

激战世界杯

2018俄罗斯世界杯





蔡 承 翰

張 仁 樵

蕭 資 峻

田 碩

目錄



資料介紹



分析議題



解決問題





資料介紹



來源資料



Preview (first 100 rows) Column Metadata Column Metrics								
date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland	FALSE
1873-03-08	England	Scotland	4	2	Friendly	London	England	FALSE
1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland	FALSE
1875-03-06	England	Scotland	2	2	Friendly	London	England	FALSE
1876-03-04	Scotland	England	3	0	Friendly	Glasgow	Scotland	FALSE
1876-03-25	Scotland	Wales	4	0	Friendly	Glasgow	Scotland	FALSE
1877-03-03	England	Scotland	1	3	Friendly	London	England	FALSE
1877-03-05	Wales	Scotland	0	2	Friendly	Wrexham	Wales	FALSE
1878-03-02	Scotland	England	7	2	Friendly	Glasgow	Scotland	FALSE
1878-03-23	Scotland	Wales	9	0	Friendly	Glasgow	Scotland	FALSE
1879-01-18	England	Wales	2	1	Friendly	London	England	FALSE
1879-04-05	England	Scotland	5	4	Friendly	London	England	FALSE
1879-04-07	Wales	Scotland	0	3	Friendly	Wrexham	Wales	FALSE
1880-03-13	Scotland	England	5	4	Friendly	Glasgow	Scotland	FALSE
1880-03-15	Wales	England	2	3	Friendly	Wrexham	Wales	FALSE
1880-03-27	Scotland	Wales	5	1	Friendly	Glasgow	Scotland	FALSE
1881-02-26	England	Wales	0	1	Friendly	Blackburn	England	FALSE
1881-03-12	England	Scotland	1	6	Friendly	London	England	FALSE



來源資料



Preview (first 100 rows) Column Metadata Column Metrics			
date dateTime	Null 0%	Mean Jun 24th 89	<div>Nov 30th 72</div> <div>Jun 9th 18</div>
home_team string	Null 0%	Unique 241	<div>Brazil (1%) Argentina (1%) Germany (1%) Mexico (1%) England (1%)</div>
away_team string	Null 0%	Unique 242	<div>Uruguay (1%) Sweden (1%) England (1%) Hungary (1%) Paraguay (1%)</div>
home_score numeric	Null 0%	Mean 1.74	<div>0.00 31.0</div>
away_score numeric	Null 0%	Mean 1.18	<div>0.00 21.0</div>
tournament string	Null 0%	Unique 95	<div>Friendly (42%) FIFA World Cup... (18%) UEFA Euro qual... (6%) African Cup of... (4%) FIFA World Cup (2%)</div>
city string	Null 0%	Unique 1.79k	<div>Kuala Lumpur (1%) Bangkok (1%) Doha (1%) Budapest (1%) London (1%)</div>
country string	Null 0%	Unique 263	<div>USA (3%) France (2%) Malaysia (2%) Germany (1%) England (1%)</div>
neutral boolean	Null 0%		<div>True 24% False 76%</div>

<https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017/data>



資料異常



##	date	home_team	away_team	home_score	away_score	total_goals
## 1	2001-04-11	Australia	American Samoa	31	0	31

https://en.wikipedia.org/wiki/Australia_31%E2%80%930_American_Samoa



分析議題

動機

時逢四年一次的世界盃，愛好足球運動的本小組員對於本屆世界盃冠軍獎落誰家爭論不休，為此我們分析世界盃隊伍交鋒時的勝率來找出誰才是最有希望奪冠的隊伍。





假 設

1.南美洲與歐洲的國家
足球實力較強

3.參加世界盃的國家
是不是最強的32個
國家

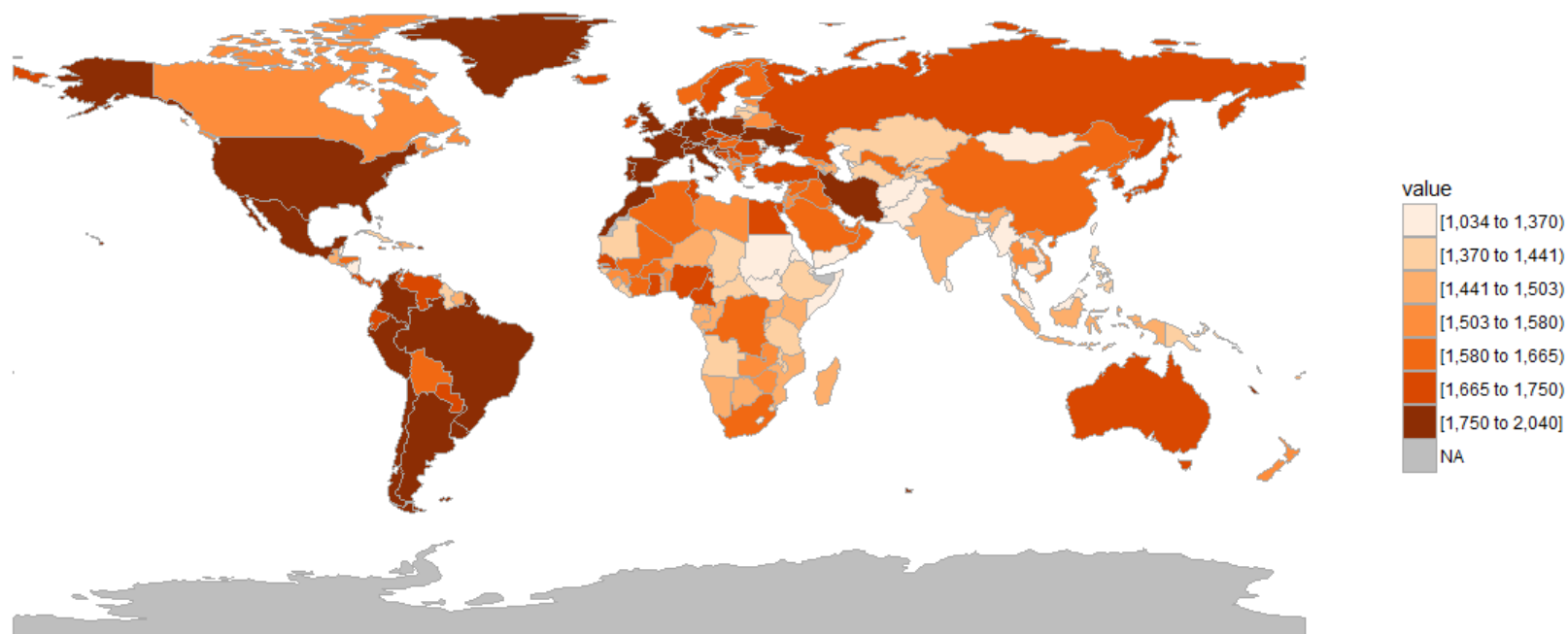


2.我們較看好巴西、
德國、西班牙等強隊
能否奪冠



分析結果

足球實力面量圖

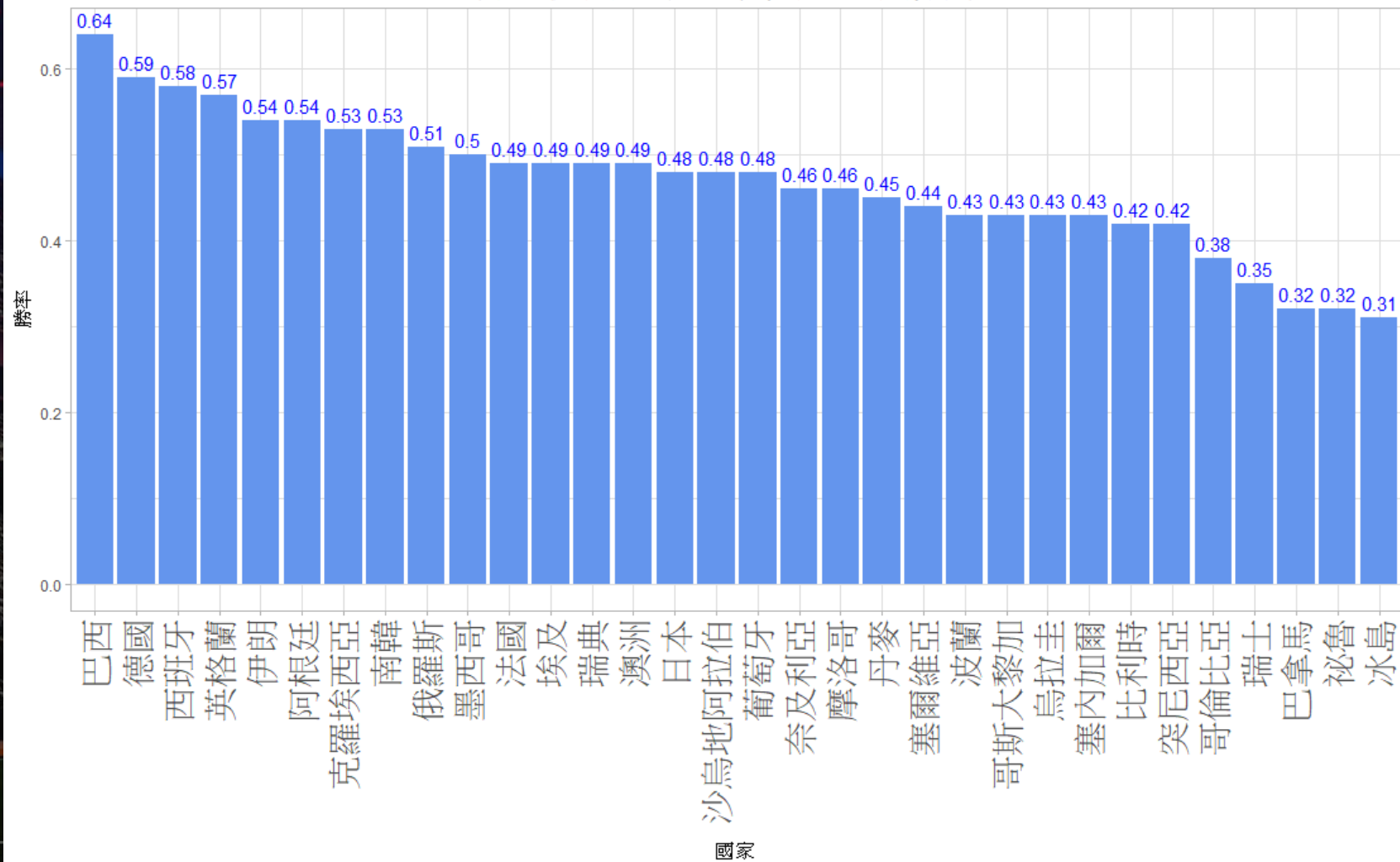


由此面量圖可看出世界足球實力的強國大多分布在歐洲及南美洲，甚至南美洲最弱的國家都能比上亞洲最強的國家。



分析結果

世界盃參賽國家勝率(所有比賽)條狀圖

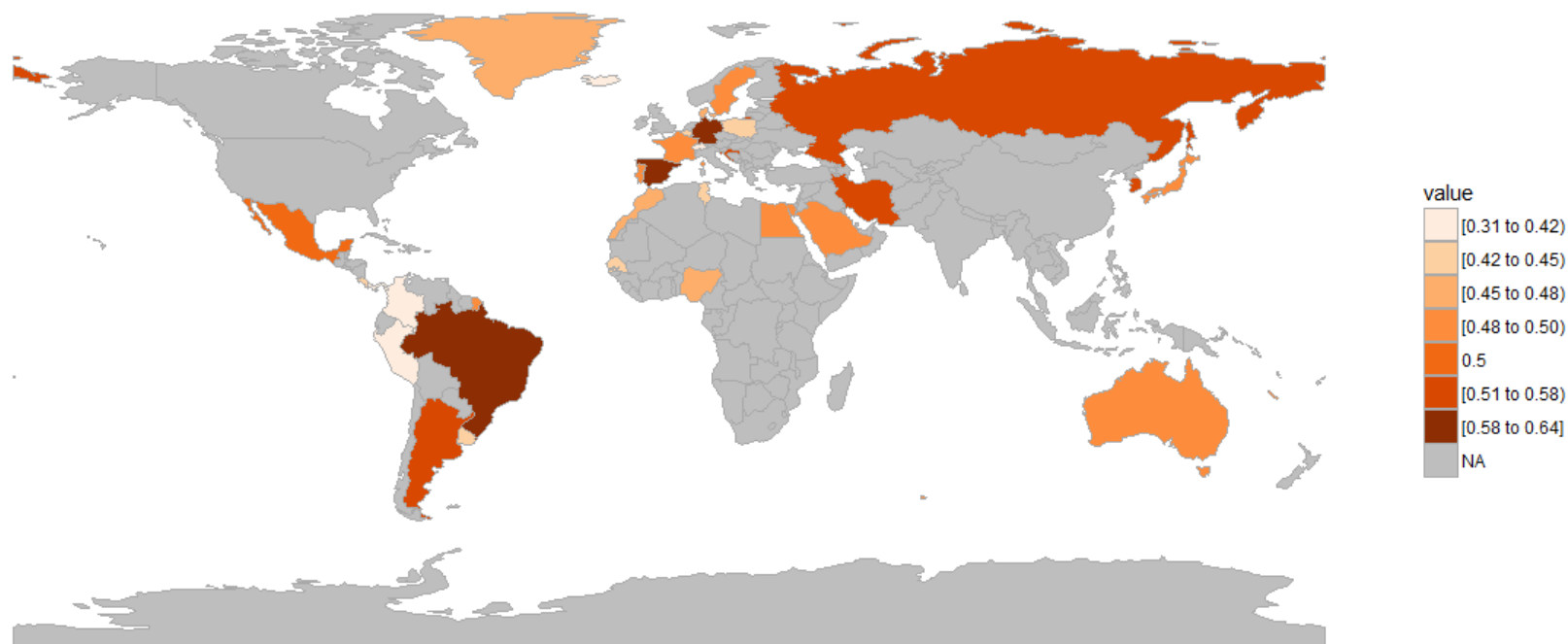


由此圖可看出所有比賽勝率最高的前三名(巴西、德國、西班牙)，剛好都是前三屆世界盃的冠軍，而超乎我們想像的伊朗居然排在第五，我們猜測可能是他所在亞洲的緣故。



分析結果

世界盃參賽國家勝率(所有比賽)面量圖

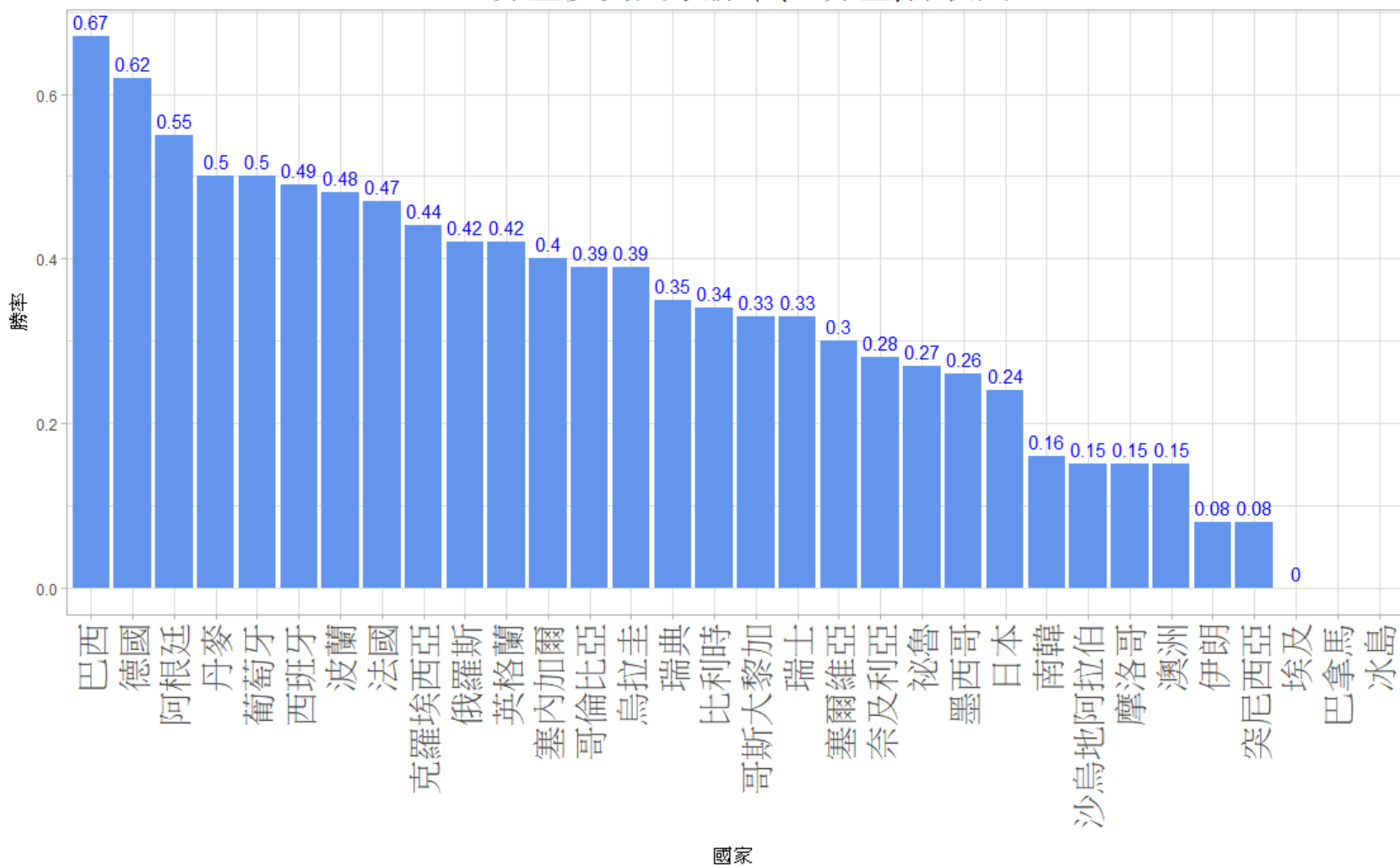


此圖是由上頁的長條圖是轉換而成的面量圖，可以較直觀的看出，此次參加世界盃的所有國家的勝率分布情形。



分析結果

世界盃參賽國家勝率(世界盃)條狀圖

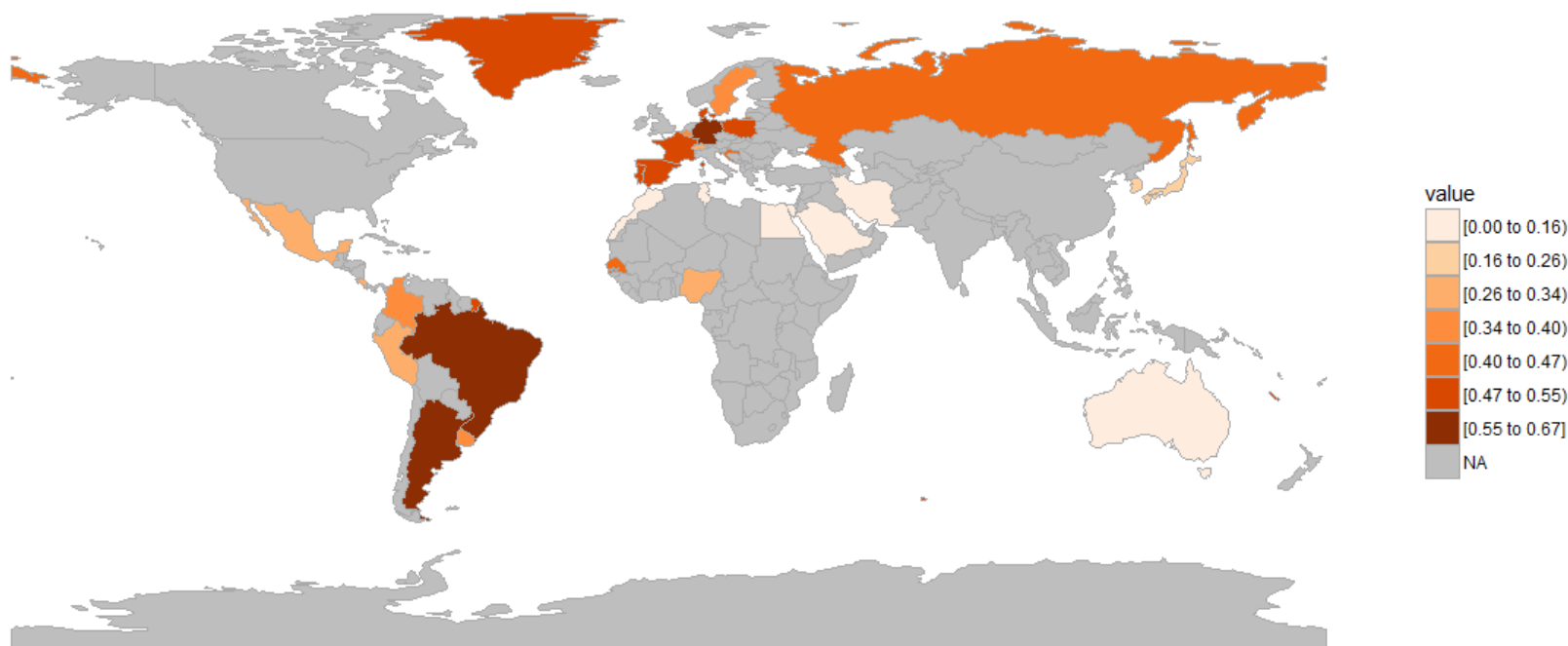


由此圖可以看出，在所有比賽中勝率較高的西班牙、伊朗都掉到較後面的名次，並且我們可以發現埃及到現在還從未在世界盃取勝，而巴拿馬和冰島則是首次參賽。



分析結果

世界盃參賽國家勝率(世界盃)面量圖

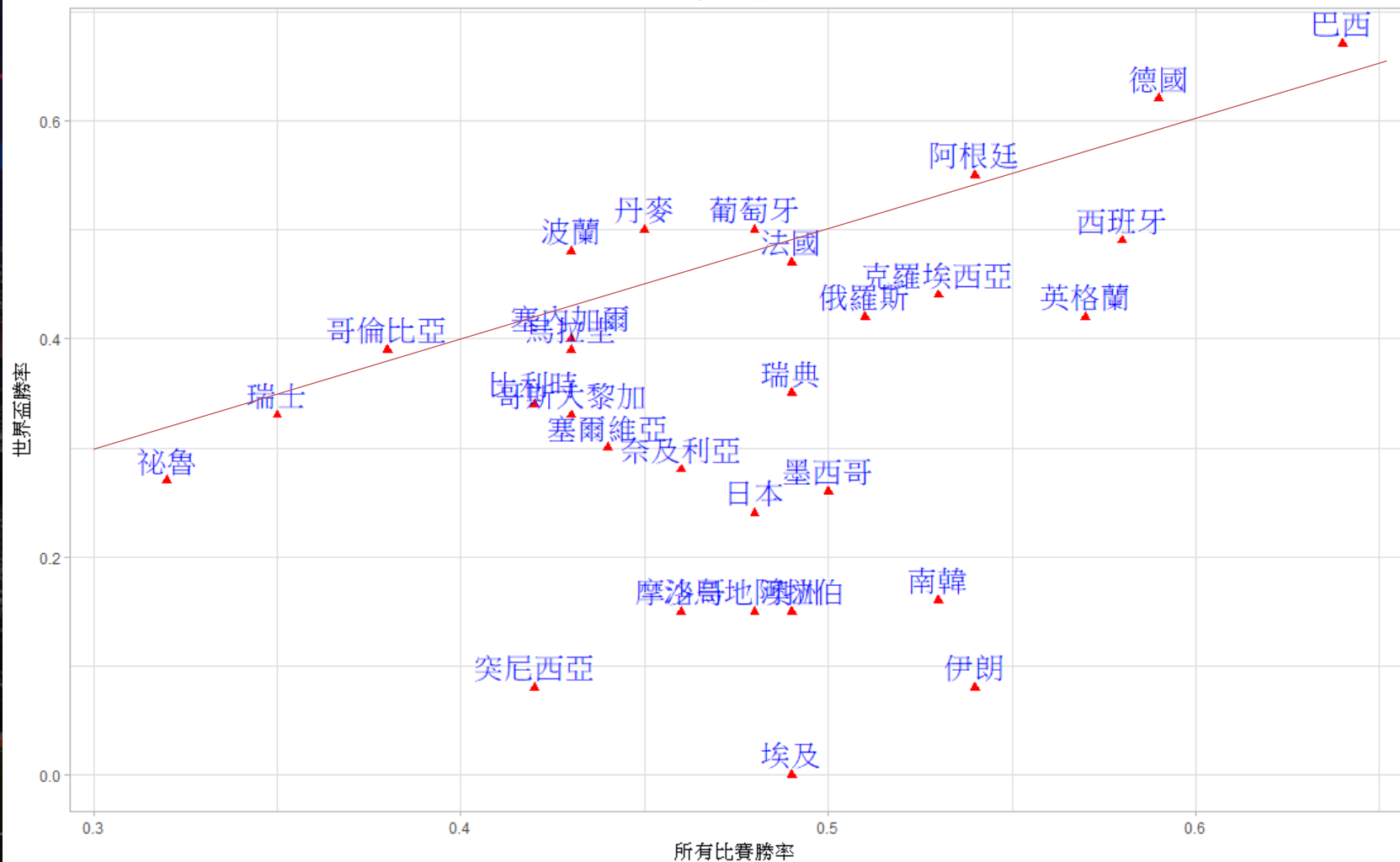


此圖是由上頁的長條圖是轉換而成的面量圖，可以較直觀的看出，此次參加世界盃的所有國家的歷屆世界盃勝率分布情形。



分析結果

勝率比較圖



由此圖可以看出紅線之上的國家是在世界盃發揮較好的國家，而紅線之下則是平常很強，但一到世界盃就會腳軟的國家。



Elo 介紹



● 玩家的Elo等級由一個數字表示，每場比賽之後，勝者將從失敗者中獲得積分。獲勝者和失敗者的評分之間的差異決定了比賽后獲得或失去的總分數。




1. 如果高評分者獲勝，只有少數評分從低評分者身上獲得
2. 如果低評分者獲勝，許多評分點將被轉移。
3. 如果平局，評分較低者也將獲得較高評分者幾分。

● 我們在後面主要將會用到

1. `elo.calc()`
2. `elo.prob()`



分析結果

	Team	Elo積分
	Netherlands	1843.180
	Italy	1835.915
	Chile	1824.524

左邊三個國家是這次並未參加世界盃，但Elo積分卻大於世界盃所有參賽國家平均者。



模擬世界盃



●基本上可以使用`sample()`函數及`elo.prob`回傳的值來選擇俄羅斯和沙烏地阿拉伯之間的隨機贏家，但選擇俄羅斯的機率為60%，並在比賽完後直接更新Elo評分。

●這樣就可以模擬整個比賽從小組賽到決賽。如果你重複很多次，你會得到每個團隊的詳細奪冠機率，或其進入淘汰賽的機率。這基本上就是`FiveThirtyEight`這樣的網站為他們的體育預測所做的事情。



小組賽

A組





小組賽

B組





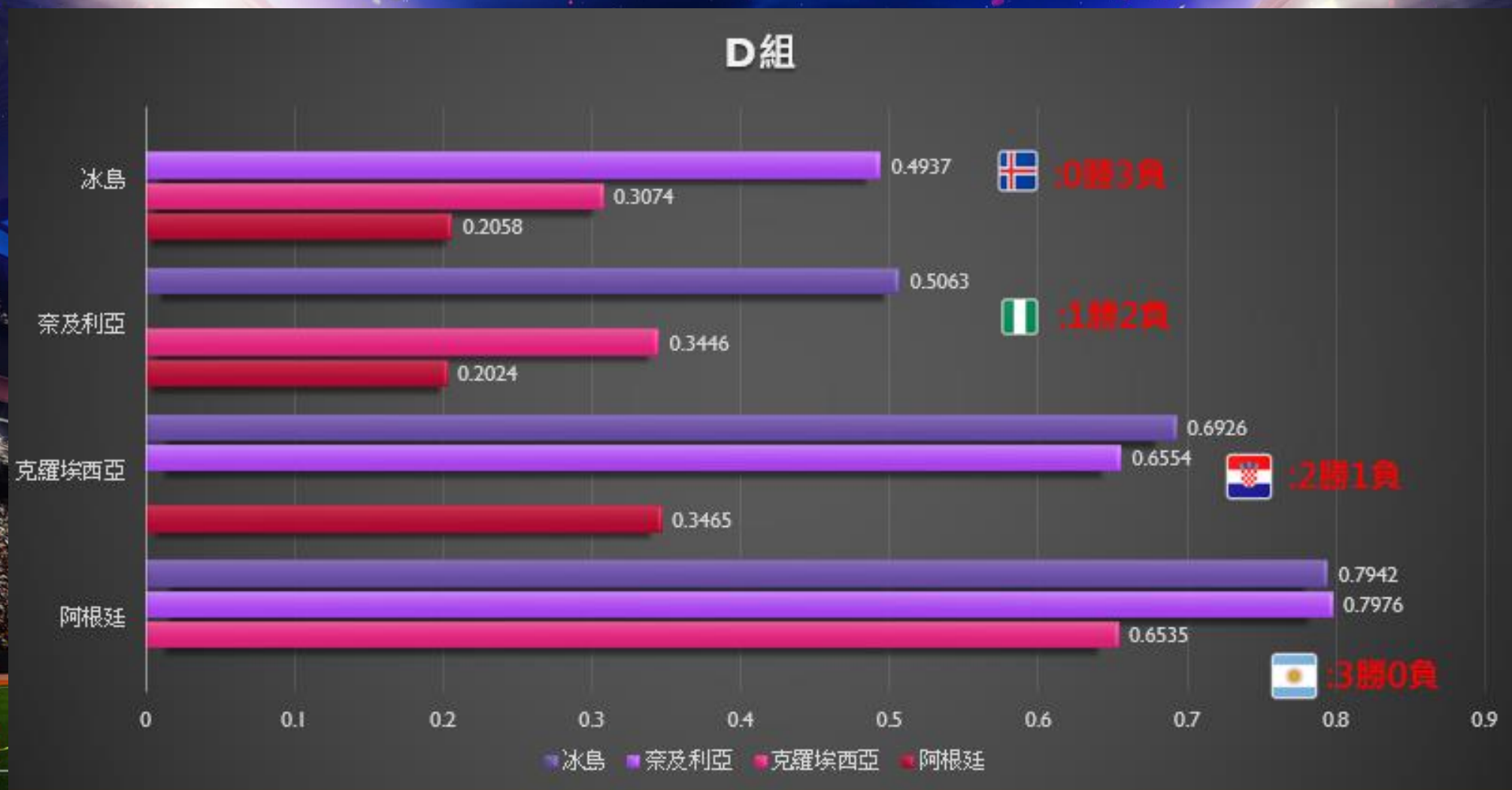
小組賽

C組





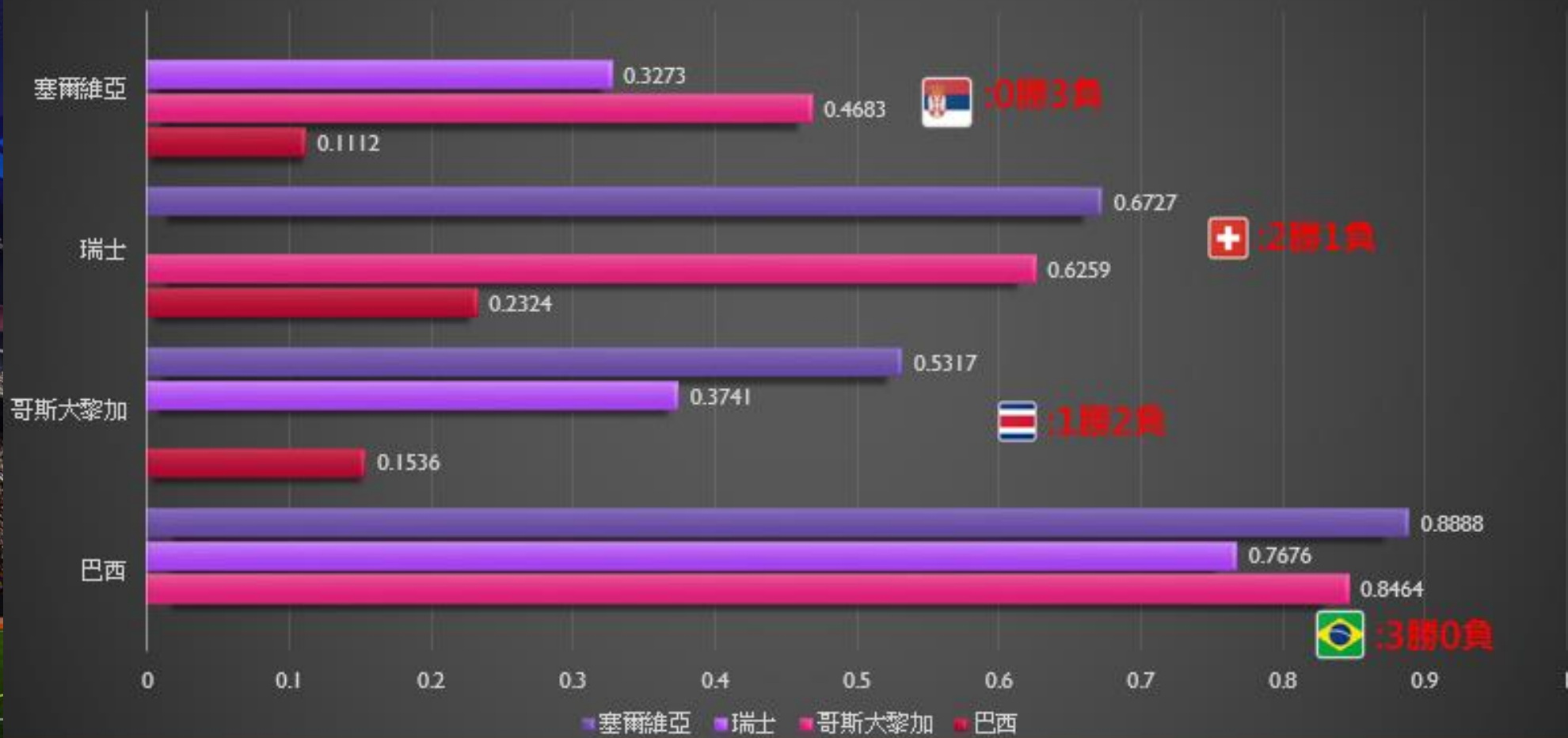
小組賽





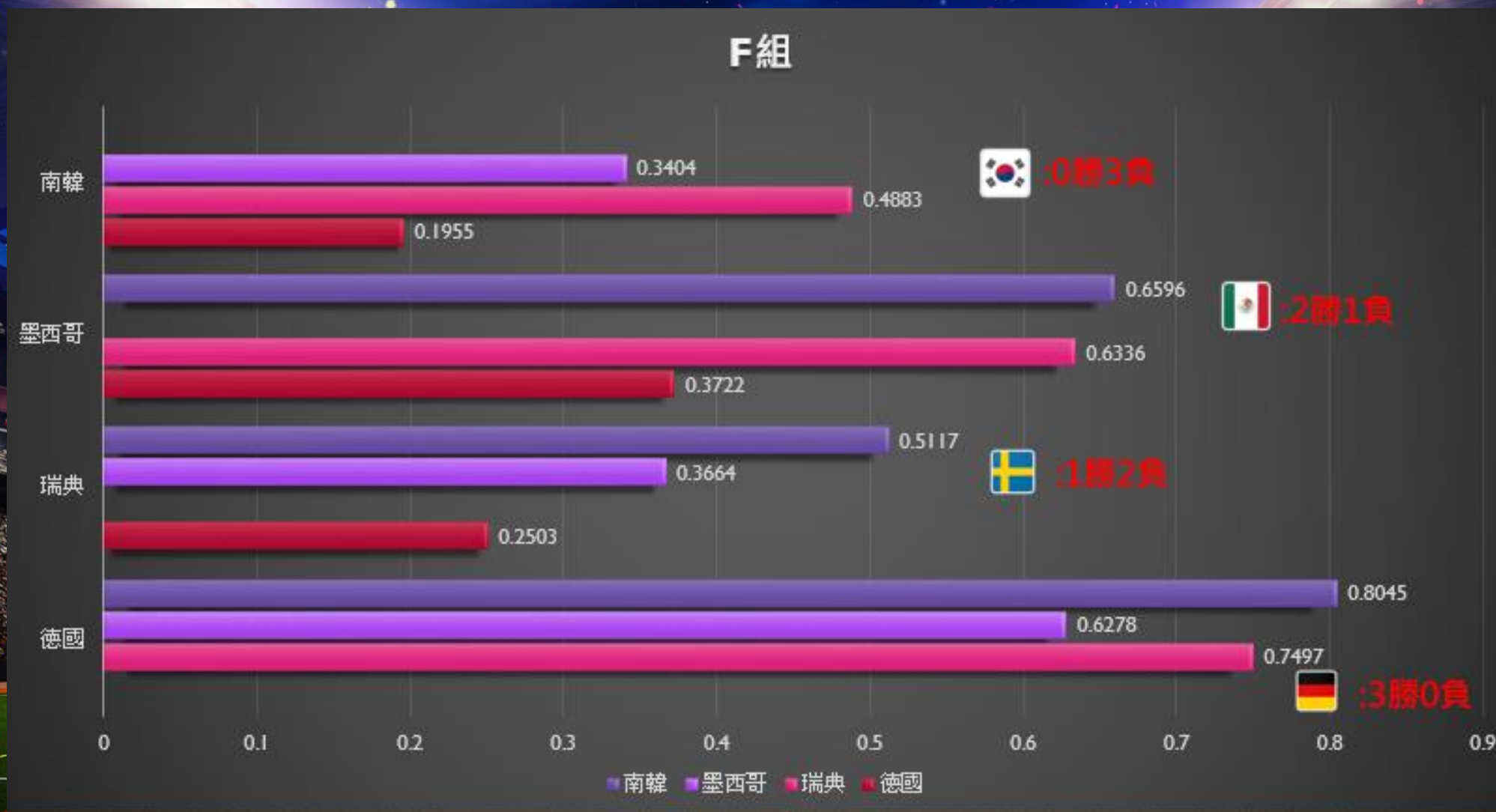
小組賽

E組





小組賽





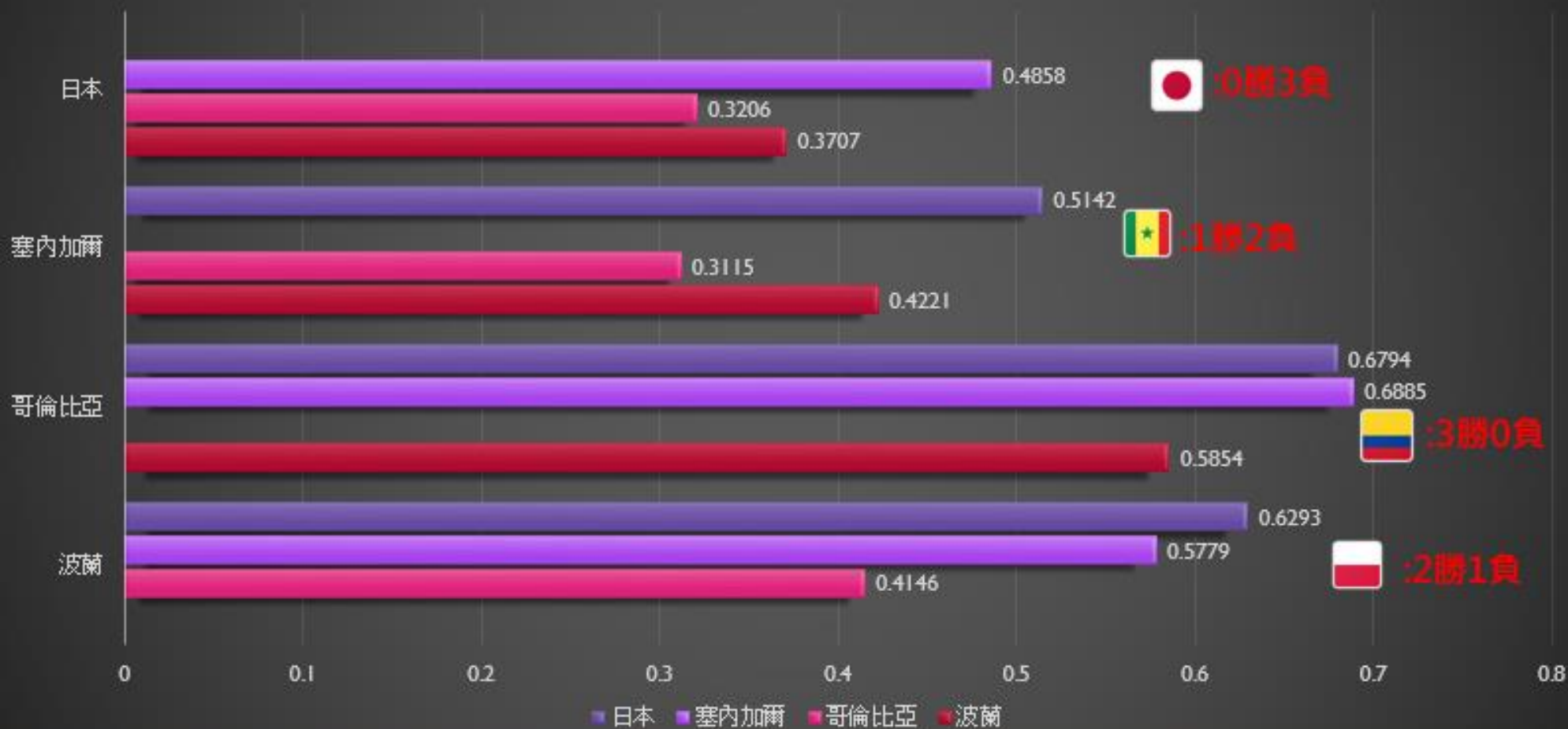
小組賽





小組賽

H組





淘汰賽



此圖為我們模擬完小組賽後，所繼續模擬的淘汰賽情形，紅色的數字代表那一隊在前面比賽勝出的機率，例如:葡萄牙上方的0.5725代表葡萄牙贏烏拉圭的機率為57.25%。

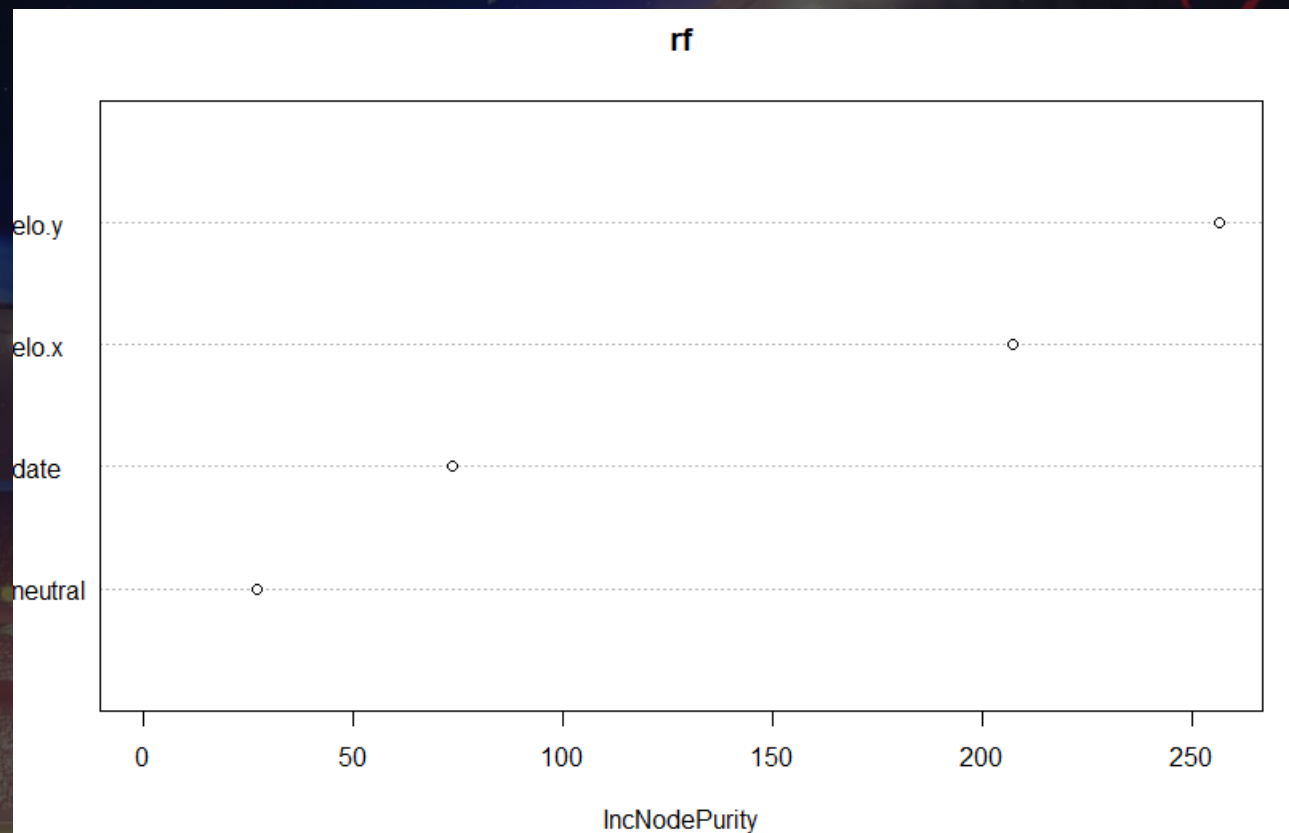


Random Forest

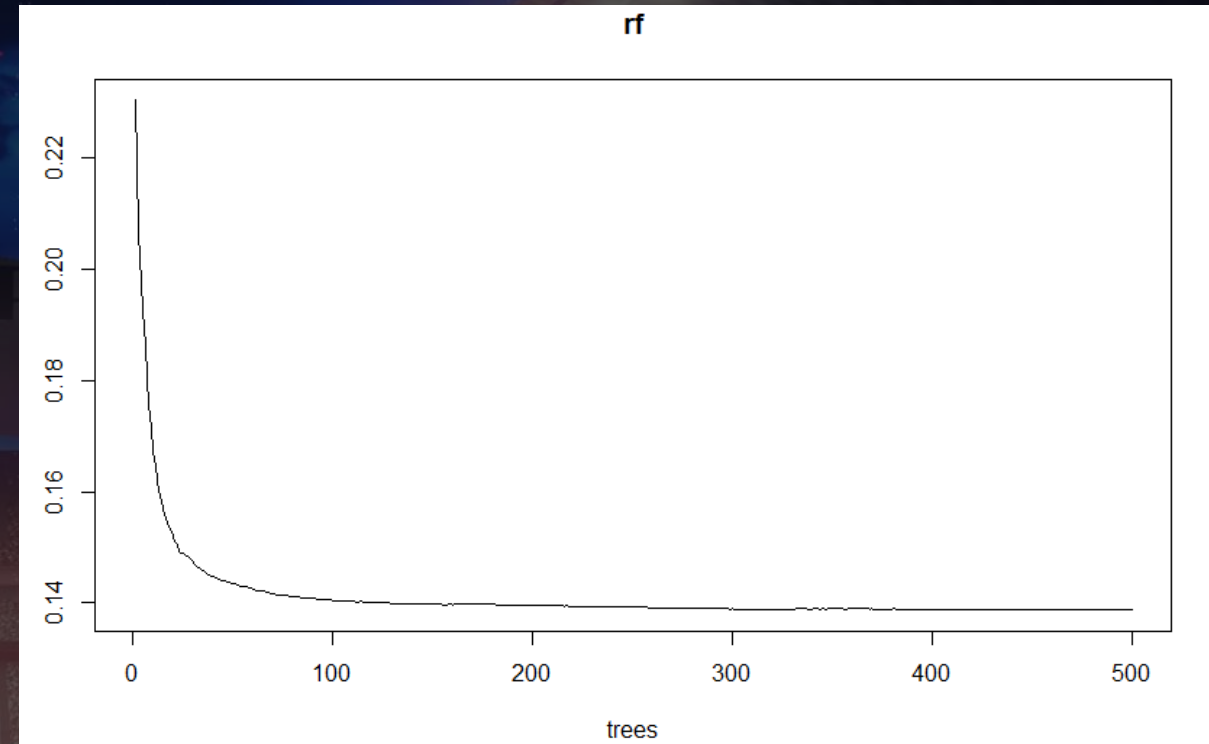
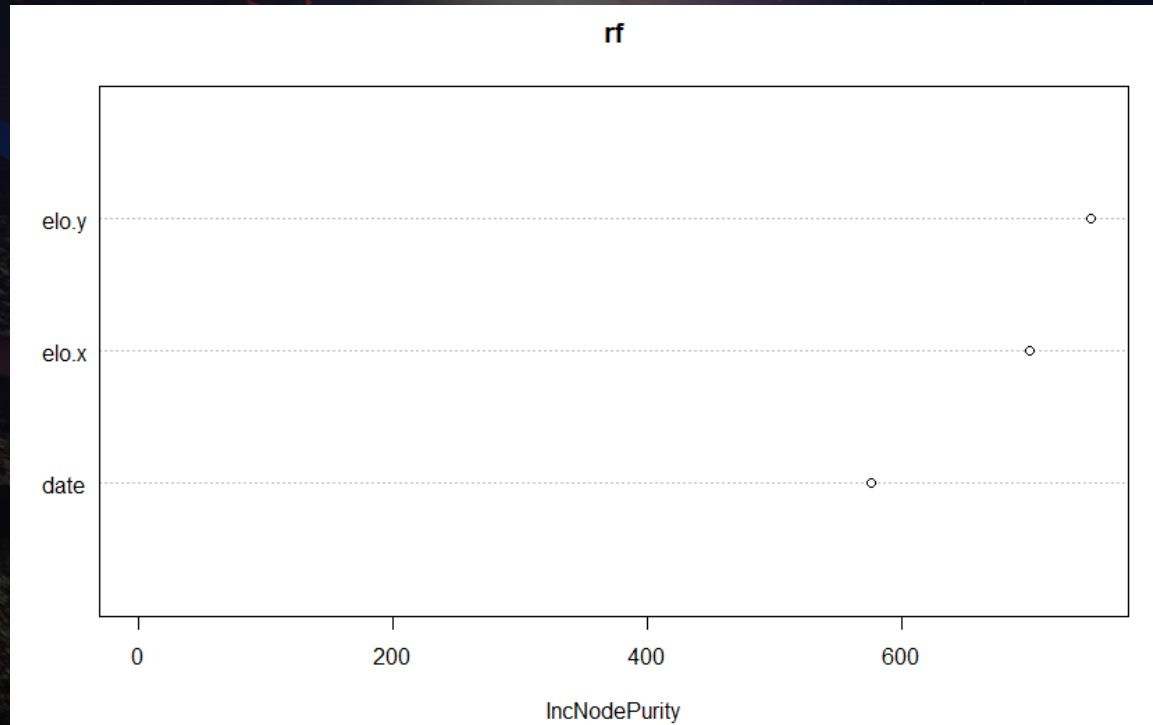
合并资料

训练测试比8 : 2

```
rf<-  
randomForest(result~elo.x+  
elo.y+date+neutral,  
data=train1)  
> varUsed(rf)  
[1] 35352 34925 36645  
7316      varImpPlot(rf)
```



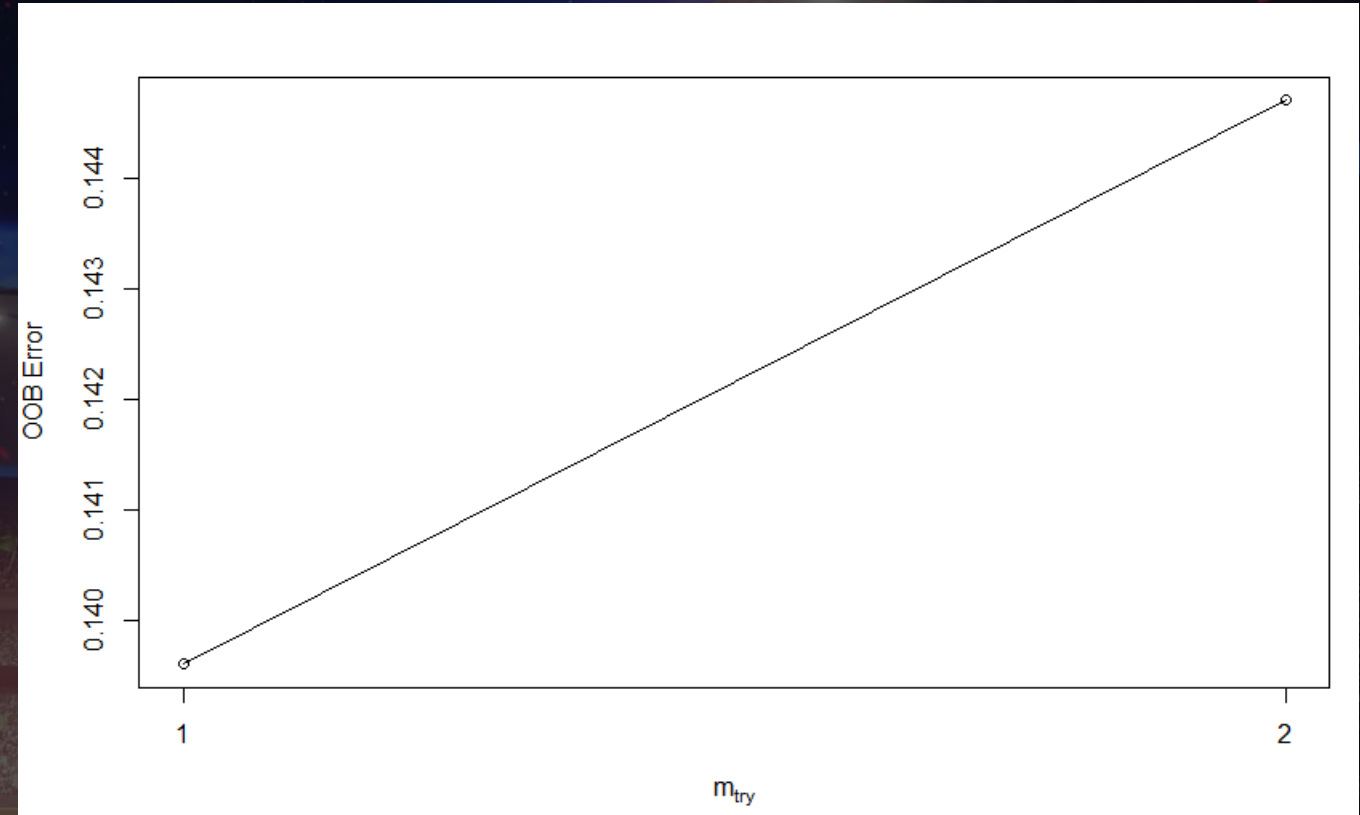

```
rf<-randomForest(result~elo.x+elo.y+date,data=train1)
```





Tune

```
tuneRF(train1[,c(3,9,11)],train1[,8],  
stepFactor = 0.5,  
plot = T,  
ntree=300,  
trace=T,  
improve = 0.05)
```





Result

```
> for (a in 1:14000) {  
+   if(p1[a,1]>0.66)  
+     p1[a,1]=1  
+   else if(p1[a,1]<0.33)  
+     p1[a,1]=0  
+   else  
+     p1[a,1]=0.5  
+ }  
> www=0  
> for(a in 1:14000)  
+   if(train1[a,8]==p1[a,1])  
+     www=www+1  
> www/14000  
[1] 0.829
```

```
> for (a in 1:3008) {  
+   if(p2[a,1]>0.66)  
+     p2[a,1]=1  
+   else if(p2[a,1]<0.33)  
+     p2[a,1]=0  
+   else  
+     p2[a,1]=0.5  
+ }  
> qqq=0  
> for(a in 1:3008)  
+   if(test1[a,8]==p2[a,1])  
+     qqq=qqq+1  
> qqq/3008  
[1] 0.7672872
```




解決問題



解決問題

入選八強之國家的航空公司
可以增開飛往俄羅斯的航班

俄羅斯飯店業者可以對巴西、
德國、西班牙、法國等強勁
隊伍國家祭出住宿優惠或是
包車接送服務



俄羅斯政府也可以增加
地鐵的西班牙文，葡萄牙
文等之指標與導覽

對於有購買運動彩券習慣的人
也可以參考此分析調整下注之
金額



課程建議



- 1.建議作業能改成分多次點，然後把每次的量減少
- 2.希望能把課開在下午，一早起來打程式腦袋有點不夠清醒
- 3.因為上課都實作，所以可以建議再多個助教應付同學



分析心得



這次的期末報告分析，剛好是我們幾個男生比較懂的體育，因此再做的時候感覺就比較有趣，此外讓我們更加懂得去處理比賽數據這種型態的資料，並繪製成圖形。

在Elo方面，讓我們知道了原來在R裡面，還有那麼方便的package可以用來直接分析比賽。

在Random Forest裡面，讓我們學習到如何調整相當不友善的因素，及慢慢的調整自己的模型。

最後，在世界盃比完兩輪後，我們發現我們的模型並沒有非常準，應該多考慮更多因素的是否會引響到，比較更加了解到要預測體育賽事是一件多麼困難的事！



參考資料



1.原始資料來源

<https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017/data>

2.資料異常新聞查詢

https://en.wikipedia.org/wiki/Australia_31%E2%80%9330_American_Samoa

3.Elo介紹

https://en.wikipedia.org/wiki/Elo_rating_system#Most_accurate_K-factor



分 工



蔡承翰

- 資料介紹
- 製作PPT
- 小組賽圖表繪製

張仁樵

- 圖表分析
- Elo模擬世界盃
- 製作PDF

蕭資峻

- 動機&假設
- 能解決的問題

田碩

- RandomForest
模型