4R\_Esercitazione1

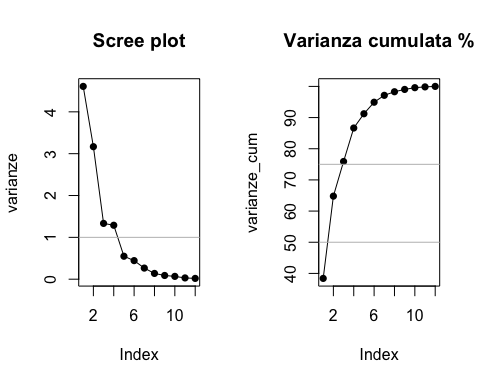
Francesca Pecorari

2024-01-19

# Studio di “Dataset\_whiskey”

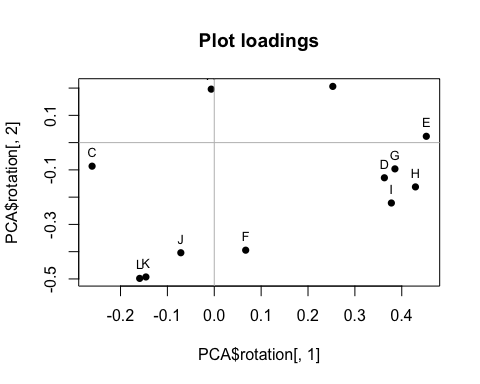
## PCA

### Screeplot e grafico della varianza cumulata



Dai grafici deduciamo che sono sufficienti 3 componenti per descrivere la varianza del dataset. Osservando lo Scree plot notiamo un gomito tra la terza e la quarta componente, le ultime a mostrare varianza > 1. Dal plot della varianza cumulata notiamo che tra le 2 e le 3 componenti sono sufficienti a spiegare tra il 50 % e il 75 % della varianza del dataset.

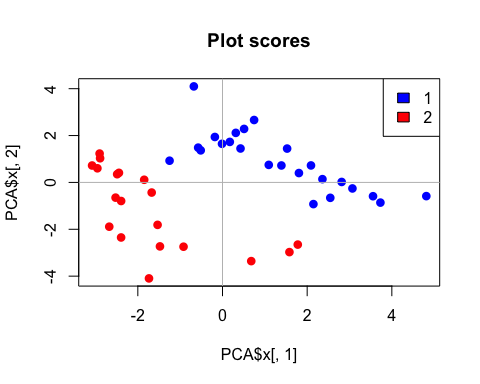
### Plot loadings



Interpretando il grafico, si potrebbe affermare che una minima parte delle variabili originali contribuisce poco alla direzione delle nostre componenti principali, solo le variabili A e E mostrano loadings molto vicini a zero.

Si nota che le variabili L, K e D, G sembrano contribuire in modo simile alla direzione delle componenti, per il resto le variabili non sembrano avere valori di loading simili e questo indica, in generale, che non sono correlate e che avranno effetti diversi sulle componenti.

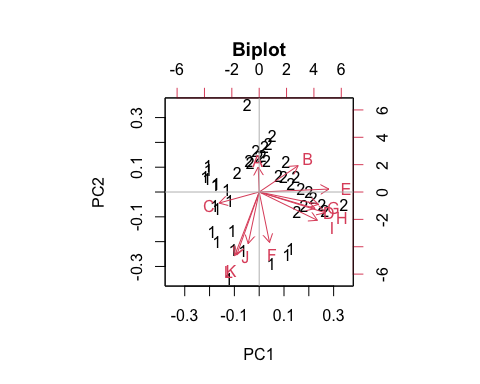
### Plot scores



Il grafico degli Score rappresenta le osservazioni nel nuovo sistema di variabili.

Non ci sono osservazioni molto vicine all’origine, questo indica che non hanno valori prossimi alla media per la maggior parte delle variabili. Potrebbero essere presenti dei valori estremi o outliers per entrambe le categorie di whiskey e la maggior parte delle osservazioni per ogni categoria sembrano essere dissimili rispetto alle osservazioni dell’altra.

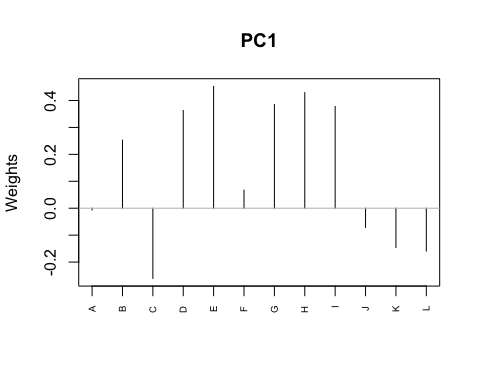
### Biplot



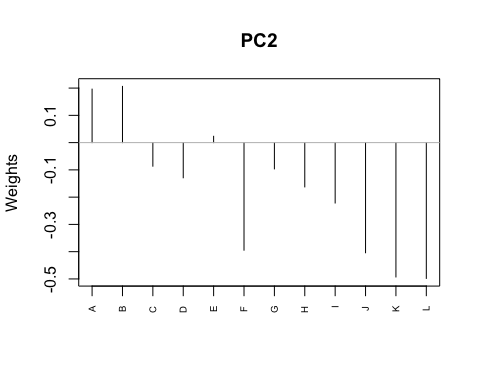
Si nota che i vettori per le variabili L, K, D e I sembrano essere i più lunghi, questo indica che la variabilità delle suddette è ben spiegata dalle componenti. Al contrario, per la variabile A la variabilità sarà meglio spiegata da altre PC non utilizzate per costruire questo grafico. La variabile E sembra contribuire quasi esclusivamente alla PC1, mentre la variabile A sembra contribuire alla PC2.

### Grafici del peso delle variabili sperimentali per ogni PC scelta

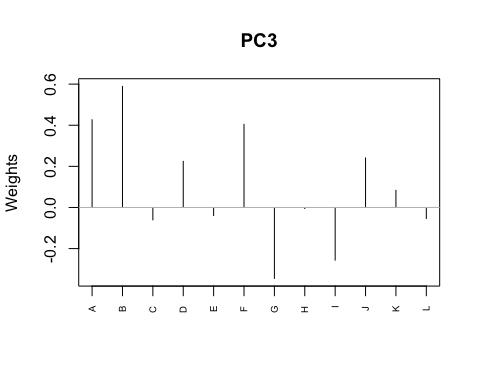
Le variabili sperimentali che hanno peso maggiore per la PC1 sono: C, E, G, I, H



Le variabili sperimentali che hanno peso maggiore per la PC2 sono: F, J, K, L

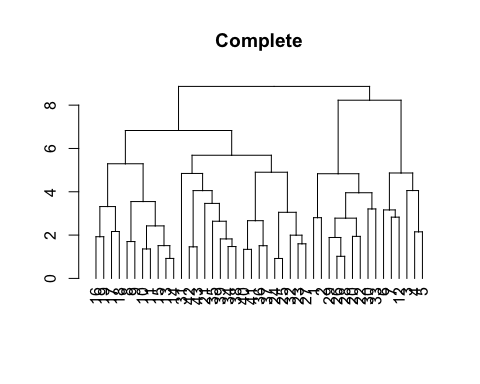


Le variabili sperimentali che hanno peso maggiore per la PC3 sono: B e F

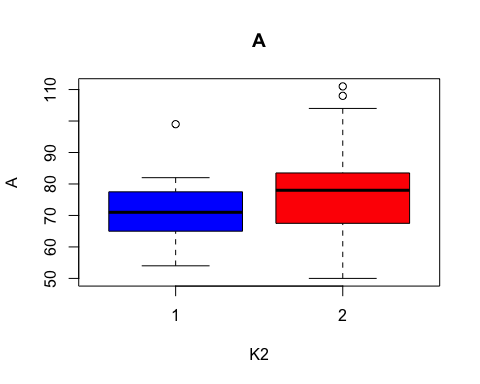


## HIERARCHICAL CLUSTERING

### Dendrogramma



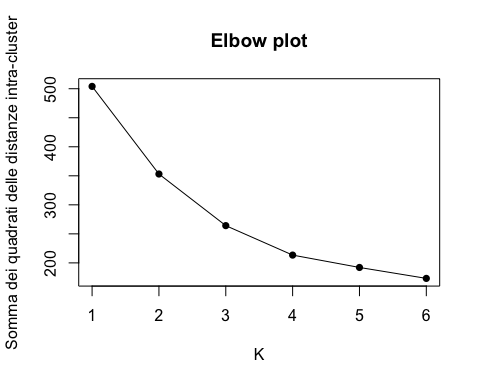
Nell’interpretazione del dendrogramma si può tenere conto della presenza di un numero definito di raggruppamenti che possa aiutare all’interpretazione dei dati. Si può scegliere di “tagliare” i rami a una certa distanza e le osservazioni saranno raggruppate separatamente in base a questa distanza.



La prova è stata eseguita su due cluster e vediamo che, ad esempio per la variabile A, si ottengono distribuzioni abbastanza simili, con un piccolo numero di outliers.

## K-MEANS CLUSTERING

### Elbow plot



Nell’ interpretazione dell’elbow plot si cerca un “gomito” in cui la discesa della somma dei quadrati degli errori inizia a rallentare significativamente; idealmente si sceglie il numero di cluster appena prima del punto di gomito. Nel nostro grafico, a seconda di un’interpretazione soggetiva, si potrebbero scegliere sia il valore 3, sia il valore 2, tuttavia, sappiamo che i nostri dati sono divisi in due categorie, quindi sarebbe probabilmte opportuno scegliere 2 come numero ottimale di cluster.