CORRELATION IS CAUSATION

(NOT)

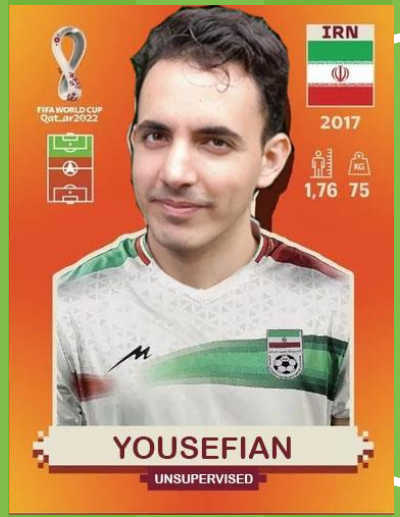
Global Football Diversity:

Exploring the Impact of Multinational Players in World Cups

PLAYERS

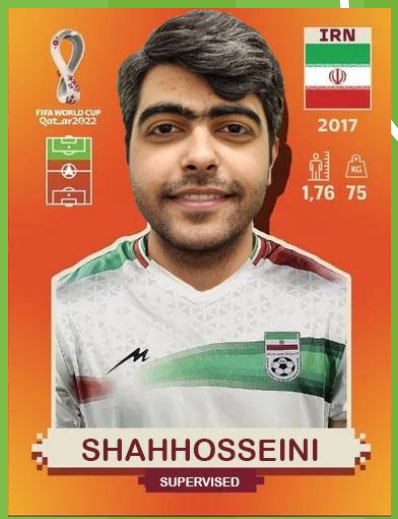
ARG
2005
1,70 72
BONIFACINI
UNSUPERVISED

IRN
2017
1,76 75
AMIRIAN
UNSUPERVISED


IRN
2017
1,76 75
YOUSEFIAN
UNSUPERVISED

ITA
2005
1,70 72
LUPO
SUPERVISED

IRN
2017
1,76 75
SABBAGH
SUPERVISED


IRN
2017
1,76 75
SHAHHOSSEINI
SUPERVISED

CONTENTS

1. Introduction
2. Datasets
3. Exploratory Data Analysis
4. Unsupervised results
5. Supervised results
6. Conclusion

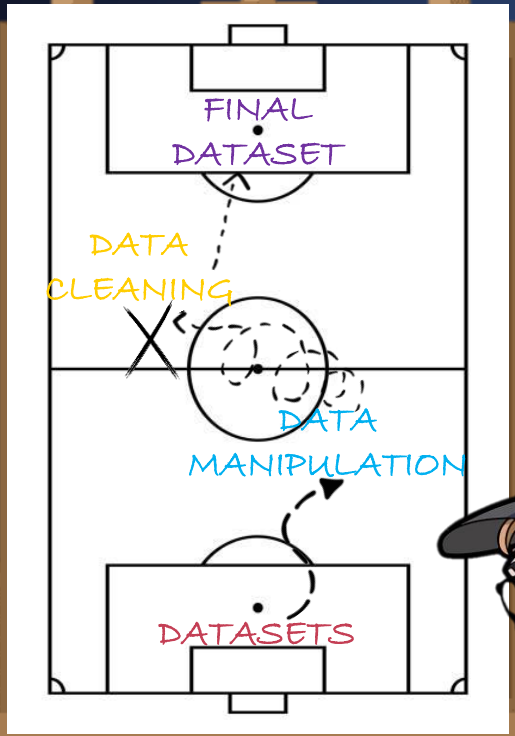
INTRODUCTION

- **Objective:** show that the inclusion of players with nationalities different from their country of origin affects the performance of national teams in the FIFA World Cup.

- **Question:** Does the presence of players who represent a country different from their country of origin influence the performance of national teams in the FIFA World Cup?

- **Methodology:**
  - Unsupervised learning
  - Supervised learning

DATASETS



name	country	year	foreign	concat
Ángel Fernando Bossio Garcilández (Ángel Bossio, "La mara...	Argentina	1930	0	Ángel Fernando Bossio Garcilández (Ángel Bossio, "La mara...
Juan Botasso	Argentina	1930	0	Juan Botasso Argentina
Roberto Eugenio "Cherro" Cerro (Roberto Cherro)	Argentina	1930	0	Roberto Eugenio "Cherro" Cerro (Roberto Cherro) Argentina
Alberto Chividini	Argentina	1930	0	Alberto Chividini Argentina
Attilio José Demaria ("Attilio Demaria")	Argentina	1930	0	Attilio José Demaria ("Attilio Demaria") Argentina
José Della Torre	Argentina	1930	0	José Della Torre Argentina
Juan Evaristo	Argentina	1930	0	Juan Evaristo Argentina
Marino "Mario" Evaristo	Argentina	1930	0	Marino "Mario" Evaristo Argentina
Manuel ("Nolo") Ferreira (Captain)	Argentina	1930	0	Manuel ("Nolo") Ferreira (Captain) Argentina
Luis Felipe Monti	Argentina	1930	1	Luis Felipe Monti Argentina
Ramón Alfredo Muttis	Argentina	1930	0	Ramón Alfredo Muttis Argentina
Rodolfo Orlandini	Argentina	1930	0	Rodolfo Orlandini Argentina
Fernando Paternoster	Argentina	1930	0	Fernando Paternoste
Natalio Perinetti	Argentina	1930	0	Natalio Perinetti Arg

Players

Rows: 10.137
Columns: 5

year	country	rank	total_points
1994	Algeria	42	37
1998	Algeria	54	40
2002	Algeria	77	473
2006	Algeria	72	511
2010	Algeria	61	548
2014	Algeria	32	749
2018	Algeria	43	818
2022	Algeria	38	1471
1994	Angola	110	12
1998	Angola	69	35
2002	Angola	56	529
2006	Angola	72	527
2010	Angola		
2014	Angola		

FIFA rank

Rows: 596
Columns: 5

home_team	away_team	home_score	away_score	Year	outcome
Argentina	France	3	3	2022	t
Croatia	Morocco	2	1	2022	w
France	Morocco	2	0	2022	w
Argentina	Croatia	3	0	2022	w
Morocco	Portugal	1	0	2022	w
England	France	1	2	2022	l
Croatia	Brazil	1	1	2022	t
Netherlands	Argentina	2	2	2022	t
Morocco	Spain	0	0	2022	t
Portugal	Switzerland	6	1	2022	w
Japan	Croatia	1	1	2022	t
Brazil	Korea Republic	4	1	2022	w
			1	2022	w
			0	2022	w

WC Matches

Rows: 964
Columns: 6

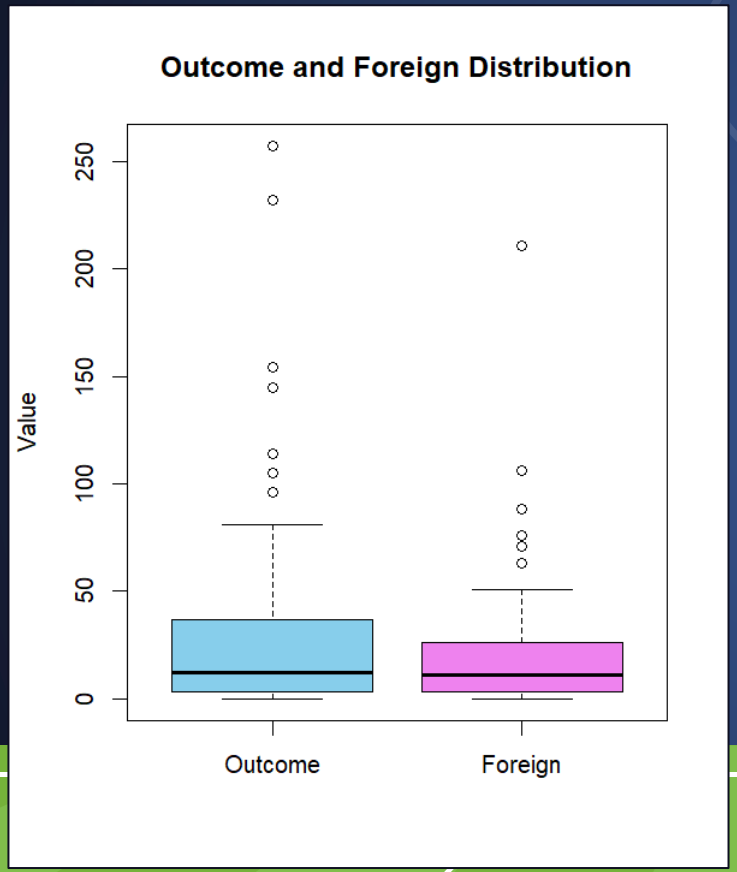
FINAL DATASET

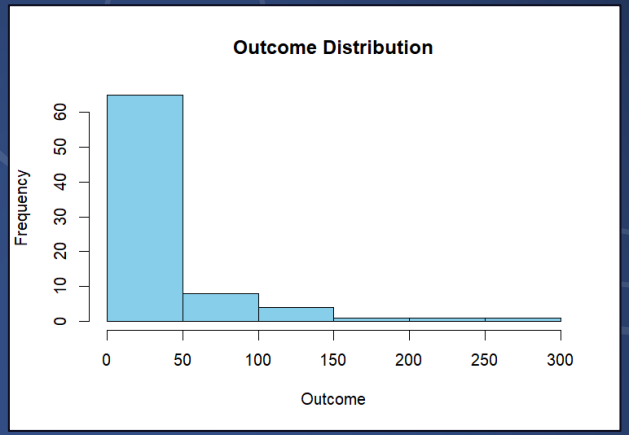
$$outcome = \sum(wins * 3) + (ties * 1) + (losses * 0)$$

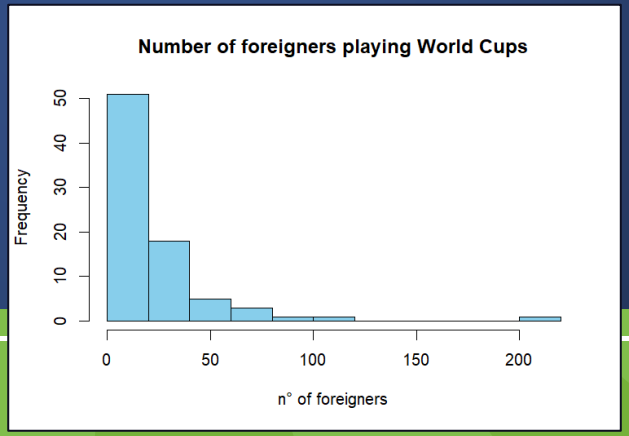
Year	home_team	away_team	home_score	away_score	outcome	home_foreign	home_rank	home_points	away_foreign	away
2010	Algeria	Slovenia	0	1	l	17	61	548	23	
2014	Algeria	Russia	1	1	t	16	32	749	22	
2006	Angola	Portugal	0	1	l	2	72	527	7	
1994	Argentina	Nigeria	2	1	w	1	9	55	1	
1994	Argentina	Bulgaria	0	2	l	1	9	55	0	
1994	Argentina	Greece	4	0	w	1	9	55	2	
1998	Argentina	Japan	1	0	w	2	11	58	1	
1998	Argentina	Croatia	1	0	w	2	11	58	22	
1998	Argentina	England	2	2	t	2	11	58	6	
1998	Argentina	Jamaica	5	0	w	2	11	58	8	
2002	Argentina	England	0	1	l	1	4	754	9	
2002	Argentina	Nigeria	1	0	w	1	4	754	1	
2006	Argentina	Mexico	2	1	w	3	4	856	3	

```
> str(matches_foreign)
'data.frame':   424 obs. of  12 variables:
 $ Year        : int  2010 2014 2006 1994 1994 1994 1998 1998 1998 1998 ...
 $ home_team   : chr  "Algeria" "Algeria" "Angola" "Argentina" ...
 $ away_team   : chr  "Slovenia" "Russia" "Portugal" "Nigeria" ...
 $ home_score  : int  0 1 0 2 0 4 1 1 2 5 ...
 $ away_score  : int  1 1 1 1 2 0 0 0 2 0 ...
 $ outcome     : chr  "l" "t" "l" "w" ...
 $ home_foreign: int  17 16 2 1 1 1 2 2 2 2 ...
 $ home_rank   : int  61 32 72 9 9 9 11 11 11 11 ...
 $ home_points : int  548 749 527 55 55 55 58 58 58 58 ...
 $ away_foreign: int  23 22 7 1 0 2 1 22 6 8 ...
 $ away_rank   : int  58 16 12 13 24 33 24 30 14 48 ...
 $ away_points : int  570 954 789 53 46 41 51 50 56 42 ...
```

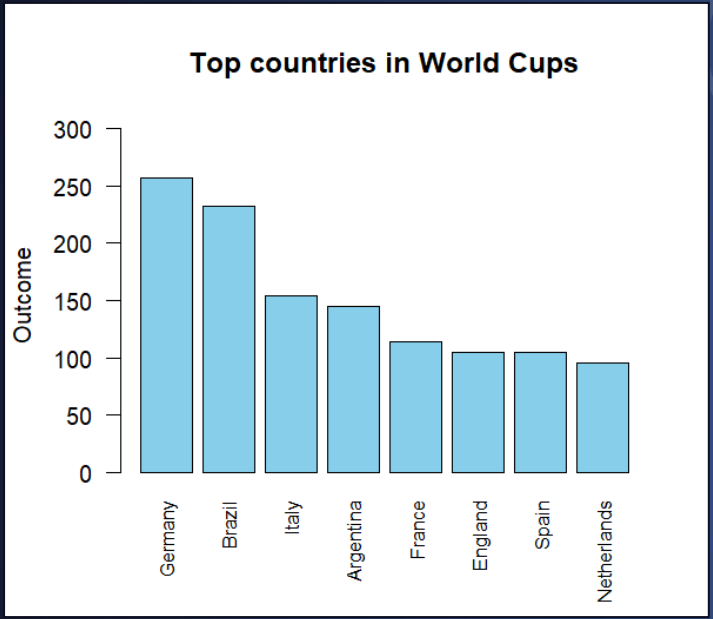
EDA – DISTRIBUTION

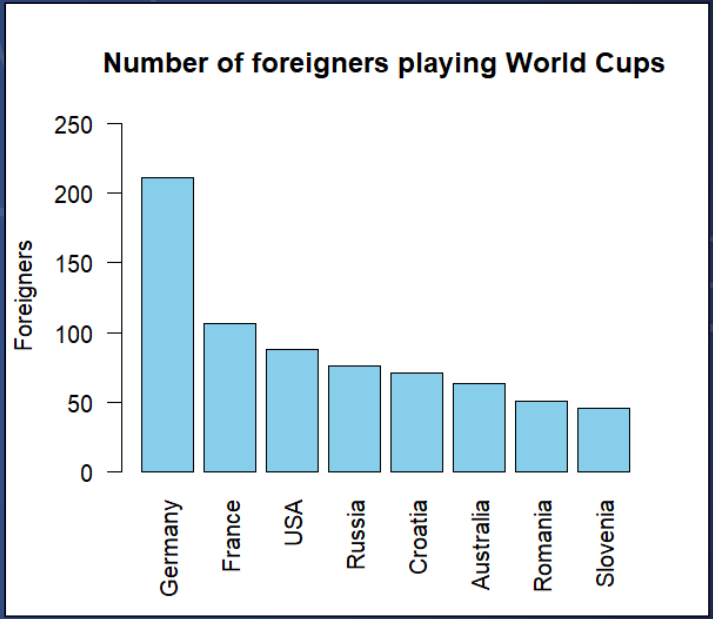






EDA – TOP 8

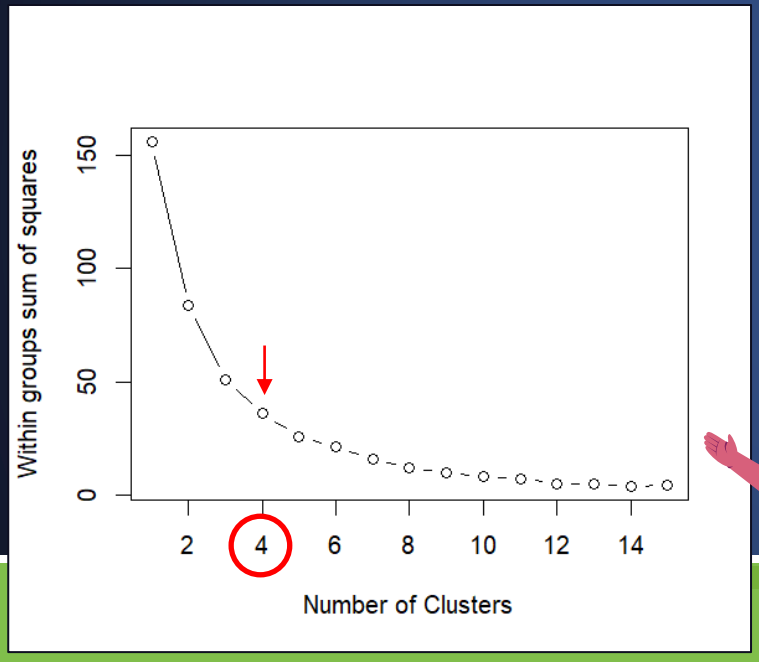

Top countries in World Cups
Outcome
300
250
200
150
100
50
0
Germany
Brazil
Italy
Argentina
France
England
Spain
Netherlands

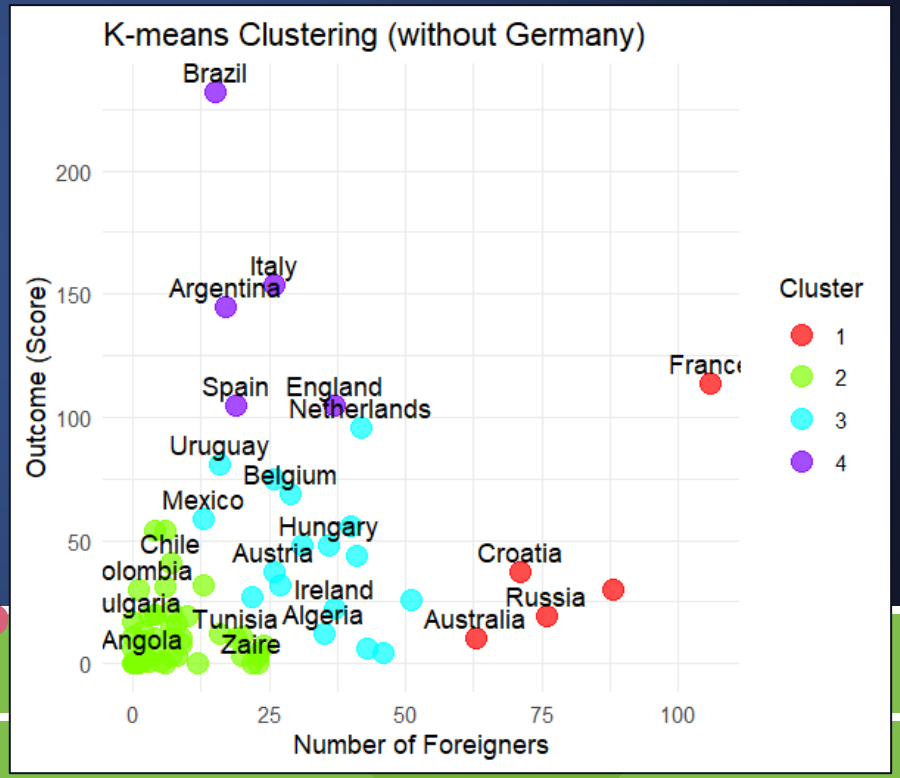

Number of foreigners playing World Cups
Foreigners
250
200
150
100
50
0
Germany
France
USA
Russia
Croatia
Australia
Romania
Slovenia

UNSUPERVISED

K-MEAN

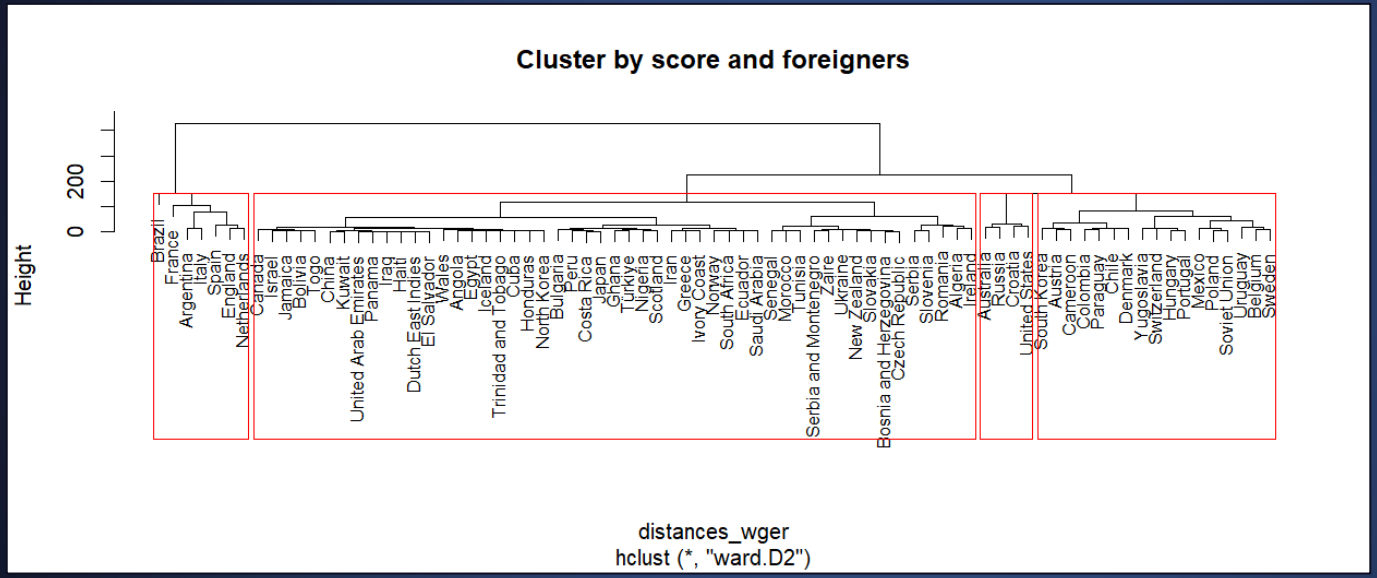
- **K = 4**
- Germany is an outlier in both variables, so we excluded

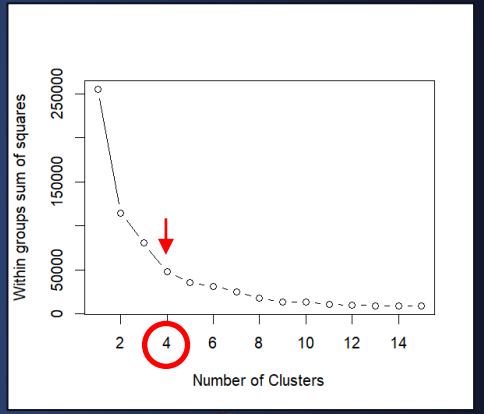




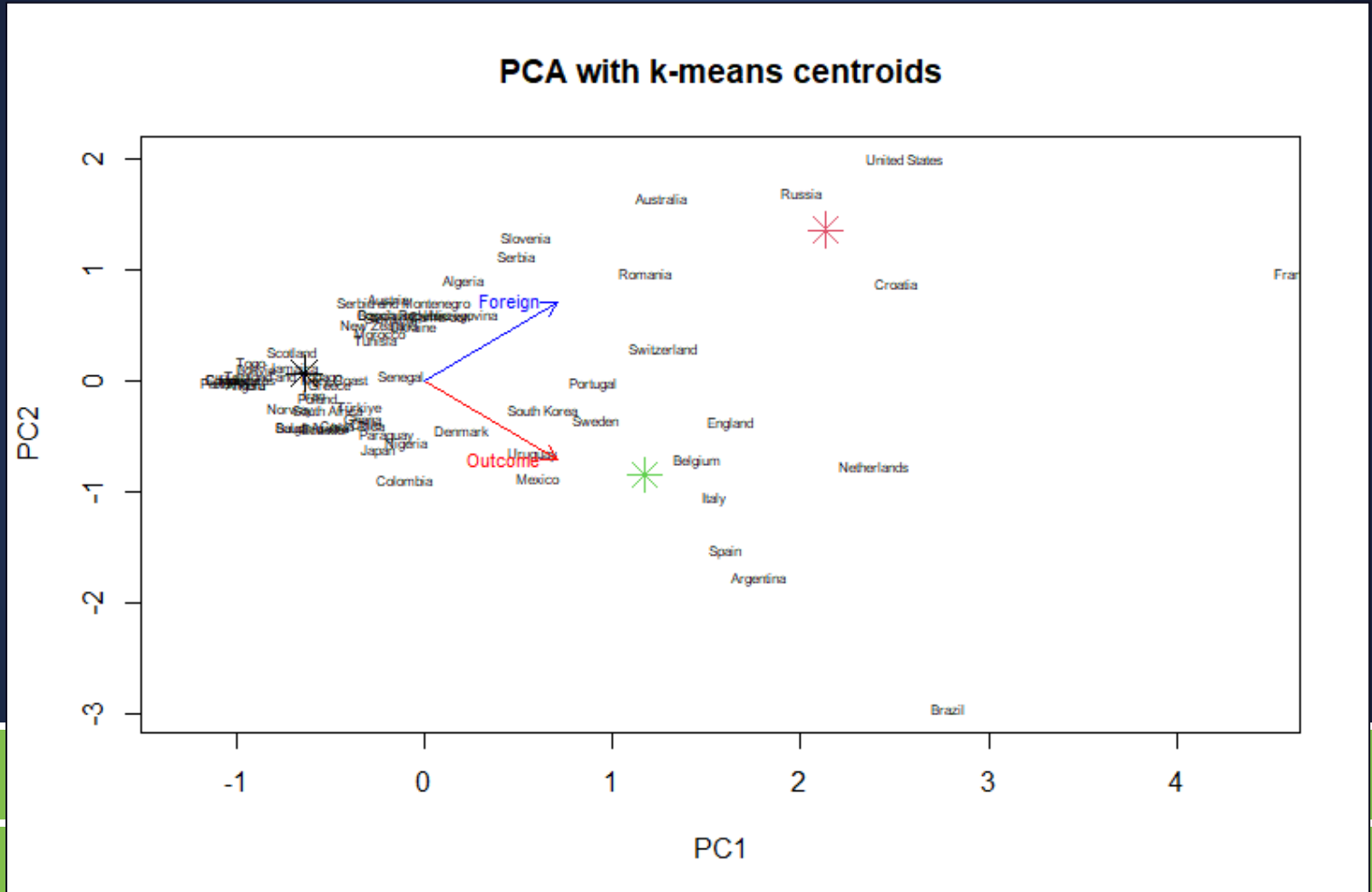
DENDOGRAM

- **K = 4**
- Germany is an outlier in both variables, so we excluded





PCA


PCA with k-means centroids
PC2
PC1
2
1
0
-1
-2
-3
-1
0
1
2
3
4
United States
Australia
Russia
Slovenia
Serbia
Romania
Croatia
Algeria
Foreign
Switzerland
Scotland
Senegal
Portugal
South Korea
Sweden
England
Denmark
Outcome
Belgium
Netherlands
Japan
Colombia
Mexico
Italy
Spain
Argentina
Brazil

SUPERVISED

Decision Tree
Random Forest
KNN

WITH TIES
WITHOUT TIES

FEATURE SELECTION

HOME_TEAM	AWAY_TEAM
HOME_SCORE	AWAY_SCORE
HOME_RANK	AWAY_RANK
HOME_POINTS	AWAY_POINTS
HOME_FOREIGN	AWAY_FOREIGN

FEATURE SELECTION

HOME_TEAM	AWAY_TEAM
HOME_SCORE	AWAY_SCORE
HOME_RANK	AWAY_RANK
HOME_POINTS	AWAY_POINTS
HOME_FOREIGN	AWAY_FOREIGN

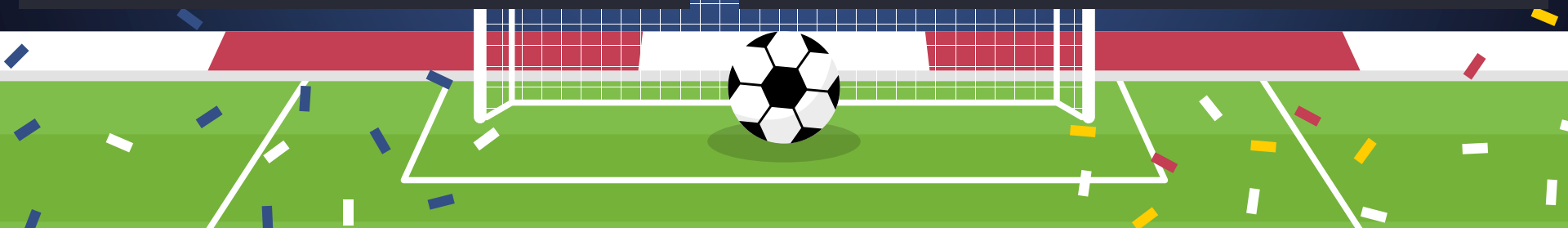
Decision Tree

No Foreign Players

```
Overall Statistics

               Accuracy : 0.4699
                 95% CI : (0.3593, 0.5826)
    No Information Rate : 0.4458
    P-Value [Acc > NIR] : 0.36900

                  Kappa : 0.1651

 Mcnemar's Test P-Value : 0.00757
```

Foreign Players

```
Overall Statistics

               Accuracy : 0.506
                 95% CI : (0.394, 0.6176)
    No Information Rate : 0.4458
    P-Value [Acc > NIR] : 0.16016

                  Kappa : 0.2145
```



No Foreign Players

Decision Tree	Random Forest
46.99%	39.76%

Foreign Players

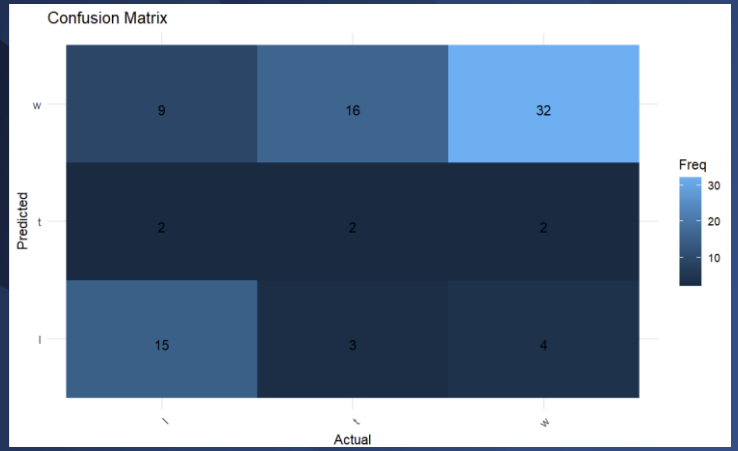
Decision Tree	Random Forest
50.60%	46.99%

WITHOUT TIES

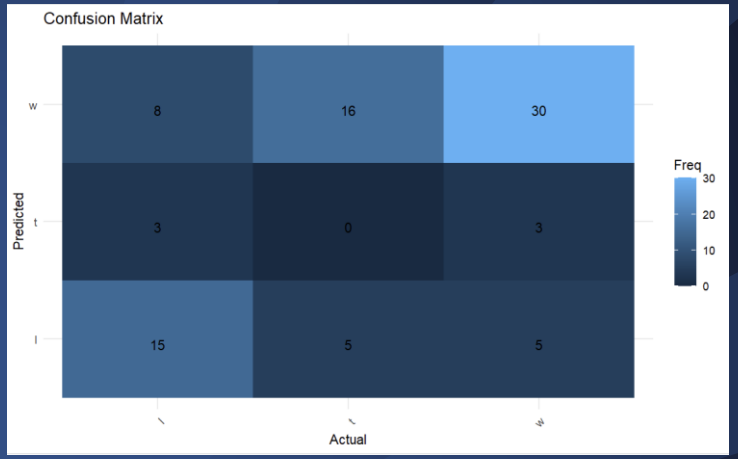
Decision Tree

```
Overall Statistics

               Accuracy : 0.6094
                 95% CI : (0.4793, 0.729)
    No Information Rate : 0.5781
    P-Value [Acc > NIR] : 0.3544

                  Kappa : 0.2188
```

KNN




```
Overall Statistics

               Accuracy : 0.5647
                 95% CI : (0.4528, 0.672)
    No Information Rate : 0.4471
    P-Value [Acc > NIR] : 0.019389

                  Kappa : 0.2768

 Mcnemar's Test P-Value : 0.002485
```




```
Overall Statistics

               Accuracy : 0.5412
                 95% CI : (0.4296, 0.6498)
    No Information Rate : 0.4471
    P-Value [Acc > NIR] : 0.05129

                  Kappa : 0.2469

 Mcnemar's Test P-Value : 0.02012
```

CONCLUSIONS

- Presence of foreign players might have an impact over World Cup matches

- Need of more complex features

- A new reality that is worth to be analysed

THANKS!

"Statistics are just a pattern of information that we have. There are players who make the team play good without the statistics. If you perform to your maximum, if you perform to your best and you helped your team-mates and make the process better, it is enough."

Pep Guardiola