Using PCA to See Which Countries have Better Players for World Cup Games

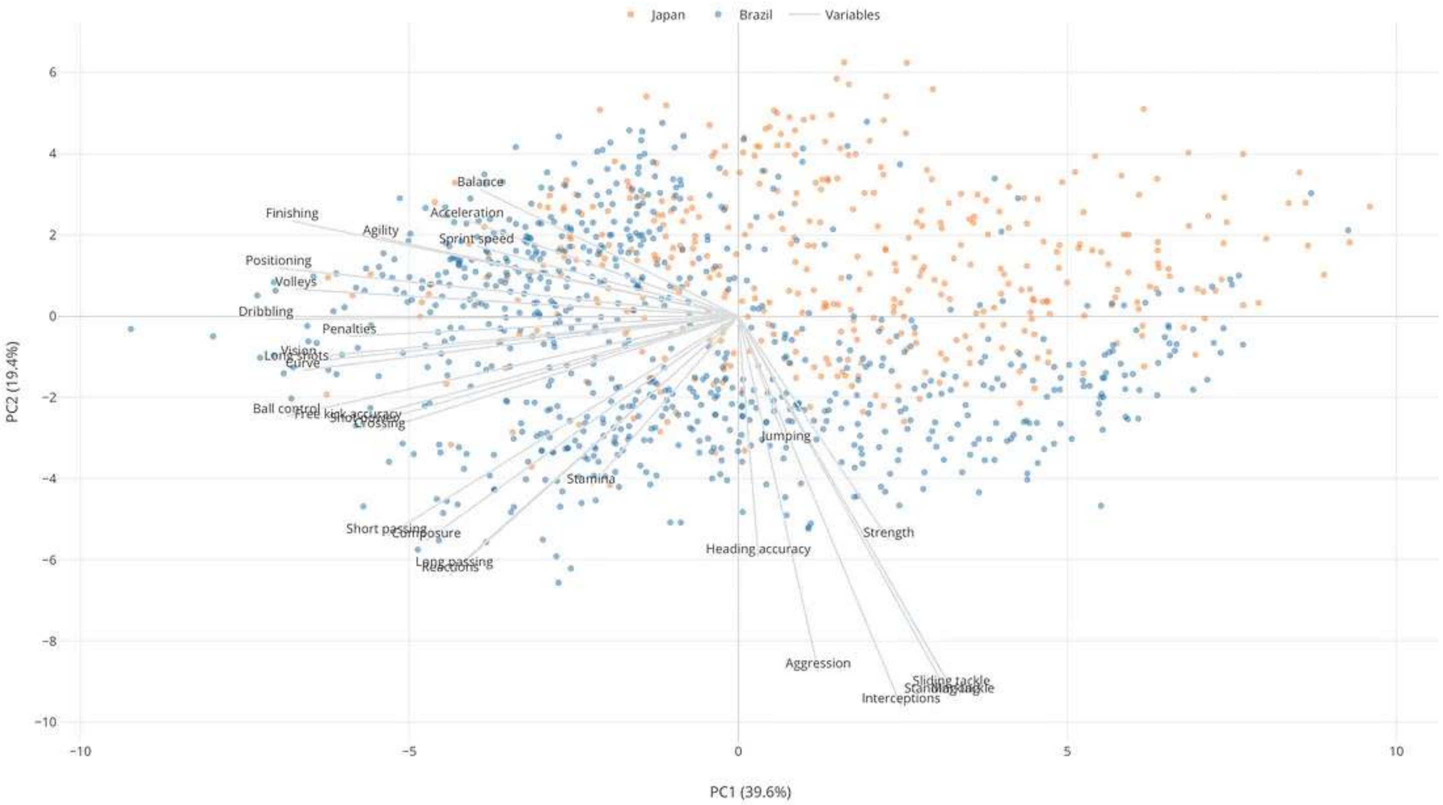
A practical use case of Principal Component Analysis (PCA) algorithm

Kan Nishida · Follow

Published in learn data science

11 min read · Jul 14, 2018

Listen Share More



This is the second post of "An Introduction to PCA (Primary Component Analysis)" series. You can directly start from this post, but if you are not familiar with PCA I'd suggest you take a look at the introduction post first.

An Introduction to Principal Component Analysis (PCA) with 2018 World Soccer Players Data

What is PCA and Why is it Useful in Real World?

blog.exploratory.io

Today, I'm going to use the same PCA algorithm, which I used in the to reduce the original dimensionality of the soccer player skill measures down to just two newly created dimensions (or components) so that I can place the original soccer player measures and the players from a given pair of two countries together on the two dimensional space (X axis and Y axis). I'm hoping that this will help us understand how the players from the two countries perform on those skill measures

to reduce the dimensionality of these variables so that I can have just two artificially created new dimensions, which should carry as much information as the original data has, to visualize how the soccer players from a given pair of two countries perform on the skill measures and compare the two countries.

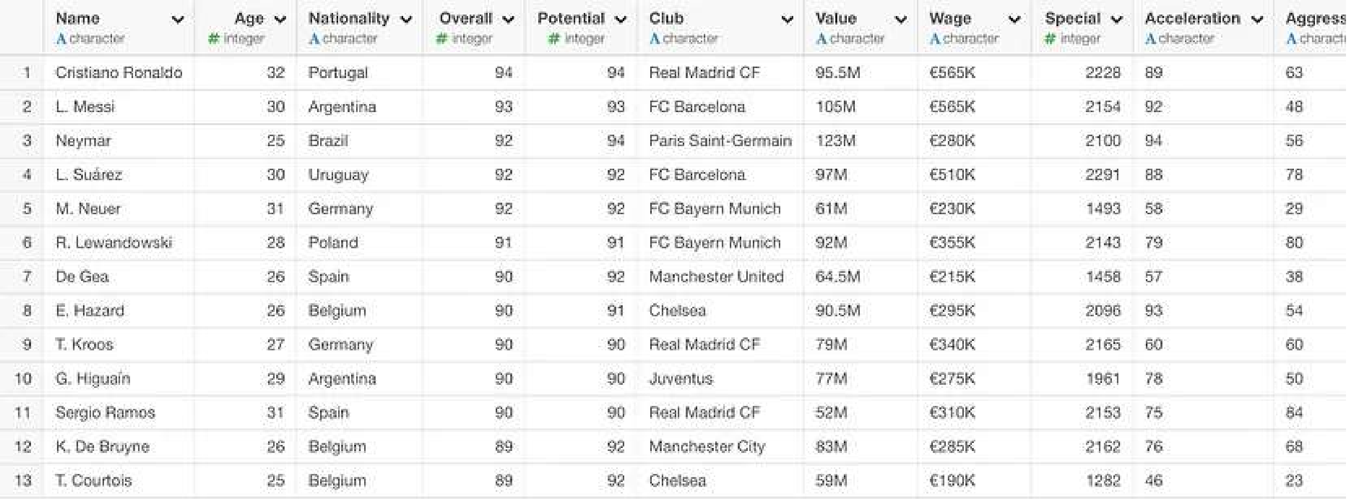
For example, how the players from Brazil and Japan are different in terms of how they score on the skill measures? Can we tell if Brazil is stronger than Japan by looking at the information the PCA algorithm generates?

Let's take a look.

Data

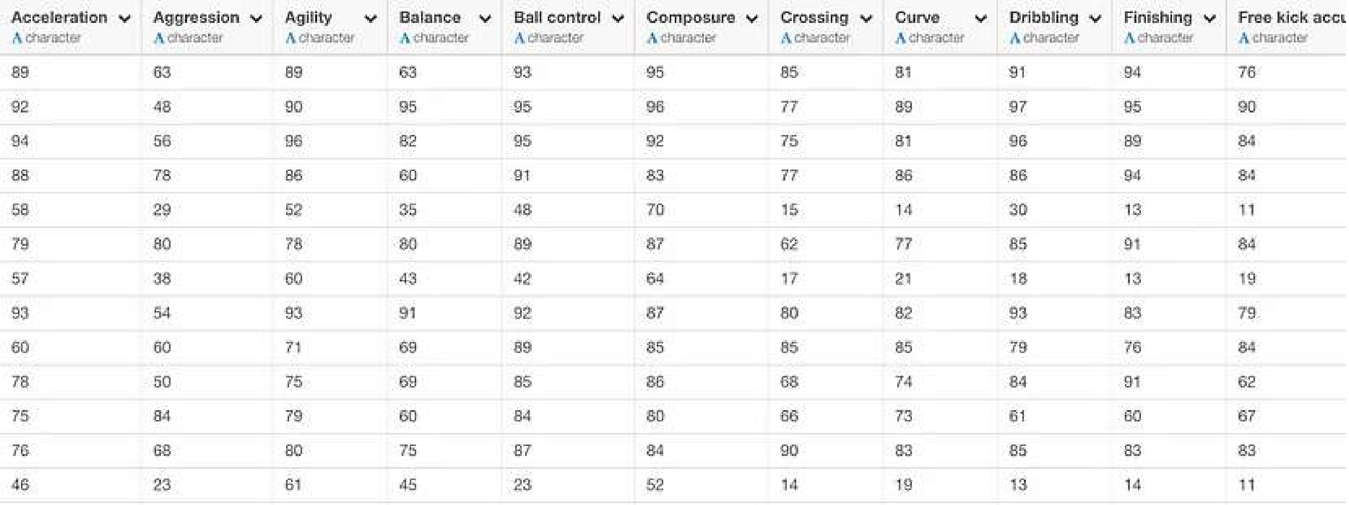
Just like the previous post, I'm going to use the same 2018 FIFA Soccer Player data that has the following information.

- Player personal data like Nationality, Photo, Club, Age, Wage, Salary etc.
- **Player skill measures** such as Dribbling, Aggression, GK Skills etc.
- Playing position related data.



	Name A character	Age # integer	Nationality A character	Overall # integer	Potential # integer	Club A character	Value A character	Wage A character	Special # integer	Acceleration A character	Aggress A characte
1	Cristiano Ronaldo	32	Portugal	94	94	Real Madrid CF	95.5M	€565K	2228	89	63
2	L. Messi	30	Argentina	93	93	FC Barcelona	105M	€565K	2154	92	48
3	Neymar	25	Brazil	92	94	Paris Saint-Germain	123M	€280K	2100	94	56
4	L. Suárez	30	Uruguay	92	92	FC Barcelona	97M	€510K	2291	88	78
5	M. Neuer	31	Germany	92	92	FC Bayern Munich	61M	€230K	1493	58	29
6	R. Lewandowski	28	Poland	91	91	FC Bayern Munich	92M	€355K	2143	79	80
7	De Gea	26	Spain	90	92	Manchester United	64.5M	€215K	1458	57	38
8	E. Hazard	26	Belgium	90	91	Chelsea	90.5M	€295K	2096	93	54
9	T. Kroos	27	Germany	90	90	Real Madrid CF	79M	€340K	2165	60	60
10	G. Higuaín	29	Argentina	90	90	Juventus	77M	€275K	1961	78	50
11	Sergio Ramos	31	Spain	90	90	Real Madrid CF	52M	€310K	2153	75	84
12	K. De Bruyne	26	Belgium	89	92	Manchester City	83M	€285K	2162	76	68
13	T. Courtois	25	Belgium	89	92	Chelsea	59M	€190K	1282	46	23

There are 34 player skill related measures (variables).



Acceleration A character	Aggression A character	Agility A character	Balance A character	Ball control A character	Composure A character	Crossing A character	Curve A character	Dribbling A character	Finishing A character	Free kick accu A character
89	63	89	63	93	95	85	81	91	94	76
92	48	90	95	95	96	77	89	97	95	90
94	56	96	82	95	92	75	81	96	89	84
88	78	86	60	91	83	77	86	86	94	84
58	29	52	35	48	70	15	14	30	13	11
79	80	78	80	89	87	62	77	85	91	84
57	38	60	43	42	64	17	21	18	13	19
93	54	93	91	92	87	80	82	93	83	79
60	60	71	69	89	85	85	85	79	76	84
78	50	75	69	85	86	68	74	84	91	62
75	84	79	60	84	80	66	73	61	60	67
76	68	80	75	87	84	90	83	85	83	83
46	23	61	45	23	52	14	19	13	14	11

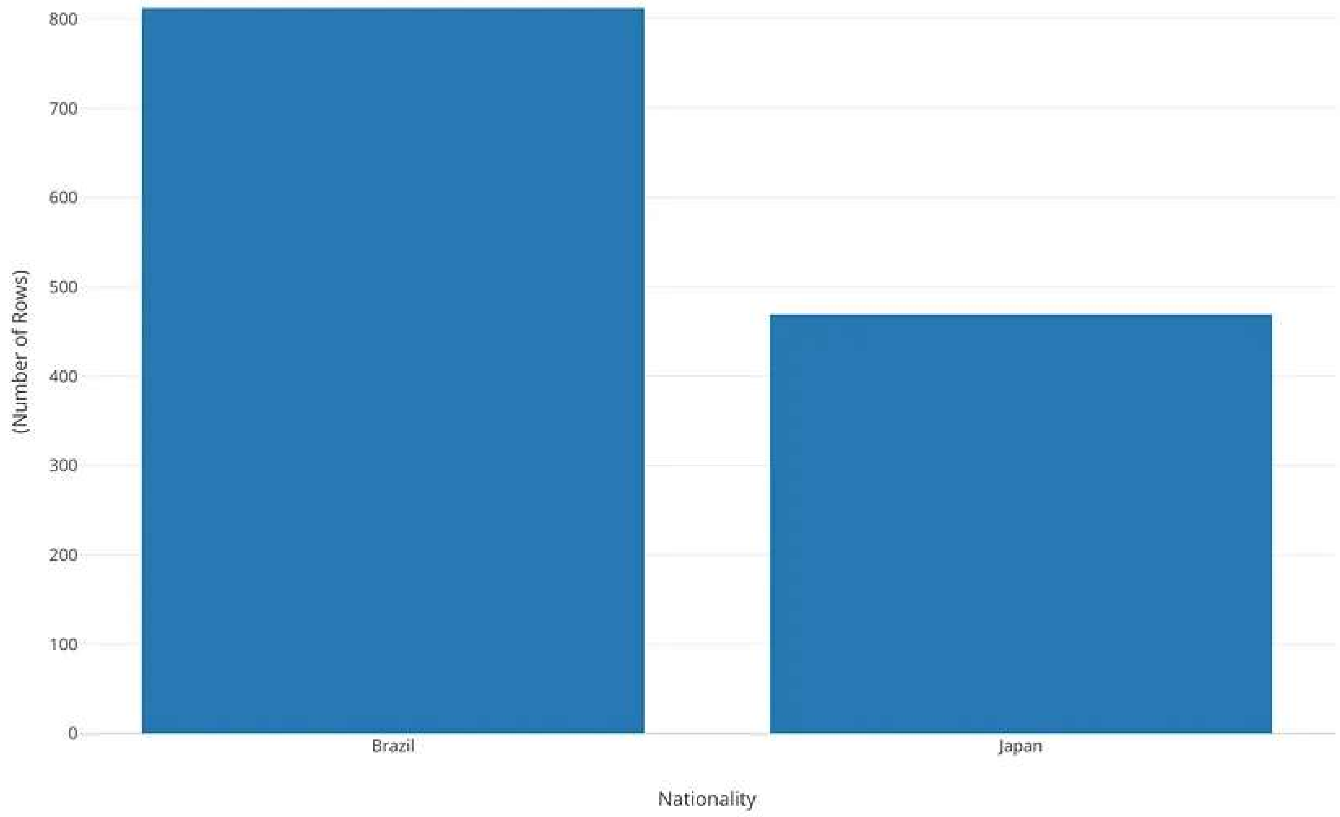
I'm going to use PCA to reduce the dimensionality of these variables so that I can have just two artificially created new dimensions, which should carry as much information as the original data has, to visualize how the soccer players from a given pair of two countries perform on the skill measures and compare the two countries.

I'm going to start with Brazil and Japan because they can show the difference clearly in terms of the player skills, which helps us understand how to interpret the result easier.

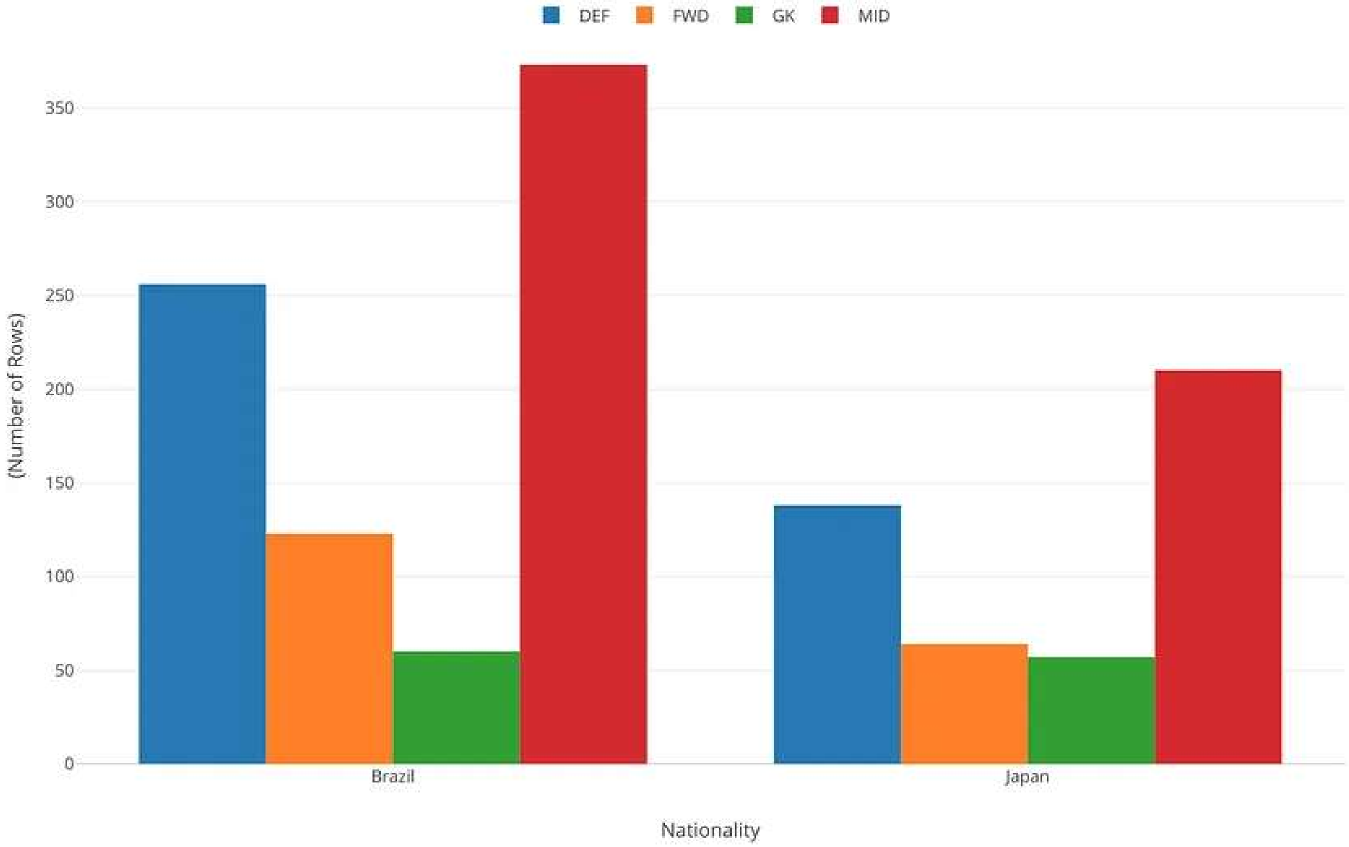
Once we set the ground, then we can move onto the teams that actually played the world cup games against each other.

Comparing Two Countries — Brazil and Japan

Let's say we want to compare the players from Brazil and Japan. We have 812 Brazilian players and 469 Japanese players in this data set.



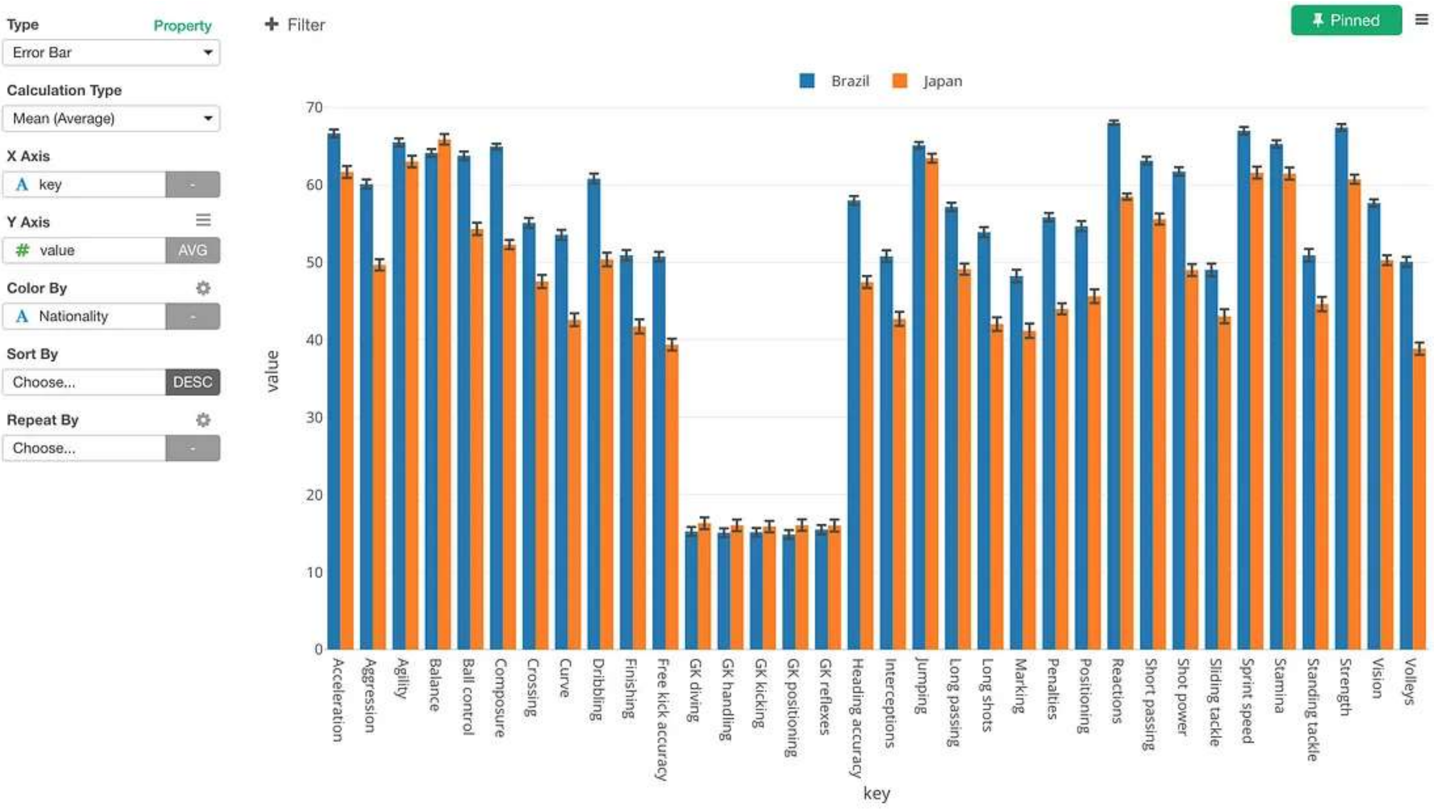
And here is a chart that I have broken down by the positions.



It’s interesting to see that the number of Goal Keepers (Green) is pretty much the same between Brazil and Japan while other positions are significantly different in terms of the number of the players. But that’s not really the point of this post.

Comparing with Average

Now, we can take a look at the average scores of all the skill measures for the two countries with Error Bar chart and compare the two countries side by side. The blue is Brazil and the orange is Japan.



Type: Error Bar

Calculation Type: Mean (Average)

X Axis: key

Y Axis: value AVG

Color By: Nationality

Sort By: Choose... DESC

Repeat By: Choose...

(The black bar-ish things at the top of blue/orange bars indicate 'standard errors'.)

Brazilian players score much higher on almost all the measures except for the ones for Goal Keepers. Again, there is something about Japanese Goal Keepers here… 🤔

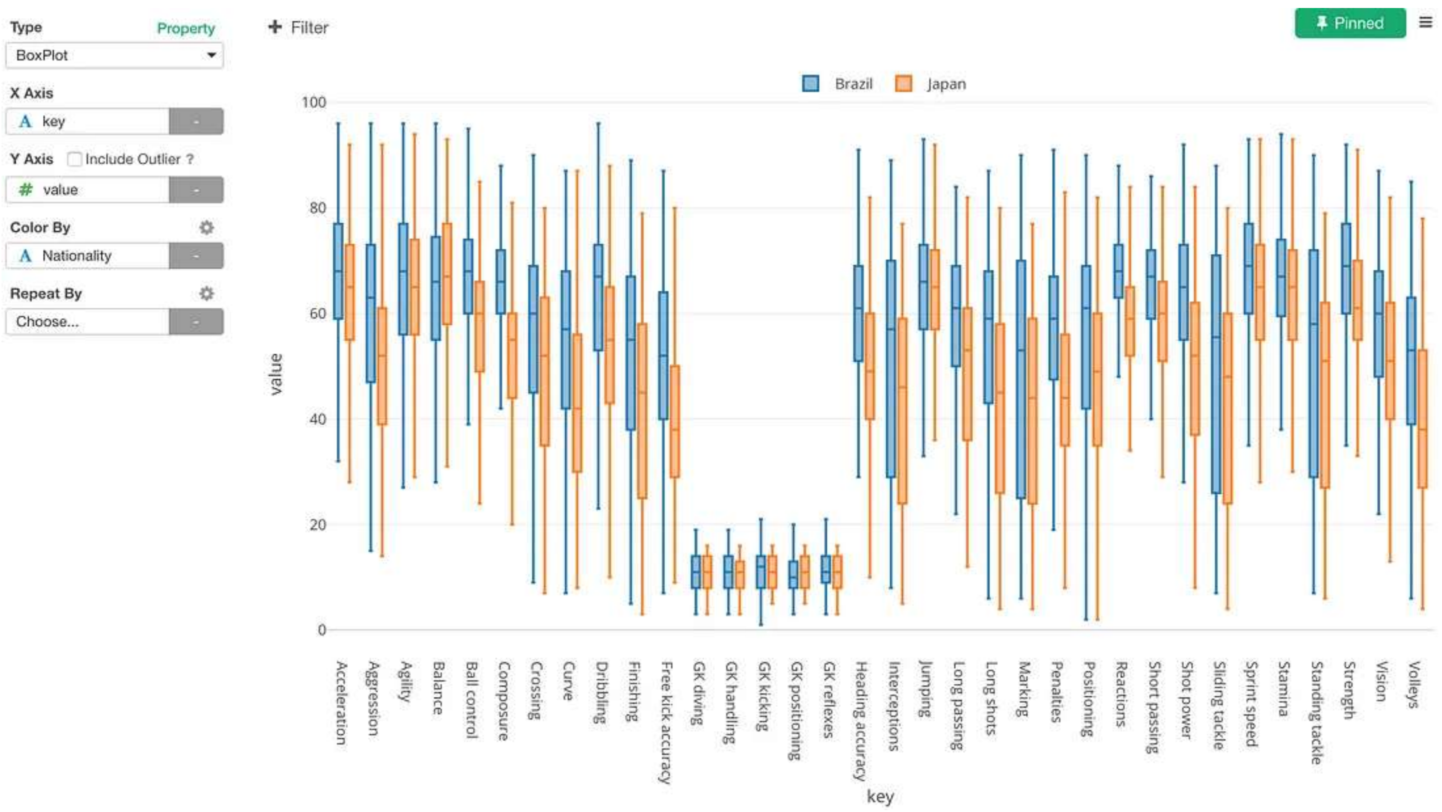
Anyway, it's clear that Brazilian players are a lot better than Japanese players in general by looking at the average scores of these skills.

But is it?

Do the average scores tell us the full story? The average is known to be highly influenced by a tiny set of outliers.

Comparing with Distribution

So, we can take a look at the distributions of the skill measures instead of the summarized value, which is the average in this case, by using the Boxplot chart like below.



Type: BoxPlot

X Axis: key

Y Axis □ Include Outlier ?: value

Color By: Nationality

Repeat By: Choose...

Here, again, the players from Brazil are better than the ones from Japan in general. But some measures like **Balance** and **Jumping** are pretty much the same between the two countries. And when it comes to the Goal Keeper skill measures, it's not like Japanese players are better, they are pretty much the same as Brazilian players.

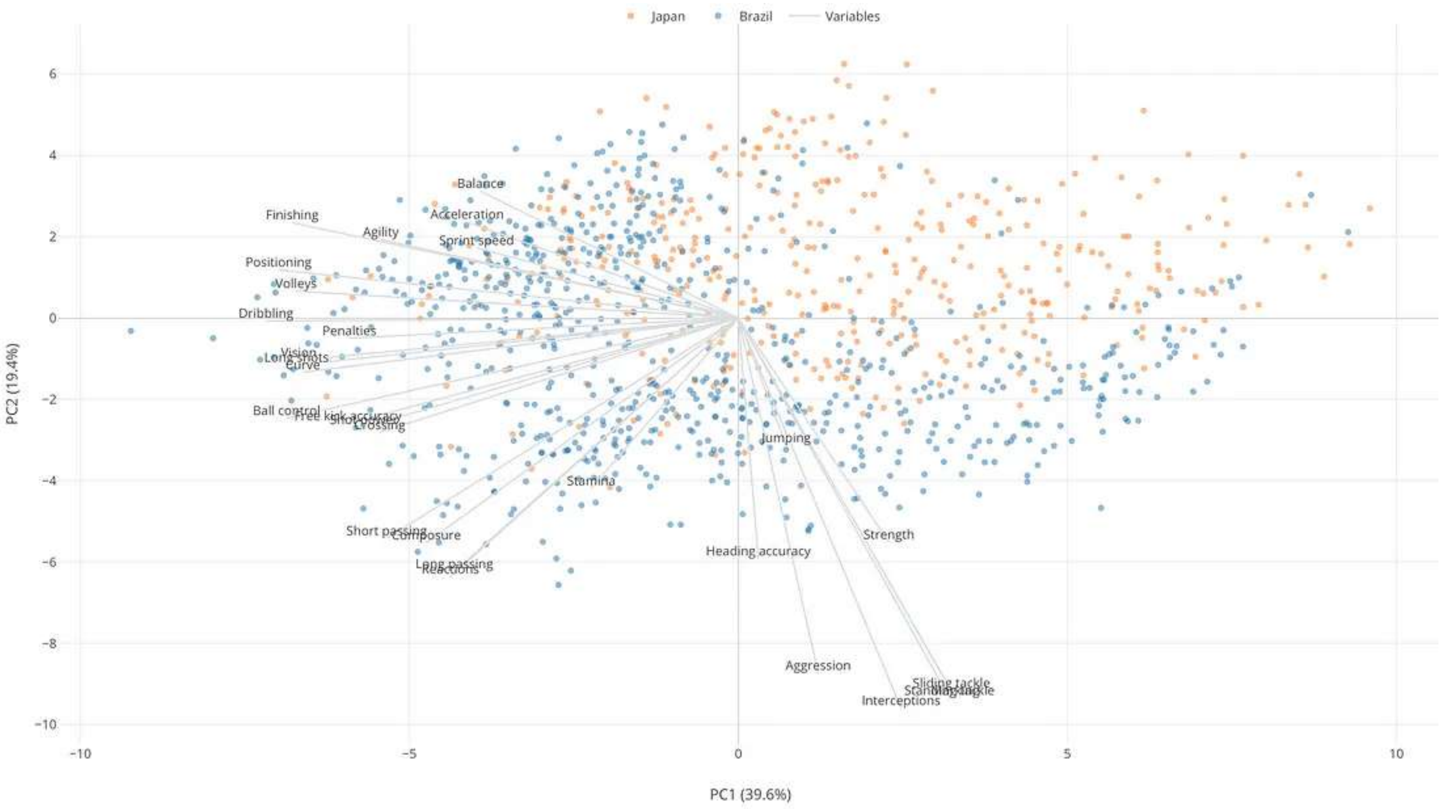
Now, we start getting a better sense about how the players from Brazil and Japan are scoring on the skill measures. But, we still don't know who are the ones making some of the measures higher or lower. These charts shown above are missing a sense of how the individual players score on the skill measures and how they make the characteristics of the two countries different.

This is when we want to use the PCA algorithm.

Comparing with PCA

I have run PCA against the data that I have filtered to keep only non-Goal Keeper players who are either Brazilian or Japanese.

Here is the Biplot chart. The blue dots are the Brazilian and the orange dots are the Japanese players.



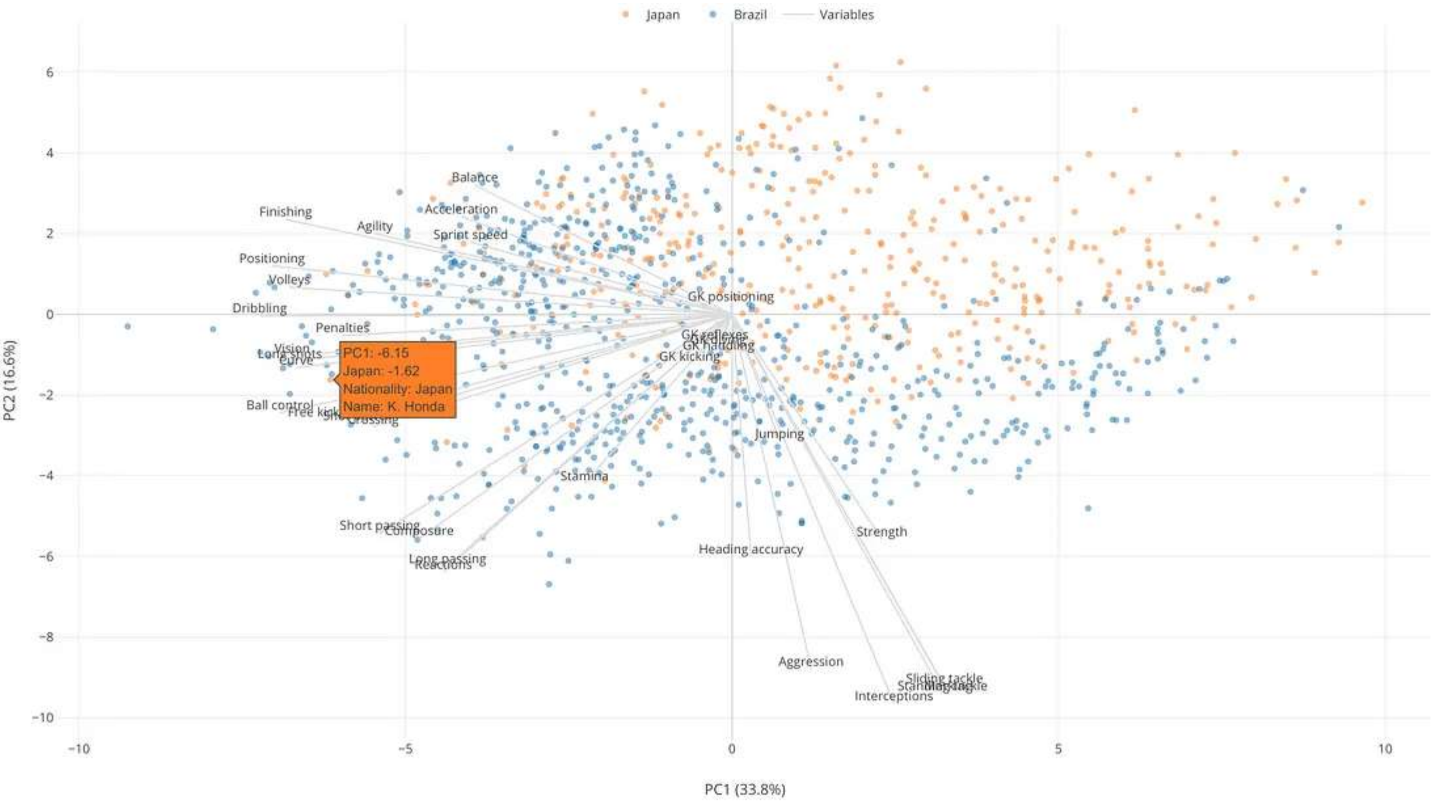
Two things are very clear.

First, most of the Japanese players (Orange) are at the right hand side top quarter area, which means that they are scoring low on all measures.

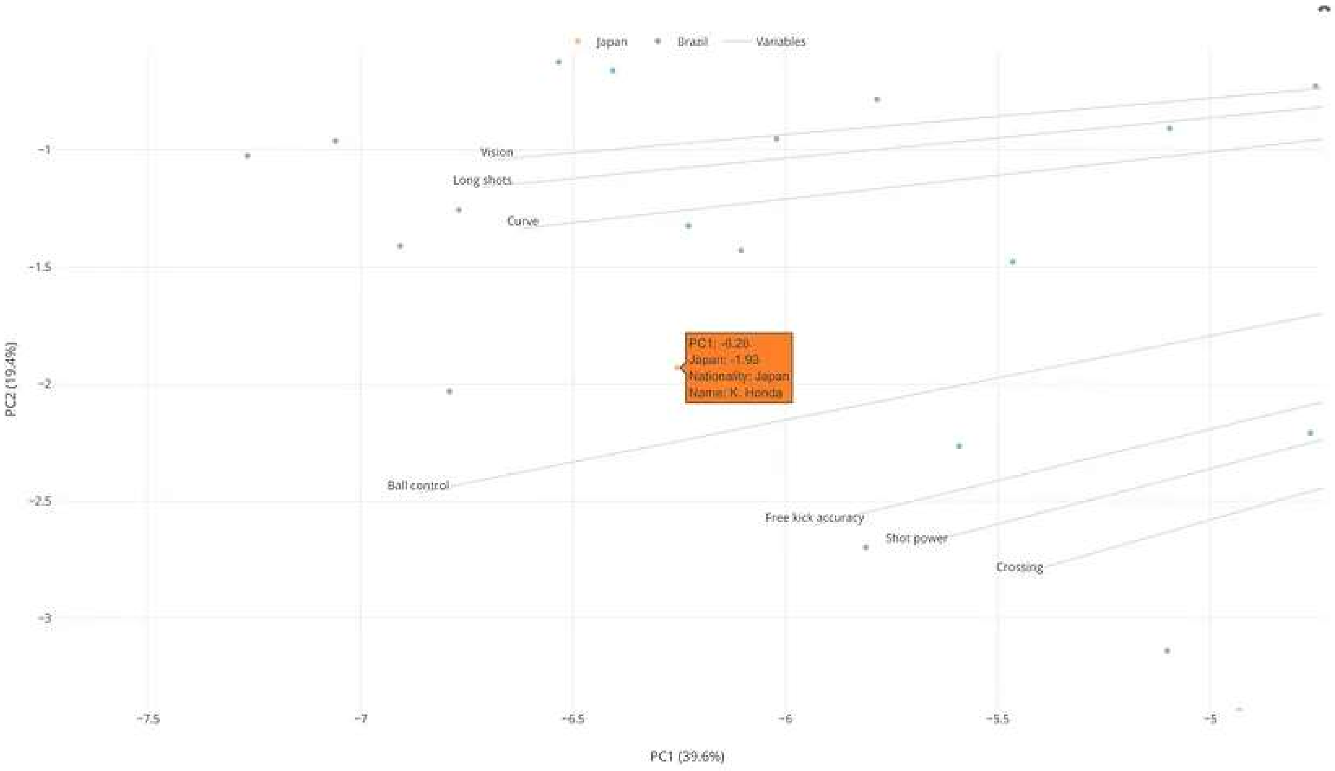
Second, Brazilian players (Blue) are scoring high on almost all the measures.

But, here is the cool thing about PCA. Just because Brazilian players are scoring higher on almost all the measures, it doesn’t mean that there are no Japanese players who score high on some of the measures.

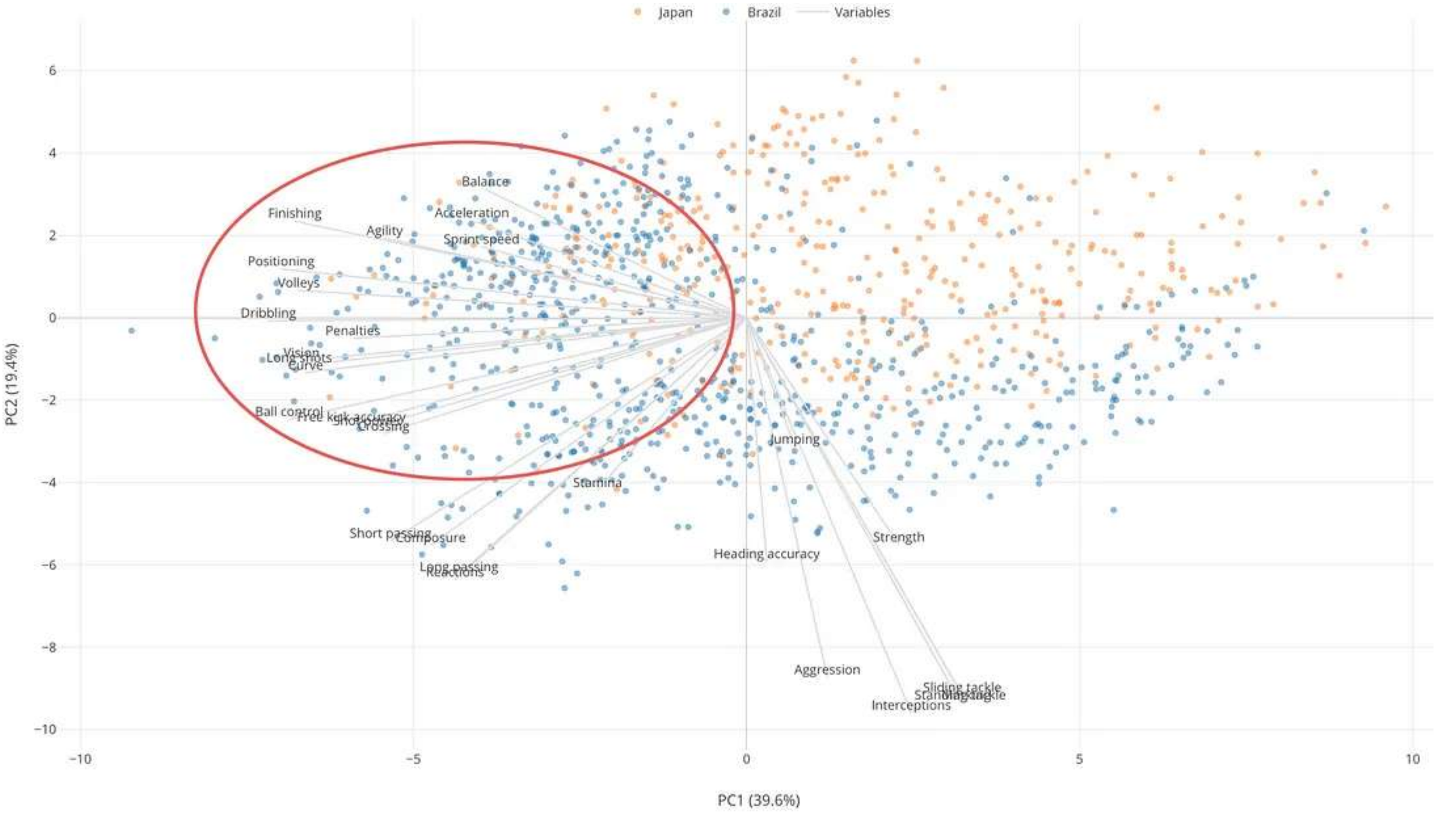
For example, **Keisuke Honda** is scoring pretty high on some of the measures.



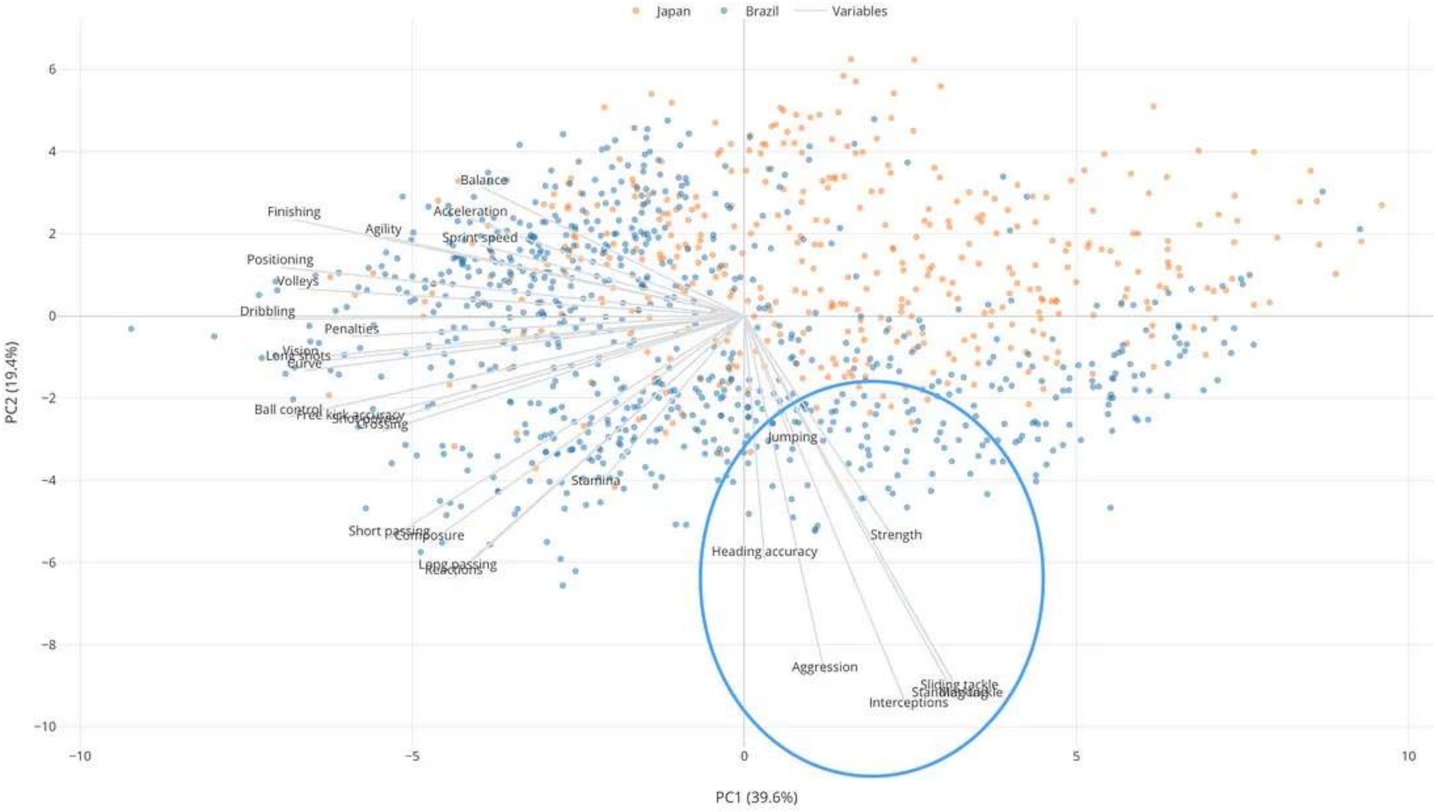
Here is a zoomed-in version that has expanded the area around **Honda.**



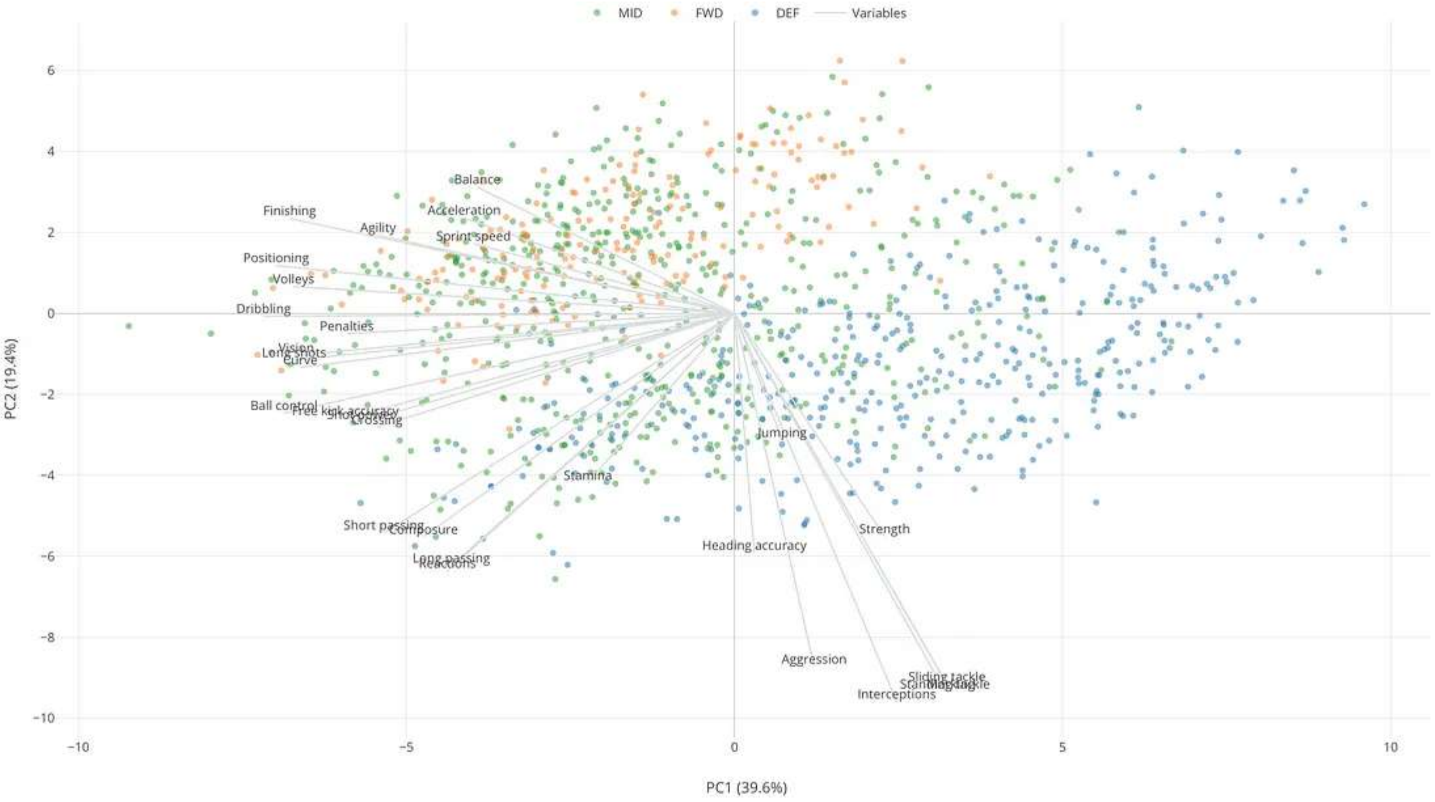
And there are some Japanese players mixed with Brazilian players in the red circled area below.



On the other hand, Japanese players are not scoring well on the measures in the blue circled area below.



What are these measures? Instead of listing those measures one by one, we can switch the color assignment from Nationality to Position like below.



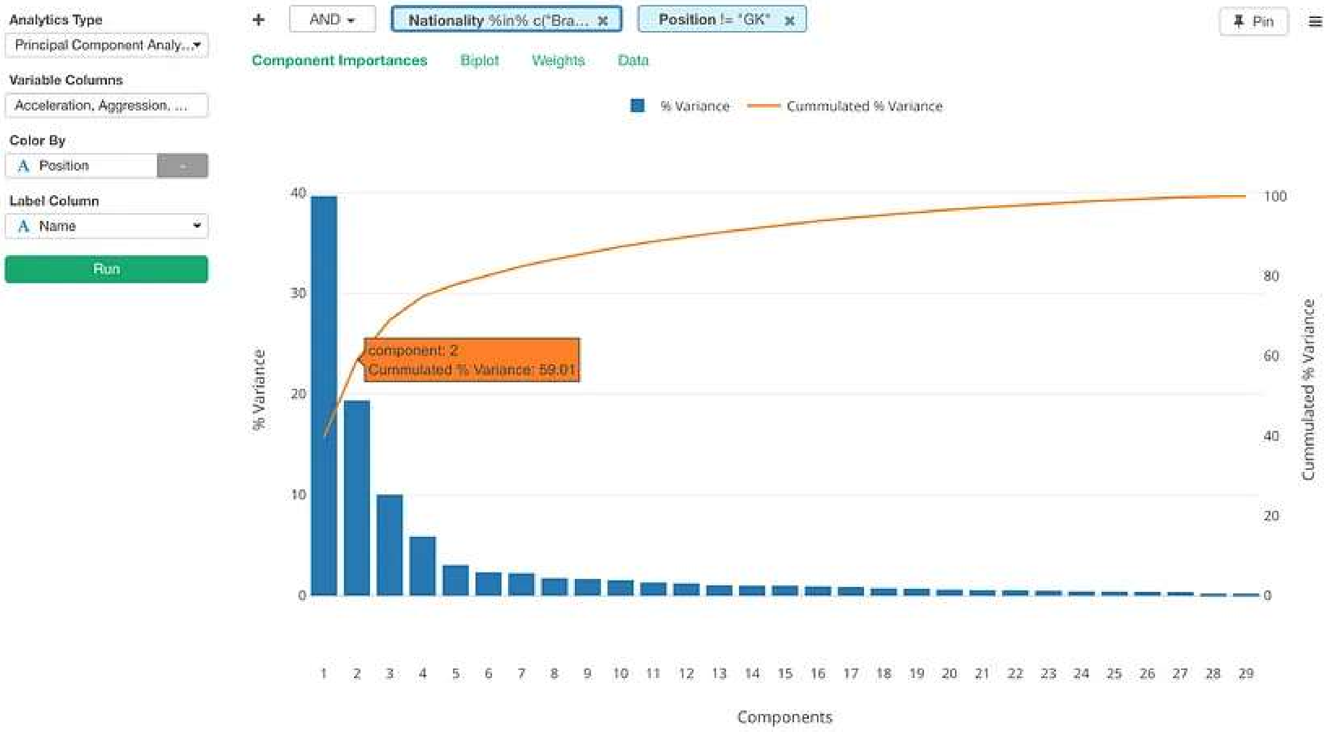
From this view, we can say that the measures some of the Japanese players are scoring well along with the Brazilian players have the characteristics of the Forward and the Mid Field players.

And the measures only Brazilian players are scoring well have the characteristics of the Defense players.

So, here is what we have found:

- Brazilian players are better than Japanese players in general if we judge them based on how they score on the skill measures.
- There are some Japanese players who are scoring on some of the measures as high as the top Brazilian players. They tend to be the offense side of players such as Forward.
- When it comes to the skills that are required for Defense, Japanese players are not doing well compared to Brazilian players.

This information might not be as accurate as we hope because these two dimensions (X and Y) that are used for the charts above have captured only about 60% of the original information.



Analytics Type: Principal Component Analy...

Variable Columns: Acceleration, Aggression, ...

Color By: Position

Label Column: Name

AND

Nationality %in% c("Bra...

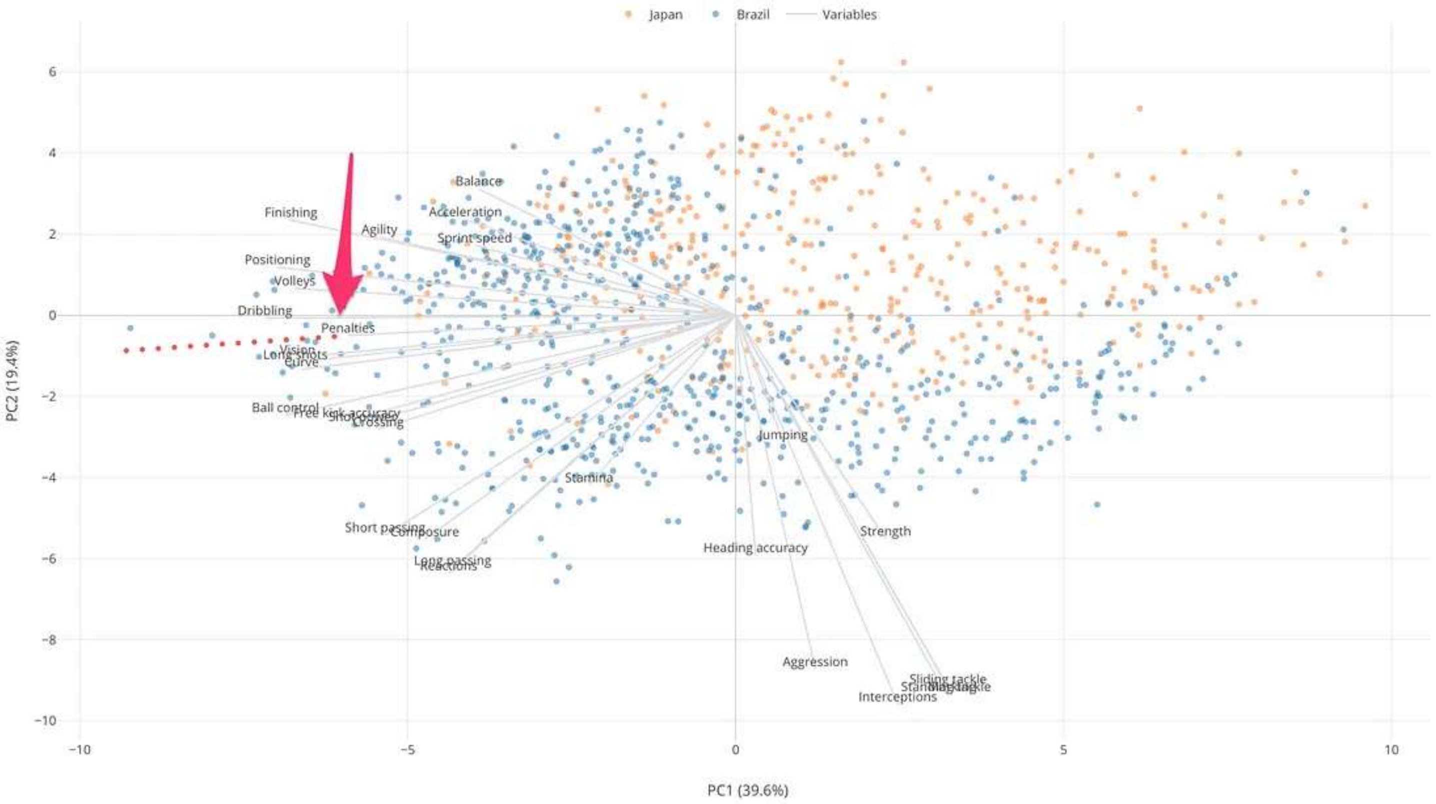
Position != "GK"

component: 2
Cummulated % Variance: 59.01

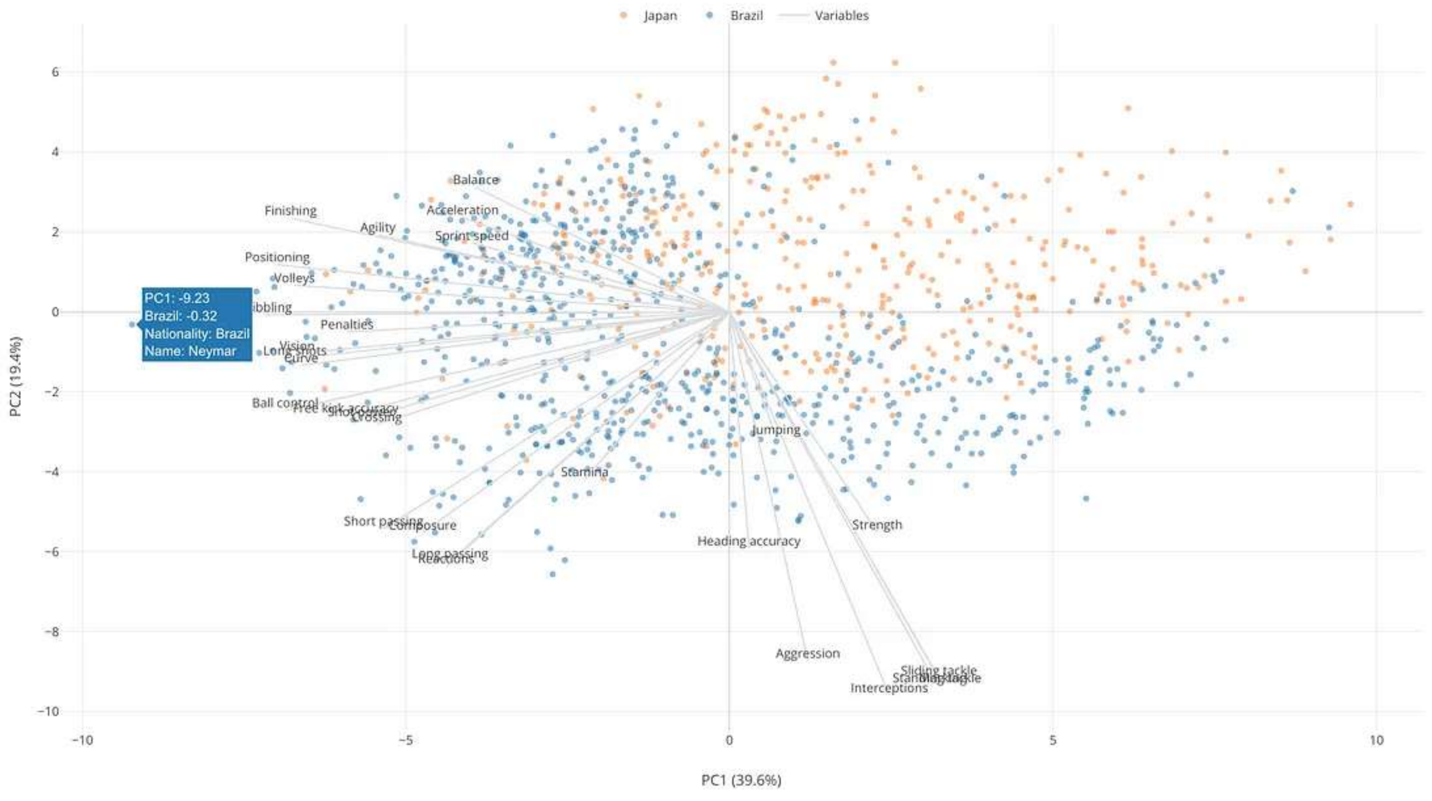
But, you can still get a good understanding of how the players are on those skills and compare the characteristics of the two countries.

Oh, you want to know something funny?

Who is the Brazilian player that is at the most left hand side? This player is on a line that we extend Penalties measure to.



That’s Neymar from Brazil.

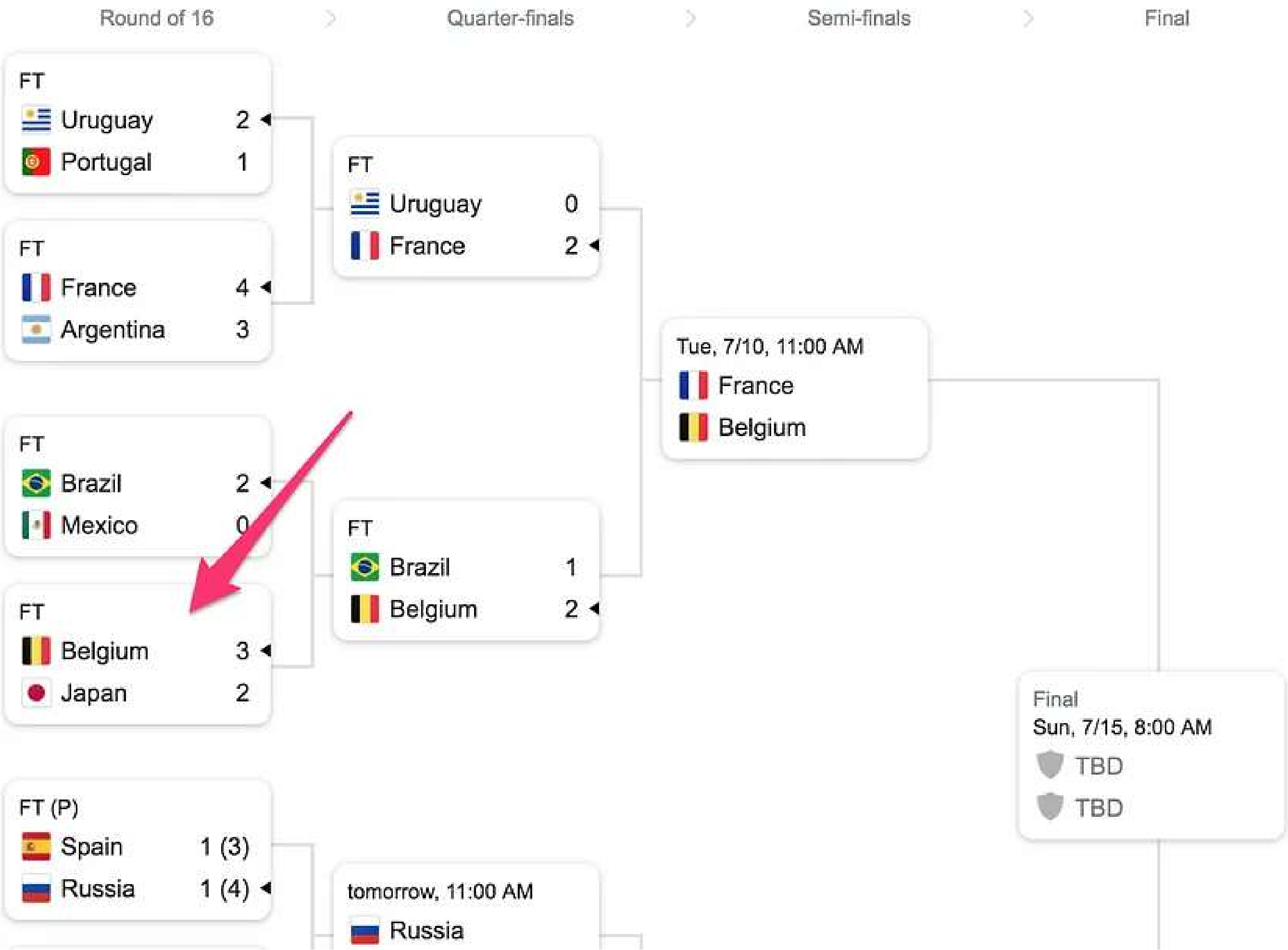


And I’m sure some of you are familiar with his performance to claim the penalties.

Now, let's take a look at three games that have actually happened at 2018 World Cup.

Belgium vs. Japan

First, let's take a look at the players from Belgium and Japan by using PCA.



Round of 16 > Quarter-finals > Semi-finals > Final

FT
Uruguay 2
Portugal 1

FT
France 4
Argentina 3

FT
Brazil 2
Mexico 0

FT
Belgium 3
Japan 2

FT (P)
Spain 1 (3)
Russia 1 (4)

FT
Uruguay 0
France 2

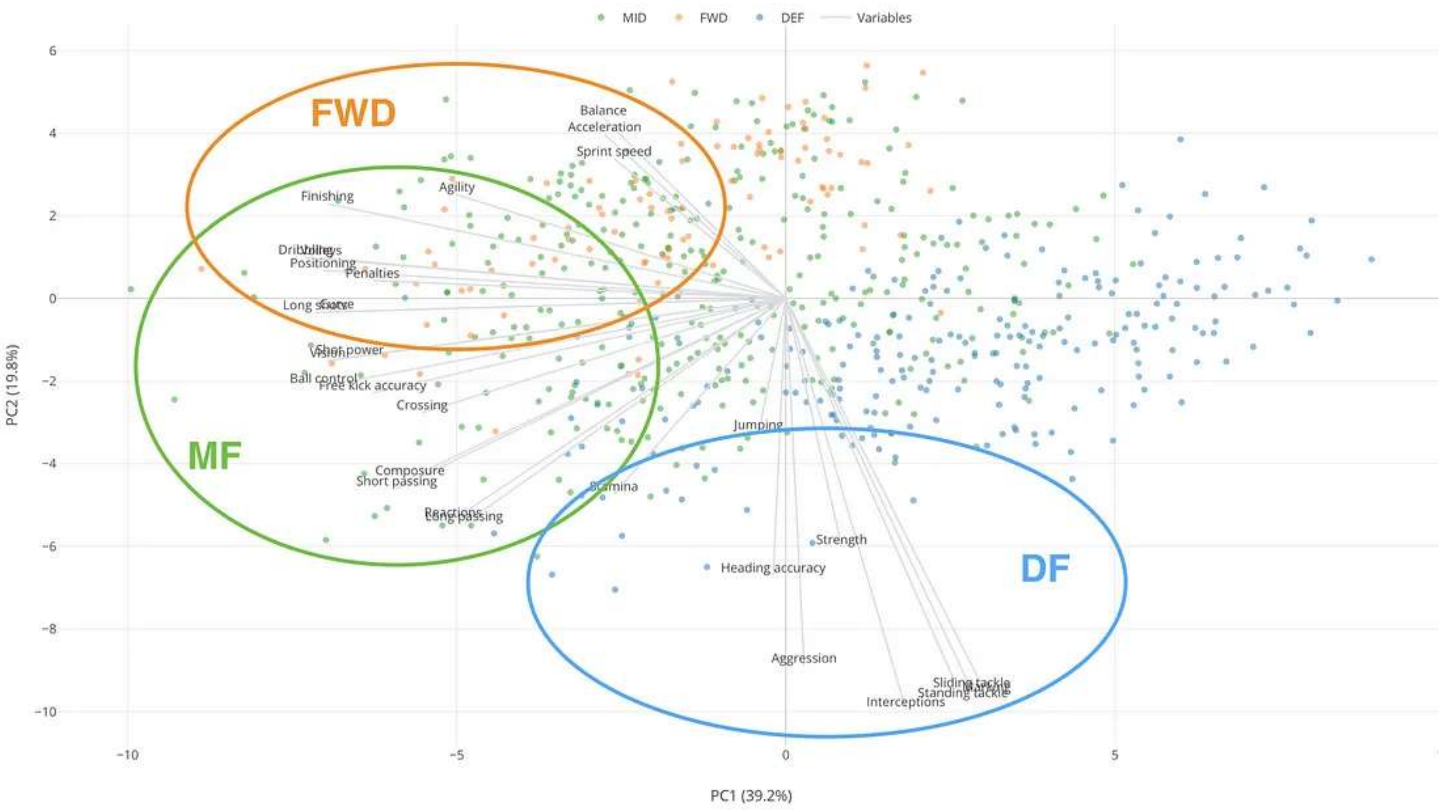
FT
Brazil 1
Belgium 2

tomorrow, 11:00 AM
Russia

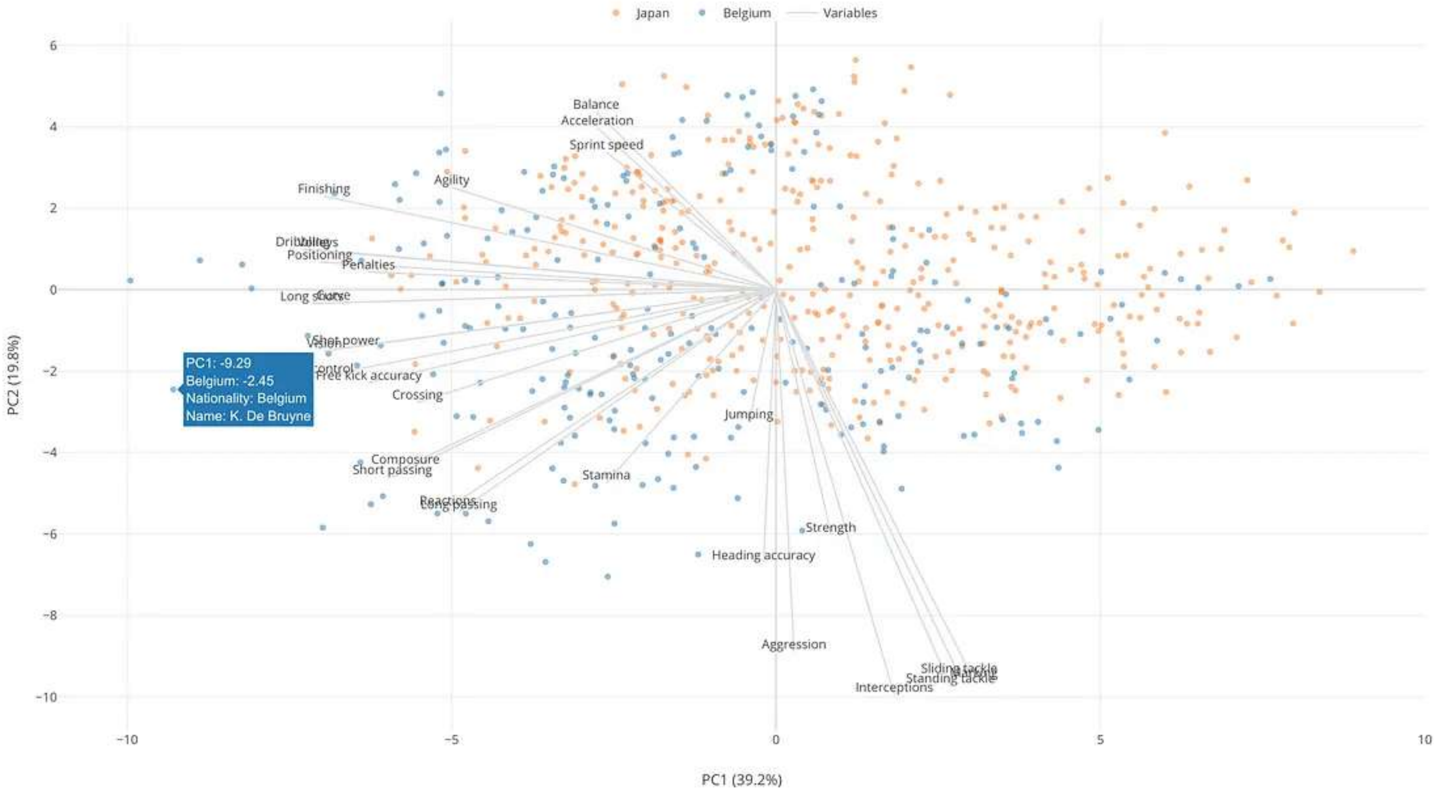
Tue, 7/10, 11:00 AM
France
Belgium

Final
Sun, 7/15, 8:00 AM
TBD
TBD

Japan has lost to Belgium, and if we run PCA for all the players from the two countries we get something like below.

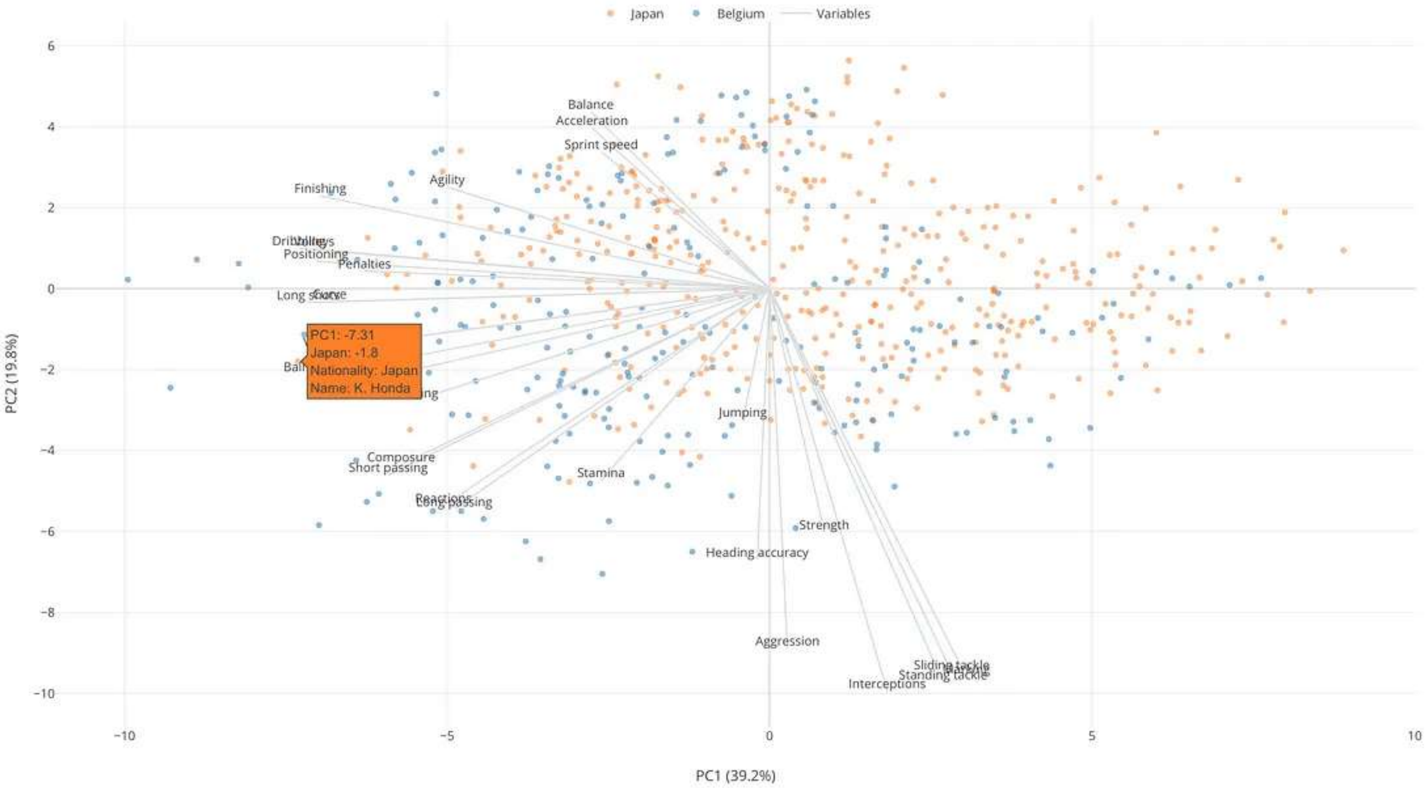


Now, if we change the color to Nationality, we can see that the players from Belgium (Blue) tend to score higher on the Mid Field player skills and the Defense player skills.



Just like we saw when we compared Brazilian and Japanese players, more Japanese players (Orange) are at the right hand side top quarter.

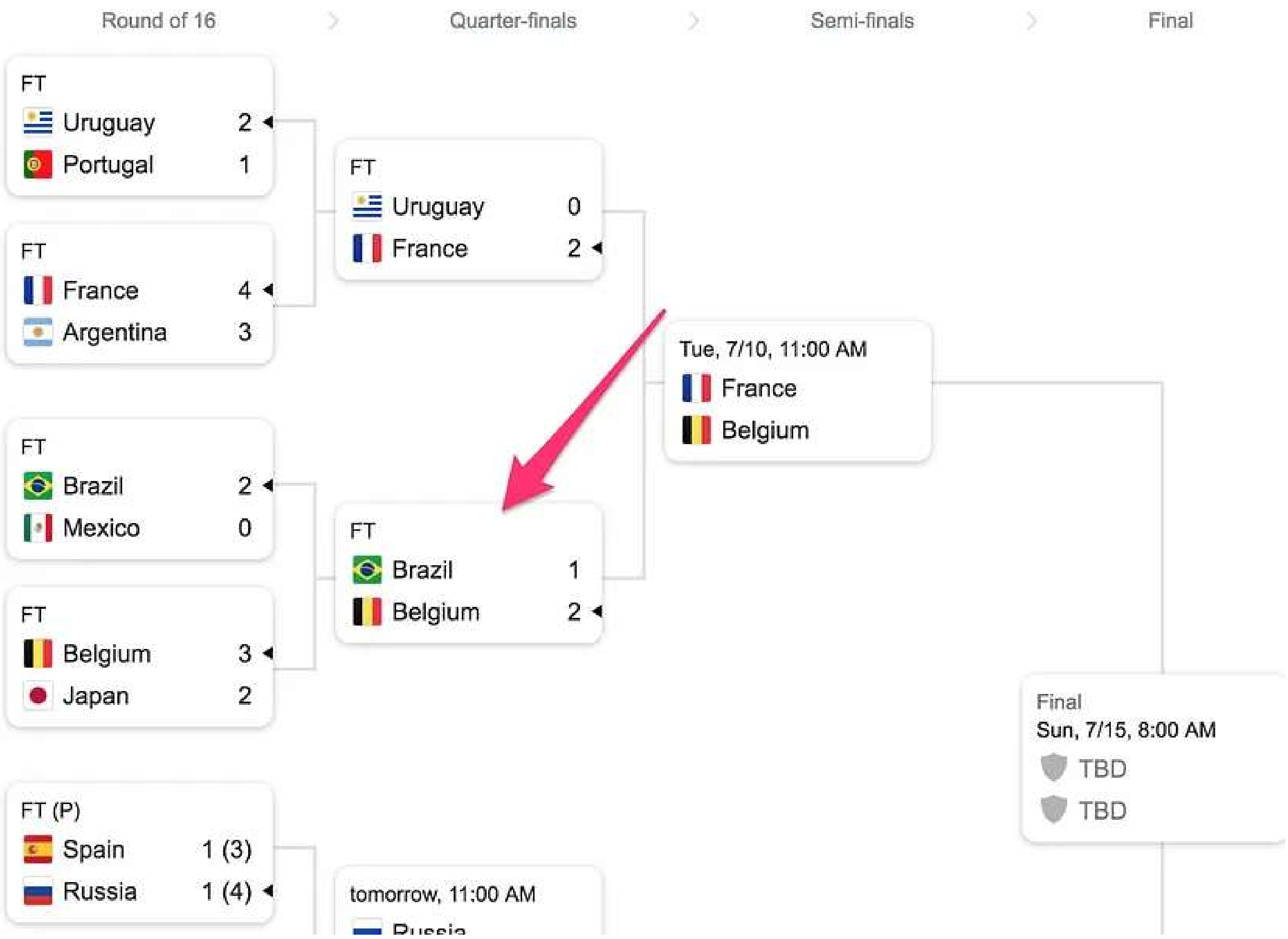
This goes to show that Belgium players tend to be measured better on most of the skills.



So, it is actually amazing that the Japan team played almost evenly with the Belgium team at the World Cup.

Belgium vs. Brazil

Now, let's take a look at the players from Belgium and Brazil with PCA. Belgium beat Brazil at the quarter final.



Round of 16 > Quarter-finals > Semi-finals > Final

FT
Uruguay 2
Portugal 1

FT
France 4
Argentina 3

FT
Brazil 2
Mexico 0

FT
Belgium 3
Japan 2

FT (P)
Spain 1 (3)
Russia 1 (4)

FT
Uruguay 0
France 2

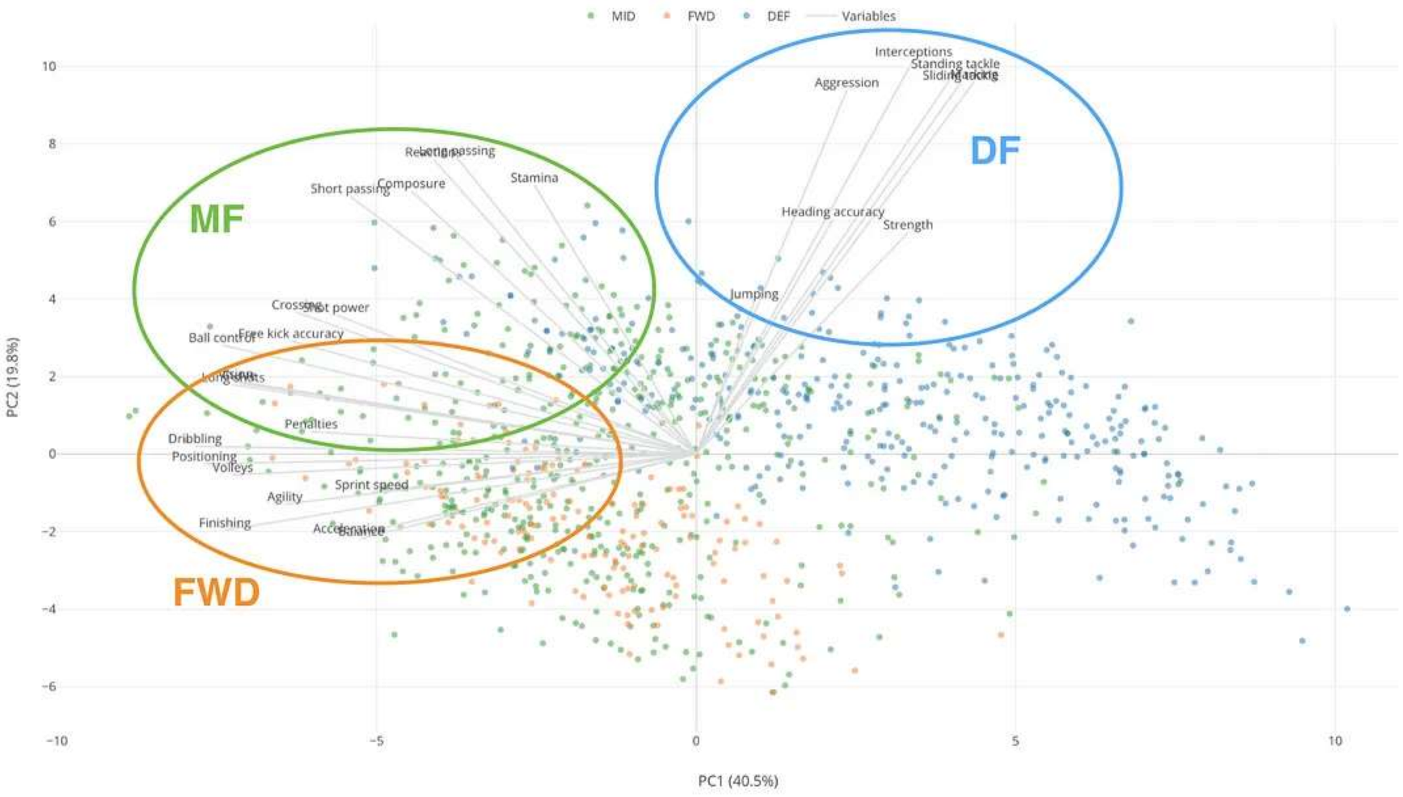
FT
Brazil 1
Belgium 2

tomorrow, 11:00 AM
Russia

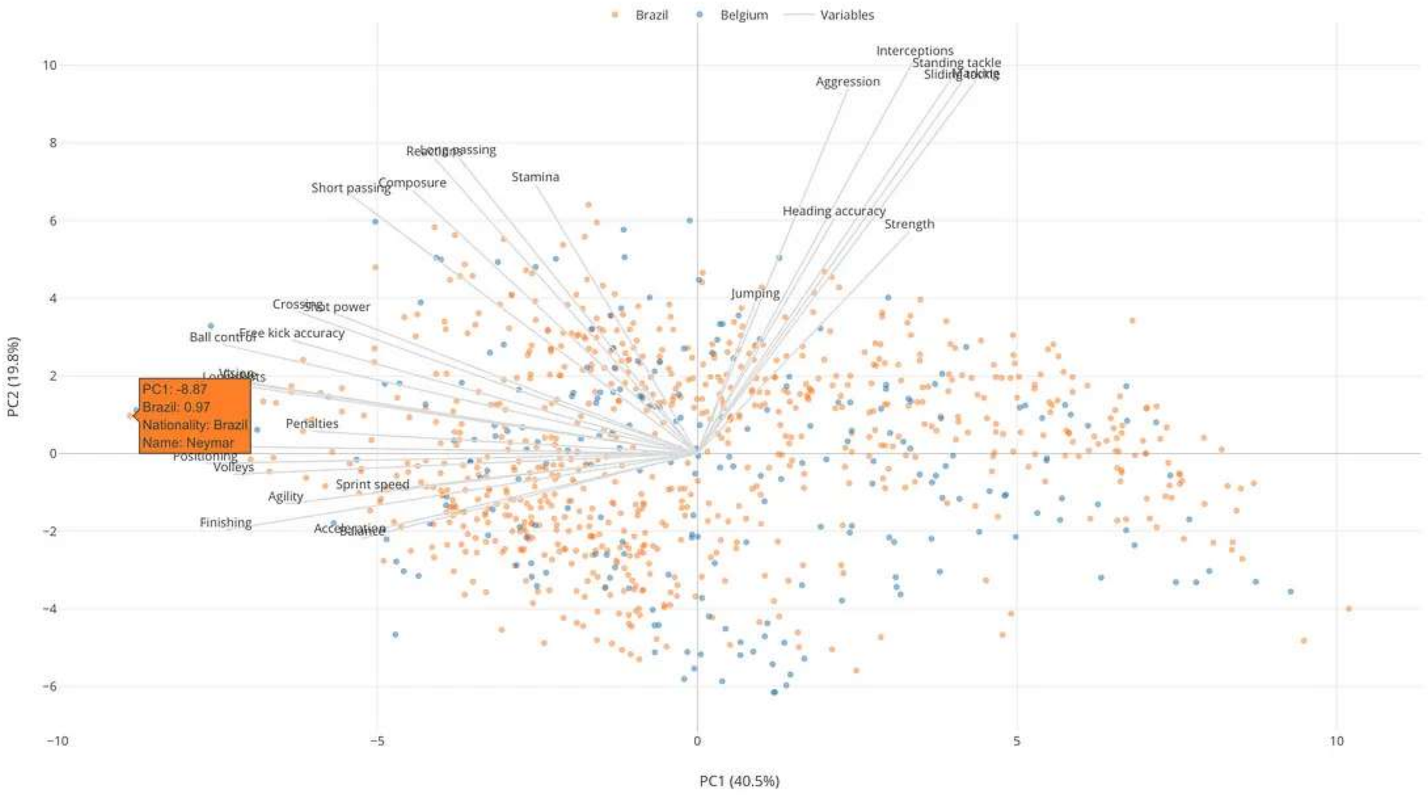
Tue, 7/10, 11:00 AM
France
Belgium

Final
Sun, 7/15, 8:00 AM
TBD
TBD

First, let’s understand the two reduced dimensions first. The defense players tend to be at the right hand side (Blue dots) and mid fielders are at the left hand side (Green dots) and forward players tend to be at the left hand side bottom area (Orange dots).



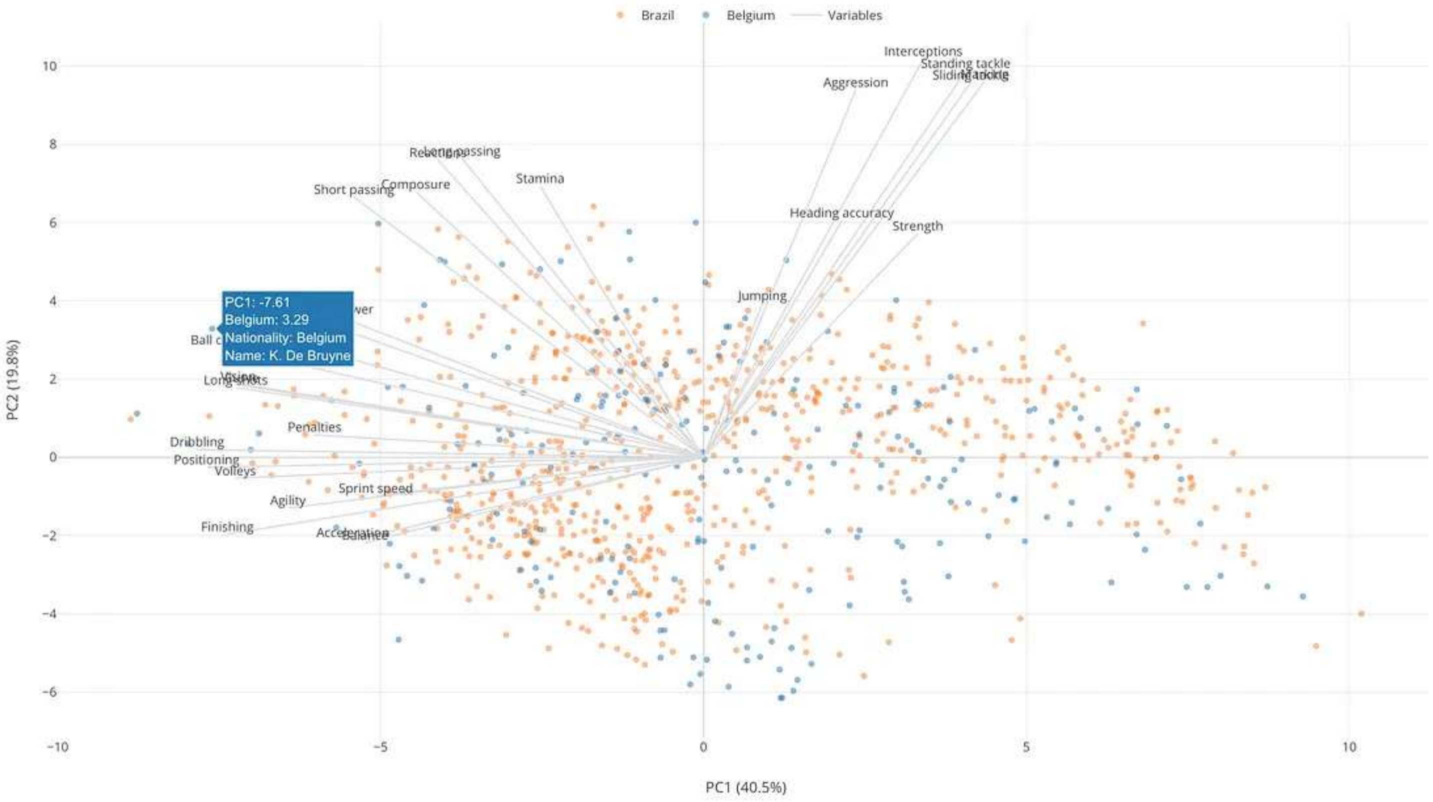
If we change the color to Nationality, we get something like below.



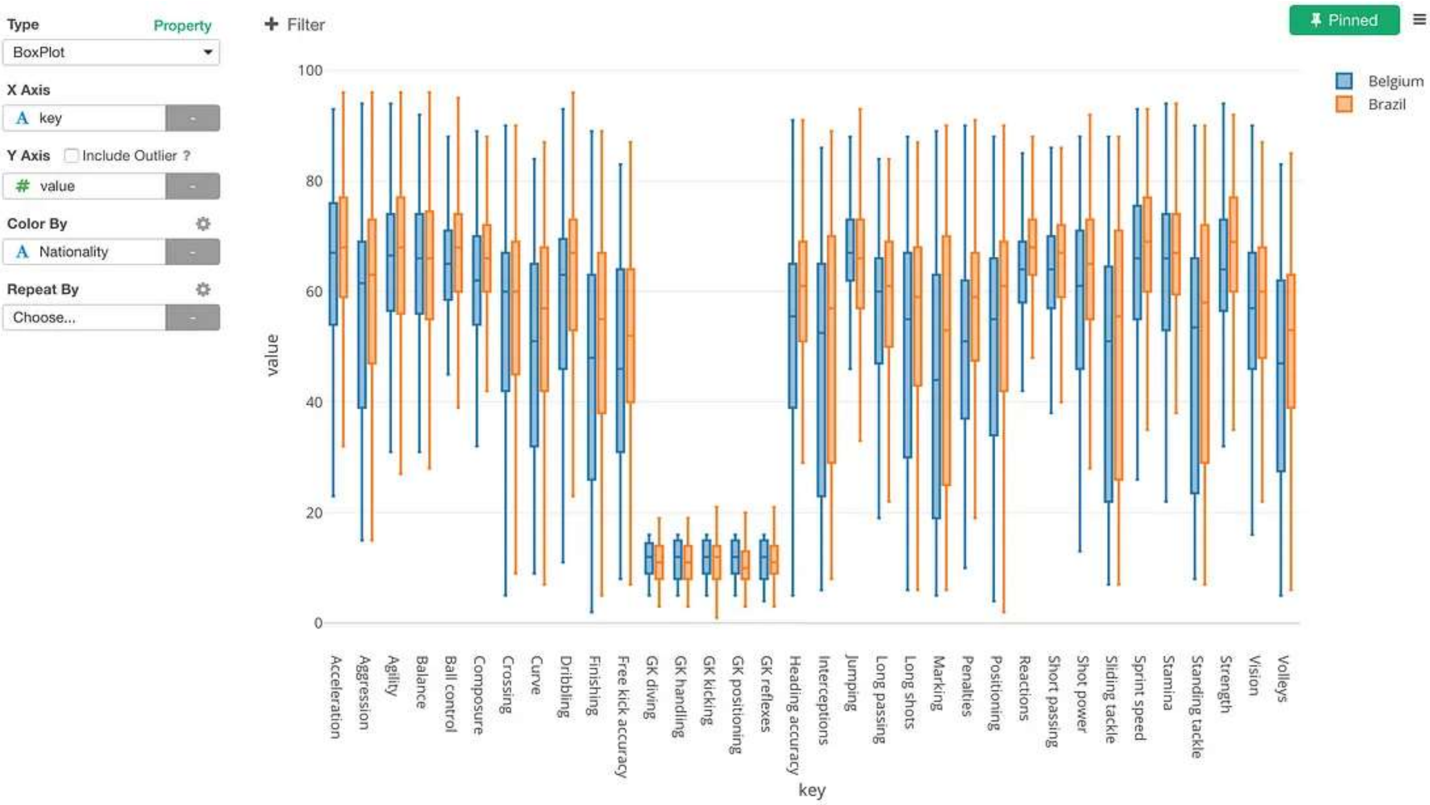
Here, we can see more Belgium players (Blue) at right hand side bottom area and we can see more Brazilian players (Orange) scoring higher on the Mid field players and Forward players skills.

We can see Neymar from Brazil at the very left side, which means he scores high on Dribbling, Volleys, Positioning, Long shots, Vision, and Penalty! 😉

On the other hand, you can see De Bruyne from Belgium on the direction of Ball Control, Free Kick Accuracy, Shot Power, Long Shot, Vision.



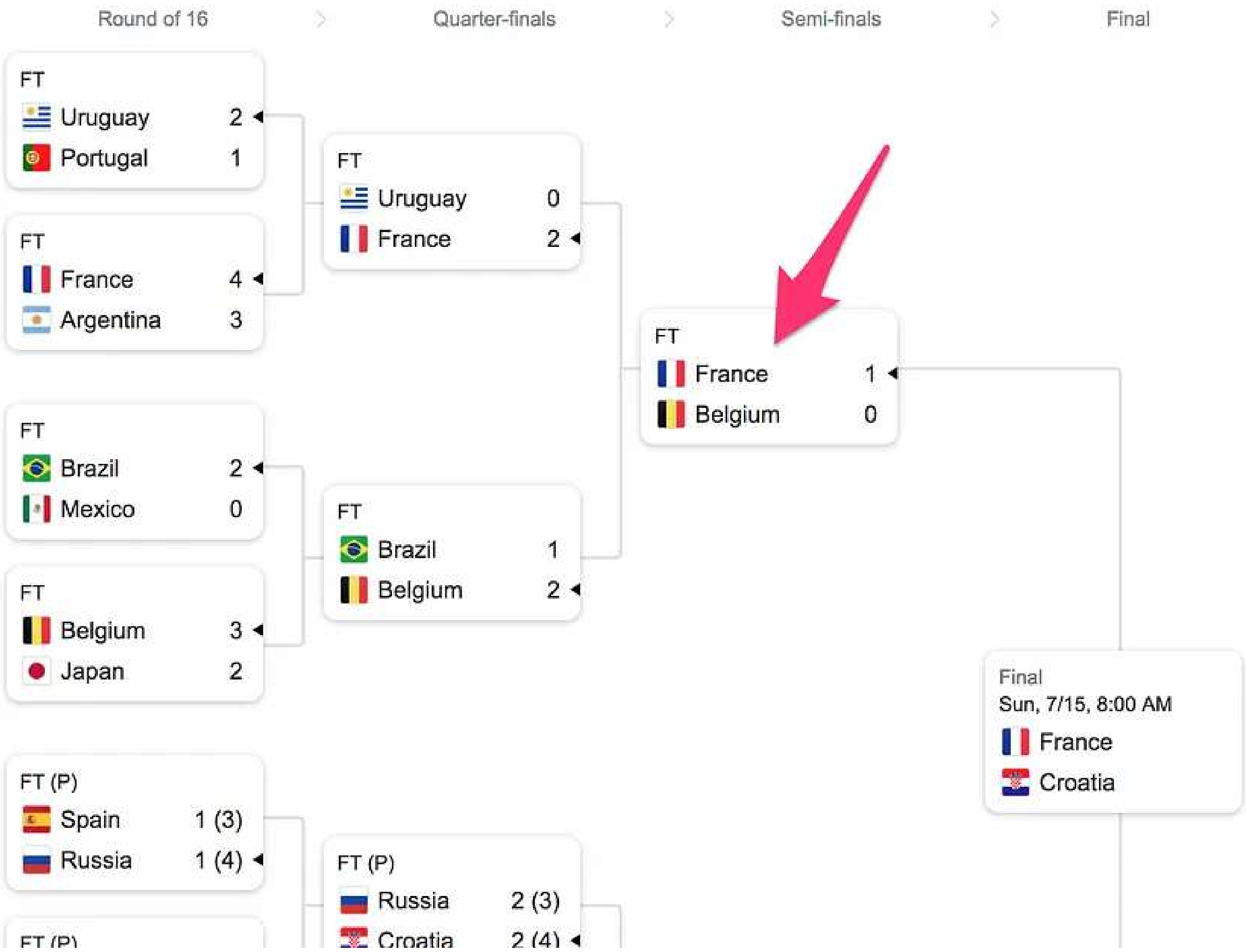
Just to see the distribution of all the skill measures for the two teams, we get something like below with Boxplot chart.


Type Property
BoxPlot
X Axis
key
Y Axis □ Include Outlier ?
value
Color By
Nationality
Repeat By
Choose...

It is amazing that Brazilian players are scoring higher on almost all the skills in general, yet they have lost the game to Belgium.

Belgium vs. France

Let's take a look at the players from Belgium and France. France beat Belgium at the Semi-final.



Round of 16 › Quarter-finals › Semi-finals › Final

FT
Uruguay 2
Portugal 1

FT
France 4
Argentina 3

FT
Brazil 2
Mexico 0

FT
Belgium 3
Japan 2

FT (P)
Spain 1 (3)
Russia 1 (4)

FT (P)

FT
Uruguay 0
France 2

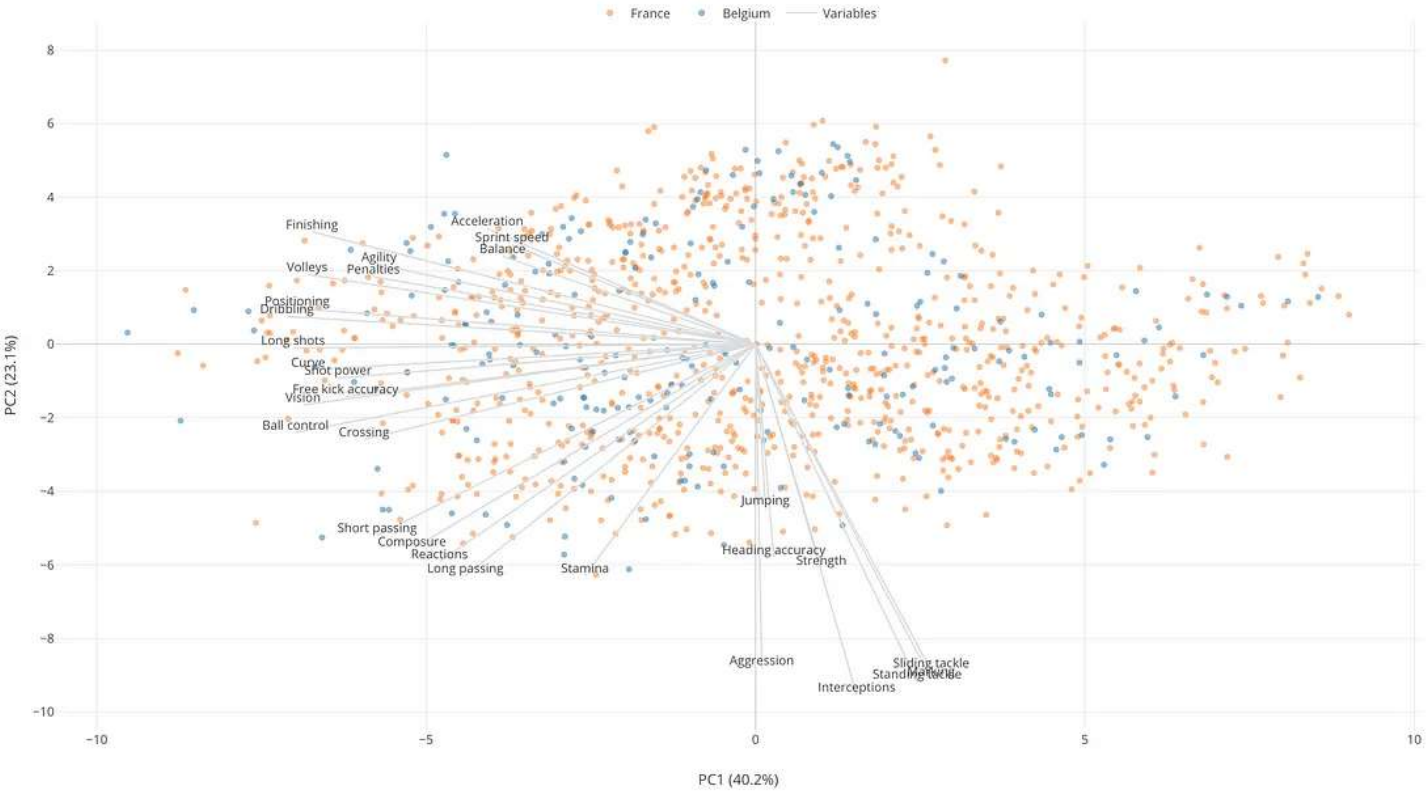
FT
Brazil 1
Belgium 2

FT (P)
Russia 2 (3)
Croatia 2 (4)

FT
France 1
Belgium 0

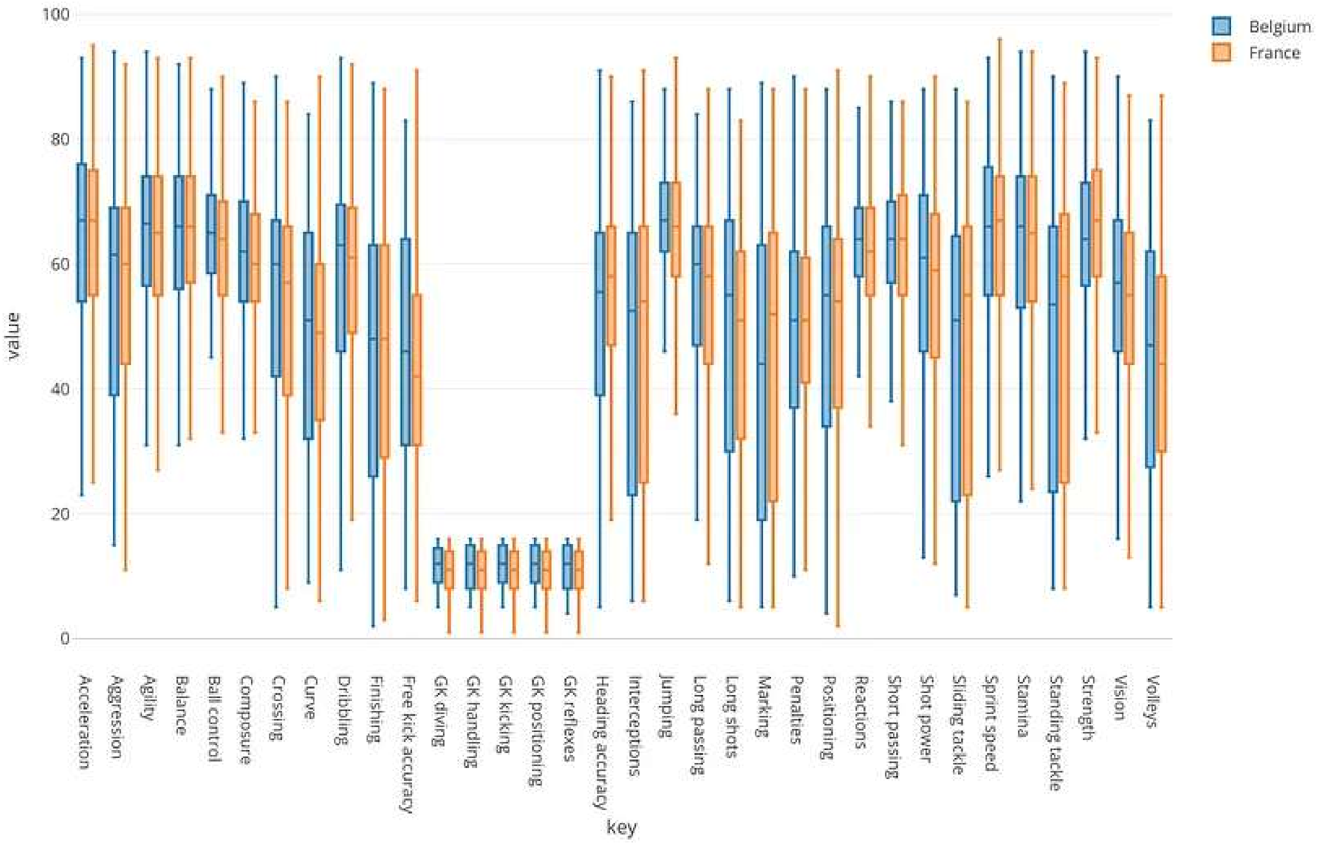
Final
Sun, 7/15, 8:00 AM
France
Croatia

And here's the result of running PCA against the players from France and Belgium.



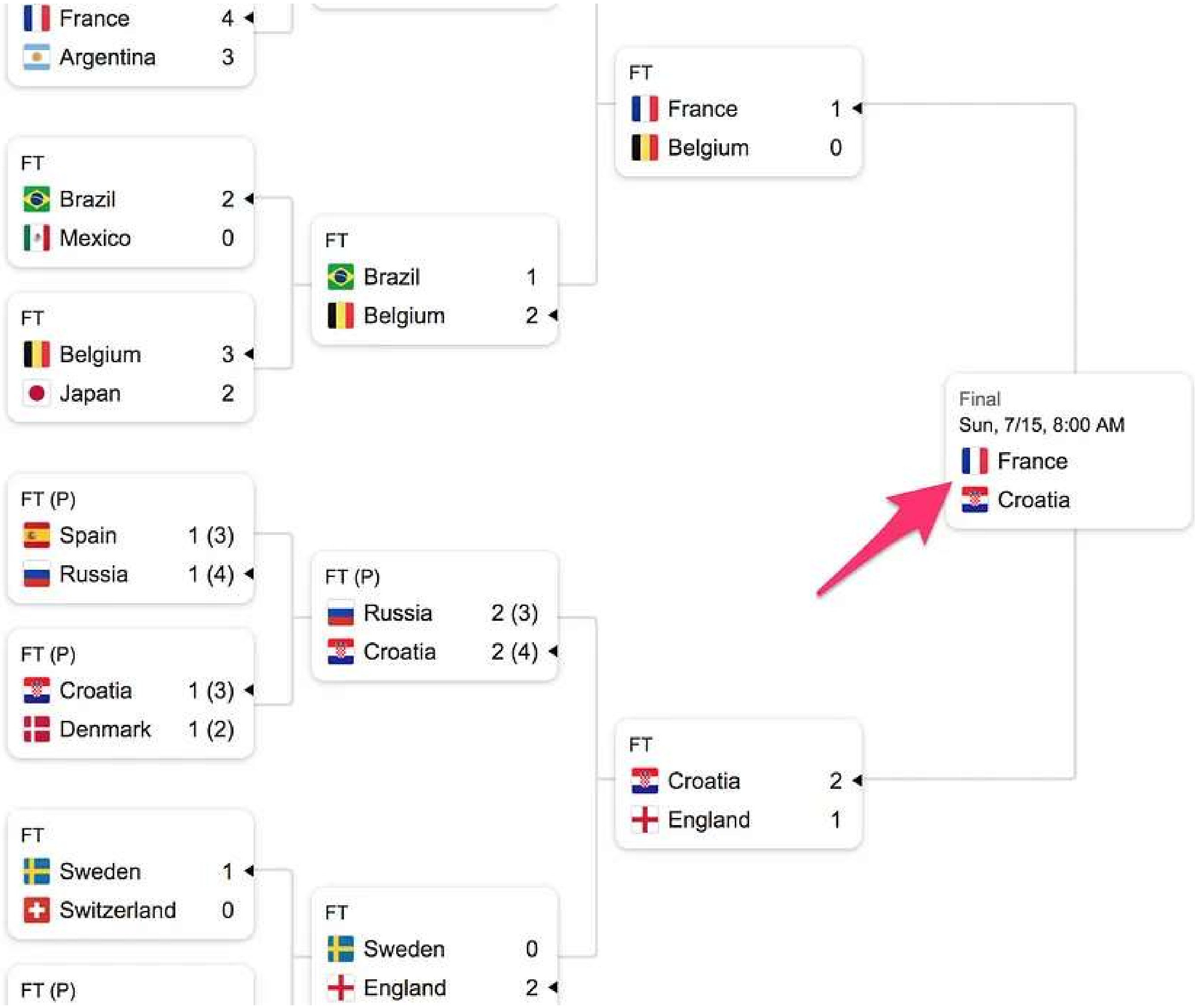
The players from these two countries are spread everywhere in a same way. It is hard find a clear difference.

If we look at the Boxplot, again it’s hard to tell the difference between the two teams.



Croatia vs. France

Here is the last one and it's about the final match. As of writing (2018/07/10), we don't know who is going to win yet.



France 4
Argentina 3

FT
Brazil 2
Mexico 0

FT
Belgium 3
Japan 2

FT (P)
Spain 1 (3)
Russia 1 (4)

FT (P)
Croatia 1 (3)
Denmark 1 (2)

FT
Sweden 1
Switzerland 0

FT (P)

FT
Brazil 1
Belgium 2

FT (P)
Russia 2 (3)
Croatia 2 (4)

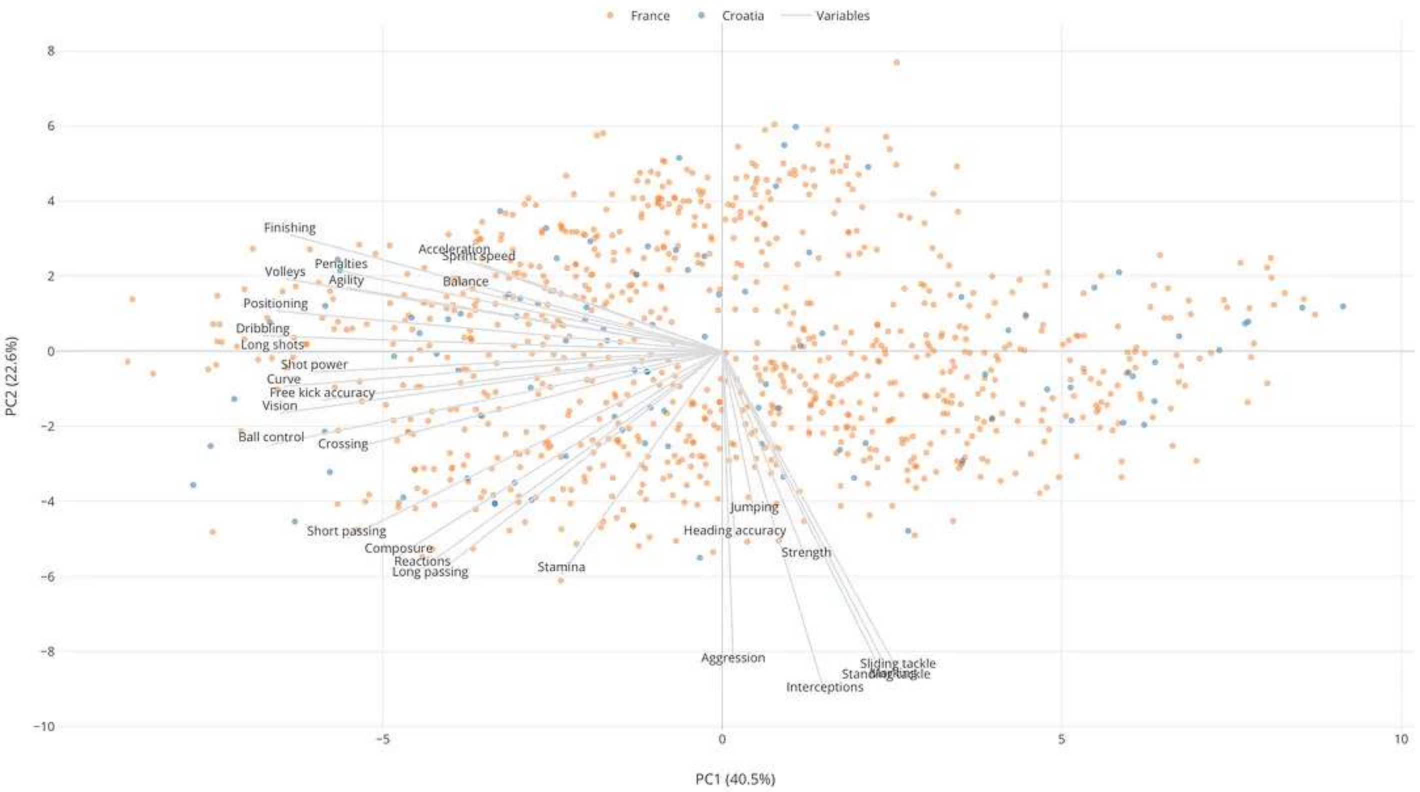
FT
Sweden 0
England 2

FT
France 1
Belgium 0

FT
Croatia 2
England 1

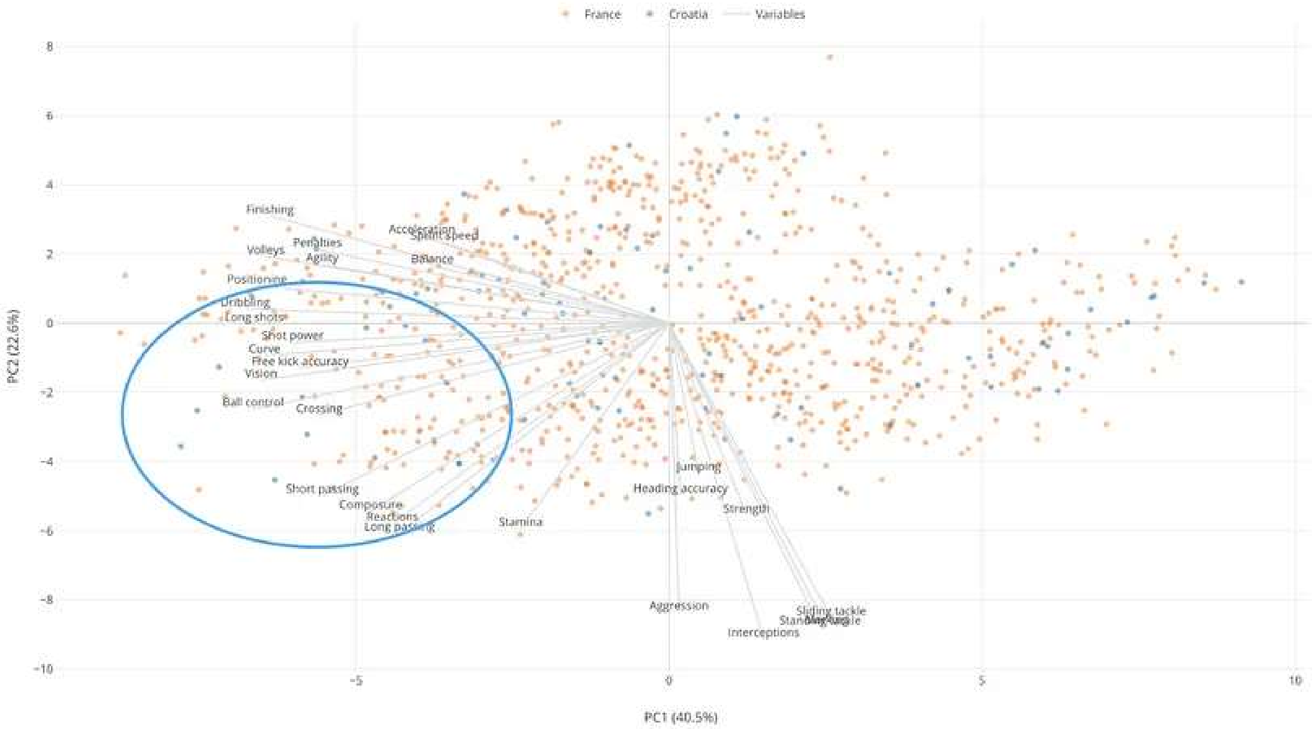
Final
Sun, 7/15, 8:00 AM
France
Croatia

Here's the Biplot chart by PCA. The blue dots are the Croatian and the orange dots are the French players.



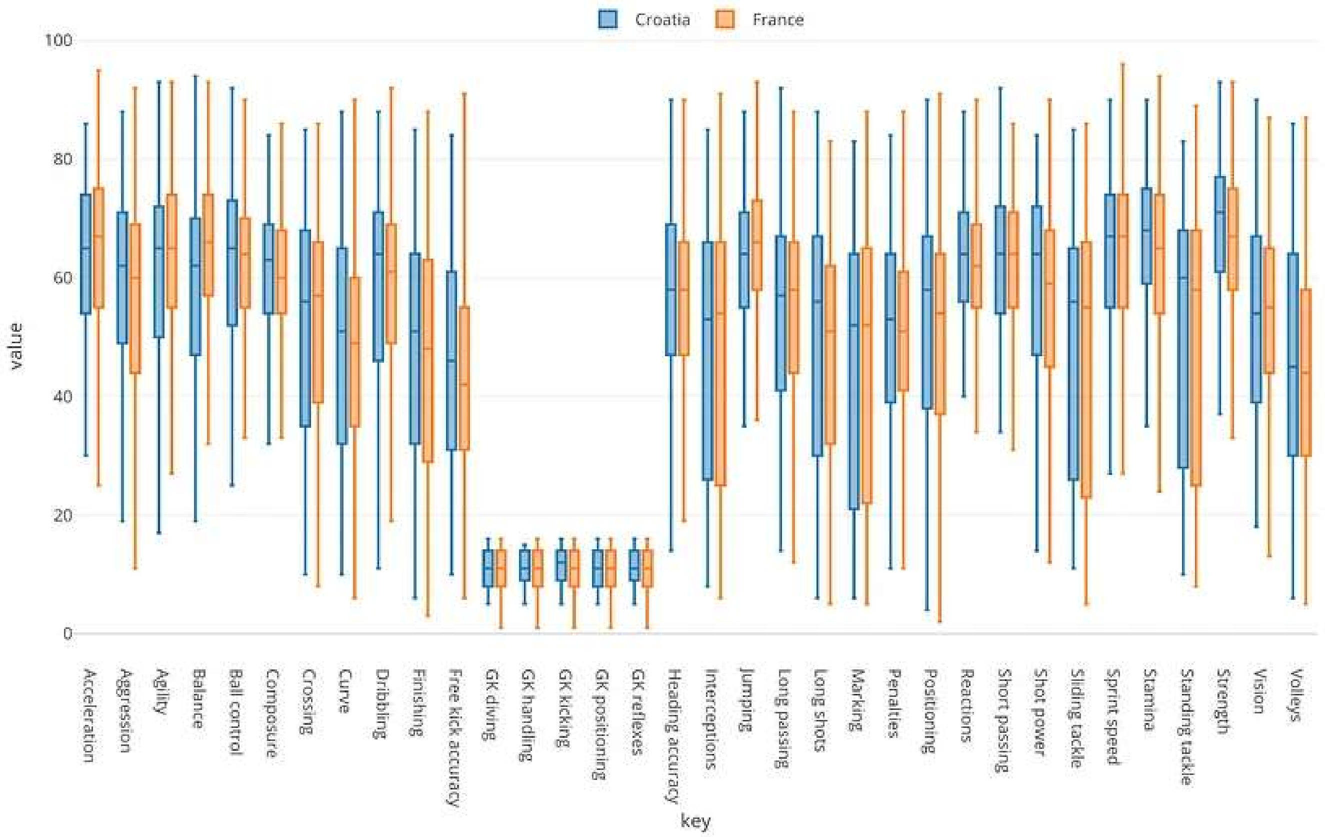
There are much less Croatian players compare to the French.

Now, the Croatian players are scoring well especially on Crossing, Ball Control, Vision, Free Kick, and Accuracy, which are at the left hand side.



And the French players are scoring well on the Forward skills such as Finishing, Volleys, Positioning, Dribbling, Long Shot, and the Defense skills such as Jumping, Strength, Sliding Tackle, Standing Tackle, etc.

Here’s a Boxplot chart comparing the distributions of the measures between the two countries.



We can see that the top players from France are scoring better than the top players from Croatia. But when we compare the 50% of the players at the middle between Croatia and France, the Croatian players are scoring better.

So it looks like both teams have a good chance to win the championship!

Summary

Having PCA in your toolbox will give you an effective and an intuitive way to understand the relationships that are hidden in the data.

In this post, by using the PCA algorithm, we could not only find out which skill measures are similar to one another, but also visualize how the soccer players are scoring on those measures and compare them by their countries and their field positions.

Happy PCA!

Try it for yourself!

If you want to try this out quickly, you can download the data from here, import it in Exploratory Desktop, and follow the steps.

If you don’t have Exploratory Desktop yet, you can sign up from here for 30 days free trial!

Exploratory

Machine Learning, AI, Statistics algorithms are not just for Data Scientists or Engineers. You need them to find hidden…

exploratory.io

Learn Data Science without Programming

If you are interested in learning various powerful Data Science methods ranging from Machine Learning, Statistics, Data Visualization, and Data Wrangling without programming, go visit our Booster Training home page and enroll today!

Data Science Booster Training

Data Science is not just for Data Scientists. It is for Everybody. Start learning Data Science without Programming!

exploratory.io

Data Science Machine Learning Statistics Data Visualization