading us to predict who would win this year's World Cup.

Task 1

- **Which countries' have the strongest defense? Midfield? Offense?**

The first task we wanted to compute for this project was to find the countries with the best defense, midfield, and offense. The metric that we used to quantify this statistic is the 'mean score' that is given to us in the dataset. This metric equates to the average FIFA game score of the 4 highest ranked players of the desired position of the field. The reason for selecting this is deemed through our understanding of the development at EA Sports'. The dominant gaming company contains the most accurate and the best source for "live" and up-to-date statistics on player performances with over 30 different grading factors that make-up a players' overall rating. As fans and players of the video game we firmly believe this is credible to use in our analysis.

away_team	away_team_mean_defense_score
Brazil	86.086364
Spain	85.427193
Italy	85.124742
Germany	84.671429
England	84.640426
France	84.132258
Argentina	83.353636
Portugal	82.896154
Belgium	81.153409
Netherlands	80.681250

home_team	home_team_mean_defense_score
Brazil	85.785185
Spain	85.585366
Italy	84.960902
England	84.733607
Germany	84.700694
France	84.047887
Portugal	82.887705
Argentina	82.646721
Belgium	81.304717
Netherlands	80.776423

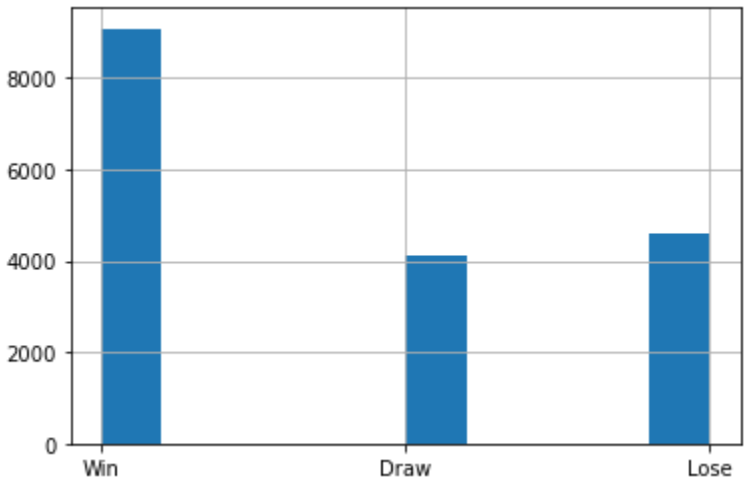
	overall defense average
Brazil	85.94
Spain	85.51
Italy	85.04
England	84.69
Germany	84.69
France	84.09
Argentina	83.00
Portugal	82.89
Belgium	81.23
Netherlands	80.73

In the above tables we can see the mean defense score is broken down into home and away. From here an overall average of both was conducted to conclude a top 10 of overall defense averages.

Task 2

- **Is there such a thing called 'Home-Team' Advantage?**

Identifying if there is an advantage for home teams was crucial to this analysis as the FIFA World Cup is played at a neutral site for 31 of the 32 teams that participate.For the second task we took the cleaned data and added two filters to reduce randomness. The first filter was removing the matches played on a neutral site because there is no actual home team in these matches. The second filter was to remove matches that ended in a shoot out because these are high pressured situations in which no team has an advantage. These added filters reduced the matches from 23,921 to 17,801 and the home team won 50.99% (9,077), lost 25.91% (4,614), and drew 23.11% (4,110).



The histogram above tells us that there is an advantage to playing at home as you are twice as likely to win than lose. To greater support our findings we looked at total home wins vs. away wins for teams:

Out[46]:

home_team	home_team_result
USA	289
Japan	207
Qatar	207
Saudi Arabia	203
France	189
Korea Republic	188
Oman	186
Thailand	186
United Arab Emirates	181
South Africa	180
Germany	173
England	167
Poland	164
Egypt	161
Tunisia	159
Estonia	159
Bahrain	158

Out[56]:

away_team	away team result
Zambia	176
Finland	175
Estonia	169
Sweden	162
Costa Rica	160
Paraguay	154
Norway	148
Uruguay	147
Lithuania	146
Jamaica	146
Syria	142
Germany	142
Brazil	140
Trinidad and Tobago	140
Turkey	139
Czech Republic	136

Looking at Team USA we can see that they have the most wins out of all teams but for total away wins they don't even crack the top 30. Having such a disparity in performance between home and away supports our findings that there is an advantage for the home team.

Task 3

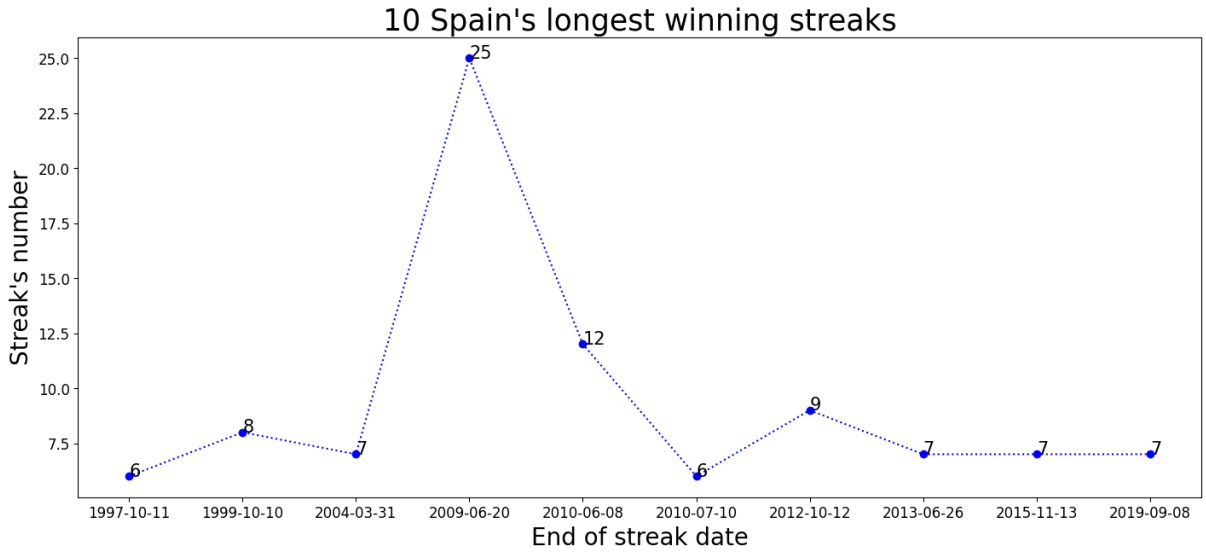
- **What country has the longest winning streak?**

Why is it important to know who has the longest winning streak? The main goal is to figure out if there is a correlation between having long winning streaks and winning important competitions. The analysis we computed on our dataset allows us to show the following table:

Teams	Longest hot streak	Last win of the streak
Spain	25	2009-06-20
Italy	15	2021-07-10
Brazil	14	1997-12-12
France	14	2004-02-18
Australia	13	1997-07-06
Morocco	12	2021-12-07
Belgium	12	2020-09-07
Saudi Arabia	11	2001-07-23
Sweden	11	2001-10-07
Russia	11	1996-03-27

From the created table, Spain has the longest winning streak with 25 games, the streak ended on the 20th of June 2009. We decided to focus on Spain's longest winning streak to understand what happens to this team when they play well regularly.

The following plot shows the 10 longest winning streaks of Spain over the last 25 years:



From 2004 to 2012, Spain had some long streaks without Lose. Knowing this, we can easily

make a link between this fact and their European and World titles. Indeed, they won the European Cup in 2008 and 2012, and also the World Cup in 2010. At this time, many journalists agree that they played probably the most beautiful soccer of all time with these incredible middlefield players : Xavi, Iniesta and Busquets.

To conclude this task, Spain has the longest winning streak thus far. We can also see that having a long winning streak and being regular allows teams to win the most important titles.

Task 4

- **Dropping:**

Task 5

- **Dropping:**

Task 6

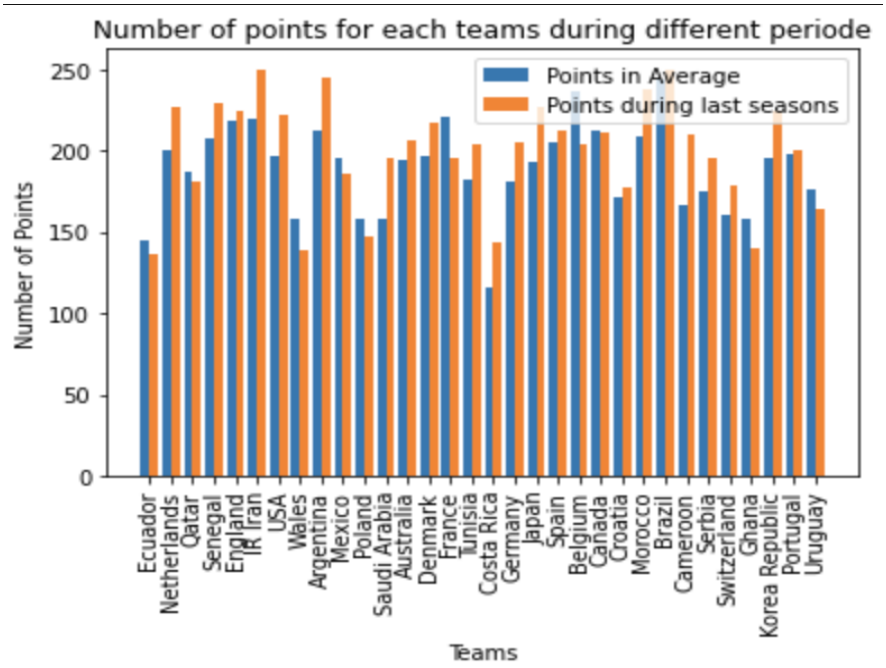
- **Who will win this World Cup ?**

One of the most important parts of this project is to find who will win this world cup. To answer this question, we have many steps.

Knowing if a team is playing good or bad this last season. This step is interesting because a strong team can have some weak moments and cannot win every time. To find if a team is in a hot or cold season, we'll compare the number of points gained this season with the number of points gained this last five seasons (+3 points for a win, +1 point for draw and +0 for loose). Why last five years ? Because if it's more, it's unlikely that it's the same team and it makes sense

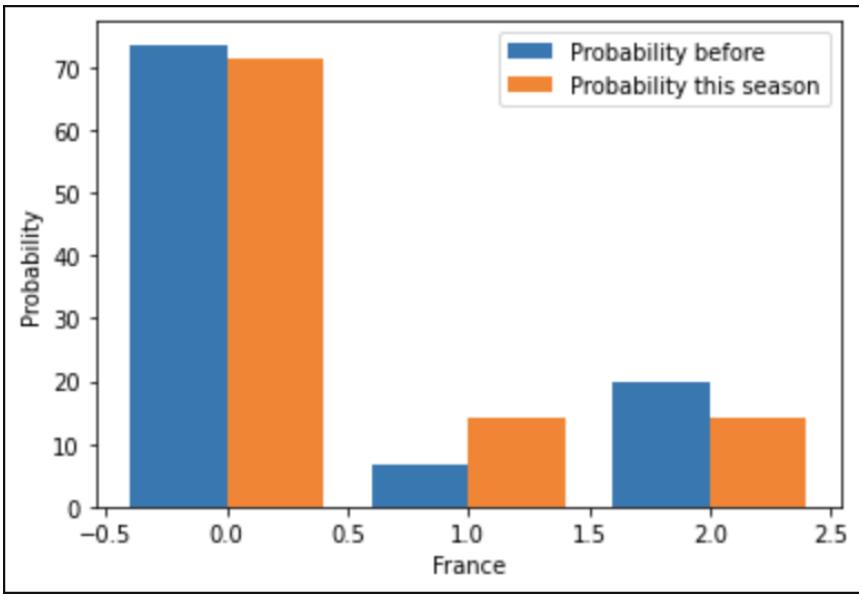
that the performance is not the same. So we decided five as the limit that the team is still the same. And of course there are less games in one season than in five seasons so we put all points in the same base.

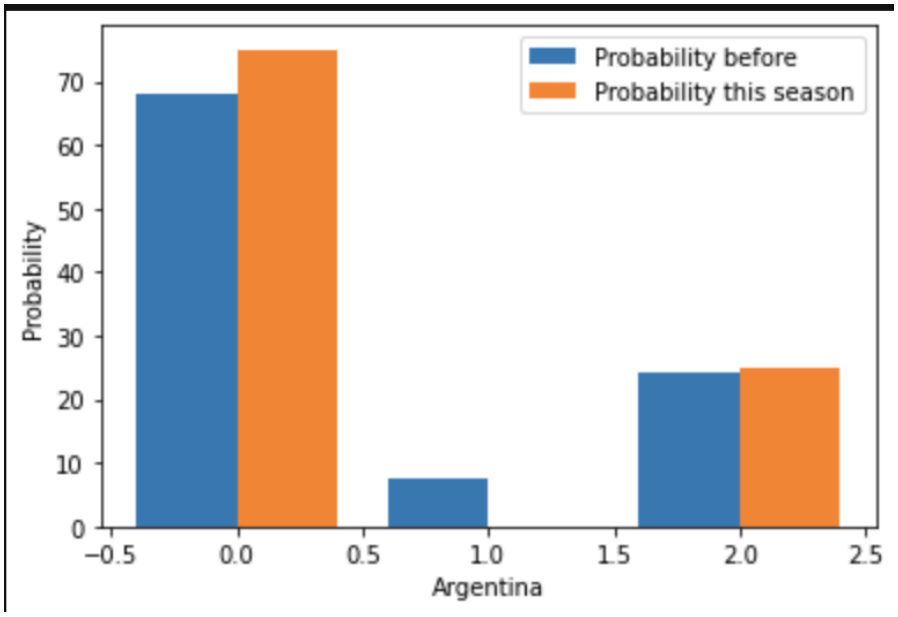
	TEAM	forme
59	Ecuador	Regular
134	Netherlands	Hot
154	Qatar	Regular
163	Senegal	Hot
62	England	Regular
89	IR Iran	Hot
198	USA	Hot
207	Wales	Cold
8	Argentina	Hot
124	Mexico	Regular
151	Poland	Regular
161	Saudi Arabia	Hot
11	Australia	Regular
55	Denmark	Hot
71	France	Cold
193	Tunisia	Hot
48	Costa Rica	Hot
75	Germany	Hot
97	Japan	Hot
174	Spain	Regular
19	Belgium	Cold
36	Canada	Regular
49	Croatia	Regular
129	Morocco	Hot
27	Brazil	Regular
35	Cameroon	Hot
164	Serbia	Hot
182	Switzerland	Hot
76	Ghana	Cold
102	Korea Republic	Hot
152	Portugal	Regular
202	Uruguay	Regular



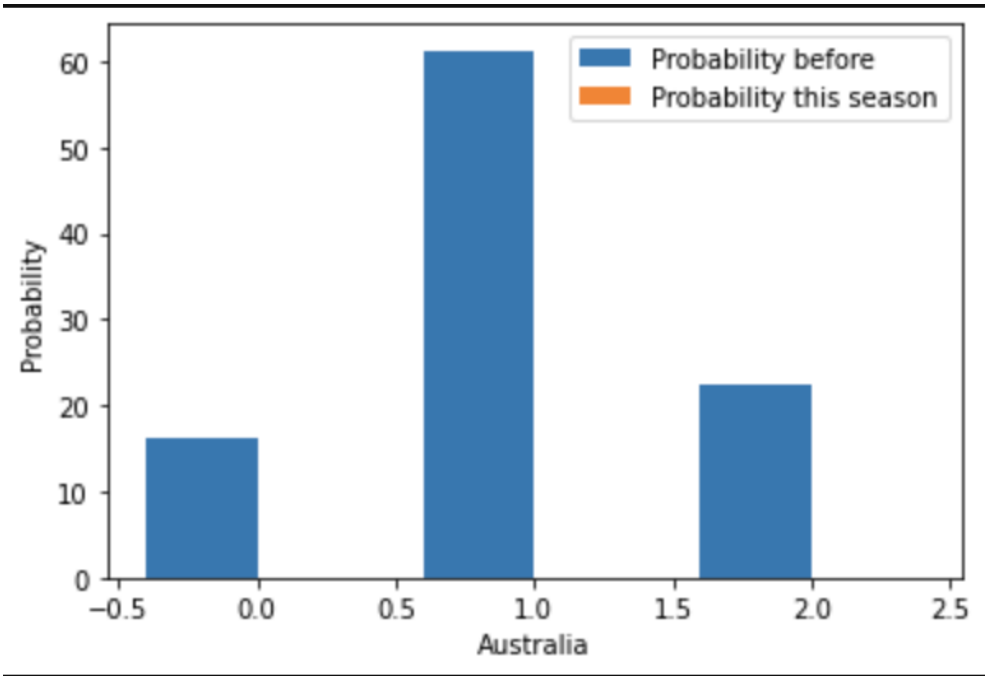
The next step is to know the probability of winning for each category of team. We have more than 200 teams in this dataset, but not all teams are in the same level, some are extremely strong, others are really bad. So we need to create a bin that will create categories for each team (Brazil in World Class, Australia in Bad…).

After this, we check all games of one team against all teams in the same bin and then find the probability of winning for this team. This is interesting to know because even if two teams are in the same bin, it doesn't mean they have the same probability of winning against a weak team. This is an example of France and Argentina (both world class) against a weak team like Australia. The first bar is the percentage of winning, the second is to Lose and last is a Draw. In Blue it's the average (last five years) and in orange it's this season so then we will just merge these two periods together.





Now we need to do the exact same thing for the weak team (Australia) against a World Class (like France) because France has a 72% of winning against Australia but if we look the graph for Australia, they have 61% of losing against a team like France so we need to merge the winning probability of France with the losing probability of Australia and we get the result off this one game.



From here we can simulate all the games we want, so this way we did for France against Australia, we will do the exact same thing for all games. So we created a simulation that will show us all the games that we will have. With this, we can know who will have enough points to pass the group stage and then we continue.

This is all the simulation using the algorithm of who is most likely to win one game. If we use

the algorithm for the last game France VS Argentina, the simulation will say that the winner will be France.

THIS IS THE LAST PAGE:

ANY REFERENCES or SOURCES can go [HERE]