Tesina Individuale

Descrizione Dataset

Il dataset selezionato è bfi: 25 elementi di personalità che rappresentano 5 fattori. Contiene i dati di 2800 soggetti. I primi 25 elementi sono organizzati da cinque fattori putativi: Agreeableness (gradevolezza), Conscientiousness (coscienziosità), Extraversion (estroversione), Neuroticism (nevroticismo) e Openness (apertura). I dati sono stati raccolti utilizzando una scala di risposta a 6 punti: 1 Very Inaccurate (Molto impreciso), 2 Moderately Inaccurate (Moderatamente impreciso), 3 Slightly Inaccurate (Leggermente impreciso), 4 Slightly Accurate (Leggermente accurato), 5 Moderately Accurate (Moderatamente accurato), 6 Very Accurate (Molto accurato). Sono incluse anche tre variabili demografiche aggiuntive (sesso, istruzione ed età).

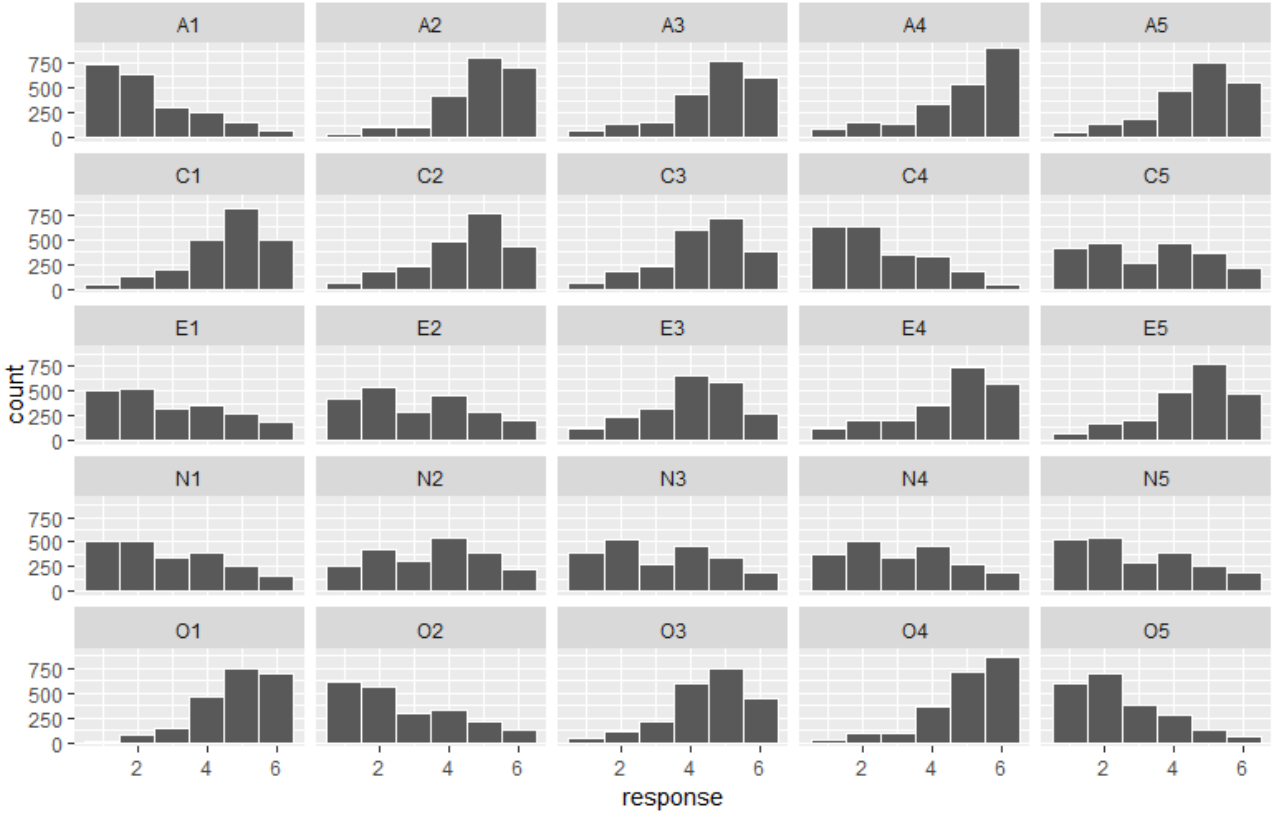
E' possibile esaminare la struttura del data frame utilizzando la funzione str():

```
> str(bfi)
'data.frame':   2800 obs. of  28 variables:
 $ A1       : int  2 2 5 4 2 6 2 4 4 2 ...
 $ A2       : int  4 4 4 4 3 6 5 3 3 5 ...
 $ A3       : int  3 5 5 6 3 5 5 1 6 6 ...
 $ A4       : int  4 2 4 5 4 6 3 5 3 6 ...
 $ A5       : int  4 5 4 5 5 5 5 1 3 5 ...
 $ C1       : int  2 5 4 4 4 6 5 3 6 6 ...
 $ C2       : int  3 4 5 4 4 6 4 2 6 5 ...
 $ C3       : int  3 4 4 3 5 6 4 4 3 6 ...
 $ C4       : int  4 3 2 5 3 1 2 2 4 2 ...
 $ C5       : int  4 4 5 5 2 3 3 4 5 1 ...
 $ E1       : int  3 1 2 5 2 2 4 3 5 2 ...
 $ E2       : int  3 1 4 3 2 1 3 6 3 2 ...
 $ E3       : int  3 6 4 4 5 6 4 4 NA 4 ...
 $ E4       : int  4 4 4 4 4 5 5 2 4 5 ...
 $ E5       : int  4 3 5 4 5 6 5 1 3 5 ...
 $ N1       : int  3 3 4 2 2 3 1 6 5 5 ...
 $ N2       : int  4 3 5 5 3 5 2 3 5 5 ...
 $ N3       : int  2 3 4 2 4 2 2 2 2 5 ...
 $ N4       : int  2 5 2 4 4 2 1 6 3 2 ...
 $ N5       : int  3 5 3 1 3 3 1 4 3 4 ...
 $ O1       : int  3 4 4 3 3 4 5 3 6 5 ...
 $ O2       : int  6 2 2 3 3 3 2 2 6 1 ...
 $ O3       : int  3 4 5 4 4 5 5 4 6 5 ...
 $ O4       : int  4 3 5 3 3 6 6 5 6 5 ...
 $ O5       : int  3 3 2 5 3 1 1 3 1 2 ...
 $ gender   : int  1 2 2 2 1 2 1 1 1 2 ...
 $ education: int  NA NA NA NA NA 3 NA 2 1 NA ...
 $ age      : int  16 18 17 17 17 21 18 19 19 17 ...
```

Essendo i valori mancanti una piccola porzione del data frame, ho preferito eliminarli con la funzione na.omit(). Attraverso la funzione View() ho notato alcune incongruenze, risultava che alcuni soggetti avessero livelli di educazione non congrui

alla loro età. Ho creato 5 dataset separati, uno per ciascun livello di educazione, selezionando le età che ritenevo più appropriate. Infine ho ricomposto il dataset. Dopo questa pulizia, il numero totale di osservazioni è diminuito da 2800 a 2160.

Istogrammi



Analizzando gli istogrammi rispetto ai primi 25 elementi del dataset, è possibile osservare quali siano state le risposte più frequenti. La maggioranza dei soggetti, per il fattore A (gradevolezza), afferma di non essere indifferente ai sentimenti degli altri (A1), bensì si informa (A2), sa come confortare gli altri (A3) e li fa sentire a proprio agio (A5).

Secondo le risposte date rispetto al fattore C (coscienziosità), gran parte dei soggetti dichiara di essere esigente nel proprio lavoro (C1), non farebbero le cose a metà (C4), bensì le porterebbero a termine seguendo un piano (C2 & C3). Osserviamo una frequenza di risposta proporzionata alla variabile C5, la perdita di tempo sarebbe quindi una caratteristica comune fra i soggetti. Varrebbe lo stesso per chi parla poco (E1) e trova difficoltà nell'approcciarsi con altre persone (E2). Nonostante ciò, la maggior parte dei soggetti affermerebbe di saper fare amicizia facilmente (E4).

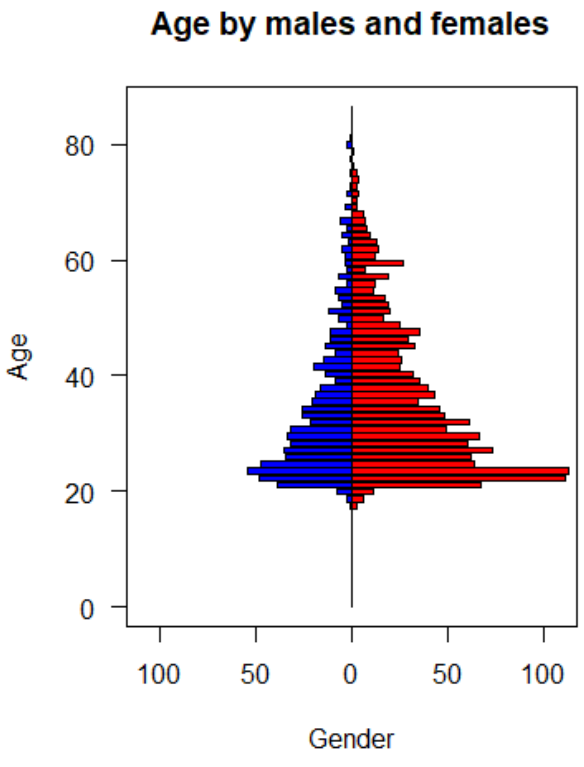
Si nota ancora una frequenza di risposta proporzionata per tutti gli elementi del fattore N (nevroticismo): poca tolleranza (N1 & N2), frequenti sbalzi di umore (N3), tristezza (N4) e attacchi di panico (N5) sono caratteristiche comuni fra i soggetti.

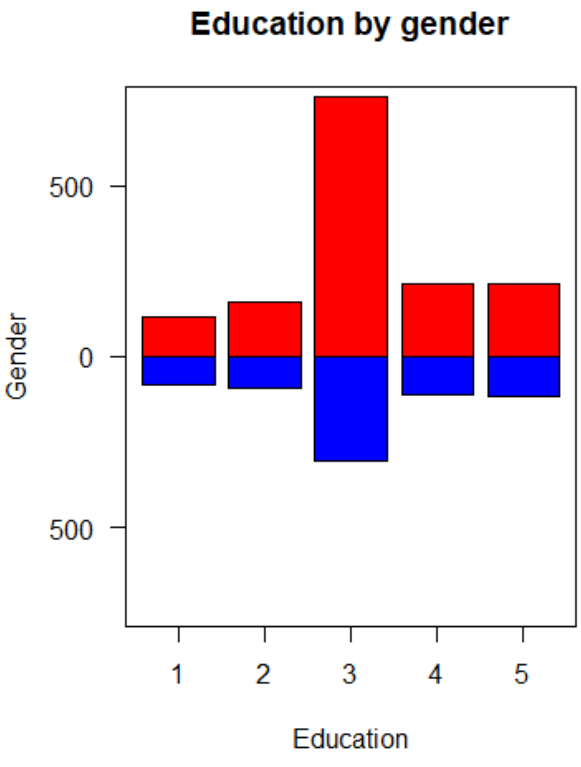
Osservando il quinto fattore (apertura), notiamo come prevalgano soggetti pieni di idee (O1), che affrontano letture complicate (O2), riflettono sulle cose (O4) e indagano

a fondo sugli argomenti (O5).

Coppie di Grafici a Barre Basati su Due Gruppi

Attraverso la libreria psych è stato possibile utilizzare la funzione bi.bars() per creare i seguenti grafici:

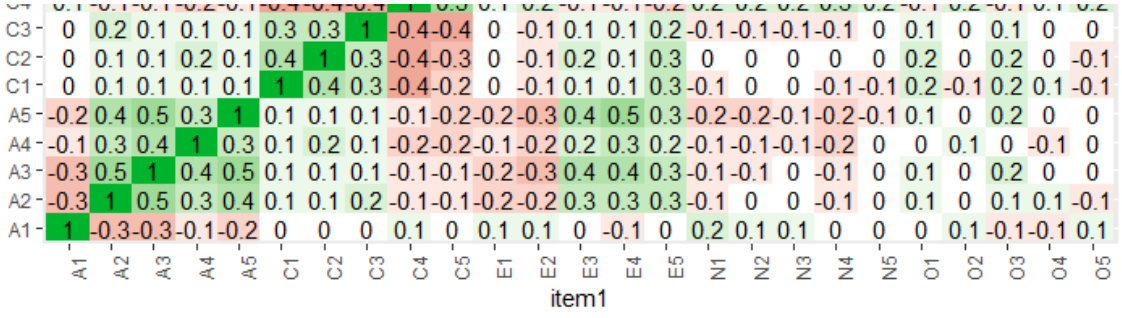




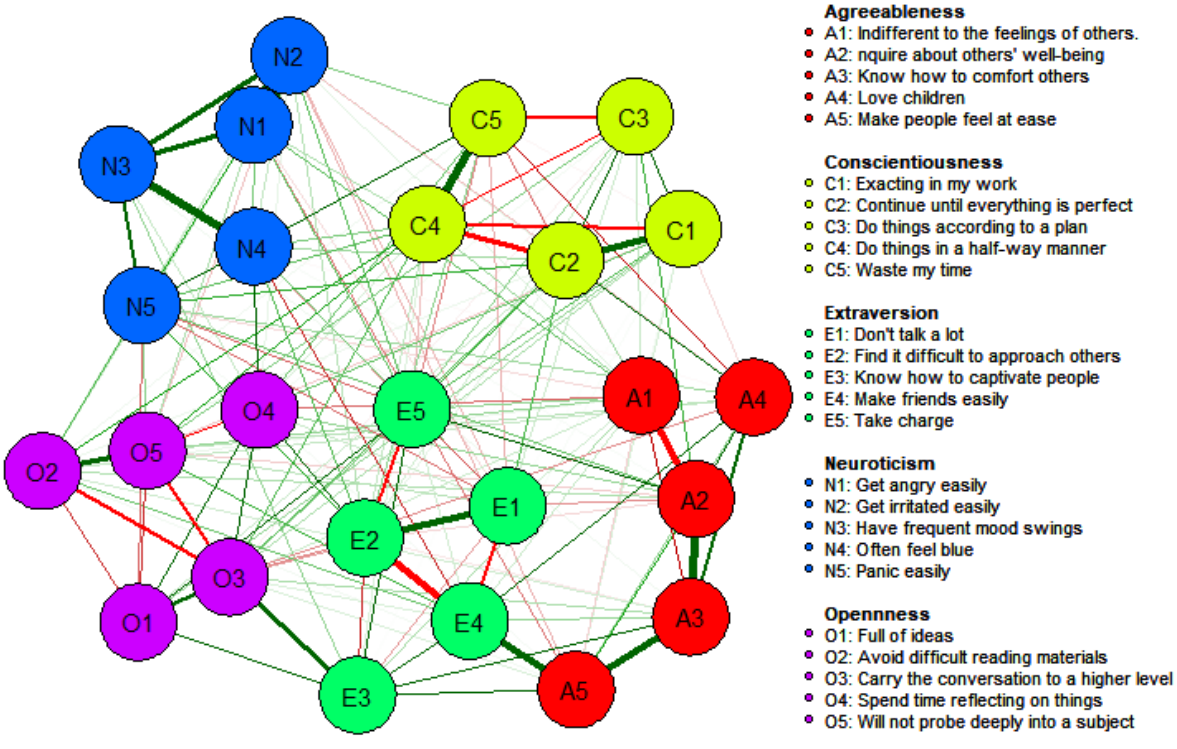
Osserviamo come la maggior parte dei soggetti che appartiene al gender "Females", abbia tra i 20 e i 30 anni e frequenti il college (livello 3 = "some college"). Lo stesso vale per il gender "Males".

Correlazione

Successivamente ho costruito un correlogramma sui primi 25 elementi.



Ho semplificato l'analisi utilizzando la libreria qgraph e ho creato un network che spiega la correlazione tra le variabili.



E' facilmente intuibile come le linee verdi rappresentino una correlazione positiva mentre quelle rosse una negativa, lo spessore indica invece quanto la correlazione sia significativa o meno.

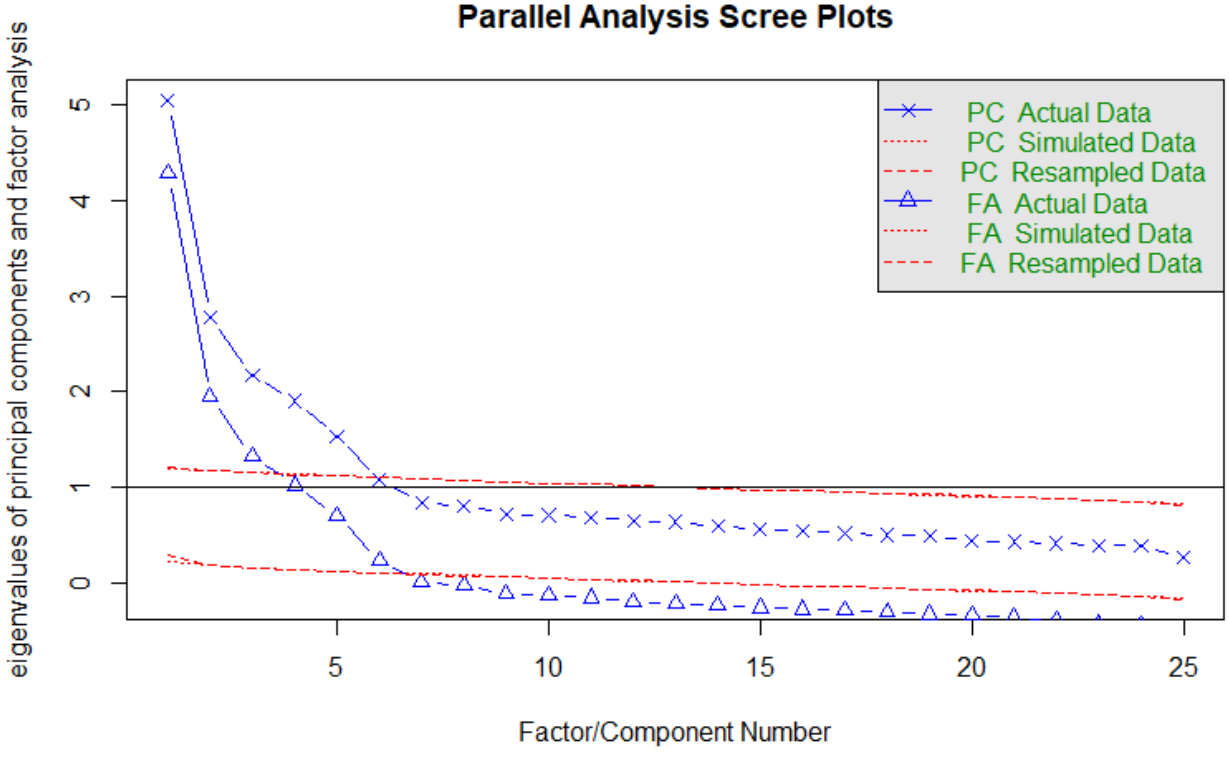
L'unico fattore interamente correlato positivamente è N (nevroticismo). Tutti gli elementi di questo fattore sono fortemente correlati positivamente a N3, ovvero i soggetti affetti da frequenti sbalzi di umore tendono ad avere una bassa tolleranza, ad essere tristi e a soffrire di attacchi di panico.

Per quanto riguarda il fattore C (coscienziosità) possiamo notare una significativa correlazione positiva tra le variabili C4 e C5 (i soggetti che tendono a fare le cose a metà sono gli stessi che temporeggiano) e tra C1 e C2 (coloro che sono esigenti nel proprio lavoro e lo portano a termine alla perfezione); naturalmente le variabili C1 e C2 sono significativamente correlate negativamente con C4.

All'interno del fattore A (gradevolezza) osserviamo una sola correlazione fortemente negativa, quella tra le variabili A1 e A2 (i soggetti che "si informano sulle altre persone" non saranno indifferenti ai sentimenti di questi ultimi). Gli elementi restanti sono correlati positivamente tra di loro (coloro che si informano sanno come confortare gli altri e li fanno sentire a proprio agio). Da notare come A5 sia significativamente correlato positivamente con l'elemento E4 (I soggetti che fanno sentire le persone a proprio agio tendono a fare amicizia facilmente). Quest'ultimo elemento risulta fortemente correlato negativamente con E1 ed E2 (coloro che parlano poco e fanno fatica ad approcciare gli altri non riescono naturalmente a fare amicizia facilmente). Infine possiamo notare la correlazione positiva fra le variabili O2 e O5 (chi evita letture complicate non indaga a fondo sugli argomenti).

Parallel Analysis

Prima di procedere con la PCA e la Factor Analysis, ho utilizzato la Parallel Analysis per determinare il numero di fattori e di componenti.

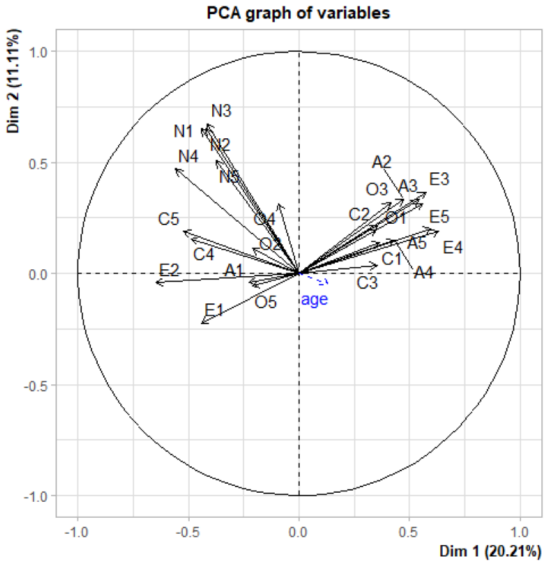


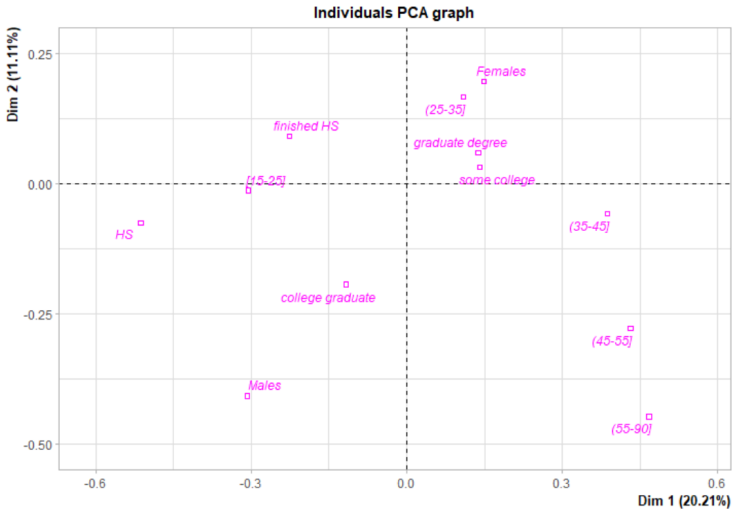
L'output indica quanti fattori/componenti è possibile estrarre in base a dove le linee blu incrociano le linee tratteggiate rosse. L'analisi parallela suggerisce 6 fattori e 5 componenti.

Principal Component Analysis

Anzitutto ho cambiato i numeri delle variabili gender ed education con i loro rispettivi valori. Successivamente ho costruito una colonna contenente i range delle età e ho

usato la funzione PCA sul nuovo dataset, specificando 5 come il numero di componenti.





Nel primo grafico è possibile riconoscere le relazioni viste precedentemente con la matrice di correlazione e il network.

Sovrapponendo il secondo grafico al primo è possibile confrontare le variabili principali con le qualitative supplementari. Il gender "Females" è quello che si avvicina di più al gruppo di caratteristiche positive: si informano (A2), sanno come confortare gli altri (A3), li fanno a sentire a proprio agio (A5), fanno amicizia più facilmente (E4), sono esigenti nel loro lavoro (C1), lo portano perfettamente a termine (C2) e seguono un piano (C3). Da notare la vicinanza al più alto livello di educazione. Al contrario, il gender "Males" non parlerebbe molto (E1), farebbe più fatica ad approcciarsi agli altri (E2), sarebbe indifferente ai sentimenti degli altri (A1) e tenderebbe ad impegnarsi di meno (O5).

A metà tra i due gender troviamo le caratteristiche del fattore N (nevroticismo), ciò sta ad indicare che i problemi emotivi non sono una caratteristica di un solo gender, bensì possono affliggere entrambi allo stesso modo. Questi problemi si riscontrano soprattutto tra i ragazzi di età tra i 15-25 anni che frequentano il liceo, i quali farebbero fatica ad approcciare con gli altri (E2).

Con l'avanzare dell'età, già a 25-35 anni, le caratteristiche tendono a migliorare: ci si impegna di più e si spreca meno tempo, si tiene di più agli altri e si trova il modo di approcciare con gli altri e fare amicizia facilmente.

Per quanto riguarda il livello di educazione, ovviamente i soggetti che sono esigenti nel loro lavoro, si impegnano e non perdono tempo sono gli stessi che frequentano il college o hanno ottenuto un graduate degree. L'unica incongruenza la si nota con il livello di "college graduate" vicino alla caratteristica O5, tipica di chi non approfondisce gli argomenti.

Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
A1	-0.227	1.016	0.051	-0.040	0.057	0.002	0.179	1.476	0.032
A2	0.473	4.436	0.224	0.336	4.062	0.113	-0.239	2.624	0.057
A3	0.543	5.827	0.294	0.333	3.983	0.111	-0.295	4.010	0.087
A4	0.432	3.695	0.187	0.150	0.806	0.022	-0.229	2.417	0.052
A5	0.596	7.041	0.356	0.201	1.456	0.040	-0.307	4.335	0.094
C1	0.365	2.634	0.133	0.139	0.693	0.019	0.511	12.051	0.262
C2	0.350	2.419	0.122	0.198	1.406	0.039	0.463	9.895	0.215
C3	0.355	2.492	0.126	0.039	0.055	0.002	0.342	5.376	0.117
C4	-0.487	4.688	0.237	0.155	0.869	0.024	-0.442	8.990	0.195
C5	-0.522	5.400	0.273	0.192	1.325	0.037	-0.251	2.907	0.063
E1	-0.443	3.888	0.196	-0.224	1.807	0.050	0.288	3.834	0.083
E2	-0.646	8.262	0.417	-0.041	0.062	0.002	0.231	2.460	0.053
E3	0.575	6.544	0.331	0.367	4.848	0.135	-0.088	0.354	0.008
E4	0.634	7.955	0.402	0.189	1.280	0.036	-0.371	6.345	0.138
E5	0.559	6.175	0.312	0.314	3.555	0.099	0.111	0.573	0.012
N1	-0.440	3.824	0.193	0.653	15.337	0.426	0.041	0.079	0.002
N2	-0.419	3.477	0.176	0.650	15.229	0.423	0.104	0.497	0.011
N3	-0.414	3.400	0.172	0.674	16.372	0.455	0.054	0.137	0.003
N4	-0.559	6.185	0.312	0.471	7.990	0.222	0.117	0.629	0.014
N5	-0.374	2.775	0.140	0.507	9.272	0.257	-0.094	0.404	0.009
O1	0.356	2.513	0.127	0.216	1.674	0.046	0.343	5.422	0.118
O2	-0.211	0.885	0.045	0.112	0.454	0.013	-0.462	9.842	0.214
O3	0.415	3.413	0.172	0.320	3.691	0.103	0.271	3.385	0.073
O4	-0.091	0.165	0.008	0.316	3.599	0.100	0.291	3.897	0.085
O5	-0.212	0.892	0.045	-0.057	0.118	0.003	-0.418	8.062	0.175

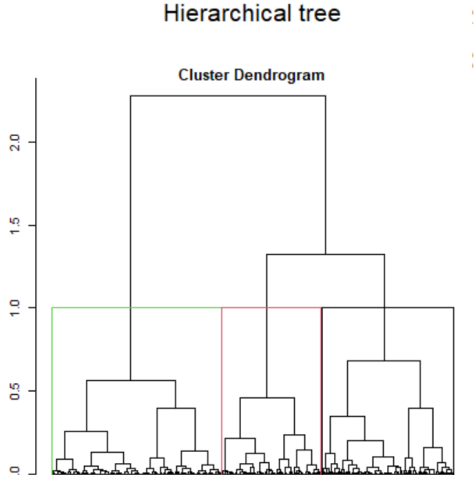
La prima componente principale aumenta con l'aumento dei punteggi di A3 (sapere come confortare gli altri), A5 (fare sentire le persone a loro agio), E3 (sapere come affascinare le persone), E4 (fare amicizia facilmente), E5 (prendere in mano [una situazione]) e la diminuzione dei punteggi di C5 (perdere tempo), E2 (trovare difficoltà nell'approcciare con gli altri), N4 (sentirsi tristi). Ciò suggerisce che i primi cinque criteri variano insieme. Se uno aumenta, anche i restanti tendono ad aumentare. Vale lo stesso per gli altri tre criteri.

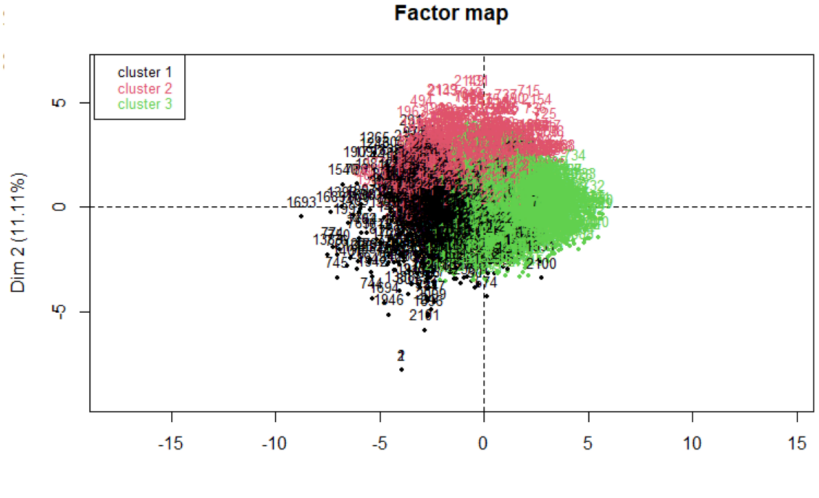
La seconda componente principale aumenta con l'aumento di 4 valori: N1 (arrabbiarsi facilmente), N2 (irritarsi facilmente), N3 (avere frequenti cambiamenti di umore) e N5 (provare panico facilmente).

La terza componente principale aumenta con uno solo dei valori, l'aumento di C1 (essere esigenti nel proprio lavoro).

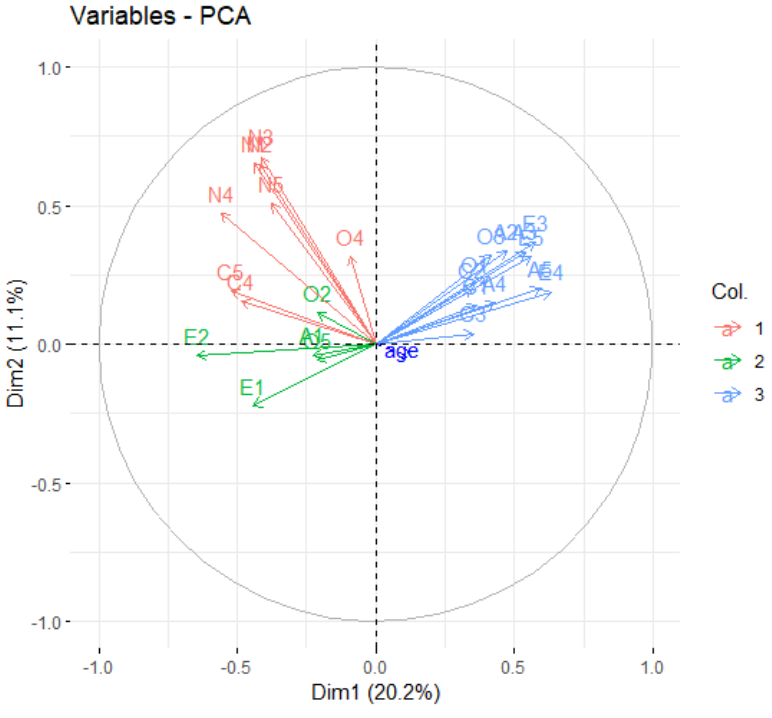
Hierarchical Clustering on Principal Components

Successivamente a PCA, ho utilizzato HCPC per vedere il numero di clusters. Impostando nb.clust=-1, l'albero viene tagliato automaticamente al livello suggerito.





Il numero di clusters proposto è 3. Impostando un seed casuale, è stato possibile osservare la possibile partizione direttamente sul grafico della PCA.

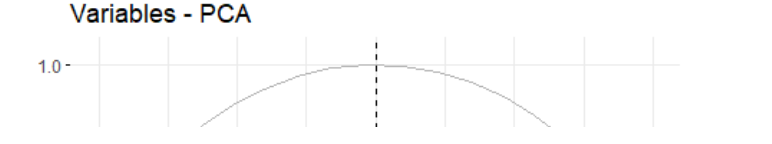


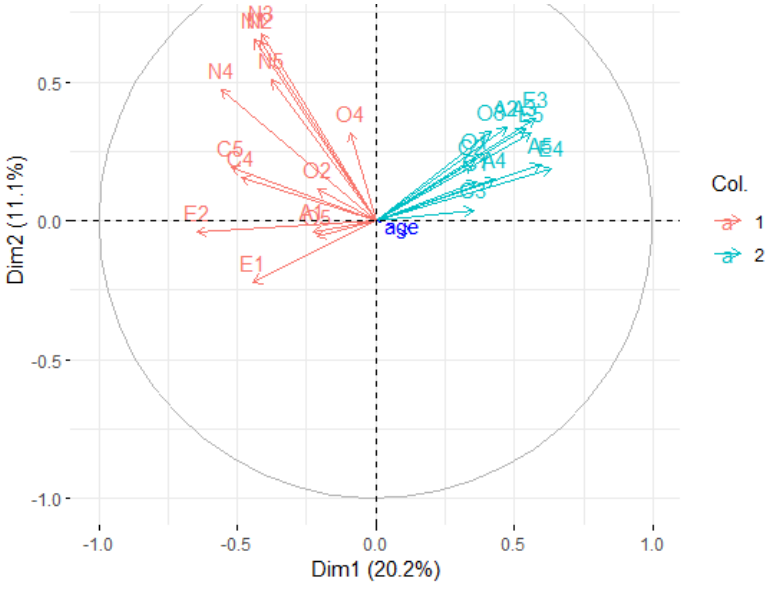
Il cluster rosso corrisponde al gruppo di caratteristiche di soggetti con problemi emotivi. Il cluster verde sono caratteristiche di individui che non si impegnano e hanno problemi a relazionarsi (includerei in esso anche C4 e C5). L'ultimo cluster riguarda coloro che si impegnano in ciò fanno e sono sociali.

NbClust

```
*******************************************************************
* Among all indices:
* 9 proposed 2 as the best number of clusters
* 7 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 2 proposed 7 as the best number of clusters
* 2 proposed 9 as the best number of clusters
* 1 proposed 11 as the best number of clusters
* 1 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2


*******************************************************************
```

NbClust suggerisce 2 come numero di clusters. E' quindi evidente la distinzione tra caratteristiche positive e negative.





Factor Analysis

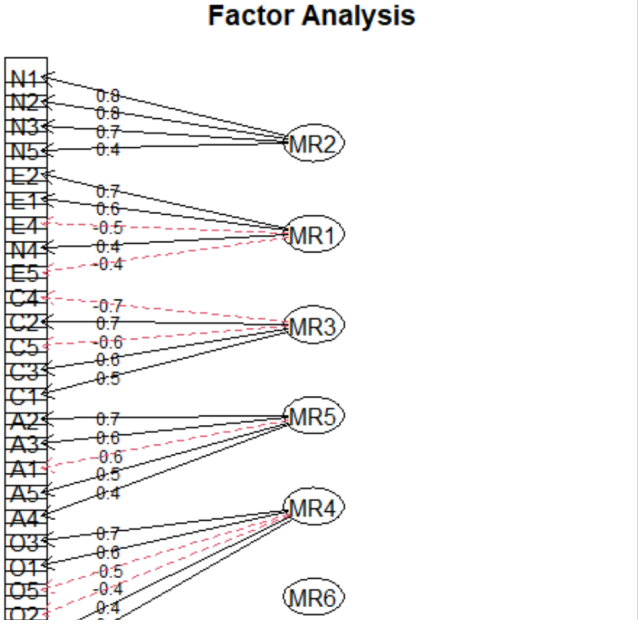
Prima di procedere con la FA ho utilizzato il test Kaiser-Meyer-Olkin (KMO) per misurare quanto fossero adatti questi dati per l'analisi fattoriale. Il valore minimo accettabile per KMO è 0,6. Tuttavia, l'ideale è superiore a 0,8.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = bfi.cleaned[1:25])
Overall MSA =  0.85
MSA for each item =
  A1   A2   A3   A4   A5   C1   C2   C3   C4   C5   E1   E2   E3   E4   E5   N1   N2   N3   N4
0.73 0.83 0.87 0.87 0.90 0.84 0.79 0.85 0.82 0.86 0.84 0.88 0.89 0.88 0.89 0.78 0.78 0.86 0.89
  N5   O1   O2   O3   O4   O5
0.86 0.85 0.78 0.82 0.78 0.75
```

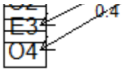
Essendo un dataset basato sull'analisi fattoriale, naturalmente il test restituisce un valore significativo.

Precedentemente, l'analisi parallela aveva restituito 6 come numero di possibili fattori.

```
Loadings:
   MR2    MR1    MR3    MR5    MR4    MR6
A1                      -0.552          0.316
A2                       0.685
A3                       0.610
A4                       0.412
A5                       0.459
C1                0.539
C2                0.654
C3                0.558
C4               -0.691
C5               -0.560
E1          0.598
E2          0.685
E3         -0.337                  0.386
E4         -0.510
E5         -0.392
N1   0.843
N2   0.831
N3   0.687
N4   0.431  0.438
N5   0.447
O1                                  0.561
O2                                 -0.433
```

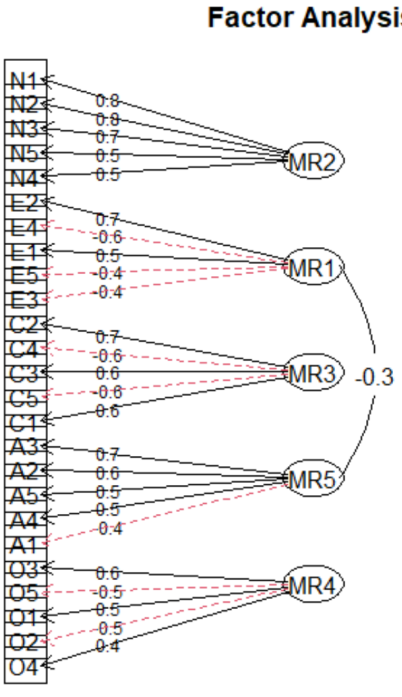


```
O3                      0.677
O4          0.351       0.363
O5                     -0.494   0.349
```



Lo scopo di questi "Loadings" è quello di avere ogni elemento caricato su un fattore separato, e non multiplo, e raggrupparli nei fattori in un modo che abbia senso. Notiamo come gli elementi non siano perfettamente raggruppati, perciò ho deciso di provare con un numero di fattori pari a 5.

```
Loadings:
   MR2    MR1    MR3    MR5    MR4
A1                      -0.417
A2                       0.639
A3                       0.663
A4                       0.452
A5                       0.538
C1                0.552
C2                0.658
C3                0.572
C4               -0.649
C5               -0.566
E1          0.547
E2          0.673
E3         -0.425
E4         -0.573         0.323
E5         -0.427
N1  0.836
N2  0.784
N3  0.704
N4  0.464  0.416
N5  0.481
O1                                0.514
O2                               -0.483
O3                                0.638
O4          0.338                 0.367
O5                               -0.543
```



Finalmente gli elementi sono stati raggruppati correttamente secondo 5 fattori. Ovviamente questi ultimi corrispondono a quelli descritti all'inizio: Agreeableness (gradevolezza), Conscientiousness (coscienziosità), Extraversion (estroversione), Neuroticism (nevroticismo) e Openness (apertura).

Infatti fu proprio questo tipo di analisi che originariamente portò gli psicometristi e i ricercatori della personalità a concludere che ci sono cinque dimensioni principali nelle differenze interpersonali.