DATA Analytics Case Competition 2017

Smith School

Information Systems

Praneesh Khanna, Sumit Bansal, Namwoo Park

Content

- Dataset and data exploration
- How to choose the first team players
- 14 players in each team – result
- How to decide the winner country
- How could we do better?

Dataset: variables

Name	Age	Nationality	Overall	Potential	Value	Acceleration	Aggression	Agility	Balance
L. Messi	30	Argentina	93	93	105000	92	48	90	95

Ball.control	Composure	Crossing	Curve	Dribbling	Finishing	Free.kick accuracy	GK.diving	GK.handling
95	96	77	[illegible]	[illegible]	[illegible]	[illegible]	6	11

GK.kicking	GK.positioning	GK.reflexes	Heading.accuracy	Interceptions	Jumping	Long.passing	Long.shots
15	14	[illegible]	[illegible]	[illegible]	[illegible]	87	88

Marking	Penalties	Positioning	Reactions	Short.passing	Shot.power	Sliding.tackle	Sprint.speed
13	74	93	95	88	85	26	87

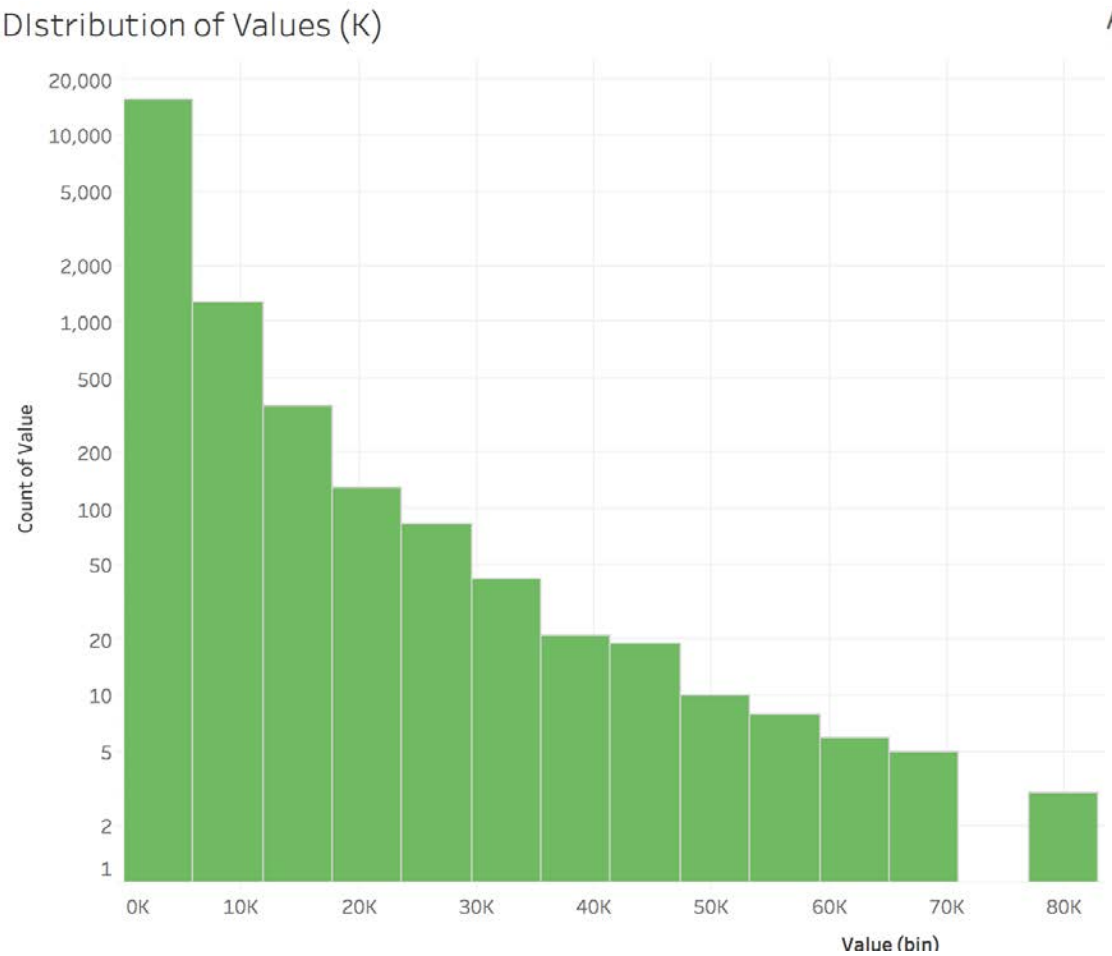
Stamina	Standing	Brazil	Spain	Argentina	Portugal	CM	LAM	LB	LCB
73	28	Enlgland	Belgium	France	Germany	84	92	57	45

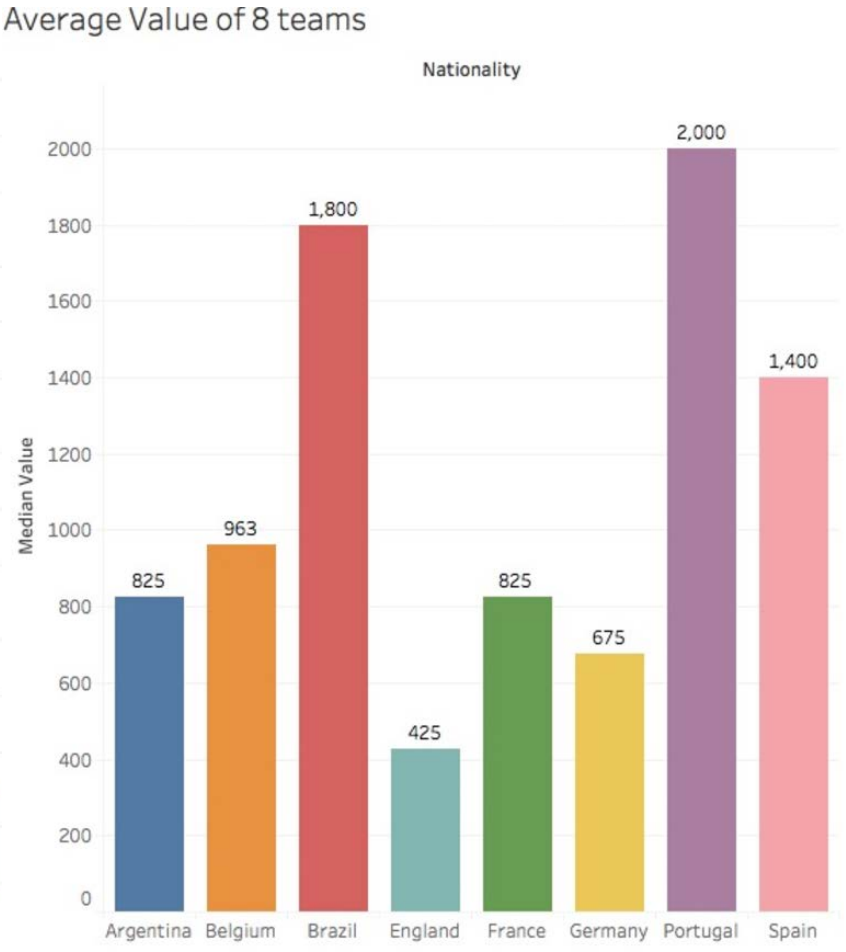
LCM	LDM	LF	LM	LS	LW	LWB	Preferred.Positions	RAM	RB	RCB	RCM	RDM	RF	RM	RS	RW	RWB	ST
84	59	92	90	88	91	62	RW	92	57	45	84	59	92	90	88	91	62	88

75 Variables

17,981 rows

Data exploration





Dataset: 4 categories

1. Basic Information

Name	L. Messi
Age	30
Nationality	Argentina
Club	FC Barcelona
Preferred Position	RW

2. Value Information

Overall	93
Potential	93
Value	105000K
Wage	565K

3. Position Ratings

CAM	92
CB	45
CDM	59
CF	93
CM	84
LAM	92
LB	57
LCB	45
LCM	84
LDM	59
LF	92
LM	90
LS	88
LW	91
LWB	62
RAM	92
RB	57
RCB	45
RCM	84
RDM	59
RF	92
RW	91
RWB	62
ST	88

RW 91

4. Skill Ratings

Acceleration	92
Aggression	48
Agility	90
Balance	95
Ball.control	95
Composure	96
Crossing	77
Curve	89
Dribbling	97
Finishing	95
Free.kick.accuracy	90
GK.diving	6
GK.handling	11
GK.kicking	15
GK.positioning	14
GK.reflexes	8
Heading.accuracy	71
Interceptions	22
Jumping	68
Long.passing	87
Long.shots	88
Marking	13
Penalties	74
Positioning	93
Reactions	95
Short.passing	88
Shot.power	85
Sliding.tackle	26
Sprint.speed	87
Stamina	73
Standing.tackle	28
Strength	59
Vision	90
Volleys	85

Dataset: 4 categories

1. Basic Information

Name	L. Messi
Age	30
Nationality	Argentina
Club	FC Barcelona
Preferred Position	RW

2. Value Information

Overall	93
Potential	93
Value	105000K
Wage	565K

3. Position Ratings

CAM	92
CB	45
CDM	59
CF	92
CM	84
LAM	92
LB	57
LCB	45
LCM	84
LDM	59
LF	92
LM	90
LS	88
LW	91
LWB	62
RAM	92
RB	57
RCB	45
RCM	84
RDM	59
RF	92
RM	90
RS	88
RW	91
RWB	62
ST	88

14 positions

4. Skill Ratings

Acceleration	92
Aggression	48
Agility	90
Balance	95
Ball.control	95
Composure	96
Crossing	77
Curve	89
Dribbling	97
Finishing	95
Free.kick.accuracy	90
GK.diving	6
GK.handling	11
GK.kicking	15
GK.positioning	14
GK.reflexes	8
Heading.accuracy	71
Interceptions	22
Jumping	68
Long.passing	87
Long.shots	88
Marking	13
Penalties	74
Positioning	93
Reactions	95
Short.passing	88
Shot.power	85
Sliding.tackle	26
Sprint.speed	87
Stamina	73
Standing.tackle	28
Strength	59
Vision	90
Volleys	85

How to choose 14 players: Regression

How to choose 14 players: Stepwise Regression

LB RB CB CDM LWB RWB CM LM RM LW RW CAM CF ST

Dependent variable

Value	105000K

~

Acceleration	92
Aggression	48
Agility	90
Balance	95
Ball.control	95
Composure	96
Crossing	77
Curve	89
Dribbling	97
Finishing	95
Free.kick.accuracy	90
GK.diving	6
GK.handling	11
GK.kicking	15
GK.positioning	14
GK.reflexes	8
Heading.accuracy	71
Interceptions	22
Jumping	68
Long.passing	87
Long.shots	88
Marking	13
Penalties	74
Positioning	93
Reactions	95
Short.passing	88
Shot.power	85

34 Independent variable

How to choose 14 players: Skill Selection

CF

- Finishing
- Reaction
- Ball Control
- Short Passing
- Vision

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -29128.712   3203.586  -9.093  < 2e-16 ***
Age             -367.095     73.192  -5.015 8.92e-07 ***
Special          -39.534      8.937  -4.424 1.34e-05 ***
Aggression        71.720     22.981   3.121 0.001973 **
Ball.control     213.270     73.184   2.914 0.003826 **
Curve             54.499     32.908   1.656 0.098713 .
Dribbling        124.038     71.003   1.747 0.081638 .
Finishing        307.699     45.406   6.777 6.20e-11 ***
Interceptions     81.418     32.296   2.521 0.012203 *
Jumping           53.202     25.319   2.101 0.036425 *
Long.shots        78.460     44.715   1.755 0.080301 .
Reactions        224.180     45.141   4.966 1.13e-06 ***
Short.passing    194.867     53.309   3.655 0.000302 ***
Sliding.tackle   112.422     37.158   3.025 0.002690 **
Sprint.speed      76.824     38.699   1.985 0.048008 *
Stamina           76.454     30.403   2.515 0.012421 *
Vision           117.139     43.175   2.713 0.007038 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4172 on 310 degrees of freedom
Multiple R-squared:  0.617,     Adjusted R-squared:  0.5972
F-statistic: 31.21 on 16 and 310 DF,  p-value: < 2.2e-16
```

How to choose 14 players: Skill Selection

LB

- Reaction
- Standing Tackle
- Sliding Tackle
- Interception
- Heading Accuracy

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -16812.408    709.439 -23.698  < 2e-16 ***
Age               -245.209     15.083 -16.258  < 2e-16 ***
Balance              8.756      6.105   1.434 0.151689
Composure           25.013      8.255   3.030 0.002479 **
Finishing           31.637      5.814   5.442 5.95e-08 ***
Heading.accuracy    35.936      7.862   4.571 5.17e-06 ***
Interceptions       48.162     12.189   3.951 8.05e-05 ***
Reactions           89.732     11.399   7.872 5.78e-15 ***
Sliding.tackle      53.219     16.539   3.218 0.001313 **
Sprint.speed        11.452      7.056   1.623 0.104760
Stamina             23.724      6.998   3.390 0.000713 ***
Standing.tackle     61.226     17.729   3.453 0.000565 ***
Vision              16.012      7.042   2.274 0.023085 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2338 on 1936 degrees of freedom
Multiple R-squared:  0.4858,    Adjusted R-squared:  0.4826
F-statistic: 152.4 on 12 and 1936 DF,  p-value: < 2.2e-16
```

How to choose 14 players: Skill Criteria Table

	cam_mid_attack	cb_mid_defend	cdm_mid_defend	cf_attack	cm_mid_mid	lb_left_defend	lm_left_mid
1	Short.passing	Sliding.tackle	Short.passing	Finishing	Short.passing	Reactions	Reactions
2	Reactions	Standing.tackle	Reactions	Ball.control	Reactions	Standing.tackle	Short.passing
3	Ball.control	Reactions	Standing.tackle	Dribbling	Ball.control	Sliding.tackle	Ball.control
4	Finishing	Interception	Sliding.tackle	Reactions	Dribbling	Interception	Crossing
5	Vision	Vision	Heading.accuracy	Short.passing	Ball.control	Heading.accuracy	Dribbing

lw_left_attack	lwb_left_defend	rb_right_defend	rm_right_md	rw_right_attack	rwb_right_defend	st_attack
Reactions	Reactions	Reactions	Reactions	Reactions	Reactions	Finishing
Ball.control	Interceptions	Sliding.tackle	Short.passing	Ball.control	Short.passing	Reactions
Composure	Crossing	Standing.tackle	Ball.control	Acceleration	Heading.accuracy	Vision
Special	Driblling	Interceptions	Dribbling	Short.passing	Dribbling	Composure
Agility	Ball.control	Short.passing	Crossing	Free.kick.accuracy	Strength	Positioning

How to choose 14 players: New Rating

CF D. Mertens

CF
92

*

Finishing	85
Reactions	85
Ball Control	88
Short.passing	81
Vision	82

Sum

= New Rating 38,732

How to choose 14 players: New tables

165 Countries

14 positions

2,310 Tables

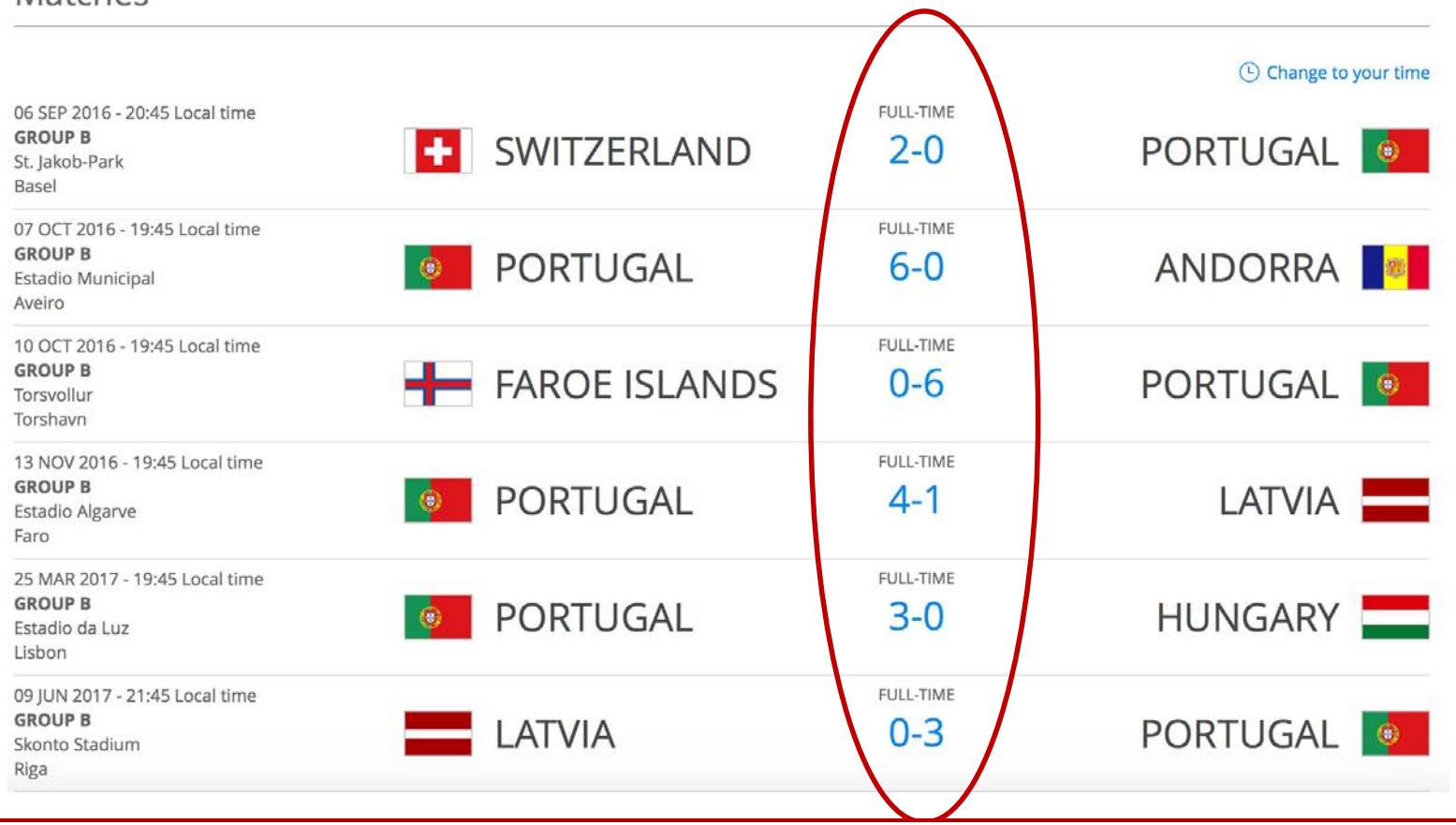
	Nationality	Name	CF	Preferred.Positions	Rating
	<chr>	<chr>	<int>	<chr>	<dbl>
1	Argentina	L. Messi	92	RW	43240
2	Argentina	S. Agüero	87	ST	37932
3	Argentina	P. Dybala	86	ST CAM	37582
4	Argentina	G. Higuaín	84	ST	35532
5	Argentina	A. Gómez	83	ST LW CF	34777
6	Argentina	J. Pastore	81	RW LW CM	33858
7	Argentina	A. Di María	82	LW RW	33456
8	Argentina	L. Vietto	80	ST	33120
9	Argentina	I. Piatti	80	CF LM LW	32720
10	Argentina	N. Gaitán	80	LM	32560
11	Argentina	D. Perotti	80	CF LW	32080
12	Argentina	A. Correa	80	RW RM CAM ST	32080
13	Argentina	E. Lamela	79	LM RM	31521
14	Argentina	M. Icardi	79	ST	31284
15	Argentina	R. Palacio	79	ST	31284
16	Argentina	E. Banega	78	CM CAM	31122
17	Argentina	M. Zárate	78	LM CAM CF ST	31044
18	Argentina	M. Lanzini	78	LM CAM	30888
19	Argentina	P. Piatti	78	LW RM LM	30420
20	Argentina	A. Fernández	77	CM	30107

How to choose 14 players: Selected Players

	Argentina	Belgium	Brazil	England	France	Germany	Portugal	Spain
Total Ratings								
ST	S. Agüero	R. Lukaku	Jonas	H. Kane	K. Benzema	T. Müller	Cristiano Ronaldo	Diego Costa
CF	A. Gómez	D. Mertens	Raffael	D. Alli	A. Lacazette	M. Kruse	Nani	Iago Aspas
CAM	P. Dybala	K. De Bruyne	Coutinho	R. Barkley	D. Payet	M. Özil	João Mário	Isco
LW	A. Di MarÃ-a	E. Hazard	Neymar	R. Sterling	A. Griezmann	M. Reus	Quaresma	Pedro
RW	L. Messi	T. Bongonda	Willian	R. Sterling	O. Dembélé	L. Sané	João Cancelo	Deulofeu
CM	E. Banega	R. Nainggolan	Rafinha	A. Lallana	P. Pogba	T. Kroos	Pizzi	Iniesta
LM	N. Gaitán	Y. Carrasco	Alex Sandro	Oxlade-Chamberlain	F. Ribéry	J. Draxler	R. Guerreiro	David Silva
RM	J. Pastore	T. Hazard	Marlos	R. Barkley	F. Thauvin	G. Castro	Bernardo Silva	Isco
CDM	J. Mascherano	M. Fellaini	Casemiro	E. Dier	B. Matuidi	S. Khedira	Danilo Pereira	Javi Martínez
LWB	C. Ansaldi	J. Lukaku	Filipe Luís	D. Rose	L. Kurzawa	J. Hector	Ricardo Pereira	Jordi Alba
RWB	P. Zabaleta	T. Meunier	Dani Alves	K. Walker	D. Sidibé	J. Kimmich	João Cancelo	Azpilicueta
CB	M. Musacchio	T. Alderweireld	Thiago Silva	G. Cahill	A. Laporte	M. Hummels	Pepe	Sergio Ramos
LB	M. Rojo	J. Lukaku	Marcelo	L. Baines	P. Evra	J. Hector	Eliseu	Jordi Alba
RB	P. Zabaleta	J. Vertonghen	Marquinhos	N. Clyne	S. Umtiti	J. Boateng	Nélson Semedo	Carvajal

How to decide winner country

FIFA homepage: Match Result



How to decide winner country

Countries		Rating		Goal	
France	Russia	477759	420673	1	1
Argentina	brazil	480948	489647	1	1
Argentina	columbia	480948	410064	1	0
Argentina	uruguay	480948	414959	1	0
Brazil	uruguay	489647	414959	2	2
Germany	Norway	487086	381803	3	0
Spain	Italy	492540	469940	1	1
⋮		⋮		⋮	

Difference of Ratings

Difference of Goals

How to decide winner country: Regression

Dependent variable

diff_Goal
0
1
2
0
1
0
3
0
0
3
-2
5
-1
-3
1
0
1

~

30 rows

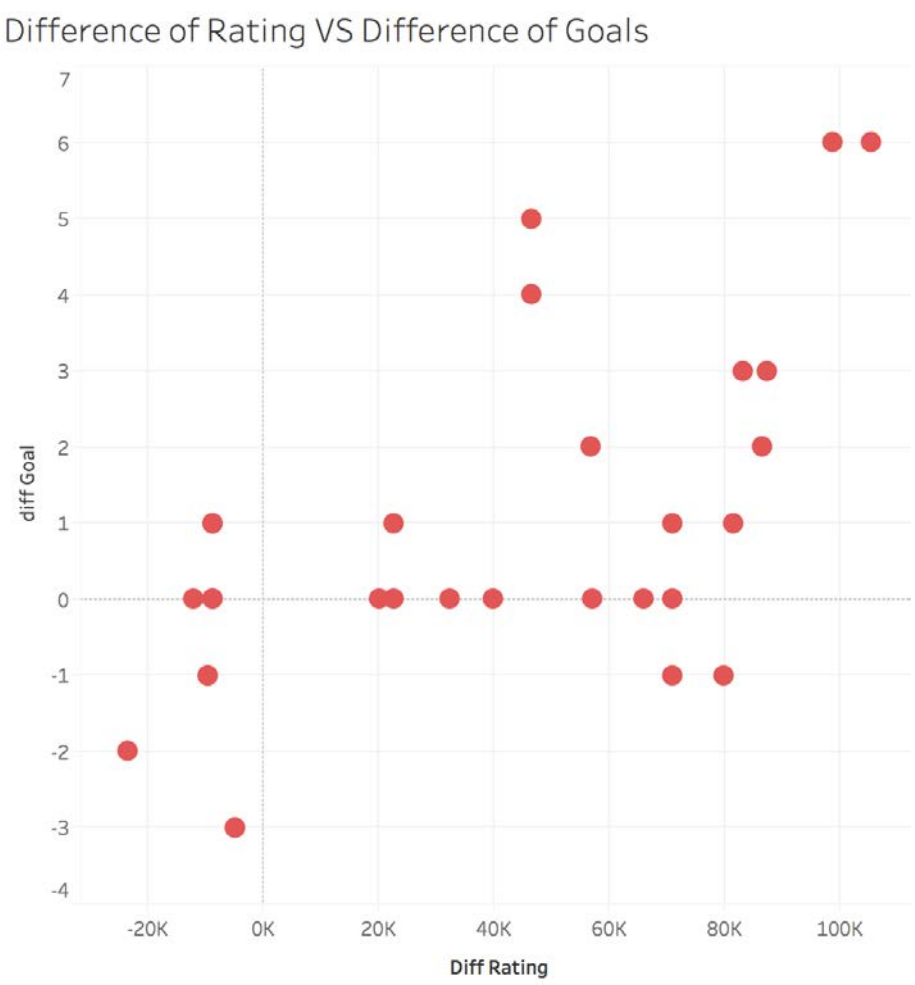
diff_rating
57086
81591
86542
-8699
70884
65989
87402
39845
22600
79901
-23292
46620
-9428
-4895
32377
-12075
20302

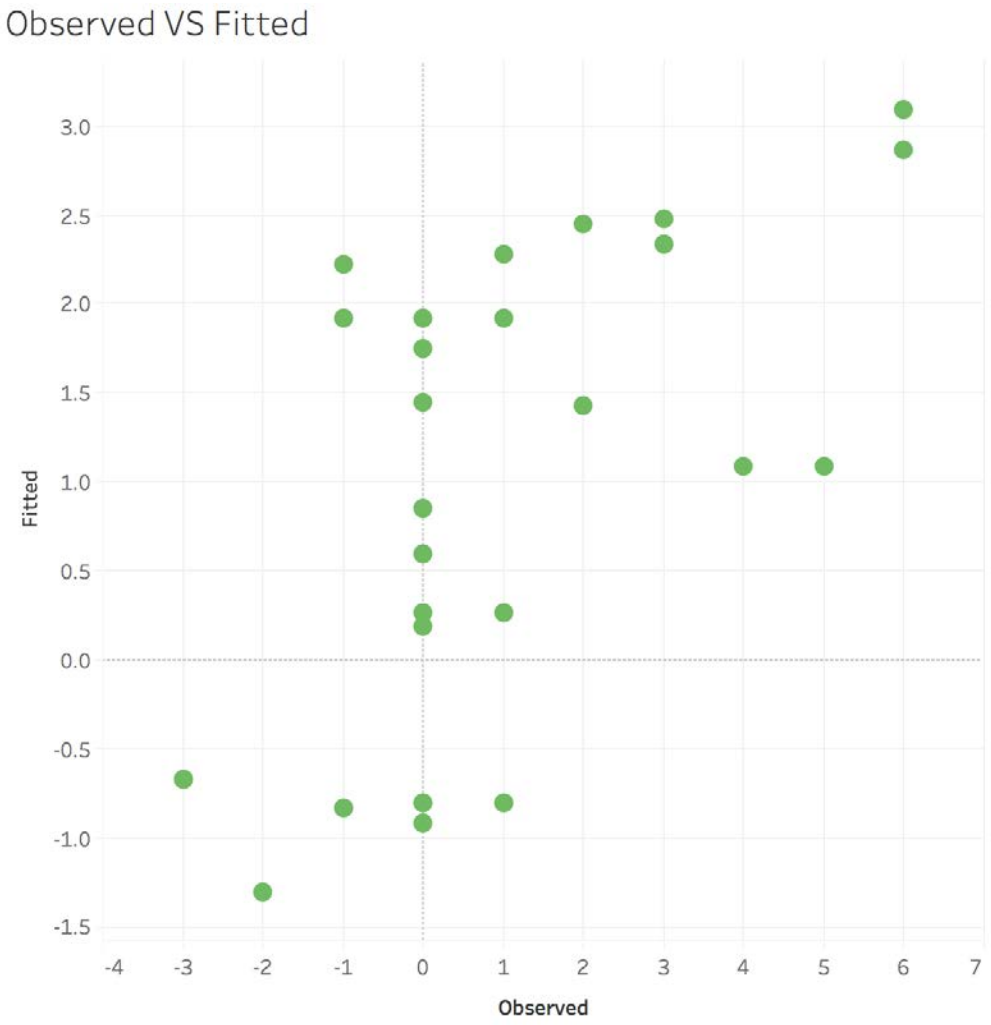
Independent variable

How to decide winner country: regression result

```
Call:
lm(formula = diff_Goal ~ diff_rating, data = reg)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2231 -1.1894 -0.1792  0.8038  3.9139

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.066e-01  4.668e-01  -1.085 0.287091
diff_rating  3.416e-05  8.454e-06   4.041 0.000376 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.844 on 28 degrees of freedom
Multiple R-squared:  0.3684,    Adjusted R-squared:  0.3458
F-statistic: 16.33 on 1 and 28 DF,  p-value: 0.0003762
```

How to decide winner country: Scatter plot





How to decide winner country: Result

	Countries		Diff of Ratings	Diff of Goals
Quarterfinal	Brazil	England	29173	0.48995859
	Spain	Belgium	17361	0.08646067
	Argentina	France	3189	-0.39765485
	Portugal	Germany	-19793	-1.18271997

Germany is likely to win

	Countries		Diff of Ratings	Diff of Goals
Semifinal	Brazil	Spain	-2893	-0.60541597
	France	Germany	-9327	-0.82520141

	Countries		Diff of Ratings	Diff of Goals
Final	Spain	Germany	5454	-0.32028245

How could we do better?

Difference of Goals ~

- Performance (Ratings)
- Home ground
- Recent Condition
- Country factor
- ...

Need factors that significantly impact the result of the soccer match

END

Thank you!