Tesina di analisi dei dati multidimensionali

Nicola Lavecchia

Matricola : 212169

Metodo di analisi: Analisi in componenti principali

L'analisi del dataset che ho scelto di usare descrive quantitativamente le statistiche delle squadre partecipanti al mondiale di calcio del 2006, cercherò di capire come variano le statistiche di ogni squadra in base alla posizione in classifica. Nel calcio alcune statistiche difficilmente determinano il piazzamento in classifica perciò sarà utile estrarre dal dataset le componenti che spiegano al meglio questo fenomeno.

Fonte: https://www.kaggle.com/iamsouravbanerjee/fifa-football-world-cup-dataset

Il dataset è costituito da 26 osservazioni e da 10 variabili.

Le squadre sono posizionate in ordine di classifica, dalla vincente (la nostra Italia) fino all'ultima classificata ovvero gli U.S.A.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Team	Games Played	Win	Draw	Loss	Goals Scored	Goals Conceded	Goal Difference	Average Ball Possession	Corner Kiks	Total shoots	
2	Italy	7	5	2	0	12	2	10	58	42	74	
3	France	7	4	3	0	9	3	6	61	40	89	
4	Germany	7	5	1	1	14	6	8	45	36	82	
5	Portugal	7	4	1	2	7	5	2	55	25	75	
6	Brazil	5	4	0	1	10	2	8	47	21	81	
7	Argentina	5	3	2	0	11	3	8	58	40	80	
8	England	5	3	2	0	6	2	4	45	36	70	
9	Ukraine	5	2	1	2	5	7	-2	50	50	69	
10	Spain	4	3	0	1	9	4	5	62	30	72	
11	Switzerlan	4	2	2	0	4	0	4	51	35	65	
12	Netherlan	4	2	1	1	3	2	1	56	34	64	
13	Ecuador	4	2	0	2	5	4	1	47	27	54	
14	Ghana	4	2	0	2	4	6	-2	52	34	52	
15	Sweden	4	1	2	1	3	4	-1	56	31	58	
16	Mexico	4	1	1	2	5	5	0	45	40	49	
17	Australia	4	1	1	2	5	6	-1	47	28	45	
18	South Kore	3	1	1	1	3	4	-1	41	42	34	
19	Paraguay	3	1	0	2	2	2	0	48	34	32	
20	Ivory Coas	3	1	0	2	5	6	-1	53	35	40	
21	Czech Rep	3	1	0	2	3	4	-1	42	32	31	
22	Poland	3	1	0	2	2	4	-2	51	35	39	
23	Croatia	3	0	2	1	2	3	-1	48	12	21	
24	Angola	3	0	2	1	1	2	-1	54	14	15	
25	Tunisia	3	0	1	2	3	6	-3	57	20	31	
26	Iran	3	0	1	2	2	6	-4	46	23	15	
27	United Sta	3	0	1	2	2	6	-4	43	18	15	

Per descrivere al meglio il dataset ho deciso di usare l'ACP poiché basa la sua utilità nel riassumere il numero di variabili fino ad ottenere la componente che maggiormente descrive il fenomeno. L'obiettivo è quello di ridurre le informazioni ridondanti senza perdere informazione, se le variabili originali sono correlate tra loro la perdita di informazione sarà minore.

Si prova quindi a massimizzare la varianza delle singole variabili tramite l'ACP cercando il peso da attribuire alle variabili di partenza col fine di concentrarle nelle componenti principali.

Si visualizzano ora, attraverso il comando del software R "summary()", le informazioni principali sul dataset:

```
 Games.Played        Win              Draw             Loss        Goals.Scored
Min.   :3.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   : 1.000
1st Qu.:3.000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.: 3.000
Median :4.000   Median :1.500   Median :1.000   Median :1.500   Median : 4.500
Mean   :4.231   Mean   :1.885   Mean   :1.038   Mean   :1.308   Mean   : 5.269
3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.: 6.750
Max.   :7.000   Max.   :5.000   Max.   :3.000   Max.   :2.000   Max.   :14.000
Goals.Conceded Goal.Difference  Average.Ball.Possession  Corner.Kiks
Min.   :0.00   Min.   :-4.000   Min.   :41.00           Min.   :12.00
1st Qu.:2.25   1st Qu.:-1.000   1st Qu.:46.25           1st Qu.:25.50
Median :4.00   Median :-0.500   Median :50.50           Median :34.00
Mean   :4.00   Mean   : 1.269   Mean   :50.69           Mean   :31.31
3rd Qu.:6.00   3rd Qu.: 4.000   3rd Qu.:55.75           3rd Qu.:36.00
Max.   :7.00   Max.   :10.000   Max.   :62.00           Max.   :50.00
 Total.shoots
Min.   :15.0
1st Qu.:32.5
Median :53.0
Mean   :52.0
3rd Qu.:71.5
Max.   :89.0
```

Si effettuano ora i controlli preliminari:

- 1° controllo: Le variabili sono di tipo quantitativo
- 2° controllo: Relazione lineare tra tutte le variabili

  Calcolando la matrice di varianza e covarianza si riesce ad individuare la relazione tra le variabili invarianti alle trasformazioni di scala.

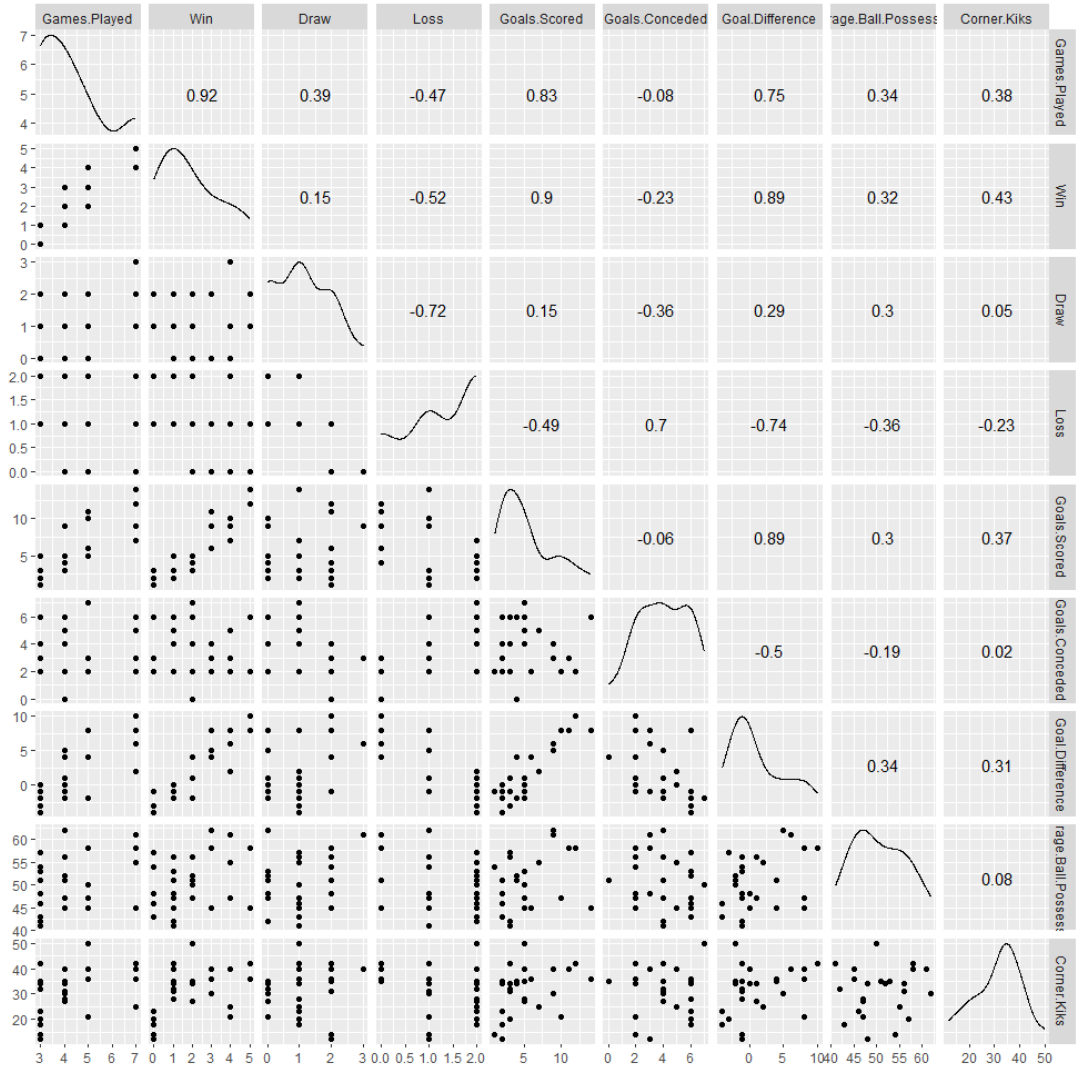
```
                      Games.Played        Win       Draw       Loss
Games.Played             1.9446154  1.9876923  0.4707692 -0.5138462
Win                      1.9876923  2.4261538  0.2046154 -0.6430769
Draw                     0.4707692  0.2046154  0.7584615 -0.4923077
Loss                    -0.5138462 -0.6430769 -0.4923077  0.6215385
Goals.Scored             4.0553846  4.9523077  0.4692308 -1.3661538
Goals.Conceded          -0.2000000 -0.6400000 -0.5600000  1.0000000
Goal.Difference          4.2553846  5.5923077  1.0292308 -2.3661538
Average.Ball.Possession  2.7938462  2.9230769  1.5723077 -1.7015385
Corner.Kiks              4.9261538  6.1969231  0.4276923 -1.6984615
Total.shoots            27.0000000 32.5600000  4.3200000 -9.8800000
                      Goals.Scored Goals.Conceded Goal.Difference
Games.Played             4.0553846          -0.20         4.255385
Win                      4.9523077          -0.64         5.592308
Draw                     0.4692308          -0.56         1.029231
Loss                    -1.3661538           1.00        -2.366154
Goals.Scored            12.3646154          -0.36        12.724615
Goals.Conceded          -0.3600000           3.28        -3.640000
Goal.Difference         12.7246154          -3.64        16.364615
Average.Ball.Possession  6.2061538          -2.04         8.246154
Corner.Kiks             12.0338462           0.36        11.673846
Total.shoots            66.4800000          -9.24        75.720000
                      Average.Ball.Possession Corner.Kiks Total.shoots
Games.Played                         2.793846   4.9261538        27.00
Win                                  2.923077   6.1969231        32.56
Draw                                 1.572308   0.4276923         4.32
Loss                                -1.701538  -1.6984615        -9.88
Goals.Scored                         6.206154  12.0338462        66.48
Goals.Conceded                      -2.040000   0.3600000        -9.24
Goal.Difference                      8.246154  11.6738462        75.72
Average.Ball.Possession             35.101538   4.5784615        59.24
Corner.Kiks                          4.578462  85.4215385       114.52
Total.shoots                        59.240000 114.5200000       532.72
>
```

- 3° controllo: Correlazione almeno moderata tra le variabili
  Si scelgono le variabili correlate fra di loro per ridurre la perdita di informazioni

```
> cor(Mondiali)
                       Games.Played        Win        Draw       Loss
Games.Played             1.00000000  0.9151096  0.38763629 -0.4673930
Win                      0.91510964  1.0000000  0.15083855 -0.5236845
Draw                     0.38763629  0.1508385  1.00000000 -0.7170275
Loss                    -0.46739305 -0.5236845 -0.71702754  1.0000000
Goals.Scored             0.82703690  0.9041868  0.15322492 -0.4928052
Goals.Conceded          -0.07919107 -0.2268736 -0.35504575  0.7003725
Goal.Difference          0.75434334  0.8875217  0.29214138 -0.7419183
Average.Ball.Possession  0.33816025  0.3167510  0.30472472 -0.3642879
Corner.Kiks              0.38221473  0.4304604  0.05313503 -0.2330977
Total.shoots             0.83887530  0.9056820  0.21491536 -0.5429670
                       Goals.Scored Goals.Conceded Goal.Difference
Games.Played             0.82703690    -0.07919107       0.7543433
Win                      0.90418677    -0.22687356       0.8875217
Draw                     0.15322492    -0.35504575       0.2921414
Loss                    -0.49280522     0.70037248      -0.7419183
Goals.Scored             1.00000000    -0.05652952       0.8945435
Goals.Conceded          -0.05652952     1.00000000      -0.4968343
Goal.Difference          0.89454348    -0.49683429       1.0000000
Average.Ball.Possession  0.29789908    -0.19012103       0.3440610
Corner.Kiks              0.37028017     0.02150710       0.3122320
Total.shoots             0.81912693    -0.22104747       0.8109764
                       Average.Ball.Possession Corner.Kiks Total.shoots
Games.Played                        0.33816025  0.38221473    0.8388753
Win                                 0.31675100  0.43046044    0.9056820
Draw                                0.30472472  0.05313503    0.2149154
Loss                               -0.36428785 -0.23309772   -0.5429670
Goals.Scored                        0.29789908  0.37028017    0.8191269
Goals.Conceded                     -0.19012103  0.02150710   -0.2210475
Goal.Difference                     0.34406101  0.31223197    0.8109764
Average.Ball.Possession             1.00000000  0.08361281    0.4332140
Corner.Kiks                         0.08361281  1.00000000    0.5368440
Total.shoots                        0.43321399  0.53684405    1.0000000
```

Si noti che le variabili come vittorie, goal fatti e tiri totali siano fortemente correlati al numero di partite giocate, infatti una squadra vincendo arriverà nelle fasi finali del torneo e giocherà più partite, inoltre giocando di più avrà la possibilità di tirare di più e di conseguenza fare più goal.
Si può invece notare che variabili come possesso palla o calci d'angolo abbiano poco a che fare con il numero di partite giocate in quanto si tratta di variabili prese in media.

Si visualizza ora lo scatterplot che evidenzia la correlazione tra le variabili originarie



La correlazione tra le variabili originarie ci suggerisce che è possibile andare avanti con l'analisi cercando di ridurre il numero di variabili eliminando la ridondanza che si crea tra queste ultime.

Ho deciso di condurre l'analisi utilizzando le variabili standardizzate per una serie di motivi:

- Le varianze delle variabili originarie presentano alcune importanti differenze di scala, si veda la varianza dei goal fatti contro quella dei goal concessi.
- Le variabili con varianza maggiore tendono a dominare nelle componenti principali più importanti.
- In questo modo le variabili originarie saranno ugualmente importanti nell'analisi che si va ad affrontare

Andiamo a calcolare su R le componenti principali tramite il comando princomp, nel caso delle variabili standardizzate poniamo l'argomento cor = TRUE e andiamo a visualizzare le componenti tramite il comando summary, visualizziamo insieme i pesi attribuiti alle componenti tramite l'argomento loadings = TRUE

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4    Comp.5     Comp.6
Standard deviation     2.3405475 1.3281484 0.93476874 0.89661517 0.8255120 0.49465356
Proportion of Variance 0.5478163 0.1763978 0.08737926 0.08039188 0.0681470 0.02446821
Cumulative Proportion  0.5478163 0.7242141 0.81159337 0.89198524 0.9601322 0.98460046
                           Comp.7      Comp.8 Comp.9 Comp.10
Standard deviation     0.32443291 0.220768388      0       0
Proportion of Variance 0.01052567 0.004873868      0       0
Cumulative Proportion  0.99512613 1.000000000      1       1
```

```
Loadings:
                      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
Games.Played           0.376  0.179  0.155         0.334  0.499  0.315         0.465
Win                    0.397  0.212         0.159         0.238  0.128  0.517 -0.519
Draw                   0.185 -0.489  0.304 -0.394  0.515                -0.281 -0.290
Loss                  -0.323  0.440  0.127  0.133         0.361  0.275 -0.594 -0.263
Goals.Scored           0.375  0.248         0.199  0.167 -0.516  0.111 -0.255  0.373
Goals.Conceded        -0.168  0.545  0.446 -0.144  0.369 -0.354 -0.204  0.222 -0.192
Goal.Difference        0.401        -0.218  0.238        -0.290  0.188 -0.321 -0.430
Average.Ball.Possession  0.200 -0.160  0.766  0.146 -0.545         0.151
Corner.Kiks            0.198  0.272 -0.155 -0.814 -0.371         0.233
Total.shoots           0.391  0.176               -0.146  0.273 -0.801 -0.276
                      Comp.10
Games.Played           0.349
Win                   -0.390
Draw                  -0.218
Loss                  -0.198
Goals.Scored          -0.497
Goals.Conceded         0.256
Goal.Difference        0.572
Average.Ball.Possession
Corner.Kiks
Total.shoots
```

I loadings sono i coefficienti applicati alle variabili originarie per determinare le componenti principali.

Essi sono preimpostati con un cut off pari a .1 se non si specifica nulla, sono infatti stati eliminati i pesi inferiori a 0.1

Dall'output si nota come le prime 3 componenti spiegano più dell'81% della variabilità totale che non è una percentuale molto alta ma comunque accettabile.

Analizziamo meglio le prime 3 componenti con un cutoff superiore, per esempio cutoff = .3 andando a nascondere i pesi con valore assoluto inferiore a 0.3.

```
Loadings:
                        Comp.1 Comp.2 Comp.3
Games.Played             0.376
Win                      0.397
Draw                            -0.489  0.304
Loss                    -0.323  0.440
Goals.Scored             0.375
Goals.Conceded                   0.545  0.446
Goal.Difference          0.401
Average.Ball.Possession                 0.766
Corner.Kiks
Total.shoots             0.391
```

Cerchiamo di capire cosa rappresentano le prime tre CP cercando di “assegnargli un nome”.

Nella prima CP sembrano essere coinvolte le squadre della parte medio/alta della classifica in quanto si vedono unità come partite giocate, vittorie, sconfitte(in negativo), tiri totali e differenza reti.

Nella seconda CP vediamo alti valori per quanto riguarda sconfitte e goal subiti, perciò questa CP è dominata dalle squadre della parte bassa della classifica.

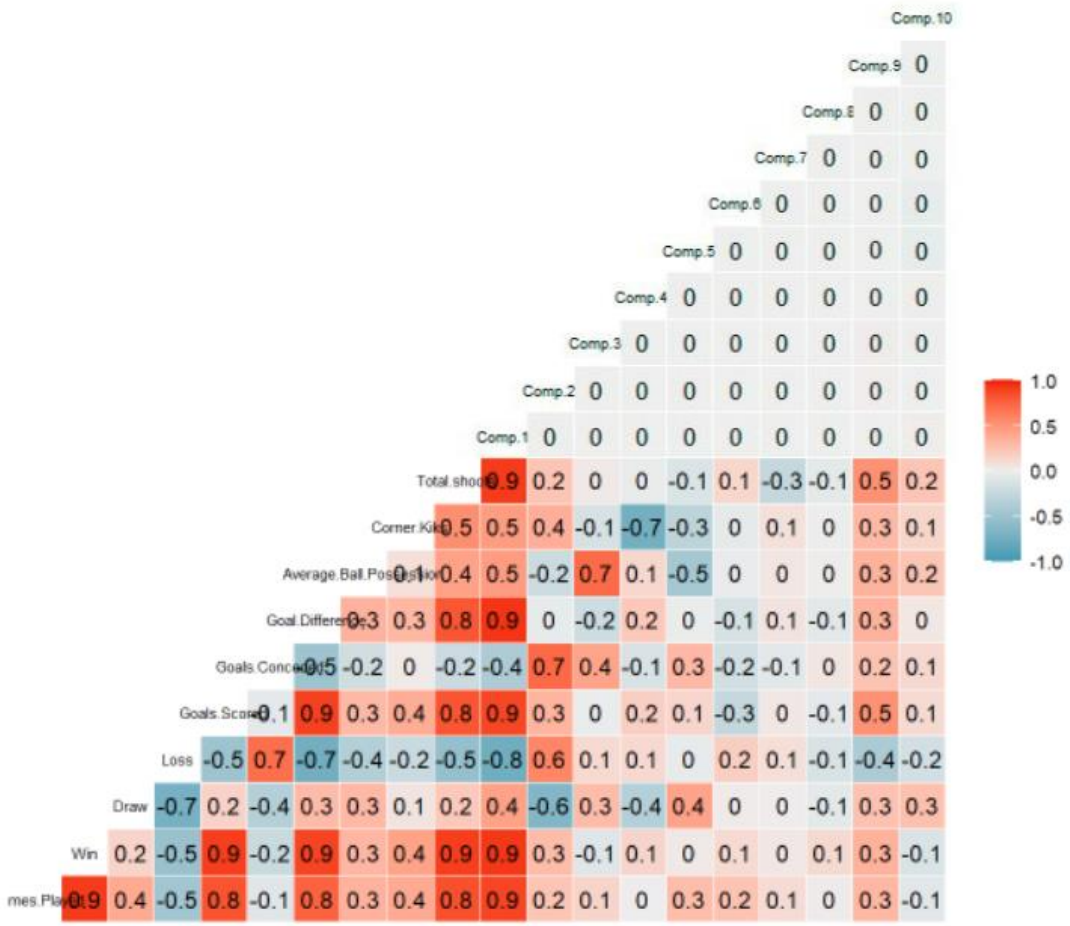
La terza CP riassume quasi sicuramente le squadre con un buon gioco(possesso palla) ma con una difesa mediocre avendo subito abbastanza goal, squadre che solitamente finiscono nella parte medio/bassa della classifica.

Cerchiamo ora una conferma di quanto detto sopra tramite gli score o punteggi, essi sono le coordinate dei punti originali proiettate sulle componenti principali

Al fine di interpretare le componenti principali stimate calcoliamo la correlazione dei punteggi delle PC con ciascuna delle variabili originarie.

```
> cor(Mondiali, Mondiali1.PCA$scores)
                            Comp.1      Comp.2      Comp.3       Comp.4      Comp.5
Games.Played             0.8800914  0.23837556  0.14477934  0.005336358  0.27581319
Win                      0.9291645  0.28129556 -0.08902674  0.142216594  0.02696651
Draw                     0.4324410 -0.64997754  0.28389333 -0.353404450  0.42496545
Loss                    -0.7567524  0.58389276  0.11837134  0.118854510 -0.03486206
Goals.Scored             0.8778315  0.32961887 -0.01989546  0.178317932  0.13754128
Goals.Conceded          -0.3934090  0.72433965  0.41670583 -0.129478943  0.30450026
Goal.Difference          0.9391702 -0.03776818 -0.20385172  0.212967571 -0.01676805
Average.Ball.Possession  0.4677113 -0.21189116  0.71610633  0.131350872 -0.45019774
Corner.Kiks              0.4636964  0.36062160 -0.14495110 -0.729901490 -0.30605748
Total.shoots             0.9161923  0.23438822  0.02973885 -0.026779035 -0.12079136
                            Comp.6       Comp.7       Comp.8     Comp.9      Comp.10
Games.Played             0.24672377  0.102320617  0.014551015  0.3115492 -0.07625414
Win                      0.11791240  0.041500413  0.114114088  0.3448197 -0.12601724
Draw                     0.02247780  0.008969136 -0.062068554  0.2534331  0.25926347
Loss                     0.17861683  0.089085333 -0.131153983 -0.4101533 -0.17230568
Goals.Scored            -0.25531423  0.036116781 -0.056238296  0.5193693  0.11247345
Goals.Conceded          -0.17509950 -0.066046642  0.049096794  0.2421979  0.12162101
Goal.Difference         -0.14353667  0.060962857 -0.070864797  0.3430230  0.04331656
Average.Ball.Possession -0.03367041  0.049094640  0.008058547  0.3454994  0.22889571
Corner.Kiks             -0.04122419  0.075693625 -0.008400682  0.2986694  0.14056635
Total.shoots             0.13496618 -0.259856531 -0.060995729  0.5233099  0.23508796
```

Confermiamo graficamente i dati con l'uso della funzione ggcorr()



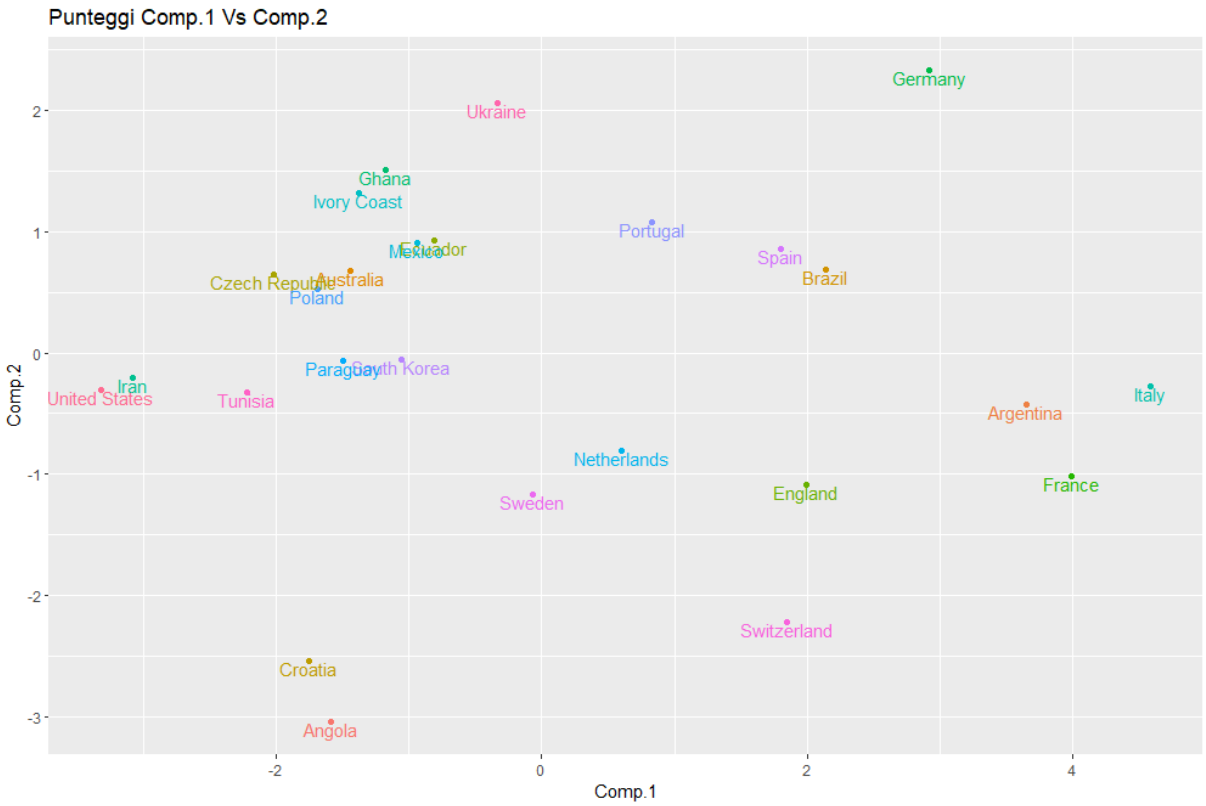
La tabella mostra che la prima CP è fortemente correlata, anche negativamente, con tutte le variabili, potremmo dire che rappresenta quasi completamente le squadre arrivate nelle prime posizioni, vediamo infatti dei valori positivi per quanto riguarda i tiri totali, i goal fatti e di conseguenza la differenza reti, vittorie e partite giocate, tutte variabili che caratterizzano in assoluto le squadre vincenti.

La seconda CP potrebbe invece rappresentare le squadre della parte bassa della classifica, vediamo infatti valori positivi per quanto riguarda sconfitte e goal concessi.

La terza CP sembra rappresentare le squadre che conducono uno sterile possesso di palla, sterile nel senso che non viene concretizzato come si può notare dallo score 0 della riga goal fatti, queste squadre nonostante siano riuscite a tenere palla durante le partite non sono riuscite a vincere molte partite, già nel dataset si poteva infatti notare come anche la zona medio/bassa della classifica abbondasse di squadra con un buon possesso palla.

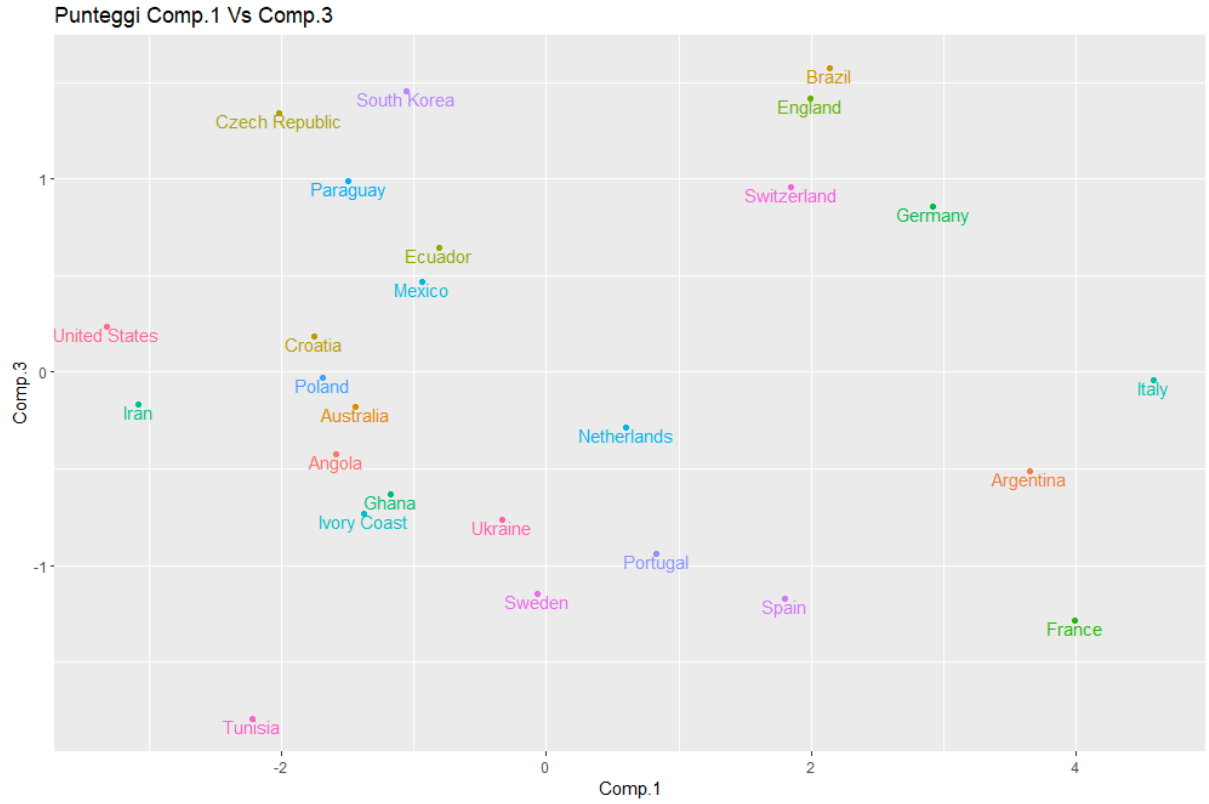
Si può notare una sorta di conferma dell'analisi dei pesi attraverso gli scores che vanno ad evidenziare letteralmente le stesse caratteristiche.

Si riproduce ora lo scatterplot delle prime 3 CP utlizzando la funzione ggplot()

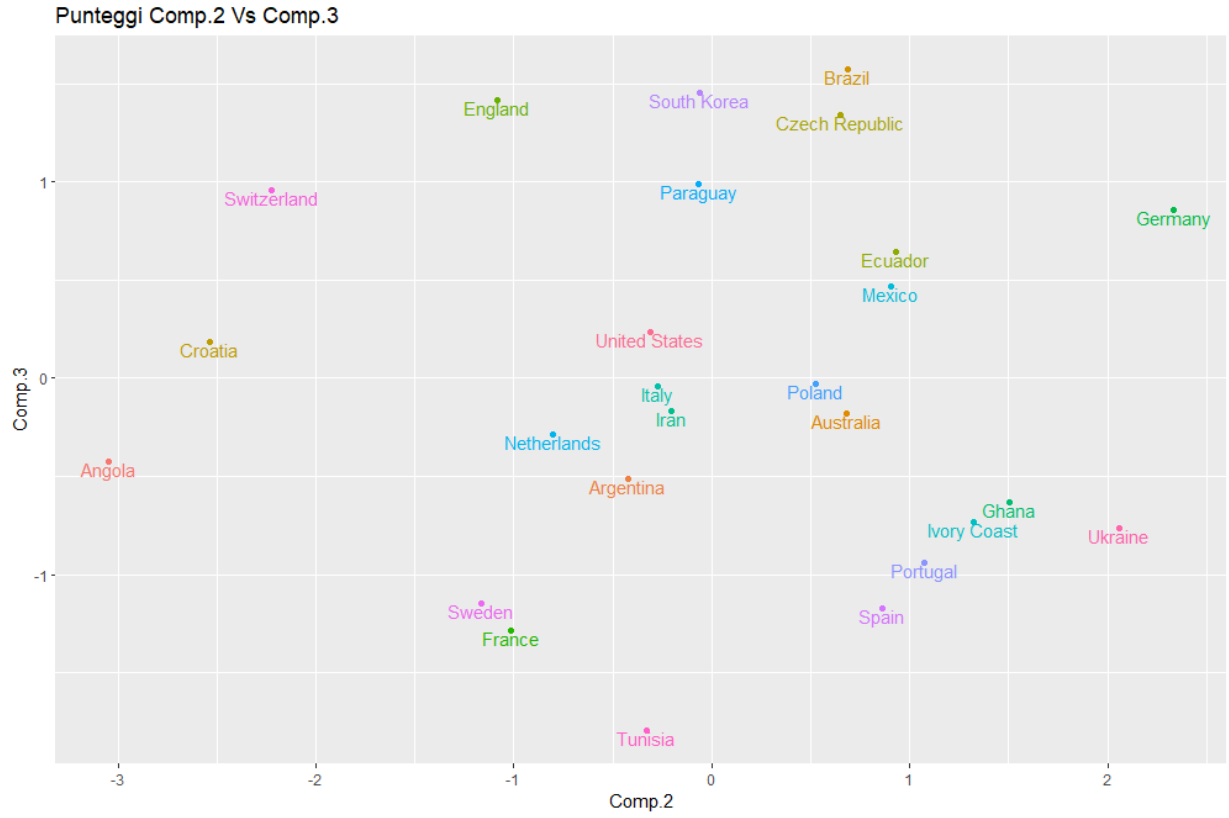


Vediamo come Italia, Francia e altre squadre della parte alta della classifica vengano rappresentate esclusivamente dalla prima componente principale come abbiamo detto fino a questo punto, c'è però un'anomalia ovvero la Germania che viene rappresentata sia dalla prima che dalla seconda componente, ciò si spiega per il fatto che questa squadra è stata

l’unica ad aver giocato tutte le partite a disposizione(7), avendo subito 6 goal, un numero che si allinea alle squadre di bassa classifica che però hanno subito quei goal in poche partite essendo subito state eliminate dalla competizione.



Si nota la moltitudine di squadre rappresentate in negativo dalla prima componente e in positivo dalla terza, sono quelle squadre di medio/bassa classifica che pur giocando mediocremente sono riuscite a portare a casa qualche pareggio e non solo sconfitte, sono ben rappresentate in questo grafico anche le squadre finite ultime che si relazionano molto negativamente con la prima CP e quasi per nulla con la terza, eccezion fatta per la Tunisia che viene rappresentata negativamente anche dalla terza componente poiché in relazione alle partite giocate ha fatto abbastanza goal da non rientrare tra le squadre che giocano senza concretizzare. Finisce comunque tra le ultime per i molti goal concessi.

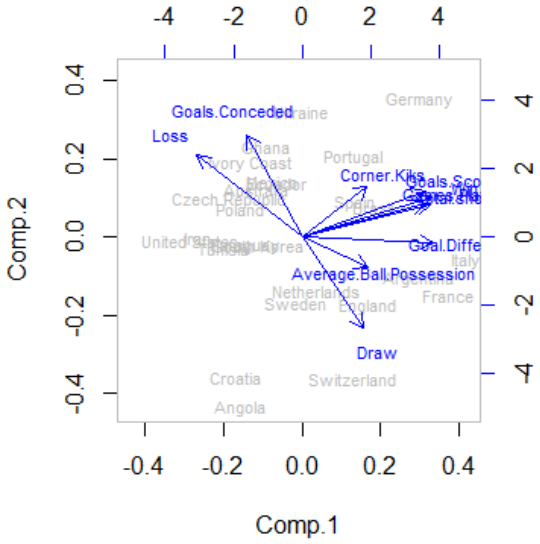


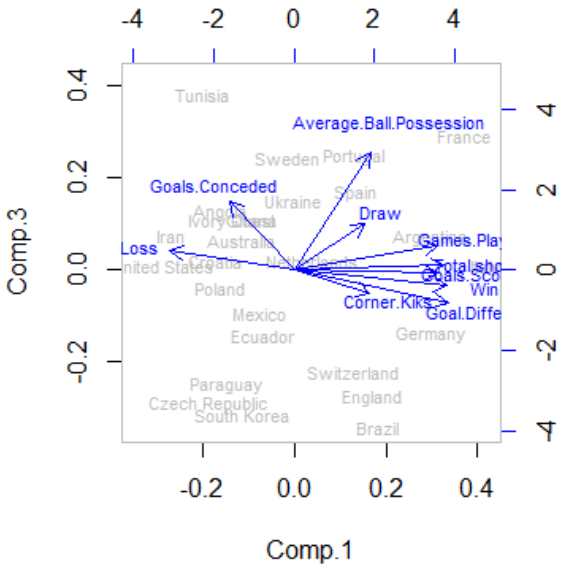
Stavolta le squadre di alta classifica sono, giustamente, nella parte centrale del grafico, il ché sta a significare che non sono rappresentate da seconda e terza CP, fa nuovamente eccezione la Germania sempre a causa dei tanti goal subiti. Il grafico, stavolta molto dispersivo, ci restituisce le medesime informazioni che sono state dette fino ad ora.

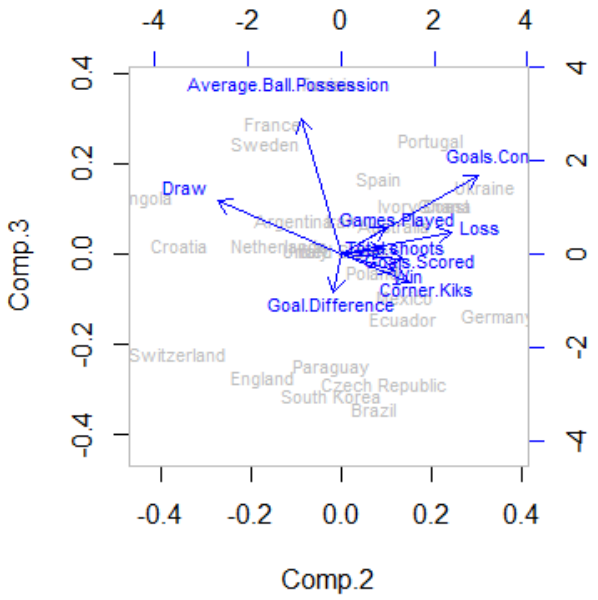
Un'altra rappresentazione grafica dei risultati di una PCA in cui sono rappresentati sia gli score sia i loading è il biplot nel quale troviamo le CP sotto forma di trasformazioni lineari delle variabili originarie.

La particolarità di questi grafici si nota guardando le unità vicino l'origine che rappresentano valori prossimi alle medie, al contrario i punti lontani rappresentano valori che si discostano dalla media.

Si mettono in confronto le prime tre componenti prese a due a due.


Comp.2
Comp.1
Goals.Conceded
Loss
Corner.Kiks
Goal.Diffe
Average.Ball.Possession
Draw
Germany
Portugal
Italy
France
Netherlands
Sweden
England
Croatia
Switzerland
Angola
Poland
Czech Republic


Comp.3
Comp.1
Tunisia
Average.Ball.Possession
France
Sweden
Portugal
Goals.Conceded
Ukraine
Spain
Draw
Iran
Australia
Loss
Games.Pla
Win
Poland
Corner.Kiks
Goal.Diffe
Mexico
Ecuador
Germany
Switzerland
Paraguay
Czech Republic
South Korea
England
Brazil



Al fine di un'analisi più approfondita ho preferito tenere inizialmente 3 CP per poi decidere quale fosse effettivamente il numero corretto da tenere in considerazione per una ACP più efficace.

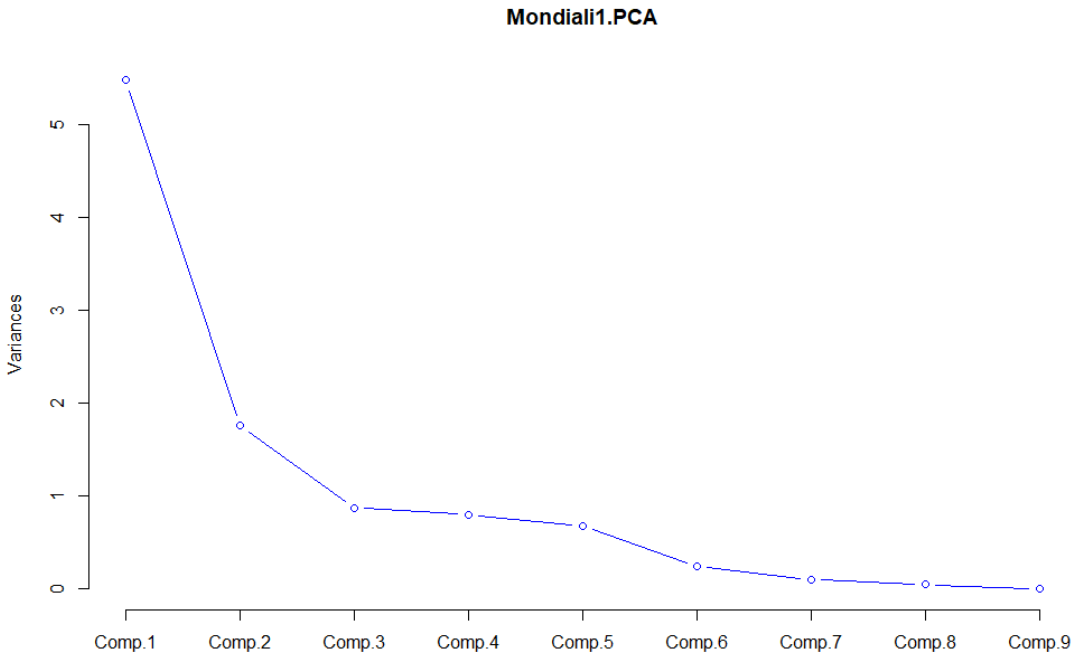
Si cerca di risolvere il principale dubbio, ovvero il numero di CP da tenere in considerazione. Vediamo tra i principali criteri quali sono soddisfatti:

- Varianza Spiegata: si sceglie il numero di componenti in grado di riprodurre una grande percentuale di variabilità complessiva riferita alle variabili originarie. Sono suggeriti valori compresi tra il 70% e il 90%

  In questo caso prendendo 2 CP si spiega circa il 72% di variabilità, prendendone 3 invece si spiega circa l'81%

- Scree plot: Un grafico che rappresenta in senso decrescente la porzione di variabilità, o meglio gli autovalori(asse y) spiegata dalle CP (asse x), se c'è molta differenza tra i primi autovalori e i successivi allora questo grafico mostrerà un gomito tra k e k+1 componenti principali, seguendo questo criterio sarebbe da scegliere la componente k.

  Il gomito si presenta tra la seconda e la terza CP perciò andrebbe scelta la seconda.



- Regola di Kaiser: Regola che si applica solamente alle variabili standardizzate. Si considerano solamente gli autovalori maggiori di 1 poiché secondo questa regola una CP dovrebbe spiegare almeno una variazione pari al valore medio di una singola variabile standardizzata. Secondo alcuni studiosi il limite di 1 è un po' alto e potrebbe essere opportuno un limite di 0.7.

  Anche secondo la regola di Kaiser il numero di CP da tenere in considerazione per questo data set è 2.

  Essendo l'obiettivo della ACP quello di riassumere al meglio l'informazione ottenuta dai dati originali è preferibile condurre questa analisi eliminando la terza componente e lasciandone solo 2.
  Nel caso di questo data set eliminiamo la componente che risultava essere un po' più confusionaria, nel senso che non dava la certezza di descrivere una precisa porzione della classifica finale. Le prime due CP infine descrivono rispettivamente e quasi perfettamente le squadre in alta e in bassa classifica.