UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA

Scuola di Economia e Statistica

Corso di laurea in Economia e Commercio

Tecniche di classificazione di statistica multivariata applicate al mercato automobilistico

Relatore: Prof. Manuela Cazzaro

Tesi di Laurea di:

Andrea Piancone

Matr. N. 812250

Anno Accademico 2018/2019

Indice

Indice delle figure

Indice delle tabelle

Introduzione

In questo lavoro si è deciso di approfondire una particolare tecnica di analisi statistica multivariata: l'analisi dei gruppi.

Questa tecnica ha lo scopo di raggruppare le unità statistiche in gruppi che siano il più possibile omogenei al loro interno e il più possibili eterogenei fra di essi.

Si è deciso di applicare l'analisi dei gruppi al mercato automobilistico a causa della grande importanza che i veicoli rivestono nelle scelte di consumo degli individui. Il data set è stato costruito cercando modelli di autoveicoli con un prezzo compreso tra i 9000€ e i 30000€ affinché si potesse procedere con un confronto omogeneo. Sugli autoveicoli oggetto dell'analisi sono state rilevate le seguenti variabili: *prezzo, lunghezza, altezza, larghezza, passo, volume del bagagliaio, capacità del serbatoio, massa in ordine di marcia, consumi urbani, emissioni di diossido di carbonio, accelerazione, velocità massima, cilindrata, potenza, coppia massima*.

Per svolgere l'analisi operativa del data set è stato impiegato il software R.

La struttura del lavoro si articola nel seguente modo: nel primo capitolo viene descritto il data set, indicando la fonte da cui sono stati raccolti i dati relativi ai veicoli e le trasformazioni ad esso apportate. Nel secondo capitolo viene descritta da un punto di vista teorico l'analisi dei gruppi, illustrando brevemente sia le tecniche di aggregazione gerarchiche che il metodo non gerarchico delle k-medie. Nel terzo ed ultimo capitolo si applica tale tecnica di statistica multivariata al data set precedentemente descritto, impiegando sia i metodi gerarchici che il metodo non gerarchico delle k-medie, al fine di individuare quale criterio di aggregazione delle unità statistiche porta alla partizione che meglio si addice all'analisi. Il capitolo si conclude con la descrizione e la caratterizzazione dei gruppi in base alle prestazioni, alle caratteristiche del motore e della carrozzeria delle automobili contenute in ogni gruppo della partizione ritenuta ottimale.

Capitolo 1

Descrizione del data set

1.1 Costruzione del data set

Il data set è stato costruito con lo scopo di indagare le prestazioni e le caratteristiche della carrozzeria e del motore di 57 automobili, comprese in una fascia di prezzo tra 9000€ e 30000€, selezionate tra station wagon, suv, due volumi e tre volumi. Questa scelta è motivata dal fatto che i veicoli dotati di queste caratteristiche sono i più adeguati all'uso quotidiano senza che il loro prezzo sia eccessivamente elevato.
Le unità statistiche e le variabili di cui si compone il data set sono state reperite consultando le schede tecniche delle automobili disponibili sul seguente sito web: "*https://www.automoto.it/listino*".

1.2 Variabili

Le variabili scelte ai fini dell'analisi osservate sulle unità statistiche, sono in totale 15 e sono le seguenti:

1) *Prezzo*.

Questa variabile indica il prezzo di vendita delle automobili.
Il prezzo (in euro) è una delle variabili a cui i consumatori attribuiscono molta importanza al momento dell'acquisto di una vettura.
Banalmente, il prezzo di un'automobile è influenzato dalla qualità delle sue parti componenti e dai materiali utilizzati per realizzare la carrozzeria e gli interni. Tuttavia, il prezzo non dipende solo dalle caratteristiche intrinseche del veicolo, ma è influenzato anche delle politiche di marketing adottate dalla casa automobilistica.

2) *Lunghezza dell'automobile.*
3) *Altezza dell'automobile.*
4) *Larghezza dell'automobile.*
5) *Passo dell'automobile.*

Lunghezza, larghezza, altezza e passo (in centimetri), dove con il termine passo si indica la distanza tra l'asse di una ruota anteriore e l'asse di una ruota posteriore, sono variabili dalle quali dipende la manovrabilità del veicolo e la comodità dei passeggeri e del conducente. In particolare, le automobili caratterizzate da valori più elevati rispetto alle variabili sopracitate garantiscono una maggiore comodità per il conducente ed i passeggeri, tuttavia rendono più difficoltosa la guida del veicolo, specialmente negli spazi piccoli come gli ambienti urbani.

6) *Volume del bagagliaio.*

Questa variabile rappresenta la capienza del bagagliaio in litri ed è una caratteristica a cui i consumatori (in special modo le famiglie), attribuiscono una grande rilevanza: tanto più il bagagliaio è voluminoso tanti più oggetti come buste della spesa e valigie possono essere trasportate comodamente.

7) *Capacità del serbatoio.*

Tale variabile indica la capienza del serbatoio del veicolo e si misura in litri. Da essa dipende l'autonomia dell'automobile: serbatoi più capienti permettono, a parità di consumi per chilometro, di poter percorrere distanze maggiori prima di dover essere costretti a recarsi ad un distributore per riempire nuovamente il serbatoio.

8) *Massa in ordine di marcia.*

La massa in ordine di marcia, espressa in kilogrammi (kg), viene calcolata aggiungendo al peso a vuoto del veicolo i suoi fluidi ed una zavorra di 75 kg che rappresenta il peso medio di un guidatore.

Dalla massa dell'automobile dipendono determinate prestazioni del veicolo, come i consumi e le emissioni: generalmente a masse più elevate corrispondono consumi ed emissioni maggiori.

9) Consumi urbani.

Questa variabile misura il consumo in litri di carburante ogni 100 chilometri percorsi (l/100km) in ambiente urbano.
I consumi sono una caratteristica di un veicolo a cui i consumatori assegnano una grande importanza poiché influenza i costi legati all'utilizzo dell'autoveicolo: automobili caratterizzate da consumi più elevati richiedono rifornimenti di carburante più frequenti, aumentando così le spese relative all'uso del veicolo.

10) Emissioni di diossido di carbonio (CO_2).

Questa variabile indica le emissioni di CO_2 espresse in grammi per ogni chilometro percorso (g/km).
Le emissioni di un veicolo dipendono da diversi fattori come la tipologia di motore ed il carburante con cui è alimentato il veicolo.
È una variabile a cui i consumatori attribuiscono sempre più importanza, a causa della crescente attenzione degli individui verso le tematiche della sostenibilità ambientale, in quanto il diossido di carbonio è uno dei principali gas responsabili dell'effetto serra. Di conseguenza consumatori maggiormente sensibili alle tematiche ambientali, sono più propensi ad acquistare automobili con minori emissioni.

11) Accelerazione.

Questa variabile indica quanti secondi impiega l'automobile partendo da zero chilometri orari a raggiungere i cento chilometri orari (100km/h).
Pertanto, veicoli che assumono valori più alti su tale variabile sono vetture caratterizzate da una minore accelerazione.

12) Velocità massima.

Questa variabile esprime la velocità massima raggiungibile dall'autovettura in chilometri orari (km/h).
Dunque, tanto più questa variabile è elevata, tanto maggiore sarà la velocità massima raggiungibile dal veicolo.

13) Cilindrata.

La cilindrata indica il volume del motore espresso in centimetri cubi (cc). La cilindrata di un veicolo è data dalla somma dei volumi di ciascun cilindro di cui si compone il motore.
Al crescere della cilindrata, cresce la quantità di aria aspirata dal motore e aumenta la dose di carburante che si mescola con l'aria aspirata, sprigionando così un maggior quantitativo di energia. Pertanto, le automobili con cilindrata più elevata tenderanno ad essere anche quelle più potenti.

14) Potenza.

Questa variabile indica la potenza massima, ossia l'energia che il motore è in grado di generare dalla combustione della miscela di aria e carburante (benzina o diesel o altri combustibili come il metano).
Il limite di questa variabile è che la potenza massima viene raggiunta a velocità elevate e difficilmente raggiungibili con uno stile di guida "normale".
Le unità di misura tradizionali della potenza massima di un motore sono i cavalli o i kilowatt (all'interno del data set questa variabile è espressa in kilowatt).

15) Coppia massima.

La coppia massima espressa in Newton per metri (Nm) rappresenta la forza prodotta dal motore mediante la combustione. Questa forza mette in moto i pistoni, che a loro volta mettono in movimento l'albero motore.
Dalla coppia massima dipende l'accelerazione del veicolo: automobili dotate di una maggior coppia motrice godono di una migliore ripresa anche a regimi bassi e quindi possiedono una accelerazione più elevata.

Di seguito nella Tabella 1 è riportato un estratto del data set riguardante le prime 41 unità statistiche.

Tabella 1: estratto del data set

Modello	Prezzo	Lunghezza	Altezza	Larghezza	Passo	Bagagliaio	Serbatoio	Massa	Consumi	Emissioni	Accelerazione	Velocità	Cilindrata	Potenza	Coppia
Peugeot 208 100 Allure	20180.00	397.00	146.00	174.00	254.00	285.00	45.00	1155.00	4.00	90.00	10.50	188.00	1560.00	75.00	250.00
Lancia Ypsilon 0.9	18850.00	384.00	152.00	168.00	239.00	202.00	40.00	1165.00	6.00	107.00	13.10	169.00	875.00	62.00	145.00
Audi A1 1.4 TFSI 125	23700.00	397.00	142.00	174.00	247.00	270.00	45.00	1155.00	6.00	115.00	8.80	204.00	1395.00	92.00	200.00
Volkswagen Golf 1.0	20700.00	426.00	149.00	179.00	262.00	380.00	50.00	1206.00	6.00	108.00	11.90	180.00	999.00	63.00	175.00
Kia Picanto 1.0	15050.00	360.00	148.00	160.00	240.00	255.00	35.00	1051.00	6.00	104.00	10.10	180.00	998.00	74.00	172.00
Volkswagen Golf 1.4	23250.00	426.00	149.00	179.00	262.00	380.00	50.00	1258.00	7.00	127.00	10.60	195.00	1395.00	81.00	200.00
Volkswagen Polo 1.6	22800.00	405.00	145.00	175.00	256.00	351.00	40.00	1280.00	4.00	105.00	11.20	185.00	1598.00	70.00	250.00
Lancia Ypsilon 1.3	18850.00	384.00	152.00	168.00	239.00	245.00	40.00	1125.00	5.00	95.00	12.00	170.00	1248.00	57.00	200.00
Kia Rio 1.2	15100.00	406.00	145.00	172.00	258.00	325.00	37.00	1156.00	6.00	109.00	12.90	173.00	1248.00	62.00	122.00
Volkswagen Polo 2.0	25650.00	407.00	144.00	175.00	255.00	305.00	40.00	1355.00	8.00	134.00	6.70	237.00	1984.00	147.00	320.00
BMW Serie 1 116d 3p	26100.00	433.00	142.00	176.00	269.00	360.00	52.00	1395.00	5.00	111.00	10.30	200.00	1496.00	85.00	270.00
Mini 1.5 Cooper	26200.00	398.00	142.00	173.00	257.00	278.00	44.00	1265.00	4.00	103.00	9.40	203.00	1496.00	85.00	270.00
Citroen C4 Cactus	20750.00	417.00	149.00	173.00	260.00	348.00	45.00	1155.00	4.00	94.00	10.70	184.00	1560.00	73.00	254.00
Hyundai i20 1.2	15600.00	404.00	147.00	173.00	257.00	326.00	50.00	1050.00	7.00	119.00	13.60	170.00	1248.00	55.00	122.00
Volkswagen up!	17750.00	360.00	148.00	164.00	241.00	251.00	35.00	1070.00	6.00	110.00	8.80	196.00	999.00	85.00	200.00
SEAT Ibiza 1.6	21280.00	406.00	144.00	178.00	256.00	355.00	40.00	1262.00	5.00	102.00	10.00	195.00	1598.00	85.00	85.00
Toyota Yaris 1.5	23450.00	394.00	151.00	170.00	251.00	286.00	36.00	1090.00	3.00	75.00	11.80	165.00	1497.00	54.00	111.00
Volvo V40 T2	24650.00	437.00	144.00	180.00	265.00	324.00	62.00	1415.00	8.00	137.00	10.40	190.00	1969.00	90.00	220.00
Honda Civic 1.6	23900.00	452.00	143.00	180.00	270.00	478.00	46.00	1411.00	4.00	93.00	10,00	201.00	1597.00	88.00	300.00
Peugeot 308 SW Allure	29200.00	458.00	146.00	180.00	273.00	610.00	53.00	1394.00	4.00	99.00	9.70	205.00	1499.00	96.00	300.00
Ford Fiesta Active 1.0	21250.00	404.00	148.00	174.00	249.00	303.00	42.00	1179.00	6.00	110.00	9.70	200.00	998.00	103.00	180.00
Citroen C1 VTi 72	10150.00	347.00	146.00	162.00	234.00	196.00	35.00	915.00	4.00	93.00	14.10	157.00	998.00	53.00	93.00
Ford Focus Station Wagon 1.5	27750.00	467.00	182.00	182.00	265.00	608.00	47.00	1413.00	4.00	97.00	10.30	194.00	1498.00	88.00	300.00
Dacia Logan MCV 1.5	11800.00	449.00	155.00	173.00	263.00	573.00	50.00	1165.00	4.00	90.00	11.80	173.00	1461.00	66.00	220.00
Peugeot 108 72	13430.00	348.00	146.00	162.00	234.00	180.00	35.00	935.00	4.00	95.00	15,7	160.00	998.00	53.00	93.00
Mercedes-Benz Classe A 200	29990.00	444.00	146.00	180.00	273.00	370.00	43.00	1355.00	6.00	133.00	8.20	225.00	1332.00	120.00	250,00
Dacia Sandero 1.5	11300.00	406.00	152.00	173.00	259.00	320.00	50.00	1160.00	4.00	90.00	14.60	164.00	1461.00	55.00	200.00
Fiat Tipo Tipo 1.3 Mjt	17400.00	453.00	150.00	179.00	264.00	520.00	45.00	1280.00	5.00	108.00	11.70	180.00	1248.00	70.00	200.00
Hyundai i30 1.6	23850.00	434.00	146.00	180.00	265.00	395.00	50.00	1338.00	4.00	95.00	12,20	186.00	1582.00	70.00	280.00
Alfa Romeo MiTo 1.4 78	13900.00	406.00	145.00	172.00	251.00	270.00	45.00	1155.00	7.00	130.00	13.00	165.00	1368.00	57.00	115.00
Fiat Tipo Tipo 1.4 4	14650.00	453.00	150.00	179.00	264.00	520.00	45.00	1225.00	8.00	133.00	11.50	185.00	1368.00	70.00	127,00
Peugeot 308 Allure	26150.00	425.00	146.00	180.00	262.00	420.00	53.00	1279.00	5.00	123.00	9.80	205.00	1199.00	96.00	230.00
Fiat Tipo Station Wagon 1.6	23650.00	457.00	151.00	179.00	264.00	550.00	45.00	1405.00	4.00	98.00	9.70	200.00	1598.00	88.00	320.00
Alfa Romeo MiTo 1.3	16900.00	406.00	145.00	172.00	251.00	270.00	45.00	1225.00	4.00	83.00	12.50	180.00	1248.00	66.00	200.00
Dr Zero 1.0	9500.00	356.00	152.00	160.00	234.00	180.00	35.00	1050.00	8.00	142.00	13.50	150.00	998.00	51.00	93.00
Fiat 500 1.3 Multijet 95	19350.00	357.00	149.00	163.00	230.00	185.00	35.00	980.00	4.00	89.00	10.70	180.00	1248.00	70.00	200.00
Ford Focus 1.0 EcoBoost	20000.00	438.00	145.00	182.00	265.00	375.00	52.00	1322.00	5.00	107.00	12.10	186.00	999.00	74.00	170.00
Fiat Panda 1.2 Easy	12390.00	365.00	155.00	164.00	230.00	225.00	37.00	1015.00	5.00	125.00	14,50	164.00	1242.00	51.00	102.00
Peugeot 208 82 5	15430.00	397.00	146.00	174.00	254.00	285.00	50.00	1120.00	6.00	109.00	12.20	178.00	1199.00	60.00	118.00
Mini 1.5 One 75	19000.00	398.00	142.00	173.00	257.00	278.00	40.00	1270.00	5.00	123.00	13.40	172.00	1499.00	55.00	160.00
Renault Twingo SCe	15100.00	360.00	155.00	165.00	249.00	188.00	35.00	970.00	5.00	108.00	14.50	151.00	999.00	51.00	91.00

Nella Tabella 1 alcune variabili vengono indicate con un nome abbreviato, in particolare con la colonna chiamata "*bagagliaio"* si intende il volume del bagagliaio, con la colonna chiamata "*serbatoio"* si fa riferimento alla capacità del serbatoio, con la colonna chiamata "*massa"* si intende la massa in ordine di marcia, la colonna chiamata "*consumi"* si riferisce alla variabile consumi urbani, con la colonna chiamata "*emissioni"* si fa riferimento alle emissioni di CO_2, la colonna intestata "*velocità"* fa riferimento alla variabile velocità massima ed infine con la colonna "*coppia"* intende la variabile coppia massima.

1.3 Trattamento delle variabili

Un primo esame necessario nel caso in cui si considerano diverse variabili ai fini di un'analisi statistica multivariata è quello di analizzare l'ordine di grandezza e la variabilità dei caratteri considerati. Infatti, i risultati dell'analisi multivariata possono essere (falsamente) guidati da particolari variabili con eccessivi ordini di grandezza e variabilità. Le variabili in entrata di una tecnica di statistica multivariata devono essere quindi "omogenee" in ordini di grandezza e variabilità. Con lo scopo quindi di stabilire se sia necessario standardizzare i dati, per ciascuna variabile è stato determinato il valore minimo, la mediana, la media aritmetica, il valore massimo e lo scarto quadratico medio. Queste misure di sintesi sono riportate nella Tabella 2.

Tabella 2: statistiche descrittive

Variabili	**Minimo**	**Mediana**	**Media**	**Massimo**	**Scarto Quadratico Medio**
Prezzo	9500	19000	19301	29990	5443,5
Lunghezza	274	406	403,6	467	36,72
Altezza	142	148	149,8	182	7,59
Larghezza	160	174	173,10	186	6,7
Passo	187	257	252,8	273	15,1
Volume del bagagliaio	168	305	331,3	610	109,25
Capacità del serbatoio	28	45	43,63	62	7,21
Massa in ordine di marcia	865	1179	1198	1590	176,25
Consumi urbani	3	5	5,316	10	15,2
Emissioni di CO_2	75	108	110,8	176	15,2
Accelerazione	6,7	11,7	11,65	16,7	1,96
Velocità Massima	150	180	181,7	237	17,16
Cilindrata	875	1368	1324	1984	252,53
Potenza	44	70	74,23	147	20,35
Coppia massima	85	200	187,9	350	79,4

Dalla lettura della Tabella 2 si nota come le variabili presentano valori molto diversi tra loro sotto l'aspetto della variabilità e dell'unità di misura. Di conseguenza ai fini di un'analisi dei gruppi significativa è necessario procedere alla standardizzazione delle variabili originali, in maniera tale da ottenere nuove variabili a media nulla e varianza unitaria che sono tra loro confrontabili.

La standardizzazione dei dati con il software R richiede che all'interno di una matrice vuota, venga caricata la matrice dei dati standardizzati determinata mediante il comando `scale`.

Capitolo 2

Aspetti teorici dell'analisi dei gruppi

2.1 Panoramica sull'analisi

L'analisi dei gruppi è una tecnica di statistica multivariata che consente di aggregare le unità statistiche simili tra loro in gruppi, rispetto ai valori delle *p* variabili osservate sulle *n* unità statistiche. I gruppi devono rispettare le seguenti proprietà:

- coesione interna: i gruppi devono essere il più possibile omogenei al loro interno;
- separazione esterna: i gruppi devono essere il più possibile eterogenei fra di loro.

Per svolgere un'analisi dei gruppi si possono utilizzare due metodologie: i metodi gerarchici e metodi non gerarchici.
I metodi gerarchici danno luogo ad una famiglia di partizioni detta gerarchia. Ogni partizione contiene uno o più gruppi, dove i gruppi devono essere non vuoti, esclusivi ed esaustivi.
Al contrario, i metodi non gerarchici danno luogo ad un'unica partizione in cui il numero di gruppi è fissato a priori.

2.1.1 Metodi gerarchici

I metodi gerarchici si dividono in metodi aggregativi e metodi scissori.
I metodi gerarchici aggregativi partono da una situazione iniziale in cui tutte le unità statistiche sono un gruppo a sé stante, ad ogni passo della procedura le unità statistiche vengono raggruppate, per poi giungere ad una situazione finale in cui tutte le unità statistiche sono comprese in un unico gruppo.

Per svolgere un'analisi dei gruppi con i metodi gerarchici aggregativi innanzitutto è necessario costruire la matrice delle distanze fra le unità statistiche. Per calcolare le distanze fra le unità statistiche sono state utilizzate la distanza Manhattan e la distanza Euclidea, le quali appartengono alla famiglia di distanze di Minkowski.

Si definisce con d_{ik} la distanza tra l'unità statistica u_i e l'unità statistica u_k, i; k =1…,*n*, mentre con il termine x_{ij} si intende l'intensità della j-esima variabile osservata sulla i-esima unità statistica, i=1,…,*n*, j=1,…*p*.

Di seguito vengono riportate le formule della distanza Manhattan (detta anche distanza City Block) e della distanza Euclidea:

- Distanza Manhattan: $d_{ik} = d(u_i, u_k) = \sum_{j=1}^{p} |x_{ij} - x_{kj}|$
- Distanza Euclidea: $d_{ik} = d(u_i, u_k) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2}$

Una volta costruita la matrice di distanze, è necessario scegliere uno dei seguenti criteri di raggruppamento tra gruppi.

- Metodo del legame singolo

  La distanza tra due gruppi C_1 e C_2 consiste nel minimo delle distanze tra ciascuna delle unità statistiche di un gruppo e ognuna delle unità statistiche dell'altro gruppo:

  $$d(C_1, C_2) = \min(d_{ik}) \text{ per } i \in C_1, k \in C_2.$$

- Metodo del legame completo

  La distanza tra due gruppi C_1 e C_2 consiste nel massimo delle distanze tra ciascuna delle unità statistiche di un gruppo e ognuna delle unità statistiche dell'altro gruppo:

  $$d(C_1, C_2) = \max(d_{ik}) \text{ per } i \in C_1, k \in C_2.$$

- Metodo del legame medio

  La distanza tra due gruppi C_1 e C_2 di numerosità rispettivamente n_1 e n_2 consiste nella media aritmetica delle distanze tra ciascuna delle unità statistiche di un gruppo e ognuna delle unità statistiche dell'altro gruppo:

$$d(C_1, C_2) = \frac{1}{n_1 n_2} \sum_i \sum_k d_{ik} \text{ per } i \in C_1, k \in C_2.$$

- Metodo del centroide

  La distanza tra due gruppi C_1 e C_2 è data dalla distanza (Euclidea o Manhattan, a seconda della distanza scelta per costruire la matrice delle distanze) tra i centroidi $\bar{x}_1$ e $\bar{x}_2$ dei due gruppi:

$$d(C_1, C_2) = d(\bar{x}_1, \bar{x}_2).$$

Ciascuno di questi criteri di aggregazione porta ad una partizione diversa delle unità statistiche.

Per valutare la bontà di una partizione si impiegano indici che si basano sulla scomposizione della devianza totale delle *p* variabili. Tale devianza può essere decomposta in devianza totale nei gruppi e devianza totale fra i gruppi. Gli indici che verranno presentati sono: l'indice Pseudo F, l'indice R^2 e l'indice Root Mean Square Standard Deviation (RMSSTD).

Per la trattazione teorica di tali indici si impiegherà la seguente notazione: T indica la matrice devianze e codevianze delle *p* variabili; W indica la matrice devianze e codevianze nei gruppi; infine B rappresenta la matrice devianze e codevianze fra i gruppi.

- Indice Pseudo F

L'indice Pseudo è dato dalla seguente formula:

$$\text{Pseudo F} = \frac{Tr(B)}{G-1} \Big/ \frac{Tr(W)}{n-G}.$$

In generale, con Tr(·) si intende l'operatore matematico *traccia,* che corrisponde alla somma degli elementi posti sulla diagonale principale di una matrice quadrata.

Tr(B) indica la traccia della matrice devianze e codevianze fra i gruppi e corrisponde alla devianza totale fra i gruppi, Tr(W) indica la traccia della matrice devianze e codevianze nei gruppi e corrisponde alla devianza totale nei gruppi, infine G indica il numero di gruppi e *n* il numero di unità statistiche totale.

L'indice Pseudo F assume valori compresi nell'intervallo [0; +∞) e misura la separazione esterna al variare del numero di gruppi, tenendo conto del compromesso necessario tra separazione esterna e numero di gruppi. Può dunque essere interpretato come un indice della bontà della classificazione all'interno di una partizione. Un valore elevato dell'indice Pseudo F segnala un'alta variabilità fra i gruppi, ciò è indice che la partizione considerata è di buona qualità in quanto presenta una maggiore separazione esterna tra i gruppi. Al contrario un basso valore basso dell'indice Pseudo F segnala una ridotta variabilità fra i gruppi, ciò indica che la partizione è di scarsa qualità a causa di una bassa separazione esterna fra i gruppi. Dunque, si preferisce la partizione che presenta il maggior valore dell'indice Pseudo F, considerato il compromesso necessario tra sintesi della classificazione e numerosità dei gruppi.

- Indice R^2

L'indice R^2 è dato dal rapporto tra la devianza totale fra i gruppi e la devianza totale delle *p* variabili.

$$R^2 = \frac{\mathrm{Tr(B)}}{\mathrm{Tr(T)}}.$$

L'indice R^2 assume valori compresi nell'intervallo [0; 1]. Tale indice assume il valore uno nella partizione banale G=*n*, in questo caso la variabilità dei caratteri considerati si deve esclusivamente alle differenze tra i gruppi, mentre assume il valore zero nella partizione banale G=1, in questa situazione la variabilità dei caratteri considerati è dovuta esclusivamente dalle differenze all'interno

dell'unico gruppo. Tanto più l'indice R^2 si avvicina ad uno tanto più nella partizione di riferimento la variabilità fra i gruppi è elevata e quindi maggiore è la coesione esterna.
Si noti che l'indice R^2 non considera il compromesso necessario tra numero dei gruppi e l'esigenza della coesione interna e della separazione esterna.

- Root mean square standard deviation (RMSSTD)

L'indice denominato RMSSTD è definito come

$$\text{RMSSTD}_g = \sqrt{\frac{W_g}{p(ng-1)}}.$$

Dove con g=1...G si indica il generico gruppo g, con w_g si indica la devianza totale delle p variabili all'interno del g-esimo gruppo, con n_g si indica la numerosità del g-esimo gruppo. L'indice RMSSTD assume valori compresi nell'intervallo [0; +∞), ed è un indice che non valuta la partizione nel suo complesso, ma misura la coesione interna al singolo gruppo.
Tanto più l'indice RMSSTD è prossimo a zero, tanto più la variabilità delle p variabili all'interno del gruppo g è bassa, quindi maggiore è la coesione interna del g-esimo *gruppo*.
Con il software R l'analisi dei gruppi svolta con i metodi gerarchici aggregativi richiede che a partire dalla matrice dei dati standardizzati, venga calcolata la matrice delle distanze tramite il comando `dist`, scegliendo il tipo di metrica da impiegare (Euclidea o Manhattan).
Una volta calcolata la matrice delle distanze, mediante l'istruzione `hclust` si implementa la cluster analysis secondo uno dei seguenti criteri di aggregazione: metodo del legame singolo, metodo del legame completo, metodo del legame medio e metodo del centroide.

2.1.2 Metodi non gerarchici: il metodo delle k-medie

I metodi di classificazione non gerarchici producono un'unica partizione in cui il numero di gruppi è fissato a priori.
Il più famoso algoritmo di classificazione non gerarchico è il metodo delle k-medie e si articola nelle seguenti fasi:

1) il primo passo consiste nello scegliere i g poli iniziali e allocare le unità statistiche al polo più vicino secondo una certa metrica, costruendo così la partizione iniziale.
2) per ogni unità statistica si calcola la distanza con il centroide di ciascun gruppo e se la distanza più piccola non è ottenuta in corrispondenza del centroide del gruppo di appartenenza, allora l'unità statistica deve essere assegnata al gruppo corrispondente al centroide più vicino e aggiornare il centroide sia del vecchio che del nuovo gruppo di appartenenza.
3) si ripete la fase 2 fino a quando l'algoritmo non giunge a convergenza, ciò accade in due casi:
    a) non si verifica alcuna variazione dei gruppi rispetto all'iterazione precedente;
    b) si raggiunge un certo numero di iterazioni fissato all'inizio della procedura.

La metrica impiegata per calcolare la distanza tra unità statistiche e centroidi dei g gruppi è solitamente quella Euclidea, questo perché è possibile dimostrare, che il metodo delle k-medie (svolto mediante la distanza Euclidea) ha come obiettivo la ricerca della partizione a g gruppi che minimizza la devianza nei gruppi.

Capitolo 3

Applicazione dell'analisi dei gruppi

3.1 Aggregazione delle unità statistiche con il metodo delle k-medie

Con il software R l'algoritmo di classificazione delle k-medie viene eseguito mediante il comando `kmeans,` indicando il numero di gruppi con i quali realizzare l'aggregazione ed individuando i poli iniziali mediante il comando `set.seed(123)`.

Per individuare il numero di gruppi ottimale si può utilizzare un metodo grafico che consiste nel cercare un "gomito" nel grafico rappresentante la devianza totale nei gruppi al variare del numero di cluster. Tale grafico è riportato nella Figura 1 e si nota (seppur in maniera non così evidente) che il gomito si trova in corrispondenza della partizione con tre gruppi, ciò significa che in base a questo criterio essa è quella ottimale.

Nella Figura 2 invece è riportato il valore dell'indice Pseudo F al variare del numero di gruppi. Si noti che si trascurerà la partizione con due gruppi che, seppur presenta il valore più alto dello Pseudo F, non è ottimale in quanto è una partizione caratterizzata da un'eccessiva sintesi e quindi non consente di caratterizzare in maniera significativa i gruppi. La partizione con tre gruppi è quella che presenta il valore più elevato dell'indice Pseudo F, di conseguenza si può ritenere la partizione con tre gruppi quella ottimale per un'aggregazione delle unità statistiche secondo l'algoritmo di classificazione non gerarchico delle k-medie.

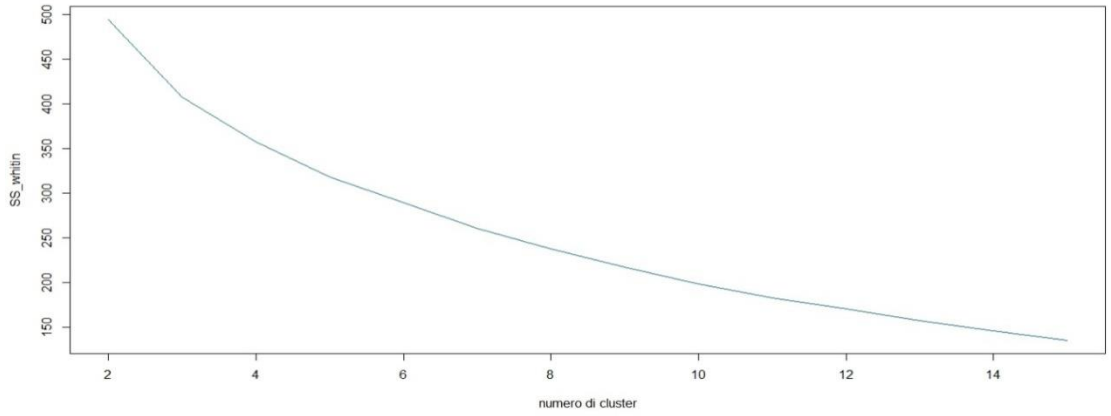

Figura 1: devianza totale nei gruppi al variare del numero di cluster

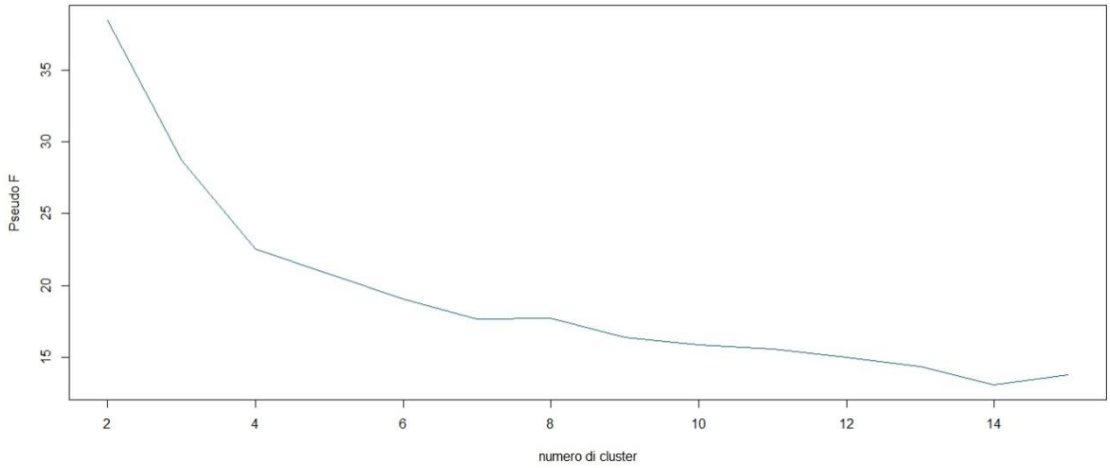

Figura 2: indice Pseudo F al variare del numero di gruppi

Per visualizzare il risultato della procedura si è impiegata la funzione `fviz_cluster` disponibile nel pacchetto `factoextra`.

Il comando `fviz_cluster` per agevolare la visualizzazione grafica dell'output di un'analisi dei gruppi estrae le Componenti Principali dal data set, per poi rappresentare i risultati di una cluster analysis in un sistema cartesiano a due dimensioni in cui l'asse delle ascisse rappresenta la prima componente principale, mentre l'asse delle ordinate rappresenta la seconda componente principale.

L'analisi delle Componenti Principali è una tecnica di statistica multivariata che ha lo scopo di ridurre il numero delle *p* variabili presenti all'interno di un data set

creando delle nuove variabili (dette Componenti Principali) che non sono direttamente osservabili e sono incorrelate tra loro. Le prime due Componenti Principali sono in grado di rappresentare l'intero data set se riproducono una quota sufficientemente elevata della variabilità totale delle variabili osservate. Nel caso in esame la prima componente principale riproduce il 56,10% della variabilità totale, mentre la seconda componente principale riproduce il 13,10% della variabilità totale, quindi le prime due Componenti Principali replicano il 69,10% della variabilità totale delle variabili osservate. Questa quota viene considerata sufficiente.

Poiché i metodi non gerarchici richiedono che il numero dei gruppi venga fissato a priori, il metodo delle k-medie è stato eseguito scegliendo di aggregare le unità statistiche prima in 4 gruppi e poi in 3 gruppi. (come suggerito dall'analisi esplorativa in Figura 1 e in Figura 2)

Nella Figura 3 è rappresentata la partizione composta da 3 gruppi, nella Figura 4 la partizione composta da 4 gruppi.

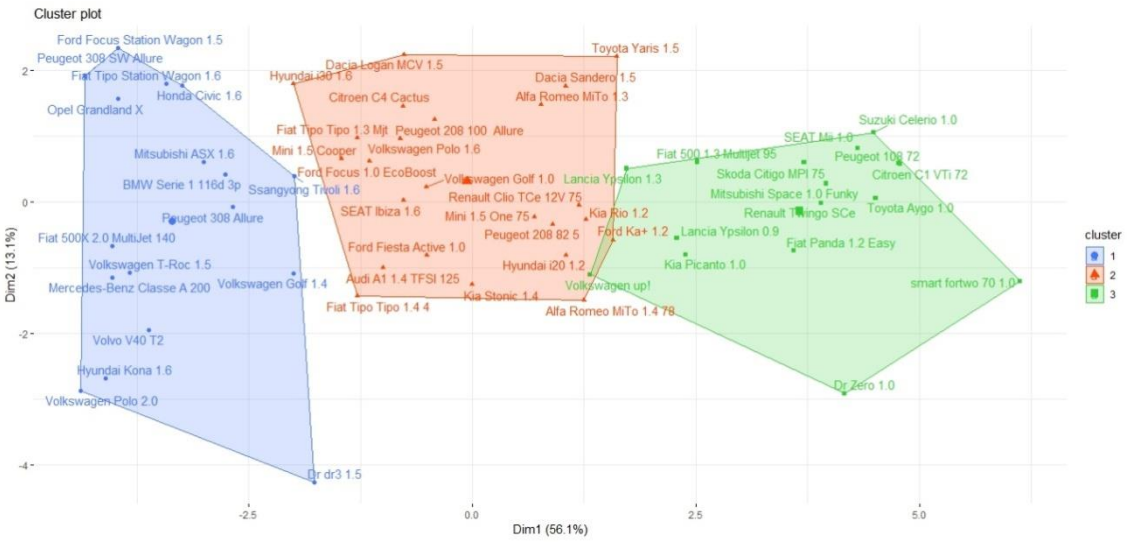

Figura 3: cluster plot con metodo delle k-medie con 3 gruppi

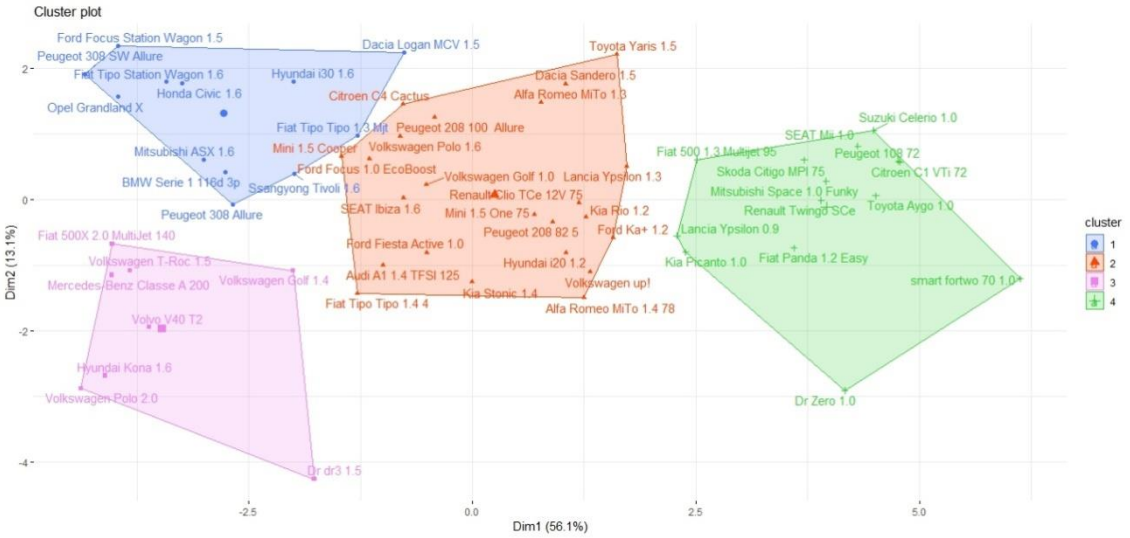

Figura 4: cluster plot con metodo delle k-medie con 4 gruppi

I risultati ottenuti tramite l'applicazione del metodo non gerarchico delle k-medie verranno analizzati in dettaglio quando si confronteranno con i risultati ottenuti mediante i metodi gerarchici aggregativi.

3.2 Aggregazione delle unità statistiche con i metodi gerarchici

3.2.1 Scelta della partizione

L'analisi gerarchica ha dato luogo a otto gerarchie dette anche "famiglie di partizioni" derivanti dall'applicazione del metodo del legame singolo, del metodo del legame medio, metodo del legame completo e del metodo del centroide, sia con la distanza Euclidea che con la distanza Manhattan.

Per individuare all'interno di ciascuna gerarchia la partizione ottimale è stato calcolato l'indice Pseudo F tramite la funzione `NbClust` per le partizioni a tre, quattro e cinque gruppi, in quanto in un'analisi dei gruppi è necessario giungere ad un compromesso tra separazione esterna, coesione interna e numero di gruppi.

In base all'indice Pseudo F si sceglie la partizione che presenta il valore più elevato dell'indice, in quanto segnala il maggior grado di separazione esterna tra i gruppi, tenendo conto del numero di cluster.

I valori dell'indice Pseudo F sono riportati nella Tabella 3 e nella Tabella 4.

Tabella 3: indice Pseudo F nelle partizioni realizzate con la distanza Manhattan

PSEUDO F			
Numero di gruppi / Metodi	**3**	**4**	**5**
Legame singolo	2,7467	3,2767	2,9880
Legame completo	26,6469	21,3944	18,2964
Legame medio	18,1569	21,1732	18,9679
Metodo del centroide	3,1349	2,9797	11,3863

Tabella 4: indice Pseudo F nelle partizioni realizzate con la distanza Euclidea

PSEUDO F			
Numero di gruppi / Metodi	**3**	**4**	**5**
Legame singolo	3,1542	3,2767	3,4055
Legame completo	24,665	18,6479	17,1765
Legame medio	18,6175	19,3276	15,8021
Metodo del centroide	3,5660	3,5571	3,4055

Con il metodo del legame completo realizzato sia con la distanza Manhattan che con la distanza Euclidea, l'indice Pseudo F individua come tre il numero ottimale di gruppi, con valori dello Pseudo F piuttosto simili tra loro, evidenziando una variabilità fra i gruppi simile per entrambe le partizioni. Considerando il metodo del legame medio l'indice Pseudo F individua un numero ottimale di gruppi pari a quattro sia che per costruire la matrice di distanze sia stata impiegata la distanza Manhattan o la distanza Euclidea. I valori dello Pseudo F sono simili tra le due partizioni, evidenziando una variabilità fra i gruppi simile per entrambe le partizioni. In merito al metodo del legame singolo, nel caso della distanza Euclidea l'indice individua come partizione ottimale quella con cinque gruppi, mentre con la distanza Manhattan la partizione ottimale è quella con quattro

gruppi. Ancora una volta si osservano valori piuttosto prossimi dell'indice Pseudo F tra le due partizioni, evidenziando una variabilità fra i gruppi simile per entrambe le partizioni. Infine, considerando il metodo del centroide l'indice Pseudo F individua un numero di gruppi differente a seconda che sia stata utilizzata o la distanza Manhattan o la distanza Euclidea. In particolare, nel primo caso l'indice segnala che il numero ottimale di gruppi è 5, mentre nel secondo caso il numero ottimale di gruppi è 3. Inoltre, il valore dell'indice Pseudo F in questo caso differisce notevolmente tra le due partizioni, segnalando una variabilità fra i gruppi significativamente differente tra le due partizioni ottimali generate col metodo del centroide. Questo significa che il metodo del centroide, rispetto agli altri criteri di aggregazione, è molto sensibile al tipo di distanza impiegata.

3.2.2 Dendrogramma

Per giungere alla scelta della partizione ottimale ottenuta secondo i metodi gerarchici aggregativi, oltre a valutare l'indice Pseudo F si userà un altro importante strumento caratteristico dell'analisi dei gruppi ossia il dendrogramma.
Il dendrogramma è un grafico nel quale sull'asse delle ascisse vengono rappresentate le unità statistiche, mentre sull'asse delle ordinate sono indicate le distanze alle quali i gruppi si uniscono durante la procedura. Attraverso il taglio del dendrogramma si individuano i gruppi di unità statistiche che si formano ai diversi livelli di distanza.
Per ogni partizione ritenuta ottimale dall'analisi dell'indice Pseudo F si è proceduto a rappresentare il relativo dendrogramma, colorando con lo stesso colore le unità statistiche appartenenti allo stesso gruppo ed evidenziando i gruppi che si sono formati con un rettangolo, con lo scopo di individuare il metodo che meglio si addice ai dati.

Per la costruzione ed il taglio del dendrogramma si utilizza il comando `cutree`, mentre per la sua visualizzazione si usa il comando `fviz_dend` contenuto nel pacchetto `factoextra`.

3.2.2.1 Analisi del dendrogramma: metodo del centroide

Il metodo del centroide svolto sia con la distanza Euclidea che con la distanza Manhattan non ha portato a risultati soddisfacenti.
Dalla lettura di entrambi i dendrogrammi (Figura 5 e Figura 6), si nota innanzitutto come il grafico presenta diverse intersezioni tra i rami del diagramma ad albero, questo significa che la distanza alla quale due gruppi si sono uniti ad una fase dell'analisi, è minore rispetto alla distanza alla quale si sono aggregati i due gruppi nella fase precedente. Inoltre, la lettura di entrambi i dendrogrammi prodotti con il metodo del centroide non consente di identificare chiaramente i gruppi che si sono formati, tant'è che nemmeno il programma è in grado di tracciare i rettangoli che delimitano ciascun gruppo. Ulteriore segnale della pessima qualità del metodo del centroide, per l'analisi dei dati in oggetto, è che il programma non è in grado di colorare il dendrogramma relativo al metodo del centroide svolto con la distanza Euclidea.
Si può osservare inoltre come ad un dendrogramma di pessima qualità si associa un basso valore dello Pseudo F (con la distanza Euclidea, in corrispondenza della partizione ottimale l'indice assume un valore pari a 3,5660, mentre con la distanza Manhattan un valore pari a 11,3863).
In forza di queste osservazioni si conclude che il metodo del centroide realizzato sia con la distanza Manhattan che con la distanza Euclidea non è in grado di cogliere i gruppi che sono presenti all'interno del data set e di conseguenza si è deciso di scartare questo criterio di aggregazione delle unità statistiche.

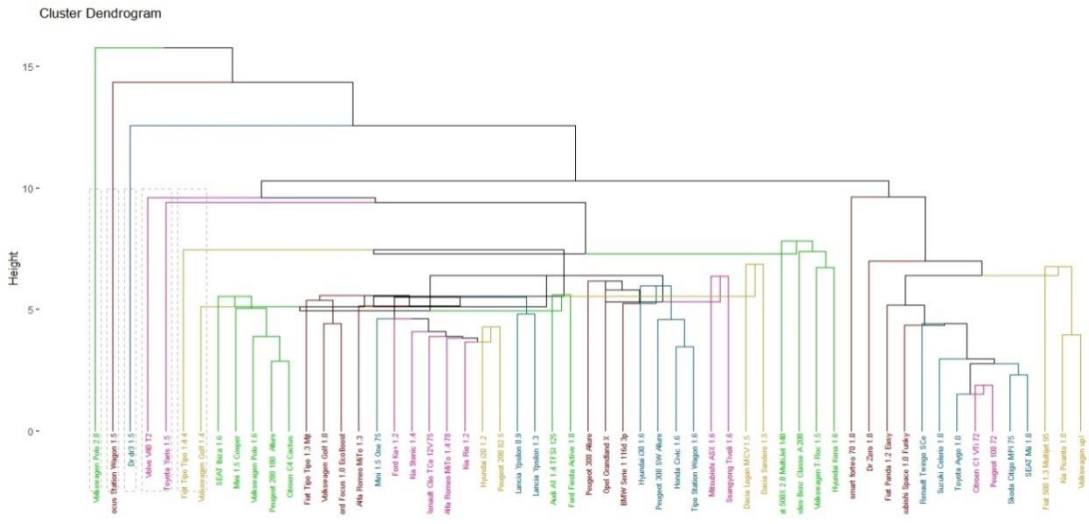

Figura 5: dendrogramma con metodo del centroide e distanza Manhattan

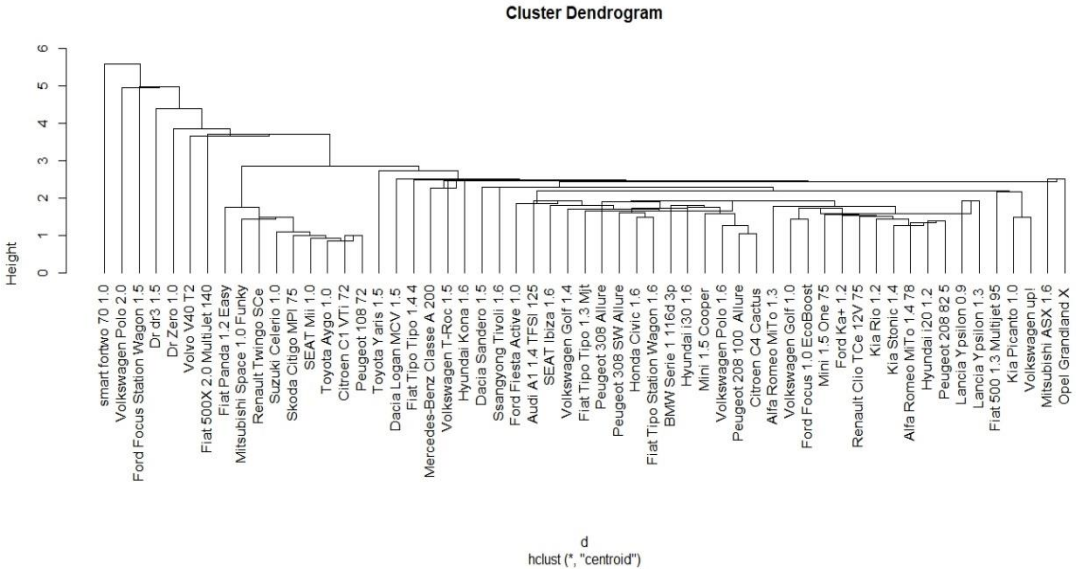

Figura 6: dendrogramma con metodo del centroide e distanza Euclidea

3.2.2.2 Analisi del dendrogramma: metodo del legame singolo

Il metodo del legame singolo svolto sia con la distanza Euclidea che con la distanza Manhattan ha portato a risultati deludenti.

Il metodo del legame singolo condotto con la distanza Euclidea (Figura 7) ha portato alla formazione di cinque gruppi, di cui quattro contengono una sola unità statistica evidenziando così la loro particolarità ed è presente un gruppo contenente la quasi totalità delle unità statistiche.

Questa situazione non si concilia con il fatto che in un'analisi dei gruppi si devono evitare situazioni in cui ci sono gruppi contenenti un numero eccessivamente elevato di unità statistiche e gruppi con pochissime unità statistiche al loro interno.

Inoltre, una partizione così composta non consente di caratterizzare in maniera significativa le unità statistiche appartenenti al gruppo più numeroso.

A risultati analoghi si giunge applicando il metodo del legame singolo con la distanza Manhattan (Figura 8). La partizione ottimale suggerita dall'indice Pseudo F è composta da quattro gruppi e ancora una volta è presente un gruppo contenente la maggior parte delle unità statistiche e tre gruppi contenenti una sola unità statistica.

Anche con il metodo del legame singolo si osserva come ad un dendrogramma di pessima qualità si associa un basso valore dello Pseudo F, segnalando una scarsa eterogeneità fra i gruppi.

In conclusione, per queste ragioni si ritiene opportuno ai fini dell'analisi scartare anche il metodo del legame singolo.

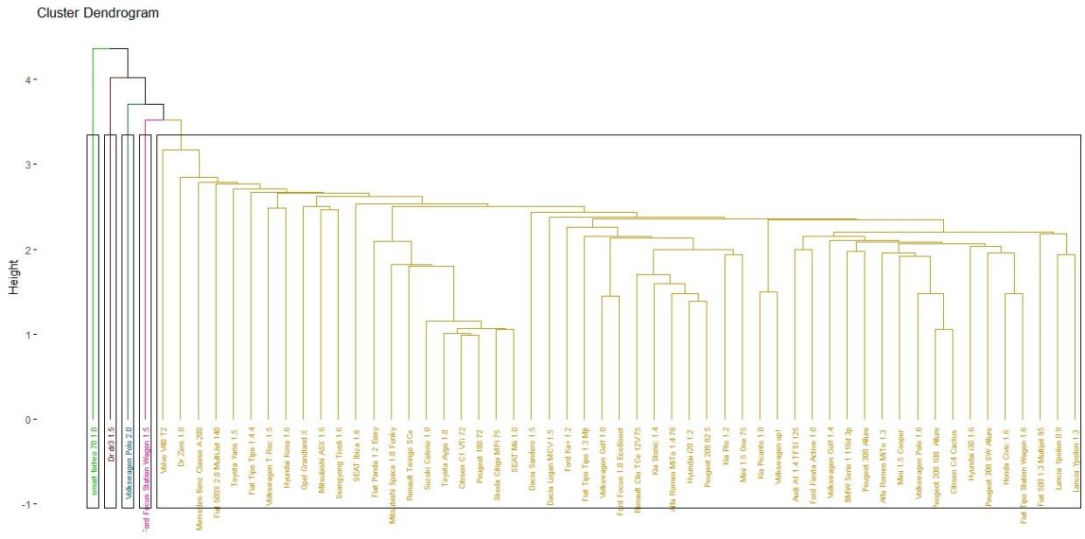

Figura 7: dendrogramma con metodo del legame singolo e distanza Euclidea

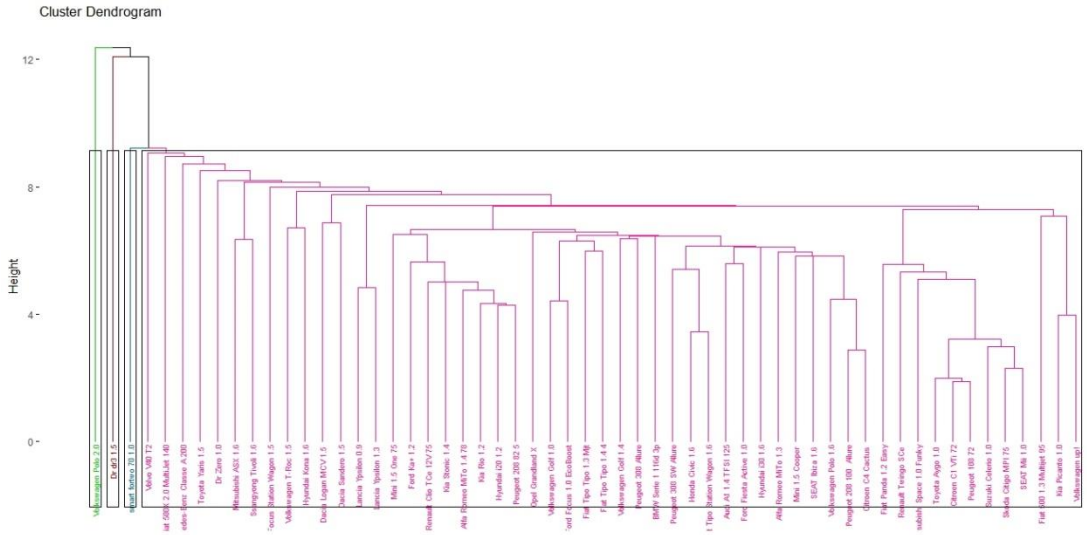

Figura 8: dendrogramma con metodo del legame singolo e distanza Manhattan

3.2.2.3 Analisi del dendrogramma: metodo del legame medio

Il metodo del legame medio realizzato sia con la distanza Manhattan (Figura 10) che Euclidea (Figura 9) ha portato alla formazione di quattro gruppi. Si noti che un gruppo, in entrambi i casi, è costituito da un'unica unità statistica che corrisponde al modello corrispondente al suv Dr. Dr3 1.5 evidenziando la sua particolarità rispetto agli altri modelli, mentre gli altri tre gruppi presentano una numerosità bilanciata tra loro in entrambe le partizioni.

Si è deciso di scartare il metodo del legame medio per le seguenti ragioni: in primo luogo le due partizioni con quattro gruppi generate con il metodo del legame medio differiscono significativamente con la rispettiva partizione con quattro gruppi ottenuta con il metodo delle k-medie (Figura 4). Infatti, se l'obiettivo dell'analisi dei gruppi è quello di identificare i "reali" gruppi sottostanti ai dati questi dovrebbero emergere indipendentemente dalla tecnica di aggregazione impiegata. In secondo luogo, il valore dell'indice Pseudo F in corrispondenza delle due partizioni ottimali assume valori inferiori rispetto quelli che l'indice assume in corrispondenza delle due partizioni ottimali ottenute con il metodo del legame completo (maggiori dettagli saranno riportati nel prossimo paragrafo). Poiché l'indice Pseudo F è una misura della bontà della partizione

significa che le due partizioni ottenute con il metodo del legame completo sono migliori in termini di separazione esterna tra i gruppi rispetto al metodo del legame medio.

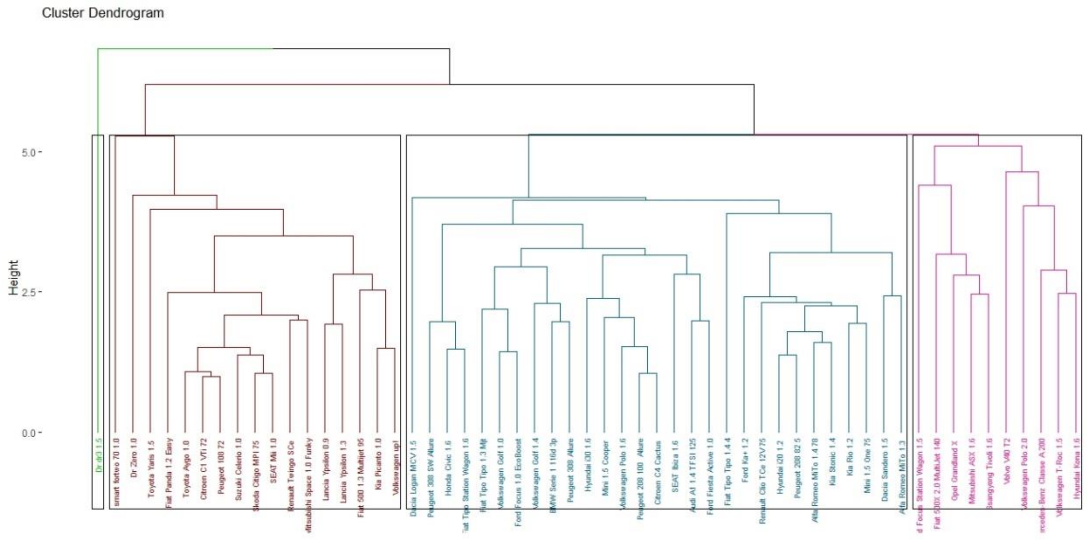

Figura 9: dendrogramma con metodo del legame medio e distanza Euclidea

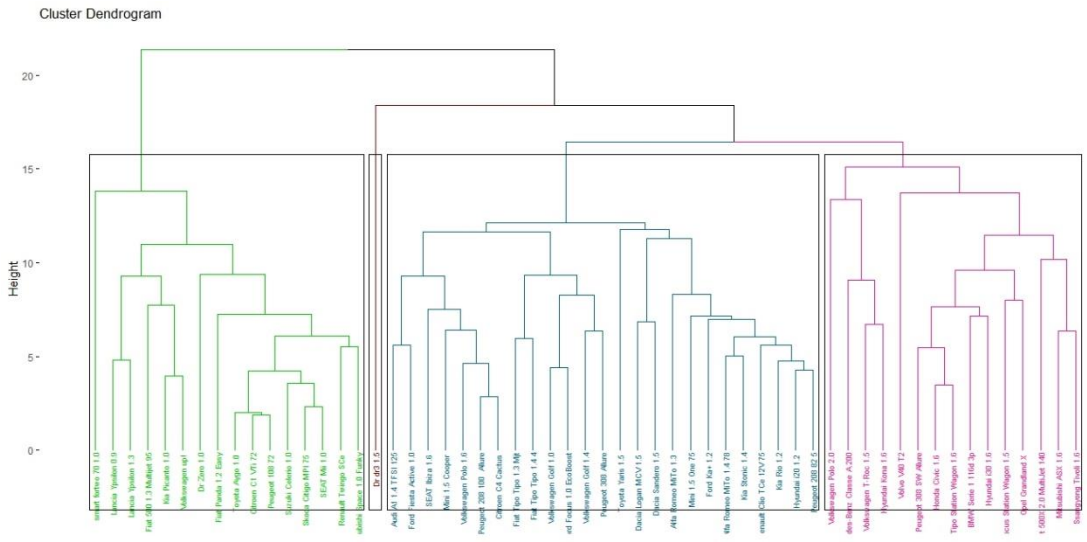

Figura 10: dendrogramma con metodo del legame medio e distanza Manhattan

3.2.2.4 Analisi del dendrogramma: metodo del legame completo

Il metodo del legame completo realizzato sia con la distanza Euclidea (Figura 11) che con la distanza Manhattan (Figura 12), porta alla selezione di tre gruppi che risultano graficamente ben distinti e di numerosità accettabile.
Si è deciso di non conservare le partizioni con quattro gruppi ottenute con la distanza City Block e la distanza Euclidea, non solo per il valore inferiore dell'indice Pseudo F, ma anche per il fatto che, tagliando il dendrogramma ad un'altezza tale da ottenere quattro gruppi si nota come i due gruppi di numerosità maggiore rimangono inalterati mentre il gruppo di numerosità più piccola (evidenziato in rosso in entrambi i dendrogrammi) si scompone in due gruppi.
Il metodo del legame completo, seppur presenti una certa sensibilità alla metrica impiegata (confrontando le due partizioni si riscontra che 16 unità statistiche su 57 si allocano in gruppi differenti) è quello che meglio si presta alla aggregazione dei dati raccolti per le seguenti ragioni. Innanzitutto, fra tutte le partizioni generate con i metodi gerarchici aggregativi, le due ottenute con il metodo del legame completo sono quelle che più si avvicinano alla partizione con tre gruppi realizzata con l'algoritmo di classificazione non gerarchico delle k-medie (Figura 3). Infatti, i tre gruppi ottenuti con i due metodi sono tra loro simili anche se non perfettamente sovrapponibili a causa di unità statistiche che si allocano in gruppi diversi a seconda del metodo utilizzato.
Questa analogia è ancor più vera per la partizione ottenuta con il criterio del legame completo realizzato con la distanza Manhattan: solo 8 unità statistiche su 57 si allocano in gruppi diversi. Confrontando le partizioni si osserva che il gruppo che nella Figura 12 è evidenziato in blu coincide quasi perfettamente con il gruppo che nella Figura 3 è colorato in verde. Il cluster relativo al metodo gerarchico oltre a contenere tutte le unità statistiche comprese nel gruppo generato col metodo delle k-medie, contiene due modelli di automobili in più che sono la Toyota Yaris 1.5 e la Ford Ka 1.2. Al contrario, il gruppo che nella Figura 12 è colorato in rosso presenta una maggior instabilità: a questo gruppo con il metodo delle k-medie si uniscono altre sei unità statistiche formando il

gruppo che nella Figura 3 è evidenziato in rosso. Queste autovetture col metodo del legame completo, realizzato con la distanza Manhattan invece sono incluse nel gruppo verde della Figura 12.

Altro punto a favore del metodo del legame completo risiede nell'indice Pseudo F. Esso assume valori più elevati nelle partizioni generate con il metodo del legame completo rispetto alle partizioni generate con gli altri criteri di aggregazione, ciò segnala partizioni che presentano una maggior variabilità fra i gruppi rispetto alle altre partizioni.

In forza di queste osservazioni si sceglie il metodo del legame completo realizzato con entrambe le distanze in quanto sembrerebbe il metodo che meglio riesce a cogliere i gruppi presenti all'interno del data set, pertanto si è scelto di adottare questo metodo per procedere all'aggregazione delle unità statistiche.

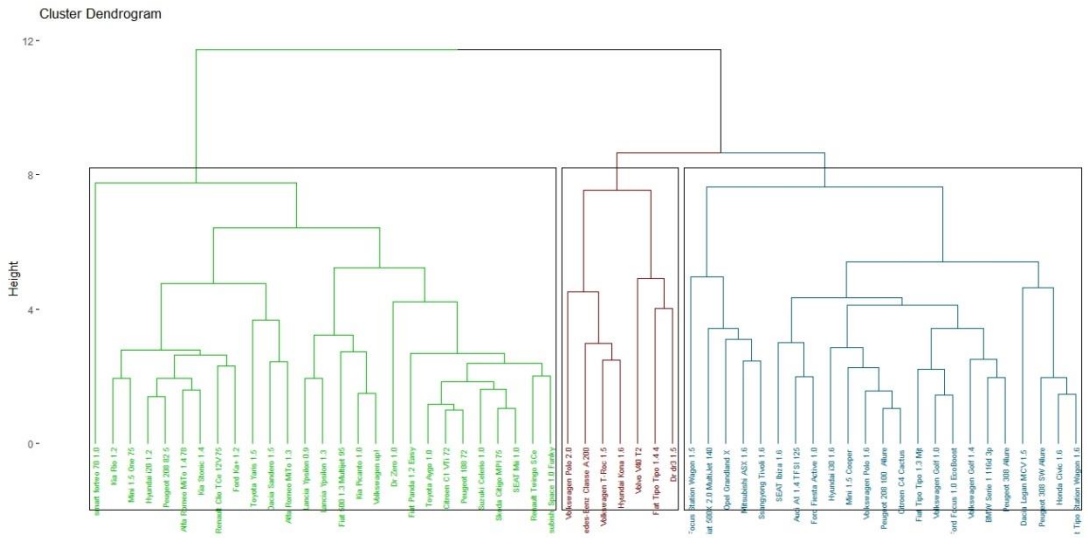

Figura 11: dendrogramma con metodo del legame completo e distanza Euclidea

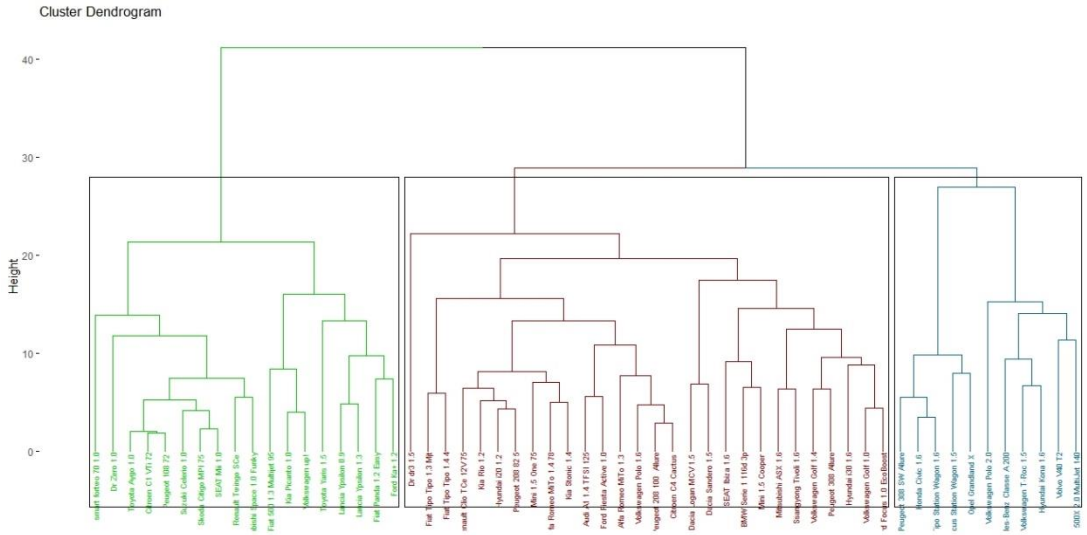

Figura 12: dendrogramma con metodo del legame completo e distanza Manhattan

3.3 Caratterizzazione dei gruppi

In questo paragrafo si analizzano le differenze tra le due partizioni ottenute col metodo del legame completo impiegando sia la distanza Euclidea che la distanza Manhattan. Successivamente si caratterizzeranno i gruppi della partizione ottenuta con la distanza Manhattan in quanto presenta un valore dell'indice Pseudo F superiore rispetto alla partizione ottenuta con la distanza Euclidea, inoltre si avvicina maggiormente alla partizione ottenuta con il metodo delle k-medie.

3.3.1 Confronto fra la partizione ottenuta con la distanza Manhattan e la distanza Euclidea

Un primo esame che si può effettuare consiste nel confrontare la numerosità dei gruppi delle due partizioni ottenute con il metodo del legame completo. Questo confronto è svolto nella Tabella 5.

Tabella 5: Numerosità dei gruppi

	Distanza Manhattan	**Distanza Euclidea**
Gruppo 1	28	23
Gruppo 2	18	27
Gruppo 3	11	7

Successivamente, si possono individuare quali sono le unità statistiche che compongo i gruppi nelle due partizioni, al fine di scoprire quali autovetture cambiano gruppo a seconda della metrica impiegata.

Per realizzare tale confronto si è deciso di utilizzare come punto di partenza i gruppi ottenuti con la distanza Manhattan e poi di verificare le differenze con la partizione alternativa, ossia quella ottenuta con la distanza Euclidea.

Nella Tabella 6 è riportata la partizione ottenuta con la distanza Manhattan, mentre nella Tabella 7 è riportata la partizione ottenuta con la distanza Euclidea.

Nella Tabella 6 sono stati utilizzati dei colori per identificare i gruppi che si sono formati e nella Tabella 7 sono state evidenziate le unità statistiche con gli stessi colori utilizzati nella Tabella 6.

Nelle due partizioni ottenute con il metodo del legame completo nessun gruppo rimane inalterato. In particolare, i veicoli Honda Civic 1.6, Ford Focus Station Wagon 1.5, Fiat Tipo Station Wagon 1.6, Peugeot 308 SW Allure, Fiat 500X 2.0 Multijet, Opel Grandland X, con la distanza Euclidea sono collocati nel gruppo 1, mentre impiegando la distanza City Block queste autovetture sono allocate nel gruppo 3. Il gruppo 2 della partizione ottenuta con la distanza Euclidea comprende tutti i veicoli appartenenti al gruppo 2 della partizione ottenuta con la distanza Manhattan, a queste unità statistiche, usando la distanza Euclidea, si aggiungono i seguenti veicoli: Kia Rio 1.2, Dacia Sandero 1.5, Hyundai i20 1.2, Alfa Romeo MiTo 1.4, Alfa Romeo MiTo 1.3, Peugeot 208 82 5, Kia Stonic 1.4, Renault Clio TCe 12V 75, Mini 1.5 One 75. Questi veicoli usando la distanza Manhattan sono inseriti nel gruppo 1. Infine, i veicoli Fiat Tipo 1.4 e DR dr3 1.5 se si impiega la distanza Euclidea vengono inseriti nel gruppo 3, mentre se si usa la distanza Manhattan sono allocati nel gruppo 1.

In totale sono 16 le unità statistiche che con il metodo del legame completo, a seconda che venga impiegata la distanza Euclidea o la distanza Manhattan, si allocano in gruppi differenti. Ciò segnala una certa sensibilità del metodo del legame completo al tipo di distanza impiegata.

Tabella 6: Partizione con la distanza Manhattan

Gruppo 1	Peugeot 208 100 Allure, Audi A1 1.4 TFSI 125, Volkswagen Polo 1.6, Kia Rio 1.2, BMW serie 1 116d 3p, Citroen C4 Cactus, Ford Fiesta Active 1.0, Dacia Sandero 1.5, Hyundai i30 1.6, Fiat Tipo tipo 1.4, Volkswagen Golf 1.0, Volkswagen Golf 1.4, Mini 1.5 Cooper, Hyunadi i20 1.2, SEAT Ibiza 1.6, Dacia Logan MCV 1.5, Fiat Tipo tipo 1.3 Mjt, Alfa Romeo Mito 1.4, Peugeot 308 Allure, Alfa Romeo Mito 1.3, Ford Focus 1.0 EcoBoost, Peugeot 208 82 5, Kia Stonic 1.4, Dr dr3 1.5, Renault Clio TCe 12V 75, Mini 1.5 One 75, Mistubishi ASX 1.6, Ssangyong Tivoli 1.6
Gruppo 2	Kia Picanto 1.0, Volkswagen up!, Toyota Yaris 1.5, Peugeot 108 72, Dr Zero 1.0, Lancia Ypsilon 0.9, Lancia Ypsilon 1.3, Citroen C1 Vti 72, Fiat 500 1.3 Multijet 95, Renault Twingo Sce, SEAT Mii 1.0, Ford Ka+ 1.2, Mitsubishi Space 1.0 Funky, Fiat Panda 1.2 Easy, Skoda Citigo MPI 75, Smart fortwo 70 1.0, Toyota Aygo 1.0, Suzuki Celerio 1.0
Gruppo 3	Honda Civic 1.6, Ford Focus Station Wagon 1.5, Fiat Tipo Station Wagon 1.6, Volkswagen Polo 2.0, Volvo V40 T2, Peugeot 308 SW Allure, Mercedes Benz Classe A 200, Fiat 500X 2.0 Multijet 140, Hyundai Kona 1.6, Volkswagen T-Roc 1.5, Opel Grandland X

Tabella 7: Partizione con la distanza Euclidea

Gruppo 1	Peugeot 208 100 Allure, Audi A1 1.4 TFSI 125, Volkswagen Polo 1.6, BMW serie 116d 3p, Citroen C4 Cactus, Honda Civic 1.6, Ford Fiesta Active 1.0, Ford Focus Station Wagon 1.5, Hyundai i30 1.6, Fiat Tipo Station Wagon 1.6, Volkswagen Golf 1.0, Volkswagen Golf 1.4, Mini 1.5 Cooper, SEAT Ibiza 1.6, Peugeot 308 SW Allure, Dacia Logan MCV 1.5, Fiat Tipo tipo 1.3 Mjt, Peugeot 308 Allure, Ford Focus 1.0 EcoBoost, Fiat 500X 2.0 Multijet, Mitsubishi ASX 1.6, Opel Grandland X, Ssangyong Tivoli 1.6
Gruppo 2	Kia picanto 1.0, Kia Rio 1.2, Volkswagen Up!, Toyota Yaris 1.5, Puegot 108 72, Dacia Sandero 1.5, Dr Zero 1.0, Lancia Ypsilon 0.9, Lancia Ypsilon 1.3, Hyundai i20 1.2, Citroen C1 VTi 72, Alfa Romeo MiTo 1.4, Alfa Romeo MiTo 1.3, Fiat 500 1.3 Multijet 95, Peugeot 208 82 5, Reanult Twingo SCe, Kia Stonic 1.4, SEAT Mii 1.0, Renault Clio TCe 12V 75, Ford Ka+ 1.2, Mitsubishi Space 1.0 Funky, Fiat Panda 1.2 Easy, Mini 1.5 One 75, Skoda Citigo MPI 75, Smart fortwo 70 1.0, Toyota Aygo 1.0, Suzuki Celerio 1.0
Gruppo 3	Fiat Tipo tipo 1.4, Volkswagen Polo 2.0, Volvo V40 T2, Mercedes Benz Classe A 200, DR dr3 1.5, Hyundai Kona 1.6, Volkswagen T-Roc 1.5

Nella Tabella 8 sono riportati i valori dell'indice Root Mean Square Standard Deviation (RMSSTD) calcolato per ogni gruppo delle due partizioni ottenute con il metodo del legame completo. Poichè l'indice RMSSTD esprime una misura della variabilità all'interno del *g*-esimo gruppo esso è preferibile che assuma un valore il più piccolo possibile. Un valore basso di tale indice segnala una scarsa variabilità all'interno del *g*-esimo gruppo, indicando così una elevata coesione interna al singolo gruppo poiché le unità statistiche in esso contenute assumono valori simili rispetto alle *p* variabili sulle quali è stata effettuata l'analisi dei gruppi.

Dalla lettura della Tabella 8 si osserva che l'indice RMSSTD assume valori inferiori nei gruppi della partizione ottenuta con la distanza Manhattan rispetto alla partizione alternativa ottenuta con la distanza Euclidea. Questo significa che i gruppi contenuti nella partizione realizzata con la distanza Manhattan contengono veicoli più omogenei tra loro rispetto ai gruppi appartenenti alla partizione realizzata con la distanza Euclidea.

Tabella 8: indice RMSSTD

Distanza / Gruppi	Distanza Euclidea		Distanza Manhattan	
	Devianza totale	**RMSSTD**	**Devianza totale**	**RMSSTD**
Gruppo 1	170,692	0.719	202,917	0,708
Gruppo 2	195,816	0,708	109,982	0,657
Gruppo 3	72,460	0,897	109,865	0,856

Nella Tabella 9 sono riportati i valori dell'indice Pseudo F ed R^2 per ognuna delle due partizioni ottenute con il metodo del legame completo:

Tabella 9: Indice R^2 e Pseudo F

Indici / Distanza	Devianza totale	Devianza totale nei gruppi	Devianza totale fra i gruppi	R^2	Pseudo F
Distanza Manhattan	840	422,764	417,236	0,4967	26,6469
Distanza Euclidea	840	438,968	401,032	0,4774	24,6650

Dalla Tabella 9 si nota che la partizione ottenuta mediante la distanza Manhattan (a parità di numero di gruppi) presenta un valore dell'indice R^2 superiore rispetto al valore assunto nella partizione con tre gruppi ottenuta con la distanza Euclidea. Questo significa che la partizione ottenuta con la distanza Manhattan presenta una maggior variabilità fra i gruppi, quindi una maggior coesione esterna, rispetto alla partizione alternativa ottenuta con la distanza Euclidea. La medesima informazione viene fornita dall'indice Pseudo F: esso assume un valore più elevato nella partizione ottenuta con la distanza Manhattan rispetto alla partizione ottenuta con la distanza Euclidea.

Unendo i risultati riportati nella Tabella 8 e nella Tabella 9 si conclude che la partizione con tre gruppi ottenuta mediante la distanza Manhattan presenta un maggior grado di separazione esterna fra i gruppi rispetto alla partizione ottenuta

con la distanza Euclidea. Inoltre, i gruppi contenuti nella partizione realizzata con la distanza Manhattan presentano una maggior coesione interna rispetto ai gruppi contenuti nella partizione alternativa ottenuta con la distanza Euclidea.
In forza di queste osservazioni e della maggiore somiglianza con la partizione con tre gruppi ottenuta con il metodo non gerarchico delle k-medie si preferisce mantenere la partizione con tre gruppi realizzata con il metodo del legame completo ottenuta con la distanza Manhattan.

3.3.2 Etichettatura dei gruppi della partizione ottimale

Un primo esame sull'analisi svolta consiste nello studio della composizione dei gruppi della partizione ottenuta con la distanza Manhattan in base alla carrozzeria dei veicoli. A tal fine alle unità statistiche è stata una attribuita una delle due etichette: "suv e station wagon" e "due volumi e tre volumi". Nella Figura 13 è riportato un grafico a barre rappresentante la composizione dei gruppi in base alla carrozzeria dei veicoli. Dalla lettura del grafico si osserva che il primo gruppo è costituito in gran parte da due volumi e tre volumi ed una piccola percentuale di suv e station wagon che corrispondono ai seguenti veicoli: Ssangyong Tivoli 1.6, Mitsubishi ASX 1.6, Dr dr3 1.5. e Kia Stonic 1.4. Il secondo gruppo comprende esclusivamente veicoli appartenenti alla categoria due volumi e tre volumi. Il terzo gruppo comprende sia due volumi e tre volumi che suv e station wagon, dove la seconda categoria prevale di poco sulla prima. Le automobili appartenenti alla categoria due volumi e tre volumi nel terzo gruppo sono: Volkswagen Polo 2.0, Peugeot 308 SW Allure, Mercedes Benz Classe A 200, Volvo V40 T2 e Honda Civic 1.6.

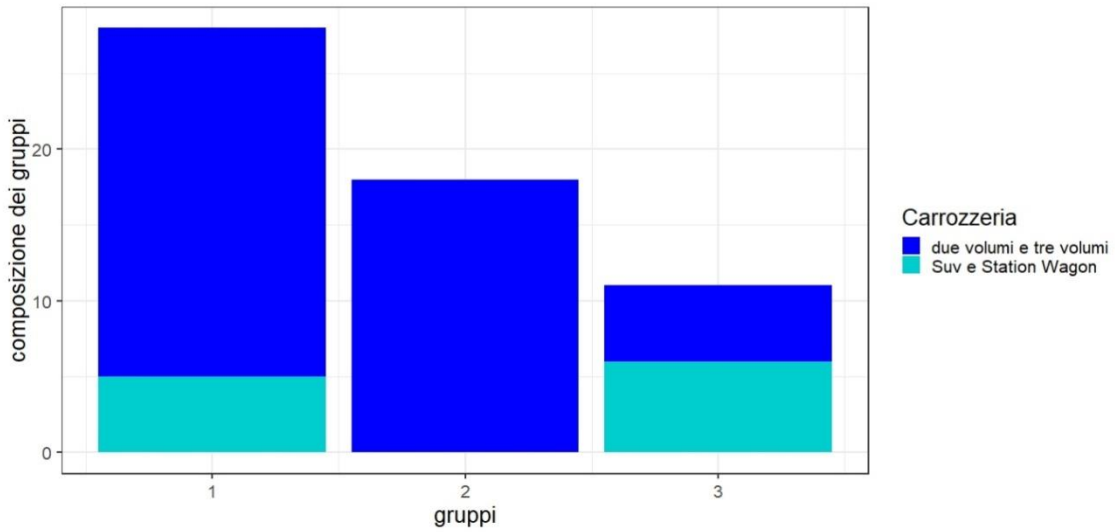

Figura 13: composizione dei gruppi

Per caratterizzare i cluster, poiché l'analisi dei gruppi è stata svolta con i dati standardizzati, si può usare il vettore delle medie per ciascun gruppo. Poiché il centroide dell'intero data set dei dati standardizzati è un vettore nullo, la caratterizzazione dei gruppi avverrà nella seguente maniera: se una variabile nel centroide del singolo gruppo assume un valore positivo allora essa assume valori superiori alla media dell'intero data set, al contrario se una variabile all'interno del centroide del singolo gruppo assume valori negativi allora questa assume valori inferiori alla media. Infine, se una variabile nel centroide del singolo gruppo assume valori prossimi a zero allora significa che la variabile in questione assume valori attorno alla media.

3.3.2.1 Caratterizzazione del primo gruppo

Focalizzando l'analisi sulla partizione ottenuta con la distanza Manhattan il primo gruppo comprende i seguenti veicoli: Peugeot 208 100 Allure, Audi A1 1.4 TFSI 125, Volkswagen Polo 1.6, Kia Rio 1.2, BMW serie 1 116d 3p, Citroen C4 Cactus, Ford Fiesta Active 1.0, Dacia Sandero 1.5, Hyundai i30 1.6, Fiat Tipo tipo 1.4, Volkswagen Golf 1.0, Volkswagen Golf 1.4, Mini 1.5 Cooper, Hyundai i20 1.2, SEAT Ibiza 1.6, Dacia Logan MCV 1.5, Fiat Tipo tipo 1.3 Mjt, Alfa

Romeo Mito 1.4, Peugeot 308 Allure, Alfa Romeo Mito 1.3, Ford Focus 1.0 EcoBoost, Peugeot 208 82 5, Kia Stonic 1.4, Dr dr3 1.5, Renault Clio Tce 12V 75, Mini 1.5 One 75, Mistubishi ASX 1.6, Ssangyong Tivoli 1.6.
Il centroide del primo gruppo è riportato nella Tabella 10.

Tabella 10: centroide del primo gruppo

Prezzo	-0,01630378
Lunghezza	0,3615988
Altezza	-0,29087804
Larghezza	0,37888312
Passo	0,40460470
Volume del bagagliaio	0,21715368
Capacità del serbatoio	0,45524891
Massa in ordine di marcia	0,26974115
Consumi urbani	0,10485498
Emissioni di CO_2	0,06467094
Accelerazione	-0,04938370
Velocità massima	0,08102810
Cilindrata	0,13866856
Potenza	-0,05535583
Coppia massima	0,052083202

Il primo gruppo comprende veicoli caratterizzati da un prezzo attorno alla media dell'intero data set, le caratteristiche relative alla carrozzeria (volume del bagagliaio, capacità del serbatoio, massa in ordine di marcia, lunghezza, larghezza e passo) assumono valori vicino alla media, ad eccezione dell'altezza che invece è inferiore alla media. Le variabili relative al motore ossia la cilindrata, la coppia massima e la potenza assumono valori prossimi alla media. Infine, le prestazioni del veicolo ossia la velocità, l'accelerazione, le emissioni di CO_2 e i consumi urbani presentano valori ancora una volta vicini alla media.
In conclusione, le autovetture incluse nel primo gruppo presentano valori praticamente su tutte le variabili impiegate nell'analisi attorno alla media.

3.3.2.2 Caratterizzazione del secondo gruppo

Il secondo gruppo è composto dalle seguenti autovetture: Kia Picanto 1.0, Volkswagen up!, Toyota Yaris 1.5, Peugeot 108 72, Dr Zero 1.0, Lancia Ypsilon 0.9, Lancia Ypsilon 1.3, Citroen C1 Vti 72, Fiat 500 1.3 Multijet 95, Renault Twingo Sce, SEAT Mii 1.0, Ford Ka+ 1.2, Mitsubishi Space 1.0 Funky, Fiat Panda 1.2 Easy, Skoda Citigo MPI 75, Smart fortwo 70 1.0, Toyota Aygo 1.0, Suzuki Celerio 1.0.
Il vettore delle medie del secondo gruppo è riportato nella Tabella 11.

Tabella 11: centroide del secondo gruppo

Prezzo	-0,8092808
Lunghezza	-1,1639434
Altezza	-0,1055065
Larghezza	-1,2396091
Passo	-1,1236211
Volume del bagagliaio	-0,9678643
Capacità del serbatoio	-1,1163090
Massa in ordine di marcia	-1,1246445
Consumi urbani	-0,3014396
Emissioni di CO_2	-0,3818186
Accelerazione	0,7790581
Velocità massima	-0,8751906
Cilindrata	-0,9080681
Potenza	-0,7895065
Coppia massima	-0,8718081

Il secondo gruppo comprende i veicoli più economici in quanto le unità statistiche incluse in questo gruppo hanno prezzi inferiori alla media. Inoltre, sono compresi i veicoli di dimensioni inferiori dotati di un passo, una larghezza e una lunghezza inferiori alla media, mentre l'altezza di questi veicoli è attorno alla media. La massa in ordine di marcia di questi veicoli è più bassa della media,

possiedono un bagagliaio di volume inferiore rispetto alla media e la capacità del serbatoio è anch'essa sotto la media. Le caratteristiche legate al motore ossia la cilindrata, la potenza e la coppia massima sono inferiori alla media, questo segnala motori di dimensioni inferiori e meno performanti rispetto ai motori dei veicoli appartenenti agli altri due gruppi. L'accelerazione di questi veicoli è superiore alla media, mentre la velocità massima è inferiore alla media. Questo significa che i veicoli compresi in questo gruppo impiegano più secondi per raggiungere la velocità massima ed una volta raggiunta essa è inferiore rispetto a quella dei veicoli appartenenti agli altri gruppi. I veicoli compresi in questo gruppo sono quelli meno inquinanti: il valore delle emissioni di CO_2 è inferiore rispetto alla media, così come i consumi urbani. Riassumendo, il secondo gruppo si compone di veicoli economici, di minori dimensioni, caratterizzati da motori deboli e con minori emissioni di sostanze inquinanti.

3.3.2.3 Caratterizzazione del terzo gruppo

Il terzo gruppo è costituito dai seguenti modelli: Honda Civic 1.6, Ford Focus Station Wagon 1.5, Fiat Tipo Station Wagon 1.6, Volkswagen Polo 2.0, Volvo V40 T2, Peugeot 308 SW Allure, Mercedes Benz Classe A 200, Fiat 500X 2.0 Multijet 140, Hyundai Kona 1.6, Volkswagen T-Roc 1.5, Opel Grandland X.
Nella Tabella 12 è riportato il centroide del terzo gruppo.

Tabella 12: centroide del terzo gruppo

Prezzo	1,3657782
Lunghezza	0,9725913
Altezza	0,5677699
Larghezza	1,0640214
Passo	0,8087499
Volume del bagagliaio	1,0310231
Capacità del serbatoio	0,6678720
Massa in ordine di marcia	1,1537135
Consumi urbani	0,2263613
Emissioni di CO_2	0,4601772
Accelerazione	-1,1491183
Velocità massima	1,2258767
Cilindrata	1,1329552
Potenza	1,4328255
Coppia massima	1,2940202

Il terzo gruppo comprende i veicoli più costosi: il prezzo è superiore rispetto alla media. I veicoli appartenenti a questo gruppo sono dotati di un bagagliaio capiente con un volume superiore alla media, così come il serbatoio: esso ha una capienza superiore alla media. I veicoli di questo gruppo sono quelli di dimensioni superiori, essi dispongono di una larghezza, altezza, passo e lunghezza superiori alla media, inoltre sono anche più pesanti: la loro massa in ordine di marcia è superiore alla media. In merito al motore, essi dispongono di un motore dotato di una cilindrata superiore alla media, la potenza da esso generata è maggiore della media così come la coppia massima. L'accelerazione è inferiore alla media, ma la velocità massima è superiore alla media, questo significa che tali veicoli impiegano meno tempo a raggiungere la velocità massima, ed una volta raggiunta questi veicoli mantengono una velocità superiore rispetto ai veicoli compresi negli altri gruppi. Inoltre, i veicoli in questo

gruppo sono quelli più inquinanti poiché sia le emissioni che i consumi urbani sono superiori alla media.
In sintesi, i veicoli di questo gruppo sono quelli più costosi, con un serbatoio e bagagliaio voluminosi, di dimensioni superiori ma anche più inquinanti e più pesanti.

3.3.2.4 Sintesi della caratterizzazione dei gruppi

Nella Tabella 13 ad ogni gruppo viene assegnata una etichetta e viene riportata una sintesi delle caratteristiche di ciascuno di essi.

Tabella 13: Sintesi delle caratteristiche dei gruppi

Gruppo 1	**Berline non sportive e crossover SUV:** i veicoli appartenenti a questo gruppo sono caratterizzati da un prezzo non eccessivamente elevato, le dimensioni della carrozzeria non sono né ingombranti né piccole ed il loro serbatoio è abbastanza capiente così come il loro bagagliaio.
Gruppo 2	**Utilitarie economiche:** in questo gruppo sono comprese le utilitarie dotate di un prezzo modico, di dimensioni contenute, con un serbatoio poco capiente e un bagagliaio poco spazioso, dotate di un motore di bassa cilindrata. Inoltre, i veicoli in questo gruppo hanno consumi contenuti e basse emissioni di diossido di carbonio.
Gruppo 3	**Suv, station wagon e berline sportive:** i veicoli di questo gruppo sono costosi, di dimensioni e peso elevato, il loro serbatoio e bagagliaio è capiente, il motore è dotato di un'alta cilindrata, ma le emissioni ed i consumi urbani sono elevati.

Si può concludere che il primo gruppo comprende veicoli che, grazie al loro prezzo non troppo elevato, alle dimensioni della carrozzeria contenute ma non troppo piccole, al loro bagagliaio e serbatoio sufficientemente capienti e grazie ai consumi contenuti sono le auto che meglio si adattano alle esigenze di una famiglia mediamente numerosa con un budget non troppo elevato.
Nel secondo gruppo rientrano veicoli che, grazie al loro prezzo e alle dimensioni contenute sono quelli che meglio si adattano ai consumatori che cercano un veicolo che si possa guidare con facilità nel traffico cittadino senza dover

disporre di un budget elevato, tuttavia a causa della loro carrozzeria di piccole dimensioni e di un bagagliaio poco capiente possono essere inadatte a soddisfare le esigenze di una famiglia. Al tempo stesso però, le automobili comprese in questo gruppo, potrebbero essere la seconda automobile di un nucleo famigliare grazie al loro prezzo modesto e alle loro dimensioni contenute.

Infine, nel terzo gruppo rientrano i veicoli di maggiori dimensioni dotati di un ampio bagagliaio, questo gli rende idonei a soddisfare i bisogni dei nuclei famigliari più numerosi, ma al tempo stesso gli rende più scomodi da guidare negli ambienti cittadini. I veicoli di questo gruppo grazie alle loro elevate prestazioni in termini di velocità e potenza del motore sono idonei a soddisfare le esigenze di chi cerca anche una guida più sportiva.

Conclusioni

L'elaborato è stato realizzato con lo scopo di applicare ad un caso reale la tecnica di statistica multivariata dell'analisi dei gruppi. Il data set oggetto dell'analisi è stato costruito reperendo i dati relativi agli autoveicoli dal sito internet: "*https://www.automoto.it/listino*". Per l'applicazione dell'analisi dei gruppi è stato utilizzato il software statistico R. L'analisi dei gruppi è stata svolta sia con il metodo non gerarchico delle k-medie che con i metodi gerarchici aggregativi. In particolare, i metodi gerarchici sono stati svolti con il metodo del legame singolo, il metodo del legame completo, il metodo del legame medio ed il metodo del centroide. Ciascuno di questi metodi è stato realizzato sia calcolando la distanza tra le unità statistiche con la distanza Manhattan che con la distanza Euclidea. Fra tutte le partizioni ottenute con i vari metodi gerarchici si è scelto di mantenere la partizione con tre gruppi ottenuta con il metodo del legame completo a partire da una matrice di distanze Manhattan, in quanto la partizione così ottenuta è quella che meglio riesce a cogliere i gruppi presenti all'interno del data set. Questa scelta è motivata dalla forte somiglianza della partizione ottenuta con il metodo del legame completo partendo da una matrice di distanze di Manhattan, con l'analoga partizione ottenuta con il metodo delle k-medie. Questa situazione ha dato forza alla convinzione che i gruppi identificati fossero davvero quelli esistenti nei dati osservati. Della partizione scelta è stata analizzata la numerosità dei gruppi, le caratteristiche dei cluster, la composizione dei gruppi in base alla carrozzeria del veicolo nonché è stata misurata la loro coesione interna.
L'analisi dei gruppi ha portato alla formazione di tre gruppi con le seguenti caratteristiche:

- Gruppo 1: Veicoli con un prezzo ed emissioni contenute, le cui dimensioni nella media gli rendono adatti sia alla guida urbana che a soddisfare le esigenze anche delle famiglie più numerose.

- Gruppo 2: Veicoli economici con basse emissioni e di piccole dimensioni con motori di piccola cilindrata e poco potenti.
- Gruppo 3: Veicoli costosi e maggiormente inquinanti, con grandi dimensioni della carrozzeria e dotati di un motore potente e di alta cilindrata.

Appendice

Di seguito vengono riportati i codici in R utilizzati per lo svolgimento dell'analisi.

In primo luogo, vengono caricate le librerie contenenti le funzioni necessarie allo svolgimento dell'analisi, oltre a quelle già contenute nella versione base di R.

```
R> library(NbClust)
   Library(Factoextra)
   Library(GGplot2)
   Library(dplyr)
```

La prima libreria consente di calcolare l'indice Pseudo F, la seconda contiene le funzioni necessarie per la rappresentazione grafica dei risultati dell'analisi dei gruppi, la terza consente di migliorare le rappresentazioni dei grafici realizzabili con il linguaggio R, infine la quarta contiene le funzioni necessarie al calcolo dell'indice Root Mean Square Standard Deviation.

```
R> Cars<-read.csv2("DataCars.csv")
```

Viene caricato il data set in formato .csv all'interno di un dataframe denominato "Cars".

```
R> row.names(Cars)<-Cars$Modello
   Cars[1]<-NULL
```

Vengono modificati i nomi delle righe, sostituendo la numerazione progressiva con il modello del veicolo, per poi eliminare la prima colonna del dataframe che accoglie i nomi dei modelli degli autoveicoli.

```
R> Zcars<-scale(Cars)
   Zcars<-as.data.frame(Zcars)
```

Viene caricata all'interno di una matrice vuota la matrice dei dati standardizzati per poi convertire la matrice stessa in un dataframe nominato "Zcars".

```
R> Dist_M<-dist(Zcars, method="manhattan")
```

Viene costruita la matrice delle distanze mediante la distanza Manhattan. Per realizzare la matrice delle distanze tramite la distanza Euclidea è necessario cambiare il campo `method` inserendo `euclidean` al posto di `manhattan`.

```
R> Zcars_complete_M<-hclust(Dist_M, method="complete")
```

Si realizza l'analisi dei gruppi mediante i metodi gerarchici aggregativi. In primo luogo, è necessario specificare la matrice di distanze da cui partire, per poi indicare nel campo `method` il criterio con cui procedere all'aggregazione delle unità statistiche. Inserendo nel campo `method` le diciture `complete`, `centroid`, `average` e `single` è stato possibile procedere all'aggregazione delle unità statistiche secondo il metodo del legame completo, del centroide, medio e singolo.

```
R> ResNbClust <-NbClust(data=Zcars, distance="manhattan", min.nc=2,
                        max.nc=5, method="complete", index="ch")
```

Viene calcolato l'indice Pseudo F specificando nel campo `index` il valore `ch`, per valutare la partizione ottimale all'interno di una famiglia di partizioni. Per selezionare una certa famiglia di partizioni bisogna specificare nel campo `distance` quale distanza usare per costruire la matrice delle distanze e specificare nel campo `method` il criterio di aggregazione. Inoltre, è necessario specificare il numero minimo (`min.nc`) ed il numero massimo (`max.nc`) di cluster.

```
R> fviz_dend(Zcars_Complete_M, k=3,
             palette=c("green3", "darkred", "deepskyblue4"),
             rect=TRUE, cex=0.5, rect_border="black", rect_lty=1 )
```

Per visualizzare il dendrogramma è necessario inserire il risultato della procedura gerarchica aggregativa nella funzione `fviz_dend`, specificando il numero di gruppi da visualizzare nel dendrogramma (`k=3`). Si è deciso di evidenziare i gruppi con dei colori diversi tra loro (`palette`) e con una linea continua (`rect_lty=1`).

```
R> Cluster_Complet_M <-cutree(Zcars_complete_M, k=3)
   Cluster_Complet_M <-as.data.frame(Cluster_Complet_M)
```

A partire dal risultato dell'aggregazione secondo un metodo gerarchico aggregativo desiderato si seleziona la partizione ottimale per visualizzare a quale gruppo appartiene ogni unità statistica. Per agevolare questa visualizzazione si è deciso di inserire il risultato della funzione `cutree` in un dataframe.

Questa operazione può essere replicata per ogni partizione richiamandole in maniera adeguata.

```
R > by(Zcars, Cluster_Complet_M, colMeans)
```

Viene calcolato il vettore delle medie per ciascun gruppo della partizione selezionata. Questa operazione può essere replicata per ogni partizione richiamandole in maniera adeguata.

```
R> Zcars[16]<-Cluster_Complet_M[1]
   colnames(Zcars)[16]<-"Gruppo"
   VarPrezzo<-Zcars %>%
   group_by(Gruppo) %>%
   summarise(
   n=n(),
   sd_prezzo=var(Prezzo))
   VarPrezzo<-diag(VarPrezzo$sd_prezzo)
```

Al dataframe “Zcars” che accoglie la matrice dei dati standardizzati viene aggiunto un nuovo attributo chiamato “Gruppo” che può assumere tre valori (1, 2, 3) a seconda del gruppo di appartenenza dell’unità statistica.

Successivamente per ogni gruppo, a partire dalla matrice dei dati standardizzati, viene calcolata la varianza del carattere ”Prezzo”. Il risultato di questa operazione viene inserita in una matrice diagonale chiamata “VarPrezzo”.

Ripetendo opportunamente questa operazione anche per le altre variabili e sommando le matrici diagonali che vengono generate si ottiene una matrice sulla cui diagonale principale giacciono le varianze totali di tutte le *p* variabili per ciascuno dei tre gruppi.

In maniera analoga si ripete questa procedura per la partizione ottenuta con il metodo del legame completo tramite la distanza Euclidea.

Lo scopo di questa operazione è il calcolo della devianza totale nei gruppi, la quale è necessaria per calcolare l’indice Root Mean Square Standard Deviation per ogni gruppo delle due partizioni generate con il metodo del legame completo.

```
R> set.seed(123)
   km.res<-kmeans(Zcars, k=3, nstart=25)
   fviz_cluster(km.res, data=Zcars, repel=TRUE, ellipse=TRUE,
                ellipse.type=”convex”, ggtheme=theme_minimal(),
                palette=c("royalblue1", “orangered”, “limegreen”))
```

Il metodo delle k-medie richiede che venga stabilito a priori il numero di gruppi con cui realizzare l’aggregazione ed individuare i poli iniziali. La prima informazione deve essere inserita dall’utente, i secondi invece vengono generati casualmente mediante il comando `set.seed(123).`

I risultati dell'aggregazione secondo il metodo delle k-medie vengono rappresentati in un piano cartesiano dove ogni gruppo viene identificato con un colore specifico (`palette`).

```
R> ks <- 2:15
   ssw<-numeric(length(ks))
   for(i in seq_along(ks))
   {
    set.seed(123)
    ssw[i] <-kmeans(Zcars, ks[i], nstart=25)$tot.withinss
   }
   plot(x=ks, y=ssw, type="l", xlab="numero_di_cluster",
        ylab="SS_whitin", "col="deepskyblue4")
```

Viene definito il numero di cluster con cui provare l'aggregazione per poi inserire in un vettore la devianza totale nei gruppi di ogni partizione generata tramite il metodo delle k-medie durante l'esecuzione del ciclo.

In un grafico a due dimensioni viene rappresentato l'andamento della devianza totale nei gruppi al variare del numero di gruppi.

```
R > ks<- 2:15
    pseudof<-numeric(length(ks))
    for(i in seq_along(ks)
    {
     set.seed(123)
     pseudof[i]<- NbClust(Zcars, distance="euclidean", min.nc=ks[i],
                 max.nc=ks[i], method="kmeans", index="ch")$All.index
   }
   plot(x=ks, y="Pseudo_F", type="l", xlab="numero_di_cluster",
        ylab="Pseudo F", "col="deepskyblue4")
```

In maniera simile all'interno di un vettore viene inserito il valore dell'indice Pseudo F per ogni partizione generata con il metodo delle k medie durante l'esecuzione del ciclo, per poi rappresentare in un grafico a due dimensioni l'andamento dell'indice Pseudo F al variare del numero di gruppi.

Bibliografia

Sergio Zani, Analisi dei dati statistici II, Giuffrè ed., Milano, 2000

Brian Everitt, Torsten Hothorn, An Introduction to Applied Multivariate Analysis with R, Springer ed., New York, 2011

Zani Sergio, Cerioli Andrea, Analisi dei dati e data mining per le decisioni aziendali, Giuffrè ed, Milano, 2007

Sitografia

https://cran.r-project.org/doc/contrib/Frascati-FormularioStatisticaR.pdf

https://cran.r-project.org/web/packages/factoextra/factoextra.pdf

https://cran.r-project.org/web/packages/NbClust/NbClust.pdf

https://cran.r-project.org/package=ggplot2/ggplot2.pdf

https://cran.r-project.org/web/packages/dplyr/dplyr.pdf

https://it.wikipedia.org/wiki/Passo_(veicoli)

https://www.quartamarcia.it/che-cose-il-peso-in-ordine-di-marcia/

https://www.businessonline.it/articoli/cose-la-cilindrata-di-unauto-come-si-misura.html

https://www.alvolante.it/da_sapere/motori-potenza-e-coppia-341409

https://www.automoto.it/listino

https://cda.psych.uiuc.edu/multivariate_fall_2012/systat_cluster_manual.pdf

http://www.riani.it/ADM/lucidi/Estratto_libro_regr.pdf