FIFA WORLD CUP WINNER PREDICTION

PROBABILITY AND STATISTICS

20F-0219
20F-0244
20F-0267
20F-0270

R-CODE

```r
#setting working directory
setwd("C:/Users/moghe/Desktop/Prob/fifa-world-cup-2022-prediction-main/data")

#installing other packages
install.packages('dplyr')
library('dplyr')

#reading data
df_historical_data=read.csv('FIFA_WorldCup_Matches.csv',TRUE,",")
df_roundOf16=read.csv('RoundOf16.csv',TRUE,",")

#to display all rows
options(max.print=100000)

#displaying data
print(df_historical_data)
print(df_roundOf16)

#Team Strength Calculation
#separating data of home and away teams
# 'select' will store only those columns whose number is given
df_home = select(df_historical_data,1,4,5)
df_away = select(df_historical_data,2,4,5)

#renaming columns
#because goalsScored is actually conceded by away team
colnames(df_home)=c("Team","GoalsScored","GoalsConceded")
colnames(df_away)=c("Team","GoalsConceded","GoalsScored")

#Row Binding
df_binded = rbind(df_home,df_away)

#Team strength
#average of goals scored and conceded by team
df_team_strength=df_binded%>%group_by(Team)%>%
  summarise(GoalScored = mean(GoalsScored),
            GoalConceded = mean(GoalsConceded))

#displaying data
print(df_team_strength)

#Points Prediction
predict_Winner<-function(home,away)
{
  #getting indices of countries
  # 'which' returns index from dataframe
  HomeRowIndex=which(df_team_strength$Team==home)
  AwayRowIndex=which(df_team_strength$Team==away)
```

```
  #getting their scored and conceded goals average in home and away
  scored=as.double(df_team_strength[HomeRowIndex,2])
  #conceded=as.double(df_team_strength[AwayRowIndex,3])

  #lamb_home = as.double(scored*conceded)
  lamb_home = as.double(scored)

  scored=as.double(df_team_strength[AwayRowIndex,2])
  #conceded=as.double(df_team_strength[HomeRowIndex,3])

  #lamb_away = as.double(scored*conceded)
  lamb_away = as.double(scored)

  prob_draw = prob_home = prob_away = 0

  for(x in 0:11){
    for(y in 0:11){
      p = dpois(x, lamb_home) * dpois(y, lamb_away)

      if (x == y){
        prob_draw = p + prob_draw
      }
      else if (x > y){
        prob_home = p+prob_home
      }
      else{
        prob_away = p+prob_away
      }
    }
  }

  points_home = 3 * prob_home + prob_draw
  points_away = 3 * prob_away + prob_draw

  #winner of match based on points
  if(points_home>points_away)
    return(home)
  else
    return(away)
}

#round of 16
for(x in 1:8)
{
  if(x==1)
    df_quarterFinal=NULL

  df_quarterFinal=rbind(df_quarterFinal,
                        c(predict_Winner(df_roundOf16[x,1],df_roundOf16[x,2])))
}
```

```
print(df_quarterFinal)

#quarter finals
for(x in seq(from=1, to=8, by=2))
{
  if(x==1)
    df_semiFinal=NULL

  df_semiFinal=rbind(df_semiFinal,

c(predict_Winner(df_quarterFinal[x,1],df_quarterFinal[x+1,1])))
}
print(df_semiFinal)

#Semifinals
for(x in seq(from=1, to=4, by=2))
{
  if(x==1)
    df_Final=NULL

  df_Final=rbind(df_Final,
                 c(predict_Winner(df_semiFinal[x,1],df_semiFinal[x+1,1])))
}
print(df_Final)

#Finals
Winner=predict_Winner(df_Final[1,1],df_Final[2,1])
print(Winner)
```

INTRODUCTION

Probability

- Introduction
- Statistical Analysis
- Players Analysis
- Prediction
- Team

FIFA WORLD CUP PREDICTION MODEL

Many people call football "the unpredictable game" because a football match has different factors that can change the final score.That's true ... to some extent. It's hard to predict the final score or the winner of a match, but that's not the case when it comes to predicting the winner of a competition.
So, now we are making a model to predict the World Cup 2022.

How are we going to predict the matches?

There are different ways to make predictions. We have decided to give a chance to the Poisson distribution. Why? Well, let's ha a look at the definition of the Poisson distribution.

The Poisson distribution is a discrete probability distribution that describes the number of events occurring in a fixed time interval or region of opportunity.

If we think of a goal as an event that might happen in the 90 minutes of a football match, we could calculate the probability of the number of goals that could be scored in a match by Team A and Team B. But that's not enough. We still need to meet the assumptions of the Poisson distribution.

- The number of events can be counted (a match can have 1, 2, 3 or more goals)
- The occurrence of events is independent (the occurrence of one goal should not affect the probability of another goal)
- The rate at which events occur is constant (the probability of a goal occurring in a certain time interval should be exactly t same for every other time interval of the same length)
- Two events cannot occur at exactly the same instant in time (two goals can't occur at the same time)

Probability

- Introduction
- Statistical Analysis
- Players Analysis
- Prediction
- Team

Now we can say that it's possible to use the Poisson distribution to calculate the probability of the number of goals that could scored in a match. Here's the formula of the Poisson distribution.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

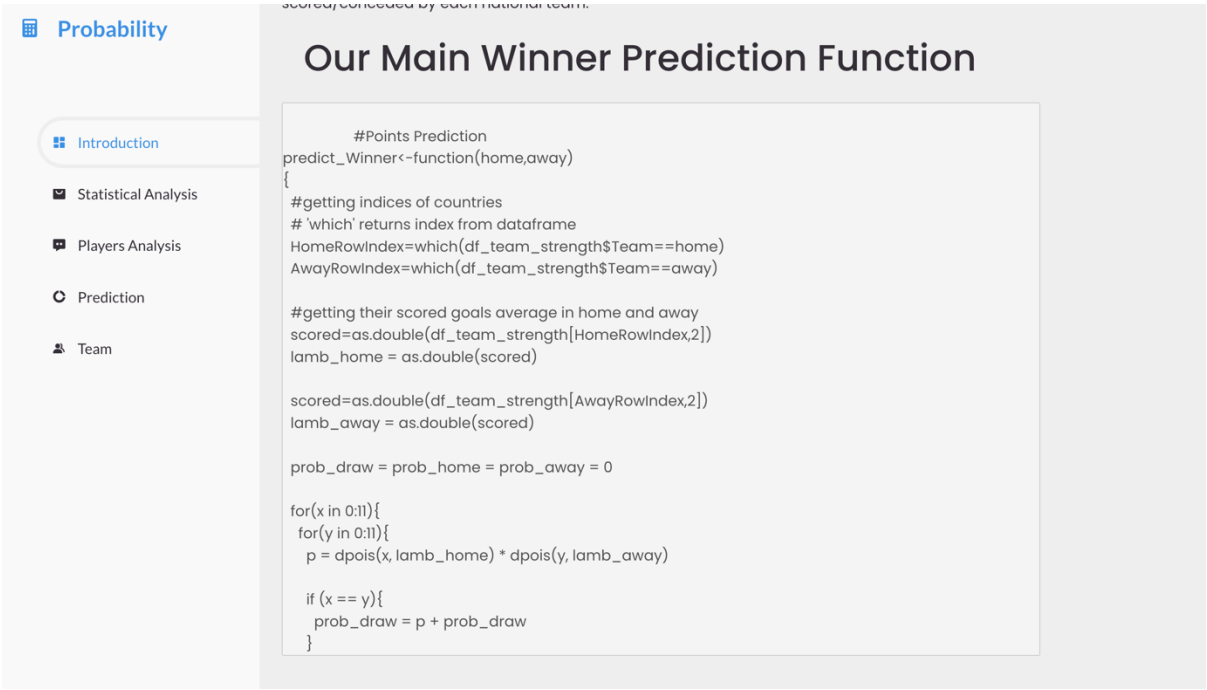
To make the predictions We considered:

lambda: median of goals in 90 minutes (Team A and Team B)

x: number of goals in a match that could be scored by Team A and Team B To calculate lambda, we need the average goals scored/conceded by each national team.

Our Main Winner Prediction Function

```
#Points Prediction
```


Probability
Introduction
Statistical Analysis
Players Analysis
Prediction
Team
Our Main Winner Prediction Function
#Points Prediction
predict_Winner<-function(home,away)
{
#getting indices of countries
# 'which' returns index from dataframe
HomeRowIndex=which(df_team_strength$Team==home)
AwayRowIndex=which(df_team_strength$Team==away)
#getting their scored goals average in home and away
scored=as.double(df_team_strength[HomeRowIndex,2])
lamb_home = as.double(scored)
scored=as.double(df_team_strength[AwayRowIndex,2])
lamb_away = as.double(scored)
prob_draw = prob_home = prob_away = 0
for(x in 0:11){
for(y in 0:11){
p = dpois(x, lamb_home) * dpois(y, lamb_away)
if (x == y){
prob_draw = p + prob_draw
}

STATISTICAL ANALYSIS

```r
#this library will help in the animation of the graphs
#install.packages("gganimate")
#this library will help convert graphs to embedded html
#install.packages("htmlwidgets")

library(htmlwidgets) #--> convert the graph into a html page
library(gganimate) # --> to animate the graphs
library(ggplot2)
library(dplyr)
library(gapminder)
library(ggthemes)
library(tidyr)
library(lubridate)
library(gifski)
library(av)
library(esquisse)

#average age of players
ggplot(avgPlayer_Data) +
 aes(x = Country, fill = Country, weight = AverageAge) +
 geom_bar() +
 scale_fill_viridis_d(option = "magma",
 direction = 1) +
 labs(x = "Country", y = "Average Age", title = "Average age of players regarding
their country") +
 theme_light()

#market value graph
ggplot(marketValue_Data) +
 aes(x = Name, y = `Value (Million in Euros)`, fill = Name) +
 geom_col() +
 scale_fill_hue(direction = 1) +
 labs(x = "Countries") +
 theme_gray()

#most appear graph
ggplot(MostAppear_Data) +
 aes(x = Player, y = Appearances, fill = Player) +
 geom_col() +
 scale_fill_hue(direction = 1) +
 labs(x = "Player Name", y = "Num# of Appearences ") +
 ggthemes::theme_tufte()

#win percentage
ggplot(winPercent_Data) +
 aes(x = Country, y = `%in wining`) +
 geom_col(fill = "#112446") +
 theme_gray()
```

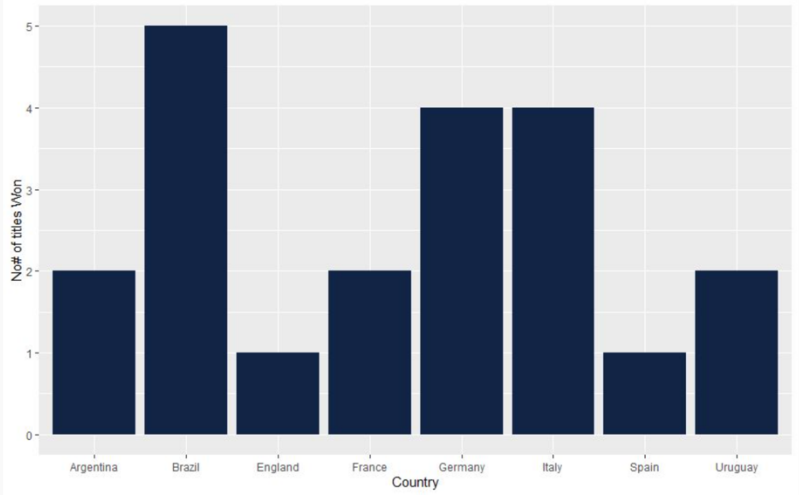
```r
#most titles
ggplot(mostTitles_Data) +
 aes(x = Country, y = `No# of titles Won`) +
 geom_col(fill = "#112446") +
 theme_gray()

#num of goals
Prob_Data %>%
 filter(Year >= 2018L & Year <= 2018L) %>%
 ggplot() +
 aes(x = HomeTeam, y = TotalGoals, fill = HomeTeam) +
 geom_col() +
 scale_fill_manual(values = c(Algeria = "#F8766D",
Angola = "#F47862", Argentina = "#F07B57", Australia = "#ED7E4C", Austria =
"#E98141", Belgium = "#E58337",
Bolivia = "#E2862C", `Bosnia and Herzegovina` = "#DE8921", Brazil = "#DA8C16",
Bulgaria = "#D78E0C",
Cameroon = "#D39101", Canada = "#CD9400", Chile = "#C79600", China = "#C09800",
Colombia = "#BA9B00",
`Costa Rica` = "#B49D00", Croatia = "#AD9F00", Cuba = "#A7A200", `Czech Republic` =
"#A1A400", Czechoslovakia = "#9AA700",
Denmark = "#94A900", `East Germany` = "#88AB04", Ecuador = "#79AC09", Egypt =
"#6BAE0F", England = "#5CAF14",
`FR Yugoslavia` = "#4EB11A", France = "#3FB31F", Germany = "#31B425", Ghana =
"#22B62A", Greece = "#13B730",
Haiti = "#05B935", Honduras = "#00BA3E", Hungary = "#00BB48", Iceland = "#00BB52",
Iran = "#00BC5C",
Iraq = "#00BD67", Israel = "#00BD71", Italy = "#00BE7B", `Ivory Coast` = "#00BF85",
Jamaica = "#00BF8F",
Japan = "#00C099", Mexico = "#00C0A2", Morocco = "#00BFA9", Netherlands =
"#00BFAF", `New Zealand` = "#00BEB6",
Nigeria = "#00BDBD", `North Korea` = "#00BCC3", `Northern Ireland` = "#00BBCA",
Norway = "#00BBD1", Panama = "#00BAD8",
Paraguay = "#00B9DE", Peru = "#03B7E4", Poland = "#0DB5E6", Portugal = "#16B2E9",
`Republic of Ireland` = "#20AFEC",
Romania = "#29ACEF", Russia = "#33A9F1", `Saudi Arabia` = "#3DA6F4", Scotland =
"#46A3F7", Senegal = "#50A1FA",
Serbia = "#599EFC", `Serbia and Montenegro` = "#649AFE", Slovakia = "#7096FE",
Slovenia = "#7C92FE",
`South Africa` = "#888EFD", `South Korea` = "#948AFD", `Soviet Union` = "#A086FC",
Spain = "#AC82FC",
Sweden = "#B87DFC", Switzerland = "#C479FB", Togo = "#D075FB", `Trinidad and
Tobago` = "#DB71FA", Tunisia = "#DE70F4",
Turkey = "#E26EEF", Ukraine = "#E66CE9", `United Arab Emirates` = "#E96BE4",
`United States` = "#ED69DE",
Uruguay = "#F067D9", Wales = "#F466D3", `West Germany` = "#F764CE", Yugoslavia =
"#FB62C8", Zaire = "#FF61C3"
)) +
 labs(x = "Countries") +
 theme_minimal()
```

Probability

- Introduction
- Statistical Analysis
- Players Analysis
- Prediction
- Team

FIFA WORLD CUP STATISTICAL ANALYSIS

Total No of title won by the Teams

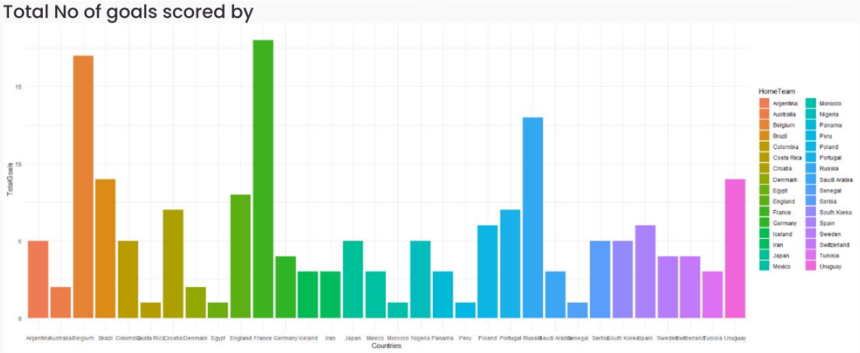


Total No of goals scored by

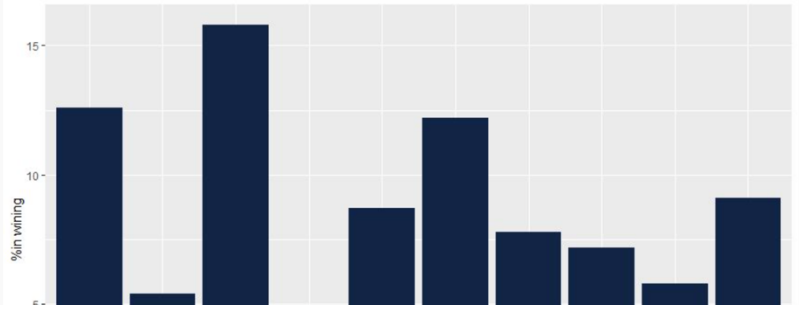
Probability

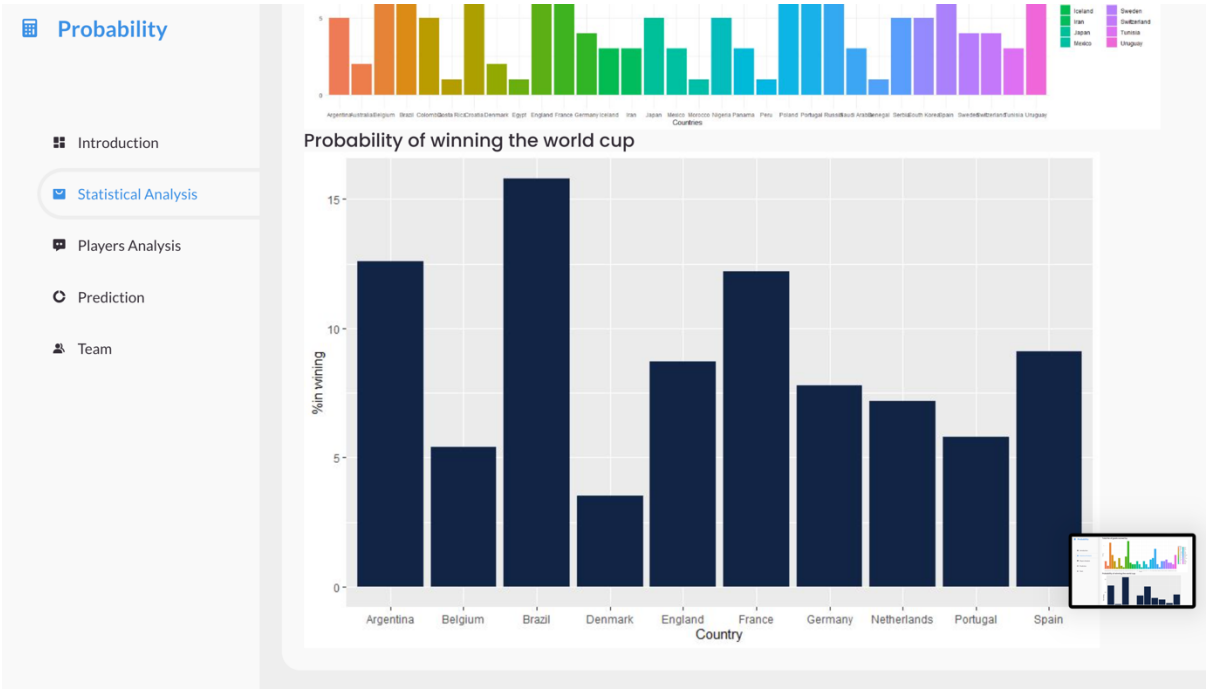
- Introduction
- Statistical Analysis
- Players Analysis
- Prediction
- Team

Total No of goals scored by

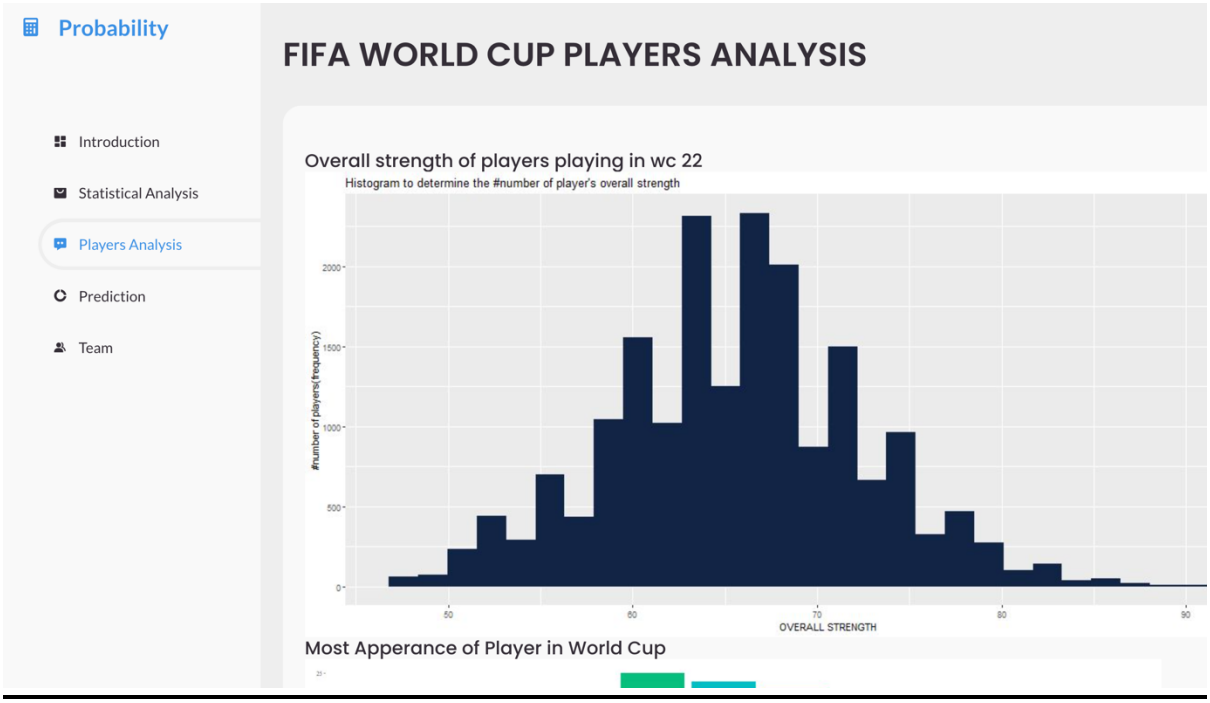


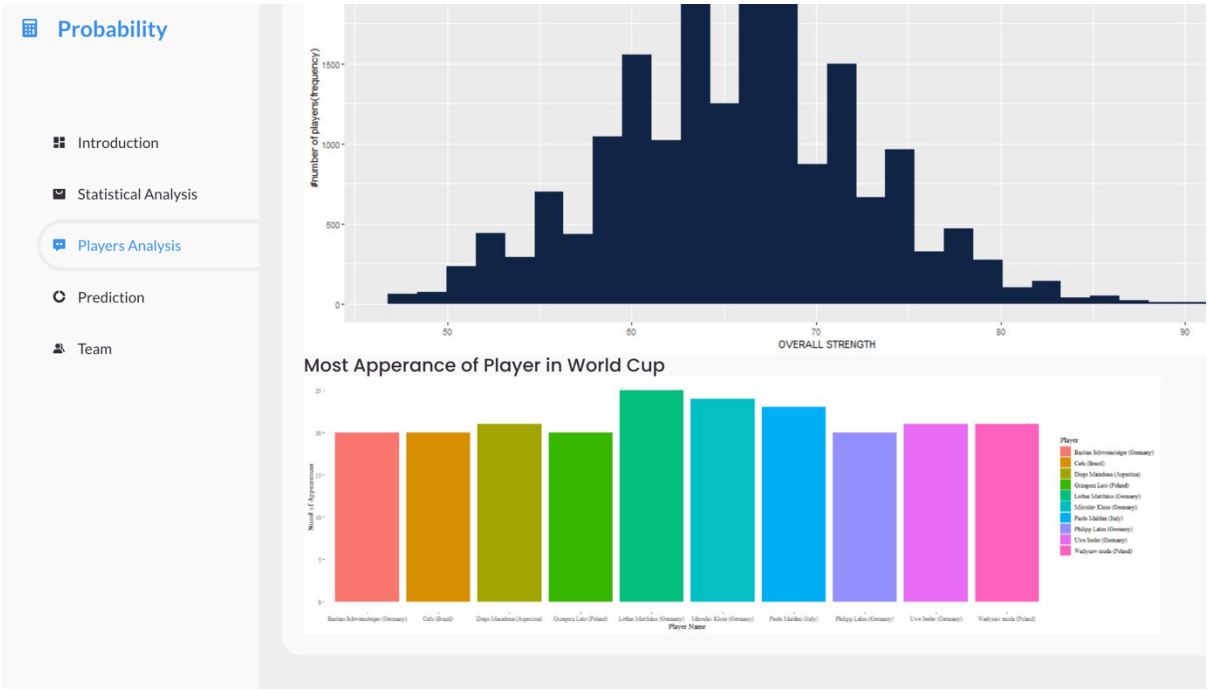
Probability of winning the world cup





PLAYERS ANALYSIS




Probability
Introduction
Statistical Analysis
Players Analysis
Prediction
Team
#number of players(frequency)
OVERALL STRENGTH
Most Apperance of Player in World Cup
Player Name

PREDICTION


Probability
Introduction
Statistical Analysis
Players Analysis
Prediction
Team
PREDICTE THE FIFA WORLD CUP 2022
PREDICT QUATER FINALS
PREDICT SEMI FINALS
PREDICT FINAL


Probability
Introduction
Statistical Analysis
Players Analysis
Prediction
Team
PREDICTE THE FIFA WORLD CUP 2022
PREDICT QUATER FINALS
home score away year winner
Germany Match 58 Brazil 2022 ?
Netherlands Match 57 Argentina 2022 ?
Spain Match 60 Portugal 2022 ?
England Match 59 France 2022 ?
Predicting the QuaterFinals.....
PREDICT SEMI FINALS
PREDICT FINAL


Probability
Introduction
Statistical Analysis
Players Analysis
Prediction
Team
PREDICTE THE FIFA WORLD CUP 2022
PREDICT QUATER FINALS
RESULT
DONE
home score away year winner
Germany Match 58 Brazil 2022 Brazil
Netherlands Match 57 Argentina 2022 Netherlands
Spain Match 60 Portugal 2022 Portugal
England Match 59 France 2022 France
PREDICT SEMI FINALS