

ESERCIZIO IN R

CLUSTER ANALYSIS: DATASET Metro.csv

Problema da risolvere:

Il processo di segmentazione della clientela si colloca nell'ambito di un progetto di marketing che ha come target finale l'introduzione di opportune e graduali politiche costituite principalmente da un sistema di offerta e da comportamenti comunicativi di vendita orientati alla gestione relazionale della clientela (Customer Relationship Management).

I dati a disposizione:

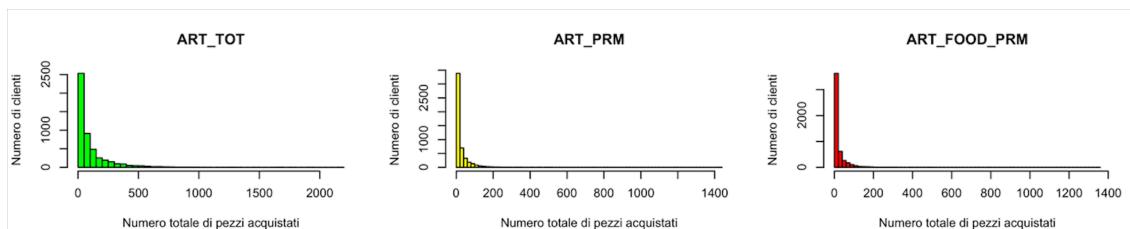
I dati a disposizione riguardano le rilevazioni, effettuate attraverso la tecnologia scanner, degli acquisti che avvengono giornalmente in ogni magazzino da parte di tutti gli iscritti. Agli iscritti, all'atto del tesseramento, viene rilasciata una card contraddistinta da un codice di otto cifre (tipo 000*****), detto *customer number*, che rappresenta l'identificativo del cliente. Ogni riga definisce, pertanto, il comportamento di acquisto di un determinato cliente sulla base di una serie di variabili, delle quali le più importanti sono:

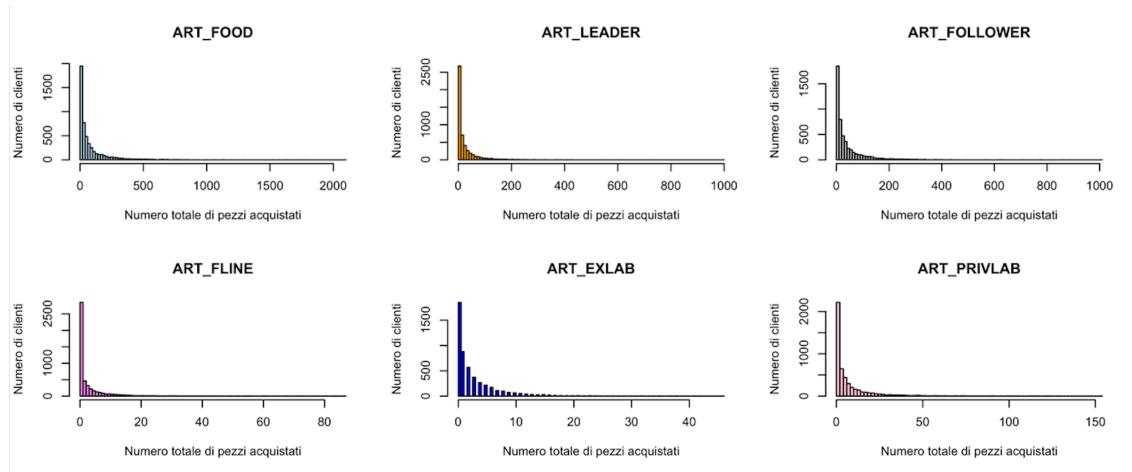
- CUST NO: numero della tessera;
- ART TOT: numero totale di pezzi acquistati;
- ART PRM: numero totale di pezzi acquistati in promozione;
- ART FOOD PRM: numero totale di pezzi acquistati in promozione Food;
- ART FOOD: numero totale di pezzi acquistati Food;
- ART LEADER: numero totale di pezzi acquistati senza marca specifica;
- ART FOLLOWER: numero totale di pezzi acquistati con nuovo layout;
- ART FLINE: numero totale di pezzi acquistati di prima linea (super qualità);
- ART EXLAB: numero totale di pezzi acquistati precedentemente in promozione;
- ART PRIVLAB: numero totale di pezzi acquistati precedentemente senza promozione.

Obiettivi:

1. RAPPRESENTARE GRAFICAMENTE LE DISTRIBUZIONI DI FREQUENZA DELLE VARIABILI E DEDURRE EVENTUALI INFORMAZIONI SULLA SIMMETRIA/ASIMMETRIA DELLE DISTRIBUZIONI

```
par(mfrow=c(3,3))for(i in 1:9)  
hist(metro[,i],breaks=75,main=colnames(metro)[i],xlab="Numero totale di pezzi  
acquistati",ylab="Numero di clienti",  
col=c("green","yellow","red","lightblue","orange","gray","violet","blue","pink"  
)[i]) # rappresentazione grafica delle distribuzioni relative ai pezzi  
acquistati in base ai comportamenti dei clienti
```





Sulle rappresentazioni grafiche delle distribuzioni di frequenza delle variabili, dove a ciascun grafico corrisponde uno dei comportamenti dei clienti (in totale 9 comportamenti, ossia 9 variabili), sull'asse delle ascisse viene indicato il numero dei pezzi comprati in relazione al determinato comportamento e sull'asse delle ordinate il numero di clienti che acquistano una determinata quantità di prodotti in relazione al comportamento.
Dal grafico si evince una distribuzione asimmetrica (moda, media e mediana non sono uguali): in questo caso si tratta di un'asimmetria a destra, ossia un'asimmetria positiva, cioè la media è maggiore della mediana.

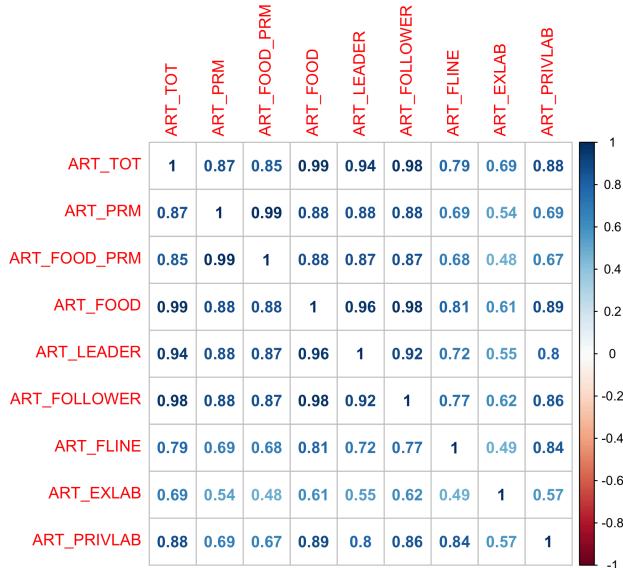
```
> summary(metro) # statistiche di sintesi delle variabili coinvolte
      ART_TOT      ART_PRM      ART_FOOD_PRM      ART_FOOD      ART_LEADER      ART_FOLLOWER
Min.   : 1.0   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
1st Qu.: 17.0  1st Qu.: 3.00   1st Qu.: 2.00   1st Qu.: 9.00   1st Qu.: 2.00   1st Qu.: 5.00
Median : 49.5  Median : 10.00  Median : 7.00   Median : 34.00  Median : 9.00   Median : 18.00
Mean   : 109.0 Mean   : 26.68  Mean   : 22.53  Mean   : 85.18  Mean   : 26.76  Mean   : 44.17
3rd Qu.: 128.0 3rd Qu.: 28.00  3rd Qu.: 23.00  3rd Qu.: 97.00  3rd Qu.: 29.00  3rd Qu.: 52.00
Max.   :2155.0 Max.   :1427.00 Max.   :1351.00 Max.   :2093.00 Max.   :999.00  Max.   :1006.00
      ART_FLINE      ART_EXLAB      ART_PRIVLAB
Min.   : 0.000  Min.   : 0.000  Min.   : 0.000
1st Qu.: 0.000  1st Qu.: 0.000  1st Qu.: 1.000
Median : 1.000  Median : 1.000  Median : 3.000
Mean   : 3.719  Mean   : 2.891  Mean   : 9.062
3rd Qu.: 4.000  3rd Qu.: 4.000  3rd Qu.: 10.000
Max.   :87.000  Max.   :46.000  Max.   :154.000
```

La stessa deduzione può essere fatto attraverso l'analisi dell'output del comando "summary(metro)", nel dettaglio tramite un confronto tra media e mediana; infatti la media relativa a ciascuna variabile (comportamento) è maggiore della mediana della stessa variabile (comportamento). Anche da questo si evince una distribuzione asimmetrica a destra.

2. VALUTARE SE SUSSISTONO RELAZIONI TRA LE VARIABILI ATTRAVERSO LA VISUALIZZAZIONE GRAFICA DELLA MATRICE DI CORRELAZIONE E COMMENTARE I RISULTATI OTTENUTI

Dopo aver caricato il pacchetto Corrplot, necessario per la visualizzazione grafica della matrice di correlazione tra le variabili quantitative, come lo sono in questo caso, e dopo aver costruito quest'ultima (`cormetro<-cor(metro)`), è possibile averne una visualizzazione grafica:

```
corrplot(cormetro, method="number") # rappresentazione grafica della matrice di correlazione
```



Dalla matrice si evincono i gradi di correlazione tra coppie delle variabili di comportamento del dataset: il grado di correlazione è rappresentato del numero in ciascuna cella e dal colore che esso assume. Più il colore è chiaro, ovvero vicino al bianco, allora il grado di correlazione sarà più vicino allo zero; più è blu, allora il grado di correlazione sarà alto e positivo; più è rosso, allora il grado di correlazione sarà basso e negativo, come indicato nella barra laterale alla matrice.

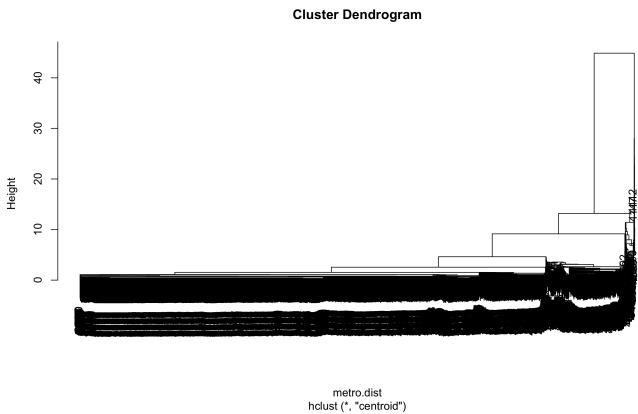
Nella matrice di correlazione del dataset metro.csv è possibile osservare che tra nessun comportamento vi è una relazione negativa: i gradi di correlazione sono tutti positivi, per questo motivo la scala cromatica entro cui si attengono i colori relativi ai gradi va dall'azzurro chiaro al blu scuro.

Ad esempio tra la variabile di comportamento ART_TOT (prima riga) e la variabile ART_FOOD (quarta colonna) è pari a 0.99, quindi il grado di correlazione positivo è molto forte, di conseguenza di colore blu scuro. Invece tra la variabile di comportamento ART_EXLAB (ottava riga) e la variabile ART_FOOD_PRM (terza colonna) è pari a 0.48, quindi il grado di correlazione positivo è relativamente debole, di conseguenza di colore azzurro.

Ovviamente nelle coppie della stessa variabile di comportamento, il grado di correlazione sarà massimo, ossia pari a 1, e di colore blu scuro.

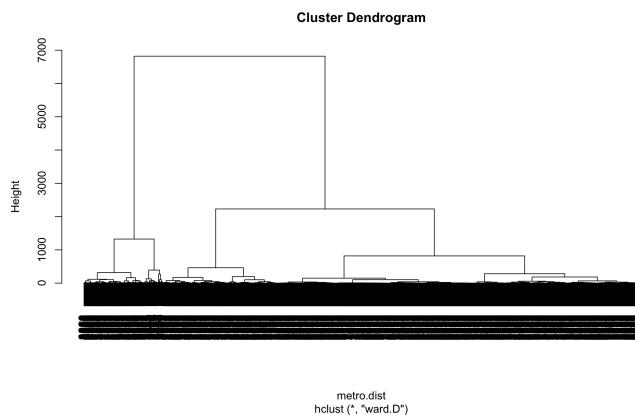
3. IMPLEMENTARE I METODI DI CLASSIFICAZIONE GERARCHICA CHE UTILIZZANO CONGIUNTAMENTE COME INPUT LA MATRICE DEI DATI E LA MATRICE DELLE DISTANZE E COMMENTARE I RISULTATI OTTENUTI SULLA BASE DEI RISPETTIVI DENDOGRAMMI. DOPO AVER STABILITO IL METODO MIGLIORE, ATTRIBUIRE UN SIGNIFICATO AI GRUPPI SELEZIONATI

Dopo aver creato la matrice delle distanze euclidee (con i comandi `metro.dist<-daisy(metro, metric="euclidean", stand=TRUE)`), dove "daisy" è la funzione che permette di misurare le distanze euclidee (grazie alla specificazione "metric") e `as.matrix(metro.dist)`) si procede con l'implementazione dei metodi di classificazione gerarchica, basati sulla matrice dei dati e sulla matrice delle distanze: il metodo del centroide e il metodo di Ward.



```
metro.hc.cen<-
hclust(metro.dist,method="centroid") #
implementazione dei metodi di
classificazione gerarchici secondo il
metodo del centroide

plot(metro.hc.cen) # visualizzazione del
dendrogramma
```



```
metro.hc.ward<-
hclust(metro.dist,method="ward") #
implementazione dei metodi di
classificazione gerarchici secondo il
metodo di Ward

plot(metro.hc.ward) # visualizzazione del
dendrogramma
```

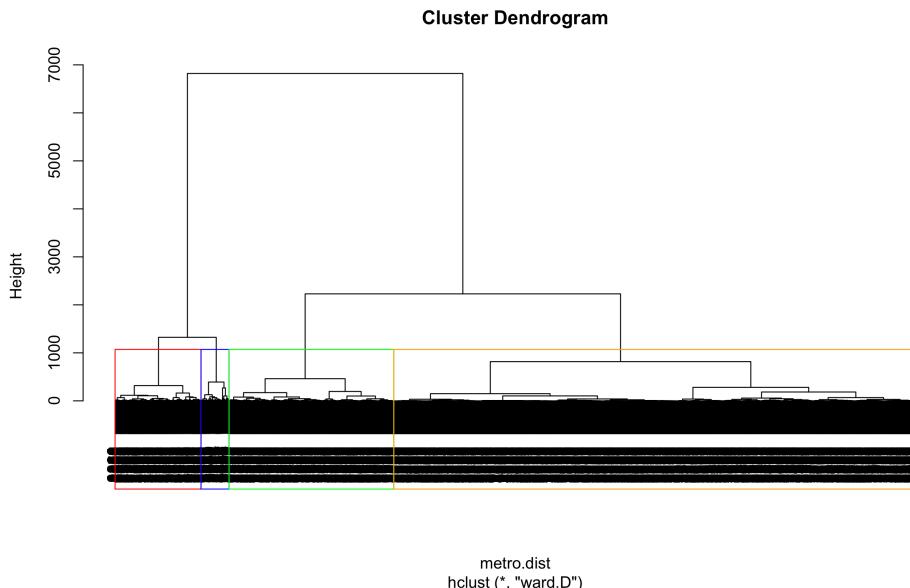
Nel primo grafico in cui viene utilizzato il **metodo del centroide**: la distanza tra due gruppi (C_1 e C_2), di numerosità n_1 e n_2 è calcolata come la distanza tra i rispettivi centroidi o baricentri (tipicamente le medie aritmetiche). Per centroide si intende il vettore formato dai valori medi delle variabili considerate nell'analisi dei gruppi.

Il dendrogramma sviluppato secondo questo metodo non risulta essere bilanciato, ma in essi vi sono associazioni continue senza ramificazioni distinte.

Il **metodo i Ward** massimizza la funzione obiettivo: esso riunisce i gruppi dalla cui fusione deriva il minimo incremento marginale della devianza interna (within).

Dal dendrogramma tratto dall'implementazione del metodo di classificazione gerarchica di Ward, dove le ramificazioni sono più chiare e bilanciate rispetto al dendrogramma costruito con il metodo del centroide.

```
rect.hclust(metro.hc.ward,k=4,border=c("red","blue","green","orange")) #
suddivisione del dendrogramma in quattro gruppi sulla base del metodo di Ward
```



Dopo aver aggregato le unità statistiche in relazione ai quattro segmenti (con il comando `metro.hc.ward.segment<-cutree(metro.hc.ward,k=4)`), è possibile eseguire l'analisi di ciascuno dei gruppi:

```
> table(metro.hc.ward.segment) # dimensione di ogni gruppo
metro.hc.ward.segment
 1   2   3   4
3259 175 1029 537
> # funzione per la determinazione delle medie di gruppo
> seg.mean<-function(data,groups){
+ aggregate(data,list(groups),FUN=mean)
+ }
> seg.mean(metro,metro.hc.ward.segment) # medie delle variabili nei quattro clusters
  Group.1 ART_TOT ART_PRM ART_FOOD_PRM ART_FOOD ART_LEADER ART_FOLLOWER ART_FLINE ART_EXLAB ART_PRIVLAB
 1       1 32.14667  6.911322  5.127647 22.4799  6.406566 12.38846  0.8646824 0.9039583 2.514268
 2       2 699.36571 211.371429 194.342857 603.6171 203.371429 294.46286 28.8971429 13.6285714 65.714286
 3       3 136.64626 30.417881 24.023324 100.4014 30.629738 54.18173 4.0359572 4.9241983 9.907677
 4       4 330.11546 79.316574 69.329609 267.5717 85.337058 136.33147 12.2253259 7.5530726 28.713222
```

Sono gruppi sbilanciati: le unità statistiche sono prevalentemente allocate nel primo gruppo (predomina il primo gruppo), che nel dendogramma corrisponde al cluster più a destra (delineato dal colore arancione).

Dalle medie di pezzi acquistati è possibile evincere quali comportamenti incidono maggiormente sugli acquisti effettuati dai clienti. Innanzitutto le medie più alte in ciascuno dei cluster sono più alte in relazione alla prima variabile (ART_TOT) in quanto essa rappresenta il numero totale di pezzi acquistati in ogni gruppo da parte dei clienti.

- Nel primo gruppo vengono principalmente acquistati pezzi in relazione alla variabile ART_FOOD, ossia il numero di pezzi acquistati nella categoria Food.
- Nel secondo gruppo vengono principalmente acquistati pezzi in relazione alla variabile ART_FOLLOWER, ossia il numero di pezzi acquistati con un nuovo layout.
- Nel terzo gruppo vengono principalmente acquistati pezzi in relazione alla variabile ART_FOOD, ossia il numero di pezzi acquistati nella categoria Food.
- Nel quarto gruppo vengono principalmente acquistati pezzi in relazione alla variabile ART_FOOD, ossia il numero di pezzi acquistati nella categoria Food.

Si può dedurre che la variabile di comportamento che incide maggiormente sugli acquisti dei clienti è quella relativa alla categoria Food.

4. IMPLEMENTARE IL METODO DELLE K-MEDIE E STABILIRE, SULLA BASE DEL GRAFICO BASATO SUL CRITERIO DI ELBOW, UN NUMERO ADEGUATO DI POSSIBILI CONFIGURAZIONI DI CLUSTER; VALUTARE LA BONTÀ DI CLASSIFICAZIONE DI CIASCUNA CONFIGURAZIONE DI

CLUSTER CONSIDERATA ATTRAVERSO L'INDICE DI SILHOUETTE E L'R²; ATTRIBUIRE UN SIGNIFICATO AI GRUPPI OTTENUTI.

Dopo aver standardizzato il dataset per il calcolo della distanza euclidea (`metrostan<-scale(metro)`), è possibile implementare il metodo delle k-medie, fissati i seeds iniziali a 200 e il numero massimo di gruppi a 20.

`withinss<-sapply`

`(1:k.max,function(k){kmeans(metrostan,k,nstart=50,iter.max=20)$tot.withinss})`

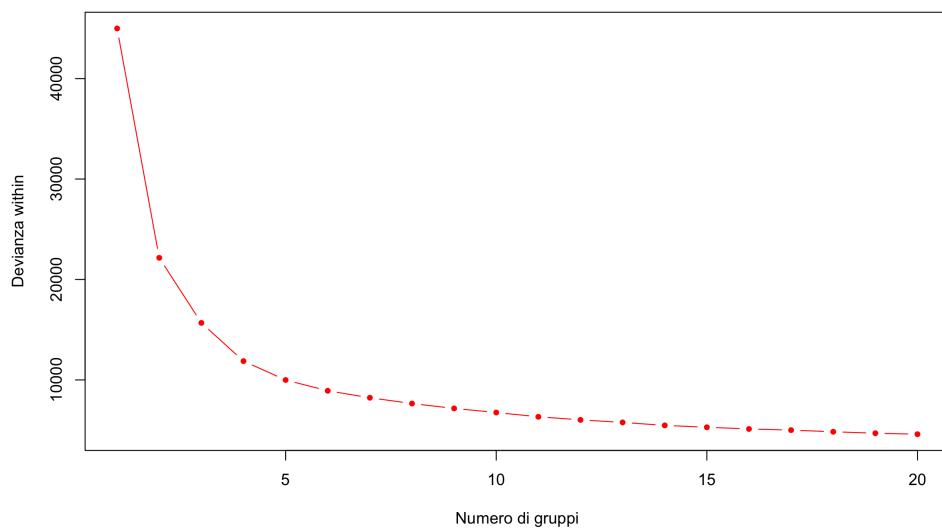
In questo caso il metodo delle k-medie è applicato con numero di iterazioni, ossia di trasferimenti, pari a 20 (che possono essere visualizzate richiamando l'oggetto “withinss”) e numero di numero di volte in cui i semi iniziali vengono ricampionati pari a 50.

```
> withinss
[1] 44991.000 22156.404 15679.004 11873.005 9991.274 8919.217 8224.618 7649.187 7163.103 6755.642
[11] 6331.150 6020.921 5770.585 5471.916 5284.780 5119.086 4999.714 4840.469 4693.675 4595.920
```

È a questo punto possibile costruire il grafico basato sul metodo Elbow: esso riporta sulle ascisse la quantità di clusters in cui viene suddivisa la clientela e sulle ordinate il valore di devianza `within` del relativo numero di gruppi. Il numero ottimale di clusters viene definito in funzione del contributo di ogni cluster aggiuntivo rispetto alla devianza interna: quando il contributo associato ad un ulteriore gruppo è minimo, allora il grafico, che esprime l'andamento della devianza in termini di gruppi, si riduce, avvicinandosi sempre più allo 0, formando una sorta di gomito.

Si procede con la costruzione del grafico grazie al comando:

```
plot(1:k.max,wss,type="b",pch=20,xlab="Numero di gruppi",ylab="Devianza within",col="red")
```



Il grafico riporta sulle ascisse il numero massimo di gruppi, fissato a priori, e sull'asse delle ordinate il valore della devianza associato a ciascun numero di gruppi: man mano che il numero di gruppi aumenta, diminuisce la devianza. Il numero ottimale di gruppi è identificato nel momento in cui un incremento unitario di un gruppo, comporta modificazioni della devianza minime.

È quindi possibile scegliere in base al grafico il numero di cluster pari a 5 e, caricato in R il pacchetto per la valutazione della configurazione di cluster ottenuta (pacchetto “factoextra”):

```

> kmeans5
K-means clustering with 5 clusters of sizes 152, 12, 3438, 417, 981

Cluster means:
          ART_TOT   ART_PRM ART_FOOD_PRM   ART_FOOD ART_LEADER ART_FOLLOWER   ART_FLINE ART_EXLAB ART_PRIVLAB
1 679.04605 183.934211 168.171053 582.0658 199.184211 284.21711 24.5855263 13.217105 63.565789
2 1551.33333 630.500000 591.583333 1366.7500 511.750000 693.91667 48.0833333 21.500000 98.000000
3 33.63147 7.390052 5.301629 22.3918 6.247818 12.32519 0.9589878 1.330425 2.534322
4 359.42446 88.004796 77.091127 292.1799 93.071942 149.26139 13.2206235 8.016787 31.139089
5 160.75433 36.470948 30.209990 124.5637 37.823649 65.97757 5.5749235 4.352701 13.019368

Within cluster sum of squares by cluster:

```

Available components:

```
[1] "cluster"      "centers"       "totss"        "withinss"      "tot.withinss" "betweenss"     "size"  
[8] "iter"         "ifault"
```

In questo modo si ottengono 5 cluster di dimensioni 152, 12, 3438, 417 e 981, dove predomina il gruppo 3.

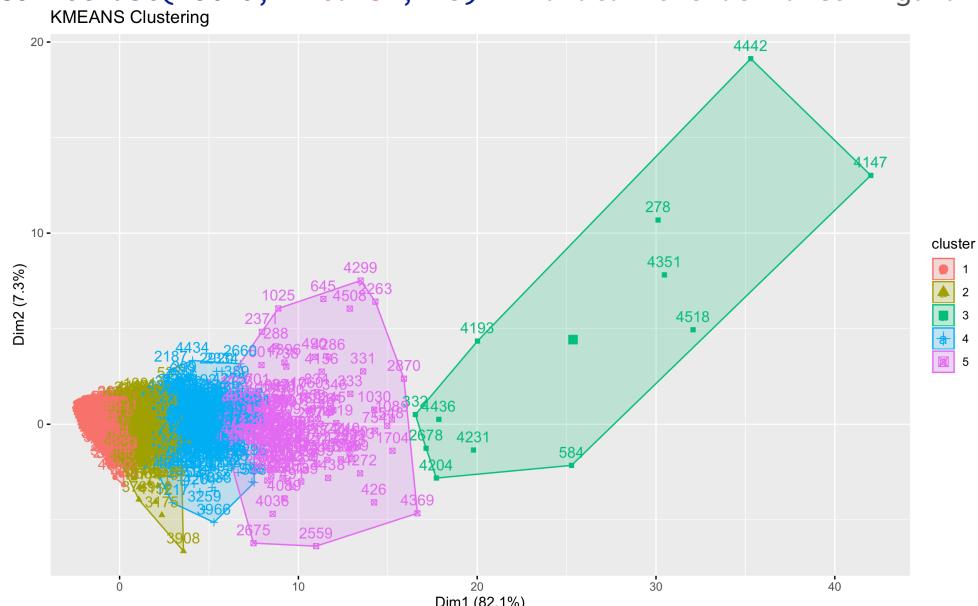
L'indice di bontà R^2 è pari all'89,3%. Il suo valore deve essere compreso tra 0 e 1 ($R^2 \in [0, 1]$): poiché esso è il risultato del rapporto tra la devianza between (ossia la varianza tra i gruppi) e la devianza totale, dove la devianza between è data dalla somma calcolata su tutte le variabili delle devianze ponderate delle medie di gruppo rispetto alla corrispondente media generale; la devianza totale è pari alla somma delle devianze delle singole variabili rispetto alla corrispondente media generale.

Se esso è vicino a 1 significa che la partizione nei cluster è ottimale, poiché all'interno dei gruppi le unità statistiche sono omogenee (quindi le unità statistiche all'interno di ciascun cluster sono simili tra loro) e che tra i vari gruppi c'è alta eterogeneità (quindi tra i singoli gruppi sono molto diversi tra loro); se, invece, esso è vicino a 0 significa che vi è bassa omogeneità all'interno dei gruppi e bassa eterogeneità tra i gruppi. Dunque più R^2 è vicino a 1, maggiore è la bontà dell'analisi di classificazione.

In questo caso R^2 è pari all'89,3%, ossia a 0,893, per cui, essendo vicino all'1, la classificazione nei 5 cluster può essere considerata molto buona, ed è quindi rispettata la condizione necessaria di alta omogeneità all'interno dei gruppi e alta eterogeneità tra i gruppi.

È necessaria un'ulteriore analisi della bontà della configurazione dei cluster al fine di poterne trarre utilità. Per eseguire questa valutazione, si deve fare ricorso all'indice di Silhouette.

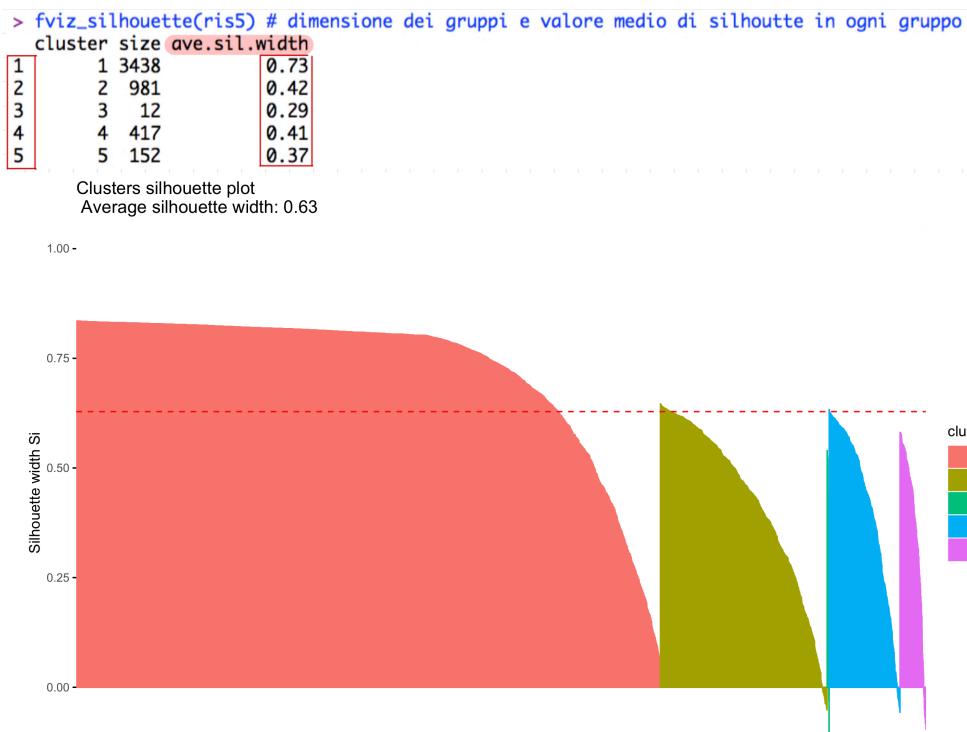
In una prima fase si analizza il grado di concentrazione delle osservazioni all'interno dei singoli cluster: `ris5<-eclust(metro, "kmeans", k=5)` # valutazione della configurazione



Nei primi tre gruppi (da sinistra a destra) è possibile osservare che in gruppi di dimensioni minori sono concentrate numerosissime osservazioni e molto vicine tra loro; invece nei gruppi successivi, ma soprattutto nel quinto, una dimensione maggiore del gruppo corrisponde ad una maggiore distanza tra le singole osservazioni, implicando un numero ridotto di osservazioni contenute nello specifico cluster. Questo è necessario per garantire un'elevata omogeneità all'interno dei gruppi e conseguentemente un'elevata eterogeneità tra i gruppi.

```
fviz_silhouette(ris5) # dimensione dei gruppi e valore medio di silhouette in ogni gruppo
```

Il comando fornisce sia la configurazione grafica, sia un'analisi dei singoli cluster circa la loro dimensione e la loro silhouette media.



L'indice di silhouette permette di verificare se l'allocazione delle singole unità statistiche sia valida. Il suo valore va da -1 a 1 (indice di silhouette $\in [-1, 1]$): se è vicino ad 1 significa che l'oggetto è ben allocato nel suo cluster; se, invece, è vicino a -1 allora l'oggetto non è ben allocato.

L'allocazione delle unità statistiche nei vari cluster è considerata buona quando tutte le unità presentano un indice di silhouette alto, ossia vicino ad 1.

In questo caso viene riportato il valore medio di silhouette all'interno di ciascun cluster (0.73, 0.42, 0.29, 0.41, 0.37). Poiché nessuno dei valori è negativo, l'allocazione delle singole unità statistiche può essere considerata buona, ma, dal momento che i valori non sono tutti vicinissimi all'1, essa non può essere considerata ottimale.

Dal grafico si evince il valore medio di silhouette per l'intero dataset (pari a 0.63) ed è possibile avere una rappresentazione generale dei singoli valori di silhouette per ciascuna unità statistica e osservare quanti sono negativi (ossia nei gruppi 2, 3, 4, 5 i valori al di sotto dell'asse delle ascisse). Con il comando `sil5<-ris5$silinfo$width` è possibile osservare il valore di silhouette corrispondente a ciascuna unità statistica, sarà poi possibile visualizzare le posizioni delle posizioni delle unità statistiche con valore di silhouette negativa, che sono 59, con il comando

`neg_sil_index5<-which(sil5[, 'sil_width']<0)` e potranno poi essere visualizzate con il comando `sil5[neg_sil_index5,]`.

Su un dataset di 5000 osservazioni, a seguito della suddivisione in 5 clusters con il metodo delle k-medie, soltanto 59 risultano avere un indice di silhouette negativo e per questo motivo la suddivisione può essere considerata soddisfacente.